INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Movement and Concentration of People in Nightlife Areas: A Comprehensive Study of Urban Nightscapes

José Pedro Correia Seno

Master's degree in Integrated Business Intelligence Systems

Supervisor:

Doctor João Carlos Amaro Ferreira, Assistant Professor with Habilitation

ISCTE - Instituto Universitário de Lisboa

Co-Orientador(a):

Master Bruno Alexandre Mateus Francisco, Assistant Professor,

ISCTE - Instituto Universitário de Lisboa

September, 2024

TECNOLOGIAS
E ARQUITETURA

Movement and Concentration of People in Nightlife Areas: A Comprehensive Study of Urban Nightscapes

José Pedro Correia Seno

Master's degree in Integrated Business Intelligence Systems

Supervisor:

Doctor João Carlos Amaro Ferreira, Assistant Professor with Habilitation

ISCTE - Instituto Universitário de Lisboa

Co-Orientador(a):

Master Bruno Alexandre Mateus Francisco, Assistant Professor,

ISCTE - University Institute of Lisbon

September, 2024

Resumo

Esta dissertação analisa os padrões de mobilidade dos indivíduos nos bairros noturnos de Lisboa, utilizando dados anónimos recolhidos a partir de dispositivos móveis pessoais. Estes dados foram obtidos através de um acordo entre a operadora que captou os dados e a Câmara Municipal de Lisboa que propôs este desafio de criar informação pertinente para promover uma melhor segurança, movimentação oportunidades de negócio aos frequentadores das noites de Lisboa em zonas de diversão. Estas zonas são utilizadas por diferentes pessoas por diversas razões, sendo a sua gestão e utilização um ponto fulcral no desenvolvimento de uma cidade moderna e sustentável. Esta investigação centrou-se na análise dos dados providenciados e no estudo da possibilidade do uso de uma Rede Neuronal para prever o número de pessoas num local num tempo específico e, como tal, foi executada uma análise sobre o uso de um modelo para executar previsões de séries temporais sobre os dados obtidos pela operadora. O resultado foi positivo, nesta dissertação uma análise das movimentações em zonas de diversão noturnas de Lisboa onde se chegou a várias conclusões relativamente ao comportamento dos utilizadores destas zonas e a possibilidade de utilizar um modelo baseado em redes neuronais para prever o número de pessoas no tempo e no espaço é possível com uma mínima margem de erro.

Palavras-chave: Mobilidade; Dados; Lisboa; LSTM; Previsão.

Abstract

This dissertation analyzes the mobility patterns of individuals in Lisbon's nightlife districts, utilizing anonymous data collected from personal mobile devices. This data was obtained through an agreement between the operator that captured the data and the Lisbon City Council, which proposed this challenge of creating relevant information to promote better security and business opportunities for those who go to Lisbon's nightlife in entertainment areas. These areas are used by different people for different reasons, with their management and use being a focal point in the development of a modern and sustainable city. This investigation focused on analysing the data provided and studying the possibility of using a Neural Network to predict the number of people in a location at a specific time and, as such, an analysis was carried out on the use of a model to perform predictions of time series on the data obtained by the operator. The result was positive, in this dissertation an analysis of movements in nightlife areas in Lisbon, where several conclusions were reached regarding the behaviour of users of these areas, and the possibility of using a model based on neural networks to predict the number of people over time and space is achievable with a minimal margin of error.

Keywords: Mobility; Data; Lisbon; LSTM; Prediction.

Index

Figure Index

Table Index

CHAPTER 1

Introduction

Recent advances in information and communications technology (ICT) have enabled unprecedented insights into real-world urban dynamics by capturing and analysing vast amounts of data from everyday digital interactions. Through mobile data, IoT sensors, and AI-driven analytics, ICT now allows urban planners to monitor crowd flows, predict trends, and optimize resource allocation in real time. These tools paint a dynamic picture of urbanism, reflecting how people move, interact, and utilize public spaces, ultimately supporting smarter, more responsive city planning and sustainable urban development. Naturally, studies about the behaviour of societies through the means of data analysis are increasing and getting more important as we evolve. An effective method for enhancing public safety, transit, and urban planning in cities is mobile data analysis. local planners may optimize public transportation timetables, create safer pedestrian paths, and improve the accessibility of local services by examining location data from mobile devices. This information includes foot traffic, popular routes, and congestion locations. For example, tracking movement patterns can assist with real-time traffic signal timing adjustments, which can improve air quality and lessen traffic congestion. Furthermore, information about nocturnal activity may be used to better distribute law enforcement resources, improving public safety in high-traffic regions.

However, to safeguard individual privacy and foster public confidence, the use of mobile data for urban insights requires strict data protection procedures. To secure the users identity, anonymization, encryption, and adherence to data protection laws like the General Data Protection Regulation, (GDPR), are crucial. Clear usage restrictions and transparency in data collecting further guarantee the protection of residents' personal information, highlighting the idea that privacy is not sacrificed for urban development. According to the European Data Protection Board [1], data protection is recognized as a fundamental right under the EU Charter. This right not only upholds individual privacy but also influences other key rights, including freedom of expression, thought, and assembly. With growing public awareness of privacy and data protection, mishandling personal data can severely impact an organization’s reputation and erode the trust it has built over time. After ensuring that user data is safely protected, it can and should be used to help enhance modern cities to become safer, have better and faster transportation without the need of giving up on sustainability, and have economic growth by rightfully using the generated data to make informed decisions when making plans to improve the city.

Mobile phones have become essential to daily life, serving not only as communication tools but also as hubs for accessing information, managing personal and professional tasks, and interacting with

the broader digital world. Their ubiquity means that they capture an extraordinary amount of location-based data, providing a real-time picture of how people move, gather, and engage with different spaces in a city.

Human mobility is one of the most important ecological and social challenges of the 21st century [2]. Rapid urbanization and increased movement across cities place significant pressure on infrastructure, leading to greater energy demand, air pollution, and waste production. This mobility not only affects ecosystems but also impacts public health, as traffic congestion and pollution reduce urban air quality. On the social side, mobility shapes how communities form, interact, and evolve. Increased urban concentration affects housing, job markets, and social services, demanding innovative solutions to ensure inclusivity and access.

Human mobility prediction faces significant challenges due to the inherent unpredictability and complexity of human behaviour, along with various technical and ethical considerations. The great degree of variability in human mobility, which is impacted by a wide range of uncontrollable elements such as social events, environmental circumstances, and individual motives, being one of the main challenges. Public holidays, sporting events, concerts, and abrupt weather changes, for instance, can significantly alter mobility patterns, making it challenging for models to appropriately account for both habitual and impulsive actions. In this complex landscape, machine learning offers promising solutions. By leveraging mobile generated data, we can begin to tackle these challenges, as they provide a wealth of real-time, geolocated data, allowing for the tracking of individual movements across various environments. This data can help provide detailed movement profiles for different times of the day, identifying trends in how people move through the city. By aggregating this data, cities can gain insights into peak travel times, popular routes, and areas of congestion, which are crucial for developing more accurate predictive models. Furthermore, machine learning algorithms can be trained on historical mobility data to recognize patterns and anomalies. These algorithms can learn to adapt to new data inputs, allowing for continuous improvement in prediction accuracy over time.

1.1. Overview

This thesis comes from a challenge proposed by The Lisbon Urban Data Laboratory (LxDataLab) which is an initiative of the Lisbon Urban Management and Intelligence Center of Lisbon, which aims to use the data generated in the city of Lisbon to generate analytical solutions capable of solving real problems and improve the services provided to those who live, work and visit [3], like creating traffic indicators with the aim of contributing to decision-making regarding mobility and promoting more sustainable mobility, using trained neural networks to assess its ability to predict green roofs with orthophotos, understanding commuter movements in the city of Lisbon, as well as the influence of the

school calendar and he weather on these movements, and providing a fundamental overview of night-time movement and thus contribute to a more informed and impartial management of Lisbon's night-time entertainment areas, a topic that will be further explored in this thesis.

The increasing touristification of major urban districts is both a cause and a result of the expansion of the tourism-driven nighttime leisure economy in many European towns. This tendency, also referred to as the touristification of "the night," has had several detrimental effects on the environment, culture, economy, society, and space. The deterioration of the region's historical and architectural legacy is one of these effects as is the case of 'Cais do Sodré' for example. in a study published in 2023, Jordi Nofre, João Carlos Martins, Manuel García-Ruiz, and Guilherme Teixeira Costa conducted observational fieldwork between 2015 and 2019, highlighting the urgent need for new policies in economic, cultural, and urban governance to safeguard the cultural heritage of nightlife in urban settings [4]. Concerned remarks were made as well about cities like Amsterdam and Barcelona [5], Budapest [6] and London [7] to name a few. As such, this particular challenge aims to use the data of national mobile devices and roaming to study the movement and permanence of people between the nightlife areas of the city of Lisbon, as well as leveraging machine learning to find a way to predict the development of crowds during the usage of these areas, in attempts to respond to the interests of the various actors, namely urban planners, merchants, residents and users.

1.2. Motivation

The capital of Portugal is a very attractive point of interest for tourism and is getting increased popularity all over the world, in 2023 the World Travel Awards [8] gave Lisbon the award for 'Europe's Leading City Destination', so as residents and tourists alike wander the streets of Lisbon a trail of data is left behind to be collected and further analysed to track patterns in the behaviour of the crowds and gain insight into how the city of Lisbon works.

The possibility of sensing and predicting the movements of crowds in modern cities is of fundamental importance for improving urban planning, urban mobility, urban safety, and tourism activities [9].

Today, with the increase of data generated by people and the instruments we have at our disposal to build data driven tools we can analyse and predict the behaviour of large groups of people focusing on gathering information and powerful insights to help people make informed decisions to improve security, mobility, commercial success and infrastructure planning and sustainability for everyone involved in the nightlife experience.

The motivation for this thesis comes from the opportunity to improve the city by using the data gathered through a mobile operator, to analyse the movements of people in zones of nightlife activities

and forecast their subsequent steps, giving city-planners, stakeholders and even user of these zones tools to make informed decisions.

1.3. Objectives

The goal of this dissertation is to study the zones designated for nighttime activities in order for various stakeholders to use the information gathered to promote security, transportation, and business opportunities by gaining insights on night life patterns and a possible way to predict the density of people in a given area at a certain point in time.

This research proposes to:

1. Identify Nightlife Areas in Lisbon and its Active Periods.
2. Analyse the Current State of Nightlife and Movement Patterns.
3. Evaluate the viability of the Machine Learning to make time series predictions on the number of people in the streets of nightlife areas.

1.4. Methodology

In this study CRISP-DM [10] (Cross Industry Standard Process for Data Mining) was used to guide the analysis of the mobility of people in zones of nighttime activities. CRISP-FM is a framework designed to organize and elaborate a data mining project. This methodology is used in this dissertation for its strong guidance in data analysis, by having a set of phases that allow the researcher to follow a path that will lead him to a better understanding of the data at hand. As CRISP-DM is an adaptable framework, the data mining process of this study was divided in four stages: 1) Business understanding; 2) Data understanding; 3) Data preparation; 4) Visualization and decision-supporting dashboards; these stages were indispensable to respond to the objectives proposed in this thesis.

1.4. Dissertation Outline

This dissertation includes six chapters:

Chapter 1 refers to the introduction of the study.

Chapter 2 refers to the related work in this field where the literature review aims to summarize the state of the art in this point in time and contextualize this study.

Chapter 3 refers to the extraction of knowledge from the data provided, using the CRISP-DM methodology to describe and analyse the movement of people in nightlife areas.

Chapter 4 refers to the Data Visualization, in this chapter the aim is to illustrate the analyse made to the data and answer some important questions about the nightlife areas and their development through the night.

Chapter 5 refers to Time Series Prediction where the viability of the usage of Machine Learning to predict the density of people in a certain location during the nighttime period in these zones of Lisbon will be studied.

Chapter 6 refers to final conclusions and review of the study as well as future work that can be done in this field.

CHAPTER 2

Related Work

2.1 Literacy Review

In 2003, while the study of human crowd motion was still in its early stages, Roger L. Hughes [11] emphasized the importance of understanding the intersection between civil engineering and behavioural science, by comparing large crowds with fluids, as they present very similar characteristics, however, unlike a classical fluid, a crowd possesses the capacity for thought. There is a persistent myth that crowds are inherently irrational and erratic, suggesting that their behaviour is unpredictable, Hughes remarked that the behavioural science could be described as "fluid mechanics with an unclassical fluid, a thinking fluid that we are only just beginning to understand".

In 2009 R. Giffinger, and H. Gudrun [12] by recurring to literature and round-table-discussions found that, economy, people, governance, mobility, environment and quality of living are six characteristics that define a 'smart city' and that 'smart' implies the willingness to always be improving in these six fields in order to enhance its performance in terms of urban development.

As this thesis focuses on the movement and concentration of people there is a need to understand mobility, which is "the ability to move or be moved freely and easily" according to the Oxford Dictionary of English [13]. G. Lyons [11] named this era we live in as the 'era of smart' referring to the use of 'smart' as a sort of prefix in everyday objects we normally use such as 'Smart TV', 'Smart Phone', 'Smart Watch', 'Smart City' and so on in his article "Getting smart about urban mobility"- Aligning the paradigms of mart and sustainable" published in 2018, Lyons's concerns about labelling these terms with "smart" to enable technology development for the sake of profits, without having consideration for the sustainability issues that may be provoked by it led him to think about the definition of 'smart urban mobility' and what it really should mean, concluding that sustainability and affordability should not be sacrificed in the pursuit of smarter urban mobility.

Through the years studies using data from cellular networks to find useful information about social dynamics have been made. These studies, obviously, come from the mobile phone becoming an everyday item used by almost everyone and closely carried through the day. Therefore, we can get a complete view of the positions of people considering the location of the fixed antennas each device is connected to, in 2018 C. Balzotti, A. Bragagnini, M. Briani, and E. Cristiani [14], tried to understand how the mass of people distribute on large areas by using data extracted from coarse-grained aggregated cellular network data without tracking mobile devices individually. This data consisted of the density profiles of people in a given area at various instants of times making it possible for the authors to

extract the main directions followed by people, understanding how the mass of people distributes through space and time.

Personal mobile phones can act as "human probes"—individual sensors that track and document behavioural data—because of their widespread distribution and sensing capabilities. When combined, this data can provide insights into how human behaviour affects how cities function and how different urban elements interact .In 2015 Santi Phithakkitnukoon, Teerayut Horanont, Apichon Witayangkurn, Raktida Siri, Yoshihide Sekimoto, Ryosuke Shibasaki [2], employed a large-scale mobile sensing technique to examine the behaviour of tourists. They were able to analyse the behaviour of tourists and show applications that can be helpful for urban planners, transportation management, and tourism authorities by using GPS position traces of mobile phone users in Japan that were gathered over a year.

Predicting the movement patterns and population density of particular areas over time is known as crowd flow forecasting. It's a crucial tool for resource allocation, public safety improvement, and urban space management, especially in places like transit stations, event spaces, and tourist destinations. Authorities can manage pedestrian flow, anticipate crowd congestion, and stop overcrowding-related incidents with the aid of accurate forecasting. Machine learning (ML) and data analytics developments have greatly enhanced crowd flow forecasting. For example in 2018 Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, Xiuwen Yi, and Tianrui Li [15] proposed a deep learning-based approach to collectively forecast crowd flows by designing an ST-ResNet model designed with an end-to-end architecture tailored to the unique characteristics of spatio-temporal data. To capture the distinct properties of crowd traffic, the model included multiple branches of residual convolutional units, each dedicated to modelling specific spatial aspects of traffic flow. By dynamically aggregating the output from three separate residual networks, ST-ResNet assigns varying weights to different branches and regions, allowing for a more flexible representation of crowd movement. Additionally, the model incorporates external factors, such as weather conditions and the day of the week, to enhance the accuracy of crowd flow predictions for each region.

The literature in the field of crowd has been significantly shaped by methods that leverage deep learning approaches, with deep neural networks demonstrating outstanding performance across various forecasting scenarios, including mobility and telecommunications. These advanced models are particularly effective at capturing complex patterns and dependencies in large datasets, making them well-suited for predicting crowd movement and density in dynamic urban environments.. In 2019 Alessandro Crivellari and Euro Beinat [16] aimed to tackle a particular case of the “trajectory prediction problem” by analysing large-scale mobility traces of short-term foreign tourists, these researchers used an LSTM neural network model trained on pre-processed location sequences to capture movement patterns and predict subsequent destinations in a tourist's trajectory. Their findings suggest

that recurrent neural network architectures, such as LSTMs, hold significant promise in the field of human mobility. Given the success of current applications, further exploration of these models could lead to advancements in trajectory analysis and motion behaviour studies, potentially supporting a wide range of urban planning and tourism management tasks.

In 2020, Utkarsh Singh, Jean-François Determe, François Horlin, and Philippe De Doncker [17] applied long short-term memory (LSTM) neural networks to forecast crowd densities based on data from Wi-Fi sensors at a large organized event in Brussels. Their goal was to provide event organizers with timely insights into potentially critical crowd densities. The study found that forecast accuracy varied across different areas, likely due to the varying predictability of each area's time series data. Overall, their proposed crowd forecasting system demonstrated suitability for monitoring large-scale public events.

Time series prediction uses historical data to forecast future values in a sequential dataset. This type of prediction is widely used in fields like finance, weather forecasting, and urban planning, where understanding future trends is essential for decision-making. Time series data is often characterized by temporal dependencies, seasonal variations, and occasional anomalies, making it a complex yet valuable study area.

CHAPTER 3

Application of CRISP-DM to the Current Problem

3.1 Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives [19].

The primary objectives of this project are as follows:

1. **Identify Nightlife Areas in Lisbon and Active Periods:** Determine the specific areas in Lisbon that are popular for nightlife and identify the periods during which these areas are most active.
2. **Analyse the Current State of Nightlife and Movement Patterns:** Examine the current dynamics of nightlife in Lisbon and analyse how people move and distribute themselves throughout the night in these areas.
3. **Evaluate the Viability of LSTM for Predictive Modelling:** Utilize a Long Short-Term Memory (LSTM) network, a type of Recurrent Neural Network (RNN), to build a model capable of predicting the movement of people over time and space within these nightlife areas. analyse the viability of the LSTM model for this specific predictive task based on the available data, to accomplish these objectives, the project will leverage Python and its extensive libraries for data analysis, followed by the application of LSTM networks for predictive modelling.

The process involves several key steps:

1. **Data Collection and Preparation:** Gather and preprocess data relevant to nightlife activities in Lisbon. Ensure the data is clean, consistent, and suitable for analysis and modelling.
2. **Exploratory Data Analysis:** Perform detailed exploratory analysis to understand the underlying patterns and trends in the data and visualize the data to gain insights into the distribution and movement of people in nightlife areas.
3. **Model Development and Evaluation:** Develop a LSTM model to forecast future movement patterns and evaluate the model's performance to determine its viability for the predictive task at hand.
4. **Analysis of LSTM Model Viability:** Conduct a thorough analysis to assess whether the LSTM model is a suitable and effective tool for predicting movement in nightlife areas.

Lastly, compare the LSTM model's predictions against actual data to evaluate accuracy and reliability.

By systematically addressing these objectives, this thesis aims to deliver a comprehensive analysis of nightlife dynamics in Lisbon and assess the effectiveness of LSTM networks for predictive modelling in this context.

3.2 Data Understanding

The data understanding phase begins with the initial collection of data and involves activities aimed at gaining familiarity with the dataset. This stage includes identifying any issues related to data quality, uncovering initial insights, and identifying subsets of data that may reveal patterns or relationships. These preliminary findings allow researchers to form hypotheses about potential hidden information and guide further analysis.

There is a close link between Business Understanding and Data Understanding, the formulation of the data mining problem and the project plan require at least some understanding of the available data [19].

The data at hand was provided by the Lisbon City Council (Câmara Municipal de Lisboa) and collected by the mobile operator Vodafone. For obvious privacy reasons, the data collected does not track the movement of individual devices, instead, it describes a specific point in time in a specific location of Lisbon, collecting the number of devices in said location and some other relevant information, that can then be explored to gather insightful knowledge.

The city of Lisbon was divided in squares of 200x200 meters and information was aggregated in periods of time of 15 minutes. This paper is focused on the year of 2023, so the data collected spans from January 2023 through December 2023

The data provided is divided in two datasets, one of them contains information about the number of mobile devices that get in, stay and get out per square in the period of 5 minutes and the second dataset describes each square of the city of Lisbon, such as the perish in which the square is, coordinates of the location and an id to label the square and later associate it with the previous dataset.

Table 1 - Mobile operator variables list dataset

ID	Variable Name	Variable Description Variable	Type
1	Grid_ID	Grid number. Lisbon's metropolitan region is divided into 3743 squares of 200 by 200 meters.	Nominal
2	Datetime	Time and date of occurrence	Datetime
3	C1	No. of distinct terminals counted on each grid cell during the 5-minute period - Measured every 5 min.	Metric
4	C2	No. of distinct terminals in roaming counted on each grid cell during the 5-minute period– Measured every 5 min.	Metric
5	C3	No. of distinct terminals that remained in the grid cell counted at the end of each 5-minute period.	Metric
6	C4	No. of distinct terminals in roaming that remained in the grid cell counted at the end of each 5-minute period.	Metric
7	C5	No. of distinct terminals entering the grid	Metric
8	C6	No. of distinct terminals leaving the grid. The calculation is made using	Metric
9	C7	No. of entries of distinct terminals, in roaming, in the grid.	Metric
10	C8	No. of exits of distinct terminals, in roaming, in the grid.	Metric
11	C9	No. of distinct terminals with active data connection in the grid cell – Measured every 5 min.	Metric
12	C10	No. of distinct terminals, in roaming, with active data connection in the grid cell – Measured every 5 min.	Metric
13	C11	No. of voices calls originating from the grid cell.	Metric
14	C12	No. of entries into Lisbon along the 11 main roads.	Metric
15	C13	No. of exits from Lisbon along the 11 main roads.	Metric
16	D1	Top 10 origin Countries of the devices in roaming.	Metric
17	E1	No. of voice calls that ended in the grid within the 5-minutes.	Metric
18	E2	Average download speed per grid within the 5-minutes.	Metric
19	E3	Average load speed per grid within the 5-minutes.	Metric
20	E4	Peak download speed on the grid within the 5-minutes.	Metric
21	E5	Peak upload speed on the grid within the 5-minutes.	Metric
22	E6	Top 10 apps used on the grid within the 5-minutes.	Metric
23	E7	Lowest permanence period on the grid within the 5-minutes.	Metric
24	E8	Average permanence on the grid within the 5-minutes.	Metric
25	E9	Maximum permanence period on the grid within the 5-minutes.	Metric
26	E10	Count of devices sharing the internet connection in the grid within the 5-minutes.	Metric

Table 2- Geographical variables datase

ID	Variable Name	Variable Description Variable	Type
1	grelha_id	Grid number. Lisbon's metropolitan region is divided into 3743 squares of 200 by 200 meters.	Nominal
2	dicofre	Identification of the district. Assigned by administrative entities.	Nominal
3	entity_id	Identification of the data source where the information was generated.	Nominal
4	entity_type	Identification of the data source where the information was generated.	Nominal
5	freguesia	District to which the largest area where the grid belongs.	Nominal
6	freguesias	Districts in which the grid is inserted.	Nominal
7	latitude	Centroid Latitude.	Metric
8	longitude	Centroid Longitude.	Metric
9	objectid	Id of the object in the database.	Nominal
10	position	Grid in geometry format.	Metric
11	wkt	Grid in WKT format.	Metric

3.3 Data Preparation

The data preparation phase encompasses all steps needed to build the final dataset from raw data, ready for input into modelling tools. This phase often requires iterative processes, with tasks executed in various sequences. Key tasks include selecting relevant tables, records, and attributes, cleaning data to address quality issues, creating new attributes, and transforming data into formats suitable for modelling.

The accuracy and efficacy of any data analysis depend significantly on the quality of the data preparation process. In this chapter, the steps taken to preprocess and prepare the dataset used in this study are detailed. This involves cleaning the data for potential null values and outliers like festive days or abnormal day that may affect the study by tending the results to numbers that may misrepresent the actual average behaviour of the trend, filtering the data to leave out unimportant variables that have no correlation with the target variable and data that is out of the scope of this study to have a leaner and easy to process dataset , transforming the data by creating new variables that will help gain more insight on the topic at hand, and structuring the data to ensure its suitability for analysis and modelling.

3.3.1 Data Selection

The initial phase of data preparation involved the meticulous selection of relevant data points from a vast pool of raw data. This step was crucial to ensure that the dataset accurately reflected the pedestrian movement patterns within the targeted nightlife zones of Lisbon. The following criteria and processes were employed to achieve this:

The primary dataset provided by the Mobile Operator that details the movement of various pedestrians through the city of Lisbon had a lot of information that was of no use for this case study. In order to have a clean and stable dataset from which the conduction of this study was viable a number of steps were taken.

Only the columns 'Grid_ID', 'Datetime', 'C3', and 'C4' were retained from the raw CSV files to focus on the essential data required for analysis.

The 'Datetime' column was converted to a datetime format to facilitate precise time-based filtering and analysis. Data was then filtered to include only those records that fell within the hours typically associated with nightlife activities, from 18:30 to 07:00 on the following day, this time range captures most of the nighttime activity and early morning movements, excluding data points that fall outside these hours.

The filtered data from individual CSV files was then concatenated into a single DataFrame. This consolidation process involved, given the large size of the dataset, that each CSV file was read in chunks of 10,000 rows to optimize memory usage and processing time. Each filtered chunk was appended to a list, which was later concatenated into a comprehensive DataFrame.

To enhance the dataset with spatial context, the consolidated DataFrame was merged with the location data from the Excel file provided by the LxDataLab[3]. The 'grelha_id' column in the location data was renamed to 'Grid_ID' to match the primary dataset and facilitate merging. An inner join was performed on the 'Grid_ID' column, combining the pedestrian movement data with the corresponding geographical information.

The final processed dataset, enriched with location information and additional features, was saved to a new CSV file. This dataset serves as the foundation for subsequent analysis and modelling efforts, ensuring a clean, relevant, and comprehensive data source. Each record in the dataset includes information such as the timestamp, grid ID, coordinates (latitude and longitude), the number of people detected, and the corresponding administrative district (freguesia).

By adhering to these data selection criteria and processes, we ensured that the dataset accurately reflects the pedestrian movement within Lisbon's nightlife zones, providing a solid foundation for meaningful analysis and insights.

3.3.2 Data Cleaning and Filtering

After the selection of the data needed to perform relevant data extraction the dataset built still had various anomalies and irrelevant data points that needed to be addressed.

As this study focus on the movement of pedestrians throughout the nightlife of Lisbon a decision was made to focus on specific administrative districts, namely "Santa Maria Maior", "Misericórdia", "Santo António", "Arroios", and "Estrela", any records outside these districts were excluded. These

districts were selected by previous knowledge of the Lisbon nightlife scene as it is well known that the primary locations for these types of activities occur mainly in the district of “Misericórdia” with areas such as “Bairro Alto” and “Cais do Sodré”, and in the district of “Estrela” with “Santos” being a well-established hang out spot for a cheap beer and a starting point for the night. The other districts “Santa Maria Maior”, “Santo António” and “Arroios” were selected for the number of people registered through the hours established previously as nightlife hours, and as such deserve to be explored.

The timestamp column was converted to a datetime format to facilitate time-based filtering and analysis. Records were filtered to include only those that fell within the typical nightlife days and hours in Lisbon (Thursdays from 22:00 onwards, Fridays and Saturdays for the entire duration, Sundays until 07:00).

Specific dates (e.g., public holidays and outlier dates) were excluded to maintain consistency. Some timestamps were registered with an unusual number of people in a given place. In some of those dates there are simple explanations for the numbers reported, for example New Year’s Eve will have more people in the streets than usual, but other dates are outliers with no specific reason found that had to be cut off from the dataset.

3.3.3 Feature Engineering

To enhance the dataset and facilitate more granular analysis, features were engineered, such as an 'Hour' column, a day of the week column and a “Month” column that were derived from the timestamp to enable hourly, daily and monthly aggregation and analysis of pedestrian movement. The number of residents and tourists belonged to two different variables, ‘C3’ and ‘C4’ respectfully, as the total of people in the streets is the aim of this study both columns were added to create a new one that will be explored in the next chapter.

3.3.4 Aggregation and Summarization

To gain insights into pedestrian movement patterns, data was aggregated at various levels. Aggregation by district was imperative to compare the districts where the nightlife is more prominent amongst them.

Hourly, daily and monthly aggregation of the data is important to detail the full picture of the urban nightlife of Lisbon and how it evolves through time through various levels of granularity.

Time periods were created to categorize the nightlife into different parts that result in different behaviours of movements, by aggregating the data into these time zones it’s possible to compartmentalize patterns and trends and give them reason to be and analyse how they movement of people registered in each district differs between one another. This step was really important for

the next chapter as it involves the building of data visualization tools like graphs and tables, to help paint a picture of the movements of people in zones of nighttime activities in Lisbon.

CHAPTER 4

Data Visualization

Data visualization is a fundamental aspect of data analysis, transforming complex datasets into accessible visuals, such as charts, graphs, and maps, to reveal trends, outliers, and patterns. This practice is essential within the data science process, as it makes data easier to understand and actionable for a diverse audience, including those outside the immediate subject matter. By converting intricate data into visual formats, data visualization enhances accessibility and interpretability, helping analysts and stakeholders to quickly identify key patterns, trends, and anomalies. This is particularly vital for big data, where the vast quantity of information can be overwhelming; effective visualization techniques allow for streamlined insights, making data more manageable and valuable.

For this chapter, analysis tools that provide interactive visualizations and business intelligence capabilities, will be leveraged to achieve a comprehensive Data Analysis, using the data previously prepared to create visualizations to analyse trends and patterns.

4.1 District overview

The data provided gives geographic coordinates of the 3743 grids of 200 x 200 meters and the 24 districts they belong to, as such, a map overview of the city of Lisbon is easy to achieve in Figure 1 by using these attributes.

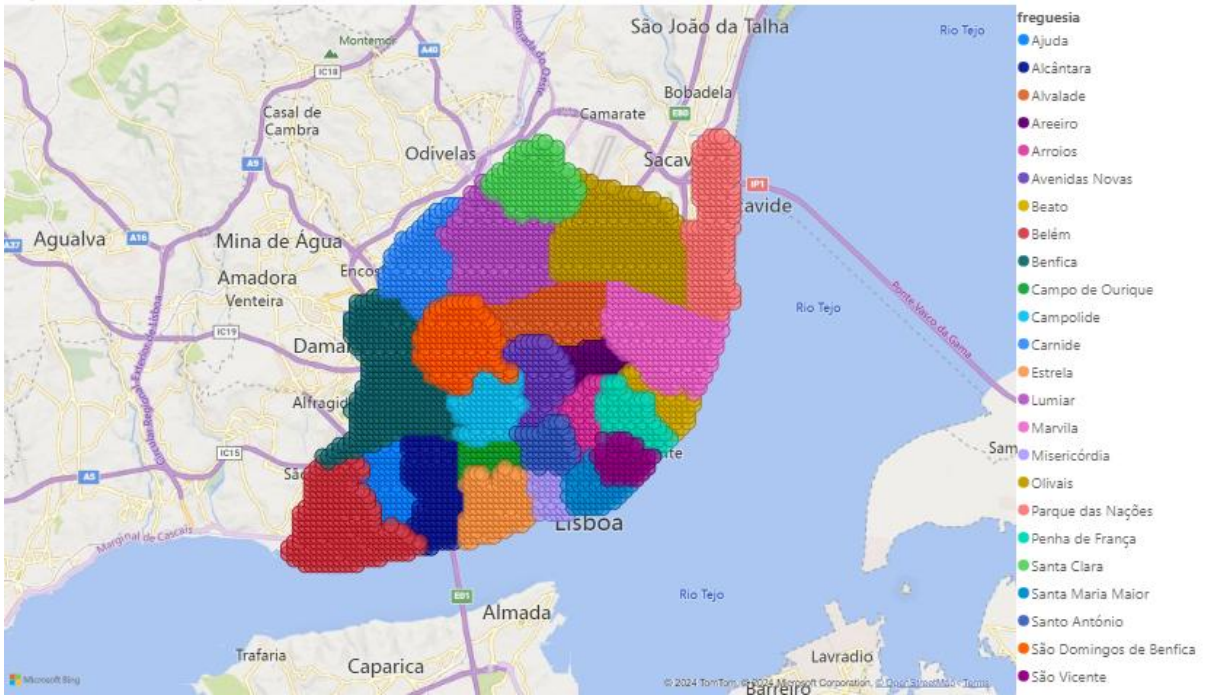

Figure 1 - Map of Lisbon's Districts

This study however is focused on the zones that have prominent nighttime movement as seen in Figure 2, most importantly the ones which contain zones that are designated for nighttime activities “Arroios” is not known for its night life, however, is where a lot of dislocated university students find affordable housing and in recent years it became home for several of Portugal’s migrant communities, explaining the high daily traffic of people in this district observed in Figure 3, “Estrela” is the less frequented district of this group but “Santos” as been for years a place where people gather for nighttime activities and “Madragoa” is gaining a new life with its gentrification, which for now is benefiting, the locals by attracting tourists and residents alike to local businesses, both “Santos” and “Madragoa” are historic neighbourhoods of “Estrela”, “Misericórdia” is with no surprise the most visited district for nighttime activities, by having several restaurants all over the district and the most popular and affordable bars in “Bairro Alto” and “Cais do Sodré”, “Santa Maria Maior” is the downtown of Lisbon it has various points of interest for tourism, various methods of transportation for people who are looking to commute from work to their homes overlapping with people that are starting to gather to begin nighttime activities in restaurants or bars in the area, explaining why it is the second most visited district of the list, lastly, “Santo António”, surrounded by the top 3 most visited districts, despite having some integral points of interest of Lisbon in “Parque Eduardo VII”, “Marquês de Pombal” and “Avenida da República” people are very likely to pass by it when coming or going to indulge in nighttime activities as well thus having such a high number of visitors during this time period.

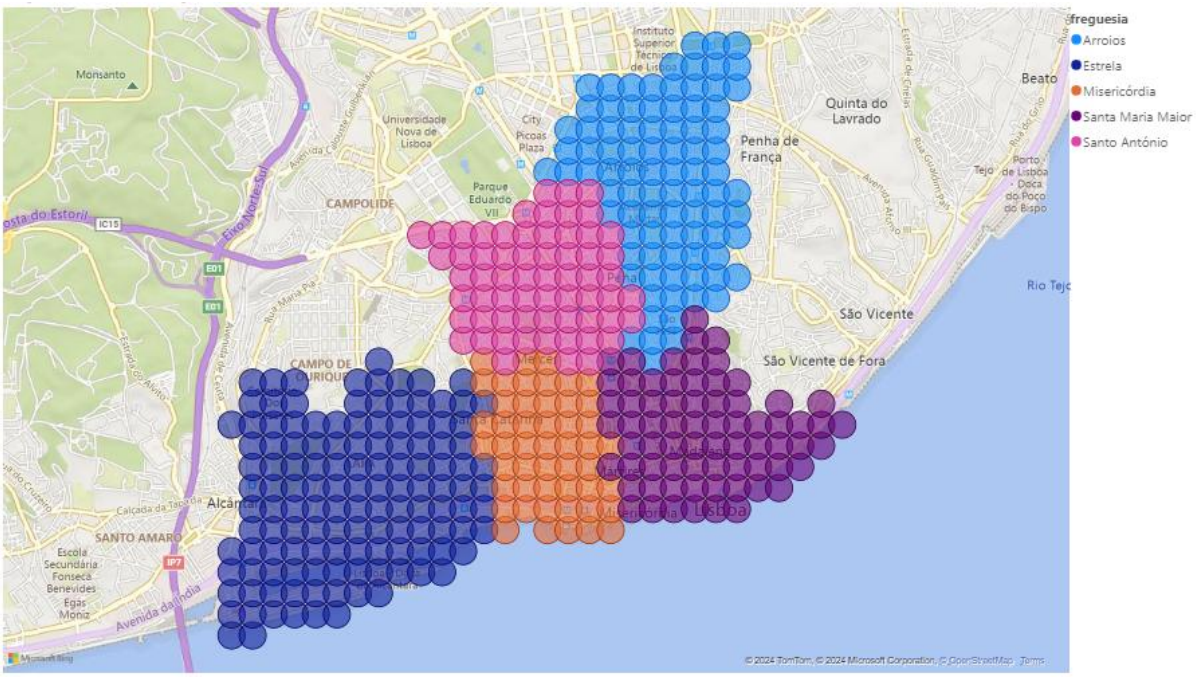

Figure 2 - Map of Lisbon's Nighttime Districts

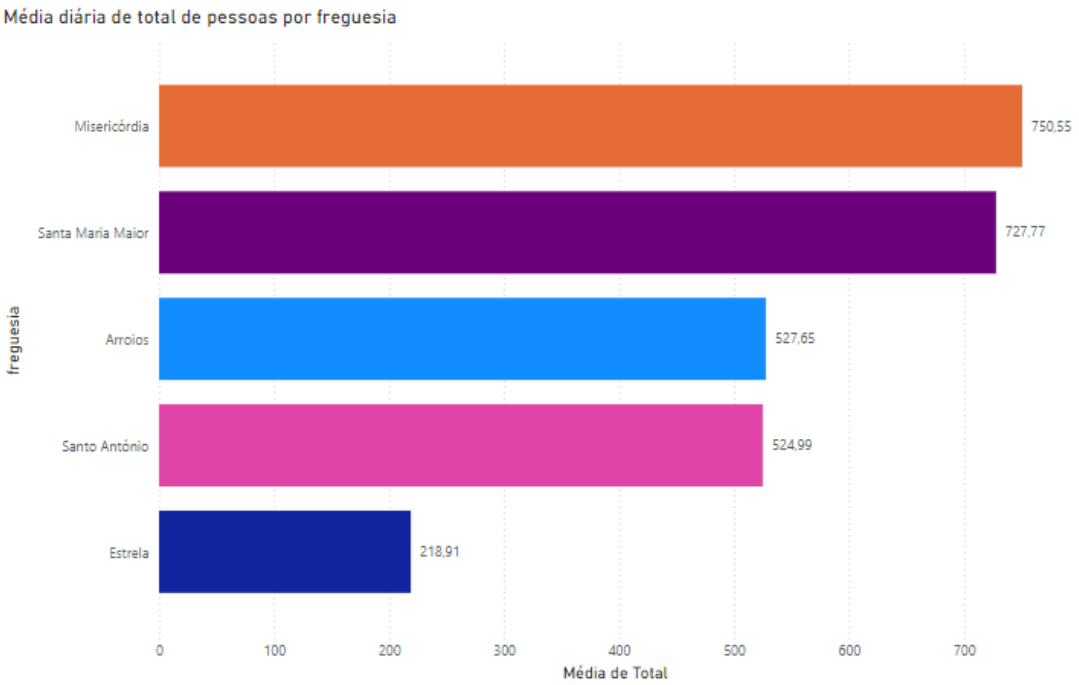

Figure 3 – Average of the total of people by district in a 15-minute span of time

4.2 Monthly Overview

In figure 4 it's possible to observe the evolution of the mean of people in a grid of each of the districts through out the year of 2023, the trend shows that the number of people in the streets is steadily increasing from the beginning of the year until May, with no high remarks and then in June in the beginning of the holyday and festival seasons as well as exam season for students the trend goes significantly down recording it's lowest point in July until the end of the year. In August the trend largely increases, and Lisbon's nightlife becomes full of life, subsequently having its best months in September and October before wrapping up the year in November and December with some impressive numbers still, showing that on average the second semester of the year has a more populated nightlife.

As for the districts it was registered that 'Misericórdia' and 'Santa Maria Maior' consistently registered higher average numbers of people by grid. These areas are evidently the most popular nightlife destinations and where it can be found more crowded places. Their prominence underscores their importance as central hubs for nightlife activities in Lisbon.

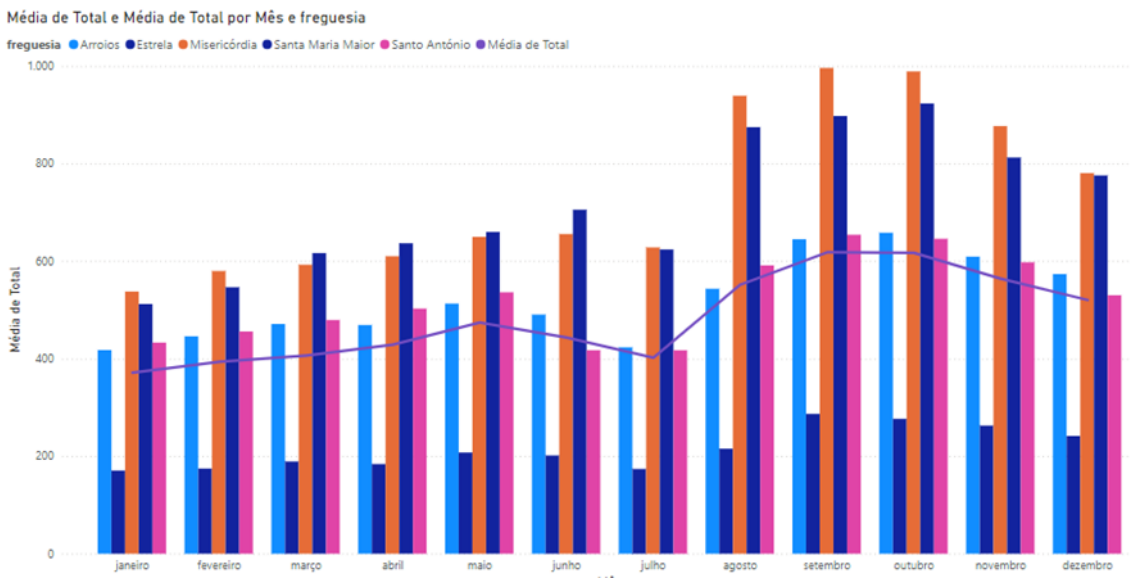

Figure 4 - Monthly Nightlife Activity by District

4.3 Daily and Hourly Overview

In Figure 5 it's feasible to see how an average weekend in the nighttime districts takes form. From Thursday to Saturday night the evenings assume the same shape more or less and the district of "Santa Maria Maior" stays busy every night, then during Fridays and Saturdays it's possible to see the district of "Misericórdia" getting a big wave of people in later hours, of the night, really proving how it really is the target place for nighttime activities.

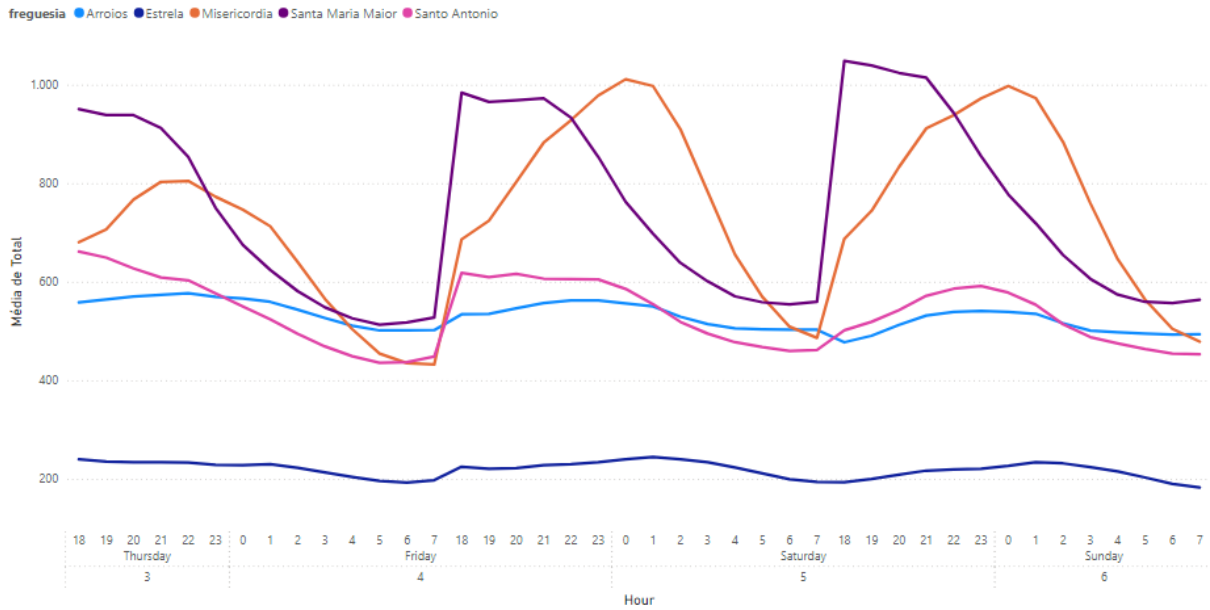

Figure 5 – Average evolution of the average number of people in a grid from a district over the weekend

In Figure 6 from an hourly perspective during a single day, on average, it was registered that overall, most people start nighttime activities in the district of 'Santa Maria Maior', where there is a variety of restaurants and bars that ease people into the night. Then, as the night sets in the district of 'Misericórdia' gets the most attention being the top spot for nighttime activities until 4 in the morning when most of the bars close and people star making their way home.

The districts of 'Arroios', 'Santo António' e 'Estrela' may not be the most frequented but maintain a stable number of people in the streets through the nighttime.

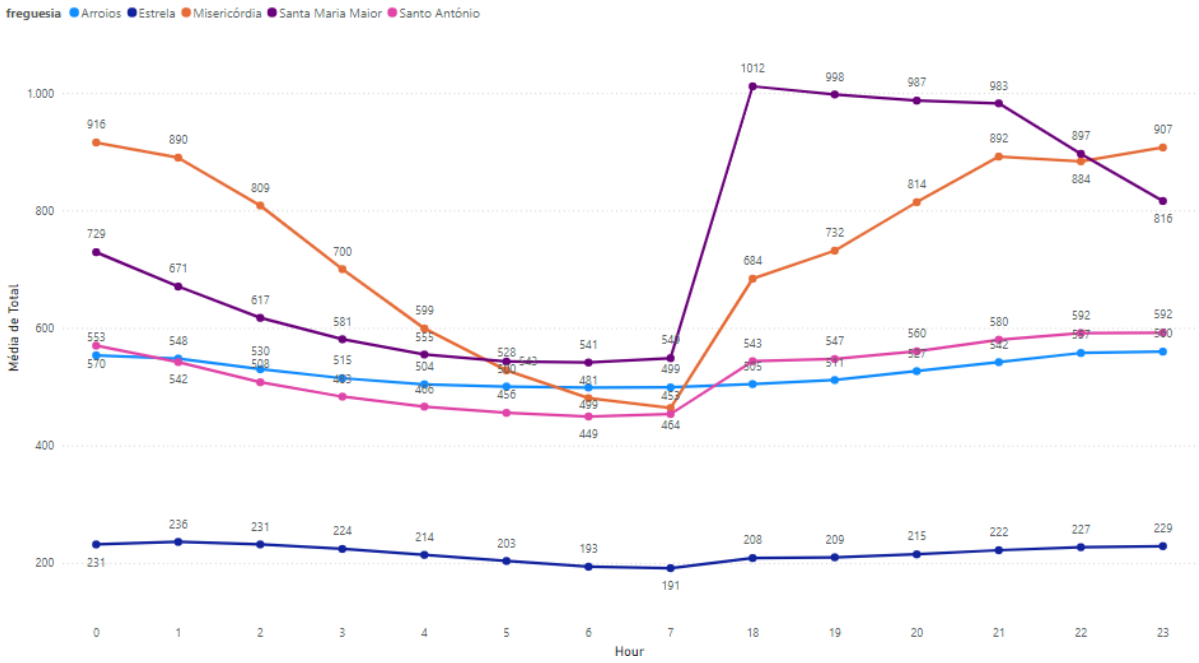

Figure 6 - Mean Distribution of people by district trough nighttime.

4.4 Time Periods Analysis – Heatmap Evolution

For the following section, the data was aggregated in three distinct Time Periods, to get a better understanding of the nighttime concentration of people in zones designated for nighttime activities.

The table 3 shows the designation of each period.

Table 3 – Time Periods

Time Period	**Hours (from – to)**
Dinner Hours	18:00 – 23:00
Busy Hours	23:00 – 4:00
After Party Hours	4:00 – 7:00

4.4.1 Dinner Hours

This time period refers the evening, more specifically to the hours between 18:00 and 23:00 the city experiences a marked increase in activity as people gather for late afternoon drinks and dinner.

Figure 7 highlights a significant concentration of people in the downtown area. The heatmap highlights particularly dense gathering in the districts of 'Santa Maria Maior' and 'Misericórdia', areas well known for their popular nightlife and dining establishments. While these districts exhibit the highest foot traffic, there is also noticeable movement in other surrounding areas. This likely represents people commuting home or exploring dining options in less central districts, which, despite being less prominent, still offer a variety of appealing options for dining experiences and social spaces.

Figure 7 – Density of people in nighttime districts during the dining period

Figure 8 zooms in on the zones with more concentration to explore the more crowded streets in this time period. Notable hotspots include "Praça do Rossio" and the streets leading to "Praça do Comércio", such as "Aurélia Street". Additionally, areas around "Bairro Alto" and "Praça Luís de Camões" are highlighted as well, known for their lively nightlife scenes and diverse food offerings. These locations attract a steady flow of people, further confirming their role as central hubs for dining and socializing during the evening hours.

Figure 8 - Zoom in grids with high density of people in the dinner period

Table 4 provides a breakdown of the movement of people across the nighttime districts during this time period. The overall traffic peaks between 20:00 and 21:00, reflecting typical dinner times when people gather at restaurants and bars across the city.

This table reinforces the trends seen in figure 7 and figure 8, where the heat maps showed similar concentrations of people in these districts. Together, these data points paint a comprehensive picture of the dynamics of Lisbon's nightlife, highlighting the prominence of certain districts over others in terms of foot traffic during the peak dinner and evening hours.

Table 4– Hourly evolution of the average number of people in a grid by district in the dinner hours

Hour	**Arroios**	**Estrela**	**Misericordia**	**Santa Maria Maior**	**Santo Antonio**	**Total**
18	521,96	218,43	684,5	995,14	591,86	602,378
19	528,85	217,87	725,14	981,62	591,01	608,898
20	542,33	221,04	801,41	977,62	594,72	627,424
21	553,86	225,86	866,02	966,64	595,5	641,576
22	558,82	227,25	891,2	910,25	598,06	637,116
Total	541,164	222,09	793,654	966,254	594,23	623,4784

4.4.2 Busy Hours

This time period refers to the hours between 23:00 and 3:00, the hours when most people start to go to the streets, bars and clubs to enjoy a night out with friends.

The concentration of people does remain in the downtown area, however, is not as dense as in the previous time period and there's a bigger concentration of people in the district of "Misericórdia" as shown in Figure 9. This figure clearly shows the movement of the crowd from one period to another when people transition to more party focus locations in Lisbon. The density of people around the downtown area has a clear decrease excluding some spots here in there where it's possible to see some activities taking place in other districts.

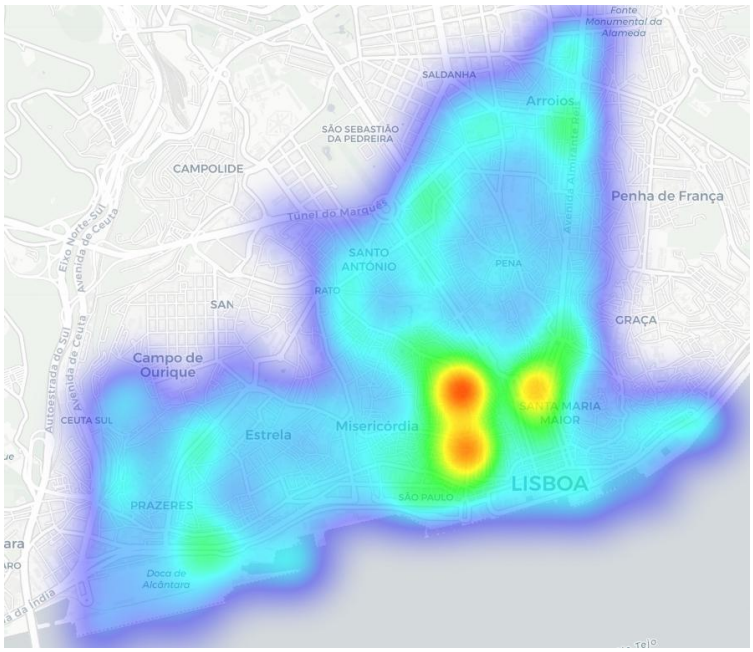

Figure 9 - Heatmap of Party time

Figure 10 zooms in once again in this denser mass of people and shows that in this period of time the main spot where people gather to have fun during the nighttime is in "Bairro Alto" separating itself from the rest of the zones designated for nighttime activities. With less density but still with above average numbers there's "Martim Moniz" and "Cais do Sodré", the first on being a hub for migrant communities to gather and the ladder being a famous place for nighttime activities but not having the sufficient average number of people to leave a significant mark in this heatmap.

Figure 10 - Zoom in grids with high density of people in the party period.

Table 5 clearly remarks “Misericórdia” as the most crowded district in these hours, having a significantly higher average of people passing by its streets as all of the other districts start to decline in numbers after 23:00 and “Misericórda” still has its peak at midnight. Overall the number of people does start to decline as people start to leave these areas to return home.

Table 5– Hourly evolution of the average number of people in a grid by district in the Busy Hours

Hour	**Arroios**	**Estrela**	**Misericordia**	**Santa Maria Maior**	**Santo Antonio**	**Total**
23	557,44	227,67	908,6	820,01	590,9	620,924
0	553,51	231,18	918,26	737,9	571,06	602,382
1	548,22	235,65	893,04	679,36	543,78	580,01
2	529,51	231,19	810,75	624,29	508,97	540,942
Total	547,17	231,4225	882,6625	715,39	553,6775	586,0645

4.4.3 After Party Hours

This period of time describes the hours between 3:00 and 7:00, when almost all the bars in “Bairro Alto” close and some people try and find other options in clubs thar are open until the morning or try to find a way to get home.

As Figure 11 shows, the density of people is now gone, the people who hadn’t yet left the zones of nighttime activities, start to go away from the previous hotspot, “Bairro Alto” and start to disperse, leaving trails of movement around “Misericórdia”, the main hotspot now is in the district of “Santa Maria Maior”, that sees significantly more movement comparing to the other districts.

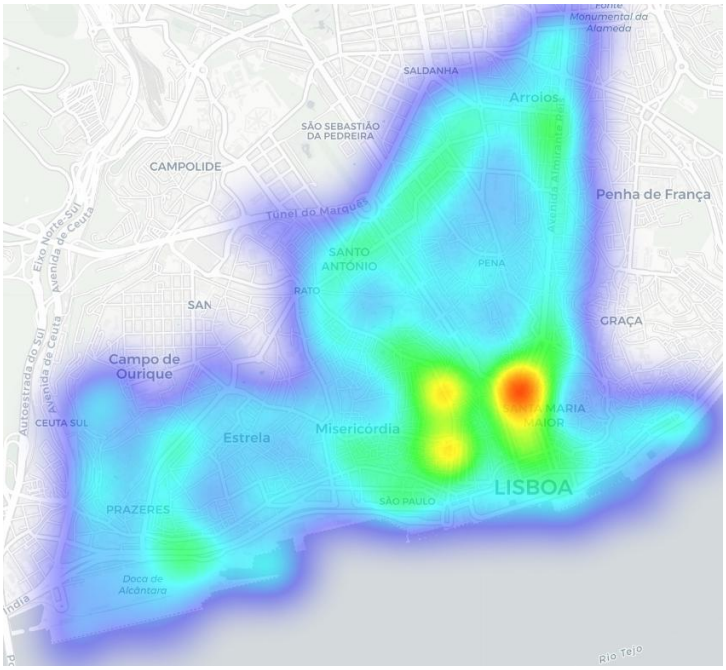

Figure 11 - Heatmap of After Party Period

Having a last look in the grids that describe the most density of people still in the streets, Figure 12 shows that the “Bairro Alto” streets are emptying and Santa Maria Maior is becoming the district with more movement again, mainly in the streets that lead to Areeiro where some cheaper hospitality options are located, and a lot of businesses start to open to begin the day.

Despite having a much bigger average than any other district in the “beginning of the end” of this las period in nighttime activities, “Misericórdia” starts to drop significantly in it’s average, leaving “Santa Maria Maior” to regain the top spot in terms of average of people per cell of district as previously stated and as shown in table 6, that helps paint the full picture of this time period.

Overall, after a significant drop in number of people from 3:00 to 4.00, derived from the drop in number of people in “Misericórdia”, in comparison to the other passages of hours, the average stagnates as a new day begins.

Figure 12 - Zoom in grids with high density of people in the after-party period.

Table 6– Hourly evolution of the average number of people in a grid by district in the After Party Hours

Hour	**Arroios**	**Estrela**	**Misericordia**	**Santa Maria Maior**	**Santo Antonio**	**Total**
3	513,9	223,34	701,19	584,41	483,56	501,28
4	504,6	213,72	600,03	556	466,79	468,228
5	499,93	202,99	527,75	542,76	455,15	445,716
6	499,16	193,65	481,52	542,29	449,82	433,288
7	499,34	191,31	464,95	549,64	453,98	431,844
Total	503,386	205,002	555,088	555,02	461,86	456,0712

CHAPTER 5

Time Series Prediction

In the realm of predictive modelling, time series analysis has emerged as a critical tool for understanding and forecasting temporal patterns across various domains. This chapter delves into the application of Long Short-Term Memory (LSTM) networks, a specialized type of recurrent neural network (RNN), to predict the number of people at specific locations over time. The ability to accurately forecast crowd dynamics is not only valuable for optimizing resource allocation and infrastructure planning but also for enhancing public safety and service delivery in real-time. Traditional time series models often struggle with complex, non-linear dependencies present in sequential data. However, LSTMs, with their unique capacity to capture long-term dependencies, offer a promising alternative.

This chapter explores the viability of using LSTM networks in the context of predicting human presence in a given location, drawing on a dataset of temporal patterns to evaluate model performance. By investigating the efficacy of LSTM models in capturing the intricate temporal relationships inherent in crowd behaviour, this research aims to contribute to the growing body of knowledge on advanced time series forecasting methods and their practical applications in urban planning and management.

5.1 Long Short-Term Memory (LSTM)

LSTM was first proposed in 1997 (Sepp Hochreiter, Jürgen Schmidhuber) motivated by an analysis of error flow in existing RNNs (Hochreiter et al., 2001), which found that long time lags were inaccessible to existing architectures [20].

The basic LSTM unit is composed of a cell, an input gate and an output gate, and a forget gate later implemented by Felix A. Gers, Jürgen Schmidhuber and Fred Cummins in 2000 [22], to make the LSTM cell learn to reset itself at appropriate times, thus releasing internal resources.

LSTM networks are structured around a chain of repeating modules of neural network layers, each consisting of a memory cell that maintains information over time. The key innovation of LSTMs is the ability to regulate the flow of information through these cells using the three main gates referenced before.

The forget gate (Figure 11 [23]) decides which information from the previous cell state should be discarded. It takes the input from the current time step x_t and the previous hidden state h_{t-1}, and passes them through a sigmoid function. The output is a number between 0 and 1, where 0 indicates "completely forget" and 1 means "completely retain".

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

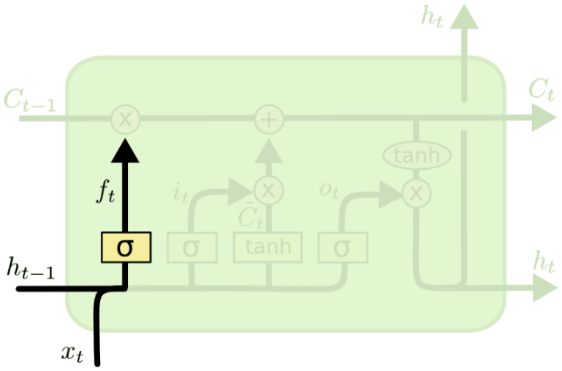

Figure 11 - Forget Gate diagram.

The input gate (Figure 12 [23]) determines which information will be updated in the cell state. It consists of two parts, the input modulation gate where it decides which values to update based on the input and previous hidden state, passed through a tanh function, and the Input gate where It determines how much of the modulated input should affect the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C$$

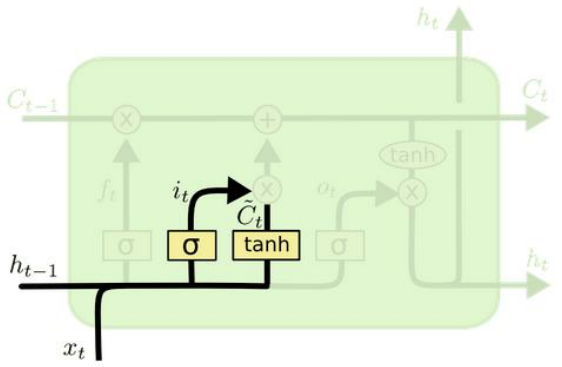

Figure 12 - Input Gate diagram.

The new cell state C_t (Figure 13 [23]) is then updated as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

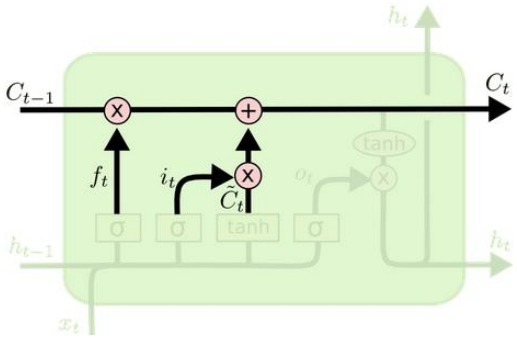

Figure 13 - New Cell State Diagram.

The output gate (Figure 14 [23]) controls the output at each time step, which is based on the updated cell state and the previous hidden state. The output of this gate is also passed through a sigmoid function.

$$O_t = \sigma(W_O \cdot [h_{t-1}, x_t] + b_O)$$

The hidden state h_t (Figure 14 [23]) for the next time step is then calculated by applying the tanh function to the updated cell state and multiplying by the output gate:

$$h_t = O_t * \tanh(C_t)$$

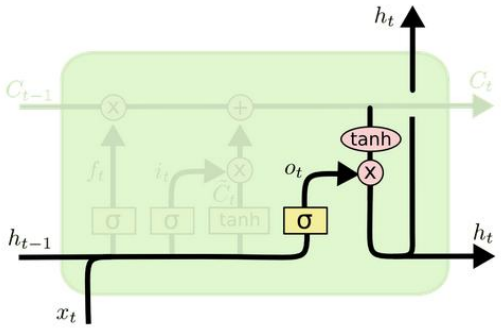

Figure 14 - Output Gate and Hidden State Diagram.

The combination of these gates allows LSTMs to retain valuable information for long periods, making them particularly effective for tasks like time series prediction, where understanding long-term dependencies is crucial.

5.2 Data Preparation for LSTM

For the predictive analysis using LSTM (Long Short-Term Memory) neural networks, additional preparation steps were taken in order to create a Time-Series Dataset from the dataset already cleaned and filtered used to previously describe and analyse the data.

The LSTM only considers the time variable in the dataset, so in order to predict the number of people in a given zone the data needs to be filtered by Grid_ID so that we can feed the model with the variation of people through the period of time that is being studied, that being the year of 2023. In order to accomplish this, the data was then filtered by the Grid_ID 625, an high density grid that covers some crowded streets in Bairro Alto, this grid will serve as a use case in this research to study the viability of the use of LSTM to predict the number of people in a given location.

The variables that describe the date and time and the total amount of people were isolated. The graph on Figure 15 shows the evolution of people in grid 625 through the night of January 6 to January 7.

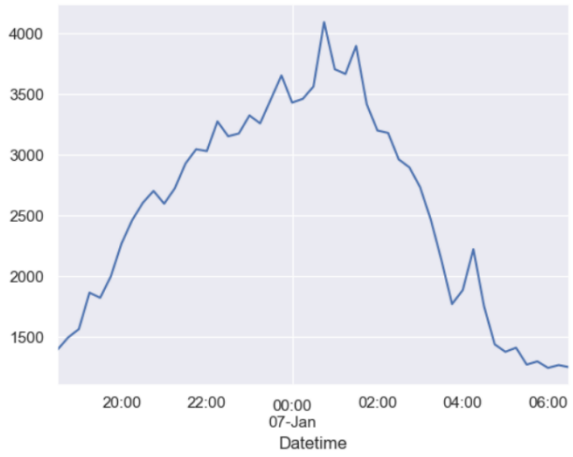

Figure 15 - Evolution of the number of people in grid 625 on the night of January 6

The dependent variable was then scaled to a specific range between 0 and 1, to normalize the data and guarantee performance quality. Then, in order to prepare the data for a time series model, sequences of data from the dataset are created and reshaped to be used in the LSTM model that will expect input data in 3D shape. Then the data is split in training and test sets while maintaining the sequence order. This is essential for time series forecasting models because the temporal order of data must be preserved to reflect the sequential nature of the data.

5.3 Modulation

Using Python's Keras library a LSTM model was built, summarized in Table 3, the model starts with two LSTM layers, the first layer processes sequences of length 10, the length of the sequences created before, and outputs sequences with 50 units. The second layer processes the output of the first layer and produces a fixed-size output. Two dropout layers are used for regularization to prevent overfitting, placed after each LSTM layer and two dense layers finish the model, the first one reduces the dimensionality from 50 to 25 and the second layer produces a single output.

Table 7- LSTM model summary

Layer	**Output shape**
LSTM	(None, 10, 50)
Dropout	(None, 10, 50)
LSTM	(None, 50)
Dropout	(None, 50)
Dene	(None, 25)
Dense	(None, 1)

When compiling the model, the configuration for training the model was set up. This configuration includes specifying the optimizer, the loss function, and optionally metrics to evaluate the model's performance. The optimizer controls how the model's weights are updated during training, Adam (Adaptive Moment Estimation) is a popular optimization algorithm that has adaptive learning rates, it adjusts the learning rates of individual weights based on estimates of first and second moments of gradients, is efficient and is suitable for most tasks, including sequence prediction. The loss function measures how well the model's predictions match the true values, during training the model tries to minimize this loss, MSE (Mean Squared Error) calculates the average of the squares of the differences between predicted values and actual values and it penalizes larger errors more heavily than smaller errors. Is commonly used for regression tasks where you want to predict continuous values. It is less suitable for classification tasks, where metrics like cross-entropy loss are more appropriate.

5.4 Evaluation

Now that the LSTM model has been developed in the preceding sections and tuned to capture the intricate patterns within the dataset, is now crucial to rigorously assess its performance.

The evaluation process is multifaceted, encompassing several key aspects to ensure that the model not only fits the training data well but also generalizes effectively to unseen data. It is important to thoroughly evaluate the model and review the steps executed to create it, to be certain the model

properly achieves the objectives proposed. At the end of this phase, a decision on the possibility of the usage of the model should be made.

5.4.1 Evaluation Metrics

Evaluation metrics are essential tools for assessing trained models, particularly in forecasting, as they measure how well a model fits the training data and indicate the expected accuracy of its predictions. By quantifying the performance, these metrics provide insights into both the strengths and weaknesses of a model, guiding improvements and helping to select the most reliable model for making future predictions.

With X_i being the predicted i^{th} value and Y_i the actual i^{th} value, the following expressions were used to evaluate the model:

- **RMSE –** this is one of the most common evaluation metrics when it comes to assessing a forecasting model's performance it is obtained by applying the square root to the Mean Square Error, measuring the average magnitude of the errors between predicted and actual values, giving more weight to larger errors due to the squaring process. RMSE is expressed in the same units as the target variable, making It easier to interpret, lower values indicate better model performance, with fewer smaller errors. With m being the number of data points, the expression is as follows:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}(\boldsymbol{Xi} - \boldsymbol{Yi})^{\mathbf{2}}}$$

- **MAPE –** Mean Absolute Percentage Error is another important metric used to evaluate the performance of regressing models, especially in time series forecasting. MAPE expresses the error as a percentage making it easier to interpret in relative terms across different scales, it gives the average absolute error as a percentage of the actual values. Lower MAPE values indicate better model performance. With m being the number of data points, the expression is as follows:

$$\boldsymbol{MAPE} = \frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\left|\frac{\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}}_{\boldsymbol{i}}}{\boldsymbol{Y_i}}\right| * \mathbf{100}$$

5.4.2 Model Evaluation

To fit the model the dataset, previously prepared, was divided in three sets, a training set with 70% of the data to train the model, a validation set composed of 15% of the data used to tune

hyperparameters and monitor the model's performance during training to avoid overfitting and a test set of 15% of the data as well for final unbiased evaluation of the model.

The model was trained for 40 epochs with a batch size of 32, meaning that during each epoch the data will be divided into batches of 32 samples.

The training took 84 seconds, during this training, both the training and validation losses decrease sharply during the first few epochs, indicating that the model is quickly learning and improving its performance on both sets. After about five epochs, both sets stabilize at a much lower value, what suggests that the model as converged and is making minimal improvements with each additional epoch. The validation loss fluctuates more than the training loss, especially between the middle epochs (around 20 – 32), this fluctuation is common and indicates that while the model generalizes well, there are slight variations due to the differences between the two sets. Importantly, the validation loss does not increase significantly, suggesting that the model is not overfitting, which is a good sign that the model generalizes well to unseen data.

Figure 16 illustrates the variation of the loss values for both training and validation sets during the 40 epochs.

Regarding the RMSE values in Figure 17, the graph shows that most predictions made by the model have relatively small errors, suggesting that the model is generally accurate with a mean value of RMSE of 336.199. While the model occasionally made larger errors, these instances are less common. The model's tendency to make relatively accurate predictions more often than not is a good sign, however, there's a presence of outliers with higher RMSE values.

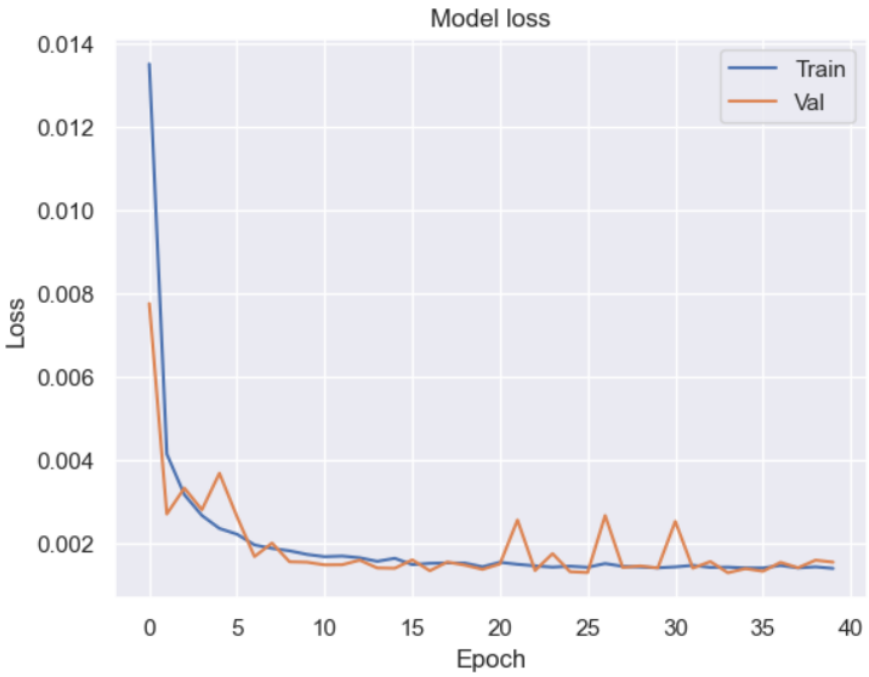

Figure 16 - Variation of loss values during the model training for training and validation sets

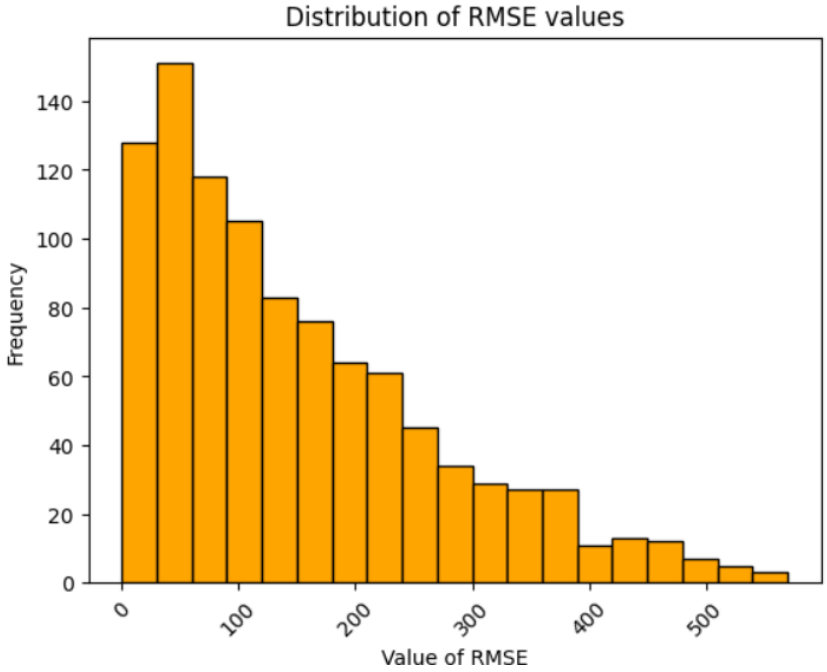

Figure 17 - Distribution of RMSE values

For the MAPE evaluation, Figure 18 shows the distribution of MAPE values illustrating that most of these values are clustered between 0% and 10%, with a sharp peak around 5%, which indicates a prominent level of accuracy with a mean value of 5.5%. The concentration of low MAPE values suggests that the model is highly effective at making accurate predictions for most cases, with most errors being small in proportion to the actual values and significant prediction errors are not common.

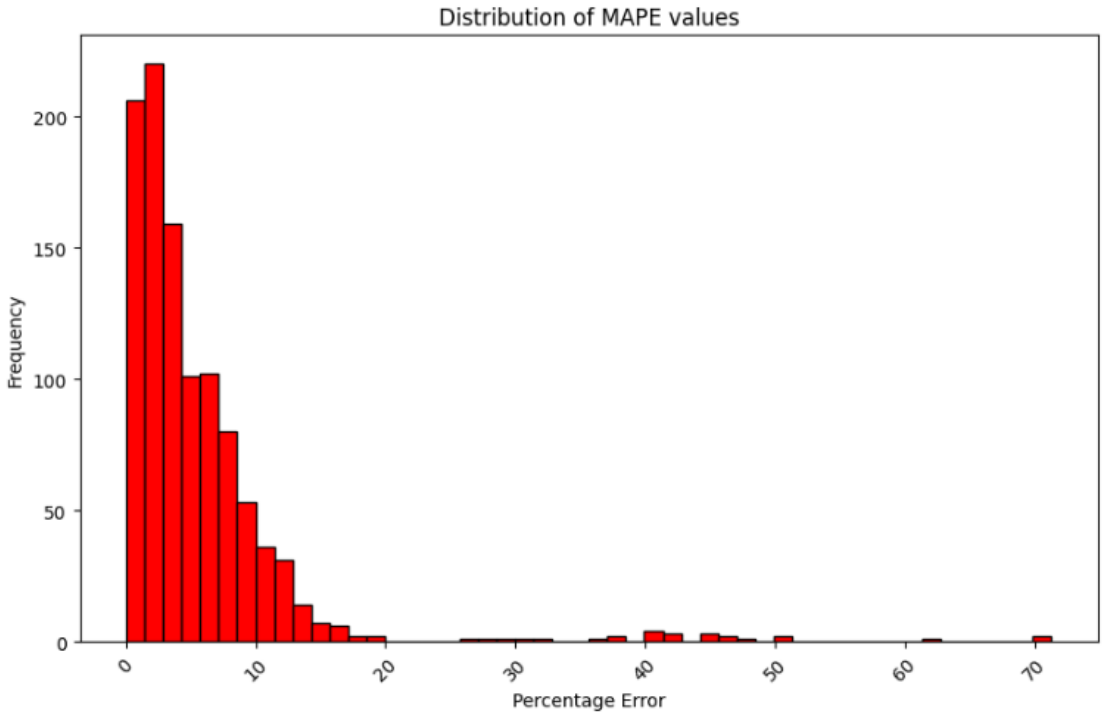

Figure 18 - Distribution of MAPE values

Figure 19 shows the predictions for November 11th of 2023 versus the actual number of people that were in Grid 625. The predictions are perfectly accurate, but they show a slight 'lag', meaning that the value of the steps is much closer to the real value of the time step that came before. This behaviour may occur in LSTMs, especially in time series with high auto correlation or when the model cannot capture well the data dynamics.

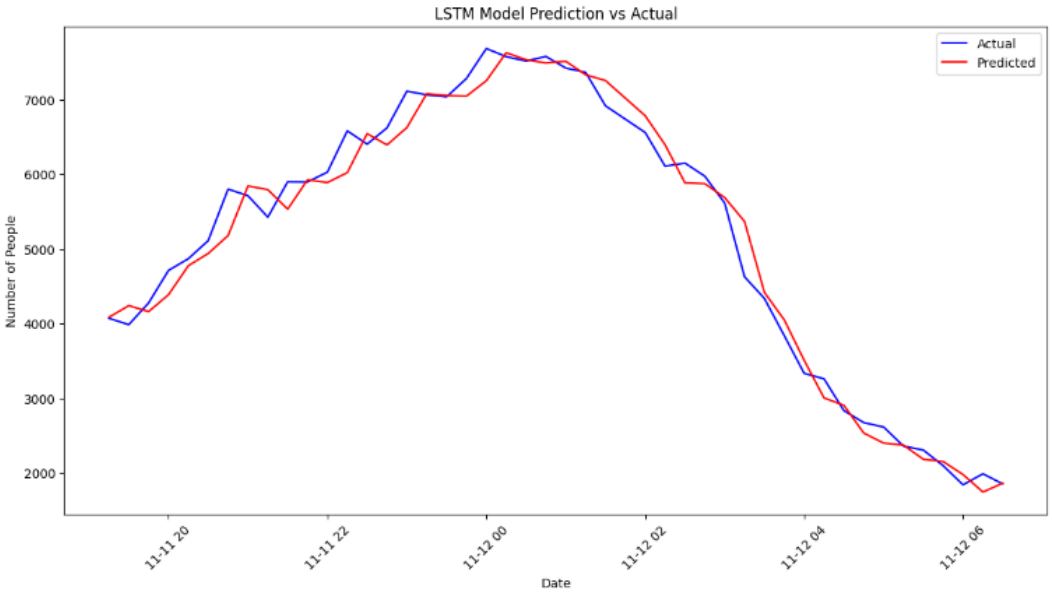

Figure 19 – Zoom in on model predictions on test dataset vs actual values.

However, in Figure 20, as we zoom out and get a full picture of the predictions made on the testing dataset, never seen before by the model, there are a lot of spikes in the data, each spike represents a night in grid 625. Recalling the pre-processing phase, the data was cut to show only the hours that were of interest for this project, leaving behind data that would help smooth out these spikes shown in the figure, as the number of people in the grid would gradually evolve without changing abruptly.

LSTMs rely on sequential patterns to make predictions. By cutting out the data, the model is being provided with a smaller and less smooth sequence to learn from. This makes it harder for the model to capture the underlying daily patterns and lead to sharper predicted spikes when it encounters sudden changes.

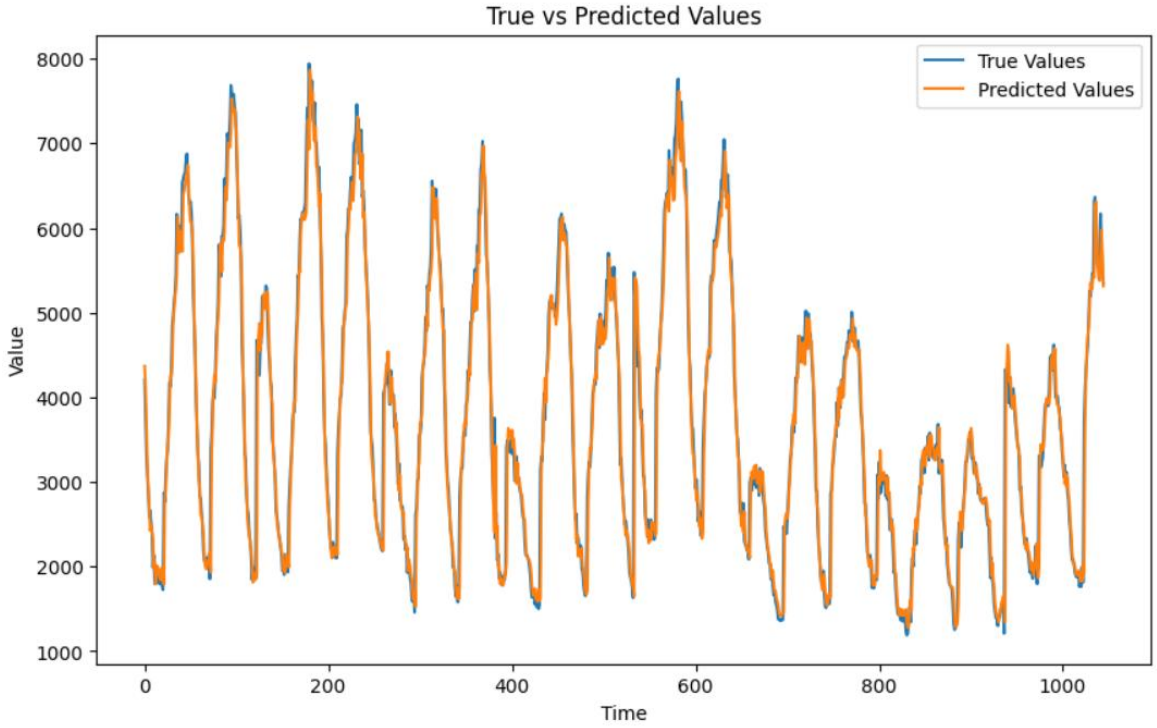

Figure 20– Complete model predictions on test dataset vs actual values

CHAPTER 6

Conclusions and Future Work

6.1 Conclusion

Understanding urban mobility is crucial for the efficient functioning and sustainability of smart cities. These smart cities are now more than ever producing significant data every day, which helps create new and innovative opportunities to explore how people behave and gather information to make sustained decisions on how to improve these cities for its population and visitors. For businesses, especially those in hospitality, retail and entertainment, urban mobility data is a key driver for success. By analysing movement patterns, business owners can maximize its exposure to people and use it to drive engagement and gather new customers. As for environmental sustainability, urban mobility can play a big part as well. Traffic congestion and inefficient transportation systems contribute significantly to air pollution in cities. By analysing how people move and promoting the use of public transport, cycling, or walking, cities can reduce their carbon footprint. This data-driven approach allows for more effective planning of green transportation initiatives, leading to lower emissions and a cleaner, healthier urban environment. In this way, understanding urban mobility not only supports economic growth and social well-being but also contributes to achieving sustainability goals in cities worldwide.

This dissertation aims to analyse designated zones of nighttime activities and gather information for various stakeholders to promote security, transportation as well as business opportunities by gaining insights on night life patterns and predict the density of people in a given area at a certain point in time. This study was conducted based on the data gathered through the year 2023 during the pre-determined nighttime activities timetable, from 6 pm to 7 am, the study focused in five districts “Arroios”, “Estrela”, “Misericórdia”, “Santa Maria Maior” e “Santo António”, as these were the districts that showed more movement during the established timetable.

Some key findings were gathered by analysing the data, during the year of 2023 it was found that the nightlife gets more movement in the second semester of the year, having its peak movement in the month of September. “Misericórdia” and “Santa Maria Maior” consistently registered a higher average of people passing by every month, showing their importance as central hubs of nightlife activities for both locals and tourists.

From an hourly perspective, “Santa Maria Maior” is the most frequented district in the beginning of the night, with a mix of people commuting from work to home and people starting to gather for drinks and dinners in this diversified district, but as soon as the night sets people tend to explore the streets of “Misericórdia” from “Bairro Alto” to “Cais de Sodré” the options to have a good time with friends are endless.

The main focus of this dissertation lies in the possibility of predicting the movement of people through the nighttime. A solution for this problem was studied by leveraging machine learning, a branch of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. For this dissertation, a Long Short-Term Memory (LSTM) network was used to study the possibility of performing time series forecasting based on the gathered data and predict the number of people in a certain part of Lisbon at a given time.

The model built is able to predict the fluctuation of people very well despite some deviations of the true number of people in the test set. The mean value of RMSE registered was 336.199, a small number as the number of people in the testing set fluctuated between 1000 and 8000 people. These deviations in the predictions were caused by focusing this study on a specific timetable, by cutting the data for this specific purpose spikes were created in the dataset, making the number of people change abruptly where a continuous flow of people increasing and decreasing should be. As LSTMs rely on sequential patterns to make predictions, by being provided with this dataset, it is harder for the model to capture the underlying daily patterns, which leads to sharper predicted spikes when it encounters sudden changes. Despite these deviations, the model is still perfectly accurate and can predict the volume of people in a given area at a given time with a small margin of error that can be easily hide by categorizing the number of people to be expected with these deviations in mind.

This dissertation tackles the objectives proposed by giving succinct information about the distribution of people through the night of Lisbon, as well as some mobility patterns during this time, and giving a solution that can help make informed decisions about security, transportation and business engagement tactics, by predicting the density of people in a given area at a given time, making it easy for everybody to follow or avoid large crowds.

6.2 Future work

In the future, exploring the mobility of people through full days, treating the data, and feeding it to the model might improve the model's ability to capture daily and weekly patterns. Maintaining a continuous data flow in future research would be of interest to enhance prediction accuracy. Exploring more complex and advanced models and compare their effectiveness with LSTMs, integrate external factors to the mode such as events, holidays, or weather conditions to improve prediction accuracy as well.

It would be very interesting to make this case study in different cities in order to find and compare the behaviours of people in different scenarios and make informed decisions that could lead to better

improvements in city planning as well as providing tools to guide governments and their people through crowded places in order to have more security and better mobility services.

References

[1] «Data protection benefits for you | European Data Protection Board». Acedido: 28 de setembro de 2024. [Em linha]. Disponível em: https://www.edpb.europa.eu/sme-data-protection-guide/data-protection-benefits-for-you_en

[2] S. Phithakkitnukoon, T. Horanont, A. Witayangkurn, R. Siri, Y. Sekimoto, e R. Shibasaki, «Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan», *Pervasive Mob. Comput.*, vol. 18, pp. 18–39, abr. 2015, doi: 10.1016/j.pmcj.2014.07.003.

[3] «LxDataLab», LISBOA ABERTA. Acedido: 2 de março de 2024. [Em linha]. Disponível em: https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/apresentacao

[4] J. Nofre, J. C. Martins, M. Garcia-Ruiz, e G. Teixeira Costa, «Una aproximación geográfica a los impactos de la turistificación del ocio nocturno en el barrio de Cais do Sodré, Lisboa», *Estud. Geográficos*, vol. 84, n.º 294, p. e129, jun. 2023, doi: 10.3989/estgeogr.2023135.135.

[5] A. Araya López, «Policing the "Anti-Social" Tourist. Mass Tourism and "Disorderly Behaviors" in Venice, Amsterdam and Barcelona.» University of Salento, 2020. doi: 10.1285/I20356609V13I2P1190.

[6] G. Olt, M. K. Smith, A. Csizmady, e I. Sziva, «Gentrification, tourism and the night-time economy in Budapest's district VII – the role of regulation in a post-socialist context», *J. Policy Res. Tour. Leis. Events*, vol. 11, n.º 3, pp. 394–406, set. 2019, doi: 10.1080/19407963.2019.1604531.

[7] C. Colomb e J. Novy, Eds., *Protest and Resistance in the Tourist City*, 0 ed. Routledge, 2016. doi: 10.4324/9781315719306.

[8] «World Travel Awards», World Travel Awards. Acedido: 29 de setembro de 2024. [Em linha]. Disponível em: https://www.worldtravelawards.com/

[9] A. Cecaj, M. Lippi, M. Mamei, e F. Zambonelli, «Sensing and Forecasting Crowd Distribution in Smart Cities: Potentials and Approaches», *IoT*, vol. 2, n.º 1, pp. 33–49, jan. 2021, doi: 10.3390/iot2010003.

[10] P. Chapman, *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS, 2000. [Em linha]. Disponível em: https://books.google.pt/books?id=po7FtgAACAAJ

[11] G. Lyons, «Getting smart about urban mobility – Aligning the paradigms of smart and sustainable», *Transp. Res. Part Policy Pract.*, vol. 115, pp. 4–14, set. 2018, doi: 10.1016/j.tra.2016.12.001.

[12] R. Giffinger e G. Haindl, «Smart cities ranking: an effective instrument for the positioning of cities?», em *5th International Conference Virtual City and Territory, Barcelona, 2,3 and 4 June 2009*, Centre de Política de Sòl i Valoracions, jun. 2009, pp. 703–714. doi: 10.5821/ctv.7571.

[13] *Oxford Dictionary of English*, 3.ª ed. Oxford University Press, 2010. doi: 10.1093/acref/9780199571123.001.0001.

[14] C. Balzotti, A. Bragagnini, M. Briani, e E. Cristiani, «Understanding Human Mobility Flows from Aggregated Mobile Phone Data», *IFAC-Pap.*, vol. 51, n.º 9, pp. 25–30, 2018, doi: 10.1016/j.ifacol.2018.07.005.

[15] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, e T. Li, «Predicting citywide crowd flows using deep spatio-temporal residual networks», *Artif. Intell.*, vol. 259, pp. 147–166, jun. 2018, doi: 10.1016/j.artint.2018.03.002.

[16] A. Crivellari e E. Beinat, «LSTM-Based Deep Learning Model for Predicting Individual Mobility Traces of Short-Term Foreign Tourists», *Sustainability*, vol. 12, n.º 1, p. 349, jan. 2020, doi: 10.3390/su12010349.

[17] U. Singh, J.-F. Determe, F. Horlin, e P. D. Doncker, «Crowd Forecasting Based on WiFi Sensors and LSTM Neural Networks», *IEEE Trans. Instrum. Meas.*, vol. 69, n.º 9, pp. 6121–6131, set. 2020, doi: 10.1109/TIM.2020.2969588.

[18] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, e H. Zhang, «Deep Learning with Long Short-Term Memory for Time Series Prediction», *IEEE Commun. Mag.*, vol. 57, n.º 6, pp. 114–119, jun. 2019, doi: 10.1109/MCOM.2019.1800155.
[19] R. Wirth e J. Hipp, «CRISP-DM: Towards a Standard Process Model for Data Mining».
[20] Y. Li e H. Cao, «Prediction for Tourism Flow based on LSTM Neural Network», *Procedia Comput. Sci.*, vol. 129, pp. 277–283, 2018, doi: 10.1016/j.procs.2018.03.076.
[21] A. G. R. Garcia, «Tourism flow forecasting for inbound European travel», masterThesis, 2023. Acedido: 30 de setembro de 2024. [Em linha]. Disponível em: https://repositorio.iscte-iul.pt/handle/10071/30719
[22] F. A. Gers, J. Schmidhuber, e F. Cummins, «Learning to Forget: Continual Prediction with LSTM», *Neural Comput.*, vol. 12, n.º 10, pp. 2451–2471, out. 2000, doi: 10.1162/089976600300015015.
[23] C. Olah, «Understanding LSTM Networks», colah's blog. Acedido: 11 de agosto de 2024. [Em linha]. Disponível em: https://colah.github.io/