



Ana Sofia Lopes Afonso

Licenciada

Predição da Procura de Bicicletas para Otimizar o Rebalanceamento das Bicicletas Partilhadas em Lisboa

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática

Orientador: João Carlos Gomes Moura Pires, Professor Auxiliar, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa
Co-orientadores: Carlos Augusto Isaac Piló Viegas Damásio, Professor Associado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa
Nuno Miguel Soares Datia, Professor Adjunto, Instituto Superior de Engenharia de Lisboa

Júri

Presidente: Doutor Pedro Abílio Duarte de Medeiros
Arguente: Doutora Ana Paula Pereira Afonso
Vogal: Doutor João Carlos Gomes Moura Pires



FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2021

Predição da Procura de Bicicletas para Otimizar o Rebalanceamento das Bicicletas Partilhadas em Lisboa

Copyright © Ana Sofia Lopes Afonso, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade NOVA de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

AGRADECIMENTOS

Ao longo de todo o processo de desenvolvimento desta dissertação recebi apoio não só a nível académico, como também a nível pessoal, sendo que gostaria de mencionar algumas pessoas sem as quais a sua realização não teria sido possível.

Em primeiro lugar, gostaria de agradecer à Câmara Municipal de Lisboa e à Empresa de Mobilidade e Estacionamento de Lisboa (EMEL) pelo fornecimento dos dados que estão na base da dissertação, bem como pelo esclarecimento de questões durante o período de desenvolvimento.

Gostaria de agradecer ao meu orientador, o Professor João Moura Pires, por todo o apoio e tempo disponibilizado, bem como pelo seu conhecimento e por me ter guiado ao longo de todo o processo, esclarecendo as dúvidas e obstáculos que foram surgindo.

Gostaria ainda de agradecer ao Professor Fernando Birra e ao Professor Carlos Damásio pelo apoio disponibilizado durante o projeto e ainda ao Professor Nuno Datia que também acompanhou o trabalho.

Aos meus colegas de mestrado, principalmente à Mafalda Camilo, João Lucas Dias, Sofia Nascimento e também à Joana Barradas e Inês Messias, com quem partilhei esta experiência e que também me ajudaram sempre que necessário.

Finalmente, gostaria de agradecer aos meus pais, Paula e Rui, que sempre me apoiaram em todo o meu percurso académico e ao longo da minha vida.

RESUMO

Com o desenvolvimento urbano nas cidades, os sistemas de bicicletas partilhadas têm vindo a ser cada vez mais utilizados como forma de evitar o trânsito causado pelos automóveis, sendo uma opção que promove a mobilidade sustentável, contribuindo para redução do tráfego automóvel e diminuição da poluição nas grandes cidades.

A utilização destes sistemas apresenta problemas de desequilíbrio na disponibilidade de bicicletas e docas para alugar e devolver bicicletas, respetivamente, o que torna o serviço insatisfatório. O rebalanceamento das estações na rede tem sido a solução utilizada para tentar minimizar este problema, através da redistribuição de bicicletas pela rede para evitar estações vazias ou cheias, que impossibilitam o aluguer e entrega das mesmas, respetivamente. No entanto, este processo pode ser melhorado, para evitar viagens desnecessárias por parte dos funcionários e, por outro lado, para diminuir o tempo em que uma estação se encontra vazia ou cheia.

De forma a otimizar este processo, podem ser utilizadas técnicas de previsão da quantidade de bicicletas nas estações sendo possível perceber quais as que necessitam de rebalanceamento e quando este é necessário. Com base em dados disponibilizados pela Câmara Municipal de Lisboa relativos às bicicletas que se encontram nas estações bem como às viagens que ocorreram na rede de bicicletas partilhadas no ano de 2018, foi efetuado o estudo dos padrões diários de enchimento e esvaziamento para cada estação. Foram utilizados algoritmos de *clustering* para agrupar as estações da Gira de acordo com variáveis temporais e espaciais, de forma a encontrar estações com características e padrões semelhantes de rácio, partidas e chegadas, localização e número de docas. Foi levada a cabo uma análise com o objetivo de detetar o rebalanceamento nos dados, de forma a poder remover essa influência dos dados e ser possível o desenvolvimento de modelos preditivos sem qualquer influência de rebalanceamento. Foram desenvolvidos modelos preditivos específicos para cada estação, tendo sido efetuado um processo de validação aquando do qual foram utilizados diversos parâmetros, tendo sido escolhidos aqueles que minimizam o erro.

Finalmente, foi levada a cabo uma análise dos momentos em que não há viagens na rede, de forma a compreender se tal se deve à inexistência de clientes ou ao facto das estações estarem cheias ou vazias.

Palavras-chave: Sistemas de Bicicletas Partilhadas, Análise de Dados, Séries Temporais, Técnicas de Clustering, Rebalanceamento, Previsão da procura de bicicletas, Visualização Interativa de Dados.

ABSTRACT

With urban development in cities, shared bicycle systems are increasingly used as a way to avoid traffic caused by cars, being an option that promotes sustainable mobility and contributes for traffic and pollution reduction in urban areas.

These systems present imbalance problems when it comes to bicycle and dock availability, which disable users from renting or returning bicycles, respectively, causing the service to be unsatisfactory. The bike rebalancing process in the system has been the current go to solution in order to try to solve the imbalance issue. This consists of redistributing bikes across stations in the network so that each station can have an optimal quantity of bicycles, which prevents stations from becoming empty or full which, in turn, disables the users to rent or deliver bikes, respectively. However, this process can be improved to avoid unnecessary rebalancing trips and minimize the amount of time stations are imbalanced.

In order for this process to be optimized, forecasting techniques can be used to predict the amount of bikes in each station, making it possible to understand which stations require rebalancing, before such an event occurs.

Based on data provided by the Lisbon City Council regarding the number of bicycles found at the stations, as well as trips that occurred on the shared bicycle network in 2018, the study of daily filling standards was carried out in order to understand when the stations are full or empty.

In order to characterize Gira stations, *clustering* algorithms were used to group them according to temporal and spatial variables, in order to find stations with similar characteristics and patterns of ratio, departures and arrivals, location and number of docks. An analysis was carried out in order to detect rebalancing in the data, in order to be able to remove this influence from the data and to enable the development of predictive models without any rebalancing influence. Specific predictive models were developed for each station, after a validation process during which several parameters were used, and those whose metrics had a lower value were chosen.

Finally, an analysis was carried out in order to understand if the moments when there are no bicycle rentals on the network is due to a lack of customers or the fact that the stations are either full or empty.

Keywords: Shared Bicycle Systems, Data Analysis, Time Series, Clustering Algorithms, Rebalancing, Bicycle Demand Forecasting, Interactive Data Visualization

ÍNDICE

Lista de Figuras	xv
Lista de Tabelas	xxi
Siglas	xxiii
1 Introdução	1
1.1 Contexto e Motivação	1
1.2 Descrição do Problema	3
1.3 Contribuições	4
1.4 Estrutura do Documento	5
2 Trabalho Relacionado e Estado da Arte	7
2.1 Introdução	7
2.2 Trabalho Relacionado	8
2.2.1 Bicicletas Partilhadas	8
2.2.2 Rebalanceamento	9
2.2.3 <i>Forecasting</i> da procura de bicicletas	13
2.2.4 <i>Clustering</i> em Sistemas de Bicicletas Partilhadas	19
2.2.5 Conclusões	22
2.3 Técnicas de Clustering	23
2.3.1 Bipartite Clustering	23
2.3.2 Agrupamento Hierárquico	23
2.3.3 K-means	24
2.4 Técnicas de Forecasting	25
2.4.1 Métodos de Smoothing	26
2.4.2 Long Short Term Memory (Long Short Term Memory (LSTM))	26
2.4.3 Random Forest (RF)	27
2.5 Avaliação da Precisão do <i>Forecasting</i>	28
3 Abordagem e Metodologia	31
3.1 CRISP-DM	31
3.2 Estudo de Padrões das Viagens e Estações	36

3.3	Elaboração de Modelos Preditivos	37
3.4	Avaliação de Resultados	38
3.5	Tecnologias Utilizadas	38
3.5.1	Tableau Prep	39
3.5.2	Tableau Desktop	39
3.5.3	Diagrams.net	40
3.5.4	Python e outras Bibliotecas	40
4	Análise Exploratória de Dados	43
4.1	Apresentação dos Dados	43
4.1.1	Estações Gira	44
4.1.2	Viagens Gira	47
4.1.3	Dados Meteorológicos	50
4.2	Agrupamento	51
4.2.1	Agrupamento Espacial	52
4.2.2	Agrupamento Temporal	58
4.3	Caracterização Espacial e Temporal das Estações Gira	68
4.3.1	Interactive Visualization Modeling Language	68
4.3.2	Saldo de Viagens	72
4.3.3	Diferença de Bicicletas nas Estações	77
4.4	Deteção do Rebalanceamento	83
5	Modelos Preditivos	89
5.1	Dados	89
5.2	Análise dos Dados	92
5.3	Construção e Otimização dos modelos	96
5.3.1	Naïve Forecasting	97
5.3.2	Média Móvel	99
5.3.3	Long Short Term Memory (LSTM)	100
5.4	Resultados	102
5.5	Análise de Plateaus	106
6	Conclusões e Trabalho Futuro	115
	Bibliografia	117
A	Relatório de Qualidade de Dados	121
A.1	Introdução	121
A.2	Dados das Estações Gira 2018	122
A.2.1	Descrição dos Dados	122
A.2.2	Exploração dos Dados	123
A.2.3	Qualidade dos Dados	134

A.3	Dados das Viagens Gira 2018	136
A.3.1	Descrição dos Dados	136
A.3.2	Exploração dos dados	137
A.3.3	Qualidade dos Dados	145
B	Processamento dos Dados	147
B.1	Estações Gira	147
B.1.1	Remoção de atributos	148
B.1.2	Morada	148
B.1.3	Localização das Estações (WKB Geometry)	152
B.1.4	Data de update	154
B.1.5	Data de início das Estações	156
B.1.6	Número de Docas	162
B.2	Viagens Gira	163
C	Relatório de Análise dos Dados IPMA	167
C.1	Descrição dos Dados	168
C.2	Exploração dos Dados	169
C.2.1	Precipitação	169
C.2.2	Humidade	172
C.2.3	Radiação	174
C.2.4	Temperatura	177
C.2.5	Vento (Intensidade)	179
C.2.6	Vento (Direção)	182
C.3	Qualidade dos Dados	183
C.4	Preparação dos Dados	185

LISTA DE FIGURAS

1.1	Rede de Estações da Gira. Retirado de ¹	2
2.1	Estação com Bicicletas da Gira em Lisboa. Retirado de	8
2.2	Camião para efetuar o Rebalanceamento na rede da CitiBike. Retirado de	10
2.3	Partidas e Chegadas por hora numa estação (a) do grupo 1, (b) do grupo 2 e (c) do grupo 3. (τ - hora do dia; ν - número cumulativo de bicicletas que chegaram ou partiram de uma estação durante um período de 7 meses. Retirado de	12
2.4	Viagens diárias entre zonas a intra-zonas em Wenzhou, China. Retirado de [39].	17
2.5	Resultados do agrupamento utilizando clustering bipartido em dados da CitiBike, Nova Iorque. Retirado de [38].	19
2.6	Resultados do agrupamento utilizando K-means em dados de Wenzhou, na China. Retirado de [39].	20
2.7	Arquitetura de uma célula LSTM. Retirado de [43].	27
3.1	Fases do processo CRISP-DM. Retirado de [18].	32
4.1	Localização das Estações e respetivos identificadores após o processamento.	46
4.2	Evolução temporal do número de viagens ao longo dos meses de 2018.	49
4.3	Registos de viagens nos dias do mês de Julho de 2018.	49
4.4	Estações Meteorológicas em Lisboa em 2018.	50
4.5	Método do Cotovelo e Métrica do <i>Silhouette Score</i> para o Agrupamento por Localização.	53
4.6	Grupos obtidos após o agrupamento espacial com base na localização.	54
4.7	Método do Cotovelo e Métrica do <i>Silhouette Score</i> para os diferentes valores de k do Agrupamento por Localização e Número de Docas.	55
4.8	Grupos obtidos após o agrupamento espacial com base na localização e no número de docas.	56
4.9	<i>Boxplot</i> dos grupos formados tendo em conta a localização e docas.	57
4.10	Vista aproximada das estações 446, 104 e 105, pertencentes ao grupo 5, localizadas no Parque das Nações e em Entrecampos.	58
4.11	Método do Cotovelo e Métrica do <i>Silhouette Score</i> para os diferentes valores de k do Agrupamento por Rácio.	60

4.12 <i>Dashboard</i> com os grupos obtidos a partir do agrupamento tendo em conta o rácio médio ao longo do dia.	60
4.13 Método do Cotovelo e Métrica do <i>Silhouette Score</i> para os diferentes valores de k do Agrupamento tendo em conta o valor médio de partidas ao longo do dia.	63
4.14 <i>Dahsboard</i> com os grupos obtidos a partir do agrupamento tendo em conta o valor médio de partidas ao longo do dia.	64
4.15 Método do Cotovelo e Métrica do <i>Silhouette Score</i> para os diferentes valores de k do Agrupamento tendo em conta o valor médio de chegadas ao longo do dia.	65
4.16 <i>Dahsboard</i> com os grupos obtidos a partir do agrupamento tendo em conta o valor médio de chegadas ao longo do dia.	66
4.17 Localização de algumas estações com o valor médio de partidas mais elevado próximas de Estações de Metro e Centro Vasco da Gama.	67
4.18 Componentes da Linguagem IVML que podem ser utilizados para a construção dos diagramas.	70
4.19 Variáveis visuais da linguagem IVML.	70
4.20 Tipos de Gráfico da linguagem IVML.	71
4.21 Tipos de interação da linguagem IVML.	71
4.22 Tipos de ação da linguagem IVML.	71
4.23 Diagrama de Interações para o <i>dashboard</i> de caracterização espacial e temporal do Saldo Médio de Viagens das Estações Gira.	72
4.24 Diagrama de <i>Tooltips</i> para o <i>dashboard</i> de caracterização espacial e temporal do Saldo Médio de Viagens das Estações Gira.	73
4.25 <i>Dashboard</i> de caracterização espacial e temporal do Saldo Médio de Viagens para os dias úteis.	74
4.26 Localização das estações cujas <i>tooltips</i> se encontram na Figura 4.27	75
4.27 <i>Tooltips</i> do <i>dashboard</i> relativo ao Saldo de Viagens, acessíveis através da ação de <i>hover</i> das estações no mapa.	76
4.28 <i>Dashboard</i> de caracterização espacial e temporal do Saldo Médio de Viagens para os fins-de-semana.	77
4.29 Diagrama de Interações para o <i>dashboard</i> de caracterização espacial e temporal do Saldo Médio de Bicicletas das Estações Gira.	78
4.30 Diagrama de <i>Tooltips</i> para o <i>dashboard</i> de caracterização espacial e temporal do Saldo Médio de Bicicletas das Estações Gira.	78
4.31 <i>Dashboard</i> de caracterização espacial e temporal do Saldo Médio da Diferença de Bicicletas para os dias úteis.	79
4.32 <i>Tooltips</i> do <i>dashboard</i> relativo ao Saldo de Bicicletas, acessíveis através da ação de <i>hover</i> das estações no mapa.	81
4.33 Localização das estações cujas <i>tooltips</i> se encontram na Figura 4.32.	81
4.34 <i>Dashboard</i> de caracterização espacial e temporal do Saldo Médio da Diferença de Bicicletas para os dias de fim-de-semana.	82

4.35 Dashboard para Análise e Detecção do Rebalanceamento nas Estações da Rede Gira, com ênfase na estação 446.	85
4.36 Análise e Detecção do Rebalanceamento na Estação 446 entre 10 e 12 de Setembro.	87
4.37 Análise e Detecção do Rebalanceamento na Estação 446 entre 14 e 17 de Setembro.	88
5.1 Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana para a Estação 101.	92
5.2 Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana, Fins-de-semana e feriados para a Estação 101.	93
5.3 Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana para a Estação 218.	94
5.4 Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana, Fins-de-semana e feriados para a Estação 218.	95
5.5 Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana para a Estação 101.	95
5.6 Média da Humidade (%) de janeiro a outubro de 2018.	96
5.7 Média da Precipitação (mm) de janeiro a outubro de 2018.	96
5.8 Média da Intensidade do Vento (m/s) de janeiro a outubro de 2018.	96
5.9 Média da Radiação (W/m^2) de janeiro a outubro de 2018.	96
5.10 Média da Temperatura (C°) de janeiro a outubro de 2018.	96
5.11 Legenda dos meses.	96
5.12 Dashboard para analisar os resultados do <i>naïve forecasting</i> por Estação, com ênfase na estação 101.	98
5.13 Dashboard para analisar os resultados do <i>naïve forecasting</i> por Estação, com ênfase na estação 446.	99
5.14 Dashboard para analisar os resultados do Moving Average por Estação, com ênfase na estação 101.	100
5.15 Dashboard para analisar os resultados do LSTM univariável single-step para a estação 101.	102
5.16 Dashboard para analisar os resultados do LSTM univariável multi-step para a Estação 101.	103
5.17 Dashboard para analisar os resultados do LSTM multivariável single-step para a Estação 101.	103
5.18 Dashboard para analisar os resultados do LSTM multivariável multi-step para a Estação 101.	104
5.19 Visualização das melhores parametrizações para cada valor de target.	104
5.20 <i>Dashboard</i> com a previsão efetuada para a Estação 101.	106
5.21 <i>Dashboard</i> de Análise dos <i>plateaus</i> nas Estações Gira.	108

5.22	<i>Dashboard</i> de Análise dos <i>plateaus</i> nas Estações Gira com ordenação das estações pelas que estiveram mais cheias.	109
5.23	<i>Dashboard</i> de Análise dos <i>plateaus</i> nas Estações Gira com realce das estações com identificador 1XX.	111
5.24	<i>Dashboard</i> de Análise dos <i>plateaus</i> nas Estações Gira com ênfase na estação 309 e respetiva <i>tooltip</i>	111
5.25	<i>Dashboard</i> de Análise dos <i>plateaus</i> nas Estações Gira com ênfase nas estações próximas da estação 309.	112
5.26	<i>Dashboard</i> de Análise dos <i>plateaus</i> nas Estações Gira com ênfase na estação 471 e respetiva <i>tooltip</i>	113
A.1	Distribuição do atributo Id planeamento. As etiquetas representam a contagem de valores distintos em cada um dos intervalos.	124
A.2	Localização das Estações e respetivos identificadores.	124
A.3	Distribuição do atributo Id planeamento. As etiquetas representam a contagem de valores distintos em cada um dos intervalos.	125
A.4	Relação entre a morada e o identificador das estações.	125
A.5	Estações que apresentam mais de um valor distinto de Well-known binary (WKB).	126
A.6	Gráfico de barras para o tipo de serviço níveis.	127
A.7	Gráfico de barras para a distribuição do número de bicicletas.	128
A.8	Gráfico de barras para a distribuição do número de docas.	129
A.9	Estações com mais de um valor para o número de docas e respetivo número.	129
A.10	Número de docas em função do tempo.	130
A.11	Distribuição para o rácio do número de bicicletas e número de docas.	130
A.12	Distribuição do Estado.	131
A.13	Distribuição do Estado.	131
A.14	Número de Registos da Data e Hora de <i>Update</i> por mês.	133
A.15	Número de Registos da Data e Hora de <i>Update</i> por dia.	133
A.16	Evolução temporal da Data de início ao longo dos meses de 2018.	138
A.17	Viagens no mês de julho.	138
A.18	Viagens no mês de setembro.	139
A.19	Evolução temporal da Data de fim ao longo dos meses de 2018.	139
A.20	Valores nulos e não nulos para a distância.	140
A.21	Distribuição dos valores não nulos da distância.	141
A.22	Distribuição dos valores para a estação de início.	141
A.23	Distribuição dos valores para a estação de fim.	142
A.24	Relação entre o Radio-frequency Identification (RFID) e o tipo de bicicleta.	142
A.25	Exemplo de um MULTILINESTRING representativo do percurso efetuado numa viagem.	143
A.26	Valores nulos e não nulos para o atributo <i>WKB Geometry</i>	143

A.27 Valores nulos e não nulos para o atributo Número de Vértices.	144
A.28 Distribuição dos Valores Não Nulos para o Número de Vértices.	144
A.29 Distribuição do tipo de bicicleta.	145
B.1 Estações que apresentam mais de um identificador ou morada.	149
B.2 Estação cuja morada é “103 - Jardim de Água” no website da Gira.	150
B.3 Estações cuja designação é igual à exceção do identificador na mesma.	151
B.4 Estações no Marquês de Pombal e Saldanha.	151
B.5 Registos diários da Estação 999 ao longo de 2018.	152
B.6 Localização aproximada da Estação 213.	154
B.7 Localização aproximada da Estação 405.	154
B.8 Localização aproximada da Estação 441.	154
B.9 Localização aproximada da Estação 448.	154
B.10 Duplicados para a Estação 101 no dia 23 de fevereiro.	156
B.11 Visão geral da data de início dos registos para cada Estação presente no conjunto de dados.	157
B.12 Registos para cada estação ao longo do tempo, cujo estado era <i>repair</i> em destaque.	158
B.13 Visão geral da data de início dos registos para cada Estação presente no conjunto de dados.	159
B.14 Estações do Grupo 1 cujos registos tiveram início em fevereiro.	159
B.15 Estações do Grupo 2 cujos registos tiveram início em março.	160
B.16 Estações do Grupo 3 cujos registos tiveram início em maio.	161
B.17 Estações do Grupo 4 cujos registos tiveram início em agosto.	162
B.18 Número de estações cujo número de docas foi zero em cada mês.	163
B.19 Análise temporal para cada estação dos registos cujo número de docas é zero.	163
B.20 Percurso efetuado na viagem com início na estação 110.	164
B.21 Localização da estação de fim da viagem com início na estação 110.	164
B.22 Percurso efetuado na viagem com início na estação 474.	165
B.23 Localização da estação de fim da viagem com início na estação 474.	165
B.24 Distribuição da duração das viagens em formato HH:MM:SS.	165
C.1 Estações Meteorológicas em Lisboa em 2018.	169
C.2 Distribuição temporal dos valores inválidos (-990) medidos para a precipitação na Estação 762 em 2018.	170
C.3 Distribuição da Precipitação para a Estação 535 em 2018.	171
C.4 Distribuição da Precipitação para a Estação 579 em 2018.	171
C.5 Distribuição da Precipitação para a Estação 762 em 2018.	171
C.6 Distribuição Temporal da Precipitação Total para as Estações Meteorológicas de Lisboa.	172

C.7 Distribuição temporal dos valores inválidos (-990) medidos para a humidade na Estação 762 em 2018.	173
C.8 Distribuição da Humidade para a Estação 535 em 2018.	173
C.9 Distribuição da Humidade para a Estação 579 em 2018.	173
C.10 Distribuição da Humidade para a Estação 762 em 2018.	174
C.11 Distribuição Temporal da Humidade para as Estações Meteorológicas de Lisboa.	174
C.12 Distribuição temporal dos valores inválidos (-990) medidos para a radiação na Estação 762 em 2018.	175
C.13 Distribuição da Radiação para a Estação 579 em 2018.	176
C.14 Distribuição da Radiação para a Estação 762 em 2018.	176
C.15 Distribuição Temporal da Radiação para as estações meteorológicas de Lisboa em 2018.	176
C.16 Distribuição temporal dos valores inválidos (-990) medidos para a temperatura na Estação 762 em 2018.	177
C.17 Distribuição da Temperatura para a Estação 535 em 2018.	178
C.18 Distribuição da Temperatura para a Estação 579 em 2018.	178
C.19 Distribuição da Temperatura para a Estação 762 em 2018.	178
C.20 Distribuição Temporal da Temperatura para todas as estações em 2018.	179
C.21 Contagem dos valores inválidos e nulos para cada mês e dia na estação 535 ao longo do ano de 2018. As células que apresentam (*) representam valores nulos, as restantes são valores inválidos.	180
C.22 Contagem dos valores inválidos e nulos para cada mês e dia na estação 762 ao longo do ano de 2018.	180
C.23 Distribuição da Intensidade do Vento para a Estação 535 em 2018.	181
C.24 Distribuição da Intensidade do Vento para a Estação 579 em 2018.	181
C.25 Distribuição da Intensidade do Vento para a Estação 762 em 2018.	181
C.26 Distribuição Temporal da Intensidade do Vento para todas as estações no ano de 2018.	182
C.27 Contagem dos valores inválidos para cada mês e dia na estação 535 ao longo do ano de 2018.	183
C.28 Distribuição da Direção do Vento para a Estação 535 em 2018.	183
C.29 Distribuição da Direção do Vento para a Estação 579 em 2018.	183
C.30 Distribuição da Direção do Vento de acordo com os pontos cardeais e colaterais para a Estação 535 em 2018.	184
C.31 Distribuição da Direção do Vento de acordo com os pontos cardeais e colaterais para a Estação 579 em 2018.	184
C.32 Distribuição Temporal da Direção do Vento para as Estações Meteorológicas de Lisboa no ano de 2018.	184

LISTA DE TABELAS

4.1	Atributos relativos ao conjunto de dados das Estações Gira e respetiva descrição.	45
4.2	Nome dos atributos relativos às viagens e respetiva descrição.	48
4.3	Tipo de dados, descrição e unidades dos atributos que irão constar do conjunto de dados final.	51
4.4	Exemplo do conjunto de dados utilizado para o agrupamento temporal do comportamento das estações com base no seu rácio de ocupação.	59
4.5	Exemplo do conjunto de dados utilizado para o agrupamento temporal do comportamento das estações com base no valor médio de partidas por hora. .	62
4.6	Exemplo do raciocínio do conjunto de dados relativo à Soma Cumulativa do Número de Bicicletas e Saldo.	85
5.1	Exemplo do conjunto de dados obtido com a contagem do número de partidas em intervalos de 15 minutos.	90
5.2	Exemplo do conjunto de dados após o pivot obtido com a contagem do número de partidas em intervalos de 15 minutos.	91
5.3	Exemplo do conjunto de dados após a soma cumulativa diária com <i>reset</i>	91
5.4	Exemplo do conjunto de dados para analisar os <i>plateaus</i> para a Estação 101. .	107
5.5	Exemplo do conjunto de dados para com a Soma do <i>Plateau</i> para a Estação 101.	107
5.6	Descrição dos graus de ocupação criados.	108
A.1	Atributos e respetiva descrição.	122
A.2	Atributos, tipologia e exemplos de valores para os mesmos.	123
A.3	Registos duplicados nos dados.	134
A.4	Nome dos atributos e respetiva descrição.	136
A.5	Nome dos atributos, tipologia e exemplos de valores possíveis.	137
B.1	Morada, Identificador, Número de Docas, Data de Update e Número de Registos para a estação com identificador 421 e 115, mostrando incongruências nos dados.	148
B.2	Estações da Gira para as quais existe mais de um valor distinto de WKB. . . .	153
B.3	Estações que apresentam apenas um valor de localização que são coordenadas a zero, morada dessa estação e informação acerca da estação estar visível no website ou não.	153

B.4	Data e hora, número de estações e estações onde foram encontrados registos duplicados nos dados.	155
B.5	Grupos de Estações pela data de início dos registos.	158
B.6	Data de início selecionada para cada estação.	161
B.7	Identificadores das estações cujos registos foram removidos do conjunto de dados.	162
B.8	Viagens cujo atributo da Estação de fim apresenta valor a nulo.	164
C.1	Descrição dos conjuntos de dados da meteorologia.	168
C.2	Descrição das unidades de cada tipo de dados da meteorologia.	168
C.3	Estações meteorológicas de Lisboa e respetiva informação.	169
C.4	Valores dos registos duplicados presentes nos dados relativos à precipitação.	170
C.5	Contagem de registos com valor -990 para cada estação meteorológica relativamente à precipitação.	170
C.6	Valores dos registos duplicados presentes nos dados relativos à precipitação.	172
C.7	Contagem de registos com valor -990 para cada estação meteorológica relativamente à Humidade.	172
C.8	Valores dos registos duplicados presentes nos dados relativos à radiação.	175
C.9	Contagem de registos com valor -990 para cada estação meteorológica relativamente à radiação.	175
C.10	Valores dos registos duplicados presentes nos dados relativos à temperatura.	177
C.11	Contagem de registos com valor -990 para cada estação meteorológica relativamente à temperatura.	177
C.12	Valores dos registos duplicados presentes nos dados relativos à intensidade do vento.	179
C.13	Contagem de registos com valor -990 e nulo para cada estação meteorológica relativamente à intensidade do vento.	179
C.14	Valores dos registos duplicados presentes nos dados relativos à direção do vento.	182
C.15	Contagem de registos com valor inválido (-990) para cada estação meteorológica relativamente à direção do vento.	182
C.16	Tipo de dados, descrição e unidades dos atributos que irão constar do conjunto de dados final.	186

SIGLAS

ANN Artificial Neural Network.

BPNN Back Propagation Neural Network.

CML Câmara Municipal de Lisboa.

CRISP-DM Cross Industry Standard Process for Data Mining.

EMEL Empresa de Mobilidade e Estacionamento de Lisboa.

IPMA Instituto Português do Mar e da Atmosfera.

IVML Interactive Visualization Modeling Language.

LSTM Long Short Term Memory.

MAE Mean Absolute Error.

MAPE Medium Absolute Percentage.

MASE Mean Absolute Scaled Error.

MSE Mean Squared Error.

NRMSE Normalized Root Mean Squared Error.

RF Random Forest.

RFID Radio-frequency Identification.

RMSE Root Mean Squared Error.

RMSLE Root Mean Squared Logarithmic Error.

RNN Recurrent Neural Network.

SBP Sistema de Bicicletas Partilhadas.

SLA Service-level Agreement.

WKB Well-known binary.

WKT Well-known text.

INTRODUÇÃO

Os sistemas de bicicletas partilhadas têm vindo a tornar-se cada vez mais habituais nas grandes cidades ao longo dos últimos anos, devido a serem uma opção sustentável que permite evitar o trânsito nas grandes cidades, promovendo ainda um estilo de vida mais ativo para os seus utilizadores.

O aumento do número de interessados neste tipo de serviço tem causado alguns problemas no que diz respeito à disponibilidade de bicicletas e docas nas estações, o que tem suscitado o interesse para a resolução deste problema, de forma a que o serviço seja mais satisfatório.

1.1 Contexto e Motivação

A presente dissertação é um desafio proposto no âmbito de uma parceria entre o Laboratório de Dados Urbanos de Lisboa (*LxDataLab*)¹, a [Câmara Municipal de Lisboa \(CML\)](#) e a Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa. Os desafios lançados pelo *LxDataLab* têm como objetivo responder a necessidades do Município através da criação de soluções de análise e visualização de dados que contribuam para a melhoria do planeamento e gestão da cidade de Lisboa.[25, 26]

O desafio proposto, pretende que se encontre uma solução para otimizar as rotas de rebalanceamento das bicicletas partilhadas em Lisboa, mais concretamente, da rede de bicicletas da Gira², um projeto lançado pela [Empresa de Mobilidade e Estacionamento de Lisboa \(EMEL\)](#)³ em 2017, que disponibiliza um serviço de bicicletas partilhadas em Lisboa com estações localizadas em vários pontos da cidade[10].

Nos sistemas de bicicletas partilhadas com estações, quando uma bicicleta é alugada,

¹<https://lisboainteligente.cm-lisboa.pt/lxdatalab/>

²<https://www.gira-bicicletasdelisboa.pt/>

³<https://www.emel.pt/pt/>

o utilizador retira-a da estação e, após este ter efetuado a viagem pretendida, devolve a bicicleta na mesma estação ou noutra estação pertencente à rede. As estações são locais devidamente identificados e com docas onde os utilizadores podem colocar ou retirar bicicletas. Cada doca tem capacidade para uma bicicleta e uma estação tem capacidade dependente do seu número de docas, que pode variar de estação para estação.

O aluguer das bicicletas é levado a cabo utilizando uma aplicação móvel da Gira, na qual os utilizadores se encontram registados e onde podem visualizar o mapa da rede de estações com informação relativa à capacidade atual de cada uma, para decidir a qual irão alugar ou devolver uma bicicleta. O pagamento deste serviço é efetuado após a viagem ter terminado e a bicicleta se encontrar devidamente acondicionada numa doca. As empresas disponibilizam várias formas de pagamento, bem como alternativas para utilizadores frequentes, tais como passes mensais ou anuais [6, 11].

Na Figura 1.1 encontram-se as Estações Gira existentes neste momento, embora a rede de bicicletas tenha vindo a ser expandida desde 2017, ano em que a Gira foi inaugurada. Cada ícone representado no mapa representa uma estação e simultaneamente a ocupação da mesma, ou seja, o número de bicicletas que se encontram de momento na estação relativamente ao número de docas que esta contém. Como referido, encontram-se em vários locais da cidade de Lisboa, havendo uma concentração visível no centro da cidade e na zona do Parque das Nações, local onde se deu início do projeto. Este não é o único Sistema de Bicicletas Partilhadas (SBP) em Lisboa, no entanto, é o único que tem estações. A Gira tem tido sucesso em Lisboa, tendo chegado a 1 milhão de viagens em 2019 e, para além disso, nessa altura contava com mais de 19 mil passes anuais e 8800 passes diários vendidos [30].

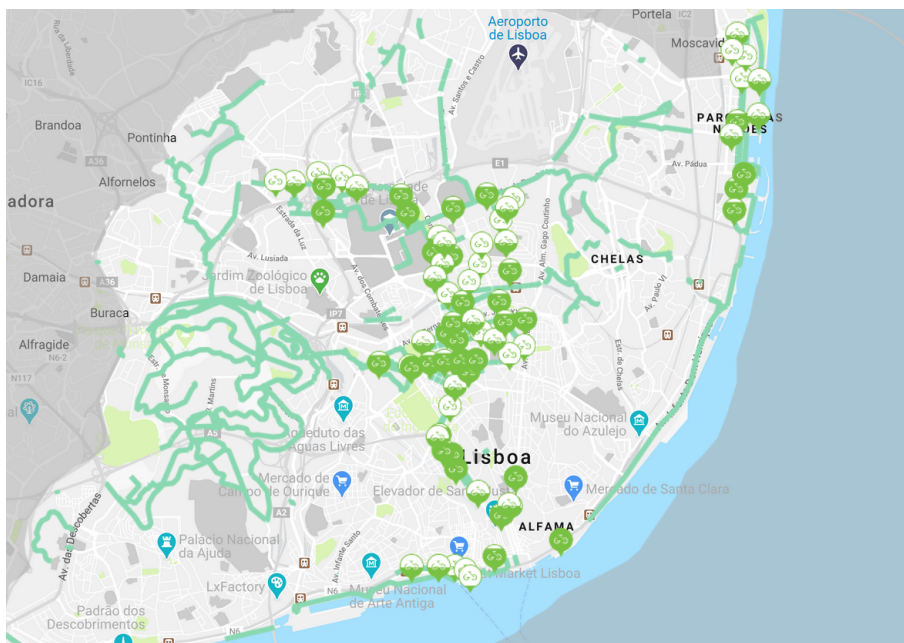


Figura 1.1: Rede de Estações da Gira. Retirado de ⁴.

A utilização das bicicletas partilhadas é cada vez mais adotada pelos cidadãos, sendo utilizada em deslocações para o trabalho, como complemento de transporte no último quilómetro ou apenas como forma de lazer. No entanto, devido à crescente popularidade deste tipo de serviço, o número de utilizadores tem vindo a aumentar, o que apresenta alguns desafios no que diz respeito à disponibilização de bicicletas ou existência de docas livres nas estações para permitir o seu aluguer ou devolução de bicicletas, respetivamente. Esta situação conduz a um estado de desequilíbrio da rede de bicicletas, no qual algumas estações se encontram cheias e outras se encontram vazias, sendo necessário efetuar a redistribuição das bicicletas pelas estações, de forma a garantir que nenhuma estação se encontra vazia ou cheia. As estratégias utilizadas para tal podem incluir os clientes destes sistemas ou veículos para transportar as bicicletas pelas estações, de modo a que cada estação tenha um número ideal de bicicletas. O rebalanceamento recorrendo a viaturas apresenta como desafios principais a otimização das rotas das mesmas, a determinação da quantidade ótima de bicicletas para cada estação e a altura do dia em que esse rebalanceamento deverá ser efetuado. [42]. Neste momento, o processo carece de otimização, nomeadamente no que diz respeito à predição da quantidade de bicicletas existentes em cada estação numa dada altura, uma vez que esta informação irá possibilitar a criação de estratégias de rebalanceamento que o irão tornar mais eficaz, evitando viagens desnecessárias.

1.2 Descrição do Problema

O rebalanceamento dos SBP com estações é necessário para possibilitar o restabelecimento do equilíbrio da rede, garantindo que todas as estações possuem uma quantidade de bicicletas ideal e, desta forma, melhorar o serviço disponibilizado aos clientes. Este processo é levado a cabo de diferentes formas, sendo necessária a previsão da procura de bicicletas nas estações, assim como a previsão de quando esse rebalanceamento é necessário, sendo possível criar estratégias de rebalanceamento mais eficazes, pois estima-se não só a quantidade de bicicletas, mas também o momento em que este procedimento tem de ser efetuado. Tal irá diminuir o número de situações em que o utilizador não tem bicicletas ou docas disponíveis para poder efetuar o aluguer ou devolução de uma bicicleta, respetivamente.

Os dados facultados no âmbito desta dissertação dizem respeito às estações da rede de bicicletas partilhadas da Gira, às viagens que ocorreram na rede e a meteorologia na cidade de Lisboa no ano de 2018. Os dados relativos às estações contêm informação acerca do número de bicicletas que se encontravam em cada estação num dado momento, permitindo perceber qual a ocupação de cada estação. Os dados das viagens permitem perceber o movimento das bicicletas entre as estações ao longo do tempo, tendo informação relativa à estação onde a viagem teve início e onde esta terminou. Relativamente à

⁴<https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/>

informação meteorológica, esta é a percentagem da humidade, precipitação, intensidade do vento, temperatura, direção do vento e radiação numa granularidade temporal horária.

Com base nos dados disponibilizados e após uma análise da qualidade destes, pretende-se efetuar a sua análise exploratória bem como outros tipos de análises. Nomeadamente, pretende-se analisar os padrões de ocupação e desocupação das estações, mais concretamente, quando é que estas se encontram vazias ou cheias, identificar quando ocorrem mais viagens e quando é que não há viagens. É ainda pretendido estudar o comportamento de cada estação tendo em conta diferentes granularidades temporais, bem como detetar grupos de estações que exibem o mesmo tipo de perfil de ocupação, não só diariamente, mas tendo em conta os dias da semana, com enfoque na diferença de ocupação das estações entre dias da semana e entre os fins-de-semana e feriados.

Tendo em conta estas análises, o objetivo é efetuar a predição dos níveis de bicicletas nas estações da Gira ao longo do tempo, de forma a perceber quantas bicicletas irão estar em cada estação, possibilitando a criação de uma estratégia de rebalanceamento da rede antes desse rebalanceamento ser necessário.

1.3 Contribuições

O resultado que se pretende obter após a presente dissertação inclui algumas contribuições, que serão descritas de seguida.

Relatório de qualidade de dados

Este relatório pretende descrever e efetuar a análise dos dados facultados pelo Lx-DataLab, que dizem respeito a dados das viagens efetuadas na rede de bicicletas partilhadas da Gira e dados relativos à ocupação das estações da rede Gira no ano de 2018. Foi ainda efetuado um relatório de qualidade dos dados e processamento relativo aos dados meteorológicos do [Instituto Português do Mar e da Atmosfera \(IPMA\)](#).

Relatório de processamento de dados

Após ser efetuada a primeira análise aos dados e aferição qualidade dos mesmos, é necessário efetuar o pré-processamento destes, tendo por base as conclusões retiradas. Como tal, este relatório especifica o processamento realizado aos dados das viagens e das estações da Gira.

Criação de *dashboards* para análise exploratória dos padrões das estações e viagens da Gira

Recorrendo aos dados fornecidos, mais concretamente, das viagens, estações e meteorologia foi efetuada uma análise exploratória, para a qual foram criados *dashboards*

de forma a poder detetar grupos de estações tendo em conta diversos parâmetros, nomeadamente grupos espaciais e temporais, bem como para detetar padrões no comportamento das estações, tendo em conta diferentes granularidades temporais. Para algumas análises foi necessário formatar e criar novos dados, para responder a questões formuladas no início do processo.

Foi ainda desenvolvido um *dashboard* com o objetivo de tentar detetar o rebalanceamento nos dados relativos à quantidade de bicicletas que se encontram nas estações Gira.

Previsão da procura de bicicletas

Utilizando os dados mencionados acima para criar um conjunto de dados adequado ao que se pretendia, em conjunto com outros tipos dados como, por exemplo, dados meteorológicos, foram aplicadas técnicas e algoritmos de previsão de forma a efetuar a previsão da quantidade de bicicletas que irão estar nas estações da rede ao longo do tempo. Foi levado a cabo um processo de validação e, depois, foi efetuada a previsão para cada estação tendo em conta os melhores parâmetros.

1.4 Estrutura do Documento

Este documento encontra-se estruturado em cinco capítulos, que se encontram descritos de seguida:

Introdução

É o presente capítulo, que apresenta uma breve introdução do tema, descreve o contexto e motivação da dissertação e descreve qual o problema a resolver. Para além disso, são apresentadas as contribuições para o período desta dissertação e a estrutura do documento.

Trabalho Relacionado e Estado da Arte

O estado da arte e trabalhos relacionados com a temática desta dissertação serão abordados no Capítulo 2. Neste caso, os tópicos abordados foram as bicicletas partilhadas, rebalanceamento, previsão de séries temporais e da procura de bicicletas, algoritmos de agrupamento e algoritmos de previsão.

Abordagem e Metodologia

O capítulo 3, denominado Abordagem e Metodologia, descreve a abordagem utilizada para proceder à resolução do problema da presente dissertação, referindo as técnicas utilizadas aquando do processamento de dados e do desenvolvimento do agrupamento e dos modelos preditivos.

Análise Exploratória de Dados

No capítulo 4 é elaborada uma descrição dos conjuntos de dados utilizados nesta dissertação, bem como o processamento que foi necessário efetuar aos mesmos de forma a poderem ser utilizados.

Encontram-se ainda descritas as análises efetuadas utilizando os dados processados, nomeadamente o processo de agrupamento espacial e temporal, a caracterização espacial e temporal das Estações Gira tendo em conta as suas viagens e quantidade de bicicletas nas estações e ainda a deteção do rebalanceamento nas estações da rede Gira.

Modelos Preditivos

No capítulo 5 capítulo encontra-se o processo de criação do novo conjunto de dados desenvolvido para proceder à previsão, o processo de desenvolvimento dos modelos preditivos, a avaliação dos modelos na fase de validação, bem como a fase da seleção dos melhores parâmetros, sendo que, no final, são analisados os resultados obtidos. É ainda levada a cabo uma análise para a qual foi criado um *dashboard* com o objetivo de avaliar os *plateaus* existentes nos dados, de forma a compreender o porquê destes ocorrerem.

Conclusões e Trabalho Futuro

Será feita a conclusão do trabalho no capítulo 6, bem como algumas possibilidades de trabalho futuro que possa ser efetuado utilizando estes dados, ou dentro desta temática.

TRABALHO RELACIONADO E ESTADO DA ARTE

Neste capítulo será levado a cabo o estudo do trabalho relacionado e do estado da arte acerca de diversas temáticas e técnicas existentes que sejam pertinentes para a problemática que se pretende abordar na dissertação. De forma a pesquisar trabalhos relacionados, foi necessário efetuar uma pesquisa recorrendo a algumas palavras chave, sendo que esse processo será descrito na secção 2.1. Seguidamente, serão apresentados os trabalhos relacionados relativos aos temas relevantes para a presente dissertação, nas temáticas de sistemas de bicicletas partilhadas, agrupamento, séries temporais e *forecasting*.

2.1 Introdução

Este capítulo apresenta os resultados da pesquisa efetuada dividida em algumas secções que abordam as bicicletas partilhadas, o rebalanceamento e quais os tipos abordados na literatura, grafos, séries temporais, o *forecasting* de séries temporais, avaliação da precisão do *forecasting*, *forecasting* da procura de bicicletas e *clustering* de séries temporais.

De forma a encontrar informação relevante acerca dos tópicos supramencionados, foi levada a cabo uma pesquisa, recorrendo a motores de busca como o *ScienceDirect*, *Research Gate*, *IEEE Xplorer*, *Google Scholar* e *Springer Link*. As palavras-chave utilizadas para encontrar informação relativa às temáticas mencionadas foram as seguintes: "*rebalancing bike-sharing*", "*dynamic rebalancing bike-sharing*", "*bike-sharing rebalancing system*" e "*rebalancing strategy*", "*time series forecasting*" e "*forecasting bike sharing*". A partir destas foi possível encontrar um número considerável de artigos de interesse, sendo que a pesquisa apresentou como restrição a data da publicação e o número de citações.

2.2 Trabalho Relacionado

2.2.1 Bicicletas Partilhadas

Os SBP têm vindo a crescer cada vez mais ao longo dos anos a nível mundial, sendo que em 2017 estes sistemas já tinham sido aplicados em 1173 cidades e contavam com cerca de 1.9 milhões de utilizadores ativos [12].

Desde a sua primeira geração com origem em Amesterdão no ano de 1965, este tipo de serviço evoluiu e hoje em dia incorpora tecnologia nas suas bicicletas e estações [27], através da utilização de GPS, *wi-fi* e sensores que permitem a recolha de dados, possibilitando a consulta de informação em tempo real acerca do número de bicicletas em cada estação, auxiliar os utilizadores a encontrar a estação mais próxima e melhorar a qualidade do serviço disponibilizado [5].

Empresas como a CitiBike¹ em Nova Iorque, a Mobike² na China, a Velov³ em Lyon ou a Gira⁴ em Lisboa disponibilizam estas bicicletas nos centros urbanos e possibilitam o seu aluguer, bem como a consulta de informação em tempo real pelos seus utilizadores através de aplicações móveis. Em Portugal existem algumas empresas de bicicletas partilhadas, sendo que a Gira se diferencia por apresentar um sistema com estações com docas, podendo-se observar uma das estações da rede da Gira na Figura 2.1 e todo o sistema de estações na Figura 1.1. O aumento da popularidade deste meio de transporte



Figura 2.1: Estação com Bicicletas da Gira em Lisboa. Retirado de ⁵.

deve-se ao facto de promover a mobilidade sustentável, apresentar baixo impacto ambiental, diminuir o tráfego e poluição causada pelos automóveis e resolver o problema do “último-quilómetro” criado pelos sistemas de transportes nas áreas de menor densidade

¹<https://www.citibikenyc.com/>, consultado a 02/2020

²<https://mobike.com/global/>, consultado a 02/2020

³<https://velov.grandlyon.com/>, consultado a 02/2020

⁴<https://www.gira-bicicletasdelisboa.pt/>, consultado a 02/2020

⁵<http://lisboa.convida.pt/poi/see-do/gira-bicicletas-de-lisboa-10314/71>, consultado a 02/2020.

de tráfego [5, 7].

DeMaio [8] aborda a história destes sistemas, afirmando que as bicicletas partilhadas contribuíram para o aumento do número de pessoas que utilizam a bicicleta como meio de transporte. No entanto, este aumento não teve como consequência o decréscimo de utilização de outros meios de transportes públicos. A indústria dos sistemas de bicicletas partilhadas tem apresentado um maior crescimento na Europa, onde a procura por este tipo de serviços é superior.

Existem estudos comparativos efetuados sobre sistemas de bicicletas partilhadas, como é o caso do trabalho [27], onde são comparadas todas as gerações destes sistemas desde a sua criação. Ao longo deste estudo foi elaborada uma *framework* para permitir efetuar a comparação dos dados de sistemas de bicicletas partilhadas, com o objetivo de compreender melhor a magnitude de cada SBP, tendo sido selecionados cinco destes sistemas para esse propósito. Os dados reunidos acerca destes sistemas incluem informações acerca do número de estações de cada sistema, o número de bicicletas e o custo de operação, dados relativos aos preços praticados por cada sistema e ao modelo de negócio de cada um, nomeadamente sobre o proprietário e o operador de cada serviço.

Outros estudos têm sido realizados acerca do modelo de negócio destes sistemas, como é o exemplo do trabalho [12], no qual foi examinado um projeto piloto de um SBP numa universidade da Indonésia, tendo sido realizados questionários a 105 estudantes que participaram como utilizadores do sistema e tendo-se observado a forma como estes utilizaram o sistema. Após este processo, os dados foram mapeados de forma a ilustrar o modelo de negócio, tendo sido desenvolvido um Business Model Canvas (BMC) para ilustrar o fluxo de informação e identificar *stakeholders* do projeto. Foram ainda determinados riscos potenciais do negócio, tendo sido elaborado um plano de mitigação do risco e uma análise SWOT. O objetivo deste estudo foi providenciar uma referência para os sistemas de bicicletas em desenvolvimento no país, possibilitando delinear estratégias de negócio e possibilitar a criação de mais sistemas de bicicletas partilhadas.

2.2.2 Rebalanceamento

A utilização das redes de bicicletas partilhadas leva à ocorrência de um desequilíbrio na quantidade de bicicletas disponíveis nas estações da rede, que ocorre quando estas se encontram vazias ou cheias, o que impossibilita, respetivamente, o aluguer de bicicletas e a sua devolução. Estes problemas ocorrem, principalmente, devido ao facto da procura por bicicletas variar tendo em conta a altura do dia e o local, tendo-se constatado, por exemplo, que existe uma maior procura por bicicletas em zonas residenciais durante a manhã e uma rápida ocupação das docas em zonas comerciais [5]. Com o desenvolvimento deste tipo de serviços, existe o interesse de melhorar os mesmos e, uma vez que o principal fator de insatisfação por parte dos clientes está relacionado com a indisponibilidade de

⁶<https://bikeportland.org/2016/03/23/portlands-bike-share-plan-gives-major-leeway-to-private-operator-178818> consultado a 02/2020

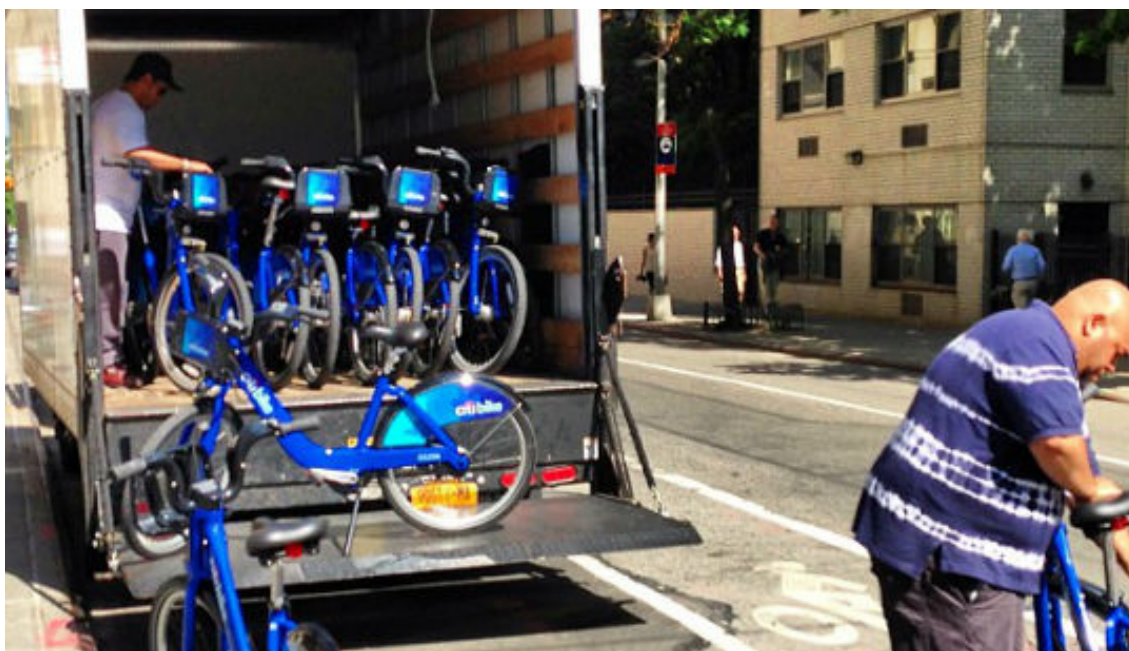


Figura 2.2: Camião para efetuar o Rebalanceamento na rede da CitiBike. Retirado de ⁶.

bicicletas ou de docas livres, é importante solucionar o problema do rebalanceamento. Desta forma, têm sido aplicadas várias estratégias de forma a efetuar a redistribuição das bicicletas pela rede, sendo que algumas se baseiam na utilização de camiões e outras utilizam os clientes para o fazer, denominando-se de rebalanceamento a este processo. Na Figura 2.2, é possível observar um camião no qual estão a ser colocadas bicicletas de forma a serem redistribuídas pela rede da empresa CitiBike em Nova Iorque, sendo um exemplo de um tipo de rebalanceamento. O processo de rebalanceamento consiste na realocação de bicicletas pela rede, sendo que são retiradas bicicletas de estações que se encontram quase cheias ou cheias para estações que se encontram vazias ou com menos bicicletas, de forma a que as estações tenham um número ideal de bicicletas [42].

A análise espaço-temporal dos padrões de rebalanceamento dos SBP tem sido um tópico de interesse, tendo sido abordada no trabalho [2], no qual foram analisados 9 casos de estudo em Londres, Nova Iorque, Viena, Luxemburgo, São Francisco, entre outros. Os SBP em estudo foram contactados para recolher informações relativamente às estratégias de rebalanceamento aplicadas, o número de veículos de rebalanceamento disponíveis, o número de bicicletas que são redistribuídas e que informações são utilizadas para facilitar o processo de rebalanceamento. Depois, foi elaborada uma análise estatística e visual dos dados relativos a cada SBP entre 2011 e 2014, tendo sido considerados apenas os dados relativos aos dias úteis, devido ao facto da utilização do sistema e rebalanceamento serem mais intensos durante este período da semana. Verificou-se que a maioria dos funcionários não faz uma distinção entre a localização ou o tipo de falha do sistema (se a estação está vazia ou cheia), no entanto, é importante perceber se as falhas ocorridas se devem a viagens potencialmente perdidas. Para além disso, também se verificou que estações

localizadas perto de comboios utilizam uma quantidade desproporcionada de recursos. A análise destes sistemas permitiu perceber que a ocorrência de falhas nos mesmos aumenta com a sua utilização e que o aumento do rebalanceamento em estações próximas de transportes públicos acaba por originar menos viagens. Adicionalmente, aumentar o número de docas e bicicletas em estações com muita utilização resulta num aumento de viagens, resultando na necessidade de aumentar o rebalanceamento, o que acarreta custos. Constatou-se que o processo de rebalanceamento costuma ser efetuado como consequência da procura nos períodos da manhã e da tarde exceder a capacidade das estações e não devido à acumulação de bicicletas a longo prazo.

No que diz respeito à elaboração de estratégias de rebalanceamento, no trabalho [7] é analisado um caso real em Reggio Emilia, na Itália, no qual foram utilizados dados das viagens efetuadas pelos utilizadores do sistema durante 7 meses, incluindo a data e hora, bem como a estação de início e fim da viagem. Este SBP é composto por 13 estações e cerca de 100 bicicletas, existindo um armazém com bicicletas que é utilizado no processo de redistribuição. A redistribuição de bicicletas ocorre apenas durante a noite, utilizando apenas um veículo que visita cada estação exatamente uma vez. Utilizando os dados, foi avaliado o fluxo de bicicletas ao longo do dia para cada estação, tendo sido computada a diferença entre as bicicletas que chegam à estação e as que são retiradas da mesma. Com base nesta análise, as estações foram divididas em três grupos e foram criados gráficos de linhas com informação do número de partidas e chegadas ao longo das horas do dia para uma estação de cada um dos grupos, que se podem observar na Figura 2.3. Neste trabalho foi criado um digrafo cujos nós são as estações e o armazém. Cada estação tem um pedido que poderá significar a retirada ou adição de bicicletas da mesma. Foram apresentadas quatro formulações *Mixed Integer Linear Programming* (MILP) para resolver o problema do rebalanceamento que, devido a apresentarem um número exponencial de restrições, foram resolvidas utilizando um algoritmo de *branch-and-cut*. Depois, foram recolhidos dados de 22 sistemas de bicicletas partilhadas nas quais foram aplicados os mesmos modelos mencionados anteriormente. Os resultados permitiram perceber que quando o número de estações de um SBP e a capacidade do veículo de rebalanceamento aumentam também se torna mais difícil de resolver.

O trabalho [42] aborda uma estratégia de rebalanceamento centrada no utilizador, que tem por base a utilização de incentivos para motivar o utilizador a colocar ou retirar uma bicicleta na estação mais adequada. O objetivo é que a rede de bicicletas partilhadas se mantenha equilibrada, contendo a quantidade ideal de bicicletas em cada estação. Primeiro, foi definido o problema da seguinte forma: dada uma estação da rede na qual o cliente aluga uma bicicleta, o SBP irá determinar se deve incentivar o utilizador a deslocar-se ou não a outra estação. Se necessário, o sistema de bicicletas deverá recomendar uma estação. A metodologia utilizada consistiu em dois tópicos: calcular o estado ótimo de cada estação e selecionar a estação ótima. Para determinar qual o estado ótimo de cada estação, utilizou-se, para além de probabilidades, o modelo *Random Walk*. O objetivo é que a estação possa manter o estado ótimo pelo maior tempo possível, significando que

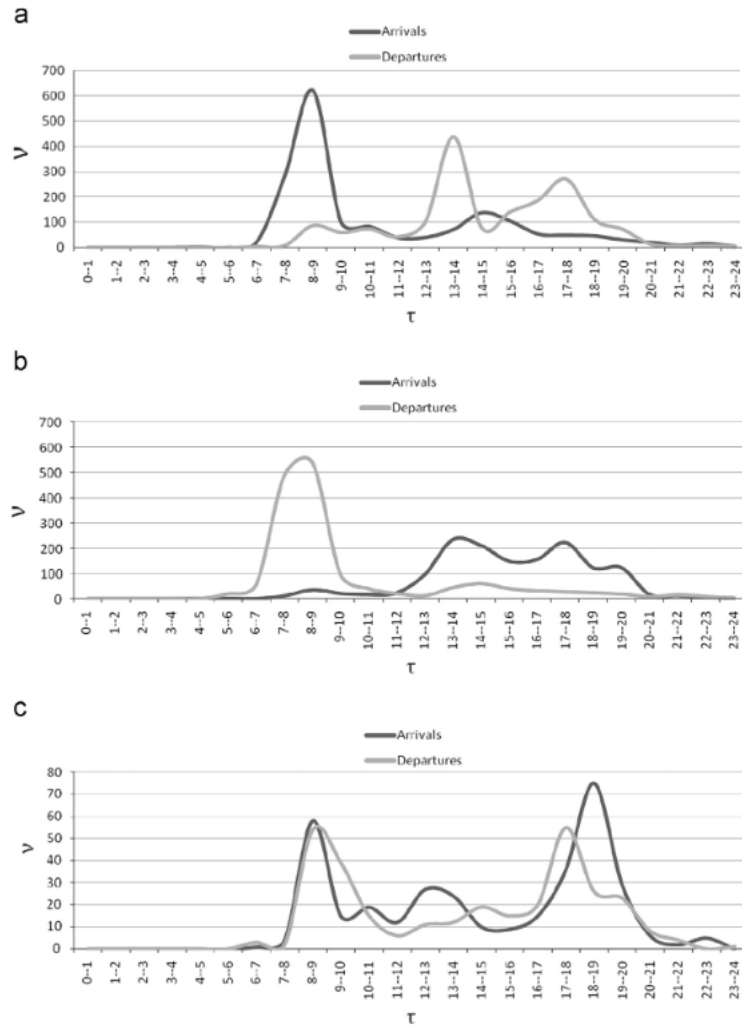


Figura 2.3: Partidas e Chegadas por hora numa estação (a) do grupo 1, (b) do grupo 2 e (c) do grupo 3. (τ - hora do dia; v - número cumulativo de bicicletas que chegaram ou partiram de uma estação durante um período de 7 meses. Retirado de [7])

deverá ter a menor possibilidade de ficar vazia ou cheia, logo é necessário calcular esta probabilidade. Foi ainda tido em conta o processo de chegada à estação para alugar ou retirar bicicletas de acordo com um processo de Poisson, tendo sido incluídas as probabilidades do cliente alugar ou devolver uma bicicleta. Para determinar qual a estação ótima é necessário resolver primeiro dois problemas: a estação selecionada deve encontrar-se o mais próxima possível da estação pretendida por parte do cliente, uma vez que, caso se encontre longe, o cliente poderá não querer deslocar-se até à mesma, invalidando o processo e, em segundo lugar, o incentivo deverá ter um impacto positivo no sentido da estação ficar mais próxima do seu estado ótimo. Tendo em conta estes dois problemas, aquando da escolha de uma estação pelo cliente, a aplicação do [SBP](#) deve incentivá-lo a alugar uma bicicleta noutra estação caso a estação escolhida não se encontre no seu estado ótimo. No final deste trabalho, foi concluído que a abordagem desenvolvida apresentou os

melhores resultados no que diz respeito à redução do desequilíbrio do sistema mantendo os custos do processo muito baixos, comparativamente aos métodos mais utilizados de rebalanceamento estático (colocar as estações no seu estado ótimo uma vez por dia) e rebalanceamento dinâmico.

No trabalho [5] é proposta uma estratégia de rebalanceamento dinâmica que, tendo em conta dados históricos do SBP de Nova Iorque, prevê a ocupação das estações, permitindo decidir quando é necessário efetuar a sua redistribuição, utilizando um processo de *Birth-Death* e teoria de grafos para a seleção do caminho a utilizar aquando do rebalanceamento. O SBP foi definido como um grafo completamente ligado, em que os nós representam as estações e são caracterizados pela sua capacidade e as ligações entre as estações são duplas e estão pesadas de acordo com a distância. Primeiro foi necessário determinar o estado ótimo da estação para cada bloco de tempo e identificar que estações estão quase cheias ou quase vazias, de forma a determinar as rotas dos camiões. O problema do rebalanceamento dinâmico foi modelado como um problema de otimização num grafo, no qual são tidos em conta o número de camiões disponíveis, a sua capacidade, quais as estações que precisam de ser rebalanceadas e o tempo de viagem dos veículos. Os resultados obtidos permitiram perceber que o modelo definido é realista e que o rebalanceamento dinâmico pode ser superior ao estático, permitindo aos funcionários encarregues do processo a decisão sobre a priorização dos custos ou da qualidade do serviço. Para além disso, o rebalanceamento dinâmico pode permitir reduzir as falhas durante a hora de ponta durante a manhã nos dias úteis, quando há mais clientes a utilizar o sistema e uma maior demanda de bicicletas. Os autores concluíram ainda que a estimação incorreta dos padrões de procura poderia ser melhor no caso de serem adicionados dados meteorológicos aos dados existentes.

2.2.3 *Forecasting da procura de bicicletas*

A necessidade de levar a cabo o rebalanceamento dos SBP deve-se ao desequilíbrio na disponibilidade de bicicletas e de docas livres nas estações da rede, sendo necessário retirar bicicletas de algumas estações e colocá-las noutras, permitindo que os clientes possam continuar a efetuar o aluguer e devolução das mesmas.

Alguns trabalhos mencionados anteriormente focaram-se em estratégias de rebalanceamento utilizando teoria de grafos e algoritmos matemáticos. No entanto, existem trabalhos na literatura em que é levada a cabo a previsão da procura de bicicletas para a resolução do problema de rebalanceamento, visto que, se a previsão for precisa, poderá melhorar a qualidade do mesmo [42].

No trabalho [41], a procura é estimada utilizando correlação temporal e espacial nas situações em que um cliente se desloca a uma estação para alugar uma bicicleta e não existem bicicletas disponíveis nessa mesma estação, sendo que tal é algo que não é detetável nos dados, sendo o objetivo do estudo. Para tal, foram desenvolvidos três modelos com os seguintes objetivos: estimar a procura com base na correlação espacial, estimar a

procura com base na correlação temporal, integrar os modelos e estimar a procura tendo em conta a correlação espacial e temporal. Para estimar a procura com base na correlação temporal foi conduzida uma correlação de *Pearson* em dados dos meses de julho e agosto para encontrar os dias com a correlação mais elevada. Depois foi combinada a utilização real da estação na hora anterior e a diferença do valor estimado no momento. Para a estimação com base na correlação espacial constatou-se que estações adjacentes apresentam semelhanças e que o padrão de utilização tem uma elevada correlação espacial. A partir das semelhanças entre as estações adjacentes e a estação referência são obtidos os pesos a atribuir, que são combinados com a mudança de utilização das estações. No final, a procura pelo utilizador no momento é estimada combinando a utilização real da estação na hora anterior e a taxa estimada da diferença de utilização na hora atual. Integrando a procura espacial e temporal, a proporção de correlação temporal será maior relativamente à espacial em estações localizadas em zonas residenciais ou perto de escolas, do que em estações localizadas perto de atrações turísticas, no entanto, a proporção de correlação espacial e temporal foi a mesma para a condução das experiências de forma a simplificar. Nas experiências apenas foi selecionada uma estação da rede localizada em Washington, sendo que os dados apenas dizem respeito à procura das estações, à sua localização e ao estado das estações durante um período de um mês. Após a fase de experiências, foi conduzido o mesmo procedimento para as restantes estações, tendo sido obtida a média para cada hora e a média de todas as estações. Utilizando o Erro Médio como métrica para avaliar os resultados, estes foram mais estáveis para a correlação temporal de toda a rede comparativamente a apenas uma estação, sendo que os valores do erro diminuem quando a utilização da rede também diminui (durante a noite). Relativamente à correlação espacial e temporal, os valores o erro são muito mais estáveis, no entanto, mesmo o valor mínimo do erro é de 45.5%.

Zhang et. al [43] desenvolveu um estudo com dados de bicicletas e outros transportes públicos de Ningbo na China durante o período de um mês. Os dados da rede de bicicletas diziam respeito às viagens ocorridas na rede, com informação da origem e destino de cada uma. Relativamente aos dados adicionais, foram utilizados dados dos autocarros acerca de quando e onde um passageiro entra num dado autocarro, havendo ainda informação relativamente às paragens dos autocarros e informação GPS dos autocarros que indicam a localização dos mesmos. Para os dados do metro, também está disponível a informação de quando e onde um passageiro entrou numa estação e ainda informação relativa à localização e identificação das estações do metro. Para o caso em estudo, a metodologia foi aplicada apenas para uma estação de bicicletas partilhadas, tendo sido necessário extrair potenciais utilizadores de acordo com 6 procedimentos: procurar linhas de transportes públicos que estejam perto de uma estação de bicicletas partilhadas, procurar estações de transportes públicos que estejam diretamente ligados a uma estação de bicicletas partilhadas, extrair o tempo de viagem entre estações de transportes públicos e a estação de bicicletas, agrupar as estações de transportes públicos, contar o número de passageiros e a utilização das bicicletas partilhadas e, finalmente, colocar os dados na rede neuronal

e efetuar a previsão. Para a previsão foi utilizada uma rede neuronal [LSTM](#), tendo sido previsto apenas um único *time step*, tendo em conta os 180 minutos anteriores (3 horas), com a informação relativa às bicicletas partilhadas numa granularidade temporal de 15 minutos. Para determinar a eficácia da metodologia aplicada, esta foi comparada com três *baselines*, tendo-se constatado que o [LSTM](#) teve o melhor desempenho, encontrando-se mais próximo dos dados reais, especialmente durante as horas de ponta. A medição do desempenho da metodologia implementada foi efetuada recorrendo ao [Root Mean Squared Error \(RMSE\)](#) e [Mean Absolute Error \(MAE\)](#) que, caso apresentem valores mais baixos, significa que a previsão foi mais precisa. Os resultados mostram que a [LSTM](#) apresenta melhores resultados para a previsão de séries temporais e que os resultados são melhores quando incluem informação relativa aos transportes públicos do que quando esta influência não é incluída. Para além disso, concluiu-se que, quando o horizonte de previsão é maior, os erros também aumentam. No entanto, a previsão para 1 hora utilizando [LSTM](#) é melhor relativamente a previsões utilizando outros algoritmos para os próximos 15 minutos, mostrando que as [LSTMs](#) são melhores a captar tendências e a capturar características temporais.

Ruffieux et. al [32] fez um estudo recorrendo a dados dos sistemas de bicicletas partilhadas de seis cidades europeias, no qual foi testado um algoritmo de CNN (Convolutional Neural Networks) em comparação com outros como [Random Forest \(RF\)](#). Os dados recolhidos foram dados da rede de bicicletas, dados meteorológicos (temperatura, humidade, vento, nuvens) e dados relativos a feriados. Para as experiências foram utilizados apenas dados do [SBP](#) na Bélgica e apenas num intervalo temporal entre março e julho pois havia problemas de falta de dados e neste intervalo temporal a rede de estações manteve-se constante. A fase de modelação dividiu-se em três abordagens: estações independentes e individuais, havendo um modelo para cada estação; estações individuais e parcialmente independentes, considerando estações vizinhas aquelas que estão a menos de 3km de distância e estações dependentes, criando apenas um modelo para todas as estações. Após a aplicação do [RF](#) e analisando os resultados dos três tipos de modelos, o que obteve melhores resultados foi o que considera que as estações vizinhas influenciam a estação para a qual está a ser efetuada a previsão. Depois, foi implementado uma CNN, tendo-se concluído que, embora os resultados sejam piores relativamente aos do [RF](#), o aumento do erro relativamente ao aumento do horizonte previsto é menor, o que pode indicar potencial na previsão de horizontes maiores. Para avaliar a precisão dos modelos, foram utilizadas as métricas [RMSE](#) e [Normalized Root Mean Squared Error \(NRMSE\)](#).

No ano de 2018, o mesmo autor desenvolveu o trabalho [31], no qual se pretende otimizar as operações diárias de rebalanceamento em três Sistemas de bicicletas partilhadas da Suíça recorrendo a algoritmos de *machine learning*. Para efetuar a previsão, foi criado um sistema para efetuar o pré-processamento dos dados para que seja possível aplicar o algoritmo pretendido, sendo possível fazer um modo de avaliação no qual se utiliza 80% dos dados para treino e os restantes 20% (dados mais recentes) para teste. As experiências efetuadas no trabalho fizeram a previsão tendo em conta vários horizontes

de 15, 30, 60, 120, 360 e 720 minutos, tendo sido selecionados dois algoritmos para o estudo: Random Forest [RF](#) e Multi-Layer Perceptron (MLP). Para avaliar os modelos foi utilizada a métrica [RMSE](#), sendo que os resultados revelaram que o algoritmo MLP foi o que teve melhor desempenho, tendo apresentado os valores mais baixos de [RMSE](#). No entanto, a diferença de desempenho entre ambos os algoritmos decresce com o aumento do horizonte de previsão, verificando-se que o [RF](#) apresenta melhores resultados para previsões num horizonte de 720 minutos.

Chen et. al [4] efetuou a previsão da procura para cada estação do [SBP](#) CitiBike utilizando dados com informação relativa às viagens ocorridas na rede. Em adição a estes dados, também foram utilizados dados meteorológicos com informação de temperatura, intensidade do vento, humidade, precipitação e tipo de meteorologia (uma escala de classificação da meteorologia). Para a previsão foi utilizada uma [Recurrent Neural Network \(RNN\)](#) bi-direcional para prever o número de bicicletas que saem e entram nas estações da rede na próxima hora utilizando dados das viagens com um intervalo temporal de 5 minutos, a informação meteorológica e informação em *one hot encoding* do ano, mês, dia da semana, hora e quarto de hora. Como *baseline* foram utilizados os métodos de Ordinary Least-Squares Regression (OLS), Random Forest [RF](#) e Feedforward Neural Network (FNN) recorrendo às métricas [MAE](#), [RMSE](#), [Medium Absolute Percentage \(MAPE\)](#) e [Root Mean Squared Logarithmic Error \(RMSLE\)](#) para avaliar os resultados. A análise dos resultados mostra que os modelos RNN utilizados têm melhores resultados relativamente à *baseline*.

Na abordagem proposta por Xu et. al [39], foram utilizados dados de um [SBP](#) em Wenzhou, na China, para prever a procura de bicicletas no mesmo. Os dados foram relativamente às viagens ocorridas na rede durante o ano de 2016, havendo informação relativamente à estação de início e fim, bem como a respetiva data e hora, assim como o código associado ao cliente que fez o aluguer e entrega da bicicleta. Adicionalmente, foram utilizados dados meteorológicos e informação relativa a dias úteis e feriados. Nos dados utilizados foi necessário remover as viagens efetuadas por motivos de manutenção, ou seja, viagens para efeitos de rebalanceamento das estações, que se encontravam indicadas como tendo o valor de "M" no atributo onde está a informação relativa ao código do cliente. Foi levada a cabo uma análise dos dados das viagens e dos dados relativos às estações da rede, tendo sido detetados padrões temporais e espaciais para cada tipo de dados mencionado. Foi analisada a correlação dos dados das viagens com os dados meteorológicos e dados sobre os dias úteis, tendo-se constatado os dias úteis e a chuva são os fatores que mais influenciam a utilização diária da rede. A análise dos padrões temporais foi efetuada agrupando as viagens de acordo com a sua duração, sendo que cerca de 87% tem duração menor que 30 minutos. A análise espacial foi efetuada separadamente para dias úteis e dias não úteis, dividindo toda a rede em zonas com a cor a representar o número de viagens dentro dessa zona e mapeando os fluxos entre as estações com a espessura a indicar o número de viagens entre as zonas, sendo possível observar as visualizações em questão na Figura 2.4. Esta análise permitiu perceber que o padrão de

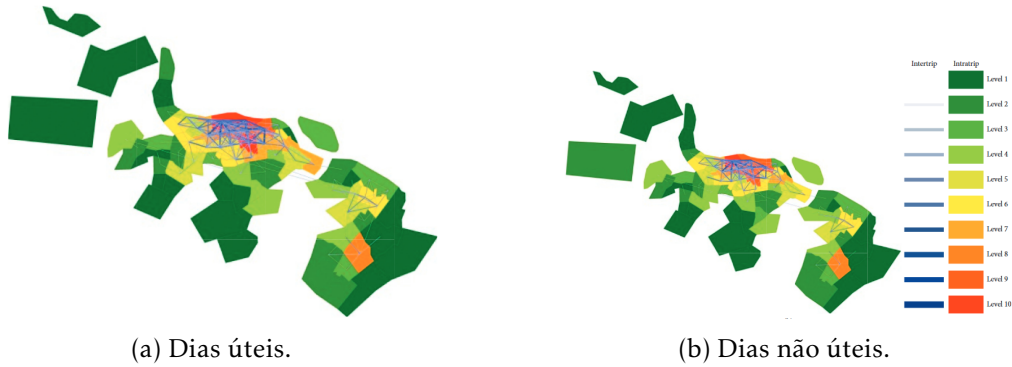


Figura 2.4: Viagens diárias entre zonas a intra-zonas em Wenzhou, China. Retirado de [39].

viagens varia de acordo com dias úteis e não úteis e também que há um maior número de viagens nos dias úteis. A análise dos dados das estações apresenta informação horária acerca do número de partidas e do número de chegadas que ocorreram em cada estação, tendo sido analisados estes dados tendo em conta dias úteis e não úteis e dias em que choveu, tendo-se concluído que se deve ser feita a distinção entre dias úteis e não úteis aquando da previsão. Os padrões espaciais permitiram perceber que, de facto, a rede de bicicletas é usada para efetuar o percurso de casa para o trabalho e vice-versa, tendo-se concluído que deve ser feita uma distinção entre estações e entre dias úteis e não úteis. Após a análise dos dados, as estações foram agrupadas de acordo com a sua taxa horária de partidas e chegadas, tendo sido utilizado o algoritmo *K-means*. Para efetuar a previsão, foram utilizadas [Back Propagation Neural Network \(BPNN\)](#) para cada tipo de estação, bem como para dias úteis e não úteis, tendo sido utilizada uma rede com três camadas. Para avaliar a precisão dos modelos dado cenários diferentes, foram considerados dados com distinção entre dias úteis e não úteis, dados com distinção entre estações e dados com distinção entre estações, dias úteis e não úteis. A avaliação da precisão das previsões foi efetuada recorrendo às métricas [RMSE](#), [NMRSE](#), [MAE](#) e [NMAE](#), tendo-se constatado que a distinção de estações, dias úteis e não úteis pode melhorar a precisão da previsão, pois foi o cenário com os valores de erro mais baixos.

No trabalho [38] foram utilizados dados das viagens ocorridas na CitiBike em Nova Iorque, dados de utilização de táxis e dados meteorológicos de granularidade temporal horária relativos à temperatura, humidade, velocidade do vento, visibilidade e classificação categórica do estado do tempo. As estações da rede foram agrupadas utilizando um algoritmo de agrupamento bipartido, de forma a efetuar a previsão tendo em conta os grupos obtidos que, de acordo com os autores, facilita a previsão. Aos dados foi aplicado um modelo denominado *Similarity Weighted K-Nearest-Neighbor* (SWK) com diferentes variações, bem como diferentes variações de *Artificial Neural Networks* (ANN), utilizando apenas dados das viagens, dados das viagens e meteorológicos, dados das viagens e táxis e a junção dos três conjuntos. Analisando os valores do erro e do erro médio, constatou-se

que os modelos alterados a partir do SWK obtiveram melhor desempenho relativamente ao original. No que diz respeito aos modelos ANN, estes apresentam uma melhor desempenho no geral, sendo que os piores resultados são obtidos recorrendo apenas aos dados das viagens e os melhores resultados foram obtidos quando utilizada a informação relativa às viagens, táxis e meteorologia, indicando que, quanto mais informação se utilizar melhor a precisão das previsões. A CitiBike também foi o [SBP](#) abordado no trabalho [40], tendo sido utilizados o modelo *Ordinary Least Squares linear regression model* (OLS) e a regressão de *Poisson* para prever a procura de bicicletas. Neste estudo foram utilizados dados relativamente ao sistema de bicicletas, dados meteorológicos e informação acerca dos feriados, tendo sido efetuado o agrupamento espacial das estações da rede antes de proceder à elaboração dos modelos. Foi efetuado o agrupamento das estações recorrendo a um novo método baseado em *bottom-up hierarchical clustering* que considera restrições espaciais. Após este processo, os modelos foram implementados nos dados com e sem o agrupamento, tendo sido previsto o número total de bicicletas que serão alugadas na próxima hora. A avaliação dos modelos teve em conta a sua precisão, de acordo com o erro, bem como relativamente ao tempo que demoraram. Analisando os resultados obtidos, concluiu-se que o agrupamento das estações pode melhorar a precisão do modelo. Comparando os modelos utilizados, concluiu-se que o modelo linear apresentou melhores resultados relativamente ao modelo não linear.

No trabalho [44] é efetuada a previsão do fluxo de bicicletas na rede de Nova Iorque, tendo em conta a influência de pontos de interesse, a distância entre estações, o grau de uma estação (o número de ligações entre estações ao longo de um intervalo de tempo) e dados meteorológicos. Para a recolha de pontos de interesse estes foram divididos em grupos de acordo com o tipo em que se enquadram, no caso de serem pontos de restauração, lojas, edifícios governamentais, entre outros. O fluxo de cada estação foi definido como sendo o número total de bicicletas que chegaram a uma estação num dado intervalo de tempo. Para cada estação foi calculada a sua *attractiveness*, que é definida tendo por base o fluxo, o grau da estação e a distância. Se um utilizador estiver disposto a percorrer uma maior distância para se deslocar a uma estação, tal significa que esta é atrativa para esse utilizador. Da mesma forma, se uma estação tiver um fluxo de bicicletas e grau elevados, essa estação também é atrativa. Nesse trabalho foram utilizadas algumas técnicas como *baseline* tais como ARMA, HA e ANN. Foram ainda agrupadas as estações utilizando um algoritmo de *bipartite clustering* tal como o utilizado em [38] para depois ser utilizado o método de *Attractiveness* proposto. Os resultados obtidos mostraram que o modelo proposto apresentou melhor desempenho relativamente aos modelos utilizados como *baseline* e que os resultados após o procedimento de *clustering* foram superiores ao modelo proposto sem o *clustering*.

2.2.4 *Clustering* em Sistemas de Bicicletas Partilhadas

O *clustering* de séries temporais permite eliminar ruído e reduzir a dimensão de conjuntos de dados de grandes dimensões, bem como efetuar a extração de padrões destes dados para possibilitar a análise dos mesmos e a aplicação de algoritmos.

Este envolve três aspetos principais, sendo estes a medição de similaridade, computação de protótipos e o algoritmo de *clustering* utilizado, que normalmente efetua a partição tendo por base as semelhanças encontradas na série temporal, que são calculadas mediante uma métrica de similaridade. O agrupamento de dados tem sido aplicado na temática dos sistemas de bicicletas partilhadas, tanto a nível espacial como temporal, de forma a possibilitar a análise destes sistemas e obter mais informação relativamente aos padrões diários de utilização dos mesmos.

Foram encontrados alguns trabalhos onde são utilizados algoritmos de *clustering* para agrupar as estações dos sistemas de bicicletas partilhadas em análise previamente à elaboração da previsão da procura de bicicletas [38, 40, 44].

No primeiro trabalho [38], as estações da rede da CitiBike em Nova Iorque foram agrupadas recorrendo a um algoritmo de *clustering* bipartido, que agrupa as estações da rede com base na sua localização geográfica e padrões de transição, cujos grupos resultantes se encontram na Figura 2.5. Após esse agrupamento, foram efetuadas as previsões para cada grupo de estações obtido.

As estações da rede da CitiBike em Nova Iorque voltaram a ser agrupados no trabalho

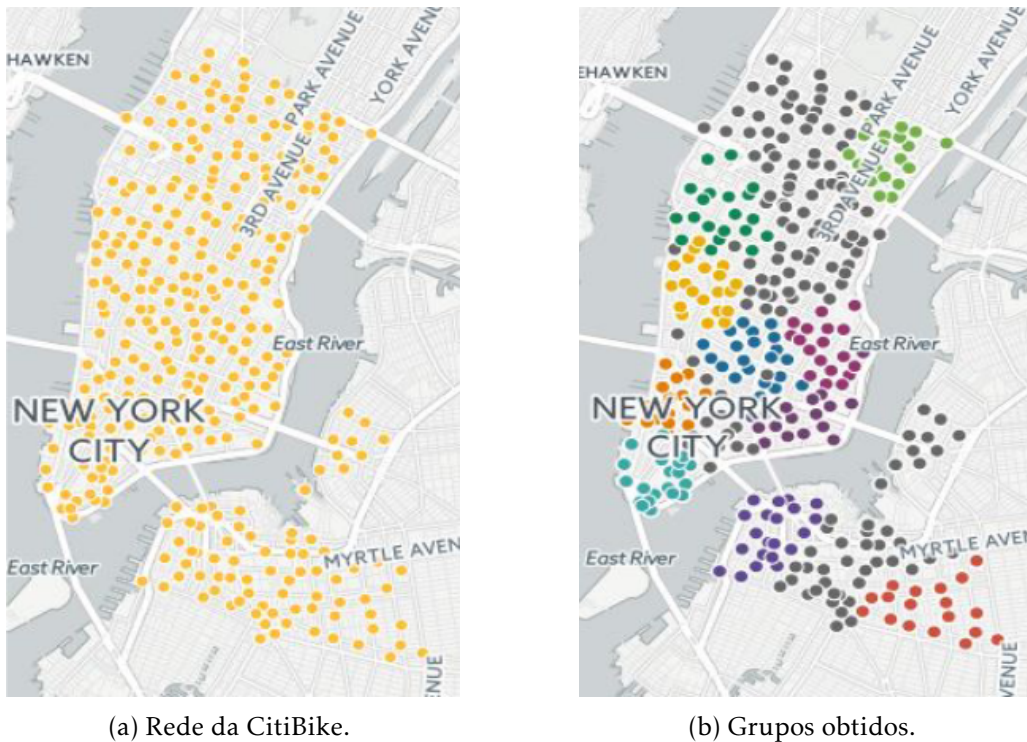
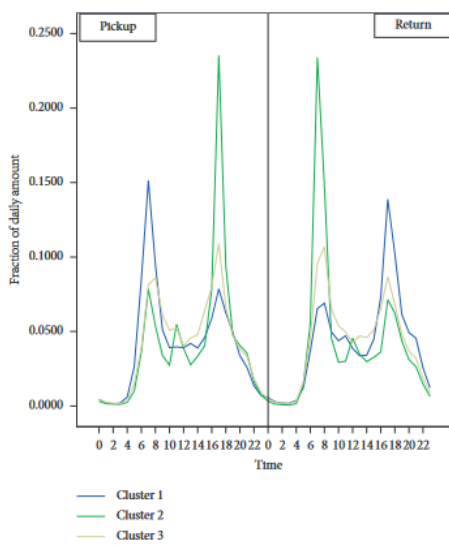


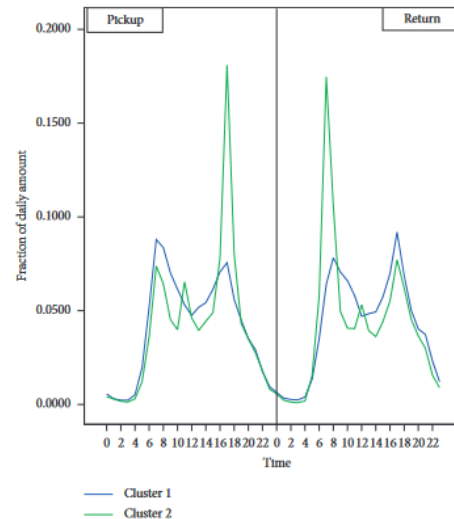
Figura 2.5: Resultados do agrupamento utilizando clustering bipartido em dados da CitiBike, Nova Iorque. Retirado de [38].

[40], desta vez recorrendo a um algoritmo baseado em *bottom-up hierarchical clustering* considerando restrições espaciais, tendo sido utilizadas como *features* a média horária de aluguer de bicicletas de cada estação e a localização da estação. De forma a estandardizar a média horária de aluguer de bicicletas, esta foi dividida pela capacidade de cada estação. As estações foram agrupadas primeiramente, de acordo com as restrições espaciais e depois foi também efetuado o agrupamento tendo apenas em conta os atributos não espaciais. Após o agrupamento, foi efetuada a previsão para cada grupo, mas também para cada estação de forma individual, sendo que um dos objetivos do estudo era comparar a previsão por grupos e por estação. Os resultados permitiram perceber que agrupar as estações resultou numa maior precisão do modelo utilizado neste trabalho.

Em [44] também foram utilizados dados da CitiBike em Nova Iorque e as estações desta rede foram agrupadas recorrendo a um algoritmo de *bipartite clustering* igual ao utilizado em [38]. Neste caso, foram obtidos 23 grupos e, utilizando o modelo de atratividade proposto no trabalho, foi calculado o erro para cada grupo. O modelo proposto teve melhor desempenho para as estações após o agrupamento, relativamente às estações sem este processo. O algoritmo de agrupamento *K-means* foi implementado no trabalho [39], no qual foram utilizados dados das viagens do SBP de Wenzhou, na China. As estações do sistema foram agrupadas de acordo com o rácio de aluguer e devoluções por hora, ou seja, foi efetuado o agrupamento temporal. A análise dos *clusters* apenas incluiu dados em dias que não ocorreu precipitação, tendo sido analisados de forma separada viagens em dias úteis e dias não úteis. Para proceder ao agrupamento, foram utilizados dados com 24 valores, cada um representativo de uma hora, que representavam tanto a média de alugueres e de devoluções, havendo 48 valores no total, que representavam o padrão de utilização nos dias úteis e nos dias não úteis. Foram determinados o valor de



(a) Dias úteis.



(b) Dias não úteis.

Figura 2.6: Resultados do agrupamento utilizando K-means em dados de Wenzhou, na China. Retirado de [39].

k e os centros iniciais tendo em conta a distância euclideana e, após este processo, foram utilizadas a coesão e separação para validar os grupos obtidos. A coesão representa o grau de similaridade entre os pontos que compõem um grupo e a separação indica o grau de distinção dos pontos em grupos diferentes. Tendo em conta a métrica de validação escolhida, foi determinado o número ótimo de grupos (valor de k) a selecionar. Com base nos resultados obtidos o número de k selecionado para os dias úteis foi o 3 e, para os dias não úteis, o 2. Os grupos obtidos podem ser visualizados na Figura 2.6.

No trabalho de Li. et al [23], recorrendo a dados relativos às viagens de um SBP em Chicago, é efetuada a deteção de padrões e são propostos métodos para a representação, redução da dimensionalidade e *clustering* de séries temporais. A representação de séries temporais tem impacto não só na eficiência como também na precisão da solução a que se chega. Os dados do SBP de Chicago utilizados neste trabalho continham informação relativa às viagens, mais concretamente, à estação de início e de fim, bem como a informação temporal relativa ao início e término de cada viagem, relativos a um período de três meses. Vários estudos demonstraram que o padrão de utilização e a procura de bicicletas varia em dias úteis e não úteis, notando-se que há uma maior procura durante a semana e menor nos dias não úteis e, como tal, foram analisados ambos os cenários separadamente. Os dados foram agregados em intervalos de 60 minutos, por ser o mais utilizado nos trabalhos consultados e, de seguida, foi reduzida a dimensão dos dados 4 vezes recorrendo a DWT, significando que, em vez de um dia ser representado por 24 pontos representativos das horas do dia, este é representado por apenas 6 pontos na série transformada. O algoritmo *K-means* foi utilizado para efetuar o agrupamento temporal da procura por hora recorrendo a diferentes partições iniciais, tendo sido medido o desempenho do agrupamento recorrendo ao *Davies-Bouldin index* e escolhidas algumas combinações que foram analisadas posteriormente. O agrupamento com 5 grupos foi analisado de forma mais aprofundada, com alguns grupos a apresentarem uma baixa utilização, outros maior e picos visíveis em horas de ponta. Após este processo, foram testadas outras combinações para os resultados dos 5 grupos, tendo sido utilizados diferentes algoritmos e diferentes técnicas de redução de dimensionalidade. Foram implementados os algoritmos de *clustering* hierárquico, *K-means*, *K-medoids* e *K-shape* e utilizaram-se as técnicas da média, *Partition around medoids* (PAM), distância euclidiana, *DTW barycenter averaging* (DBA) e *shape extraction* (SE). Analisando os resultados obtidos com a métrica de DB, constatou-se que os melhores resultados foram obtidos aquando da utilização de DTW relativamente à distância euclidiana. O agrupamento utilizando DBA apresentou melhores resultados relativamente ao PAM. O algoritmo *K-shape* apresentou o resultado mais baixo, ou seja, o melhor. No entanto, o valor de DB de ambos os algoritmos não difere muito e o *K-shape* demora 5 vezes mais relativamente ao *K-means*, mostrando que o *K-means* é uma solução de baixo custo e eficiente.

2.2.5 Conclusões

A pesquisa efetuada que incidiu, sobretudo, nas temáticas das bicicletas partilhadas, rebalanceamento, *forecasting* de bicicletas e *clustering* em sistemas de bicicletas partilhadas que permitem conhecer melhor os sistemas existentes no mundo, bem como quais as tentativas de rebalanceamento, *forecasting* e *clustering* que estão a ser levadas a cabo nos mesmos.

No que diz respeito às bicicletas partilhadas, os trabalhos relacionados permitiram perceber que estes sistemas têm ido um desenvolvimento ao longo dos anos que levou ao aumento do número de pessoas que utilizam a bicicleta como meio de transporte. Para além disso, foi providenciada uma referência para outros sistemas de bicicletas que pode permitir delinear estratégias de negócio aquando da criação de novos sistemas de bicicletas partilhadas.

Com base nos trabalhos de rebalanceamento consultados, percebe-se que a estratégia mais utilizada para proceder ao mesmo, quando este é dinâmico, passa pela previsão do número de bicicletas nas estações e, só depois destes resultados, se procede à criação de estratégias para levar a cabo o rebalanceamento da rede. É de salientar que a abordagem para tratar os dados provenientes de sistemas bicicletas partilhadas com ênfase na criação de estratégias de rebalanceamento se baseia na construção de digrafos representativos da rede de estações. Nestes, os vértices representam as estações e os arcos representam as ligações entre estas, sendo que o peso dos arcos e os vértices poderão conter informação variada, dependendo do objetivo ou da abordagem levada a cabo num dado artigo.

Os trabalhos que abordam o *forecasting* da procura de bicicletas utilizam, no geral, dados relativos às viagens e dados meteorológicos, havendo ainda um trabalho no qual é utilizada informação sobre pontos de interesse na cidade e ainda informação sobre os utilizadores do sistema de bicicletas partilhadas e acerca da utilização do metro, autocarro e comboio. O processo de análise dos dados, bem como a consequente previsão da procura de bicicletas são processos considerados importantes para auxiliar no processo de rebalanceamento da rede de bicicletas, para o qual se recorreram a várias técnicas como [RF](#), [BPNN](#), [LSTM](#), entre outras. A abordagem que é levada a cabo nos artigos lidos passa pela utilização de algumas técnicas, sendo que uma delas é a proposta e as restantes atuam como uma *baseline* de forma a que seja possível efetuar uma comparação entre os resultados obtidos, mediante os valores de erro das métricas escolhidas para esse propósito. Só após este processo são formuladas conclusões acerca das mesmas.

Em alguns trabalhos consultados, a previsão da procura de bicicletas foi efetuada após um processo de *clustering* que, recorrendo a técnicas distintas, permitiu agrupar não só as estações da rede de bicicletas em análise, como também efetuar o agrupamento temporal em alguns casos, permitindo caracterizar as estações relativamente ao seu padrão de utilização diário. Tal indica que é interessante abordar e aplicar algoritmos de *clustering* neste tipo de dados, pois tal permite encontrar semelhanças entre as estações relativamente a diferentes variáveis, permitindo caracterizar o sistema de bicicletas partilhadas e formular

análises relativamente ao seu comportamento.

2.3 Técnicas de Clustering

Os algoritmos de *clustering* têm vindo a ser utilizados em alguns trabalhos relacionados com os sistemas partilhados de bicicletas, de forma a poder encontrar grupos de estações mediante diversos tipos de informação. Nesta secção serão descritas algumas técnicas de *clustering* abordadas nos trabalhos relacionados no âmbito dos sistemas de bicicletas partilhadas e da previsão da procura de bicicletas e algumas que podem ser interessantes para o caso em estudo nesta dissertação, mesmo não tendo sido abordadas no trabalho relacionado.

2.3.1 Bipartite Clustering

Nos trabalhos [38] e [44], foi utilizado um algoritmo de *clustering* bipartido para agrupar as estações da rede CitiBike em Nova Iorque proposto por [24]. O processo iterativo deste algoritmo, combina dois fatores das estações - localização e transição - sem ser necessário alocar um peso para cada fator. Em cada iteração são repetidos três passos:

- *Geo-clustering*: Neste passo as estações são agrupadas tendo em conta a sua localização, sendo que este passo é conduzido primeiramente para o grupo total de estações que compõem o SBP. Nas restantes iterações, este processo é levado a cabo para as estações em cada grupo obtido após o processo de *T-clustering*.
- Geração da matriz T: Com base nos grupos obtidos a partir do *geo-clustering*, é gerada uma matriz T para cada estação que descreve o seu padrão de transição. Esta matriz é composta por sete linhas, que correspondem a sete intervalos temporais que dividem o dia em vários blocos horários fazendo a distinção entre dias da semana e fins-de-semana e feriados. Cada entrada da matriz tem um número correspondente à probabilidade de uma bicicleta ser adicionada numa das estações do seu grupo, se tiver sido retirada de uma dessas estações no mesmo *time slot*. Esta probabilidade é estimada a partir dos dados das bicicletas.
- *T-clustering*: Após obter um conjunto de T-matrizes, as estações são agrupadas tendo em conta as suas matrizes, utilizando o *K-means*.

Estes três passos são repetidos até os grupos obtidos atingirem a convergência no passo 1 ou se chegue ao threshold K.

2.3.2 Agrupamento Hierárquico

Os métodos de *clustering* hierárquico agrupam os dados numa hierarquia de grupos na qual, com o aumento do nível da hierarquia, os grupos são construídos através da junção de grupos de níveis mais abaixo. Este tipo de agrupamento pode ser aglomerativo ou

divisivo, podendo seguir uma abordagem *bottom-up* ou *top-down*, respetivamente. Neste caso, o algoritmo utilizado no trabalho [38] era uma modificação do tipo de *clustering* hierárquico aglomerativo, que utiliza uma estratégia *bottom-up*, que consiste na formação de um grupo por parte de cada objeto e na consequente junção desses grupos num grupo de maior dimensão, até todos os pontos se encontrarem num único grupo ou as condições para o término do algoritmo serem atingidas. Os agrupamentos são efetuados de acordo com uma medida de similaridade entre ambas as partes e o processo ocorre ao longo de várias iterações, sendo que em cada uma ocorre a junção de dois grupos, onde cada grupo contém, no mínimo, um objeto, significando que existem no máximo n iterações, com o número n a representar o número total de objetos a agrupar [13].

No trabalho [38], o algoritmo de *clustering* utilizado foi baseado em *clustering* hierárquico com uma estratégia *bottom-up*, tendo sido proposto em [3]. Este algoritmo de *clustering* hierárquico utiliza duas matrizes de dissimilaridade com restrições geográficas, sendo que no caso do sistema de bicicletas partilhadas foi utilizada a média horária do número de alugueres e a localização das estações. O algoritmo proposto utiliza duas matrizes de dissimilaridade e um parâmetro de *mixing*. As matrizes de dissimilaridade podem ser ou não distâncias euclidianas e o peso atribuído às observações pode ser não uniforme. A primeira matriz apresenta as dessemelhanças das *features* e a segunda apresenta as dessemelhanças das restrições utilizadas. A parâmetro de *mixing* indica a importância das restrições no processo de agrupamento. O critério minimizado em cada fase é a combinação do critério de homogeneidade calculado para ambas as matrizes. O valor do parâmetro de *mixing* determinado deverá aumentar a contiguidade espacial sem deteriorar a qualidade da solução.[3]

2.3.3 K-means

Este algoritmo foi proposto nos anos 60 e ainda é um dos mais utilizados para agrupamento, pois é fácil de implementar, simples e eficiente, permitindo obter bons resultados. Tendo um conjunto de dados com pontos de n dimensões, o *K-means* agrupa o mesmo em K grupos, encontrando uma partição de forma a que o erro quadrado entre a média empírica de um grupo e os pontos constituintes de outro grupo seja minimizado. O objetivo é minimizar a soma do erro quadrado em todos os K grupos [21].

Os três passos principais deste algoritmo são os seguintes: selecionar uma partição inicial com K grupos, repetir os passos seguintes até a qualidade dos grupos estabilizar; gerar uma nova partição atribuindo cada padrão ao centro mais próximo; calcular os novos centróides. Este algoritmo requer 3 parâmetros: o número de *clusters* (K), a inicialização e a métrica de distância, sendo que o parâmetro mais crítico é o número de *clusters*. Normalmente, o algoritmo é computado para diferentes valores de K , sendo selecionada a partição mais adequada. Diferentes inicializações podem dar origem a diferentes grupos no final do processo, pois este algoritmo apenas converge para um mínimo local, podendo-se ultrapassar este problema correndo o algoritmo para um valor de K com

diferentes partições iniciais, escolhendo aquela com o menor erro quadrado. Uma vez que se costuma utilizar a métrica Euclidiana para computar a distância entre pontos e centróides, o *K-means* tende a encontrar grupos de forma circular nos dados [21].

Para selecionar o valor de K adequado, a heurística mais utilizada é o método do cotovelo, que consiste em calcular a soma do quadrado dos pontos de cada grupo e a distância média. Quando o valor de k é igual a 1, este valor será maior e tende a descer com o aumento do número de grupos considerado. Tendo por base um gráfico para cada valor de k e para os valores da soma do quadrado de cada grupo, procede-se à deteção do cotovelo, que será a zona em que este valor diminuiu mais.

Para avaliar os resultados obtidos, também é muito utilizado o método do *Silhouette Score*, que se baseia na distância média dentro de cada grupo e na distância média entre grupos, podendo apresentar valores entre -1 e 1, sendo que quanto mais próximo o valor for de -1, pior é a qualidade dos grupos obtidos e quando mais próximo de 1, melhor a qualidade dos mesmos, significando que se encontram bem separados. Um valor próximo de 0 significa que os grupos obtidos não são significativos. Outras métricas de validação como a de Davies-Bouldien e S_Dbw também costumam ser utilizadas. A primeira, utilizada em [23], consiste no cálculo da média da similaridade de cada grupo com o grupo que é mais semelhante a este. Quanto menor for o valor desta métrica, melhor é a qualidade dos grupos obtidos, melhor é a separação entre eles e melhor é o resultado. A métrica S_Dbw [39] tem em conta a variância dentro de cada grupo (intra-cluster) e a densidade entre grupos, ou seja inter-cluster. Uma variância baixa indica que os grupos são compactos, uma densidade baixa indica que uma boa separação dos grupos e, consequentemente, um valor baixo da métrica S_Dbw indica um bom resultado.

2.4 Técnicas de Forecasting

O *forecasting* ou previsão de séries temporais é uma técnica analítica preditiva através da qual se prevêem os valores futuros de uma série temporal tendo por base os seus valores passados, bem como outros inputs adicionais que podem ter influência na variável que está a ser prevista [22].

Os trabalhos relacionados relativos consultados acerca dos sistemas de bicicletas partilhadas abordam a análise e previsão de séries temporais, com o objetivo de detetar padrões de utilização e efetuar a previsão da procura de bicicletas na rede, tendo sido conduzidas experiências nos mesmos envolvendo vários tipos de técnicas de aprendizagem automática, entre outras.

Nesta secção serão abordadas algumas técnicas de *forecasting* que foram utilizadas nos trabalhos relacionados consultados, que podem ter interesse para este trabalho.

2.4.1 Métodos de Smoothing

Neste tipo de métodos, a observação prevista é conseguida através de uma função das observações passadas, tendo em conta a granularidade temporal do conjunto de dados utilizado. Para tal, é necessário utilizar a noção de um número representativo do período de tempo a considerar, que significará o número de observações que será utilizado para calcular o valor previsto tendo em conta a série temporal em análise [22].

O *naïve forecasting* é um dos exemplos de uma técnica de *smoothing*, na qual o valor previsto é igual ao valor anterior da série temporal. Esta técnica pode ser adaptada para ter em conta a sazonalidade do conjunto de dados em casos em que os dados existentes são relativos a vários anos, sendo possível prever, por exemplo, um valor relativo ao mês de julho tendo em conta o valor do mês de julho do ano anterior.

Outra técnica é a da média móvel, que consiste em calcular a média de um número de observações passadas para obter a previsão, havendo ainda a possibilidade de adicionar um valor de peso em cada uma das observações, que permite que as observações mais recentes contribuam mais para a previsão relativamente às observações mais antigas, o que é denominado de *Weighted Moving Average* [22].

2.4.2 Long Short Term Memory (LSTM)

Uma rede neuronal recorrente (RNN) permite ter em conta a ordem de inputs, sendo adequado para a utilização em dados de séries temporais, nos quais a ordem dos dados tem uma grande importância.

Numa RNN, o output da rede depende não só da informação que lhe está a ser apresentada, mas também da informação anterior disponível na mesma, ou seja, existe uma memória dos inputs anteriores que persiste no estado interno da rede e influencia o seu output. Em cada passo, a informação da camada atual e da anterior é processada e atualiza o seu estado mediante uma função de ativação, efetuando uma previsão. Este tipo de rede neuronal apresenta alguns problemas, tais como o *vanishing gradient*, que ocorre quando os valores de um gradiente são demasiado baixos, o que apresenta como consequência o modelo parar a aprendizagem ou demorar muito tempo a aprender, o que costuma ocorrer quando existem muitos *timesteps*. O outro problema é o do *exploding gradient*, que ocorre quando o algoritmo atribui muita importância aos pesos utilizados [4].

Estes problemas foram solucionados com a criação da rede neuronal Long Short Term Memory (LSTM), que é um tipo de rede neuronal recorrente (RNN), proposta por Hochreiter e Schmidhuber em 1997 [14] que, tendo por base gradientes, é uma forma mais rápida para aprendizagem de maior quantidade de informação num curto espaço de tempo.

A arquitetura deste tipo de redes é constituída por células de memória (Figura 2.7), que são compostas por um *input gate*, um *output gate* e um *forget gate*, que permitem selecionar que informação entra na célula, sendo compostas por uma operação de multiplicação e uma função sigmóide que imprime números entre zero e um, que descrevem quanto da informação deve ser passada, com o zero a significar que nada deve passar e um valor

igual a um a significar que toda a informação deve ser passada para a célula [4, 14]. Mais

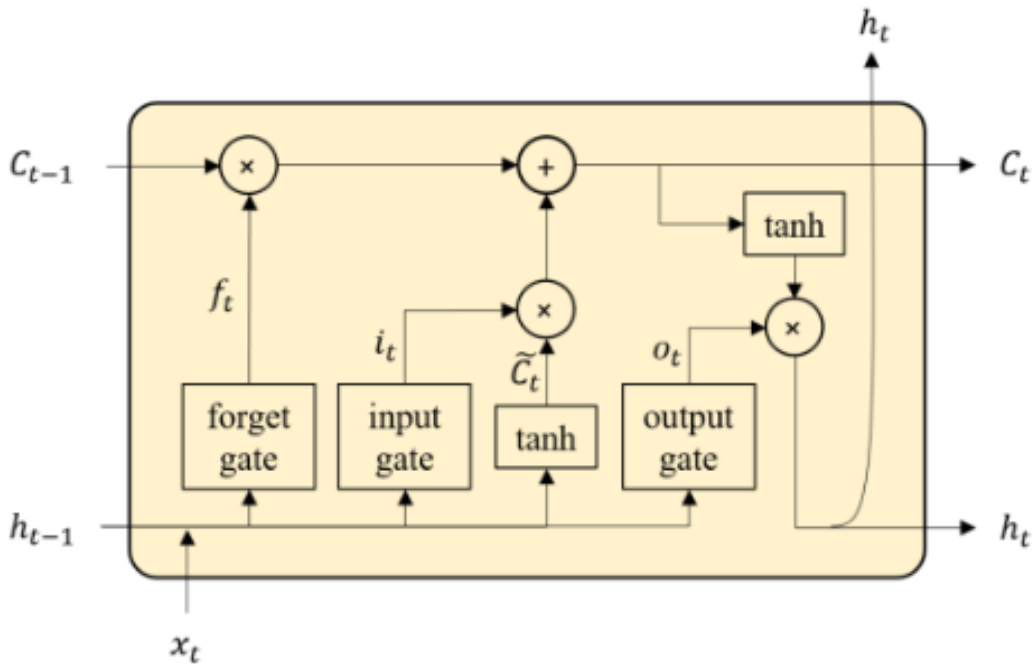


Figura 2.7: Arquitetura de uma célula LSTM. Retirado de [43].

concretamente, o *forget gate* escolhe se a informação que entra na célula é relevante e deve ser lembrada, ou se é irrelevante, sendo que nesse caso deve ser esquecida. O *input gate* permite que a informação entre na célula e que a informação seja aprendida e, por fim, o *output gate* passa a informação atualizada do *timestamp* atual para o seguinte [43].

2.4.3 Random Forest (RF)

O algoritmo *Random Forest* é um classificador que consiste numa coleção de estruturas em forma de árvore, que pode ser utilizado para classificação ou regressão. Esta coleção de árvores atuam como classificadores e são independentes e distribuídas de forma idêntica, em que cada árvore representa um voto imparcial na classe mais popular num dado input [1].

Uma árvore de decisão é composta por três componentes principais: os nós de decisão, nós folha e o nó de raiz. O conjunto de dados de treino é dividido em ramos e esta sequência continua até se obter um nó folha. Estes nós representam os atributos que são utilizados na previsão. Após a construção das árvores de classificação tendo por base o conjunto de dados original, uma árvore final de decisão é construída tendo por base a relevância dos preditores, que pode ser utilizada para prever as observações.

No que diz respeito à utilização de *Random Forest* para regressão, estas são formadas através da construção de árvores dependendo de um vetor aleatório de forma a que o preditor da árvore toma valores numéricos em vez de etiquetas de classe. Os valores que se

obtem das árvores são numéricos e é necessário assumir que as observações do conjunto de dados para treino é independente da distribuição do vetor aleatório. O preditor *Random Forest* é formado a partir da média das árvores criadas e os requisitos para que a regressão seja precisa é a existência de uma baixa correlação entre os resíduos e as árvores com um erro de generalização baixo [1].

Esta técnica é adequada e útil para previsão e, devido à Lei dos números grandes, é robusta no que diz respeito ao *overfitting*, pois este não ocorre e é diminuído com o aumento do número de árvores. É ainda possível utilizar este algoritmo em conjuntos de dados de grande dimensão com rapidez. No entanto, as observações utilizadas devem ser independentes, o que é uma desvantagem [1].

2.5 Avaliação da Precisão do *Forecasting*

Quando é efetuada a previsão utilizando modelos de *forecasting*, é necessário avaliar esses modelos e aferir a precisão dos seus resultados, o que é conseguido através do cálculo de métricas que permitem não só retirar conclusões acerca do modelo, como também comparar diferentes modelos.

De forma a evitar ter de esperar para confirmar os dados (visto que a previsão é feita para valores futuros), costuma reservar-se uma porção dos dados como um conjunto de validação, com o objetivo de testar a precisão do modelo sem o expor a dados de teste, através do cálculo do erro, que é dado pela diferença entre o valor real e o valor previsto.

Na literatura encontram-se alguns exemplos da utilização de métricas como [MAE](#) e [RMSE](#) na avaliação dos métodos utilizados no *forecasting* da procura em [SBP](#) como em [40], que leva a cabo a previsão da procura a curto-prazo da procura de bicicletas utilizando algoritmos de aprendizagem automática, no qual se recorreu à métrica [MAE](#) para avaliar a precisão dos modelos preditivos.

Em [38] utilizam-se tanto a taxa de erro (ER), como a média dos erros na avaliação do desempenho. Para avaliar um método que utiliza redes neuronais [43], foram utilizados dois tipos de métricas na validação: [RMSE](#) e [MAE](#), sendo que Ruffieux et al. [31] também utilizou [RMSE](#) para validar as técnicas de aprendizagem automática usadas na previsão da utilização de [SBP](#). De forma a prever a procura ao nível das estações num [SBP](#) [4], efetuou-se a avaliação do modelo proposto recorrendo ao [MAE](#), [Mean Squared Error \(MSE\)](#), [Mean Absolute Scaled Error \(MASE\)](#), [MAPE](#), [RMSE](#) e [RMSLE](#), que utiliza o logaritmo dos valores, sendo pouco afetado por *outliers*, sendo que esta última métrica é também utilizada em [44] juntamente com a taxa de erro.

De seguida, serão descritas as métricas utilizadas nos trabalhos consultados para quantificar a precisão dos resultados dos modelos preditivos [22].

Média do Erro Absoluto ([MAE](#))

Esta métrica mede a média da magnitude dos erros num conjunto de previsões.

Após obter o valor do erro, que é dado fazendo a subtração do valor previsto ao valor real, somam-se todos os valores para o erro e dividem-se pelo total de previsões efetuadas como se observa na equação 2.1.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.1)$$

Média do Erro Quadrado (MSE)

Esta métrica mede a média dos quadrados dos erros, ou seja, a média do quadrado da diferença entre os valores previstos e os valores reais. Uma vez que se baseia no quadrado da distância euclidiana, o seu valor é sempre superior a zero, com o erro a decrescer conforme se aproxima do zero. O MSE tem as mesmas unidades que o quadrado da quantidade que está a ser estimada.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2.2)$$

Raiz Quadrada do Erro Médio (RMSE)

Esta métrica calcula a média do erro à semelhança do MSE, no entanto, o somatório das diferenças entre os valores previstos e os reais é elevado ao quadrado e, após a média ser obtida, é efetuada a raiz quadrada desse valor.

Esta métrica é muito utilizada na literatura, mas é muito afetada pela existência de *outliers*, penalizando modelos que apresentam erros baixos e altos comparativamente a outro modelo que possa apresentar um valor de erro mais constante. O RMSE encontra-se nas mesmas unidades que a quantidade que está a ser estimada. A equação 2.2 mostra a fórmula para calcular este erro.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.3)$$

Mean Absolute Percentage Error (MAPE)

A média absoluta da percentagem do erro é uma métrica que avalia a precisão de um modelo de acordo com a percentagem do erro absoluto. O MAPE é útil para efetuar a comparação entre diferentes modelos preditivos por ser independente no que diz respeito à sua escala. No entanto, este apresenta uma desvantagem em conjuntos de dados onde valores iguais a zero ocorrem, o que inclui as séries temporais, uma vez que, neste caso, este dá origem um valor infinito do erro.

Na fórmula seguinte, o y_i é representativo do valor real e o $\hat{y}_i$ representa o valor previsto. Após efetuar a diferença entre estes valores, o resultado é dividido pelo valor

real. O valor absoluto obtido é somado para todos os pontos previstos e dividido pelo número total de valores (n).

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.4)$$

Root Mean Squared Logarithmic Error (**RMSLE**)

Esta métrica é o logaritmo da raiz quadrada do erro médio é uma modificação do **RMSE**, na medida em que é efetuada a diferença do logaritmo entre o valor previsto ($\hat{y}_i$) e o valor real (y_i). Esta diferença faz com que esta métrica seja pouco afetada pela existência de *outliers*, pois o valor é drasticamente reduzido devido ao logaritmo.

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(\hat{y}_i + 1) - \log(y_i + 1))^2} \quad (2.5)$$

Mean Absolute Scaled Error (**MASE**)

Esta métrica é independente da escala e permite ultrapassar a desvantagem do **MAPE** comparando os valores previstos com o *naïve forecast*, onde o próximo ponto dos dados apresenta o mesmo valor que o anterior. Este erro é menor que um se o valor previsto for melhor do que o *naïve forecast* e superior a um se for pior. A fórmula deste tipo de erro é a seguinte:

$$MASE = \frac{\frac{1}{n} \sum_{i=1}^T |y_i - \hat{y}_i|}{\frac{1}{T-1} \sum_{t=2}^T |Y_t - Y_{t-1}|} \quad (2.6)$$

ABORDAGEM E METODOLOGIA

Neste capítulo será descrita a abordagem levada a cabo ao longo desta dissertação de forma a poder realizar as contribuições previstas e atingir os objetivos pretendidos.

Para além da descrição da abordagem, referindo conclusões retiradas da análise de dados, serão ainda enumeradas algumas tecnologias que foram utilizadas ao longo do desenvolvimento da dissertação, quer no que diz respeito à análise exploratória de dados, quer na parte da previsão.

3.1 CRISP-DM

A abordagem definida para esta dissertação teve como base a metodologia [Cross Industry Standard Process for Data Mining \(CRISP-DM\)](#), um modelo que é atualmente o mais utilizado para mineração de dados, constituído por um conjunto de passos que permitem planear, organizar e implementar projectos de *data science*.[\[18\]](#)

Esta metodologia foi a escolhida por ser a mais utilizada para a mineração de dados, por ser possível modificar a sequência pela qual se leva a cabo cada fase e também pela flexibilidade inerente ao processo, que permite visitar uma parte do mesmo, conforme necessário. As seis fases do modelo [CRISP-DM](#), que representam o ciclo de vida de um projeto de *data science*, encontram-se na figura [3.1](#). Como referido anteriormente, a sua sequência pode ser alterada ao longo da realização do projeto, no entanto, as setas presentes no diagrama indicam e serão descritas de seguida.

Entendimento do Negócio

Primeiramente, deve-se entender o que a organização espera atingir com base no processo de mineração de dados, sendo ideal discutir este tópico com as pessoas de interesse, de forma a poder perceber e determinar quais as razões pelas quais o projeto é necessário,

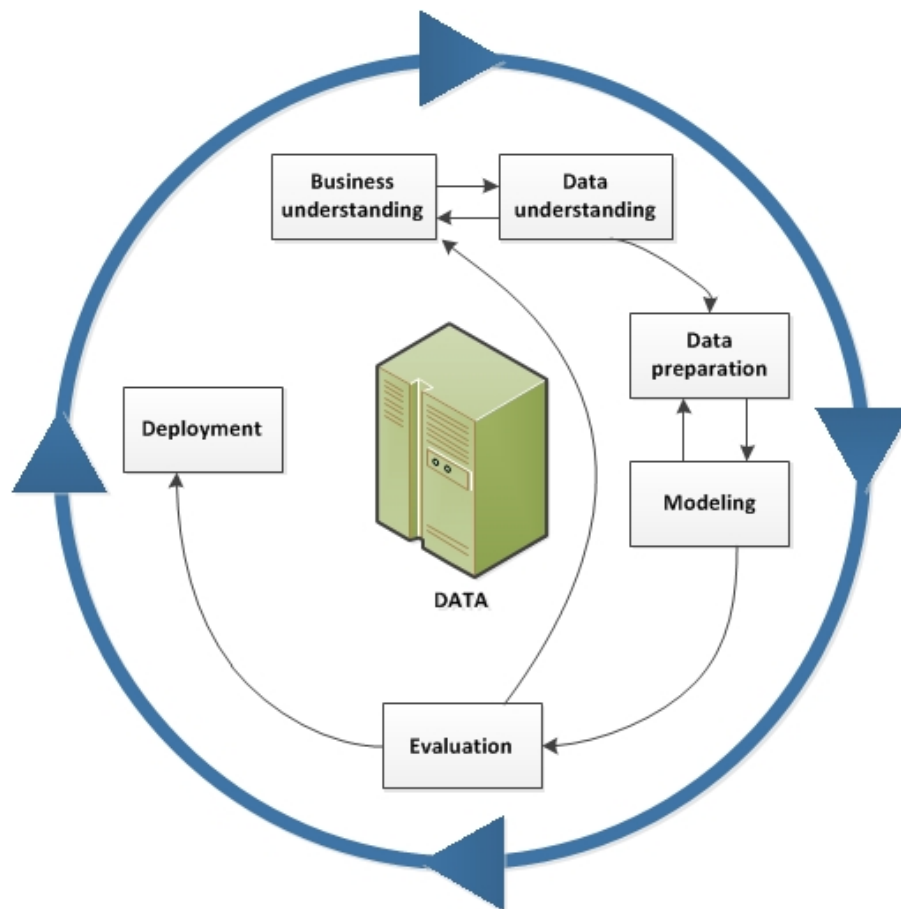


Figura 3.1: Fases do processo CRISP-DM. Retirado de [18].

ou seja, quais os objetivos e requerimentos do mesmo.

Para tal, é necessário determinar quais os objetivos que se pretende atingir após a mineração dos dados, o que é levado a cabo recolhendo informação acerca das características do negócio de forma a poder determinar a organização estrutural, descrever o problema que se pretende resolver e a solução atual que está a ser utilizada para a resolução desse problema, bem como as suas vantagens e desvantagens.

Após este processo, é feita uma avaliação da situação, nomeadamente dos riscos e contingências, bem como a disponibilidade de recursos e, finalmente, selecionam-se as tecnologias e ferramentas a utilizar. [15].

Entendimento dos Dados

A fase de entendimento dos dados envolve analisar os dados que estão disponíveis no âmbito do projeto, sendo importante para evitar problemas durante as fases seguintes do projeto. A análise que deverá ser realizada inclui a exploração dos dados utilizando gráficos, tabelas, ou outras visualizações, de forma a determinar a sua qualidade e tomar decisões relacionadas com os dados nas fases seguintes [16].

Este processo ocorre ao longo de quatro passos principais, que serão descritos de seguida:

1. Recolha dos dados

Nesta fase, pretende-se aceder aos dados, que podem ser conjuntos de dados já existentes, que foram cedidos por alguma organização, dados que foram comprados ou dados adicionais, no caso de os dados já existentes não serem suficientes para corresponder ou atingir os objetivos delineados na fase inicial.

2. Descrição dos dados

A descrição dos dados deve basear-se na quantidade e qualidade dos dados, avaliando os dados disponíveis e em que estado estes se encontram. Deve ser efetuada a descrição da quantidade dos dados, especificamente o número de registos e dados estatísticos dos mesmos, quais os atributos de cada conjunto de dados utilizado, qual o tipo de dados (que pode ser numérico, booleano, categórico, entre outros).

3. Exploração dos dados

Após a descrição dos dados, estes deverão ser explorados através da utilização de técnicas de visualização de dados, como gráficos, tabelas, ou outro tipo de visualização, de forma a identificar possíveis relações entre atributos, ou retirar conclusões acerca das transformações que poderão ser efetuadas na próxima fase.

4. Verificação da qualidade dos dados

Esta fase tem como objetivo detetar inconsistências nos dados, mais concretamente, valores em falta, tais como valores em branco, valores nulos ou valores incorretos.

Ao longo desta fase, todas as informações acerca dos dados deverão ser registadas num relatório de qualidade de dados, onde conste não só a descrição dos mesmos, bem como a exploração efetuada e consequentes inconsistências encontradas nos mesmos, de forma a perceber que tipo de operações é necessário efetuar para que os dados possam ser utilizados futuramente.

Preparação dos Dados

Esta fase é a mais importante e a que consome mais tempo em todo o processo, podendo ser constituída pelos passos seguintes:

1. Seleção dos dados

São selecionados as linhas e atributos que se pretende incluir no conjunto de dados final.

2. Limpeza dos dados

Tendo em conta o relatório de qualidade de dados, deverão ser efetuadas operações nos dados para remover linhas ou atributos no caso de existir dados em falta, ou,

caso seja possível, poderão ser preenchidos os dados com valores estimados. No caso de dados incorretos, esses deverão ser substituídos ou corrigidos.

3. Construção de dados

Por vezes, pode ser necessário proceder-se à construção de novos dados, ou até vantajoso para poder realizar análises numa fase futura do processo, o que pode ser feito derivando novos atributos ou através da criação de novas linhas nos dados.

4. Integração de dados

A integração de dados pode ser feita através da junção de dois conjuntos de dados com registos semelhantes e atributos diferentes ou através da união de dados com atributos semelhantes mas registos diferentes.

5. Formatação de dados

Este passo pode ser necessário para que os dados fiquem no formato adequado para algum algoritmo, ou caso seja necessário ordenar os dados de alguma forma. Deverá ser tido em conta o tipo de modelos ou algoritmos que se pretende utilizar e se esses requerem uma formatação prévia dos dados para que possam ser implementados.

Modelação

A fase de modelação necessita normalmente, de múltiplas iterações, consistindo na implementação de vários modelos com diversos parâmetros e na otimização dos mesmos. Poderá ser necessário repetir o passo de formatação para corresponder às necessidades dos modelos a utilizar. Os modelos deverão ser avaliados para averiguar se é necessário repetir ou experimentar modelos diferentes [19].

Avaliação

Os modelos desenvolvidos deverão ser avaliados, tendo em conta não só os seus resultados mas também os objetivos que foram definidos na fase de entendimento do negócio [20].

Desenvolvimento

Na última fase do processo planeia-se e monitoriza-se o desenvolvimento dos resultados, completando o processo ou produzindo um relatório ou documento final [17].

No caso da presente dissertação, o processo **CRISP-DM** foi ligeiramente modificado de forma a ser aplicado ao objetivo final, que é a criação dos modelos preditivos. Primeiramente, a fase de entendimento do negócio consistiu na realização de uma reunião com a **CML** de forma a poder obter informações acerca da solução atual para o problema do rebalanceamento, dos dados que foram facultados e dos objetivos pretendidos. A solução

de rebalanceamento utilizada pela EMEL tem por base o conhecimento e experiência dos funcionários. O processo de rebalanceamento encontra-se em funcionamento durante todo o dia, não havendo um período preferencial para o mesmo, nem quantidades de bicicletas específicas que são colocadas ou retiradas das estações. Para proceder ao rebalanceamento, são utilizadas viaturas, mais concretamente 5 veículos com capacidade para 6 bicicletas e 3 reboques que podem transportar 12 bicicletas cada. Após esta fase, foram formuladas algumas questões às quais se pretende responder ao longo do processo de desenvolvimento desta dissertação:

- Existem grupos espaciais de estações em Lisboa?
- Existem estações com um comportamento temporal semelhante?
- Qual o comportamento diário das estações tendo em conta o seu padrão de viagens (partidas e chegadas)?
- Qual o comportamento diário das estações tendo em conta o seu padrão de bicicletas?
- É possível detetar o rebalanceamento efetuado pelos funcionários nos dados?
- Existem estações na quais se nota uma maior influência de rebalanceamento?
- Como é o comportamento de ocupação e desocupação das estações nos dias de semana e em fins-de-semana e feriados?
- Qual o impacto da meteorologia na utilização da rede da Gira, ou seja, no número de viagens efetuadas?
- Quantas bicicletas estão em cada estação ao longo do tempo?
- Quando não ocorrem viagens na rede, tal é devido a não haverem clientes ou ao facto de as estações se encontrarem cheias/vazias?

Após a formulação de algumas perguntas acerca dos dados facultados, é necessário entender os dados, o que consiste, primeiramente, na definição dos conjuntos de dados que vão ser utilizados no âmbito do projeto. Neste caso, foram fornecidos dados relativos às viagens ocorridas na rede da Gira, bem como dados relativos às estações com informação relativa ao número de bicicletas que se encontravam em cada estação e ao número de docas (capacidade) de cada uma. Para além disso, foram facultados dados meteorológicos de precipitação, intensidade do vento, radiação e temperatura, sendo que serão ainda criados novos dados para auxiliar à criação de modelos preditivos, como informação relativa aos dias da semana, fins-de-semana e feriados, bem como informação relativa à hora do dia.

Depois da recolha dos dados, foi necessário efetuar a descrição, exploração e verificação da qualidade da sua qualidade, tendo resultado desta análise um relatório de

qualidade de dados das viagens e estações, que pode ser consultado no Apêndice A, bem como um relatório de análise da qualidade dos dados meteorológicos, que se encontra no Apêndice C. Com base nas conclusões retiradas da análise da qualidade de dados, foi necessário efetuar a preparação dos dados, que incluiu a seleção, formatação e limpeza dos dados que, no caso dos dados da viagens e estações se encontra no Apêndice B e no caso dos dados meteorológicos se encontra no Apêndice C. Os conjuntos de dados mencionados serão apresentados na secção 4.1, onde é efetuada uma breve descrição de cada um, bem como um resumo dos problemas encontrados e consequentes operações de processamento que foram efetuadas.

A fase de modelação dividiu-se em etapas diferentes, nas quais não só foram levadas a cabo análises sobre os dados, para as quais foram elaborados *dashboards*, como também foram desenvolvidos modelos preditivos, cujo processo se encontra descrito nas secções 3.2 e 3.3, respetivamente. Para auxiliar o processo de desenvolvimento dos *dashboards*, foi desenvolvida, com outros colegas, uma linguagem gráfica denominada *Interactive Visualization Modeling Language (IVML)* cujo objetivo é planear e documentar o processo de criação de um *dashboard*, não só no que diz respeito às visualizações que se pretendem criar, mas também no que diz respeito às interações e *tooltips* existentes no mesmo.

A análise dos resultados obtidos utilizando os modelos efetuados, foi efetuada através do cálculo de métricas das previsões para cada estação, tendo sido desenvolvido um *dashboard* para facilitar a sua análise.

3.2 Estudo de Padrões das Viagens e Estações

Nesta fase pretendeu-se determinar a existência de grupos de estações tendo por base diferentes atributos, tais como apenas a localização das estações, o número de docas e a localização, a média do rácio ao longo do dia num intervalo temporal horário, as partidas médias ao longo das horas do dia e ainda as chegadas médias ao longo do dia, de forma a tentar encontrar resposta para as questões que se seguem:

- Existem grupos espaciais de estações em Lisboa?
- Existem grupos que relacionem a localização e o número de docas das estações?
- Existem grupos temporais de estações, tendo em conta o seu rácio médio diário?
- Qual o comportamento diário das estações tendo em conta o seu padrão de viagens (partidas e chegadas)?

A resposta a estas perguntas irá permitir perceber se existem grupos de estações com semelhança no número de docas e proximidade entre si, ou se existem estações cujo número de docas as diferencie mesmo que se encontrem mais distantes geograficamente, ou se os grupos de estações tendo em conta a localização coincidem com os grupos existentes tendo em conta o identificador de cada estação. Adicionalmente, utilizando tanto

o rácio médio ao longo do dia, bem como o valor médio de partidas e chegadas para o agrupamento, é possível ter um perfil diário da ocupação das estações, bem como, no caso das viagens (partidas e chegadas), da utilização da rede, o que permite compreender se existem estações com um comportamento diário semelhante e caracterizar a dinâmica diária da rede de estações da Gira.

Para proceder ao agrupamento de estações, o algoritmo *K-means* foi o utilizado, uma vez que também foi utilizado em alguns dos trabalhos consultados [32, 39]. Foi ainda necessário formatar os dados para que se encontrem no formato pretendido para o algoritmo e para que tenham a informação necessária. Para analisar os resultados do agrupamento, estes serão observados tendo em conta a vertente geográfica, bem como a vertente temporal, quando aplicável, de forma a poder associar, por exemplo, padrões diários de rácio a certas localizações da cidade de Lisboa, bem como o padrão diário de partidas e chegadas, permitindo perceber se existe um fluxo diário tendo em conta a altura do dia e as estações da rede.

Recorrendo à criação de *dashboards*, pretende-se ainda analisar os padrões de comportamento diários das estações, não só relativamente ao número de bicicletas, mas também às viagens, de forma a perceber o padrão de ocupação e desocupação das mesmas e o padrão de viagens ao longo do dia. Desta forma, pode-se averiguar se existem padrões similares entre as estações, bem como proceder à identificação das estações que apresentam de forma sistemática falta de docas livres ou de bicicletas e onde estas se encontram localizadas.

Uma vez que os dados relativos ao número de bicicletas nas estações incluem o processo de rebalanceamento, ou seja, a remoção e adição de bicicletas nas estações encontra-se nos dados, é importante tentar perceber se é possível detetar esse processo, pois a sua presença nos dados e, no caso positivo, se é possível removê-lo dos dados, de forma a que estes não se encontrem adulterados e reflitam a verdadeira necessidade de rebalanceamento, o que é imperativo para proceder à elaboração de modelos preditivos, uma vez que o objetivo dos modelos é permitir determinar e detetar a necessidade de rebalanceamento nas estações da rede.

3.3 Elaboração de Modelos Preditivos

A predição da procura de bicicletas irá focar-se principalmente na elaboração de um modelo preditor para cada estação. Após efetuar a análise da deteção do rebalanceamento, mediante os resultados obtidos foi necessário proceder à criação de novos dados sem a influência do rebalanceamento que, quando presente, significa que qualquer previsão levada a cabo utilizando os dados em causa iria incluir essa influência, não irá ser fidedigna, uma vez que nunca iria ter em conta o comportamento real das estações. Os novos dados criados foram elaborados recorrendo aos dados das viagens, nos quais o rebalanceamento não se encontra presente, visto estes conterem informação das viagens efetuadas na rede, tendo sido necessário utilizar em simultâneo informação relativamente às partidas e às

chegadas ao longo do tempo, para cada estação.

Utilizando estes dados, pretende-se desenvolver um modelo preditivo para cada estação a partir da implementação de alguns algoritmos, nomeadamente *Naïve Forecasting*, que servirá como um *benchmark* para a posterior avaliação de resultados, a Média Móvel e vários tipos de *LSTM*, nomeadamente univariável e multivariável [43], sendo que o primeiro método utiliza apenas os dados relativos às bicicletas e o segundo utiliza variáveis adicionais, o que costuma produzir melhores resultados. Para a previsão multivariável serão utilizados dados meteorológicos, informação acerca de fins-de-semana e feriados, acerca da hora e do dia da semana. Estas previsões terão ainda outro fator que será variável, a componente temporal, pois serão realizadas para diferentes intervalos temporais como, por exemplo, para um horizonte de quinze minutos, uma hora e três horas. A escolha destes períodos temporais tem como objetivo não só a comparação dos resultados através de métricas, como também a realização da previsão de acordo com um horizonte temporal previamente discutido em reunião com a EMEL, na qual um horizonte de três horas foi indicado como sendo útil para a operação.

Para proceder à elaboração destes modelos, será efetuado primeiro um período de validação, no qual serão experimentados os algoritmos com parâmetros diferentes, de forma a perceber quais os melhores parâmetros a utilizar para cada uma das 74 estações. Após a análise dos resultados será selecionado, para cada estação, o melhor modelo, ou seja, aquele que apresentou os melhores resultados. A utilização de mais do que um algoritmo tem como objetivo efetuar um estudo comparativo dos resultados obtidos, cuja metodologia será descrita de seguida.

3.4 Avaliação de Resultados

Esta fase será recorrente ao longo do processo da elaboração dos modelos preditivos, tendo como objetivo avaliar a precisão dos modelos escolhidos para a sua realização.

Na literatura utilizam principalmente as métricas *MAE*, *RMSE* e *MSE* [4, 31, 43, 44], sendo essas as que serão selecionadas para este trabalho. A seleção destas métricas advém do facto de estas serem as mais utilizadas no trabalho relacionado encontrado e descrito no capítulo 2. Poderão haver mudanças nas métricas utilizadas caso assim se justifique.

3.5 Tecnologias Utilizadas

Nesta secção serão descritas as tecnologias utilizadas no desenvolvimento desta dissertação, que permitiram levar a cabo a análise e processamento de dados, o desenvolvimento de diagramas e *dashboards* e a elaboração de modelos preditivos. As tecnologias utilizadas foram o *Tableau Prep* e *Tableau Desktop* para o desenvolvimento de visualizações, tratamento e exploração de dados, o *Python* para a elaboração do agrupamento e dos modelos preditivos, incluindo a utilização de um IDE e outras bibliotecas e o *Diagrams.net* na elaboração dos diagramas dos *dashboards* realizados.

3.5.1 Tableau Prep

O *Tableau Prep* é um software da empresa *Tableau* que permite fazer operações de limpeza, processamento e formatação de dados, de uma forma visual e direta, tornando mais fácil o processo inicial da análise de dados. Através da utilização de uma interface visual, que proporciona uma visão mais completa e abrangente dos dados, bem como recursos que facilitam o processo de limpeza, automatização, administração e limpeza dos dados [37].

Este software é composto por dois produtos: o *Prep Builder* e o *Prep Conductor*. O *Prep Builder* permite combinar, formatar e efetuar a limpeza dos dados para que estes possam ser analisados através de três vistas coordenadas, que permitem visualizar os dados por colunas e linhas, perfis de cada coluna, bem como todo o processo de preparação de dados, que é efetuado passo a passo por fluxos nos quais é possível implementar as operações pretendidas. O *Prep Conductor* facilita a publicação e partilha de fluxos através do *Tableau Server* ou *Tableau Online*, permitindo ainda a automatização e programação de fluxos para que a sua execução seja efetuada a uma hora pretendida.

Devido à vista por fluxos, é muito fácil editar valores nos dados, efetuar *joins* ou outras operações aos dados e observar, de forma imediata, a mudança nos dados, tornando o processo de processamento e limpeza de dados muito intuitivo, sem ser necessário escrever código. Para além disso, é ainda possível observar os dados no *Tableau Desktop* ou obter uma extração dos dados num dado passo do fluxo, de forma a consultar rapidamente os dados e partilhá-los, auxiliando a preencher a lacuna entre o processo de preparação de dados e sua análise [36].

Durante esta dissertação, o *Tableau Prep* foi utilizado na fase inicial de limpeza e processamento de dados, quer no que diz respeito aos dados das viagens, como aos dados das estações, nos quais era necessário efetuar algumas operações de limpeza, como eliminação de colunas e linhas, correção de valores, entre outros. Para além disso, foi ainda utilizado para efetuar outras operações nos dados como criar novas colunas de dados para efetuar análises nos mesmos.

A escolha desta tecnologia deve-se ao facto de, para além da interface visual permitir efetuar o processamento de dados de forma rápida e intuitiva, como referido anteriormente, visto não ser necessário escrever código, este software facilita a transição para o *Tableau Desktop* de forma a poder efetuar a criação de visualizações utilizando os dados processados.

3.5.2 Tableau Desktop

O *Tableau Desktop* é um software da *Tableau* que permite a ligação a vários tipos de dados, permitindo a criação de visualizações dos mesmos e refinar as mesmas através de operações de filtragem, agrupamento e ordenação.

Através da sua interface visual é possível efetuar a junção de dados e criar visualizações com mapas, gráficos de linhas, entre outros, utilizando a funcionalidade de *drag*

and drop, de forma a efetuar perguntas aos dados. Este software permite ainda criar cálculos em dados já existentes e efetuar correlações, regressões e análises de tendência. e *dashboards* interactivos, através da funcionalidade de *drag and drop*, que permitem detetar padrões visuais nos dados, permitindo a tomada de decisões com base nos dados [34].

A partir da combinação de múltiplas visualizações, é possível construir um único *dashboard* interativo, permitindo visualizar uma variedade de dados em simultâneo, que pode incluir filtros, páginas web, imagens, parâmetros, botões, entre outros, de forma a proporcionar uma vista mais completa dos dados. É ainda possível criar uma história utilizando uma sequência de visualizações, de forma a criar uma narrativa, fornecer contexto ou demonstrar como as decisões se relacionam com os resultados [35].

3.5.3 Diagrams.net

O *Diagrams.net*, previamente denominado *Draw.io*, é um *software open-source* para a criação de diagramas, havendo uma versão *online* e uma para *desktop*.

Este software permite criar diagramas UML, diagramas de sequência, fluxogramas e diagramas de classes, que já se encontram disponíveis, podendo-se gravar os mesmos em diversos formatos como PNG, HTML, JPEG, SVG e ainda num formato que permite editar o diagrama no *browser* novamente. Para além das bibliotecas de formas disponibilizados para a criação dos diagramas, é ainda possível criar a nossa própria biblioteca com formas criadas por nós, que é guardada num ficheiro de formato XML que pode ser carregado quando se pretende utilizar a mesma. Permite ainda a integração com diversas aplicações como *Google Drive*, *GitHub*, *GitLab* e *Dropbox*, existindo ainda uma aplicação *Office 365* para o Word, Excel e Power Point [9].

Nesta dissertação foi criada uma biblioteca com diversas formas para permitir a criação de diagramas dos vários *dashboards* realizados, utilizando a linguagem *IVML*. Este software foi escolhido para tal uma vez que, para além de ser de acesso fácil, também permite a criação de uma biblioteca personalizada, o que foi necessário neste caso. Para além disso, permite guardar facilmente os diagramas não só em diversos formatos, mas também em outras aplicações, como o *Google Drive*.

3.5.4 Python e outras Bibliotecas

A linguagem *Python* é fácil de aprender e muito poderosa, apresentando estruturas de dados eficientes e uma abordagem simples à programação orientada por objetos, fazendo desta a linguagem ideal para o desenvolvimento rápido de aplicações em diversas áreas, sendo utilizada para processamento e visualização de dados, desenvolvimento web, desenvolvimento de jogos, aprendizagem automática e inteligência artificial [29].

Foi utilizado o ambiente de desenvolvimento *PyCharm*, da *JetBrains*, especialmente criado para programar em linguagem Python, embora também suporte linguagens como *JavaScript*, *Cython*, *SQL*, *AngularJS*, entre outros. Este permite efetuar inspeções ao código,

correr, testar e efetuar o *deployment* de aplicações, possuindo um controlo de versões integrado, possuindo ainda uma consola *Python* interativa, bem como múltiplos pacotes científicos como o incluindo *matplotlib*, *NumPy*, *tensorflow*, entre outros [28].

A linguagem *Python* foi utilizada nesta dissertação através do ambiente *PyCharm* para efetuar algumas operações de processamento de dados e para o desenvolvimento de programas nos quais foram utilizados algoritmos de *clustering*, bem como para a realização de algoritmos preditivos, devido ao facto da facilidade de aprendizagem e devido ao facto do *PyCharm* incluir muitas bibliotecas que foram úteis ao longo do desenvolvimento deste projeto.

ANÁLISE EXPLORATÓRIA DE DADOS

Neste capítulo irão ser descritos os conjuntos de dados utilizados, serão enunciadas as conclusões retiradas da análise exploratória efetuada e o consequente processamento que foi necessário levar a cabo nos dados. Para além disso, serão ainda descritos os conjuntos de dados construídos no âmbito da implementação de algoritmos de agrupamento, bem como na condução de análises aos dados para as quais foram elaborados *dashboards*.

Este capítulo encontra-se organizado numa primeira secção 4.1 de Apresentação dos Dados, na qual são descritos os conjuntos de dados utilizados no âmbito desta dissertação, bem como o processo de análise exploratória e processamento de dados que foram levadas a cabo de forma a poder realizar análises utilizando os mesmos.

Na secção 4.2, são descritos os procedimentos de agrupamento realizados nos dados, nomeadamente agrupamento espacial e temporal, bem como os dados utilizados para o efeito, sendo levada a cabo uma análise dos resultados obtidos.

De seguida, foi efetuada a Caracterização Espacial e Temporal das Estações Gira através dos dados das estações e também das viagens, tendo sido elaborados alguns *dashboards*, que são descritos e apresentados na secção 4.3. Desta secção consta ainda uma breve introdução à linguagem *IVML*, que foi utilizada no processo de elaboração destes *dashboards*.

Finalmente, na secção 4.4, é descrita a análise elaborada com o objetivo de detetar o rebalanceamento nos dados das estações, para a qual foi elaborado um *dashboard*, que também é apresentado.

4.1 Apresentação dos Dados

No âmbito desta dissertação foram fornecidos dados do Sistema de Bicicletas Partilhadas da Gira em Lisboa, sendo necessário explorar os mesmos, de forma a conhecê-los. Os dados em questão foram fornecidos pela *CML*, mais concretamente pelo *LxDataLab*, o

Laboratório de Dados Urbanos de Lisboa, no âmbito do desafio proposto para esta dissertação. Estes dizem respeito às estações e viagens ocorridas no [SBP](#) da Gira ao longo do ano de 2018. Para além destes, foram fornecidos ainda conjuntos de dados meteorológicos provenientes do [IPMA](#) (Instituto Português do Mar e da Atmosfera), com informação relativa a precipitação, intensidade do vento, temperatura, radiação, direção do vento e humidade ao longo do ano de 2018.

De forma a ser possível utilizar os dados mencionados, foi necessário proceder à análise exploratória dos mesmos, bem como ao seu processamento, sendo que tal foi levado a cabo de acordo com a metodologia [CRISP-DM](#). Esta metodologia é um processo de mineração de dados que estabelece uma abordagem composta por vários passos para analisar os dados, tendo sido realizadas as fases de *Data Understanding* e *Data Preparation*.

A primeira fase mencionada consiste em perceber os dados disponíveis, efetuando a sua descrição, exploração e análise da qualidade dos mesmos, através da utilização de tabelas, gráficos, entre outros, sendo um passo crítico, de forma a poder evitar problemas durante a fase de processamento. A segunda fase é a mais demorada, podendo consistir em juntar dados, derivar novos atributos, substituir ou corrigir valores em branco, incorretos ou valores em falta, entre outros.

Aquando da análise exploratória dos dados fornecidos relativos às bicicletas Gira, foi elaborado um relatório que pode ser consultado no Apêndice [A](#), onde foi efetuada a análise detalhada de ambos os conjuntos de dados. Após esta análise, foi ainda efetuado o processamento, cujo relatório se encontra no Apêndice [B](#). No que diz respeito aos conjuntos de dados meteorológicos, também foi realizado um relatório que aborda ambas as fases de análise exploratória e processamento, que se encontra no Apêndice [C](#).

4.1.1 Estações Gira

Os dados relativos às Estações Gira contêm informação acerca do número de bicicletas que se encontravam numa dada estação e informação temporal relativa à data e hora a que se deu o registo dos dados, sendo que estes dizem respeito ao ano de 2018.

Na Tabela [4.1](#) encontram-se as informações relativas aos atributos deste conjunto de dados e o seu significado. Três destes atributos são identificadores numéricos: o identificador referente a cada um dos registos que é único e os identificadores de cada estação. Os identificadores das estações (Id Expl e Id Planeamento) são números de três algarismos compreendidos entre 101 e 999. Estes identificadores estão disponíveis no *website* da Gira ¹ de forma a que os seus utilizadores possam encontrar as estações, havendo quatro grupos distintos para números a começar por 1XX, 2XX, 3XX e 4XX. Dentro de cada grupo, a numeração das estações não segue uma ordem específica, ou seja, apesar de existirem 8 estações no grupo 2XX, a numeração das mesmas tem início no 208 e termina no 225, por exemplo.

Na Figura [4.1](#), é possível observar as estações do [SBP](#) da Gira, bem como uma relação

¹<https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/>

Atributo	Descrição
Identificador	Identificador referente ao número da leitura
Id Expl	Identificador da Estação Gira
Id planeamento	Identificador da Estação Gira
Designação comercial	Morada da Estação e Identificador
WKB Geometry	Representação geométrica em texto
Tipo Serviços Níveis	Indica o tipo de serviço
Número de Bicicletas	Número de bicicletas na estação
Número de Docas	Número de docas existente numa estação
Rácio	Rácio entre número de bicicletas e número de docas
Estado	Estado da Estação
Data de Update	Data e hora do registo

Tabela 4.1: Atributos relativos ao conjunto de dados das Estações Gira e respetiva descrição.

entre a localização das estações e o identificador, uma vez que as estações cujo identificador começa por 1XX se localizam no Parque das Nações, as que o identificador começa por 2XX localizam-se na zona da Baixa, na Avenida da Liberdade apresentam o identificador a começar por 3XX e no centro da cidade e zona de Telheiras todas as estações apresentam um identificador de tipo 4XX. Relativamente ao número de estações, o grupo 1XX é constituído por 12 estações o grupo 2XX contém 8 estações, o grupo da Avenida da Liberdade (3XX) conta com 7 estações e o grupo 4XX tem 47 estações.

Dois atributos são de formato de texto: a designação comercial, que é uma junção entre o identificador da estação e a sua localização (por exemplo, a rua) e o Estado, que pode apresentar três valores possíveis: *active*, *repair* ou *stock*, representando, respetivamente, que uma estação se encontra em funcionamento, para reparação ou que não se encontra visível na aplicação da Gira.

O atributo *WKB Geometry* estava no formato *WKB*, tendo sido efetuada a conversão para *Well-known text (WKT)*, o que permitiu perceber que o atributo é um multiponto com apenas um ponto contendo as coordenadas de cada estação.

Três atributos são numéricos: o número de bicicletas, o número de docas e o rácio. O atributo relativo ao número de bicicletas indica quantas bicicletas se encontravam disponíveis numa estação aquando da data e hora a que o registo foi efetuado e o número de docas dá informação acerca das docas de cada estação, sendo que cada doca tem capacidade para uma bicicleta, significando que este número indica a capacidade de cada estação. O rácio é uma medida da ocupação de cada estação numa dada data e hora, sendo que é a razão entre o número de bicicletas e o número de docas. Os valores deste atributo estão compreendidos entre 0 e 1, sendo que o zero indica que a estação se encontra vazia e o 1 indica que a estação apresenta um número de bicicletas igual à sua capacidade.

O único atributo temporal presente nos dados é a data de *update* e diz respeito à data

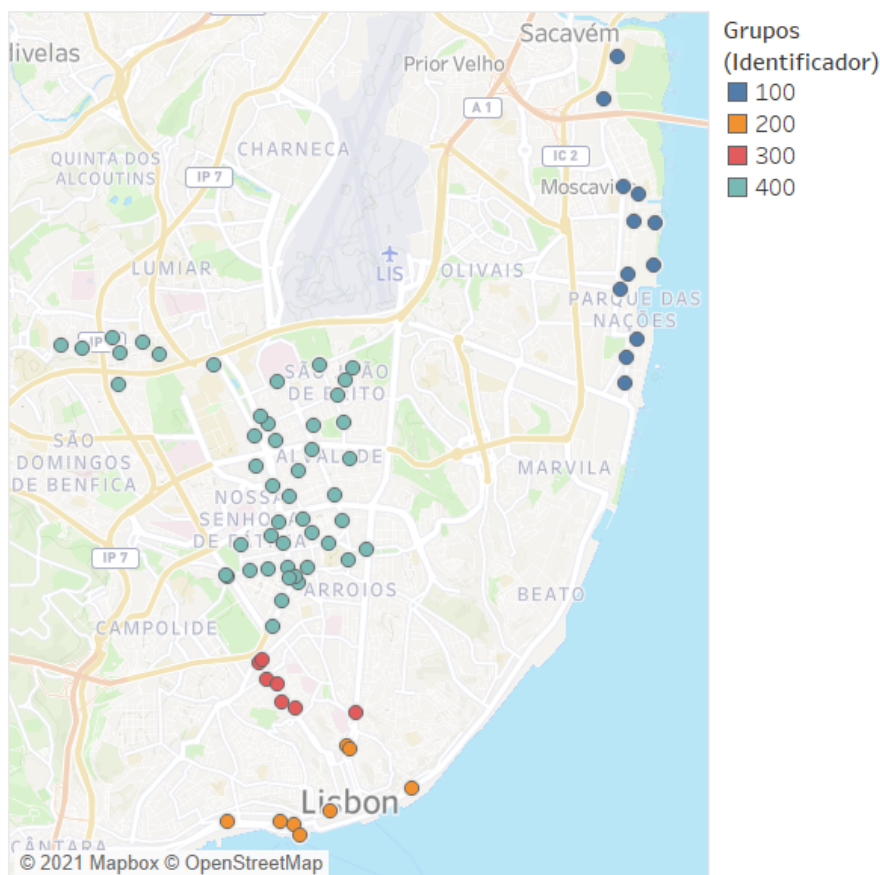


Figura 4.1: Localização das Estações e respetivos identificadores após o processamento.

e hora em que os registos foram efetuados, sendo que estes estão compreendidos entre 23 de Fevereiro e 18 de Dezembro com uma granularidade temporal de 5 minutos.

A análise exploratória destes dados, cujo relatório pode ser consultado no Apêndice A permitiu chegar a algumas conclusões relativamente à sua qualidade, tendo sido detetados alguns problemas. Com base nos problemas detetados, foi necessário proceder ao seu processamento, que pode ser consultado no Apêndice B, de forma a ser possível utilizar os dados no âmbito desta dissertação.

Optou-se por remover um dos atributos identificadores das estações, uma vez que se constatou que estes eram iguais (Id Expl e Id Planeamento), bem como o atributo “Tipos Serviços Níveis”, devido a este não ser relevante para a análise em questão, pois apenas dá informação relativa ao contrato de prestação de serviços associado à Gira. Removeram-se ainda os dados cujo identificador era 999, por este não apresentar relevância para os dados.

Foi efetuada a remoção dos dados relativos aos meses de Fevereiro e Março, devido à existência de duplicados nos mesmos. A escolha de remover estes dados por completo deve-se ao facto de, no mês de Fevereiro, apenas existirem dados de um dia e no mês da

Março, apenas existem dados relativos a 2 dias, o que significa que estes não são representativos dos meses referidos e, como tal, foram removidos.

Os valores do atributo geográfico *WKB Geometry* foram convertidos para formato *WKT*, de forma a extrair as coordenadas da localização de cada estação. Na análise exploratória, já se havia verificado que existia mais do que um valor deste atributo para algumas estações, significando que foi necessário confirmar a localização das estações após a conversão dos valores, de forma a introduzir nos dados a localização correta de cada estação. Esse processo, que se encontra descrito de forma mais aprofundada na sub-secção B.1.3, foi levado a cabo utilizando o *website* da Gira², de forma a fazer uma referência cruzada entre a localização que constava no *website* com os valores que constavam nos dados para cada estação. Foi ainda necessário corrigir a Designação Comercial das estações, uma vez que haviam estações para as quais existia mais do que um valor deste atributo, sendo que tal foi efetuado tendo como referência a designação comercial que consta do *website* da Gira.

Ainda relativamente ao processamento dos dados, foi necessário determinar uma data de início para cada estação, devido ao facto de existirem valores incongruentes no que diz respeito ao número de docas, que se constatou ser variável aquando do início dos registos de cada estação. Tal poderá ter ocorrido por as estações ainda se encontrarem em fase de preparação e não estarem disponíveis para utilização. Recorrendo a uma visualização, analisaram-se os dados relativos a cada estação, tendo também em conta o estado da estação, o número de docas e o número de bicicletas. Selecionou-se como data de início a data e hora a partir da qual o número de docas era o correto para cada estação e quando o seu estado passava a estar ativo. Os dados anteriores às datas de início selecionadas para cada estação foram removidos. Utilizando esta visualização foi ainda possível constatar que haviam 22 estações que permaneceram sempre em estado “stock”, com zero docas e zero bicicletas, tendo sido efetuada a sua remoção dos dados. O processo descrito acima poderá ser consultado na subsecção B.1.5 no Apêndice B. Após este processo, restaram as 74 estações presentes na figura 4.1, cujos dados serão utilizados ao longo do decorrer da dissertação.

4.1.2 Viagens Gira

Os dados relativos às viagens são relativos ao ano de 2018 e contêm informação relativamente às viagens que ocorreram na rede de bicicletas partilhadas da Gira nesse intervalo de tempo. A tabela 4.2 apresenta os atributos presentes nos dados e o significado dos mesmos. O identificador é único em cada linha do conjunto de dados, representando cada viagem de forma unívoca. As data de início e de fim são dois atributos temporais, dando informação relativamente à data e hora a que uma viagem teve início e quando esta terminou. O atributo da distância indica qual a distância percorrida na viagem ocorrida, encontrando-se em metros, com valores compreendidos entre 0 e 28 quilómetros.

²<https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/>

Atributo	Descrição
Identificador	Identificador único do registo
Data de início	Data e hora de início da viagem
Data de fim	Data e hora de fim da viagem
Distância	Distância percorrida aquando da viagem
Estação de início	Estação de início da viagem
Estação de fim	Estação de fim da viagem
RFID da bicicleta	Identificador da bicicleta
WKB Geometry	Pontos com coordenadas representativos da viagem
Número de Vértices	Número de vértices do percurso
Tipo de Bicicleta	Tipo de bicicleta

Tabela 4.2: Nome dos atributos relativos às viagens e respetiva descrição.

Os atributos relativos à estação de início e estação de fim apresentam informação relativa ao identificador da estação em que a viagem começou, bem como à estação em que a viagem terminou. O **RFID** da bicicleta é um código identificador eletrónico correspondente a cada bicicleta, que se encontra associado a cada registo, permitindo perceber que bicicleta foi utilizada em cada viagem. Existe ainda um atributo em formato **WKB**, um formato binário, que após conversão para formato **WKT**, se constatou ser um *multilinestring*, ou seja, uma sequência de pontos, sendo que representa o percurso efetuado na viagem. O atributo relativo ao número de vértices indica o número de pontos presentes no atributo **WKB** descrito anteriormente. Finalmente, o atributo relativo ao tipo de bicicleta é categórico, assumindo dois valores possíveis: E ou C, que representam, respetivamente, se uma bicicleta é elétrica ou clássica.

A distância indica qual a distância percorrida aquando do percurso de bicicleta, encontrando-se nos dados em metros. Constatou-se que as distâncias estão compreendidas entre 0 e 28 quilómetros, no entanto, este atributo apresentava quase 50% de valores nulos. Analisando apenas os valores não nulos das distâncias, conclui-se que quase a totalidade das viagens foi entre 0 e 5 quilómetros.

A estação de início indica a localização do início de uma viagem, sendo que o número corresponde ao identificador da Estação. De acordo com o website da Gira ³, a bicicleta elétrica tem uma autonomia até 25 km e permite pedalar até uma velocidade de 25 km/h. Aquando da análise exploratória, que pode ser consultada no Apêndice A, foi possível perceber melhor os dados, a sua qualidade e tomar decisões relativamente ao processamento a efetuar. Analisando os atributos da data de início e fim, é perceptível que, ao longo do ano de 2018, o número de viagens vai aumentando em cada mês, exceptuando no último

³<https://www.gira-bicicletasdelisboa.pt/conhece-a-bicicleta/>

mês em que há registos de viagens, outubro, uma vez que apenas existem registos até dia 15 desse mês, como se pode observar na Figura 4.2. Para além disso, observou-se uma diferença no número de viagens relativamente aos dias da semana, havendo um maior número de viagens durante a semana e um menor número de viagens aos fins-de-semana e feriados. A Figura 4.3 apresenta o número de viagens ocorridas na rede da Gira nos dias do mês de Julho, observando-se que existe sempre um decréscimo com duração de dois dias, representativo dos fins-de-semana desse mês.

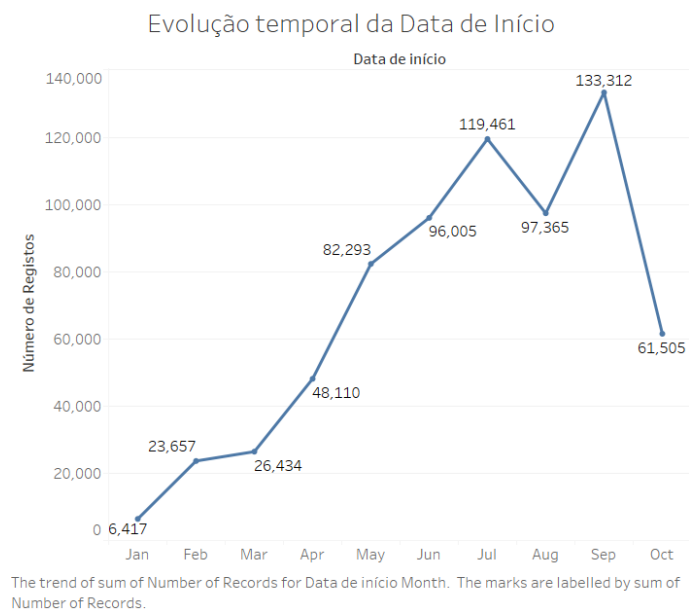


Figura 4.2: Evolução temporal do número de viagens ao longo dos meses de 2018.

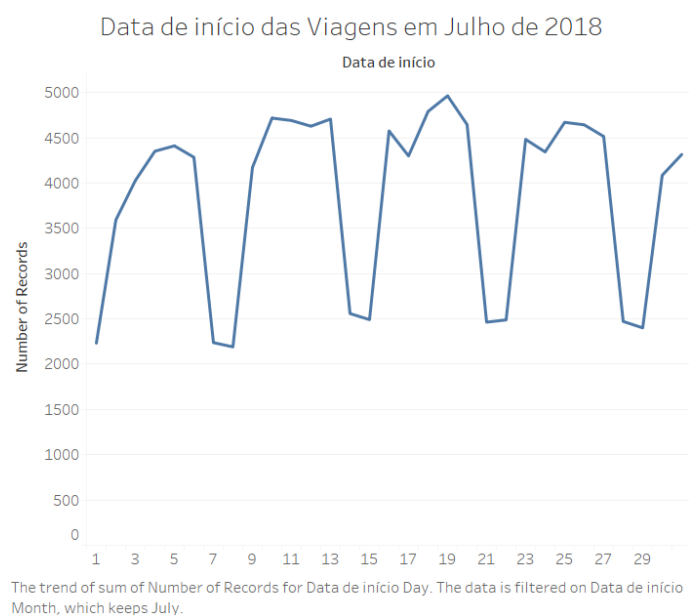


Figura 4.3: Registos de viagens nos dias do mês de Julho de 2018.

Os atributos relativos ao percurso efetuado na viagem, bem como ao número de vértices foram removidos dos dados, devido a mais de 50% destes serem valores nulos e também devido ao facto de não serem de interesse para a análise e trabalho que se pretende efetuar nesta dissertação. Para além disso, o atributo referente ao tipo de bicicleta, que poderia apresentar interesse, teve de ser removido devido ao facto de, para cada **RFID** de uma bicicleta, esta se encontrar classificada simultaneamente como clássica e elétrica, o que está incorreto. Foram encontrados dois registos cujo valor do atributo da estação de fim era nulo, tendo-se procedido à sua correção, recorrendo ao atributo do percurso (*WKB Geometry*) que não era nulo nestas duas linhas. Desta forma, foi possível descobrir qual era o identificador correto para ambos os registos, tendo sido efetuada a sua correção.

Foi calculado um novo atributo utilizando a data e hora de início e fim, de forma a calcular a duração das viagens, tendo sido detetados caso de viagens cuja duração é negativa e igual a zero, que foram removidos dos dados. Em adição a este processamento, foi utilizado o conjunto de dados com informação relativa aos identificadores e localização (latitude e longitude) de cada estação juntamente com os dados da viagens, de forma a obter informação geográfica das mesmas.

4.1.3 Dados Meteorológicos

Foram facultados pela Câmara Municipal de Lisboa dados meteorológicos do **IPMA** relativos ao ano de 2018. Estes eram compostos por seis ficheiros com granularidade horária e informação relativa à humidade, precipitação, radiação solar, temperatura, direção do vento e intensidade do vento medida em três estações meteorológicas localizadas em Lisboa. Na Figura 4.4 é possível observar a localização das estações meteorológicas.

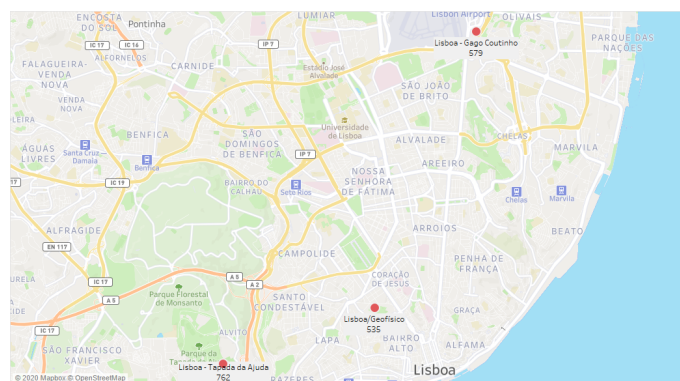


Figura 4.4: Estações Meteorológicas em Lisboa em 2018.

No relatório do apêndice **C** é possível consultar a análise exploratória, bem como o processamento efetuado aos dados de forma mais aprofundada. Aquando da análise exploratória referida, verificou-se que, por vezes, as estações não se encontravam a funcionar, ou ocorreu algum erro na recolha dos dados, o que era sinalizado pelo valor -990 ou um valor a nulo. Verificou-se que nos casos da Radiação, Intensidade do Vento e Direção do Vento, não haviam dados relativos à estação 535 ou estes eram escassos, havendo uma

predominância de valores inválidos. Para além disso, também se verificou um número de registos com valores inválidos nas restantes estações e para os diferentes conjuntos de dados, embora em menor proporção.

Devido à existência destas lacunas nos dados, bem como ao facto de Lisboa ser um concelho com uma área geográfica relativamente pequena, verificou-se aquando da análise se os valores de cada atributo diferiam ou não entre as três estações, tendo-se verificado que estes eram muito semelhantes. Como tal, de forma a colmatar a falta de dados em algumas estações, que se verificou nunca ser simultânea nas três estações para todos os atributos, optou-se por calcular, para cada atributo, a média dos valores das estações meteorológicas, sendo que, no caso de não existirem dados relativos a uma estação, este valor não entrava para a média, utilizando-se apenas as estações restantes. Após este processo, foi construído um novo conjunto de dados cujas colunas constam da tabela 4.3.

Atributo	Descrição	Unidades
Timestamp	Data e hora do registo	DD/MM/YYYY HH:mm
Humidade	Média da Humidade do Ar	%
Precipitação	Média da Precipitação Total	mm
Radiação Solar	Média da Radiação Solar	W/m ²
Temperatura	Média da Temperatura	°C
Rumo do Vento	Média do Rumo do Vento	°
Intensidade do Vento	Média da Intensidade do Vento	m/s

Tabela 4.3: Tipo de dados, descrição e unidades dos atributos que irão constar do conjunto de dados final.

4.2 Agrupamento

Neste trabalho, recorreu-se a técnicas de *clustering* para analisar algumas características das estações da rede da Gira e das viagens ocorridas. Primeiramente, foi levado a cabo o *clustering* espacial, para o qual foram utilizados dados geográficos das estações, bem como informação relativamente ao número de bicicletas médio em cada estação. De seguida, utilizando dados tanto os dados das estações como das viagens, foram implementados algoritmos de agrupamento para encontrar grupos de estações tendo em conta alguns atributos como o rácio, bem como as partidas e chegadas diárias.

Nesta secção serão descritos os procedimentos de agrupamento, bem como os respetivos parâmetros utilizados, quais os dados que foram criados para proceder ao procedimento de agrupamento e, de seguida, serão analisados os resultados obtidos.

4.2.1 Agrupamento Espacial

O agrupamento espacial teve como objetivo encontrar grupos de estações semelhantes tendo em conta diferentes atributos, nomeadamente a sua localização e o seu número de docas, de forma a responder às questões formuladas na fase de entendimento do negócio:

- "Existem grupos espaciais de estações em Lisboa?"
- "Existem grupos de estações que relacionem a localização e o número de docas?"

Para proceder ao agrupamento tendo em conta a localização das estações e o número de docas, foi utilizado um conjunto de dados em que cada linha diz respeito a uma estação, sendo composto por quatro atributos: o identificador único da estação, a latitude, a longitude e o número de docas de cada estação.

Recorreu-se ao algoritmo de agrupamento *K-means* para obter grupos de estações sem haver a exclusão de nenhuma das estações, uma vez que este algoritmo não deteta ruído nos dados, ou seja, não considera nenhum grupo de *outliers* nos resultados do processo de agrupamento. Pretende-se comparar os grupos de estações obtidos tendo em conta a localização com os grupos de estações existentes na Gira (por número de identificador), assim como os grupos tendo em conta o número de docas e a localização. Este algoritmo requer apenas um parâmetro k , que é representativo do número de grupos, sendo necessário avaliar qual o valor de k a utilizar, o que foi levado a cabo utilizando o método do cotovelo e a métrica *silhouette score*. Para o método do cotovelo foi utilizada a inércia, que mede a soma das distâncias ao quadrado das amostras ao centro do grupo mais próximo, seleccionando-se o k no qual a diminuição da inércia começa a ser linear, formando um "cotovelo". O *silhouette score* é uma métrica utilizada para avaliar a qualidade dos agrupamentos obtidos tendo em conta as distâncias intra e inter cluster, encontrando-se entre -1 e 1. Um resultado próximo de 1 é o ideal pois indica que a atribuição dos grupos está correta, um *silhouette score* de 0 indica que os grupos se encontram sobrepostos e um valor negativo ou próximo de -1 indica que estes estão incorretos.

4.2.1.1 Agrupamento por Localização

No agrupamento tendo em conta apenas a localização, foi obtido um gráfico relativo ao método do cotovelo, que se encontra na Figura 4.5a, o valor ideal do k poderá ser 4 ou 5, sendo necessário averiguar os valores do *silhouette score*. Na Figura 4.5b, é possível observar o gráfico relativo aos valores para o *silhouette score* que mostram que o valor seleccionado para o k deverá ser o 4, uma vez que é o que apresenta o valor mais elevado para esta métrica. Como tal, seleccionou-se um valor de k de 4, que significa que irão existir quatro grupos.

Os quatro grupos obtidos no agrupamento podem ser observados no mapa da Figura 4.6 e serão analisados de seguida:

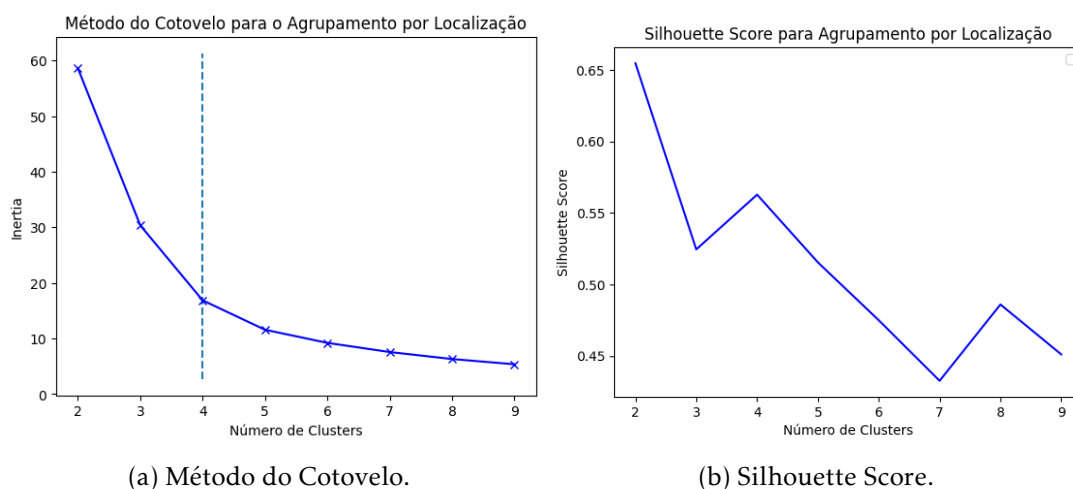


Figura 4.5: Método do Cotovelo e Métrica do *Silhouette Score* para o Agrupamento por Localização.

Grupo 0

Este grupo é constituído por 38 estações, sendo aquele que apresenta o número mais elevado e localiza-se na zona centro de Lisboa, mais concretamente entre a zona do Saldanha e Alvalade.

Grupo 1

O Grupo 1 é composto por 16 estações que se localizam, principalmente, nas zonas da Avenida da Liberdade e Baixa.

Grupo 2

As 12 estações pertencentes ao Grupo 2 localizam-se na zona do Parque das Nações.

Grupo 3

O último grupo é constituído por 8 estações localizadas perto da zona do Campo Grande e Telheiras.

Os grupos obtidos são quatro à semelhança dos grupos existentes das Estações da Gira, que são definidos pelos seus identificadores, no entanto, apenas o grupo do Parque das Nações se mantém, como se pode constatar observando a Figura 4.1, talvez por estar mais afastado de todas as outras estações.

Os grupos das estações com identificadores começados por 2XX e 3XX equivalem ao grupo 1 obtido no agrupamento e o grupo das estações com identificadores 4XX representa os grupos 3 e 0 do agrupamento obtido utilizando o *K-means*. Existem muitas estações cujo identificador é 4XX, que se encontram localizadas na zona Centro de Lisboa e na zona de Telheiras, sendo que as estações em Telheiras se encontram afastadas das que estão localizadas na zona Centro, o que pode justificar o facto de, no agrupamento obtido, estas terem sido colocadas em dois grupos distintos. No que diz respeito às estações com

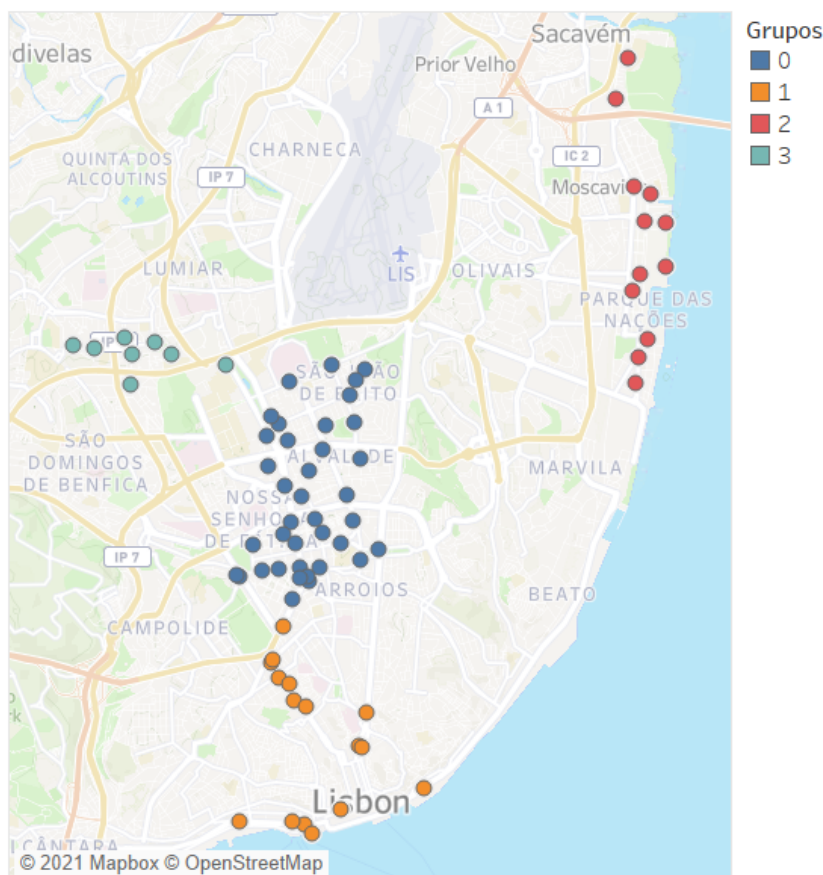


Figura 4.6: Grupos obtidos após o agrupamento espacial com base na localização.

identificadores 2XX e 3XX, estas estão localizadas, respetivamente, na Avenida da Liberdade e zona da Baixa, tendo sido colocadas no mesmo grupo mediante a sua localização, o que poderá ser devido à proximidade entre ambos estes locais.

No geral, o grupo com maior número de estações é aquele cujas mesmas se localizam no centro da cidade, englobando muitas das estações com identificador 4XX, sendo que as restantes foram incluídas nos restantes três grupos, cujo número de estações por grupo é muito menor relativamente ao grupo 0.

4.2.1.2 Agrupamento por Localização e Docas

Para o agrupamento tendo em conta a localização e o número de docas foram obtidos os gráficos para o método do cotovelo e para o *silhouette score*, que podem ser observados nas Figuras 4.7a e 4.7b. Foi escolhido utilizar ambos estes atributos e não apenas o número de docas pois pretende-se agrupar estações mais próximas de acordo com o seu número de docas, de forma a tentar compreender se existe alguma relação do número de docas na disposição de estações na rede como, por exemplo, dispor estações com maior capacidade perto de outras com menor capacidade, por exemplo. Assim, pode ser possível perceber se é possível, mais tarde, utilizar estações vizinhas a uma dada estação para auxiliar no

processo de rebalanceamento.

A partir do método do cotovelo, o valor óptimo para o k poderá ser 4, 5 ou 6, uma vez que, a partir da observação do gráfico, o cotovelo não se encontra definido de forma clara. Como tal, para auxiliar na decisão do valor de k , foram consultados os valores da métrica *silhouette score* para cada um dos valores mencionados, tendo sido determinado que, embora o valor mais elevado da métrica seja no valor de k igual a 8, este não é claramente o cotovelo no gráfico respetivo. Por essa razão, o valor adequado de k deverá ser o 6, visto que é o valor mais elevado entre os três valores considerados inicialmente.

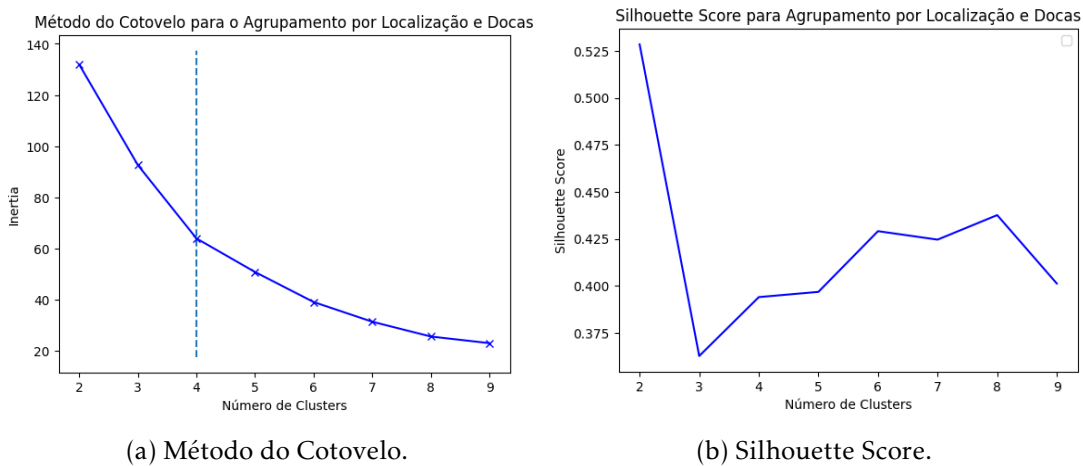


Figura 4.7: Método do Cotovelo e Métrica do *Silhouette Score* para os diferentes valores de k do Agrupamento por Localização e Número de Docas.

Analisando os grupos formados após o processo de agrupamento que se encontram na Figura 4.8, observa-se que existe alguma sobreposição dos grupos a nível geográfico, sendo necessário analisar os valores do número de docas relativo a cada um dos grupos.

Como tal, foi criado um *boxplot* (Figura 4.9) que permite analisar, para cada grupo, o número de docas das estações que o compõem, permitindo obter mais informação acerca dos grupos obtidos no que diz respeito ao número de docas, em adição à informação geográfica proporcionada pelo mapa. A análise dos grupos obtidos encontra-se descrita abaixo:

Grupo 0

As 29 estações pertencentes a este grupo encontram-se maioritariamente na zona centro de Lisboa, desde Arroios a São João de Brito, havendo ainda uma estação perto do Martim Moniz. Numa primeira observação apenas do mapa, constata-se que este grupo se encontra sobreposto com o Grupo 3, no entanto, observando o *boxplot* da Figura 4.9 é evidente que ambos os grupos se diferenciam no número de docas, uma vez que, no caso do Grupo 0, o número de docas destas estações se encontra entre 10 e 17, sendo superior no caso do Grupo 3, como se irá descrever de seguida.

Grupo 1

O grupo 1 é constituído por 12 estações localizadas na zona da Avenida da Liberdade

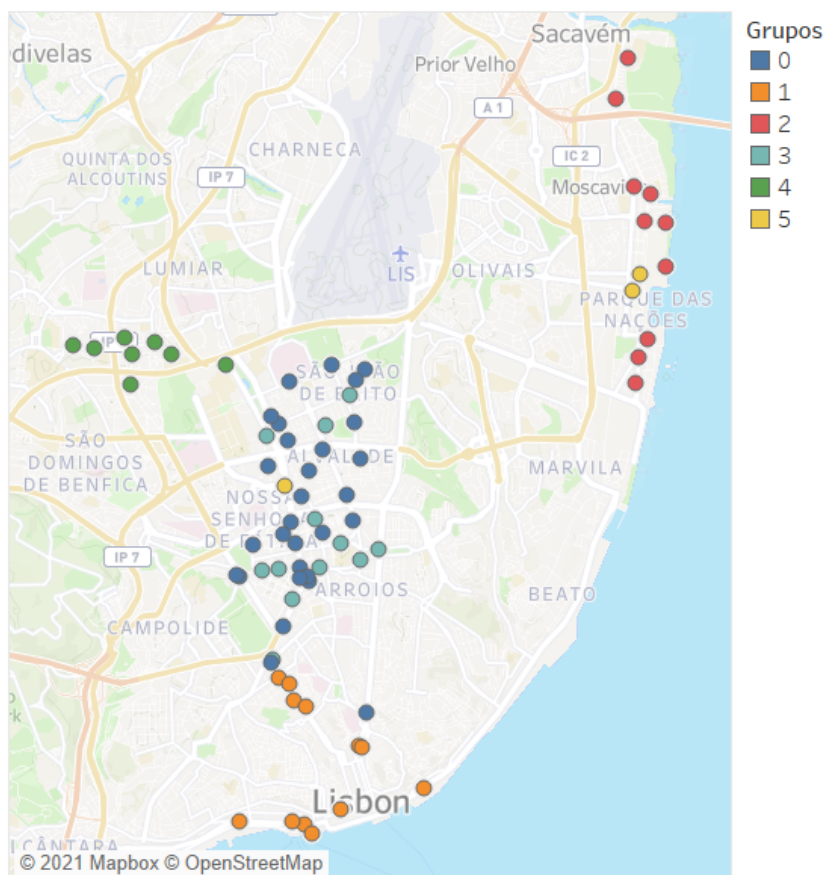


Figura 4.8: Grupos obtidos após o agrupamento espacial com base na localização e no número de docas.

e Baixa, ou seja, na zona sul de Lisboa. O número de docas destas estações encontra-se compreendido entre 16 e 24.

Grupo 2

As estações do grupo 2 encontram-se localizadas na zona do Parque das Nações, sendo 10 no total e com valor de docas compreendido entre 10 e 20. Estes valores de número de docas são semelhantes aos do grupo 1, no entanto, estas estações não ficaram no mesmo grupo pois a sua localização é diferente.

Grupo 3

O grupo 3 é composto por 12 estações que se encontram localizadas na zona centro de Lisboa, sendo que este grupo se encontra sobreposto com o Grupo 0, como referido anteriormente. A diferença entre estes grupos é no número de docas, cujos valores estão compreendidos entre 19 e 33 no grupo 3, sendo superiores.

Grupo 4

As 8 estações do grupo 4, localizadas na zona de Telheiras, apresentam os valores do

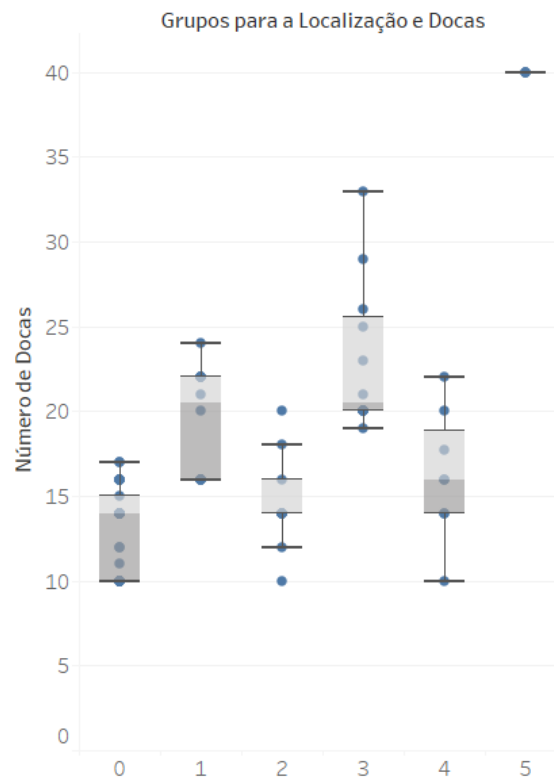


Figura 4.9: *Boxplot* dos grupos formados tendo em conta a localização e docas.

número de docas entre 10 e 22, sendo um intervalo semelhante aos grupos 0 e 2, com a diferenciação dos grupos a ser devido à componente geográfica.

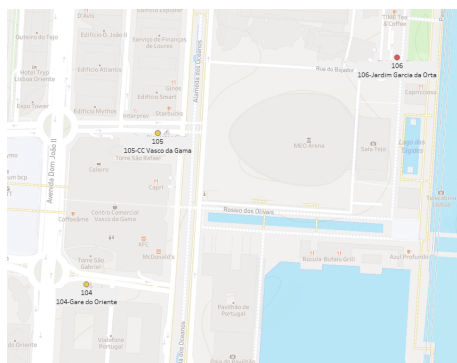
Grupo 5

O grupo 5 apenas é composto por 3 estações em localizações diferentes, como se pode observar na sub-Figura 4.10a, sendo que duas estações estão na zona do Parque das Nações, na Gare do Oriente e perto do Centro Vasco da Gama e outra no centro da cidade, ao lado da Estação de Entrecampos (sub-Figura 4.10b). Observando o *boxplot* da Figura 4.9, é perceptível que o número de docas deste grupo, que é de 40 para todas as estações, é muito mais elevado relativamente ao número de docas das restantes estações, sendo essa a razão pela qual estas estações ficaram no mesmo grupo apesar de uma delas se encontrar mais distante geograficamente.

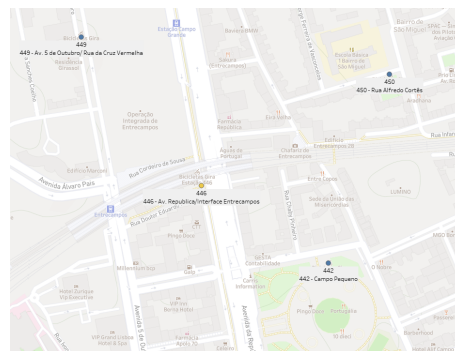
Para além disso, tendo em conta a sua localização, percebe-se porque é que estas contêm mais docas, uma vez que se encontram localizadas em locais de muito movimento, como o Centro Vasco da Gama e as Estações da Gare do Oriente e Entrecampos.

Os grupos obtidos apresentam dimensões semelhantes, entre 8 e 12 estações, à exceção de um deles que apenas é composto por 3, o que é devido ao facto do número de docas que as constituem ser muito elevado comparativamente às restantes.

A partir da análise efetuada dos mesmos, é perceptível que, apesar das estações de alguns grupos se encontrarem sobrepostas geograficamente, tal ocorreu devido ao facto



(a) Estações do Parque das Nações.



(b) Estação de Entrecampos.

Figura 4.10: Vista aproximada das estações 446, 104 e 105, pertencentes ao grupo 5, localizadas no Parque das Nações e em Entrecampos.

do seu número de docas ser distinto. Verifica-se ainda o oposto, estações cujo número de docas se encontra no mesmo intervalo foram colocadas em grupos distintos devido ao facto de não se encontrarem próximas geograficamente. No geral, existem três intervalos do número de docas principais entre 10 e 24 (grupos 0, 1 e 2), 19 e 33 (grupo 3) e o grupo 5 cujo número de docas é 40 para todas as estações que o constituem. Os grupos 0, 1 e 2, cujos valores do número de docas são semelhantes, encontram-se distantes a nível geográfico, com um dos grupos localizado no centro de Lisboa, outro na zona da Avenida da Liberdade e Baixa e o último no Parque das Nações.

Em suma, estes grupos permitem perceber que existem muitas estações com número de docas mais elevado no centro de Lisboa (a estação de Entrecampos do grupo 5 e as estações do grupo 3) e que duas das estações com maior capacidade se encontram muito próximas no Parque das Nações (as estações 104 e 105 do grupo 5). Tal permite perceber que existe uma necessidade das estações apresentarem um maior número de docas no centro da cidade e em localizações com maior movimento.

4.2.2 Agrupamento Temporal

O *clustering* baseado na componente temporal apresenta como objetivo responder à questão: "Existem estações com um comportamento temporal semelhante?", tendo sido decidido analisar o comportamento horário de alguns atributos, nomeadamente, a média do rácio (razão entre o número de bicicletas e a capacidade da estação), a média do valor de partidas e a média do valor das chegadas. A granularidade temporal escolhida foi de 15 minutos no caso do rácio e 1 hora no caso das partidas e chegadas, devido ao facto de os valores de partidas e chegadas serem baixos para uma granularidade temporal de 15 minutos.

Para proceder à realização do agrupamento, foi necessário construir novos conjuntos de dados que possam traduzir o comportamento que se pretende analisar, sendo necessário um conjunto de dados com o rácio médio, outro para as partidas e um terceiro para as

chegadas. Estes conjuntos de dados foram criados utilizando apenas a informação relativa aos dias úteis, uma vez que, como referido aquando da apresentação dos dados, bem como no trabalho relacionado, tende a existir uma diferença na utilização das redes partilhadas de bicicletas nos fins-de-semana e feriados e nos dias úteis.

4.2.2.1 Agrupamento por Rácio Médio Diário

No que diz respeito ao rácio, foi construído um conjunto de dados em que cada linha dizia respeito a uma estação, identificada pelo seu identificador único e colunas representativas dos intervalos horários ao longo do dia numa granularidade temporal de 15 minutos. Para cada estação, o conjunto de dados apresentava o valor médio do rácio relativo a cada intervalo temporal referido. Este valor médio do rácio foi calculado tendo em conta os meses de atividade de cada estação, de forma a que fosse possível comparar as estações. Na tabela 4.4 encontra-se um exemplo do conjunto de dados referido.

Identificador	00:00	00:15	00:30	00:45	..	23:45
101	0,5640	0,5784	0,5849	0,5880	..	0,5917
102	0,6383	0,6399	0,6455	0,6508	..	0,6361
..	..	..	..	..	..	..
490	0,4894	0,4999	0,5104	0,5208	..	0,4780

Tabela 4.4: Exemplo do conjunto de dados utilizado para o agrupamento temporal do comportamento das estações com base no seu rácio de ocupação.

Foi utilizado o algoritmo de agrupamento *K-means*, de forma a poder agrupar todas as estações tendo em conta o comportamento diário de cada estação tendo em conta o seu rácio sem haver a exclusão de estações, sendo que o objetivo deste agrupamento é perceber quais os grupos de estações com o mesmo padrão de utilização diária.

De forma a encontrar o valor óptimo de k , utilizou-se o método do cotovelo e a métrica *silhouette score* para avaliar os grupos. Visualizando o gráfico relativo ao método do cotovelo 4.11a constata-se que o valor óptimo para o k poderá ser o 4 ou o 5 e, consultando o gráfico para o *silhouette score* 4.11b, observa-se que o valor mais alto entre ambos é no 5, sendo esse o valor seleccionado.

Para analisar os grupos obtidos foi elaborado um *dashboard* (Figura 4.12) com um mapa com todas as estações e os respetivos grupos e uma visualização na qual cada grupo é representado por uma linha com a média do rácio ao longo do dia. Desta forma, é possível observar em simultâneo o comportamento diário do rácio, bem como em que local as estações desse grupo se encontram.

Observando os grupos resultantes do agrupamento com base na média do rácio de 15 em 15 minutos é possível observar o seguinte:

Grupo 0

As estações pertencentes ao grupo 0 apresentam-se mais vazias durante o dia, sendo que o número médio de bicicletas nestas começa a decrescer por volta das 7 da manhã,

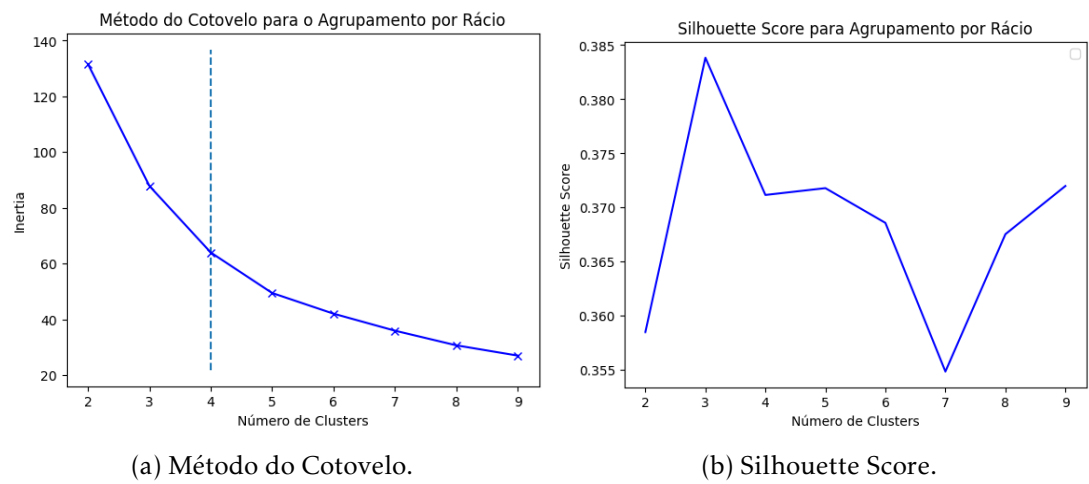


Figura 4.11: Método do Cotovelo e Métrica do *Silhouette Score* para os diferentes valores de k do Agrupamento por Rácio.

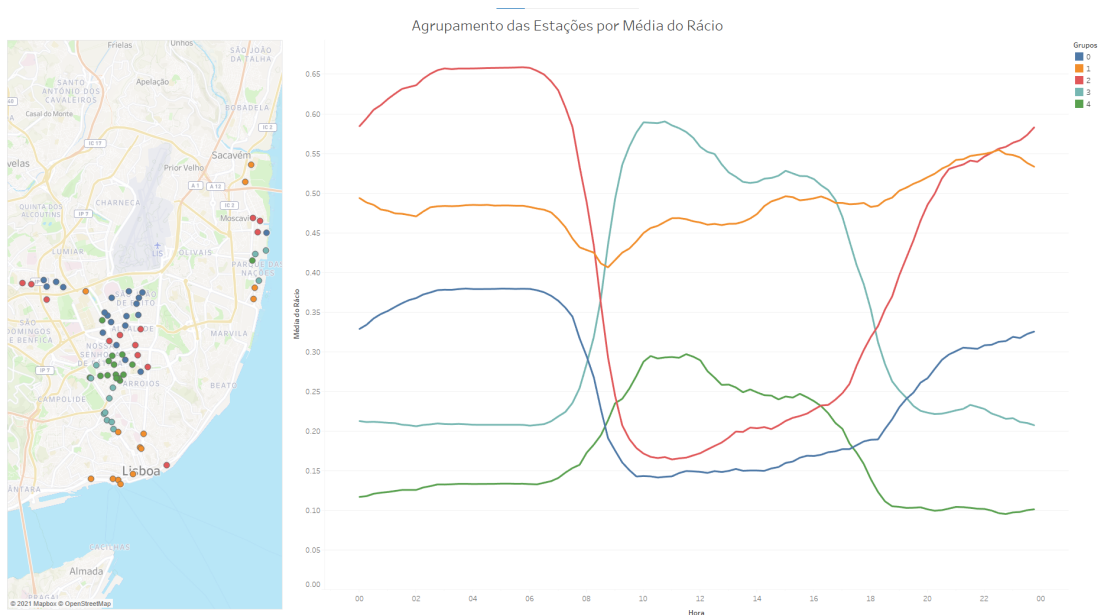


Figura 4.12: *Dashboard* com os grupos obtidos a partir do agrupamento tendo em conta o rácio médio ao longo do dia.

mantendo-se mais ou menos constante ao longo do dia, subindo novamente ao final da tarde, por volta das 18 horas. Geograficamente, a maioria das estações deste grupo encontra-se no centro de Lisboa e na zona de Telheiras, havendo uma única estação na zona do Parque das Nações com este padrão diário. Os valores de rácio diários oscilam entre 0.37 e 0.12, atingindo o valor mais alto nas horas de menor atividade (entre as 2 e as 7 da manhã).

Grupo 1

O grupo 1 apresenta um padrão que se mantém constante ao longo do dia e, observando o mapa, constata-se que se encontram isoladas de outras estações, no caso da zona centro de Lisboa e no Parque das Nações, podendo não ser tão utilizadas. Algumas estações neste grupo localizam-se na zona da Avenida da Liberdade e também na zona da Baixa, encontrando-se próximas, podendo significar que, ao longo do dia, não são muito utilizadas ou que a sua utilização é igual no que diz a ocupação e desocupação das estações. No geral, os valores de rácio deste grupo variam entre 0.35 e 0.65 ao longo do dia.

Grupo 2

Este grupo apresenta um comportamento semelhante às estações do grupo 0, embora neste os valores do rácio sejam entre 0.45 e 0.8, significando que estas estações se encontram quase cheias durante os períodos do início e final do dia. O decréscimo no rácio é mais acentuado, bem como o seu aumento e estes ocorrem às mesmas horas que no grupo 0. Relativamente à localização das estações deste grupo, estas encontram-se dispersas pela cidade.

Grupo 3

O grupo 3 apresenta um comportamento inverso aos grupos 0 e 2, pois as estações encontram-se mais vazias nas primeiras e últimas horas do dia e mais cheias durante o dia. Verifica-se um aumento no número de bicicletas a partir das 7 horas e um decréscimo pelas 17 horas, sendo que as estações do grupo 3 apresentam valores médios de rácio entre 0.35 e 0.65. Relativamente à localização das estações deste grupo, estas encontram-se, principalmente, na zona da Avenida da Liberdade e Parque das Nações.

Grupo 4

Neste grupo também se verifica que o comportamento é inverso ao dos grupos 0 e 2, com as estações a encontrarem-se mais cheias durante o dia e mais vazias no início e final do mesmo. Relativamente aos valores de rácio, estes são inferiores relativamente aos do grupo 4, sendo estações que se mantêm com um valor de rácio inferior a 0.5 durante todo o dia, ou seja, estas estações nunca se encontram a mais de metade da sua capacidade. Relativamente à localização das estações deste grupo, estas encontram-se na zona Centro e no Parque das Nações.

Os resultados obtidos no agrupamento permitem compreender como se comportam as estações da rede ao longo das horas do dia, sendo perceptível que, por exemplo, durante

o período da manhã as viagens que tendem a ser iniciadas nas estações dos grupos 0 e 2, cujo número de bicicletas diminui, tendem a terminar nos grupos 3 e 4, cujo número de bicicletas tende a aumentar também nesta altura, ocorrendo o oposto durante o período da tarde. Como os dados utilizados têm em conta apenas os dias úteis, tal poderá mostrar o percurso efetuado pela maioria dos utilizadores da rede aquando da sua deslocação para o trabalho e para casa durante a tarde.

4.2.2.2 Agrupamento por valor médio de Partidas e Chegadas

Para proceder ao agrupamento temporal tendo em conta os valores de partidas e chegadas de cada estação, foi utilizado o conjunto de dados relativo às viagens, tendo-se criado dois novos conjuntos de dados para esse efeito.

Como se pretendia avaliar o comportamento diário do número de partidas e chegadas, cada linha do conjunto de dados corresponde a uma estação e as colunas representam uma hora, na qual estão os valores médios de partidas ou chegadas (dependendo do conjunto de dados). É possível consultar um exemplo do conjunto de dados para as partidas médias na tabela 4.5. Uma vez que os dados de cada estação variam em termos do seu

Identificador	00:00	..	20:00	21:00	22:00	23:00
101	0,26	..	1,34	0,89	0,60	0,42
102	0,31	..	1,29	1,04	0,72	0,44
..	..	..	..	..	..	..
490	1,05	..	2,37	2,14	2,50	1,69

Tabela 4.5: Exemplo do conjunto de dados utilizado para o agrupamento temporal do comportamento das estações com base no valor médio de partidas por hora.

intervalo temporal, havendo estações com dados de junho a outubro e outras com dados de janeiro a outubro, foi necessário garantir que os valores eram comparáveis, de forma a que uma estação não apresentasse mais viagens apenas por se encontrar em atividade há mais tempo. Como tal, os conjuntos de dados apresentam a média do valor de partidas e chegadas, obtido através da soma do número de partidas ou chegadas a dividir pelo número de dias de atividade de uma estação, significando que a média foi calculada para cada estação de acordo com o período temporal dos dados relativos à mesma.

Agrupamento por Valor Médio de Partidas

Relativamente ao agrupamento do valor médio diário de partidas, determinou-se que o valor ótimo de k era 4, tendo em conta o método do cotovelo, cujo gráfico se encontra na Figura 4.13a e o *silhouette score* para o diferentes valores de k , com o respetivo gráfico na Figura 4.13b.

Para analisar os resultados, foi elaborado um *dashboard* (Figura 4.14) composto por um mapa com todas as estações da rede Gira, no qual a cor representa os grupos obtidos no processo de agrupamento e por um gráfico de linhas no qual o eixo do X representa

as horas ao longo do dia e o eixo do Y representa o número médio de partidas. Cada linha da visualização representa o valor médio de partidas de todo o grupo, tendo sido somado os valores médios de partidas a dividir pelo número de estações nesse grupo, de forma a que grupos com mais estações não apresentassem valores superiores de partidas relativamente a grupos com uma menor quantidade de estações, o que também permite facilitar a análise do comportamento de cada grupo. Observando os grupos obtidos, é possível retirar as seguintes conclusões:

Grupo 0

Este grupo é composto apenas por uma estação, cujo padrão de partidas se destaca das restantes estações por apresentar valores superiores, sendo que esta estação é a 446, localizada na Avenida da República/Estação de Entrecampos, o que poderá justificar a maior quantidade de partidas.

Grupo 1

O grupo 1 é composto por estações cujo padrão de partidas apresenta valores mais baixos, sendo composto por 45 estações que, observando o mapa, se percebe que estão localizadas na zona da periferia da rede de bicicletas da Gira, o que pode justificar o facto de não serem tão utilizadas.

Grupo 2

O grupo 2, composto por apenas 6 estações, apresenta valores mais elevados relativamente aos valores das estações do grupo 1. As estações que compõem este grupo são a estação 105, localizada no Centro Vasco da Gama, a estação 307 no Marquês de Pombal, as estações 417, 420 e 421, perto do Instituto Superior Técnico na Alameda e a estação 481 no Campo Grande, que fica perto não só da Estação do Campo Grande como também da Cidade Universitária.

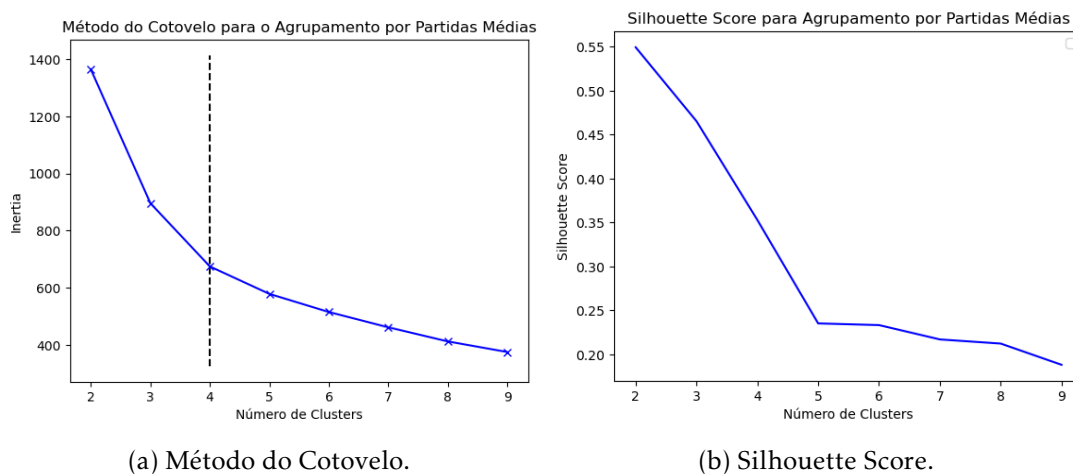


Figura 4.13: Método do Cotovelo e Métrica do *Silhouette Score* para os diferentes valores de k do Agrupamento tendo em conta o valor médio de partidas ao longo do dia.

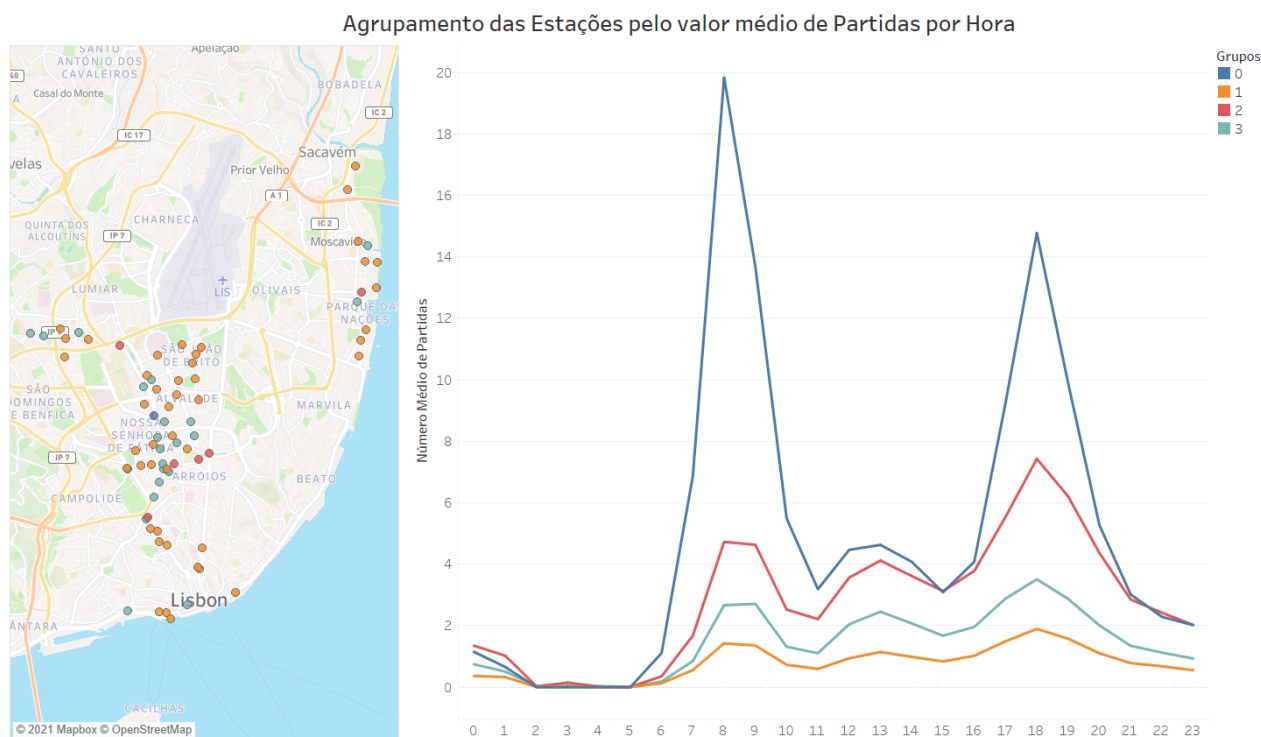


Figura 4.14: *Dahsboard* com os grupos obtidos a partir do agrupamento tendo em conta o valor médio de partidas ao longo do dia.

De um modo geral, o número de docas destas estações é mais elevado, sendo que a estação 105 tem 40 docas (o número máximo de docas na rede), a estação 307 tem 37 docas e as restantes têm entre 20 e 26 docas.

Grupo 3

O último grupo, constituído por 22 estações, apresenta valores superiores aos do grupo 1, no entanto o padrão diário é semelhante. A maioria destas estações encontra-se na zona Centro, no entanto, existem 2 que se encontram no Parque das Nações, 2 na zona da Baixa e 3 na zona de Telheiras.

Em todos os grupos é visível um aumento no valor de partidas por três vezes ao longo do dia, com picos durante a manhã (entre as 7 e 9 horas), hora do almoço (entre as 11 e as 14h) e ao final da tarde (entre as 17 e as 20 horas).

De uma forma geral, a localização das estações cujos grupos apresentam um valor superior de partidas ao longo do dia (o grupo 0 e o grupo 2), é maioritariamente no centro de Lisboa. As estações da periferia da rede encontram-se, maioritariamente, nos grupos 3 e 1, que apresentam valores diários de partidas mais baixos. Relativamente ao grupo 2, estas estações apresentam um número de docas maior relativamente às outras estações da rede de bicicletas e encontram-se localizadas em zonas de grande movimento da cidade de Lisboa (o centro Vasco da Gama e o Instituto Superior Técnico), o que pode justificar o facto do valor médio de partidas ser elevado.

Agrupamento por Valor Médio de Chegadas

No que diz respeito às chegadas, o valor óptimo de k foi também o 4, tendo sido selecionado com base no método do cotovelo (Figura 4.15a) e no valor do *silhouette score* (Figura 4.15b). Os grupos resultantes podem ser observados no *dashboard* da Figura 4.16, tanto no mapa como no gráfico de linhas, no qual cada linha diz respeito à média de todas as estações pertencentes a esse grupo, de forma a facilitar a visualização e análise dos grupos. Analisando o *dashboard*, é possível retirar as seguintes conclusões:

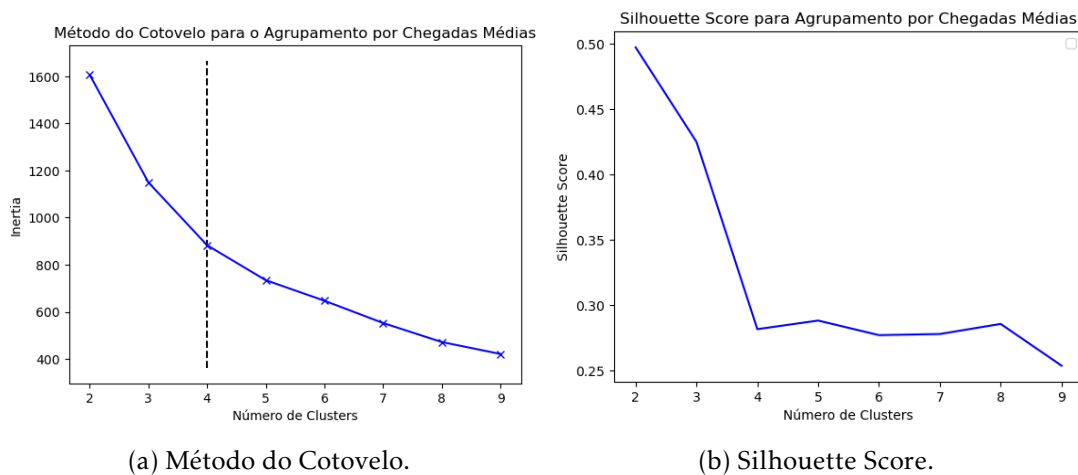


Figura 4.15: Método do Cotovelo e Métrica do *Silhouette Score* para os diferentes valores de k do Agrupamento tendo em conta o valor médio de chegadas ao longo do dia.

Grupo 0

Este grupo composto por 13 estações localizadas na zona centro de Lisboa, à exceção de duas estações localizadas no Parque das Nações e outra no Largo de Santos. As estações deste grupo apresentam valores médios de chegadas elevados principalmente durante a manhã, notando-se que no pico da tarde estes valores não são tão elevados.

Grupo 1

O grupo 1 é composto, mais uma vez, por estações que apresentam valores baixos de chegadas ao longo do dia, sendo o grupo que contém mais estações (35), cuja localização está dispersa por toda a rede.

Grupo 2

As 23 estações que compõem o grupo 2 apresentam valores médios de chegada superiores aos do grupo 1, apresentando um padrão diário semelhante. A localização destas estações encontra-se dispersa pela cidade, com 3 estações na zona do Parque das Nações, 5 na zona de Telheiras, 8 na zona da Baixa e as restantes no centro de Lisboa no Areeiro e Alvalade.

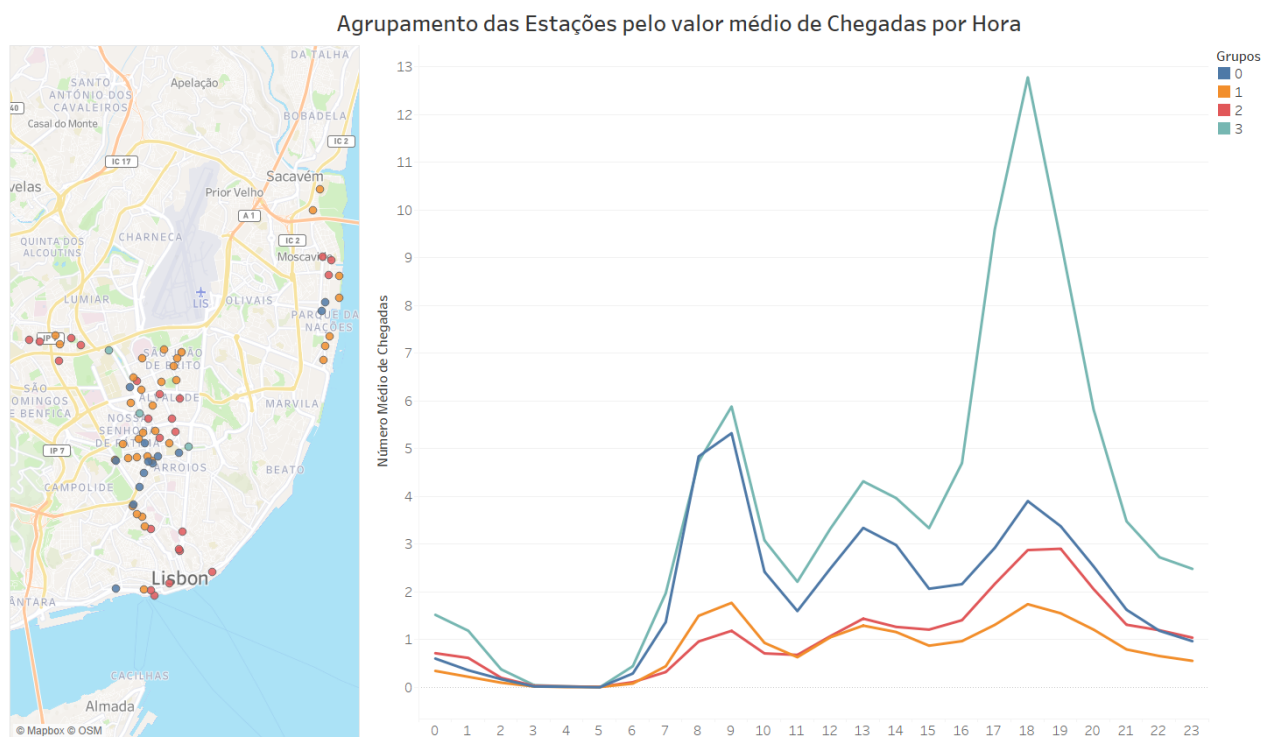


Figura 4.16: Dashboard com os grupos obtidos a partir do agrupamento tendo em conta o valor médio de chegadas ao longo do dia.

Grupo 3

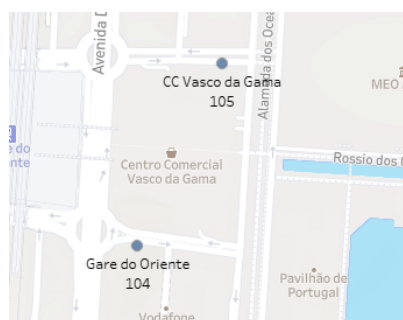
O último grupo é composto por apenas 3 estações que apresentam picos no valor das chegadas na manhã e tarde, com o pico da tarde a apresentar valores mais elevados relativamente ao grupo da manhã. As estações deste grupo apresentam valores mais elevados durante todo o dia e localizam-se no centro de Lisboa, mais concretamente, na Estação de Entrecampos, na Alameda D. Afonso Henriques, perto do Instituto Superior Técnico, e no Campo Grande, perto da Estação do Campo Grande e da Cidade Universitária.

Em todos os grupos se nota um padrão médio de chegadas com três picos ao longo do dia, sendo que o primeiro ocorre durante a manhã (entre as 7 e as 10), o segundo ocorre a meio do dia entre as 12 e as 15 horas e o último ocorre à tarde, tendo início pelas 16 horas e terminando pelas 21 horas.

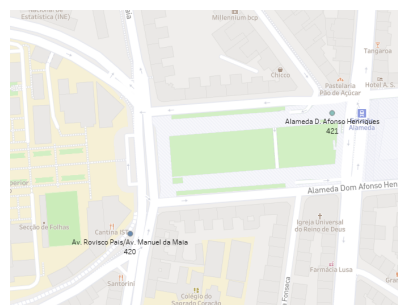
Para além disso, tanto o grupo 3 como o grupo 2 apresentam valores superiores para a média de chegadas durante a tarde, comparativamente ao bloco da manhã, enquanto que o comportamento inverso é observado nos restantes grupos, nos quais o pico da manhã apresenta um valor superior da média de chegadas.

É ainda notório que as estações que pertencem aos dois grupos que apresentam os valores mais elevados se encontram próximas de Estações de Comboio ou Metro, ou perto de Centros Comerciais e Universidades, como é o exemplo das Estações de Santos, São Sebastião, Marquês de Pombal, Picoas, Alameda, Instituto Superior Técnico, Cidade

Universitária e Centro Vasco da Gama, como se pode observar na Figura 4.17.



(a) Centro Vasco da Gama.



(b) Alameda.



(c) São Sebastião.



(d) Picoas.

Figura 4.17: Localização de algumas estações com o valor médio de partidas mais elevado próximas de Estações de Metro e Centro Vasco da Gama.

4.3 Caracterização Espacial e Temporal das Estações Gira

No âmbito desta dissertação foram utilizados dois conjuntos de dados principais, sendo que um destes é relativo às viagens e o outro é relativo às bicicletas que se encontram nas estações da rede. Para compreender melhor ambos os conjuntos de dados, bem como o comportamento diário de cada estação da rede da Gira, foram elaborados dois *dashboards*, que permitem caracterizar espacial e temporalmente as estações gira e responder às questões:

- "Qual o comportamento diário das estações tendo em conta o seu padrão de viagens (partidas e chegadas)?"
- "Qual o comportamento diário das estações tendo em conta o seu padrão de bicicletas?"

Para auxiliar à criação dos *dashboards* necessários a esta análise, foi criada em colaboração com outros colegas, como referido no final da secção 3.1, uma linguagem denominada **IVML** (Interactive Visualizaton Modeling Language), cuja utilização permitiu a elaboração de diagramas que permitem descrever o processo de *design* do *dashboard*, mais concretamente, as visualizações que irão constar do mesmo, bem como as interações e *tooltips* que se pretende adicionar. Nesta secção será feita uma breve introdução a esta linguagem, de forma a possibilitar o entendimento dos diagramas que serão posteriormente apresentados.

Depois, foram descritas as análises realizadas. Estas têm como objetivo avaliar o saldo de uma estação de duas formas diferentes, recorrendo não só a informação relativa à ocupação da estação (ao número de bicicletas presentes em cada estação) naquela data e hora, mas também a informação relativa ao número de partidas e chegadas ocorridas num intervalo de tempo, ou seja, a relação entre as partidas e chegadas de bicicletas a uma estação, que representam, respetivamente, a remoção e adição de bicicletas. Para tal, foi necessário adicionar criar novas colunas no conjunto de dados, de forma a poder ter a informação pretendida.

4.3.1 Interactive Visualization Modeling Language

A linguagem denominada Interactive Visualization Modeling Language (**IVML**) foi criada devido à necessidade de documentar o processo de design de *dashboards*, visto não existir nenhuma linguagem gráfica para a construção de diagramas que especifiquem e descrevam *dashboards* e a visualização interativa de dados.

Esta linguagem tem como objetivo, para além da documentação do processo, possibilitar o planeamento do que irá constar do *dashboard* e quais as interações e *tooltips* que se pretendem incluir neste, tendo em conta a pergunta ou perguntas a que esse *dashboard* pretende responder, antes de iniciar a sua construção. Desta forma, é possível representar o *dashboard* estruturalmente e de forma dinâmica, descrevendo os seus componentes, bem

como a relação entre estes e as interações que é possível realizar, respetivamente.

Utilizando esta linguagem, é possível construir dois tipos de diagramas utilizando esta linguagem: diagramas de interações e diagramas de *tooltips*.

Diagrama de Interações

Este diagrama descreve as visualizações e outros componentes que irão estar presentes no *dashboard*, como as visualizações, botões, filtros, entre outros. Relativamente às visualizações, são descritas as variáveis visuais utilizadas nas mesmas, bem como os dados e os tipos de gráfico. Para as interações, o diagrama permite demonstrar a ação necessária para esta ter início, bem como o componente onde começa e qual ou quais os componentes que são afetados por essa interação. É ainda referido quais os dados que são alterados durante este processo.

Diagrama de *Tooltips*

Este tipo de diagrama existe para descrever as *tooltips* existentes no *dashboard*, nomeadamente qual a informação que estas contêm, na forma dos dados que estão a ser mostrados nas mesma, qual o componente onde é possível aceder a cada *tooltip* e qual a ação que é necessário ser realizada para a visualizar.

Para possibilitar a construção dos diagramas definidos acima, foram criados vários componentes que auxiliam o processo, que podem ser observados na Figura 4.18. Os componentes principais são os de *Dashboard*, de uma Visualização e de *Tooltip* (sub-Figuras 4.18a, 4.18b e 4.18c, respetivamente). Depois existem ainda os componentes de Filtro, Parâmetro, Legenda e Botão, que são normalmente adicionados ao *dashboard* para permitir a realização de ações sobre os dados (sub-Figuras 4.18g, 4.18d, 4.18e e 4.18f, respetivamente). Para facilitar a compreensão do diagrama, todos os componentes apresentam, no canto superior esquerdo, um código composto por uma letra e um número, que indicam o tipo de componente e o seu número, respetivamente. Por exemplo, se existirem dois componentes de tipo Visualização, um terá o código V1 e o outro V2, de forma a poder não só distinguir entre ambos os componentes, mas também para facilitar a compreensão do diagrama. Para além disso, os componentes também apresentam diferenças na sua forma e cor, com os componentes secundários como o filtro, botão, legenda e parâmetro, serem de cor cinzenta, o componente de visualização é azul e os dois componentes de *tooltip* são brancos mas apresentam formas distintas. No que diz respeito à legenda, parâmetro e filtro, estes não podem ser distinguidos pela cor e forma, logo utiliza-se o código no canto superior esquerdo. O componente relativo ao botão é o único cuja forma é um hexágono. Quando se insere um componente de visualização, é possível indicar qual o título da mesma, os dados que constam da visualização, e ainda as variáveis visuais e o tipo de gráfico.

As variáveis visuais podem ser de três tipos: cor, tamanho e forma, com a forma a

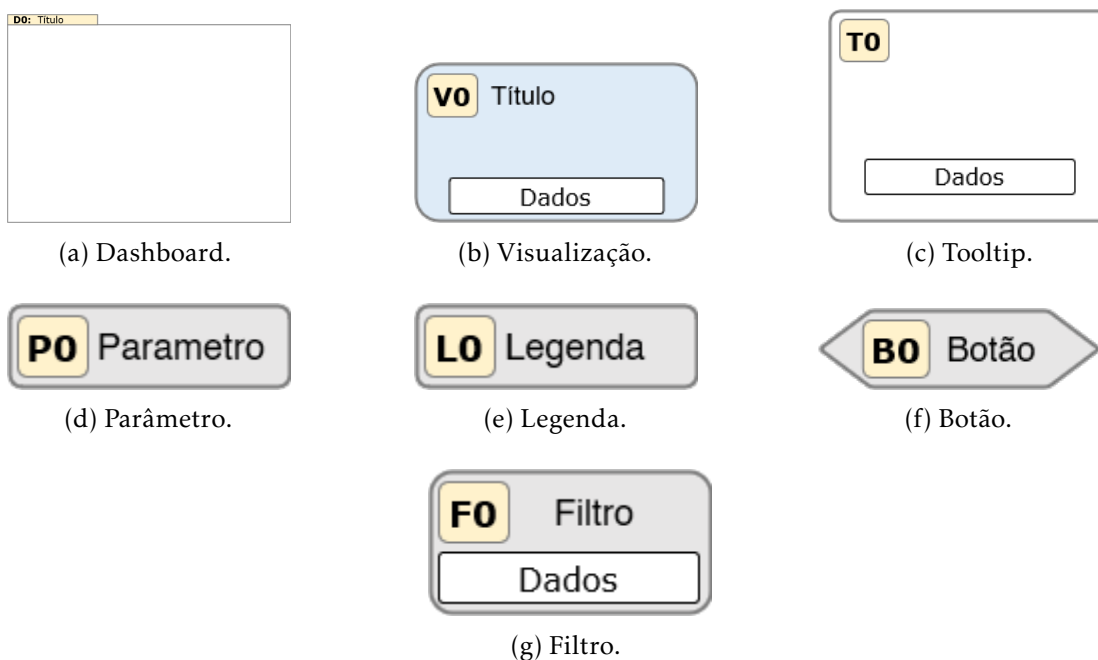


Figura 4.18: Componentes da Linguagem IVML que podem ser utilizados para a construção dos diagramas.

poder ser círculo, linha, ou várias formas. Um exemplo das variáveis visuais encontra-se na Figura 4.19.

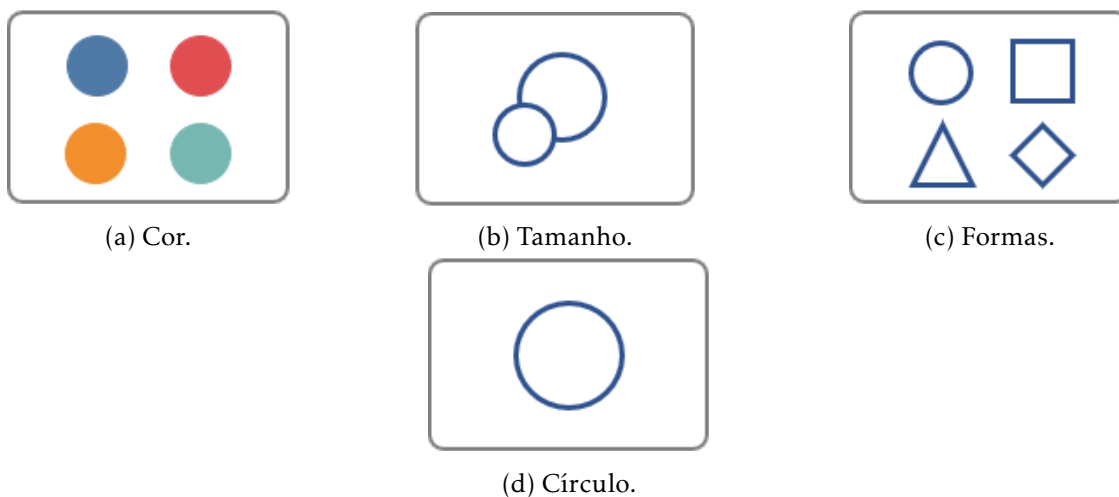


Figura 4.19: Variáveis visuais da linguagem IVML.

O tipo de gráfico pode ser de linhas, tabela, gráfico de barras, gráfico de barras empilhadas, gráfico de área, *hexabins*, mapa, *heatmaps*, entre outros, sendo possível observar alguns exemplos na Figura 4.20.

Como referido anteriormente, esta linguagem permite descrever as interações que é possível efetuar no *dashboard*, sendo que, para este caso, apenas se utilizaram as ações de filtragem e destaque, que se podem observar na Figura 4.21. Num diagrama, as interações

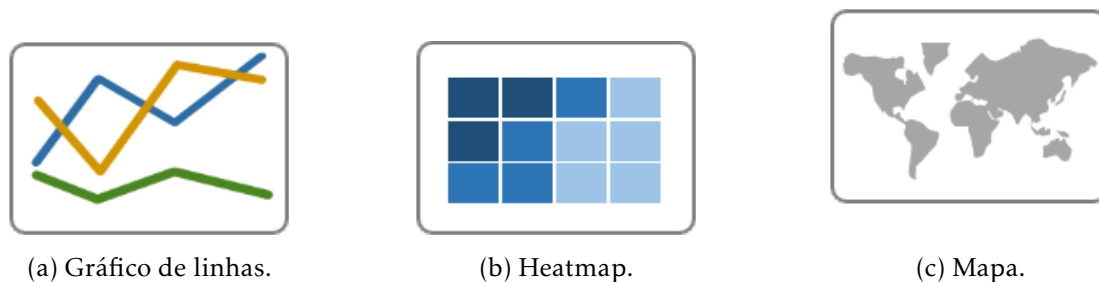


Figura 4.20: Tipos de Gráfico da linguagem IVML.

são introduzidas dentro de uma forma de paralelogramo, para as distinguir, na qual se encontra o ícone representativo do tipo de interação (filtragem ou realce), bem como o tipo de dados que estão a ser filtrados. Cada interação terá de estar ligada a um componente de saída, a partir do qual é desencadeada a interação e podem a um ou mais de um componente de entrada, que são influenciados pela interação em causa.

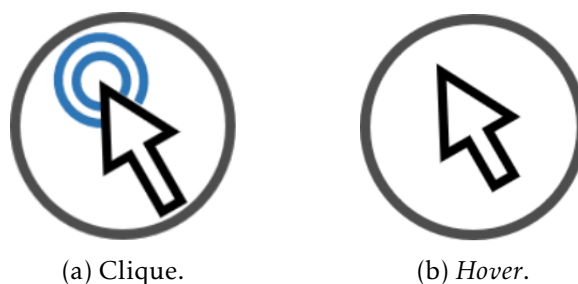


Figura 4.21: Tipos de interação da linguagem IVML.

As ações que é possível efetuar são, entre outras, clique e *hover*, tendo sido criados ícones para descrever cada uma delas, que se encontram na Figura 4.22. Os ícones relativos a este linguagem foram inspirados por alguns existentes no Tableau Desktop, bem como pelas ações e interações que é possível realizar nos *dashboards*.

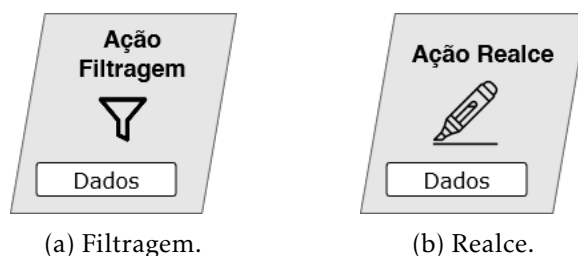


Figura 4.22: Tipos de ação da linguagem IVML.

Esta linguagem encontra-se definida e explicada de forma mais aprofundada num relatório acerca da mesma, do qual constam alguns exemplos da sua utilização.

4.3.2 Saldo de Viagens

Este *dashboard* apresenta como objetivo a caracterização relativa ao número de viagens ocorridas, tendo sido calculado o número de partidas e chegadas ocorridas para cada estação, utilizando a informação relativa à estação de início (partida) e à estação de fim (chegada) presente nos dados das viagens.

Para cada intervalo de 5 minutos, foram calculados o número de chegadas e o número de partidas e depois foi calculado o Saldo, através da subtração das partidas às chegadas (chegadas - partidas), sendo que as partidas foram consideradas como tendo um sinal negativo, uma vez que significam que houve a remoção de uma bicicleta da estação. De forma a que os valores fossem comparáveis, calculou-se a média dos mesmos, fazendo a sua soma e consequente divisão pelo número de dias correspondente ao intervalo temporal relativo a cada estação, uma vez que este é variável, pois há estações com dados de abril a outubro e outras cujos dados apenas são entre junho e outubro, por exemplo. Finalmente, foi efetuada a soma cumulativa da média do saldo ao longo das horas do dia.

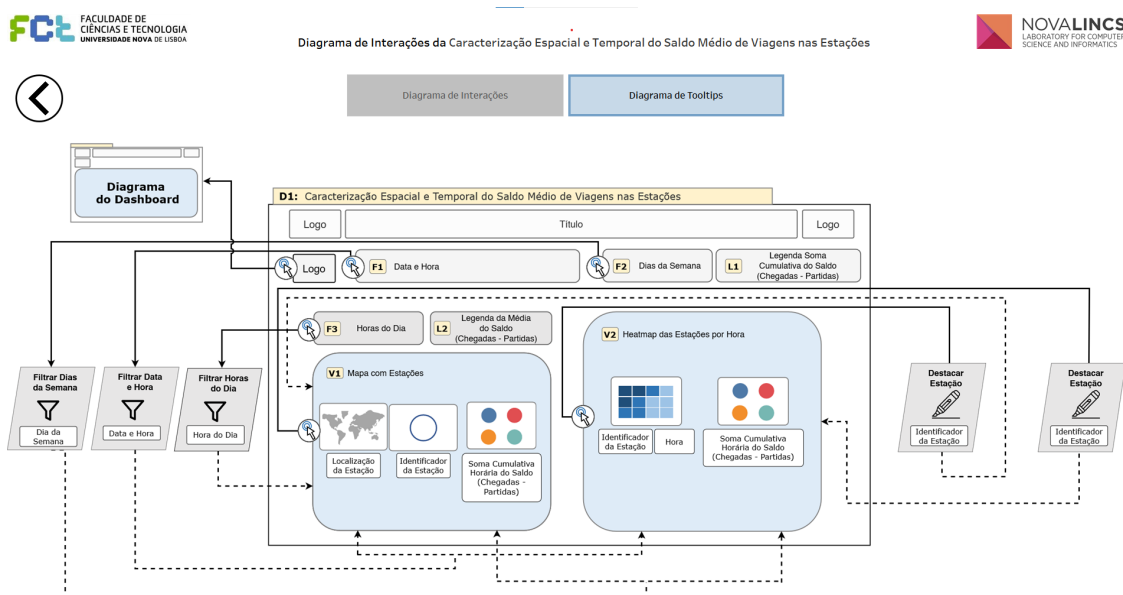


Figura 4.23: Diagrama de Interações para o *dashboard* de caracterização espacial e temporal do Saldo Médio de Viagens das Estações Gira.

Para auxiliar o processo de construção do *dashboard* foram desenvolvidos diagramas recorrendo à linguagem *IVML*, mais concretamente um diagrama de interações e um diagrama de *tooltips*, que descrevem, respetivamente, as visualizações existentes no *dashboard* e as interações que se podem realizar no *dashboard*, bem como que visualizações contêm *tooltips* e que informação estas contêm. Estes diagramas podem ser observados nas Figuras 4.23 e 4.24.

O *dashboard* desenvolvido pode ser observado na Figura 4.25, sendo composto por um mapa onde se encontram as estações representadas por círculos, com a cor a mapear a média do Saldo (chegadas - partidas) e um *heatmap* no qual cada linha representa uma

4.3. CARACTERIZAÇÃO ESPACIAL E TEMPORAL DAS ESTAÇÕES GIRA

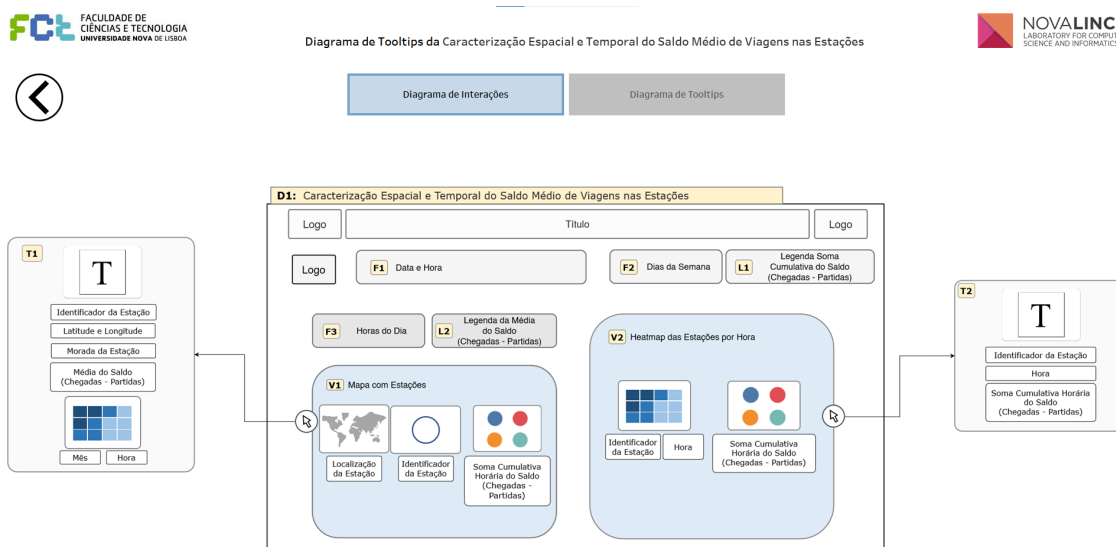


Figura 4.24: Diagrama de *Tooltips* para o *dashboard* de caracterização espacial e temporal do Saldo Médio de Viagens das Estações Gira.

estação, as colunas representam as horas do dia e as cores representam a soma cumulativa da média do Saldo calculada, ou seja, a média do Saldo é somada ao longo do dia. No *heatmap*, a ordenação das colunas (relativas às horas) está de forma a que o dia acabe às 2 da manhã, visto esta ser a hora a partir da qual há menos atividade nas estações. A ordenação das estações foi efetuada de forma decrescente com base no valor da variância do Saldo ao longo do dia, resultando numa ordenação em que as estações em que as estações que se encontram no cimo do *heatmap* são aquelas cujo valor médio do saldo mais varia e as que se encontram abaixo apresentam uma variação cada vez menor ao longo do dia. Analisando este *heatmap*, é possível perceber, a partir dos valores das viagens, que estações apresentam um perfil de falta de bicicletas e quais apresentam um perfil de excesso de bicicletas no final do dia, o que permite compreender em que estações teriam mais tendência a necessitarem de reposição de bicicletas ou de retirada de bicicletas no final do dia, respetivamente.

Relativamente à interactividade do *dashboard*, é possível seleccionar um intervalo temporal pretendido, bem como os dias da semana que se pretende analisar, sendo possível analisar, por exemplo, apenas a segunda-feira, apenas o fim-de-semana, ou os dias da semana (excluindo fins-de-semana). Esta interação permite analisar o comportamento das estações em dias específicos e em intervalos temporais específicos, de forma a perceber se existem dias em que o comportamento de uma estação é diferente, por exemplo.

Quando se coloca o rato sobre uma das estações no mapa, é ainda possível observar informação adicional acerca da mesma, como as suas coordenadas, o valor médio do saldo e ainda um *heatmap* com a média cumulativa diária do Saldo ao longo de todos os meses

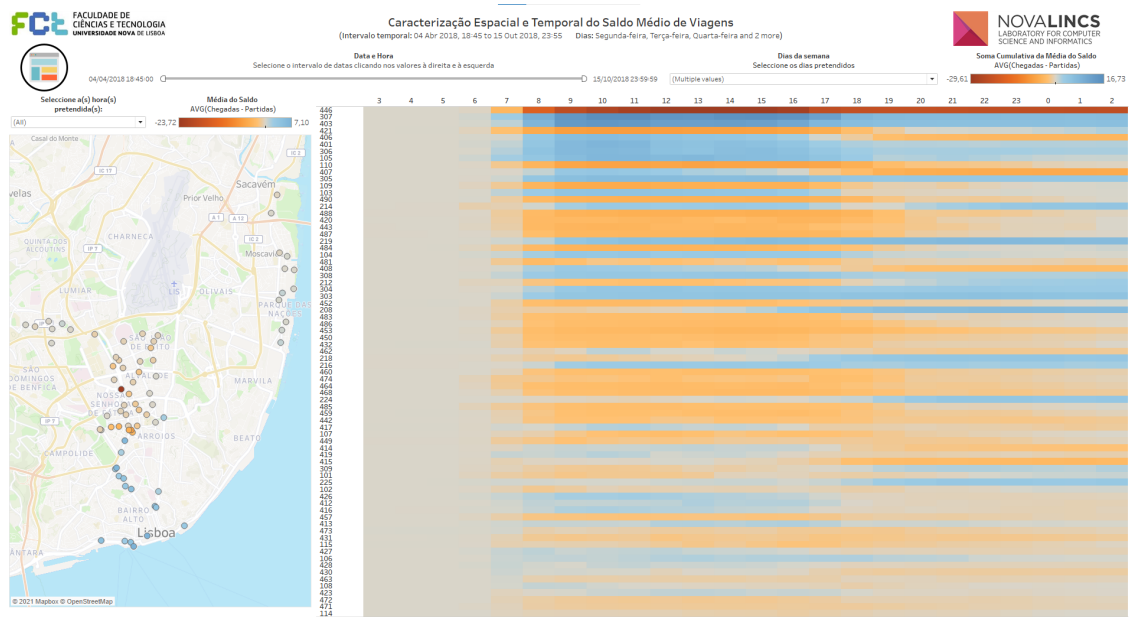


Figura 4.25: Dashboard de caracterização espacial e temporal do Saldo Médio de Viagens para os dias úteis.

do ano de 2018, o que permite analisar se houve uma mudança no padrão do comportamento diário ao longo dos meses ou se este se manteve, por exemplo. Um exemplo de uma tooltip pode ser observado na Figura 4.27a. Ainda relativamente ao mapa, como referido, é possível observar as estações coloridas de acordo com o seu valor médio do saldo, sendo que na Figura 4.25 este valor médio diz respeito a todas as horas do dia. No entanto, é possível selecionar a hora ou horas para as quais se pretende observar este valor, podendo escolher-se, por exemplo, intervalos específicos do dia, o que permite perceber se existem diferenças na média do saldo ao longo do dia. Tal permite analisar cada estação apenas tendo em conta o valor do saldo, uma vez que esta medida não é a soma cumulativa, sendo apenas o saldo médio na hora ou horas selecionadas. Para facilitar a análise das estações, é possível clicar nas estações pretendidas no mapa para as realçar no heatmap e vice-versa. Finalmente, é possível clicar no botão localizado no canto superior esquerdo para aceder aos diagramas do *dashboard*, sendo possível consultar o diagrama de interações (Figura 4.23), bem como o diagrama de tooltips (Figura 4.24).

Na Figura 4.25 é possível observar o *dashboard* com a soma cumulativa diária do Saldo relativa aos dias da semana, tendo sido excluídos o sábado e o domingo. Observando o *heatmap*, verifica-se na legenda que os valores da soma cumulativa da Média do Saldo estão compreendidos entre -29 e 16, significando que, como existe um maior intervalo entre ambos os valores, percebe-se rapidamente através da observação do *heatmap* que a maioria das estações apresenta valores de soma cumulativa menores, visto que as cores não são tão fortes como, por exemplo, as observadas para a primeira estação (446). Verificando o mapa, constata-se o que foi referido, ou seja, a maioria das estações, localizadas no centro de Lisboa e no Parque das Nações, apresentam a cor "bege", que representa

valores próximos do zero, ou seja, a meio da escala. A estação com cor vermelha escura é a 446, a primeira estação no *heatmap*, que apresenta a soma cumulativa sempre negativa ao longo do dia a partir das 7 da manhã.

As restantes estações na Avenida da Liberdade e na Baixa, esta tonalidade é azul, sendo representativa de um valor positivo da soma cumulativa, indicando que, no geral, nestas estações tendem a ocorrer mais chegadas do que partidas, sugerindo uma necessidade de efetuar a remoção de bicicletas para que estas não fiquem cheias, o que impossibilita a entrega de mais bicicletas na estação. Consultando as *tooltips* de algumas estações, é possível obter mais informação acerca do seu padrão de comportamento diário ao longo dos meses, sendo possível observar as *tooltips* para as estações 103, 208, 309 e 446 na 4.27. Para referência, a localização destas estações pode ser consultada na Figura 4.26.

Como referido anteriormente, estas *tooltips* apresentam informação adicional sobre a estação pretendida, tal como o Saldo Cumulativo ao longo do dia para os meses do ano de 2018 para os dias úteis. Observando o mapa, obtemos a informação de que a estação 446 apresenta uma média negativa do saldo, enquanto que as restantes apresentam este valor positivo, com ênfase na estação 208, que apresenta o valor mais elevado, sendo possível concluir que no geral a estação 208 apresenta um saldo maioritariamente positivo, significando que existe a necessidade de retirar bicicletas da mesma. Analisando o comportamento desta estação ao longo dos meses (neste caso entre junho e outubro) na Figura 4.27b, tal é confirmado, visto que em todos os meses o saldo cumulativo é positivo quase sempre a partir das 11 da manhã, sendo negativo apenas entre as 6 e 10 da manhã.

As estações 103 e 309, cujo tom de azul é mais claro e, por isso, significando que o seu valor médio do saldo é mais baixo, apresentam variações na soma cumulativa do saldo ao

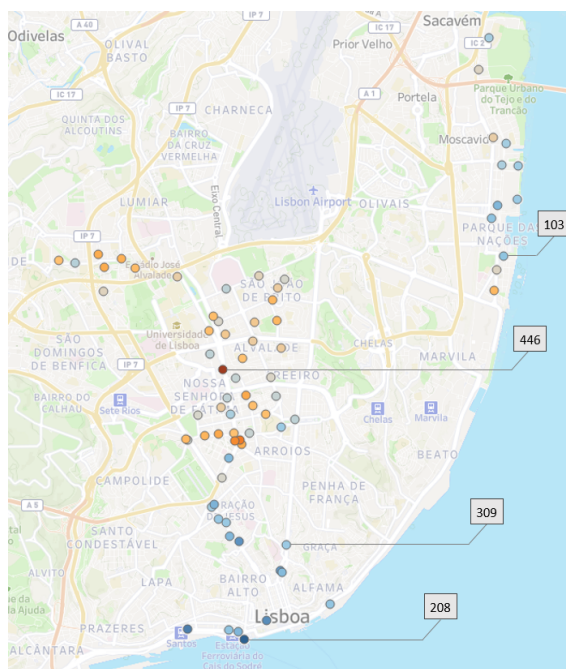


Figura 4.26: Localização das estações cujas *tooltips* se encontram na Figura 4.27

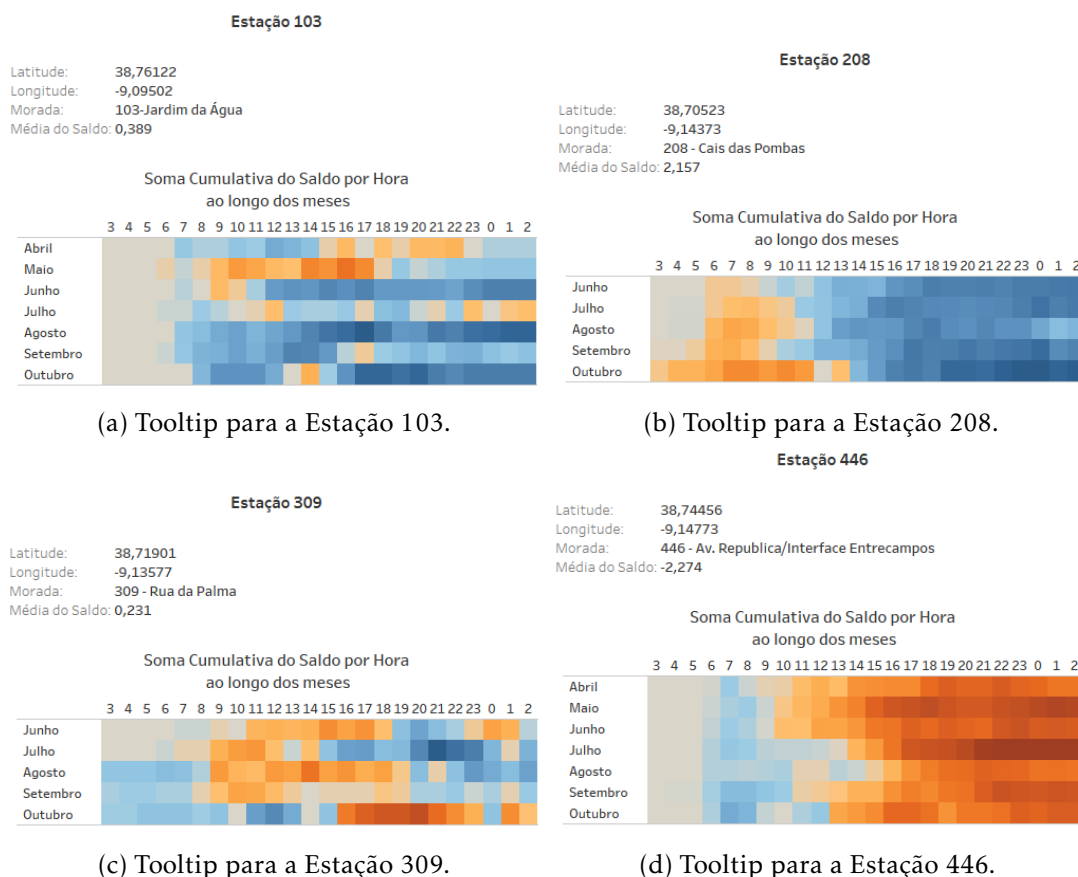


Figura 4.27: *Tooltips* do *dashboard* relativo ao Saldo de Viagens, acessíveis através da ação de *hover* das estações no mapa.

longo dos meses, apresentando comportamentos diferentes. No caso da estação 103, na Figura 4.27a, os primeiros meses são os que mais se diferenciam dos restantes, apresentando saldo negativo, no entanto, a partir de Junho o valor médio da soma cumulativa do saldo é quase sempre positivo, havendo sempre um maior número de chegadas relativamente ao número de partidas. A estação 309 (Figura 4.27c) apresenta mais meses com valores médios de saldo negativos, enquanto que a estação 446 (Figura 4.27d) apresenta, de forma geral, um valor de soma cumulativa negativo a partir das 10 da manhã em todos os meses, apresentando um comportamento semelhante ao longo de todos os meses.

Analisando a média da soma cumulativa do saldo para apenas os dias do fim-de-semana (Figura 4.28), é possível perceber que os valores da soma cumulativa são baixos durante estes dias, encontrando-se compreendidos entre -2,479 e 2,388, o que indica que a rede da Gira não é tão utilizada durante os fins-de-semana. Na literatura já havia sido verificado que, no geral, os sistemas de bicicletas partilhadas costumam ser mais utilizados para efetuar viagens de casa para o trabalho e vice-versa e, embora também fossem utilizadas por motivos de lazer, este tipo de utilização não é tão verificada, o que é compreendido através da análise e comparação do comportamento da rede nestes dois tipos de dia (dias úteis e não úteis). Verifica-se ainda que não há tantas estações com

4.3. CARACTERIZAÇÃO ESPACIAL E TEMPORAL DAS ESTAÇÕES GIRA

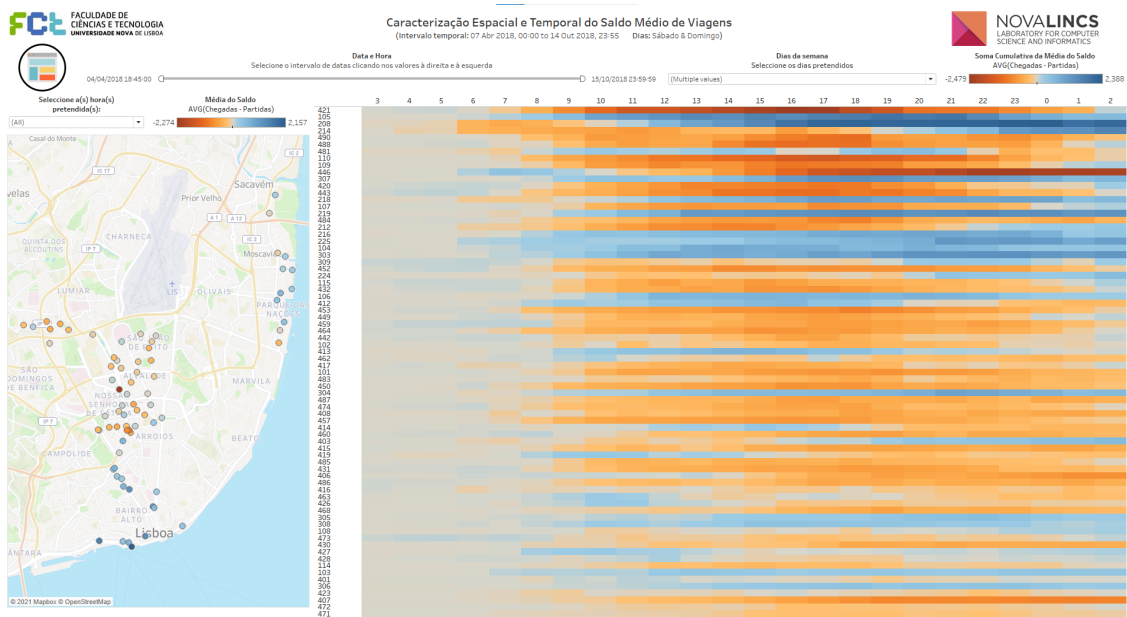


Figura 4.28: Dashboard de caracterização espacial e temporal do Saldo Médio de Viagens para os fins-de-semana.

valores próximos de zero, embora tal também se deva ao facto dos valores serem baixos. No entanto, a zona da baixa continua a apresentar uma média do saldo positiva, enquanto que no centro da cidade, se verifica o contrário. É ainda de salientar que, enquanto que durante a semana a estação 446 é aquela onde ocorre uma maior variação do saldo, aos fins-de-semana tal não sucede. Como esta estação se localiza na estação de Entrecampos, tal pode ser justificado devido ao facto de, durante estes dias, esta estação não ser tão utilizada devido a não ser um dia de trabalho para a maioria das pessoas.

4.3.3 Diferença de Bicicletas nas Estações

De forma a analisar o comportamento diário de cada estação relativamente ao número de bicicletas disponíveis, foi elaborado outro *dashboard*, para o qual foi calculada uma nova coluna nos dados com a diferença do número de bicicletas, ou seja, tendo em conta o atributo com informação acerca do Número de Bicicletas em cada estação, foi efetuado o cálculo da diferença desse valor ao longo do tempo. Para obter valores comparáveis, foi efetuada a média desse valor, fazendo a sua soma a dividir pelo número de dias de cada estação e, no *heatmap*, foi efetuada a sua soma cumulativa para cada hora do dia.

Para auxiliar o processo de construção do *dashboard* foram desenvolvidos diagramas recorrendo à linguagem *IVML*, mais concretamente um diagrama de interações e um diagrama de *tooltips*, que descrevem, respetivamente, as visualizações existentes no *dashboard* e as interações que se podem realizar no mesmo, bem como que visualizações contêm *tooltips* e que informação estas contêm. Estes diagramas podem ser observados nas Figuras 4.23 e 4.24.

O *dashboard* desenvolvido, que pode ser observado na Figura 4.31, apresenta uma

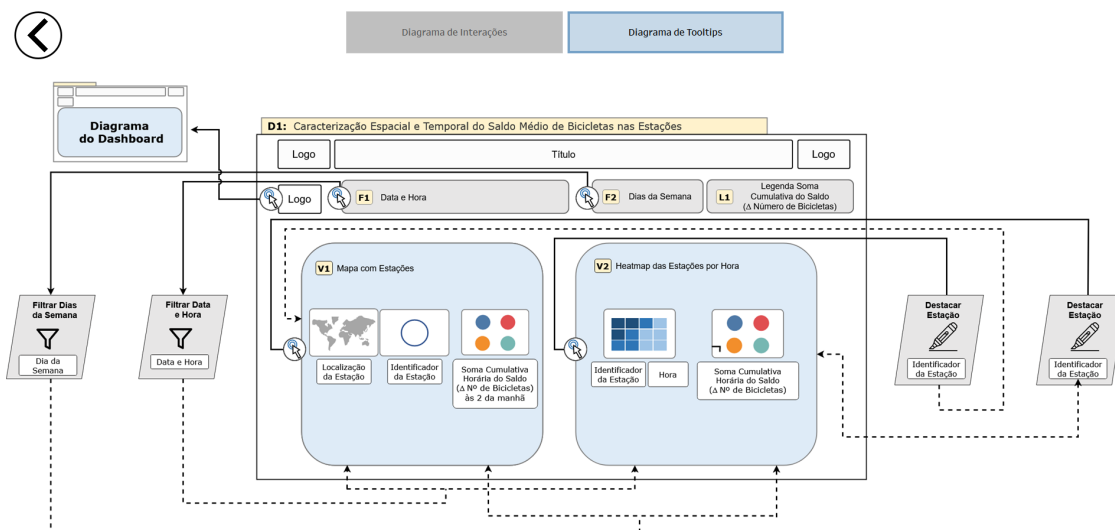


Figura 4.29: Diagrama de Interações para o *dashboard* de caracterização espacial e temporal do Saldo Médio de Bicicletas das Estações Gira.

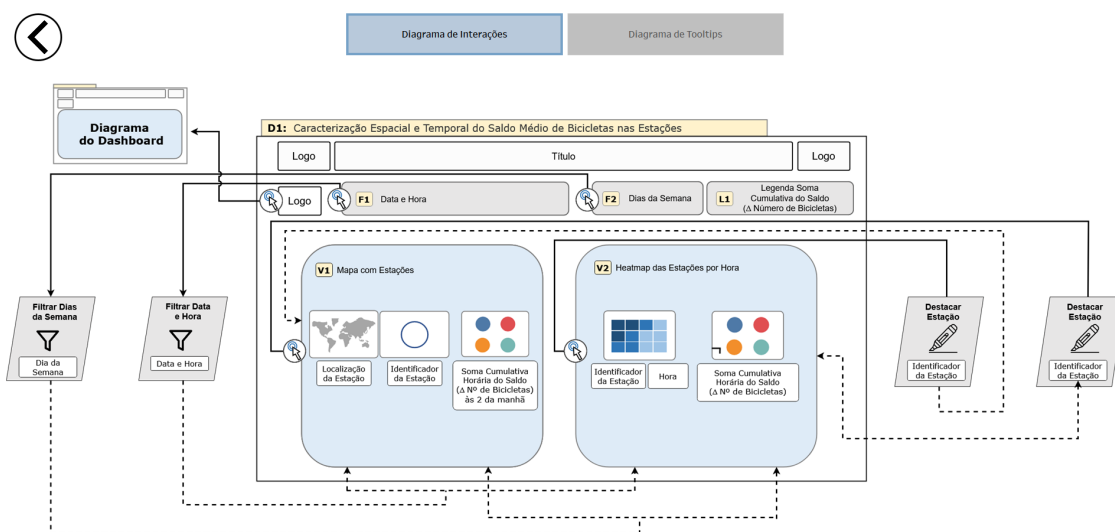


Figura 4.30: Diagrama de *Tooltips* para o *dashboard* de caracterização espacial e temporal do Saldo Médio de Bicicletas das Estações Gira.

estrutura idêntica ao da sub-secção 4.3.2, sendo composto por um mapa no qual se encontram as estações representadas por círculos, com a cor a mapear a média da diferença de bicicletas e um *heatmap* em que cada linha representa uma estação e cada coluna representa uma hora do dia, sendo que a cor mapeia a soma cumulativa da média da diferença de bicicletas ao longo do dia. As horas do dia encontram-se ordenadas das 3 da manhã às 2 da manhã, de forma a que o padrão diário tenha início na hora em que a

4.3. CARACTERIZAÇÃO ESPACIAL E TEMPORAL DAS ESTAÇÕES GIRA

utilização da rede é menor, o que ocorre durante a noite. Relativamente às estações do *heatmap*, representadas em cada linha, a sua ordenação é decrescente tendo em conta a variância da média da diferença de bicicletas, significando que as estações com a maior variância deste valor irão encontrar-se no cimo do *heatmap* e as estações nas quais este valor é menor irão encontrar-se na parte inferior. Com esta ordenação, é possível perceber quais as estações nas quais existe mais oscilação no número de bicicletas ao longo do dia e quais se mantêm quase constantes, o que pode significar que são auto-suficientes ou que é efetuado o rebalanceamento nas mesmas com frequência. Desta forma, a partir deste *dashboard* é possível perceber em que estações existe mais remoção de bicicletas e em quais existe mais adição das mesmas. No entanto, é importante referir que, neste caso, o comportamento observado tem em conta as ações de rebalanceamento efetuadas pelos funcionários, uma vez que o que está a ser avaliado é a soma cumulativa da diferença do número de bicicletas presentes numa estação e estes dados contêm essas ações.

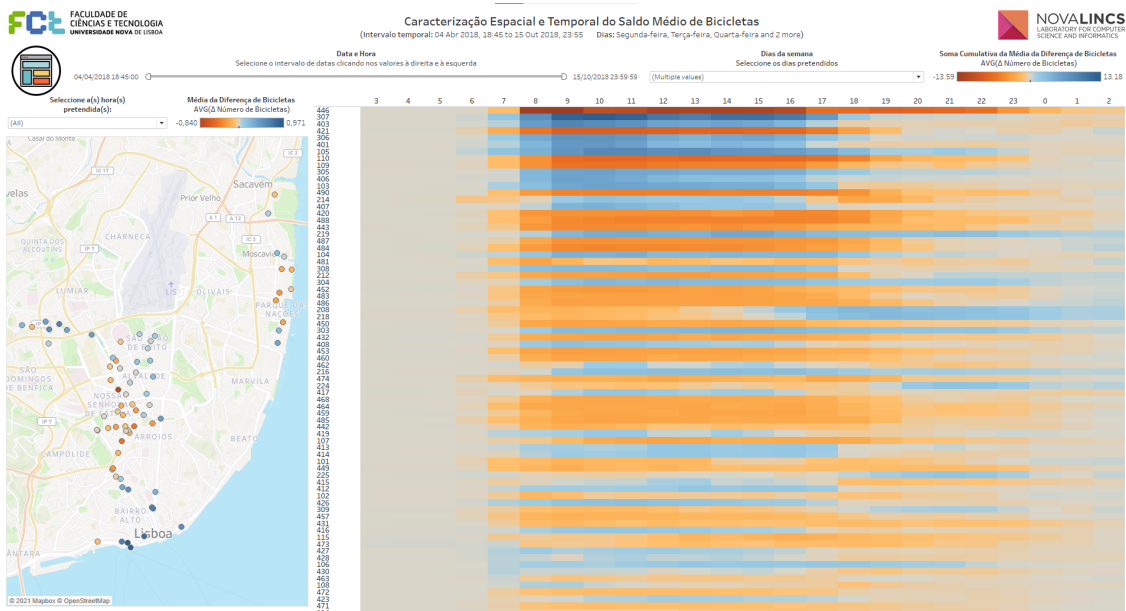


Figura 4.31: Dashboard de caracterização espacial e temporal do Saldo Médio da Diferença de Bicicletas para os dias úteis.

Para além do mapa e do *heatmap* que compõem este *dashboard*, este permite escolher um intervalo temporal que se pretende analisar a partir do *slider*, selecionar uma ou várias horas para as quais se pretende analisar o valor médio da Diferença de Bicicletas no mapa para todas as estações e ainda selecionar os dias que se pretende analisar, sendo possível escolher apenas a Segunda-feira, Sábado e Domingo, todos os dias úteis, entre outras combinações. Tal permite analisar diferentes dias separadamente, de forma a perceber qual o comportamento das estações nos mesmos. Fazendo *hover* sobre uma estação no mapa, é possível consultar o comportamento diário de cada estação ao longo dos meses, através de uma *tooltip* que contém, entre informações adicionais sobre essa estação como a morada, latitude e longitude, um *heatmap* relativo à estação em causa com a soma cumulativa

diária da diferença de bicicletas ao longo de todos os meses. Finalmente, é possível clicar no botão localizado no canto superior esquerdo para aceder aos diagramas do dashboard, sendo possível consultar o diagrama de interações (Figura 4.29), bem como o diagrama de tooltips (Figura 4.30).

O *dashboard* da Figura 4.25 é relativo apenas aos dias úteis, ou seja, de segunda a sexta-feira. Observando o mapa, é visível que tanto na zona da baixa como na zona de Te-lheiras as estações apresentam valores positivos de diferença de bicicletas, enquanto que no centro de Lisboa e no centro do grupo do Parque das Nações estas estações apresentam valores negativos para este valor, indicando que, no geral, estas têm mais tendência a ter falta de bicicletas, pois ocorre mais vezes a remoção de bicicletas, seja devido a viagens de utilizadores ou devido a ações de rebalanceamento.

Observando o *heatmap*, os valores da soma cumulativa da média da diferença de bicicleta encontram-se compreendidos entre -13,59 e 13,18, sendo perceptível que existem algumas estações que apresentam valores de soma cumulativa próximos de ambos os extremos da escala, apresentando uma tonalidade mais forte, não havendo muitas estações com valores próximos de zero. É possível perceber que a estação 446 é a que apresenta uma maior variância do valor médio da diferença de bicicletas, cujo valor é negativo. A segunda estação é a 307, localizada na zona da Baixa, que apresenta valores opostos à estação 446, ou seja, apresenta valores positivos para a diferença de bicicletas, indicando que, ao longo do dia costumam ser adicionadas bicicletas a esta estação, embora este valor diminua a partir das 18, altura em que este valor fica próximo de zero e, consequentemente, colorido de bege. O mesmo tipo de comportamento também foi observado na estação 446, na qual os valores também se aproximam do zero a partir do final do dia, neste caso, a partir das 23 horas. Observando o dashboard apenas com os dados relativos às viagens para a estação 446, na Figura 4.25, é visível que esta estação também apresenta valores da Soma Cumulativa do Saldo negativos, no entanto, estes mantêm-se até as 2 da manhã, pois estes não têm a influência do rebalanceamento. Tal pode sugerir que ocorre o rebalanceamento desta estação no final do dia, uma vez que no caso do *dashboard* da diferença de bicicletas os valores são negativos durante o dia mas ficam próximos de zero no final do mesmo, indicando que tende a haver a adição de bicicletas à estação no final do dia.

Observando as *tooltips* na Figura 4.32 relativas a algumas estações, de forma a obter mais informação relativamente ao seu comportamento ao longo dos meses, é possível observar que este é muito variável de estação para estação. Para referência relativamente à localização de cada uma das estações, esta encontra-se no mapa da figura 4.33. No caso da estação 103 (figura 4.32a), localizada na zona do Parque das Nações, o seu comportamento é semelhante em todos os meses, com a soma cumulativa da diferença de bicicletas a tomar valores positivos a partir das 7 da manhã e mudando para valores negativos de forma consistente a partir das 18 horas. O comportamento da estação 208 (figura 4.32b) varia ligeiramente ao longo dos meses, apresentando valores negativos nas horas da manhã, que vão ficando positivos a partir das 12 horas. Observa-se, no geral, algumas

4.3. CARACTERIZAÇÃO ESPACIAL E TEMPORAL DAS ESTAÇÕES GIRA

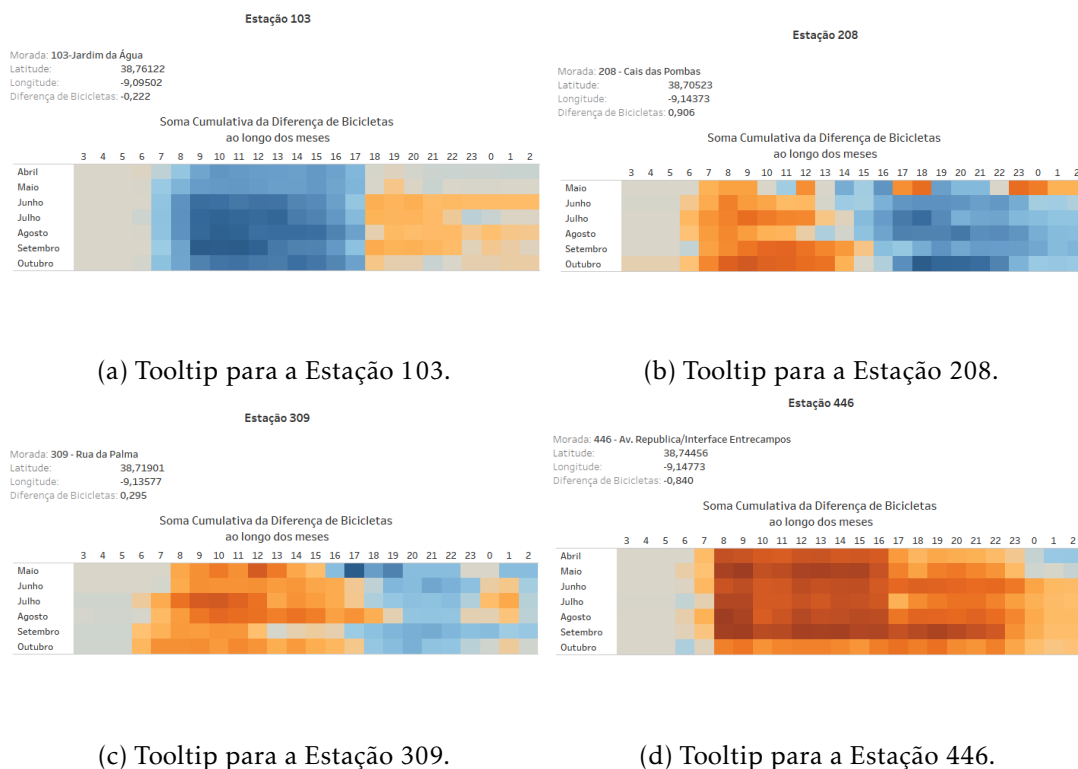


Figura 4.32: *Tooltips* do *dashboard* relativo ao Saldo de Bicicletas, acessíveis através da ação de *hover* das estações no mapa.

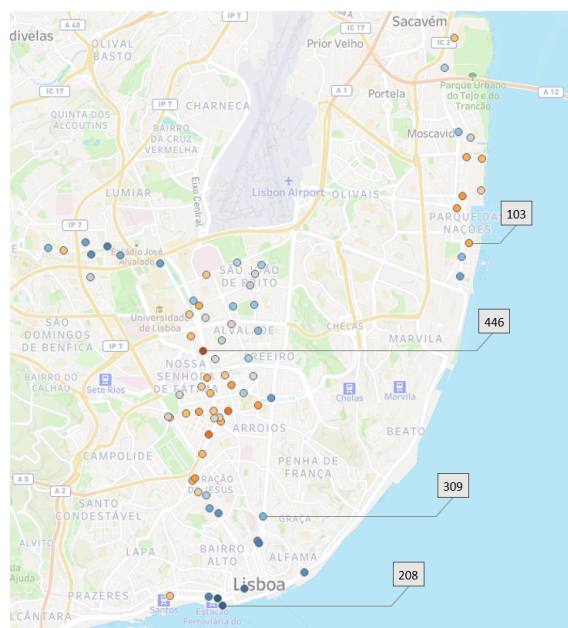


Figura 4.33: Localização das estações cujas *tooltips* se encontram na Figura 4.32.

diferenças entre os meses, nomeadamente relativamente à hora em que os valores passam de negativos para positivos. A estação 309 (figura 4.32c) apresenta um comportamento semelhante à estação 208, embora apenas fique com valores positivos da soma cumulativa da diferença de bicicletas a partir das 18 horas. Relativamente à diferença entre os meses, esta não é muito significativa. Finalmente, a estação 446 (figura 4.32d) apresenta valores negativos durante quase todo o dia, notando-se apenas que os valores se aproximam do zero a partir das 17 horas, embora no mês de Setembro tal comece a acontecer apenas por volta das 23 horas e, no mês de abril, a diferença de bicicletas atingiu valores positivos por volta da meia noite, o que coincide com o que foi observado anteriormente a partir da observação do *heatmap* com todas as estações. Observando a *tooltip* relativa às viagens, é perceptível que os valores da soma cumulativa do saldo ficam sempre negativos até ao final do dia, sugerindo, mais uma vez, que poderá haver rebalanceamento nesta estação, justificando a diferença existente entre as bicicletas e as viagens.

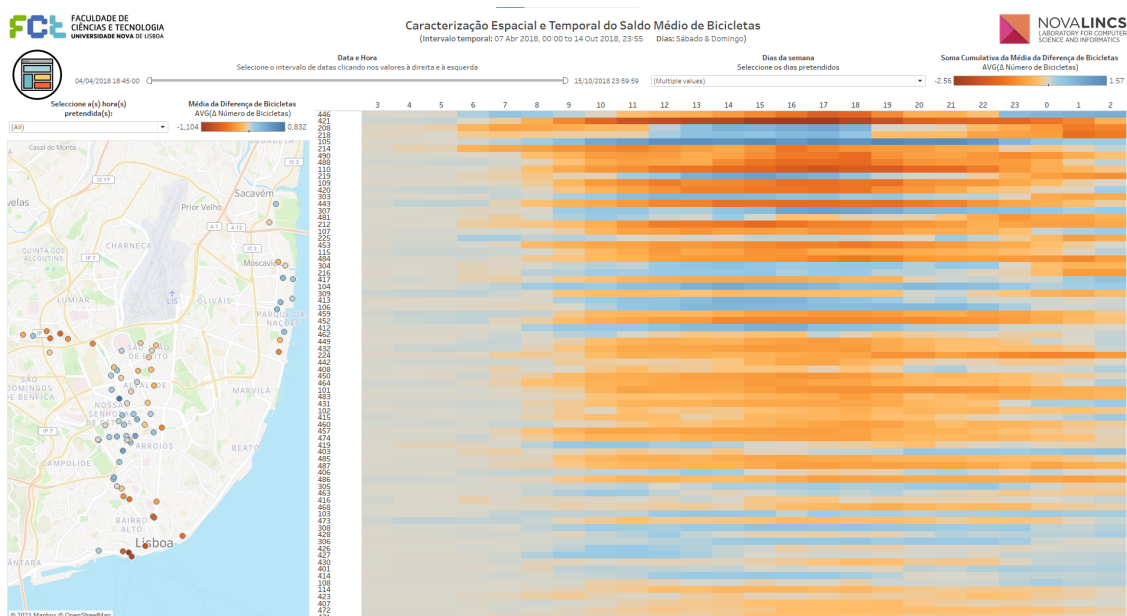


Figura 4.34: Dashboard de caracterização espacial e temporal do Saldo Médio da Diferença de Bicicletas para os dias de fim-de-semana.

Analisando apenas os dados relativos aos fins-de-semana, na figura 4.34, percebe-se que os valores da diferença de bicicletas são menores, estando compreendidos entre -2.56 e 1.57, o que significa que existe uma menor variação na diferença de bicicletas e, consequentemente, menor utilização da rede de bicicletas partilhadas. Para além disso, o comportamento da estação 446 é diferente relativamente aos dias da semana, visto que o valor da diferença de bicicletas atinge valores positivos durante a noite, a partir das 23 horas e, observando o mapa, percebe-se que a média desta diferença é, no geral, positiva. Para além disso, a zona da Baixa apresenta valores médios negativos, o que é o comportamento inverso ao que ocorre durante os dias úteis e o centro da cidade encontra-se com tons azuis, indicando uma diferença positiva de bicicletas. Tal sugere que, enquanto

que durante a semana existe uma maior tendência para sair do centro da cidade para a periferia, no fim-de-semana, existem mais utilizadores a deslocarem-se para o centro da cidade. Tal permite concluir que existe um comportamento diferente das estações nos dias úteis e nos dias não úteis, o que pode estar relacionado com a razão pela qual a rede é utilizada, havendo não só uma menor utilização nos fins-de-semana, mas também um fluxo da periferia para o centro.

4.4 Detecção do Rebalanceamento

O conjunto de dados relativo ao número de bicicletas presente em cada estação descrito na sub-secção 4.1.1 dá informação acerca do número de bicicletas que se encontram nas estações da rede Gira numa data e hora. No entanto, estes dados não permitem distinguir se a alteração no número de bicicletas numa dada estação foi devido à ocorrência de viagens ou se foi uma consequência de uma ação de rebalanceamento na mesma, visto que essa informação não está presente nos dados, nem foi possível obter essa informação.

Uma vez que se pretende efetuar a previsão do número de bicicletas em cada estação, é necessário perceber quando ocorreu o rebalanceamento nas estações, para que se possam responder às perguntas que se seguem:

- "É possível detetar o rebalanceamento efetuado pelos funcionários nos dados?"
- "Existem estações nas quais se constata uma maior influência de rebalanceamento?"

Desta forma, os dados podem ser representativos da utilização de cada estação ou, pelo menos, será possível ter acesso a dados sem a influência do rebalanceamento nos mesmos.

Relativamente ao processo de rebalanceamento efetuado nas estações, sabe-se que existem 5 veículos e 3 reboques para o propósito, sendo que cada veículo pode transportar 6 bicicletas e cada reboque tem capacidade para 12 bicicletas. A operação de rebalanceamento funciona durante todo o dia, não havendo um período do dia preferencial e esta é levada a cabo com base na manutenção e na experiência de operação por parte dos funcionários. Quanto ao número de bicicletas colocadas ou retiradas nas estações e a hora, tal é variável de forma a garantir que existem bicicletas nas estações. Tendo por base estas informações, é possível perceber que o rebalanceamento efetuado é, no mínimo, de 6 bicicletas, caso todos os veículos se encontrem em estações diferentes, podendo ser mais no caso de vários veículos se deslocarem à mesma estação.

Numa tentativa de remover a influência do rebalanceamento dos dados das estações, foi levada a cabo uma análise utilizando os dados das estações e das viagens em simultâneo, sendo que, para tal, foi necessário proceder à construção de um novo conjunto de dados.

Nos dados relativos às viagens, foi calculada a soma do número de partidas e chegadas de 5 em 5 minutos, tendo em conta os atributos da data de início e data de fim de

uma viagem, respetivamente, ou seja, a data de início de uma viagem numa dada estação representa a ocorrência de uma partida ou remoção de uma bicicleta dessa estação e a data de fim de uma viagem indica que a viagem terminou nessa estação e, como tal, houve a chegada ou devolução de uma bicicleta a essa estação. Utilizando os dados das estações, cujos registos representam o número de bicicletas em cada estação em intervalos de tempo de 5 em 5 minutos, foi efetuada a junção dos dados relativos às viagens, com a soma efetuada para as partidas e chegadas, de acordo com o atributo temporal referido.

Uma vez que o objetivo é tentar reproduzir os dados relativos às estações sem a influência do rebalanceamento, é necessário fazer uma combinação entre o número de bicicletas que se encontram numa dada estação e o número de partidas e chegadas que ocorreram. Para tal, foi calculada uma nova coluna nos dados que representa a diferença entre partidas e chegadas (Chegadas - Partidas), à qual se denominou de Saldo, tendo sido atribuído um sinal negativo a todas as partidas, uma vez que estas simbolizam a remoção de uma bicicleta da estação, resultando numa diminuição do número de bicicletas disponíveis e atribuiu-se um sinal positivo a todas as chegadas, por estas representarem a adição de bicicletas a uma estação.

Utilizando os valores relativos à coluna do Saldo, foi adicionada outra coluna na qual se procedeu à soma dos valores do Saldo ao número de bicicletas disponíveis em cada estação. No entanto, se estas operações fossem efetuadas ao longo do tempo, não iria ser possível observar qual o padrão diário da estação, pois seria como se as estações tivessem capacidade infinita. Como tal, na nova coluna calculada, o valor foi reiniciado todos os dias às 2 da manhã, passando a corresponder ao número de bicicletas que se encontravam na estação naquela altura, de forma a tentar minimizar este problema. Esta hora foi selecionada devido ao facto de ser no final do dia e a uma hora de baixa utilização da rede.

Na tabela 4.6 é possível observar um exemplo do conjunto de dados. Neste caso, para simplificar, a granularidade temporal utilizada foi horária apenas para auxiliar à explicação. Às 2 da manhã é consultado o número de bicicletas que se encontravam na estação e, de seguida, tendo em conta os valores do saldo, é adicionado esse valor ao número de bicicletas até chegar novamente às 2 da manhã, altura em que este número é reiniciado para a quantidade de bicicletas que se encontravam na estação. A coluna relativa ao número de bicicletas apenas apresenta o valor relativo às 2 da manhã pois é apenas esse que é utilizado no cálculo do novo atributo.

Em suma, esta nova coluna, denominada de Soma Cumulativa do Número de Bicicletas e Saldo, apresenta o valor correspondente ao número de bicicletas numa dada estação todos os dias às 2 da manhã, ao qual é somado de forma cumulativa o valor do Saldo que, por sua vez, é obtido efetuando a diferença entre as Chegadas e partidas (Chegadas - Partidas). Uma vez que o valor do saldo pode ser positivo ou negativo, se este for subtraído diariamente ao número de bicicletas que se encontravam na estação numa dada hora, o resultado obtido irá reproduzir os dados das estações, indicando uma aproximação do número de bicicletas que se encontram na estação tendo por base o número de viagens efetuadas, com a diferença de que estes valores não irão estar limitados à capacidade da

4.4. DETEÇÃO DO REBALANCEAMENTO

Estação	Hora	Número de Bicicletas	Saldo	Soma Cumulativa do Número de Bicicletas e Saldo
101	02:00	15	0	15
101	03:00		0	15
101	04:00		0	15
101	...		...	...
101	18:00		5	20
101	19:00		4	24
101	20:00		2	26
101	21:00		-3	23
101	22:00		-1	22

Tabela 4.6: Exemplo do raciocínio do conjunto de dados relativo à Soma Cumulativa do Número de Bicicletas e Saldo.

estação.

Com base na informação gerada, foi elaborado um dashboard para detetar o rebalanceamento em cada estação da rede da Gira, que pode ser observado na figura 4.35, com o objetivo de analisar e detetar o rebalanceamento nas estações da rede da Gira. Este é

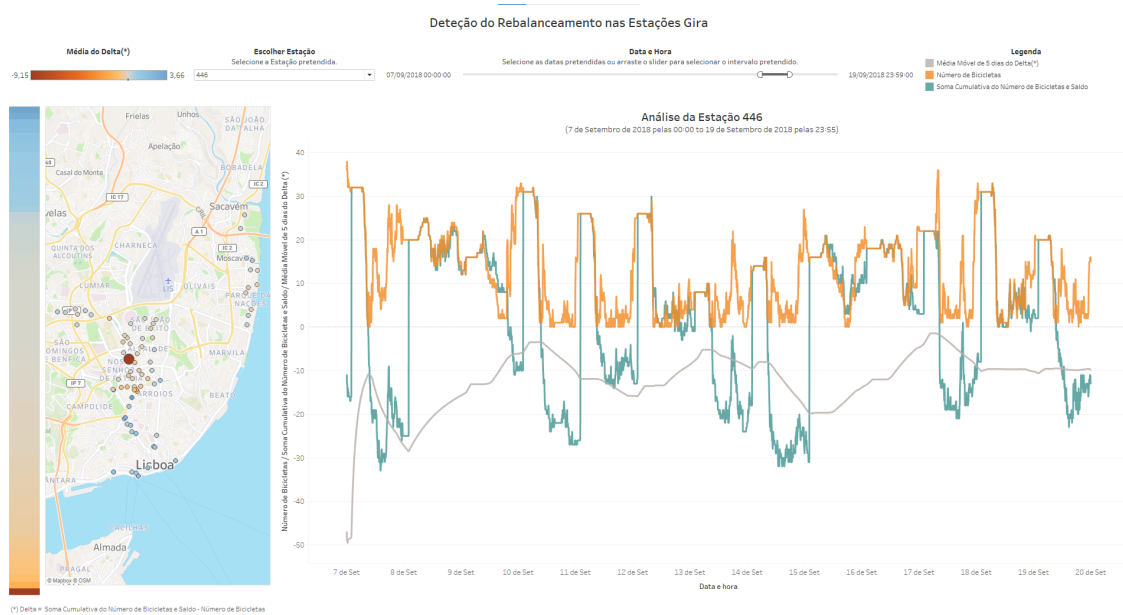


Figura 4.35: Dashboard para Análise e Detecção do Rebalanceamento nas Estações da Rede Gira, com ênfase na estação 446.

composto por três visualizações principais: a primeira, localizada à esquerda, é um *heatmap* em que cada quadrado é representativo de uma estação, cuja cor indica a média geral do Delta, uma medida obtida efetuando a média da subtração do Número de Bicicletas à Soma Cumulativa do Número de Bicicletas e Saldo (Soma Cumulativa do Número de Bicicletas e Saldo - Número de Bicicletas). O mapa localizado à direita do *heatmap* referido

mostra todas as estações da rede que podem ser alvo de análise, com a cor a representar, mais uma vez, a média do valor de Delta. A última visualização, localizada à esquerda, é um gráfico de linhas com informação relativa ao Número de Bicicletas (laranja) que se encontrava na estação, a Soma Cumulativa do Número de Bicicletas e Saldo (verde-água) e a média móvel de 5 dias (cinzento) do Delta (Soma Cumulativa do Número de Bicicletas e Saldo - Número de Bicicletas).

O *heatmap* permite efetuar uma interpretação geral e rápida da diferença entre o número de bicicletas e o sinal criado, sendo que, quando este valor é negativo tal significa que o sinal relativo à Soma Cumulativa é, no geral, menor relativamente ao número de bicicletas que se encontravam na estação e, caso este valor seja positivo, tal indica que o número de bicicletas é menor relativamente ao sinal da Soma Cumulativa, indicando que sem rebalanceamento a estação ficaria mais cheia do que o que é verificado se se observar o número de bicicletas. Este sinal identifica, principalmente, as estações nas quais existe uma maior diferença entre o sinal do número de bicicletas e o sinal criado, podendo indicar em que estações há uma maior influência de rebalanceamento nos dados das Estações, visto que, se essas estações forem alvo de rebalanceamento com maior frequência irá haver uma discrepância entre ambos os sinais. Para além disso, o sinal criado não contempla o número de docas existentes nas estações e, como tal, estes valores poderão ultrapassar o número de docas e tomar valores abaixo de zero.

À semelhança do *heatmap*, o mapa também permite interpretar de forma geral a diferença do sinal criado e do número de bicicletas e, para além disso, a componente geográfica permite observar as estações de uma forma diferente, sendo possível por exemplo, verificar que na zona da Avenida da Liberdade e Baixa, tende a existir uma diferença positiva entre ambos os sinais, indicando que, provavelmente, o rebalanceamento que é efetuado consiste na remoção de bicicletas. Já no centro de Lisboa, a situação é a inversa, indicando que, provavelmente, o rebalanceamento se baseia na adição de bicicletas.

Utilizando o *dropdown* localizado no cimo do *dashboard* é possível seleccionar uma estação para analisar no gráfico de linhas, sendo ainda possível modificar o intervalo de tempo que se encontra em análise utilizando o *slider*. Na figura 4.35, a análise está a ser efetuada para a estação 446 entre 7 e 19 de Setembro. Para além de se observarem diferenças entre a Soma Cumulativa e o Número de Bicicletas, uma vez que o sinal relativo à Soma Cumulativa atinge valores negativos na ordem das -30 bicicletas, é visível que a média móvel dos último 5 dias se mantém em valores negativos, indicando que o padrão observado também é verificado nos cinco dias anteriores. Analisando os sinais mais detalhadamente (4.36), é visível que, no dia 10 de Setembro pelas 23h40, enquanto que a Soma Cumulativa se mantém constante, o Número de Bicicletas existentes na estação aumenta de zero (às 23h40) para 26 (às 00h00), indicando a adição de 26 bicicletas à estação, o que não se verifica no sinal da Soma Cumulativa, indicando que este aumento no número de bicicletas não foi resultante de viagens efetuadas por clientes, indicando a existência de uma ação de rebalanceamento. Este padrão volta a repetir-se no dia seguinte, por volta das 22h20 de dia 11 de Setembro, quando também se verifica um aumento no

número de bicicletas, que passa de 1 para 26 pelas 01h50 de dia 12. Verificando o sinal da Soma Cumulativa, embora se verifique que houveram algumas viagens, a sua totalidade não resultaria numa adição de tantas bicicletas à estação, indicando, mais uma vez, um possível rebalanceamento.

Analisando os dias 15 e 16 de Setembro para a estação 446, que correspondem ao sábado e ao domingo, cuja informação se encontra na figura 4.37, é possível verificar que os sinais são muito semelhantes, verificando-se um decréscimo no número de viagens ocorridas. No entanto, existe novamente um aumento no número de bicicletas pelas 21 horas de sábado que é incongruente com o sinal da Soma Cumulativa sinalizado na figura com o número 1 e, mais uma vez, no domingo pelas 23 horas, sinalizado com o número 2. Verifica-se ainda, que na segunda-feira de manhã existe um aumento no número de bicicletas por volta das 8 da manhã (número 3 na figura), com a estação a ficar quase totalmente cheia (encontravam-se 36 bicicletas na estação, sendo que esta possui 40 docas), que também sugere que terá ocorrido o rebalanceamento da estação.

Apesar da análise efetuada mostrar alguns indícios de ações de rebalanceamento como os que foram referidos acima, esta não foi conclusiva, não sendo possível quantificar o número de bicicletas introduzidas ou retiradas das estações e em que registo horário tal ocorreu, não permitindo efetuar a remoção do rebalanceamento dos dados de forma fidedigna, sem haver possível adulteração da informação.

De acordo com a experiência de algumas pessoas que utilizavam a rede com alguma frequência, foi referido que eram avistadas carrinhas com bicicletas perto das estações na hora de mais fluxo durante a manhã e que as bicicletas vão sendo adicionadas ou

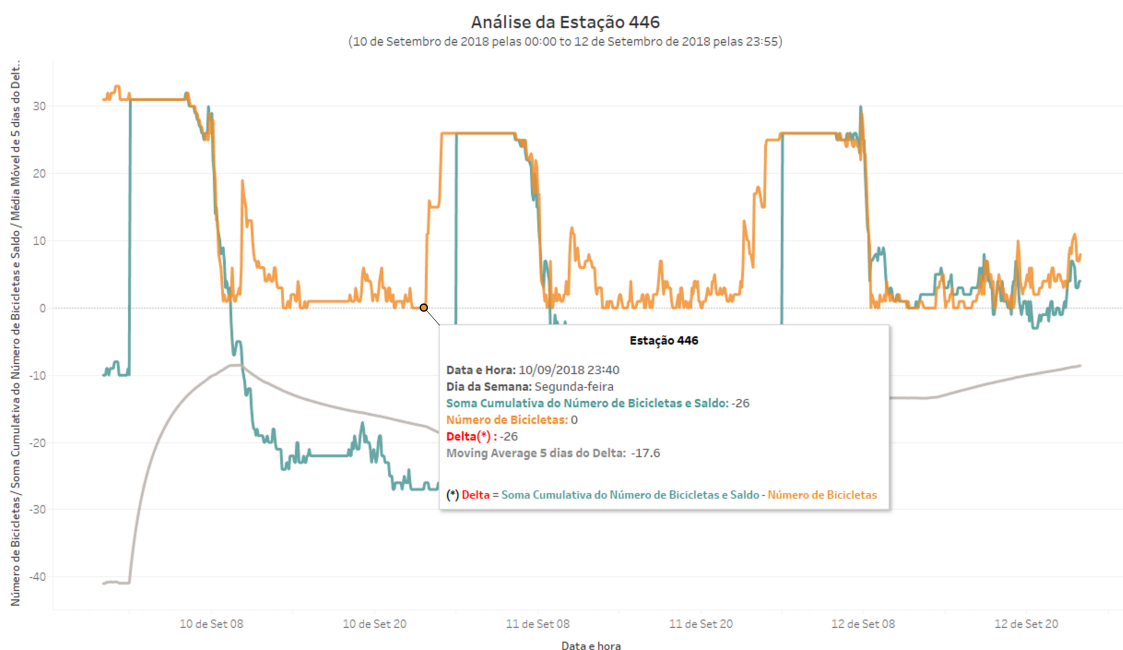


Figura 4.36: Análise e Detecção do Rebalanceamento na Estação 446 entre 10 e 12 de Setembro.

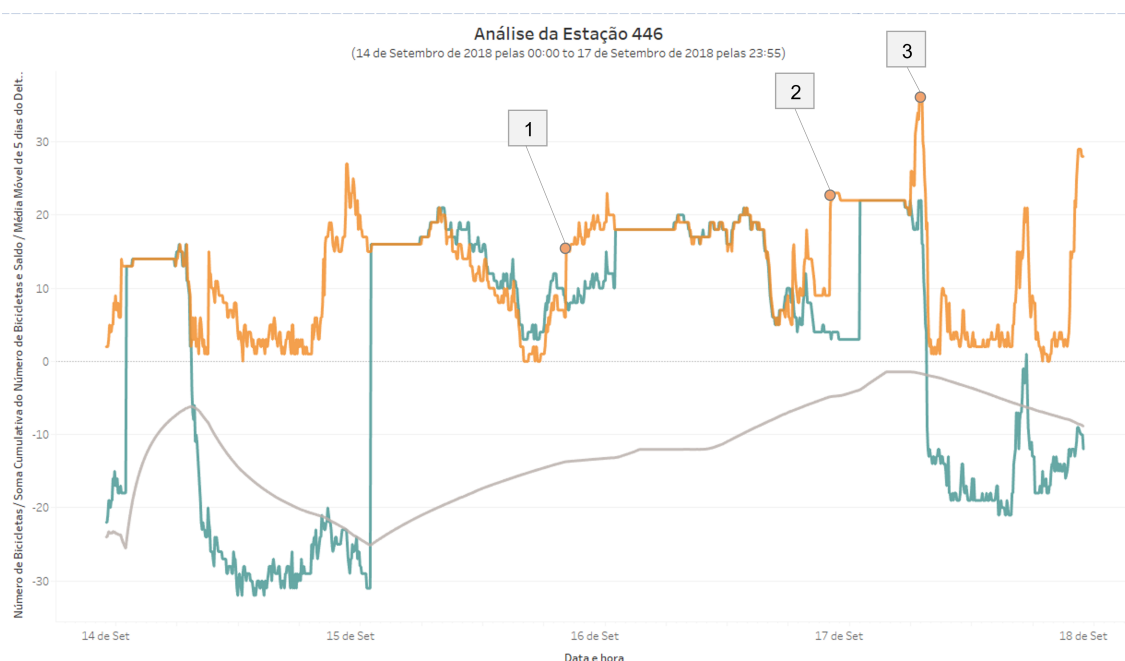


Figura 4.37: Análise e Detecção do Rebalanceamento na Estação 446 entre 14 e 17 de Setembro.

removidas conforme necessário. Para além disso, as informações fornecidas pela Câmara Municipal de Lisboa também indicam que o processo de rebalanceamento não apresenta um período do dia preferencial e que a quantidade de bicicletas colocadas ou retiradas é variável. Ora, durante a hora de maior fluxo nas estações, devido ao facto de haver uma maior utilização das estações para alugar ou colocar bicicletas, tal irá dar origem a uma variação do número de bicicletas nas estações num curto espaço de tempo no caso de o número de partidas e chegadas ser muito dispar, o que também torna mais difícil detetar pequenas alterações causadas pela inserção ou remoção de bicicletas por parte de funcionários que estejam a rebalancear essa mesma estação, tornando complicada a detecção da quantidade de bicicletas que foram inseridas ou removidas da estação durante essa altura.

Por estas razões, conclui-se que não é possível remover a influência do rebalanceamento dos dados das estações de forma fidedigna, pois este é variável não só no que diz respeito à quantidade de bicicletas removidas ou inseridas mas também relativamente à hora a que ocorreu, sendo ainda mais complicado detetá-lo nas horas de maior utilização da rede. Desta forma, apenas os dados das viagens não apresentam a influência do rebalanceamento, uma vez que estes apenas dizem respeito às viagens que ocorrem na rede de bicicletas.

MODELOS PREDITIVOS

Para otimizar o rebalanceamento nas estações da Gira é necessário prever o número de bicicletas que irão encontrar-se em cada estação numa determinada data e hora sendo que, para tal, é necessário decidir quais os dados a utilizar, bem como quais as variáveis externas que irão ser utilizadas de forma a fornecer mais informação.

Após este processo, é necessário proceder ao desenvolvimento de modelos preditivos, aplicando técnicas de *forecasting* e estatística que permitam não só efetuar uma otimização dos modelos preditivos, como também providenciar um *benchmark* para comparar resultados e analisar os mesmos.

Neste capítulo, serão descritos os dados utilizados para proceder à realização dos modelos preditivos, justificando o porquê de estes terem sido utilizados, tendo por base as análises realizadas no Capítulo 4. Depois, serão descritos os algoritmos utilizados, bem como as fases levadas a cabo ao longo deste processo, tendo sido efetuado uma primeira fase de validação e otimização e só depois desta os modelos finais.

Finalmente, devido ao facto de os dados conterem alguns *plateaus* cuja causa é desconhecida, foi realizada uma análise destes momentos em que não há viagens registadas na rede, de forma a tentar perceber se tal se deve às estações se encontrarem cheias ou vazias, ou se é devido à inexistência de viagens, ou seja, de clientes a utilizar a rede de estações da Gira.

5.1 Dados

Aquando da análise exploratória de dados, foi efetuada uma análise para proceder à deteção do rebalanceamento nos dados das Estações (Secção 4.4), cujos resultados não permitiram detetar com certeza os momentos em que ocorreu o rebalanceamento, nem quantificar a quantidade de bicicletas que foram colocadas ou retiradas das estações. Tal

significa que esses dados não poderão ser utilizados para proceder à previsão do número de bicicletas em cada estação, uma vez que o sinal que estaria a ser previsto incluiria a influência do rebalanceamento, não mostrando a necessidade real de bicicletas nas estações. Como tal, é necessário proceder à criação de um novo conjunto de dados que não apresente esta influência e possa representar adequadamente cada estação da rede Gira.

Este novo conjunto de dados foi criado utilizando os dados das viagens ocorridas na rede, que incluem informação acerca da data e hora início e fim de uma viagem, bem como do identificador da estação onde esta viagem começou e terminou. De seguida será descrito o processo levado a cabo para a criação do conjunto de dados. Primeiramente, foram calculadas tanto o número de partidas como o número de chegadas ocorridas em intervalos temporais de 15 minutos, tendo sido criados dois conjuntos de dados provisórios: um que representa a contagem de chegadas e outro que representa a contagem de partidas. O conjunto de dados das partidas tem três colunas com a estação de início, a data de início da viagem e a contagem em intervalos de 15 minutos. O conjunto de dados das chegadas também tem três colunas com a estação de fim, data de fim da viagem e a contagem de chegadas em intervalos de 15 minutos. Um exemplo dos conjuntos de dados pode ser consultado na tabela 5.1. Como se observa na tabela 5.1, o conjunto de

Data e Hora	Estação de Início	Número de Partidas
2018-01-25 15:45	101	2
2018-01-25 16:45	101	1
2018-01-25 18:15	101	2
2018-01-25 19:15	101	1

Tabela 5.1: Exemplo do conjunto de dados obtido com a contagem do número de partidas em intervalos de 15 minutos.

dados obtido apresenta lacunas, devido ao facto de não terem havido viagens em todos os intervalos de 15 minutos. Como tal, foi criado um conjunto de dados com uma frequência de 15 minutos, tendo sido feito um *join* e uma operação de *pivot* do conjunto de dados das partidas e outro para as chegadas, o que permitiu a obtenção de dois conjuntos de dados em que cada linha representa um intervalo de 15 minutos e cada coluna representa uma estação. Na tabela 5.2 encontra-se um exemplo do conjunto de dados das partidas após estas operações, sendo que, nos casos em que não houveram chegadas, os valores nulos foram preenchidos por zeros, sendo visível que há alguns valores a zero, devido ao facto de não terem ocorrido chegadas.

Depois, foi efetuada a subtração do conjunto de dados das partidas ao das chegadas, tendo em conta a data e hora, ou seja, se numa estação houveram 5 partidas no dia 2 de Junho pelas 20 horas e nessa mesma hora ocorreram 2 chegadas, o saldo será igual a -3. A razão pela qual se atribuiu um sinal negativo às partidas é pelo facto de uma partida representar a remoção de uma bicicleta de uma estação, resultando num decréscimo do número de bicicletas nessa mesma estação. As chegadas representam a adição de

Data e Hora	101	102	103	104	105
2018-01-25 15:30	0	0	0	0	1
2018-01-25 15:45	2	1	0	1	0
2018-01-25 16:00	0	0	0	0	1
2018-01-25 16:15	0	0	0	1	0
2018-01-25 16:30	0	0	0	0	0
2018-01-25 16:45	1	0	1	0	2

Tabela 5.2: Exemplo do conjunto de dados após o pivot obtido com a contagem do número de partidas em intervalos de 15 minutos.

bicicletas numa estação, tendo sido consideradas, por isso, como valores positivos. Após este processo, foi efetuada a soma cumulativa diária neste conjunto de dados, tendo-se escolhido como *reset* as 2 horas da manhã, devido ao facto de ser uma hora com baixa utilização do serviço e, como tal, representando uma baixa quantidade de viagens ocorridas. Tal significa que, a cada dia, com início às 2 horas, os valores do saldo vão sendo somados até à 1h45, reiniciando novamente às 2 da manhã, hora na qual ficam a zeros ou, caso tenha ocorrido alteração do saldo, ficam com esse respetivo valor.

Este conjunto de operações resultou num único conjunto de dados com 75 colunas, sendo que uma delas é a data e hora e as restantes dizem respeito às estações, com o valor do saldo cumulativo com reset às 2 da manhã de cada dia em cada data e hora para cada estação. As operações efetuadas aos dados apresentam como objetivo recriar o conjunto de dados das bicicletas que não pôde ser utilizado devido ao facto de incluir as operações de rebalanceamento embora, neste caso, este conjunto de dados não tenha em conta a capacidade das estações, no sentido em que o saldo de uma estação pode atingir valores negativos e que sejam superiores à capacidade da estação, ou seja, é como se cada estação tivesse uma capacidade infinita. A tabela 5.3 apresenta um exemplo do conjunto de dados descrito acima.

Data e Hora	101	102	103	104	105
2018-01-25 15:30	-2	0	0	0	1
2018-01-25 15:45	-1	1	-1	1	-1
2018-01-25 16:00	-1	0	-1	1	0
2018-01-25 16:15	-1	0	-2	2	0
2018-01-25 16:30	-2	0	-2	1	0
2018-01-25 16:45	-2	0	-1	0	2

Tabela 5.3: Exemplo do conjunto de dados após a soma cumulativa diária com *reset*.

Para além dos dados relativos às estações, com informação relativa ao saldo cumulativo diário, foram fornecidos no âmbito desta dissertação dados do IPMA, que serão utilizados para o desenvolvimento dos modelos preditivos, mais concretamente, dados relativamente à velocidade do vento, radiação, precipitação, humidade e temperatura na zona de Lisboa. Estes dados foram analisados relativamente à sua qualidade e foi efetuada um processamento, que se encontra no relatório do Apêndice C, tendo sido efetuada uma

breve descrição de ambas as operações na secção 4.1.3 do Capítulo 4. Em adição a estes dados, foi ainda adicionada informação relativamente à hora, dias da semana e feriados, sendo que uma coluna é numérica e representa a hora, outra coluna também numérica representa os dias da semana e, para distinguir entre dias da semana, fins-de-semana e feriados, foi criado um atributo booleano que distingue entre ambos.

A utilização de dados meteorológicos e dados relativos aos feriados, hora e dias da semana foi observada em diversos trabalhos relacionados com o *forecasting* da procura de bicicletas, tendo sido utilizada não só para analisar os dados numa primeira instância, mas também como variáveis externas aquando da previsão[38][40][4][32][39].

5.2 Análise dos Dados

Os dados relativos à soma cumulativa do saldo foram criados com base nos dados das viagens e correspondem ao ano de 2018, mais concretamente, entre os meses de janeiro e outubro, embora este intervalo temporal varie de acordo com as estações, uma vez que, para algumas estações, apenas há dados de Maio a Outubro, por exemplo. Tal deve-se ao facto de 2018 ter sido o primeiro ano da Gira, tendo ocorrido a adição de estações na rede ao longo desse ano.

De forma a poder analisar os dados a utilizar e a sua relação com os dias da semana, quer no que diz respeito aos 7 dias da semana, como também a fins-de-semana e feriados, foi desenvolvido um *dashboard*, que pode ser observado nas Figuras 5.1 e 5.2.

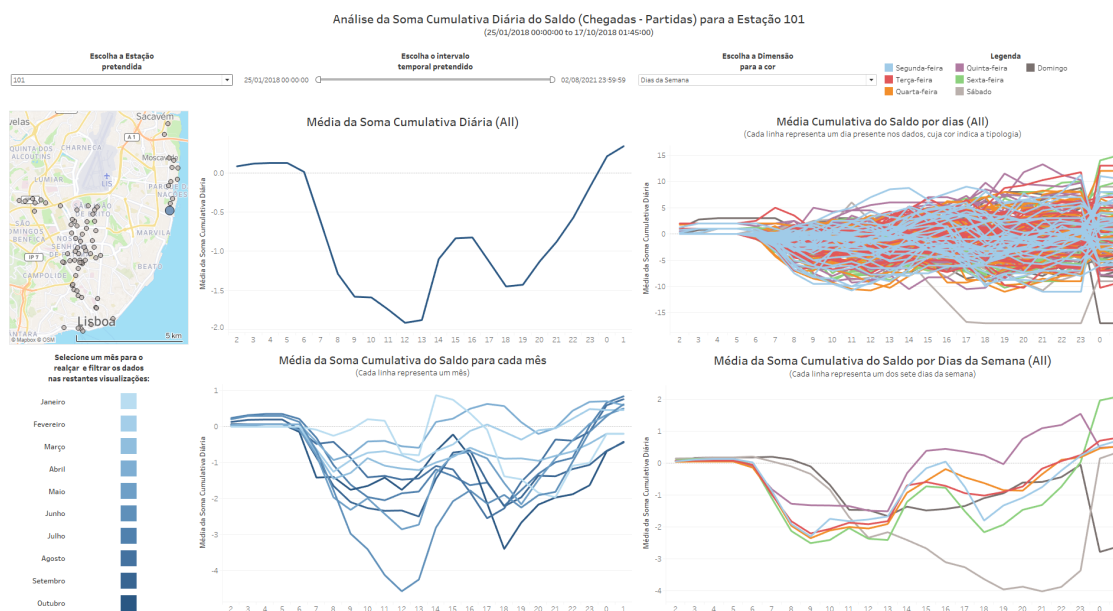


Figura 5.1: Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana para a Estação 101.

Na Figura 5.1, é possível observar os dados relativos à Estação 101, percebendo-se que esta estação existiu desde janeiro, existindo dados de janeiro a outubro. O *dashboard*

apresenta um mapa, no qual é possível observar todas as estações da rede, encontrando-se a azul a estação a ser analisada de momento, neste caso, a 101. Ao lado do mapa, encontra-se uma visualização com a Média da Soma Cumulativa Diária para a Estação 101, sendo a média de todos os meses para os quais existem dados. Em baixo desta visualização, é possível analisar a média diária para cada mês, sendo que o mês está mapeado para a cor, que se encontra em tons graduais de azul de forma a simbolizar a componente temporal dos dados. Nas visualização do lado direito e na parte superior do *dashboard*, cada linha representa um dia presente no conjunto de dados e a respetiva média da soma cumulativa ao longo desse dia, com a cor a indicar qual o dia da semana. Na visualização abaixo, existem apenas 7 linhas que caracterizam os sete dias da semana, apresentando a média de da soma cumulativa para todos os dias da semana do conjunto de dados, o que permite perceber se existe um padrão num dado dia da semana ao longo de todos os meses. Neste caso, é visível que ao Sábado existe sempre um decréscimo na soma cumulativa do Saldo, enquanto que nos restantes dias os valores seguem um padrão semelhante ao longo do dia.

No mesmo dashboard, é possível selecionar outra dimensão para a cor, no *dropdown* localizado no cimo do *dashboard*, que permite observar quais os dias que são dias úteis e quais os que são fins-de-semana ou feriados, obtendo-se as visualizações da direita na Figura 5.2. Neste caso, na visualização dos dias da semana, é possível existirem mais do que 7 linhas, no caso de ter existido um feriado que é simultaneamente um dia útil, de forma a ser possível comparar o padrão, sendo que, neste caso, é visível que existe uma diferença clara entre os dias úteis e dias que são fins-de-semana ou feriados. Nas Figuras

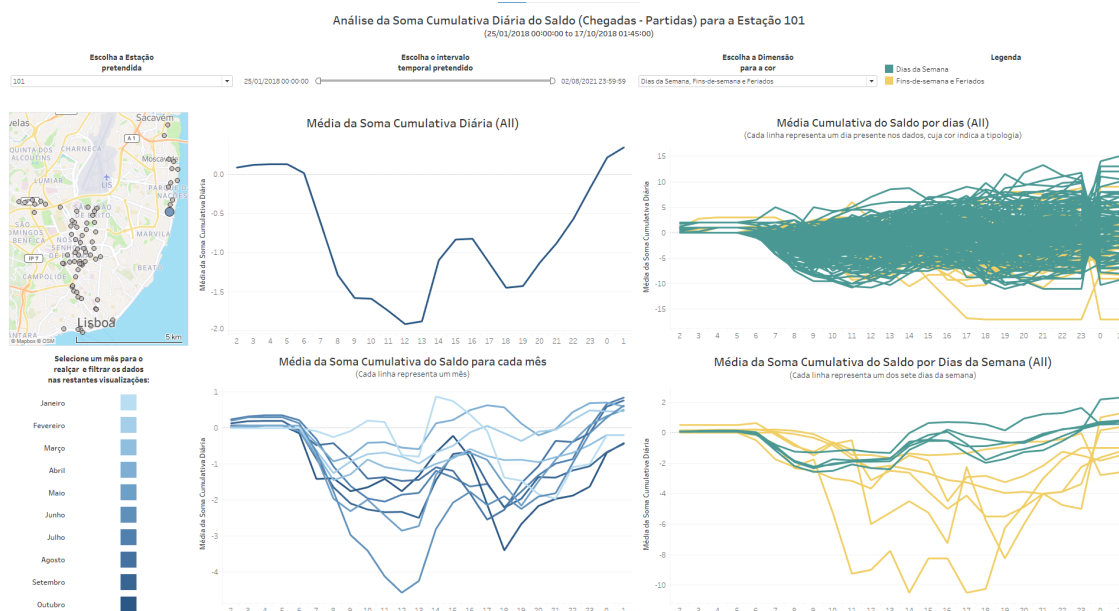


Figura 5.2: Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana, Fins-de-semana e feriados para a Estação 101.

5.3 e 5.4 encontra-se o dashboard descrito anteriormente mas, desta vez, com ênfase na

estação 218. É possível constatar que, para esta estação, o intervalo temporal de dados apenas está compreendido entre Maio e Outubro. Durante esses meses, o padrão da soma cumulativa diária foi semelhante à exceção do mês de Maio, em que houve um decréscimo da soma cumulativa entre as 19 e as 22 horas. Para além disso, a variação entre o fim-de-semana e os dias úteis não é tão diferente como no caso da Estação 101, sendo que este se mantém semelhante em todos os dias da semana. Analisando apenas os dias úteis e os feriados tal muda um pouco, visto que se verifica que, quando existe um feriado ou fim de semana o padrão de utilização varia um pouco mais relativamente aos dias úteis, cujo padrão se mantém praticamente inalterado. Tal poderá indicar que, em dias que são feriados, esta estação poderá ser utilizada para propósito de lazer.

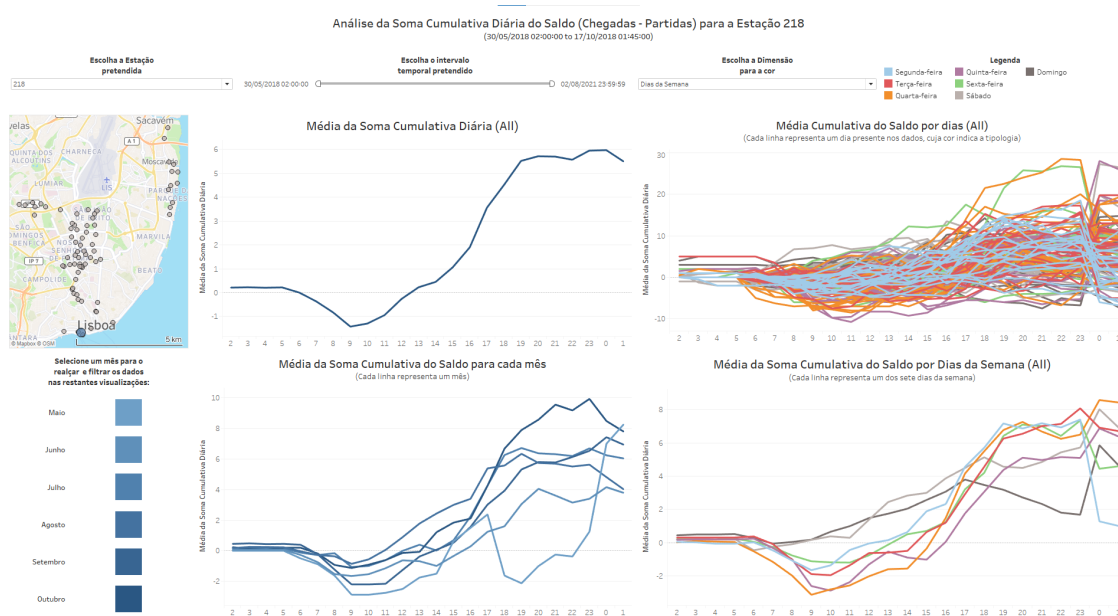


Figura 5.3: Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana para a Estação 218.

De forma a analisar os dados meteorológicos e das estações, foi criado o *dashboard* da Figura 5.5, no qual é possível consultar um mapa com todas as estações, uma visualização com a soma cumulativa diária ao longo dos meses para apenas uma estação e, na visualização em baixo à direita, é possível escolher qual o tipo de dados meteorológicos a consultar, observando também o seu padrão diário ao longo dos meses. Em ambas as visualizações o eixo do X representa as horas e cada linha representa um mês, diferenciado pela cor através de uma paleta de cores contínua de forma a simbolizar a componente temporal. Clicando nos meses da legenda à direita é possível realçar o mês ou meses pretendidos em ambas as visualizações. Neste caso, é possível observar que a temperatura apresenta um padrão diário semelhante, havendo um aumento da mesma por volta das 10 horas e um decréscimo pelas 20 horas. Esta varia também ao longo dos meses, sendo que os meses mais frios são os de inverno e os mais quentes os de verão.

5.2. ANÁLISE DOS DADOS

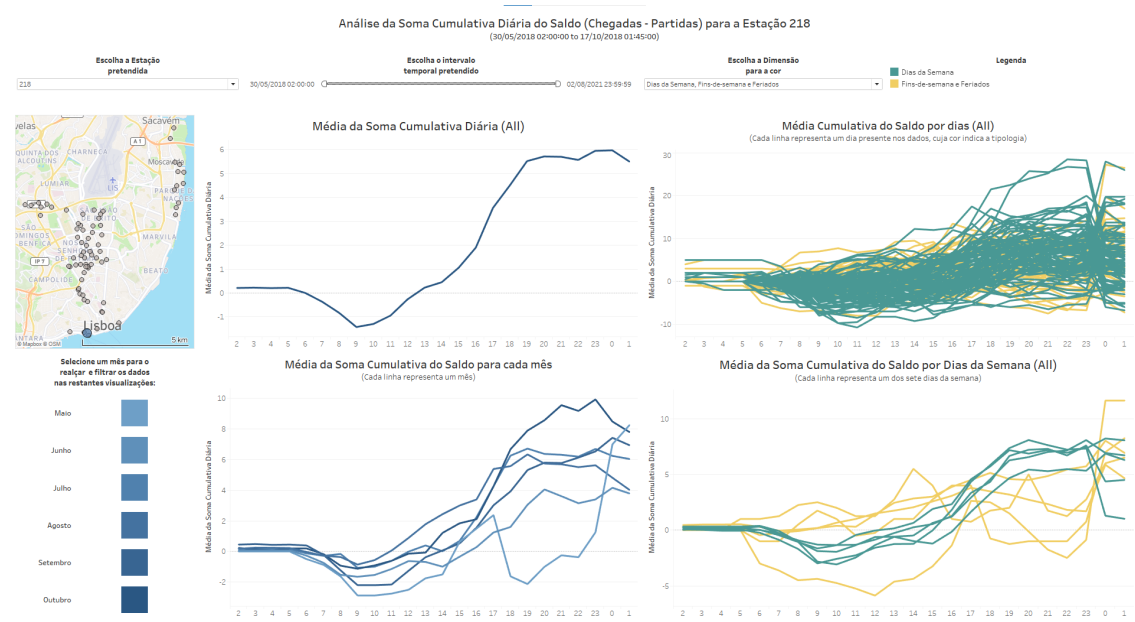


Figura 5.4: Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana, Fins-de-semana e feriados para a Estação 218.

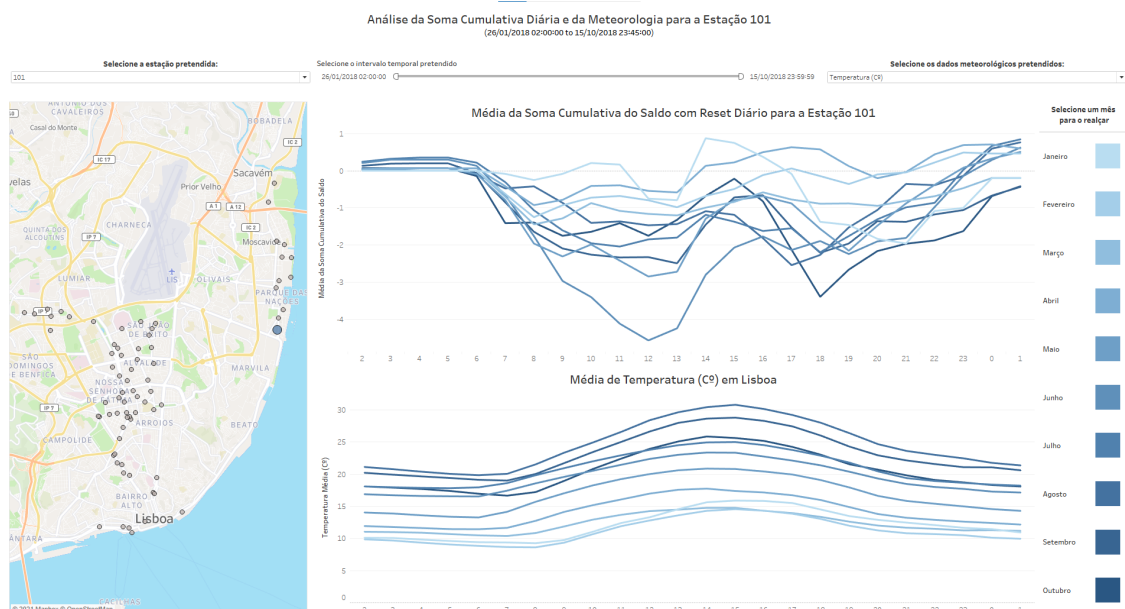


Figura 5.5: Dashboard de Análise da Média da Soma Cumulativa Diária e Dias da Semana para a Estação 101.

Nas Figuras 5.6, 5.7, 5.9, 5.8 e 5.10 é possível observar as visualizações para os restantes dados meteorológicos, nomeadamente a humidade, precipitação, intensidade do vento e radiação, bem como a legenda de cores para os meses utilizada (Figura 5.11). Verifica-se que, à exceção da precipitação, existe um padrão diário semelhante nos dados meteorológicos.

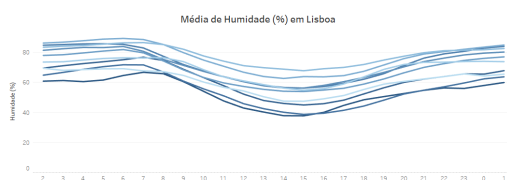


Figura 5.6: Média da Humidade (%) de janeiro a outubro de 2018.

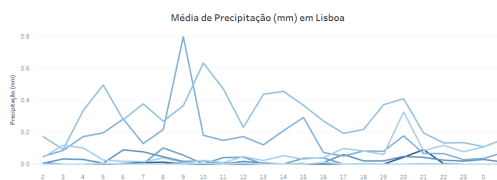


Figura 5.7: Média da Precipitação (mm) de janeiro a outubro de 2018.

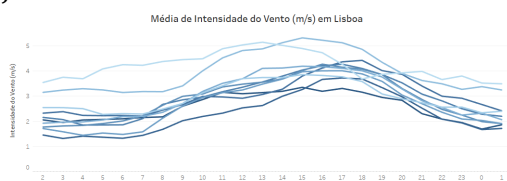


Figura 5.8: Média da Intensidade do Vento (m/s) de janeiro a outubro de 2018.

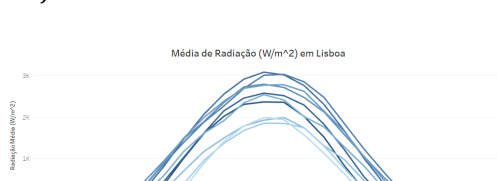


Figura 5.9: Média da Radiação (W/m^2) de janeiro a outubro de 2018.

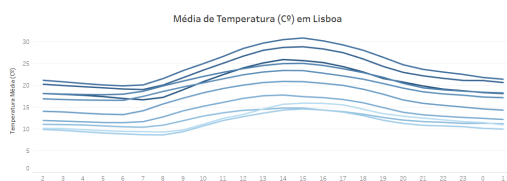


Figura 5.10: Média da Temperatura ($^{\circ}\text{C}$) de janeiro a outubro de 2018.

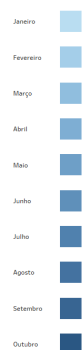


Figura 5.11: Legenda dos meses.

5.3 Construção e Otimização dos modelos

No âmbito desta dissertação foram desenvolvidos modelos preditivos específicos para cada estação, tendo-se utilizado redes neurais recorrentes, mais concretamente, Long Short Term Memory (LSTM). Para além disso, foram utilizadas as técnicas de *Naïve Forecasting* e *Moving Average* como *baseline*, de forma a poder comparar resultados. A escolha destes modelos deve-se ao facto de a técnica de LSTM ter sido utilizada em trabalhos relacionados consultados, tendo-se obtido bons resultados recorrendo à mesma. Para além disso, optou-se por utilizar também outras técnicas como *baseline* por ser também essa a

abordagem utilizada nos trabalhos mencionados. As técnicas de *naïve forecasting* e *Moving Average* permitem de forma simples estabelecer uma *baseline* de performance para os modelos a criar com o LSTM, ou seja, pretende-se utilizá-las como referência para atingir resultados melhores com a implementação do LSTM.

Foram utilizados os dados relativos às bicicletas, bem como dados meteorológicos e de dias da semana, fins-de-semana e feriados, cuja descrição e análise foi efetuada nas secções anteriores. Para proceder à criação dos modelos, foi efetuada a divisão do conjunto de dados em três sub-conjuntos, representativos do conjunto de treino, validação e teste. Estes três conjuntos de dados serão amostras do conjunto principal, que serão utilizados para treinar o modelo, avaliar o modelo enquanto se experimentam diferentes hiperparâmetros e perceber qual o modelo mais indicado a utilizar para efetuar a previsão final, no conjunto de teste. A divisão dos conjuntos foi efetuada tendo em conta uma percentagem, à semelhança do trabalho [31], o que se deve ao facto de o intervalo temporal de cada estação ser diferente, ou seja, há estação cujos dados são de janeiro a outubro e outras com dados de Abril e a Outubro. Neste caso, o conjunto de treino correspondeu a 80% dos dados, o conjunto de validação corresponde a 10% dos dados e o de teste também a 10%. Esta divisão recorrendo a uma percentagem costuma ser adotada para este tipo de projetos, tendo sido utilizado uma semelhante no trabalho [31]. Para além disso, uma vez que cada estação apresenta um período temporal diferente, desta forma é possível uniformizar a divisão do dataset e garantir que, tendo em conta os dados iniciais relativos a cada estação, a mesma percentagem é utilizada para treino, validação e teste. Os conjuntos de treino e validação serão utilizados para uma primeira experimentação com diferentes parâmetros e diferentes tipos de dados, sendo efetuada a previsão para o conjunto de validação e avaliado o desempenho dos modelos. Depois, serão selecionados os melhores parâmetros para cada estação e será efetuada a previsão no conjunto de teste e os dados de validação e treino serão unidos, tornando-se no conjunto de dados de treino.

Os dados relativos a cada estação encontram-se compreendidos, no máximo, entre janeiro e outubro de 2018 e, no mínimo, entre julho e outubro, o que levanta alguns problemas no que diz respeito à remoção da sazonalidade dos mesmos visto que, para além de não se possuir um ano completo dos dados, apenas existem dados relativos a um ano, sendo que, para ocorrer esta remoção, seriam necessários dados de, pelo menos, dois anos, uma vez que esta é removida utilizando o ano anterior para remover a sazonalidade no ano seguinte, utilizando este último para a previsão. Como tal, esta sazonalidade não foi removida. De seguida, serão descritos os processos utilizados para proceder à previsão do saldo em cada estação e serão mostrados os resultados relativos a algumas estações.

5.3.1 Naïve Forecasting

Esta técnica consiste em atribuir ao valor que se pretende prever o valor da última observação, ou seja, se o conjunto de dados tiver uma granularidade temporal de 15 minutos e se pretender prever o saldo para os próximos 15 minutos, esse valor será igual ao valor

do saldo nos 15 minutos anteriores.

Foi efetuada uma adaptação do método de *naïve forecasting*, recorrendo aos dados do saldo cumulativo com reset, ou seja, apenas dados das viagens, tendo sido utilizado o valor anterior do conjunto de validação para ir prevendo o próximo valor. Para além disso, foram efetuadas mais previsões para diferentes valores de n , sendo o n representativo de quantos *timesteps* serão recuados para obter a previsão. O objetivo desta variação é permitir comparar os resultados deste método com os resultados do método LSTM, no qual será previsto o valor do saldo para diferentes horizontes temporais. Neste caso, o conjunto de dados apresenta uma granularidade temporal de 15 minutos e os valores de n utilizados no *naïve forecasting* foram 1, 4, 8 e 12, que significam, respetivamente, 15 minutos, 1, 2 e 3 horas. Tal significa que, por exemplo, se for efetuado o *naïve forecasting* com um valor de n de 4, o valor previsto será igual ao valor do saldo na hora anterior.

Neste caso, o algoritmo foi efetuado para os conjuntos de treino e validação, de forma a avaliar o erro do conjunto de validação, sendo possível observar os resultados para a Estação 101 na Figura 5.12. Os dados relativos à estação 101 estão compreendidos entre

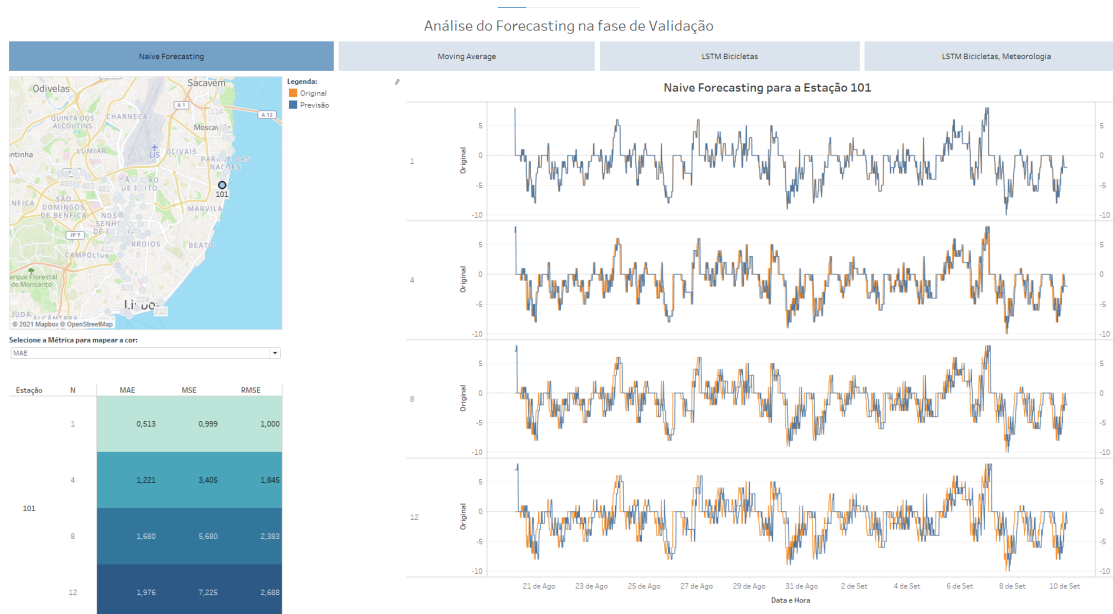


Figura 5.12: Dashboard para analisar os resultados do *naïve forecasting* por Estação, com ênfase na estação 101.

janeiro e outubro e, após a divisão do conjunto de dados em treino e validação (80% e 10%), sendo os restantes 10% para o conjunto de teste, o método foi aplicado no conjunto de validação que ficou compreendido entre Agosto e Setembro. O *dashboard* mostra um mapa com as estações, uma tabela com os valores das métricas para os diferentes tipos de *naïve forecasting* (tendo em conta o valor de n) e, finalmente, uma visualização onde se pode observar o valor real e o valor previsto, ou seja, o conjunto de validação no qual foi efetuada a previsão, também para os diferentes valores de N . As métricas utilizadas foram MSE, MAE e RMSE. Relativamente à previsão, é possível observar no gráfico de

linhas que a previsão vai ficando mais afastada do valor real conforme o aumento do N , uma vez que o aumento deste valor significa que a previsão é efetuada tendo em conta o valor do saldo num *timestep* mais distante no passado. A tabela mostra que o valor do erro também aumenta com o valor de N , com o **RMSE** a aumentar de 1, no caso do N igual a 1 para 2,68 para um valor de N igual a 12.

Observando outra estação como, por exemplo, a 446 (Figura 5.13), que se encontra localizada na Estação de Entrecampos, é possível observar que os valores de erro são muito mais elevados, o que também é uma consequência do facto dos valores da soma cumulativa do saldo também o serem. Para além disso, o erro vai aumentando conforme o aumento do valor de N , devido ao facto de ser uma previsão num futuro mais longínquo relativamente à do N igual 1, que é apenas para os próximos 15 minutos, o que não seria útil num cenário real. Tal permite perceber que o erro no caso da estação 446 também é superior ao da estação 101 pois esta estação apresenta valores do saldo superiores, devido ao facto de ser uma das estações mais utilizadas da rede da Gira.

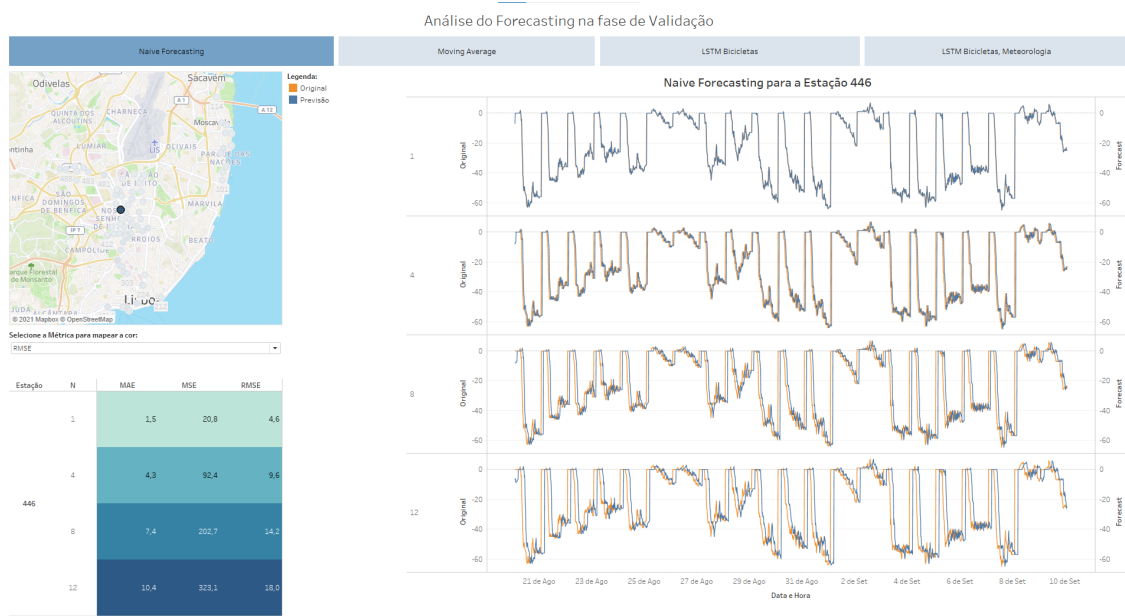


Figura 5.13: Dashboard para analisar os resultados do naïve forecasting por Estação, com ênfase na estação 446.

5.3.2 Média Móvel

O método da média móvel consiste em calcular uma série de médias de subconjuntos dos dados, apresentando como parâmetro o n , que representa o número de valores ou *timesteps* que irão ser utilizados para o cálculo dessa média. Esta técnica é denominada de técnica de smoothing, resultando numa suavização do sinal, uma vez que este passa a ser o resultado da média dos últimos n valores dos dados.

Utilizando os dados apenas relativos ao saldo cumulativo com reset, nomeadamente o conjunto de validação, foi sendo aplicado para cada *timestep* deste conjunto o método

da média móvel para calcular a previsão, tendo sido implementado este algoritmo para diferentes valores de N , de forma a poder comparar os resultados com o método de LSTM que será aplicado. Os valores escolhidos foram 4, 8, 12, 16 e 24 que, nos dados utilizados, que se encontram numa granularidade temporal de 15 minutos, representam uma média móvel de 1, 2, 3, 4 e 6 horas, respetivamente. Tal significa que a previsão é efetuada tendo em conta a média dos valores da soma cumulativa do saldo nas horas anteriores, dependendo do valor de n selecionado.

Foi também criado um dashboard para observar os resultados e as respetivas métricas utilizadas, sendo possível observar não só os valores do erro MAE, MSE e RMSE como também o gráfico de linhas correspondente a cada n com o valor original e o valor previsto. Na Figura 5.14, podem ser observados os resultados relativos à estação 101, sendo notório que, com o aumento do valor de N , o valor das métricas também sofre um aumento, o que é devido ao facto de se utilizarem valores cada vez mais longínquos para efetuar a previsão. Comparando os valores das métricas do moving average com o naive forecasting para os mesmos valores de n , observa-se que esta técnica apresentou valores de erro mais baixos, o que pode ser devido ao facto de a previsão ser baseada numa média de valores, em vez de ser apenas resultante do valor real nos n *timesteps* anteriores, resultando numa aproximação do valor previsto ao valor real.

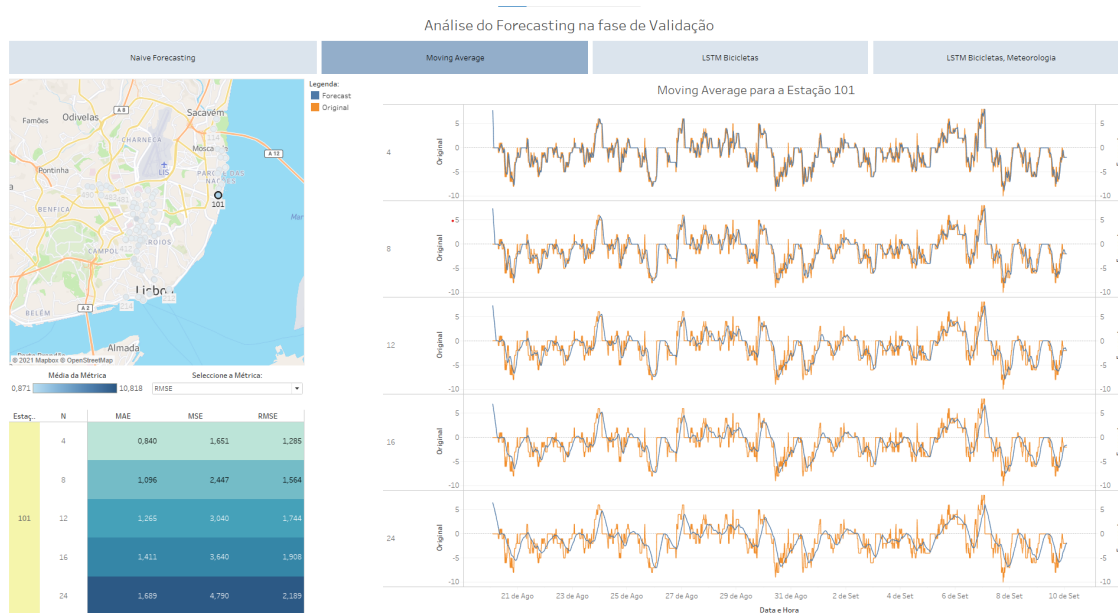


Figura 5.14: Dashboard para analisar os resultados do Moving Average por Estação, com ênfase na estação 101.

5.3.3 Long Short Term Memory (LSTM)

Até agora, nas técnicas implementadas, apenas foram utilizados dados relativos ao saldo cumulativo com reset das estações, no entanto, para a técnica de Long Short Term Memory (LSTM), foram utilizados diferentes tipos de dados, não só dados apenas relativos ao saldo

(univariável), como também os dados meteorológicos e informação relativamente aos dias da semana, à hora e informação relativamente a fins-de-semana e feriados, mencionados nas secções 5.1 e 5.2.

Para além da utilização de dados univariável e multivariável, o objetivo foi efetuar uma fase de validação, utilizando o conjunto de treino para treinar os modelos e o conjunto de validação para medir a performance dos mesmos, mediante a utilização de métricas. Nesta fase de validação, pretende-se efetuar uma experiência com diversos parâmetros não só do algoritmo, como também dos dados, de forma a encontrar o melhor modelo para cada estação.

Antes de proceder à construção dos modelos, foi necessário colocar os dados num formato aceite pela rede neuronal que, neste caso, é tridimensional, sendo composta por: [samples, timesteps, features]. As amostras são o número de sequências, os timesteps é um valor selecionado que irá servir como a "memória" que é consultada para efetuar a previsão, sendo que, neste caso, cada timestep corresponde a 15 minutos, devido a esta ser a granularidade temporal dos dados e as features representam o número de variáveis, ou seja, no caso de apenas se utilizar os dados das viagens, o número de features é 1. As sequências de dados formadas são duas, sendo que uma delas contém os dados que vão ser introduzidos na rede neuronal para o treino e a outra, normalmente denominada de label, é utilizada para confirmar a previsão efetuada pela rede, sendo depois utilizada para o cálculo do erro e consequente quantificação da performance do modelo. Neste caso, foram efetuados dois tipos de previsão: multi-step e single-step, sendo que este termo se refere às labels, ou seja, no caso de a previsão ser single-step, apenas será feita a previsão de um único valor e, no caso de ser multi-step, vários valores são previstos. No caso da previsão single-step, apenas um timestep é previsto, tendo sido efetuada a previsão de um único timestep nos próximos 15 minutos, na próxima hora e nas próximas 3 horas. A previsão multi-step foi efetuada com valores de 4 e 12 para as labels, correspondendo, de acordo com os dados utilizados, a uma e três horas, respetivamente, significando que, comparativamente à previsão single-step, neste caso foi efetuada a previsão do saldo durante a próxima hora (4 timesteps) e durante as próximas 3 horas (12 timesteps). Relativamente à janela de memória utilizada, esta foi de 24 timesteps, ou seja, correspondente a 6 horas, significando que a previsão efetuada teve em conta dados históricos das últimas 6 horas, quer na previsão single-step, quer na previsão multi-step.

Os parâmetros utilizados para proceder à otimização foram o número de neurónios, o batch size, ou seja, o número de amostras que são introduzidas na rede de cada vez, o número das labels, ou seja, o número de valores que serão previstos e, no caso dos dados utilizados, a quantidade de valores a serem previstos, dependendo se a previsão é de apenas um timestep ou de múltiplos.

Para analisar os resultados foi elaborado um dashboard que permite visualizar não só as estações de forma geográfica, como também os valores das métricas e a previsão conseguida, sendo composto por um mapa, no qual se encontram as estações da rede,

uma tabela com as métricas da previsão, bem como um gráfico de linhas no qual é possível observar o valor real e a previsão. Na Figura 5.15 encontra-se o dashboard relativo à previsão com apenas os dados relativos às bicicletas, neste caso, para a estação 101, tendo sido efetuado LSTM single-step, ou seja, apenas se preveu um valor. É possível observar que o valor do erro decresceu imenso relativamente às técnicas aplicadas anteriormente.

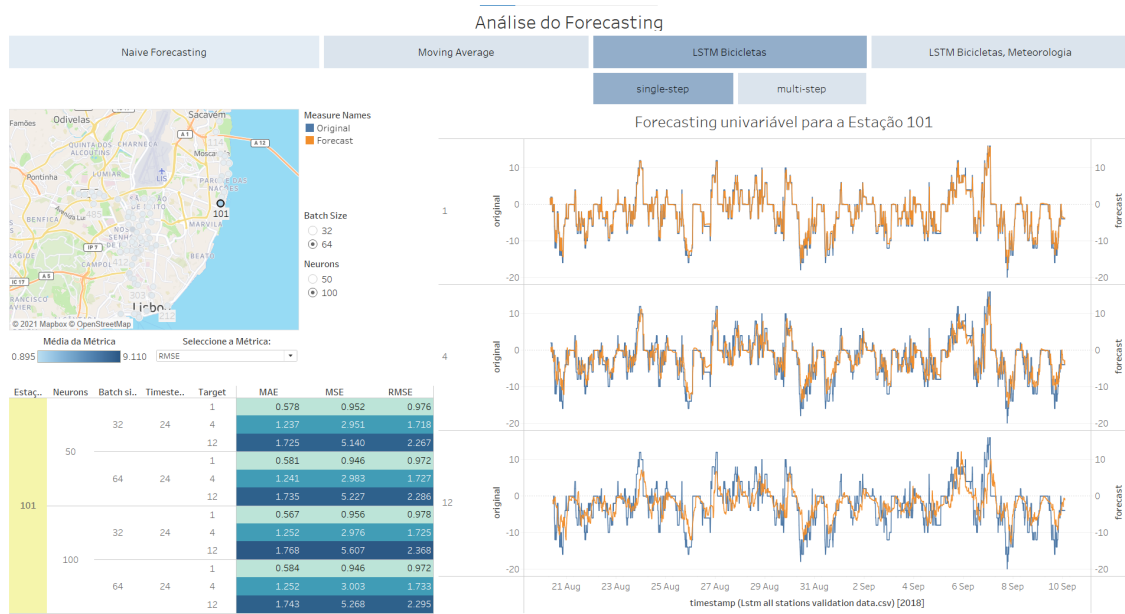


Figura 5.15: Dashboard para analisar os resultados do LSTM univariável single-step para a estação 101.

Na Figura 5.16 encontra-se o dashboard relativo à previsão com apenas os dados relativos às bicicletas, fazendo a previsão de múltiplos valores para a estação 101, sendo possível observar um ligeiro aumento do erro conforme se vai aumentando o número de valores a prever. No geral, os melhores resultados foram conseguidos utilizando um conjunto de dados multivariável, que contém informação não só relativa às bicicletas, como também relativa à meteorologia, dias úteis e não úteis e ainda à hora. O facto do conjunto de dados ser multivariável e, por isso, existir mais informação, permitiu efetuar a previsão recorrendo a mais informação, permitindo a obtenção de melhores resultados, como se pode observar nas Figuras 5.17 e 5.18, onde se encontram exemplos para a estação 101.

5.4 Resultados

Após a fase de validação, foi verificado, para cada estação, qual o valor mais baixo da métrica RMSE, bem como os parâmetros utilizados que permitiram atingir a mesma.

Utilizando esses parâmetros, foi efetuada a previsão para o conjunto de teste, sendo que cada rede neuronal foi treinada utilizando os dados de treino, tendo sido utilizado o conjunto de validação e utilizou-se o conjunto de teste para a previsão. No *dashboard* da Figura 5.19 é possível observar as melhores parametrizações para algumas estações,

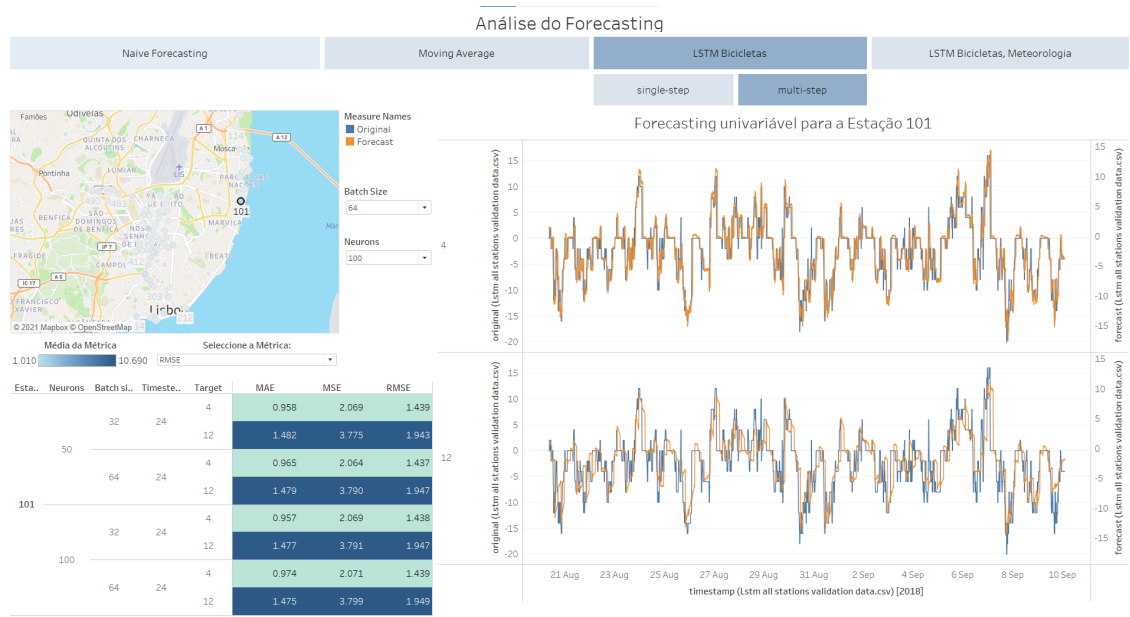


Figura 5.16: Dashboard para analisar os resultados do LSTM univariável multi-step para a Estação 101.

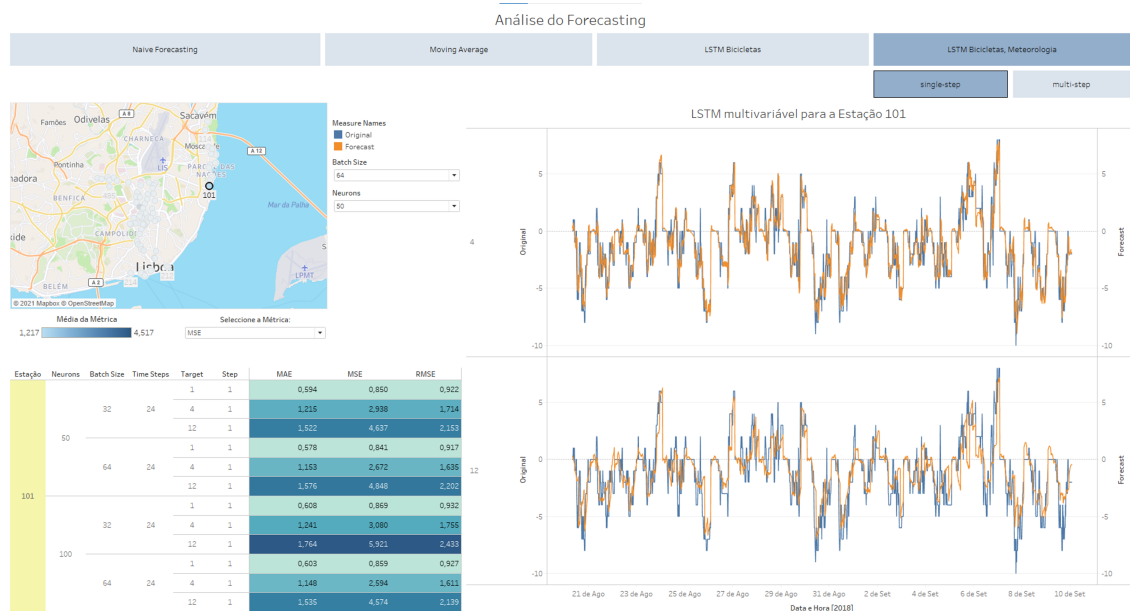


Figura 5.17: Dashboard para analisar os resultados do LSTM multivariável single-step para a Estação 101.

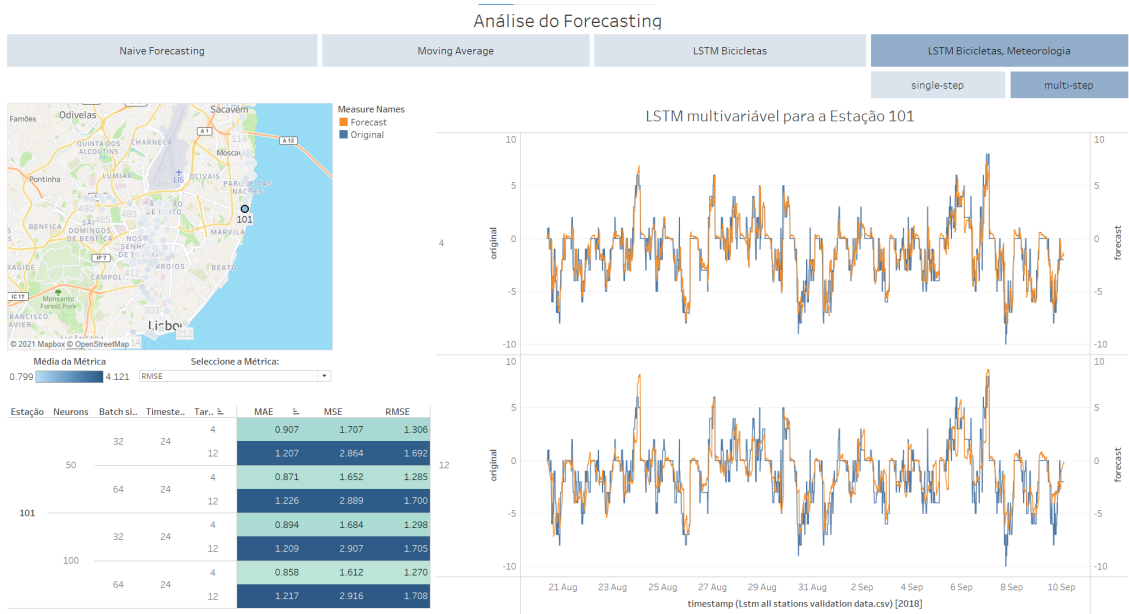


Figura 5.18: Dashboard para analisar os resultados do LSTM multivariável multi-step para a Estação 101.

sendo que foram estas as que foram utilizadas para efetuar a previsão final.

Ao longo da fase de validação verificou-se que, com o aumento do número de *labels* a prever, o erro aumentava ligeiramente, no entanto, é mais útil que a previsão seja efetuada para mais horas e, como tal, foi elaborada a previsão com a melhor parametrização para cada *label*, de forma a poder comparar resultados.

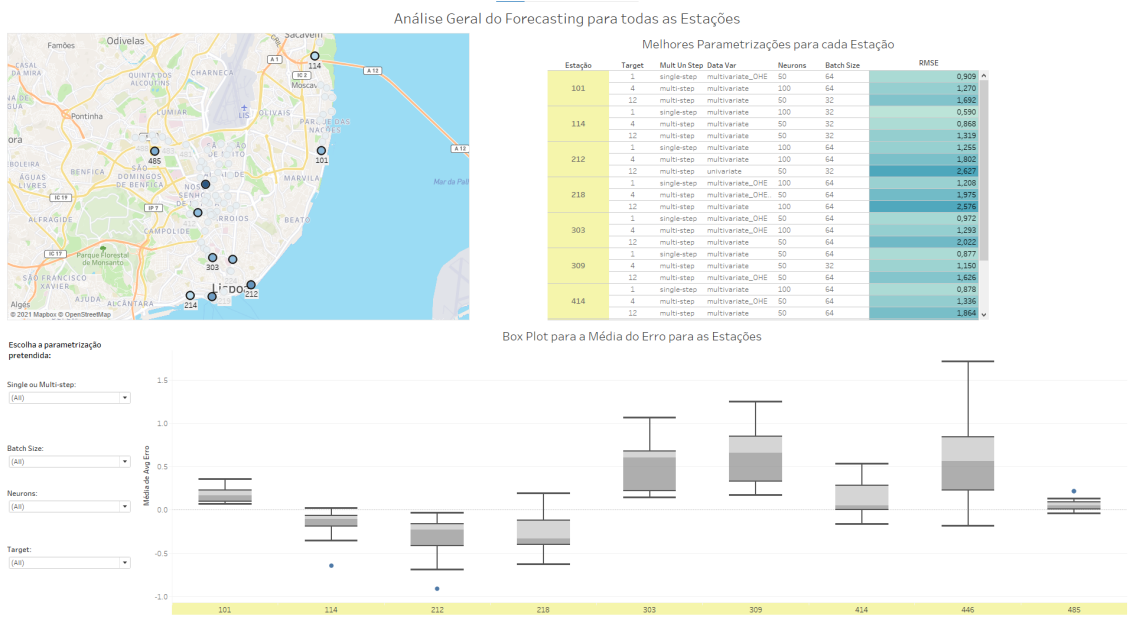


Figura 5.19: Visualização das melhores parametrizações para cada valor de target.

Como referido, esta previsão foi efetuada utilizando como conjunto de treino a junção

do conjunto de dados de treino e validação, efetuando a previsão no conjunto de teste, o que, no caso da Estação 101, significa que o conjunto de treino foi de janeiro a setembro e o conjunto de teste encontra-se compreendido entre Setembro e Outubro.

Na Figura 5.20 é possível observar um dashboard com os resultados da previsão para a estação 101 nos três horizontes de previsão (15 minutos, 1 hora e 3 horas), sendo perceptível que a previsão se encontra muito próxima do valor real, cujos os valores da métrica **RMSE** podem ser observados na tabela do dashboard da mesma figura. Embora os valores do **RMSE** sejam ligeiramente superiores aos da fase de validação, este aumento pode ser devido ao facto de os dados que estão a ser previstos serem relativos ao mês de Setembro e não existir informação nos dados relativa ao mês de Setembro do ano anterior. Para além disso, na análise dos dados na secção 5.2, é perceptível que, nos meses de Setembro e Outubro, a utilização da rede aumentou muito relativamente aos meses anteriores, o que também pode ter influenciado a previsão, embora, observando o gráfico, esta seja muito semelhante aos valores reais do saldo cumulativo.

Aquando do processo de previsão, houveram algumas dificuldades, nomeadamente devido ao conjunto de dados inicial, relativo às bicicletas, que continha informação relativa ao rebalanceamento efetuado por funcionários no mesmo, o que influenciava o valor de bicicletas que se encontravam nas estações numa dada hora e, como consequência, poderiam contribuir para uma previsão do número de bicicletas tendo em conta as operações de rebalanceamento que já eram efetuadas na rede. Tal não iria corresponder à necessidade real de bicicletas, o que resultaria numa previsão pouco fidedigna. Embora tenha sido criado outro conjunto de dados utilizando informação relativa às viagens, este diz respeito ao saldo cumulativo diário, não sendo relativo à quantidade de bicicletas e, para além disso, os dados são apenas relativos a alguns meses do ano de 2018, o que também não permitiu a deteção de sazonalidade nos mesmos. No geral, dadas as dificuldades encontradas devido aos conjuntos de dados, foram obtidos bons resultados, visto que os valores de **RMSE** obtidos, que se encontram nas mesmas unidades do conjunto de dados, são baixos, representando um erro de 1 a 2 pontos do saldo cumulativo.

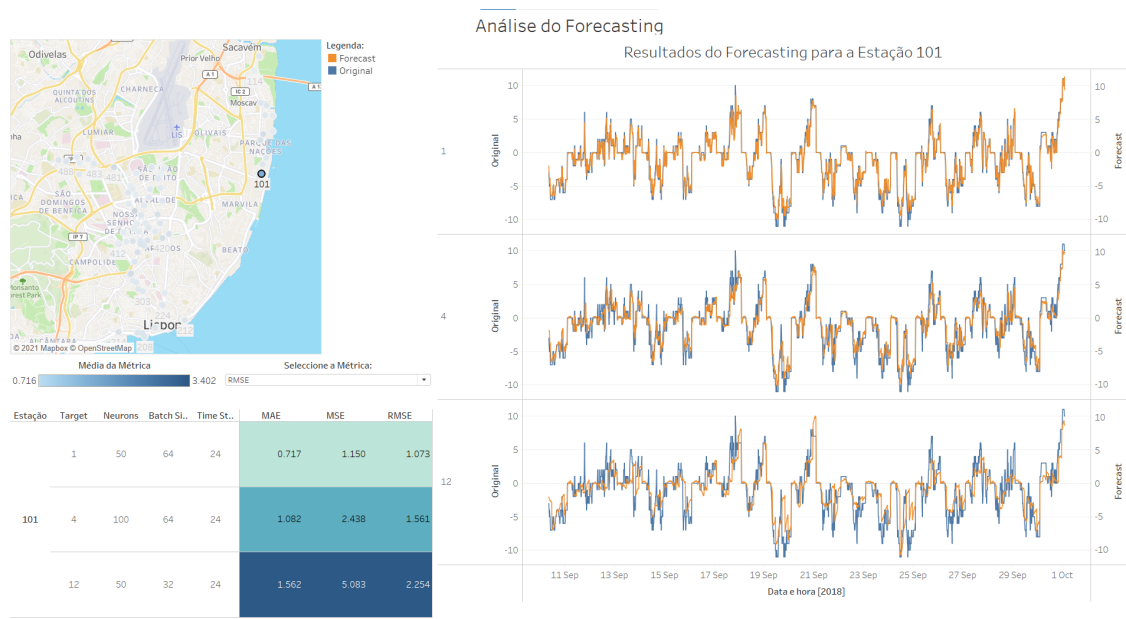


Figura 5.20: *Dashboard* com a previsão efetuada para a Estação 101.

5.5 Análise de Plateaus

Para proceder à previsão da procura de bicicletas nas estações da rede da Gira, foi criado um conjunto de dados tendo por base o saldo de bicicletas em intervalos de 15 minutos, que foi conseguido efetuando a subtração entre o número de partidas e chegadas para cada intervalo temporal.

Estes dados apenas contam com a influência das viagens ocorridas na rede, ou seja, cada estação funciona como se fosse de capacidade ilimitada e não há informação relativamente à sua ocupação, não sendo possível perceber se esta se encontra cheia ou vazia. Uma vez que os dados contabilizam apenas o saldo (chegadas - partidas), tal significa que se tiverem ocorrido um número igual de partidas e chegadas no mesmo intervalo de 15 minutos, tal irá resultar num valor de saldo igual a zero.

Uma vez que o conjunto de dados é a soma cumulativa diária desse Saldo, os momentos em que o saldo é igual a zero irão significar que a soma cumulativa desse valor se irá manter inalterada, resultando num *plateau*, ou numa linha plana. Este *plateau* ocorre porque o valor do saldo é igual a zero, o que ocorre quando o número de partidas e chegadas é igual. No entanto, este poderá ser igual porque ocorreu um número igual superior a zero de partidas e chegadas, ou porque estas não ocorreram, ou seja, porque houveram zero partidas e zero chegadas. É a última situação que tem interesse nesta análise, ou seja, as ocasiões em que não ocorreram partidas nem chegadas numa dada estação, o que poderá ser devido ao facto de não existirem clientes para utilizar a rede, ou pode ser devido ao facto da estação se encontrar cheia ou vazia, significando que não é possível entregar ou alugar bicicletas, respetivamente.

Desta forma, esta análise pretende responder à questão: "Quando não ocorrem viagens

na rede, tal é devido a não haverem clientes ou ao facto de as estações se encontrarem cheias/vazias?", pois é importante tentar compreender o que ocorre nestes momentos.

Foi criado um novo atributo para avaliar os *plateaus* que ocorrem nos dados. Este novo atributo é de tipo booleano, sendo que, quando não ocorreram partidas nem chegadas, é atribuído o valor de zero e, quando ocorreu pelo menos uma partida ou uma chegada, é atribuído o valor de 1. Um exemplo pode ser observado na tabela 5.4, representando, para a estação 101, as colunas de data e hora, partidas, chegadas e *plateau*, respetivamente.

Data e Hora	Partidas	Chegadas	Plateau
2018-03-16 08:00	0	0	0
2018-03-16 08:15	0	0	0
2018-03-16 08:30	0	1	1
2018-03-16 08:45	0	0	0
2018-03-16 09:00	1	0	1
2018-03-16 09:15	0	0	0
2018-03-16 09:30	2	0	1
2018-03-16 09:45	3	0	1

Tabela 5.4: Exemplo do conjunto de dados para analisar os *plateaus* para a Estação 101.

Tendo como base este novo atributo, foi criado um *dashboard* para analisar os *plateaus* numa granularidade temporal horária, significando que, uma vez que o conjunto de dados original apresenta granularidade temporal de 15 em 15 minutos, existem quatro observações por hora, significando que podem existir, no máximo, 4 ocasiões em que não existem *plateaus*, ou seja, nas quais existe, pelo menos, uma partida ou uma chegada numa dada estação. Desta forma, uma vez que a análise será feita numa granularidade temporal horária e se pretende analisar o *plateau*, utilizou-se a soma horária dos valores deste atributo para o efeito. Na tabela 5.4 encontra-se um exemplo da soma do valor do *Plateau*, sendo possível verificar que às 9 horas este é 3, significando que houveram viagens em três dos intervalos temporais de 15 minutos.

Para esta análise, apenas serão considerados valores da Soma do *Plateau* que sejam menores ou iguais a um, significando que, ao longo de uma hora, apenas houveram viagens num único bloco de 15 minutos ou não ocorreram viagens.

Data e Hora	Soma do Plateau
2018-03-16 08:00	1
2018-03-16 09:00	3
2018-03-16 10:00	1
2018-03-16 11:00	0

Tabela 5.5: Exemplo do conjunto de dados para com a Soma do *Plateau* para a Estação 101.

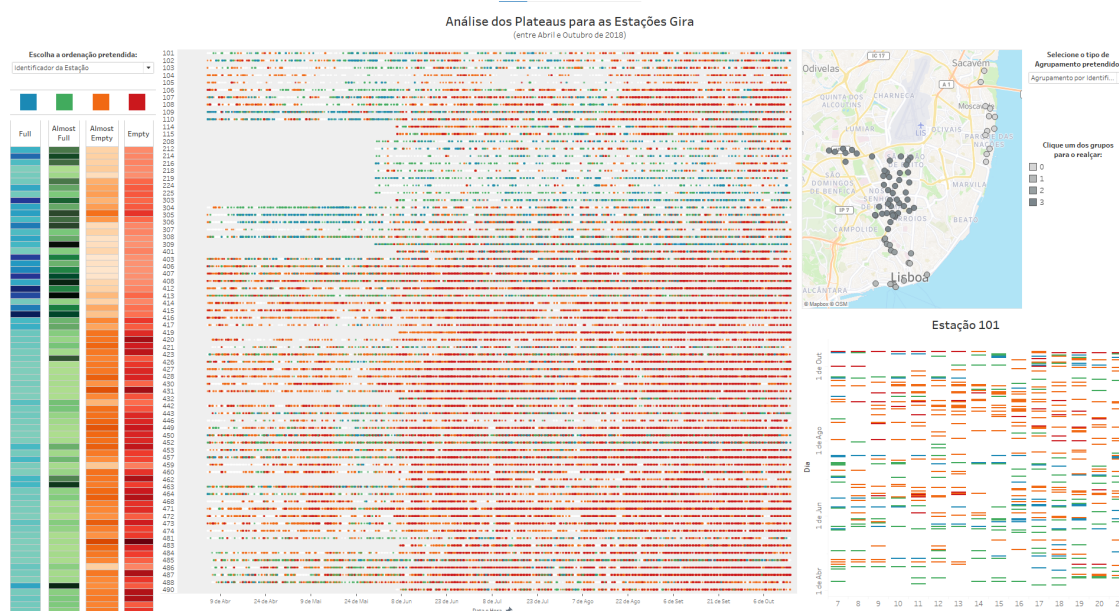
De forma a poder avaliar se a existência dos *plateaus* é devido à inexistência de clientes ou devido a uma dada estação se encontrar cheia ou vazia, foi utilizada a informação relativamente ao número de bicicletas existentes em cada estação e ao número de docas

de cada estação. Nos momentos em que o valor da soma do *plateau* é igual a zero ou a 1, a informação relativamente ao número de bicicletas foi consultada, de forma a avaliar a ocupação das estações, tendo sido criada uma escala de ocupação com quatro graus: *Full*, *Almost Full*, *Almost Empty* e *Empty*, cuja descrição se encontra na tabela 5.6. Quando uma estação se encontra cheia (*Full*), não é possível ocorrer a entrega de bicicletas ou chegadas a essa estação e, por sua vez, quando a estação está vazia (*Empty*), não é possível haver o aluguer de bicicletas ou partidas nessa estação. Os graus de *Almost Full* e *Almost Empty* representam que a estação apenas tem 1 ou 2 docas livres ou que apenas existem 1 ou 2 bicicletas, respetivamente.

Graus de Ocupação	Descrição
Full	Não existem docas livres na estação.
Almost Full	Existem 1 ou 2 docas livres na estação.
Almost Empty	Existem apenas 1 ou 2 bicicletas na estação.
Empty	Não existem bicicletas na estação.

Tabela 5.6: Descrição dos graus de ocupação criados.

Utilizando esta informação, foi criado o *dashboard* da Figura 5.21, que pretende analisar de diferentes formas a ocorrência dos momentos de *plateau* dos dados. As quatro visualizações à esquerda no *dashboard*, são quatro colunas, uma para cada grau de ocupação, que representam, para cada estação, o valor de uma métrica criada com o objetivo de classificar cada estação de acordo com cada grau de ocupação. Esta métrica é a soma de

Figura 5.21: *Dashboard* de Análise dos *plateaus* nas Estações Gira.

todas as linhas do conjunto de dados em que o grau de ocupação é classificado como pertencendo, por exemplo, ao estado *Full* a dividir pelo intervalo temporal correspondente a

cada estação. Fazendo este processo para todos os graus de ocupação obteve-se uma métrica comparável relativamente a cada grau de ocupação que permite comparar, dentro de cada grau, todas as estações relativamente à quantidade de vezes que estiveram cheias, vazias, quase cheias ou quase vazias de uma forma simples e rápida. Relativamente às cores seleccionadas, embora a escala dos graus de ocupação seja do tipo discreto, foi escolhida uma paleta de cores que representasse uma ordem entre cada grau de ocupação, desde o estado Cheio até ao estado Vazio. Para cada grau de ocupação, foi escolhida uma escala de cores contínua de acordo com a cor original seleccionada para cada grau da escala. Para além destas quatro colunas permitirem avaliar todas as estações e ter uma vista geral das mesmas de uma forma simples, é ainda possível escolher a ordenação das estações como, por exemplo, ordenar de forma crescente por identificador da estação, como se observa na Figura 5.21 ou ordenar de forma decrescente tendo em conta qualquer um dos graus de ocupação, como se pode observar na Figura 5.22, em que as estações se encontram ordenadas de forma decrescente de acordo com a métrica do grau de ocupação *Full*. Desta forma, é possível observar qual a estação cuja métrica é a maior e ainda perceber como esta se comporta tendo em conta outras métricas.

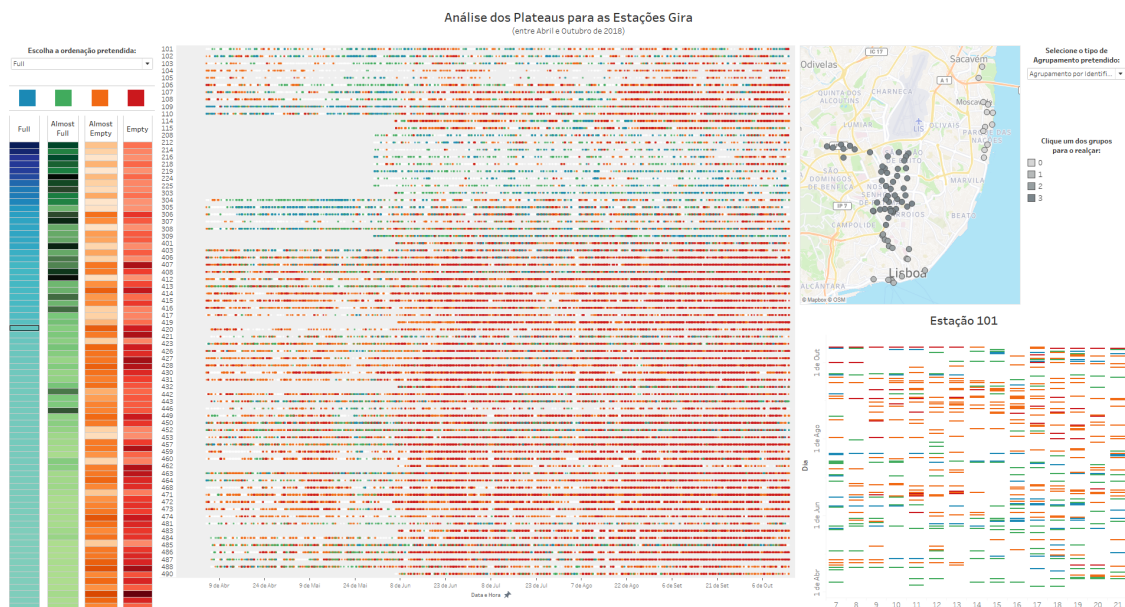
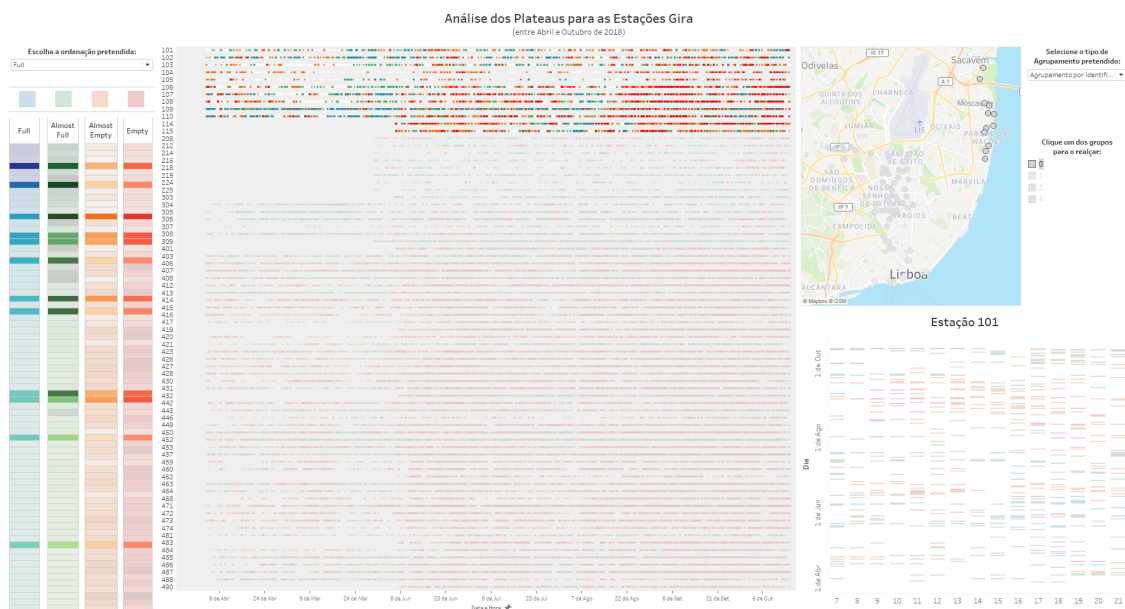


Figura 5.22: *Dashboard* de Análise dos *plateaus* nas Estações Gira com ordenação das estações pelas que estiveram mais cheias.

A visualização imediatamente à direita pretende analisar de forma mais específica relativamente à componente temporal de todas as estações em simultâneo, sendo que cada linha deste *heatmap* representa uma estação ao longo do tempo, com as cores a representarem os quatro graus de ocupação. Como referido anteriormente, o intervalo temporal das estações varia devido ao facto de algumas só terem sido adicionadas mais tarde, o que é evidente pelas zonas a cinzento. Por exemplo, os dados das estações 115, 208 e 212 têm início mais tarde do que as estações 101 e 102. Nesta visualização, é possível perceber o

padrão dos graus de ocupação ao longo do tempo, sendo ainda possível realçar um grau de ocupação à escolha clicando na legenda de cores à esquerda.

Na zona da direita do *dashboard*, encontram-se duas visualizações: um mapa e um *heatmap*. O mapa permite escolher uma estação da rede Gira, cujo perfil pode ser observado no *heatmap* imediatamente abaixo, sendo que, neste caso, este é específico para a estação selecionada, com o eixo do X a representar as horas do dia e o eixo dos Y a representar todos os dias do conjunto de dados de forma contínua, sendo possível avaliar o perfil de uma estação de forma mais detalhada. Finalmente, é possível escolher diferentes tipos de agrupamentos no mapa, nomeadamente de acordo com o identificador, de acordo com apenas a localização, localização e docas, de acordo com apenas o perfil de rácio ou nenhum tipo de agrupamento. Após a escolha de um tipo de agrupamento, é possível clicar num dos grupos de forma a realçar as estações do mesmo nas visualizações à esquerda, permitindo analisar apenas as estações desse grupo, como se pode visualizar na Figura 5.23.



ent Em adição às interações referidas aquando da descrição das visualizações presentes no *dashboard*, é possível selecionar uma ou várias estações nas colunas à esquerda, sendo que vai selecionar as estações em causa no mapa e também nas visualizações dos *heatmaps*. Selecionando a legenda de cores acima de cada grau de ocupação, é possível realçar em ambos os *heatmaps* os momentos em que esse grau de ocupação ocorreu. Para obter mais informação acerca de uma estação relativamente aos seus graus de ocupação, é possível colocar o rato por cima da estação pretendida numa das quatro colunas, de forma a ter acesso a uma *tooltip* que apresenta o *ranking* dessa mesma estação para cada grau de ocupação, sendo possível verificar, por exemplo, se uma estação é a que está mais vezes cheia, mas a que está menos vezes vazia. Também é possível selecionar uma ou mais estações à escolha no mapa, de forma a realçá-la nas colunas relativas aos graus de ocupação e no *heatmap* com todas as estações e para filtrar a ou as estações escolhidas no *heatmap* da direita.

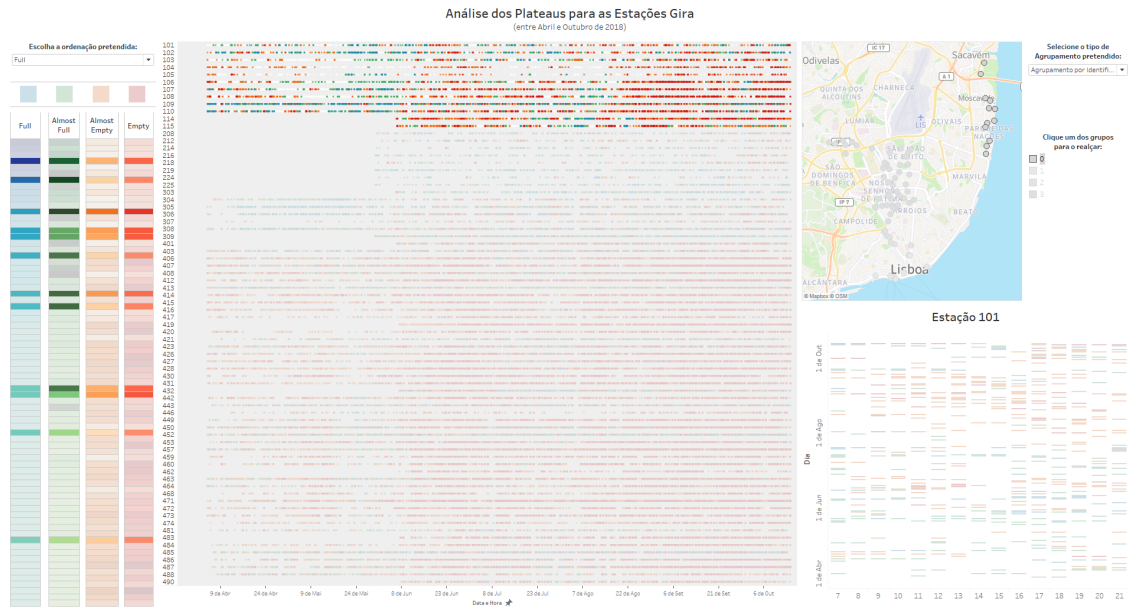


Figura 5.23: *Dashboard* de Análise dos *plateaus* nas Estações Gira com realce das estações com identificador 1XX.

Ordenando as estações de acordo com a métrica do grau de ocupação *Full*, a estação 309 é aquela cujo valor da métrica é o mais elevado, como é possível observar na Figura 5.24. Colocando o rato sobre a mesma, é possível averiguar que, para além de ser a esta-

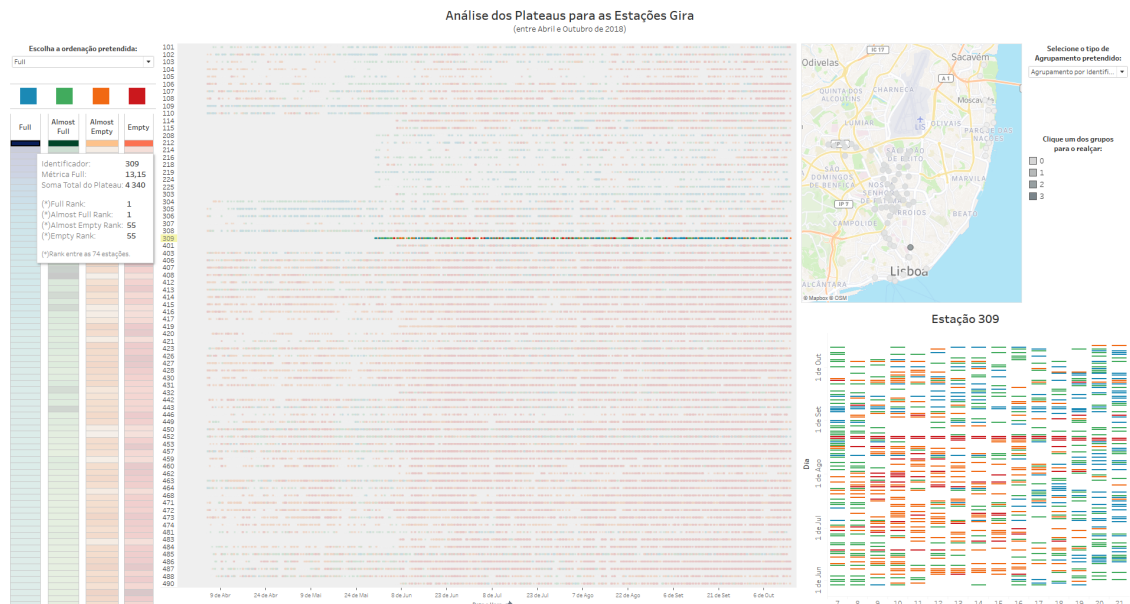


Figura 5.24: *Dashboard* de Análise dos *plateaus* nas Estações Gira com ênfase na estação 309 e respetiva *tooltip*.

ção que se encontra mais vezes cheia, também é aquela que se encontra mais vezes quase cheia, ou seja, a primeira do grau de ocupação *Almost Full*. Nos dois graus de ocupação

que restam, esta ocupa o lugar nº55 em 74 estações, significando que esta estação tende a receber mais bicicletas. Para além disso, é evidente que os *plateaus* ocorridos nos dados podem ser devidos, na maioria das vezes, ao facto da estação estar cheia, o que significa que, mesmo que hajam utilizadores na rede que se desloquem à estação 309, não conseguem colocar uma bicicleta nesta estação, sendo necessário deslocarem-se a uma estação mais próxima para o fazer. Analisando o *heatmap* relativo à estação 309 verifica-se que, de facto, a maioria dos *plateaus* ocorridos apresentam cor verde ou azul, que indicam que a estação permanece num estado cheio ou quase cheio. Uma vez que a estação se encontra cheia com muita frequência, o comportamento dos clientes aquando da intenção de devolver uma bicicleta na estação 309 é deslocarem-se a uma estação próxima a esta e, analisando o mapa, existem três estações próximas da mesma: as estações 224 e 225 localizadas no Martim Moniz e a estação 303 localizada perto do metro da Avenida. Realçando estas estações no mapa, é possível realçá-las também nas restantes visualizações e, observando a Figura 5.25, é possível perceber que também estas estações se encontram mais vezes cheias e quase cheias e poucas vezes vazias ou quase vazias. Dado o que foi analisado anteriormente, este acontecimento pode ser uma consequência do perfil da estação 309, sendo que há mais bicicletas a ser entregues nestas três estações como consequência da estação 309 se encontrar cheia. O *heatmap* geral com todas as estações também permite perceber que a estação 224, por exemplo, apresenta muitos espaços a branco, significando que não são *plateaus*, o que indica que esta estação teve poucos momentos de *plateau* relativamente às outras, que apresentam menos pontos a branco.

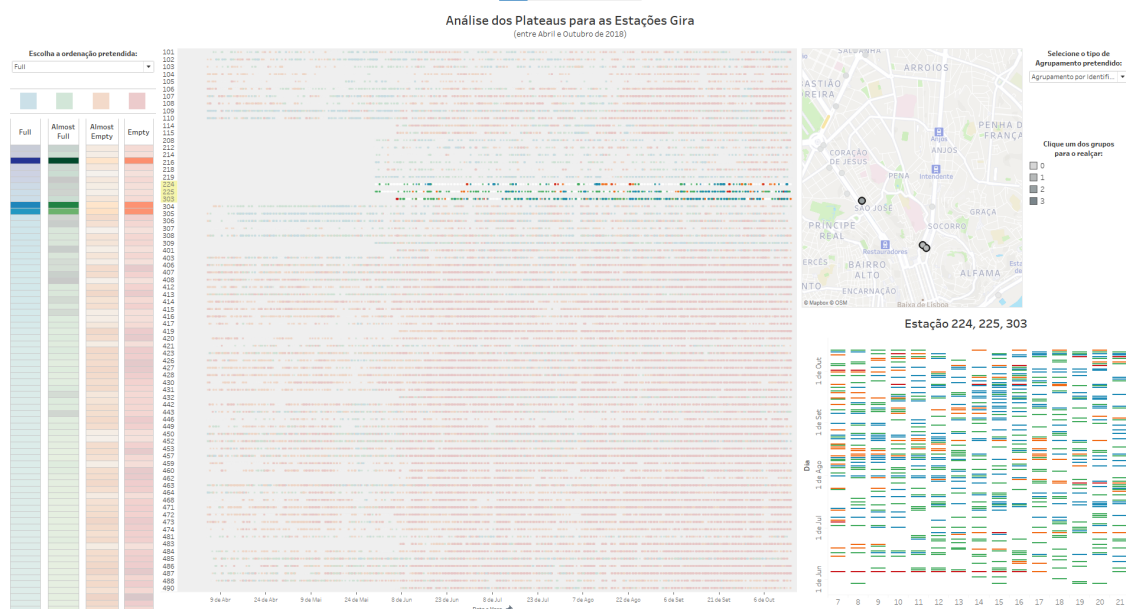


Figura 5.25: *Dashboard* de Análise dos *plateaus* nas Estações Gira com ênfase nas estações próximas da estação 309.

Ordenando as estações pelo grau de ocupação Empty, observa-se na Figura 5.26 que a estação com o maior valor para esta métrica é a 471, localizada perto da Escola Básica

de São João de Brito. Tendo em conta a informação presente na *tooltip*, percebe-se que, embora esta estação se encontre muitas vezes cheia e quase cheia, esta raramente está vazia ou quase vazia, ocupando lugares próximos do 74. Tal indica que existem mais partidas do que chegadas nesta estação, e que os *plateaus* ocorridos nesta estação podem ser devidos ao facto de não existirem bicicletas para alugar. O *heatmap* relativo à estação 471 encontra-se quase completamente preenchido com *plateaus* de cor laranja e vermelha, indicando que, quando estes ocorrem, na maioria das vezes tal é devido a falta de bicicletas na estação.

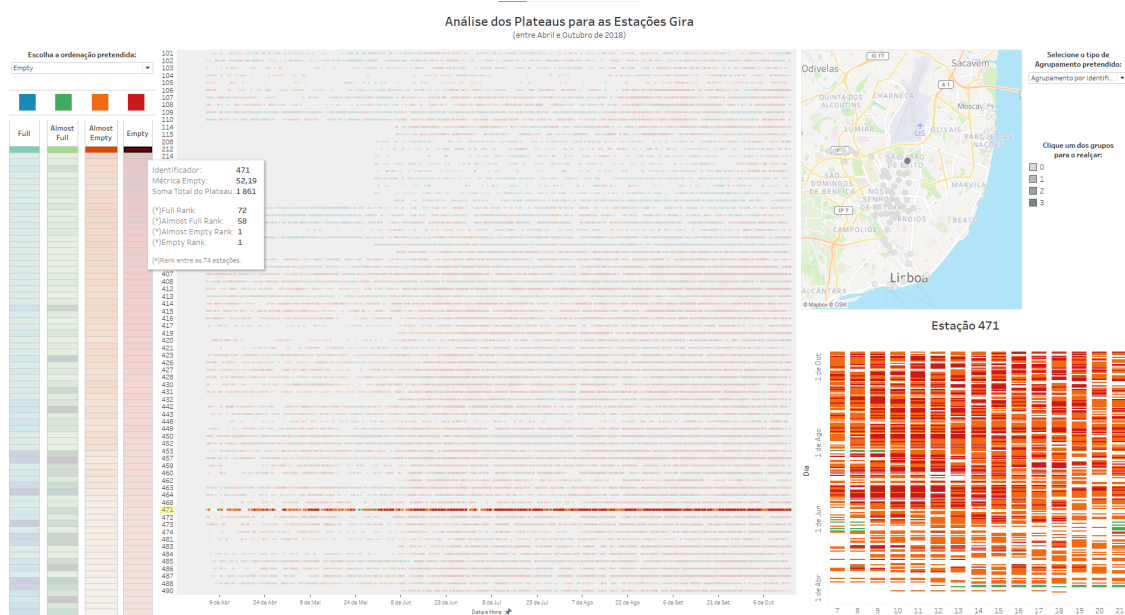


Figura 5.26: *Dashboard* de Análise dos *plateaus* nas Estações Gira com ênfase na estação 471 e respetiva *tooltip*.

Esta análise permitiu concluir que, de facto, existem muitos momentos nos dados em que a inexistência de viagens se deveu ao facto das estações se encontrarem cheias ou vazias. Observando o *heatmap* geral das estações, é perceptível que foram mais os momentos de falta de bicicletas nas estações do que aqueles em que uma estação se encontrava cheia, indicando que existe, no geral, falta de bicicletas nas estações da rede da Gira.

CONCLUSÕES E TRABALHO FUTURO

Foram cumpridos os objetivos principais desta dissertação propostos no capítulo 1, tendo ainda sido desenvolvidas algumas análises que não se encontravam planeadas no início da mesma.

Houve a necessidade de criar um novo conjunto de dados para efetuar a previsão, uma vez que os dados originais tinham influência do rebalanceamento. Não foi possível remover a mesma e não foi nem corrigida essa influência nos dados, nem foram facultados dados adicionais ou corrigidos de forma a ultrapassar este obstáculo. Como consequência deste facto, foi ainda necessário efetuar uma análise de plateaus, uma vez que, mesmo com o novo conjunto de dados criado, houve a necessidade de perceber a que se deve a ausência de viagens.

Foram desenvolvidas ferramentas para análise dos dados facultados, bem como agrupamentos que permitem descrever os mesmos e detetar padrões de enchimento e esvaziamento das estações ao longo do dia, permitindo caracterizar as estações da rede. Foram ainda desenvolvidos modelos preditivos que permitem prever, com uma baixa margem de erro, a quantidade aproximada (através do saldo) de bicicletas que estarão em cada estação.

Desta dissertação resultaram ainda relatórios de análise de qualidade dos dados facultados, nomeadamente das estações, das viagens e dos dados meteorológicos e novos conjuntos de dados, nomeadamente os que foram utilizados para as análises espaciais e temporais, bem como o conjunto de dados utilizado para os modelos preditivos.

No que diz respeito ao trabalho futuro, seria interessante utilizar mais dados, de vários anos e sem influência do rebalanceamento, para elaborar novos modelos preditivos, pois tal permitiria remover sazonalidade dos dados, bem como uma melhor previsão. Poderia ainda ser incluída informação relativamente ao tipo de bicicleta, ou seja, se esta é elétrica ou clássica na previsão, incluir o tempo que as bicicletas demoram a ser carregadas, pois

isso também poderá influenciar o número de bicicletas que se encontram na estação.

BIBLIOGRAFIA

- [1] L. Breiman. “Random Forests”. English. Em: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324). URL: <http://dx.doi.org/10.1023/A%3A1010933404324>.
- [2] C. M. de Chardon, G. Caruso e I. Thomas. “Bike-share rebalancing strategies, patterns, and purpose”. Em: *Journal of Transport Geography* 55 (2016), pp. 22–39. ISSN: 0966-6923. DOI: <https://doi.org/10.1016/j.jtrangeo.2016.07.003>. URL: <http://www.sciencedirect.com/science/article/pii/S0966692316300114>.
- [3] M. Chavent, V. Kuentz-Simonet, A. Labenne e J. Saracco. “ClustGeo: an R package for hierarchical clustering with spatial constraints”. Em: *Computational Statistics* 33.4 (2018), 1799–1822. ISSN: 1613-9658. DOI: [10.1007/s00180-018-0791-1](https://doi.org/10.1007/s00180-018-0791-1). URL: <http://dx.doi.org/10.1007/s00180-018-0791-1>.
- [4] P. Chen, H. Hsieh, X. K. Sigalingging, Y. Chen e J. Leu. “Prediction of Station Level Demand in a Bike Sharing System Using Recurrent Neural Networks”. Em: *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. 2017, pp. 1–5. DOI: [10.1109/VTCSpring.2017.8108575](https://doi.org/10.1109/VTCSpring.2017.8108575).
- [5] F. Chiariotti, C. Pielli, A. Zanella e M. Zorzi. “A Dynamic Approach to Rebalancing Bike-Sharing Systems”. Em: *Sensors* 18.2 (2018). ISSN: 1424-8220. DOI: [10.3390/s18020512](https://doi.org/10.3390/s18020512). URL: <https://www.mdpi.com/1424-8220/18/2/512>.
- [6] *Citi Bike - Planos de assinatura*. <https://account.citibikenyc.com/access-plans>. Acedido em 2021-01-26.
- [7] M. Dell’Amico, E. Hadjicostantinou, M. Iori e S. Novellani. “The bike sharing rebalancing problem: Mathematical formulations and benchmark instances”. Em: *Omega* 45 (2014), pp. 7–19. ISSN: 0305-0483. DOI: <https://doi.org/10.1016/j.omega.2013.12.001>. URL: <http://www.sciencedirect.com/science/article/pii/S0305048313001187>.
- [8] P. DeMaio. “Bike-sharing: History, Impacts, Models of Provision, and Future”. Em: *Journal of Public Transportation* 12 (dez. de 2009). DOI: [10.5038/2375-0901.12.4.3](https://doi.org/10.5038/2375-0901.12.4.3).
- [9] *Diagrams.net - Security-first diagramming for teams*. <https://www.diagrams.net/>. Acedido em 2021-09-18.

- [10] EMEL lança Bike Sharing. <https://www.gira-bicicletasdelisboa.pt/noticias/emel-lanca-bike-sharing/>. Acedido em 2020-01-26.
- [11] Gira - Bicicletas de Lisboa. Passes e Tarifários. <https://www.gira-bicicletasdelisboa.pt/passes-e-tarifarios/>. Acedido em 2020-01-25.
- [12] O. Guyandi, G. Wang, T. Sriwidadi, H. Prabowo e R. K. Sari. “Assessing bike sharing business model”. Em: *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*. Nov. de 2017, pp. 89–94. DOI: [10.1109/ICICoS.2017.8276343](https://doi.org/10.1109/ICICoS.2017.8276343).
- [13] J. Han, M. Kamber e J. Pei. “10 - Cluster Analysis: Basic Concepts and Methods”. Em: *Data Mining (Third Edition)*. Ed. por J. Han, M. Kamber e J. Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 443–495. ISBN: 978-0-12-381479-1. DOI: <https://doi.org/10.1016/B978-0-12-381479-1.00010-1>. URL: <https://www.sciencedirect.com/science/article/pii/B9780123814791000101>.
- [14] S. Hochreiter e J. Schmidhuber. “Long Short-term Memory”. Em: *Neural computation* 9 (dez. de 1997), pp. 1735–80. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [15] IBM - CRISP-DM Business Understanding. <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-business-understanding>. Acedido em 2021-09-07.
- [16] IBM - CRISP-DM Data Understanding. <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-data-understanding>. Acedido em 2021-09-07.
- [17] IBM - CRISP-DM Deployment. <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-deployment>. Acedido em 2021-09-07.
- [18] IBM - CRISP-DM Help Overview. <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview>. Acedido em 2021-09-07.
- [19] IBM - CRISP-DM Modeling. <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-modeling>. Acedido em 2021-09-07.
- [20] IBM - CRISP-DM Modeling. <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-evaluation>. Acedido em 2021-09-07.
- [21] A. K. Jain. “Data clustering: 50 years beyond K-means”. Em: *Pattern Recognition Letters* 31.8 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), pp. 651–666. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2009.09.011>. URL: <https://www.sciencedirect.com/science/article/pii/S0167865509002323>.
- [22] V. Kotu e B. Deshpande. “Chapter 12 - Time Series Forecasting”. Em: *Data Science (Second Edition)*. Ed. por V. Kotu e B. Deshpande. Second Edition. Morgan Kaufmann, 2019, pp. 395–445. ISBN: 978-0-12-814761-0. DOI: <https://doi.org/10.1016/B978-0-12-814761-0.00012-5>. URL: <http://www.sciencedirect.com/science/article/pii/B9780128147610000125>.

- [23] D. Li, Y. Zhao e Y. Li. “Time-Series Representation and Clustering Approaches for Sharing Bike Usage Mining”. Em: *IEEE Access* 7 (2019), pp. 177856–177863. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2958378](https://doi.org/10.1109/ACCESS.2019.2958378).
- [24] Y. Li, Y. Zheng, H. Zhang e L. Chen. “Traffic Prediction in a Bike-Sharing System”. Em: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL '15. New York, NY, USA: Association for Computing Machinery, 2015. ISBN: 9781450339674. DOI: [10.1145/2820783.2820837](https://doi.org/10.1145/2820783.2820837). URL: <https://doi.org/10.1145/2820783.2820837>.
- [25] LxDataLab - Desafios Propostos. <https://lisboainteligente.cm-lisboa.pt/lxi-iniciativas/lx-data-lab/>. Acedido em 2020-01-18.
- [26] LxDataLab - O conhecimento ao serviço do cidadão. Apresentação. <https://lisboainteligente.cm-lisboa.pt/lxdatalab/apresentacao/>. Acedido em 2020-01-18.
- [27] T. Mátrai e J. Tóth. “Comparative Assessment of Public Bike Sharing Systems”. Em: *Transportation Research Procedia* 14 (2016). Transport Research Arena TRA2016, pp. 2344 –2351. ISSN: 2352-1465. DOI: <https://doi.org/10.1016/j.trpro.2016.05.261>. URL: <http://www.sciencedirect.com/science/article/pii/S2352146516302678>.
- [28] PyCharm. <https://www.jetbrains.com/pycharm/>. Acedido em 2021-09-03.
- [29] Python. <https://docs.python.org/3/tutorial/>. Acedido em 2021-09-03.
- [30] Público - Bicicletas Gira de Lisboa ultrapassam um milhão de viagens. <https://www.publico.pt/2019/01/03/local/noticia/bicicletas-gira-lisboa-ultrapassam-milhao-viagens-1856519>. Acedido em 2020-01-26.
- [31] S. Ruffieux, E. Mugellini e O. Abou Khaled. “Bike Usage Forecasting for Optimal Rebalancing Operations in Bike-Sharing Systems”. Em: *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. 2018, pp. 854–858. DOI: [10.1109/ICTAI.2018.00133](https://doi.org/10.1109/ICTAI.2018.00133).
- [32] S. Ruffieux, N. Spycher, E. Mugellini e O. A. Khaled. “Real-time usage forecasting for bike-sharing systems: A study on random forest and convolutional neural network applicability”. Em: *2017 Intelligent Systems Conference (IntelliSys)*. 2017, pp. 622–631. DOI: [10.1109/IntelliSys.2017.8324359](https://doi.org/10.1109/IntelliSys.2017.8324359).
- [33] K. Ryden. “OpenGIS® Implementation Specification for Geographic information - Simple feature access - Part 1:Common architecture”. Em: 2005.
- [34] Tableau Desktop Answer questions at the speed of thought with Tableau Desktop. <https://www.tableau.com/products/desktop>. Acedido em 2021-09-03.
- [35] Tableau Desktop Welcome to Tableau Desktop. <https://www.tableau.com/learn/get-started/creator>. Acedido em 2021-09-03.
- [36] Tableau Prep - Overview - Scaling your prepped data just got easier. Consultado em <https://www.tableau.com/products/prep>. Acedido em 2021-09-03.

- [37] *Tableau Prep - What is Tableau Prep?* Consultado em <https://www.tableau.com/learn/get-started/prep>. Acedido em 2021-09-03.
- [38] N. Thi Hoai Thu, L. Trung Thanh, C. Thi Phuong Dung, N. Linh-Trung e H. Vu Le. “Multi-source data analysis for bike sharing systems”. Em: *2017 International Conference on Advanced Technologies for Communications (ATC)*. Out. de 2017, pp. 235–240. DOI: [10.1109/ATC.2017.8167624](https://doi.org/10.1109/ATC.2017.8167624).
- [39] X. xu, Y. Zhirui, J. Li e M. Xu. “Understanding the Usage Patterns of Bicycle-Sharing Systems to Predict Users’ Demand: A Case Study in Wenzhou, China”. Em: *Computational Intelligence and Neuroscience 2018* (set. de 2018), pp. 1–21. DOI: [10.1155/2018/9892134](https://doi.org/10.1155/2018/9892134).
- [40] H. Yang, X. Zhang, L. Zhong, S. Li, X. Zhang e J. Hu. “Short-term demand forecasting for bike sharing system based on machine learning”. Em: *2019 5th International Conference on Transportation Information and Safety (ICTIS)*. 2019, pp. 1295–1300. DOI: [10.1109/ICTIS.2019.8883732](https://doi.org/10.1109/ICTIS.2019.8883732).
- [41] X. Yao, X. Shen, T. He e S. H. Son. “Demand Estimation of Public Bike-Sharing System Based on Temporal and Spatial Correlation”. Em: *2018 4th International Conference on Big Data Computing and Communications (BIGCOM)*. Ago. de 2018, pp. 60–65. DOI: [10.1109/BIGCOM.2018.00016](https://doi.org/10.1109/BIGCOM.2018.00016).
- [42] P. Yi, F. Huang e J. Peng. “A Rebalancing Strategy for the Imbalance Problem in Bike-Sharing Systems”. Em: *Energies* 12 (jul. de 2019), p. 2578. DOI: [10.3390/en12132578](https://doi.org/10.3390/en12132578).
- [43] C. Zhang, L. Zhang, Y. Liu e X. Yang. “Short-term Prediction of Bike-sharing Usage Considering Public Transport: A LSTM Approach”. Em: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. Nov. de 2018, pp. 1564–1571. DOI: [10.1109/ITSC.2018.8569726](https://doi.org/10.1109/ITSC.2018.8569726).
- [44] Y. Zhou e Y. Huang. “Context Aware Flow Prediction of Bike Sharing Systems”. Em: *2018 IEEE International Conference on Big Data (Big Data)*. 2018, pp. 2393–2402. DOI: [10.1109/BigData.2018.8621918](https://doi.org/10.1109/BigData.2018.8621918).



RELATÓRIO DE QUALIDADE DE DADOS

A.1 Introdução

No âmbito desta dissertação foram facultados dois conjuntos de dados pelo *LxDataLab* da Câmara Municipal de Lisboa, que deverão ser utilizados no decorrer da dissertação.

Os dados facultados dizem respeito ao [SBP](#) de Lisboa, denominado Gira, e são relativos às Estações da rede de bicicletas, bem como às Viagens efetuadas na mesma durante o ano de 2018. De forma a perceber os dados, é necessário levar a cabo a sua análise exploratória, para que seja possível analisar a sua qualidade.

Para tal, utilizou-se a metodologia [CRISP-DM](#), que permite uma abordagem estruturada para projetos que envolvem conjuntos de dados, sendo considerada robusta. Neste caso, a fase levada a cabo foi a de Compreensão dos Dados, que se decompõe em três sub-tópicos: Descrição dos Dados, Exploração dos Dados e Verificação da sua qualidade.

Este relatório encontra-se dividido em duas secções distintas, sendo que cada uma diz respeito a um dos conjuntos de dados. Em cada secção, é elaborada a descrição dos dados, indicando os atributos e os seus nomes, bem como o seu significado semântico e tipologia. De seguida, é elaborada a exploração dos atributos apresentados na fase anterior. Finalmente, é analisada a qualidade dos dados, identificando problemas nos mesmos e possíveis soluções a tomar na próxima fase, que será de pré-processamento dos mesmos.

Os ficheiros facultados são Folhas de Cálculo do Excel (formato .xlsx) e foram analisados recorrendo ao *Tableau Prep Builder*, de forma a poder ter uma ideia geral dos dados e ao *Tableau*, para poder construir visualizações que pudessem ajudar a compreender melhor os mesmos.

A.2 Dados das Estações Gira 2018

A.2.1 Descrição dos Dados

Os dados contêm 11 atributos e 6,237,806 linhas, sendo relativos aos dados das estações da Gira localizadas em Lisboa e foram retirados ao longo do ano de 2018.

De forma a proceder à descrição dos mesmos, organizaram-se os atributos numa tabela [A.1](#), que explica qual o seu significado semântico. Para além disso, especificaram-se ainda as tipologias e foi dado um exemplo de valores que constam dos dados para cada atributo, o que se pode observar na tabela [A.2](#).

Observando as tabelas, constata-se que seis atributos são numéricos e cinco são não numéricos. Destes últimos, dois deles apresentam escalas simbólicas que descrevem as estações. O tipo de serviço apresenta como valores possíveis A e B e o estado apresenta como valores possíveis active, repair e stock, que descrevem o estado da estação num dado momento.

Os restantes atributos não numéricos descrevem a designação comercial das estações, cujos valores são dados pelo identificador da estação e pela sua localização (rua); a localização da estação ([WKB Geometry](#)) e a hora a que o registo foi efetuado.

Os numéricos dão informação acerca de atributos das estações à hora em que foi efetuada a leitura como, por exemplo, o número de bicicletas, o número de docas daquela estação, o rácio entre as bicicletas e as docas e o identificador da estação. O identificador identifica cada registo, sendo único.

Atributo	Descrição
Identificador	Identificador referente ao número da leitura
Id Expl	Identificador da Estação Gira
Id planeamento	Identificador da Estação Gira
Designação comercial	Morada da Estação e Identificador
WKB Geometry	Representação geométrica em texto
Tipo de Serviço	Indica o tipo de serviço
Número de Bicicletas	Número de bicicletas na estação
Número de Docas	Número de docas existente numa estação
Rácio	Rácio entre número de bicicletas e número de docas
Estado	Estado da Estação
Data de Update	Data e hora do registo

Tabela A.1: Atributos e respetiva descrição.

Atributo	Tipologia	Exemplo
Identificador	Numérico	1000, 15000
Id Expl	Numérico	211, 500
Id planeamento	Numérico	211, 500
Designação comercial	Simbólico	105-CC Vasco da Gama
WKB Geometry	Simbólico	0104000020E61000..
Tipo de Serviço	Simbólico	A,B
Número de Bicicletas	Numérico	0, 2, 15
Número de Docas	Numérico	12, 40
Rácio	Numérico	0.2, 0.4, 1
Estado	Simbólico	active, repair, stock
Data de Update	Simbólico	23/02/2018, 22:09:16

Tabela A.2: Atributos, tipologia e exemplos de valores para os mesmos.

A.2.2 Exploração dos Dados

Identificador

Este atributo é um identificador para cada leitura efetuada, tendo-se verificado que não existiam valores duplicados. No total existem 6 237 806 linhas, que correspondem ao número de registos efetuados, estando compreendidos entre 157 e 6 236 098.

Id expl

Este atributo corresponde ao identificador das Estações Gira, havendo 99 valores únicos compreendidos entre 101 e 999.

Observando a figura [A.1](#), é possível constatar que os números relativos às Estações não apresentam valores seguidos, existindo 14 valores distintos compreendidos entre 100 e 200, mas 60 entre 400 e 500. É de notar a existência de apenas um identificador 1 e 78 449 leituras do identificador 999.

O projeto Gira arrancou no parque das nações com algumas estações cujo número dos identificadores estão entre 100 e 199. Após este primeiro conjunto de estações, têm sido adicionadas mais estações ao longo da cidade. Analisando a localização das estações e o seu identificador, é visível uma relação entre ambos, como se observa na figura [A.2](#).

Constata-se que existem 4 grupos de estações que se encontram relacionados com o identificador, observando-se o grupo das estações 20X na zona da Baixa, o grupo das estações cujo identificador é 30X mais perto do Príncipe Real e o conjunto das estações 40X, que é o maior, na zona mais central da cidade, que se estende desde São Sebastião da Pedreira até ao Campo Grande.

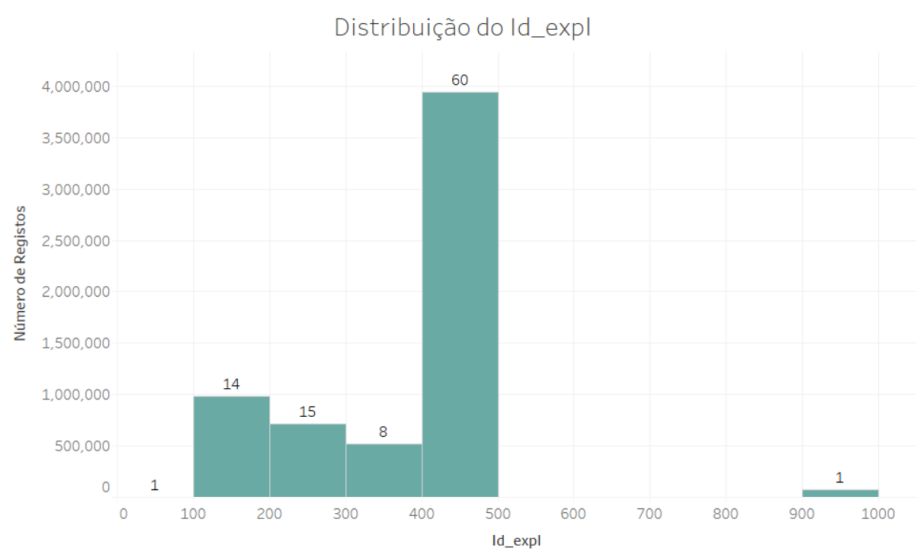


Figura A.1: Distribuição do atributo Id planeamento. As etiquetas representam a contagem de valores distintos em cada um dos intervalos.

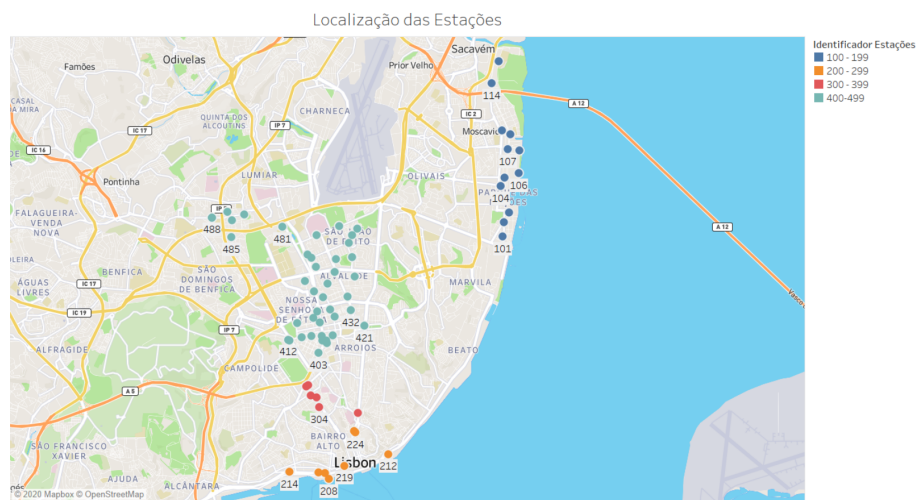


Figura A.2: Localização das Estações e respetivos identificadores.

Id planeamento

Este atributo corresponde ao identificador das Estações Gira, havendo 99 valores únicos compreendidos entre 101 e 999.

Observando a figura A.3, constata-se que os valores para o Id planeamento apresentam a mesma distribuição que os valores do Id expl, analisado anteriormente.

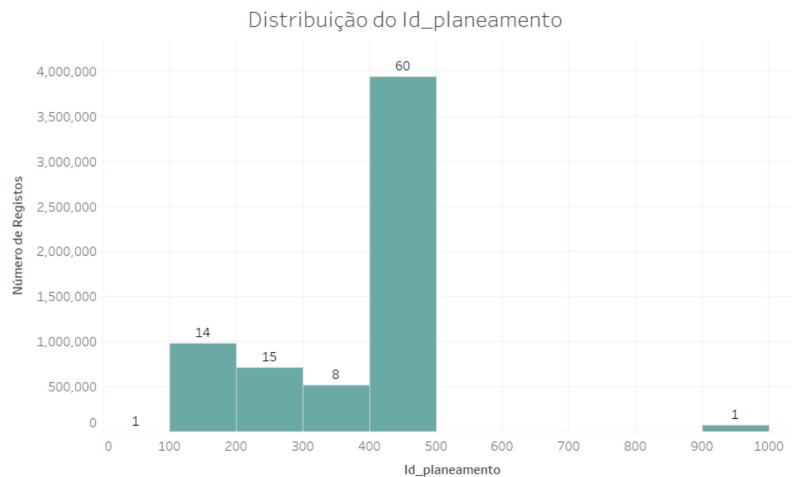


Figura A.3: Distribuição do atributo Id planeamento. As etiquetas representam a contagem de valores distintos em cada um dos intervalos.

Designação Comercial

A designação comercial diz respeito a uma combinação entre o identificador da estação e a sua localização.

Existem 105 valores distintos para este atributo, o que não se encontra de acordo com os 99 valores de identificador das estações. Tal pode indicar que existem valores de morada incorretos.

Na figura A.4 encontram-se as estações para as quais a morada apresenta mais de um identificador ou cujo identificador apresenta mais de uma morada. É possível observar que ambas as situações ocorrem, nomeadamente no caso da estação cujo identificador é 488, existem, três designações comerciais diferentes.

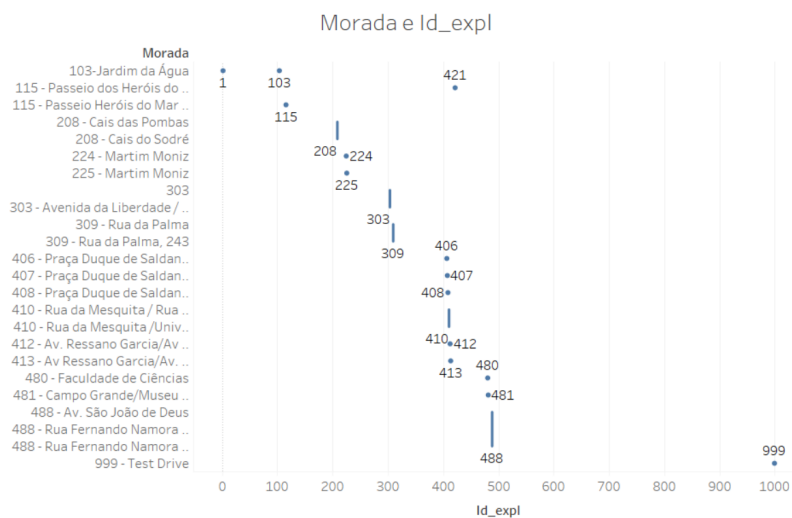


Figura A.4: Relação entre a morada e o identificador das estações.

WKB Geometry

O formato **WKB** é uma representação de um objeto geométrico utilizando *bytes*, o que permite guardar o mesmo em *SQL*. [33]

Este conjunto de dados apresenta um atributo com este formato, que fornece informação relativa à localização de cada estação, existindo 119 valores únicos para o mesmo. Uma vez que este formato é binário, de forma a perceber melhor os dados, será necessário efetuar a conversão destes para **WKT**.

Este formato também é uma representação de um objeto geométrico, que pode ter vários formatos como pontos, linhas, multipontos, entre outros. Neste caso, a conversão dos valores permitiu perceber que se trata de um multiponto composto por apenas um ponto com duas coordenadas representativas da latitude e longitude de cada estação Gira, tendo o seguinte formato: `MULTIPOINT(longitude latitude)`, `MULTIPOINT(-9.096439 38.759156)`.

Verificou-se que 10.27% (640.355 linhas) dos valores dizem respeito a um multiponto com as coordenadas a zero, o que deverá ser corrigido. Uma vez que é possível ter acesso à localização das estações através do *website* da Gira¹, é possível proceder à correção destes valores através da correspondência entre a localização das estações no *website* e os valores presentes nos dados.

Observando a figura A.5, é possível verificar que existem 48 estações com dois ou mais valores de localização, indicando que existem erros neste atributo, sendo que, tal como referido anteriormente, existe também um valor que representa um multiponto cujas coordenadas são zero.

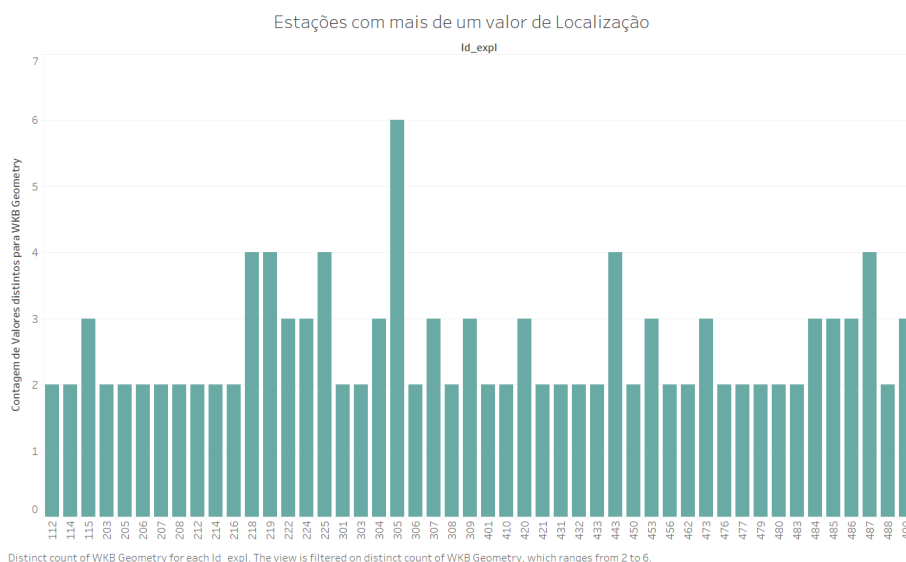


Figura A.5: Estações que apresentam mais de um valor distinto de **WKB**.

¹<https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/>

Tipo de Serviço

Este atributo é do tipo categórico que apresenta três valores: null, A e B, que representam o tipo de serviço para cada estação (uma estação corresponde a apenas um destes valores) e apenas era utilizado para calcular o [Service-level Agreement \(SLA\)](#), algo fundamental quando existe um contrato de prestação de serviços e que permite especificar os serviços que o contratante pode esperar do fornecedor. De momento, este valor já não é utilizado, não apresentando relevância para os dados.

Na figura A.6 é possível observar o gráfico que mostra, para cada valor do tipo de serviço, a sua frequência e percentagem relativamente ao total de valores.

Com base na figura em questão, é possível constatar que este atributo apresenta 0.55% de valores nulos, o que corresponde a 34 494 linhas.

Após uma observação mais aprofundada destas linhas, constata-se que existem muita linhas com valores duplicados. Estes correspondem a leituras que foram registadas à mesma hora para várias estações.

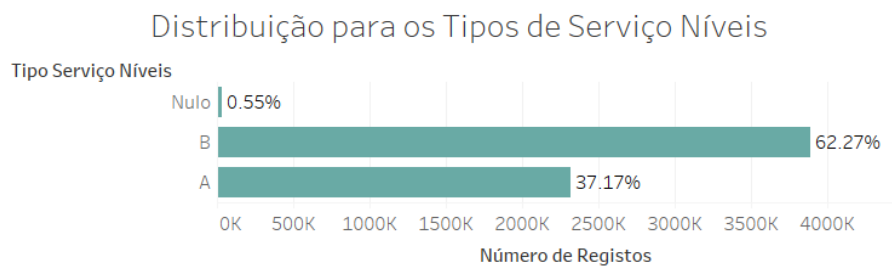


Figura A.6: Gráfico de barras para o tipo de serviço níveis.

Número de Bicicletas

Este atributo diz respeito ao número de bicicletas existentes numa estação da Gira a uma dada altura do dia.

A partir do gráfico de barras na figura A.7, observa-se que o número de bicicletas está compreendido entre 0 e 41 e a sua distribuição é enviesada para a esquerda, havendo uma quantidade muito superior de registos cujo número de bicicletas é zero.

O número mínimo de bicicletas numa estação é zero e o número máximo encontra-se limitado pela capacidade de cada estação, que é variável, sendo visível um decréscimo do número de leituras respetivo à quantidade de bicicletas conforme o número destas aumenta.

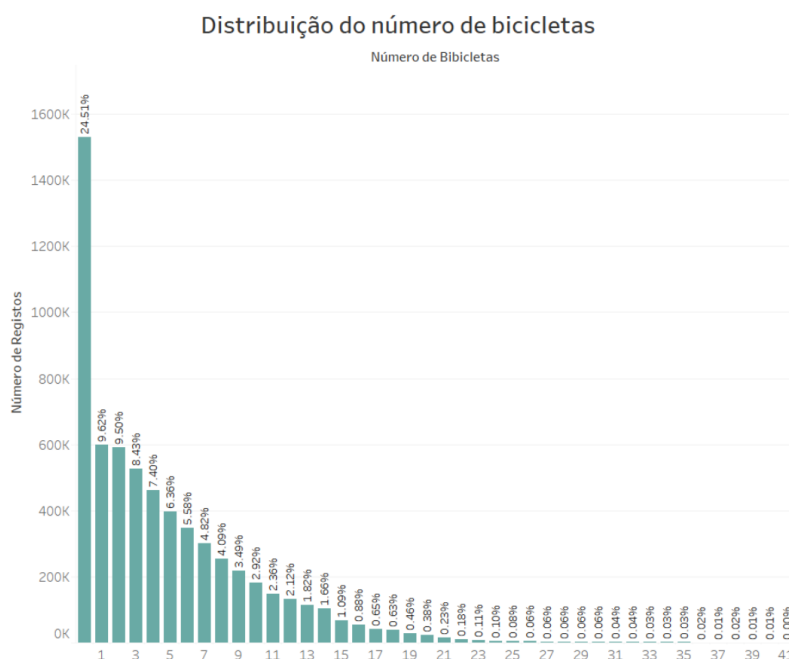


Figura A.7: Gráfico de barras para a distribuição do número de bicicletas.

Número de Docas

O número de docas representa as docas existentes numa estação Gira e, neste caso, os valores estão compreendidos entre 0 e 66.

Observando o gráfico A.8, verifica-se a existência de registos cujo número de docas é zero, o que está incorreto, visto que as estações têm sempre um número de docas superior a zero. Isto pode revelar que existem problemas nos dados. Para além disso, também foi possível averiguar que apenas existe uma leitura cujo número de docas é 66, o que também poderá estar incorreto.

Seria interessante determinar se existem estações para as quais existe mais de um valor distinto para o número de docas. Desta forma, a partir da figura A.9, foi possível determinar que há 51 estações com mais de um valor distinto para o número de docas, quais estas estações e o seu número. Observa-se que as estações 224 e 225 apresentam o maior número distinto de valores para o número de docas (11) e que a maioria das estações apresenta dois valores distintos.

De forma a perceber melhor a variação do número de docas, observou-se esta evolução ao longo do tempo para as estações que apresentavam mais do que dois valores distintos para este atributo, cuja figura A.10 demonstra. Observando o mesmo, é possível perceber que o valor do número de docas para todas as estações começa a zero, dando-se um aumento deste número no mesmo dia, sendo que algumas apresentam aumentos também ao longo do tempo, como é o exemplo da estação 206. A estação 307 apresenta um pico para o valor do número de docas, que depois vai descendo, até estabilizar em 33 docas.

É necessário decidir como efetuar a limpeza destes dados, uma vez que estes podem

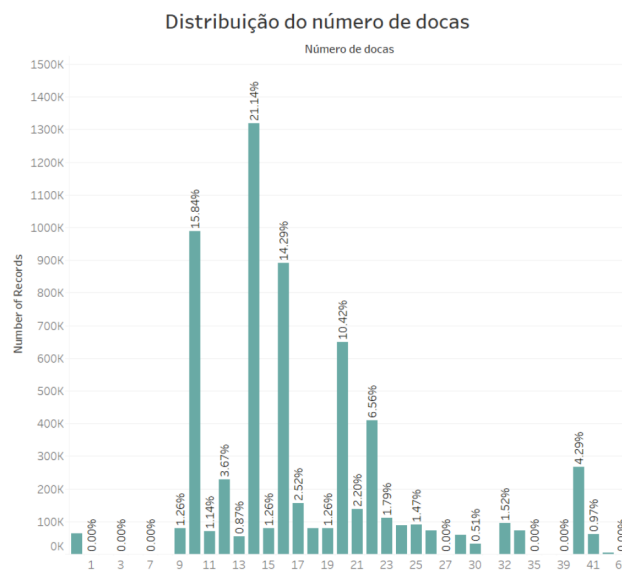


Figura A.8: Gráfico de barras para a distribuição do número de docas.

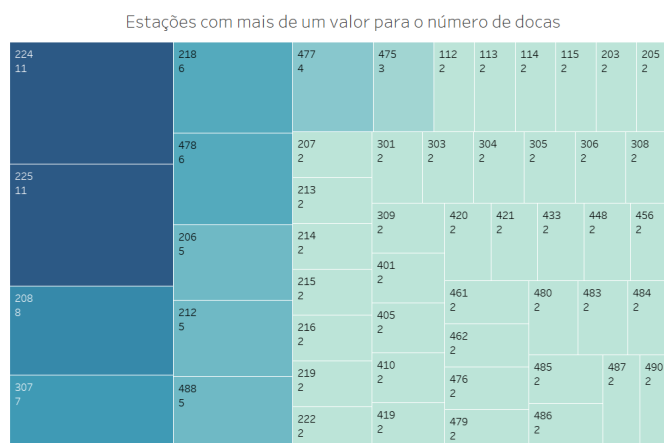


Figura A.9: Estações com mais de um valor para o número de docas e respetivo número.

modificar as conclusões retiradas destes, visto que o valor do número de docas terá influência no valor do rácio das estações. Este gráfico também permite observar que as leituras referentes às estações não começaram todas na mesma altura, havendo estações como, por exemplo, 475, 477 e 478, cujos registos começam apenas em finais de agosto. As estações com dois valores distintos para o número de docas apresentavam todas o zero como um dos valores, à exceção da estação 419.

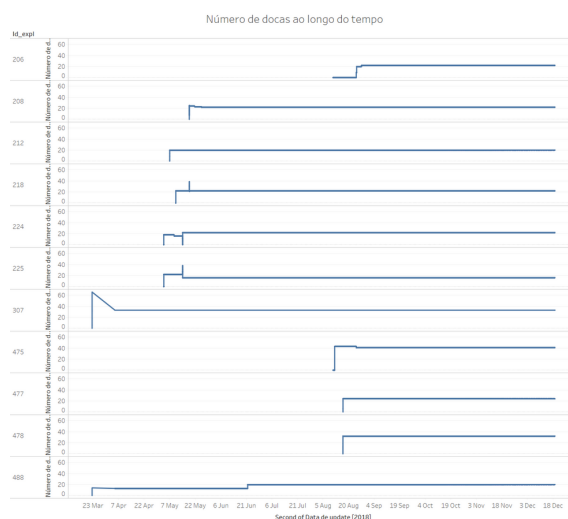


Figura A.10: Número de docas em função do tempo.

Rácio

Este atributo diz respeito à divisão do número de bicicletas pelo número de docas de uma estação, sendo uma métrica que permite saber a ocupação da estação na altura do registo. Os valores referentes a este atributo estão compreendidos entre 0 e 1, representando que uma estação está vazia ou cheia, respetivamente. A sua distribuição, que pode ser observada na figura A.11, demonstra que cerca de 33% dos registos apresenta um rácio de zero, sendo que mais de metade dos valores apresentam um rácio entre 0 e 0,3.

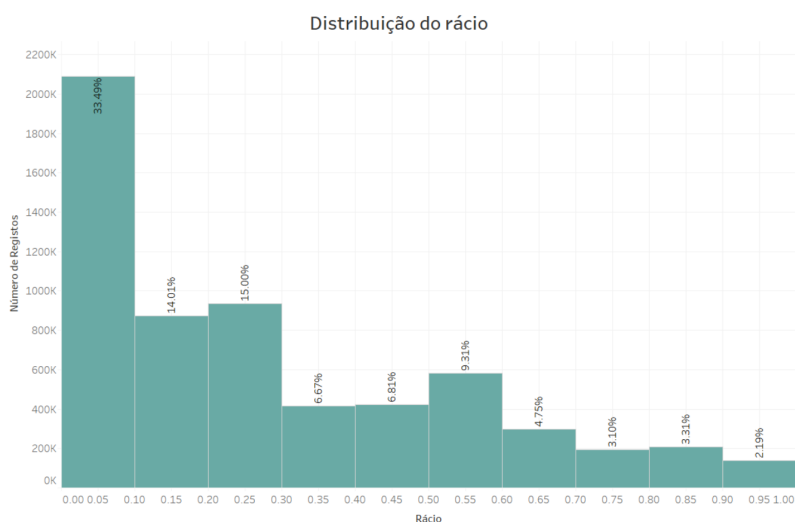


Figura A.11: Distribuição para o rácio do número de bicicletas e número de docas.

Estado

O atributo relativo ao Estado é categorial, sendo constituído por uma lista de três valores: "Active", "Stock" ou "Repair". Estes compõem uma escala nominal, na qual cada um apresenta um significado distinto e, para além disso, os valores não apresentam qualquer tipo ordem associada.

O estado representa o estado da estação aquando da data da leitura, podendo estar activa ("Active"), significando que se encontra em funcionamento, em reparação ("Repair"), não sendo possível a sua utilização e em "Stock", não se encontrando disponível na aplicação da Gira para utilização.

A distribuição do Estado encontra-se na figura A.12, sendo visível que em 83,48% o estado era "Active", existindo 15% destas com Estado a "stock", significando que não se encontravam visíveis na aplicação e apenas 1,39% destas se encontravam em reparação.



Figura A.12: Distribuição do Estado.

Para além disso, observou-se a relação entre o número de bicicletas e o estado da Estação (Figura A.13), tendo sido considerados apenas os estados "stock" e "repair". Verifica-se que quase a totalidade dos registos cujo estado é "stock" e "repair" apresentam um número de bicicletas igual a zero.

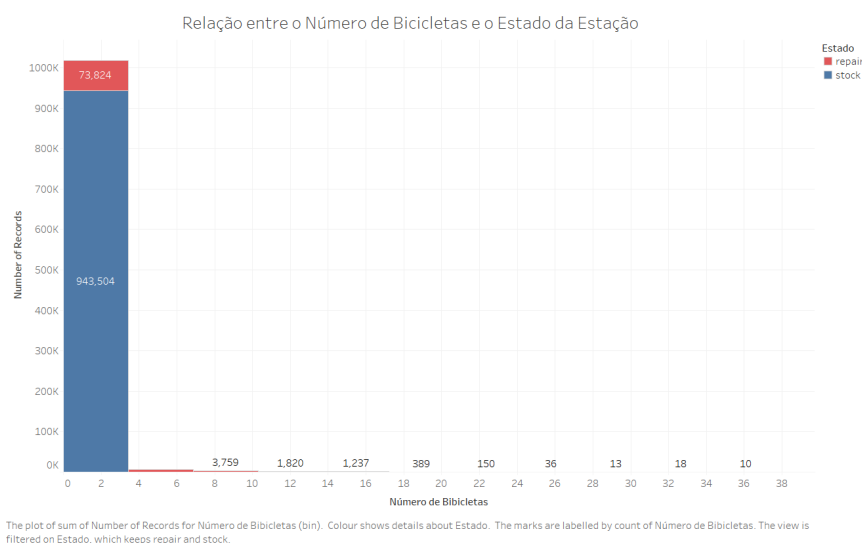


Figura A.13: Distribuição do Estado.

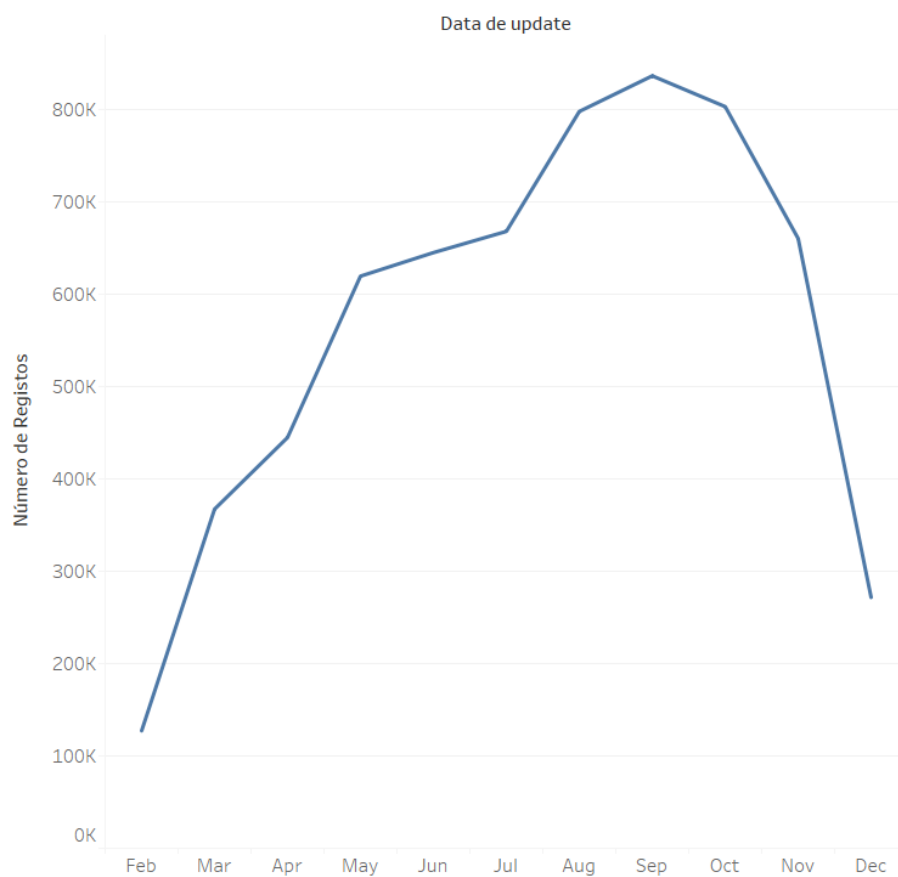
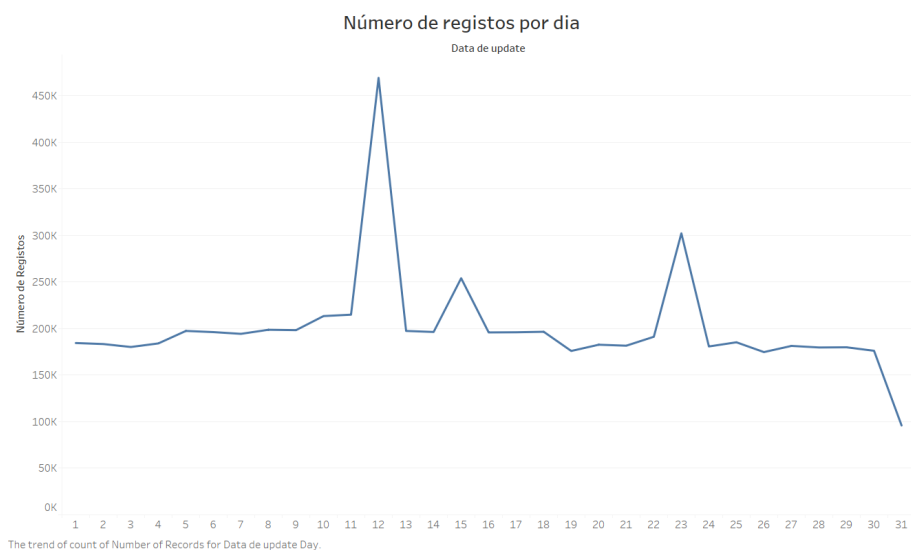
Data de Update

A data de *update* corresponde à data e hora em que cada registo foi efetuado, sendo o seu formato do tipo DATE TIME (YYYY-MM-DD HH:MM:ss).

Foram feitas leituras entre 23 de fevereiro e 18 de dezembro, sendo que a quantidade efetuada não foi constante de mês para mês, como se observa na figura [A.14](#). É evidente que no mês de fevereiro a quantidade de registos é a menor, o que também é devido ao facto de as leituras apenas terem tido início a dia 23. Após este mês, é evidente o aumento do número de registos até ao pico de leituras em setembro. Como se constatou anteriormente, há estações cujos registos apenas tiveram início em finais de agosto, o que também poderá influenciar a quantidade de leituras. Observando a figura [A.15](#), constata-se que as leituras efetuadas por dia se mantêm relativamente semelhantes, com exceção de dois picos observados nos dias 12, 15 e 23. O dia 31 apresenta menos leituras devido ao facto de alguns meses não terem 30 dias.

A observação dos picos de registos na imagem [A.14](#) levou a uma análise mais aprofundada destes valores, tendo-se verificado a existência de duplicados nos registos referentes à data de *update*. É de salientar que estes estão de acordo com os picos observados na imagem referida, existindo mais de 100 000 duplicados no dia 23 de fevereiro, vários duplicados no dia 12 de março a diferentes horas e também para o dia 15 de março. Estas datas e contagem dos valores pode ser observada na Tabela [A.3](#), sendo possível verificar que são todos nos meses de fevereiro e março, embora o dia 23 de fevereiro seja aquele em que se verifica o maior valor. Estes registos duplicados dizem respeito a várias estações e o número de bicicletas também varia, embora os registos tenham sido efetuados à mesma hora.

Evolução temporal do número de registos para as estações

Figura A.14: Número de Registos da Data e Hora de *Update* por mês.Figura A.15: Número de Registos da Data e Hora de *Update* por dia.

Data e Hora	Número de Registos
23/02/2018 22:09:16	127,193
12/03/2018 10:55:03	18,100
12/03/2018 10:55:04	42,233
12/03/2018 10:55:05	36,203
12/03/2018 10:55:06	36,204
12/03/2018 10:55:07	36,204
12/03/2018 10:55:08	42,238
12/03/2018 10:55:09	36,198
12/03/2018 10:55:10	18,097
15/03/2018 11:55:12	5,157
15/03/2018 19:05:14	6,927
15/03/2018 21:45:12	10,078
15/03/2018 21:45:13	5,039
15/03/2018 22:00:13	10,072
15/03/2018 22:05:13	5,035
15/03/2018 22:15:13	10,066
22/03/2018 08:15:12	6,364
22/03/2018 08:15:13	6,364
26/03/2018 11:00:13	2,002
27/03/2018 13:10:13	1,688

Tabela A.3: Registos duplicados nos dados.

A.2.3 Qualidade dos Dados

Verificou-se a existência de alguns problemas de qualidade nos dados relativos às Estações, sendo que alguns dizem respeito a valores nulos, valores a zero, valores incorretos e ainda duplicados.

No que diz respeito a valores nulos, apenas se verificou a sua presença no atributo relativo ao Tipo de Serviço, embora seja em apenas 0,55% do totalidade dos dados. Devido ao facto deste atributo não ser importante para a análise que se pretende levar a cabo, o mesmo poderá ser removido do conjunto de dados.

Os atributos `id_expl` e `id_planeamento` apresentam uma leitura cujo valor é 1, que se encontra incorreto, devido a não existir uma estação com esse valor. Para além disso, existem algumas leituras cujo identificador é 999, que deverão ser removidas de forma a aumentar a qualidade dos dados.

No que diz respeito à designação comercial, verificou-se a existência de inconsistências entre este atributo e o identificador da estação, existindo Moradas com mais de um identificador associado e identificadores com mais de uma Morada associada. Poderá ser possível efetuar a correção destes dados através da averiguação da localização da estação

(atributo *WKB Geometry*), de forma a aumentar a qualidade dos dados.

O atributo *WKB Geometry* apresenta não só um valor que corresponde a coordenadas a zero, mas também vários valores de coordenadas para a mesma estação, o que indica a existência problemas de inconsistência e de validação nos dados, visto que existem coordenadas erradas. No entanto, estas incongruências poderão ser corrigidas através da realização de uma comparação dos valores deste atributo com a informação referente às Estações e a sua localização que se encontra no *website* da Gira. Desta forma, será elaborado um ficheiro do qual irão constar os identificadores das Estações, o seu valor para o *WKB* e, para além disso, o seu valor de *WKT*, latitude e longitude respetivas. Desta forma, será mais fácil fazer operações sobre os dados utilizando a latitude e longitude.

Foram detetados dados incorretos no atributo correspondente ao número de docas, sendo que existem não só estações cujo número de docas é zero, mas também estações que apresentam mais do que um número distinto de docas ao longo do tempo. Após uma análise mais detalhada constatou-se que essas variações acontecem, muitas vezes, no mesmo dia, evidenciando um erro no registo dos dados, que poderá ser removido de forma a aumentar a sua qualidade. Para além disso, existem estações no qual o número de docas aumenta ao longo do tempo, o que poderá ser devido a terem sido inseridas mais docas na estação e, como tal, esses dados não serão alterados.

Foram encontrados duplicados nos dados, nomeadamente, registos para uma mesma estação à mesma hora, sendo que todos ocorreram nos meses de fevereiro e março. Fazendo uma análise mais aprofundada, constata-se que, por exemplo, para uma estação, à mesma hora, existem registos cujo número de bicicletas é diferente, não sendo possível determinar qual das leituras foi a correta. Como tal, de forma a aumentar a qualidade dos dados, estes duplicados terão de ser removidos. Para além disso, constatou-se que existem muito menos leituras para o mês de fevereiro, sendo aconselhável não ter em conta as leituras do mesmo, removendo-as dos dados.

A.3 Dados das Viagens Gira 2018

A.3.1 Descrição dos Dados

Este conjunto de dados diz respeito às viagens efetuadas no [SBP](#) da Gira no ano de 2018, sendo composto por 10 atributos e 694 559 linhas.

Na tabela [A.4](#) observa-se o nome dos atributos e a sua descrição. Os detalhes relativos à tipologia dos atributos e exemplos de valores possíveis encontram-se na tabela [A.5](#).

Observando as tabelas, constata-se que cinco dos atributos são numéricos e os restantes são não numéricos. O tipo de bicicleta é um atributo simbólico cujos valores podem ser E ou C, dependendo da bicicleta ser elétrica ou clássica. O [RFID](#) da bicicleta é um código identificador de cada bicicleta e o *WKB Geometry* diz respeito ao percurso efetuado em cada viagem, cujos valores de exemplo podem ser observados na tabela [A.5](#). Os restantes atributos não numéricos dizem respeito à data de início e término das viagens, tendo o formato de data e hora.

Os atributos numéricos descrevem as viagens, nomeadamente a sua distância percorrida durante as mesmas, o número de vértices do percurso, que está relacionado com o atributo geométrico referido anteriormente e as estações de início e término da viagem, que permitem saber o local onde a viagem começou e terminou, através do identificador único da estação em causa. O identificador é único e permite identificar cada viagem efetuada, ou seja, cada registo do conjunto de dados.

Atributo	Descrição
Identificador	Identificador único do registo
Data de início	Data de início da viagem
Data de fim	Data de fim da viagem
Distância	Distância percorrida aquando da viagem
Estação de início	Estação de início da viagem
Estação de fim	Estação de fim da viagem
RFID da bicicleta	Identificador da bicicleta
WKB Geometry	Pontos com coordenadas representativos da viagem
Número de Vértices	Número de vértices do percurso
Tipo de Bicicleta	Tipo de bicicleta

Tabela A.4: Nome dos atributos e respetiva descrição.

Atributo	Tipologia	Exemplo
Identificador	Numérico	425 533, 423 939
Data de início	Não numérico	25/01/2018, 15:38:26
Data de fim	Não numérico	25/01/2018, 16:01:51
Distância	Numérico	6,173.85669, 1,418.8180066
Estação de início	Numérico	101, 446, 105
Estação de fim	Numérico	453, 105, 481
RFID da bicicleta	Não Numérico	e2db9e6dca, 62db9c0f2a
WKB Geometry	Não numérico	0105000020E610000001000..
Número de Vértices	Numérico	64, 18, 20
Tipo de Bicicleta	Não Numérico	E, C

Tabela A.5: Nome dos atributos, tipologia e exemplos de valores possíveis.

A.3.2 Exploração dos dados

Identificador

Este atributo é o identificador de cada registo, apresentando 694 559 valores únicos, ou seja, o mesmo número que as linhas do presente conjunto de dados. Os seus valores estão compreendidos entre 16 e 701 887.

Data de início

Este atributo diz respeito à data de início da viagem, ou seja, a data e hora a que uma bicicleta foi alugada numa das estações da rede. Os valores deste atributo encontram-se entre 25 de janeiro e 16 de outubro de 2018.

Observando a evolução temporal ao longo dos meses de 2018 na figura A.19, é possível averiguar que houveram poucas viagens no mês de janeiro, uma vez que os registos das viagens só começam a dia 25. O número de registos começa a aumentar até ao mês de julho, que é o segundo mês com maior quantidade de viagens. Observa-se um decréscimo do número de viagens em agosto, devido, provavelmente, ao facto de ser um mês no qual as pessoas costumam estar de férias, resultando numa menor utilização do SBP. Observa-se um pico no mês de setembro, sendo o mês com a maior quantidade de viagens e, novamente, um decréscimo em outubro, devido ao facto das leituras pararem no dia 16.

Analisando as viagens nos meses de julho e setembro separadamente (Figuras A.17 e A.18), verifica-se um padrão em ambas as figuras que, após análise, se deve aos dias de semana e fins-de-semana. O número de viagens efetuadas apresenta sempre um decréscimo nos fins-de-semana e um aumento durante a semana, sendo que o número de viagens ao domingo é sempre menor do que ao sábado. Enquanto que, durante a semana, o número de viagens ronda as 5000 no mês de julho e os 6000 no mês de setembro, durante os fins-de-semana, este valor decresce para valores entre os 2000 e os 3000.

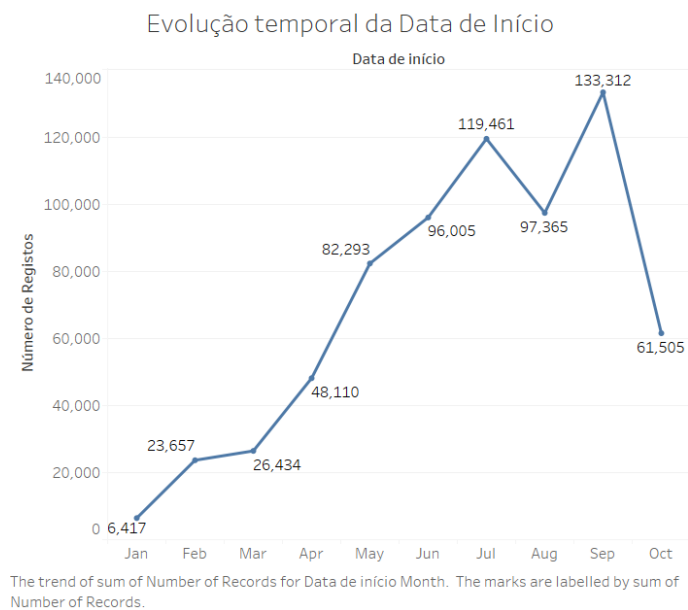


Figura A.16: Evolução temporal da Data de início ao longo dos meses de 2018.

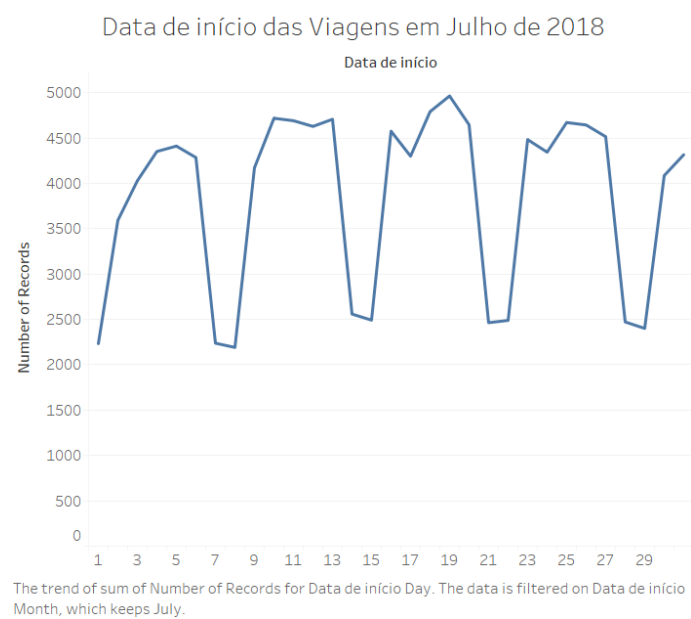


Figura A.17: Viagens no mês de julho.

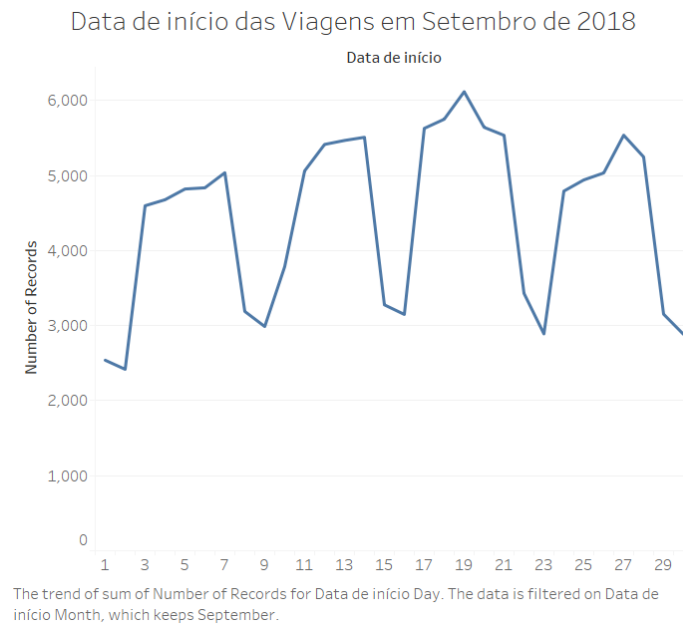


Figura A.18: Viagens no mês de setembro.

Data de fim

A data de fim diz respeito à data e hora do final de uma viagem, ou seja, a altura em que a bicicleta foi colocada numa das estações da rede após a viagem ter sido realizada.

Uma vez que quando se faz um aluguer de uma bicicleta, também é necessário efetuar a sua devolução, a distribuição deste atributo e da data de início deverão ser semelhantes em termos da evolução temporal mensal.

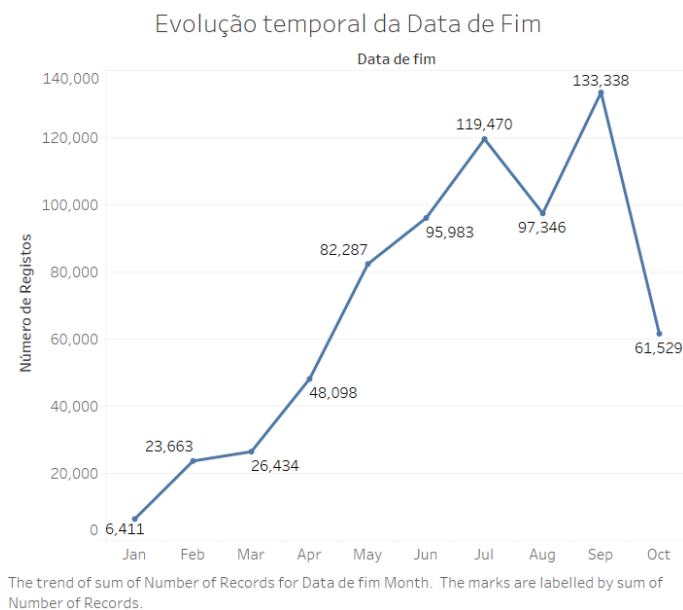


Figura A.19: Evolução temporal da Data de fim ao longo dos meses de 2018.

Distância

O atributo da distância diz respeito à distância das viagens efetuadas em metros. Existem 340 445 valores únicos para a distância, que estão compreendidos entre 0 e 28 quilómetros, não tendo em conta os valores nulos.

A figura A.20 o número de registos cujos valores são nulos e não nulos, sendo possível verificar que quase metade dos valores são nulos, significando que não há informação relativamente à distância percorrida aquando das viagens. Após observar as linhas de forma mais aprofundada, é possível averiguar que quando este valor é nulo, o número de vértices também é nulo, quase na totalidade, assim como os valores de *WKB Geometry*. No entanto, os valores relativos à estação de início, estação de fim e duração da viagem não são nulos. Estes valores nulos podem significar que ocorreu um erro na leitura dos dados (possivelmente devido a algum problema no GPS) ou o utilizador pode ter colocado a bicicleta no carro e ter-se deslocado para outro local dessa forma.

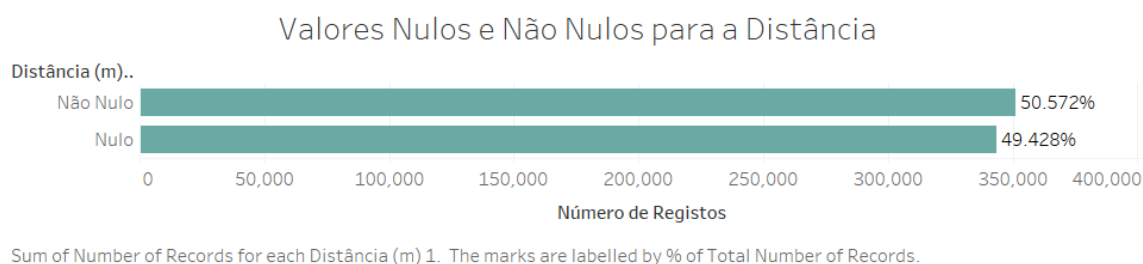


Figura A.20: Valores nulos e não nulos para a distância.

Dado que o objetivo de análise destes dados não se relaciona com a análise das viagens em si, o facto deste atributo conter muitas linhas nulas não é algo que prejudique a análise em si, no entanto, tal facto revela problemas de qualidade nos dados.

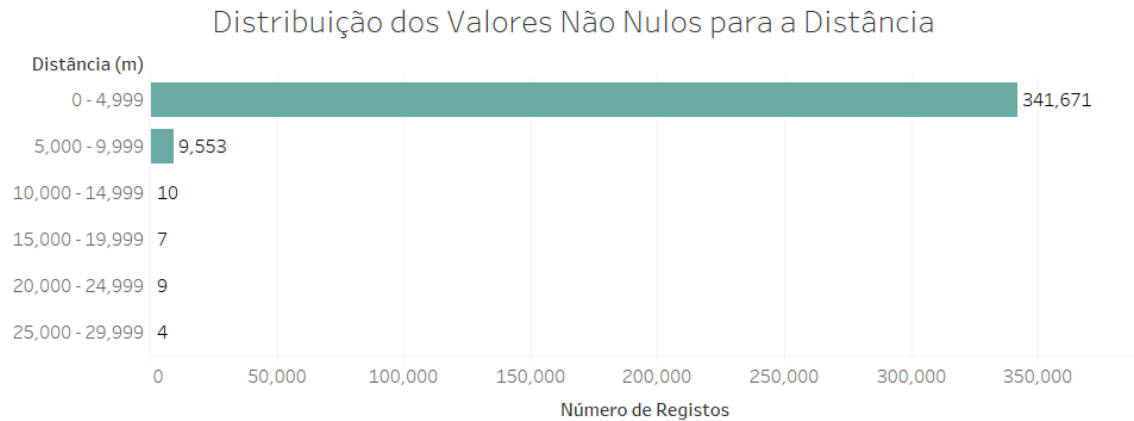
Para além dos valores nulos, também se verificou que existiam 5 811 linhas com distância a zero, o que também não está correto.

Analisando os valores não nulos (Figura A.21), constata-se que a maioria das viagens se encontra entre 0 e 4,99 quilómetros, sendo viagens de curta distância, pouco mais de 9000 viagens entre 5 e 9,99 quilómetros e apenas 30 viagens cuja distância foi superior a 10 quilómetros, sendo que a viagem cuja distância é maior de 28 quilómetros.

Estação de início

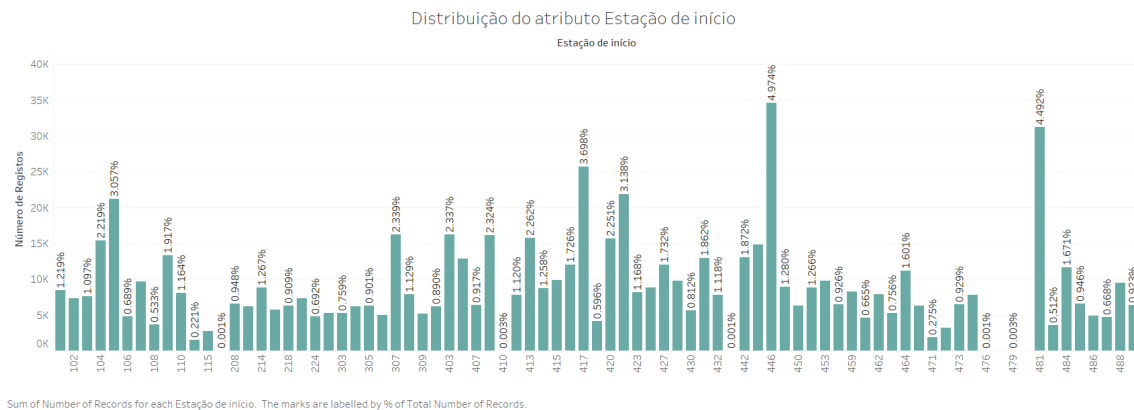
Este atributo diz respeito à Estação Gira da qual a bicicleta foi alugada, sendo correspondente ao Identificador da Estação que se encontra no website da Gira. Existem 81 valores únicos para este atributo, significando que foram alugadas bicicletas de 81 estações.

Observando a figura A.22, existem algumas estações cuja quantidade de viagens é quase inexistente e outras que apresentam uma grande quantidade de viagens, indicando que são estações com elevada utilização, tais como a 481 e 446.



Sum of Number of Records for each Distância (m). The marks are labelled by sum of Number of Records. The view is filtered on Distância (m), which excludes Nulo.

Figura A.21: Distribuição dos valores não nulos da distância.



Sum of Number of Records for each Estação de início. The marks are labelled by % of Total Number of Records.

Figura A.22: Distribuição dos valores para a estação de início.

Estação de fim

A estação de fim diz respeito à Estação do Sistema de Bicicletas da Gira na qual a bicicleta foi entregue, aquando do término de uma viagem. À semelhança do atributo "Estação de início", este também representa o Identificador da Estação. Este atributo contém 82 valores únicos, significando que se deu a entrega de bicicletas em 82 estações.

A distribuição destes valores mostra também que há estações que quase não apresentam viagens (433, 477, entre outras) e estações que se destacam por apresentarem muitas viagens relativamente às restantes estações. Neste último caso, constata-se que as estações 481 e 446 apresentam, novamente, um número mais elevado de viagens.

Comparando ambas as figuras (A.22 e A.23), é de notar que o mesmo grupo de estações apresenta uma maior percentagem de viagens, significando que essas estações são muito utilizadas tanto para aluguer como para entrega de bicicletas. Finalmente, existem duas linhas cuja estação de fim é nula, no entanto, após observação destas linhas, o seu valor de **WKB Geometry** não é nulo e contém um multilinestring do caminho efetuado, sendo possível efetuar a correção destas linhas.

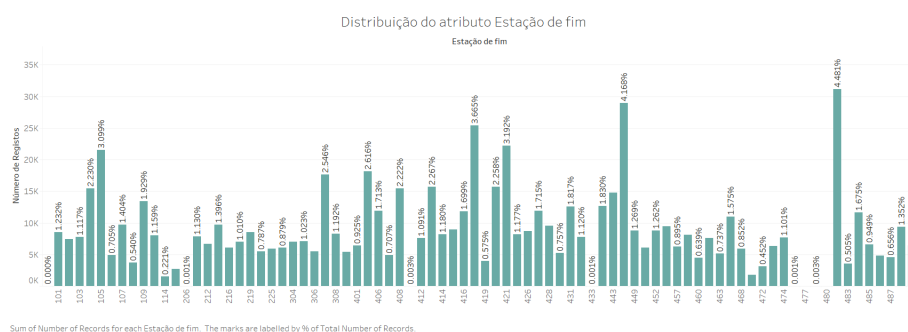


Figura A.23: Distribuição dos valores para a estação de fim.

RFID da Bicicleta

O **RFID**, denominado identificador de rádio frequência, é um identificador que utiliza campos magnéticos para localizar etiquetas ligadas a objetos de forma automática, sendo utilizada para identificar objetos e podem conter informação que é guardada de forma eletrônica.

Neste caso, cada **RFID** é único de uma bicicleta, significando que o número de valores únicos deste atributo dá informação acerca do número de bicicletas existentes que, neste caso, é 763.

De forma a descobrir a quantidade de bicicletas elétricas e clássicas, contaram-se os valores distintos do **RFID** tento em contra o tipo de bicicleta, cujo gráfico se pode observar na figura A.24. No entanto, constatou-se que haviam 763 valores únicos para o tipo E, o mesmo valor para o tipo C e 89 valores únicos a nulo, significando que todas as bicicletas são classificadas nos registos como clássicas e elétricas e 89 delas tiveram esta classificação a nulo.

Tal apresenta um problema, pois deixa de ser possível retirar conclusões acerca de quais as bicicletas mais utilizadas, visto que não há forma de determinar quais as bicicletas são elétricas e quais são clássicas através do seu identificador. A causa para o erro nestes dados terá ainda de ser apurada.

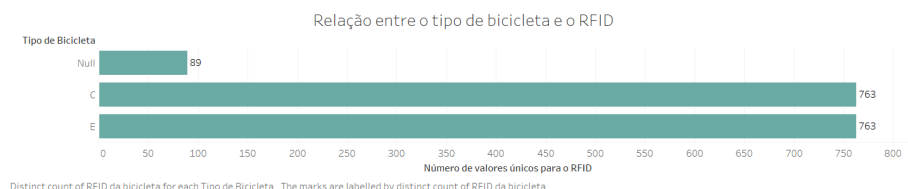


Figura A.24: Relação entre o **RFID** e o tipo de bicicleta.

WKB Geometry

Estes valores dizem respeito a um formato binário que, após ter sido convertido, se constatou representar um tipo MULTILINESTRING, que representa uma sequência de pontos que, neste caso, representam o caminho efetuado pela bicicleta aquando da viagem. Um

exemplo de um percurso pode ser observado na figura A.25.

Observando os valores deste atributo (Figura A.26), constata-se que quase metade destes valores são nulos, significando que quase metade das viagens não apresentam este atributo. No entanto, neste caso, o atributo em questão não é de grande importância para o trabalho que se pretende efetuar e, por isso, as linhas cujos valores se encontram a nulo não serão removidas.

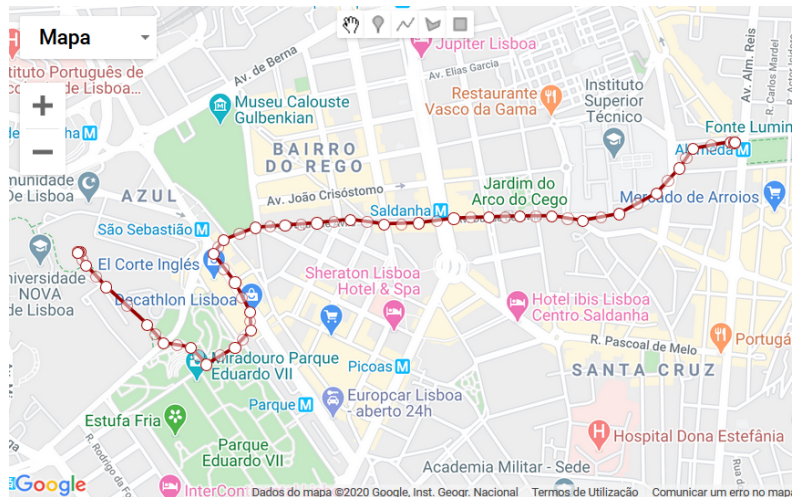


Figura A.25: Exemplo de um MULTILINESTRING representativo do percurso efetuado numa viagem.

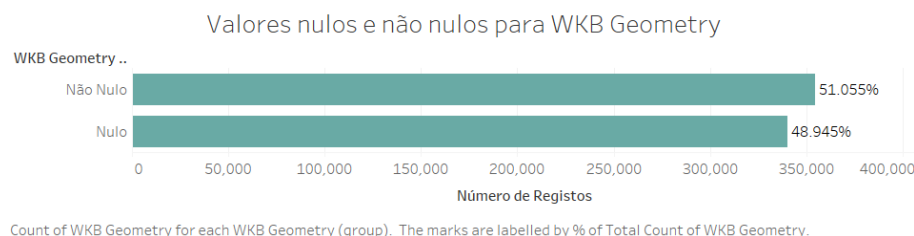


Figura A.26: Valores nulos e não nulos para o atributo *WKB Geometry*.

Número de Vértices

Este atributo diz respeito ao número de vértices que compõem a MULTILINESTRING do atributo *WKB Geometry*, dizendo respeito ao número de pontos que constitui a sequência representativa do percurso efetuado pelo cliente na sua viagem. Existem 743 valores únicos para este atributo, que estão compreendidos entre 1 e 1015.

Na figura A.27 é possível observar que quase 50% dos valores são nulos, significando que quase metade das linhas do conjunto de dados não possui informação do número de vértices do percurso. Tal poderia ser corrigido utilizando o atributo relativo ao *WKB Geometry*, no entanto, este atributo também apresenta uma enorme quantidade de valores nulos, embora o número de valores nulos do Número de Vértices seja um pouco menor

(48.956%) do que os valores nulos do *WKB Geometry* (48.945%).

O número de vértices apresenta a maioria dos seus valores entre 0 e 33 vértices, como se pode observar na figura A.28, sendo que os restantes valores se encontram distribuídos entre este valor e 1000 vértices.

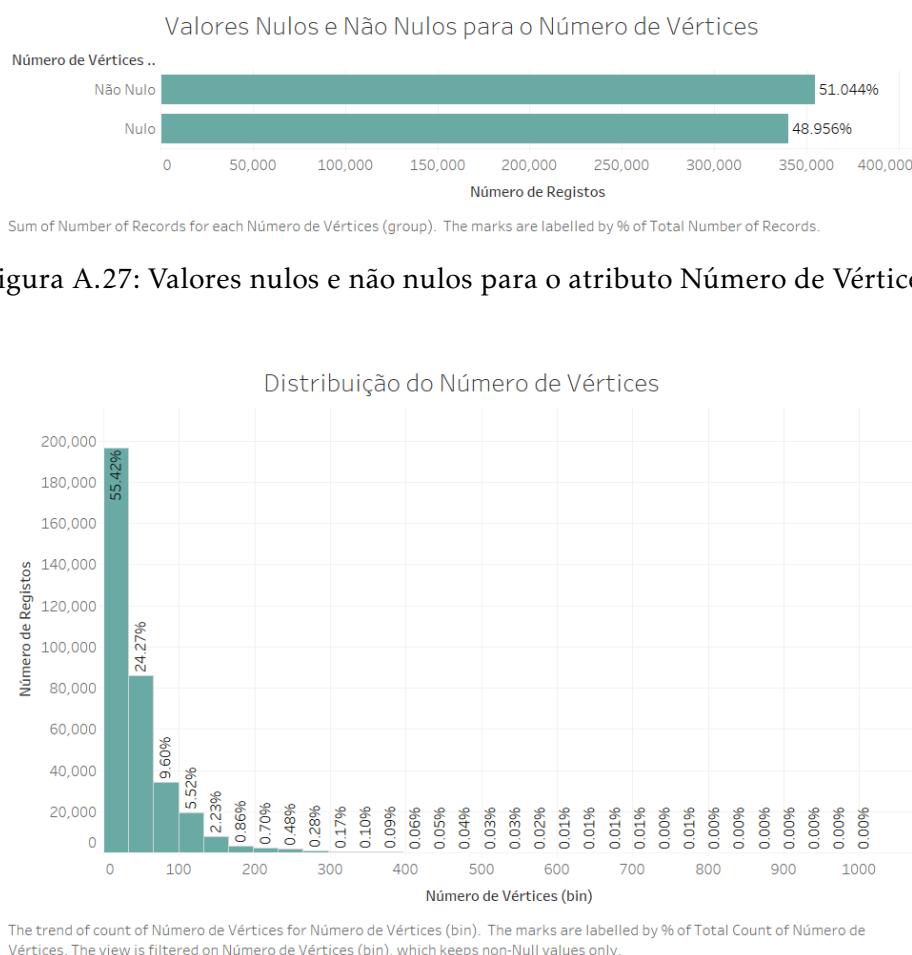


Figura A.28: Distribuição dos Valores Não Nulos para o Número de Vértices.

Tipo de Bicicleta

O tipo de bicicleta é um atributo categórico que pode tomar os valores E ou C, que significam que a bicicleta utilizada para efetuar a viagem é Elétrica ou Clássica, respetivamente.

A distribuição deste atributo, que pode ser observada na figura A.29, permite constatar que existem 97 valores nulos presentes nos dados, embora não represente uma percentagem significativa destes. Para além disso, é possível concluir que a maioria dos registos diz respeito a viagens cujo tipo de bicicleta é E (elétrica), correspondendo a quase 80% dos registos.

No entanto, como se constatou anteriormente através da exploração do atributo relativo ao *RFID* das bicicletas, o mesmo *RFID* de uma bicicleta apresenta valores C e E, ou seja, esse *RFID* é registado nos dados como sendo de uma bicicleta elétrica e clássica.

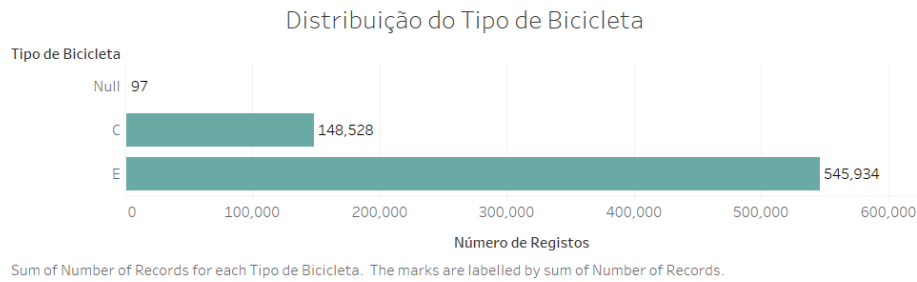


Figura A.29: Distribuição do tipo de bicicleta.

A.3.3 Qualidade dos Dados

Aquando da análise exploratória dos dados verificaram-se alguns problemas na qualidade dos mesmos, que dizem respeito a valores nulos, incorretos e valores a zero.

O atributo relativo à distância apresentava quase metade dos valores a nulo, o que é um problema na qualidade dos dados. Devido a este valor ser elevado, não é aconselhável a remoção destas linhas, uma vez que o atributo em causa também apresenta importância para a análise que se pretende levar a cabo, sendo que esta irá incidir na informação relativa à estação de fim e início, bem como às horas respetivas, dados que não apresentam valores nulos.

No entanto, este atributo revela-se importante para a análise a efetuar pelo João Dias, que irá encontrar padrões nos dados em questão. Como tal, de forma a aumentar a qualidade dos dados, deverão ser estimadas as distâncias das viagens, o que pode ser feito de duas formas.

A distância poderá ser estimada utilizando, por um lado, os dados relativos à Estação de Início e à Estação de Fim, pois ambos este atributos dão informação acerca das estações de início e término da viagem, cujas localizações podem ser determinadas facilmente consultando o mapa das Estações Gira. No entanto, as distâncias determinadas utilizando este método serão em linha reta de uma estação à outra, não sendo possível determinar ao certo o percurso feito pela bicicleta.

Uma forma diferente de estimar a distância seria determinar esta distância utilizando não só as estações, mas também a duração da viagem, que poderá ser obtida através dos atributos da Data de início e Data de fim, que permitem saber quando ocorreu o início e fim de uma viagem. A partir da duração da viagem, e admitindo que o percurso foi efetuado a uma velocidade normal no que diz respeito a uma bicicleta, pode ser estimada a distância percorrida. Essa estimativa será efetuada pelo João Dias, uma vez que é de interesse para a problemática que irá ser resolvida pelo mesmo, cujo objetivo é detetar padrões de utilização do Sistema de Bicicletas Partilhadas da Gira.

Os valores relativos à Estação de Fim apresentavam duas linhas cujo valor era nulo. De forma a aumentar a qualidade dos dados poderia proceder-se à remoção dessas linhas ou à sua correção. Neste caso, ambas as linhas contêm valores não nulos e válidos para o

atributo de *WKB Geometry*, por isso, será possível efetuar a correção destas linhas, preenchendo a estação na qual ocorreu a entrega da bicicleta.

O atributo *WKB Geometry* apresentava quase metade dos valores a nulo, significando que não é possível ter acesso ao percurso efetuado pela bicicleta. Neste caso, tal como para a distância, não é vantajoso proceder à remoção destas linhas, visto que se iria estar a eliminar quase metade das linhas do conjunto de dados. Para além disso, este atributo não é algo que possa ser estimado ou corrigido, por isso, estes valores irão manter-se.

Como o número de vértices diz respeito ao número de pontos na sequência dada pelo percurso da bicicleta, este atributo encontra-se relacionado com o atributo *WKB Geometry*, apresentando também quase metade dos valores a nulo. Neste caso, também não é vantajoso eliminar estas linhas e não é possível estimar este valor nem corrigi-lo utilizando outros atributos. Como tal, estes valores irão manter-se.

O tipo de bicicleta apresentava alguns valores nulos, no entanto, são apenas 97 e não apresentam utilidade para a análise, não sendo necessária a sua remoção. O problema mais grave deste atributo é o facto de cada bicicleta ser classificada como clássica e elétrica, sendo que cada valor (E e C) apresenta valores únicos iguais ao número total de bicicletas. Tal indica que a informação relativa ao tipo de bicicletas se encontra errada e, por isso, este atributo não poderá ser tido em conta para futuras análises, caso não se consiga corrigir esta informação.

PROCESSAMENTO DOS DADOS

Após a fase de análise exploratória dos conjuntos de dados fornecidos no âmbito da dissertação, foi necessário proceder ao seu processamento, tendo-se recorrido à metodologia **CRISP-DM**, uma abordagem composta por vários passos de forma a proceder à mineração de dados.

Tendo em conta as conclusões retiradas do processo de análise exploratória dos dados e qualidade de dados, procedeu-se à fase de *Data Preparation*. Esta fase é considerada a mais demorada de todo o processo de mineração de dados, podendo incluir junção de dados, remoção de atributos, correção ou substituição de valores incorretos ou em falta, entre outros.

Neste capítulo será descrito o processamento de dados relativo aos conjuntos de dados das Estações Gira e das Viagens Gira, cujo relatório de análise de qualidade de dados se encontra no Apêndice A. Esta foi levada a cabo tendo em conta as conclusões retiradas da qualidade dos dados e o capítulo encontra-se dividido em duas secções, cada uma representando um conjunto de dados. Cada secção encontra-se dividida em sub-secções indicativas das operações de processamento que foram realizadas.

B.1 Estações Gira

Na fase de análise exploratória de dados foi efetuada a análise dos dados relativos às Estações da rede de bicicletas partilhadas Gira, que consistem em registos da quantidade de bicicletas presente em cada estação da rede numa data e hora, ao longo do ano de 2018. Com base nas conclusões retiradas relativamente à qualidade destes dados, foi necessário efetuar algum processamento nestes dados, nomeadamente remoção de atributos, correção de valores em alguns atributos, remoção de linhas nos dados e construção de novos dados.

B.1.1 Remoção de atributos

Tendo por base a análise exploratória efetuada, foi tomada a decisão de remover alguns atributos dos dados.

No caso dos atributos Id_expl e Id_planeamento, constatou-se que estes eram iguais e, por isso, uma das colunas foi removida dos dados, de forma a não ter informação repetida e igual. Para além deste, também foi removido o atributo “Tipo de Serviço Níveis”, uma vez que este atributo apenas dá informação acerca do serviço do contratante, não sendo relevante para a análise que se pretende levar a cabo.

O atributo WKB Geometry também será removido após ser efetuada a conversão do mesmo para latitude e longitude, sendo que irá ser construído um novo conjunto de dados onde consta, para cada estação, a sua localização.

B.1.2 Morada

Na análise preliminar efetuada no período de preparação, verificou-se que alguns identificadores das Estações estavam associados a mais do que uma morada e vice-versa, sendo necessária uma análise mais aprofundada para poder corrigir esses valores.

Na Figura B.1, observam-se os identificadores e moradas em questão. Para cada um destes valores, será necessário confirmar qual a morada ou identificador correto sendo que, para efetuar a análise em questão, se recorreu ao website da Gira¹.

No caso da Morada “103 - Jardim de Água”, existem dois identificadores: 1 e 103, sendo que o primeiro apenas está presente em uma linha e o segundo em 78518 linhas. No website da Gira, esta estação apresenta o identificador 001, como se observa na figura B.2, no entanto, este irá ser alterado para 103, visto que é esse o identificador que fica de acordo com o que está na Morada.

No caso do identificador 115, a morada respetiva também está presente em 1,856 linhas cujo identificador é 421. Observando os valores de localização destas linhas, verifica-se que é um multiponto com coordenadas a zero, não sendo possível perceber se o que está incorreto é o identificador ou a morada. Analisando o número de docas destes registos (Tabela B.1), verificou-se que, para todos o valor do número de docas e do número de bicicletas é zero e, por isso, deve-se proceder à remoção das mesmas.

Morada	Identificador	Número de Docas	Data de update	Número de Registos
115 - Passeio dos Heróis do Mar/Rua Capitão Cook	421	0	15/03/2018 19:05:14	1856
115 - Passeio Heróis do Mar (Colégio Pedro Arrupe)	115	0 12	15/03/2018 11:55:12 *	1985 72009

Tabela B.1: Morada, Identificador, Número de Docas, Data de Update e Número de Registos para a estação com identificador 421 e 115, mostrando incongruências nos dados.

¹<https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/>, Consultado em 02/04/2020.

Identificadores das Estações e Moradas

Id_expl	Morada	
1	103-Jardim da Água	1
103	103-Jardim da Água	78,518
115	115 - Passeio Heróis do Mar (Colégio Pedro Arrupe)	73,994
208	208 - Cais das Pombas	53,429
	208 - Cais do Sodré	3,001
224	224 - Martim Moniz	60,732
225	225 - Martim Moniz	60,733
303	303	3
	303 - Avenida da Liberdade / Rua das Pretas	51,293
309	309 - Rua da Palma	55,306
	309 - Rua da Palma, 243	18,560
410	410 - Rua da Mesquita / Rua Dr. Júlio Dantas	1,658
	410 - Rua da Mesquita / Universidade Nova de Lisboa	30,626
421	115 - Passeio dos Heróis do Mar/Rua Capitão Cook	1,856
	421 - Alameda D. Afonso Henriques	73,904
488	488 - Av. São João de Deus	1
	488 - Rua Fernando Namora n35 / Rua António Quadro	43,631
	488 - Rua Fernando Namora N35 / Rua António Quadro	28,381

Count of Number of Records broken down by Id_expl and Morada. The view is filtered on Morada, which keeps 17 of 105 members.

Figura B.1: Estações que apresentam mais de um identificador ou morada.

Para a Estação 208, existem duas moradas diferentes, “208 - Cais das Pombas” e “208 - Cais do Sodré”, sendo que a primeira está presente em 53,429 linhas e a segunda em 3,001. Consultando o website da Gira, verifica-se que a estação 208 apresenta como designação “208 - Cais das Pombas” e, como tal, este campo irá ser corrigido para que todas as linhas tenham este valor.

A Estação 303 apresenta três linhas nas quais a sua morada é “303”, que deverão ser corrigidas para “303 - Avenida da Liberdade/Rua das Pretas”. Aquando da verificação no website verificou-se que esta estação não aparece no mesmo neste momento, podendo significar que já não existe ou que não está em funcionamento.

No que diz respeito à Estação 309, esta apresenta 55,306 com a morada “309 - Rua da Palma” e 18,560 linhas com a morada “309 - Rua da Palma, 243”, sendo que esta última morada deverá ser corrigida para coincidir com a informação disponível online.

A estação 410 apresenta como designação “410 - Rua da Mesquita/Universidade Nova de Lisboa” e, como tal, as linhas cuja designação é “410 - Rua da Mesquita / Rua Dr. Júlio

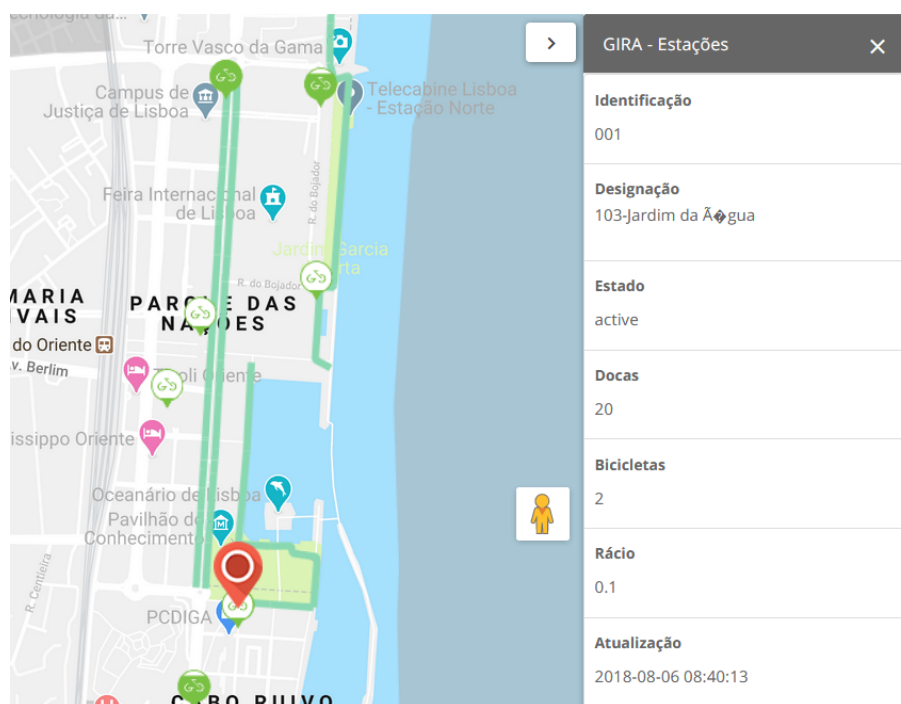


Figura B.2: Estação cuja morada é “103 - Jardim de Água” no website da Gira.

Dantas” deverão ser modificadas para o valor referido.

A estação 488 apresenta três designações diferentes, no entanto, a que consta do website é diferente destas, devendo ser alteradas para “488 - Rua Fernando Namora / Rua António Quadros”.

A estação 109 apresenta moradas diferentes nos dados e no website (“109 - Alameda dos Oceanos-Rua do Zambeze” e “109 - Alameda dos Oceanos / Rua Ilha dos Amores”, respetivamente), sendo que a morada foi modificada para estar de acordo com o website, sendo que a confirmação utilizando o atributo da localização mostra que a localização é igual à que se encontra atualmente no website. A morada da Estação 212 apresentava a morada incorreta (faltavam apenas algumas letras no nome), o que foi corrigido. Também se procedeu à correção da morada da Estação 419 (faltavam apenas algumas letras no nome). A estação 215 encontra-se no website, no entanto, a sua designação é “215 - Av. 24 de julho” e não “215 - Rua D. Luís I”, tendo-se alterado a designação da estação para ficar correta.

Ainda dizendo respeito à morada, é de referir que existem moradas iguais para estações diferentes, como se observa na Tabela B.3, sendo que, após observar o mapa do website da Gira² ao facto da sua localização ser muito próxima, . Um exemplo de estações cuja designação é a mesma pode ser observado na figura B.4, na qual as estações 406, 407 e 408 na subfigura B.4a encontram-se localizadas na Praça Duque de Saldanha, tendo a mesma morada, bem como as Estações 307 e 308 (subfigura B.4b).

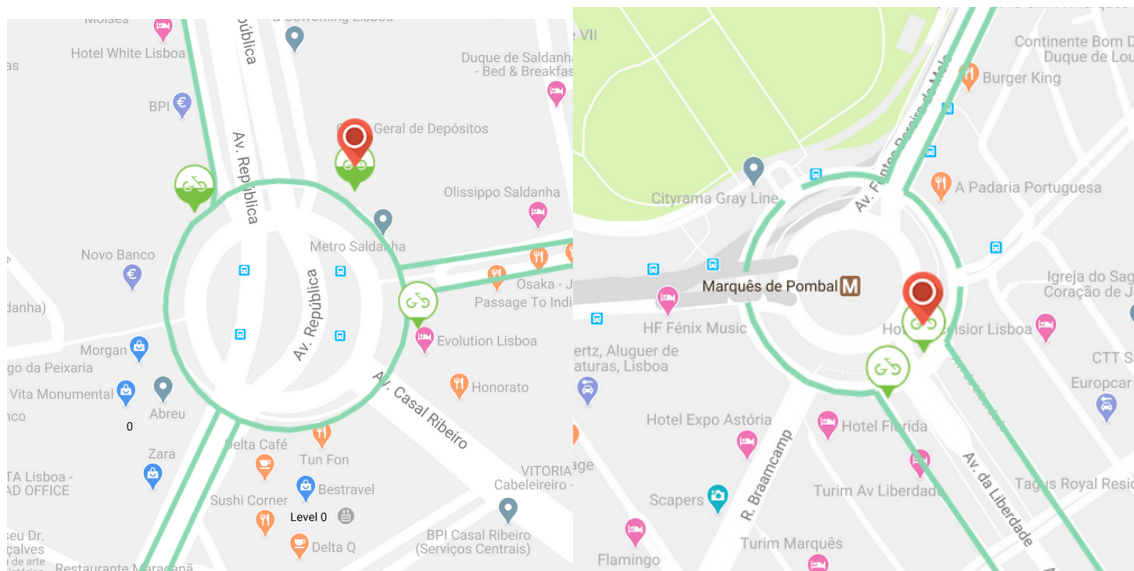
²<https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/>, Consultado a 22/08/2020.

Identificadores das Estações e Moradas

Id_expl	Morada	
224	224 - Martim Moniz	60,732
225	225 - Martim Moniz	60,733
307	307 - Marquês de Pombal	72,014
308	308 - Marquês de Pombal	70,829
406	406 - Praça Duque de Saldanha	78,508
407	407 - Praça Duque de Saldanha	78,509
408	408 - Praça Duque de Saldanha	78,511
412	412 - Av. Ressano Garcia/Av António Augusto Aguiar	78,515
413	413 - Av Ressano Garcia/Av. António Augusto Aguiar	78,514

Count of Number of Records broken down by Id_expl and Morada. The view is filtered on Morada, which keeps 9 of 105 members.

Figura B.3: Estações cuja designação é igual à exceção do identificador na mesma.



(a) Estações 406, 407, 408 na Praça Duque de Saldanha. (b) Estações 307 e 308 no Marquês de Pombal.

Figura B.4: Estações no Marquês de Pombal e Saldanha.

B.1.2.1 999 - Test Drive

Na análise realizada aquando da preparação constatou-se a existência de uma estação cujo identificador é 999 e a morada "999 - Test Drive", que diz respeito a 78,499 linhas do conjunto de dados.

Os registos relativos a esta estação ocorreram de 23 de fevereiro a 18 de dezembro, sendo que o número de bicicletas foi sempre zero e o número de docas foi sempre 9. O número de registos diários varia, havendo um pico de 2891 registos no dia 23 de fevereiro e outro de 6163 a 12 de março de 2018, como se observa na figura B.5.

Estas linhas foram removidas do conjunto de dados devido a servirem apenas como teste.

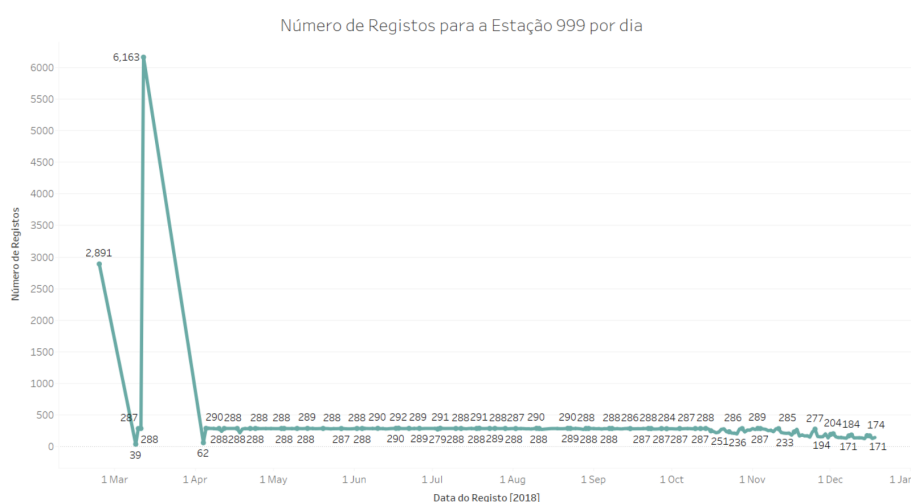


Figura B.5: Registos diários da Estação 999 ao longo de 2018.

B.1.3 Localização das Estações (WKB Geometry)

Nos dados, a localização das estações é representada pelo atributo **WKB Geometry**, que é um formato binário para guardar formatos geométricos, tendo sido observado na fase de preparação que 48 estações apresentavam mais de um valor para este parâmetro, sendo necessário proceder à sua correção.

Para tal, foi efetuada a verificação de qual dos valores era o correto para cada uma destas estações, através da conversão do valor WKB para WKT, seguida da observação da localização no mapa e, finalmente, procedeu-se à confirmação da localização utilizando o website da Gira e a morada da Estação.

No que diz respeito às Estações com mais de um valor para o atributo **WKB Geometry**, como é possível observar na tabela B.2, 29 destas apresentavam dois valores, 13 apresentavam três valores distintos, 5 estações apresentam quatro valores distintos para o WKB e 1 apresenta seis valores distintos.

As estações com dois valores distintos apresentavam quase sempre um deles como sendo um multiponto de coordenadas a zero, à exceção das Estações 431 e 450, cujos

Estações	Valores Distintos de WKB
112 114 203 205 206 207 208	2
212 214 216 301 303 306 308	
401 210 421 431 432 433 450	
456 462 476 477 479 480 483	
488	
115 222 224 304 307 309 420	3
453 473 484 485 486 490	
218 219 225 443 487	4
305	6

Tabela B.2: Estações da Gira para as quais existe mais de um valor distinto de WKB.

valores eram coordenadas válidas, mas ligeiramente diferentes, sendo que neste caso foi escolhida a que era a correta. Para as estações com três ou mais valores distintos, todos estes valores foram avaliados e foi escolhido o correto. Os erros presentes neste atributo diziam respeito a valores relativos a outras estações ou valores perto da estação correta, mas que não eram o valor correto.

Nove estações apenas apresentavam um valor de WKB, no entanto, após efetuada a conversão, verificou-se que este dizia respeito a coordenadas a zero. De forma a poder introduzir informação referente à localização destas estações, recorreu-se ao website da Gira, mais concretamente ao mapa, no entanto, verificou-se que algumas estações não se encontravam visíveis no mesmo. Na Tabela B.3, encontram-se as 9 estações referidas, com informação relativa à visibilidade no website da Gira. Para as estações que se encontravam visíveis (113, 215, 461, 475 e 478), foi utilizada essa informação para proceder à correção dos dados.

Identificador	Morada	Visível no Website
113	113 - Alameda dos Oceanos/ Passeio do Levante	Sim
213	213 - Av. Infante Dom Henrique / Santa Apolónia	Não
215	215 - Av. 24 de julho	Sim
405	405 - Rua Pedro Nunes / Av. 5 de outubro	Não
441	441 - Rua Dr. Eduardo Neves / Av. 5 de outubro	Não
448	448 - Av. São João de Deus	Não
461	461 - Rua Professora Teresa Ambrósio	Sim
475	475 - Av. Professor Gama Pinto / Cantina Velha	Sim
478	478 - Alameda da Universidade / Faculdade de Letras	Sim

Tabela B.3: Estações que apresentam apenas um valor de localização que são coordenadas a zero, morada dessa estação e informação acerca da estação estar visível no website ou não.

Para a restantes, de forma a proceder à correção dos dados relativos à localização, optou-se por aproximar a sua localização tendo em conta o atributo da Morada, que contém a rua onde cada estação se encontra. Como se observa na tabela B.3, a morada da estação é uma junção do identificador da mesma seguida da rua ou ruas onde esta se

encontra.

Desta forma, no caso da Estação 213, cuja designação é “213 - Av. Infante Dom Henrique / Santa Apolónia”, esta foi aproximada mediante entre ambas as ruas, como se pode observar na figura B.6. A localização da estação 405 foi registada num local próximo da intersecção entre as ruas “Av. 5 de outubro” e “Rua Pedro Nunes” (Figura B.7).



Figura B.6: Localização aproximada da Estação 213.



Figura B.7: Localização aproximada da Estação 405.

No caso da estação 441 (Figura B.8), foi aproximada uma localização entre as ruas “Rua Dr. Eduardo Neves” e “Av. 5 de outubro”. O mesmo procedimento foi efetuado para a estação 448 (Figura B.9), localizada na Avenida São João de Deus e para a estação 461, localizada na Rua Professora Teresa Ambrósio.



Figura B.8: Localização aproximada da Estação 441.



Figura B.9: Localização aproximada da Estação 448.

B.1.4 Data de update

Constatou-se que existem duplicados nos dados, mais concretamente, registos efetuados para uma mesma estação à mesma hora, minuto e segundo, cujo número de bicicletas registado é diferente.

Tal ocorre para várias estações e também em diferentes dias, sendo possível observar quais as estações e respetivas datas na Tabela B.4.

Data e Hora	Número de Estações	Identificadores das Estações
23/02/2018	44	101 102 103 104 105 106 107 108 109 110 403 406 407 408 412 413 414 415 416 417 423 426 427 428 430 431 442 446 449 450 452 453 457 459 460 463 464 468 471 472 473 474 481 999
12/03/2018	44	101 102 103 104 105 106 107 108 109 110 403 406 407 408 412 413 414 415 416 417 423 426 427 428 430 431 442 446 449 450 452 453 457 459 460 463 464 468 471 472 473 474 481 999
15/03/2018	10	115 219 304 305 306 309 420 421 484 485
22/03/2018	4	307 486 487
26/03/2018	1	308
27/03/2018	1	443

Tabela B.4: Data e hora, número de estações e estações onde foram encontrados registos duplicados nos dados.

Observando a tabela, é perceptível que os registos duplicados se verificam apenas nos meses de fevereiro e março e, consultando os dados nos meses de fevereiro e março, relativamente ao primeiro mês só existem registos no dia 27 e no caso do mês de março, para 8 dias (9, 10, 11, 12, 15, 22, 26 e 27).

Analisando os registos duplicados detetados de forma mais detalhada, verificou-se que, para uma mesma estação, existem vários valores diferentes para o número de bicicletas no mesmo timestamp, indicando que houve algum problema na recolha dos dados, sendo possível consultar um exemplo referente aos duplicados detetados na estação 101 na Tabela B.10. Esta mostra os valores relativos ao número de bicicletas registados para a estação 101 no dia 23 de fevereiro, sendo possível observar que foram registados 16 valores diferentes na mesma data e hora e, para cada valor, este foi registado também várias vezes, como se observa na coluna mais à direita, que representa o número de registos. Tal ocorre para todas as estações e em todos os dias de fevereiro e março para os quais há registos, o que significa que não é possível considerar os dados mesmo após a remoção destes duplicados, visto que, mesmo depois dessa remoção, ainda ficaríamos com mais de um valor de número de bicicletas para uma dada estação na mesma hora, não sendo possível determinar qual o valor correto. Para além disso, também existem muito dias em ambos os meses para os quais não há registos.

Tendo em conta todos estes aspetos, a melhor decisão é eliminar os dados relativos a fevereiro e março pois, desta forma, não só irão ser removidos os duplicados encontrados, mas também os dados relativos a esses meses para os quais existem também registos em falta.

Second of Data de up..	Id_expl	Número ..	Número ..
23/02/2018 22:09:16	101	1	16
		2	16
		3	16
		4	16
		5	16
		6	16
		7	16
		8	16
		9	16
		10	16
		11	16
		12	16
		13	16
		14	16
		15	16
		16	16

Figura B.10: Duplicados para a Estação 101 no dia 23 de fevereiro.

B.1.5 Data de início das Estações

Na análise exploratória dos dados, constatou-se que o número de docas não era único em todas as estações, havendo estações que apresentavam mais de um valor para o número de docas ao longo do ano de 2018. Tal pode significar a existência de registos incorretos nos dados, ou a adição/remoção de docas das estações da rede, sendo necessária uma análise mais aprofundada, tendo também em conta o atributo do Estado de cada estação, de forma a perceber qual a janela temporal de interesse para cada estação.

Como tal, analisou-se, para cada estação, a evolução do número de docas ao longo do ano de 2018. Esta análise permitiu averiguar a variação do número de docas para cada estação, bem como verificar quando teve início o registo para cada estação da rede. Este início poderá significar que a adição de Estações na rede, ou a colocação de sensores em estações já existentes para se proceder ao registo dos dados da rede.

A figura B.11 mostra o número de docas ao longo do tempo para cada estação, sendo que as estações estão ordenadas de acordo com a data de início dos registos de forma decrescente e a cor mapeia o atributo de Estado, que poderá estar como ativa, para reparação ou em *stock*. É de referir que, neste visualização, ainda não foram removidos os dados referentes a fevereiro e março, de forma a ter uma visão geral da data de início do registo para todas as estações.

Observando a figura percebe-se que existem, sensivelmente, quatro grupos distintos de estações tendo em conta o mês de início dos registos e que os registos terminam todos no dia 16 de dezembro, não havendo a remoção de Estações da rede. Relativamente ao Estado de cada estação, para as que se encontram em modo *stock*, tal apenas ocorre no

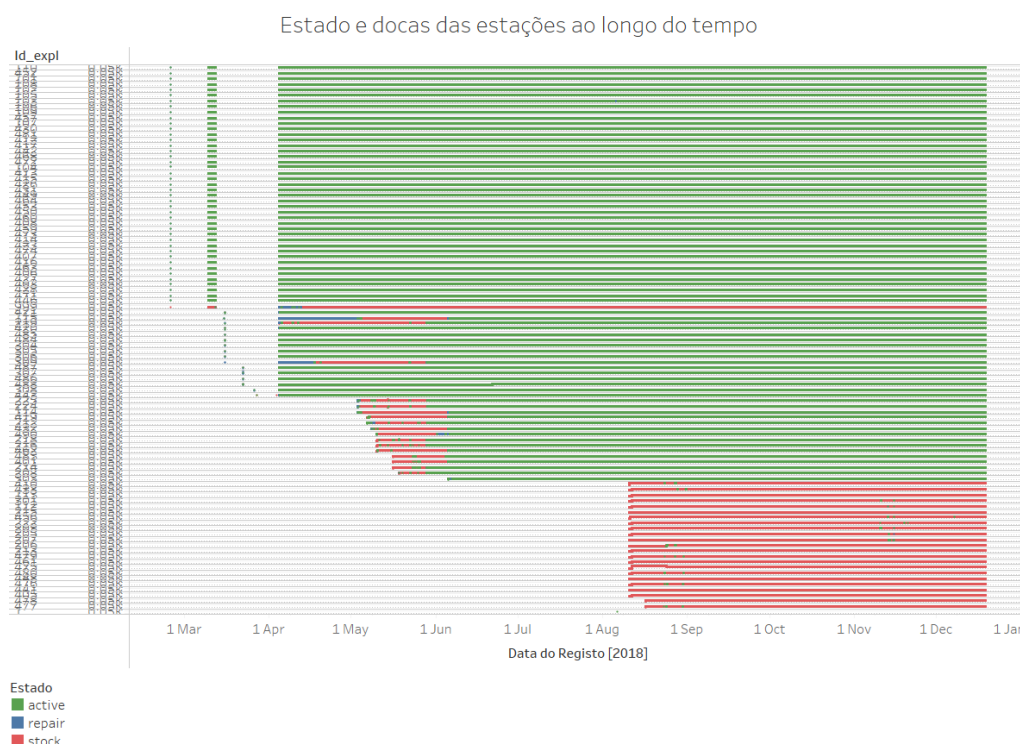


Figura B.11: Visão geral da data de início dos registos para cada Estação presente no conjunto de dados.

início dos registos de cada estação. Aquando deste estado, já se havia verificado na análise efetuada anteriormente que a estação apresenta um número de bicicletas a zero, podendo significar que este estado representa uma estação que ainda não se encontra em utilização. O estado de *repair* ocorre em várias estações de forma esporádica ao longo do ano (Figura B.12), no entanto, aquando da sua ocorrência o número de bicicletas não passa a ser sempre zero como no caso do estado *stock*.

Para determinar a data de início, as estações foram divididas em quatro grupos de acordo com o a data inicial dos registos sendo que, no caso do primeiro grupo, os registos tiveram início em fevereiro, no segundo tiveram início em março, no terceiro grupo em maio (e uma das estações em junho, embora tenha sido colocada no grupo de maio) e o grupo final no mês de agosto.

Na Tabela B.5 encontram-se as estações que compõem cada grupo de acordo com a data de início dos registos, sendo que o grupo 1 respetivo a fevereiro contém 43 Estações, o grupo de março contém 16 estações, o terceiro grupo é composto por 15 estações e o último grupo, de agosto é composto por 22 Estações, perfazendo as 96 estações presentes no conjunto de dados. Observando estes grupos no mapa (Figura B.13), é possível perceber que as estações que foram sendo adicionadas complementam as que já existiam num dado local, como se observa, por exemplo, na zona do Parque das Nações, ou na zona da Baixa.

Analisando cada grupo separadamente, pretende-se determinar qual a data de início que será considerada para cada estação, o que irá depender do estado da estação. Quando

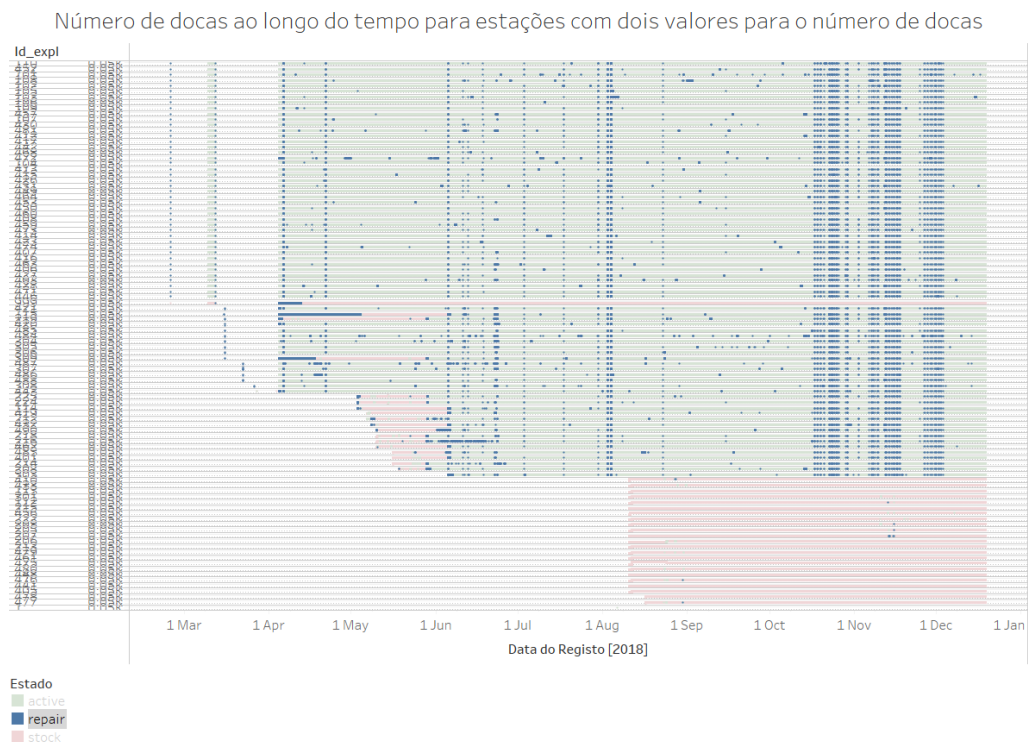


Figura B.12: Registos para cada estação ao longo do tempo, cujo estado era *repair* em destaque.

Grupo 1 (fevereiro)					Grupo 2 (março)		Grupo 3 (maio)		Grupo 4 (agosto)		
101	403	423	452	473	115	443	114	419	112	405	479
102	406	426	453	474	219	484	208	432	113	410	480
103	407	427	457	481	304	485	212	462	203	433	
104	408	428	459		305	486	214	483	205	448	
105	412	430	460		306	487	216	490	206	456	
106	413	431	463		307	488	218		207	461	
107	414	442	464		308		224		213	475	
108	415	446	468		309		225		215	476	
109	416	449	471		420		303		222	477	
110	417	450	472		421		401		301	478	

Tabela B.5: Grupos de Estações pela data de início dos registos.

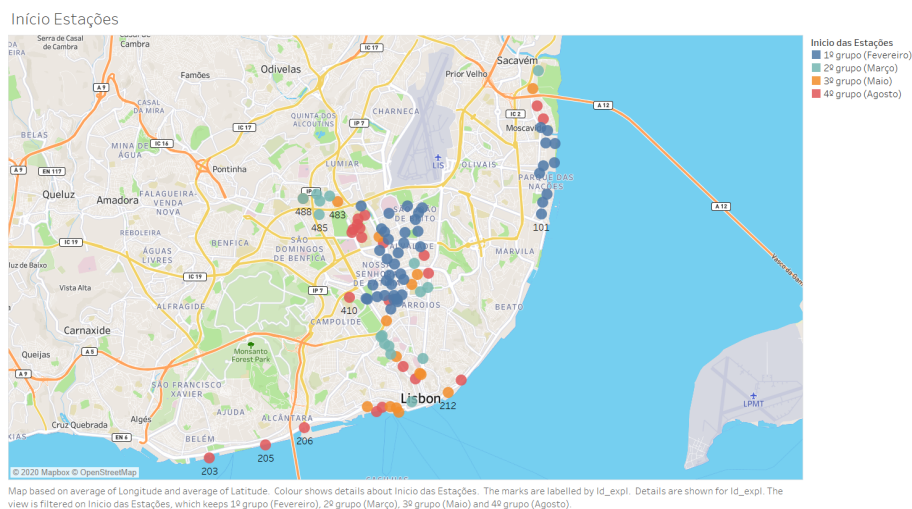


Figura B.13: Visão geral da data de início dos registos para cada Estação presente no conjunto de dados.

a estação se encontra em estado de *stock*, o seu número de bicicletas é sempre zero, no entanto, quando a estação se encontra em modo *repair* ou *active*, esta poderá ter bicicletas. Para esta análise, fará apenas sentido considerar as alturas nas quais uma estação se encontra ativa, no entanto, já se constatou que, no início do registo de uma estação, é frequente que o estado da mesma vá alternando entre *active*, *stock* e *repair*, sendo necessário averiguar se é necessário remover estes dados.

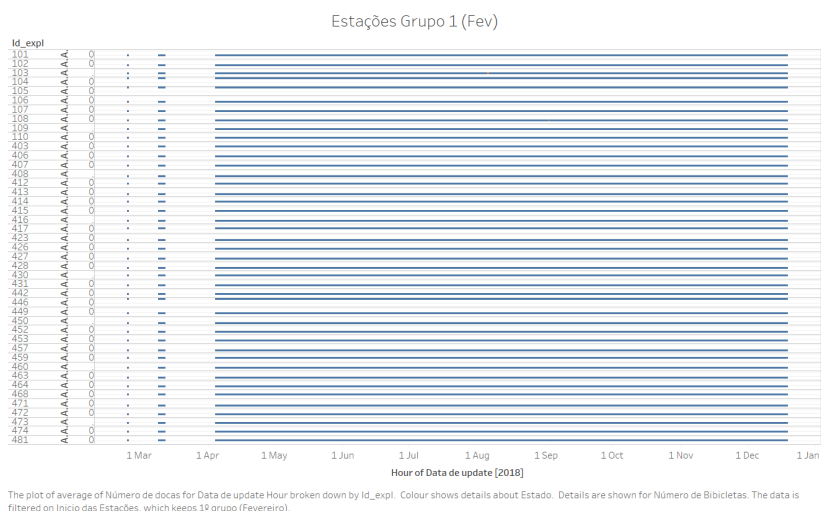


Figura B.14: Estações do Grupo 1 cujos registos tiveram início em fevereiro.

No caso do grupo 1 (Figura B.14, cuja data de início dos registos é fevereiro), é evidente que os dados relativos tanto a fevereiro como a março não são consistentes, como foi analisado na fase exploratória de dados e em maior profundidade na subsecção B.1.4, sendo necessária a sua remoção. Após este período, todas as estações apresentam registos com início no dia 4 de abril, sendo essa a data de início seleccionada.

Durante os registos destas estações, estas apenas se encontraram no estado *active* e *repair*, sendo que, quando se encontravam no estado de reparação, o número de bicicletas não era zero e o número de docas é o correto para cada estação.

Analisando o grupo 2 na figura B.15, cujos registos tiveram início em março, existe também uma interrupção nos registos que tem início de novo no dia 4 de abril, devendo ser esta a data de início considerada para a estações pertencentes a este grupo.

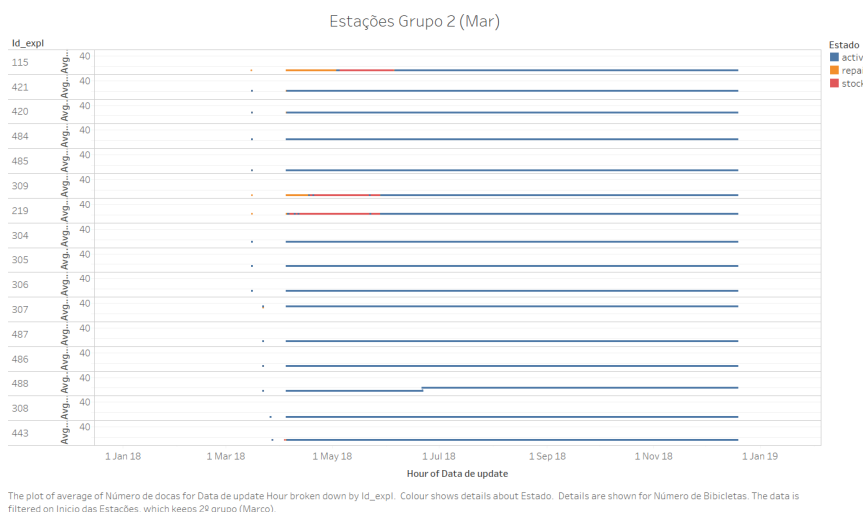


Figura B.15: Estações do Grupo 2 cujos registos tiveram início em março.

No caso de algumas estações, aquando dos seus registos iniciais, o seu estado vai-se alterando entre *stock*, *repair* e *active*, até que este permanece em estado *active* até ao final de dezembro, havendo ainda algumas leituras em que o Estado é *repair*. Uma vez que nos dias iniciais se verificou que o número de bicicletas nas estações em causa também era zero, estas linhas serão removidas, sendo considerada uma nova data de início para as estações tendo em conta este fator. Existem 6 estações nesta situação, para as quais a data de início considerada será diferente. No caso da estação 115, a data de início considerada será 6 de junho, para as Estações 420, 421 e 443 a data será 5 de abril, 29 de maio para as Estações 219 e 309 e 6 de junho para a Estação 115. É ainda interessante referir que se observa um aumento do número de docas para a Estação 488 no dia 21 de junho, que passa de 13 para 20 docas, significando que foram adicionadas docas a esta Estação (que se encontra localizada na rua Fernando Namora).

No grupo de Estações cujos registos têm início em maio (Figura B.16, as estações apresentam, no geral, um período inicial no qual o seu Estado é *stock* ou *repair* e o número de bicicletas a zero, havendo também a necessidade de modificar a data de início considerada para estas estações. Também se observa uma variação no número de docas aquando dos registos iniciais que estabiliza quando as estações ficam ativas. Analisando cada estação em separado, foram determinadas as novas datas de início para cada estação, que deverão ser 6 de junho no caso das Estações 114, 419, 432, 490, 462, 483 e 401, 29 de maio para as Estações 224, 225, 212, 218, 216, 21 e 208 e 7 de junho para a Estação 303. Na tabela

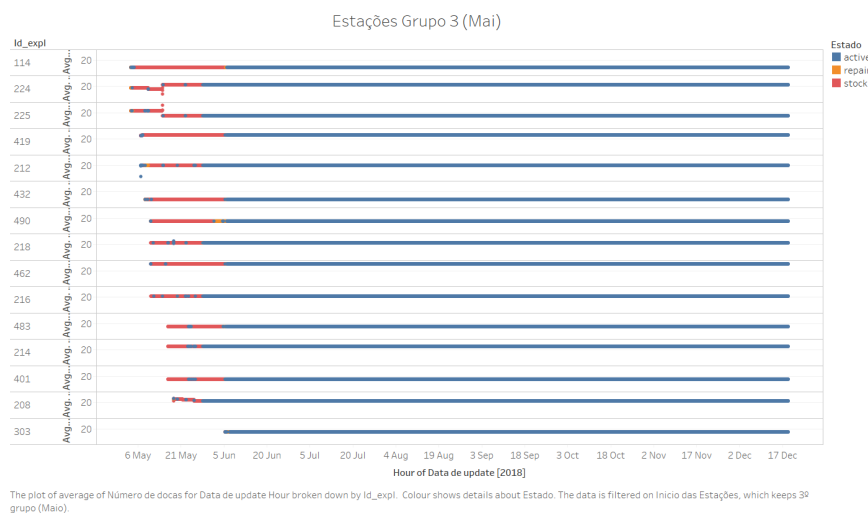


Figura B.16: Estações do Grupo 3 cujos registos tiveram início em maio.

B.6 encontram-se todas as datas de início e respetivos identificadores.

Data de início	Estações
4 de abril	101 102 103 104 105 106 107 108 109 110
	304 305 306 307 308 403 406 407 408 412
	413 414 415 416 417 423 426 427 428 430
	431 442 446 449 450 452 453 457 459 460
	463 464 468 471 472 473 474 481 484 485
	486 487 488
5 de abril	420 421 443
29 maio	208 212 214 216 218 219 224 225 309
6 de junho	114 115 303 401 419 432 462 483 490

Tabela B.6: Data de início selecionada para cada estação.

Analisando o grupo de estações de agosto, que se pode observar na figura B.17 é visível que todas as estações se mantêm em Estado de *stock* desde o período de agosto até dezembro de 2018, sendo que o seu número de bicicletas é sempre zero à exceção de raras ocasiões em que o estado se encontrava como *active*. Por este motivo, estas estações serão removidas dos dados, podendo-se consultar os identificadores das mesmas na tabela B.7.

Estações Removidas dos Dados

112 113 203 205 206
207 213 215 222 301
405 410 433 448 456
461 475 476 477 478
479 480

Tabela B.7: Identificadores das estações cujos registros foram removidos do conjunto de dados.

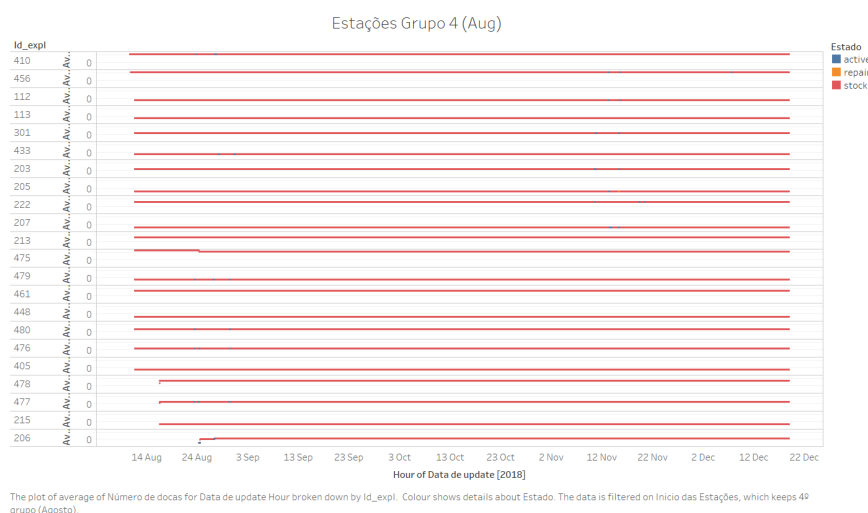


Figura B.17: Estações do Grupo 4 cujos registros tiveram início em agosto.

B.1.6 Número de Docas

Aquando da análise exploratória observou-se que haviam 62 899 linhas cujo número de docas era zero, significando que é necessária uma análise mais aprofundada das mesmas.

Na figura B.18, observa-se o número de estações nas quais há registros cujo número de docas foi zero ao longo dos meses, aparecendo ainda abaixo o número de registros respectivo para cada mês. É possível observar que estes registros se encontram distribuídos ao longo do ano de 2018 e por várias estações, não tendo ocorrido apenas num mês ou numa determinada estação e, para além disso, nos meses de março e agosto existe um maior número de estações cujo número de docas foi registado como zero, relativamente aos outros meses.

Analisando os dados, verificou-se ainda que, nas linhas que apresentam um número de docas igual a zero, o número de bicicletas também é sempre zero, o que ocorre na totalidade das estações para as quais o número de docas foi registado como sendo zero. Estas ocorrências foram analisadas de forma mais aprofundada, tendo-se construído o dashboard da figura B.19, na qual se observa a janela temporal de registros relativa a cada estação na visualização da esquerda e, na visualização da direita, apenas constam os registros em

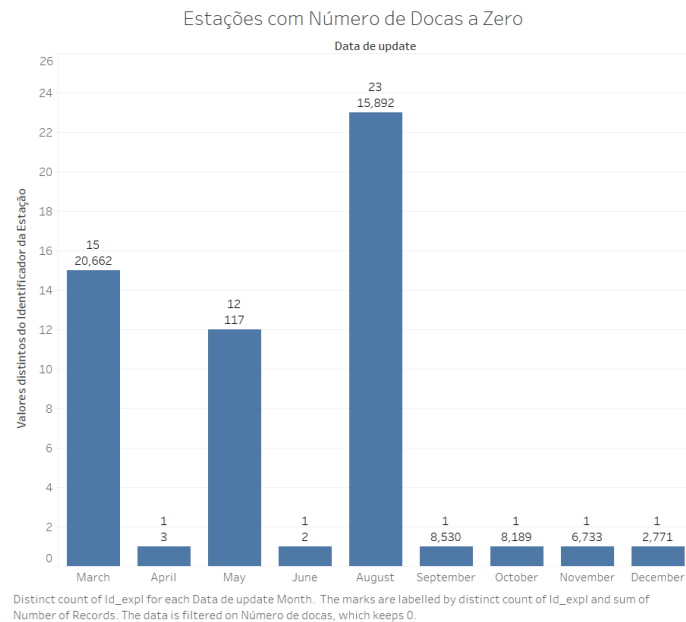


Figura B.18: Número de estações cujo número de docas foi zero em cada mês.

que o número de docas é zero. No dashboard é possível observar o realce dos registos de cada estação cujo número de docas é zero, sendo possível perceber que estes registos ocorreram aquando do início dos registos dessa mesma estação, encontrando-se compreendidos nos dados que serão removidos aquando da seleção da data de início considerada para cada estação. Desta forma, serão removidos aquando desse processamento.

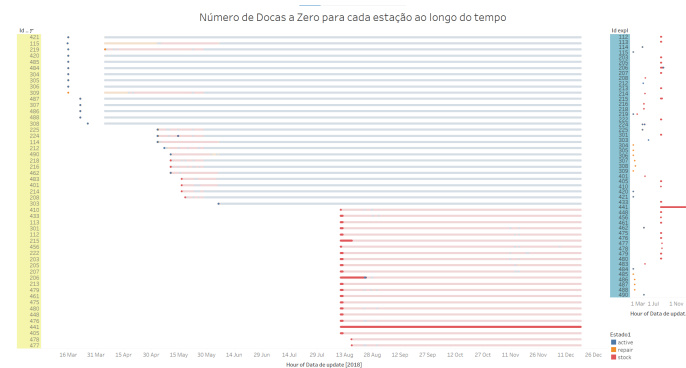


Figura B.19: Análise temporal para cada estação dos registos cujo número de docas é zero.

B.2 Viagens Gira

Aquando da análise exploratória constatou-se que o atributo WKB Geometry, que era uma multilinestring com alguns pontos representativos do percurso efetuado pela bicicleta numa viagem, bem como o atributo relativo ao número de vértices dessa multilinestring apresentavam cerca de 50% de valores nulos. Visto que estes não apresentam interesse para o trabalho a ser desenvolvido, optou-se por proceder à remoção destes atributos do

conjunto de dados.

O atributo relativo ao tipo de bicicleta também foi removido dos dados. Este poderia apresentar dois valores possíveis: E ou C, caso a bicicleta utilizada numa viagem fosse elétrica ou clássica, respetivamente, o que poderia ser interessante para a análise. No entanto, verificou-se que, para cada identificador representativo de uma bicicleta, este era simultaneamente classificado como se a bicicleta fosse elétrica e clássica, significando que este atributo se encontra com informação incorreta.

No que diz respeito ao atributo de Estação de fim, que apresenta os identificadores das estações nas quais as viagens terminaram, existiam dois registos com valor nulo. Analisando estes registos, constatou-se que havia informação relativa ao percurso efetuado pela bicicleta (atributo WKB Geometry), bem como à estação de início de ambas as viagens, que era 110 e 474, respetivamente, como se pode observar na tabela B.8. Desta forma, converteu-se o atributo para formato WKT e observou-se o percurso no mapa para ambas as viagens, que pode ser consultado nas figuras B.20 e B.22.

Data de início	Data de fim	Distância (m)	Estação de início	Estação de fim
05/02/2018, 18:26:22	05/02/2018 22:35:25	7200	474	<i>null</i>
12/02/2018 19:30:35	12/02/2018 20:00:14	6485	110	<i>null</i>

Tabela B.8: Viagens cujo atributo da Estação de fim apresenta valor a nulo.

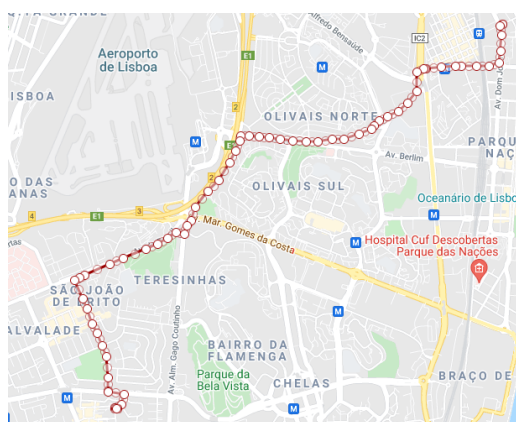


Figura B.20: Percurso efetuado na viagem com início na estação 110.

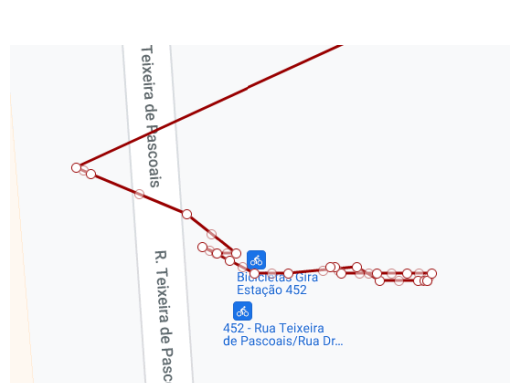


Figura B.21: Localização da estação de fim da viagem com início na estação 110.

Utilizando os atributos da data de início e data de fim, que contêm informação relativamente à data e hora a que uma viagem começou e terminou, respetivamente, procedeu-se ao cálculo da duração das viagens efetuando a diferença entre os atributos referidos. Analisando os resultados obtidos, constatou-se que existiam o que permitiu determinar a existência de durações negativas nas viagens (viagens com duração de -1 minuto). Neste caso, o atributo relativo à duração das viagens não apresenta muito interesse, no entanto, poderá averiguar-se a razão pela qual estas viagens apresentam estes valores, verificando se é necessária a remoção destas linhas dos dados.

o atributo relativo à duração da viagem foi calculado no formato HH:MM:SS, sendo

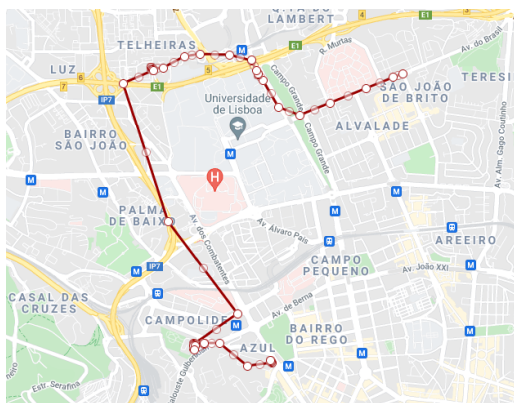


Figura B.22: Percurso efetuado na viagem com início na estação 474.



Figura B.23: Localização da estação de fim da viagem com início na estação 474.

possível observar a sua distribuição na figura B.24 constatar que a maioria das viagens teve uma duração entre 0 segundos e 09:59 minutos e 25% das viagens tiveram duração compreendida entre 10 e 19:59 minutos. Existem algumas viagens com duração muito elevada, supondo-se que não tenha sido entregue logo a bicicleta de imediato. No caso das viagens em que o atributo da distância não é nulo verifica-se que as viagens de duração elevada também apresentam distâncias muito elevadas, o que corrobora a teoria de que o cliente ficou com a bicicleta na sua posse durante aquele tempo. Como tal, essas viagens serão mantidas nos dados. No que diz respeito às viagens com duração negativa, existem viagens em que a estação de início e de fim são iguais e casos em que estas são diferentes, o que poderá indicar que as datas de início e fim se encontram incorretas, embora não haja informação relativamente à forma de recolha destes valores de data e hora. Para efeitos de análise dos dados, estas 3782 viagens foram removidas dos dados.

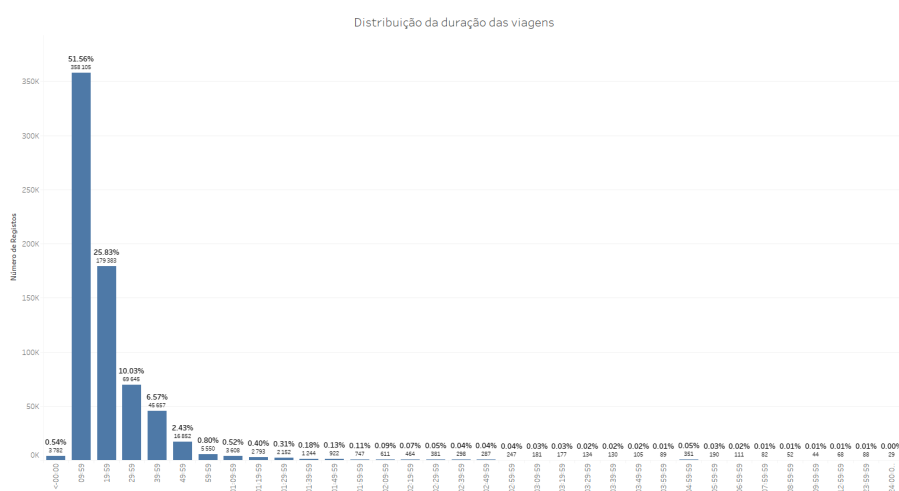


Figura B.24: Distribuição da duração das viagens em formato HH:MM:SS.

Os dados relativos às viagens apenas contêm os identificadores das estações onde uma viagem teve o seu início e fim, não havendo informação relativa à localização das estações.

Uma vez que foi construído um conjunto de dados contendo os identificadores das estações e respetiva localização em WKB, WKT, latitude e longitude, também seria apropriado utilizá-lo para agregar aos dados das viagens. No entanto, aquando da operação de *join*, constatou-se que não havia informação relativa a algumas estações. Como tal, foi necessário adicionar essa informação ao ficheiro original, consultando-se o *website* da Gira de forma a verificar a localização das estações que faltam. No entanto, algumas das estações não se encontravam visíveis no *website* e, como tal, não foi possível obter essa informação.



RELATÓRIO DE ANÁLISE DOS DADOS IPMA

No âmbito desta dissertação foram facultados, pelo *LxDataLab* (Câmara Municipal de Lisboa), vários conjuntos de dados de cariz meteorológico, de forma a serem utilizados na mesma.

Estes dados meteorológicos são provenientes do IPMA, sendo compostos por seis ficheiros de formato .xlsx com informação relativa à humidade média do ar, precipitação total, radiação solar, temperatura média do ar, intensidade média do vento e rumo médio do vento.

Para que seja possível compreender os dados, é necessário efetuar uma análise exploratória dos mesmos e, de seguida, analisar a sua qualidade.

Para proceder à análise dos dados, utilizou-se a metodologia [CRISP-DM](#), de forma a abordar de forma mais estruturada este processo, sendo que a fase levada a cabo consiste na de Compreensão dos Dados, que inclui a Descrição, Exploração e Verificação da Qualidade dos Dados.

O presente relatório encontra-se dividido em quatro secções: na primeira secção será elaborada a descrição dos dados, de seguida será efetuada a exploração de dados, a terceira secção será para a qualidade de dados e, finalmente, a última secção irá tratar da preparação de dados. Na secção de Descrição dos dados é efetuada a descrição geral dos seis conjuntos de dados meteorológicos, indicando os atributos presentes em cada conjunto de dados, o tipo de dados e as unidades de cada atributo. Nesta fase também é feita uma exploração relativa às estações meteorológicas que foi necessária à compreensão dos dados, na qual é feita a descrição das estações, bem como da sua localização.

A análise foi levada a cabo utilizando o *Tableau Prep Builder*, de forma a ter uma ideia geral dos mesmos, o *Tableau*, para construir visualizações que possibilitem a compreensão dos dados e recorrendo à linguagem Python, de forma a fazer operações sobre os dados.

C.1 Descrição dos Dados

Os dados provenientes do IPMA são compostos por seis ficheiros de formato .xlsx com informação relativa à humidade média do ar, precipitação total, radiação solar, temperatura média do ar, intensidade média do vento e rumo médio do vento no ano de 2018 numa granularidade temporal horária.

Após consultar cada conjunto de dados, constatou-se que todos apresentam os mesmos atributos, que podem ser consultados na tabela C.1. Estes são Ano, Mês, Dia e Hora, que apresentam informação relativa a data e hora em que a leitura foi efetuada e os atributos 535, 579 e 762 dizem respeito aos valores medidos para cada estação, sendo que cada número diz respeito a uma estação meteorológica existente em Lisboa.

Uma vez que não havia informação disponível no que diz respeito às unidades de cada atributo recorreu-se ao website do IPMA¹ para averiguar as mesmas, que podem ser consultadas na tabela C.2. A Humidade do Ar é medida em percentagem, a precipitação em mm, a radiação solar é medida em W/m², a temperatura em graus centígrados, o rumo do vento encontra-se em graus e a intensidade do vento em m/s.

Atributo	Tipo de Dados	Descrição
Ano	Numérico	Ano no qual foi efetuada a leitura
Mês	Numérico	Mês no qual foi efetuada a leitura
Dia	Numérico	Dia no qual foi efetuada a leitura
Hora	Numérico	Hora na qual foi efetuada a leitura
535	Numérico	Valor medido na estação 535
579	Numérico	Valor medido na estação 579
762	Numérico	Valor medido na estação 762

Tabela C.1: Descrição dos conjuntos de dados da meteorologia.

Atributo	Unidades
Humidade do Ar	%
Precipitação	mm
Radiação Solar	W/m ²
Temperatura	° C
Rumo do Vento	° (graus)
Intensidade do Vento	m/s

Tabela C.2: Descrição das unidades de cada tipo de dados da meteorologia.

As leituras que constam de cada conjunto de dados são respetivas a três estações meteorológicas de Lisboa, que se podem observar na Tabela C.3 e cuja localização pode ser observada na figura C.1. Uma vez que não havia informação relativamente ao que cada coluna significava, de forma a perceber que era referente a estações meteorológicas, foi necessário consultar o website do IPMA², do qual se retirou a informação presente

¹<https://www.ipma.pt/pt/otempo/prev.localidade.hora/>, Consultado em 22/08/2020

²<https://www.ipma.pt/pt/enciclopedia/redes.observacao/meteo/index.jsp>, Consultado a 20/08/2020.

Nome	Número	Tipo	Latitude	Longitude	Altitude	Município	Início de Funcionamento
Lisboa / Geofísico	535	Estação Automática Principal (EMA I)	38,719	-9,149	77	Lisboa	12-07-1999
Lisboa - Gago Coutinho	579	Estação Automática Principal (EMA I)	38,766	-9,127	104	Lisboa	01-01-1996
Lisboa - Tapada da Ajuda	762	Estação Automática Climatológica (EMA II)	38,709	-9,182	62	Lisboa	06-03-2009

Tabela C.3: Estações meteorológicas de Lisboa e respetiva informação.

na tabela referida anteriormente. Estas medições foram feitas no período entre a meia noite de 1 de janeiro de 2018 e as 23:00 de 31 de janeiro de 2018 nas três Estações Meteorológicas existentes em Lisboa, sendo de granularidade temporal horária.

De forma a ser possível proceder à análise exploratória dos dados, bem como da qualidade dos mesmos, os atributos Ano, Mês, Dia e Hora foram unidos de forma a ficarem num formato Data e Hora (DD/MM/YYYY HH:00), tendo sido adicionada esta nova coluna a cada um dos conjuntos de dados em análise.

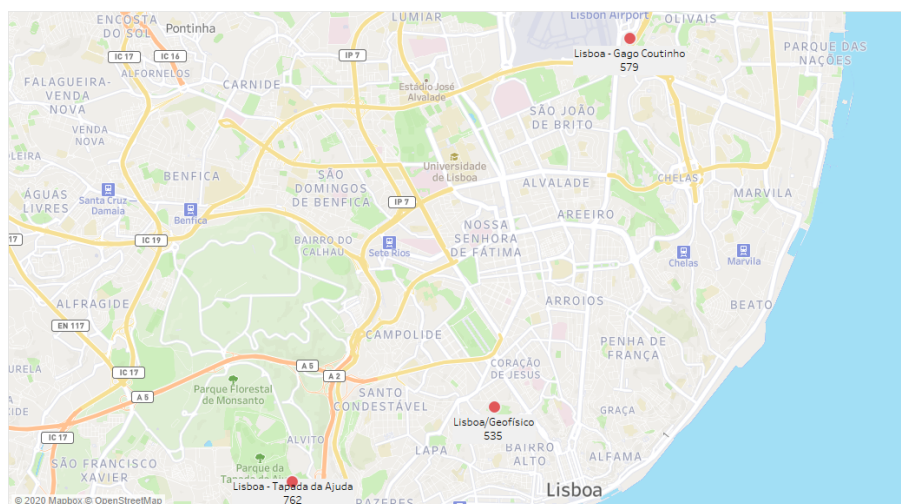


Figura C.1: Estações Meteorológicas em Lisboa em 2018.

C.2 Exploração dos Dados

C.2.1 Precipitação

O primeiro diz respeito à precipitação total, contendo 8760 registos ao longo do intervalo temporal referido acima, para cada uma das três estações meteorológicas. Constatou-se a existência de 2 registos duplicados nos dados para cada estação no dia 25 de março às 2 da manhã, cujos valores podem ser consultados na tabela C.4.

Para além disso, foram ainda observados registos cujo valor era -990, indicando uma falha na leitura ou a ocorrência de outro problema que originou falhas na recolha de medições. Este evento ocorreu nas três estações, havendo um número mais elevado destes valores na estação 762, como é possível consultar na tabela C.5. As medições de valor -990 relativas à Estação 535 ocorreram no dia 8 de outubro entre as 9 e as 11 horas, para

Timestamp	Estação 535	Estação 579	Estação 762
25/03/2018 02:00:00	1,5 2,1	0	0

Tabela C.4: Valores dos registos duplicados presentes nos dados relativos à precipitação.

a Estação 579 tiveram lugar no dia 18 de maio entre as 10 e as 12 horas, bem como no dia 20 de setembro pelas 14 horas. No caso da Estação 762, ocorreram ao longo do ano nos meses de fevereiro, abril, julho, agosto, setembro e dezembro, como é possível observar na Figura C.2. Nesta encontra-se representado um *heatmap* no qual as linhas são os meses e as colunas são os dias e a cor mapeia o número de registo inválidos num dado dia. A ausência de cor representa a inexistência de valores inválidos. Outra irregularidade detetada nos dados foi a inexistência de registos relativos a 25 de março de 2018 à 1 da manhã para as três estações.

Estação 535	Estação 579	Estação 762
3	4	515

Tabela C.5: Contagem de registos com valor -990 para cada estação meteorológica relativamente à precipitação.

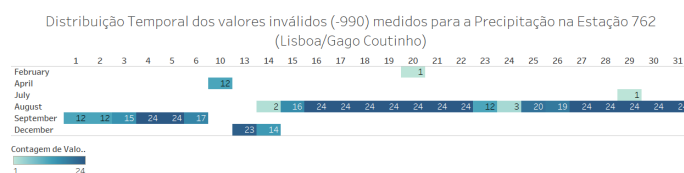


Figura C.2: Distribuição temporal dos valores inválidos (-990) medidos para a precipitação na Estação 762 em 2018.

Os valores -990 foram filtrados antes de proceder à análise da distribuição da precipitação, cujos histogramas podem ser consultados nas figuras C.3, C.4 e C.5. Estes são representativos dos valores de precipitação (mm) para cada uma das estações meteorológicas de Lisboa no ano de 2018, cujo eixo representando o número de registos é logarítmico devido ao elevado número de registos de valor zero (correspondentes à ausência de precipitação).

Observando a distribuição da precipitação total medida em cada uma das Estações verifica-se que esta é semelhante nas três estações de Lisboa, sendo enviesada para a esquerda, havendo apenas alguma variação nas medições.

Na figura C.6 é possível observar a distribuição temporal da precipitação total ao longo do ano de 2018 para as três estações meteorológicas, verificando-se que mesmo ao longo do ano as medições de precipitação são semelhantes independentemente da estação na qual foram efetuadas as medições.

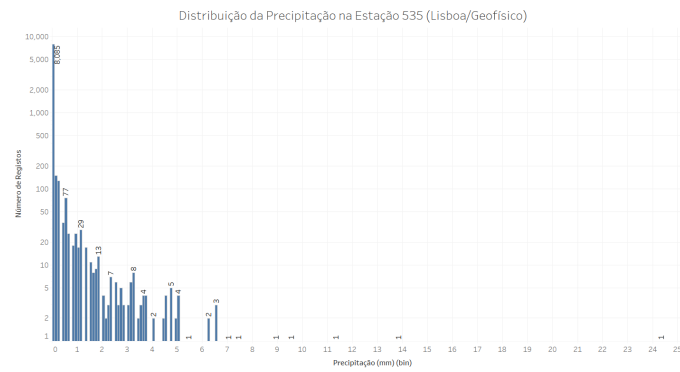


Figura C.3: Distribuição da Precipitação para a Estação 535 em 2018.

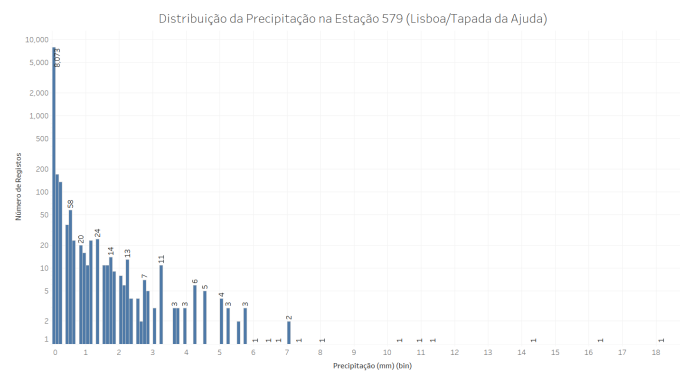


Figura C.4: Distribuição da Precipitação para a Estação 579 em 2018.

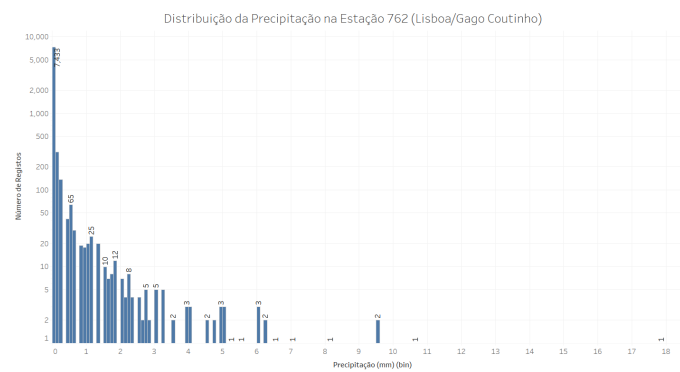


Figura C.5: Distribuição da Precipitação para a Estação 762 em 2018.

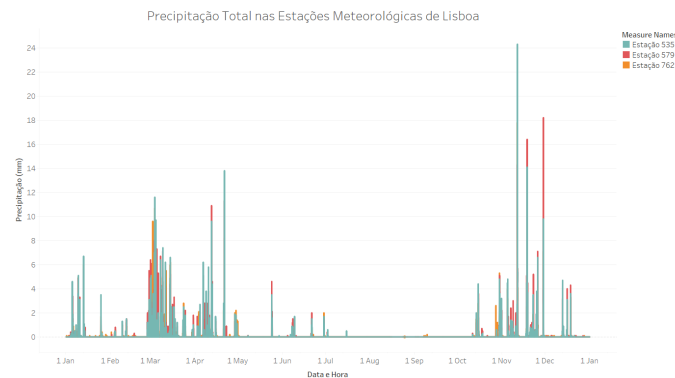


Figura C.6: Distribuição Temporal da Precipitação Total para as Estações Meteorológicas de Lisboa.

C.2.2 Humidade

Os dados relativos à humidade contêm 8760 registos, tendo sido registados valores de humidade (%) nas três estações meteorológicas mencionadas anteriormente.

À semelhança dos dados da precipitação, também neste caso foram detetados 2 registos duplicados no dia 25 de março pelas 2 da manhã, bem como valores iguais a -990, significando falhas na leitura dos valores. Foi ainda detetada a inexistência de um registo relativo a 25 de março de 2018 à 1 da manhã.

As tabelas C.6 e C.7 mostram, respetivamente, os registos duplicados e o número de registos onde houve uma falha na leitura. No que diz respeito aos registos de valor -990 verifica-se que existem apenas 2 registos com este valor para as Estações 535 e 579, mas existem 451 no caso da Estação 762. No caso da estação 535 os registos de valor -990 ocorreram a 8 de outubro às 9 e 10 horas, para a estação 579 ocorreram a 18 de maio às 11 horas e a 20 de setembro às 14 horas e, finalmente, é possível observar através da figura C.7 que, para a estação 762, os registos com este valor ocorreram entre 14 de agosto e 31 de agosto e entre 1 e 6 de setembro.

Timestamp	Estação 535	Estação 579	Estação 762
25/03/2018 02:00:00	96	94	93
	98	95	96

Tabela C.6: Valores dos registos duplicados presentes nos dados relativos à precipitação.

Estação 535	Estação 579	Estação 762
2	2	451

Tabela C.7: Contagem de registos com valor -990 para cada estação meteorológica relativamente à Humidade.

De forma a observar a distribuição da percentagem de humidade para cada estação foram efetuados três histogramas (Figuras C.8, C.9, C.10), sendo possível verificar que a distribuição da humidade nestas três estações é semelhante e mais enviesada para a

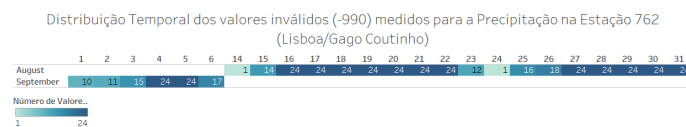


Figura C.7: Distribuição temporal dos valores inválidos (-990) medidos para a humidade na Estação 762 em 2018.

direita com a maioria dos registos a apresentar uma percentagem de humidade superior a 50%.

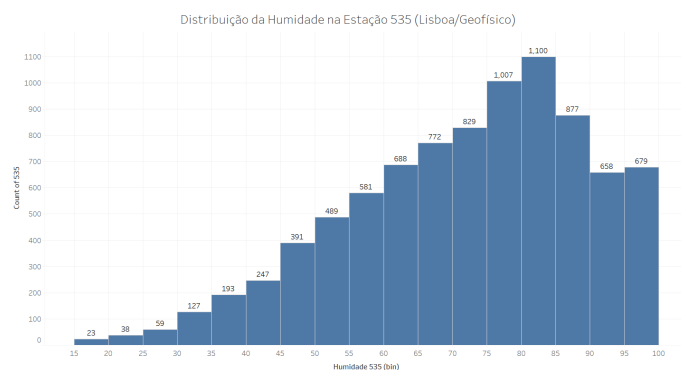


Figura C.8: Distribuição da Humidade para a Estação 535 em 2018.

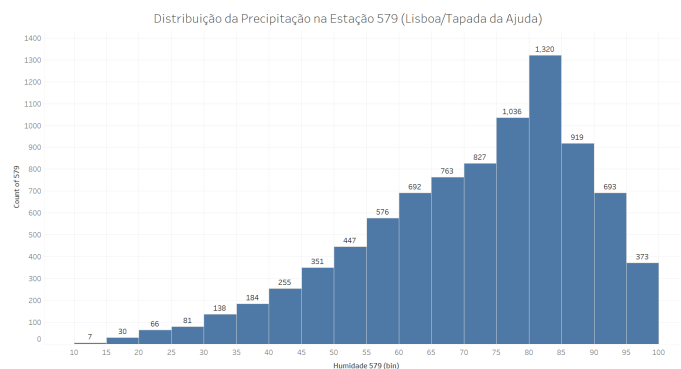


Figura C.9: Distribuição da Humidade para a Estação 579 em 2018.

Na figura C.11 é ainda possível observar a distribuição temporal da percentagem de humidade ao longo do ano de 2018 para as três estações, após filtragem dos valores inválidos, sendo que é evidente que os valores de humidade registados ao longo do ano são semelhantes para as três estações.

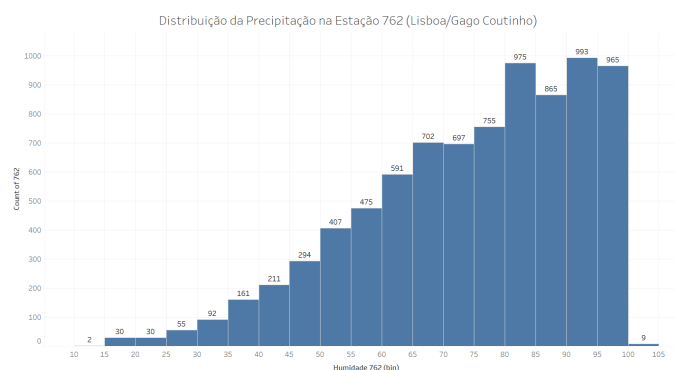


Figura C.10: Distribuição da Humidade para a Estação 762 em 2018.

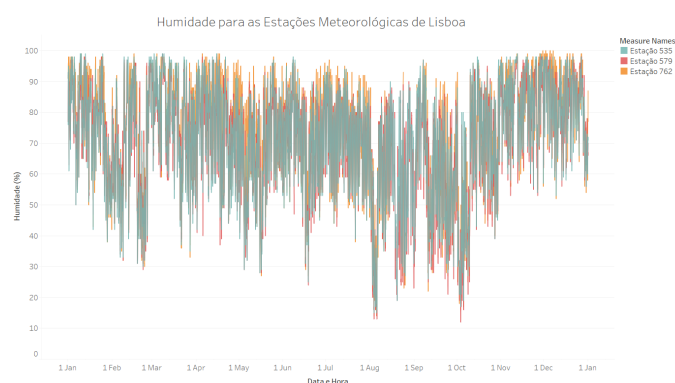


Figura C.11: Distribuição Temporal da Humidade para as Estações Meteorológicas de Lisboa.

C.2.3 Radiação

Os valores de radiação foram medidos no intervalo temporal referido acima, existindo 8760 registos.

Tal como nos atributos analisados anteriormente, também existem 2 registos duplicados no dia 25 de março às 2 da manhã (Tabela C.8).

Para além disso, constatou-se que a estação 535 apenas apresenta valores inválidos (-990) para a radiação, o que pode ser devido a uma falha ou a própria estação poderá não dispor do equipamento para efetuar este tipo de medição. Na estação 762 também existem 468 valores inválidos, o que é uma quantidade considerável, sendo que quase a totalidade destas falhas ocorreram no mês de setembro. Estes valores podem ser consultados na Tabela C.9. No que diz respeito à distribuição temporal dos valores inválidos relativos à Estação 579, estes ocorreram a 9 de maio às 13 horas, 18 de maio entre as 10 e as 12 horas, a 22 de maio às 11 horas e a 20 de setembro às 14 horas. Uma vez que existem muitos valores inválidos relativos à estação 762, foi elaborado um *heatmap* (Figura C.12 cujas linhas representam os meses, as colunas representam os dias e cada célula tem uma cor de acordo com o número de registos cujo valor é inválido para cada dia de cada mês, sendo possível averiguar que a maioria dos valores inválidos ocorreram entre 14 de agosto

e 6 de setembro, tendo havido ainda um valor inválido em fevereiro, maio e julho.

Foi ainda detetada a inexistência de um registo relativo a 25 de março de 2018 à 1 da manhã.

Timestamp	Estação 535	Estação 579	Estação 762
25/03/2018 02:00:00	-990	0	0

Tabela C.8: Valores dos registos duplicados presentes nos dados relativos à radiação.

Estação 535	Estação 579	Estação 762
8760	6	468

Tabela C.9: Contagem de registos com valor -990 para cada estação meteorológica relativamente à radiação.

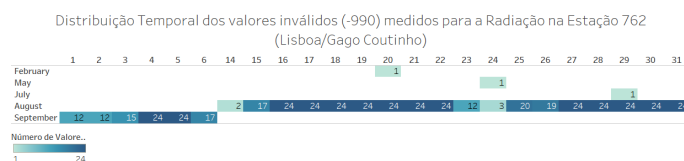


Figura C.12: Distribuição temporal dos valores inválidos (-990) medidos para a radiação na Estação 762 em 2018.

Relativamente aos valores inválidos observados na Estação 579 ocorreram a 9 de maio pelas 13 horas, 18 de maio às 10 horas, 22 de maio às 11 horas e 20 de setembro às 14 horas. No caso da Estação 762 ocorreram a 20 de fevereiro à 1 da manhã, 24 de Maio às 9 horas, 29 de julho às 14 horas e, finalmente, entre 14 de agosto às 17 horas e 6 de setembro às 16 horas.

De forma a analisar a distribuição da radiação foram filtrados todos os valores inválidos, sendo que esta pode ser observada nas Figuras C.13 e C.14 é possível observar a distribuição dos valores de radiação para as Estações 579 e 762, sendo notória a semelhança de distribuição destas duas estações. Apesar de haverem muitos valores de radiação entre 0 e 200, para valores superiores a este existe um número de registos semelhante para cada intervalo.

Na figura C.15 é ainda possível observar a distribuição temporal da da quantidade de radiação ao longo do ano de 2018 para as estações 579 e 762, após filtragem dos valores inválidos. A estação 535 não foi incluída por ter apenas valores inválidos. Observando a figura, é notória a semelhança dos valores da percentagem de humidade registados ao longo do ano para ambas as estações.

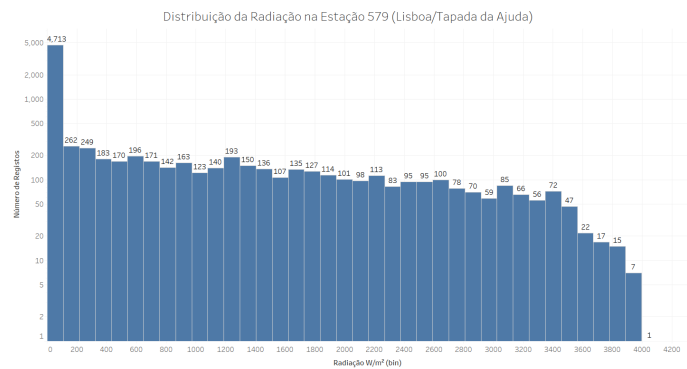


Figura C.13: Distribuição da Radiação para a Estação 579 em 2018.

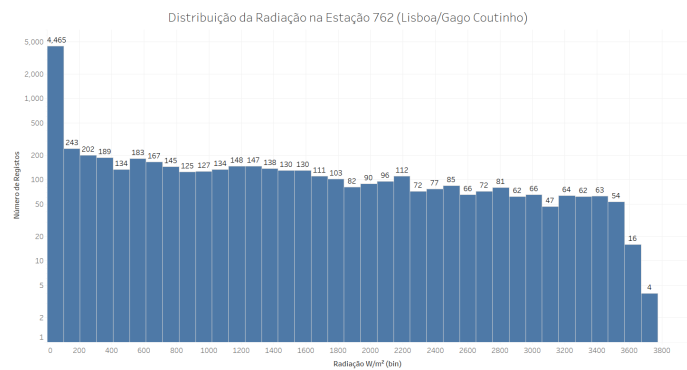


Figura C.14: Distribuição da Radiação para a Estação 762 em 2018.

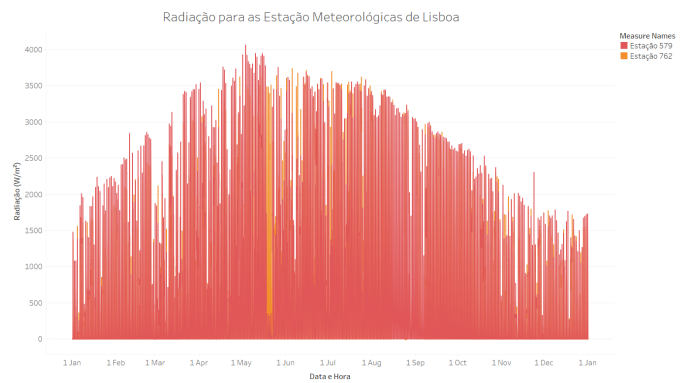


Figura C.15: Distribuição Temporal da Radiação para as estações meteorológicas de Lisboa em 2018.

C.2.4 Temperatura

Existem 8760 registos relativos à temperatura no ano de 2018, tendo sido medidos nas três estações meteorológicas de Lisboa. Foram encontrados dois registos duplicados, cujos detalhes podem ser consultados na tabela C.10, bem como registos com um valor de -990 indicando uma falha nas leituras, sendo que este número varia para cada estação, podendo ser consultado na Tabela C.11.

Timestamp	Estação 535	Estação 579	Estação 762
25/03/2018 02:00:00	11	10,3	10,8
	12,1	11,5	11,8

Tabela C.10: Valores dos registos duplicados presentes nos dados relativos à temperatura.

Estação 535	Estação 579	Estação 762
2	2	451

Tabela C.11: Contagem de registos com valor -990 para cada estação meteorológica relativamente à temperatura.

Os valores inválidos registados na Estação 535 ocorreram a 8 de outubro às 9 e 10 horas e na Estação 579 ocorreram dia 8 de maio pelas 11 horas e 20 de setembro pelas 14 horas. No caso da Estação 762 os valores inválidos ocorreram de forma intermitente entre 14 de agosto e 6 de setembro, podendo observar-se com maior detalhe na visualização da figura C.16.

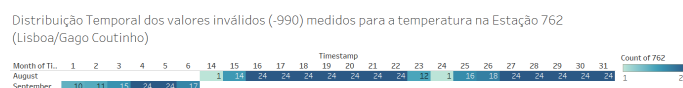


Figura C.16: Distribuição temporal dos valores inválidos (-990) medidos para a temperatura na Estação 762 em 2018.

Esta visualização é um *heatmap* representativo dos Meses e Dias nos quais a medição foi inválida (valor de -990), sendo que a cor mapeia a contagem de registos inválidos medidos para cada dia. Um valor de 24 registos inválidos significa que todas as leituras efetuadas nesse dia desse mês foram inválidas. Como referido anteriormente verifica-se que as leituras inválidas ocorreram todas nos meses de agosto e setembro, sendo que, de 16 a 22 de agosto, de 27 a 31 de agosto e em 4 e 5 de setembro todos os registos foram inválidos.

Analisando a distribuição da temperatura após a remoção dos valores inválidos detetados é possível (Figuras C.17, C.18, C.19, verifica-se que esta é semelhante para as três estações, sendo enviesada para a esquerda, havendo uma maior quantidade de registos abaixo dos 25°C, sendo que a maioria dos registos se encontra no intervalo entre 10 e 25 °C em todas as estações.

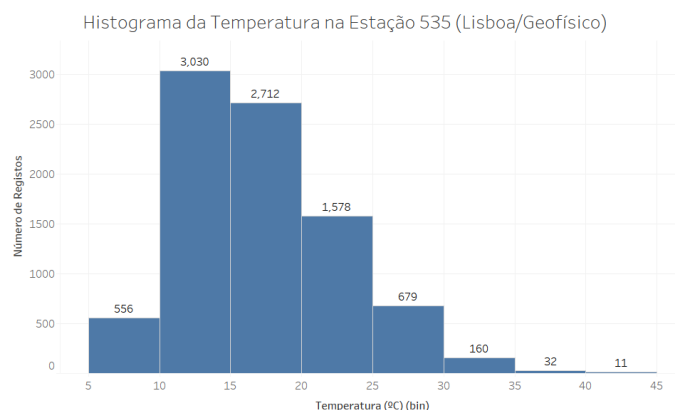


Figura C.17: Distribuição da Temperatura para a Estação 535 em 2018.

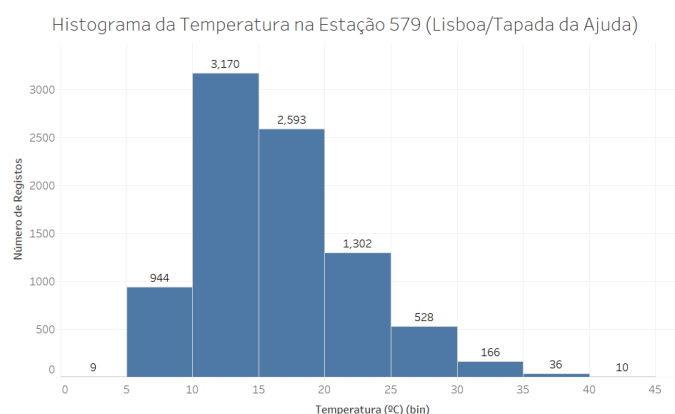


Figura C.18: Distribuição da Temperatura para a Estação 579 em 2018.

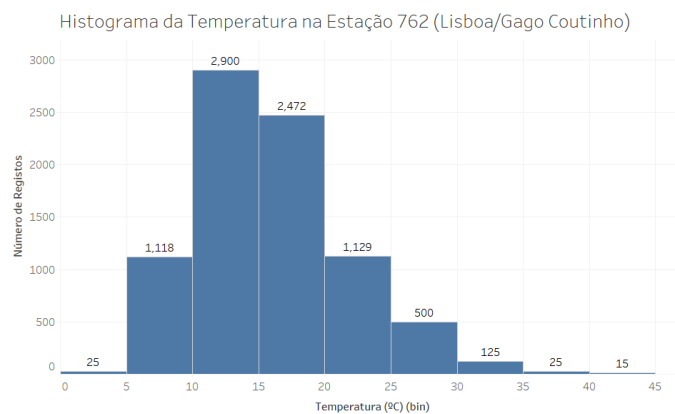


Figura C.19: Distribuição da Temperatura para a Estação 762 em 2018.

Na figura C.20 é possível observar a distribuição temporal da temperatura ao longo do ano de 2018 para as três estações, verificando-se que as medições efetuadas por cada estação foram muito semelhantes ao longo de todo o ano. É evidente que nos meses de janeiro, fevereiro, março, novembro e dezembro as temperaturas registadas foram menores, na ordem 5 a 17 °C, enquanto que a partir de abril se dá um aumento gradual da mesma, até agosto, quando se dá um pico na ordem dos 40 a 45 °C, havendo de novo uma descida gradual nos meses de setembro e outubro para temperaturas na ordem dos 20 a 35°C.

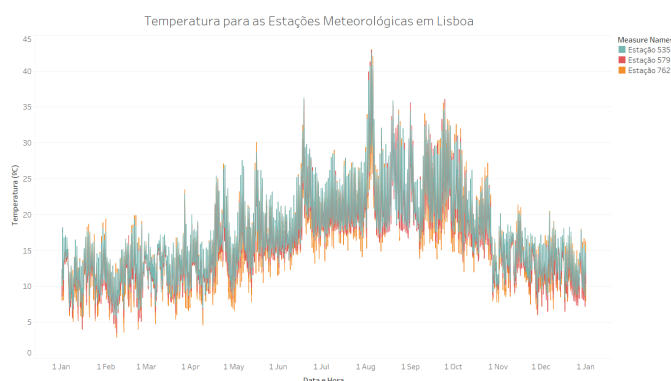


Figura C.20: Distribuição Temporal da Temperatura para todas as estações em 2018.

C.2.5 Vento (Intensidade)

Relativamente à intensidade do vento, existem 8760 registos para cada estação, tendo sido observados 2 registos duplicados (C.12). Para além disto, verificou-se a existência de registos cujo valor é de -990 para as três estações, bem como registos de valor nulo para a estação meteorológica 535 (Tabela C.13).

Timestamp	Estação 535	Estação 579	Estação 762
25/03/2018 02:00:00	-990	4	3,1
		5,9	4,1

Tabela C.12: Valores dos registos duplicados presentes nos dados relativos à intensidade do vento.

Valores	Estação 535	Estação 579	Estação 762
-990	8378	72	451
Nulos	343	0	0

Tabela C.13: Contagem de registos com valor -990 e nulo para cada estação meteorológica relativamente à intensidade do vento.

De forma a observar a distribuição dos valores inválidos e nulos foi criada a visualização da figura C.21, um *heatmap* no qual as linhas representam os meses, as colunas

representam os dias e, em cada dia, a cor representa o número de registos cujo valor é nulo (caso tenha um asterisco) ou inválido. Observando o *heatmap* é possível perceber os valores válidos e nulos foram registados entre 21 de maio e 7 de junho (apresenta o * e é menor de 24) e que os valores inválidos registados ocorreram todos antes e depois deste intervalo temporal.

Foi elaborado o mesmo *heatmap* para representar os 451 valores inválidos referentes à estação 762, sendo possível consultar o mesmo na figura C.22. No caso da estação 579, os valores inválidos ocorreram entre os dias 18 e 21 de Maio.

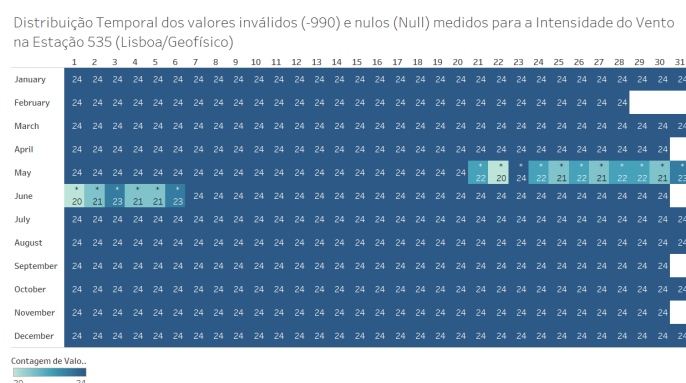


Figura C.21: Contagem dos valores inválidos e nulos para cada mês e dia na estação 535 ao longo do ano de 2018. As células que apresentam (*) representam valores nulos, as restantes são valores inválidos.



Figura C.22: Contagem dos valores inválidos e nulos para cada mês e dia na estação 762 ao longo do ano de 2018.

Para observar a distribuição da Intensidade do Vento relativa às três estações meteorológicas não foram tidos em conta os valores inválidos nem nulos, sendo possível consultar a distribuição mencionada nas figuras C.23, C.24, C.25.

A distribuição da intensidade do vento é enviesada para a esquerda no caso de todas as estações, embora a quantidade de registos em cada bin não seja semelhante entre elas. A distribuição observada para os valores referentes à Estação 535 são os que mais diferem, provavelmente devido ao facto de haverem muito poucos registos válidos medidos nessa estação.

Na figura C.26 é possível observar a distribuição temporal da intensidade do vento para todas as estações no ano de 2018, sendo notória a escassez de registos relativos à estação 535. Para as duas estações restantes, os valores da intensidade do vento medidos foram semelhantes, embora tenham sido medidos valores superiores da intensidade na estação 762 do que na estação 579, o que também é corroborado pelos histogramas, nos

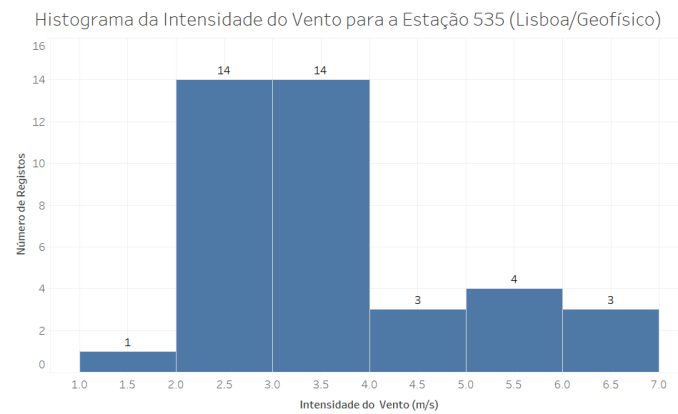


Figura C.23: Distribuição da Intensidade do Vento para a Estação 535 em 2018.

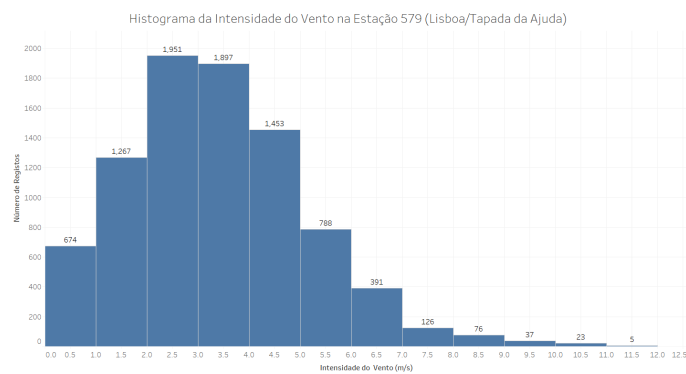


Figura C.24: Distribuição da Intensidade do Vento para a Estação 579 em 2018.

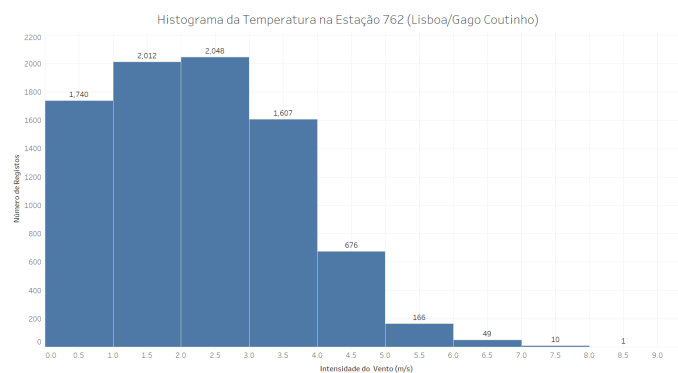


Figura C.25: Distribuição da Intensidade do Vento para a Estação 762 em 2018.

quais os valores máximos de intensidade medidos foram 8 e 11, havendo ainda cerca de 65 valores acima de 9.0 m/s relativo à estação 579.

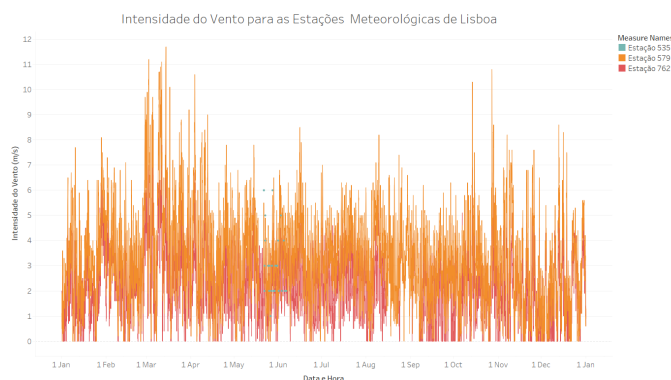


Figura C.26: Distribuição Temporal da Intensidade do Vento para todas as estações no ano de 2018.

C.2.6 Vento (Direção)

Existem 8760 registos para cada estação relativos à direção do vento em graus. À semelhança dos conjuntos de dados analisados anteriormente, também se verificou a existência de dois registos duplicados no dia 25 de março de 2018 pelas 2 da manhã, cujos valores da direção do vento podem ser consultados na Tabela C.14.

Timestamp	Estação 535	Estação 579	Estação 762
25/03/2018 02:00:00	-990	240	-990
		257	

Tabela C.14: Valores dos registos duplicados presentes nos dados relativos à direção do vento.

Para além disso, verificou-se que 100% dos valores da direção do vento relativos à Estação 762 e 96% dos valores na Estação 535 são -990, podendo significar que houveram problemas na leitura dos valores na Estação 535 e que não se efetuou este tipo de leitura na Estação 762. Na estação 579, apenas 0.84% (74 registos) dos registos correspondem ao valor -990. Esta informação pode ser consultada na Tabela C.15.

Estação 535	Estação 579	Estação 762
8378	74	8760

Tabela C.15: Contagem de registos com valor inválido (-990) para cada estação meteorológica relativamente à direção do vento.

Relativamente à distribuição temporal dos valores inválidos encontrados nos dados, no que diz respeito à estação 535 estes encontram-se distribuídos entre 1 de janeiro e 21 de maio e entre 6 de junho e 31 de dezembro, como se pode observar na figura C.27. Na

estação 579 os valores inválidos ocorreram entre 18 de maio às 10 horas e 21 de maio às 11 horas.

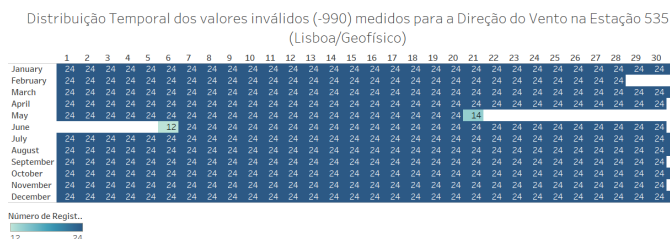


Figura C.27: Contagem dos valores inválidos para cada mês e dia na estação 535 ao longo do ano de 2018.

Devido ao referido anteriormente, apenas será analisada a distribuição da direção do vento para as Estações 535 e 579 após uma filtragem dos valores -990. A direção do vento encontra-se em graus, compreendida entre 0 e 360°, sendo possível observar a sua distribuição nas figuras C.28 e C.29. Uma vez que estes valores correspondem a pontos cardeais (Norte, Sul, Este e Oeste) e pontos colaterais (Nordeste, Sudoeste e Sudeste), foram elaboradas visualizações onde os valores da direção do vento em graus foram divididos em classes correspondentes aos mesmos (Figuras C.30 e C.31).

É notório que a distribuição da direção do vento não é semelhante na Estações em causa, o que poderá também ser devido ao facto de, no caso da Estação 535 haverem muito poucos registos.

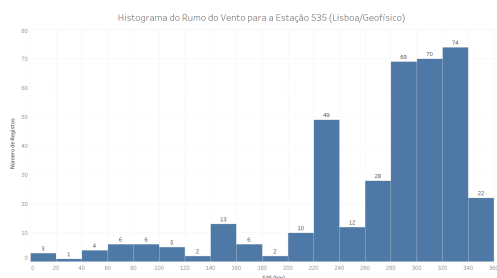


Figura C.28: Distribuição da Direção do Vento para a Estação 535 em 2018.

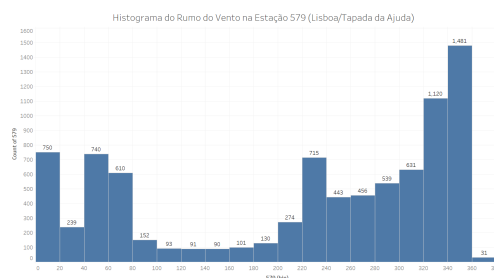


Figura C.29: Distribuição da Direção do Vento para a Estação 579 em 2018.

Observando a figura C.32, verifica-se que os registos relativos à Estação 535 foram entre 21 de maio e 6 de junho. Relativamente à Estação 579, os registos estão compreendidos entre 1 de janeiro e 31 de dezembro de 2018, sendo que as leituras em falta foram entre 18 e 21 de maio pelas 10 e 11 da manhã, respetivamente.

C.3 Qualidade dos Dados

Após a descrição e exploração dos dados meteorológicos verificou-se a existência de alguns problemas relativamente à sua qualidade no que diz respeito a valores inválidos, nulos e duplicados.

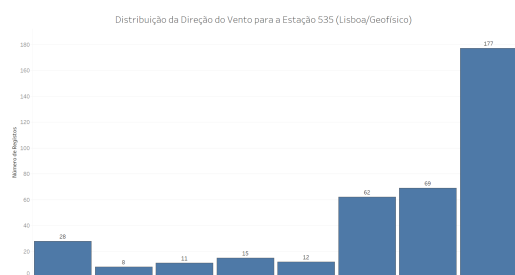


Figura C.30: Distribuição da Direção do Vento de acordo com os pontos cardeais e colaterais para a Estação 535 em 2018.

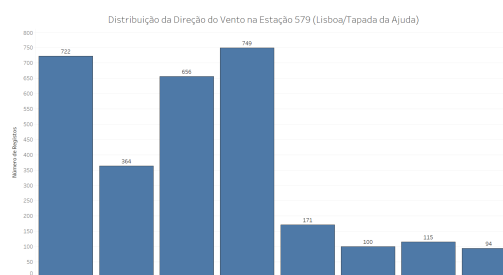


Figura C.31: Distribuição da Direção do Vento de acordo com os pontos cardeais e colaterais para a Estação 579 em 2018.

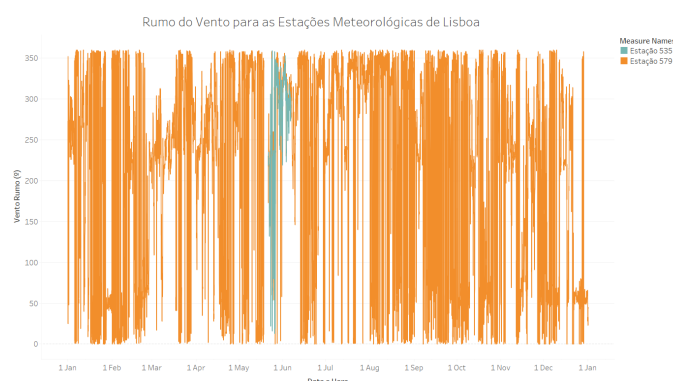


Figura C.32: Distribuição Temporal da Direção do Vento para as Estações Meteorológicas de Lisboa no ano de 2018.

Em todos o conjuntos de dados meteorológicos foi verificada a existência de valores duplicados no dia 25 de março pelas 2 da manhã. Esta linha duplicada em todos os ficheiros deve-se à mudança de hora que, em 2018, ocorreu dia 25 de março à 1 da manhã, sendo que nessa altura o fuso horário adiantou uma hora, o que poderá ter originado este erro nos dados. Verificou-se, no entanto, que não ocorreu o mesmo no dia 28 de outubro quando o fuso horário mudou novamente. Estes registos duplicados têm como consequência a inexistência de um registo referente a 1 hora da manhã do mesmo dia, razão pela qual a hora da primeira linha relativa ao duplicado será modificada de 2 horas para 1 hora, eliminando simultaneamente o duplicado e colmatando a falta de registos para a 1 da manhã do dia 25 de março.

Foram observados valores inválidos em todos os conjuntos de dados, sendo que este valor varia de estação para estação e em cada conjunto de dados. No caso da precipitação verificou-se que haviam apenas 3 e 4 valores inválidos nas estações 535 e 579, respetivamente, e 515 valores inválidos para a estação 762. No conjunto de dados relativo à humidade houveram mais valores inválidos para a estação 762 (451 valores inválidos), do que no caso das estações 535 e 579 (2 para ambas). Para a radiação, a estação 535 apenas continha valores inválidos, a estação 579 continha 6 valores inválidos e a estação 762 continha 468 valores inválidos. No que diz respeito à temperatura, houveram poucos

valores inválidos: 2 nas estações 535 e 579 e 451 valores inválidos para a estação 762. Nos dados relativos à intensidade do vento a estação 535 apresentou quase a totalidade dos valores inválidos ou nulos (mais concretamente 8378 valores inválidos e 343 valores nulos dos 8760 totais presentes no conjunto de dados). O conjunto de dados relativos à direção do vento foi o que apresentou mais valores inválidos, sendo que não haviam dados válidos relativos à estação 762, existiam 8378 valores inválidos para a estação 535 e 74 valores inválidos para a estação 579.

Devido à quantidade de valores nulos e inválidos, bem como ao facto dos dados não apresentarem muitas diferenças entre as estações, o conjunto de dados final irá conter, para cada atributo, a média dos valores de todas as estações, de forma a que não hajam valores nulos ou inválidos no conjunto de dados final. Tal será possível, uma vez que em nenhum dos conjuntos de dados existem valores inválidos ou nulos em simultâneo nas três estações.

C.4 Preparação dos Dados

Nesta fase de preparação dos dados irá ser levada a cabo a seleção e limpeza dos dados, tendo por base a análise efetuada nas fases anteriores de descrição, exploração e análise da qualidade dos dados.

Os dados selecionados correspondem ao período completo relativo ao ano de 2018, visto que é esse o período de interesse para a análise que se pretende levar a cabo nesta dissertação.

Foi tomada a decisão de, para cada conjunto de dados, efetuar a média dos valores medidos para as três estações (ou duas, nos casos em que uma das estações não apresenta valores válidos) devido à existência de valores inválidos em todas as estações. Desta forma, evita-se a utilização de técnicas de interpolação para preencher os dados e, para além disso, como já foi referido, no geral, os valores são semelhantes nas três estações, o que significa que não existe interesse em manter dados referentes às mesmas.

Uma vez que os dados meteorológicos vão ser utilizados como dados auxiliares aos dados das estações Gira, poderia ser interessante escolher a estação meteorológica da qual se utiliza os dados de acordo com a estação Gira a analisar (de acordo com a proximidade entre estas), no entanto, como os dados são semelhantes entre as estações, tal não apresenta nenhuma vantagem, sendo melhor criar apenas uma coluna que represente cada atributo computando a média dos valores das estações existentes. Para além disso, a área de Lisboa é relativamente pequena, o que também pode ser uma razão para a semelhança entre os valores das estações meteorológicas.

Desta forma, o conjunto de dados final irá conter uma coluna representativa da data e hora (DD/MM/YYYY HH:00) e mais seis colunas, sendo que cada uma será representativa da média dos seguintes atributos: Humidade do Ar, Precipitação, Radiação Solar, Temperatura, Rumo do Vento e Intensidade do Vento. Os atributos referidos podem ser visualizados na tabela [C.16](#).

Atributo	Tipo de Dados	Descrição	Unidades
Timestamp	Data/Hora	Data e hora do registo	DD/MM/YYYY HH:mm
Humidade	Numérico	Média da Humidade do Ar	%
Precipitação	Numérico	Média da Precipitação Total	mm
Radiação Solar	Numérico	Média da Radiação Solar	W/m ²
Temperatura	Numérico	Média da Temperatura	°C
Rumo do Vento	Numérico	Média do Rumo do Vento	°
Intensidade do Vento	Numérico	Média da Intensidade do Vento	m/s

Tabela C.16: Tipo de dados, descrição e unidades dos atributos que irão constar do conjunto de dados final.

Para obter o conjunto de dados final, eliminou-se primeiramente os duplicados encontrados, tendo-se modificado um dos registos para a hora anterior, que estava simultaneamente em falta em todos os conjuntos de dados. De seguida, foi calculada, para cada conjunto de dados, a média tendo em conta os dados das três estações, após a substituição de todos os valores inválidos (-990) por nulos. Desta forma, a média que foi obtida teve em conta apenas os valores não nulos. Foi criada uma nova coluna respetiva à média para cada conjunto de dados e, de seguida, foi efetuada a junção das mesmas, de forma a ter o conjunto de dados final, que contém uma coluna relativa a cada atributo e uma relativa ao *timestamp*, como foi indicado na tabela C.16.

Este procedimento foi levado a cabo recorrendo à linguagem *Python* e a bibliotecas como *numpy*, *pandas* e *datetime*, uma vez que estas são muito utilizadas para processamento de dados e devido à facilidade de utilização.