iscte

**UNIVERSITY
INSTITUTE
OF LISBON**

Analysis of the Tourists behavior in Lisbon using Data from a Mobile Operator

Bruno Alexandre Mateus Francisco

Master's degree in integrated business Intelligence Systems

Supervisor
Doctor Ricardo Santos Faro Marques Ribeiro, Associate Professor
ISCTE-University Institute of Lisbon

Co-Supervisor
Doctor Fernando Manuel Marques Batista, Associate Professor
ISCTE-University Institute of Lisbon

October 2022

Department of Information Science and Technology

Analysis of the Tourists behavior in Lisbon using Data from a Mobile Operator

Bruno Alexandre Mateus Francisco

Master's degree in integrated business Intelligence Systems

Supervisor
Doctor Ricardo Santos Faro Marques Ribeiro, Associate Professor
ISCTE-University Institute of Lisbon

Co-Supervisor
Doctor Fernando Manuel Marques Batista, Associate Professor
ISCTE-University Institute of Lisbon

October 2022

To my beloved Parents, Wife and Mother-in-Law

"We can only see a short distance ahead, but we can
see plenty there that needs to be done"
Alan Turing, *"Computing Machinery and Intelligence", **1950***

Acknowledgements

Bruno Alexandre Mateus Francisco

Resumo

Esta dissertação, tem como objetivo fornecer a todas as entidades envolvidas no turismo em Lisboa, uma ferramenta de análise geoespacial, estatistica e longitudinal baseada em dados fornecidos por um operador móvel, que permita obter conhecimento sobre os comportamentos e hábitos dos turistas e vistantes da Cidade de Lisboa. A principal intenção é a de fornecer informação que permita aos decisores basearem as suas escolhas em dados e factos reais e não em conhecimento empirico. Através do trabalho desenvolvido no decorrer desta dissertação foi possível criar conhecimento em relação aos seguintes tópicos: a) as naacionalidades que mais visitam Lisboa - Alemanha, Espanha, França, Reino Unido, Itália e Brasil; b) as reguesias mais visitadas de Lisboa - Santa Maria Maior, Misericórdia, e São Vicente; c) as áreas populares da Cidade: Castelo, Alfama, Baixa Pombalina e São Vicente de Fora; d) a forma um evento como a web summit altera as movimentações dos turistas/visitantes e os locais visitados; e) durante a madrugada e a noite existe um decréscimo de turistas/visitantes originários da Europa. Tendo por base as análises desenvolvidas, as entidades gestoras do Turismo, passam a ter a possibilidade de promover de tomar ações que permitam potenciar os locais menos visitados e fazer uma melhor gestão dos locais de maior afluência, evitando aglomerados cuja gestão pode tornar-se problemática e que podem traduzir-se numa experiência menos agradável para os visitantes.

O trabalho baseia-se num conjunto de dados fornecido através de um acordo entre a Câmara Municipal de Lisboa e um operador móvel, consistindop em eventos gerados de forma completamente anonimizada através da sinalização trocada entre a rede móvel e o terminal dos utilizadores, estando cada uma das métricas associada à data/hora e à localização em que foi coletada, sendo a partir daqui tipificar o comportamento dos turistas e visitantes da Cidade de Lisboa. O desenvolvimento da investigação incidiu em três fases: 1) criar conhecimento sobre o negócio do turismo e compreensão dos dados disponíveis no sentido de perceber se estariam ou não aptos a dar resposta às nossas questões de investigação; 2) os dados foram trabalhados no sentido de serem preparados e adaptados às nossas necessidades - os dados que nos foram entregues tinham informação relativa tanto a telefones móveis pertencentes a utilizadores Portugueses, como a estrangeiros e tendo em conta que o nosso foco foi no segundo grupo, uma parte da informação foi descartada; 3) foi desenvolvido o trabalho de visualização e análise e foi nesta fase que conseguimos extrair o valor do nosso trabalho e conseguimos fornecer esta análise aos tomadores da decisão.

Do ponto de vista de entregáveis, além do presente documento, o trabalho desenvolvido consiste igualmente de um conjunto de análises, dashboards e visualizações desenvolvidas na ferramenta Microsoft Power BI. As análises e as informações fornecidas foram devidamente validadas por especialistas da Câmara Municipal de Lisboa, através de apresentações e questionários aos decisores e utilizadores da solução desenvolvida.

Keywords: Turismo, Lisboa, Comportamento de Viagem, Mobilidade Inteligente, Redes de Transporte, Big Data, Análise de Dados, Redes Móveis, Orientado a Dados

Abstract

This Dissertation aims to provide all entities involved in tourism in Lisbon with a geospatial, statistical, and longitudinal analysis tool based on data provided by a mobile operator, which allows obtaining knowledge about the behaviours and habits of tourists and visitors of the city. from Lisbon. The main intention is to provide information that allows decision-makers to base their choices on real data and facts. Through the work developed during this dissertation it was possible to create knowledge in relation to the following topics: a) the nationalities that most visit Lisbon - Germany, Spain, France, United Kingdom, Italy and Brazil; b) the most visited parishes in Lisbon - Santa Maria Maior, Misericórdia, and São Vicente; c) the popular areas of the City: Castelo, Alfama, Baixa Pombalina and São Vicente de Fora; d) the way an event like the web summit changes the movements of tourists/visitors and the places visited; e) during the early morning and night there is a decrease in tourists/visitors from Europe. Based on the developed analyses, the Tourism management entities now have the possibility to promote and take actions that allow to enhance the less visited places and to better manage the places of greater affluence, avoiding agglomerations whose management can become problematic and that can translate into a less pleasant experience for visitors.

The work is based on a dataset provided through an agreement between the Municipality of Lisbon and a mobile operator, consisting of events generated in a completely anonymized way through the signaling exchanged between the mobile network and the user's terminal, each of the metrics associated with the date/time and the location where it was collected, and from here to typify the behavior of tourists and visitors to the City of Lisbon. The development of the research focused on three phases: 1) creating knowledge about the tourism business and understanding the available data in order to understand whether or not they would be able to answer our research questions; 2) the data was worked on in order to be prepared and adapted to our needs - the data given to us had information regarding both mobile phones belonging to Portuguese users and foreigners and taking into account that our focus was on the second group , a part of the information was discarded; 3) the visualization and analysis work was developed and it was in this phase that we were able to extract the value of our work and were able to provide this analysis to the decision makers.

From the point of view of deliverables, in addition to this document, the work developed also consists of a set of analyses, dashboards and visualizations developed in the Microsoft Power BI tool.

The analyses and information provided were duly validated by specialists from the Lisbon Municipal Council, through presentations and questionnaires to decision-makers and users of the developed solution.

Keywords: Tourism, Lisbon, Travel Behavior, Smart Mobility, Transportation Networks, Big Data, Data Analytics, Mobile Networks, Data driven

Index

List of Figures

List of Tables

CHAPTER 1

1.1. Introduction

Cities are complex environments and there is a huge number of challenges that need to be addressed to provide everyone a better experience in a city like Lisbon. In our current work, we will focus mainly on tourists.

Addressing these challenges assumes a decisive role, especially now that tourist activity begins to recover after the pandemic period. As of the second half of 2021, the recovery started, and in May (last data available), there was a recovery of around 162% [2], compared to the same month of 2021. Even so, we continue with a negative variation, compared to May 2019.

Considering the big data generated and available these days, it is perfectly unthinkable that this digital asset is not used and that is exactly what we intend with our work – to use big data to respond to our challenges.

1.2. Framework

According to [1], the Portuguese capital receives about 4.5 million tourists every year. Considering that Lisbon has approximately 504 thousand residents, the city receives around 9 tourists per resident, the same is to say that it receives nine times the fixed population. As a term of comparison, cities like Prague, Barcelona and London receive between 4 and 5 tourists per resident. Looking into this ratio and considering that when compared with any of the referred cities, Lisbon is much smaller, it is easy to understand that all the stakeholders need to have a deep knowledge of the behaviors and movements along the city, to provide the tourists, the best possible experience. Considering this number of visitors, the empirical knowledge is not enough to manage and define the strategy for hotels, local accommodations, stores, restaurants, transports, museums, security and all the areas we can remember when we think about tourism. By providing to stakeholders like City Council and the entities responsible for tourism with the necessary information, we can have a very positive impact on the decision-making process, mitigating the risk of incorrect actions that can lead to an unpleasant stay in Lisbon and to the decrease of the financial income. The best ambassadors that Lisbon can have abroad are the previous visitors.

This work was developed in partnership with the Lisbon City Council [3], more specifically with the Lx Data Lab [4], which provided us with the data obtained through an established contract with a mobile operator.

According to the General Secretariat for the Economy [5], the weight of tourism on the Portuguese Gross Domestic Product is around 19%, being the 5th country in the world where the contribution of tourism has the greatest weight. This fact is particularly relevant given the number of people that this sector employs in its various aspects, so any negative variations in this indicator have an extremely harmful influence, not only in economic terms, but also in social terms. Unfortunately, it was not possible to confirm from any source what percentage of financial income derived from tourism is generated in Lisbon. Of course, we cannot just and only focus on the financial part, which is not always in the best interests of the "customers", that, in our case, are the tourists. It is important that whenever you visit Lisbon, you can be sure that you will find a safe city, properly sanitized, with a good transport network, enough accommodation, and framed with the most visited places and events of interest that can properly complete all points of interest, such as monuments, gastronomy, climate and so on. We believe that from the moment we provide decision-makers with the data, they will be able to create a transport network that meets the demand of tourists, and they will be able to better train police authorities and all professionals working in tourism.

1.3. Objectives

Considering that it was not possible through the literature review to find the information available in previous works, with this work, we intend to create knowledge in relation to:

1. Where do the top visitors of Lisbon come from.
2. What are the most visited areas of Lisbon.
3. Where are the tourists during mealtimes and where they sleep.
4. Event analysis.
5. Variation of Tourists by day periods and Lisbon Parishes.

To achieve this, we apply a data science approach with CRISP-DM [18] using past data of mobile operators, where we use cellular grid areas with information about tourists' nationality (number only due to GDPR rules), and time stamps in periods of five minutes.

The remainder of this document demonstrates, mostly through visualizations, that it is fundamental to understand how do tourists "behave" in the city of Lisbon. The focus of the analysis will be on the evolution over the five months of data we have available, considering the number of people and origin, with particular attention to the most represented countries and continents. Attention will also be paid to the places where tourists spend the most time and where the largest numbers of visitors are found, where they sleep and where they are at mealtimes (for the time being, our datasets do not make it possible to specify the commercial establishments). Finally, the influence of events in Lisbon will be highlighted, in relation to volumes and origins by comparison with a baseline that will always be the same period of the previous or subsequent week, to understand how events such as the Web Summit or the games Football League of Champions League change the usual panorama of Tourism in Lisbon.

1.4.   Motivation

Lisbon is currently one of the most visited European capitals by tourists of various nationalities, receiving around 4.5 million visitors a year. However, their behavior and patterns are not yet properly typified. Nonetheless, the huge amount of behavioral data currently being generated by different sources, such as mobile phones, makes it now possible to automatically track the movement of tourists in Lisbon. For this purpose, several vectors of analysis must be considered, such as the number of tourists by country of origin, the way they generally travel, where they go, what they do, where they spend more time, the preferred places to visit, and where they stay during mealtimes and sleeping time. The objective of this work is to carry out a geo-spatial, longitudinal, and statistical analysis through the creation of dashboards to help public policy makers, like, in this case, the Lisbon Municipal Council or the National tourist's office. Using this information, makes it possible to provide a better experience to tourists by giving the best information possible to the stakeholders, which ultimately leads to higher financial gains and an understanding of tourism through past data.

Several aspects of Lisbon, such as security, historical interest, gastronomy, the beauty of the city, the existence of several hubs of knowledge and developments, among others, lead to the fact that we are witnessing a growing trend of tourists and visitors.

Bearing this scenario in mind, the motivation to develop this academic work is related to the fact that we have identified a possibility to respond to the challenge of Lx Data Lab to analyze and modulate the behavior of tourists in Lisbon, based on the available data. by a mobile operator, making it possible to give the work developed a practical applicability. It is our

intention to provide managers and decision-makers with valid information that can be used in the management of resources destined for tourism in areas such as safety, hygiene, transport, logistics, commerce, hotels, and restaurants. We believe that given the size of the tourism business in Lisbon, management based on empirical knowledge is clearly insufficient, so data-based decision making is essential to create a positive financial impact on the business and, consequently, improve life. of all those who depend on Tourism, from the moment we can give visitors an excellent and exemplary experience in Lisbon.

1.5. Dissertation challenges

During this dissertation, three major challenges were encountered:

1. Understanding the data – given the amount of data available there was the need to carefully study all the available dimensions in order understand how it could be used to provide the answers to our research questions and objectives mentioned in sub-chapter 1.3.

2. Data enrichment – given our goals, there was the need to enrich the data with additional information: parishes, latitude/longitude for each grid, separate days between weeks and weekends, holidays, all this information was needed to our work.

3. Data volume – before data cleaning there was the need to handle with 40 Gigabytes of information and given the computing capacity available the strategy was to break the dataset into smaller chunks

1.6. Dissertation Outline

This dissertation is composed of five Chapters in which the introductory chapter (Chapter 1) is included. The structure of the document is presented below, as well as a high-level description of each of the chapters.

Chapter 2: in addition to the introduction, it comprises a description of the research techniques used for the literature review, identification of research topics and work already developed in areas like the one we set out to create knowledge.

Chapter 3: framework and description of the methodology followed in the way we approach the knowledge of the business and the data, the description of the data used in the work and the visualizations developed.

Chapter 4: development, description and in-depth analysis of the work carried out. In this chapter it is possible to find the answers to the questions to which we set out to create knowledge as well as the validation of everything that has been accomplished.

Chapter 5: the conclusions we reached are described, as well as the relative proposal of new developments to the work carried out.

CHAPTER 2

2. Introduction

During the next chapter of this dissertation, the techniques used to carry out research related to the topic in relation to which we intend to develop knowledge and the state of the art of works developed in relation to mobility in cities and people's behavior will be described. The topics on which the research focused are mainly on the use of data from mobile operators and sensors to study the behavior of people in general on the cities.

2.1. Related work

The high rate of use of mobile phones combined with mobility makes the data generated through the signaling exchanged between the terminal and the network a tool through which analyzes can be carried out that make it possible to identify patterns and behaviors related to mobility. According to the GSM Association [19] there were 460 million mobile subscribers in Europe in 2021, covering around 86% of the population; according to data provided by ANACOM [20], in Portugal there are around 13 million active sim cards. Therefore, this information can effectively be one of the best probes that exist for the analyses. It should also be noted that since roaming within the European Union started to have costs like those in the country of origin, people began to enjoy their mobile services much more when they are visiting another European country, which now allows them to have a sufficient volume of data and potential also to carry out research work related to tourism using this information. Therefore, the use of such a kind of data has been applied to study the way how people move in the cities.

Mariem Fekih et al [5], 2021, explored and made use of data generated through signaling exchanged between a mobile operator in order to create Origin-Destination matrices, with the main objective of assessing whether the amount of data generated through signaling can or cannot be a reliable source of analysis of the commuting movements of individuals, proposing a system capable of transforming the data generated by the mobile network into flows that allow typifying the Origin-Destination, having these validated through inquiries made by the local authority responsible for transport. This study was conducted in Rhône in the French Alps and the data used were provided by the mobile operator Orange. Through this work, the authors showed that these data can be used to estimate the pendulum movements, having been possible to prove through the questionnaires that the conclusions are valid. It was also possible to prove that this method can be automated and that after a few days, it may be possible to typify these

movements successfully, being possible to avoid the constant questionnaires. Thus, the group of researchers proved that the use of this information allows the means of transport to be optimized, benefiting users through optimizations that allow a higher quality of service.

In the article "Enhancing pedestrian mobility in Smart Cities using Big Data", Ebony Carter et al [6], 2020, proposes the use of different datasets generated through sensors installed on the Internet of Things network of the city of Melbourne to improve accessibility and the sustainability of the Municipality. The datasets used for the study include diverse information, for example: about parking, mobility, departures and arrivals at the airport and pedestrian traffic, having been estimated that in a period of 24 hours, data of around 650 thousand people. The results and analyzes were produced through heatmaps and various graphics, which allowed interpreting and contextualizing the analyses. Through the work carried out, it was possible to characterize pedestrian movements in Melbourne, and it was proposed to City officials that they continue to develop the sensor network and Internet of Things, since it is an essential source for the continued development of knowledge. necessary information on pedestrian movements to improve sustainability and accessibility.

Continuing the study of the work carried out using the data generated by the signaling generated by cellular networks, Claudio Badii et al [7], 2021, developed several metrics that allow us to perceive whether a given individual is or is not in mobility and if he is in motion, how you are doing it (on foot, by bicycle or in your own or public means of motorized transport), with the objective of sending each person personalized messages that can raise awareness of issues related to sustainable mobility and healthy living habits. To achieve their goals, the authors created a multi-class classifier that proved to be more accurate than resorting to a hierarchical approach and able to handle and manage data in real time. The developed solution was implemented in Antwerp and Helsinki.

Data generated through social networks are also an important source of information for the topic of mobility and behaviors in cities, Saqib Ali Haidery et al [8], 2020 made use of data from Weibo, a Chinese social network in the sense of analyze and typify the number and density of Weibo users in the city of Shanghai using estimation techniques with one or more variables. With this work, it was possible to use the referred data to conduct different vectors of analysis: points of concentration of people and from their location develop personalized recommendations and typify the different volumes in the 10 districts of Shanghai. From this information, it is possible to develop disaster mitigation plans as well as manage security and emergency resources.

Chiara Mizzi et al [9], 2018, in the article "Unraveling pedestrian mobility on a road network using ICTs data during major tourist events", also used the data provided by the Italian mobile operator TIM to study the characteristics of pedestrian mobility on the road network, using the City of Venice as an example to study the impact of tourists on the lives of local citizens as well as preserving the city's cultural heritage. After having worked and transformed the data, they developed an algorithm capable of reconstructing pedestrian movements through the streets of Venice and from there, they were able to distinguish mobility patterns between tourists and locals. Additionally, it allowed stakeholders to be given important and relevant information for decision-making based on data and not on empirical knowledge.

Entering in the field of machine learning, Jaeseong Jeong et al [10], 2021, presents a model in which people's mobility predictions are made from the traffic seen from the core side and the 5G radio. With this approach, it is possible to analyze mobility in real time and with data being generated in real time. The contribution of this work focused on the concept of NWDAF (Network data analytics function), having proposed an approach a predictive model capable of adapting the 5G network to each user and their needs, allowing to give the user a better experience regarding internet connection speeds and the reduction of latencies.

Continuing the study of the state of the art, the work developed by PENGZHAN GUO et all [11], 2021 is also relevant, who in their article "Route Optimization via Environment-Aware Deep Network and Reinforcement Learning" studied and developed an adaptive system that allows optimizing taxi services, proposing a deep learning system capable of optimizing routes on which these vehicles provide service, identifying the "optimal path", especially in abnormal cases and unexpected situations. For this case study, the researchers used data from "yellow taxis" circulating in New York City in the pre- and post-Covid period. The model created was able to detect anomalous events such as unexpected concentrations of people and, from there, adapt its recommendations on the best route to follow between two points, having been able to increase the weekly profitability of the vehicles by 98%.

Martin ŠAUER et al [12], 2021, developed knowledge about the intra-regional flow of tourists in Central Europe and its implications, having listed that understanding and typifying these movements is essential for strategic planning and sustainable development, particularly at the level of the most visited cities. In carrying out this work, data provided by entities responsible for tourism in various cities in Germany, Austria and the Czech Republic were used and from these data it was possible to conclude that the factors that most influence the distribution of tourists are: air connections, the attractiveness of the chosen destination and the

size of the tourism market in the place where visitors come from, given that, as in the case of Lisbon, the Germans are the ones who most influence tourism.

Continuing with the study of work done in relation to mobility, Xin Lao et all[13], 2021, in the article entitled "Comparing Intercity Mobility Patterns among Different Holidays in China: a Big Data Analysis" made use of data provided by Tencent, a Chinese Internet-related service company to model mobility patterns between Chinese cities on holidays, identifying the differences between different more traditional festive seasons such as Spring Festival, Tomb-sweeping Day, Dragon Boat Festival, and Mid-Autumn Festival and the less traditional ones in China, like Christmas and New Year's Eve. Through the work carried out in this article, they were able to prove that: a) movements are different depending on the type of holidays, b) the cities of Pearl River Delta and Xi'an are those from which more people leave for their hometowns during the traditional festive periods and c) during less traditional holidays, travel is mainly for recreational reasons, unlike traditional holidays where people travel mainly for cultural and traditional reasons.

Xin Li et al, 2018 [14], through the article entitled "Position prediction system based on spacial-temporal regularity of object mobility" proposed the creation of a system that makes predictions regarding the mobility of a given object, using historical data of the referred object that could be any type of "connected device" with GPS, from a mobile phone, to a car or any type of wearable - from the past data, the proposed model is able to predict which will be the next positions occupied by the referred object, and each possible position to be occupied in space is classified according to a score calculated on historical data, with the one with the highest classification being displayed. It should be noted that when evaluating the accuracy of the model proposed by Xin Li et al, it obtained accuracy rates 44% higher than those obtained with an algorithm based on Markov time series, thus proving the capacity of its model for predictions.

In the article entitled "The path of least resistance explaining tourist mobility patterns in destination areas using Airbnb data", Umut Turk et all [15], 2021, resorted to data provided by the Airbnb local accommodation platform to identify which are the 25 most attractive tourist destinations the world, having as motivation to do so the fact that a lack of knowledge on the topic was identified. Initially, an assessment was made of the quality of Airbnb's offer and prices in each of the locations and subsequently the prices and quality of the transport network were evaluated in each of the locations studied. The authors confirmed that the asking price for local accommodation is directly related to its geographic location and the quality of public

transport to the places of interest in each city that one of the reasons that most weigh in choosing accommodation is the proximity to good transport, especially in cities like Berlin and Frankfurt.

2.2.   Search and inclusion tecniques

To better understand the state of the art and the work already developed in relation to the use of data provided by mobile operators to typify patterns and behaviors of tourists we carried out systematic research in the Scopus database. The search strategy was based on a single query with numerous studies focuses. Using this method, we have the capability to count the number of articles while considering the topic and population under study. It is important to mention that we have only considered papers in our research. For a first selection on the articles, we have started by analyzing the title and the abstract, the entire text was evaluated in certain cases if that information was insufficient. The title, author, year, journal, subject, keywords, and abstract were all included in the data. Based on the data synthesis and analysis results mentioned above, a qualitative evaluation was conducted.

The notion of "Data Analysis" or "Behavior Analysis," the target population of "Smart cities," the "Cellular network," the "Tourist*" or the "Roaming," and the study's "Mobility" context were all exhaustively looked for in Scopus.

It became evident after reading all the publications that there has been a significant increase in behavioral study on visitor migration globally in recent years.

The data was handled and stored using Microsoft Excel and Zotero. Title, author, year, journal, topic, keywords, and abstract were among the details provided. Based on the results for data synthesis and analysis mentioned above, a qualitative assessment was conducted.

The study's target demographic, "Smart cities," the "Cellular network," "Tourism," or "Roaming," as well as the concepts of "Data Analysis" and "Behavior Analysis," were all taken into consideration when searching Scopus for published works on the subject. The research was conducted by looking up relevant articles in Scopus that addressed the study's concept, target audience, and context, shown in Table 1. The database was used for the query, and the same limitations and filters were applied.

Table 1 - Keywords definition

Concept	Population	Context	Limitations
Data Analysis Behavio#r Analysis	smart cities cellular network Touris* Roaming	Mobility	Only journal papers, articles, and reviews from 2018 to 2022
453106 Documents	220301 Documents	642769 Documents	
3156			
44 Documents			

This is demonstrated by the 44 documents returned by the query (Concept AND Population AND Context AND Limitations) when we use the keywords from each column. The same s to say that when we combine the concept with the population, the context and add the limitation we obtain 44 documents.

After completing a manual approach to identify the key subjects for their research questions and specify the outcomes, 10 publications were identified and described on the previous sub-chapter. Our study's systematization considered the year, the region, the RQ subject, and a succinct description. The 10 studies that were reviewed were selected based on the standards. The trend line in Figure 1, developed by the dissertation author, reveals that the subject we're looking at is becoming more and more popular, underscoring its significance.
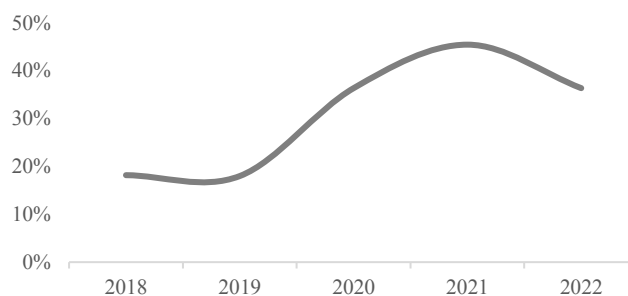
Figure 1 - Evolution of relevant studies per year

Given that the goal of this study is to identify how tourist behavior analysis and tourism mobility are used in smart cities, Table 2 and Figure 2 provide theoretical explanations of the topics mentioned in each of the papers that were evaluated, with a focus on the use of mobile phones and tourist behavior analysis when using mobile devices. Figure 2 demonstrates how most studies examined how people used mobile phones and other ICT infrastructure and their behavior (ICT). Our research is based on both ideas since we not only analyze human behavior

utilizing Lisbon's communication infrastructure as an operator, but also grasp it and develop a plan to satisfy their needs.
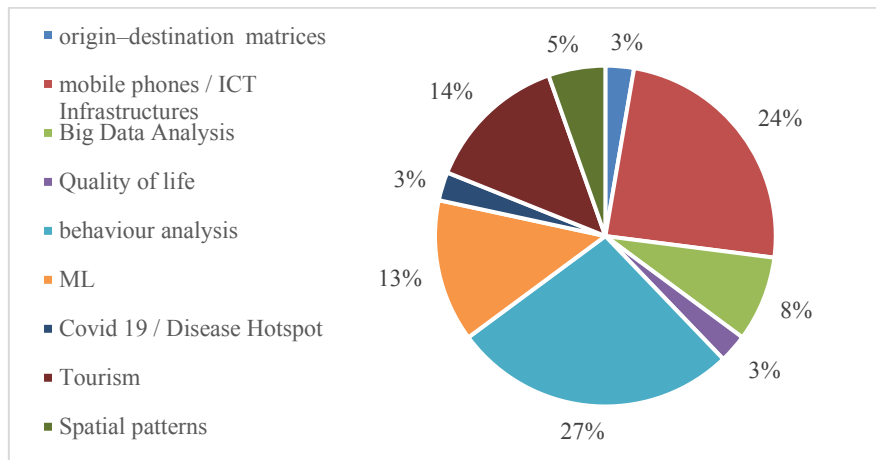
Figure 2 - Relative weight by document subject

Table 2 provides a summary of a more thorough analysis of this review. The problems were plainly stated; therefore, it was unnecessary to ask the publications' authors for clarification. Since the studies' results were categorized based on their inclusion or exclusion in the research, they are not mutually exclusive.

Table 2 - Summary review analysis

Topic	Reference
(1) origin–destination matrices	[3]
(2) mobile phones / ICT Infrastructures	[3]– [11]
(3) Big Data Analysis	[3], [5], [10]
(4) Quality of life	[12]
(5) behaviour analysis	[4], [8], [10], [11], [13]
(6) ML	[4], [6], [7], [9], [13]
(7) Covid 19 / Disease Hotspot	[13]
(8) Tourism	[11], [13]– [15]
(9) Spatial patterns	[14], [15]

CHAPTER 3

3. Knowledge extraction approach

This chapter is composed of five Sections in which the approach to the data used to produce this dissertation will be described. Section 3.1 describes the business understanding phase, that is, it demonstrates how we approach the business to be analyzed, and what it took to obtain knowledge about tourism and its specifics. In the next Section (3.2), we describe the work done in terms of data understanding. Given the volume of data and the amount of information made available to us, it was necessary to know the data to guarantee that we would be able to work with them to extract the greatest possible value from them. Section 3.3, related to data preparation, describes in high-level terms the transformations and enrichments made to the data in the pre-processing phase, with the objective of preparing the datasets with everything we consider necessary to develop the present work. Section 3.4 describes the process of creating the visualizations developed and delivered within the scope of this dissertation. On the last Section of this dissertation (3.5) is described the methodology followed on the knowledge extraction phase. of the present dissertation.

3.1. Business understanding

In this first stage, the emphasis is on comprehending the project's requirements and goals from a business standpoint. Using this knowledge, a data mining issue definition and a rough project schedule are then created to meet the goals.

Understanding the project's goals and requirements is the focus of the business understanding phase and it is divided in 4 sub-tasks. Except for the third task, the remaining three tasks in this phase are fundamental project management procedures that apply to most projects:

    a) Define business objectives

    b) Assess situation

    c) Define data mining goals

    d) Create a project plan

In our specific case, and considering that we have no knowledge regarding the tourism business, in addition to what is common sense, we mostly resort to the help of the Lisbon city Council and the LX Data Lab. This was the way found to ensure that we were able to obtain enough knowledge to allow us to interpret the data and the results obtained, this would not be

possible without knowing the business or, at least, the analyzes would be more superficial and eventually with less added value.

3.2. Data description and understanding

Mobile operators permanently collect data related to the use that Customers and users make of the service, whether for technical, billing or even legal issues. In this sense, whether the Customer is using 2G, 3G, 4G or 5G, the network always needs to collect various information and metrics so that it is possible, for example, to continue a telephone call or a mobile data session. It should be noted that when we refer to "data from a mobile operator", we are not referring to the mobile data service, but to specific information regarding all the signaling exchanged between the user's mobile equipment and the network that is serving it. Although an operator may have several networks, depending on the technology used, there is always a need for a base station that is in a certain location and that serves a certain geographical area to be serving the Customer, this base station interacts with the mobile phone and, at the same time, interacts with the core of the network, through this interface with the core it collects events such as a voice call, an internet session or a written message, as well as events related to mobility. such as handover and location updates or a network attach/detach. In this dissertation, the analyzes and work developed were based on the signaling exchanged between each of the roaming users and the network of a mobile operator in the City of Lisbon in the 5 months between September 2021 and January 2022.

As a complement to the set of data provided by the mobile operator and to enrich it, a metadata file was also used, which made it possible to associate each event with a geographic location.

The data made available by Lisbon city Council (Câmara Municipal de Lisboa) is supplied under an agreement with a mobile operator and is generated using the information provided by the cellular network and the mobile devices of each user. The information contained in the dataset is duly anonymized for legal and privacy reasons. In this way, it is not possible in any way to make any specific analysis of a particular user. There is not even any key that relates a given user to an event, and it is only possible to carry out analyzes involving volumes.

All the data available is aggregated in 3743 grids of 200x200 meters, being collected in periods of 5 minutes. Due to privacy constraints, if a certain grid doesn't have at least 10 users in the 5 minutes frame, it won't be reported. Data is made available on the big data platform, for a period of about 45 minutes after being collected. This means that we can have a maximum of 1 hour delay between the collection and the availability of the data. However, it is important

to say that for the scope of the present work, we will use a snapshot of the data and, therefore, we will not be leverage of the online data stream. Although we won't be using them all in our project, Table 3 presents the 24 indicators/dimensions available in the data provided by the Mobile Operator.

In addition to the dataset that contains the data provided by the city Council through the agreement established with Mobile Operator, a dataset that contains information related to each of the 3743 grids was also used. These are the data that allow us to geo-reference the main dataset since it contains the coordinates of the centroid of each grid, the parish, or parishes in which the grid is inserted, the name, the geometry and the WKT. With this information and using the "Grid_ID" key, it becomes possible to insert the events in the space and, from there, trigger our analyses. That said, Table 4 describes the data available in the dataset with the geographic information and a detailed description of each of the dimensions.

Table 3 - Mobile operator dataset variables

ID	Variable Name	Variable Description	Variable Type
0	Grid_ID	Number of the grid There are 3743 squares of 200 by 200 meters to cover the metropolitan area of Lisbon	Nominal
1	Datetime	Time and date of occurrence	Datetime
2	C1	Number of distinct terminals counted on each grid cell during the 5-minute period – Measured every 5 minutes	Metric
3	C2	Number of distinct terminals in roaming counted on each grid cell during the 5-minute period– Measured every 5 minutes	Metric
4	C3	No. of distinct terminals that remained in the grid cell counted at the end of each 5-minute period	Metric
5	C4	No. of distinct terminals in roaming that remained in the grid cell counted at the end of each 5-minute period	Metric
6	C5	No. of distinct terminals entering the grid	Metric
7	C6	Terminals leaving the grid – These are the distinct terminals that left the grid. The calculation is made using the previous 5-minute interval as reference, also considering the crossings of the grid in the same interval	Metric
8	C7	Number of entries of distinct terminals, in roaming, in the grid	Metric
9	C8	Number of exits of distinct terminals, in roaming, in the grid	Metric
10	C9	Total no. of distinct terminals with active data connection in the grid cell – Measurement every 5 minutes	Metric
11	C10	Total no. of distinct terminals, in roaming, with active data connection in the grid cell – Measurement every 5 minutes	Metric
12	C11	No. of voices calls originating from the grid cell	Metric
13	C12	Entering the city: No. of devices that for 5 minutes enter the 11 street Sections considered for analysis. For this purpose, a Section of track is a route with	Metric

14	C13	Entering the city: No. of devices that for 5 minutes enter the 11 street Sections considered for analysis. For this purpose, a Section of track is a route with	Metric
15	D1	Top 10 origin Countries of the devices in Roaming	Metric
16	E1	Number of voice calls that ended in the grid within the 5-minutes	Metric
17	E2	Average download speed per grid within the 5-minutes	Metric
18	E3	Average load speed per grid within the 5-minutes	Metric
19	E4	Peak download speed on the grid within the 5-minutes	Metric
20	E5	Peak upload speed on the grid within the 5-minutes	Metric
21	E6	Top 10 apps used on the grid within the 5-minutes	Metric
22	E7	Lowest permanence period on the grid within the 5-minutes	Metric
23	E8	Average permanence on the grid within the 5-minutes	Metric
24	E9	Maximum permanence period on the grid within the 5-minutes	Metric
25	E10	Count of devices sharing the internet connection in the grid within the 5-minutes	Metric

Table 4 - WKT dataset variables

Variable Name	Variable Description	Variable Type
grelha_id	Number of the grid There are 3743 squares of 200 by 200 meters to cover the metropolitan area of Lisbon	Nominal
dicofre	Identification of the parish. Assigned by administrative entities.	Nominal
entity_id	Identification of the data source where the information was generated.	Nominal
entity_type	Identification of the data source where the information was generated.	Nominal
freguesia	Parish to which the largest area where the grid belongs	Nominal
freguesias	Parishes in which the grid is inserted.	Nominal
latitude	Centroid Latitude	Metric
longitude	Centroid Longitude	Metric
objectid	id of the object in the database	Metric
position	Grid in Geometry format	Metric
wkt	Grid in WKT format	Metric

After collecting the data for our study, we carefully examined it and investigated each variable to understand its potential and how we could increase the added value of this research. As previously mentioned, our key objective is to comprehend how tourists move around. To do so, the Lisbon city Council provided us with a dataset about people's movement in the city of Lisbon (both roaming and non-roaming), based on mobile phone data produced. The mobile operator extrapolated the data to create the currently accessible dataset to provide a more accurate depiction of the mobility of all individuals who moved around Lisbon between September 2021 and January 2022.

The Mobile Operator dataset involves a total of 26 variables and approximately 126 million records, separated into monthly-based files, as described in Table 5.

Table 5 - Number of records per month (mobile operator dataset)

MONTH	NUMBER OF RECORDS
SEPTEMBER 2021	17 233 318
OCTOBER 2021	32 627 337
NOVEMBER 2021	21 619 292
DECEMBER 2021	33 121 658
JANUARY 2022	22 344 624

3.3. Data preparation

Data selection - roaming is the term used to describe when a cell phone is registered in a network different from the home network, with this we can assume that "roamers" are tourists or visitors. Data is gathered every 5 minutes from a grid of 3743 cells, each one measuring 200 × 200 meters, Figure 3 reveals the Lisbon areas, grouped by color, and the grid used by the operator. Due to privacy constrains, no values are provided for cells containing less than 10 devices.

Data cleaning - duplicated or NULL rows were deleted and as part of the dataset's first cleanup.

Feature selection -After cleaning the dataset, we have selected the most relevant variables to our goal. We then constructed a dataset using solely the roaming (tourists) users' mobility data from the dataset. It was crucial to distinguish between weekends and holidays in our dataset because the focus of our research is on people's movement. We did this by using the holidays library, only marking the holidays that fell on the weekdays since the weekends will not significantly affect mobility in general. To have information about the parish and the latitudes and longitudes of each Grid ID, we linked the dataset with a reference data supplied by the council. After that. it was feasible to create new columns from the intersection of this dataset to make it easier for Power BI to analyze and visualize the resulting data later. A column with Lisbon zones was created, in which the 24 parishes of Lisbon were divided into 5 unique zones:

- North zone: Santa Clara, Lumiar, Carnide, São Domingos de Benfica, Benfica;
- Western zone: Alcântara, Ajuda, Belém;
- Center zone: Campolide, Alvalade, Avenidas Novas, Santo António, Arroios, Areeiro;
- Historic downtown Area: Campo de Ourique, Estrela, Misericórdia, Santa Maria Maior, São Vicente, Penha de França;
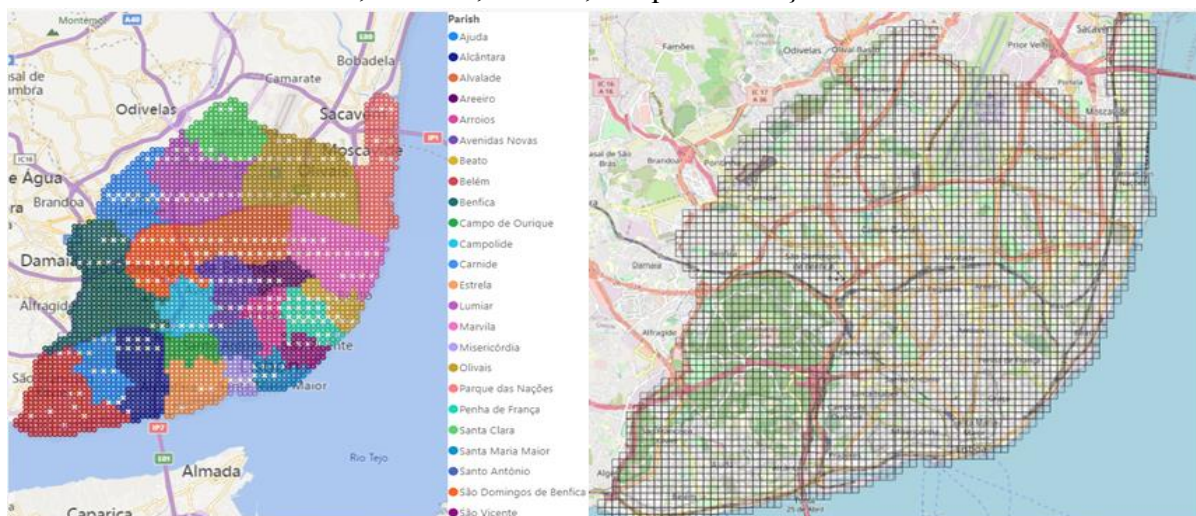- Eastern zone: Beato, Marvila, Olivais, Parque das Nações.

Figure 3 - Lisbon parishes and operator grids

20

3.4. Visualization

This process was oriented to the dashboard visualization of geographic and temporal data to obtain a clear image that would be able to help us understand the data and address the questions we set out to answer. In the course of our work, we realized that the result would be as rich as the more information we were able to provide to stakeholders. It is important to mention that to develop the entire visualization layer developed for this dissertation, it was necessary to obtain knowledge regarding the tool used.

The creation of the visualizations presented was the most challenging part of all the work developed, considering that there was an awareness that a large part of the answers to which we proposed to respond and in relation to the areas to which we intended to create knowledge, would largely depend on measure of the quality of what was achieved at this stage.

3.5. Methodology

During this dissertation on the behavior of tourists in the city of Lisbon, we followed the CRISP-DM methodology, which stands for Cross Industry Standard Process for Data Mining. The objective of CRISP-DM is to provide those who work with data with a structured framework that allows planning and executing a data mining project. This methodology has proven to be a good way to explore and extract value from data of all types and for that reason, we chose to follow it. However, CRISP-DM is not, nor does it intend to be, a set of rules to be strictly followed, leaving spaces for adaptation by the working groups. That said and considering our objectives and the fact that it is an academic work, we made some adaptations to the methodology to better serve our intentions. Therefore, and in a summarized way, we chose to develop our data exploration in four phases, which we believe are necessary for us to be able to answer our questions and provide decision-makers with information: 1) Business understanding; 2) Data understanding; 3) Data preparation 4) Visualization and dashboards to support decision.

CHAPTER 4

4. Insights and visualizations

In this chapter, we present the results of the analyzes carried out as well as the answer to the questions in relation to which we intend to create knowledge. That said, in Section 4.1, the origins of tourists and visitors to Lisbon will be described, focusing particularly on the countries and geographic areas where the largest number of people come from. Section 4.2 describes, according to the data worked on, which are the areas of the city most visited by tourists, making an analysis at the level of parishes and aggregation by areas of Lisbon as well as the specific places where there is a greater concentration of visitors. In Section 4.3, the places where tourists are at the typical times of the main meals - lunch and dinner will be described. Sections 4.4 and 4.5 describe the analysis made to two events, the 2021 Web Summit and the Benfica-Bayern match that took place in September 2021, in these Sections, it was analyzed how the events changed the usual patterns of tourists in Lisbon. In Section 4.6, the variation in the number of tourists per hour and time of day was analyzed, here it is possible to find the evolutions considering the dimensions described and the origins of the tourists, being verifiable that the evolutions change depending on the places from where are from. Section 4.7 describes the variation and evolution of visitors in each of the parishes of Lisbon, where it is possible to verify the volumes by parish at each moment of the day, making it clear that there are parishes that are mainly dormitories, in contrast to others that have more tourists during the day. The last Section (4.8) describes the validation and evaluation made by Lx Data Lab to the analyzes and work developed in this dissertation, being demonstrated through the answers given to a questionnaire made to those present at the meeting where we presented the conclusions. In all Sections, the way in which the data were processed to lead to the analyzes and conclusions described is described. The various Figures presented above serve to support the work developed in this dissertation and help decision makers and evaluators understand them. "A picture is worth a thousand words".

4.1. Where do the top visitors of Lisbon come from?

Considering that the present dissertation is dedicated to analyzing the behavior of Tourists and Visitors of Lisbon, we started our analysis with the typification of their origin and their evolution over the 5 months of data studied. only the country but also the geographic area of which they are native. This analysis allowed us to have information on where the tourists and

visitors come from in the capital, as well as to have the perception of how they evolved, that is, to understand if the largest number of Visitors always comes from the same countries/geographical areas or, on the other hand, if there are relevant variations. For this, the average value of mobile phones roaming was used in each of the 200x200 meters squares in which the data used are divided, having been grouped by country and geographic area.

To carry out the analysis of the geographic areas, we used the dimension that was created through the grouping made through the mapping between the Countries and the Area to which they belong. In the case of analysis by Country and considering the high number of possibilities, we chose to analyze the six most relevant origins in each of the months (top 6).

Knowledge about the origin of those who visit Lisbon is a very important piece of information since, from here, it is possible, for example, to act in terms of advertising campaigns abroad by the Tourism authorities, to promote Lisbon in countries whose inhabitants choose the capital of Portugal less as a destination to visit. This information may also be relevant to adapt the tourist offer in Lisbon to its visitors. It is known that habits change depending on the origin and to that extent, there may be simple actions that are to the liking of the visitors, making the experience in the city a little more pleasant.

Additionally, it may be possible to do something like targeted advertising and develop a differentiated tourist offer by country of origin. It is known, for example, that the countries of southern Europe tend to have less financial capacity than those of the center and north, to that extent and once they have obtained knowledge on this topic, entrepreneurs in the catering, hotel and tourist activities will be able to adapt its products according to the visitors.

Having said that, we begin the description of the analysis related to this issue, starting, starting with the evolution by country of origin of tourists and visitors in Lisbon. Therefore, and as described in Figure 4, we found that in the 5 months studied (September, October, November, December 2021, and January 2022), the top 6 of countries from which visitors to Lisbon do not change, being the Germany, Spain, France, Brazil, United Kingdom, and Italy.

However, as it is also possible to verify through the same graph, it is important to verify that the position that each of the countries occupies in the volume of Visitors changes over the analyzed months.

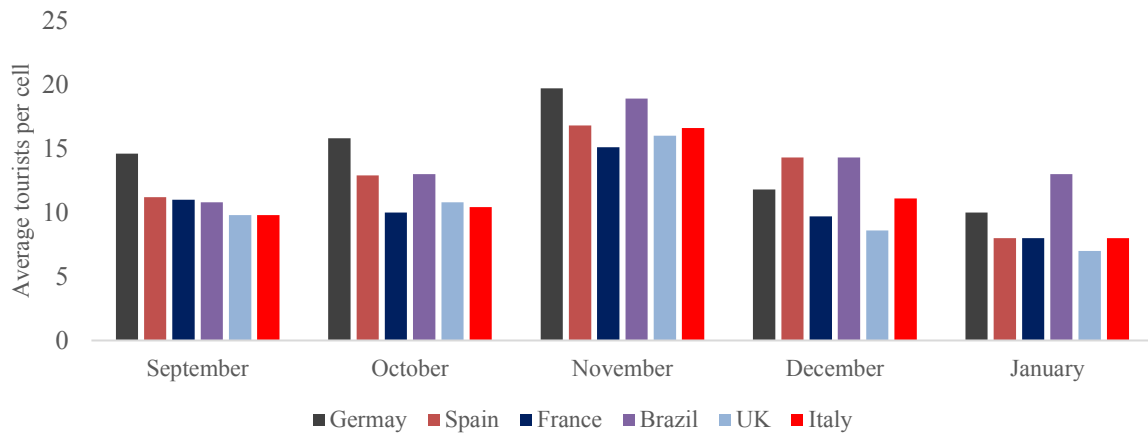Tourists evolution (average per grid) top 6 countries

Figure 4 -Top 6 Lisbon tourists' citizenship

Continuing with the analysis, this time at the level of Geographic Areas, the 5 most relevant were considered, to carry out an analysis like the one made for the countries. Therefore, and as can be seen through the graph in Figure 5, the geographic areas that most contribute with tourists and visitors in Lisbon are South America, Central Europe, the Alpine Countries and Northern Europe. We can also confirm that although there are changes of position over the months, these geographic areas are those that consistently provide the most visitors to the city of Lisbon. It is possible to verify through the same Figure that the geographic area that contributes most with visitants to Lisbon is South America, mainly through visitors from Brazil. However, it is important to mention that, if all Europeans are added together, they will clearly be the ones that contribute most to Lisbon visitors. However, we consider that they are people with different habits and, therefore, it made sense for the working group to proceed with this division.

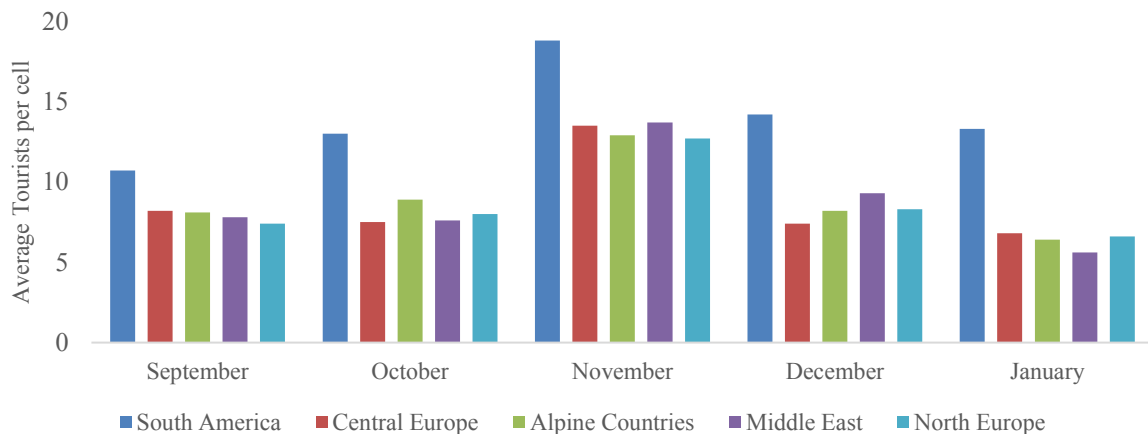Tourists evolution (average per grid) top 5 areas

Figure 5 - Evolution of Lisbon tourists by geographic area

4.2. What are the most visited areas of Lisbon?

As previously mentioned, we grouped the parishes of Lisbon and our first analysis of the most visited areas of the city falls precisely on this grouping. Not surprisingly, the areas most visited by tourists are the historic center and the city center of Lisbon (Figure 6), representing the average number of visitors in each of the 200 × 200 meters grids. Even so, it is interesting to check the average number of visitors in each of the grids and their evolution over the months.
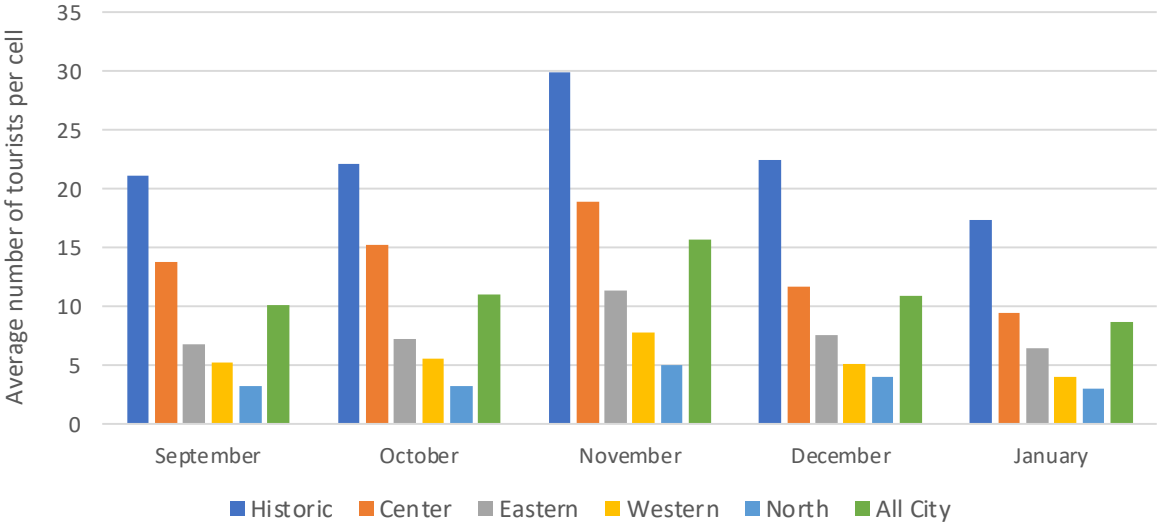

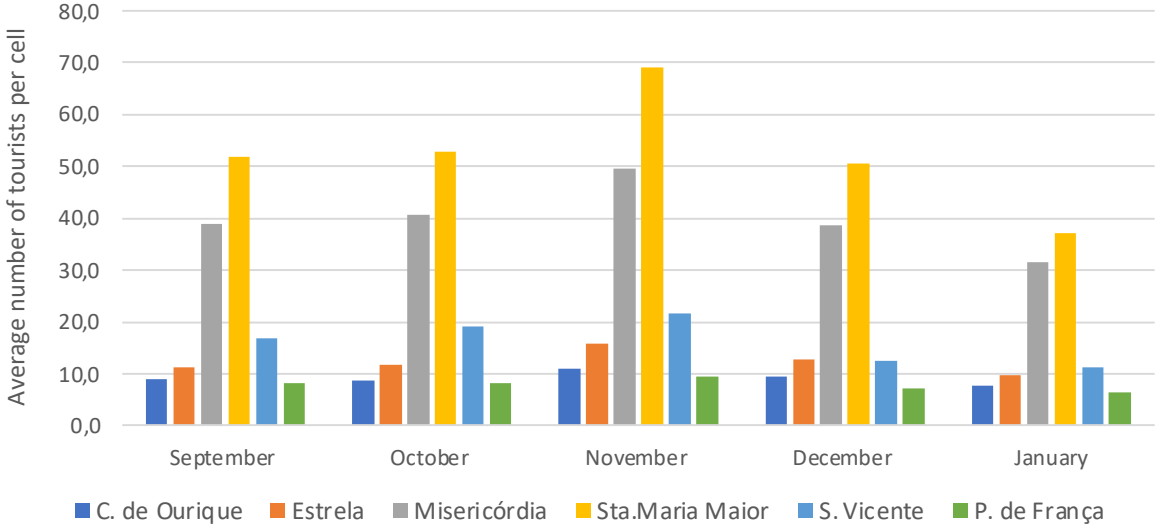*Figure 6 - Evolution of Lisbon tourists by month and city area*

Figure 7 - Evolution of Lisbon tourists by month in historical parishes

According to the available data and through Figure 7 and focusing on parishes in the historic center of Lisbon that receive the most foreign visitors we can see that Santa Maria Maior and Misericórdia are by far the most visited ones. This insight is consistent with the fact that it is in these parishes that very emblematic monuments of the city are located.

Through a georeferenced analysis, using the centroids included in the dataset of the mobile operator, we were able to understand the exact locations in the parishes of the historic center of Lisbon where Tourists travel the most. Analyzing Figure 8 that shows a heatmap of the tourists/visitor's concentration, it is possible to see highlight in Castelo, Alfama, Baixa Pombalina and São Vicente de Fora. These visualizations provide useful insights to decision-makers detailed information at street and hourly level.

Continuing with the analysis of the parishes most visited by tourists in Lisbon, it is still very important to characterize those belonging to the city center. Among the 6, and as we can see in Figure 9, representing the average number of visitors in each grid cell (200×200 meters), there are 3 that stand out for clearly having several tourists well above the others and they are Santo António, Arroios, and Avenidas Novas. Through the geographic analysis in Figure 10 representing the heatmap of city center visitors concentration it is possible to clearly perceive that Avenida da Liberdade and the Saldanha area are where more tourists move in the parishes of the center of the city.
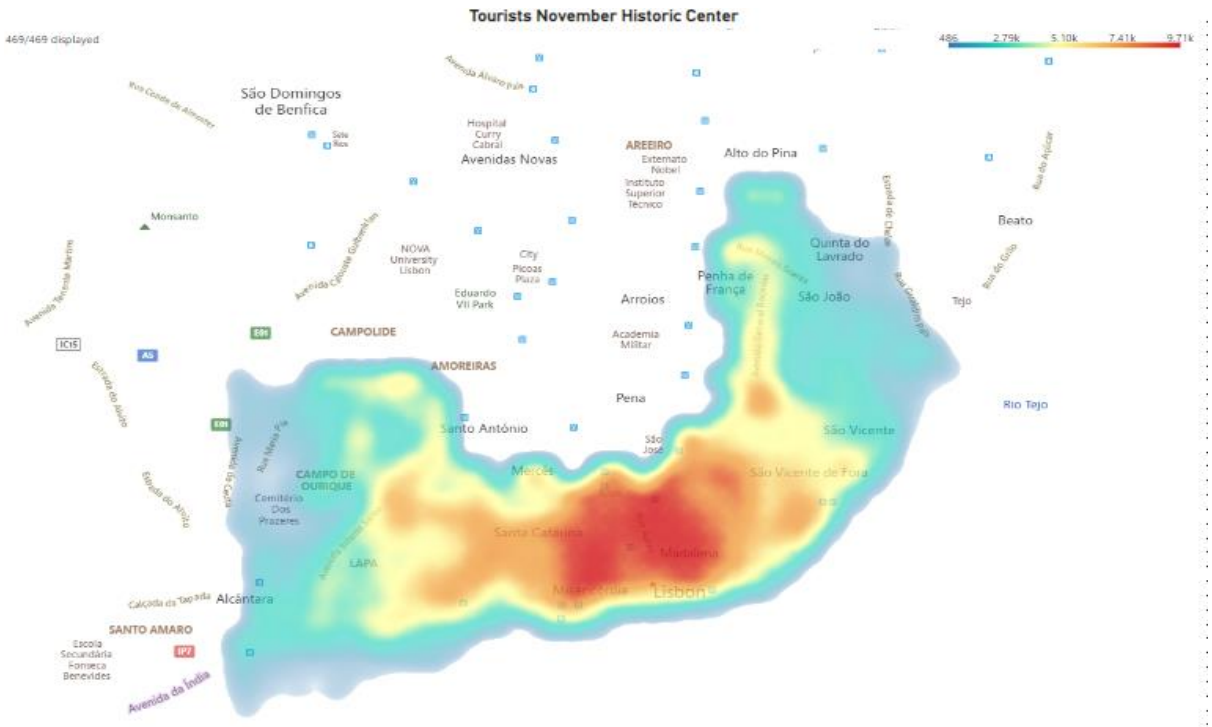

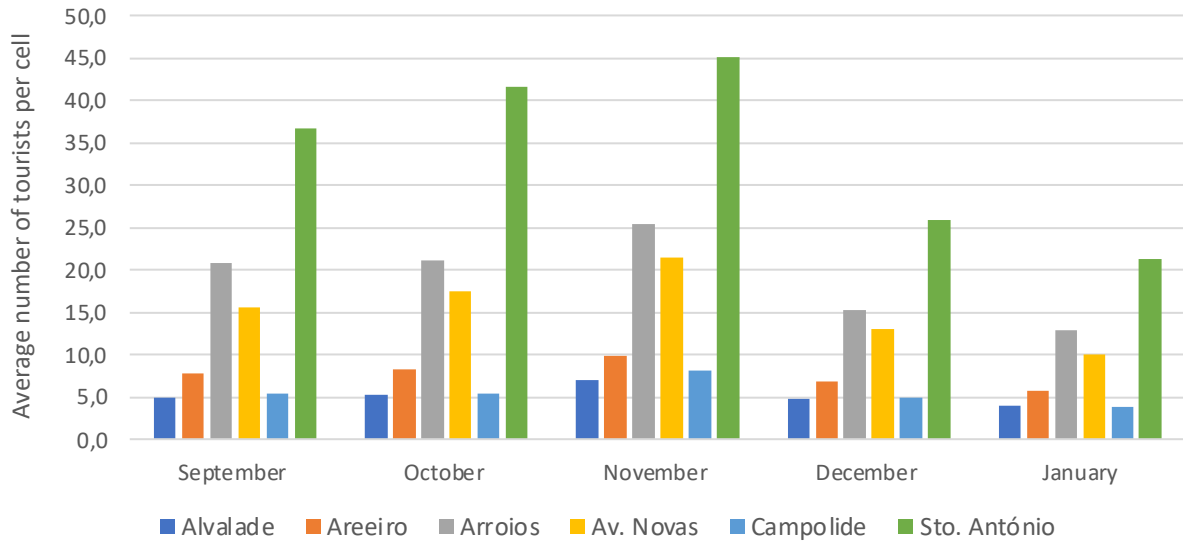*Figure 8 - Heatmap of Lisbon tourists in historical parishes (November)*

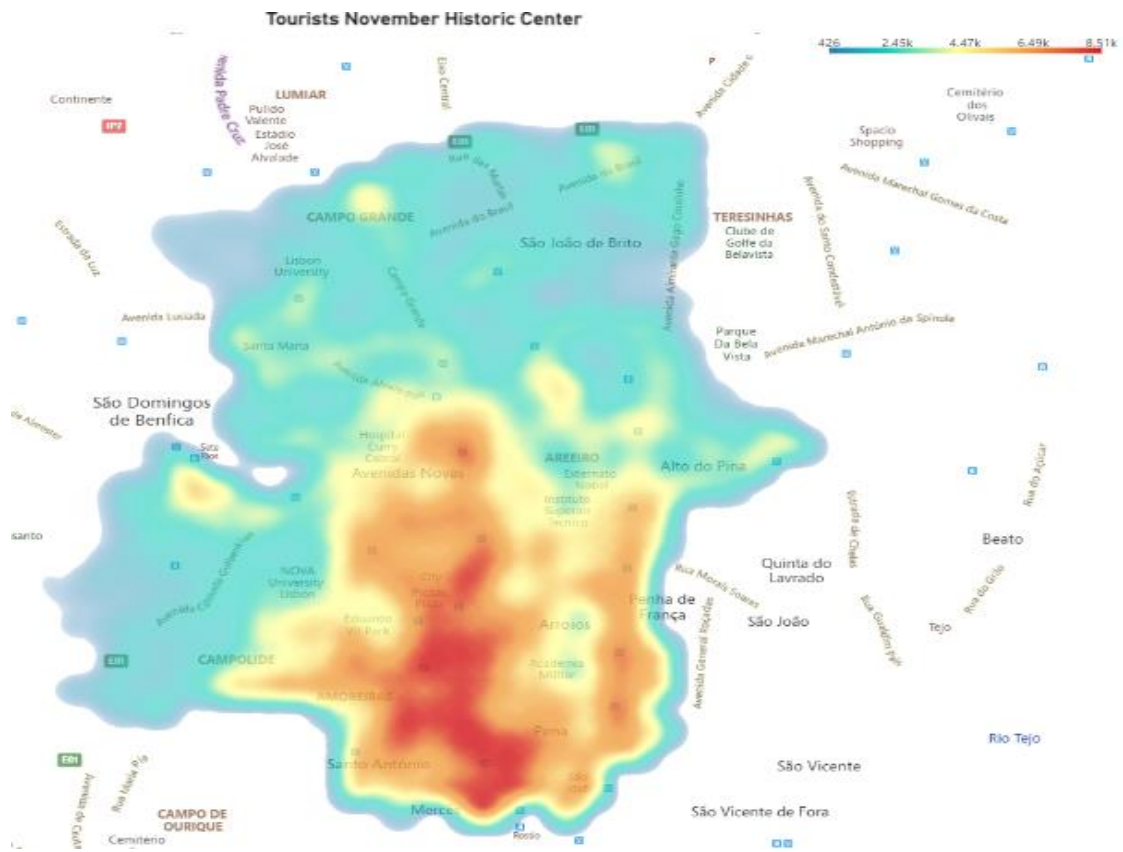*Figure 9 - Evolution of Lisbon tourists by month in city center parishes*

Figure 10 - Heatmap of Lisbon tourists in city center parishes (November)

4.3. Where are the Tourists during the mealtimes and where do they sleep?

Using the centroids present in the dataset of the mobile operator again and applying a time filter to the data, considering that the lunch period is between 12 am and 2.30 pm and dinner is between 7.30 pm and 10pm, we created the visualization in Power BI (Figure 11) showing the concentration of tourists in those periods. Carrying out the exercise for the month of September, even though the data are prepared to do so for any of the months, we can see that, except for the west zone of Lisbon (Belém/Alcântara), visitors have lunch and dinner in the same places in the city, which ends up making sense since it does not make much sense for them to travel to have their meals, unless the restaurant it is also a point of interest.

The data and visualizations are prepared to a level of detail that allows you to go down to the street and time level also allowing data from other months to be analyzed or possibly other time intervals, making it possible to carry out specific analyzes according to the user's needs.
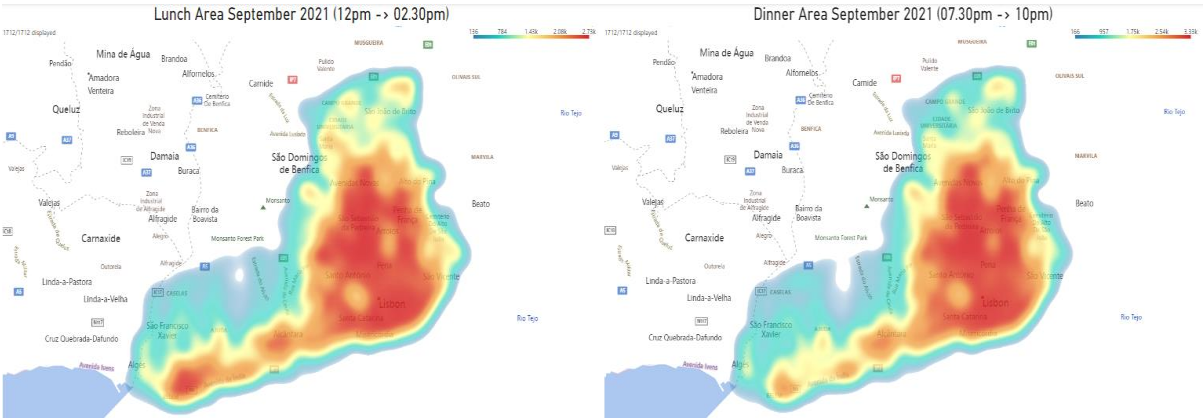

Figure 11 - Heatmap of Lisbon tourists at lunch and dinner time (September)

4.4. Event Analysis – Web Summit 2021

As is well known, the number of major international events has been growing in Lisbon, bringing an even greater number of visitors to the city, in addition to the already large number of tourists. However, the behavior pattern of someone who visits the city on business is different from someone who visits for recreation. To that extent, it makes sense that we can understand not only the differences between both profiles, but also understand how an event of international dimensions and with the capacity to bring thousands of visitors alters or not the usual pattern

of tourism in Lisbon. Through the analysis of the 2021 web summit, we intend to understand whether, during the days of the event, there were significant changes in relation to the places visited in the parish of Parque das Nações, as well as the origin of the visitors. Through this information, it may be possible to better manage transport and security and emergency means in the specific area of the event.

For the present work, we believe that it would make perfect sense to compare the volumes and origins on the days of the event, with the same days of the previous or subsequent week, this being the most correct way we found to make the comparison. Therefore, we developed the analysis of the Web Summit that took place between the 1st and 4th of November 2021, and for the counterpart days of the week before and the week after the event. Through Figure 12, it is possible to see that on the days of the event, tourists were mostly concentrated in the Lisbon International Fair while on the same days of the counterpart weeks were more spread out above by the points of interest of the parish, namely in the Oceanário. It is also possible to see that the number of Tourists on the days of the event almost tripled compared to the same period last year. It is also possible to verify that, in percentage terms, the geographic areas of origin of the visitors are also quite different. During the Web Summit week about 21% of the Tourists were from South America while in the previous and following week were from Eastern Europe.

Considering the analysis carried out, it was possible to verify that in the period in which the 2021 web summit took place, there was a change in the normal patterns of tourism in Parque das Nações. The changes were visible in terms of the areas that are usually visited, but also in terms of the origin of visitors to this parish of Lisbon. Bearing this information in mind, the managing entities of the usual points of interest in Parque das Nações can take actions aimed at promoting visits during the events of the coming years and, in this way, also leverage from a financial and indirect point of view.
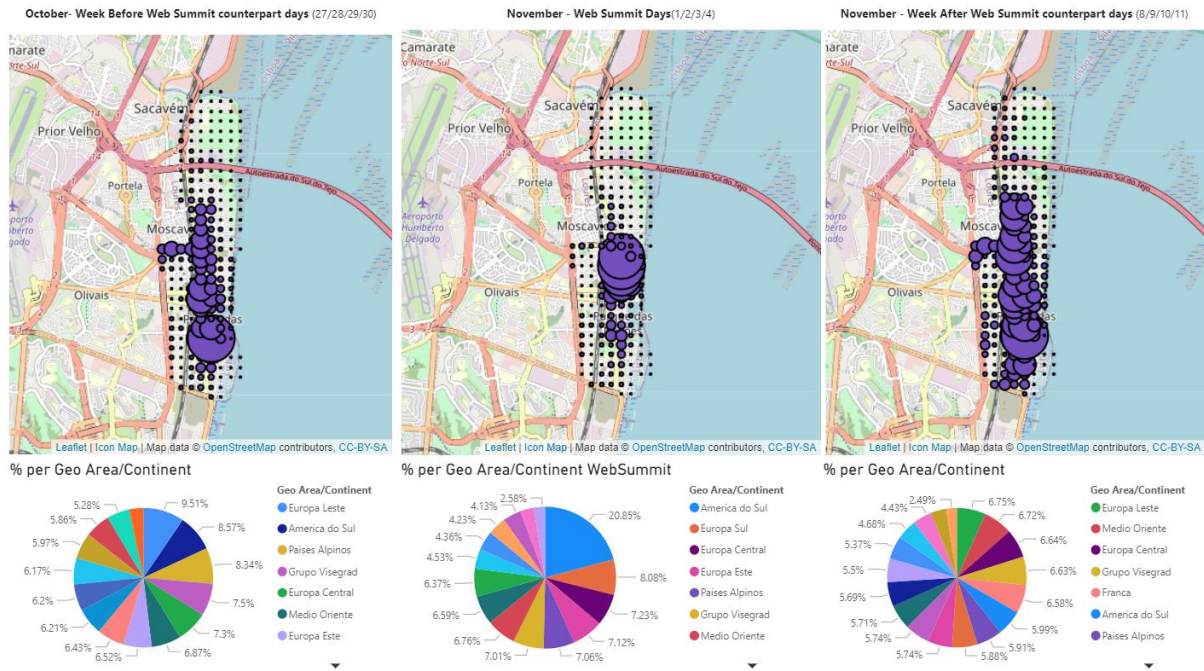
Figure 12 - Web Summit analysis

4.5.   Event Analysis – Champions League match

The influence of a sporting event was also analyzed, in this case, a Champions League football match held on October 20th - Benfica - Bayern. For this case, we only use visitors from Germany, as Bayern Munich is a German Club. In this analysis, we applied a time filter between 07 pm and 10pm, in the parishes of Benfica, Carnide, and São Domingos de Benfica. In Figure 11, it is noticeable that, compared to the same day and time of the previous and the after week, the number of Germans in the analyzed parishes grew almost 20 times and that their concentration was almost exclusively in the centroids of Estádio da Luz. As with the web summit, the Parish of Benfica can also boost financial income, promoting, for example, the typical restoration of large-scale match venues.

   The fact that there is a much higher than usual volume of German tourists in the area surrounding the stadium on the day of a Champions League game against a German team, does not represent anything new, being exactly what would be expected from find in this analysis. The exercise intends to demonstrate the capability of a tool like the one that was developed to help quickly and in real time take on events that may require, for example, the intervention of security forces.
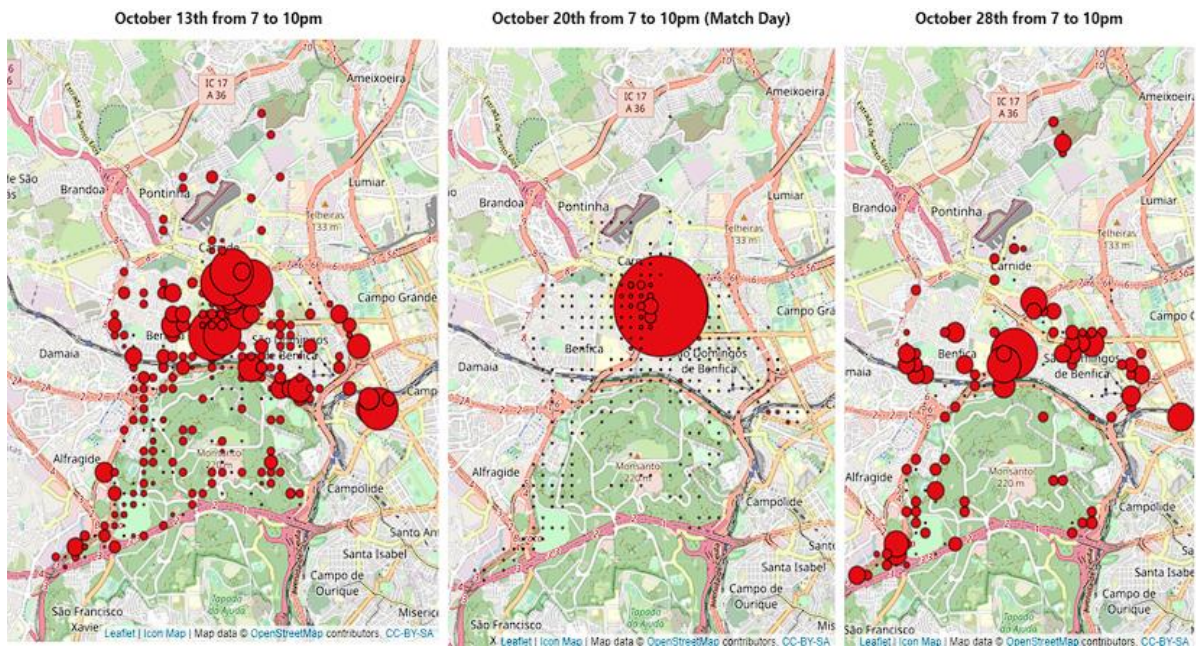
31

Figure 13 - Benfica vs Bayern

4.6.   Variation of tourists by day periods and hours

During the data analysis, it was possible to notice that in the period between 10 pm and 6 am, there is a decrease in the number of tourists in the city, as can be seen through Figure 14 (count of tourists in the city of Lisbon with an hourly aggregation). This topic seemed particularly relevant to us, and it would be important to provide the city Council with additional information. Several possibilities were considered, such as, for example, the fact that accommodation in Lisbon is too expensive and some visitors look for cheaper alternatives in the metropolitan cities of Lisbon, people who make a stopover at Lisbon airport, but end up not staying overnight, professionals who travel just for a meeting, an isolated event and who end up returning, or, eventually, visitors who are not in Lisbon, but who visit the city. We questioned the City Council about the existence of a known explanation for this fact, without success. This loss of tourists had already been effectively identified by the Municipality; however, they had no further details. From here, we realized that it would be of all relevance to look for answers in the available data.
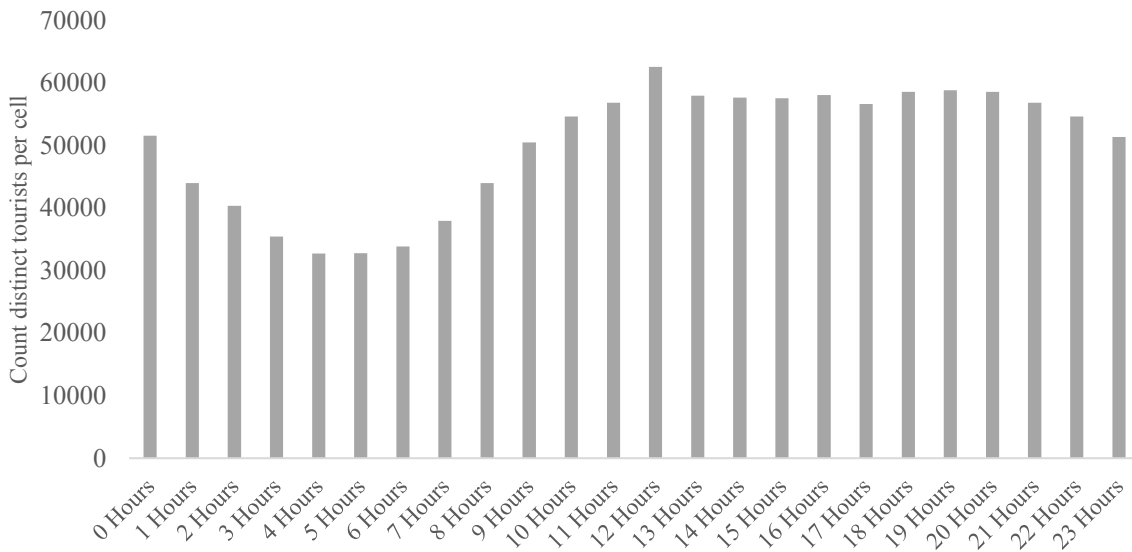
Evolution of tourists by hour

Figure 14 - Hourly evolution of tourists in Lisbon

The first approach was to understand whether this behavior was different depending on whether it was a weekday or a weekend. Through Figures 15 and 14, where we have the hourly variation segregated by day of the week, it is possible to see that regardless of the day, this phenomenon is always verified and therefore, according to the available data, the day has no influence as the decrease of tourists is observed in regular weekdays but also on weekends, meaning that we needed to carry on with our investigation and analysis in order to be able to retrieve some deeper information regarding this topic and be able to get the most out of our available data.
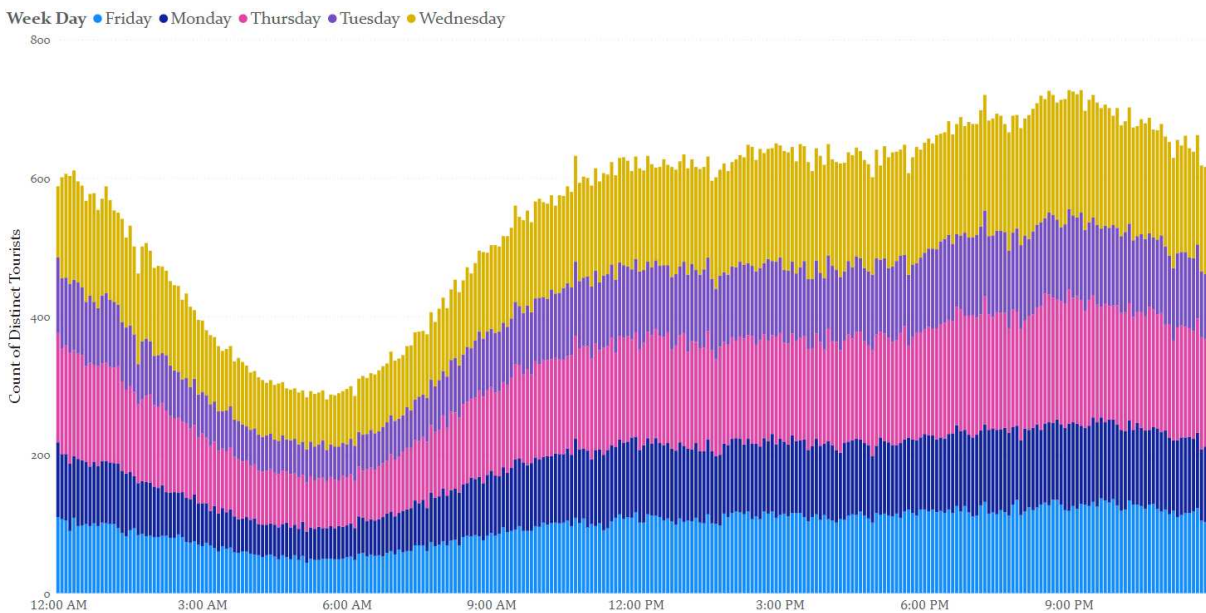

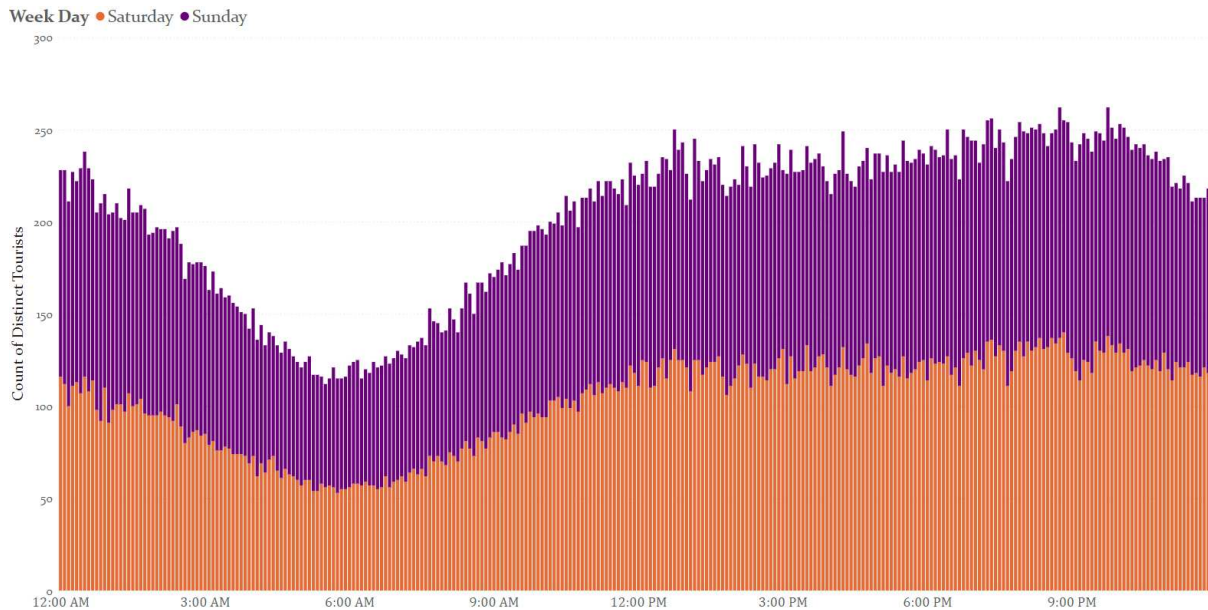*Figure 15 - Evolution of tourists in Lisbon by weekday*

Figure 16 - Evolution of tourists in Lisbon by weekend day

We considered that the topic deserved careful analysis and, therefore, the next approach was to understand whether the origin of the visitors could provide additional data regarding the decrease in the number of tourists during the mentioned period (22:00 - 06:00).

Bearing in mind that we created the "Continent" and "Hour" dimension in the data to assist us in analyzes that made sense to us to extract the greatest possible value from the data, we carried out, in a first phase, an analysis based on both created dimensions. The objective of this analysis was to understand if, among the tourists/visitors from different continents, with a "cut" by time of day, we could effectively perceive possible differences. The reason for starting an analysis at the level of the continente is related to the fact that it is simpler to start at a higher hierarchy than to proceed directly to an analysis by country, mainly for two reasons; a) it would be necessary to analyze more than two hundred countries and, b) because some of the countries have a very small and volatile number of Tourists/Visitors to Lisbon and, therefore, it would end up not adding value. In this way, we chose to start by analyzing the average values of tourists by continent and hour, keeping in mind that from there, it could be possible to identify in a more assertive way which nationalities cause the decrease of tourists in the period between 10 pm and at 6 am. It is also important to mention that in the analysis, tourists/visitors from Oceania were not considered since their number is perfectly negligible and too volatile compared to the other continents - Africa, America, Asia, and Europe.

Therefore, and after grouping the data as defined above, it was possible to produce the graph described by Figure 17. From this analysis, it was possible to perceive that the

visitors/tourists in Lisbon city that most contribute to the decrease verified between 22:00 and 06:00 are from the European continent.

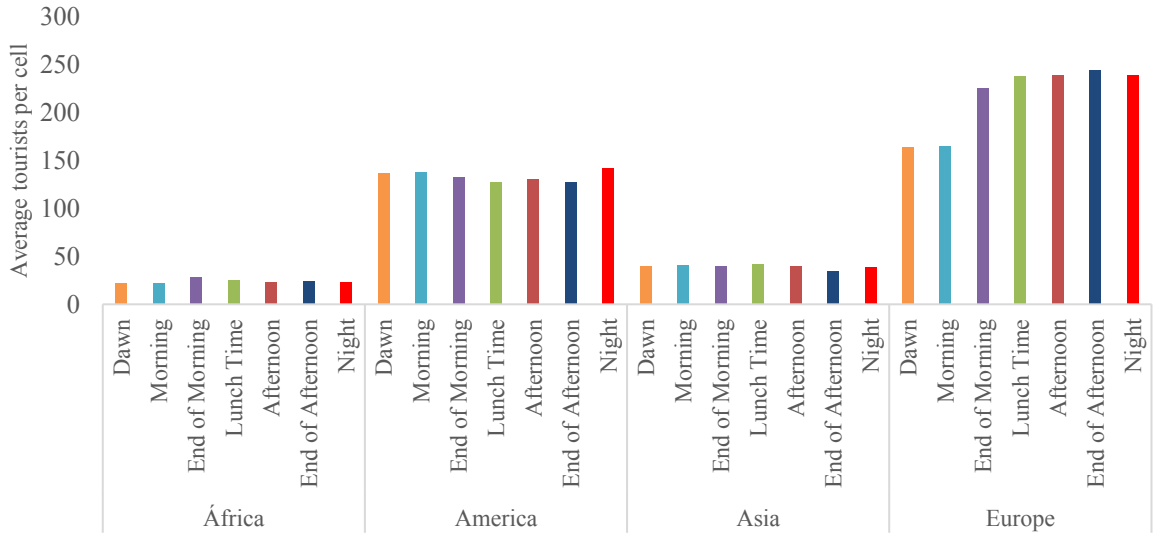Evolution of average tourists by day period

Figure 17 - Evolution of tourists by day period

Based on the analysis above, and as mentioned in the previous paragraph, it was possible to perceive that Visitors/Tourists from mainland Europe are the ones that most contribute to the decrease in Lisbon in the period between 10pm and 6am.

That said and considering that we have the data available in our dataset, we proceeded to the analysis of the Top 10 of the European countries whose inhabitants most visit Lisbon. The countries in question, and as can be confirmed through Figure 18, are Germany, Spain, France, United Kingdom, Italy, Belgium, Switzerland, Ireland, Denmark, and Austria. Still using the same graph, it is possible to see that the first 5 are notably responsible for most tourists/visitors. To produce this analysis, we used an hourly aggregation of the data, duly filtered by each of the above-mentioned countries and counting each hour the number of different visitors by each of the origins. From this analysis and as we can confirm through Figure 19 that the visitors causing the decrease between 10 pm and 6 am are Germany, Spain, United Kingdom, Italy, Ireland, France, and Belgium. On the other hand, through the same graph, we were able to confirm that visitors/tourists from Austria, Switzerland, and Denmark, despite being in lower numbers compared to other European countries, their volume is stable practically throughout the day.
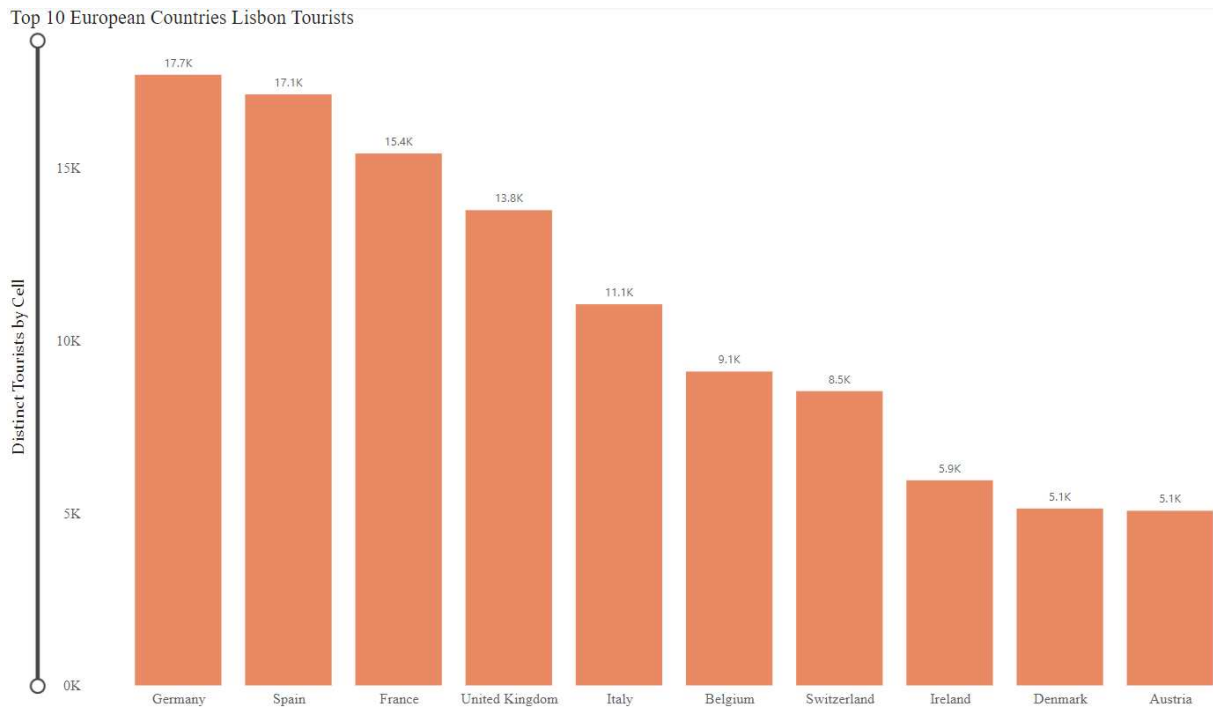
Figure 18 - Top 10 European countries Lisbon tourists

Evolution of top 10 European countries tourists by hour
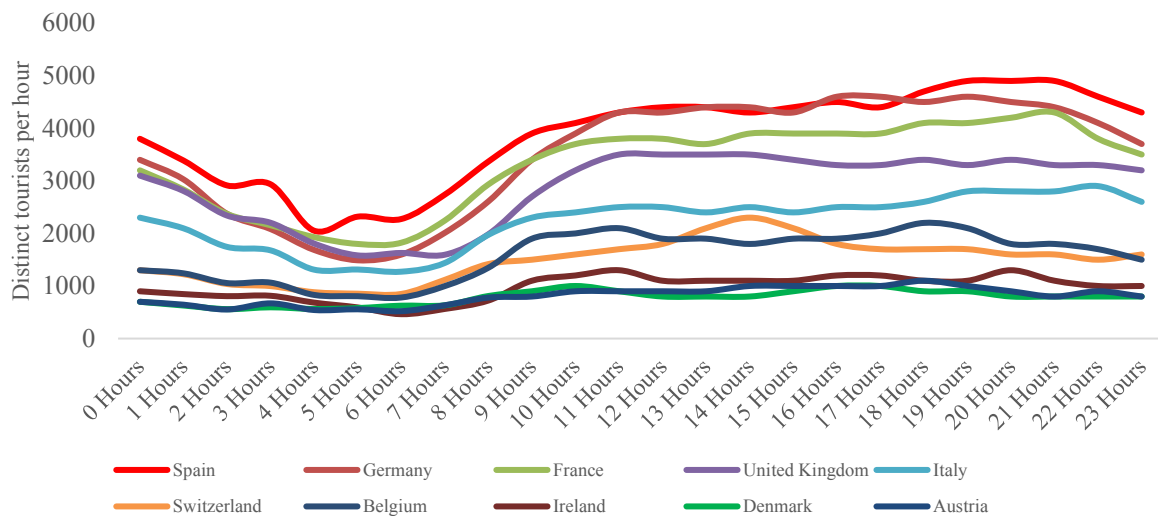
Figure 19 - Evolution of top 10 European countries tourists by hour

In the available data, there is no possibility of being sure of the reasons for this decrease, in the same way as those responsible for Lx Data Lab, did not provide us with additional information. That said, it is our understanding that the decrease may eventually be related to:

a) Tourists sleep outside Lisbon because it is cheaper

b) Visitors who are in Portugal but elsewhere and come to Lisbon for a few hours and then return to other places

c) Phones whose mobile data is turned off

d) Late departures at the airport

During this dissertation, we reported to the Lisbon City Council, through the Lx Data Lab, this decrease in Tourists between 10pm and 6am and we show the results we were able to find with the data we had available. However, it was not possible to receive concrete reasons, and as far as we were told, the topic has not yet been properly explored. It is undoubtedly relevant to find the reasons that lead to this decrease, since the city of Lisbon may be losing revenue from overnight stays and meals, a loss that can eventually be mitigated, if the actual reasons are identified.

4.7.   Variation of tourists in Lisbon parishes

The City of Lisbon has 24 parishes with very different characteristics, some of which are mostly residential and without a relevant interest from the point of view of tourism, and others - located in the center and historic center of the city where it is possible to find the places of greatest interest and concentration.

The need to analyze the movements of tourists/visitors between each of the parishes at different times of the day, seemed to us to be a topic of particular interest, insofar as it also allows us to typify the behavior of tourists, which can be interesting information in the aid in data-based decision making.

Our objective is to understand which parishes which parishes gain/loss tourists in different periods of the day, being in this way possible to identify which parishes gain or lose visitors in the night and dawn period, and from there it is possible to analyze, for example, the parishes that function mainly as a dormitory for tourists and, on the other hand, those that are lost during these periods of the day.

To carry out this analysis, the aggregation of parishes by area of the city was used, which was carried out in the data preparation phase, as well as the day moments created in the same phase of the dissertation. Therefore, the parishes were grouped as follows: a) historic center of the city: Campo de Ourique, Estrela, Misericórdia, Santa Maria Maior, São Vicente and Penha de França; b) city center: Alvalade, Areeiro, Arroios, Avenidas Novas, Campolide and Santo António; c) eastern part of the city: Beato, Marvila, Olivais and Paque das Nações; d) west Zone of the City: Ajuda, Alcantara and Belém; e) north zone of the City: Carnide, Benfica, Lumiar, Santa Clara and São Domingos de Benfica.

Starting by analyzing the Parishes of the historic center of Lisbon, and as can be seen through Figure 20, we found that the parishes of Campo de Ourique, Estrela, Penha de França and São Vicente are the ones that have a smaller number of distinct visitors during the day, however, there was an increase, albeit slight, in the number of tourists during the night and dawn period. On the other hand, the parishes of Misericódia and Santa Maria Maior are the ones that clearly stand out in terms of volume of visitors, as previously mentioned. These two parishes also have a greater number of visitors in the afternoon, late afternoon, night, and dawn compared to the morning, late morning, and lunch hours.

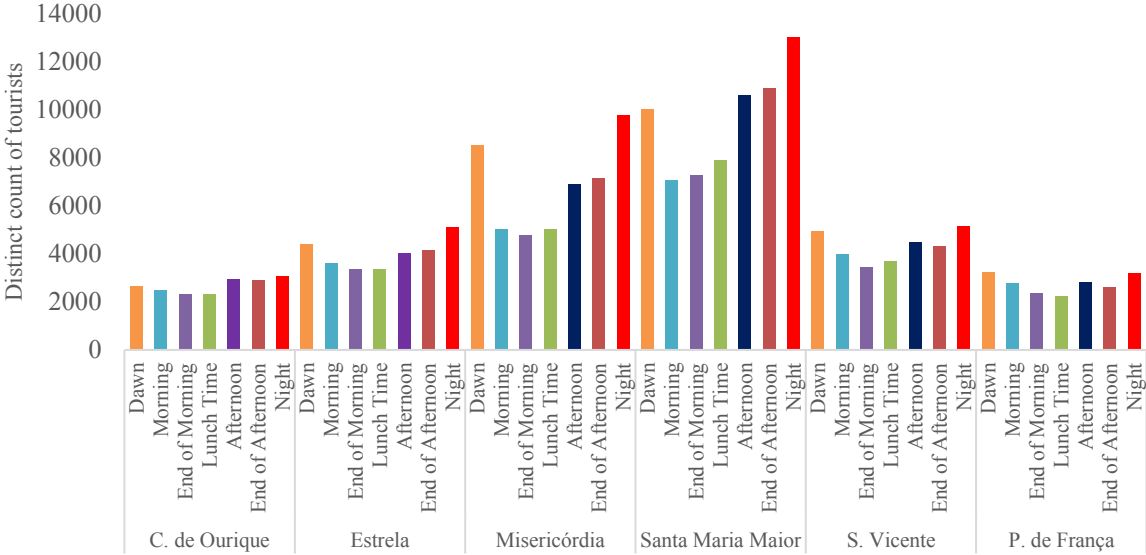Evolution of tourists in city historical parishes by day moment

Figure 20 - Evolution of Tourists in city historical parishes by day moment

Continuing the analysis by each of the Parishes of the center of Lisbon as we can confirm through Figure 21, the Parishes of this group that receive more Visitors/Tourists are Santo António, Avenidas Novas and Arroios. It is also possible to see that in these Parishes, the volume of Visitors is higher during the night and dawn periods, which makes sense since even though these places have a lot of movement because they are located in the center of the City, they do not really have major tourist attractions, contrary to what exists in the Parishes of the historic center, which ends up making these parishes serve mainly as places of passage and dormitory.

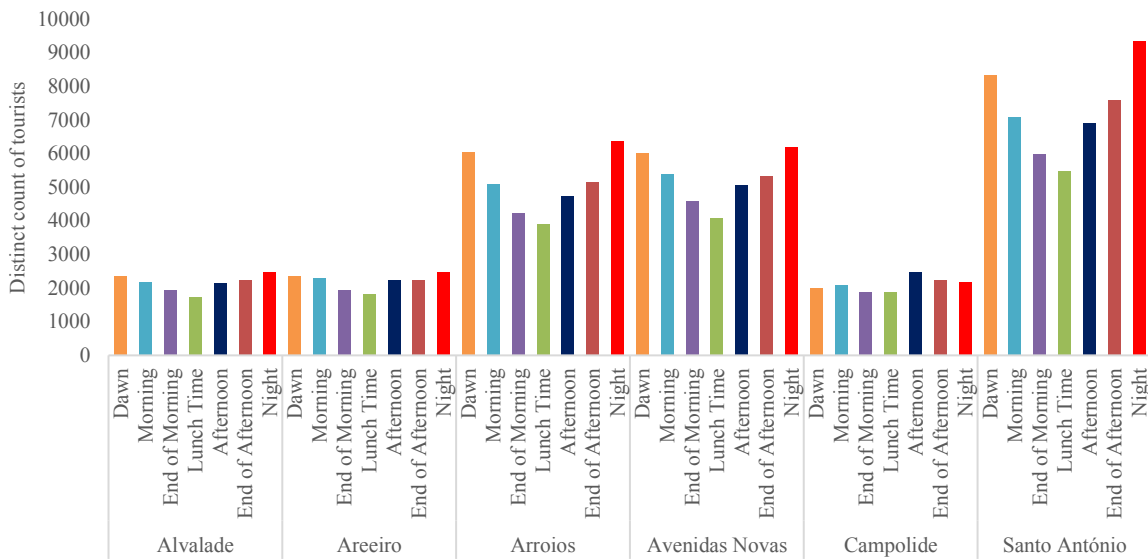Evolution of tourists in city center parishes by day moment

Figure 21 - Evolution of tourists in city center parishes by day moment

Analyzing the group of parishes of eastern Lisbon - Figure 22, we can see that Olivais is the parish that receives the most tourists and visitors from Lisbon, mainly because of the airport, it is interesting to note that most flights arrive or depart during the periods of morning and afternoon, according to the fact that it is at these times of the day that the concentration is greatest. The parish of Parque das Nações is more frequented between late morning and late afternoon, which is consistent with the hours of points of interest such as the Oceanário, Fil and the existence of several companies that receive visitors who travel in work - Parque das Nações has several multinationals, such as Microsoft and Vodafone. The remaining parishes - Marvila and Beato receive very few tourists, which makes perfect sense because they are effectively places with little interest in terms of monuments and places to visit.

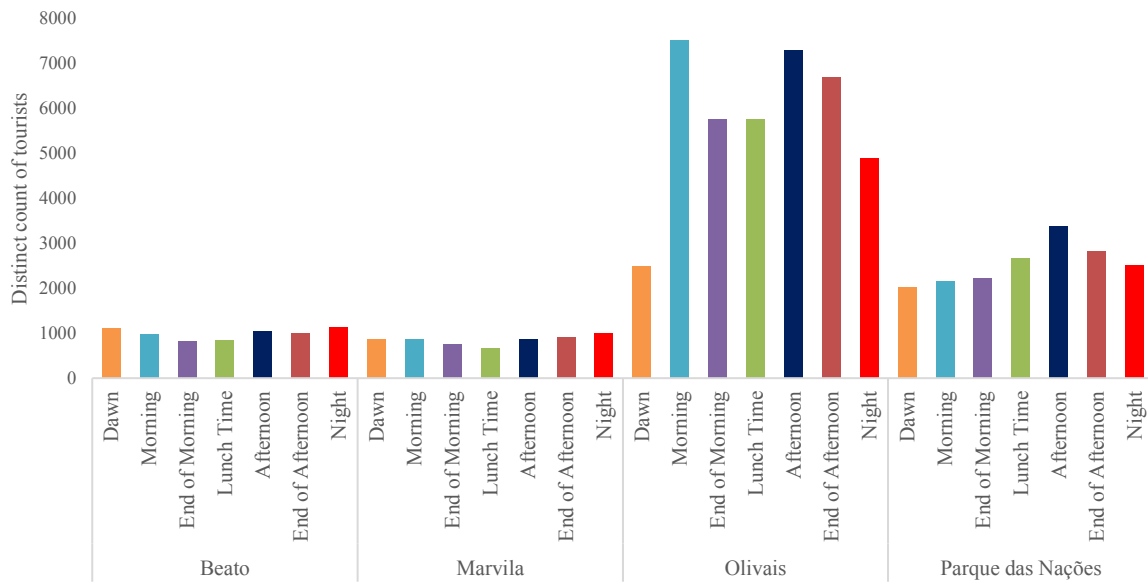Evolution of tourists in city eastern parishes by day moment

Figure 22 - Evolution of tourists in city eastern parishes by day moment

Finally, we analyze the remaining parish Groups of Lisbon – western and north zone. In the first group, which includes the parishes of Ajuda, Alcântara and Belém (Figure 23), it is possible to see that Belém is by far the most visited area, especially between late morning and late afternoon, once again, we were able to establish a relationship between the visiting hours of monuments such as the Torre de Belém and the Jerónimos Monastery, in addition, the area by the river, is quite visited during the times of day mentioned. It also makes sense to mention that there is a significant decrease in Tourists/Visitors in this parish during the dawn and morning, which shows us that there may be little supply in terms of overnight stays or, eventually, they are too expensive for most people. In the case of the Parish of Alcantara, we can also see that the highest number of tourists is verified during the afternoon and at night, which is related to the offer of restaurants and nightlife. Ajuda is the Parish of this group that receives the fewest visits, and the volume remains practically unchanged during all times of the day.

Evolution of tourists in city western parishes by day moment
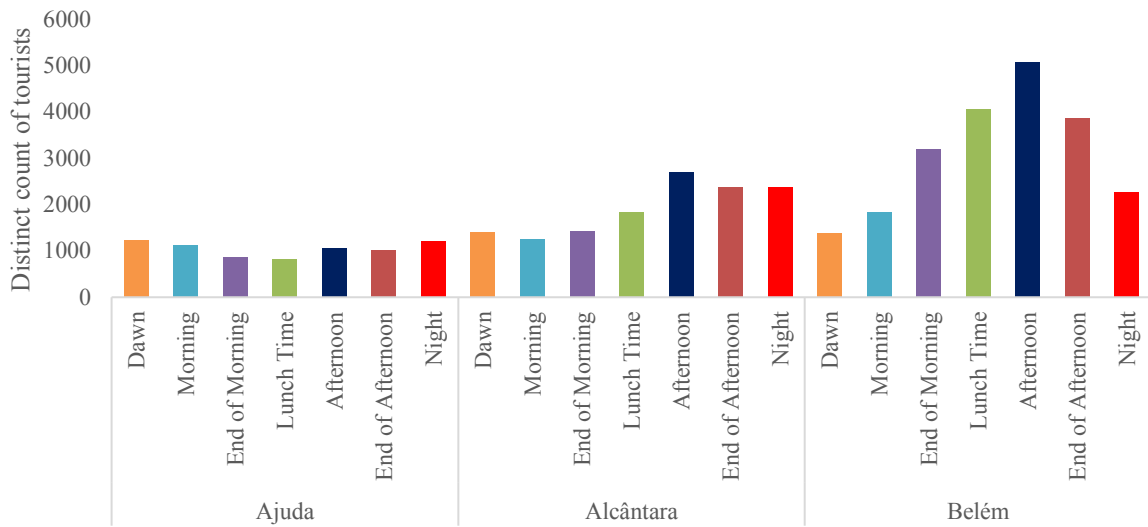
Figure 23 - Evolution of tourists in city western parishes by day moment

In the case of the North of Lisbon - Figure 24, and except for Carnide, we can see that there are more tourists at night and at dawn when compared to the rest of the day - which may reveal that they are parishes mostly used as places to sleep. Carnide is a particular case in terms of the number of visitors, due to the presence of the Colombo shopping center and the estádio da Luz.

Evolution of tourists in city north parishes by day moment

Figure 24 -Evolution of tourists in city north parishes by day moment

4.8. Work assessment

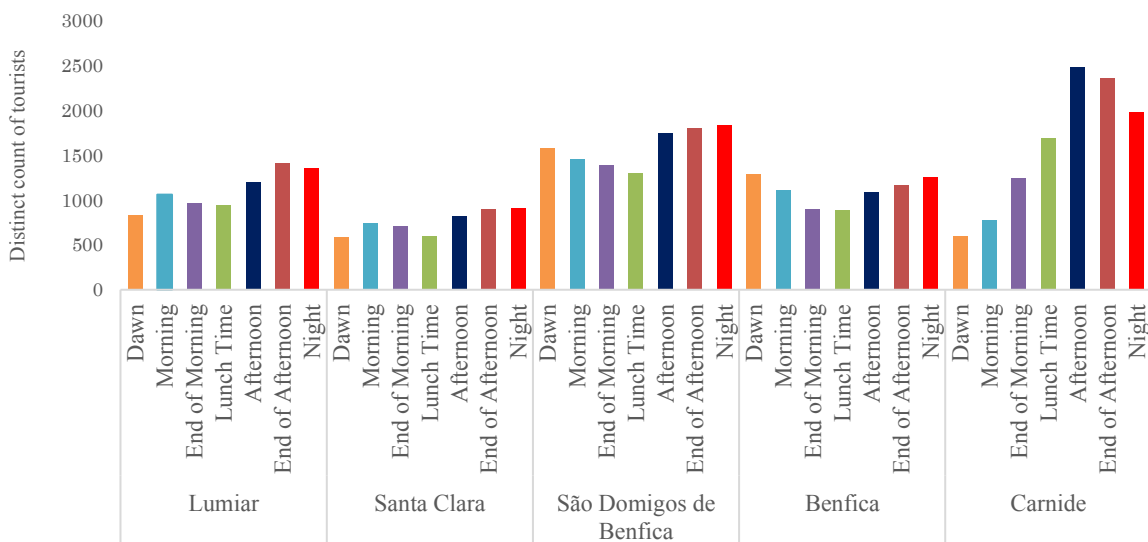As mentioned, several times during this dissertation, the work was done in collaboration with Lisbon city Council and Lx Data Lab. To this extent, and once the work was carried out in terms of visualizations and analysis, we considered that it made sense to show what was done to decision-makers and users, not only to obtain the validation of the analyses, but also to obtain feedback and understand the need for any adjustments.

During the meeting, we presented everything that was developed, with a particular focus on answering the questions we intend to answer, and the tool developed and its possibilities and capabilities. We chose to leave aside the technicalities related to the way the data were processed, since the audience's main interest was precisely in the analysis and operation of the tool. At the end of the presentation, there was a question-and-answer session that lasted approximately thirty minutes and in which it was possible to clarify some doubts. At the end of this session, Lx Data Lab, through its managers, informed us that the analyzes and visualizations obtained through the work carried out not only made sense, but also considered them important. After submitting and discussing this work, we were asked to deliver the developed application to the Lisbon City Council.

It was during this meeting that we revealed to Lx Data Lab the analysis regarding the decrease in tourists/visitors during the night period and we were informed that it was a fact that they were completely unaware of. The management of the Lx Data Lab team asked us to do some more research on this issue within the scope of this work and that any findings be shared. Therefore, and considering all that we were able to analyze with the available data, we inform them according to what was described above, in Section 4.7 of this dissertation.

To obtain quantitative and qualitative feedback on the work we have developed, we have created a simple questionnaire that we send to decision makers/users, especially to be able to sustain this dissertation and keep the record for future memory. Table 6 describes the questions that were asked in the questionnaire sent to Lx Data Lab as well as the possible answers to each one. The questions asked were as summarized as possible and without great subjectivity to obtain a representative number of answers, since it is known that questionnaires that are too long tend not to be answered or to have skewed answers.

Table 6 - Questionnaire content

Question 1	Possible answers
From the roles below, which one best describes your responsibilities at Lx Data Lab/City council	Manager / Decision Maker
	Data scientist/engineer
	Data analyst
	Technician/senior technician
	Other
Question 2	**Possible answers**
Do you carry out most of your work at Lx Data Lab or at another department of the Lisbon City Council?	Lx Data Lab
	Other city council department
Question 3	**Possible answers**
Do you consider that the use of data is an essential tool in Decision Making and Resource Management in the context of Tourism in Lisbon or do you consider that there are better ways of doing it ?	Yes, I consider it essential
	I do not consider essential
	There are better ways to do it
Question 4	**Possible answers**
Does the approach taken to the available data seem the most correct to you	Yes
	No
Question 5	**Possible answers**
Do the data shown seem useful and relevant in the Decision-Making process and to enhance a better experience for Lisbon Tourists/Visitors?	Yes
	No
Question 6	**Possible answers**
Did the results presented turn out to be innovative, relevant, or presented from a more interesting perspective than usual?	Yes
	No

From the questionnaire and has we can confirm trough Figure 25 it was possible to ascertain that, in relation to Question 1, 83% of respondents, corresponding to 5 answers, are Senior Technicians/Technicians, while 16%, corresponding to 1 person, is a Manager/Decision Maker. Regarding Question number 2, 83% of people work at Lx Data Lab while 16% work in another department of the Lisbon City Council.

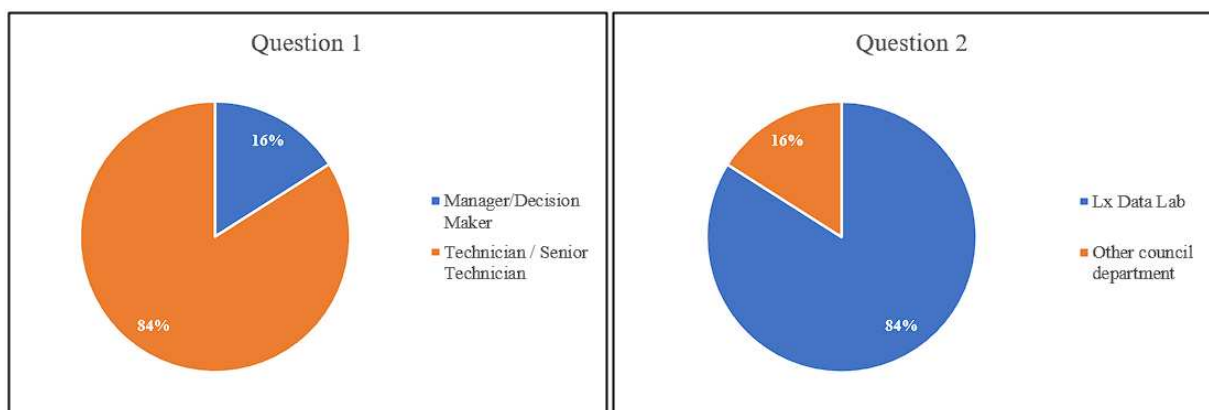*Figure 25 - Answer to questions 1 and 2*

The remaining questions (3,4,5 and 6), and as already mentioned, intended to assess the opinion of our "Customer" not only regarding the work developed in this dissertation, but also regarding the importance of the data and its analysis as an aid to decision making regarding tourism in Lisbon. Through the description above and Figure 26, it is possible to effectively

confirm not only the importance of the data, but also the validity that the Lx Data Lab and the Lisbon City Council attributed to the analyzes developed and presented since 100% of the respondents to the questionnaire consider that the use of data is essential for the management of tourism in Lisbon. Additionally, we were also able to confirm that all those who answered the questionnaire confirm that our approach to the data was correct, that the data shown allow us to improve the experience of tourists in Lisbon and that our analyzes represented not only innovation but also a perspective more interesting than usual in terms of how the data were explored and analyzed.

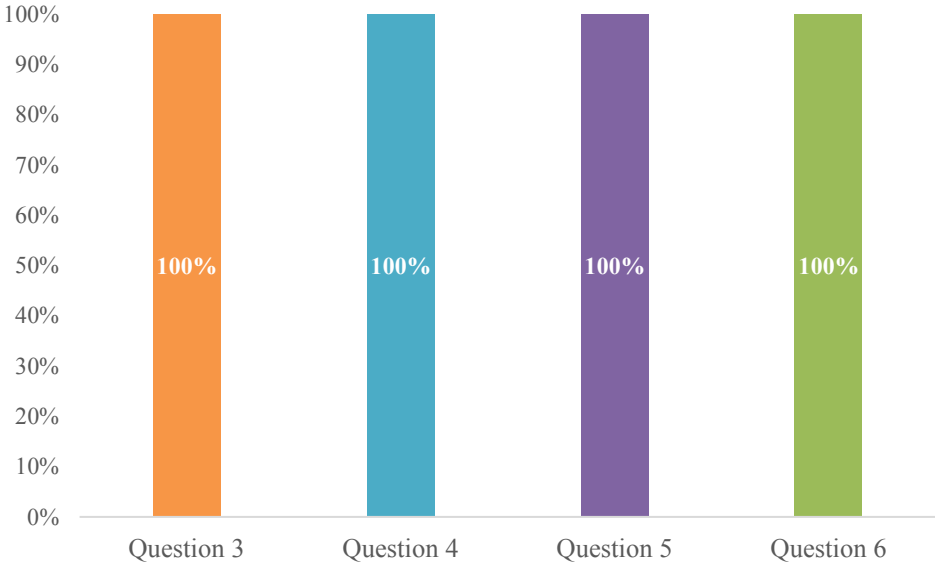*Figure 26 - Results to questions 3, 4, 5 and 6*

The absolute value of responses is not actually relevant in statistical terms; however, it is important to emphasize that those who attended the presentation are effective users of data and develop their professional activity and make their professional decisions using them and therefore, the results gain importance and validate the work that was developed added value.

CHAPTER 5

5. Conclusions and future work

5.1. Conclusions

It is essential for public policy to comprehend the spatial spread of urban tourism. As a result, in areas where there is a high concentration of tourists, local authorities might think about initiatives to improve the tourist experience, like creating pedestrian-only lanes or enlarging sidewalks, increasing the number of public spaces with free Wi-Fi, and positioning new tourist information centres, among other things.

Through the method developed and the tools used, we were able to answer the questions we wanted to. It was possible to identify the origin, and the number of Tourists in Lisbon as well as the evolution over the period under study (September 2021 to January 2022), making a separation not only by Country but also by Continent / Large Geographical Area. We were able to understand and demonstrate which areas and parishes are most visited in the City of Lisbon and the places where they are during the typical meal and sleep times. Finally, we were able to understand in a way that a major international event such as the Web Summit or a football match in the Champions League changes the panorama of tourists/visitors in Lisbon.

In that sense, properly adapting our method to handle a constant flow of data, public policy makers, like the Lisbon Municipal Council or the National Tourists Office, can take advantage of an aggregated view that in real time manages the resources of various departments ranging from transport, hygiene, and safety. By processing this data in real-time, everyone involved in the management of the city in its various vectors will be able, on the one hand, to provide a much more pleasant experience for visitors, but also to avoid security breaches and any type of unwanted events Additionally, our work can encompass an analysis tool based on real data that allows, in the medium and long term, to plan accommodation and commerce.

It is also important to mention that with the available data, we can have much more insights than the ones we refer to in this work. However, due to scope limitations, we have chosen the one that seems to have the best fit. The developed tool was designed with the possibility of using a series of filters that allows the Lisbon City Council to make its own analysis on topics that were not explored in this dissertation, making more final analysis both from a temporal and geographical point of view. If users deem it necessary, they can still use the solution developed

to load new data, if they follow the same structure, making it possible to use it in analyses with new information coming from the same source.

To process an amount of data of this order to obtain real value for the benefit not only of Tourism, but of all those who travel through Lisbon for work or leisure, it is essential that there is a large computational and analytical capacity. For the first case, there should be recourse to cloud technology, which, as is well known, although with high costs, allows processing large amounts of information, without delay and without the need to install local capacity. From an analytical point of view, it makes sense to develop machine learning mechanisms, capable of highlighting patterns and helping data analysis, providing decision-makers with automated reports and dashboards to support decision making.

Based on the present dissertation and work developed for it, a scientific article was written with the title "Analysis of the Tourists behavior in Lisbon using Data from a Mobile Operator" which was submitted and accepted for the EAI INTSYS 2022 - 6th EAI conference International Conference on Intelligent Transport Systems which will take place in Lisbon on the 15th and 16th of December 2022. The submitted article focuses mainly on a subset of the analyzes and visualizations developed by the working group that produced this dissertation.

5.2.   Future work

Considering the current real-time processing and storage capacity, we consider it a proposal for future work, in the sense of using this data in two ways:

a) Creation of a system that allows real-time analysis of georeferenced data. In this way, it would be possible to perceive possible movements and unexpected clusters of Tourists that may foresee a need for intervention by security, emergency, sanitation, transport, or other services. As an example, we can list a football game where fans of rival Clubs get involved in conflicts in an unsupervised location - the cluster of Visitors/Tourists from the countries to which the Clubs belong, may be a cause for concern and the mobilization of security forces and emergency can potentially prevent bigger problems.

b) Use of data for medium/long term planning. In this proposal, we suggest the use of data to understand how Tourists/Visitors are evolving and, from there, plan the Restoration, Hospitality and Transport resources, and it is even possible to develop Machine Learning mechanisms/applications that can predict the way in which Tourism will evolve. In this way, it is possible to guarantee that all the logistics are adjusted to the needs and there are no limitations

to the reception of a greater volume of Visitors. Additionally, this can be a mechanism that makes it possible to understand whether Lisbon is reaching the limit of ceasing to be a pleasant City for Visitors.

Both proposals aim to improve the experience of Tourists/Visitors in Lisbon and, in this way, enhance the positive impact of Tourism in the Capital, enhancing data-based decisions.

References

1. 1LISBONEIDEE Homepage, http://www.lisbonne-idee.pt/p5358-lisboa-recebe-mais-turistas-por-residente-que-londres.html, last accessed 2022/07/23.

2. Câmara Municipal de Lisboa Homepage, https://www.lisboa.pt/, last accessed 2022/07/23.

3. LxDataLab Homepage, https://lisboainteligente.cm-lisboa.pt/, last accessed 2022/07/23.

4. Secretaria Geral da Economia Homepage, https://www.sgeconomia.gov.pt/noticias/portugal-e-o-5-pais-com-mais-forte-contributo-do-turismo-para-o-pib.aspx, last accessed 2022/07/24.

5. M. Fekih, T. Bellemans, Z. Smoreda, P. Bonnel, A. Furno, and S. Galland, "A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France)," *Transportation*, vol. 48, no. 4, pp. 1671–1702, 2021, doi: 10.1007/s11116-020-10108-w.

6. E. Carter, P. Adam, D. Tsakis, S. Shaw, R. Watson, and P. Ryan, "Enhancing pedestrian mobility in Smart Cities using Big Data," *J. Manag. Anal.*, vol. 7, no. 2, pp. 173–188, 2020, doi: 10.1080/23270012.2020.1741039.

7. C. Badii, A. Difino, P. Nesi, I. Paoli, and M. Paolucci, "Classification of users' transportation modalities from mobiles in real operating conditions," *Multimed. Tools Appl.*, vol. 81, no. 1, pp. 115–140, 2022, doi: 10.1007/s11042-021-10993-y.

8. S. A. Haidery, H. Ullah, N. Ullah Khan, K. Fatima, S. Shahla Rizvi, and S. J. Kwon, "Role of big data in the development of smart city by analyzing the density of residents in shanghai," *Electron. Switz.*, vol. 9, no. 5, 2020, doi: 10.3390/electronics9050837.

9. C. Mizzi *et al.*, "Unraveling pedestrian mobility on a road network using ICTs data during great tourist events," *EPJ Data Sci.*, vol. 7, no. 1, 2018, doi: 10.1140/epjds/s13688-018-0168-2.

10. J. Jeong *et al.*, "Mobility Prediction for 5G Core Networks," *IEEE Commun. Stand. Mag.*, vol. 5, no. 1, pp. 56–61, 2021, doi: 10.1109/MCOMSTD.001.2000046.

11. P. Guo, K. Xiao, Z. Ye, and W. Zhu, "Route Optimization via Environment-Aware Deep Network and Reinforcement Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 6, 2021, doi: 10.1145/3461645.

12. M. Šauer, J. Vystoupil, M. Novotná, and K. Widawski, "Central European Tourist Flows: Intraregional Patterns and Their Implications," *Morav. Geogr. Rep.*, vol. 29, no. 4, pp. 278–291, 2021, doi: 10.2478/mgr-2021-0020.

13. X. Lao, X. Deng, H. Gu, J. Yang, H. Yu, and Z. Xu, "Comparing Intercity Mobility Patterns among Different Holidays in China: a Big Data Analysis," *Appl. Spat. Anal. Policy*, 2022, doi: 10.1007/s12061-021-09433-z.

14. X. Li *et al.*, "Position prediction system based on spatio-temporal regularity of object mobility," *Inf. Syst.*, vol. 75, pp. 43–55, 2018, doi: 10.1016/j.is.2018.02.004.

15. U. Türk, J. Östh, K. Kourtit, and P. Nijkamp, "The path of least resistance explaining tourist mobility patterns in destination areas using Airbnb data," *J. Transp. Geogr.*, vol. 94, 2021, doi: 10.1016/j.jtrangeo.2021.103130.

16. Data Science Process Alliance, https://www.datascience-pm.com/crisp-dm-2, last accessed 2022/07/24.

17. GSM Association Homepage, https://www.gsma.com/mobileeconomy/europe/, last accessed 2022/12/20

18. ANACOM Homepage, https://www.anacom.pt/streaming/ServicosMoveis1T22.pdf?contentId=1722575&field=ATTACHED_FILE, last accessed 2022/12/20