

MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Understanding Urban Tourists' Spatial Behaviour Using Roaming Data in Lisbon

Sofia Soares Vieira

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

UNDERSTANDING URBAN TOURISTS' SPATIAL BEHAVIOUR USING ROAMING DATA IN LISBON

Sofia Soares Vieira

Thesis presented as partial requirement for obtaining a master's degree in data science and
advanced Analytics, with a specialisation in Business Analytics

Supervised by

Supervisor: Miguel de Castro Neto, PhD, NOVA IMS – Universidade Nova de Lisboa

Co-Supervisor: Nuno Alpalhão, MSc, NOVA IMS – Universidade Nova de Lisboa

July 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Sofia Soares Vieira

Lisbon, 15 July 2024

ACKNOWLEDGEMENTS

Completing this thesis has been a challenging yet very rewarding journey, and it would not have been possible without the support and guidance of many individuals and institutions. I am deeply grateful to everyone who contributed to this work, and I would like to take this opportunity to express my sincere appreciation.

I express my deepest gratitude to my academic advisors, Professor Nuno Alpalhão, MSc, and Professor Miguel Neto, PhD, and Dean of NOVA Information Management School (IMS), for their invaluable guidance, support, and encouragement throughout this research journey. I am also profoundly thankful to the NOVA Innovation and Analytics Lab for providing the Vodafone dataset and to Vodafone for generously sharing their data with the University of NOVA IMS, which was essential for the completion of this study. I am very thankful for my colleagues and fellow students, particularly the student class T63 arranged by Professor Nuno, with monthly meetings and classes that provided guidance and examined the progress of our thesis. The peer reviews done by other colleagues were instrumental in refining and improving my work.

My sincere gratitude goes to my mother Carmen Soares, my father Carlos Vieira, my brother Pedro Vieira and my beloved friend David Cordeiro for their persistent support, patience, and inspiration. I also extend my appreciation to all my friends who have been a source of strength and motivation. Finally, I would like to acknowledge NOVA IMS for providing an excellent academic environment and the necessary resources that facilitated my research. Thank you all for your significant contributions to this thesis.

ABSTRACT

In the dynamic urban landscapes of today, understanding tourist movements is essential for effective city management and urban planning. Tourists generate substantial data when travelling which can be leveraged to make informed decisions by the city stakeholders and policymakers. This thesis explores the complex patterns of tourist behaviour in Lisbon, utilising mobile phone roaming data provided by Vodafone to offer a granular analysis of tourist mobility. By characterising diverse tourist groups, this study provides valuable insights that extend beyond academic interest, offering practical relevance for the Lisbon Municipality and business owners. The data-driven approach adopted aims to improve urban infrastructure, effective resource allocation, and the overall tourist experience, promoting Lisbon's evolution as a responsive and vibrant city in the face of growing visitor dynamics. The research uses the CRISP-DM framework and unsupervised learning techniques, particularly density-based clustering for its ability to identify natural groupings in the data without prior labels. The analysis revealed three distinct tourist clusters, each with unique behaviours and preferences, enabling targeted strategies to better meet the visitors' needs. These clusters differed significantly in their mobility patterns, preferred destinations, and duration of stay. Advanced techniques like those involving the K-Means algorithm and feature engineering ensured a robust understanding of tourist movements. The findings of this thesis significantly impact sustainable tourism development in Lisbon. Ultimately, this study demonstrates the transformative impact of advanced data analytics in developing resilient and visitor-friendly urban environments.

KEYWORDS

Data-Driven; Roaming Data; Clustering; K-means; Tourism.

Sustainable Development Goals (SDG):



INDEX

1. Introduction.....	9
2. Literature review	12
2.1. Urban Tourist Mobility	12
2.2. Data Mining Approaches	14
2.3. Case of Lisbon.....	17
2.4. Research Gap.....	18
3. Methodology	20
3.1. Data Understanding	20
3.2. Data Preparation	26
3.3. Modelling.....	31
4. Results and discussion	33
4.1. Clustering Results	33
4.2. Managerial Insights	37
5. Conclusions and Future Works.....	40
5.1. Conclusions.....	40
5.2. Limitations and Future Work.....	41
Annexes	47
Annexe 1 - Vodafone Data Framework	47
Annex 2 - Information provided by CML	47
Annex 3 - Vodafone Data Integration Method.....	48
Annex 4 - Vodafone Dataset Indicators.....	49

LIST OF FIGURES

Figure 1. Bar Plot for categorical Figures for Fall	23
Figure 2. Histogram for variables C2 and E9	24
Figure 3. Box plot for variables C2 and E9	25
Figure 4. Spearman Correlation Matrix for Numerical Variables	26
Figure 5. Elbow method and Variance explained Ratio, respectively.....	30
Figure 6. Inertia Plot and Average Silhouette Score, respectively	32
Figure 7. Clusters Profile	33
Figure 8. Spatial Distribution of Clusters in Lisbon	35
Figure 9. Weekdays with higher DTR per cluster	35

LIST OF TABLES

Table 1. Data Mining Algorithms by Author, featuring Advantages and Limitations	16
Table 2. Key Variables Description I	21
Table 3. Key Variables Description II	22
Table 4. Renamed columns of the dataset	28
Table 5. Final selected Variables to use in K-means	31
Table 6. Scores considered to select ideal number of clusters.....	32
Table 7. Cluster Characterisation	36
Table 8. Vodafone Data Framework	47
Table 9. Information provided by CML	47
Table 10. Vodafone Data Integration Method	48
Table 11. Vodafone Dataset Indicators	49

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
CDR	Cellular Data Records
CML	Câmara Municipal de Lisboa
CRISP-DM	Cross-Industry Standard Process for Data Mining
EU	European Union
GDPR	General Data Protection Regulation
GIS	Geographic Information System
HHMD	Heterogeneous Human Mobility Data
HTTPS	Hypertext Transfer Protocol Secure
IQR	Interquartile Range
OHC	One-hot-encoding
PCA	Principal Component Analysis
PMD	Passive Mobile Data
POI	Points of Interest
SNA	Social Network Analysis
SPADE	Sequential Pattern Discovery using Equivalence Classes
UMAP	Uniform Manifold Approximation and Projection

1. INTRODUCTION

In the current urban landscapes, the study of tourist movements stands as a key tool for interpreting the dynamic patterns of tourism activity within a city. This thesis embarks on a comprehensive exploration, that goes above conventional methodologies by leveraging data to reveal the elaborate nature of tourist behaviour. Allowing a granular understanding of tourists' mobility patterns, mapping their trajectories, and studying the underlying drivers shaping their movements across cities. One of the main goals of this pursuit is the characterisation of the diverse tourist groups, a meticulous effort aimed at constructing a broad profile of their preferences, interests, and exploration habits within the city's boundaries. However, this thesis focus extends beyond academic interest, it holds significant relevance in practical city management and urban governance. The insights gathered from the analysis of tourist mobility serve as precious knowledge, guiding city stakeholders and policymakers towards informed decision-making. The quality and relevance of the data used in this project serve as a foundation in urban planning, resource allocation, and the design of experiences adapted to enrich tourist involvement and improve city infrastructure. The synergy between understanding tourist mobility, characterising their presence, and applying this knowledge to city development presents an unprecedented opportunity for cities to evolve, adapt, and prosper in the face of visitor dynamics and urban evolution.

Understanding the spatial behaviour of urban tourists can be essential for effective city management. Lisbon, as a major tourist destination, can benefit significantly from comprehending and adapting to the patterns of tourist behaviour. This exploration gains further depth by characterising the diverse tourist types in Lisbon, considering their variations across different seasons and days of the week, and allowing the development of public policies that go towards these visitors' dynamic needs and patterns. By recognising the movement and conduct of tourists within the city landscape (Balzotti et al., 2018), Lisbon can polish its strategies, elevate service quality, and provide more gratifying experiences for its guests (Chu & Chou, 2021). Embracing these insights is a realistic chance for Lisbon to evolve and adapt, thereby shaping a more vibrant and responsive urban landscape. For all of this to be accomplished, it is necessary to identify the diverse range of tourists, and understand their unique characteristics, behaviours, and shifting trends when exploring Lisbon. By analysing visitors' demographics, preferences, and habits, whether they are culture seekers, leisure travellers, or urban adventurers, it is possible to gather instrumental insights into their interests, movement patterns, and chosen attractions.

Although tourists generate substantial data during their visits, there is a lack of scientific research addressing their movement patterns and activities within cities (Xu et al., 2021). Besides, when using mobile data, crucial gaps are impacting the validity of findings such as the reliability and completeness of the data sources and the potential inaccuracy when using data mining techniques. Likewise, historical behaviour analysis can be considered a relevant limitation since the dynamic nature of tourists' preferences and behaviours can be overlooked. The above-mentioned reduces the comprehensive understanding of tourist behaviour and delays the development of targeted strategies for urban planning and tourism management. Using a rich source such as roaming data and acknowledging the gaps mentioned can help mitigate those constraints for Lisbon.

Having said that, leveraging mobile phone roaming data appears to be a powerful tool in this pursuit. It can provide a broad understanding of visitor movements, popular locations, the duration of their stays and activity patterns throughout their visit. Through this data-driven approach, local businesses, tourism entities, and city planners can strategically optimise services, such as by customising tour packages (cultural, gastronomic, or architectural) or developing smart infrastructures such as Wi-Fi spots, rest areas, and shuttles to high-traffic touristic zones (Buhalis & Amaranggana, 2013). This approach also enables efficient resource allocation by identifying frequently visited areas where it becomes possible to employ maintenance efforts or extra staff. Moreover, it facilitates the development of personalised experiences that align with the diverse preferences of tourists, like providing recommendations for local events, dining options or even lesser-known attractions. Ultimately this enriches the overall visitor experience in Lisbon while corresponding with the city's diverse tourism landscape.

As the research progresses, there exists the possibility of finding insights that may trigger changes in the planned decisions. These revelations commonly emerge throughout the stages of data exploration and preparation. Despite that, the objectives of this study include a multi-fold approach: initially the analysis of tourists' mobility patterns through feature engineering. By introducing new variables, it is possible to elevate the analysis and facilitate data visualizations which allow the exploration of a different perspective of the data. Subsequently, it aims to identify distinct tourists' clusters in Lisbon by leveraging the most appropriate data mining algorithm or potentially a combination thereof, as clusters stand as a widely adopted method for analysing high-density areas, such as the one under consideration. In the end, the main objective remains to characterise Lisbon's tourists according to their spatial occupation, understanding their activities and helping the municipality to make more informed decisions.

The methodological approach of this project will be the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a widely adopted approach known for its structured and periodic nature in data analysis and mining. In particular, this project utilises unsupervised learning techniques, specifically focusing on data mining via density-based clustering, similar to previous applications alike Park et al., (2020). This technique offers distinct advantages when analysing tourism movements and behaviour, it stands out in detecting non-linear and complex patterns within the data, crucial when understanding the diverse and varied patterns of tourist mobility. Unlike methods assuming uniform distribution, density-based approaches are robust in high-density areas, being particularly useful and effective in analysing destinations like Lisbon. Clustering methods effectively identify specific areas of high tourist activity, providing insights into preferred locations, attractions, or less explored regions. Moreover, the flexibility in identifying clusters of various shapes and sizes is vital in analysing visitors' data with unpredictable routes. Furthermore, these methods detect outliers or unusual behaviours, helping in recognising atypical tourist activities or events. Overall, density-based clusters offer a flexible, robust, and insightful approach, enabling a comprehensive understanding of tourist behaviour and movement in Lisbon.

The roaming data, provided by Vodafone, will be used to construct a rich and representative dataset that accurately reflects a typical tourist day in Lisbon. This collaboration holds significant importance because Vodafone offers an anonymised and aggregated dataset capturing mobile terminal movements, such as mobile phones, between grid squares within a 200x200 grid at five-minute intervals. This dataset enables a detailed analysis of aggregated flow within Lisbon's urban landscape, providing essential insights into movement patterns. Vodafone is an ideal partner given its capability to provide precise and detailed insights into the movements, entry points, and departures across Lisbon. This is crucial for understanding the detailed aspects of how people move around busy Lisbon. By using unsupervised methods such as K-means on this dataset, it will be possible, for example, to determine the various types of tourists visiting Lisbon and identify the most frequented routes. The technical foundation of this exploration is Python, a powerful programming language that enables effective data preprocessing, feature engineering, and strong model development. This approach aims to identify significant clusters within the dataset that will serve as critical benchmarks for understanding and extracting valuable insights relating to the spatial behaviour of urban tourism in Lisbon.

The results are expected to contribute to the significant improvement of urban planning and tourism management in Lisbon. By characterising the various types of tourists in the city, the research seeks to uncover essential insights into Lisbon's dynamics as a poly-centric destination, contributing to the development of a smarter and more sustainable urban environment (Park et al., 2020b). Through the utilisation of clusters, the study enables the identification of recurring behaviours and preferences among visitors, exposing common routes, popular attractions, and concentrated areas of tourist activity. This segmentation of tourists based on movement patterns promotes an understanding of diverse visitor types, facilitating tailored strategies for cultural, nightlife, or historical attractions. The application of clusters further extends to efficient resource allocation, guiding local authorities, businesses, and service providers to focus efforts on frequently visited areas, thereby improving infrastructure and services for a better-quality overall visitor experience. Moreover, clusters prove valuable in anticipating and managing events, strategically promoting local businesses, and supporting urban planning for necessary infrastructure development. Ultimately, the insights from this cluster analysis provide a structured framework for stakeholders to empower strategic decision-making and optimise resources, culminating in an enriched and sustainable tourist experience in Lisbon.

2. LITERATURE REVIEW

2.1. URBAN TOURIST MOBILITY

Understanding urban tourist behaviour is a multifaceted attempt that involves analysing their spatial and temporal trajectories, along with the dynamics of their interactions with the urban environment and its attractions. It involves a full exploration of how individuals navigate, engage with, and experience diverse urban landscapes. The study of these behavioural patterns holds vast practical value. It offers insights that are revealed to be crucial for city planning, optimising attractions, and tailoring experiences that meet the desires of urban travellers. Overall, investigating urban tourist behaviour is a fundamental step towards designing cities in a way that appeals to a wide range of people visiting from various backgrounds, cultures, and interests.

Diverse methodologies and data sources, such as geotagged photos, user-generated content, and Social Network Analysis (SNA), have been instrumental in recognising the complexities of tourist mobility. An influential study by Domènech et al. (2020) stands out in this field by employing geotagged Flickr photos to reconstruct individual visitor paths in cultural cities like Toledo, Spain. By analysing these trajectories, the research enlightens the profound influence of urban environments and their usage on visitor spatial behaviour. Their advanced Geographic Information System (GIS) techniques revealed crucial insights. Distinct visitor behaviour patterns were found by identifying heavily frequented tourist locations while uncovering overlooked areas. Importantly, their findings highlighted the influence of urban environments and multifunctional zones (commercial and leisure areas) on visitor behaviour. This approach also addresses the challenges faced by destinations heavily reliant on tourism, dealing with overcrowding and the need to protect heritage assets.

Similarly, a study focused on urban tourism by Bouchon (2022) enriches the present understanding by highlighting its aim to attract visitors through leisure and business activities. It not only emphasises the economic contribution of tourism to the urban economy but also highlights the importance of diverse attractions—cultural, architectural, technological, leisure, and social—that contribute to a city's appeal. Essentially, this study analyses the varied interactions between tourism activities and cities as multifaceted entities, considering differences in population, size, functions, and geographical location. It considers how metropolitan cities attract the largest visitor numbers but might not be visibly impactful economically or physically. In contrast, smaller cities, with specialised offerings like historic or gambling attractions, may showcase more visible tourism activities.

A powerful study conducted by Van der Zee & Bertocchi (2018) substantially advances the current understanding of destination management by leveraging data-driven insights from social media. This research utilises a relational approach and SNA on user-generated content from TripAdvisor, specifically focusing on Antwerp, Belgium. The study clears up distinct review patterns among locals, Belgians, Europeans, and non-Europeans while revealing repeated behavioural trends. By comparing these review networks to the spatial distribution of attractions, hotels, and restaurants, the research features the value of SNA in converting big data into actionable knowledge. Particularly, it identifies a concentration of clustered attractions in Antwerp's historic centre, highlighting the relationship between geographical and relational proximity for international reviewers but less so for domestic ones. These findings lead to recommendations for representatives and destination managers, aiming to optimise tourist distribution across the destination.

Following the SNA application, another study conducted by Chu & Chou (2021), employs network analysis alongside Cellular Data Records (CDR) to establish a trajectory network in Taiwan. The main target of this research is to uncover the functionalities of tourism destinations by discerning the first location tourists opt to visit. The study addresses the challenge of tracking multi-destination tourists by leveraging mobile phone positioning data. It employs three centrality measures - Degree, Reach Closeness, and Betweenness - along with role analyses to interpret the network relations among 78 nodes (origins and destinations) within the tourism network. The findings present diverse functions of these nodes, some with multiple roles and others singular, which can vary based on weekdays versus weekends. Furthermore, the study characterises the tourism region into different district sizes, offering valuable insights for tourism planners to strategise and optimise tourism marketing efforts by anticipating tourist movements.

Positioning data holds significant potential for exploring space-time behavioural patterns, particularly in the context of studying tourism geography. A study in Estonia utilises Passive Mobile Data (PMD) alongside various other datasets including border records, transportation/accommodation statistics, questionnaires, and modelling exercises to explore tourism trends (Ahas et al., 2008). These methods have demonstrated their value in examining the common routes taken by tourists of specific nationalities. The study achieves correlations of up to 0.99 between positioning data and traditional accommodation statistics in frequently visited tourist regions, meaning that the information gathered from tracking tourists' movements via mobile devices aligns very closely with where tourists stay. However, regions with more tourists and limited tourism infrastructure showed lower correlations. The advantages of positioning data lie in its ability to collect information on larger spatial units as well as in less-visited areas, offering more detailed and accurate information about the location and the timing of where people are, surpassing the level of detail provided by regular tourism statistics.

In the dynamic landscape of tourism research, a profound shift towards innovative methodologies is shaping the understanding of visitor behaviours, especially within urban tourism. Reif & Schmücker (2020) embark on an ambitious journey, employing probabilistic and classification approaches to explore the spatial behaviours of tourists. The focal points encompass the identification of visitor segments, seasonal variations, and meticulous delineation of inter- and intra-destination movement patterns. By consolidating insights gathered from a range of research across different places in Germany, this study builds a solid conceptual framework that seeks to meticulously explain the specific strengths and limitations in the utilisation of PMD for tourism research. Leveraging datasets from two major mobile network operators in Germany, this research ties the gap in understanding tourists' spatial behaviours, offering priceless insights to advance the efficacy of utilising PMD for optimising tourism experiences and urban planning strategies.

Research by Xu et al. (2021) employs a multifaceted methodology centred on the analysis of mobile phone data. This study analyses the dimensions of tourist movements by introducing different mobility indicators, such as duration of stay in a city, number of trips taken, and locations visited obtained from cell phone data. Utilising Principal Component Analysis (PCA), it aims to reveal the multidimensional nature of tourist mobility patterns and how different groups of tourists behave within cities. The dataset, sourced from a major cellular operator in South Korea, provides a rich foundation of individual phone users' habits, capturing their spatial trends and durations of stay. This exploration through mobile phone data offers a unique method to scrutinise the layers of tourist

mobility within urban environments, enabling insights into the development of location-based recommendations and tourism services, particularly in transportation.

These studies collectively reveal that tourist behaviour can be influenced by several factors, such as the availability and location of tourist attractions, diverse land uses, and the natural characteristics of the city itself. For example, the concentration of cultural, architectural, and leisure attractions significantly contributes to a city's appeal to tourists. The analysis of tourist behaviour yields valuable insights into the distribution of tourists within a destination, thereby helping in the establishment of effective strategies for managing tourist flows. Overall, tourists can be considered as mere wanderers of the urban environment (Wearing & Foley, 2017).

2.2. DATA MINING APPROACHES

Data mining methodologies, particularly those leveraging mobile data, have emerged as crucial tools in comprehending and optimising tourist behaviours. The study conducted by Not (2021) analyses tourist behaviour by mining usage data gathered from mobile applications. This approach seeks to expose trends and variations in both time and space among different groups of attendees at large events. The findings reveal distinct classes of users with different travel planning behaviours, ranging from individuals seeking practical information well in advance to those seeking the information at the last minute or upon arrival at the event site. These outcomes offer valuable insights and guidelines for designers, developers of information systems, event organisers, and tourism stakeholders. It also demonstrates the significant contribution that user behaviour analysis offers to tourism management and marketing, enabling a deeper understanding of consumer behaviour and preferences that transcend standard analytics.

Moreover, Sheikhhassan et al. (2021) employ clustering algorithms and communication rules to discern patterns of tourist mobility. This not only aims to identify specific behavioural trends but also facilitates the development of adapted strategies to meet the diverse needs of tourists. The study highlights the importance of a targeted and adaptive approach in the tourism industry. Extending the application of data mining, Sharma et al. (2021) contributes to the field by utilising data mining algorithms and applications to discover data patterns. The Weka software, which is a collection of machine learning algorithms for data mining tasks, is used by this author to enable more informed and efficient decision-making within the tourism sector. The research classifies the algorithms and applications and provides the best combination to be employed in the tourism industry under the provided circumstances. This work emphasises the transformative potential of data-driven insights in optimising operational and strategic features of the industry.

The study by Sun et al. (2021) explores the potential of mobile phone data to distinguish tourists from locals in a tourism-focused city. Highlighting the following advantages of mobile phone data: its widespread coverage, temporal consistency, and cost-effectiveness compared to other sources like social media and GPS. The research presents a framework involving the extraction of hidden behaviour from mobile phone data, followed by the application of K-means clustering to distinguish between these two groups. Through an experimental study in Xiamen, China, the method successfully identifies popular tourist spots and key transportation routes. Extracting and analysing specific features over time and space in cities helps in understanding urban tourism, learning about tourist behaviour and managing city spaces closely.

When examining tourism destinations' spatial setups, a study leverages real-time data from travellers to cluster their behaviours and reveal how tourism activities are structured in those destinations. This research integrates four methodologies to improve the accuracy of predicting travel movement patterns. The initial phase involves a spatial descriptive analysis that aims to find the traveller profiles and map out the initial distribution of their movements within the destination. Moving forward, a point-density analysis examines stay points within mobile roaming data. After this examination the DBSCAN clustering method strategically groups together nearby stay points, and finally, the Sequential Pattern Discovery using the Equivalence classes (SPADE) algorithm comes into play to optimise the efficiency of mining sequential patterns by reconfiguring the original search space. The research discovers various popular zones within travel destinations and examines how these places interact spatially. Consequently, this study offers significant insights into tourism by merging spatial models with destination planning, forming the basis of tourism design (Park et al., 2020).

In a broader context, the infusion of big data technology into the tourism industry is explored by Keertika et al. (2021). By utilising social media platforms, such as Flickr or Panoramio websites, this study demonstrates how companies can utilise big data to gather information. It also shows the capability of this technology to draw insightful conclusions about tourist destinations and implement necessary measures. This highlights the potential of big data technology in providing a real-time understanding of the evolving landscape of tourist preferences and trends.

The study by Fu et al. (2019) focuses on developing a collective learning model modified for Heterogeneous Human Mobility Data (HHMD), combining diverse sources like cell phone traces, vehicle GPS trajectories, and geo-tagged posts. Effectively modelling HHMD holds vast potential for applications in city planning, transportation, urban economics, and public safety, yet it remains a considerable challenge due to its diverse origins. The proposed approach determines and quantifies urban forms within residential communities. The model captures correlations among residential areas, urban functions, temporal effects, and user mobility patterns. Incorporating LASSO analysis and significant testing, the approach uses Points of Interest (POIs) as metadata to identify urban forms and functions of residential areas. The study explores practical applications, including real estate ranking and restaurant popularity prediction, leveraging the learned insights. The efficacy of the proposed model is verified through extensive evaluations with real-world data, demonstrating its effectiveness in modelling HHMD and its successful applications to address diverse urban challenges.

Collectively, these studies feature the transformative potential of data mining applications and big data technology in revealing the difficulties in understanding tourist behaviours. Whether applied to specific domains or adopted on a broader scale within the tourism industry, these approaches pave the way for more informed decision-making and tailored strategies that align with the dynamic needs of diverse tourist segments.

To summarise all data mining approaches and methods discussed above, explore the table below with the methods, correspondent author and some advantages and limitations:

Table 1. Data Mining Algorithms by Author, featuring Advantages and Limitations

Data Mining Algorithms	Author	Advantages	Limitations
K-means Clustering	Not (2021)	Exposes trends in tourist behaviours. Offers insights for event organisers and tourism stakeholders.	Relies solely on mobile app usage data.
	Sun et al. (2021)	Provides cost-effective insights into tourist behaviour. Identifies popular spots and transportation routes.	Privacy concerns. Accuracy can vary based on signal strength and device usage.
DBSCAN	Park et al. (2020)	Improves accuracy in predicting travel patterns. Discovers popular zones within destinations. Provides insights into tourism and destination planning.	Requires integration of heterogeneous data sources.
	Sharma et al. (2021)	Enables informed decision-making in tourism. Classifies algorithms for optimal use. Emphasises data-driven insights for operational and strategic improvements.	Requires expertise in data mining and algorithm selection. Implementation challenges with data quality and availability.
SPADE	Park et al. (2020)	Optimises the efficiency of mining sequential patterns by reconfiguring the original search space. Discovers various popular zones within travel destinations. Examines how these places interact spatially.	Requires integration of heterogeneous data sources.
LASSO analysis	Fu et al. (2019)	Demonstrates effectiveness in modelling Heterogeneous Human Mobility Data (HHMD) and addressing diverse urban challenges.	May require extensive computational resources and expertise. Challenges in integrating diverse data sources.

2.3. CASE OF LISBON

Lisbon, being a major tourist destination, can benefit significantly from understanding and adapting to the needs and patterns of tourists. By determining how they move and behave (Balzotti et al., 2018), the city can optimise its strategies, improve services, and create a more satisfying and personalised experience for visitors (Chu & Chou, 2021).

Urban tourism in Lisbon has witnessed remarkable growth in recent years, solidifying its status as a desirable destination for visitors. A particular research conducts an extensive examination of tourist behaviour and habits in Lisbon (Francisco et al., 2023). This comprehensive analysis offers valuable insights into the most frequently areas visited by tourists, their dining choices, and preferred accommodations, and identifies the countries or geographic regions from which tourists visiting Lisbon originate. Beyond the anticipated findings related to the most frequented places and parishes, the research explored unexpected hints, uncovering the influence of specific events such as the Web Summit or football games on visitor behaviour and movements. These revelations broaden the comprehension of the dynamic and sometimes surprising factors that drive tourist movements within the city. By acknowledging the unanticipated elements, this study adds a layer to the complexity of understanding urban tourism dynamics.

Lisbon experienced substantial urban tourism development following its recognition as the top 'city break' destination in 2017 (Castela, 2018). A key factor in shaping tourism is the perception of residents, which subsequently influences their interaction with tourists. There is one investigation that focused on assessing the level of interaction between tourists and residents in the Alfama neighbourhood. Castela's study employed qualitative methods such as direct observation and resident interviews. The results revealed multifaceted social representations and generally positive perceptions of tourism's economic, socio-cultural, and environmental contributions. These findings propose strategies for city authorities to satisfy tourist demands while maintaining the city's distinctiveness and authenticity.

The study by Mendes (2020) explores the concept of tourism-driven gentrification within Lisbon's historically significant neighbourhoods, revealing the impacts of increased tourist activity on the city's housing market. "Tourism gentrification" here refers to the changing landscape reinforced by the traveller's interest, resulting in notable transformations in the housing dynamics of these valued areas. With the rise in tourism, these neighbourhoods witness a significant shift in property values, housing availability, and the character of the community. Accommodations, once primarily residential, are transformed to offer short-term rentals or tourist lodgings, impacting the local housing stock. Mendes' research navigates through this scenery, exposing the interaction between tourism-driven transformations and their direct implications on the housing accessibility and dynamics within Lisbon's historically rich quarters.

Understanding tourist movements not only benefits city management but also accelerates the improvement of highly appealing locations. By leveraging social media, a study connects geo-tagged photographs from Flickr to pinpoint tourists' POIs in Lisbon, Porto, and Faro while quantifying their relational aspects through user attendance at these locations (Pereira, 2023). The results demonstrate the efficacy of standard clustering techniques in identifying potential POIs. The association of these POIs via users' social media activities, particularly through photo postings, forms an elaborate network that transcends geographical proximity. Across all studied cities, the findings reveal the capturing of

historical places like churches and cathedrals, viewpoints, and beaches, indicating their importance in tourist engagement and interest.

In a more comprehensive view, an insightful exploration considers the influence of tourism on the revitalisation of central areas in Lisbon, examining its effects on residents' daily lives and the urban economy (Barata-Salgueiro et al., 2018). This research serves as a comprehensive reflection on the diverse impacts of tourism, offering valuable insights into its multifaceted effects on the city's urban landscape.

In summary, the research on tourism in Lisbon holds vital importance for the city's growth as a major tourist destination. Understanding tourist behaviours, preferences, and unexpected influences allows for strategic optimisation and personalised experiences. This body of research offers a comprehensive understanding of the multifaceted impacts of tourism on Lisbon's urban landscape, residents, and economy, guiding effective city management and a smarter and more sustainable development (Buhalis, 2020).

2.4. RESEARCH GAP

The literature on data mining approaches in analysing tourist mobility behaviours using mobile data presents valuable insights, yet certain weaknesses and drawbacks are notable in the methodologies employed. An important concern is the heavy reliance on data mining techniques for analysis and prediction. While these algorithms play a crucial role, it's essential to acknowledge their natural limitations. Data mining algorithms may not always guarantee accuracy and can potentially generate false or misleading results (Not, 2021). The forthcoming thesis should explicitly recognise the uncertainties associated with these techniques to ensure a more refined interpretation of the findings.

Another significant limitation is the predominant focus on analysing tourists based on their historical behaviour. While past behaviour is certainly informative, it may not consistently predict future actions, given the dynamic nature of tourists' preferences and circumstances (Sharma et al., 2021). A more comprehensive understanding could be achieved by integrating analyses of evolving factors that influence tourist behaviour over time. Since understanding the influences guiding tourists' choices is crucial, a meticulous examination across the varied weekdays and distinct seasons will be performed in the upcoming study. When dealing with mobile phone trajectory data, it should be considered the potential bias of a sample that is largely associated with social media and online sources to access geotagged data (Park et al., 2020).

Furthermore, a reliance on quantitative methods, such as questionnaires and data analysis, is evident in these studies. While these methods provide valuable statistical insights, they may not capture the full complexity of tourist behaviour (Sheikh Hassan et al., 2021). Incorporating qualitative methods, such as interviews and observations, could offer a more profound understanding of the motivations and preferences that quantitative data might not fully capture. To address this, leveraging both quantitative and qualitative approaches becomes crucial. While quantitative data often can be volatile, the upcoming study will combine qualitative methodologies, and explore diverse sources such as travel blogs, reviews and websites detailing local tourist spaces in Lisbon. This research aims to bridge quantitative limitations by acknowledging the data nature and potential biases.

With the increasing presence of tourists in urban spaces, mobile phone fingerprints offer a smart means to comprehend the urban ecosystem. However, mining individual geolocations from mobile communication data can bring challenges due to diverse processes, potentially resulting in incomplete trajectories (Klimek, 2019). To overcome that, the chosen approach in the upcoming thesis must blend informal and formal elements, such as unstructured and structured data, so that the datasets can be complemented with context to reconstruct aggregated patterns within urban spaces.

An essential aspect often overlooked is the explicit discussion of the limitations associated with the data sources used, such as photo-sharing websites or mobile applications. Evaluating the reliability and representativeness of these sources is crucial for drawing valid conclusions about tourist behaviour. A comprehensive assessment of data source limitations would contribute to a more robust interpretation of the research findings and foster transparency in the methodologies employed. The coming study will be supported by rigorous inspection of data quality, combined with sensitivity analysis to understand the robustness of the findings and to identify which variables have the most significant impact on the overall outcomes.

Privacy issues and ethical concerns might emerge when employing data mining algorithms within the tourism sector, due to the examination of personal data and individual behavioural trends. Consequently, the General Data Protection Regulation (GDPR) is a framework that oversees these cases regarding the acquisition and handling of personal information belonging to individuals within the European Union (EU). Specifically concerning the gathering of tourist mobile data, GDPR delegates organisations to secure explicit consent from individuals before gathering their personal information. These entities are also obligated to provide transparent and lucid details about the data's intended use and storage, while individuals retain rights to access, rectify, and delete their data. Moreover, organisations must implement security protocols to safeguard against unauthorised access or breaches. Non-compliance with GDPR may lead to substantial fines and penalties (X. Wei, 2021). The upcoming thesis will bridge this gap since the data is collected by an entity, Vodafone, which already follows all the specifications and obligations of the GDPR. In Annex 3, Table 10, find the Vodafone data integration method employed in this case.

Having said that, in this forthcoming project, the research aspires to offer substantial improvements to Lisbon's urban strategies and services by considering the multifaceted dynamics of tourist behaviour. This project aims to characterise tourist profiles around Lisbon, aiming to reveal insights into Lisbon's dynamics as a poly-centre destination. Park et al. (2020), emphasise in their study the goal to foster a smarter and more sustainable city model. By doing so, it will be possible to provide a personalised urban management approach and optimised resource allocation and services. At last, this will elevate the visitor experience, promoting sustainability and adaptability within Lisbon's urban environment.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

To successfully identify tourist segments in Lisbon and understand their characteristics, it is essential to explore and comprehend the dataset provided by Vodafone thoroughly. The data understanding phase helps in becoming acquainted with the dataset, the various variables, and values, while also highlighting any necessary adjustments for later alteration. This understanding is crucial for identifying relevant features that may characterise tourist behaviour and preferences. This type of exploration allows one to identify key variables, detect anomalies or missing data, expose patterns and trends in traveller tendencies, and validate assumptions about tourist first choices. Subsequently, there is potential to develop more accurate and insightful segmentation strategies that better capture the diverse characteristics of tourists visiting the city of Lisbon. Vodafone provided multiple files containing anonymised and pseudonymised data. This data represents mobile terminal movements of users in Lisbon from September 2021 to August 2022, find the data framework in Table 8 in Annex 1. In telecommunication contexts, terminals can refer to endpoints in a network where data is transmitted, received, or routed, such as cell towers or Wi-Fi access points. This type of data has already been used before to explore mobility patterns in cities (Park et al., 2020).

The database is a comprehensive resource designed to provide the Lisbon City Council (CML) with valuable insights into mobile terminal movements within the city. Sourced by a telecommunications company, the database captures detailed information about the interactions of mobile devices within Lisbon. This data is recorded at a high level of granularity, with movements tracked within a 200x200-meter grid. The database operates at a high temporal resolution, capturing data points at regular intervals of every 5 minutes. To safeguard privacy, the data is anonymised, ensuring that individual identifiers are removed. Despite its anonymised nature, the database retains its analytical value through aggregation, which summarises movement patterns without compromising privacy. Each week, the database generates approximately 7,500,000 data points, which means that the full dataset contains around 360 million rows, indicating a substantial volume of data that necessitates robust management and processing techniques. This wealth of data serves a critical purpose for this thesis, enabling the analysis of people flows within Lisbon. By leveraging telecommunications data, it is possible to gain insights into traffic patterns, popular areas, and mobility trends, which can lead to improvements in urban planning, transportation management, and other decision-making processes. In the end, the database plays a vital role in understanding and managing urban mobility within Lisbon, offering a valuable instrument for developing the city's liveability and efficiency.

Nonetheless, due to the limited computational capacity an approach was followed to represent in a smaller dataset as much data as possible within a year. The data was organised into four different seasons: Fall 2021 (September to November), Winter 2021 (December to February), Spring 2022 (March to May) and Summer 2022 (June to August). The overall data combined represents a full year of tourism movements in the city of Lisbon. This way it will be possible to analyse the data by season and get insights into patterns and trends that could not be obtained if using the overall data.

The dataset comprises various indicators concerning mobile terminal movements and user behaviours in Lisbon. In annex 4, table 11, find the complete set of indicators. These indicators can be leveraged into valuable insights regarding the user activity patterns and dynamics within the area

under consideration. These variables include counts of distinct terminals within the grid area, terminals in roaming mode, terminals remaining within the grid, terminals exiting the grid, and terminals with active data connections. Table 2 presents a brief explanation of some key variables that represent roaming data.

Table 2. Key Variables Description I

Key Variable	Description	Usage in this study
C.2 Number of distinct terminals, roaming, in the grid	Counts the unique mobile terminals detected in roaming mode within the specified grid area.	Provide insights into roaming users in the area, which is valuable for tourism and network management.
C.4 Number of distinct terminals that remained in the grid, roaming	Focuses on roaming terminals that remained within the grid area.	Provide insights into the behaviour of roaming users within the specified location.
C.7 Number of different terminal entries, roaming, in the grid	Number of entries made by roaming terminals into the grid area.	Provides information on the frequency of entry of roaming users into the specified location.
C.8 Number of different terminal outputs, roaming, in the grid	Similar to C.7 but focuses on the number of exits made by roaming terminals from the grid area.	Offer insights into the frequency of exits made by roaming users from the specified location.
C.10 Number of separate terminals with active data connection, roaming, in the grid	Focuses on roaming terminals with active data connections within the grid area.	Provide insights into the usage of data services by roaming users within the specified location.
C.11 Number of voice calls from the grid	Counts the number of voice calls that originated from the grid area.	Clarifies the communication patterns and demand for voice services within the specified location.

Additionally, some variables on the dataset can provide crucial information on user mobility, communication patterns, and internet usage. Analysing these indicators facilitates a deeper understanding of user behaviours, enabling informed decisions in urban planning, tourism management, and network optimisation for Lisbon. Table 3 describes a brief explanation of some variables that can elevate this thesis analysis:

Table 3. Key Variables Description II

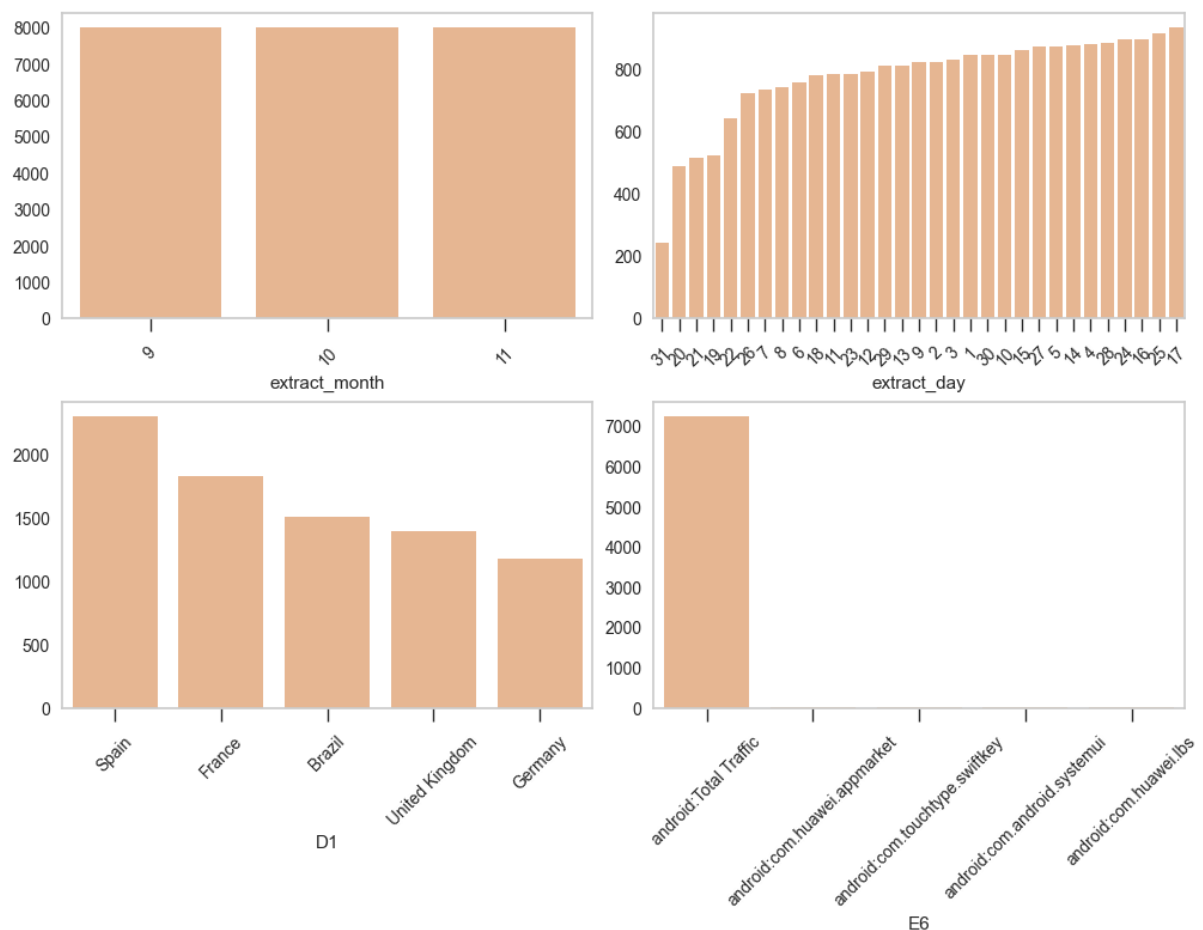
Key Variable	Description	Usage in this study
C.12 Number of entries in Lisbon for the 10 main roads	Counts the number of entries into Lisbon via the 10 main roads.	Provide insights into the traffic flow and transportation patterns in the city.
C.13 Number of departures from Lisbon for the 10 main roads	Similar to C.12 but focuses on the number of departures from Lisbon via the 10 main roads.	Help identify the outbound traffic flow and transportation patterns from the city.
D.1 Top 10 countries of origin of roaming terminal devices	Lists the top 10 countries of origin for roaming terminal devices detected within the specified area.	Present insights into the geographical distribution of roaming users.
E.1 Number of voice calls that ended in the grid	Counts the number of voice calls that terminated within the grid area.	Help comprehend the communication patterns and demand for voice services within the specified location.
E.6 Top 10 Applications (Text with names separated by;)	Lists the top 10 applications used by mobile terminals within the grid area, separated by semicolons.	Offer insights into the popular mobile applications among users in the specified location.
E.8 Duration of the average stay in the grid	Calculates the average duration of user stays within the grid area. It can	Help understand the typical length of time users spend in the specified location.
E.10 Number of devices that share the connection in the grid	Counts the number of devices sharing the connection within the grid area.	Provide insights into the network congestion level and sharing among users within the specified location.

From a quick overview of the statistics of the indicators, it can be highlighted that the top 5 countries of origin of roaming terminal devices across all seasons are mostly European countries, mainly Spain, France, the United Kingdom, and Germany, followed by Brazil. For the whole year of analysis, the most used application by roaming users (E6) is Total Traffic an Android app very popular and free, focused on travel and transportation. However other applications stand out in different seasons: in fall Huawei ID Service app, Totem Weather and Maps; in winter WhatsApp and Vodafone app; in spring WhatsApp, Facebook, and Google Play Store; and in summer WhatsApp and Google Chrome. The two indicators regarding the number of entries and departures from Lisbon through the 10 main roads (C12 and C13) only have values between September and November 2021, so these features will not be considered for the study since it is impossible to analyse them for the whole year.

At an initial data analysis stage, visualizations offer a comprehensive overview of the dataset. They provide a complete view of the data, which allows to identify the distribution, variability, and central tendency of variables. This broad understanding facilitates informed decision-making and hypothesis generation. Additionally, visualizations benefit in the detection of outliers, anomalies, and errors, which may require further investigation or data cleaning. Each type of visualization employed offers distinct advantages and insights, tailored to reveal specific facets of the data.

Bar plots are instrumental in visualizing categorical data, showcasing the frequency or distribution of different categories within a dataset (Rana, 2023). In this case, by analysing the bar plot in Figure 1, it is possible to recognise the evenly distributed data across all seasons. Considering the example of Fall 2021, each of the 3 months – September, October, and November - preserved equally 33.3% of the data. This statistic prevailed in the other seasons. Furthermore, all days of the month are represented in all seasonal data frames. Spain is the most common country of origin and Total Traffic is the most used application. One can also perceive that all applications mentioned in the variable E6 are Android Apps, which may indicate that other operating systems may have not allowed this information to be shared.

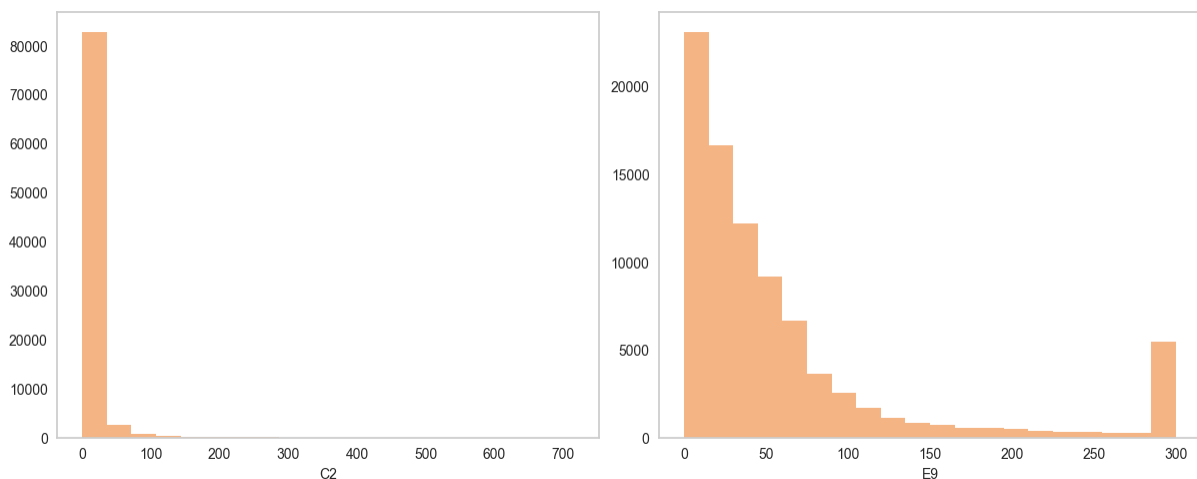
Figure 1. Bar Plot for categorical Figures for Fall



Histograms provide a graphical representation of the distribution of continuous numerical data. By visualizing the distribution of these variables, it can be identified potential outliers, central tendencies, and underlying patterns, thereby informing segmentation strategies and targeted interventions (Rana, 2023). Considering the context of this analysis, over all seasons and across most of the variables the data presents a positive skew (right-skewed), meaning that the data is not evenly distributed around the central point, instead, it exhibits a pronounced tail to the right-end side.

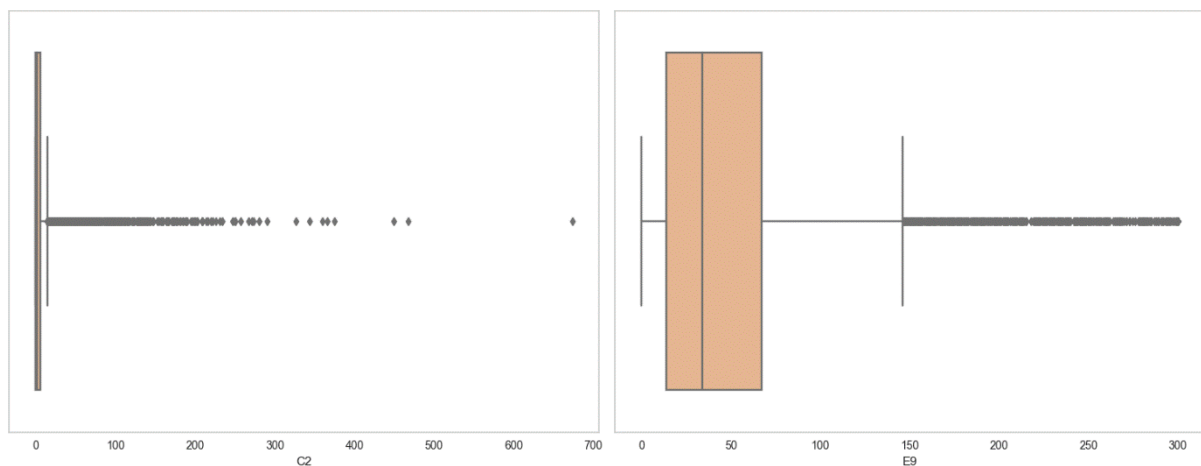
As an example, consider Figure 2 with the histogram for variables C2 (Distinct Roaming Terminals) and E9 (Maximum stay in the grid). This right-skewed data indicates that the terminal usage tends to be concentrated at lower values, with fewer occurrences of higher values, which aligns with the expectation that the majority of terminals may operate within a certain range. Understanding this skewness is crucial for resource allocation and infrastructure planning, as it suggests that the grid should be designed to accommodate peak usage periods while efficiently managing lower usage times. Additionally, the data is tightly clustered around the peak and presents a narrow spread which can suggest low variability within the variables, indicating consistent terminal usage patterns across different seasons. This finding is valuable for managing terminal demand, as it implies relatively stable usage levels over time. It also suggests that the grid infrastructure and services are effectively meeting the demand without significant fluctuations, contributing to overall operational efficiency. Overall, the frequency decreases as the number of distinct terminals increases. This trend suggests that while few terminals may experience high usage, the majority operate at lower levels of activity. Outliers are not easily identified through this visualization but box plots will present a clearer vision of the outliers.

Figure 2. Histogram for variables C2 and E9



Box plots offer a concise summary of the distribution of numerical data, highlighting key statistical analysis such as the median, quartiles, and outliers. The position of the box in the plot indicates the interquartile range of the data (McLeod, 2023). According to Figure 3, for most of the features, the box is positioned at the beginning of the graph, meaning that most data is concentrated at lower values. Features such as C1, and C3 exhibit higher medians (Q2) and third quartiles (Q3) in Fall and Spring, and C5 and C6 in Winter and Spring, indicating increased terminal activity during those periods. Understanding these seasonal variations can help in resource allocation and capacity planning, which may require additional infrastructure or operational adjustments to meet demand. The presence of outliers outside the maximum quartile (upper whisker) in several features, particularly in Winter and Spring, suggests the occurrence of unusually high values compared to the rest of the data. Potential reasons for outliers could include exceptional events, anomalies in terminal behaviour, or data recording errors. Overall, the box is skewed towards one side with a longer whisker, indicating a right-skewed distribution, where much of the data is concentrated towards lower values, but there are some extremely high values presented. This skewness in the distribution may be inherent to terminal usage patterns, where a large proportion of terminals exhibit low activity levels, while a few terminals experience exceptionally high usage.

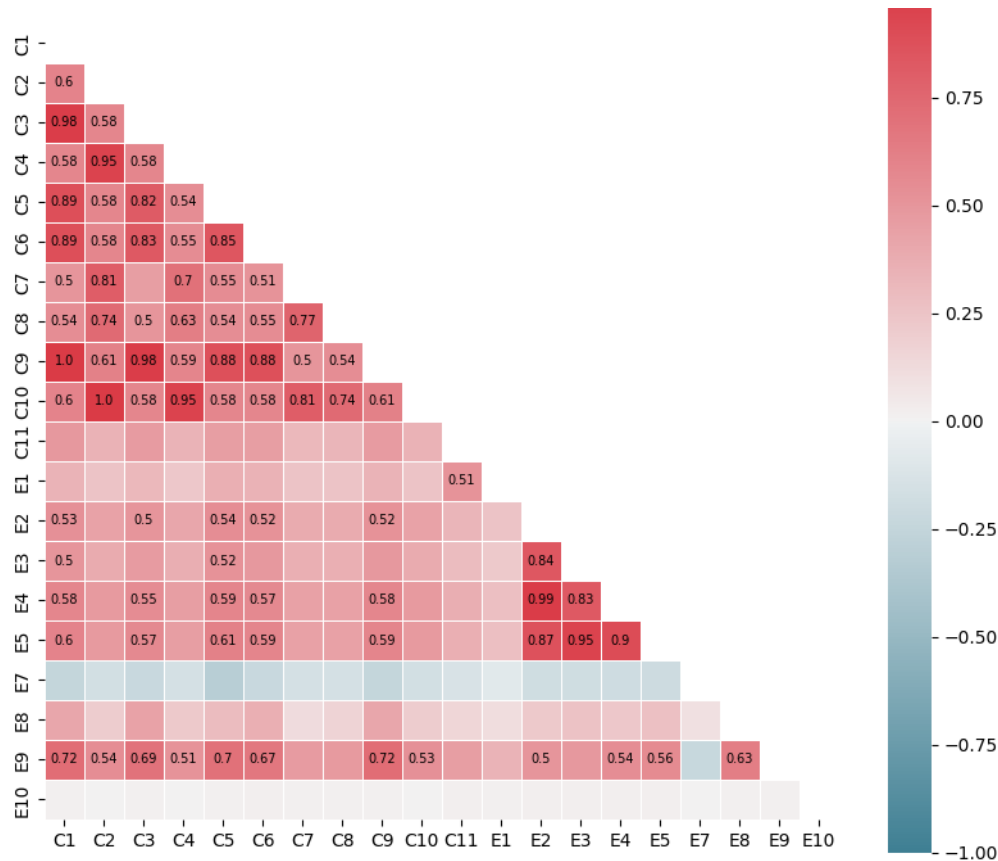
Figure 3. Box plot for variables C2 and E9



Spearman correlation matrices provide a comprehensive overview of the pairwise correlations between numerical variables, accounting for potential non-linear relationships (Frost, 2021). Figure 4 provides only the significant correlations in the data. When examining this correlation for all seasons, it can be observed positive correlations (close to 1), suggesting that as one variable increases, the other tends to increase as well. The high correlations between features related to the number of distinct terminals in the grid (C1-C6) indicate strong associations between different aspects of terminal activity, such as inputs, outputs, and roaming behaviour. These strong correlations suggest that certain terminal usage patterns tend to occur together, which could be indicative of consistent operational behaviours within the grid. The perfect correlation (1.0) between the number of distinct terminals with active data connection (C9) and the number of separate terminals with active data connection, roaming (C10) suggests a close relationship between these variables. It implies that terminals with active data connections tend to exhibit similar roaming behaviour, which could have implications for network management and data traffic optimisation. The correlation between the number of voice calls from the grid (C11) and other terminal usage features could indicate potential relationships between

voice communication patterns and terminal activity levels. Understanding these correlations can provide insights into the dynamics of communication within the grid and its impact on overall terminal usage. The high correlations between different grid rhythms (E2, E3, E4, E5) suggest coherent patterns in downstream and upstream data flows. These correlations can help identify synchronisation among data transmission and may be useful for optimising network performance and resource allocation.

Figure 4. Spearman Correlation Matrix for Numerical Variables



In summary, visualizations serve as indispensable tools in the data understanding phase of a project, offering insights into the structure, characteristics, and relationships within the dataset. By leveraging a variety of visualization techniques such as those mentioned above, it becomes possible to gain valuable insights that drive informed decision-making and further analysis regarding tourism mobility in Lisbon.

3.2. DATA PREPARATION

The Data Preparation phase involves transforming raw data into a clean, organised, and suitable format to then apply the most appropriate algorithm to the resulting dataset. This step ensures that the data is ready to be used effectively. Stepping into this process, the starting point will be outliers. Outliers are data points significantly different from the rest of the dataset and can greatly impact the results of clustering analysis. They can potentially lead to the formation of clusters that do not accurately represent the underlying structure of the data (Cohen, 2018). In the context of this thesis, outliers could represent tourists exhibiting unusually high data usage, roaming in unexpected

locations, displaying irregular roaming patterns, or staying for unusually long or short durations. These outliers must be treated appropriately to ensure the accuracy and reliability of the analysis. Addressing them is essential to mitigate their potential impact on clustering algorithms, ensuring that they do not distort statistical analyses. This approach is vital for accurately capturing typical tourist behaviours in Lisbon, as it fosters more reliable insights from the data. Considering previously explored literature, K-means will be the selected algorithm for this thesis and in this clustering method, the distance between data points is used to determine which points should be in the same cluster (Sun et al., 2021). Given its reliance on distance metrics and sensitivity to outliers, K-Means clustering is particularly vulnerable to the presence of outliers. Therefore, it's imperative to preprocess the data by removing or appropriately handling outliers to ensure the reliability of the clustering results.

Two methods were employed to deal with outliers. The Interquartile Range (IQR), a measure of statistical dispersion, delineates the middle 50% of a dataset by calculating the difference between the 75th and 25th percentiles (Q3 and Q1). Employing the IQR method for outlier detection involves establishing lower and upper bounds ($Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$), aiding in pinpointing potential outliers (Pereira, 2023). Robust to skewed data distributions, the IQR method identifies outliers based on percentiles, offering less sensitivity to extreme values. Its simplicity and effectiveness make it easily implementable and interpretable, providing a clear range for most data points, thereby serving as a valuable tool for data analysis and quality control, crucial for ensuring the accuracy and reliability of results. The Z-score, also known as the standard score, is a pivotal measure utilised in both scaling features for machine learning model training and identifying outliers within datasets. It quantifies how many standard deviations a data point deviates from the mean, with values greater than 3 often signalling extreme outliers. However, its effectiveness relies on the underlying assumption of data normality, and deviations from this assumption can impact the accuracy of Z-scores in reflecting the relative distances of data points from the mean (Chu & Chou, 2021). By applying the IQR method, a percentage below 90% of the data was kept after removing the outliers, which is a low percentage indicating a substantial loss of data. On the other hand, by applying the Z-score method, a percentage above 90% represents a more favourable ratio. Given this outcome, the Z-score method emerges as the preferred approach for handling outliers in this analysis. The immediate result after removing outliers is a cleaner and more focused dataset that better represents the typical patterns and behaviours of tourists in Lisbon, ensuring that the analysis will not be influenced by irregular data points.

To enrich the dataset, some new features were created in the feature engineering phase of the thesis. To differentiate the roaming activity during the weekend and weekdays, a feature called *Weekdays* was created. This variable provides the corresponding day of the week according to the Date of each line. Moreover, to have some input on the meteorological conditions during the stay of the roaming users, the variable *Temperature* was included. This data was collected by the “LoRa” network¹ and represents the average hourly temperature on the corresponding day in degrees Celsius. Moreover, there was some data provided by the CML including a table with the corresponding parish of the centroid of each grid and the exact spot of each terminal. In annex 2, table 9, find the range of variables in this table. This data was leveraged to add two new columns to the dataset: *Parish* which

¹ *<https://dados.gov.pt/pt/datasets/monitorizacao-de-parametros-ambientais-da-cidade-de-lisboa/>

provides the current parish for each line/grid and *Exact Location* which provides the name of the street or square of that terminal.

Moving forward, variable selection is an important step in the data preparation phase. Such selection aims to determine a set of variables that are related to the outcome and produce the most complete and accurate model (Brownlee, 2020). Metrics like the number of roaming terminals, their behaviour within specified grid areas, data usage patterns, communication activities, and network congestion levels offer valuable insights. These metrics help in understanding tourist presence, behaviour, preferences, and network management needs, contributing to more informed decision-making for tourism and network management purposes. To ensure the model's performance, features with low variance or high correlations with other variables were removed from the database. Conveniently, the removed features correspond to the variables not related to roaming data, such as C1 (N. of distinct terminals), C3 (N. of distinct terminals that remained in the grid), C5 (N. of different terminal inputs), C6 (N. of different terminal outputs), C9 (N. of terminals with active connection), E2 (Average downstream rhythm), E3 (Average upstream rhythm), E4 (Quad downstream rhythm), E5 (Quadruple peak rhythm), and E10 (N. of devices that shared connection), as they do not significantly contribute to the analysis. However, given the primary focus on leveraging roaming data, only variables related to roaming data will be considered for further analysis.

To better understand the data in the current dataset, the columns for numerical variables remaining for analysis were renamed and can be found in Table 4.

Table 4. Renamed columns of the dataset

C2	Distinct Roaming Terminals (DRT)
C4	DRT remained in grid
C7	DRT entries
C8	DRT exits
C10	DRT active connection
C11	N. voice calls
D1	Origin countries
E1	N. ended voice calls
E6	Applications
E7	Min. stay in grid
E8	Avg. stay in grid
E9	Max. stay in grid

It will only be possible to compare different variables with diverse scales or units by normalizing them, ensuring that each variable contributes equally to the analysis regardless of its scale (X. Wei, 2021). Normalizing data improves model performance by facilitating faster convergence during training and preventing features with larger scales from overshadowing those with smaller scales. It allows more stable and efficient models and increases interpretability by ensuring that all features contribute equally to the analysis, thereby simplifying the understanding of each feature's importance and associated coefficients. Additionally, it facilitates comparison by bringing all features to a similar scale, making it easier to compare their magnitudes and understand their relative importance in the model.

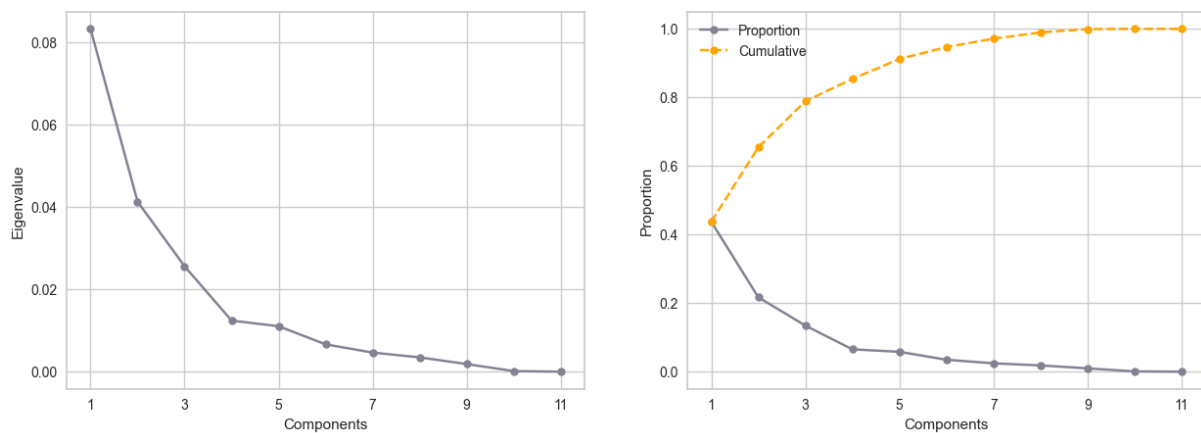
Regarding specific standardization methods, three different ones were applied to a data frame that combined all four seasons. MinMax Scaling rescales features to a specified range, typically between 0 and 1, which is useful when the minimum and maximum possible values for features are known (Xu et al., 2021). Standard Scaling scales feature to have a mean of 0 and a standard deviation of 1, assuming a Gaussian distribution of the data. Robust Scaling, on the other hand, subtracts the median and divides it by the interquartile range, making it robust to outliers and suitable for non-normally distributed data (Cosgun, 2023). For roaming data analysis, where outliers may be prevalent and feature distributions may not follow a Gaussian distribution, Robust scaling or MinMax scaling may be more appropriate. Robust scaling can effectively handle outliers, while MinMax scaling ensures that all features are within a consistent range, aiding in interpretability. MinMax scaling was the selected method to scale the numerical features of the dataset.

Categorical variables help organise data based on specific attributes, making it easier to segment and analyse information. Similarly, this type of data needs to be standardized and for that it was applied one-hot-encoding (OHC). This method transforms categorical data into numeric one by splitting the variables into multiple columns with "0" or "1" depending on which column has what value (Gefferth, 2024). Applying OHC is extremely relevant because the model K-means only accepts numerical input, so in this way, categorical variables can be used in the analysis. It can also improve model performance since it provides more information to the model. One-hot encoding was applied to variables *Origin Countries*, *Applications*, *Weekdays*, *Parishes*, and *Exact Location*.

Furthermore, dimensionality reduction techniques were also applied, to address the challenges presented by the large dataset, the new variables and the expansion of the categorical variables resulting from one-hot encoding. These techniques can reduce the redundancy in data and improve clustering performance. Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) indirectly achieve feature normalization by decorrelating and standardising principal components or preserving structure in high-dimensional data, respectively. UMAP operates by preserving both local and global structure within the data by constructing a high-dimensional graph representation and then optimising a low-dimensional projection to preserve the local and global structure of this graph (Coenen & Pearce, n.d.). It seeks to learn a low-dimensional representation of the data that preserves neighbourhood distances, effectively capturing both local and global patterns while minimising distortions in the resulting visualization. This makes it particularly useful for exploring complex datasets and finding meaningful relationships between data points.

PCA is a linear dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving most of the variability present in the original data. It is an unsupervised technique, making it suitable for datasets without labels, and it effectively addresses the challenges presented by high-dimensional data, aiming to describe the key characteristics of the data (Xu et al., 2021). PCA identifies principal components, or eigenvectors, which represent directions in the data with the highest variance, while the corresponding eigenvalues quantify the amount of variation explained by each component. Considering Figure 5 below, one can analyse the Elbow method and the explained variance ratio. By analysing the ratio, which indicates the proportion of variance explained by each principal component, one can determine the optimal number of components to retain. Using the elbow method, it was identified a point where the explained variance ratio starts decreasing at a slower rate, suggesting an ideal number of principal components. In the present analysis, retaining between 3 and 5 principal components explains above 80% of the data variation. Additionally, assessing the correlations of principal components with original features helps understand each feature's contribution to the principal components and their relationship with the original data. While UMAP is effective for preserving local and global structure in high-dimensional data, PCA's simplicity, interpretability, and ability to capture most of the variability make it a preferred choice in this scenario. PCA was performed with 4 principal components and the results show that principal component 0 is the one that best represents most of the variables in the dataset.

Figure 5. Elbow method and Variance explained Ratio, respectively.



After dealing with outliers, adding new features, selecting the best variables for the model, applying normalization techniques for numerical data, one-hot-encoding to categorical data, and reducing the dimensionality, some data exploration should be processed to ensure that the data looks the way it is expected to. Outliers and missing values were checked, and some plots previously mentioned were calculated once again to verify the information in the dataset and if there were any mistakes. Subsequently, the dataset is ready for the model application. Table 5 shows the final variables used in the model.

Numerical Variables		Categorical Variables	
- Distinct Roaming Terminals (DRT)		- Origin countries	
- DRT remained in grid		- Applications	
- DRT entries		- Week Days	
- DRT exits		- Parish	
- DRT active connection	- Min. stay in grid	- Exact Location	
- N. voice calls	- Max. stay in grid		
- N. ended voice calls	- Avg. stay in grid		
- Temperature	- PCO		

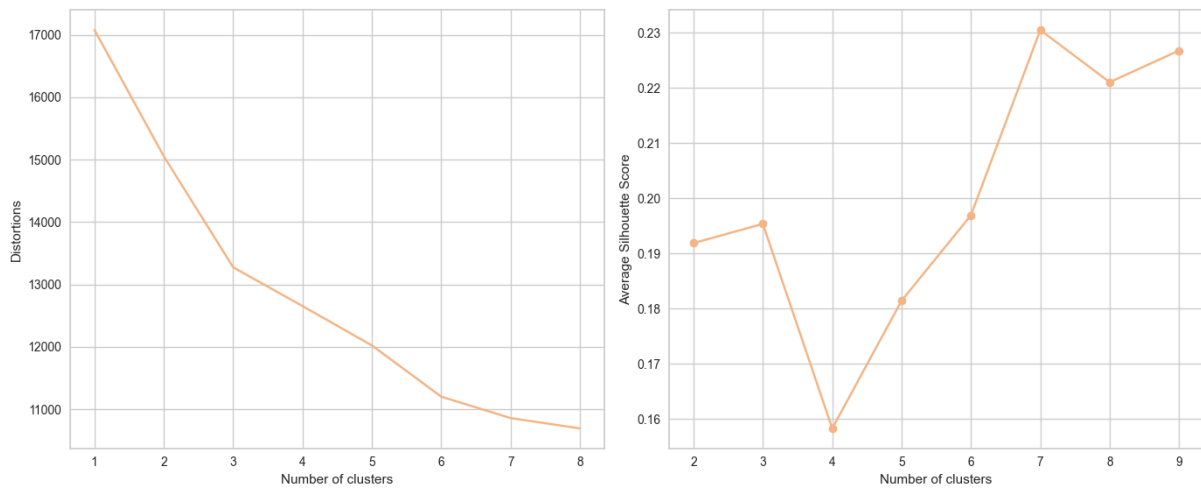
Table 5. Final selected Variables to use in K-means

3.3. MODELLING

K-means clustering is an unsupervised algorithm that groups similar data points by dividing a dataset into a number K of clusters based on their similarity. It minimises the sum of squared distances between observations and their assigned cluster centroids. The algorithm initialises K centroids randomly and iteratively assigns data points to the cluster with the nearest centroid, recalculating centroids based on the mean of assigned data points until convergence. In the case of roaming data, where the number of clusters is not predefined, K-means is preferred for its efficiency and ability to handle large datasets while providing meaningful cluster assignments (Sun et al., 2021). The variables used in the model are the ones selected in the variable selection phase, which were submitted to normalization and one-hot encoding so that are eligible for the K-means algorithm.

To pinpoint the optimal number of clusters for meaningful insights into tourist behaviours around Lisbon, two key metrics are employed: distortion and silhouette score. Distortion measures the sum of squared distances against the number of clusters and identifies the "elbow" point signifying where the distortion decreases more slowly. Meanwhile, the silhouette score measures intra-cluster similarity and inter-cluster dissimilarity, guiding towards the ideal cluster count, with higher scores indicating better clustering (Kumar, 2021). In this scenario, and as can be seen in Figure 6, distortion is declining steadily and silhouette scores show an initial improvement from three clusters, followed by fluctuations and a slight increase at seven clusters, suggesting that this configuration captures meaningful patterns in the data without introducing unnecessary complexity. Opting for three clusters strikes a balance between granularity and interpretability, capturing distinct tourist groups without excessive complexity. Alternatively, selecting seven clusters offers increased precision, potentially finding finer distinctions in tourist behaviours. Considering the previously explained options, the two solutions were applied and moving forward the best one will be selected.

Figure 6. Inertia Plot and Average Silhouette Score, respectively



After a thorough examination of multiple clustering metrics, as shown in Table 6, including cluster cardinality and magnitude (Berman, n.d.), R-squared (Pereira, 2023), silhouette score (Kumar, 2021), Davies-Bouldin score (Sheikh Hassan et al., 2021), and Calinski and Harabasz score (H. Wei, 2020), it becomes evident that the 3-cluster solution presents a more compelling choice for the dataset. Cluster cardinality and magnitude provide insights into the distribution and size of each cluster, with the 3-cluster solution exhibiting a more balanced distribution compared to the 7-cluster solution. Despite a slight imbalance in cluster sizes, the 3-cluster solution demonstrates superior performance across various evaluation criteria. The R-squared value, which measures the proportion of variance explained by the clustering solution, is higher for the 3-cluster solution compared to the 7-cluster solution, indicating better explanatory power. Similarly, the positive silhouette score for the 3-cluster solution suggests better cluster separation and cohesion compared to the negative silhouette score of the 7-cluster solution. The Davies-Bouldin score, which evaluates the quality of clustering based on cluster separation, compactness, size, and dispersion, is lower for the 3-cluster solution compared to the 7-cluster solution, indicating more distinct and compact clusters. The Calinski and Harabasz score, which measures the ratio of between-cluster variance to within-cluster variance, is higher for the 3-cluster solution compared to the 7-cluster solution, suggesting better cluster separation and cohesion. Therefore, selecting the 3-cluster solution promises to yield more interpretable and robust clustering results, facilitating a deeper understanding of underlying patterns in tourism activity in Lisbon.

Table 6. Scores considered to select ideal number of clusters.

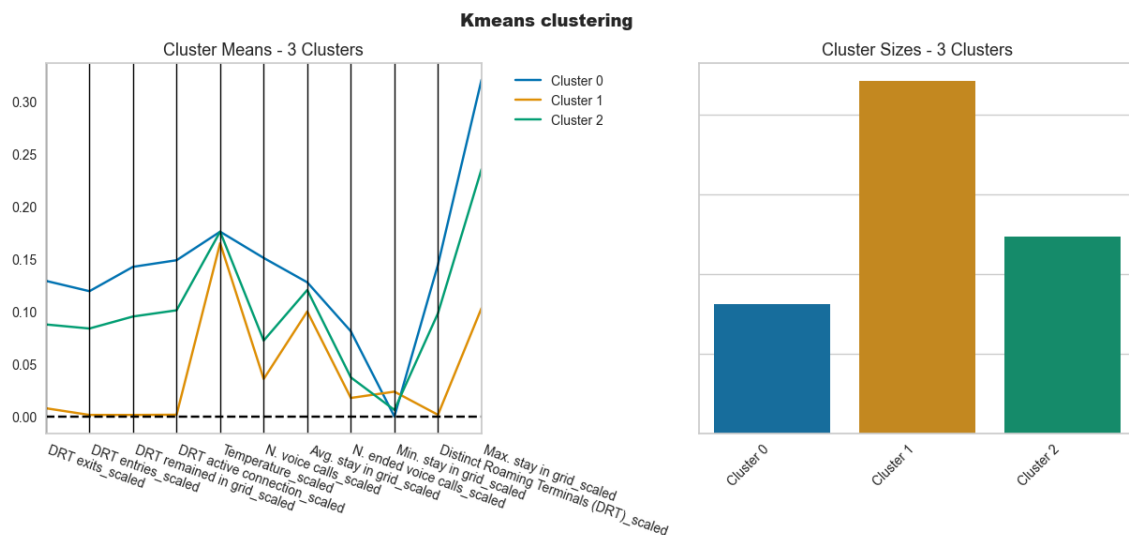
N. Clusters	R-squared	Silhouette	Davies-Bouldin	Calinski and Harabasz
3	0.1373	0.1045	4.5111	679.0863
7	0.1257	-0.0973	18.3266	204.3797

4. RESULTS AND DISCUSSION

4.1. CLUSTERING RESULTS

Considering the whole process behind the methodology phase, it is now possible to draw conclusions based on the three segments of roaming users in Lisbon. To do so, the normalized dataset was converted to its original format, maintaining the label of the clusters. The mean results were analysed for the numerical and categorical features. From Figure 7, which presents the mean results of each cluster for the numerical features and the cluster sizes, insights can be found to help in the management of public spaces or tourism-related services and to understand the behaviour of tourism activity in Lisbon as the objectives of this thesis aid. Primarily, let it be stated that the varying values across clusters indicate distinct patterns of tourism behaviour.

Figure 7. Clusters Profile



Cluster 0 is characterised by longer stays in the grid, higher entry and exit rates, and increased engagement in voice calls and active connections. It likely represents tourists who spend more time exploring the city and are actively engaged with their surroundings. These tourists may benefit from targeted services that are adapted to their longer stays and higher levels of activity. In contrast, Cluster 1, which indicates shorter stays in the grid, lower entry and exit rates, and fewer active connections, may include more short-term tourism seeking quick experiences or passing through the city. Understanding these distinctions between longer and shorter stays in the grid allows for shaped strategies in managing tourist flows, optimising resources, and enhancing the overall visitor experience in Lisbon. Cluster 2 tourists demonstrate behaviours and characteristics that lie between the opposites observed in Clusters 0 and 1. It exhibits moderate levels of stay duration in the grid, entry and exit rates, and engagement in activities such as voice calls and active connections. These tourists may represent a mix of visitors who are neither staying for extended periods nor quickly passing through the city. Instead, they may be individuals or groups with a balanced approach to tourism, looking for a combination of exploration and entertainment during their time in Lisbon. This type of tourist is likely to engage in various activities and attractions but may not be as committed to extended stays as those in Cluster 0. Similarly, their stay may not be as momentary as for tourists in Cluster 1, indicating a level of interest in experiencing the city beyond its most visited attractions.

Upon comparing the daily weather temperature across clusters, it appears that there is minimal temperature variation. The differences in mean temperature between clusters 0, 1, and 2 are quite small, with cluster 1 having the lowest mean temperature and cluster 2 having the highest. This suggests that temperature may not be a significant factor in differentiating the behaviour or characteristics of tourists over clusters, indicating that Lisbon is a tourist destination across all seasons.

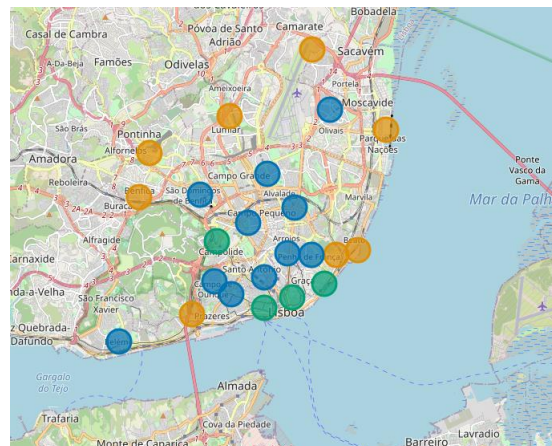
As for the results of the categorical variables, a thorough analysis was done on the mean results of these variables. Analysing the *Origin Country* of the roaming users revealed patterns in the composition and characteristics of tourists visiting Lisbon. Cluster 0 primarily consists of tourists from various European countries such as France, Germany, Italy, Spain, and the United Kingdom, indicating a diverse international mix. These tourists likely represent a significant portion of long-stay visitors exploring Lisbon's attractions and cultural legacy. Also, Cluster 1 is predominantly composed of tourists categorised under "None", which suggests a substantial proportion of missing or unspecified origin country data. This cluster may encompass short-term visitors whose origin country information was not recorded, indicating an unrelated mix of tourists with varying travel motivations and preferences. Cluster 2 exhibits similarities to Cluster 0, with tourists from countries like France, Germany, Italy, Spain, and the United Kingdom. Tourists from countries such as Brazil and Finland have relatively higher representation in this cluster. Cluster 2 likely includes tourists with diverse travel preferences, including both long-stay visitors and those seeking shorter leisure trips.

By analysing the top 10 most used *Applications* by the Lisbon tourists across clusters it is possible to get insights into tourists' digital behaviour and preferences during their stay in Lisbon. Cluster 0 presents a diverse range of application usage, with tourists accessing various applications such as navigation apps, messaging platforms like WhatsApp, and entertainment apps like Spotify. This implies that tourists in this cluster engage in varied activities, including exploration, communication, and leisure, during their visit to Lisbon. Cluster 1 is characterised by a dominant usage of the application category labelled as "None" indicating a significant proportion of users whose application usage data is missing or unspecified. This cluster may include tourists who rely less on digital applications for navigation, communication, or entertainment purposes during their stay. Cluster 2 shares similarities with Cluster 0 in terms of application usage, with tourists accessing navigation apps like Google Maps, WhatsApp, and utility apps like email. However, Cluster 2 also shows a notable usage of specific local applications such as Portugalia and Emel ePark suggesting that tourists in this cluster engage with local services during their visit. These findings highlight the diverse digital behaviour of tourists in Lisbon and feature the importance of understanding their application usage patterns to improve visitor experiences to meet their needs.

Furthermore, examining the distribution of roaming terminals across different *Parishes* revealed distinct spatial patterns of tourist activity within Lisbon. In Cluster 0, the parishes of Olivais and Belém stand out with relatively higher proportions of roaming terminals compared to other parishes. This means that tourists in this cluster are more likely to visit these areas, probably due to attractions like the Parque das Nações and all the historical landmarks in Belém. Additionally, Alcântara and Lumiar also demonstrate notable proportions of roaming terminals in this cluster, indicating tourism activity in these areas. In Cluster 1, the parishes of Benfica, Marvila, and Santa Clara appear as the top three areas with significant concentrations of roaming terminals. This indicates that tourists in this cluster may be attracted to the cultural sites in Benfica, the restored industrial zones in Marvila, and the modern developments in Santa Clara. Alvalade and São Domingos de Benfica also show considerable

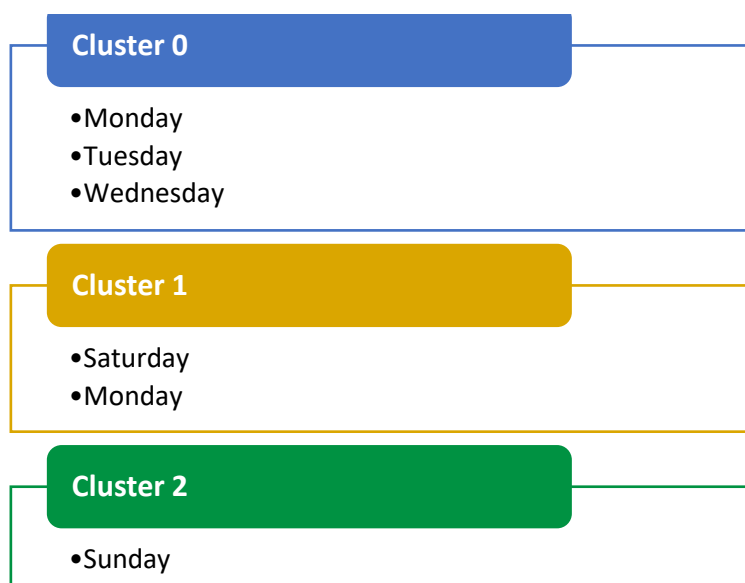
tourist activity in Cluster 1. Cluster 2 portrays a different distribution of roaming terminals, with parishes such as Belém, Avenidas Novas, and Campo de Ourique having higher proportions compared to other areas. This points out that tourists in Cluster 2 may prefer exploring cultural attractions in Belém, exclusive neighbourhoods in Avenidas Novas, and the charming area of Campo de Ourique. Overall, the analysis of roaming terminals across parishes highlights popular tourist destinations, cultural attractions, and commercial zones of interest to the visitors. Figure 8 is a map that represents the spatial distribution of the Clusters in Lisbon, consider the above pre-defined colours as legend (Blue – Cluster 0, Orange – Cluster 1 and Green – Cluster 2).

Figure 8. Spatial Distribution of Clusters in Lisbon



The distinct roaming terminals variation during the week suggests the preferred days for tourism activities in each group of tourists. Evaluating the organisation of roaming terminals across different *Weekdays* revealed that some tourists are more active at the beginning of the week and others at the weekend. These insights help stakeholders to better plan and manage tourism-related activities and services to increase the overall tourist experience in those days. Figure 9 represents the weekdays with a higher proportion of roaming terminals per cluster.

Figure 9. Weekdays with higher DTR per cluster



When analysing the *Exact Location* of tourists by segment it was feasible to conclude that Cluster 0 consists of locations that appear to be relatively less frequented. Some examples within this cluster include specific streets, residential areas, and lesser-known landmarks. On the other hand, Cluster 1 represents key tourist spaces within Lisbon. These locations demonstrate moderate to high values, indicating their significance as popular tourist spots or major landmarks. Cluster 1 likely involves well-known tourist attractions, parks, historical sites, and popular neighbourhoods, which contribute significantly to the city's tourism economy. Ultimately, Cluster 2 represents locations with mixed values, generally lower compared to Cluster 1. These areas likely experience some level of tourist activity but may not be as known. This cluster includes smaller parks, local attractions, or less famous historical sites.

Understanding the distribution of tourism activity across the clusters is crucial for strategic planning, resource allocation, and sustainable tourism management in Lisbon. By prioritising resources for marketing, infrastructure development, and visitor services in key tourist spots, tourism experts can increase the overall tourist experience and maximise the economic benefits of tourism for the city of Lisbon. Table 7 summarises the most important findings of the previously described cluster results.

Table 7. Cluster Characterisation

Clusters	Results for Numerical Features	Results for Categorical Features
Cluster 0	Longer stays in grid, higher entry and exit rates, and increased engagement in voice calls and active connections.	Tourists mostly from France, Germany, Italy, Spain, and the United Kingdom. The most used apps were Navigation Ones, WhatsApp, and Spotify. Olivais, Belém, Alcântara and Lumiar as top parishes. Monday and Tuesday are the busiest days. The exact location represents specific streets, residential areas, and lesser-known landmarks.
Cluster 1	Shorter stays in grid, lower entry and exit rates, and fewer active connections.	Short-term visitors whose origin country information and application usage data were unspecified. Benfica, Marvila, and Santa Clara emerge as the top three areas with more roaming users. Saturday and Monday stand out as the chosen days. Includes well-known tourist attractions, historical sites, and popular neighbourhoods.
Cluster 2	Moderate levels of stay duration in grid, entry and exit rates, and engagement in activities such as voice calls and active connections.	Most tourists are from Brazil and Finland. Portugalia and Emel were a lot of used apps. Belém, Avenidas Novas, Alvalade and Campo de Ourique are the chosen destinations. Sunday is the weekday with the highest proportion. Include smaller parks, local attractions, or less famous historical sites.

4.2. MANAGERIAL INSIGHTS

Tourists in Cluster 0, characterised by longer stays in the grid and higher levels of activity, stand to benefit from a range of targeted services adapted to their preferences. A longer stay in the grid could suggest that these tourists spend a relatively extended period at a specific location within the city. While this may not necessarily indicate a long-duration trip to Lisbon as a whole, it does imply a significant amount of time spent at a particular spot. Therefore, a longer stay in the grid could reflect a deeper level of engagement or interest in a particular area or activity. Considering the previous statement, the accommodation businesses could leverage the possible longer duration of stays by offering discounted rates for extended bookings or package deals that include additional services for those. Tour promoters and tourism agencies could design multi-day itineraries showcasing the diverse attractions and experiences that Lisbon has to offer (Sarkar & Majumder, 2021), including visits to lesser-known landmarks and the residential neighbourhoods highlighted in Cluster 0 analysis. Wellness centres and spas in Lisbon could adapt their services to satisfy the needs of Cluster 0 tourists, offering specialised packages for relaxation and renewal after their active days of exploration. Furthermore, restaurants and shops could implement loyalty programmes to motivate repeated visits, providing discounts or exclusive offers to Cluster 0 guests. Community engagement initiatives could also play a significant role in improving the experience of Cluster 0 tourists, for instance, local organisations could offer volunteer opportunities or cultural exchanges, allowing visitors to immerse themselves more deeply into Lisbon's community. This not only enriches the tourist experience but fosters meaningful connections between visitors and locals. For the city of Lisbon, leveraging the characteristics of Cluster 0 tourists presents an opportunity to promote sustainable tourism and economic growth. By satisfying the needs and preferences of the promising longer-staying visitors, the city can encourage spending in local businesses and foster positive relationships between tourists and residents. Investing in infrastructure and services in areas frequented by these tourists, such as the parishes of Olivais, Belém, Alcântara, and Lumiar, can increase the overall visitor experience and support local businesses. Moreover, highlighting lesser-known attractions and residential areas can distribute tourism traffic more evenly across the city, reducing overcrowding in popular tourist hotspots and promoting a more authentic and diverse visitor experience. City authorities could collaborate with local businesses and community organisations to develop marketing campaigns and initiatives that specifically target Cluster 0 tourists. These efforts could include promoting extended stay packages, organising cultural events in residential neighbourhoods, and developing and improving infrastructure and facilities in areas frequented by Cluster 0 visitors. Overall, by offering specialised services and experiences tailored to the preferences of Cluster 0 tourists, both tourists and the city of Lisbon can benefit from enriched economic opportunities, cultural exchange, and sustainable tourism practises.

Adapted strategies for managing tourist flows, optimising resources, and boosting the overall visitor experience in Lisbon can be specifically planned for Cluster 1 tourists. These tourists, characterised by shorter stays in the grid and fewer active connections, often seek quick experiences or are passing through the city. A shorter stay in the grid can suggest that the tourist either passed through the area quickly or did not spend much time engaging with the surroundings. While it doesn't necessarily imply a short-duration trip to Lisbon overall, it may indicate a more transient or less engaged visit to specific locations within the city. For tourists belonging to Cluster 1, efficient transportation options such as shuttle services or hop-on-hop-off buses, that follow a pre-defined route, can help them navigate the city's attractions instantly (Philip Stemmann, 2024). In addition, investing in express/priority passes to popular tourist sites and attractions can minimise waiting times

and ensure that visitors make the most of their limited time in Lisbon (Aaron Hurd, 2024). Short-term tour packages and compact city guides designed to accommodate their interests can further elevate their experiences, allowing tourists to explore the essence of Lisbon within their apparent shortened timeframe. Moreover, personalised customer support can speed up the process of booking restaurants, organising tours, or buying tickets for events or museums (Patrick Grieve, 2024). Self-guided walking tours can add convenience and flexibility to their itineraries, enabling them to maximise their experience in the city. In parallel, for the city of Lisbon, developing transportation infrastructure, including the options mentioned above such as shuttle services, can better serve the needs of temporary visitors and facilitate smooth mobility. Furthermore, developing a digital city guide or mobile application offering personalised recommendations and itineraries fitted to different visitor clusters (Anacleto et al., 2014), including Cluster 1 tourists, can enrich their experience. The many tourist businesses in Lisbon may consider investing in hospitality training programmes for service providers which can also equip them with the skills necessary to accommodate the diverse needs of short-term tourists effectively. Eventually, promoting flexible booking options and custom-made experiences can encourage this type of visitor to explore Lisbon even with limited time, benefiting both tourists and the city by fostering a greater visitor experience, increased satisfaction, and optimised resource management.

For tourists in Cluster 2, characterised by generally moderate levels of stay duration in the grid, entry and exit rates, and engagement in activities, a balanced approach to services can greatly improve the experience of these visitors in Lisbon. It is important to acknowledge that these tourists represent a mix of visitors with varying levels of commitment to exploring the city. Therefore, accommodation businesses should consider offering flexible booking options to fit the varying lengths of stay, accommodating both short-term and longer-term visitors (Tourism Tiger, n.d.). Tourism agencies could provide customisable itineraries that satisfy the diverse interests and preferences of this cluster of tourists, highlighting both popular attractions and lesser-known local areas. Likewise, establishments across the city can focus on providing express check-in services which can simplify the arrival process, while quick-service options, such as lunch menus or takeaway, at restaurants and cafes can accommodate their busy schedules. As suggested previously, implementing mobile app-based recommendations for nearby attractions, dining options, and activities can further facilitate their exploration of Lisbon, providing them with valuable insights and suggestions that can be adapted to their interests. Several strategies can benefit both tourists in Cluster 2 and the local economy. Diversified Marketing Efforts should be carried out to promote a diverse range of attractions, including smaller parks, local landmarks, and lesser-known historical sites, to attract Cluster 2 tourists who seek authentic and hidden experiences. Efforts can be made to boost accessibility to tourist attractions and facilities in areas such as Belém, Avenidas Novas, Alvalade, and Campo de Ourique, ensuring that visitors can easily explore these neighbourhoods and enjoy their offerings. The tourists in this cluster would also appreciate any kind of collaborative tourism initiative. These initiatives could be fostered between local businesses, tourism agencies, and community organisations to develop unique experiences and offerings and promote sustainable tourism practises. For instance, workshops, concerts, or cultural events.

Finally, implementing customer feedback mechanisms to gather insights and suggestions from Lisbon tourists enables continuous improvement of services and facilities to better meet their needs and preferences (Caemmerer & Wilson, 2010). Through all mentioned adapted strategies, both Lisbon tourists and the city itself can benefit from enriched tourism experiences, increased visitor satisfaction, and sustainable economic growth. Consequently, these tourists are likely to feel motivated to recommend this trip to friends and relatives and ideally to revisit Lisbon or even another city in Portugal. Overall, the achievement of the objective of this study is evident through the identification of three clusters of tourists in Lisbon and the insights derived from their characteristics. City authorities can use this information to adapt public policies, infrastructure development, and resource allocation to better accommodate the needs and preferences of different types of tourists. Local businesses can leverage these insights to optimise services and offerings, thereby enhancing the overall visitor experience. Additionally, tourists themselves can benefit from personalised recommendations and experiences based on their cluster profile (Buhalis & Amaranggana, 2015).

5. CONCLUSIONS AND FUTURE WORKS

5.1. CONCLUSIONS

This thesis has explored the complex behaviour of tourist mobility within urban landscapes, particularly focusing on Lisbon, a major tourist destination, as a case study. By leveraging mobile phone roaming data provided by Vodafone and employing data mining techniques, this thesis aimed to characterise diverse tourist groups, understand their spatial behaviour and provide insights to improve urban planning and tourism management strategies in Lisbon. It builds upon existing methodologies by leveraging data to offer a granular understanding of tourist movements and their implications for city management.

The study followed the CRISP-DM methodology and the initial exploration of the data revealed that the top five countries of origin of roaming terminal devices are predominantly European and that navigation applications are the most frequently used. Subsequent phases involved data preparation, feature engineering, and variable selection, aligning with the overarching goal of the thesis. Techniques such as the Z-score method for outlier removal, adding the features *Weekday*, *Temperature*, *Parish* and *Exact Location*, standardization using MinMax Scaler for numerical features and one-hot-encoding for categorical ones were instrumental in preparing the dataset for analysis and ultimately the final algorithm. Furthermore, Principal Component Analysis (PCA) helped address challenges caused by the large dataset. The resulting three-cluster solution, selected after examining multiple clustering metrics, provided a balanced distribution representing distinct tourist behaviours within Lisbon.

The present research aimed to address the gap in understanding tourist behaviour within cities, particularly Lisbon, and provide actionable insights for urban planning and tourism management. It was possible to successfully characterise three tourist groups, as well as to identify key areas of tourism activity and propose strategies to develop the visitor experience in Lisbon. By understanding the diverse patterns of tourist behaviour, city stakeholders and policymakers can develop targeted strategies to optimise resources, improve infrastructure, and tailor services to meet the needs of different tourist clusters. To better serve Cluster 0 tourists, accommodations could offer discounts for their extended stays, while tour agencies could design multi-day itineraries showcasing diverse attractions and lesser-known areas. Restaurants, shops, and wellness centres can implement loyalty programmes and specialised packages to improve visitor experiences. Additionally, investing in infrastructure in frequented areas by these tourists, such as Belém and Alcântara, and promoting lesser-known attractions can distribute tourism evenly and support sustainable growth. As for tourists in Cluster 1, characterised by shorter stays and quick experiences, efficient transportation options like shuttle services and hop-on-hop-off buses can help them navigate the city instantly. Investing in express passes for popular attractions and creating short-term tour packages or compact city guides can maximise their limited time. In addition, enhancing digital city guides, personalised customer support, and flexible booking options can further enrich their experience, ensuring a smooth and satisfying visit. Ultimately, for Cluster 2 Tourists with moderate stay durations and varied engagement levels, accommodation businesses should offer flexible booking options to suit both short-term and longer-term visitors. Tourism agencies can provide customisable itineraries that highlight both popular attractions and lesser-known areas, while establishments can offer express check-in services and quick-service dining options. Developing mobile app recommendations for nearby attractions and

dining could facilitate exploration. Furthermore, promoting diverse attractions and improving accessibility in areas like Avenidas Novas, Alvalade, and Campo de Ourique, along with fostering collaborative tourism initiatives, can enhance their overall experience and benefit the local economy.

From personalised experiences to efficient resource allocation, the insights resulting from this research can contribute to the sustainable development of tourism in Lisbon and boost the overall visitor experience. Moreover, entrepreneurs and business owners can leverage these findings to adjust marketing campaigns and make informed decisions about investments in tourism-related projects. By aligning their strategies with the identified tourist clusters, businesses can improve their competitiveness, thereby contributing to the economic growth and vitality of Lisbon's tourism industry.

In comparison to similar studies, such as the one by Francisco et al. (2023) in analysing tourist behaviour using data from a mobile operator, this research stands out for its implications in urban planning and tourism management. While previous studies may have touched upon aspects of tourist mobility, this study examines deeper by not only identifying tourist clusters but also characterising their preferences and behaviours. By doing so, it addresses a critical gap in the literature, which often lacks in-depth analysis of tourist movements within cities, especially using rich data sources like mobile phone roaming data. In the bigger picture, this study exemplifies the power of data-driven approaches in addressing complex urban challenges and shaping the future of cities as vibrant, resilient, and visitor-friendly spaces. The groundwork for a more sustainable approach to managing tourism in Lisbon is laid, ultimately enriching the city's cultural landscape and visitor experience.

5.2. LIMITATIONS AND FUTURE WORK

While this thesis provides a comprehensive examination of tourist behaviour in Lisbon, it's important to acknowledge some potential limitations that influenced the research design and outcomes. One such limitation resulted from the decision to use a sample spanning a complete year. While this approach offers numerous advantages such as ensuring data representativeness, reflecting current trends, and facilitating broad applicability of the conclusions to various stakeholders including entrepreneurs, students, and tourists, it also introduces constraints. Primarily, the choice of a year-long sample was made to maximise the quality and reliability of the data, providing a comprehensive understanding of tourist behaviour throughout different seasons and times of the year. However, this decision demanded significant computational resources to process and analyse the extensive dataset effectively. Despite optimisation efforts, constraints in computational capacity limited the extent of data analysis. With increased computational capacity, a more extensive dataset could have been analysed, potentially leading to even more accurate and multifaceted results.

Moreover, had the computational capacity not been a limiting factor, segmentation of the data by season to characterise tourist behaviour across different seasons could have been pursued, as it was intended from the beginning. This approach would have provided valuable insights into seasonal variations in tourist behaviour, enabling even more adapted strategies for each season. However, such segmentation would require additional time and resources for data preparation, processing, and analysis, which were beyond the scope of the current study.

Furthermore, while employing data mining techniques such as density-based clustering is valuable, there may exist alternative methods or additional analyses that could provide further insights

into tourist behaviour. Exploring different methodologies could strengthen the validity of the research outcomes. Additionally, tourist behaviour is naturally dynamic and subject to change over time. While this thesis may capture a snapshot of tourist behaviour during a specific period, ongoing monitoring and analysis are necessary to understand how behaviours evolve and adapt accordingly.

There are still some proposals on how to complement this research going forward. The dataset could include more data associated with tourism and its spatial distribution. For instance, specific points of interest and similar features would positively impact the definition of the clusters. Moreover, extending the spatial analysis to examine micro-level spatial patterns within neighbourhoods, attraction sites, or specific tourism hotspots. This could provide more accurate insights into areas of high tourist density, congestion, or underutilisation.

Addressing these limitations transparently and discussing potential paths for future research can enrich the scholarly contribution of this thesis and guide further advancements in understanding tourist behaviour and its implications for urban management and tourism development. As future work, it is planned to improve the current computational capacity, as it revealed to be a major challenge, then reanalyse a more extensive dataset and define the clusters by season, as intended from the start, to gain a deeper understanding of tourism activity. These improvements will provide more insights and reinforce the applicability of the findings, contributing to more informed decision-making in urban management and tourism development.

BIBLIOGRAPHICAL REFERENCES

- Aaron Hurd. (2024). How to Maximize Your Priority Pass Select Membership Rewards. *Time Stamped*.
- Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469–486. <https://doi.org/10.1016/J.TOURMAN.2007.05.014>
- Anacleto, R., Figueiredo, L., Almeida, A., & Novais, P. (2014). Mobile application to provide personalised sightseeing tours. *Journal of Network and Computer Applications*, 41(1), 56–64. <https://doi.org/10.1016/J.JNCA.2013.10.005>
- Balzotti, C., Bragagnini, A., Briani, M., & Cristiani, E. (2018). *Understanding Human Mobility Flows from Aggregated Mobile Phone Data**. 51(9), 25–30. <https://doi.org/10.1016/j.ifacol.2018.07.005>
- Barata-Salgueiro, T., Mendes, L., & Guimarães, P. (2018). Tourism and urban changes. *Tourism and Gentrification in Contemporary Metropolises*, 255–275. <https://doi.org/10.4324/9781315629759-11>
- Berman, D. (n.d.). *The Cardinality Challenge in Monitoring* | Logz.io. Retrieved May 7, 2024, from <https://logz.io/blog/cardinality-challenge-in-monitoring/>
- Bouchon, F. (2022). Urban Tourism. *Encyclopedia of Tourism Management and Marketing*, 649–652. <https://doi.org/10.4337/9781800377486.URBAN.TOURISM>
- Brownlee, J. (n.d.). *Feature Selection For Machine Learning in Python - MachineLearningMastery.com*. 2020. Retrieved May 7, 2024, from <https://machinelearningmastery.com/feature-selection-machine-learning-python/>
- Buhalis, D. (2020). Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article. *Tourism Review*, 75(1), 267–272. <https://doi.org/10.1108/TR-06-2019-0258/FULL/HT>
- Buhalis, D., & Amaranggana, A. (2013). Smart Tourism Destinations. *Information and Communication Technologies in Tourism 2014*, 553–564. https://doi.org/10.1007/978-3-319-03973-2_40
- Buhalis, D., & Amaranggana, A. (2015). Smart Tourism Destinations Enhancing Tourism Experience Through Personalisation of Services. *Information and Communication Technologies in Tourism 2015*, 377–389. https://doi.org/10.1007/978-3-319-14343-9_28
- Caemmerer, B., & Wilson, A. (2010). Customer feedback mechanisms and organisational learning in service operations. *International Journal of Operations and Production Management*, 30(3), 288–311. <https://doi.org/10.1108/01443571011024638>
- Castela, A. (2018). Impacts of Tourism in an Urban Community: The Case of Alfama. *Athens Journal of Tourism*, 5(2), 133–148. <https://doi.org/10.30958/AJT.5-2-4>

- Chu, C. P., & Chou, Y. H. (2021). Using cellular data to analyze the tourists' trajectories for tourism destination attributes: A case study in Hualien, Taiwan. *Journal of Transport Geography*, 96. <https://doi.org/10.1016/J.JTRANGE.2021.103178>
- Coenen, A., & Pearce, A. (n.d.). *Understanding UMAP*. Retrieved May 7, 2024, from <https://pair-code.github.io/understanding-umap/>
- Cohen, I. (2018, October 5). *Outliers explained: a quick guide to the different types of outliers | by Ira Cohen | Towards Data Science*. <https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6>
- Cosgun, H. (2023, August 5). *Which data scaling technique should I use ? | by Hasan Hüseyin Coşgun | Medium*. <https://medium.com/@hhuseyincosgun/which-data-scaling-technique-should-i-use-a1615292061e>
- Domènech, A., Mohino, I., & Moya-Gómez, B. (2020). Using Flickr Geotagged Photos to Estimate Visitor Trajectories in World Heritage Cities. *ISPRS International Journal of Geo-Information*, 9(11), 646. <https://doi.org/10.3390/IJGI9110646>
- Francisco, B., Ribeiro, R., Batista, F., & Ferreira, J. (2023). Analysis of the Tourist's Behavior in Lisbon Using Data from a Mobile Operator. *Lecture Notes of the Institute for Computer Sciences, Social- Informatics and Telecommunications Engineering, LNICST, 486 LNICST*, 3–22. https://doi.org/10.1007/978-3-031-30855-0_1
- Frost, J. (2021, March). *Spearman's Correlation Explained - Statistics By Jim*. <https://statisticsbyjim.com/basics/spearmans-correlation/>
- Fu, Y., Liu, G., Ge, Y., Wang, P., Zhu, H., Li, C., & Xiong, H. (2019). Representing Urban Forms: A Collective Learning Model with Heterogeneous Human Mobility Data. *IEEE Transactions on Knowledge and Data Engineering*, 31(3), 535–548. <https://doi.org/10.1109/TKDE.2018.2837027>
- Gefferth, A. (2024, February 15). *One Hot Encoding Explained | Built In*. <https://builtin.com/articles/one-hot-encoding>
- Keertika, V., Taruni, B., Bhavan, R., Begum, M. S., & Ashiya, S. (2021). Big data Analytics based Tourist Behaviour Analysis. *Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021*, 1888–1892. <https://doi.org/10.1109/ICCES51350.2021.9489047>
- Klimek, R. (2019). Towards recognising individual behaviours from pervasive mobile datasets in urban spaces. *Sustainability (Switzerland)*, 11(6). <https://doi.org/10.3390/SU11061563>
- Kumar, A. (2021, November 28). *Elbow Method vs Silhouette Score - Which is Better? - Analytics Yogi*. <https://vitalflux.com/elbow-method-silhouette-score-which-better/>
- Mcleod, S. (2023, July). *Box Plot Explained: Interpretation, Examples, & Comparison*. <https://www.simplypsychology.org/boxplots.html>

- Mendes, L. (2020). Bye Bye Lisbon: Tourism Gentrification Impacts on Lisbon's Inner-City Housing Market. *Handbook of Research on the Impacts, Challenges, and Policy Responses to Overtourism*, 136–155. <https://doi.org/10.4018/978-1-7998-2224-0.CH008>
- Not, E. (2021). Mining mobile application usage data to understand travel planning for attending a large event. *Information Technology & Tourism*, 23(3), 291–325. <https://doi.org/10.1007/S40558-021-00204-7>
- Park, S., Xu, Y., Jiang, L., Chen, Z., & Huang, S. (2020a). Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data. *Annals of Tourism Research*, 84. <https://doi.org/10.1016/J.ANNALS.2020.102973>
- Park, S., Xu, Y., Jiang, L., Chen, Z., & Huang, S. (2020b). Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data. *Annals of Tourism Research*, 84. <https://doi.org/10.1016/J.ANNALS.2020.102973>
- Patrick Grieve. (2024, April 4). *Personalised customer service: what it is and how to provide it*.
- Pereira, M. do C. D. (2023). *Using Flickr to identify and connect tourism Points of Interest: The case of Lisbon, Porto and Faro*. <https://run.unl.pt/handle/10362/149640>
- Philip Stemann. (2024, January 11). *How to Use Public Transport in Lisbon: The Ultimate Guide*.
- Rana, A. (2023, July). *Bar Charts Vs Histograms: A Complete Guide - Venngage*. <https://venngage.com/blog/bar-charts-vs-histograms/>
- Reif, J., & Schmücker, D. (2020). Exploring new ways of visitor tracking using big data sources: Opportunities and limits of passive mobile data for tourism. *Journal of Destination Marketing and Management*, 18. <https://doi.org/10.1016/J.JDMM.2020.100481>
- Sarkar, J. L., & Majumder, A. (2021). A new point-of-interest approach based on multi-itinerary recommendation engine. *Expert Systems with Applications*, 181, 115026. <https://doi.org/10.1016/J.ESWA.2021.115026>
- Sharma, P., Meena, U., & Sharma, G. K. (2021). Application of Data Mining Algorithms for Tourism Industry. *Advances in Intelligent Systems and Computing*, 1172, 481–495. https://doi.org/10.1007/978-981-15-5566-4_43
- Sheikh Hassan, M., Sadeghzadeh, M., & Faghihi, M. (2021). Analysis of Health tourist's Behaviors Using Data Mining Method (International Tourists). *Razavi International Journal of Medicine*, 9(2), 48–55. <https://doi.org/10.30483/RIJM.2021.254173.1027>
- Sun, H., Chen, Y., Lai, J., Wang, Y., & Liu, X. (2021). Identifying Tourists and Locals by K-Means Clustering Method from Mobile Phone Signaling Data. *Journal of Transportation Engineering, Part A: Systems*, 147(10). <https://doi.org/10.1061/JTEPBS.0000580>
- Van der Zee, E., & Bertocchi, D. (2018). Finding patterns in urban tourist behaviour: a social network analysis approach based on TripAdvisor reviews. *Information Technology & Tourism*, 20(1), 153–180. <https://doi.org/10.1007/S40558-018-0128-5>

- Wearing, S. L., & Foley, C. (2017). Understanding the tourist experience of cities. *Annals of Tourism Research*, 65, 97–107. <https://doi.org/10.1016/j.annals.2017.05.007>
- Wei, H. (2020, January 2). *How to measure clustering performances when there are no ground truth?* / by Haitian Wei | Medium. <https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c>
- Wei, X. (2021). Tourism Management System Based on GIS. *Journal of Physics: Conference Series*, 1881(3). <https://doi.org/10.1088/1742-6596/1881/3/032005>
- Why Consider a Flexible Booking Policy for Your Tour Business* | Tourism Tiger. (n.d.). Retrieved June 30, 2024, from <https://www.tourismtiger.com/blog/why-consider-a-flexible-booking-policy-for-your-tour-business/>
- Xu, Y., Xue, J., Park, S., & Yue, Y. (2021). Towards a multidimensional view of tourist mobility patterns in cities: A mobile phone data perspective. *Computers, Environment and Urban Systems*, 86. <https://doi.org/10.1016/J.COMPENVURBSYS.2020.101593>

ANNEXES

ANNEXE 1 - VODAFONE DATA FRAMEWORK

Table 8. Vodafone Data Framework

Framework	Vodafone provided the CML with anonymised and aggregated data on mobile device movements within a 200x200 grid every 5 minutes. This information will help analyse people's flows within Lisbon and the city's entrances and exits.
-----------	---

ANNEX 2 - INFORMATION PROVIDED BY CML

Table 9. Information provided by CML

dicofre	Unique identifier of the parish.
Freguesia	Name of the current parish of the square's centroid.
Freguesias	Names of the former parishes of the square's centroid.
geopoint do centroide	Names of the former parishes of the square's centroid.
geoshape	Establishes relationships in the database (created by NEC).
Grelha	Square size of 200x200.
Grelha_Are	Grid area of 40,000 m ² .
Grelha_Per	Grid perimeter of 800 meters.
grelha_x	X-axis coordinate, starting at 0.
grelha_y	Y-axis coordinate, starting at 0.
ID	Unique identifier for the grid.
Latitude	Latitude of the grid's center (WGS84 coordinate system).
Longitude	Longitude of the grid's centroid (WGS84 coordinate system).
nome	Common name for the area represented by the grid.
WKT	Item in Well-Known Text (WKT) format for software processing.

ANNEX 3 - VODAFONE DATA INTEGRATION METHOD

Table 10. Vodafone Data Integration Method

API Usage	NEC Corporation provides an API that Vodafone uses to deliver files.
Testing and Deployment	A test URL is available, but NEC must provide a final URL for deployment.
Authentication	HTTPS and basic authentication are used to obtain a token for API requests.
Token Usage	The token is used in each API communication after authentication.
Communication	Synchronous communication with success or failure return codes upon file receipt.
Parallel Invocation	The API can be invoked multiple times in parallel to avoid dependencies.
File Hashing	Each file delivery includes a hash (type to be defined) for NEC to validate receipt.
Error Handling	If files can't be delivered on time, none should be lost. Files will be sent from latest to oldest once the connection is reestablished.
Data Integration	Delayed files will be integrated based on the DATETIME field, not the insertion time.
Technical Solution	Vodafone and NEC will discuss solutions for delayed data, prioritizing API delivery but considering alternative methods for larger delays.

ANNEX 4 - VODAFONE DATASET INDICATORS

Table 11. Vodafone Dataset Indicators

C.1 Number of distinct terminals in the grid
C.2 Number of distinct terminals, roaming, in the grid
C.3 Number of distinct terminals that remained in the grid
C.4 Number of distinct terminals that remained in the grid, roaming
C.5 Number of different terminal inputs in the grid
C.6 Number of outputs of distinct terminals in the grid
C.7 Number of different terminal entries, roaming, in the grid
C.8 Number of different terminal outputs, roaming, in the grid
C.9 Number of distinct terminals with active data connection, in the grid
C.10 Number of separate terminals with active data connection, roaming, in the grid
C.11 Number of voice calls from the grid
C.12 Number of entries in Lisbon for the 10 main roads
C.13 Number of departures from Lisbon for the 10 main roads
D.1 Top 10 countries of origin of roaming terminal devices
E.1 Number of voice calls that ended in the grid
E.2 Average downstream rhythm of the grid, in the grid
E.3 Average grid upstream rhythm
E.4 Quad downstream peak rhythm
E.5 Quadruple peak rhythm
E.6 Top 10 Applications (Text with names separated by;)
E.7 Duration of minimum stay within the grid
E.8 Duration of the average stay in the grid
E.9 Duration of maximum stay within the grid
E.10 Number of devices that share the connection in the grid



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa