NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Mobility patterns of docked bikes vs cars in Lisbon

Assessing its impact in air quality during event days

Guilherme Lopes Henriques

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

Mobility patterns of docked bikes vs cars in Lisbon
Assessing its impact in air quality during event days

by

Guilherme Lopes Henriques

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Supervisor: Miguel de Castro Neto, PhD, NOVA Information Management School
Co-supervisor: Bruno Jardim, PhD, NOVA Information Management School

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.

[Lisbon, July 2024]

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, my co-supervisor for their invaluable guidance, support, and encouragement throughout this research. Their insights and expertise have been instrumental in shaping this thesis.

I am also profoundly grateful to my family and friends for their constant encouragement and unwavering support. Their belief in my abilities has been a source of motivation and strength during this journey.

ABSTRACT

The importance of sustainable mobility in large urban areas has become increasingly crucial for improving the well-being of city residents. This study will focus on the urban area of Lisbon exploring the impact of major events, how the air quality is affected by those events, how the traffic is affected and how the usage of bicycle sharing systems change in these days compared to the days when there are no events.

Three hypotheses were specified to be examined using the difference-in-differences model. These hypotheses are as follows: H1 – Air quality in some zones is affected by being an event day; H2 – Being an event day increases the number of jams in the peripheral zones of the event; H3 – Being an event day increases the use of bike-sharing systems.

After applying the statistical model and obtaining the results, it can be concluded that in relation to H1 in several zones an event does not worsen air quality. Regarding H2, we can say that holding an event increases the number of traffic jams in the vicinity and about H3, we can also say that holding an event increases the number of Bicycle-sharing systems used in the vicinity of the event. With these findings, which are important, it is hoped that they can help decision-makers to establish policies in order to reduce the number of traffic jams and increase the quality of life of the inhabitants in the vicinity of the event, as well as the experience of the people attending the events themselves.

KEYWORDS

Bike-sharing systems; Sustainable mobility; Air quality; Mobility patterns; Difference-in-differences

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

GPS	Global position system
LEZ	Low emission zone
PM	Particulate matter
NO2	Nitrogen Dioxide
HSR	High speed rail
CO2	Carbon Dioxide
CO	Carbon Monoxide
DID	Difference-in-differences
PT	Parallel trends
BSS	Bicycle sharing systems
BMC	Bicycle mobility chain

1. INTRODUCTION

1.1. CONTEXT

A docked bike sharing service operates through a network of docking stations spread across a city, providing users with access to bicycles for short-term use. Users can locate, unlock, and use a bike from one station, ride it to their destination, and then return it to any available docking station. These stations have docks where bikes can be locked and made available for the next user. Users unlock the bikes by using a smartphone app. Due to their extensive utilization and accessibility in numerous areas, bicycles have emerged as a viable alternative to automobiles for short trips. Selecting this form of transportation allows individuals to engage in physical activity while avoiding traffic congestion, a prevalent problem in metropolitan regions. Noteworthy for its eco-friendly attributes, the bicycle fulfils all criteria to encourage its escalating adoption.

Cars have a significant impact on air quality in the city of Lisbon. The presence of cars on the road contributes to the emission of nitrogen oxides and carbon dioxide, which are air pollutants that can have bad effects on air quality (SIlveira et al., 2019). For that reason, reducing car usage and promoting the use of electric and autonomous vehicles can help reduce the negative impact of cars on air quality in Lisbon (Fontes et al., 2022).

The urban environment is going through a significant transformation, conducted by the demands of the modern world, environmental concerns, and the need for sustainable transportation options. In this era of rapid urbanization, cities around the globe are fighting with challenges related to congestion, pollution, and the promotion of healthier and more sustainable modes of mobility. The city of Lisbon in Portugal is no exception to these trends.

1.2. OBJECTIVES

This research compares the mobility patterns of vehicles and docked bikes during event days in Lisbon and examines how each affects air quality. Like many other cities, Lisbon regularly holds a lot of different types of events but in this study only football games will be analyzed. Like in other cities, in Lisbon these events can have a significant impact on the city's traffic patterns and air quality (Olabarrieta & Laña, 2020). It is necessary to comprehend the decisions that citizens make during those events in order to make decisions that will improve urban sustainability.

The decision of comparing docked bikes and cars as modes of transportation during events in Lisbon is particularly significant in the context of the need for sustainable urban mobility. Docked bike-sharing systems have gained popularity in many cities all over the world as a sustainable and eco-friendly mean of transportation, while the use of private cars is over and over more associated with traffic congestion and air pollution, factors that reduce the quality of life of communities (Albuquerque et al., 2021; Zakiah et al., 2019). This study aims to give better insights into how people's transportation choices are influenced by the context of an event, and how these choices impact air quality during event days.

To achieve this goal, this research will use a multi-faceted approach, incorporating data extracted from traffic and environmental monitoring to provide a comprehensive view of the mobility patterns and air quality in Lisbon during event days. The findings from this research will not only contribute to the academic understanding of urban mobility but will also provide insights for city planners and policy makers who seek to promote sustainable and environmentally responsible transportation in the urban context.

In conclusion, the analysis of mobility patterns during event days in Lisbon is a significant move in the direction of understanding the connection between events, transportation choices, and air quality in a city like Lisbon. By comparing the use of docked bikes and cars, we can better understand the potential for sustainable transportation options to reduce the effects of events on air quality and community quality of life, contributing to the creation of healthier, more liveable cities.

After the literature review, that explores most relevant articles, this thesis ensues with presenting the methodology where context and datasets used are explained in detail. Subsequently, all data was treated to be prepared to use in the following sections. Following this we have the modelling section, that provides a detailed explanation to the model to be utilized, its functioning and the hypotheses outlined in the study. The results section shows two types of analyses, a visual and a statistical, for each of the hypotheses under study. Finally, the last section unfolds the discussion, conclusions, limitations and possible future work.

2. LITERATURE REVIEW

In this chapter we address work related to cars mobility patterns, air quality and bike sharing systems. We're going to see how the studies about these three topics changed over time and different types of approaches for all the topics.

2.1. Analysis of Traffic patterns

Urban transportation research has experienced substantial advancements in comprehending travel behaviour, mobility patterns, and the impacts of transportation systems on urban life quality. Huang et al. (2019) present an important study in this domain by concentrating on the travel behaviour of private car users. Leveraging a large-scale private car trajectory dataset, they employ sophisticated methodologies, including stop-and-wait analysis, DBSCAN clustering, and Markov chain modelling. Their study introduces the concept of spatial-temporal entropy rate, providing quantitative measures of individual private cars' mobility regularity. This research not only sheds light on individual travel patterns but also contributes valuable insights for traffic management and urban planning.

The study conducted by Pappalardo et al. (2013) investigated the travel patterns of cars by utilizing a comprehensive dataset of GPS trajectories from Italy. Their analysis confirms existing mobility models while revealing new insights into the dynamics of car travel. The precise estimation of the model in the study, which is based on GPS data and road sensors, improves comprehension in mobility modelling and traffic estimation, thereby contributing to more efficient urban transportation planning.

Xiao et al. (2020) examines the typical travel behaviour of private car drivers in urban environments through the analysis of trajectory data. Their innovative approach constructs a matrix of trajectory similarity by utilizing temporal-spatial distance measurement and uses Kernel Principal Component Analysis for reducing features. The study's method of classification, which is based on transfer learning, effectively identifies the regularity of travel from trajectory data that is unlabelled, offering practical applications in predicting destinations and planning routes.

Lv et al. (2021) examine the problem of private car commuting patterns using Electronic Registration Identification data. Their methodology, involves measuring spatiotemporal similarity and clustering based on spatial-temporal factors, effectively identifies private cars used for commuting, providing valuable insights into commute times and the relationship between residence and workplace. This research highlights the significance of comprehending spatiotemporal characteristics in order to manage transportation effectively.

Similarly, Zhu et al. (2021) delve into the field of human mobility and travel behaviour in private cars. Through their analysis of trajectory data from private cars, they discover important information about mobility patterns and areas with high levels of activity, highlighting the importance of understanding travel behaviour for urban planning and transportation management.

Addressing real-world challenges, Zulkifli et al. (2018) examine the impact of traffic congestion on the quality of life in the Seri Kembangan community. Their study focuses on the negative effects of congestion on daily routines and lifestyles, promoting effective traffic planning and policy implementation to improve transportation systems and improve the quality of life in the community.

Olabarrieta and Laña (2020) introduce a methodology to assess the impact of major events, in particular soccer matches in Madrid, on urban traffic. By investigating traffic dynamics during such events, the authors offer insights into managing traffic flow and congestion associated with sporting events, concerts, and festivals. Their findings reveal a notable decrease in traffic impact with distance from the stadium, noticing the importance of understanding spatial dynamics for effective urban traffic management during special events.

Currie and Walker (2011) investigated the impact of E-ZPass technology on traffic congestion and infant health outcomes in New Jersey and Pennsylvania. By doing a natural experiment, they found that E-ZPass implementation significantly reduced traffic delays improving air quality and reducing pollution levels. The study used a difference-in-differences approach and regression models to compare areas near toll plazas with those farther away, revealing a reduction in prematurity and a decrease in low birth weight. These findings highlight the health benefits of traffic congestion mitigation policies, accenting their importance not only for travel efficiency but also for public health. The research underscores the need for evidence-based policymaking and urban planning to address traffic-related pollution.

These studies collectively underscore the multifaceted nature of urban transportation research, highlighting the importance of understanding travel behaviour, mobility patterns, and the impact of transportation systems on the quality of urban life. Below is a structured summary of the articles utilized in this section of the literature review (Table 1).

Table 1 – Overview of the “Analysis of traffic patterns” studies

Title	**Authors**	**Year**
“Exploring individual travel patterns across private car trajectory data.”	Huang, Y., Xiao, Z., Wang, D., Jiang, H., & Wu, D.	2019
“Understanding the patterns of car travel.”	Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., & Giannotti, F.	2013
“Exploring human mobility patterns and travel behavior: A focus on private cars. ”	Xiao, Z., Xu, S., Li, T., Jiang, H., Zhang, R., Regan, A. C., & Chen, H	2020
“Understanding Travel Patterns of Commuting Private Cars using Big data of Electronic Registration Identification of Vehicles.”	Lv, J., Zheng, L., Ye, Y., & Ye, C.	2021
“Exploring human mobility patterns and travel behaviour: A focus on private cars.”	Xiao, Z., Xiao, H., Jiang, H., Chen, W., Chen, H., & Regan, A. C.	2021
“The effect of traffic congestion on quality of community life”	Zulkifli, C. N. A. B., & Ponrahono, Z.	2018
“Effect of soccer games on traffic, study case: Madrid.”	Olabarrieta, I. I., & Laña, I.	2020
“Traffic congestion and infant health: Evidence from E-ZPass. ”	Currie, J., & Walker, R.	2011

2.2. Air Quality

Air quality and urban transportation are connected domains that impact public health, environmental sustainability, and quality of life in urban areas. The increasing usage of cars in urban transportation significantly contributes to air pollution, creating serious health risks and environmental challenges. As cities continue to struggle with rapidly growing populations and car traffic, understanding the dynamics of air pollution becomes very important in creating effective strategies to reduce its adverse effects.

Sarroeira et al. (2023) conducted a comprehensive study in the municipality of Lisbon to understand the dynamics of air pollution. Their data science approach utilized data from 80 monitoring stations to analyze the evolution of pollutants such as carbon monoxide, nitrogen dioxide, and particulate matter. The study revealed alarming levels of nitrogen dioxide and particulate matter 10, exceeding World Health Organization recommendations.

Santos et al. (2019) investigated the impact of a Low Emission Zone on air quality in Lisbon from 2009 to 2016. Their findings indicated significant improvements in PM10 and NO2 concentrations within the LEZ zones, emphasizing the effectiveness of LEZ policies in reducing exposure to air pollutants. However, insignificant reductions in NOx and PM2.5 levels highlight the need for stricter standards and complementary measures to reduce population exposure.

Silveira et al. (2019) evaluated the impact of traffic management scenarios on air quality and health outcomes in a Portuguese urban area. Their modeling system demonstrated reductions in daily maximum NO2 concentrations, underlining the effectiveness of targeted traffic management strategies, particularly within Low Emission Zones, for improving air quality and public health.

Lopes et al. (2019) focused on assessing urban mobility strategies to reduce particulate matter concentrations along a main avenue near Lisbon, Portugal. Despite efforts to mitigate emissions from road transport, particulate matter concentrations continue to exceed European air quality limits. Their study utilized an air quality modeling system to evaluate proposed strategies, highlighting the importance of further research to identify effective mobility strategies for improving air quality.

Taborda et al. (2020) presented a platform for city councils to monitor air quality using low-cost portable sensors. Their interactive dashboard integrates air pollution data with environmental context, helping decision-makers in understanding air quality issues and promoting urban sustainability.

The investigation conducted by Jia et al. (2021) explores into the consequences of high-speed rail on CO2 emissions within urban areas of China through the utilization of a difference-in-differences (DID) methodology. Through the examination of an extensive dataset encompassing urban emissions and the progression of HSR, the study evaluates the way the initiation and enlargement of HSR networks impact the carbon footprint of cities. The outcomes indicate that HSR plays a significant role in the reduction of CO2 emissions by presenting a more environmentally friendly option compared to both cars and airplanes, particularly demonstrating more pronounced effects in larger, more polluted urban centers. Moreover, the study highlights the presence of spatial spillover effects, wherein neighboring cities experience positive outcomes in terms of reduced emissions. Consequently, these findings imply that the expansion of HSR infrastructure could serve as a strategic element within urban policies aimed at curbing CO2 emissions.

Ban and Kedagni (2022) develop a generalized difference-in-differences (DID) framework to improve the identification of average treatment effects on the treated (ATT) under the parallel trends assumption. Instead of relying solely on pre-treatment period trends, their method uses multiple data sources and pre-treatment periods. They reinterpret PT using selection bias, assuming post-treatment selection bias lies within pre-treatment biases. This approach defines an information set that includes multiple pre-treatment periods or baseline covariates, ensuring the true ATT is captured within the identified set. They also propose criteria for selection biases from a policymaker's perspective to achieve precise ATT identification. Below is a structured summary of the articles utilized in this section of the literature review (Table 2).

Table 2 – Overview of the “air quality” studies

Title	Authors	Year
“Monitoring Sensors for Urban Air Quality: The Case of the Municipality of Lisbon.”	Sarroeira, R., Henriques, J., Sousa, A. M., Ferreira da Silva, C., Nunes, N., Moro, S., & Botelho, M. D. C.	2023
“Impact of the implementation of Lisbon low emission zone on air quality.”	Santos, F. M., Gómez-Losada, Á., & Pires, J. C.	2019
“Impact of traffic management strategies on air quality and health in a Portuguese urban area.”	Silveira, C., Ferreira, J., Lopes, D., & Miranda, A. I.	2019
“Urban mobility strategies to improve local air quality: case study of Lisbon, Portugal.”	LOPES, D., FERREIRA, J., RAFAEL, S., BAPTISTA, P., FARIA, M., CANHA, N., & ALMEIDA-SILVA, M. A. R. I. N. A.	2019
“Exploring air quality using a multiple spatial resolution dashboard—a case study in Lisbon.”	Taborda, R., Datia, N., Pato, M. P. M., & Pires, J. M.	2020
“High-speed rail and CO2 emissions in urban China: A spatial difference-in-differences approach.”	Jia, R., Shao, S., & Yang, L.	2021
“Generalized difference-in-differences models: Robust bounds.”	Ban, K., & Kedagni, D.	2022

2.3. Bike Sharing Systems

In recent years, the increasing usage of bike-sharing systems has emerged as a promising solution to mitigate urban transportation emissions and improve air quality.

This segment of the literature review synthesizes the findings of several studies on bike-sharing systems, with numerous studies examining their impact on travel behaviour, congestion reduction, and environmental sustainability. Fishman et al. (2014) conducted a comprehensive investigation into the effectiveness of bike share programs in reducing car trips across urban areas. Analysing data from programs in several cities, such as Melbourne, Brisbane, Washington, D.C., London, and Minneapolis/St. Paul, the study revealed significant reductions in motor vehicle use in Melbourne and Minneapolis/St. Paul, indicating the potential of bike-sharing initiatives to reduce car dependency. Contrarily, London's program experienced an increase in motor vehicle use. This study underscores the importance of context-specific analyses to evaluate the efficacy of bike-sharing programs in different urban environments.

Feng et al. (2017) contributed valuable insights into the operational dynamics of bike-sharing stations through clustering algorithms, focusing on the Vélib’ bike-sharing system in Paris. Their analysis discovered patterns within stations, offering critical guidance for system control and redesign to increase operational efficiency and user experience. The findings of the study contributed to a wider discourse on sustainable urban mobility solutions, emphasizing the importance of data-driven strategies in optimizing bike-sharing systems.

Raposo and Silva (2022) directed attention to the environmental impact and usage patterns of the e-bike sharing system GIRA in Lisbon, offering a detailed examination of the system's performance and sustainability. By analysing trip data and weather factors, the study elucidated daily trip numbers, travel times, usage peak trends, and seasonality patterns, while also estimating the potential impact on energy consumption and greenhouse gas emissions. These insights provide valuable information for policymakers and urban planners that are trying to promote sustainable transportation alternatives in urban environments.

Kou and Cai (2018) studied travel patterns within bike-sharing systems across eight United States cities, providing insight on trip distance and duration distributions for both leisure and touristic purposes. Their findings revealed distinct distributions influenced by system size and geographical constraints, with larger systems exhibiting lognormal distributions while smaller systems displayed varied patterns. Furthermore, for longer trips in larger systems, the study revealed a power law decay in trip length and distance, providing insights into the spatial dynamics of bike-sharing travel behaviour.

Li et al. (2021) examined the spatial spillover effects of shared mobility on urban traffic congestion in 94 Chinese cities, employing spatial econometric models to clear up complex congestion patterns influenced by shared mobility, economic development, population density, and public transportation. The study reinforced the subtle relationship between bike-sharing and congestion, highlighting spatial heterogeneity in congestion relief across different regions. By elucidating these dynamics, the research contributes to a deeper understanding of shared mobility's impact on urban congestion, informing evidence-based policymaking and urban planning efforts.

Albuquerque et al. (2021) adopted a data-driven approach to understanding BSS mobility patterns in Lisbon, using a CRISP-DM data mining methodology to analyse large datasets encompassing spatiotemporal trip distribution and weather factors. Despite the challenges created by the COVID-19 pandemic, the study revealed a significant growth in bike sharing systems trip counts, reflecting increasing demand for sustainable transportation options. These findings underscore the adaptive nature of urban mobility systems and the importance of data-driven insights in shaping responsive transportation policies.

Fontes et al. (2022) addressed urban mobility challenges in Lisbon by developing a decision support system for bike-sharing docking stations, aiming to identify soft mobility hotspots and optimize mobility management systems. By leveraging clustering methods and GSM data, the study gives an innovative approach to urban mobility infrastructure, aligning with global efforts toward sustainable transportation alternatives.

Ma et al. (2019) explored the spatiotemporal activity patterns of a bicycle sharing system in Ningbo, China, employing clustering analysis to classify docking stations into distinct clusters based on temporal and spatial attributes. Their research enhances bicycle sharing operators' understanding of system usage and offers insights into travel habits and land use characteristics, facilitating improvements in service quality and system optimization.

Xin et al. (2023) introduced the concept of the bike mobility chain (BMC) as a novel perspective for understanding bike-sharing dynamics, focusing on trajectory-based research and mobility pattern discovery. By reconstructing bike-sharing trip data and analysing BMCs, the study sheds light on locality patterns and disparities in mobility statistics, offering empirical references for urban bike management and enhancing the analytical perspective in urban transport research.

In conclusion, these studies collectively contribute to the evolving discourse on bike-sharing systems and urban mobility, offering valuable insights into travel behaviour, system dynamics, environmental impact, and policy implications. By resuming findings from diverse research endeavours, this literature review provides an understanding of the multifaceted dimensions of bike-sharing systems, informing evidence-based strategies for sustainable urban transportation planning and development. Below is a structured summary of the articles utilized in this section of the literature review (Table 3).

Table 3 – Overview of the "Bike Sharing Systems" studies

Title	**Authors**	**Year**
"Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. "	Fishman, E., Washington, S., & Haworth, N.	2014
"Analysis of bike sharing system by clustering: the Vélib'case"	Feng, Y., Affonso, R. C., & Zolghadri, M.	2017
"City-Level E-Bike Sharing System Impact on Final Energy Consumption and GHG Emissions. "	Raposo, M., & Silva, C.	2022
Understanding bike sharing travel patterns: An analysis of trip data from eight cities	Kou, Z., & Cai, H.	2019

“The Spatial Effect of Shared Mobility on Urban Traffic Congestion: Evidence from Chinese Cities. Sustainability.”	Li, J., Ma, M., Xia, X., & Ren, W	2021
“Smart cities: Data-driven solutions to understand disruptive problems in transportation—The Lisbon Case Study.”	Albuquerque, V., Oliveira, A., Barbosa, J. L., Rodrigues, R. S., Andrade, F., Dias, M. S., & Ferreira, J. C.	2021
“A cluster-based approach using smartphone data for bike-sharing docking stations identification: Lisbon case study. “	Fontes, T., Arantes, M., Figueiredo, P. V., & Novais, P.	2022
“Spatiotemporal clustering analysis of bicycle sharing system with data mining approach.”	Ma, X., Cao, R., & Jin, Y.	2019
“Spatiotemporal analysis of bike mobility chain: A new perspective on mobility pattern discovery in urban bike-sharing system.”	Xin, R., Yang, J., Ai, B., Ding, L., Li, T., & Zhu, R.	2023

3. METHODOLOGY

The Cross-Industry Standard Process for Data Mining – CRISP-DM (Shearer, 2000) is a widely used methodology for guiding data mining and analytics projects. Developed by a consortium of industry experts, CRISP-DM provides a structured approach to solving complex data problems, ensuring that projects are well-defined, executed efficiently, and produce actionable results. The steps of this methodology are displayed in the figure below (Figure 1).

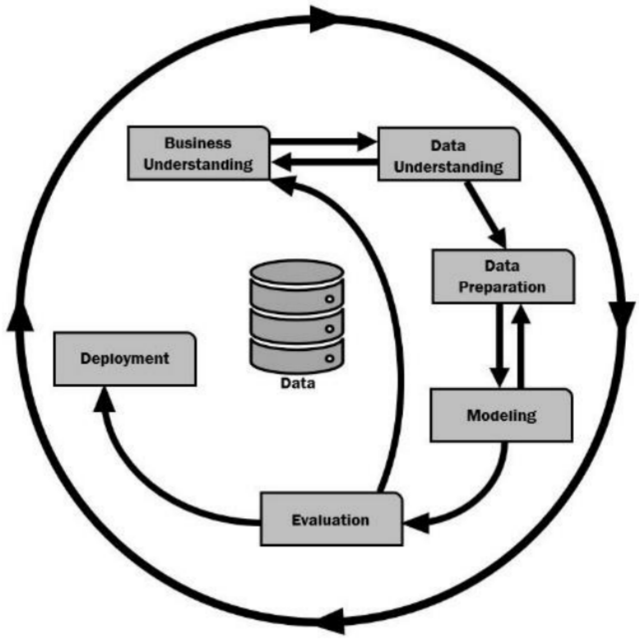

Figure 1- Steps of Crisp-Dm. Taken from Chapman, et al. (2000).

It's important to note that CRISP-DM is an iterative process, meaning that the phases are not strictly sequential. Instead, it's common for researchers to revisit earlier stages as new insights are gained or as project requirements evolve. This iterative approach allows for flexibility and continuous improvement throughout the process (Saltz, 2021).

In this study, the CRISP-DM methodology is applied to address the research objectives outlined in Chapter 1. Each phase of CRISP-DM is executed systematically, with careful consideration given to the specific requirements of the research problem. The following chapters detail the implementation of CRISP-DM in this thesis, including the techniques and tools used at each stage.

4. DATA UNDERSTANDING

4.1. Study Area

The area for this study is the municipality of Lisbon (Figure 2), more specifically the areas around the “Estádio da Luz” (Figure 3) and “Estádio de José de Alvalade” (Figure 4) to measure how the normal traffic behaviour is affected by the games of Sport Lisboa e Benfica which has the “Estádio da Luz” as its stadium and Sporting Clube de Portugal, who plays at “Estadio José de Alvalade”.

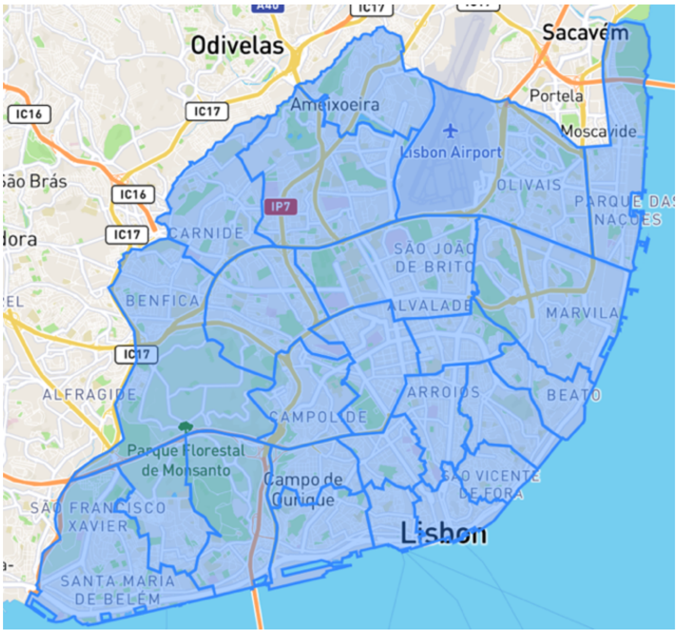

Figure 2 - Area of the municipality of Lisbon. Taken from https://withportugal.com/

Figure 3 - Surroundings of "Estádio da Luz". Taken from https://www.google.com/maps

Figure 4 - Surroundings of "Estádio José de Alvalade". Taken from https://www.google.com/maps

4.2. DATA

In order to make the analysis, we gathered information from four datasets, two related with the air quality sensors, one related with Bike Sharing Systems and the other one related with the traffic information.

4.2.1. Sensors Data

The sensor data was supplied by an open data platform from Lisbon[1].This data is very important as it will allow me to analyse throughout the study the impact that the other elements studied have on air quality.

This part of the data is made up of two datasets, one with sensors data (Table 4) with 734 lines and 12 columns, where each line represents a different sensor, and which contains general information about these sensors. The main reason and need for using this file is to allow the selection of sensors to use, depending on the type of measurement the sensor makes and its location.

[1] https://lisboaaberta.cm-lisboa.pt/index.php/pt/

Table 4 – Structure of sensor data

Field name	Description	Data type
id	row unique identifier	Numeric
address	address of the location	object
geometry	Geo location of the station	object
lat	latitude coordinate of the sensor's location	float64
lon	longitude coordinate of the sensor's location	float64
id_sensor	sensor unique identifier	object
unit	unit of measurement	object
indicator_id	identifier for the type of environmental indicator being measured	object
measure_id	identifier for the measure	object
location_id	identifier for the location	numeric
measure_description	description of the measure	object
indicator_description	description of the indicator	object

The other part of the data related with the sensors is the measurements (Table 5) with 49 million lines and 5 columns. This file represents all measurements from the 734 sensors referred above, basically this file only has the identification of the sensor, the value measured and the date when the measure occurred.

Table 5 – Structure of measurements from sensors

Field name	Description	Data type
id	row unique identifier	Numeric
id_sensor	sensor unique identifier	Object
value	value measured by the sensor	Float64
date	measurement date	Object
hour	measurement hour	Datetime

Regarding the sensor, we used the data from Table 5 to create a line chart that would give us a visualization covering all hours of a day and the corresponding average CO concentration for each hour. Within those, we chose to analyse Carbon Monoxide, because it is one of the primary pollutant gases produced by cars (Chamberlain, 2016). Unsurprisingly, as stated below in Figure 5, the times of the day when there was the highest concentration of Carbon Monoxide captured by the sensors were at 8:00h, 9:00h, 17:00h, 18:00h and 19:00h, the times when traffic in Lisbon is at its heaviest. In terms of CO, a gas produced by burning fossil fuels, the highest concentration is at 17:00h and has a value of 0.196 µg/m3.

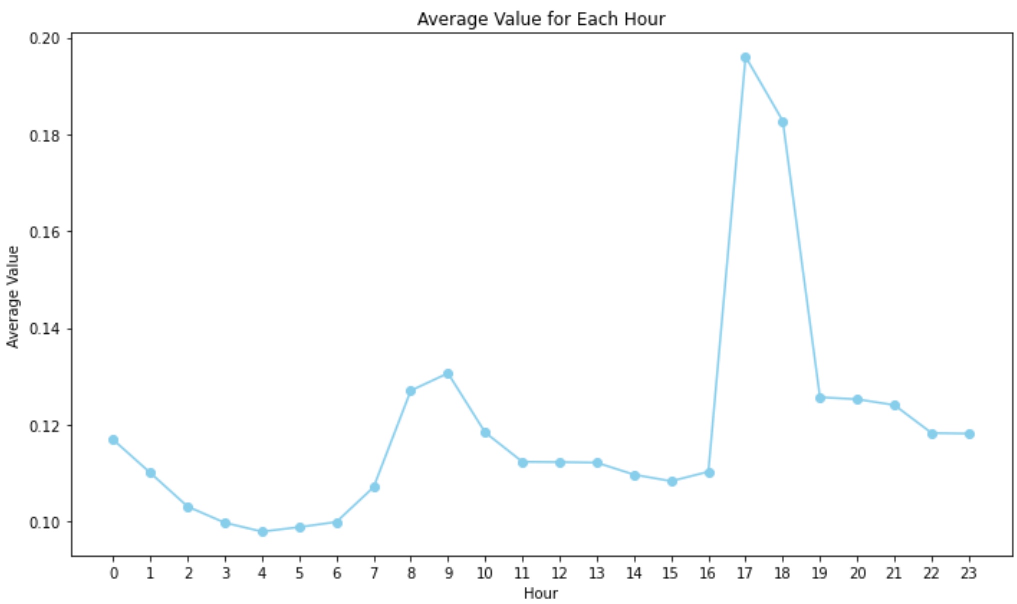

Figure 5 - Line chart with average CO value for hour

4.2.2. Bike Sharing Systems data

This data was obtained in “Câmara Municipal de Lisboa“ website, more specifically at “Lisboa Aberta”[2]. This is a portal for Lisbon's open data made available by the City Council and everyone can access this data. The file has 1.5 million rows and 7 columns, and it provides information on the hours and the day that a bicycle passed by a certain location. This information allows us to understand the mobility patterns of bicycles on the selected days and times.

Table 6 – Structure of dataset containing BSS data

Field name	Description	Data type
detectionId	row unique identifier	Numeric
tenantIdentifier	Tenant unique identifier	Object
locationID	Location of the sensor identifier	Numeric
detected	Hour and day of the detection	Numeric
direction	Direction of the bycicle	Numeric

4.2.3. Traffic data

To analyse the mobility patterns of private cars and traffic on the days of the events mentioned above, we used data from the well-known GPS Waze, this specific data can be found in Emel’s open data portal[3], provided by Waze. This dataset (Table 7) was composed of 1,048,576 rows and 15 columns. This data allowed us to understand the real impact of the events on traffic, through an examination of the traffic patterns within the selected zones and dates, focusing on certain parameters like "street", "startNode", "endNode" and "start_time".

[2] https://lisboaaberta.cm-lisboa.pt
[3] https://opendata.emel.pt

Table 7 – Structure of the dataset related with car traffic

Field name	Description	Data type
id	row unique identifier	Numeric
country	country where the jam was recorded	object
city	city where the jam was recorded	object
level	level of congestion, it can be 1,2,3,4,5	numeric
speedKMH	average speed in kilometres per hour that vehicles recorded before the congestion was recorded	numeric
length	length of congestion in metres	numeric
type	type of congestion, it can be road works, accidents, etc.	numeric
endNode	node where the congestion ended	object
speed	average speed in kilometres per hour that vehicles recorded after the congestion was recorded	numeric
roadType	type of road (1-7)	numeric
delay	delay in seconds that the congestion caused	numeric
street	street where the congestion occurred	object
startNode	node where the congestion started	object
geometry	Geo location of the congestion	object
start_time	Date and hour when the congestion started	object

5. DATA PREPARATION

Before proceeding to the analysis stage, it is crucial to address all the quality issues found in the data, such as missing values, duplicates, and other inconsistencies, ensuring the date is adequate and perceptible.

In the first dataset, which is detailed in Table 4, various types of sensors were included. However, since the focus was specifically on air quality sensors, any sensors without the letters "QA" in their "indication_id" column were excluded from the dataset. "QA" stands for air quality, and thus, only sensors relevant to this aspect were retained. This filtering process was the only modification made to this dataset. The goal was to identify and retain only the air quality sensors to determine which ones were located in relevant areas for the study. The "id" of these selected sensors was then saved to be used in conjunction with another dataset related to these sensors (Table 5).

For the same reason mentioned above, in the dataset regarding the measurements from the sensors (Table 5) we started by dropping all the rows that didn't have "QA" in the "id" column, to keep only the sensors that had air quality measures. With this step the dataset reduced from 47 million rows to 23.5 million rows. After this, the dataset was reduced by approximately half, and the duplicated values underwent scrutiny. In order to determine duplication, it was essential for all rows to be identical. Subsequently, any rows found to be duplicates were eliminated. With the duplicate checked, 17.3 million rows were eliminated from the dataset, remaining 6.1 million.

At this point, the dataset contained solely one measurement from each sensor per hour. Throughout the dataset, no values were found to be missing. While certain hours and days lacked records, the entries that were presented were complete without missing values. As stated previously, only 1 pollutant gas will be analysed, which is Carbon Monoxide the others have been removed from the dataset with the "drop" function.

As said before, the analysis will be based on two different zones, so the dataset was spilt in two, one for the peripheral zone of the "Estádio da Luz" made up of measurements from the three closest sensors to it and one for the area around the "Estádio José de Alvalade", which includes 5 sensors close to the stadium. This method will allow us to analyse each case in an easier and more organised way.

Regarding Bike sharing systems data, as said above only the locations in the surroundings of the stadiums where of interest, so all the rows that didn't have the location id "3" that is related to a sensor in "Telheiras" zone, id "13" at "Jardim do Campo Grande", id "8" at "Avenida Lusíada" and id"2" at "Avenida do Colégio Militar" were eliminated from the dataset.

Still in this dataset, the columns "detectionID" and "tenantIdentifier" were dropped because they were irrelevant for the study. The dataset was reduced, instead of having for example a hundred rows for the same day for location, it will have one row per day and per location id with a new column named "count" that have the number of detections registered in the sensor. This transformation will allow us to analyse how the number of detections change in game days compared to non-game days, this dataset didn't require much treatment because, in general, it was already very clean, so there wasn't much more to do to improve the quality of the data.

Regarding the Waze data, we started by removing the lines where the "city" column wasn't "Lisboa", as there was only interest in analysing Lisbon. About the "type" column, as it was always "none" and didn't add any information, the column was removed. As far as the "street" column is concerned, only the streets that are actually in the "Estádio da Luz" or "Estádio de Alvalade" peripheral areas are of interest, all the other streets were removed as they were of no interest to the analysis. Just like in the dataset regarding Bike Sharing System, a new column titled "count" was introduced to represent the total count of traffic jams occurring per day. Unlike the previous format where each traffic jam had a separate row, the current dataset consolidates all jams into one row per day, providing a count of jams for that specific day. This dataset was also split in two, one with the streets regarding "Estádio da Luz" and the other one with the streets regarding "Estádio José de Alvalade".

6. MODELLING

The statistical model Difference-in-Differences – DiD (Card & Krueger, 2000) was selected to assess the genuine impact of the above-mentioned events on traffic and subsequently on air quality. An overview of how this model operates and its application to address a portion of the problem examined in this thesis will be presented.

The DiD model, is a robust statistical method utilized to evaluate causal effects in observational studies, especially in situations where randomized controlled trials are not viable or ethical. It examines the variations in outcomes over time among two or more groups, with one group undergoing a treatment or intervention, while the others act as a control. The functionality of the DiD model can be delineated into three key steps, with the initial and pivotal step involving the identification of the treatment and control groups. It is imperative that these groups possess similarity in all aspects except for the treatment under scrutiny. The treatment group is subjected to the intervention, whereas the control group is not. Subsequently, employing a linear regression model, DiD scrutinizes the outcome discrepancies between the treatment and control groups. By evaluating the outcome changes over time across the groups, difference-in-differences endeavours to pinpoint time-invariant distinctions and shared trends influencing all groups. The estimation of the treatment effect is carried out by analysing the coefficient linked to the interaction term between the treatment and post-treatment indicators within the regression model. This coefficient captures the differential alteration in outcomes between the treatment and control groups subsequent to the introduction of the treatment, giving an approximation of the causal impact (Ban & Kedagni, 2022).

Overall, the use of Difference-in-differences model is the optimal choice for analysing the impact of game days on air quality, bike sharing systems usage and traffic congestion in the locations near the events because of its ability to estimate causal effects by comparing changes over time between event days and non-event days. The robustness and simplicity of the DiD model ensure that the insights collected are reliable and relevant.

In order to make the analysis as effective as possible, the treatment group will be the days when selected events are held; the control group, days when there are no events; and outcome variables will be the air quality measures for both groups.

The regression model that will be used in this thesis is:

$$\text{Y}_{it} = \beta_0 + \beta_1 Eventt + \beta_2 Control\ Variablesit + \epsilon_{it}$$

- Y_{it} is the outcome variable for day *t.*
- $Event_t$ is an indicator variable for whether it's a event day.
- Control $Variables_{it}$ represent other variables that may influence the outcome, such as weather conditions, day of the week, etc.
- β_0 is the intercept, representing the variable on days without events.
- β_1 captures the average effect of football match days on the outcome.
- β_2 represents the average effect of control variables on the outcome.
- ϵ_{it} is the error term.

This regression model allows to estimate the impact of event days on the outcome variable, while keeping an eye on for any changes over time and other factors that possibly could influence the outcome. In obtain clarified outcomes, adjustments were required for each of the air quality datasets. So had to get back to the “Data Preparation” phase of Crisp-Dm to make those adjustments (Iterative Nature). Within the two datasets associated with football games, a new column titled "is_gameday" was introduced. This column would indicate the value "1" if the dataset's team had a home match and "0" if it was a regular day without a match.

A new column was introduced in all datasets to specify whether order to conduct the model run and the date in each entry corresponded to a weekday or weekend. This was done considering that traffic tends to be more congested on weekdays compared to weekends.

As previously mentioned, the datasets were segregated based on location, with sensors positioned near the areas where events occur. However, the objective of this analysis is to determine if certain zones experience changes in air quality due to nearby events. Consequently, a distinct model will be developed for each sensor. This approach will facilitate the identification of zones that may require immediate attention to implement new strategies.

Following these modifications, the datasets are now equipped with all the necessary components to execute and generate outcomes.

In the results section of this thesis, 3 hypotheses will be tested using this model, with the aim of making the analysis clearer and more precise. The three hypotheses are:

H0: Air quality is affected by being an event day in the peripheral zones of the event.

H1: Being an event day increases the number of jams in the peripheral zones of the event.

H2: Being an event day increases the use of bike-sharing systems such as 'Giras'.

7. RESULTS AND DISCUSSION

Before delving into statistical results, a visual analysis is made to give some context, to know what to expect and to set the stage for the subsequent statistical analysis, done with difference in differences model, mentioned above.

7.1. RESULTS

For the first hypothesis, which is “Air quality is affected by being an event day in the peripheral zones of the event.”. Upon completion of the primary visual examination (Figure 6), it becomes apparent that the sensor 'QA00CO0048' situated at 'Avenida Lusíada' detects a greater amount of CO particles in the air on game days as opposed to non-game days.

Observing the 7th figure, which presents a comparative analysis of the mean concentrations of carbon monoxide in the peripheral areas of “Estádio José de Alvalade”, it is evident that two sensors, namely “QA00CO0055” situated at “Avenida General Norton de Matos” and “QA00CO0057” situated at “Campo Grande - Museu da Cidade”, exhibit a noticeable decline in air quality on days when games are held as opposed to non-game days. Upon reviewing the remaining three sensors, the data indicates a consistent level of air quality in these specific locations, irrespective of whether it is a game day or not.

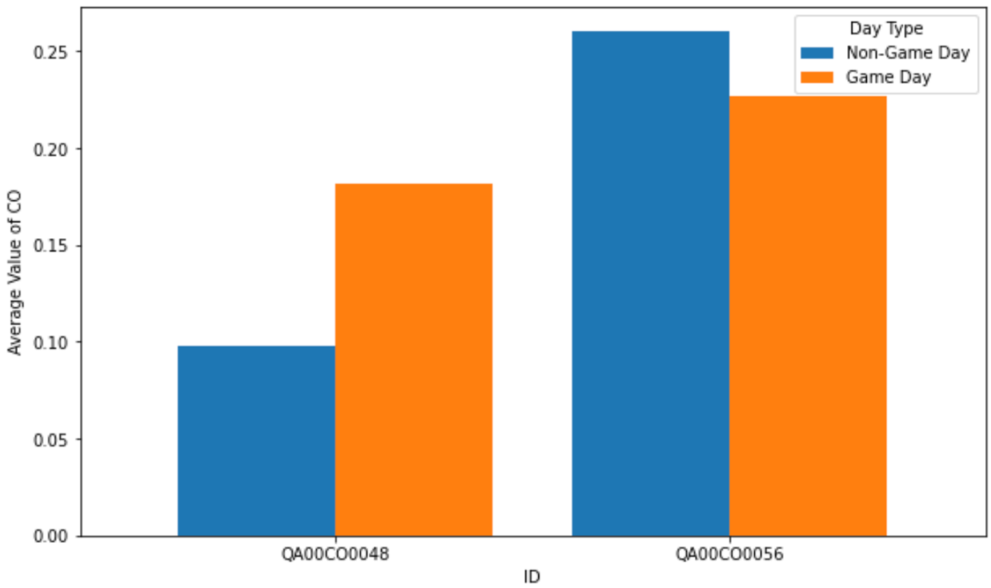

Figure 6 - Comparison between average values of CO in game days and non-game days for sensors located near “Estádio da Luz”

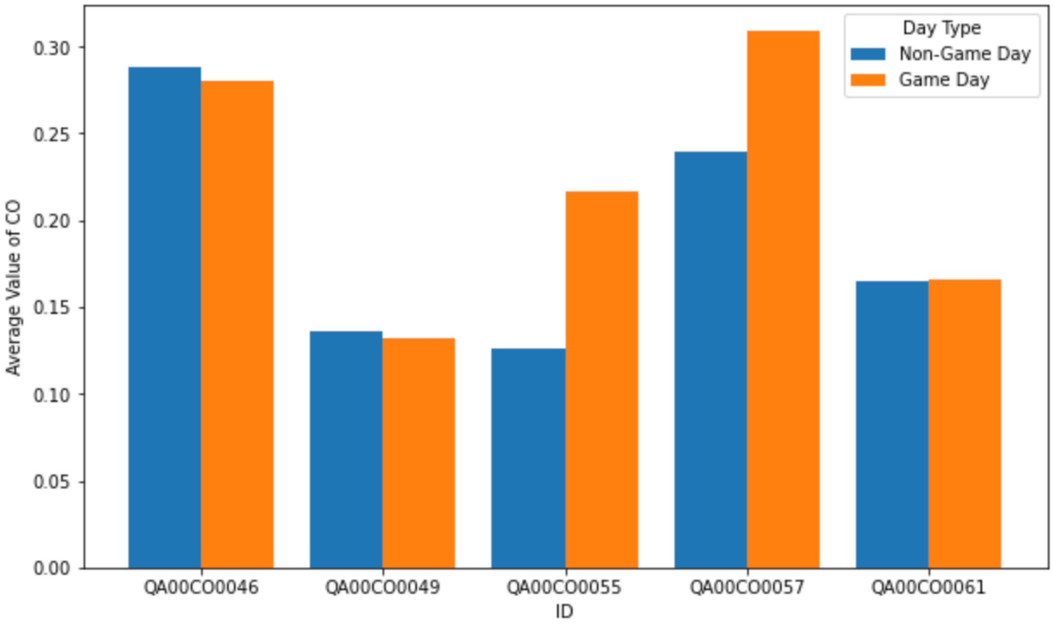

Figure 7 - Comparison between average values of CO in game days and non-game days for sensors located near “Estádio José de Alvalade”

H0 – Statistical Analysis:

Taking the dataset "Estádio da Luz", which contains three sensors monitoring Carbon Monoxide levels - "QA00CO0056" situated at "Avenida General Norton de Matos", "QA00CO0039" at "Estrada de Benfica" and "QA00CO0048" at "Avenida Lusíada" the regression outcomes of the models are as follows.

The sensor QA00CO0048 located at “Avenida Lusíada” has a coefficient for the "is_gameday" variable of 0.0841, which is statistically significant (p-value < 0.001), indicating that there is a positive association between being a game day and the dependent variable (air quality value). This implies that during match days, there is an elevated presence of carbon monoxide in the air in contrast to days when matches are not held.

Located at “Avenida General Norton de Matos” the sensor “QA00CO0056” has a coefficient for the "is_gameday" variable of -0.0291, but it's not statistically significant (p-value = 0.287). This suggests that being a Benfica game day does not have a significant impact on the dependent variable (this can be explained by the fact that it may be a location where don't pass more cars on match days than on non-match days).

For the sensor QA00CO0039 located at “Estrada de Benfica” there were no measurements on match days in 2023, which means that all the values in the model related to this variable are 0, making any analysis of this sensor meaningless.

Taking the dataset "Estádio José de Alvalade", which contains five sensors monitoring Carbon Monoxide levels - "QA00CO0055" situated at "Avenida General Norton de Matos", "QA00CO0057" at "Campo Grande - Museu da Cidade", "QA00CO0049" at "Avenida de Roma", “QA00CO0046” at “Entrecampos” and “QA00CO0061” at “Quinta das Conchas- Avenida Maria Helena Vieira da Silva” the regression outcomes of the models are as follows.

The sensor located at “Avenida General Norton de Matos” has a coefficient for 'is_gameday' variable of 0.085, which remains statistically significant (p-value=0.001), indicating a positive association between Sporting game days and the dependent variable (Air quality value). This suggests that the Air quality value tends to be higher, which implies higher concentration of the pollutant gas, on game days compared to non-game days.

The sensor QA00CO0057 located at “Campo Grande - Museu da Cidade” has a coefficient for the 'is_gameday' variable of 0.052, which remains statistically significant (p-value = 0.003), indicating a positive association between game days and the dependent variable (Air quality value). This suggests that the Air quality value tends to be higher on game days compared to non-game days.

Upon the application of the model to the sensor QA00CO0046 located at “Entrecampos” the coefficient for the 'is_gameday' variable (-0.0099) is not statistically significant (p-value = 0.400), suggesting that there is no clear association between game days and the dependent variable (Air quality value). This indicates that the Air quality value does not significantly differ between game days and non-game days.

After applying the model to the sensor QA00CO0049 located at “Avenida de Roma” the coefficient for the 'is_gameday' variable (-0.001) is not statistically significant (p-value = 0.400), suggesting that there is no clear association between game days and the dependent variable (Air quality value). This indicates that the Air quality value does not significantly differ between game days and non-game days.

For the sensor QA00CO0061 located at “Quinta das Conchas- Avenida Maria Helena Vieira da Silva” the coefficient of the 'is_gameday' variable (0.004) is not statistically significant (p-value = 0.700), suggesting that there is no clear association between game days and the dependent variable (Air quality value). This indicates that the Air quality value does not significantly differ between game days and non-game days. An overview of the results is displayed in the table below (Table 8).

Table 8 - Overview of the results obtained for H0

Sensor	P-value	Coefficient (is_gameday)
QA00CO0048	< 0.001	0.0841
QA00CO0056	0.287	-0.0291
QA00CO0055	0.001	0.085
QA00CO0057	0.003	0.052
QA00CO0046	0.4	-0.0099
QA00CO0049	0.4	-0.001
QA00CO0061	0.7	0.004

H1 – Visual Analysis:

For the second hypothesis, being an event day increases the use of bike-sharing systems. The information that can be gathered for both the graph related to the ”Estádio José de Alvalade” (figure 4) and the graph with streets on the outskirts of the ‘Estádio da Luz’ (figure 3) is that the average number of jams on match days is higher than the average number of jams on non-match days in practically all the streets, with the exception of the ‘Avenida Lusíada’ related to the ‘Estádio da Luz’. Following this analysis, it's time to see if the what’s written above can be statistically proven.

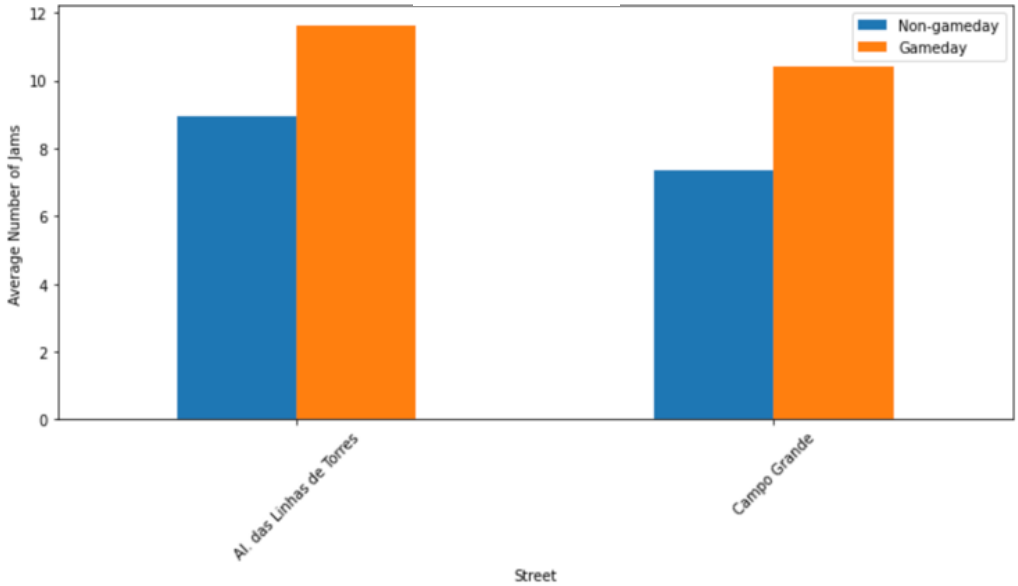

Figure 8 - Average number of jams per location

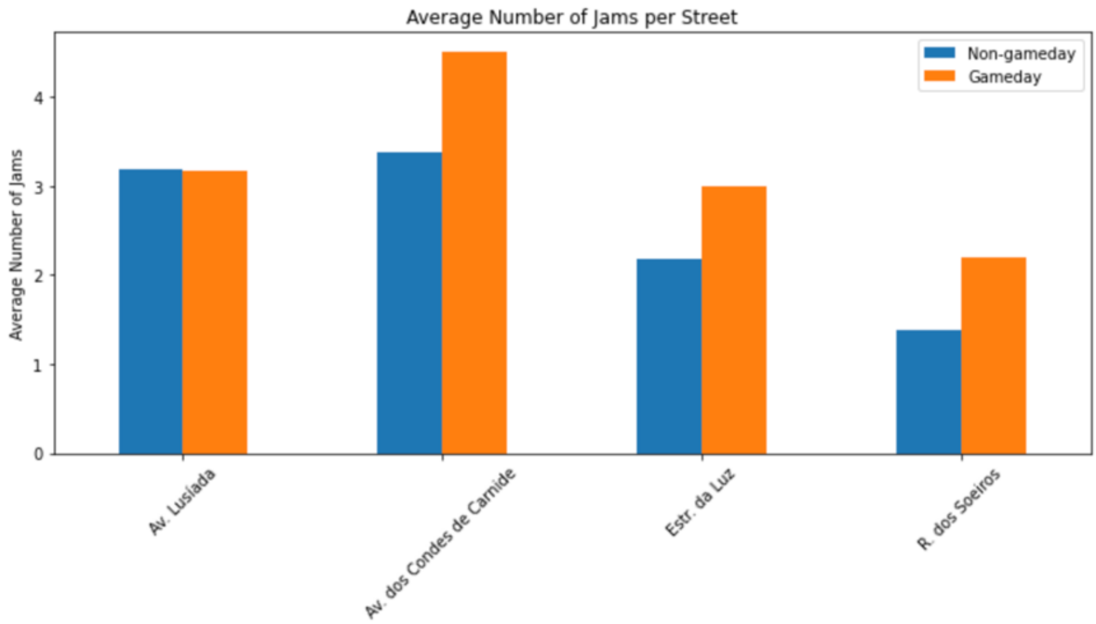

Figure 9 - Average number of jams per location

H1 – Statistical Analysis:

Applying the regression model difference in differences to dataset related with "Estádio José de Alvalade" which contains two zones "Alameda da Linhas das Torres" and "Campo Grande", the following results were obtained. The model explained approximately 40.5% of the variance in the number of traffic jams ($R^2 = 0.405$), the coefficient for match days (is_gameday) was 4.07 ($p < 0.001$). This suggests that, on average, the number of traffic jams increases by approximately 4.07 on match days compared to non-match days, holding the effect of weekdays constant. For weekdays (is_weekday) the coefficient was 8.19 ($p < 0.001$). This indicates that, on average, the number of traffic jams increases by approximately 8.19 on weekdays compared to weekends, holding the effect of match days constant.

To the dataset related with "Estádio da Luz", in order to run the model "Avenida Lusíada" was removed from this dataset and stored in another one. For the dataset without "Avenida Lusíada" , the model explained approximately 41.5% of the variance in the number of traffic jams ($R^2 = 0.405$). The coefficient for match days was 5.25, with a p-value < 0.001. This suggests that, on average, the number of traffic jams increases by approximately 5.25 on match days compared to non-match days, holding the effect of weekdays constant. Additionally, the coefficient for weekdays was 6.75, with a p-value < 0.001. This indicates that, on average, the number of traffic jams increases by approximately 6.75 on weekdays compared to weekends, holding the effect of match days constant.

After applying the model to the dataset with only "Avenida Lusíada", it explained approximately 3.2% of the variance in the number of traffic jams ($R^2 = 0.032$). The coefficient for match days was 0.78 with a p-value of 0.497, this suggests that, on average, the number of traffic jams increases by approximately 0.78 on match days compared to non-match days, holding the effect of weekdays constant. However, this effect is not statistically significant. The coefficient for weekdays was 1.71 ($p = 0.171$). This indicates that, on average, the number of traffic jams increases by approximately 1.71 on weekdays compared to weekends, holding the effect of match days constant. However, this effect is also not statistically significant.

In conclusion, and statistically proving what the visual elements above showed, being match day influences the number of jams in all the zones studied, except for 'Avenida Lusíada', which was statistically proven that match day has no influence on the number of jams, and consequently on traffic flow. An overview of the results is displayed in the table below (Table 9).

Table 9 - Overview of the results obtained for H1

Location	R-Squared	P-value	Coefficient (is_gameday)
“Estádio da Luz”	41.5%	< 0.001	5.25
“Estádio José de Alvalade”	40.5%	< 0.001	4.07
“Aveninda Lusíada”	3.2%	0.497	0.78

H2 – Visual Analysis:

For the third hypothesis, being an event day increases the use of bike-sharing systems. Two two specific zones were selected for each stadium, namely “Telheiras” and “Jardim do Campo Grande” in the proximity of “Estádio José de Alvalade” and “Avenida Lusíada” and “Avenida do Colégio Militar” for the “Estádio da Luz”. This visual analysis will illustrate the variations in number detections captured by the sensors in match days and non-match days within each designated zone.

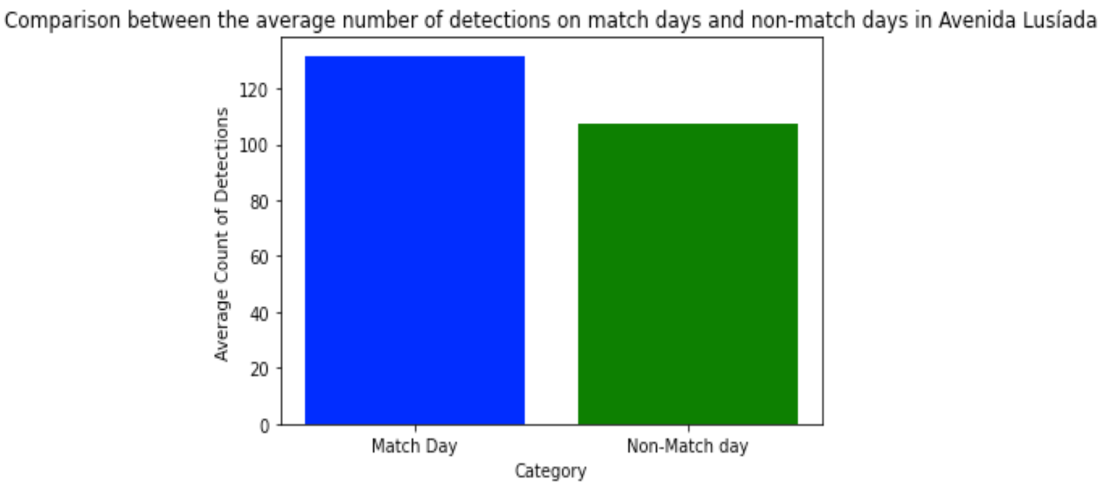

Figure 10 - Comparison between number of detections on match days and non-match days in Avenida Lusiada

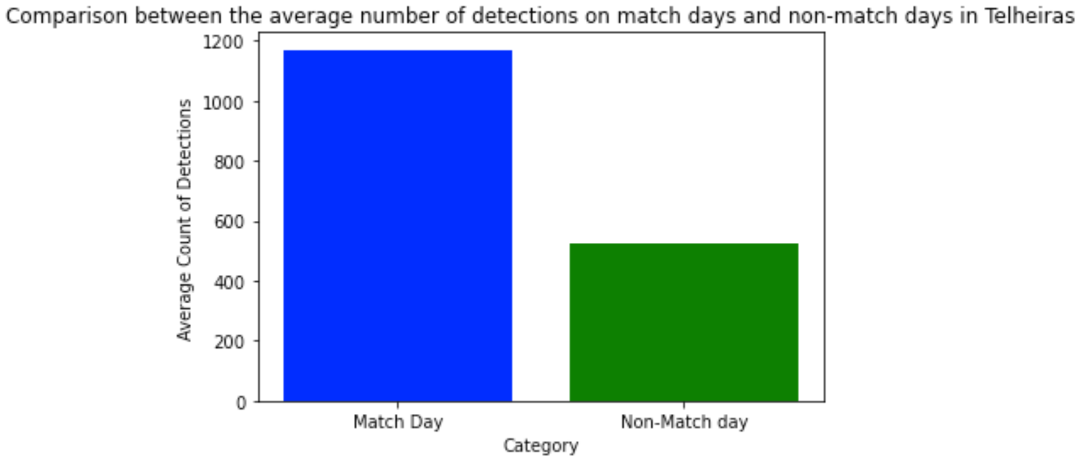

Figure 11 - Comparison between number of detections on match days and non-match days in Telheiras

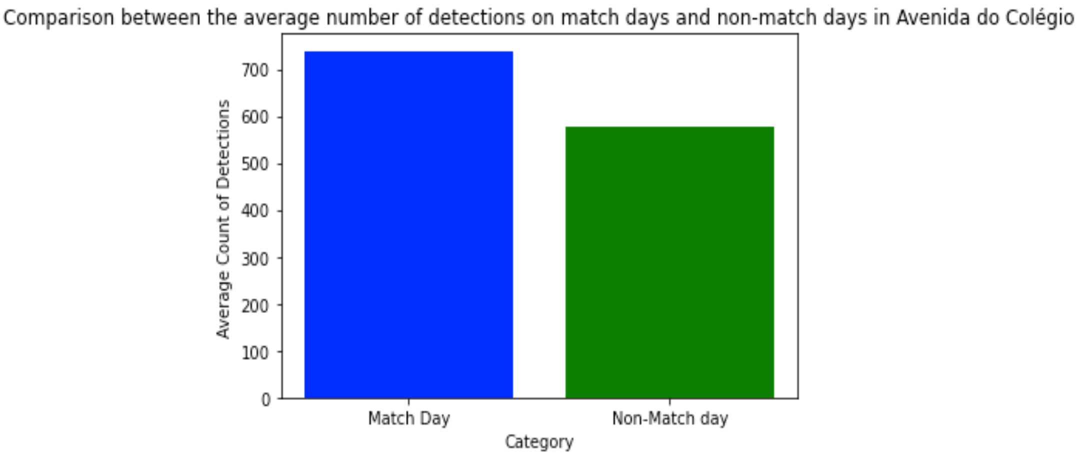

Figure 12 - Comparison between number of detections on match days and non-match days in Avenida do Colégio

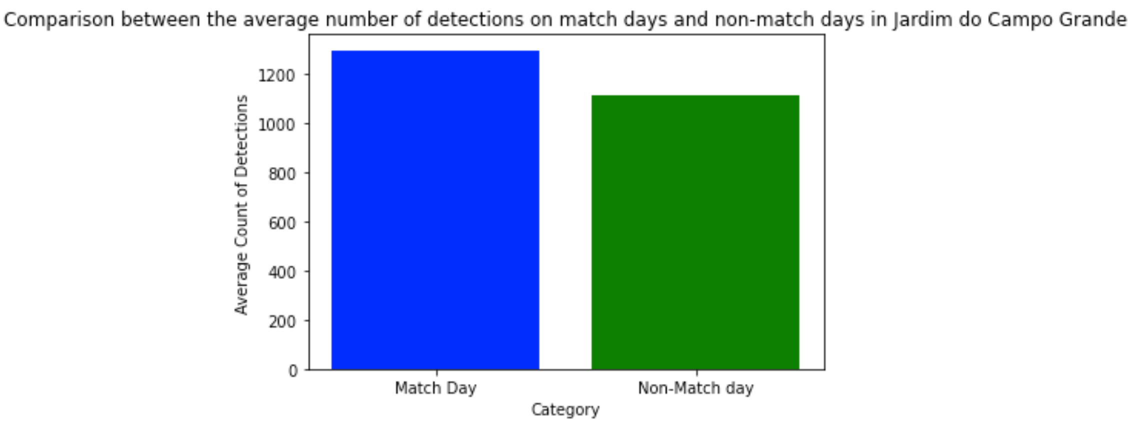

Figure 13 - Comparison between number of detections on match days and non-match days in Jardim do Campo Grande

Taking these bar graphs into account, it can be said that in any of the zones, match day increases the number of bicycles circulating in the areas surrounding the stadiums. But in order to be able to confirm this with 100 per cent certainty, it has to be proven statistically, which will be done in the 'Statistical Analysis' section below.

H2 – Statistical Analysis:

Upon the application of the model to the sensor data detections at "Jardim do Campo Grande", the following results were obtained**,** the model explained approximately 57.3% of the variance in bike counts ($R^2 = 0.573$). The coefficient for match days was 320.98 ($p = 0.003$). This suggests that, on average, bike counts increase by approximately 321 on match days compared to non-match days, holding the effect of weekdays constant. Additionally, the coefficient for weekdays was 603.33 ($p = 0.001$). This indicates that, on average, bike counts increase by approximately 603 on weekdays compared to weekends, holding the effect of match days constant.

To the "Telheiras" located sensor the model explained approximately 61.1% of the variance in bike counts ($R^2 = 0.611$). The coefficient for the variable 'is_gameday' was 748.64 ($p < 0.001$). This suggests that, on average, bike counts increase by approximately 749 on match days compared to non-match days, holding the effect of weekdays constant. For the variable 'is_weekday' the coefficient was 427.61 ($p < 0.001$).

After applying the model to the "Avenida Lusíada" located sensor, the model explained approximately 45.0% of the variance in bike counts ($R^2 = 0.450$)**.** The coefficient for match days was 34.64 ($p = 0.001$). This suggests that, on average, bike counts increase by approximately 35 on match days compared to non-match days, holding the effect of weekdays constant. Additionally, the coefficient for weekdays was 43.89 ($p < 0.001$). This indicates that, on average, bike counts increase by approximately 44 on weekdays compared to weekends, holding the effect of match days constant.

To the "Avenida do Colégio Militar" located sensor, the model explained approximately 16.3% of the variance in bike counts ($R^2 = 0.163$). The coefficient for match days was 166.95 ($p < 0.001$). This suggests that, on average, bike counts increase by approximately 167 on match days compared to non-match days, holding the effect of weekdays constant. And the coefficient for weekdays was 20.79 ($p = 0.366$).

In conclusion, it is possible to state, since it has been statistically proven, that match day has a positive influence on the number of bike-sharing detections in the areas studied on the outskirts of stadiums. An overview of the results is displayed in the table below (Table 10).

Table 10 - Overview of the results obtained for H2

Location	**R-Squared**	**P-value**	**Coefficient (is_gameday)**
“Avenida Lusíada"	45%	< 0.001	34.64
“Telheiras”	61.1%	< 0.001	748.64
“Avenida do Colégio”	16.3%	0.001	166.95
“Jardim do Campo Grande”	57.3%	0.003	320.98

7.2. DISCUSSION

In this study, the impact of event days on various urban dynamics was investigated, specifically focusing on air quality, traffic congestion, and the usage of bike-sharing systems. Using a Difference-in-Differences (DiD) regression model were tested three hypotheses to understand how event days influence these factors. The analysis revealed several key insights:

7.2.1. Hypothesis 0 (H0): Air quality is affected by being an event day

The analysis proved statistically that air quality in certain zones is significantly affected by being an event day, but in other zones, also in the surroundings of the event the air quality isn’t affected. On average, pollution levels were higher in some areas close to event venues compared to non-event days, suggesting that the influx of vehicles and increased human activity contribute to deteriorating air quality. These findings highlight the need for targeted environmental policies and temporary air quality monitoring during events. Implementing measures such as temporary car-free zones and promoting public transportation can help mitigate the adverse effects on air quality.

7.2.2. Hypothesis 1 (H1): Being an event day increases the number of jams in the peripheral zones of the event

The analysis proved statistically that the number of traffic jams significantly increased in the peripheral zones of the event on match days compared to non-match days. This indicates that event-related activities cause substantial traffic congestion not only in the immediate vicinity of the venue but also in surrounding areas. Considering these findings, city planners and event organizers should develop comprehensive traffic management plans that extend beyond the event venue to include peripheral zones. Enhanced traffic control measures, real-time traffic monitoring, and clear communication with the public about alternative routes can alleviate congestion.

7.2.3. Hypothesis 2 (H2): Being an event day increases the use of bike-sharing systems in the peripheral zones of the event

Being an event day was associated with a significant increase in the use of bike-sharing systems, in the surroundings areas of the event, what was proven statistically. The data showed a notable rise in bike usage on event days, indicating that many people use bikes as an alternative mode of transportation to avoid traffic jams. Given these findings, promoting bike-sharing systems and ensuring their availability and accessibility during events can encourage sustainable transportation. Expanding bike lanes and providing additional bike docking stations near event venues can further support this eco-friendly mode of transport.

The results of the analysis demonstrate the multifaceted impact of event days on urban environments. The combined effects on air quality, traffic congestion, and bike-sharing usage underscore the importance of integrated urban planning strategies that address these interrelated factors.

8. CONCLUSION

Through this study, a model solution was developed using the Difference in Differences approach to explore the impact of football matches in Lisbon on three distinct hypotheses. Examining the effects of these events on air quality, traffic congestion, and the utilization of sustainable transportation modes like bike-sharing systems was crucial. Additionally, analysing how the increased usage of such transportation methods could reduce traffic congestion and enhance air quality, thereby contributing to an improved quality of life for residents near these events.

An extensive analysis of the literature about the topics above was developed to fulfil the gap in study, such as how the traffic is affected by football games in Madrid (Olabarrieta & Laña, 2020), how low emission zones affect air quality (Santos et al., 2018) and how bike sharing impact on car usage(Fishman et al., 2014) , these are just three examples of a literature review that has been a key element throughout the development of this thesis.

Following Crisp-dm methodology, initially we understood the problem and the data that were going to be used. After bringing these points together, we started data preparation where all the datasets in use were cleaned, variables that were considered important were added do the datasets that were divided by location. Lastly, the model was applied and achieved the results discussed thoroughly in the previous section.

As regards the Hypothesis (H0), findings did not support completely that air quality in the surroundings of the stadiums is affected by being a football game day. However, Hypothesis 1 and 2, that specified that the number of traffic jams increased in the surroundings areas of stadiums in game days and that the usage of bike sharing system also increased in these days, were proven statistically.

9. LIMITATIONS AND FUTURE WORK

In this study, some limitations must be acknowledged. Firstly, in terms of geographical scope, this research was conducted in a single city, which may not reflect how other big cities, in Portugal or outside of Portugal function in football game days. Other limitation is the external factors, like weather conditions and local policies were not controlled for, potentially influencing the results. Additionally, the team's performance can influence the frequency to which fans go to the stadium. Therefore, it could influence the results. Also, this study period was confined to a single football season, which may not capture variations over multiple seasons or account for long-term trends.

Future work could expand the scope of analysis to include more diverse events and additional variables, this could be a great way to provide a more comprehensive understanding of urban dynamics on event days. Investigating the long-term effects of implemented strategies on air quality, traffic congestion, and bike-sharing usage could also offer valuable insights for continuous improvement in urban planning. Another hypothesis of future work could be expanding the geographical scope to have a perspective of how other cities work in these days, including more external factors, like weather conditions or local policies could be a good add-on to a future work.

BIBLIOGRAPHICAL REFERENCES

Albuquerque, V., Oliveira, A., Barbosa, J. L., Rodrigues, R. S., Andrade, F., Dias, M. S., & Ferreira, J. C. (2021). Smart cities: Data-driven solutions to understand disruptive problems in transportation—The Lisbon Case Study. Energies, 14(11), 3044.

Ban, K., & Kedagni, D. (2022). Generalized difference-in-differences models: Robust bounds.

Chamberlain, O. (2016). Carbon-monoxide (CO): A poisonous gas emitted from automobiles, its effect on human health. Adv Automob Eng, 5(151), 2.

Card, D., & Krueger, A. B. (2000). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply. American Economic Review, 90(5), 1397-1420.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide (). The CRISPDM consortium.

Currie, J., & Walker, R. (2011). Traffic congestion and infant health: Evidence from E-ZPass. American Economic Journal: Applied Economics, 3(1), 65-90.

Feng, Y., Affonso, R. C., & Zolghadri, M. (2017). Analysis of bike sharing system by clustering: the Vélib'case. IFac-Papersonline, 50(1), 12422-12427.

Fishman, E., Washington, S., & Haworth, N. (2014). Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. Transportation research Part D: transport and environment, 31, 13-20.

Fontes, T., Arantes, M., Figueiredo, P. V., & Novais, P. (2022). A cluster-based approach using smartphone data for bike-sharing docking stations identification: Lisbon case study. Smart Cities, 5(1), 251-275.

Huang, Y., Xiao, Z., Wang, D., Jiang, H., & Wu, D. (2019). Exploring individual travel patterns across private car trajectory data. IEEE Transactions on Intelligent Transportation Systems, 21(12), 5036-5050.

Jia, R., Shao, S., & Yang, L. (2021). High-speed rail and CO2 emissions in urban China: A spatial difference-in-differences approach. Energy Economics, 99, 105271.

Kou, Z., & Cai, H. (2019). Understanding bike sharing travel patterns: An analysis of trip data from eight cities. Physica A: Statistical Mechanics and its Applications, 515, 785-797.

Li, J., Ma, M., Xia, X., & Ren, W. (2021). The Spatial Effect of Shared Mobility on Urban Traffic Congestion: Evidence from Chinese Cities. Sustainability, 13(24), 14065.

LOPES, D., FERREIRA, J., RAFAEL, S., BAPTISTA, P., FARIA, M., CANHA, N., & ALMEIDA-SILVA, M. A. R. I. N. A. (2019). Urban mobility strategies to improve local air quality: case study of Lisbon, Portugal. WIT Transactions on Ecology and the Environment, 236, 257-266.

Lv, J., Zheng, L., Ye, Y., & Ye, C. (2020, July). Understanding Travel Patterns of Commuting Private Cars using Big data of Electronic Registration Identification of Vehicles. In SEKE (pp. 43-48).

Ma, X., Cao, R., & Jin, Y. (2019). Spatiotemporal clustering analysis of bicycle sharing system with data mining approach. Information, 10(5), 163.

Olabarrieta, I. I., & Laña, I. (2020, September). Effect of soccer games on traffic, study case: Madrid. In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC) (pp. 1-5). IEEE.

Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., & Giannotti, F. (2013). Understanding the patterns of car travel. The European Physical Journal Special Topics, 215, 61-73.

Raposo, M., & Silva, C. (2022). City-Level E-Bike Sharing System Impact on Final Energy Consumption and GHG Emissions. Energies, 15(18), 6725.

Saltz, J. S. (2021, December). CRISP-DM for data science: strengths, weaknesses and potential next steps. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 2337-2344). IEEE.

Santos, F. M., Gómez-Losada, Á., & Pires, J. C. (2019). Impact of the implementation of Lisbon low emission zone on air quality. Journal of hazardous materials, 365, 632-641.

Sarroeira, R., Henriques, J., Sousa, A. M., Ferreira da Silva, C., Nunes, N., Moro, S., & Botelho, M. D. C. (2023). Monitoring Sensors for Urban Air Quality: The Case of the Municipality of Lisbon. Sensors, 23(18), 7702.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, 5(4), 13-22.

Silveira, C., Ferreira, J., Lopes, D., & Miranda, A. I. Impact of traffic management strategies on air quality and health in a Portuguese urban area.

Taborda, R., Datia, N., Pato, M. P. M., & Pires, J. M. (2020, September). Exploring air quality using a multiple spatial resolution dashboard—a case study in lisbon. In 2020 24th International Conference Information Visualisation (IV) (pp. 140-145). IEEE.

Xiao, Z., Xiao, H., Jiang, H., Chen, W., Chen, H., & Regan, A. C. (2021). Exploring human mobility patterns and travel behavior: A focus on private cars. IEEE Intelligent Transportation Systems Magazine, 14(5), 129-146.

Xiao, Z., Xu, S., Li, T., Jiang, H., Zhang, R., Regan, A. C., & Chen, H. (2020). On extracting regular travel behavior of private cars based on trajectory data analysis. IEEE Transactions on Vehicular Technology, 69(12), 14537-14549.

Xin, R., Yang, J., Ai, B., Ding, L., Li, T., & Zhu, R. (2023). Spatiotemporal analysis of bike mobility chain: A new perspective on mobility pattern discovery in urban bike-sharing system. Journal of Transport Geography, 109, 103606.

Zulkifli, C. N. A. B., & Ponrahono, Z. (2018). The effect of traffic congestion on quality of community life. European Proceedings of Multidisciplinary Sciences.

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

Universidade Nova de Lisboa