NOVA IMS Information Management School

# MDDM

Master's degree Program in
**Data-Driven Marketing**

## ANALYSIS OF LISBON VISITORS' INTERNET ACCESS BEHAVIOR

Behavior analysis through the identification of clusters

Beatriz Miguel Carvalho Silva

Project Work

presented as a partial requirement for obtaining the master's degree program in Data-Driven Marketing

**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

# ANALYSIS OF LISBON VISITORS' INTERNET ACCESS BEHAVIOR

## Behavior analysis through the identification of clusters

By

Beatriz Miguel Carvalho Silva

Project Work presented as a partial requirement for obtaining the Master's Degree in Data-Driven Marketing, with a specialization in Marketing Intelligence.

**Supervisor:** Miguel de Castro Ferreira Simões Neto

**Cossupervisor:** João Bruno Morais de Sousa Jardim

January 2023

# 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.

*Beatriz Miguel Carvalho Silva*

*Lisbon, 14th of July 2023*

# ACKNOWLEDGMENTS

First and foremost, I would like to express my heartfelt gratitude to my family for their unwavering support throughout this Dissertation. They made tremendous efforts to facilitate my education and empowered me to pursue my aspirations, shaping me into the confident, determined, ambitious, and joyful individual I am today. I am truly grateful for their dedication and consider this a mission accomplished!

Particularly, I am immensely grateful to my oldest sister, Carolina Cabral, and my brother-in-law, Guilherme Cabral, for their invaluable support. They made significant contributions to this personal and academic achievement, motivating me to enhance my intellectual capabilities, especially in deepening my understanding of tourist behavior and mobility.

Furthermore, I would like to extend my heartfelt appreciation to my colleagues and friends who enhanced my social well-being with their joyful and loving presence. They have been a source of support throughout my academic journey and beyond. I would like to express a special note of gratitude to my fellow master's degree classmates, Mariana Castro, and Francisco Afonso. Their presence made this journey even more joyful, fulfilling, and enriching.

Lastly, I am deeply grateful to my friends for their unwavering support and understanding throughout this demanding endeavor. Their love, encouragement, and belief in my abilities have been a constant source of inspiration and strength.

Thank you!

# ABSTRACT

This master's thesis focuses on clustering the internet access behavior of urban visitors in the Lisbon urban area. To promote smart city development, the study aims to provide insights into visitors' behaviors while accessing the internet in Lisbon, enabling improved decision-making processes for city management, and enhancing the overall online and offline experience for visitors. The over-tourism phenomenon has put a strain on infrastructure, public transportation, and cultural heritage sites. Therefore, innovative methods are needed for effective smart city management, particularly in urban mobility. The increasing availability of Wi-Fi networks during travel has generated valuable data that can be used to develop groundbreaking approaches to understanding visitors' behaviors and mobility patterns in urban areas. This knowledge enables the analysis and clustering of urban visitors' behavior, contributing to improved decision-making processes in smart city management.

# KEYWORDS

Urban planning; Visitor's spatial behavior; Urban mobility; Travel behavior; Mobile phone data; Visitors mobility

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **ANOVA** | Analysis of Variance |
| **BDT** | Big Data Technology |
| **CHI** | Calinski-Harabsz Index |
| **CML** | City Council of Lisbon |
| **CRISP-DM** | Cross-Industry Standard Process for Data Mining |
| **DBI** | Davies-Bouldin Index |
| **DBSCAN** | Density-Based Spatial Clustering of Applications with Noise |
| **EDA** | Exploratory Data Analysis |
| **GMM** | Gaussian Mixture Model |
| **ICT** | Information and Communication Technology |
| **INE** | National Institute of Statistics |
| **K** | Cluster |
| **KNN** | K-Nearest Neighbors |
| **OPTICS** | Ordering Points to Identify the Clustering Structure |
| **PCA** | Principal Component Analysis |
| **POI** | Points of Interest |
| **RSSI** | Received Signal Strength Indicator |
| **SPSS** | Statistical Package for the Social Sciences |
| **UGC** | User-Generated Content |
| **UNWTO** | United Nations World Tourism Organization |
| **UTM** | Universal Transverse Mercator |
| **WCSS** | Within-Cluster Sum of Squares |

# 1. INTRODUCTION

## 1.1. BACKGROUND AND CONTEXT

Tourism represents a highly dynamic and particularly promising industry sector worldwide as a business of paramount importance to deliberately attract a growing number of visitors. Urban center management needs to consider the characteristics of the visitors in city centers to provide better infrastructure and services, allowing not only a better experience but also better conditions for residents (Streimikiene et al., 2021).

The act of traveling, whether it be for necessity, recreation, or as a lifestyle choice, is an integral aspect of human existence, and the effects of Covid-19 on travel demand have long since dissipated. In 2022, global tourism recorded more robust outcomes than anticipated, propelled by substantial suppressed demand and the easing or removal of travel limitations in numerous nations ("UNWTO World Tourism Barometer and Statistical Annex, January 2022," 2022). According to United Nations World Tourism Organization (UNWTO) projections for 2023, this year could see international tourist arrivals recover to between 80% and 95% of their pre-pandemic levels.

In recent years, there has been a growing concern about over-tourism and tourism phobia, which have emerged because of unsustainable mass tourism practices and social movements protesting against tourism growth. There is a need for more sustainable tourism practices and stronger governmental policies and regulations to manage and optimize infrastructures regarding tourism growth (Milano et al., 2019). The increase in the number of tourists, the emergence of new segments, and technology applied to means of transport and tourist communication are identified as the catalyzing factors of over-tourism. Tourist destinations that are experiencing overcrowding need to implement measures to manage the flow of visitors and reduce the negative impacts.

Multiple solutions have been applied to mitigate the phenomenon of over-tourism. Tourists can be distributed to the city and its surroundings through sensors, virtual reality videos, and mobile applications. Sensoring has been real-life tested by Dubrovnik City Council while creating a predictive system for the volume of visitors, and mobile applications have been developed such as the application "Play London with Mr. Bean". Other solutions such as the definition of new touristic sites and the development of new itineraries have also been tested in capital cities such as Venice, Stockholm, and Prague. Georeferencing tourists has also been a solution for urban management, communicating with tourists visiting specific attractions, and providing accurate regulation of tourist activities. Furthermore, techniques regarding sensoring, optimization of public transportation, and energy efficiency, have been extensively put into action in many countries such as Italy, Spain, Germany, Netherlands, and Denmark (García-Hernández et al., 2019). Many solutions can be applied to improve urban management by accessing Big Data.

The use of Information and Communication Technology (ICT) in Smart Cities and the practice of smart tourism can enhance efficiency, accessibility, and visitor experience, but it is also important to develop human capital and preserve a city's unique cultural identity (Encalada et al., 2017).

Furthermore, Urban Big Data refers to a vast amount of dynamic and static information created from various aspects of urban life such as infrastructure, organizations, and people. This data is collected and organized by city governments, public institutions, businesses, and individuals using advanced information technology. By analyzing and interpreting this data, it is possible to better understand how cities operate and make more informed decisions about urban management. This, in turn, can help optimize the use of resources, reduce operational costs, and promote sustainable and intelligent urban development (Pan et al., 2016).

The global transformation of cities requires the creation of an intelligent network infrastructure using advanced technologies. This can be achieved through the implementation of Big Data Technology (BDT) based on an integrated approach to the Smart City concept. The primary goal of this concept is to meet the needs of the city residents by covering all areas of urban planning and management. The concept relied heavily on intellectual innovation based on information technology and data analysis. Smart City technology is active (Ivanov & Gnevanov, 2018).

Smart City technology is actively being developed in the United States of America, and across Europe and Southeast Asia. Two main approaches to the implementation of the concept have been identified: the introduction of technology through the creation of new cities and the implementation of the concept in existing cities through local or complex projects. The use of information-communication infrastructure and technology makes it possible to adapt urban systems to existing needs, reduce resource consumption, improve the quality of services, and create new points of economic growth. Smart City technologies are increasingly becoming a part of citizens' daily lives, acting as a centralized means of obtaining services and improving the efficiency of urban services (Ivanov & Gnevanov, 2018).

## 1.2. Problem Statement and Research Questions

The increasing search for city breaks in Lisbon creates not only pressure on tourism but also raises a challenge of economic dependence on this industry. According to (Milano et al., 2019), a tourist destination should be an attractive place for both tourists and residents, not sacrificing the normal functioning of the city. However, there is a balance that should be considered, otherwise, the city could become over touristic. The term over-tourism refers to a scenario where the effects of tourism in certain places and times surpass the threshold of physical, ecological, social, economic, psychological, and political capacities (European Parliament, 2018).

The phenomenon of over-tourism has increasingly affected the city of Lisbon in recent years, with specific impacts being observed. These impacts include overcrowding and congestion in the urban center, particularly in areas surrounding popular tourist attractions, resulting in difficulty for both residents and visitors to move around and enjoy the city. Additionally, the growing number of tourists has put pressure on the city's housing market, leading to a shortage of affordable housing for residents and the conversion of residential properties into short-term rentals, thereby displacing residents. The increase in tourism has also placed pressure on the city's infrastructure and natural resources, leading to environmental impacts. Furthermore, the growing number of tourists can negatively impact the quality of life for residents, including noise pollution, increased traffic, and a loss of community cohesion. Finally, tourism can also lead to safety and security concerns, particularly in crowded or tourist-heavy areas, which may require the presence of law reinforcement or security personnel to maintain public safety (Milano et al., 2019).

Overall, the pressure of over-tourism in Lisbon is a complex issue that requires a multifaceted approach to address the problem, including the implementation of measures to better manage tourism. Such measures could implicate the setting of limits on the number of tourists allowed in certain areas, promoting sustainable tourism practices, and working to protect the rights and needs of locals. Addressing the consequences of over-tourism is imperative and requires solutions that aim to build harmonious urban areas where both tourists and residents can cohabitate, not interfering with each other's experiences in the urban centers. Indeed, innovation in urban mobility could occur through the understanding of tourists' movements when visiting urban centers. This could potentially create outstanding opportunities for urban management concerning transportation routes, help improve infrastructures in overcrowded areas, and stimulate the development of smart solutions towards providing real-time information on touristic traffic.

Ultimately, whilst prioritizing the development of sustainable touristic mobility and aptitude to create smart cities, this work project aims to sustain decision-making processes through the data gathered. Thereby, this work project has the power to promote change in urban management, as well as improve the urban visitors' and locals' access to services and reverse over-tourism's impact to overcome the city's main challenges.

This study aims to address these gaps by clustering grids based on their visitors' behavior, to allow better decision-making processes, with the information gathered from this project. By analyzing a large dataset of public data gathered from Wi-Fi connections that capture the visitors' demography, their Wi-Fi connections, and the device used for Wi-Fi connection, among other variables. The main research questions aimed to be addressed are as follows:

- How many people access specific urban areas at a certain time?
- What is the temporal behavior in internet access points?
- Which parishes present similar behaviors?

### 1.3. RESEARCH OBJECTIVES

The object selected for this work project is the urban visitors' behavior surrounding the Lisbon urban area. Also, with the increasing easiness of accessing data to pursue further investigation about visitors' mobility in urban areas – the possibilities to promote and sustain decision-making processes in urban management are broader.

As stated previously, this work project aims to outline guidelines based on real data obtained through open networks in Lisbon's urban area. To achieve this, the goal is to analyze in further detail the clusters of grids most visited based on their visitors' behaviors. This investigation should provide concrete conclusions which could be used in the decision-making process on urban management.

### 1.4. WORK PROJECT STRUCTURE

The methodology followed embraces a conceptual approach to understanding the behavior of visitors in Lisbon's urban area. The work project is structured into six chapters, including an introduction, literature review, methodology, empirical study, results and discussion, and conclusion.

The following workflow outlines the structure of the methodology used in this work project (see Figure 1).

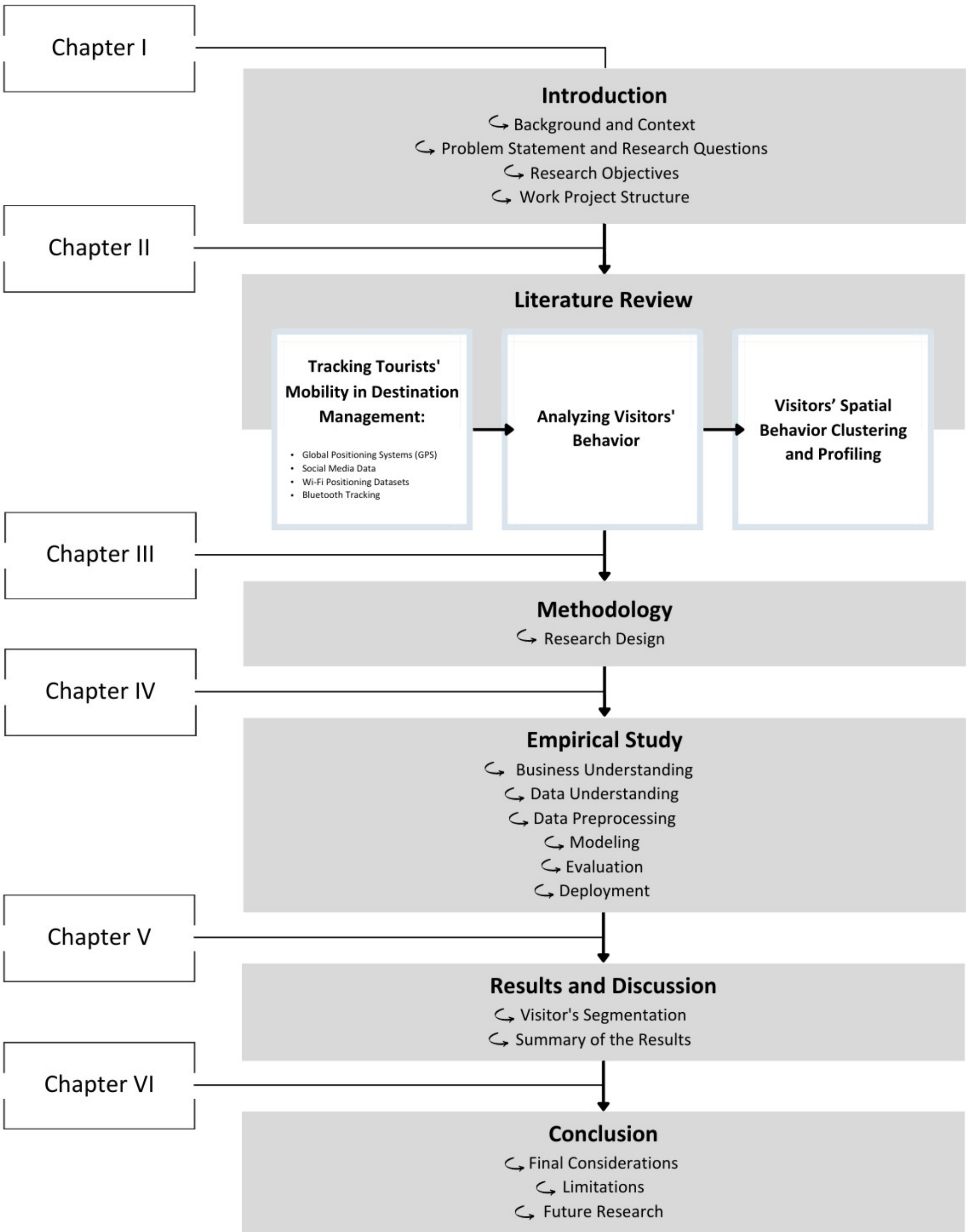


Figure 1 – Proposed Methodology Workflow: Introduction, Literature Review, Methodology, Empirical Study, Results and Discussion, and Conclusion.

# 2. LITERATURE REVIEW

## 2.1. INTRODUCTORY NOTE: VISITORS VS TOURISTS

The UNWTO's definition of tourism recognizes that visitors include both tourists and excursionists, with the key distinction being the inclusion of an overnight stay for tourists. By acknowledging that tourists represent a significant subset of visitors, it is reasonable to draw literature that specifically focuses on tourists' behavior online.

The justification for utilizing literature on tourist behavior stems from the understanding that tourists often exhibit similar behaviors compared to other types of visitors. They have similar motivations, preferences, and patterns of engagement with the destination, including the use of online resources. Previous studies on tourists' behavior online provide valuable insights into their information-seeking processes, decision-making factors, and digital engagement during their visits.

Moreover, considering tourist behavior in the literature review allows for the identification of trends, best practices, and theoretical frameworks that have been established in the context of tourism research. While the dataset may not explicitly identify individuals as tourists, the findings from previous studies on tourist behavior can still inform the understanding of visitor behavior more broadly. This approach enables the exploration of potential similarities and differences between tourists and other visitors, helping to identify common patterns and unique characteristics within the overall visitor population.

By including a literature review focused on tourists' behavior online, despite the absence of direct proof of tourists in the dataset, the study acknowledges the significance of tourists as a key visitor segment and recognizes the potential relevance of their behaviors and preferences to the broader visitor population. It demonstrates a comprehensive approach to understanding visitor behavior, incorporating insights from the tourism literature to enrich the analysis and interpretation of the dataset's findings.

## 2.2. TRACKING TOURISTS' MOBILITY IN DESTINATION MANAGEMENT

Tourism relies on people moving to different places, and the motivation for travel is important in defining tourism demand. Tourist mobility overlaps with daily mobility flow in cities, which makes it mandatory to integrate tourist mobility planning into overall urban mobility planning (La Rocca, 2015).

Understanding tourist mobility is crucial for creating sustainable tourism destinations. Tourist mobility patterns reflect preferences and can be used to segment the market, predict flow, reduce congestion, and assist in constructing tourist-friendly environments. Historically, research on spatial patterns focused on supply-side factors rather than tourist demand. However, recent studies have used digital tracking technologies, such as GPS and geotagged tweets, to examine spatial, temporal, and spatial-temporal behavior. A spatial-temporal analysis is preferred as it provides a comprehensive understanding of tourist behaviors. Furthermore, recent research has shifted from methodological contributions to practical implications for sustainable tourism destination development and management ("Mobility Patterns of International Tourists: Implications for Responsible Urban Tourism," 2021).

The study of tourist tracks has been ongoing for several decades. It began in the 1970s and 1980s with the use of mental maps, diaries, surveys, and interviews, and later moved on to non-participant observation in the 1990s. Tourist tracking is essential for destination management and tourism planning. To achieve a balanced approach, multi-method approaches are necessary to weigh the benefits and drawbacks of various tracking methods. Tourists often resist being tracked through their mobile phones due to concerns about privacy invasion. However, they are becoming more receptive to mobile applications that offer location-independent services. These applications are opening new marketing channels in tourism, and tracking technologies are increasingly important (Thimm & Seepold, 2016).

The World Tourism Organization recommends that tourism data be collected, analyzed, and made available systematically to enable better planning and management of tourism activities. Tourism observatories around the world should involve tourism agents and generate information that serves common interests. Moreover, technology has progressed significantly in recent years, resulting in faster internet connections and increased use of mobile devices. As a result, access to information has become more convenient, leading to a greater amount of data being generated compared to previous decades (Aguiar & Szekut, 2019).

Studying tourist mobility is key to developing a sustainable tourism industry transforming and adjusting the offering in tourist destinations according to the demand and patterns analyzed (Versichele et al., 2014). The following table outlines the main characteristics of the data sources used in previous studies of tourist mobility regarding usability, data collection, advantages, and disadvantages (see Table 1).

Table 1 - Data Sources for Research Analysis.

| Data Source | Usability | Data Collection | Advantages | Disadvantages | References |
|---|---|---|---|---|---|
| Global Positioning System (GPS) | Develop a more comprehensive understanding of Tourists' social-spatial patterns. | Utilizing tracking devices that leverage GPS technology. | Can precisely and continuously clarify tourist mobility within destinations. | Covers a small sample size and does not apply to indoor contexts. | (Edwards & Griffin, 2013) |
| Social Media Data | Identifying and grouping people who share similar interests, as well as studying social and economic issues in diverse locations within a city. | Using different social media. | Enable analysis of tourists on a larger scale and uncover rich contextual information (such as travel experiences). | Nonlinear data prevents policymakers from understanding tourists' flow. | (Senefonte et al., 2022) |
| Wi-Fi Positioning Datasets | Studying people's activities, and social activities, describing behavioral patterns. | Difficult to acquire. | Record the location footprints of sizable populations. | Does not collect sociodemographic data about travelers. | (Xu et al., 2021) |
| Bluetooth Tracking | Tracking tourist movements | Through Bluetooth. | This can apply to indoor contexts. | Hard to scale up due to the needed consent of individuals tracked, plus the risk of self-selection bias. | (Versichele et al., 2014) |

### 2.2.1. Global Positioning Systems (GPS) as a Data Source

Tourists' movements can be precisely tracked through Global Positioning Systems (GPS) loggers and GPS-enabled smartphones. GPS units allow researchers to continuously collect high-resolution data that can show movement in time and space over an extended duration. GPS-based tracking has been used in tourism research on numerous occasions, with varying degrees of success and in a variety of ways. However, GPS and spatial science as a data collection method and analysis have been criticized for being too reductionist and neglecting participants' subjectivities. There have also been ethical concerns regarding privacy violations resulting from close tracking movement. Despite this criticism, GPS tracking has many strengths. It is unobtrusive, highly accurate, and precise. It has the potential to provide continuous tracking of individuals, including their velocity and directionality. Moreover, if combined with surveys, it can provide rich information on the travel patterns that different types of tourists take. While GPS tracking is limited by factors such as limited battery life, signal interference, and the necessity for tourists to hold or wear a device, it remains significantly more accurate than data that can be collected through travel diaries or post-travel surveys (Hardy et al., 2017).

### 2.2.2. Social Media as a Data Source

Social media can be classified into seven categories, including forums and message boards, social networks, blogging, microblogging, media sharing, bookmarking, and review websites. Social media data can be used for various research purposes, including analyzing human behavior, public opinions, marketing and political campaigns, and transport planning. The mining process involves using techniques from various scientific domains such as computer science, data mining, machine learning, social network analysis, sociology, statistics, and mathematics (Sdoukopoulos et al., 2018).

Serna et al (2017) identify User Generated Content (UGC) as a valuable source of information for travel behavior analysis, and state that Information and Communication Technologies (ICT) offer opportunities to improve traditional survey methods while identifying sustainability issues related to urban mobility by analyzing UGC using sentiment analysis techniques. Recent research has shown an increase in the use of social media analysis, particularly in the field of urban transport exploring sustainable urban mobility and motivating change towards sustainable mobility choices.

Social media is increasingly being used as a source of data for research, including profiling or understanding users' behavior, demographics, interactions, and sentiments toward topics or products. Automated tools are being used to mine and analyze social media data to reveal new associations or predict future behaviors or outcomes. Using social media in research has several advantages including reaching larger numbers of participants, analyzing trends and associations within large data sets, reducing the costs of research, providing a channel for social research that is less prone to bias, involving citizens in the research process, and generating new channels for research dissemination. Some challenges and ethical considerations arise including the self-selecting nature of social media users, inequalities in access to social media platforms and data, difficulties in obtaining meaning from heterogeneous data, and ethical concerns related to privacy, ownership, and intellectual property (Taylor & Pagliari, 2018).

### 2.2.3. Wi-Fi Positioning as a Data Source

Zhou et al (2016) have studied the importance of understanding urban spatial structure in city planning and governance, with a focus on the use of large-scale mobility data to study human movements and interactions in urban areas (Zhou et al., 2016).

Mobile phone location data has been useful in predicting mobility, urban planning, and transportation research. However, the representativeness of sparsely sampled location data in characterizing human mobility remains unclear. Lu et al (2017) have debated the representativeness and biases of this data, particularly for sparsely sampled mobile phone location data (Lu et al., 2017). Wi-Fi positioning is made possible by using the Received Signal Strength Indicator (RSSI) technique in mobile phones with embedded wireless and open Wi-Fi access points in urban areas. This technique has a spatial resolution of several meters, but it can be improved to within one meter by combining it with the triangulation positioning technique (Wang et al., 2018).

### 2.2.4. Bluetooth Tracking as a Data Source

The use of Bluetooth tracking in the analysis of pedestrian behaviors enables researchers to capture data regarding human mobility in a large-scale dataset without biasing their inner thoughts and covers a longer period (Yoshimura et al., 2017).

This technique can track individual phones but does not provide demographic information or insight into how tourists move. The technology was used, as an example, to track attendees at the Ghent Festivities event and was found to produce more accurate results than mobile phone tracking data when tracing movement over a relatively small site such as a festival. However, this technology is limited to those with Bluetooth-enabled phones. Obtaining mobile phone data directly from telecommunications companies can overcome these limitations but is usually cost-prohibitive (Hardy et al., 2017).

Bluetooth positioning is made possible by the Bluetooth technology present in many mobile phones. With Bluetooth beacon devices installed in certain locations, the position of a mobile phone can be recorded with a time stamp when it enters or passes through such places (Wang et al., 2018).

## 2.3. ANALYZING TOURISTS' BEHAVIOR

Tourist behavior refers to the actions and decision-making processes of individuals or groups of people who are traveling for leisure or business purposes. This can include factors such as destination choice, travel itinerary, transportation methods, lodging, and activities engaged in while on a trip (Yang et al., 2017). Tourists' travel behavior is shaped by a range of factors, including personal and trip characteristics, destination attributes, and travel motivations (Valls Giménez et al., 2020). Secondly, tourist mobility can have both positive and negative impacts on destinations, encompassing economic benefits, environmental degradation, and potential overcrowding or strain on local resources (Holzmann, 2021).

Tourists often have limited knowledge of the destination and rely heavily on information from travel agents, tour operators, and online resources (Juvan et al., 2017). They also tend to have restricted flexibility in their travel plans due to factors such as time, budget, and travel arrangements and are inclined to opt for organized tours and pre-planned itineraries, with a lesser propensity for independent or spontaneous travel. Additionally, tourists' interactions with residents and local culture are typically limited, and they tend to concentrate on tourist-oriented areas and activities (Rusdi et al., 2019).

The concept of tracking travelers' behaviors refers to the collection and analysis of data on how individuals move and travel. Studies previously conducted by researchers have allowed a deeper understanding and prediction of human travel behavior to guide urban planning and transportation development, to provide both long-term guidance and short-term strategies (Wang et al., 2018). Also, according to a study by Xu et al (2021), as mobile phones become more widely adopted and equipped with location and motion sensors, they are becoming increasingly useful in collecting data on human travel behavior. This data can be extensive and dynamic, making it a valuable tool for researchers.

Yao et al. (2021) add that most research on tourist behavior has focused on inter-destination behavior, and less on intra-destination behavior. Intra-destination space refers to an enclosed space with defined boundaries, in which tourist behavior is more controllable. Several methods have been used to analyze tourist behavior within a destination, including traditional methods of live travel diaries and observations, mobile phones, and handheld GPS devices. Previous research has focused on feasibility analysis, tourist behavior characteristics, tourist behavior classification, and tourist behavior prediction. However, existing research still has some shortcomings, such as limited data sources and a lack of investigation into the relationship between indicators during tourists' behavior.

Hamid & Croock (2020), for instance, have analyzed tourist behavior patterns to predict their preferred places using clustering algorithms based on density and POI information collected from GPS data extracted from trajectories and Google Maps. The initial stage of developing a predictive model involved analyzing and preprocessing the dataset to eliminate any potential noise. Afterward, novel characteristics from the dataset were identified and extracted, including stay points and points of interest (POI). For the prediction model, the K-Nearest Neighbors (KNN) method was used to find the nearest tourism places for the tourist based on their GPS position and POIs, allowing the system to recommend places that satisfy their requests using Geopandas to implement this feature in Python (Eslamirad et al., 2023). The system uses real geolocation of tourists from their smartphones to evaluate the nearest distance between them and recommended tourist places. Furthermore, research by Wang et al (2018) has established that human travel behavior displays a high degree of regularity in terms of both spatial and temporal distributions. Other studies have found that individuals tend to have recurrent visits to specific locations with a high frequency, while also exploring other places with a certain probability.

Table 2 summarizes the key features of the data processing methods used in prior research on touristic mobility, including data processing techniques, data source techniques, dataset description, research application, and references found.

Table 2 – Data Processing Techniques from Prior Studies.

| Data Processing Techniques | Data Source | Dataset Description | Research Application | References |
|---|---|---|---|---|
| Codes found through bibliometric analysis | Scopus database | World's largest bibliographic database indexing more than 21,000 titles of international scientific publishers | Tracking Tourists behavior | Padrón-ávila & Hernández-Martín (2020) |
| K-Means Clustering | Flickr | 13 545 photos' metadata from Flickr, such as upload time, location, and user | Evaluate the suitability of photo-sharing platforms, such as Flickr, to gather relevant data on tourists' spatial movement and visitation behavior to POIs | Höpken et al. (2020) |
| Markov Chain | GeoLife dataset | GPS dataset composed of 182 users in a period of 5 years (2007-2012) | Develop a predictive model for tourists' mobility | Hamid & Croock (2020) |
| Clustering Algorithms | Mobile data | 7,989 individuals | Identify activity and types of locations visited | Chen et al. (2016) |

Azman et al. (2021) used a qualitative methodology with semi-structured interviews to identify tourists' spatial behavior and movement patterns, using GPS tracking devices and interviews with tourists. The study focused on tourists to Muar Royal Town in Johor, Malaysia, using purposive sampling with seven respondents. The targeted respondents were domestic tourists of different ages from adjacent districts or states. The collected data helped conceptualize these patterns and provided insights into tourist behavior, which provided direction in transportation planning and future infrastructure development for sustainable development.

Yao et al. (2021) suggest a framework to process open GPS trajectory data, explore indicator relationships, and identify seasonal differences in tourist behavior to address these limitations. The study conducted data analysis in four steps, including gridding the spatial data for efficient processing and visualization, using the spatial join function of ArcGIS to join location and attribute information of trajectory data to the grid and calculate tourist behavior indicators such as path length, tour time, and average speed, visualizing the data using kernel density analysis and identifying hotspots for tourist activity, and correlating analysis and analysis of variance (ANOVA) to understand the relationship between tourist behavior indicators and seasonal differences. Overall, the proposed framework allows for better control of experiments and an understanding of tourist behavior characteristics.

### 2.4. Tourists' Spatial Behavior Clustering and Profiling

Tourist clustering refers to the process of grouping tourists based on common characteristics or behaviors. This can include demographic information such as age, gender, income level, education level, and occupation, as well as travel-related characteristics such as destination preferences, travel purpose, and travel companions, among others. Clustering tourists can provide valuable insights into the different segments of the tourism market and help to inform marketing and management strategies. Furthermore, it can be used to identify patterns in the behavior of specific groups of tourists, which can be used to tailor products and services to meet their specific needs and preferences (D'Urso et al., 2021).

Tourist profiling refers to the process of creating a detailed description of a specific group of tourists based on a variety of demographic and behavioral characteristics identified in the clustering process. The profiling process can be based on both quantitative and qualitative data. After identifying the clusters, it is possible to create detailed descriptions of each group by using additional data that was not used in the initial clustering process. This can be accomplished by calculating weighted averages for quantitative variables and weighted proportions for qualitative variables (D'Urso et al., 2021).

The task of clustering and profiling tourists' behavior can present several challenges, as highlighted by D'Urso et al. (2021). One of the major challenges is obtaining accurate and reliable data on tourists' behavior, especially if they are not forthcoming with their personal information. To overcome this, it may be necessary to utilize multiple sources of data to gain a more comprehensive view of tourists' behavior. Another challenge is the diversity of tourists, who come from varying backgrounds and cultures and have different motivations for travel. As a result, creating meaningful clusters or profiles can be difficult. Advanced clustering techniques, such as mixed-membership models or hierarchical clustering, may be employed to address this issue. Privacy concerns may also arise when clustering tourists' behavior. It is essential to collect and use data ethically, and measures such as implementing strict data privacy policies and utilizing anonymization techniques should be taken to protect tourists' personal information. Furthermore, the limited nature of the data used for clustering and profiling can lead to inaccurate or incomplete results. Techniques such as data augmentation could potentially be used to overcome this issue.

Shou & Di (2018) presented a new methodology for the analysis of activity patterns using multi-person multi-day data. The authors used frequent sequential pattern mining to extract frequently occurring activity sequences and calculated similarity measures between travelers' activity patterns to group them into communities. They also used multinomial logistic regression to examine the relationship between socio-demographics and travel patterns. This methodology can be used to infer demographics based on activity patterns or to reconstruct activity patterns based on demographics and similar travelers' patterns.

Research regarding tourists' spatial behavior clustering mostly aims to identify hotspots visited by tourists. To perform a density-based spatial clustering approach for tourists' behavior, Rodríguez-Echeverría et al. (2020) included reverse geocoding to obtain municipality names for origin and destination locations, computing staying time for each trip, and converting destination locations to Universal Transverse Mercator (UTM) coordinates for distance measurement. The study recurred to a density-based clustering algorithm called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to find places visited by tourists regardless of how many times they were visited, and the Ordering Points to Identify the Clustering Structure (OPTICS) algorithm to handle the problem of heterogeneous cluster density.

On a different note, Ruiz Reina (2021) recurred to unsupervised learning using hierarchical clustering, also described as an agglomerative method. The clustering process involved three steps: data preprocessing, uncertainty modeling using Shannon's Entropy, and cluster processing based on spatial, temporal, and proximity criteria.

To pursue further research on clustering Covid-19-driven travel behavior of rural tourists in South Tyrol, Italy, Scuttari et al. (2021) used Statistical Package for the Social Sciences (SPSS) software for data analysis, including descriptive statistics, principal component analysis (PCA), and cluster analysis. Also, chi-square and z-tests were used to explore associations between categorical variables. They used the two-step clustering method to identify different types of tourists based on their reactions to the pandemic in terms of mobility behaviors, considering sociodemographic patterns and PCA factors. The final solution was chosen based on the predictors' importance, the number of clusters, and interpretative strength.

There has always been a relationship between geography and tourism. The behavior of humans in time and space is influenced by various physical, social, and economic factors. Tobler's First Law of Geography which states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1995) can also be applied in the tourism context. Švajda et al. (2018) mention the concept of friction of distance, describing the effect of distance on cultural or spatial interactions. The author noted that research into tourists' attitudes and behavior, particularly in mountain environments, is becoming more popular, with studies analyzing the time-space behavior of tourists through tracking technologies. Furthermore, his research aimed to understand the relationship between time-space constraints and spatial patterns, as well as to inform sustainable development and destination management efforts through the creation of innovative tourist typologies. The study was conducted over three days in the Priehyba area – Chopok – in 2017, using a combination of GPS loggers and structured questionnaires to investigate the spatial-temporal behavior of tourists and their characteristics. The GPS data was analyzed using the ArcGIS programming environment and the Kernel density algorithm to calculate the tourists' collective behavior. The analysis found three clusters based on two nominal variables: sport motivation and social motivation. The differences among the clusters were highly significant and could be shown either in tables or graphs.

Deeper into the research of profiles concerning tourists' behavior, Panapakidis et al. (2015) wrote an article in which they discussed load profiling in future smart grids, where every consumer is connected through a smart meter. Load profiling involves extracting typical load curves for single consumers and groups of consumers and grouping them based on similarity. The article explores different clustering algorithms and pattern representation techniques, as well as dimensionality reduction techniques to enhance clustering performance. The optimal approach is found to be {representation technique, algorithm, type of first-stage load profile} = {PCA, KM2, Ip1}.

## 2.5. Summary of Findings

Studies on smart city management recurring to big data (Amović et al., 2021), research on behavior using smartphone usage data (Rusdi et al., 2019), and studies on touristic mobility patterns (Xu et al., 2021) and (Aranburu Amiano et al., 2020), reinforce that tourists' behavior is dynamic, and it changes over time, necessitating constant monitoring and adaptation of clustering and profiling techniques. Regular monitoring and updating of techniques can help adapt to changes in tourist behavior. The advent of information and communication technologies has led to the emergence of mobile phone data as a valuable tool for gathering extensive and dynamic information on tourists' mobility patterns.

To gain a comprehensive understanding of the underlying principles governing travel behavior, a wide range of influential factors have been scrutinized. The technologies employed to process and analyze mobile phone data vary depending on the type and source of the data, as well as the research objectives, encompassing both traditional data mining techniques and innovative computational methods. By studying tourists’ mobility patterns and behavior, researchers can identify clusters of tourists and their shared characteristics based on similar behaviors and traits.

The use of mobile phone data is still in its early stages, and researchers are still focusing on data practices rather than making original contributions. The limitations of mobile phone data include intrinsic shortages in data attributes, sampling bias, and noise, which require significant pre-processing efforts. Mobile phone data are rich in spatial-temporal information but limited in socio-demographic attributes, which limits the exploration of determinants that lead to observed travel patterns (Wang et al., 2018).

Overall, several previous studies have investigated the topic of clustering tourists’ behavior based on their social media posts among several social media networks (Höpken et al., 2020). Also, previous studies have been conducted on clustering destination attractions for different origin markets through the special analysis of online reviews across different review platforms (Kirilenko et al., 2019). Apart from these studies, and while not specifically related to urban destinations, studies on internet access behavior have also contributed to knowledgeable insights on these related topics (Zulfadhilah et al., 2016). A comparative analysis of these studies reveals notable similarities and differences in terms of research methodology, clustering algorithms, and evaluation techniques used.

In terms of research methodology, previous studies have employed various approaches such as K-Means, DBSCAN, OPTICS, and PCA-MAP clustering algorithms, followed by evaluation metrics such as Silhouette Score, DBI, CHI, and visual representations of clusters. While some studies focused on understanding spatial behavior and movement patterns, others focused on identifying internet user behaviors. It is important to note that the choice of methodology can influence the outcomes and generalizability of the findings and should consider the nature of the data collected.

Furthermore, the variation in sample characteristics across studies should be considered. Previous studies have targeted different locations, populations, and considered different variables for analysis. It is worth emphasizing that the limitations of data sources vary among the previous studies found resulting in different outcomes.

Regarding the key findings, previous studies have successfully identified distinct clusters of tourists based on their travel preferences and mobility patterns using social media data analysis. Research of this sort has uncovered specific points of interest visited by each cluster. Additionally, Orama et al., (2022) have explored tourists’ mobility within clusters, identifying interconnected points of interest and patterns of movement. Many studies highlighted inconclusive results, highlighting the need for further investigation or replication.

While several studies have provided valuable insights into the clustering of tourist behavior and the clustering of internet access behavior, there are certain gaps when combining these two research topics, meaning that studies regarding clustering tourists behaviors or behaviors identified on specific grids in urban areas based on their internet accesses have not been found. By addressing this limitation, the present study aims to contribute to the existing body of knowledge and provide a more comprehensive understanding of visitors' behaviors while accessing the internet in Lisbon's urban area.

# 3. METHODOLOGY

## 3.1. Research Design

The research project will follow Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology which is a common model for data mining used in many projects. CRISP-DM is a widely used reference model for data mining and analytics projects. It provides a structured approach to the entire data mining process, from understanding the business problem to deploying the solution. The different stages involved include the six phases of the data mining life cycle which are not strictly sequential and can be revisited depending on the project's needs. The outcomes of each phase determine the next steps in the project. The outer circle in the model represents the iterative nature of data mining, where lessons learned from the process and deployed solutions can lead to new business questions.

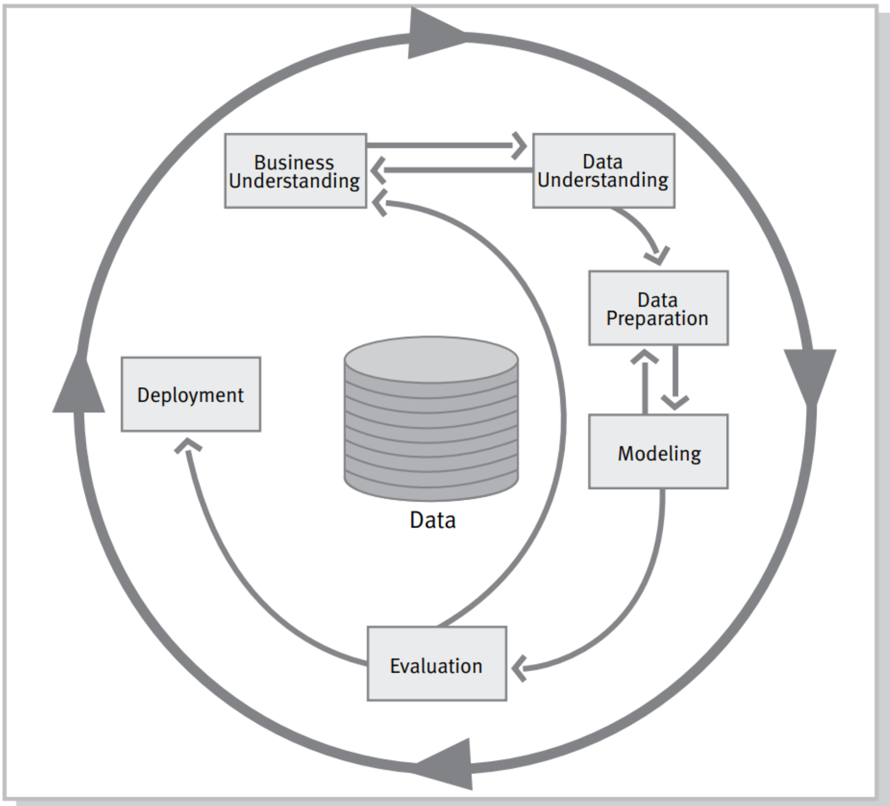


Figure 2 - CRISP-DM Reference Model for Data Mining in Visitor's Clustering.

The CRISP-DM reference model provides a flexible and iterative approach to data mining projects, which allows for changes and adjustments to be made throughout the process. This model is widely used in various industries and has become a standard for data mining projects.

The six main phases of the data mining process are business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

The business understanding phase in the CRISP-DM model is the first phase of the data mining process. It involves understanding the project objectives, requirements, and constraints from a business perspective. The goal of this phase is to establish a clear understanding of the business problem or opportunity and define the data mining goals that align with it. The key activities included in the business understanding phase are determining the project objectives, analyzing the situation and context, defining data mining goals, and producing a project plan.

The data understanding phase in the CRISP-DM model involves exploring and familiarizing yourself with the available data. The goal of this phase is to gain an understanding of the data, assess its quality, and identify any initial insights or patterns that may be relevant to the data mining goals. The main activities included in the data understanding stage are data collection getting acquainted with the data available, exploring data through preliminary exploratory data analysis (EDA) to understand the distribution, summary statistics, and relationships among variables, verifying data quality, identifying issues or challenges, documenting findings and observations, and lastly preparing a detailed description of the data, including metadata, data sources, and any data transformation.

The data preparation phase in the CRISP-DM model involves transforming and organizing the data to make it suitable for analysis. This phase focuses on preparing the data in a way that facilitates effective data mining and modeling considering the main objectives previously outlined. The activities included in this stage are data selection, data cleaning including dealing with missing values, outliers, inconsistencies or errors on the data, data integration with multiple datasets, if necessary, data transformation, feature engineering, and dimensionality reduction through principal component analysis (PCA).

In the modeling phase of the CRISP-DM model, unsupervised learning techniques are employed to discover hidden patterns, structures, or relationships within the data. While supervised learning algorithms are utilized when labeled training data is available, unsupervised learning methods are essential for exploratory analysis and pattern discovery in situations where labeled outcomes or targets are lacking.

Unsupervised learning algorithms, particularly clustering algorithms, are used to group similar data instances together based on their intrinsic characteristics. This helps identify clusters or segments within the data, revealing underlying patterns or subpopulations. Additionally, dimensionality reduction techniques, such as principal component analysis (PCA), are applied to reduce the data's dimensionality and facilitate visualization and further analysis.

During the modeling phase, the chosen unsupervised learning techniques are employed to uncover clusters, reduce dimensionality, and extract meaningful patterns. The insights gained from these techniques are evaluated, validated, and refined to ensure their effectiveness in addressing the data mining goals. This evaluation involves assessing the quality of the discovered clusters or patterns, evaluating the effectiveness of dimensionality reduction, and validating the relevance of the insights derived through unsupervised learning.

By incorporating unsupervised learning within the modeling phase of the CRISP-DM model, this research project gains the ability to discover hidden patterns, insights, or subpopulations within the data without relying on labeled training data. This approach allows for a comprehensive exploration and understanding of the data, providing valuable insights that contribute to the research objectives.

The evaluation phase in the CRISP-DM model involves assessing the results and effectiveness of the models developed in the modeling phase. It aims to determine whether the models meet the business objectives, provide valuable insights, and are suitable for deployment in a real-world environment. The evaluation phase helps to ensure that the developed models are reliable, accurate, and effective in addressing the data mining goals.

Finally, the deployment phase in the CRISP-DM model involves developing models or insights into operations within a production environment. It focuses on integrating the models into the existing systems and processes to enable real-world utilization and generate value from the data mining industry.

# 4. EMPIRICAL STUDY

## 4.1. Business Understanding

Understanding the business surrounding this specific project involves analyzing the study area, determining the business objectives, and determining the data mining objectives.

### 4.1.1. Study Area

Lisbon urban area was the area chosen for research regarding the behavior of visitors when accessing the internet in the Portuguese capital, to pursue clustering. Lisbon, the capital city of Portugal, boasts a population exceeding 500.000 people, as indicated by the initial findings of the recent Portuguese census administered by the National Institute of Statistics (INE, 2001).

Assuming that visitors in Lisbon are likely to access the internet frequently, this represents an opportunity to collect data on their behaviors. Visitors may use the internet to find information on local attractions, plan their itineraries, or communicate with fellow travelers. By examining these patterns of internet usage, we can gain a better understanding of which areas are the most popular, times of permanency, and how they cluster and interact with each other.

While incorporating data on internet usage, researchers can gain valuable insights into the behavior and preferences of different groups of visitors, which can inform strategies for improving internet services, city management, tourism management, and promoting sustainable tourism development in the city. The Lisbon urban area is highlighted below with the grids outlined and the monuments' locations identified. Figure 3 illustrates the grid's geographic location concerning the city of Lisbon.

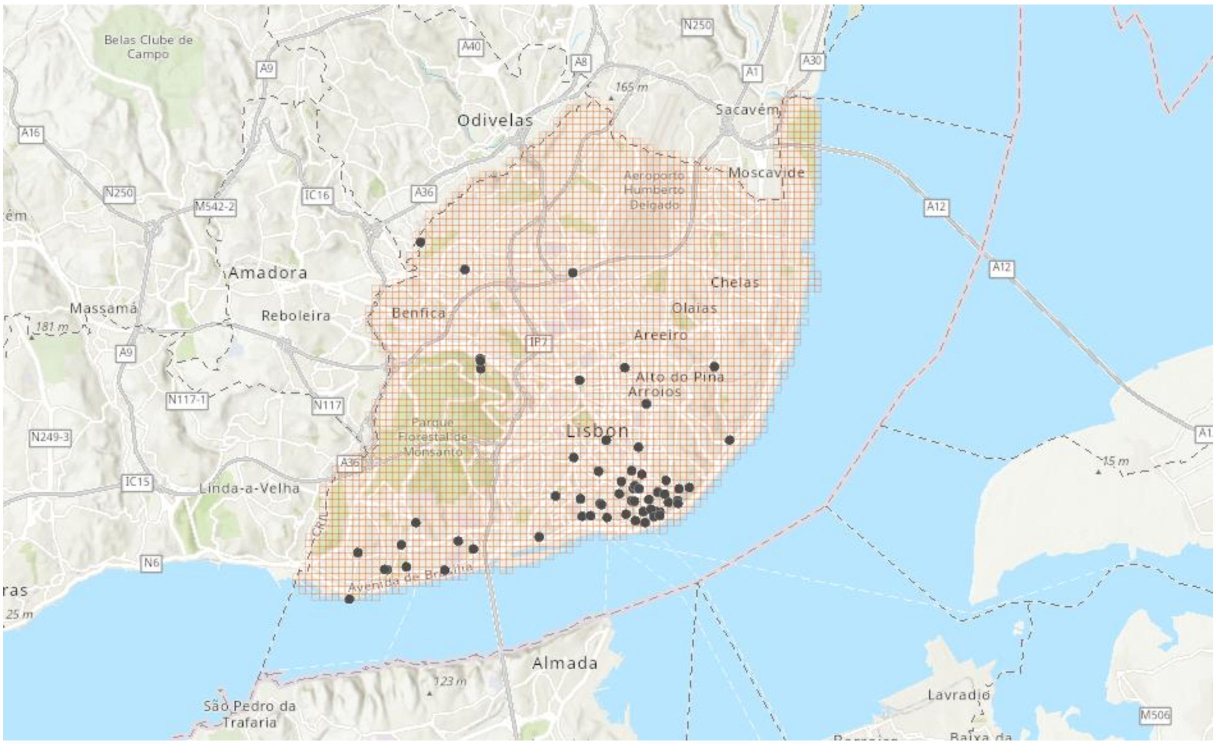


Figure 3 - Map of Lisbon with Highlighted Data Grids for Analysis.

Moreover, to enhance the precision of insights derived from the grids and narrow down the study area, a selection was made of the primary parishes where significant monuments are situated. Consequently, the scope of the study became more focused. Among the various parishes originally included in the initial datasets, the parishes selected for subsequent analysis were Belém, Alcântara, Estrela, Santa Maria Maior, Misericórdia, Ajuda, São Vicente, and Santo António. For a visual representation of the final study area, please refer to Figure 4.

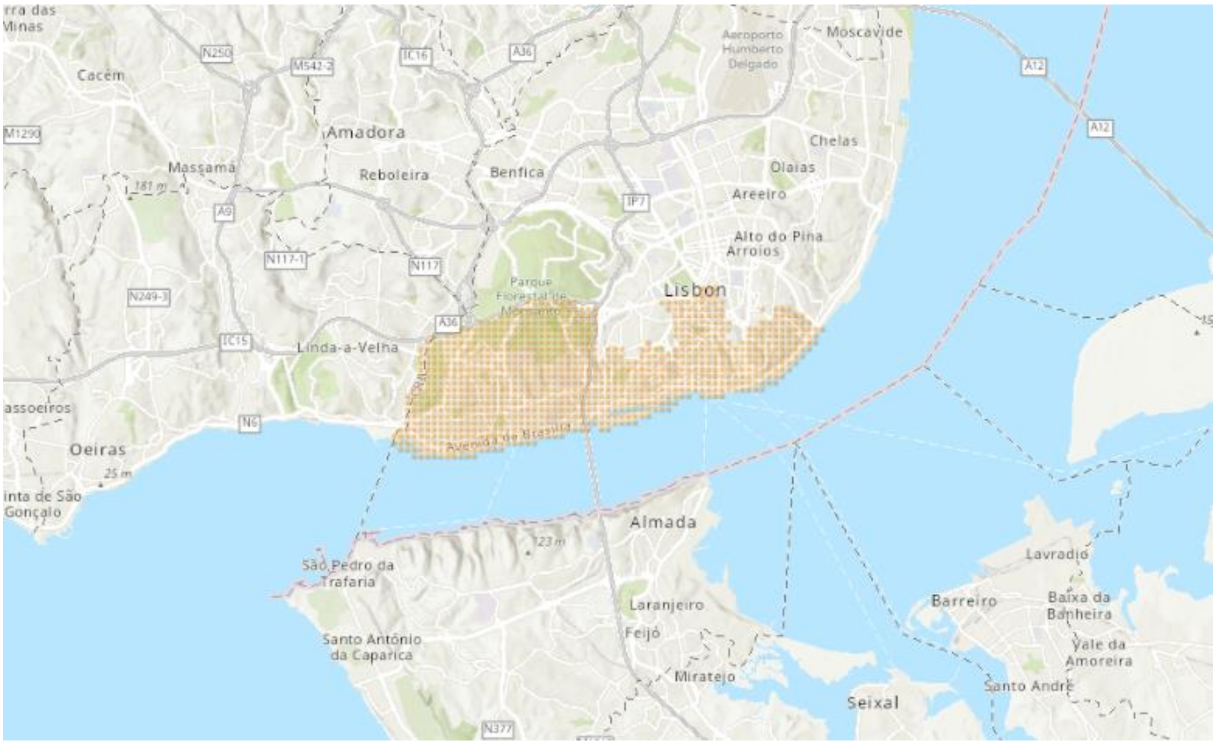


Figure 4 – Final Study Area in Lisbon with Highlighted Parishes for Analysis.

### 4.1.2. Objectives

The primary objective of this master's thesis is to gain a comprehensive understanding of people's behaviors in accessing the internet during their visits to the eight main visited parishes within the Lisbon urban area while grouping grids into clusters with similar behaviors. Studying how individuals engage with the internet while visiting specific locations can provide valuable insights into user preferences, usage patterns, and the impact of digital technologies on urban environments.

The main objectives of my thesis are as follows:

Analyze internet usage patterns.

1. This analysis will encompass various aspects, including the frequency of internet access, nationalities, understanding parishes with similar behaviors, understanding peaks of internet accessibility by month, weekdays, and hours of the day, among others.

Provide insights for urban planning and digital infrastructure development.

2. Based on the findings of this research, I will offer practical recommendations and insights for urban planners, policymakers, and stakeholders involved in the development of digital

infrastructure within the Lisbon urban area. These recommendations will focus on optimizing internet accessibility, improving digital services, and creating an environment that caters to the diverse needs and preferences of residents and visitors.

Contribute to the broader understanding of the digital landscape in urban environments.

3. By studying people's behaviors regarding internet access in the context of urban visits, this thesis aims to contribute to the existing body of knowledge on the impact of digital technologies on urban spaces. The research findings will provide valuable insights into the role of internet access in shaping urban experiences, informing urban planning strategies, and fostering the development of smart and connected cities.

## 4.2. DATA UNDERSTANDING

To conduct data mining for this master's thesis, a substantial volume of data was required for subsequent analysis. Furthermore, comprehending the available data entails more than simply its collection. It necessitated describing, exploring, and validating the quality of the data.

### 4.2.1. Data Collection

The present study utilizes available data from Wi-Fi networks as the primary source of data. The data was collected within the Lisbon urban area over a span of 12 months, specifically from September 2021 to August 2022. The datasets encompass various information, including the device used for the Wi-Fi connection, location, date of access, and other relevant details outlined in Table 3. Vodafone was responsible for collecting the data by accessing the location of users' mobile devices.

#### 4.2.1.1. Internet Access Data

Information on internet access points and user characteristics was provided by Vodafone through a direct agreement with Nova Cidade Urban Analytics LAB at Nova Information Management School. Originally, the data was gathered to furnish the City Council of Lisbon (CML) with anonymized and aggregated insights into mobile terminal movements within the city on a 200x200 grid, at 5-minute intervals, to facilitate analysis.

For data visualization and comprehension purposes, data concerning points of interest, such as monuments, within the Lisbon urban area was also collected. This dataset, obtained from Gestor Geodados of the City Council of Lisbon (CML), comprises a feature layer containing precise information regarding the location of these monuments. This data proved instrumental in determining which parishes would be included within the study area.

In total, the initial dataset encompasses 95 individual datasets, including the data collected from internet access in CSV format, a shapefile containing information on the locations of the 200x200 grids, and a shapefile containing the locations of heritage monuments within the Lisbon urban area. A detailed description of the attributes collected on each grid regarding internet access identified is given in Table 3.

Table 3 - Description of Variables in Internet Access Observations Datasets.

| Variable | Description |
|---|---|
| Grid_ID | Unique grid identifier |
| DateTime | Date of caption |
| Extract_year | Year of caption |
| Extract_month | Month of caption |
| Extract_day | Day of caption |
| C1 | Number of distinct terminals in the grid |
| C2 | Number of distinct terminals, roaming, in the grid |
| C3 | Number of distinct terminals that remained in the grid |
| C4 | Number of distinct terminals that remained in the grid, roaming |
| C5 | Number of different terminal inputs in the grid |
| C6 | Number of outputs of distinct terminals in the grid |
| C7 | Number of different terminal entries, roaming, in the grid |
| C8 | Number of different terminal outputs, roaming, in the grid |
| C9 | Number of distinct terminals with active data connection, in the grid |
| C10 | Number of separate terminals with an active data connection, roaming, in the grid |
| C11 | Number of voice calls from the grid |
| D1 | Top 10 countries of origin of roaming terminal devices |
| E1 | Number of voice calls that ended in the grid |
| E2 | The average downstream rhythm of the grid, in the grid |
| E3 | Average grid upstream rhythm |
| E4 | Quad downstream peak rhythm |
| E5 | Quadruple peak rhythm |
| E6 | Top 10 Applications (Text with names separated by;) |
| E7 | Duration of minimum stay within the grid |
| E8 | Duration of the average stay in the grid |
| E9 | Duration of maximum stay within the grid |
| E10 | Number of devices that share the connection in the grid |

#### 4.2.1.2. Grids Data

The dataset about the 200x200 grids includes a comprehensive set of attributes, which are described in Table 4. This table offers a detailed account of the specific characteristics and properties associated with the data points contained in the dataset. By providing a structured overview of the attributes, Table 4 facilitates a more thorough analysis and interpretation of the dataset, enabling researchers to gain deeper insights into its underlying patterns and dynamics.

Table 4 – Description of Variables in Grid Location Dataset for Study Analysis in Lisbon.

| Variable | Description |
|---|---|
| WKT | The well-known text representation of geometry |
| Latitude | Centery grid latitude (WGS84 coordinate system) |
| Longitude | Square centrode longitude (WGS84 coordinate system) |
| Grelha | Square size 200x200 |
| ID | Unique grid identifier |
| Grelha_Per | Grid perimeter = 800 m |
| Grelha_Are | Grid area = 40000 m2 |
| Freguesia | Designation of the current parish of the centroid of the square |
| Freguesias | Designations of the former parishes of the centroid of the square |
| Nome | The common designation of the area characterized by the grid |
| Grelha_y | Alternative Y-axis coordinate, with integers starting at 0 |
| Grelha_x | X-axis alternative coordinate, with integers starting at 0 |
| Discover | Unique identifier of the parish |

#### 4.2.1.3. Patrimony Data

Lastly, the dataset contains information regarding the patrimony in Lisbon's urban area. This data was provided in a shapefile usable afterward in ArcGIS aiming to understand which are the most visited places and the locations with the highest affluence, as well as to outline new emerging areas which could potentially start showing increasingly higher demand on visitors' end. The shapefile regarding Lisbon patrimony has the attributes outlined in Table 5.

Table 5 – Description of Variables in Points of Interest Location Dataset in Lisbon. [1]

| Variable | Description |
|---|---|
| X | Latitude |
| Y | Longitude |
| IDTIPO | Point of interest ID |
| INF_NOME | Name of the point of interest |
| INF_MORADA | Point of interest address |
| INF_TELEFONE | Point of interest phone number |
| INF_FAX | Point of interest fax number |
| INF_EMAIL | Point of interest email address |
| INF_SITE | Point of interest website |
| INF_DESCRICAO | Description of the point of interest |
| INF_MUNICIPAL | Municipality of the point of interest |

### 4.2.2. Data Description

The analysis of the dataset revealed several noteworthy observations. Firstly, there were a total of 6,331,448 observations available for analysis. The final dataset covered a range of 7,581 observations, providing a comprehensive temporal perspective. In terms of nationality, the analysis encompassed 146 different nationalities. The study area comprised eight distinct parishes, contributing to the geographical diversity of the dataset.

While examining the variables, the maximum number of terminals detected simultaneously in a single grid was 12,247, with the minimum ranging between 0 and 1. Similarly, the maximum number of terminals roaming in a grid at the same time was 12,247, with the minimum also ranging from 0 to 1. The variable representing voice calls exhibited interesting patterns as well. The maximum number of voice calls detected simultaneously in a grid was 2,393, while the minimum was 0. Additionally, the maximum number of voice calls made at the same time in a grid was 778, with a minimum of 0. In terms of duration, it was observed that the minimum stay in a grid was recorded as 0, indicating instances of very brief visits. On average, individuals spent approximately 6.31 minutes in a grid during their visits. The dataset also included instances of longer stays, with the maximum duration recorded as 300 minutes.

These observations shed light on the dataset's characteristics and provide valuable insights for further analysis. By understanding the distribution of variables and the patterns they exhibit, researchers can delve deeper into exploring the behavior and dynamics within the dataset.

[1] source: https://geodados-cml.hub.arcgis.

### 4.2.3. Exploratory Data Analysis (EDA)

To comprehensively understand the data's structure and ascertain the variables that hold relevance for subsequent analysis, extensive Exploratory Data Analysis (EDA) techniques were employed. This phase commenced with conducting descriptive statistics to summarize the data's key features. Subsequently, various visualization methods such as charts, graphs, and maps, were employed to present the data in a more accessible and comprehensive manner. These visual representations serve to facilitate a deeper understanding of the dataset and provide valuable insights into its underlying characteristics.

By observing Figure 5 it is possible to understand the months where most accesses to the internet were registered. While September 2021, October 2021, November 2023, and January 2022 registered the highest number of internet accesses, February 2022, March 2022, June 2022, and August 2022 registered the lowest number of accesses to the internet.

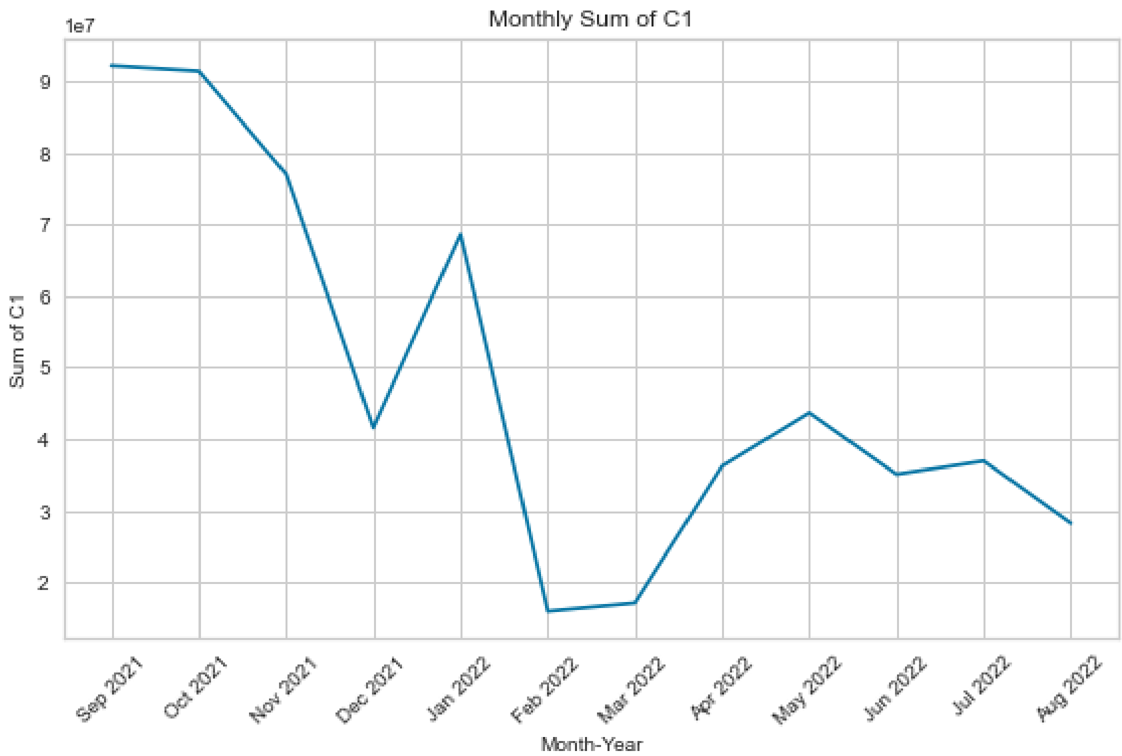


Figure 5 – Monthly Internet Accesses Graph.

Furthermore, analyzing the weekdays that registered the most affluence in internet access allowed to understand that Monday to Friday are the days of the week that register most of the accesses possibly meaning that most affluence to the urban area of Lisbon is registered on these weekdays. Also, the fact that these are business days could help justify this observation (see Figure 6).

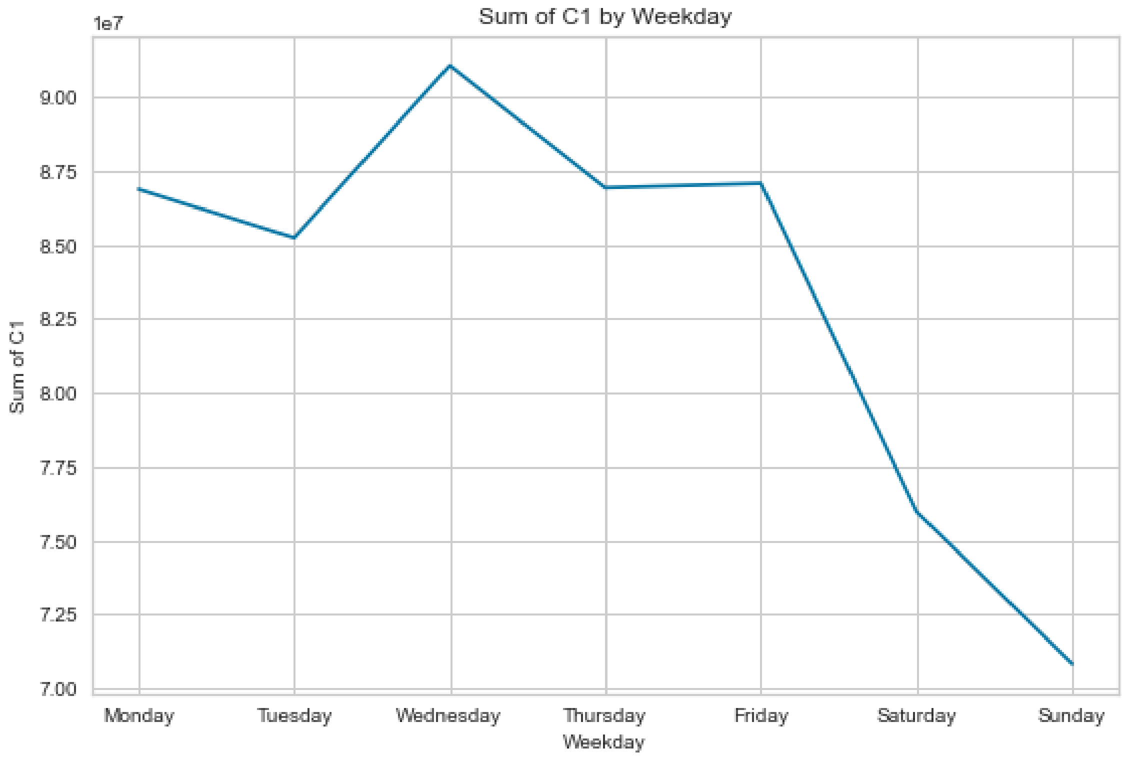


Figure 6 - Internet Accesses by Weekday Graph.

For a deeper understanding of the affluence of internet access, hours of the day were also analyzed. There was found a peak between 08h00 and 09h00 possibly justified by the time most people start their day jobs in urban areas (see Figure 7).

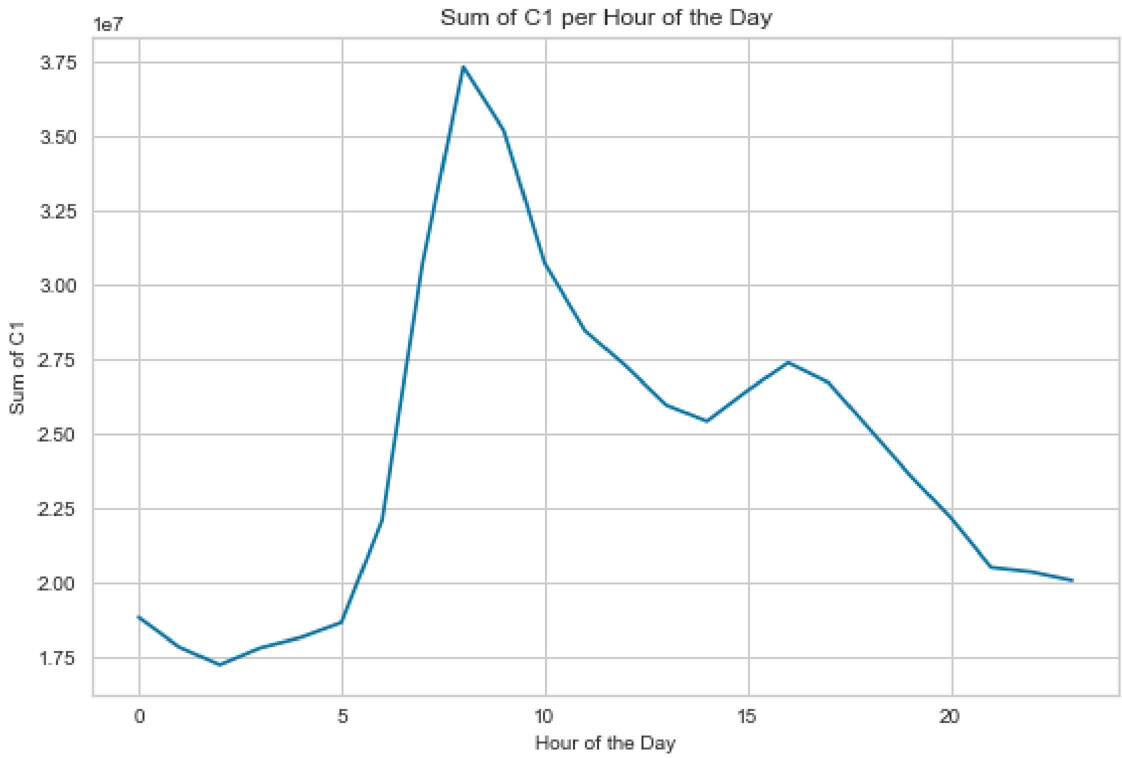


Figure 7 - Internet Accesses by Hour of the Day Graph.

Lastly, understanding the distribution of internet access by hour of the day combined with the day of the week was also an on-demand observation. To better comprehend these patterns, the heatmap from Figure 8 was created (see Figure 8).

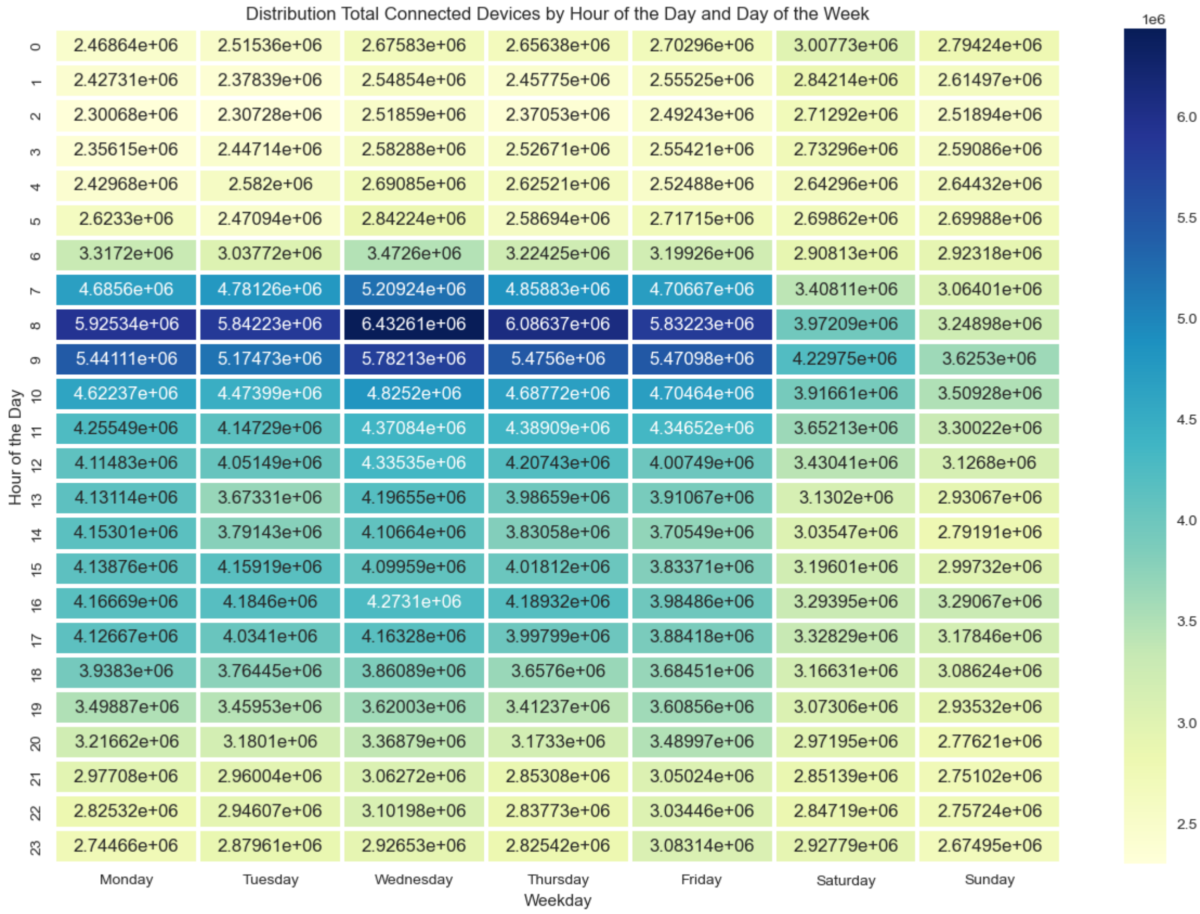


Figure 8 - Heatmap of Internet Accesses by Hour and Weekday.

By combining these two variables, it is possible to conclude that the times when most internet accesses are registered are comprehended from Monday to Thursday, between 07h00 and 10h00. It is also notable a slight increase in internet access around 16h00 on all weekdays, particularly on Wednesdays. In contrast, the periods from 21h00 to 05h00 are the ones with the lowest registration of internet access. Plus, both on Saturdays and Sundays the affluence in internet access is also lower than on other weekdays.

Furthermore, the top 10 nationalities that exhibited a remarkable presence of connected devices, signifying a strong technological presence and high levels of internet usage, were Spain, France, United Kingdom, Germany, Brazil, Italy, Switzerland, Belgium, Poland, and Ireland (see Figure 9).

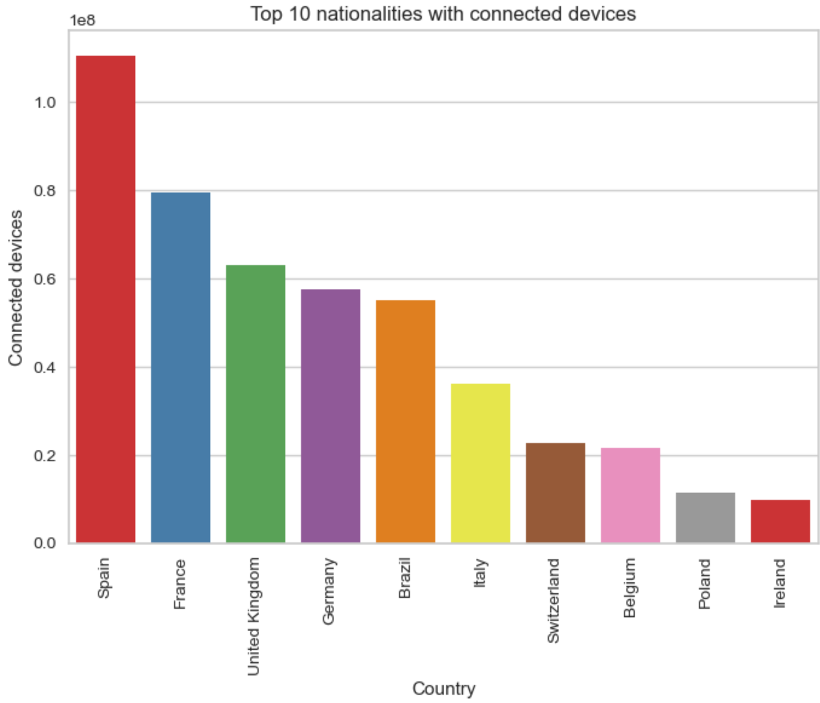


Figure 9 - Connected Devices by Top 10 Nationalities.

Conversely, the bottom 10 nationalities, included Mali, Nepal, Lesotho, Brunei Darussalam, Guinea-Bissau, Senegal, Uzbekistan, Democratic Republic of Congo, Faroe Islands, and Kosovo (see Figure 10).

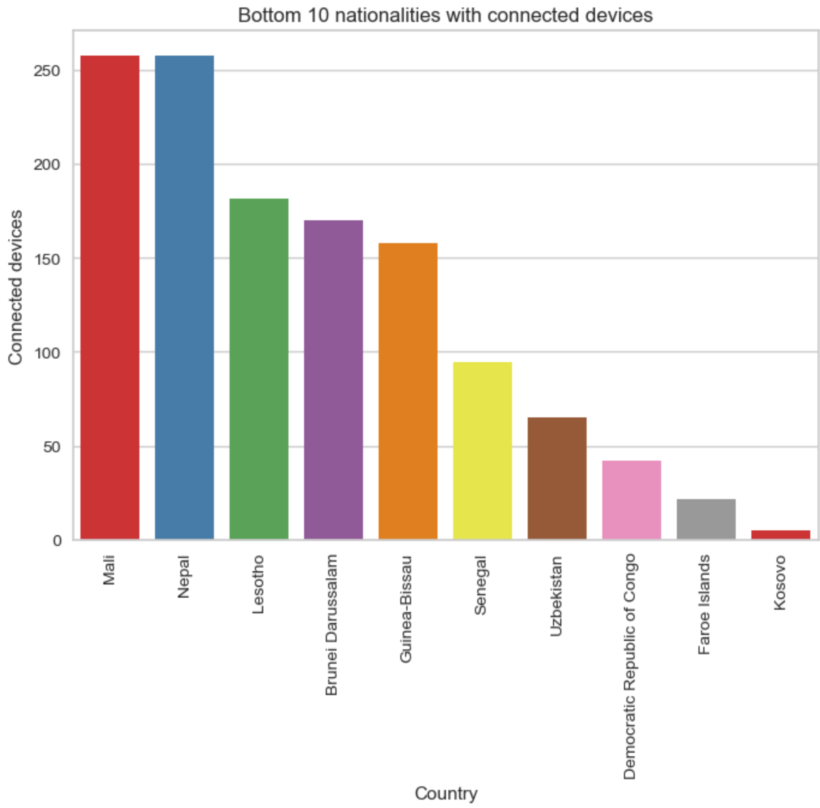


Figure 10 - Connected Devices by Bottom 10 Nationalities.

A thorough examination of the cardinality of parishes within the study area was conducted to gain insights into the diversity and representation of different geographic regions. The analysis revealed that expression for each parish included in the study: Belém, Alcântara, Estrela, Santa Maria Maior, Misericórdia, Ajuda, São Vicente, and Santo António. Each parish contributes uniquely regarding cultural heritage, socio-economic dynamics to the urban landscape of Lisbon, and unique characteristics. Understanding the cardinality of parishes allows for a more comprehensive exploration of the spatial distribution and influence of these areas on the research objectives. By encompassing a diverse range of parishes, this study ensures a comprehensive analysis of the urban area and provides a robust foundation for investigating various aspects of behavior and access to the Internet within the specific context of Lisbon. The parish with the most observations is Belém, and the one with the least observations is Santo António (see Figure 11).

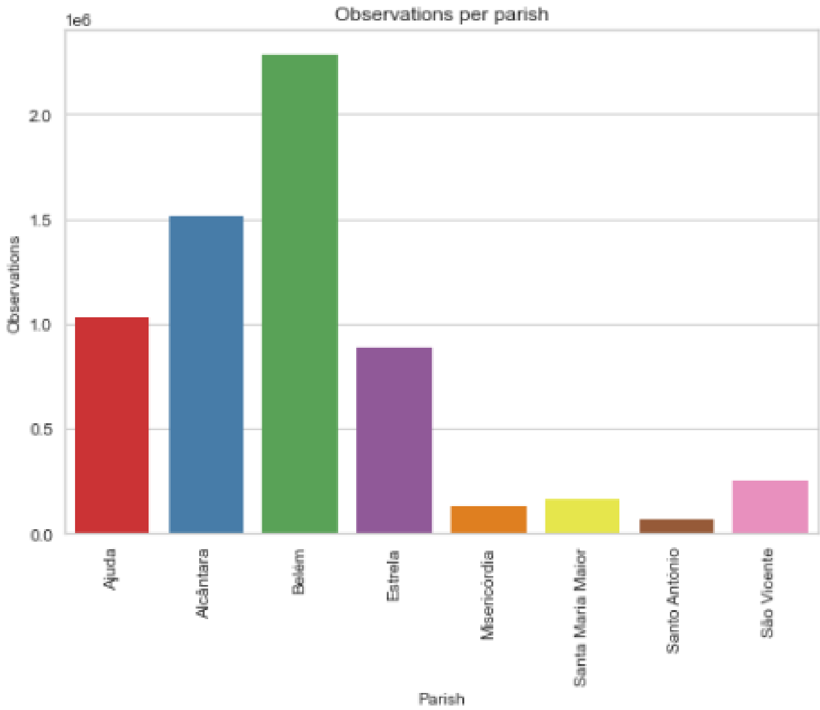


Figure 11 - Cardinality per Parish Graph.

## 4.3. Data Preprocessing

Data preparation encompasses a series of crucial steps, including data selection, cleaning, integration, and formatting, which are essential for ensuring the quality and integrity of the dataset used in this thesis project. To facilitate the data preparation process, a Python notebook utilizing the Pandas open-source library was created.

While understanding the data available, some issues were encountered that required further data preparation. To execute these data preparation tasks efficiently, a Python notebook utilizing the Pandas library was employed.

### 4.3.1. Data Cleaning

The dataset underwent rigorous scrutiny, excluding irrelevant information and refining the dataset to primarily include key variables related to tourist behavior within specific grids. These variables encompassed Grid ID, datetime, number of distinct terminals, number of voice calls, top 10 countries of origin for roaming terminal devices, number of voice calls ending in the grid, number of distinct terminals roaming in the grid, minimum stay duration, average stay duration, and maximum stay duration. At first, the top 10 accessed applications variable was also included, however, due to the number of missing values, was afterward excluded from the dataset.

The final dataset used for further analysis contains information on the variables described in Table 8.

Table 6 - Variables Used for Clustering in the Study.

| Variable | Description |
|---|---|
| Observation_ID | Unique observation identifier |
| Grid_ID | Unique grid identifier |
| Weekday | Weekday of caption |
| Hour | Hour of caption |
| Freguesia | Parish where the observation was detected |
| C1 | Number of distinct terminals in the grid |
| C2 | Number of distinct terminals, roaming, in the grid |
| C11 | Number of voice calls from the grid |
| D1 | Top 10 countries of origin of roaming terminal devices |
| E1 | Number of voice calls that ended in the grid |
| E6 | Top 10 Applications (Text with names separated by;) |
| E7 | Duration of minimum stay within the grid |
| E8 | Duration of the average stay in the grid |
| E9 | Duration of maximum stay within the grid |

### 4.3.2. Correlation Analysis

Spearman correlation was chosen as the appropriate statistical method for analyzing the relationship between the variables due to the unique characteristics of the data. The dataset employed in this study contained outliers, missing values, and exhibited a nonlinear relationship between some of the variables of interest. Spearman correlation was well-suited for this analysis as it does not rely on the assumption of linearity and is less sensitive to outliers compared to Pearson correlation. This choice was further justified by the robustness of Spearman correlation in handling missing values and its ability to rank the data, thus accommodating the ordinal nature of some variables. Furthermore, previous studies in the field have successfully utilized Spearman correlation for similar analyses, ensuring consistency with established practices (Liu & Ban, 2013).

The correlation between variables was studied to understand the relationship and dependency between them. Correlation analysis helps uncover patterns, associations, and dependencies between different variables in a dataset. The correlation coefficient varies between -1 and +1, and 0 indicates no correlation. Strong correlations are correlations with coefficients between -0.8 and -1, and between +0.8 and 1. On the other hand, weak correlations between variables are represented by coefficients between -0.2 and 0, and between 0 and +0.2 (see Figure 12).

According to the heatmap in Figure 12, some variables are correlated. Variables E1 and C11 are slightly correlated with a coefficient of +0.46, variables E9 and C1 are somewhat correlated with a coefficient of +0.56, and variables E9 and E8 are correlated with a coefficient of +0.64. These positive correlations mean that as one variable increases, the other variable tends to increase as well.

The analysis of the relationships between variables in the dataset revealed important insights. A positive correlation was observed between variables E1 and C11, as indicated by the upward trend of the data points. The tight clustering around the trend line indicated a strong correlation, although there were a few outliers suggesting anomalous data points. The computed correlation coefficient of +0.46 confirmed a slight positive correlation between E1 and C11. Further investigation of the outliers may be warranted (see Figure 12).

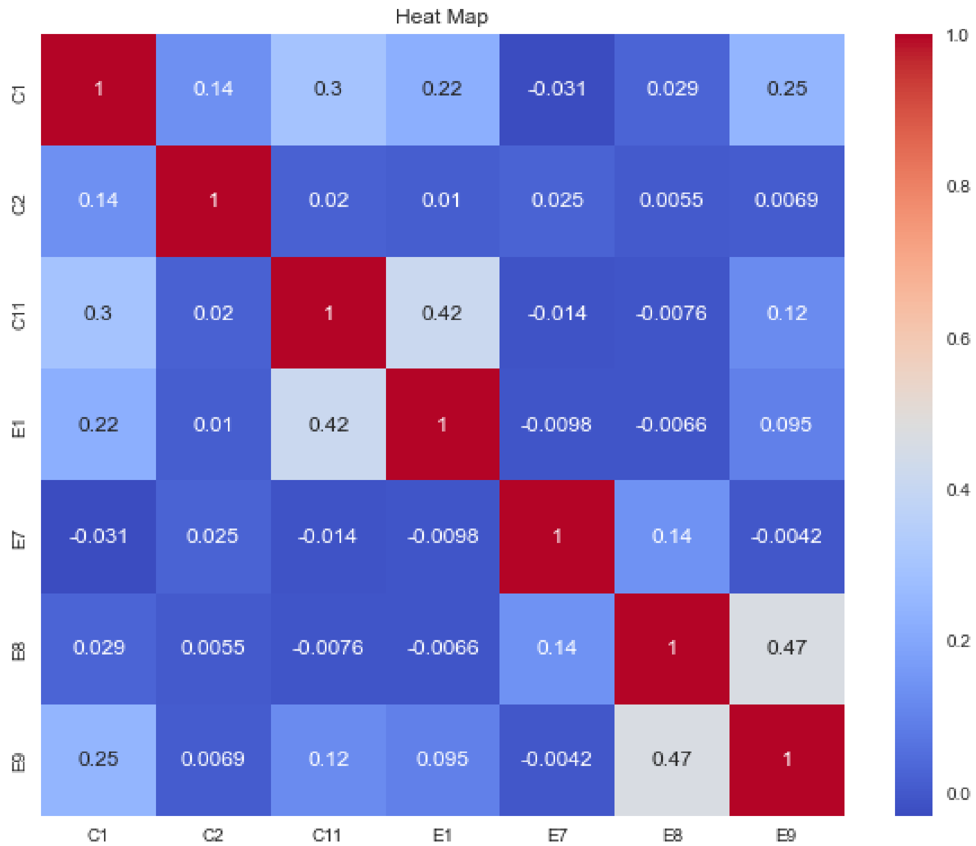


Figure 12 - Correlation Matrix of Variables Used for Clustering.

Also, a positive correlation was found between variables E9 and C1. The computed correlation coefficient of +0.56 confirmed this relationship. Specifically, it was observed that as the duration of maximum stay within the grid (E9) increased, the number of distinct terminals detected in the grid (C1) also tended to be higher. This finding suggests a potential connection between these two variables.

Lastly, a positive correlation between variables E9 and E8 was identified. The upward trajectory in the scatter plot indicated the relationship between the variables. While a few outliers were present, the computed correlation coefficient of +0.64 confirmed a significant positive correlation between E9 and E8. These findings provide valuable insights into the dataset, emphasizing the importance of these variables in the analysis.

### 4.3.3. Outliers

Outliers' analysis plays a crucial role in data exploration and analysis, enabling the identification and understanding of data points that deviate significantly from the norm. This analysis aims to exhibit extreme values or unusual patterns since the presence of outliers can impact the accuracy and reliability of the statistical analysis, leading to biased results and erroneous conclusions. By comprehensively examining outliers, it is possible to gain a deeper understanding of data patterns and make informed decisions regarding data treatment and modeling strategies.

Firstly, the analysis of the skewness and kurtosis measures was conducted for the variables. These statistical measures offered valuable insights into the shape and symmetry of the variable distributions.

Skewness measures the asymmetry of a distribution. A positive skewness indicates a longer or fatter tail on the right side of the distribution, while a negative skewness indicates a longer or fatter tail on the left side. Looking at the values, we can observe that all variables have positive skewness values, indicating right-skewed distributions. Variables C2, C11, and E7 have relatively high positive skewness values, suggesting a significant rightward skew with longer or fatter tails on the right side. Variables C1, E1, E8, and E9 also have positive skewness, but with smaller magnitudes. This indicates less right skewness, although the distributions still exhibit a slight imbalance towards the right. These skewness values provide insights into the distributional characteristics and the departure from the symmetry of the variables in the dataset. Table 7 presents the Skewness of each numerical variable.

Table 7 – Results of Variables' Skewness.

| Variable | Skewness |
|---|---|
| C1 | 7.276591 |
| C2 | 1137.527569 |
| C11 | 49.938528 |
| E1 | 35.999797 |
| E7 | 100.859831 |
| E8 | 11.508143 |
| E9 | 2.611723 |

A positive Kurtosis indicated a more peaked and heavy-tailed distribution compared to a normal distribution. Looking at the values, we can observe that variables C1 and C2 have significantly high kurtosis values, indicating that their distributions have heavy tails and are more peaked. Variables C11 and E1 also have relatively high kurtosis values, suggesting deviations from a normal distribution. On the other hand, variables E7, E8, and E9 have smaller kurtosis values closer to zero, indicating distributions that are closer to a normal distribution in terms of tail heaviness and peak. Overall, these kurtosis values provide insights into the shape and distribution characteristics of the variables in the dataset. Table 8 presents the Kurtosis of each numerical variable.

Table 8 – Results of Variables' Kurtosis.

| Variable | Kurtosis |
|---|---|
| C1 | 263.6390 |
| C2 | 1,673,955 |
| C11 | 10,452.74 |
| E1 | 5,026.202 |
| E7 | 13,820.01 |
| E8 | 157.136 |
| E9 | 6.402936 |

Additionally, to gain a visual understanding of the distributions, boxplots were used. The box plots serve as graphical representations of the distribution and statistical properties of each numerical variable in the dataset. They offer valuable insights into the characteristics and patterns of the variables, contributing to a comprehensive analysis of the data.

For variable C1, the box plot revealed a symmetric distribution with a median value of around 60. However, outliers were presented beyond the whiskers, indicating some extreme values. Moving to variable C2, the box plot displayed a wide range of values, with the upper quartile significantly higher than the median, suggesting a positively skewed distribution with a concentration of values towards the lower end. Outliers were also observed on the higher side of the distribution. Regarding variable C11, the box plot depicted a moderately skewed distribution, with the median closer to the lower quartile. Notably, outliers were present on both ends of the distribution, extending beyond the whiskers. Examining variable E1, the box plot illustrated a right-skewed distribution, with the median closer to the lower quartile. Outliers on the higher end of the distribution indicated extreme values. Similarly, variable E7 displayed a positively skewed distribution, with outliers on the higher side. Variable E8 showed a relatively symmetric distribution, centered around a median value of approximately 150, with no significant whisker extension. Finally, for variable E9, the box plot revealed a roughly symmetric distribution, with a median around 6, and outliers on both ends. These findings provide insights into the distributions and presence of outliers in the examined variables.

### 4.3.4. Analyzing Null Values

After an initial evaluation of the data, it became evident that certain data quality concerns have surfaced. For instance, variable E6 presented many missing values and some of the rows for nationalities presented either missing values or more than one nationality (see Figure 13).

Rows containing zero countries or multiple countries were systematically removed to eliminate inconsistencies that could compromise the accuracy of the analysis. Additionally, grids associated with parishes not included in the scope of the analysis were excluded to maintain the focus on the designated study area.

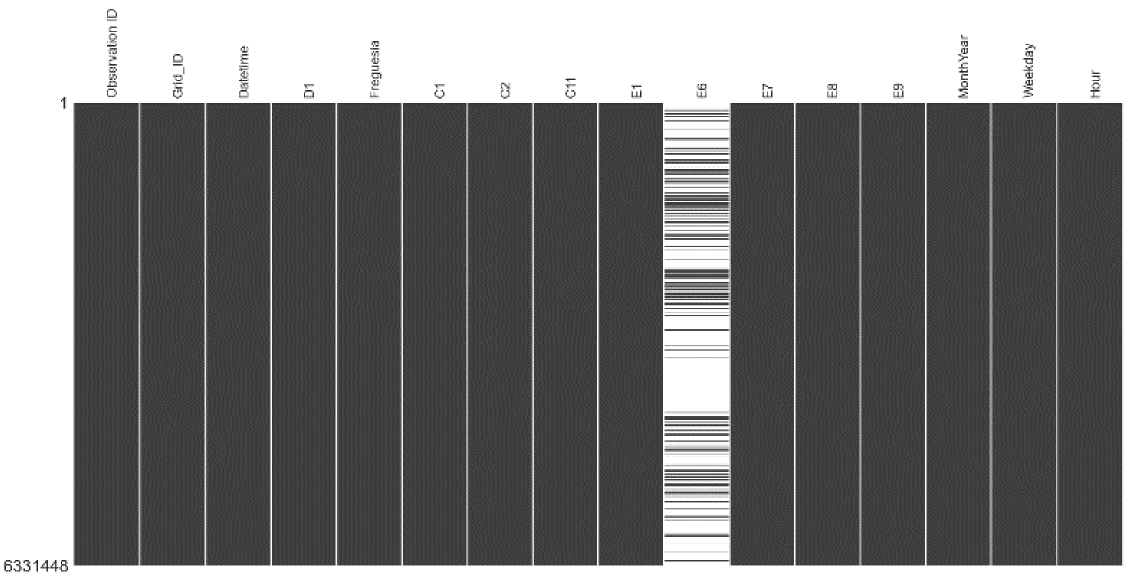


Figure 13 – Missing Values Visualization using Missingno Matrix.

### 4.3.5. Data Transformation

As part of the data preprocessing process, the datetime information was grouped by hour to enable a more comprehensive exploration of temporal patterns and trends. This grouping facilitated the aggregation of data and provided a broader perspective on the behavior and access to the internet within each hourly interval.

To increase the granularity of this information, the datetime variables were processed by converting them to the appropriate data types: the variables 'MonthYear' and 'Weekday' were transformed into string format, while the 'Hour' variable was converted to integer format. This ensured that the variables were represented in a suitable manner for subsequent analysis.

Next, categorical variables were encoded using one-hot encoding. Particularly, the variables 'Freguesia', 'MonthYear', 'Weekday', 'Hour', and 'D1' were subjected to this encoding technique. One-hot encoding creates binary indicator variables for each category within a variable. By doing so, it avoids the introduction of ordinal relationships between categories and allows the model to appropriately handle categorical information.

To address the challenge of processing large datasets, the dataset was divided into smaller chunks. This chunking approach enables the efficient handling of data by processing it in manageable portions. Each chunk was subsequently examined for non-numeric columns, which were identified using the select_dtypes() function. These non-numeric columns were then removed from each chunk to ensure that only numeric features remained for further analysis.

Following the removal of non-numeric columns, the data was normalized using the MinMaxScaler method. Normalization is a common preprocessing technique that scales the numerical features to a specific range, often between 0 and 1. By applying this transformation, the data is standardized, and potential issues related to different scales among features are mitigated. In this case, normalization was conducted on each chunk of data separately using the MinMaxScaler.

Parallel processing was employed during the normalization step to enhance computational efficiency. The 'threading' parallel backend was utilized, enabling the process to be distributed across multiple threads. By utilizing multiple threads, the normalization procedure could be conducted concurrently, leading to potential time savings. Finally, the normalized chunks were concatenated into a single array to facilitate subsequent analysis and modeling.

These preprocessing steps, including datetime processing, categorical variable encoding, chunking, non-numeric column removal, and data normalization, were performed to ensure the dataset was appropriately prepared for further analysis and modeling in the master's thesis. By adhering to these rigorous data preparation procedures, the dataset was carefully curated, ensuring its quality and integrity for subsequent analysis.

## 4.4. MODELING

This stage involves applying various techniques and algorithms to extract patterns, relationships, and insights from the data.

### 4.4.1. Feature Selection

Feature selection is a process commonly used in machine learning and data analysis that involves selecting a subset of relevant features (or variables) from a larger set of available features. For this project, feature selection was used to identify the most informative and discriminative features that contribute the most to the predictive performance of a model while reducing dimensionality and eliminating irrelevant or redundant features. By selecting a subset of features, the model can be simplified, computational resources can be saved, and overfitting can be mitigated. Feature selection helps to improve model accuracy, interpretability, and generalization capabilities by focusing on the most important features that capture the underlying patterns and relationships in the data.

Feature selection can be performed using various techniques, depending on the nature of the data and the specific requirements of the analysis. Dimensionality Reduction was the method chosen in this case using Principal Component Analysis (PCA), reducing the dimensionality of the feature space by transforming the original features into a lower-dimensional representation that retains most of the important information.

### 4.4.2. Principal Component Analysis (PCA)

The analysis of cumulative explained variance reveals new insights. The cumulative explained variance refers to the proportion of total variance in the data that can be accounted for by each component in the analysis.

In this analysis, nine components were examined. The results show that the first component explains 80.78% of the total variance in the dataset. When combined with the second component, which accounts for an additional 14.81% of the variance, the cumulative explained variance becomes 95.59% (0.955894). As we progress through the components, the explained variance continues to accumulate. The third component explains 4.18% of the variance, contributing to a cumulative explained variance of 99.77% (0.997682).

Table 9 - Cumulative Explained Variance by Component.

| | Component | Variance Explained | Cumulative Variance Explained |
|---|---|---|---|
| 0 | 1 | 0,807819 | 0,807819380799127 |
| 1 | 2 | 0,148075 | 0,9558941133782988 |
| 2 | 3 | 0,041788 | 0,9976816528422912 |
| 3 | 4 | 0,001217 | 0,9988983533169083 |
| 4 | 5 | 0,000443 | 0,9993415429374863 |
| 5 | 6 | 0,000338 | 0,9996791293693104 |
| 6 | 7 | 0,000215 | 0,9998943081421509 |
| 7 | 8 | 0,000074 | 0,9999687450420954 |
| 8 | 9 | 0,000031 | 0,9999999999999999 |

The contribution of additional components to the explained variance diminishes further. For instance, the fourth component explains only 0.12% of the total variance, resulting in a cumulative explained variance of 99.89% (0.998898). Similarly, the fifth, sixth, and seventh components explain 0.04%, 0.03%, and 0.02% of the variance, respectively, leading to a cumulative explained variance of 99.97% (0.999341). The eighth component explains 0.01% of the variance, contributing to a cumulative explained variance of 99.99% (0.999894). The ninth component makes a minimal contribution of 0.003% to the variance.

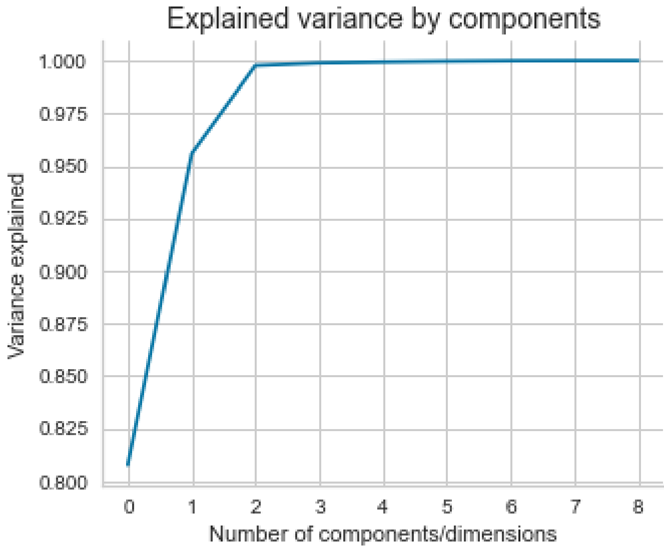


Figure 14 - Cumulative Explained Variance Plot.

Overall, the initial principal components still account for most of the variance in the data, suggesting that a reduced number of components can effectively capture the essential information. The cumulative explained variance plot demonstrates that including additional components results in diminishing returns, as these components explain progressively smaller amounts of variance.

Applying dimension reduction through principal component analysis (PCA) with three components, in this case, yields valuable results. These three components account for 99.77% of the variability in the data, providing a concise summary while retaining most of the significant information.

### 4.4.3. Subsampling

In this study, a subsampling technique was performed after analyzing the principal component analysis (PCA) and explained variance. The rationale behind this approach lies in the need to address the challenges posed by large datasets in the context of data analysis. PCA, a dimensionality reduction technique, was initially applied to identify the principal components that capture the majority of the dataset's variability. The cumulative variance explained by each component was examined to determine the optimal number of components to retain. By reducing the dimensionality, PCA allowed for a more compact representation of the data without losing critical information. Subsequently, subsampling was applied to further reduce the dataset's size, aiming to alleviate computational limitations and improve processing efficiency. By randomly selecting a subset of observations from the reduced dataset, we were able to obtain a representative sample for subsequent analyses, thereby reducing computational complexity while preserving the essential characteristics of the data. This approach ensured that the sampled dataset retained the inherent patterns and structures revealed by PCA and explained variance analysis, allowing for reliable and efficient analysis of the data.

A subsampled dataset was created by randomly selecting 500,000 data points from the encoded dataset using a specific random state for reproducibility.

### 4.4.4. Clustering

The clustering section of this master thesis explores the application of unsupervised learning techniques to identify patterns and groupings within the dataset. This section will detail the steps followed in the clustering process, including the selection of variables, determination of the optimal number of clusters, and application of the chosen clustering algorithm.

To determine the optimal number of clusters for the clustering analysis, two commonly used methods, namely the Elbow Method and the Silhouette Method, were applied.

#### 4.4.4.1. Elbow Method

In this section of the master's thesis, the Elbow method was employed to determine the optimal number of clusters (denoted as K) for the K-means clustering algorithm. The K-means algorithm is a widely used technique in unsupervised machine learning that aims to group data points into distinct clusters based on their similarities.

To apply the Elbow method, an instance of the K-means algorithm was created. The K values considered ranged from 1 to 20, capturing a diverse set of potential cluster quantities. The K-means algorithm was executed for each K value within the specified range. Consequently, the visualizer computed the sum of squared distances of data points to their assigned cluster centers, generating a plot that depicted this relationship between K and the resulting distances.

The plot aids in the identification of the "elbow" point, which signifies a significant decline in the sum of squared distances. This elbow point indicates the optimal number of clusters for subsequent steps in the analysis. By employing the Elbow method, this section of the master's thesis provides a systematic approach to determining the suitable number of clusters for the K-means clustering algorithm. This aids in ensuring meaningful and accurate clustering results in the subsequent stages of the analysis.

#### 4.4.4.2. Clustering Algorithms

Clustering algorithms may have different strengths and weaknesses when applied to different types of data or problem domains. Some algorithms may excel in handling certain types of data distributions or cluster shapes, while others may perform better in scenarios with noise or outliers. By testing multiple algorithms, it is possible to identify the one that performs the best in this context.

Moreover, testing different clustering algorithms allows us to explore the sensitivity of the results to algorithmic choices. By comparing the outcomes of different algorithms, we can gain insights into the stability and robustness of the identified clusters. This helps us make more informed decisions and draw more reliable conclusions from our clustering analysis.

Overall, testing different clustering algorithms provides a comprehensive assessment of their performance, suitability, and reliability for a given dataset and problem. It allows us to choose the most appropriate algorithm that best aligns with our goals and maximizes the quality of the clustering results. To choose the best clustering algorithm to be used, the advantages and disadvantages of each one individually was considered.

Table 10 – Advantages and Disadvantages of Clustering Algorithms. (Mahdi et al., 2021), (Alqahtani et al., 2021)

| Clustering Algorithms | Advantages | Disadvantages |
|---|---|---|
| K-Means Clustering | Simple and easy to understand, scales well to large datasets, has fast convergence, suitable for data with distinct spherical clusters. | Requires specifying the number of clusters (K) in advance, sensitive to initial centroid selection, may converge to local optima, not suitable for data with non-linear or irregular shapes. |
| Hierarchical Clustering | Does not require specifying the number of clusters in advance, can handle different types of distance metrics, and provides a visual representation of clustering hierarchy. | Computationally expensive for large datasets, sensitivity to noise and outliers, and difficult to determine the optimal number of clusters from the dendrogram. |
| Density-Based Spatial Clustering of Applications with Noise (DBSCAN) | Can discover clusters of arbitrary shape, robust to noise and outliers, and does not require specifying the number of clusters in advance. | Sensitive to the choice of distance metric and density parameters, may struggle with clusters of varying densities, not suitable for high-dimensional data. |
| Gaussian Mixture Models (GMM) | Can model clusters with different shapes and sizes, provides probabilistic cluster assignments, and handles overlapping clusters. | Sensitive to the choice of initialization, may converge to optima, computationally expensive for large datasets, and requires estimating the number of Gaussian components. |
| Mean Shift | Can identify clusters of arbitrary shape and size, adaptive bandwidth estimation, and robust to noise and outliers. | Computationally expensive for large datasets, sensitive to brand width selection, may result in irregularly shaped or fragmented clusters. |
| Spectral Clustering | Can discover non-convex and complex clusters, handles high-dimensional data, can incorporate prior knowledge through affinity matrix. | Requires affinity matrix computation, sensitive to the choice of similarity measure and clustering parameters, computationally expensive for large datasets. |

## 4.5. EVALUATION

The evaluation stage in data analysis plays a crucial role in assessing the performance and effectiveness of the analytical methods. It involves measuring the quality, accuracy, and reliability of the results obtained from the data analysis process. The evaluation stage serves to validate the models or techniques used and determine their suitability for addressing the research questions or objectives. To evaluate a clustering model, several metrics and techniques can be used. The most used methods for evaluating clustering models are the ones presented in Table 10.

Table 11 – Cluster Evaluation Techniques and Interpretations.

| Clustering Evaluation Techniques | Interpretation | References |
|---|---|---|
| Inertia or Within-Cluster Squared Sum (WCSS) | Inertia measures the compactness of clusters by calculating the sum of squared distances between each data point and its centroid. A lower inertia value indicated better clustering. Inertia scores can be used to compare different clustering models or to determine the optimal number of clusters using the elbow method. | (Vysala & Gomes, 2020) |
| Silhouette Score | The silhouette score measures how well each sample fits within its cluster and the separation between clusters. It ranges from -1 to 1, where a score closer to 1 indicated better-defined clusters. The average silhouette score can be calculated to evaluate the overall quality of the clustering. | (Rousseeuw, 1987) |

| | | |
|---|---|---|
| Davies-Bouldin Index (DBI) | Davies-Bouldin Index measures the average similarity between each cluster and its most similar cluster, while also considering the average dissimilarity between clusters. A lower Davies-Bouldin Index (DBI) score indicated better-defined and well-separated clusters, where distinct clusters have lower dissimilarity between them and higher similarity within them. | (Singh et al., 2020) |
| Calinski-Harabsz Index (CHI) | Calinski-Harabsz Index (CHI) is used to evaluate the quality of clustering algorithms. It measures the ratio of between-cluster dispersion to within-cluster dispersion, providing a higher score for well-separated and compact clusters. | (Luna-Romera et al., 2019) |
| Visual Inspection | Visual Inspection of the clustering results can provide insights into the quality and structure of the clusters. Allows to plot the data points with different colors representing the clusters and examine if they form distinct and well-separated groups. Furthermore, it is also possible to analyze cluster characteristics such as size, density, and shape to gain a better understanding of the clustering performance | (Shkaberina et al., 2022) |

To evaluate the clustering performance, the subsampled dataset was split into training and testing sets using an 80:20 ratio. The training set, represented by X_train, was used for model training, while the testing set, represented by X_test, was used for assessing the model's performance.

Various evaluation metrics were computed for different values of k (the number of clusters). These evaluation metrics included the Silhouette Score, Inertia, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). For each value of k, the K-means algorithm was applied to the transformed dataset (X_pca), and the corresponding evaluation scores were computed.

### 4.6. DEPLOYMENT

During the deployment stage of this master's thesis project, a crucial aspect involves the development of dashboards within the Power BI platform. These dashboards serve as decision support tools, intended to facilitate the practical utilization of the analytical findings obtained through the research conducted in this study.

The purpose of deploying these dashboards as decision-support tools is multifaceted. Firstly, they aim to enhance the decision-making processes within the relevant domain by providing stakeholders with a comprehensive and visually appealing representation of the underlying data and insights derived from the research. By presenting the information in a structured and intuitive manner, these dashboards assist in reducing the complexity associated with data analysis and interpretation.

Furthermore, the deployment of these dashboards seeks to promote a data-driven approach to decision-making, enabling stakeholders to base their judgments on empirical evidence rather than subjective factors. Through interactive visualizations, data exploration capabilities, and the integration of various data sources, these dashboards empower users to derive meaningful insights and make informed decisions.

While translating the research findings into actionable information, these dashboards allow the research to have a tangible impact, while providing a practical avenue for implementing the derived insights within relevant organizational contexts, helping address real-world challenges, and contributing to evidence-based decision-making.

# 5. RESULTS AND DISCUSSION

This section presents a comprehensive examination of the extensive results obtained from the study, which aim to become a good foundation for decision-making arrangements concerning the planning of urban city centers in Lisbon regarding internet accessibility. The results include the clustering of behaviors in internet access and can be reviewed below.

## 5.1. VISITORS' SEGMENTATION

This section encompasses a comprehensive analysis of the outcomes obtained from the various stages involved in visitors' segmentation. The discussion begins with a detailed examination of the k-means clustering technique, which serves as a fundamental step in the segmentation process.

The Elbow Method identified K=4 as the number of clusters that best balanced the variance explained by the clusters with the simplicity of the model. This approach allowed for a meaningful reduction in dimensionality and the identification of a suitable number of clusters for further analysis.

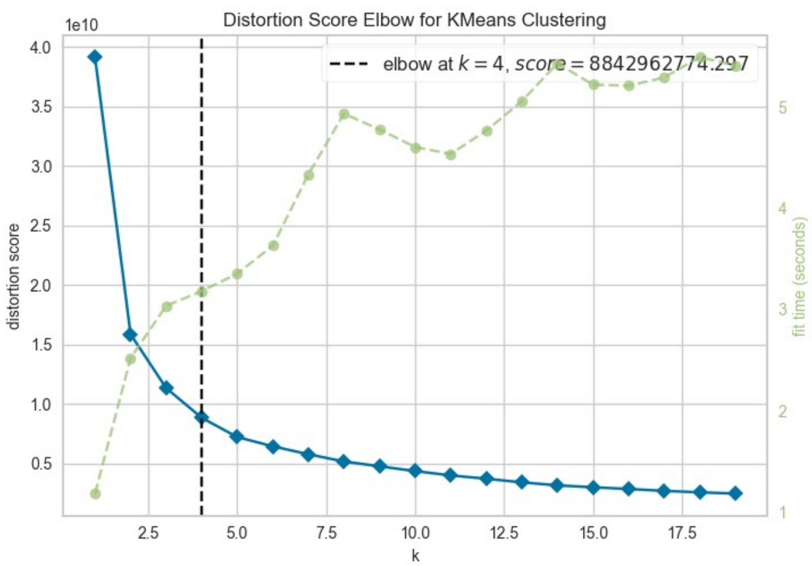


Figure 15 - Elbow Graph for Determining the Optimal Cluster Number.

### 5.1.1. K-means Clustering

Among the various clustering algorithms available, I selected K-means clustering for its suitability to the specific requirements of the analysis. K-means clustering is a popular and widely used algorithm known for its simplicity and efficiency in handling large datasets. It partitions the data into K distinct clusters by minimizing the within-cluster sum of squares, making it particularly effective in scenarios where the number of clusters is known or can be estimated. Additionally, K-means clustering performs well with numeric data and is robust to outliers. While other clustering algorithms such as hierarchical clustering, DBSCAN, GMM, Mean Shift, and Spectral Clustering offer their advantages and are suitable for different data characteristics, K-means clustering aligns well with the objectives and characteristics of the dataset under consideration, making it the preferred choice for this analysis.

Furthermore, to assess the effectiveness and distinctiveness of the generated customer segments, several metrics were employed in this study. These metrics included the silhouette score, the inertia, the Davies-Bouldin Index (DBI), the Calinski-Harabsz Index (CHI), and a visual inspection. These measures were utilized to evaluate the quality and uniqueness of the resulting segments, providing valuable insights into the segmentation performance.

Table 12 - Clustering Performance Metrics for Visitors' Segmentation.

| Number of K | Silhouette Score | Inertia | DBI | CHI |
|---|---|---|---|---|
| 2 | 0.521 | 19785548406.945 | 0.691 | 735501.156 |
| 3 | 0.423 | 14142383047.577 | 0.894 | 614243.846 |
| 4 | 0.448 | 11026598594.475 | 0.814 | 572299.828 |
| 5 | 0.394 | 9006483610.472 | 0.841 | 553533.857 |
| 6 | 0.336 | 8030173600.863 | 0.960 | 508827.512 |
| 7 | 0.355 | 7194588069.901 | 0.901 | 482944.301 |
| 8 | 0.362 | 6434594407.302 | 0.892 | 471277.667 |
| 9 | 0.371 | 5883880199.921 | 0.918 | 456814.269 |
| 10 | 0.344 | 5411807837.402 | 0.913 | 446324.237 |

The analysis of the clustering results provides valuable insights into the structure and organization of the dataset. The evaluation metrics shed light on the performance of the clustering algorithm across different numbers of clusters (K).

The Silhouette Score measures the consistency of samples within clusters, with higher values indicating better cluster cohesion. In this case, the highest Silhouette Score of 0.521 is observed when using 2 clusters. This suggests that the data points within each cluster are well-matched to their respective clusters, emphasizing the distinctiveness of the identified groups.

The Inertia, representing the sum of squared distances of samples to their closest cluster center, decreases as the number of clusters increases. This reduction in inertia implies the formation of tighter and more compact clusters. However, it is important to note that the decrease in inertia becomes less substantial as the number of clusters increases, indicating diminishing returns in terms of cluster compactness.

The Davies-Bouldin Index (DBI), measuring the average similarity between clusters, achieves its lowest value of 0.651 when using 5 clusters. This indicates a favorable level of clustering quality, where each cluster is well-separated and distinct from the others.

Furthermore, the Calinski-Harabasz Index (CHI), evaluating the ratio of between-cluster dispersion to within-cluster dispersion, shows an increasing trend as the number of clusters increases. The highest CHI value is observed when using 10 clusters, suggesting better-defined and well-separated clusters.

Based on these evaluation metrics, the selection of the optimal number of clusters involves weighing the trade-offs between cluster separation and within-cluster coherence. Considering the results provided, employing 5 clusters appears to strike a desirable balance. This configuration offers a relatively high Silhouette Score of 0.394, indicating well-matched data points within each cluster. Additionally, the DBI value is lower at 0.841, denoting distinct and well-separated clusters. Furthermore, the inertia value remains relatively low, signifying compact clusters.

It is important to note that visual inspection of the clusters plays a crucial role in understanding the characteristics and patterns within the data. By visually examining the clusters, it is possible to gain insights into their distribution, separability, and internal structure, which can further enhance the interpretation and understanding of the clustering results.

The visual inspection of clusters provides valuable insights into the structure and patterns within the data, including their distribution and separability. The visual inspection presented in Figure 22 reveals distinct clusters with well-defined boundaries, indicating that the clustering algorithm successfully grouped similar data points. One of the clusters, however, highlights the presence of outliers or discrepancies in the data points that indicate a possible need for further analysis.

For each cluster, a distinct cluster centroid was identified, surrounded by data points that belong to that specific cluster. Since the proximity of data points to their respective cluster centroid indicates the degree of similarity within the cluster, it is possible to identify that one of the clusters presents a higher dissimilarity within the cluster possible due to the presence of outliers. As for the others, they are tightly presented with data points located close to their centroid suggesting a higher level of cohesion and similarity among the data points within the cluster.

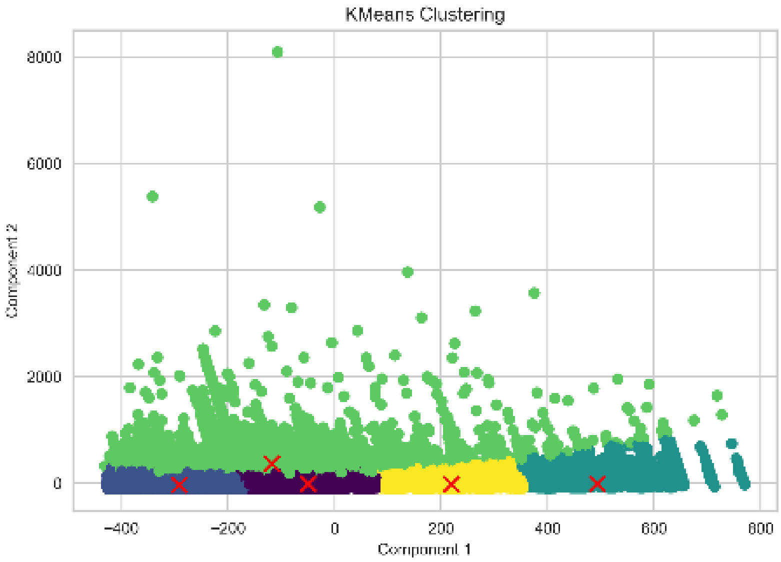


Figure 16 - Visual Inspection of K-Means Clustering Results.

While analyzing the visual inspection of clusters it is possible to identify the clusters, potentially detect outliers, and explore potential relationships or overlaps between clusters. This visual representation complements the quantitative analysis of the clustering algorithm's evaluation metrics and facilitates the interpretation and communication of the clustering results.

### 5.1.2. Description of Clusters

The following insights provide an understanding of the different behaviors and characteristics exhibited by the grids within each cluster. It can help identify patterns, preferences, and potential usage scenarios for each cluster, enabling targeted strategies or actions to improve user experiences or optimize resource allocation.

Table 13 - Label and Description of Clusters

| K | Label | Description |
|---|---|---|
| 1 | Urban Weekend | ➢ Average grid ID: 379.27<br>➢ Average distinct terminals in the grid: 82.72<br>➢ Average distinct terminals with roaming: 2.85<br>➢ Average voice calls from the grid: 1.78<br>➢ Average voice calls ending in the grid: 0.54<br>➢ Minimum stay duration: 0.06 minutes<br>➢ Average stay duration: 6.89 minutes<br>➢ Maximum stay duration: 58.16 minutes<br>➢ Presence across parishes: Belém (0.28), Alcântara (0.19), Estrela (0.19)<br>➢ Most affluent days of the week: Saturday and Sunday<br>➢ Most affluent hours of the day: 16h00 – 18h00<br>➢ Most affluent months of the year: April, May, and July<br>➢ Visitors' nationalities: Spanish, French, and English (UK) |
| 2 | Suburban Consistency | ➢ Average grid ID: 135.66<br>➢ Average distinct terminals in the grid: 69.97<br>➢ Average distinct terminals with roaming: 2.95<br>➢ Average voice calls from the grid: 1.57<br>➢ Average voice calls ending in the grid: 0.44<br>➢ Minimum stay duration: 0.11 minutes<br>➢ Average stay duration: 6.13 minutes<br>➢ Maximum stay duration: 50.90 minutes<br>➢ Presence across parishes: Belém (0.61), Alcântara (0.18) and Estrela (0.16)<br>➢ Most affluent days of the week: All days of the week (evenly distributed)<br>➢ Most affluent hours of the day: 09h00 – 11h00<br>➢ Most affluent months of the year: May, July, and October<br>➢ Visitors' nationalities: Spanish |
| 3 | Tourist and Commercial Hub | ➢ Average grid ID: 921.63<br>➢ Average distinct terminals in the grid: 68.48<br>➢ Average distinct terminals with roaming: 2.89<br>➢ Average voice calls from the grid: 1.21<br>➢ Average voice calls ending in the grid: 0.43<br>➢ Minimum stay duration: 0.14 minutes<br>➢ Average stay duration: 5.92 minutes<br>➢ Maximum stay duration: 45.14 minutes<br>➢ Presence across parishes: Alcântara (0.43) and São Vicente (0.20)<br>➢ Most affluent days of the week: Saturday (slightly higher activity)<br>➢ Most affluent hours of the day: 16h00 – 17h00<br>➢ Most affluent months of the year: July, August, and May<br>➢ Visitors' nationalities: Spanish, French, and German |
| 4 | Flexible Visitation Patterns | ➢ Average grid ID: 647.57<br>➢ Average distinct terminals in the grid: 73.29<br>➢ Average distinct terminals with roaming: 2.64<br>➢ Average voice calls from the grid: 1.58<br>➢ Average voice calls ending in the grid: 0.60<br>➢ Minimum stay duration: 0.10 minutes<br>➢ Average stay duration: 6.00 minutes<br>➢ Maximum stay duration: 49.43 minutes<br>➢ Presence across parishes: Alcântara (0.28), Belém (0.26) and Estrela (0.16)<br>➢ Most affluent days of the week: Weekends (Saturday and Sunday)<br>➢ Most affluent hours of the day: 11h00, 16h00, and 17h00<br>➢ Most affluent months of the year: October, May, and July<br>➢ Visitors' nationalities: Spanish, French, and English (UK) |

| | | |
|---|---|---|
| 5 | Seasonal Engagement | ➢ Average grid ID: 322.78<br>➢ Average distinct terminals in the grid: 468.56<br>➢ Average distinct terminals with roaming: 4.64<br>➢ Average voice calls from the grid: 8.25<br>➢ Average voice calls ending in the grid: 3.07<br>➢ Minimum stay duration: 0.03 minutes<br>➢ Average stay duration: 6.99 minutes<br>➢ Maximum stay duration: 101.15 minutes<br>➢ Presence across parishes: Belém (0.48), Alcântara (0.18) and Estrela (0.16)<br>➢ Most affluent days of the week: Thursday and Wednesday<br>➢ Most affluent hours of the day: 08h00 – 10h00<br>➢ Most affluent months of the year: October and November<br>➢ Visitors' nationalities: Spanish, French, and Brazil |

In terms of the presence across parishes, Cluster 1 – Urban Weekend – is characterized by a higher presence in Belém, Alcântara, and Estrela. Grids within this cluster are likely to be in densely populated areas such as urban centers. Most of the internet access is registered over the weekends (Saturday and Sunday), particularly during April, May, and July, between 16h00 and 18h00. The temporal distribution of this cluster's grids across different months and years is relatively uniform, indicating a consistent pattern of user activity throughout time. Similarly, the distribution across weekdays and hours does not show any significant deviations, suggesting that user behavior in this cluster remains relatively constant throughout the week and day. Observations of this cluster presented an average stay of 7 minutes within the grids, a mean of 83 distinct terminals detected in the grid, and an average of 2 voice calls registered every 5 minutes. The nationalities represented mostly are Spanish, French, and English (UK) visitors.

Cluster 2 – Suburban Consistency – stands out with a relatively lower value for the variable C1, indicating a smaller number of distinct terminals in the grids. Cluster 2 exhibits distinct characteristics in terms of parish presence. It is prominently associated with Belém, followed by Alcântara and Estrela, while no significant results were found for other parishes included in the study. These grids may be in sparsely populated areas, residential suburbs, or rural regions. In terms of temporal patterns, internet accesses within this cluster are evenly distributed across all days of the week. However, particular attention should be given to the months of July, May, and October, which exhibit higher activity. Moreover, a peak in internet access is observed between 09h00 and 11h00. When considering user behavior, the average stay within the grids of this cluster is approximately 6 minutes. The grids detect an average of 70 distinct terminals, with an average of 2 voice calls registered every 5-minute interval. In terms of nationality, visitors from Spain stand out the most. These observations provide insights into the characteristics and dynamics of Cluster 2, highlighting its significant presence in Belém and its consistent temporal patterns throughout the week.

Moreover, Cluster 3 – Tourist and Commercial Hub – is primarily characterized by its prominent presence in Alcântara, with São Vicente following closely behind. These grids may be in areas with transportation hubs, tourist attractions, or commercial centers that attract a significant number of visitors. In contrast, other parishes included in the study do not show significant results. Internet accesses within this cluster are evenly distributed across all days of the week, with a slight increase observed on Saturdays. The temporal distribution of this cluster's grids indicates a higher proportion in certain months and years, suggesting seasonal or temporal peaks in user activity, possibly aligned with tourism or specific events. The months of July, August, and May demonstrate higher activity and a peak in access occurs between 16h00 and 17h00. The distribution across weekdays shows a higher proportion on weekends, while the distribution across hours indicates higher activity during daytime and early evening hours, coinciding with typical tourist or commercial activity patterns. Regarding user behavior, the average stay within the grids of this cluster is approximately 6 minutes, with an average maximum stay of 45 minutes. The grids detect an average of 68 distinct terminals, and an average of 1 voice call is registered within every 5-minute interval. In terms of nationality, visitors from Spain stand out the most, followed by France and Germany.

Furthermore, Cluster 4 – Flexible Visitation Patterns – is characterized by a prominent presence in Alcântara and Belém, with Estrela also demonstrating noteworthy significance. Other parishes included in the study do not yield substantial results. The average number of accesses in this cluster remains relatively consistent across all days of the week, with a slight increase observed on Saturdays and Sundays. The distribution across weekdays shows a relatively higher proportion during weekends, indicating that users prefer to engage in longer visits and activities during their leisure time. The month of October stands out as the period of highest activity, closely followed by May and July. Peaks in access occur notably at 11h00, 16h00, and 17h00, reflecting the flexibility of visitation patterns in this cluster. Regarding user behavior, the average duration of stay within the grids of this cluster amounts to approximately 6 minutes, with an average maximum stay of 49 minutes. Notably, this cluster presents a longer minimum stay duration in the grids. These grids may be associated with places of interest, recreational facilities, or venues where users spend extended periods, such as parks, museums, or entertainment complexes. The grids detect an average of 73 distinct terminals, and an average of 2 voice calls are registered within every 5-minute interval. In terms of nationality, users from countries known for their cultural tourism or longer-term stays, such as European countries and some Asian countries, stand out in this cluster, including visitors from Spain exhibiting the most pronounced presence, followed by France and the United Kingdom.

Lastly, Cluster 5 - Seasonal Engagement – exhibits a notable concentration in the parishes of Belém, followed by Alcântara and Estrela. Among the included parishes, Belém stands out with the highest presence. Access patterns across the week reveal that Thursday and Wednesday attract the highest number of visitors, while weekends experience considerably fewer accesses. Notably, October and November demonstrate the highest levels of activity, while the summer months, typically associated with the high tourist season, exhibit fewer observations. A peak in access is observed during the morning hours, specifically between 8h00 and 10h00. Spain is the most prevalent nationality represented in this cluster, followed by France and Brazil. The grids within this cluster detect a significant number of distinct terminals, reaching 469 on average, reflecting a higher frequency of voice calls, indicating more active communication behavior. Additionally, an average of 8 voice calls are registered within every 5-minute interval. In terms of user behavior, the average duration of stay within the grids of this cluster is approximately 7 minutes, with an average maximum stay of 101 minutes, suggesting longer durations of data sessions, indicating users who engage in data-intensive activities or have longer browsing sessions. These findings suggest a moderate level of user engagement, with some visitors spending over 100 minutes within the grids.

### 5.1.3. Power BI Dashboard

Within the scope of my master's thesis, the Power BI dashboards were structured into three distinct sections, each serving a specific purpose. This division aims to provide a clear and organized presentation of the data and analysis conducted as part of the research.

The first dashboard will focus on observing the study area. It will provide an overview of the relevant information and insights specific to the area under investigation. This includes a geographical map of that area. The purpose of this dashboard is to offer a comprehensive understanding of the location and context in which the research is conducted.

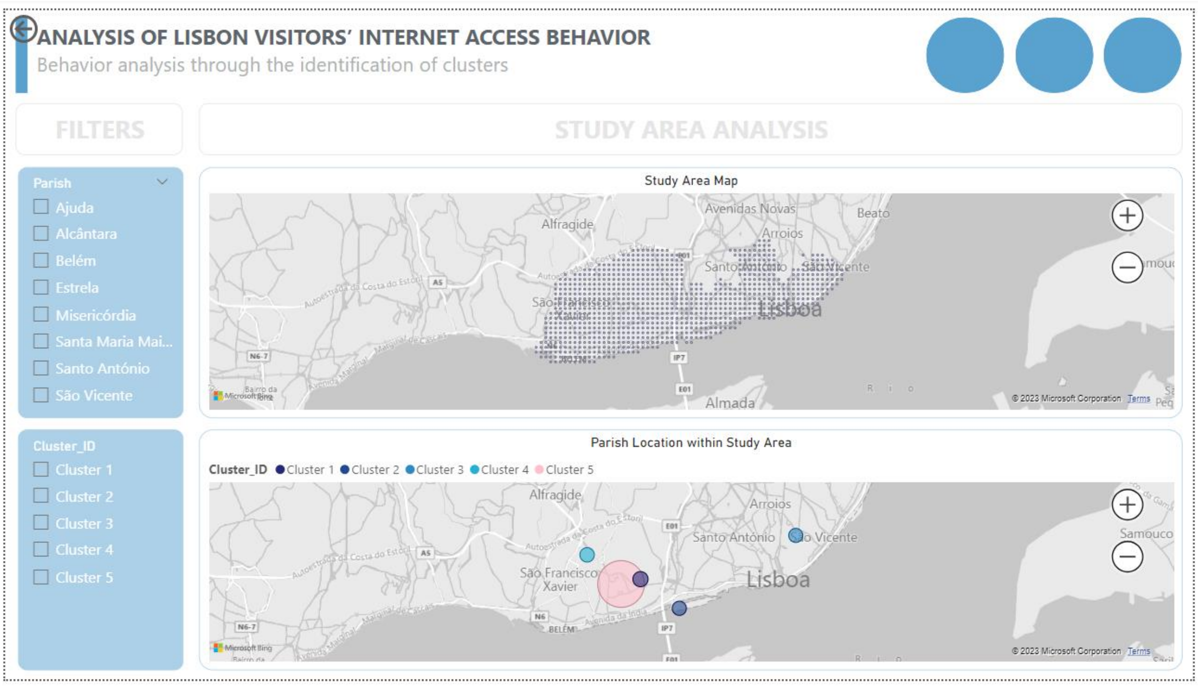


Figure 17 - Power BI Dashboard I: Study Area Visualization.

The second dashboard was dedicated to data analysis, interpretation, and exploration. It allows users to interact with the data by selecting various parameters such as parish, nationality, quarter, month, weekday, and hour. By choosing specific filters, users can refine the analysis and investigate the data from different perspectives. The selected parameters dynamically update the built-in graphs and visuals displayed below, providing the results based on the user's selections.

The visualizations available on this data analysis dashboard provide valuable insights. They include the number of observations, the top 10 nationalities represented in the data within the filters selected, a map illustrating the distribution of observations across different parishes, voice call patterns per month per parish, the average duration of stays within a specific grid per parish per month, observations categorized by weekday, month, and hour. These visualizations facilitate the identification of trends, patterns, and relationships within the dataset, enabling deeper analysis and understanding of the collected information.

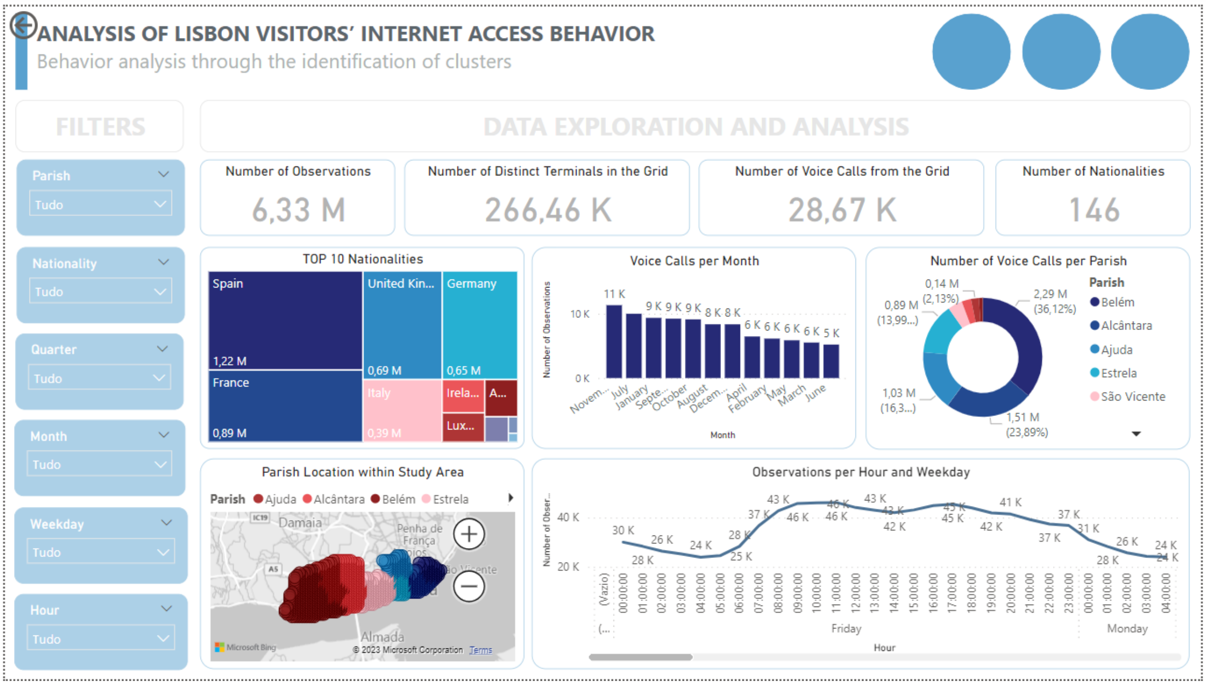


Figure 18 - Power BI Dashboard II: Data Visualization and Exploration.

The third and final dashboard focuses on cluster interpretation and exploration. This dashboard provides visualizations specifically designed to explore and analyze each cluster individually. These visualizations offer insights into the distinct characteristics and behaviors exhibited by each cluster, allowing for a more detailed examination and interpretation of the clustering results.

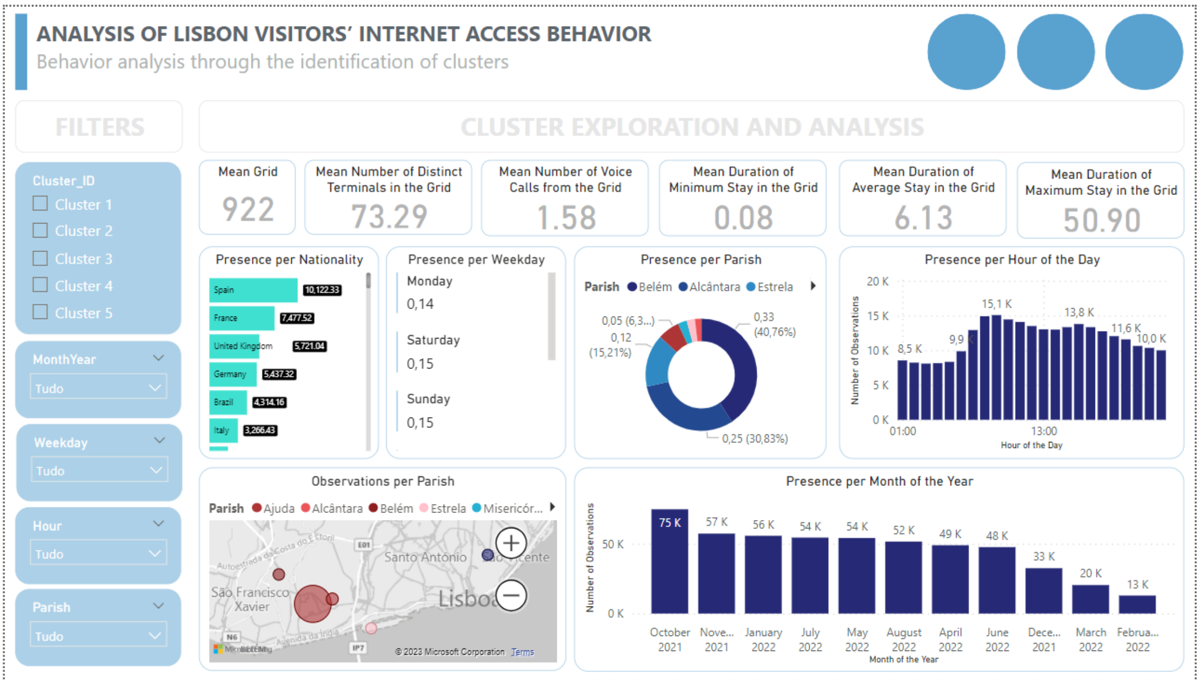


Figure 19 - Power BI Dashboard III: Cluster Interpretation and Exploration.

One of the key metrics examined in the dashboard is the mean number of distinct terminals in the grid. This metric represents the average number of unique devices detected in the grid for each cluster. By analyzing this metric, researchers can gain insights into the level of user engagement and the diversity of users within each cluster. A higher mean value suggests a larger number of unique devices accessing the grid, indicating a more active and varied user population.

Another important metric is the mean number of voice calls from the grid. This metric provides insights into the communication behavior of users within each cluster. By calculating the average number of voice calls made from the grid, researchers can understand the extent of user interactions and the level of communication activity within the clusters. A higher mean value suggests a higher frequency of voice calls, indicating more active communication patterns among users.

The dashboard also includes metrics related to the duration of user stays in the grid. These metrics include the mean minimum duration, mean average duration, and mean maximum duration of stay in the grid. These metrics provide valuable information about user engagement and behavior within each cluster. The mean minimum duration represents the average shortest duration of user visits to the grid, while the mean average duration reflects the typical length of user engagement. On the other hand, the mean maximum duration represents the average longest duration of user stays in the grid. By analyzing these metrics, researchers can identify patterns of user behavior and understand the range of durations users spend within the clusters.

Additionally, the dashboard presents visualizations and analytics on the presence of users based on various dimensions such as parish, weekday, month, and hour of the day. These visualizations help researchers identify trends and patterns in user activity. For example, the presence per parish visualization highlights the distribution of user presence across different parishes, emphasizing the parishes with the highest presence within each cluster. The presence per weekday visualization provides insights into any specific patterns or variations in user activity during different days of the week. The presence per month visualization showcases seasonal variations in user activity within each cluster, while the presence per hour of the day visualization highlights the peak hours of user activity.

Furthermore, the dashboard includes information on the top 10 nationalities represented within each cluster. This data allows researchers to understand the diversity of visitors from different countries and identify any significant trends or preferences among specific nationalities.

To facilitate analysis, the dashboard provides the option to select a specific cluster for an in-depth examination. This feature allows researchers to focus on a particular cluster and gain a deeper understanding of its unique characteristics and patterns of user behavior.

By utilizing Power BI's powerful capabilities, these dashboards provide an effective and user-friendly platform for visualizing and comprehending the data collected as part of the research. They will serve as valuable tools in presenting and communicating the research findings, facilitating informed decision-making, and contributing to a comprehensive understanding of the research subject.

#### 5.1.3.1. Cluster Insights and Patterns

Firstly, when examining the stay duration, it is evident that the clusters exhibit relatively consistent average durations, ranging from 5.92 to 6.99 minutes. This indicates that visitors in these areas tend to spend a similar amount of time within the clusters. However, there is some variation in maximum stay durations, which range from 45.14 to 101.15 minutes. This suggests that while most visitors have relatively short stays, there are occasional instances of longer visits within these clusters.

Secondly, an interesting pattern emerges in relation to the affluent days of the week. Three clusters, namely Urban Weekend, Flexible Visitation Patterns, and Tourist and Commercial Hub, highlight Saturday and Sunday as the most affluent days. This implies that weekends are particularly bustling in these areas, likely due to increased leisure and recreational activities. However, the Suburban Consistency cluster stands out by evenly distributing affluence across all days of the week, implying a consistent level of activity throughout the week.

Next, considering the affluent hours of the day, two distinct time periods emerge as common patterns across multiple clusters. The hours between 16:00 and 18:00 exhibit affluence in both the Urban Weekend and Tourist and Commercial Hub clusters, suggesting peak hours for social engagement and commercial activity. Additionally, the Seasonal Engagement cluster displays heightened activity between 08:00 and 10:00, possibly indicating morning routines or early engagement in specific activities within that cluster.

Regarding the affluent months of the year, May, July, and October stand out as months of significance in multiple clusters. The Urban Weekend and Flexible Visitation Patterns clusters both exhibit activity during these months, suggesting a seasonal preference for these periods. Similarly, the Tourist and Commercial Hub cluster includes July, August, and May as affluent months, highlighting the significance of summer months for tourist and commercial activities. On the other hand, the Seasonal Engagement cluster shows a distinct preference for October and November, indicating specific seasonal engagements during the autumn period.

Lastly, analyzing visitors' nationalities reveals commonalities across the clusters. Spanish and French visitors are present in four out of the five clusters, indicating their prominence in the region and their interest in the activities offered within these clusters. Additionally, English (UK) visitors are also present in three clusters, suggesting their active engagement in the visited areas. Notably, the Urban Weekend and Seasonal Engagement clusters share the presence of visitors from Brazil, which may indicate specific cultural or seasonal connections between these clusters and the country.

## 5.2. SUMMARY OF THE RESULTS

In conclusion, the objectives of my master's thesis, which focused on clustering visitor behavior in accessing the internet in Lisbon's urban area, have been effectively achieved. By analyzing distinct visitors' behaviors, segments were identified, enabling a better understanding of behavior and internet accessibility. The analysis and interpretation of the data were facilitated through the development of Power BI dashboards, providing valuable insights for decision-making.

The segmentation approach employed in this study has proven successful in identifying different visitor segments based on their internet access behavior. These segments can be utilized to optimize internet accessibility, enhance service provision, and improve city management by tailoring strategies and initiatives to specific target groups. The Power BI dashboards created for data visualization and interpretation have played a crucial role in presenting the findings clearly and comprehensively.

The research questions proposed at the beginning of this master thesis can be answered through the analysis of the Clusters Dashboard on Power BI. Key metrics and visualizations have been used to assess the performance of the clustering analysis and identify actionable insights. These findings contribute to a deeper understanding of visitor behavior and preferences in the context of internet access, opening avenues for further research in this area.

The development and implementation of Power BI dashboards have significantly facilitated the interpretation of the data and allowed efficient decision-making. The visualizations and interactive features of the dashboards have provided a user-friendly platform for exploring and analyzing the results, enabling stakeholders to gain valuable insights from the data.

Overall, the findings of this study, supported by the Power BI dashboards, have enhanced the understanding of visitor behavior regarding internet access in Lisbon's urban area. The insights gained from this research can be utilized by organizations and policymakers to optimize marketing strategies, improve visitor experiences, and make informed decisions in the digital landscape.

# 6. CONCLUSION

## 6.1. FINAL CONSIDERATIONS

In conclusion, this master thesis has successfully investigated the clustering of visitor behavior in accessing the internet within the Lisbon urban area, aiming to improve internet accessibility, provide better services, and enhance city management. Through an in-depth analysis of visitor behavior patterns and the utilization of Power BI dashboards for data visualization and interpretation, valuable insights have been gained.

The analysis of diverse data sources and the identification of distinct segments have provided a comprehensive understanding of visitor behavior regarding internet access. These findings have significant implications for tailoring strategies and interventions that can improve internet accessibility, optimize service delivery, and support effective city management.

By considering contextual factors such as location and time, a more holistic view of internet access dynamics within the urban area has been achieved. This understanding highlights the importance of accounting for external influences and specific user requirements when designing initiatives to enhance internet accessibility.

The outcomes of this research have practical implications for policymakers, urban planners, and service providers. The identified visitor behavior clusters and their characteristics can guide the development of targeted interventions and initiatives, ultimately leading to improved internet accessibility, enhanced service provision, and more efficient city management.

Furthermore, this study has set the stage for future research endeavors. Exploring the specific needs and preferences of each identified visitor behavior cluster can further refine strategies and interventions. Additionally, investigating the long-term impacts of improved internet accessibility on various aspects of urban life, such as economic development, social inclusion, and sustainability, would provide valuable insights for policymakers and researchers.

In summary, this master thesis has shed light on the clustering of visitor behavior in accessing the internet within the Lisbon urban area, offering insights that can be leveraged to enhance internet accessibility, improve service provision, and advance city management. The findings provide a solid foundation for future research and contribute to the establishment of a more connected and technologically advanced urban environment in the Lisbon area and beyond.

## 6.2. LIMITATIONS

While this study contributes valuable insights into the clustering patterns of internet accessing data in the Lisbon area, it is important to acknowledge certain limitations that may have impacted the findings and interpretations. These limitations are discussed below:

Data Availability

1. The analysis conducted in this study heavily relies on the availability and quality of the data obtained from Vodafone. The accuracy and completeness of the data may be subject to limitations inherent to the data collection process, including potential errors or missing values.

Some examples of the limitations across the data are the missing values found which resulted in excluding those variables from the analysis, and the fact that some observations had more than one nationality associated also resulted in removing these observations from the analysis.

Accessing further data such as social, demographic, or economic characteristics of the users accessing the internet in the Lisbon urban area could have enriched the research on this master thesis and provided further insights.

Variable Selection

2. The clustering analysis performed in this study was based on a predefined set of variables, including parishes, months, quarters, years, weekdays, hours, nationalities, duration of average stay within the grid, number of distinct terminals within the grid, and number of voice calls from the grid. While these variables provide relevant information for clustering patterns, future studies could explore additional variables to refine the clustering results.

Cluster Interpretation

3. The interpretation of the clusters generated in this study relies on the knowledge and expertise of the researcher. The assigned labels and segmentation names for each cluster are subjective and based on the available data and domain knowledge. For further research, alternative interpretations or alternative segmentation approaches could result in different cluster definitions and labels.

Generalizability

4. The clustering patterns identified in this study are specific to the internet access data from the Lisbon area during the specified period. The results may not be directly applicable to other cities or regions with different demographic, geographic, or transportation characteristics, nor too different periods as behaviors may vary. A replication of this study in different contexts or timelines is necessary to validate the generalizability of the findings.

Temporal Considerations

5. The clustering analysis performed in this study considered temporal aspects such as MonthYear, Weekday, and Hour. However, the specific time range selected for the analysis, along with potential temporal variations in internet access patterns, may have influenced the resulting clusters. Future studies could explore different time intervals or consider more sophisticated temporal analysis techniques to capture dynamic changes in transportation behavior.

External Factors

6. The clustering patterns identified in this study may be influenced by external factors that were not considered in the analysis. Factors such as weather conditions, special events, or infrastructure changes could impact internet accessibility patterns and lead to variations in the clusters. Incorporating additional contextual factors could provide a more comprehensive understanding of the clustering dynamics.

## 6.3. FUTURE RESEARCH

One crucial area for future research involves conducting a comparative analysis of clusters and internet access in other urban areas or cities. This approach allows for the identification of similarities and differences in clustering patterns and internet connectivity across different urban contexts. By exploring diverse settings, researchers can uncover unique factors influencing internet access disparities and gain insights into potential strategies for improvement. This comparative analysis provides a broader perspective on the issue, contributing to a more comprehensive understanding of cluster formation and internet access in the Lisbon urban area.

Complementing the quantitative analysis, future studies should incorporate qualitative research methods such as interviews or focus groups. These methods offer valuable in-depth insights into the experiences, challenges, and perceptions of individuals residing within different clusters regarding internet access. By understanding the lived experiences and perspectives of the affected population, researchers can uncover nuanced barriers and identify opportunities for targeted interventions. This qualitative investigation provides a human-centered approach to understanding the complexities of internet access disparities, giving voice to the individuals directly impacted by these inequalities.

An essential field for future research involves investigating the socioeconomic factors associated with clusters and internet access in the Lisbon urban area. Exploring the relationship between income levels, education, employment, and other socioeconomic indicators with cluster formation and internet connectivity would shed light on the underlying factors contributing to disparities. This analysis would provide a comprehensive understanding of the social and economic dimensions of internet access inequalities, facilitating the development of targeted strategies and policies to bridge the gap.

These suggested research areas provide avenues for further exploration and understanding of clusters and internet access in the Lisbon urban area. By undertaking these research endeavors, researchers can contribute to the development of informed policies and strategies that address the digital divide and promote equitable access to the Internet for all residents and visitors.

## BIBLIOGRAPHICAL REFERENCES

Aguiar, A. B., & Szekut, A. (2019). BIG DATA AND TOURISM: opportunities and aplications in tourism destination management. *Applied Tourism*, *4*(2). https://doi.org/10.14210/at.v4n2.p36-47

Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep time-series clustering: A review. In *Electronics (Switzerland)* (Vol. 10, Issue 23). https://doi.org/10.3390/electronics10233001

Amović, M., Govedarica, M., Radulović, A., & Janković, I. (2021). Big data in smart city: Management challenges. *Applied Sciences (Switzerland)*, *11*(10). https://doi.org/10.3390/app11104557

Aranburu Amiano, I., Plaza Inchausti, B., & Esteban Galarza, M. (2020). Identification of central urban attractions based on GPS tracking data and network analysis. *Boletin de La Asociacion de Geografos Espanoles*, *84*. https://doi.org/10.21138/bage.2840

Azman, N. A. N. M. N., Abd Rahman, N. H., Md Sawari, S. S., Abas, S. A., & Latif, S. A. A. (2021). The tourists' spatial behaviour and tourist movement pattern in Muar Johor. *Planning Malaysia*, *19*(16). https://doi.org/10.21837/PM.V19I16.970

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. In *Transportation Research Part C: Emerging Technologies* (Vol. 68). https://doi.org/10.1016/j.trc.2016.04.005

D'Urso, P., De Giovanni, L., Disegna, M., Massari, R., & Vitale, V. (2021). A Tourist Segmentation Based on Motivation, Satisfaction and Prior Knowledge with a Socio-Economic Profiling: A Clustering Approach with Mixed Information. *Social Indicators Research*, *154*(1). https://doi.org/10.1007/s11205-020-02537-y

Encalada, L., Boavida-Portugal, I., Ferreira, C. C., & Rocha, J. (2017). Identifying tourist places of interest based on digital imprints: Towards a sustainable smart City. *Sustainability (Switzerland)*, *9*(12). https://doi.org/10.3390/su9122317

Eslamirad, N., De Luca, F., Lylykangas, K. S., Ben Yahia, S., & Rasoulinezhad, M. (2023). Geoprocess of geospatial urban data in Tallinn, Estonia. *Data in Brief*, *48*, 109172. https://doi.org/10.1016/j.dib.2023.109172

García-Hernández, M., Ivars-Baidal, J., & Mendoza de Miguel, S. (2019). Overtourism in urban destinations: The myth of smart solutions. *Boletin de La Asociacion de Geografos Espanoles*, *83*. https://doi.org/10.21138/bage.2830

Hamid, R. A., & Croock, M. S. (2020). A developed GPS trajectories data management system for predicting tourists' POI. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(1). https://doi.org/10.12928/TELKOMNIKA.V18I1.13006

Hardy, A., Hyslop, S., Booth, K., Robards, B., Aryal, J., Gretzel, U., & Eccleston, R. (2017). Tracking tourists' travel with smartphone-based GPS technology: a methodological discussion. *Information Technology and Tourism*, *17*(3). https://doi.org/10.1007/s40558-017-0086-3

Holzmann, V. (2021). Sustainable smart cities challenges in project management curriculum. *Journal of Higher Education Theory and Practice*, *21*(7). https://doi.org/10.33423/JHETP.V21I7.4484

Höpken, W., Müller, M., Fuchs, M., & Lexhagen, M. (2020). Flickr data for analysing tourists' spatial behaviour and movement patterns: A comparison of clustering techniques. *Journal of Hospitality and Tourism Technology*, *11*(1). https://doi.org/10.1108/JHTT-08-2017-0059

Ivanov, N., & Gnevanov, M. (2018). Big data: Perspectives of using in urban planning and management. *MATEC Web of Conferences*, *170*. https://doi.org/10.1051/matecconf/201817001107

Juvan, E., Omerzel, D. G., & Maravić, M. U. (2017). Tourist Behaviour : An Overview of Models to Date. In *Management International Conference*.

Kirilenko, A. P., Stepchenkova, S. O., & Hernandez, J. M. (2019). Comparative clustering of destination attractions for different origin markets with network and spatial analyses of online reviews. *Tourism Management*, *72*, 400–410. https://doi.org/10.1016/j.tourman.2019.01.001

La Rocca, R. A. (2015). Tourism and mobility. Best practices and conditions to improve urban livability. Tema. *Tema. Journal of Land Use, Mobility and Environment*, *8*(3).

Liu, X., & Ban, Y. (2013). Uncovering Spatio-Temporal Cluster Patterns Using Massive Floating Car Data. *ISPRS International Journal of Geo-Information*, *2*(2), 371–384. https://doi.org/10.3390/ijgi2020371

Lu, S., Fang, Z., Zhang, X., Shaw, S. L., Yin, L., Zhao, Z., & Yang, X. (2017). Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. *ISPRS International Journal of Geo-Information*, *6*(1). https://doi.org/10.3390/ijgi6010007

Luna-Romera, J. M., Martínez-Ballesteros, M., García-Gutiérrez, J., & Riquelme, J. C. (2019). External clustering validity index based on chi-squared statistical test. *Information Sciences*, *487*, 1–17. https://doi.org/10.1016/j.ins.2019.02.046

Mahdi, M. A., Hosny, K. M., & Elhenawy, I. (2021). Scalable Clustering Algorithms for Big Data: A Review. In *IEEE Access* (Vol. 9). https://doi.org/10.1109/ACCESS.2021.3084057

Milano, C., Novelli, M., & Cheer, J. M. (2019). Overtourism and Tourismphobia: A Journey Through Four Decades of Tourism Development, Planning and Local Concerns. In *Tourism Planning and Development*. https://doi.org/10.1080/21568316.2019.1599604

Mobility Patterns of International Tourists: Implications for Responsible Urban Tourism. (2021). *Journal of Responsible Tourism Management*. https://doi.org/10.47263/jrtm.01-01-07

Orama, J. A., Huertas, A., Borràs, J., Moreno, A., & Anton Clavé, S. (2022). Identification of Mobility Patterns of Clusters of City Visitors: An Application of Artificial Intelligence Techniques to Social Media Data. *Applied Sciences*, *12*(12), 5834. https://doi.org/10.3390/app12125834

Padrón-ávila, H., & Hernández-Martín, R. (2020). How can researchers track tourists? A bibliometric content analysis of tourist tracking techniques. *European Journal of Tourism Research*, *26*. https://doi.org/10.54055/ejtr.v26i.1932

Pan, Y., Tian, Y., Liu, X., Gu, D., & Hua, G. (2016). Urban Big Data and the Development of City Intelligence. In *Engineering* (Vol. 2, Issue 2). https://doi.org/10.1016/J.ENG.2016.02.003

Panapakidis, I. P., Alexiadis, M. C., & Papagiannis, G. K. (2015). Enhancing the clustering process in the category model load profiling. *IET Generation, Transmission and Distribution*, *9*(7). https://doi.org/10.1049/iet-gtd.2014.0658

Rodríguez-Echeverría, J., Semanjski, I., Van Gheluwe, C., Ochoa, D., Ijben, H., & Gautama, S. (2020). Density-based spatial clustering and ordering points approach for characterizations of tourist behaviour. *ISPRS International Journal of Geo-Information*, *9*(11). https://doi.org/10.3390/ijgi9110686

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruiz Reina, M. Á. (2021). Spatio-temporal clustering: Neighbourhoods based on median seasonal entropy. *Spatial Statistics*, *45*. https://doi.org/10.1016/j.spasta.2021.100535

Rusdi, J. F., Salam, S., Abu, N. A., Sunaryo, B., Taufiq, R., Muchlis, L. S., septiana, T., Hamdi, K., Arianto, A., Ilman, B., Desfitriady, D., Kodong, F. R., & Vitianingsih, A. V. (2019). Dataset smartphone usage of international tourist behavior. *Data in Brief*, *27*. https://doi.org/10.1016/j.dib.2019.104610

Scuttari, A., Ferraretto, V., Stawinoga, A. E., & Walder, M. (2021). Tourist and viral mobilities intertwined: Clustering covid-19-driven travel behaviour of rural tourists in south tyrol, italy. *Sustainability (Switzerland)*, *13*(20). https://doi.org/10.3390/su132011190

Sdoukopoulos, A., Nikolaidou, A., Pitsiava-Latinopoulou, M., & Papaioannou, P. (2018). Use of social media for assessing sustainable urban mobility indicators. *International Journal of Sustainable Development and Planning*, *13*(2). https://doi.org/10.2495/SDP-V13-N2-338-348

Serna, A., Gerrikagoitia, J. K., Bernabé, U., & Ruiz, T. (2017). Sustainability analysis on Urban Mobility based on Social Media content. *Transportation Research Procedia*, *24*. https://doi.org/10.1016/j.trpro.2017.05.059

Shkaberina, G., Rezova, N., Tovbis, E., & Kazakovtsev, L. (2022). Visual Assessment of Cluster Tendency with Variations of Distance Measures. *Algorithms*, *16*(1), 5. https://doi.org/10.3390/a16010005

Shou, Z., & Di, X. (2018). Similarity analysis of frequent sequential activity pattern mining. *Transportation Research Part C: Emerging Technologies*, *96*. https://doi.org/10.1016/j.trc.2018.09.018

Singh, A. K., Mittal, S., Malhotra, P., & Srivastava, Y. V. (2020). Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 306–310. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00057

Streimikiene, D., Svagzdiene, B., Jasinskas, E., & Simanavicius, A. (2021). Sustainable tourism development and competitiveness: The systematic literature review. In *Sustainable Development* (Vol. 29, Issue 1). https://doi.org/10.1002/sd.2133

Švajda, J., Masný, M., Koróny, S., Mezei, A., Machar, I. V. O., & Taczanowska, K. (2018). Visitor profiling using characteristics of socio-demographic and spatial behavior as tools to support the management of protected mountain areas. *Geografie-Sbornik CGS*, *123*(4). https://doi.org/10.37040/geografie2018123040461

Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, *14*(2). https://doi.org/10.1177/1747016117738559

Thimm, T., & Seepold, R. (2016). Past, present and future of tourist tracking. *Journal of Tourism Futures*, *2*(1). https://doi.org/10.1108/JTF-10-2015-0045

Tobler, W. (1995). Migration: Ravenstein, thornthwaite, and beyond. *Urban Geography*, *16*(4). https://doi.org/10.2747/0272-3638.16.4.327

UNWTO World Tourism Barometer and Statistical Annex, January 2022. (2022). *UNWTO World Tourism Barometer*, *20*(1). https://doi.org/10.18111/wtobarometereng.2022.20.1.1

Valls Giménez, J. F., Pipoli, G., Rita, P., & Labairu-Trenchs, I. (2020). Tourist profiles and attitudes: a comparison between cities in a different phase of the life cycle. *International Journal of Tourism Cities*, *6*(4). https://doi.org/10.1108/IJTC-02-2019-0030

Versichele, M., de Groote, L., Claeys Bouuaert, M., Neutens, T., Moerman, I., & van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management*, *44*. https://doi.org/10.1016/j.tourman.2014.02.009

Vysala, A., & Gomes, Dr. J. (2020). *Evaluating and Validating Cluster Results*.

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, *11*. https://doi.org/10.1016/j.tbs.2017.02.005

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

Yang, L., Wu, L., Liu, Y., & Kang, C. (2017). Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from flickr. *ISPRS International Journal of Geo-Information*, *6*(11). https://doi.org/10.3390/ijgi6110345

Yao, Q., Shi, Y., Li, H., Wen, J., Xi, J., & Wang, Q. (2021). Understanding the tourists' spatio-temporal behavior using open gps trajectory data: A case study of yuanmingyuan park (Beijing, China). *Sustainability (Switzerland)*, *13*(1). https://doi.org/10.3390/su13010094

Yoshimura, Y., Amini, A., Sobolevsky, S., Blat, J., & Ratti, C. (2017). Analysis of pedestrian behaviors through non-invasive Bluetooth monitoring. *Applied Geography*, *81*. https://doi.org/10.1016/j.apgeog.2017.02.002

Zhou, M., Yue, Y., Li, Q., & Wang, D. (2016). Portraying temporal dynamics of urban spatial divisions with mobile phone positioning data: A complex network approach. *ISPRS International Journal of Geo-Information*, *5*(12). https://doi.org/10.3390/ijgi5120240

Zulfadhilah, M., Riadi, I., & Prayudi, Y. (2016). Log Classification using K-Means Clustering for Identify Internet User Behaviors. *International Journal of Computer Applications*, *154*(3), 34–39. https://doi.org/10.5120/ijca2016912076