INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Identifying Patterns in City Municipality Incident Reports by Lisbon Citizens: A Data-Driven Analysis

Pedro Miguel Silva de Sousa

Master's degree in Integrated Business Intelligence Systems

Supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso
Associate Professor,
ISCTE-IUL

Co-supervisor:
PhD José Eduardo de Mendonça Tomás Barateiro
Invited Assistant Professor,
ISCTE-IUL

September 2024

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Identifying Patterns in City Municipality Incident Reports by Lisbon Citizens: A Data-Driven Analysis

Pedro Miguel Silva de Sousa

Master's degree in Integrated Business Intelligence Systems

Supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso
Associate Professor,
ISCTE-IUL

Co-supervisor:
PhD José Eduardo de Mendonça Tomás Barateiro
Invited Assistant Professor,
ISCTE-IUL

September 2024

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Elsa Cardoso and Professor José Barateiro, for their time, dedication and valuable guidance throughout this work. Your support and advice were fundamental to the development of this dissertation.

I would also like to thank my colleagues, who made my experience at ISCTE both enriching and highly enjoyable, providing an atmosphere of friendship and continuous learning.

Finally, a very special thank you to my girlfriend, my family and my friends, for their unconditional support, motivation and, above all, their patience with me throughout this process. Without your support, this journey would have been far more difficult.

Resumo

Num cenário de crescente urbanização, as grandes cidades enfrentam o desafio de proporcionar condições ideais aos seus residentes. Este trabalho centra-se na importância da aquisição e análise de informações relativas a ocorrências urbanas, uma vez que este conhecimento é crucial na previsão de ocorrências e na eficiente alocação de recursos municipais. O objetivo primordial desta dissertação é fornecer dados e previsões que possam instrumentalizar as autoridades municipais na tomada de decisões informadas e na gestão proativa de áreas críticas da cidade.

A pesquisa baseia-se nos dados recolhidos através da aplicação "Na Minha Rua", disponibilizados pela Câmara Municipal de Lisboa. Ao explorar esta fonte de informação, pretende-se identificar padrões em ocorrências urbanas, mapeando áreas propensas a problemas específicos. Este estudo visa contribuir para uma abordagem mais eficaz na gestão urbana, possibilitando a antecipação de necessidades e a implementação das respetivas medidas preventivas.

Ao abordar a dinâmica urbana através da análise de dados, este trabalho oferece informações de valor para otimizar a segurança e a qualidade do ambiente urbano. Ao compreender e antecipar ocorrências, as autoridades municipais podem aprimorar a sua resposta aos mesmos, promovendo assim uma cidade mais segura e eficiente. Este trabalho representa um contributo para uma abordagem informada e proativa na gestão urbana, visando melhorar a qualidade de vida dos cidadãos na cidade de Lisboa.

Palavras-Chave: Gestão de Incidentes Urbanos, Ocorrências, Análise de Dados, Cidades Inteligentes, Qualidade dos Dados.

Abstract

In a scenario of escalating urbanization, major cities face the challenge of providing optimal conditions for their residents. This work focuses on the significance of obtaining and analysing information regarding urban occurrences, as such knowledge is crucial in forecasting occurrences and efficiently allocating municipal resources. The primary objective of this dissertation is to provide data and forecasts that can empower municipal authorities to make informed decisions and proactively manage critical areas of the city.

The research is based on data collected through the “Na Minha Rua” application, made available by the Lisbon City Council. By exploring this source of information, the aim is to identify patterns in urban occurrences, mapping areas prone to specific issues. This study seeks to contribute to a more effective approach in urban management, enabling the anticipation of needs and the implementation of preventive measures.

By addressing urban dynamics through data analysis, this work offers valuable insights for optimizing the safety and quality of the urban environment. By understanding and anticipating incidents, municipal authorities can enhance their response, thus promoting a safer and more efficient city. This work represents a contribution to an informed and proactive approach to urban management, aiming to improve the quality of life for citizens in the city of Lisbon.

Keywords: Urban Incident Management, Occurrences, Smart Cities, Data Analysis, Data Quality.

Index

Tables Index

Figures Index

List of abbreviations

ICT - Information and Communication Technologies

CRISP-DM - Cross-Industry Standard Process for Data Mining

DM - Data Mining

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

CML - Lisbon City Council (*Câmara Municipal de Lisboa*)

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

LSTM - Long Short-Term Memory

IPMA - Portuguese Institute for Sea and Atmosphere

CSV Comma-Separated Values

INE – Portuguese National Institute of Statistics

JF - Parish Councils (*Junta de Freguesia*)

BI - Business Intelligence

CHAPTER 1

Introduction

In the dynamic panorama of global urbanization, where over half of the world's population currently resides in urban areas, a figure projected to reach 60% by 2030 according to the United Nations [1], cities are transforming into epicentres of human activity and innovation. Historically, cities have always faced challenges related to infrastructure and resource management, dating back to the earliest urban settlements [2]. However, industrialization accelerated urbanization, increasing population density and intensifying pressure on municipal administrations to provide efficient public services and robust infrastructure [3]. Throughout the 20th and 21st centuries, urbanization became a major driver of social and economic change, leading to complex challenges ranging from urban planning to waste management and public transportation [4].

As these urban hubs expand to accommodate an ever-growing population, municipal administrations are facing increasing challenges. Identifying and addressing daily occurrences, along with the subsequent resolution, has become an essential concern in numerous cities [5], underscoring the critical need for an agile and well-informed approach to healthy urban governance.

Typical urban occurrences often include potholes in the streets, which cause damage to vehicles and pose risks to pedestrians and cyclists. Additionally, graffiti not only detracts from the urban aesthetic but also signifies neglect. Issues with public lighting further compromise safety, especially at night, while accumulated trash on the streets and overflowing dumpsters pose significant public health and environmental problems. Moreover, occurrences such as fallen trees, lack of street cleaning, and excessive noise must be quickly identified and addressed to ensure the general well-being of citizens and the smooth functioning of urban life [6].

An integral element of this approach is citizen engagement, recognized as a cornerstone of smart city governance [7]. Smart City governance can be described by the use of Information and Communication Technologies (ICT) to modernize and improve municipal management, with the goal of creating more efficient and sustainable cities [8]. The collaborative involvement of the public becomes a key factor, leveraging their insights and

experiences to create content and address challenges. In this paradigm, the efficient management of public services depends on not only the swift response of authorities to reported issues but also on the active engagement of citizens.

To facilitate greater public participation and streamline the process of addressing urban occurrences, the Lisbon City Council (CML) developed the "*Na Minha Rua*" online portal and smartphone application[1]. Through this platform, citizens report occurrences such as infrastructure issues and environmental concerns, creating a valuable database of reports, which will be the basis for this research.

Examining the information from these reports goes beyond just looking at isolated data, evolving into an important exercise in identifying patterns within reported occurrences. The emphasis on understanding these patterns is essential for proactive urban management, a strategy that extends beyond reactive responses to occurrences. The ability to discern trends and anticipate issues empowers municipal authorities to allocate resources wisely and preventatively address recurring problems.

1.1. Motivation and topic relevance

This research is driven by the need for cities to embrace new approaches in the face of growing urbanization, affecting over half of the global population [1]. By focusing on incident reporting and data analysis, the aim is to improve urban living through discerning patterns within citizen occurrence reports, providing meaningful insights for proactive urban management [9].

As cities continue to expand, the demands on municipal services and infrastructure intensify [10], needing more efficient solutions. Traditional methods of urban management, such as manual reporting of issues, routine scheduled maintenance, and paper-based record-keeping, often fall short of meeting the evolving needs of modern cities. These approaches create a gap in service delivery and responsiveness, as they can be slow, inefficient, and unable to cope with the dynamic nature of urban environments.

Cities like Lisbon deal with escalating incidents among a growing population. Examining the complexities of urban growth and, more crucially, identifying patterns within

[1] *https://naminharualx.cm-lisboa.pt/*

occurrences reported is a pressing need. Using data from the "*Na Minha Rua*" application [11], this research goes beyond isolated incident analysis. It specifically aims to uncover recurring patterns, providing a comprehensive understanding of challenges faced by Lisbon's residents, spanning infrastructure issues to environmental concerns. By decoding these patterns, the goal is to support the Municipality's decision-making at both strategic and operational levels. This analysis aims to identify areas where certain types of issues are more frequent, allowing responsible entities to allocate resources more effectively and expedite resolution processes. This approach is expected to enhance the efficiency of urban management while contributing to a safer and more sustainable urban environment for all Lisbon residents.

1.2. Objectives and Research Questions

This dissertation aims to explore urban governance through a comprehensive analysis of data from the "*Na Minha Rua*" platform, covering the period from 2019 to 2023. This extended timeframe allows us to detect both short-term variations and long-term trends, thereby deepening our understanding of the evolving dynamics within the city. The insights derived from this study will be instrumental in guiding strategic decisions and urban planning initiatives.

1.2.1. Objectives

1. Investigate the geographical distribution of occurrences, to identify localized patterns and pinpoint urban challenges;

2. Identify and analyse temporal patterns in reported occurrences, offering dynamic insights into the evolving nature of urban issues over time;

3. Investigate the nature and typology of the most frequent occurrences, providing a comprehensive understanding of systemic issues and contributing factors.

These objectives provide a foundation for a focused exploration of occurrence patterns, aiming to deliver valuable information for stakeholders to make informed decisions, optimize resources, and enhance the city's infrastructure.

1.2.2. Research Questions

1. Spatial Distribution of occurrences: How are these occurrences spatially distributed across different areas of Lisbon?

2. Emerging Patterns: What perceptible patterns emerge when analysing occurrence reports on a spatial and temporal level?

These questions guide the research, aiming to provide a comprehensive understanding of occurrence patterns in Lisbon, facilitate informed decision-making, and contribute to the proactive management of urban challenges.

1.3. Methodologic approach

In the exploration of this research, the application of the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) framework plays a key role. Renowned for its standard approach in data mining projects, CRISP-DM aims to streamline processes, reduce costs, and enhance reliability, repeatability, and manageability, ultimately elevating the efficiency of the data mining process [12]. As shown in *Figure 1* this is an iterative process.

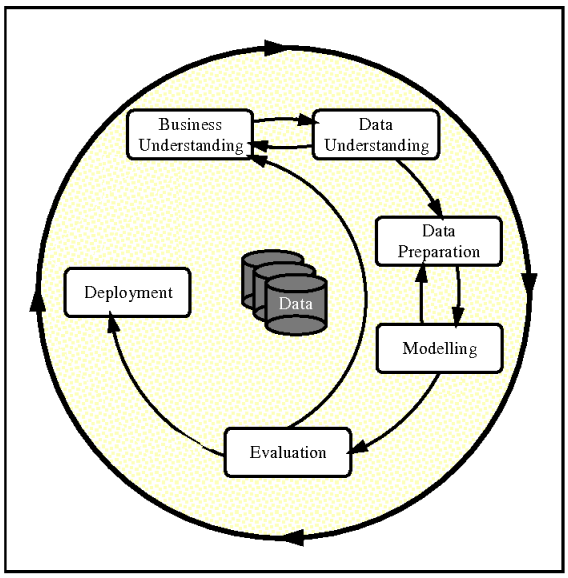

Figure 1 - CRISP-DM. Source R. Wirth and J. Hipp [13]

This study engages only with the initial five phases of the CRISP-DM methodology, diverging from the original six-phase structure. The deliberate exclusion of the deployment phase is anchored in its association with result delivery. In the context of this dissertation, the insights gained will be organized into a comprehensive report, tailored for presentation to stakeholders of the Lisbon City Council.

This methodology initiates with Business Understanding, emphasizing the comprehension of the project's primary objectives from a business perspective. The subsequent step involves Data Understanding, where the focus shifts to acquainting oneself with the available data. This phase embraces exploration to identify issues related to data quality and draw initial insights from the variables. Moving to the third phase, Data Preparation, the cleaning process begins, ensuring data quality validation and normalization, followed by integration. The fourth phase, Modelling, will be replaced by Data Exploration, as this work did not involve modelling. Instead, the focus will be on applying Data Mining (DM) techniques aligned with the initially defined objectives, using data exploration to gain insights and identify patterns in the dataset. This phase prioritizes data visualization to discern patterns and extract information crucial for achieving the proposed goals. Finally, the fifth phase was initially planned to be altered from Evaluation to Discussion, with the intention of validating the results and confirming whether the defined objectives had been successfully achieved, based on feedback from the responsible team of Lisbon City Council. However, as this feedback was not obtained in time for submission, the planned Discussion phase could not take place, leaving this phase without effect.

CHAPTER 2

Literature Review

2.1. Introduction

Modern city administration faces complex challenges related to efficient management of urban services and improving citizen's quality of life [14]. A valuable source of data for supporting decision-making is information derived from municipal occurrence reports submitted by citizens, providing direct insights into the problems faced by the local population [15].

Citizens in large cities worldwide are increasingly utilizing technologies such as smartphone applications to communicate urban occurrences to their municipalities [16]. These reports provide significant insights into citizens' concerns about urban infrastructure, public lighting, waste collection, safety, and other essential municipal services. These insights are crucial for city authorities to prioritize and address issues to enhance overall citizen satisfaction and the quality of life in urban areas [17].

The data provided by citizens is highly valuable for municipal administration. Analysing these reports can reveal patterns, trends, and problematic areas that may not be immediately apparent [18]. Understanding these trends assists municipal managers in implementing more effective interventions and continuously monitoring the quality of services provided. However, identifying patterns in these occurrence reports requires a systematic, data-driven approach. This includes analysing large volumes of unstructured data, applying data science techniques to interpret the data, and identifying meaningful patterns. For example, the study by authors Dhini A. et al. [19] utilized text clustering to identify specific topics within community reports, such as flood-related issues, transportation problems, housing and land use concerns.

2.2. Literature review methodology

The literature review for this research adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, as outlined in the work of Moher et al. [20]. PRISMA provides a structured and comprehensive approach to systematically

review and synthesize existing literature, ensuring transparency and reproducibility in the process.

The primary search for relevant studies was conducted on Scopus [21] and Web of Science [22] databases, aligning with the systematic review standards recommended by PRISMA. The main goal is to find and select literature related to occurrence reports made by citizens in the context of smart cities.

The query employed for the literature search is designed to encompass key themes related to incident or problem management, smart cities, citizen participation and data-driven approaches. The specific query utilized was: ((“incident” OR “problem” OR “risk”) AND (“smart city” OR “smart cities” OR “urban” OR “city”) AND (“report” OR “data driven” OR “data analysis”) AND (“citizen participation” OR “crowdsourcing” OR “crowdsensing”)). This selection of keywords ensures a targeted and relevant set of articles for inclusion in the review.

The literature search focused on articles, conference papers, and book chapters published between 2014 to 2024 and targeted the disciplines of computer science, engineering, decision sciences, mathematics, and business management. The target is English language publications aiming to capture the most pertinent contributions to the field.

The *Figure 2* illustrates the flowchart of the PRISMA methodology, along with the various steps leading to the final articles present in this study.

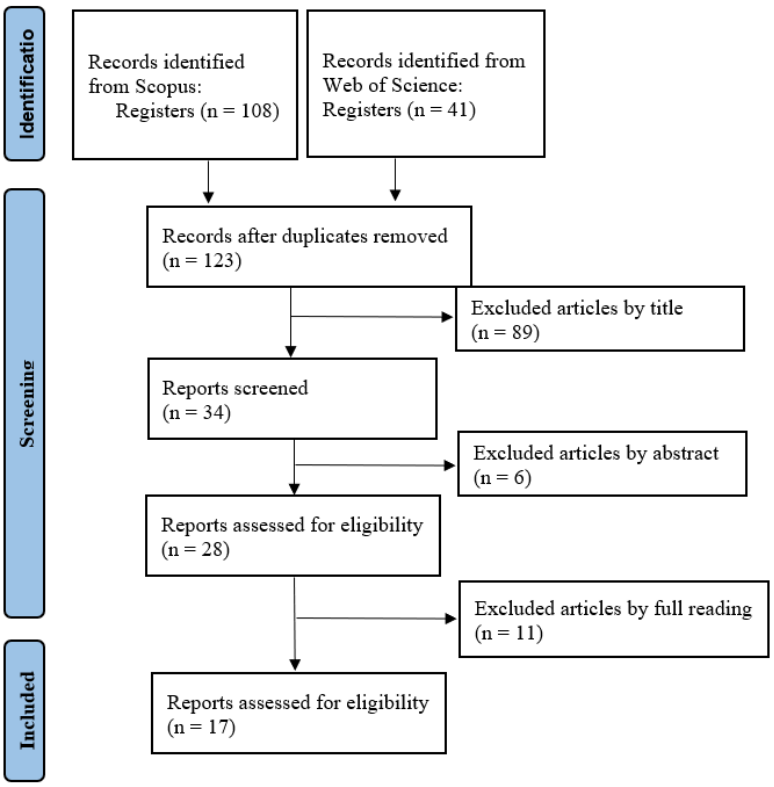

Figure 2 - PRISMA flow diagram. Adapted from PRISMA 2020 flow diagram [23]

The search process across the two databases yielded a total of 149 documents (108 from Scopus and 41 from Web of Science). Following the removal of duplicate records, the remaining pool consisted of 123 documents. A preliminary screening of document titles was then conducted, resulting in the exclusion of articles that fell outside the scope of this review, leaving a total of 34 documents. Articles are considered out of scope and excluded from the review if their titles don't align with the research related to the topics considered in the search query. By focusing on articles whose titles correspond to the relevant themes of the study, the selection process ensures a targeted examination of the existing literature. This approach facilitates the identification of articles that directly contribute to the research objectives and excludes those that fall outside the scope of the study. Subsequent examination of abstracts led to the identification of 28 documents for further consideration.

In the final stage of the systematic review, the full texts of these 28 documents were assessed in detail. This process resulted in the exclusion of 11 documents, ultimately producing a final set of 17 eligible articles for inclusion in the literature review.

2.3. Results

Presenting the results obtained, *Figure 3* shows a bar chart illustrating the distribution of selected articles by country. Indonesia leads with four articles, followed by Greece and the United States of America, each with two articles. Other countries contribute with one article each, reflecting a diverse global perspective on smart city governance and incident management.

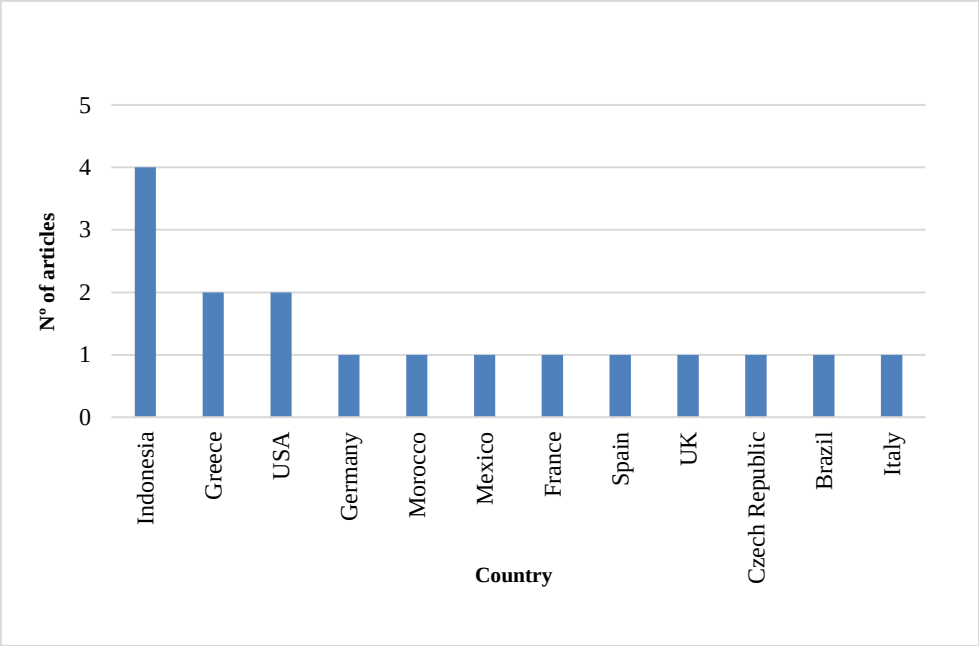

Figure 3 - Number of articles per country

Additionally, *Figure 4* shows a temporal distribution of the selected articles, showcasing publication years ranging from 2015 to 2024. In 2016, there were three articles published in this field, while 2017 saw a peak in scholarly contributions with five articles, making it the year with the most publications in this area. In recent years, the number of articles published has been lower.

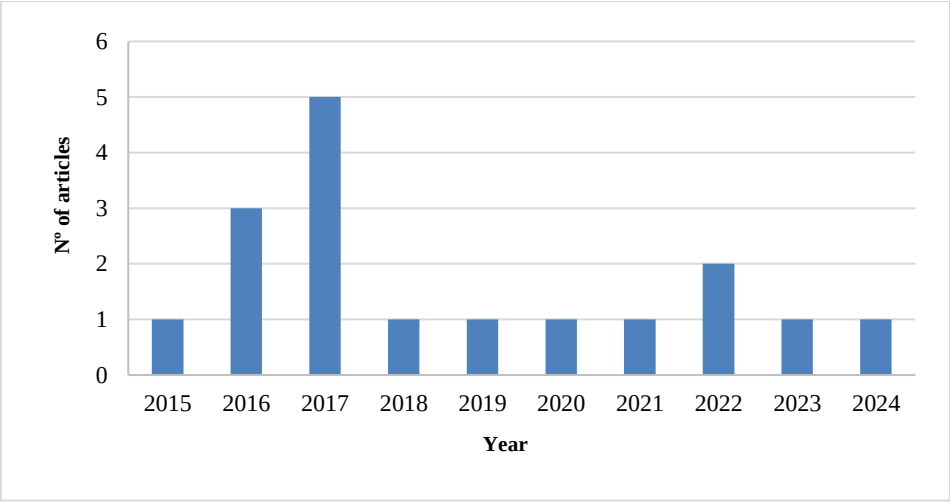

Figure 4 - Number of articles per year

As shown in *Table 1* the selected documents are organized in ascending order by title. The table includes information about the authors, article title, document type, source of publication, publication year, and country. It also provides an overview of the various applications discussed, the methods used and correspondent validation, and the number of citations for each article. The methods applied across these studies include advanced techniques such as DBSCAN [15], [24] for identifying clusters, Maximum Entropy algorithms [25], and LSTM classification [26], among other methodologies. This diverse range of approaches demonstrates the breadth of strategies employed in existing research and offers a comprehensive explanation of the techniques explored in the context of urban incident analysis.

Table 1 - Summary of the publications found in the literature search

Authors	Title	Document Type	Source Title	Year	Country	Application	Method	Evaluation	Nº of Citations
Xanthopoulos T.; Anagnostopoulos T.; Kytagias C.; Psaromiligkos Y.	A smartphone-enabled crowdsensing and crowdsourcing system for predicting municipality resource allocation stochastic requirements	Conference paper	ACM International Conference Proceeding Series	2020	Greece	Smartphone-enabled system for predicting municipality resource allocation.	LSTM classification model	Prediction accuracy	0
Yosua Grandy Ara S.; Suzianti A.	Analysis of technology adoption for real-time aspiration delivery system	Conference paper	ACM International Conference Proceeding Series	2017	Indonesia	Analyses factors affecting technology adoption of Real-Time Aspiration Delivery System.	Structural equation modelling (SEM) with UTAUT (Unified Theory of Acceptance and Use of Technology) and TTF (Task-Technology Fit) models.	Chi-Square, Importance-Performance Analysis (IPA) and T-Value Analysis	1
Tadili J.; Fasly H.	Citizen participation in smart cities: A survey	Conference paper	ACM International Conference Proceeding Series	2019	Morocco	Survey on citizen participation in smart cities for innovative solutions.	Survey	-	2
Ariya Sanjaya I.M.; Supangkat S.H.; Sembiring J.	Citizen Reporting Through Mobile Crowdsensing: A Smart City Case of Bekasi	Conference paper	Proceeding - 2018 International Conference on ICT for Smart Society: Innovation Toward Smart Society and Society 5.0	2018	Indonesia	Spatial analysis reveals distribution of citizen reports in Bekasi.	Analysis of citizen participation in providing data to SOROT (Smart Online Reporting & Observation Tools) platform	Spatial/Temporal Analysis and Report Status Analysis	8
Bousios A.; Gavalas D.; Lambrinos L.	CityCare: Crowdsourcing daily life issue reports in smart cities	Conference paper	Proceedings - IEEE Symposium on Computers and Communications	2017	Greece	Introduces CityCare and CityCareW mobile apps for citizen-centric problem reporting	Questionnaires for evaluation.	User Questionnaires	18
Dhini A.; Hardaya I.B.N.S.;	Clustering and visualization of community	Conference paper	Proceeding - 2017 3rd International Conference on	2017	Indonesia	Text clustering of public complaints on urban issues in	Text clustering to identify specific topics in community complaints	Determining the error values	1

Authors	Title	Document Type	Source Title	Year	Country	Application	Method	Evaluation	Nº of Citations
Surjandari I.; Hardono	complaints and proposals using text mining and geographic information system		Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era			Jakarta, prioritizing urban issues like drainage, roadwork, and infrastructure.			
Pistolato A.C.; Brandão W.C.	Connectcity: A collaborative e-government approach to report city incidents	Conference paper	Proceedings of the 15th International Conference	2016	Brazil	Collaborative e-government approach to improve incident reporting and classification	Usability evaluation with the SURE methodology and think-aloud method.	Users' evaluation	0
Rodríguez-García D.; García-Díaz V.; González García C.	Crowdsl: Platform for incidents management in a smart city context	Article	Big Data and Cognitive Computing	2021	Spain	Design and development of a platform for reporting and managing the incidents that are originated in the day-to-day of a city.	DSL evaluation	Users Likert Scale	2
Brus J.; Vrkoč J.; Kubásek M.	Design of decision support tools for the quality assessment of illegal dumping notifications based on crowd-sourced data	Conference paper	Environmental Modelling and Software for Supporting a Sustainable Future	2016	Czech Republic	Decision support tool designed for quality assessment of illegal dumping reports.	MaxEnt (Maximum Entropy) algorithm for illegal dump probability distribution modelling.	Accuracy	1
Zhang J.; Wang D.	Duplicate report detection in urban crowdsensing applications for smart city	Conference paper	Proceedings - 2015 IEEE International Conference on Smart City	2015	USA	Developed a duplicate report detection scheme for smart city applications	Expectation Maximization (EM) framework; ST-DBSCAN algorithm	Accuracy, F-1 Score, Precision, and Recall Rate	16

Authors	Title	Document Type	Source Title	Year	Country	Application	Method	Evaluation	Nº of Citations
Supangkat S.H.; Ragajaya R.; Setyadji A.B.	Implementation of Digital Geotwin-Based Mobile Crowdsensing to Support Monitoring System in Smart City	Article	Sustainability (Switzerland)	2023	Indonesia	Implements Digital Geotwin for 3D visualization of urban problems.	Object Positioning System (OPS) combined with triangulation, trilateration, computer vision techniques, and multi-modal sensor information	-	3
Kaluarachchi Y.	Implementing Data-Driven Smart City Applications for Future Cities	Article	Smart Cities	2022	UK	Review of data-driven smart city applications	Systematic review	-	44
De Filippi F.; Coscia C.; Boella G.; Antonini A.; Calafiore A.; Cantini A.; Guido R.; Salaroglio C.; Sanasi L.; Schifanella C.	MiraMap: A We-Government Tool for Smart Peripheries in Smart Cities	Article	IEEE Access	2016	Italy	MiraMap IT tool enhances citizen-administration interaction in urban peripheries.	Community Impact Analysis method for qualitative evaluations.	-	32
Tehrani P.F.; Pfennigschmidt S.; Kriegel U.; Billig A.; Fuchs-Kittowski F.; Meissen U.	Multidimensional report analysis in urban incident management	Conference paper	Proceedings of the 2017 4th International Conference on Information and Communication Technologies for Disaster Management	2017	Germany	Proposes a method to classify, filter, correlate, and cluster incident reports	Semantic reasoning, information retrieval, and artificial intelligence classifiers.	Fault Tolerance and Availability measurement	0
Liu, Z; Bhandaram, U; Garg, N	Quantifying spatial under-reporting disparities in resident crowdsourcing	Article	Nature Computational Science	2024	USA	Method to identify reporting delays without external ground-truth data; Spatial disparities in reporting rates linked to socioeconomic characteristics.	Poisson rate estimation	Pearson correlations	0

Authors	Title	Document Type	Source Title	Year	Country	Application	Method	Evaluation	Nº of Citations
Ruiz-Correa S.; Santani D.; Ramírez-Salazar B.; Ruiz-Correa I.; Rendón-Huerta F.A.; Olmos-Carrillo C.; Sandoval-Mexicano B.C.; Arcos-Garcia Á.H.; Hasimoto-Beltrán R.; Gatica-Perez D.	SenseCityVity: Mobile Crowdsourcing, Urban Awareness, and Collective Action in Mexico	Article	IEEE Pervasive Computing	2017	Mexico	SenseCityVity project engages youth in Guanajuato City to address urban issues.	Analysing urban problems through geolocalized images, audio, and video data.	-	27
Boumchich A.; Picaut J.; Bocher E.	Using a Clustering Method to Detect Spatial Events in a Smartphone-Based Crowd-Sourced Database for Environmental Noise Assessment	Article	Sensors	2022	France	DBSCAN method used to identify clusters in NoiseCapture database.	DBSCAN clustering method applied to detect spatial events.	Contingency-table-based analysis	2

2.3.1. Smart Cities

The concept of “smart city”, according to the authors Yosua Grandy Ara and Suzianti A. [27], can be interpreted as the utilization of networked infrastructure to improve economic and political efficiency and enable social, cultural, and urban development. This encompasses business services, housing, leisure and lifestyle services, and information and communication technologies such as mobile and fixed phones, satellite TV, computer networks, and internet services.

The selected articles [17], [27], [28] provide key insights into smart city initiatives and their impact on modern urban development. Yamuna Kaluarachchi emphasizes the need for data-driven applications to tackle challenges like climate change and population growth [29]. Similarly, Jihane Tadili and Hakima Fasly focus on fostering citizen participation and local government involvement to promote inclusivity [17]. Yosua Grandy Ara S. and Amalia Suzianti highlight the importance of aligning technology adoption with user needs for having a successful implementation [27].

2.3.2. Crowdsourcing applications for citizen reporting

Crowdsourcing applications for citizen reporting play an important role in smart cities by enhancing the connection between residents and local government entities [30]. These platforms empower citizens to report urban issues and contribute to community improvement and governance [14].

The "MiraMap" project, as discussed in the article of the authors De Filippi F et al., focuses on creating a technology platform for smart peripheries within smart cities. This initiative strengthens communication between residents, public authorities, and local institutions to improve social inclusion and community engagement [30].

Other applications like "ConnectCity" and "CrowDSL" are collaborative platforms designed to facilitate the reporting and classification of city incidents. "ConnectCity" streamlines communication between citizens and government entities for efficient incident resolution, as outlined by its authors [31]. "CrowDSL" employs a model-driven approach to enhance incident management and promote citizen engagement, according to the authors [14]. Similarly, "SenseCityVity," a mobile crowdsourcing platform in Mexico, empowers citizens to report urban issues and fosters governance and community involvement in smart cities [32].

Collectively, these platforms demonstrate the importance of crowdsourcing applications in advancing citizen reporting and engagement, which is vital for effective urban governance and the development of smart cities.

2.3.3. Methods and Application

Researchers have employed a variety of approaches to address urban issues and enhance the efficiency of smart city applications. These methods span different areas of concern, including noise pollution, illegal dumping, resource allocation, and duplicate report detection.

The author Boumchich A. et al. wrote about the “Noise-Planet” project that has developed an open-source smartphone tool to assess noise pollution levels [24]. By using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering method to identify clusters in the data collected by the users. The project aims to improve data analysis and the relevance of noise maps. In other article on clustering methods, focusing on community complaints in Jakarta [19], text mining and clustering technics where applied, highlighting key urban issues such as drainage, roadwork, and infrastructure.

Another study introduced a smartphone-enabled crowdsensing system to predict municipality resource allocation [26]. Utilizing a Long Short-Term Memory (LSTM) classification model with a 10-fold cross-validation evaluation method, the system demonstrated potential for improving urban resource distribution.

The study on illegal dumping in Prague leveraged crowdsourcing and Geographic Information System (GIS) technologies to design decision support tools for assessing spatial data quality [25]. The “ZmapujTo” application, mentioned in the study, aimed to predict dump occurrences and forward reports to city management departments.

Lastly, a study focusing on smart city applications developed an Expectation Maximization framework to detect duplicate reports. This scheme achieved significant improvements in accuracy, F-1 measure, precision, and recall rate [15].

These methods and applications demonstrate the potential of data-driven strategies to address diverse urban issues and enhance smart city development.

2.3.4. Keywords occurrence analysis

For this analysis, the bibliometric tool VOS Viewer [33] was used to construct and visualize keyword co-occurrence networks. This method creates a network where each keyword is represented as a node and each pair of co-occurring words forms a link between the nodes. The strength of the link is determined by the frequency of the keyword pair's co-occurrence [34].

Figure 5 presents the 15 keywords with the highest weights from the 17 documents analysed. The figure highlights four distinct clusters of keywords, each represented by a distinct colour. These keywords are related to the focus of this study, as they cover topics associated with smart cities, crowdsourcing, and incident management.

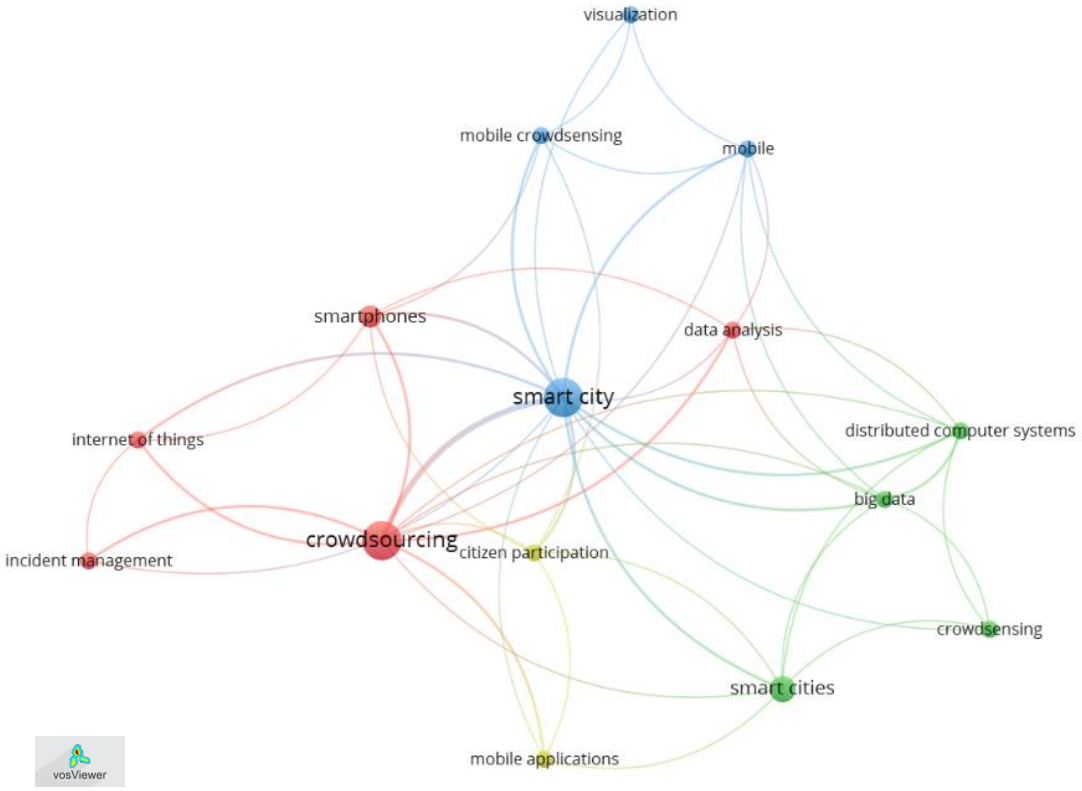

Figure 5 - Keyword occurrence network visualization from VOS Viewer

CHAPTER 3

Analytical process of occurrence reports

In this chapter, the focus is the systematic application of the initial three phases of the CRISP-DM framework to address the challenge of identifying patterns within the citizen reported occurrences.

The first phase, Business Understanding, begins by establishing a clear comprehension of the research objectives and the context of the study. Next, Data Understanding delves into the data itself, exploring its properties, quality, and relevance. This phase seeks to understand the various attributes of the dataset and gain preliminary insights into the occurrences reported, paving the way for deeper analysis. Finally, Data Preparation focuses on readying the data for analysis [35]. This involves cleaning, transforming, and structuring the data to ensure it is suitable for the subsequent phases Data Exploration and analysis.

3.1. Business Understanding

Lisbon, according to information from the Portuguese National Institute of Statistics (INE) [36], based on the census conducted in 2021, is the fourth city with the highest population density in Portugal, with an average of 5 466,5 inhabitants per square kilometre [37]. The high concentration of people and activities amplifies the impact of issues such as accumulated waste, deficient public lighting, or road potholes. Early identification and prompt response to these occurrences are essential to prevent them from escalating into larger problems, affecting the quality of life, safety, and local economy.

This dissertation is based on the challenge *"Identificação de padrões na aplicação Na Minha Rua*[2]", proposed by *LxDataLab*[3]. *LxDataLab* is an initiative of the Lisbon Urban Management and Intelligence Center, of the Lisbon City Council, aimed at using data generated in the city of Lisbon to develop analytical solutions capable of solving real problems and improving services provided to residents, workers, and visitors. It consists of a partnership between the municipality of Lisbon and various entities, including higher education institutions and scientific research institutions. The academia benefits from access to real data about the city, and the Lisbon City Council has the opportunity to test analytical and data visualization solutions capable of promoting innovation, efficiency, and proactivity in services provided to citizens [38].

The following *Figure 6* illustrates the steps required to report an urban occurrence on the "*Na Minha Rua*" online platform. The process consists of five stages, from selecting the type of problem to localising it, describing it, including photographs and finally submitting the report.

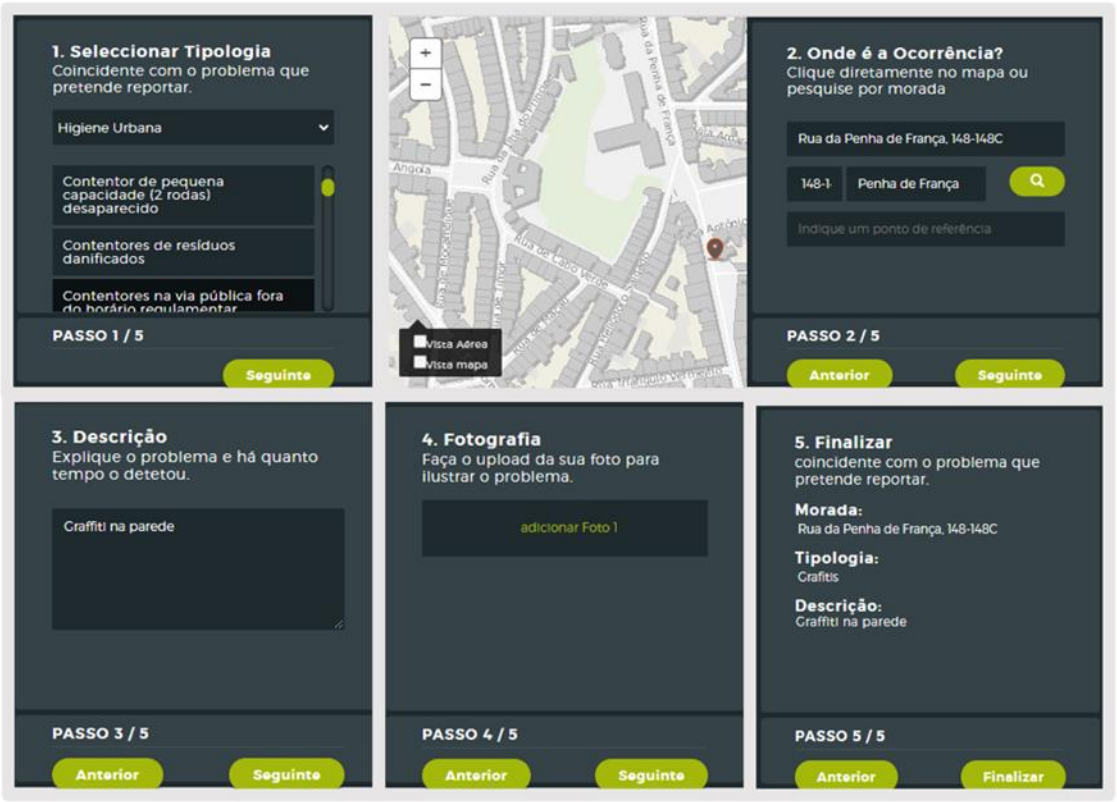

Figure 6 - Five steps to make a report on the "Na Minha Rua" platform

[2] *https://dados.cm-lisboa.pt/dataset/desafio-01-23-lxdatalab*
[3] *https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/apresentacao*

1. In the initial menu, the user must select the type of problem, which is the category that best describes the occurrence. Firstly, select the problem area, such as "*Urban Hygiene*" and then, within the chosen area, select a typology, such as "*Graffiti*";

2. The next step is to indicate the exact location of the problem. This can be done by clicking directly on the map or by entering the address in the search field;

3. Next, the user must describe the problem, giving details about the nature of the occurrence and how long ago it was detected. In the example given, the description is "Graffiti on the wall";

4. After that, there is the possibility of adding a photograph that illustrates the problem, helping the authorities to better understand the situation;

5. Finally, in the last step, the user must review all the information provided. This is the last opportunity to ensure that all the details are correct before finalising and sending the report. By clicking on "Finalise", the occurrence is submitted for analysis and action by those responsible.

The *Figure 7* below shows the interface of the "*Na Minha Rua*" portal, which allows the visualization and monitoring of reported occurrences. Users can consult a detailed list of problems, check the resolution status, and use filters to search by codes, areas and occurrence statuses.

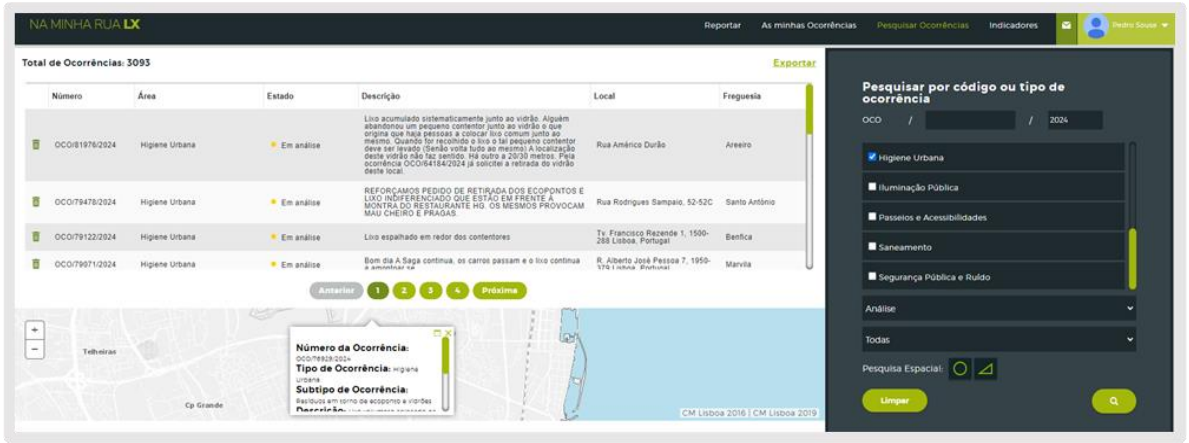

Figure 7 - "Na Minha Rua" interface

The challenge focuses on using data generated by citizens through the "*Na Minha Rua*" application to identify patterns and trends in urban occurrences. The application serves as a channel for Lisbon residents to communicate with municipal authorities about various issues affecting their city, from infrastructure concerns to public safety, among others.

Understanding the business context is important for extracting relevant information from the data. The "*Na Minha Rua*" application acts as a bridge between citizens and local governance, facilitating a direct and real-time flow of information about urban challenges. By analysing this data, municipal officials can gain a deeper understanding of the needs and concerns of Lisbon residents, enabling more efficient and targeted responses.

From a strategic perspective, identifying patterns in the data reported by citizens has several benefits. It provides municipal authorities with a clearer picture of occurrences, paving the way for proactive measures in resource allocation and service provision. This type of analysis promotes better urban planning, targeted interventions, and an overall improvement in governance and quality of life.

The challenge highlights the need for applying data analysis techniques to extract valuable insights from the data. Through the examination of these patterns, authorities can improve their ability to solve problems proactively, streamline services, and cultivate a more responsive and sustainable urban environment. By exploring these patterns, we aim to highlight some of the dynamics of Lisbon's urban context, thus making significant contributions to municipal management and planning efforts.

3.2. Data Understanding

To achieve the objectives outlined in this dissertation, it is essential to explore and understand the selected datasets. For this process, two datasets were used. One dataset was made available by *LxDataLab*, with data relating to the "*Na Minha Rua*" application, and the second with socioeconomic indicators obtained from the INE.

Exploring both datasets is crucial as it allows for the identification of potential correlations between variables. By integrating the information gathered from the "*Na Minha Rua*" dataset with the socioeconomic data from the INE, it becomes possible to discern patterns within reported occurrences. Socioeconomic factors such as population density, education, and condition of the buildings may influence the occurrence and nature of reports, providing valuable insights into their underlying causes. This approach enhances the understanding of the interplay between social conditions and human behaviour, ultimately facilitating the development of more effective strategies for incident prevention and occurrence management.

This process of familiarization with the data involves conducting exploratory tasks to gain a comprehensive understanding, identify potential data quality issues, and identify essential variables for analysis. This preliminary step is of great importance as it allows us to determine if the available datasets contain the necessary information to achieve the proposed objectives.

3.2.1. "Na Minha Rua" Dataset

This dataset includes occurrences reported by Lisbon citizens in the "*Na Minha Rua*" platform. The dataset consists of a CSV (Comma-Separated Values) file, and the data period is between 01/01/2018 and 31/03/2024.

It should be noted that this was the second set of data made available by *LxDataLab*. Initially, a dataset was provided that did not show the time at which the reports were made, only indicated the day. Given the need for this information for a more precise and detailed analysis, a request was made for the time of the reports to be made available. In response to this request, a new dataset was provided, and this is the dataset used to conduct this work.

Initial analysis revealed that the dataset consists of 765 769 records, each of which represents a single occurrence reported by time of day. The *Table 2* below shows the ten columns present in the file, along with a description of each one, these being the dataset's attributes.

Table 2 – "Na Minha Rua" Dataset description

Variable Name	Variable Description	Variable Type	First line of data
id	Occurrence identification number	Int64	663932
dt_registo	Occurrence creation date	Date-Time	01/01/2018 06:28:47
area	Occurrence area, which groups together several types of occurrences	Object	Higiene Urbana
tipo	Type of occurrence that varies according to the area of occurrence	Object	Entulhos, objetos volumosos…
longitude	Longitude coordinate of the occurrence	Float64	-9.1335238
latitude	Latitude coordinate of the occurrence	Float64	38.7328749
codsig	Internal system code	Int64	104129
servico_resp	Entity responsible for resolving the situation	Object	CML
estado	Occurrence status	Object	Resolvido
dt_estado	Date of status change	Date-Time	04/01/2018 15:41:01

It is also important to note that the dataset does not have any duplicate lines or null values, which simplifies the analysis process.

The *"dt_registration"* variable represents the day and time when the occurrence is reported on the platform; from this variable we can see the time period of the data made available. *Figure 8* shows the annual distribution of the number of occurrences reported on the "*Na Minha Rua*" platform between 2018 and 2024.

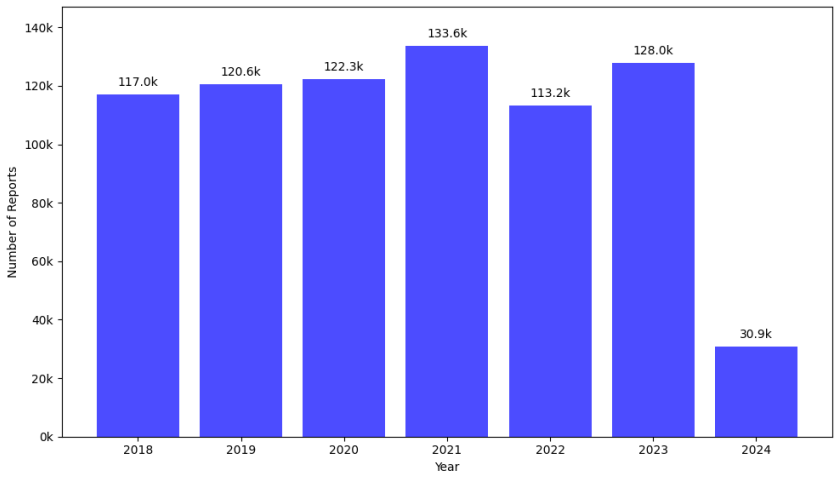

Figure 8 - Number of Reports per Year (Complete Dataset)

It can be noted that the dataset is balanced in terms of the number of occurrences reported each year, except for 2024. 2021 was the year with the highest number of occurrences reported, peaking at 133 625 records. The fact that there were local elections for Lisbon City Council in September 2021 may have had an influence on the number of reports registered that year. During election periods, there is a tendency for increased civic participation and greater attention to local issues, which can result in a higher number of occurrences reported by citizens. Voters may be more motivated to report urban problems in the hope that candidates will address these issues in their campaigns and electoral promises. This association may explain the peak observed in the number of occurrences reported in 2021.

There is then a slight decrease in 2022, with 113 221 occurrences, followed by a further increase in 2023, with 127 984 reports. In 2024, the number of occurrences is significantly lower as the dataset only contains data up to the end of March of that year.

The "*area*" column contains eight thematic areas that encompass various types of occurrences. *Figure 9* shows the distribution of occurrences reported in the "*Na Minha Rua*" platform. It is noteworthy that the majority are related to "*Urban Hygiene*" totalling 580 444 occurrences. The remaining categories, such as "*Public Lighting*" and "*Walkways and Accessibility*" have significantly lower volumes, with 40 846 and 39 071 occurrences, respectively.

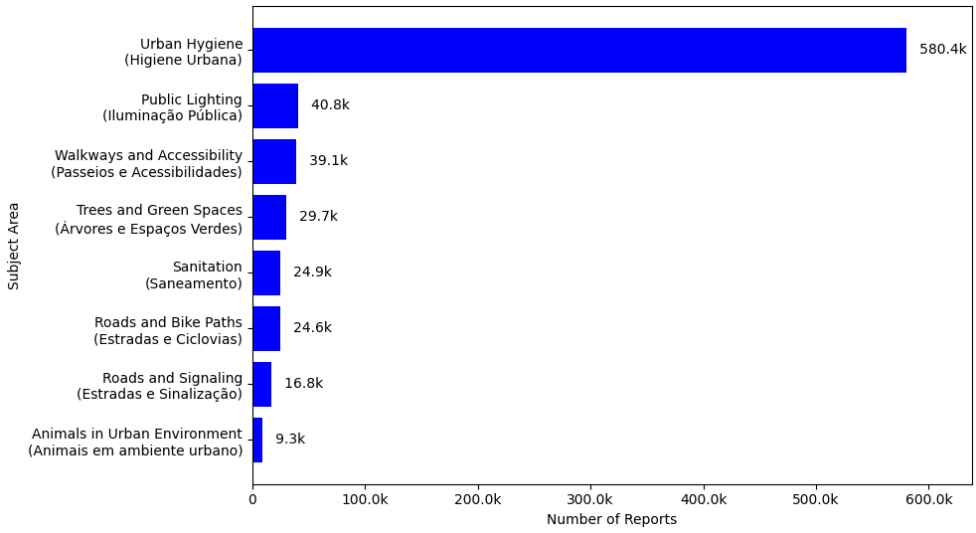

Figure 9 - Number of Reports by Subject Area

The "***type***" column provides a more detailed breakdown of the nature of the occurrence. Analysing the dataset, 251 different types were identified.

The *Table 3* and *Figure 10* below shows the 10 most reported types of occurrences, where the main subcategories stand out: "*Removal - Large Waste Collection Request*" is the most common, with 268 266 records (35,03% of the total); "*Rubble and objects abandoned on the public highway*" follows with 101 848 occurrences (13,30%). Other sub-categories, such as "*Lamp Out*" and "*Removal-Gardens-Collection Request*" account for 2,96% and 2,93% respectively. The Others category, which accounts for 32,30% of occurrences, includes the remaining less frequent problems.

Table 3 - Top 10 most reported types of occurrences

Occurence Type (in Portuguese)	**Count**	**Frequency**
Removal – Large Waste Collection Request **(***Remoção-Monstros-Pedido de recolha*)	268 266	35,03%
Rubble and objects abandoned on the public highway (*Entulhos, objetos volumosos…*)	101 848	13,30%
Lamp out **(***Candeeiro apagado*)	22 638	2,96%
Removal-Gardens - Collection Request (*Remoção-Jardins-Pedido de recolha*)	22 408	2,93%
Pests and diseases **(***Pragas e doenças*)	21 608	2,82%
Complaints regarding the daily collection of solid urban waste (*Reclamações no âmbito da recolha diária de resíduos sólidos urbanos*)	19 320	2,52%
Graffiti **(***Grafitis*)	18 472	2,41%
Street cleaning **(***Limpeza da via pública*)	16 560	2,16%
Trees, shrubs or grass – Maintenance (*Árvores, arbustos ou relva – Manutenção*)	14 564	1,90%
Selective Removal - Occasional removal of paper/cardboard (*Remoção Seletivas - Remoção pontual de papel/cartão*)	13 113	1,71%
Others	246 972	32,30%

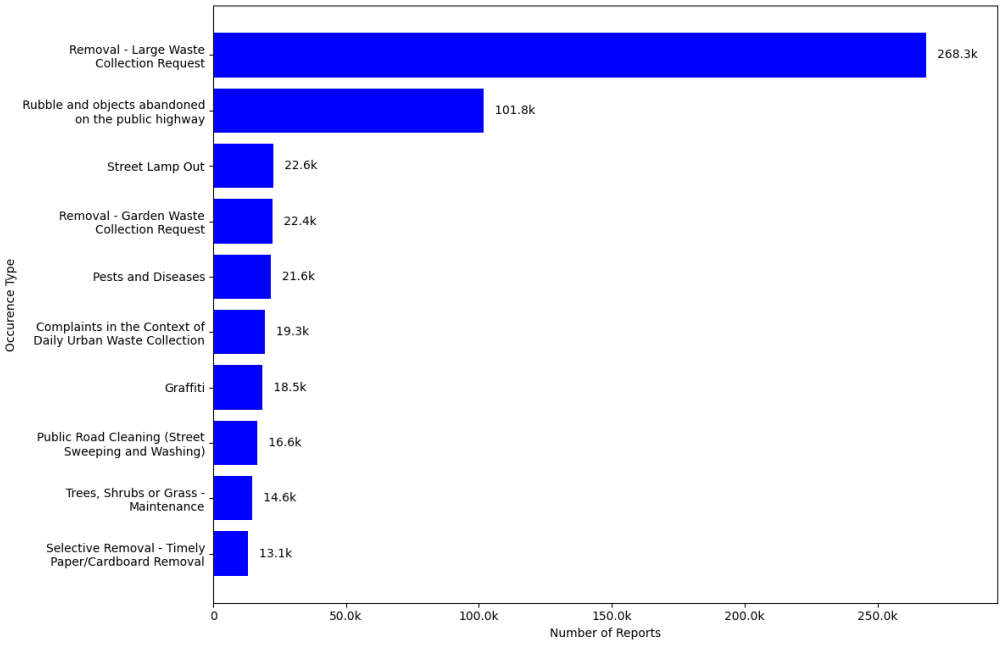

Figure 10 - Top 10 most reported types of occurrences

The "*status*" column categorises reported occurrences based on their stage of resolution. This attribute is important for assessing the efficiency of the response and management of occurrences by the competent authorities.

Table 4 below shows the percentage distribution of occurrences according to their current status. It can be seen that the majority, 95,7%, are marked as "*Resolved*", indicating a high rate of resolution of the problems reported. The remaining categories, "*Under Analysis*", "*In Progress*", and "*Registered for Resolution*", represent a small fraction of the total, which shows that the majority of occurrences are dealt with and subsequently reported as concluded.

Table 4 - Distribution of occurrences by status of resolution

Status	**Count**	**Frequency**
Resolved	732 753	95,7%
Under Analysis	19 120	2,5%
In Progress	9 317	1,2%
Registered for Resolution	4 579	0,6%
Total	**65 769**	**100%**

The "*service_resp"* attribute identifies which organisation is responsible for dealing with reported occurrences. This attribute helps to understand the distribution of responsibilities between the different organisations involved in managing and solving urban problems in Lisbon, namely *Lisbon City Council* (CML) and the *Parish Councils* (JF).

Table 5 shows that *CML* is the predominant responsible entity, accountable for 90,45% of occurrences, totalling 692 634. *JF* manage 9,31% of the reports, corresponding to 71 292 occurrences. *External Entities* are responsible for only 0,24% of cases, with 1 843 occurrences. These are probably situations that require competencies or services outside the scope of municipal and parish bodies, indicating an occasional need for a specialised intervention.

Table 5 - Responsible Entities

Responsible Entity	Count	Frequency
CML	692 634	90,45%
JF	71 292	9,31%
External Entity	1 843	0,24%
Total	**765 769**	**100%**

This distribution indicates that most occurrences are resolved by the *CML*, while the *JF* deal with a smaller but still significant proportion (approximately 10%) of local problems.

3.2.1.1 A Critical Analysis of the Classification of occurrence Types

The proper categorisation of occurrence types is fundamental to obtaining a more accurate analysis. For this reason, a critical analysis was carried out with the aim of identifying inconsistencies and redundancies in the classification of the types of occurrences reported in the available dataset, proposing improvements for clearer and more useful categorisation.

When analysing the dataset, eight different classification areas and 251 different types of occurrences were identified, which were then analysed according to the following aspects:

1. **Identification of redundancies and inconsistencies**: During the analysis, several typologies with the same or very similar names were detected in multiple areas, indicating possible redundancies and inconsistencies in their classification. Some of the typologies that stood out for appearing repeatedly in different areas can be noted in *Table 6:*

- "*Surface dips*": Found in the "*Roads and Cycleways*", "*Roads and Signalling*", and "*Walkways and Accessibility*" classification areas (see *Table 6*).

- "*Bituminous*": Present in "*Roads and Cycle Routes*", "*Roads and Signalling*", and "Pavements and Accessibility" classification areas (see *Table 6*).

- "*Hole in the carriageway*": Appears in several instances as "*Bituminous*", "*Concrete*", and "*Cubes*" in the "*Roads and Cycleways*" and "*Roads and Signalling*" areas (see *Table 6*).

Table 6 - Same typology present in several classification areas

Classification Area	**Typology**
Roads and Cycleways (*Estradas e Ciclovias*)	Surface dips (*Abatimentos superficiais*)
Roads and Signalling (*Estradas e Sinalização*)	
Walkways and Accessibility (*Passeios e Acessibilidades*)	
Roads and Cycleways (*Estradas e Ciclovias*)	Bituminous (*Betuminoso*)
Roads and Signalling (*Estradas e Sinalização*)	
Walkways and Accessibility (*Passeios e Acessibilidades*)	
Roads and Cycleways (*Estradas e Ciclovias*)	Hole in the carriageway – Bituminous (*Buraco na faixa de rodagem - Betuminoso*)
Roads and Signalling (*Estradas e Sinalização*)	
Roads and Cycleways (*Estradas e Ciclovias*)	Hole in the carriageway – Concrete (*Buraco na faixa de rodagem - Betão*)
Roads and Signalling (*Estradas e Sinalização*)	
Roads and Cycleways (*Estradas e Ciclovias*)	Hole in the carriageway – Cubes (*Buraco na faixa de rodagem - Cubos*)
Roads and Signalling (*Estradas e Sinalização*)	

In addition, the duplication of some typologies was observed, such as "*Horizontal signalling*", which appears twice due to an additional space at the end of one of the entries. This type of error led to a distinction between two identical typologies.

2. **Need to review the categorisation of classification areas**: In addition to redundancies in the typologies, the need to review and possibly reorganise the categorised areas was identified. For example, the "*Roads and Cycle Routes*" and "*Roads and Signalling*" areas have an overlap of most of the reported types related to the pavement, which may lead to some confusion when users of the "*Na Minha Rua*" platform report the occurrence. All the situations identified are shown in the *Appendix 1*.

3.2.2. INE Dataset

To complement the analysis of occurrences recorded in the "*Na Minha Rua*" platform, additional information from the INE was incorporated. INE is the organisation responsible for producing and disseminating official statistics in Portugal, providing reliable data on various social, economic and demographic aspects of the country. This data was obtained from the 2021 nationwide census [39].

The *Table 7* below shows the various detailed indicators for each Lisbon parish that will be used in this work.

Table 7 - INE indicators dataset

Parish	Population density (inhab/km2)	Foreign Nationality (%)	Higher education completed (%)	Average age (years)	Buildings in need of repair (%)	Area (Km2)	Resident population (nº)
Ajuda	4 967,36	7,28	28,3	46,89	37	2,88	14 306
Alcântara	2 731,76	11,75	40,31	45,01	48,9	5,07	13 850
Alvalade	6 237,64	5,38	56,03	45,25	36	5,34	33 309
Areeiro	12 302,33	7,86	55,88	45,02	29,8	1,72	21 160
Arroios	15 634,74	23,33	46,49	43,29	33,2	2,13	33 302
Avenidas Novas	7 779,60	9,97	61,13	44,69	39,1	2,99	23 261
Beato	4 912,50	11,84	26,08	46,37	45,9	2,48	12 183
Belém	1 586,39	7,12	56,95	45,81	23,3	10,43	16 546
Benfica	4 409,23	5,98	40,47	47,82	39	8,02	35 362
Campo de Ourique	13 418,18	11,26	47,01	44,89	49,1	1,65	22 140
Campolide	5 338,27	10,48	36,34	45,7	51,5	2,77	14 787
Carnide	4 885,64	4,76	42,31	44,92	61,9	3,69	18 028
Estrela	4 405,87	14,35	53,74	42,96	47,9	4,6	20 267
Lumiar	7 052,36	5,08	60,63	43,3	23,8	6,57	46 334
Marvila	4 983,01	5,42	14,93	45,22	48,7	7,12	35 479
Misericórdia	4 410,05	18,67	39,66	45,96	35,4	2,19	9 658
Olivais	3 977,63	5,8	31,83	47,41	26,7	8,09	32 179
Parque das Nações	4 114,34	9,55	56,07	40,44	16,6	5,44	22 382
Penha de França	10 507,38	14,93	35,02	45,03	37,8	2,71	28 475
Santa Clara	7 037,20	8,65	26	39,45	45,8	3,36	23 645
Santa Maria Maior	3 339,20	33,28	29,01	44,22	44,7	3,01	10 051
Santo António	7 422,82	17,34	54,62	44,41	36,3	1,49	11 060
São Domingos de Benfica	7 943,12	6,31	57,26	46,22	26,2	4,29	34 076
São Vicente	7 013,07	19,86	35,05	45,48	50	1,99	13 956

These indicators include:

1. *Population density* (inhab/km²): This indicator lets us know the relationship between the population and the surface area of the territory.

2. *Percentage of Foreign Nationality*: A metric that can indicate the diversity of the population and potential challenges in communication and access to services.

3. *Percentage of Higher Education Completed*: This data is useful for assessing the population's level of education, which can affect civic participation and awareness of urban issues.

4. *Average Age (years)*: Offers an insight into the age distribution of the population.

5. *Percentage of Buildings in Need of Repair*: An indicator of the state of urban infrastructure, which can be correlated with the number or type of occurrences.

6. *Area (Km2)*: Area in km2 of each parish.

7. *Resident Population*: Number of residents per parish.

Integrating this data allows for a richer and more contextualised analysis of occurrences, correlating socio-economic and demographic characteristics with the frequency of reported occurrences. This can make it easier to identify critical areas where urban interventions may be more necessary.

3.3. Data Preparation

In the Data Preparation phase, the focus is to ensure that the dataset is ready for analysis. This process involves selecting relevant data, creating new variables that can add value to the analysis and eliminating redundant or irrelevant information.

Data selection consists of identifying and keeping only the attributes that are essential to achieving the study's objectives. New variables will be created to capture additional characteristics of the occurrences that may be important for the analysis, such as temporal aggregations or derived classifications. Finally, eliminating information that is not of interest will help to simplify the dataset, improving the efficiency of the subsequent analysis and reducing the noise that could compromise the results.

3.3.1. "Na Minha Rua" Dataset

In the "*Na Minha Rua*" dataset, after the initial analysis, it was found to contain no null values or duplicates, which simplified the data preparation process. However, the analysis conducted

in the previous Data Understanding section revealed that some typologies contained unnecessary spaces at the end, which affected the consistency of the data. To solve this problem, it was necessary to eliminate these spaces using the *strip* [40] function in *Python* [41], thus making the data more uniform.

To enrich the analysis and allow for more detailed exploration, new variables were created from those existing in the initial dataset. These new variables, derived from the original data, provide an additional, aggregated view of the reported occurrences, allowing for a more detailed analysis. *Table 8* below shows the new variables created, including the name of the variable and the variable that originated it.

Table 8 - New Variables

New Variable	Original Variable	Type	Data Sample
Season	dt_registo	Object	Winter
Day of the Week	dt_registo	Object	Monday
Period of the day	dt_registo	Object	Morning
Month	dt_registo	Object	January
Year	dt_registo	Object	2018
Elapsed Time	dt_estado/ dt_registo	Timedelta	3.09:12:14
Parish	latitude/longitude	Object	Arroios
Statistical Section	latitude/longitude	Object	110656023
Weekday Indicator	dt_registo	Object	Weekday/Weekend
Holliday Indicator	dt_registo	Object	Local Holiday/ National Holliday / Nonholiday

The new variables were created as follows:

1. "*Season*": This variable was created to identify the season corresponding to the date of each occurrence. Using the date of registration (***dt_registo***), the dates were categorised into four seasons: Spring, Summer, Autumn and Winter. The seasons were defined based on the specific start dates of the season [42], namely:

    - *Spring*: 20 March to 20 June
    - *Summer*: 21 June to 22 September
    - *Autumn*: 23 September to 20 December

- *Winter*: 21 December to 19 March

To do this, a function was implemented that checked the month and day of the registration date and assigned the correct season based on these intervals. This variable makes it possible to analyse seasonal patterns in the data, such as variations in the volume of occurrences or requests that may be influenced by climatic or seasonal factors.

2. "*Day of the Week*": The variable was extracted from the registration date (*dt_registo*). The feature *strftime* [43] of the *Pandas* [44] library was used to obtain the day of the week in textual format (e.g. "Monday", "Tuesday"...). Understanding the day of the week on which reports occur can help identify patterns, such as increases in activity on specific days, which can be useful for adjusting resources and operations.

3. "*Period of the Day*": This variable was created by categorising the time of the registration date (***dt_registo***) into five time periods. Each period was defined as follows:

    - *Dawn*: 00:00 - 05:59
    - *Morning*: 06:00 - 11:59
    - *Lunchtime*: 12:00 - 13:59
    - *Afternoon*: 14:00 - 17:59
    - *Evening*: 18:00 - 23:59

    A function was developed which, based on the time extracted from the registration date, assigned the corresponding period. Dividing the day into distinct periods helps to identify in which periods throughout the day more records occur, making it easier to analyse peaks and temporal patterns.

4. "*Year*": The variable was extracted directly from the registration date (*dt_registo*) using *Pandas* to obtain the year in numerical format.

5. "*Month*": This variable was derived from the registration date (*dt_registo*) using *Pandas* to obtain the month in text format.

6. "*Elapsed Time*": This variable was calculated from the difference between the status date (*dt_estado*) and the record date (*dt_registo*). The result was converted into a measure of time (duration), representing the time elapsed from the occurrence being recorded until its

resolution in the format *dd:hh:mm:ss*. *Pandas* date subtraction method was used, which returns a *timedelta* [45] object. Measuring the time taken to resolve an occurrence or fulfil a request is important for assessing operational efficiency and identifying possible delays in the process of resolving requests or incidents.

7. "*Parish*": For the study carried out, it was essential to assign a parish to each geographical point in the "*Na Minha Rua*" dataset to facilitate geographical analysis of the various occurrences. As the dataset provided did not have a description of the parish, it was necessary to obtain this information from the *latitude* and *longitude* attributes. The process involved using a *Shapefile* [46] containing the polygons of Lisbon's parishes, obtained from the CML portal[4]. The whole process was carried out using the *Google Colaboratory* [47] environment, using *Python* libraries.

   The steps for creating the "*Parish*" variable and dealing with possible incompatibilities are described below:

   a) Loading the data:
      - Parish *Shapefile*: The *Geopandas* [48] library was used to load the *shapefile* containing the polygons of the parishes.
      - Points Dataset: The CSV file "*Na Minha Rua*" containing the *latitude* and *longitude* coordinates of the points was loaded using the *Pandas* library.

   b) Conversion of Coordinates to Geographical Objects:
      - The latitude and longitude coordinates were converted into "*Point*" objects from the *Shapely* [49] library, thus enabling geospatial manipulations.

   c) Coordinate System Definition:
      - It was ensured that both datasets were in the same *WGS84 - EPSG:4326* coordinate system before being converted to an *EPSG:3857* [50] coordinate system suitable for distance calculations.

   d) Initial assignment of parishes:

[4] https://geodados-cml.hub.arcgis.com/maps/7322f3fe2a574a97a9accfd4dcec81e0/about

- Using the *Within* [51] function from the *Shapely* library, it was checked whether each point was within one of the parish polygons. If so, the corresponding parish is assigned to the point.

e) Checking and processing points outside the polygons:

- It was found that there were 30 points that had not been assigned a parish (null values). After checking these points, it was realised that they were on the border of the polygons, as can be seen in *Figure* 11:

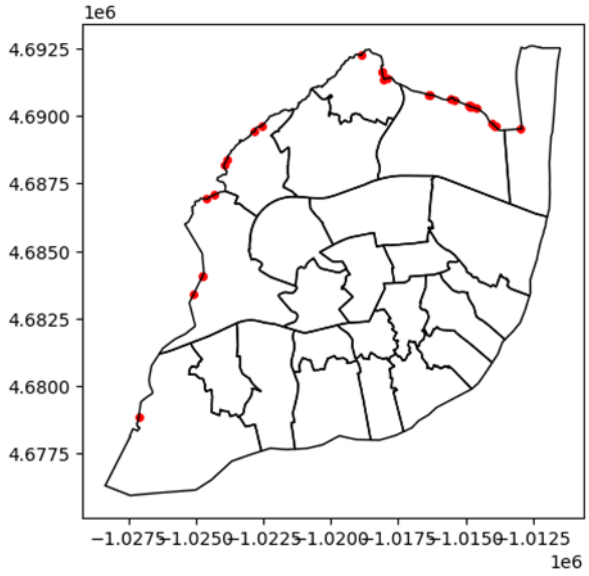

Figure 11 - Points on the border of the polygons

- For the points without an assigned parish, it was decided that the parish closest to each point would be assigned. For this task, the *Distance* [52] function of the *Geopandas* library was used, which measures the distance of each point from the polygons of the parishes, thus assigning the nearest parish.

f) Main dataset update:

- The main table has been updated to include the nearest parish for the 30 points that initially did not have a parish assigned.

8. "*Statistical Section*": to create this variable, the same process for assigning the parishes corresponding to the geographical points mentioned in point 7 above was followed. However, in this case, the focus is on assigning a statistical section to each point in the dataset. Statistical sections, as defined by INE, are territorial units corresponding to a continuous area within a parish, with approximately 300 residential dwellings [53]. These

sections provide a more granular level of detail than parishes, offering finer insights into the geographical distribution of occurrences.

To accomplish this, a *Geopackage* [54] file containing the statistical sections was obtained from the INE website and subsequently converted into a *shapefile* for use in the analysis. This new variable was only added in the final phase of the work, as it became evident that in certain parishes it was important to provide a more detailed breakdown of the data. By working at this more refined level, it will allow for more precise analyses of urban problems and the identification of trends or patterns that might be obscured when looking at data aggregated by parish. The figure *Figure 12* below shows the polygons of the various statistical sections of the city in black and the layer of parishes in overlay in blue.

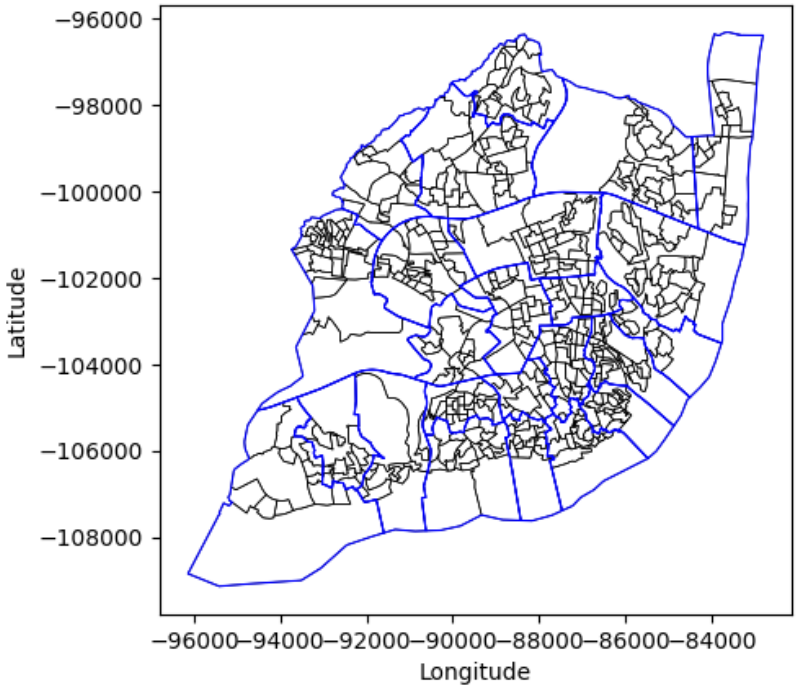

Figure 12 - Statistical sections and parishes of the city of Lisbon

9. "*Weekday Indicator*": The variable was created, which classifies each occurrence reported as "*Weekday"* or "*Weekend"*. To do this, a function was applied to the (*dt_registo)* column which checks whether the day of the week is Saturday or Sunday (classifying it as a "*Weekend"*) or Monday to Friday (classifying it as a "*Weekday"*). This function used the *dt.dayofweek* [55] method to determine the day of the week and stored the result in the new "*Weekday Indicator"* column. This variable allows for more detailed analyses of the influence of weekdays compared to weekends on reported occurrences. *Table 9* shows the values obtained.

Table 9 - Weekday Indicator

Weekday Indicator	Number of Days	Number of Reports	Average Reports per Day
Weekday	1 630	692 200	424,6626
Weekend	652	73 569	112,8359

10. "*Holiday Indicator*": In addition, the "*Holiday Indicator"* variable was created, which identifies whether the day on which the occurrence was reported is a public holiday or not. The *holidays* [56] library was used to define the national holiday calendar in Portugal, ensuring that all national holiday dates were identified, including the municipal holiday of Santo António in Lisbon (on June 13). A function was applied to the *dt_register* column to determine whether the date corresponds to a public holiday, classifying each occurrence according to this condition and storing the result in the new "*Holiday Indicator*" column. This variable makes it possible to specifically analyse how public holidays influence the frequency and types of occurrences reported. *Table 10* shows the values obtained.

Table 10 - Holiday Indicator

Holiday Indicator	Number of Days	Number of Reports	Average Reports per Day
Not Holiday	2 195	757 982	345,3221
Holiday	87	7 787	89,50575

3.3.2. Data Selection

In the Data Preparation phase, in addition to creating new variables, it was necessary to select the data that would be used to conduct this analysis. During this process, it was realised that the data for 2024 only included information up to the end of March. Therefore, in order to guarantee the temporal consistency of the analysis, it was decided to exclude the 2024 data. As a result, the total number of records went from 765 769 to 734 822, considering only the years 2018 to 2023 in their entirety.

CHAPTER 4

Data Analysis and Exploration

This chapter focuses solely on the Data Exploration phase of the CRISP-DM process. Initially, it was intended to cover both Data Exploration and Discussion, however, since feedback from the Lisbon City Council was not received in time, the Discussion phase could not be carried out. As a result, this chapter will concentrate on consolidating the insights gained through data exploration and ensuring that the results align with the objectives initially defined.

In the Data Exploration phase, Data Mining and data visualization tasks were conducted, which are an essential step to facilitate understanding and help detect relevant patterns in the data. Visualization is a powerful tool that allows complex data to be transformed into graphical representations that are easier to interpret, making it easier to identify trends, anomalies and correlations. In addition, correlation methods will be applied to detect correlations between the variables in the dataset from the "*Na Minha Rua*" platform and the socio-economic indicators from the INE data. This combined analysis will make it possible to predict the factors that may have an influence on the occurrences reported in the city of Lisbon.

4.1. Temporal Analysis of Reported Occurrences

To start the data visualization phase, it is essential to understand the distribution of occurrences over the defined time period, which covers 01/01/2018 to 31/12/2023. Visualizing the data makes it easier to interpret, allowing us to intuitively detect variations in the number of occurrences and relate them to possible external events.

Firstly, the annual variation in the number of reports is shown. In the *Figure 13* we can see the distribution of the number of occurrences reported each year. This visualization helps us understand how the volume of reports has evolved over the years. We can see a moderate variation over time, with 2021 being the year with the most occurrences reported.

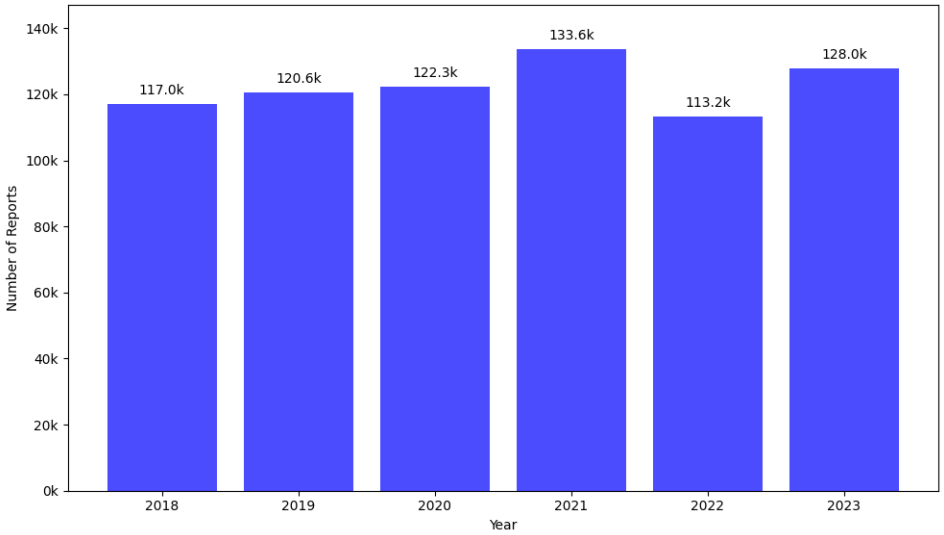

Figure 13 - Number of reports per year

Looking at the monthly evolution over the years in *Figure 14*, we can see two sharp drops in the number of occurrences during the lockdown periods decreed due to the COVID-19 pandemic [57].

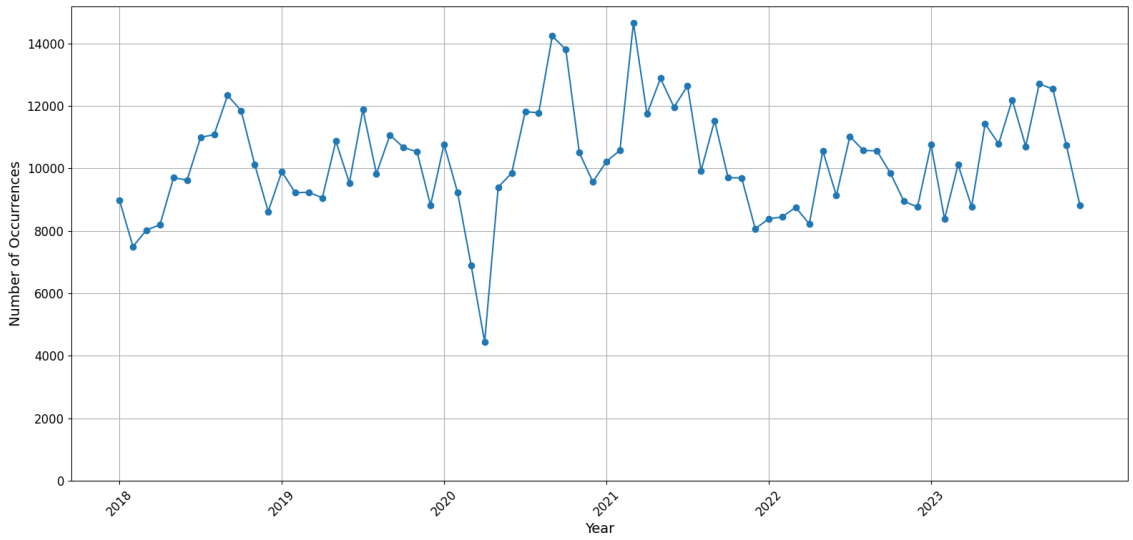

Figure 14 - Monthly evolution of occurrences (2018-2023)

In the *Figure 15* below, we focus only on the years 2020 and 2021, where we see that the first drop occurs between January and April 2020, a period in which COVID-19 cases tended to increase, leading to the first lockdown being decreed in March 2021, with a subsequent start of deconfinement from the end of April 2020 [58] (marked on the graph).

The second wave of the pandemic resulted in a further lockdown from the end of October 2020 until mid-February 2021. These periods of lockdown, marked on the graph, show a decrease in reports made, possibly due to mobility restrictions and the focus of authorities and citizens on issues related to the pandemic.

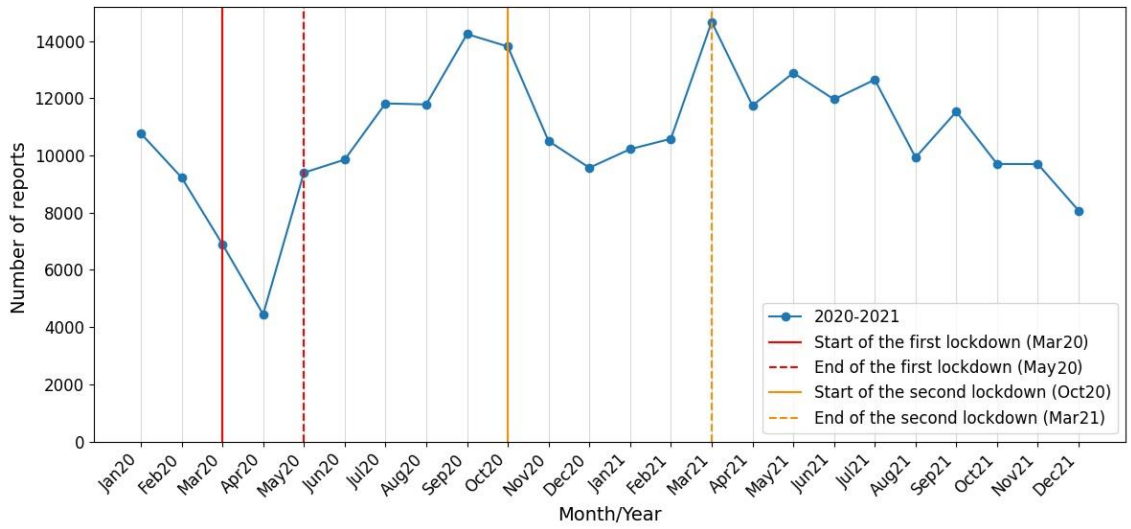

Figure 15 - Monthly evolution of reports during COVID-19 periods (2021-2022)

This pattern suggests that external events, such as the COVID-19 pandemic, have a significant impact on the number of occurrences reported, reflecting changes in citizen behaviour and urban service operations during times of crisis.

After a chronological analysis, it is also important to investigate the seasonal variation in occurrences over the years. Analysing occurrences by season allows us to see how the different weather conditions and seasonal activities of citizens can influence the number of reports. The *Figure 16* shows the number of reports made by season over the years 2018 to 2023. The coloured bars indicate the different seasons: *autumn* (green), *winter* (blue), *spring* (yellow) and *summer* (red).

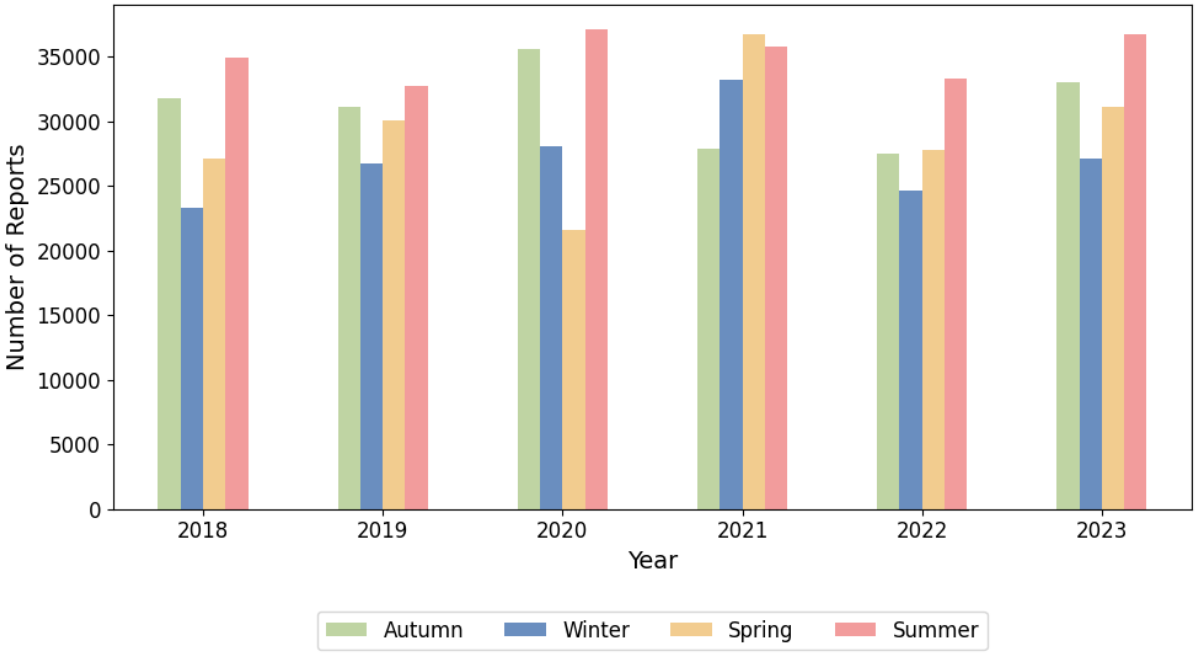

Figure 16 - Number of reports per season by year

It can be noted that, in general, *summer* and *autumn* tend to have a higher number of occurrences, possibly reflecting greater citizen activity outdoors, which makes it easier to detect urban problems during these seasons. On the other hand, *winter*, marked by more adverse weather conditions, consistently has a lower number of reports. *Spring* shows some variation over the years, with 2020 being the season with the fewest occurrences and the following year, 2021, the one with the most occurrences reported. This seasonal pattern suggests that weather conditions and the seasonal activity of citizens significantly influence the number of reports.

The *Table 11* shows the four most frequent types of occurrences in each of the seasons. It is clear that the type "*Removal-Bulky items - Collection Request"* is the most frequently reported in all seasons. *"Pests and diseases"*, on the other hand, show significant variations depending on the time of year, being more frequent in spring and summer.

Table 11 - Most frequent types of occurrences per season

Season	Occurrence Type	Count	Frequency
Autumn	Removal-Bulky items - Collection Request	66 274	35,48%
	Rubble and objects abandoned on the public highway	24 585	13,16%
	Lamp out	6 925	3,71%
	Complaints regarding the daily collection of solid urban waste	4 998	2,68%
Winter	Removal-Bulky items - Collection Request	59 664	36,60%
	Rubble and objects abandoned on the public highway	20 001	12,27%
	Removal-Gardens - Collection Request	6 600	4,05%
	Lamp out	6 329	3,88%
Spring	Removal-Bulky items - Collection Request	59 699	34,23%
	Rubble and objects abandoned on the public highway	25 400	14,56%
	Pests and diseases	4 993	2,86%
	Removal-Gardens - Collection Request	4 730	2,71%
Summer	Removal-Bulky items - Collection Request	72 856	34,60%
	Rubble and objects abandoned on the public highway	29 388	13,96%
	Pests and diseases	9 699	4,61%
	Street cleaning	6 004	2,85%

The *Table 11* above shows that *"Pests and Diseases"* tend to be more frequent in spring and summer, which is why it is important to check the number of occurrences over the years. The *Figure 17* below illustrates the number of reports related to *"Pests and Diseases"* over the twelve months from 2018 to 2023. The vertical bars represent the sum of the number of occurrences for each month in the various years, and the coloured lines represent the annual evolution of the number of occurrences of this type.

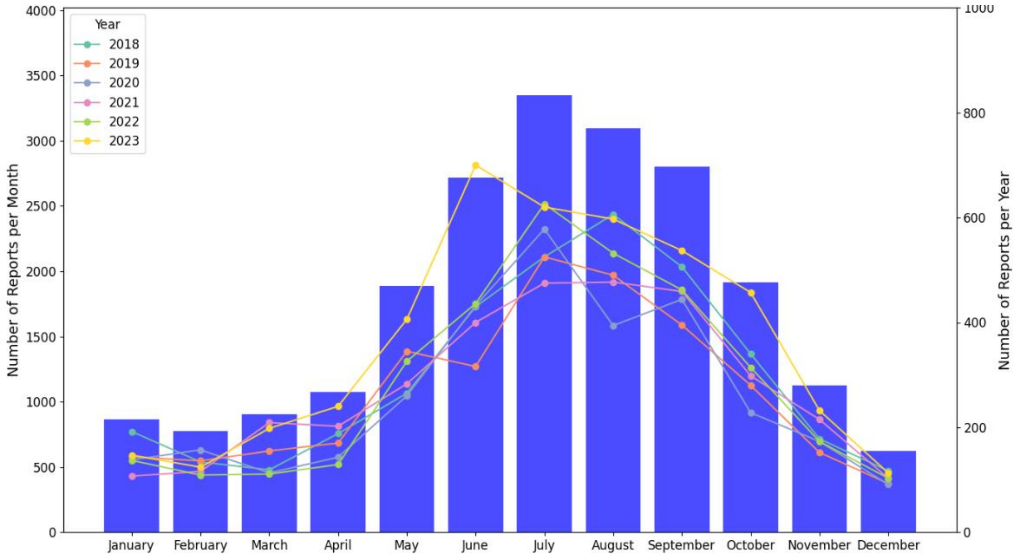

Figure 17 - Monthly and Annual number of reports for "Pests and Diseases"

Analysing the graph, it is possible to see that there is a clear seasonal pattern, with significant variations in the number of occurrences throughout the year. The hottest months have the highest number of occurrences in every year. In August, the number of reports peaks, with over 3 200 occurrences. July and September are also high, with over 2 500 reports. In contrast, the coldest months are the ones with the lowest figures, with only 843 occurrences in February. This pattern suggests a strong influence of climatic conditions and the biological activity of pests and diseases, which tend to proliferate during the warmer months.

Now observing once again the figures in the *Table 11*, we can also notice that unlike the occurrences related to *"Pests and Diseases"*, the reports related to *"Lamp out"* had the opposite effect. They were more frequent in the winter and autumn months, as can be clearly observed in the *Figure 18* below.

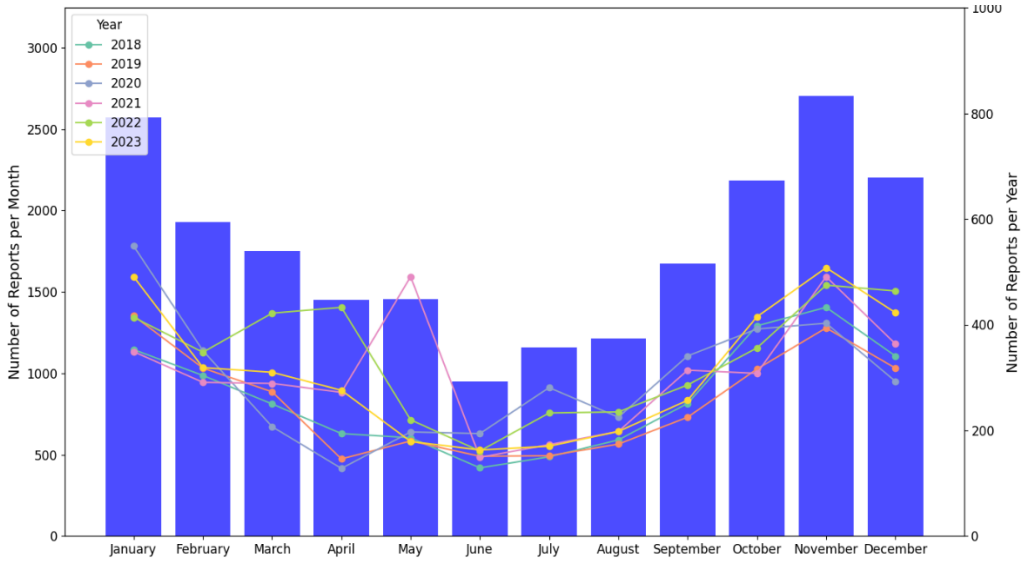

Figure 18 - Monthly and Annual number of reports for "Lamp out"

The *Figure 18* shows the monthly and annual number of reports for the *"Lamp out"* category over the years 2018 to 2023. There is a seasonal variation in reports, with a significant peak in January of each year, followed by a downward trend until June. From July onwards, there is a gradual increase in the number of reports, reaching another peak in November. This variation may be related to changes in daylight conditions throughout the year, where the winter months, with shorter days, lead to an increase in the perception of this type of occurrence and consequent reporting of these situations. In addition, the trend line for each year shows fluctuations, but maintains a similar overall pattern, highlighting the seasonal consistency of this type of occurrence.

Another important analysis is the distribution of occurrences by day of the week, as can be seen in the *Figure 19* below, which shows the aggregate of occurrences by day of the week over the period from 2018 to 2023.

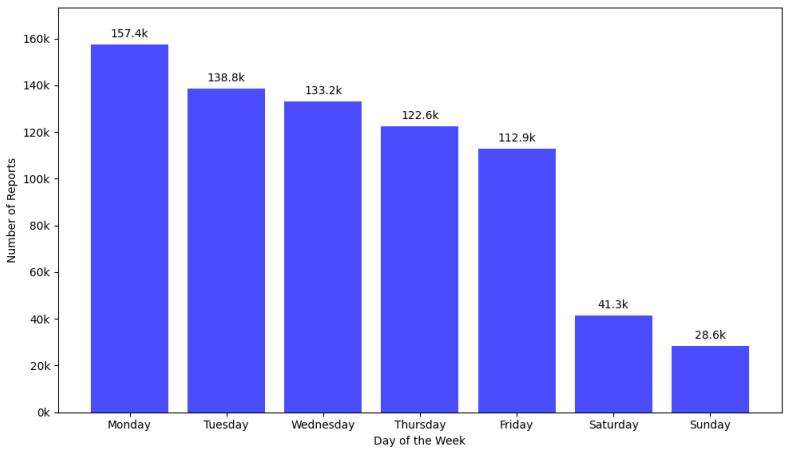

Figure 19 - Number of reports by day of the week

It can be noted that Mondays have the highest number of occurrences, with 157 425 reports. Throughout the week, the number of reports gradually decreases, with a more significant reduction at the weekend, with Sunday recording the lowest number of occurrences, with 28 560. This trend may indicate that at the beginning of the week, when citizens resume their activities, they are more attentive to urban problems, while at weekends activity decreases, possibly due to less movement and use of urban services.

In addition to analysing the various days that make up the week individually. It also proved important to create the *"Weekday Indicator"* and *"Holiday Indicator"* variables in order to understand the influence of the type of day on the number of occurrences reported. These indicators allow for a more detailed analysis of reporting patterns, helping to identify how urban activity and the availability of services vary between working days, weekends and public holidays, and how this affects the number of reports.

The *Table 12* shows the classification of occurrences into "*Weekday"* and "*Weekend"*. It can be seen that between 2018 and 2023 were recorded 664 919 occurrences on weekdays, with a daily average of 425 reports. At weekends, as observed in the previous visualization, the number of occurrences is significantly lower, with 69 903 reports and a daily average of 112.

Table 12 - Weekday Indicator

Weekday Indicator	Number of Days	Number of Reports	Average Reports per Day
Week Day	1 565	664 919	425
Weekend	626	69 903	112

This substantial difference can be attributed to the fact that during the week, there is more activity from both citizens and urban services, resulting in more occurrences being reported and dealt with. At weekends, activity decreases, reflected in the lower number of reports.

The *Table 13* shows occurrences classified as "*Not Holiday"* and "*Holiday"*. On non-holidays, between 2018 and 2023 were recorded 727 297 occurrences, with a daily average of 345 reports. On public holidays, the number of occurrences is much lower, with only 7 525 reports and a daily average of 90. The lower number of occurrences reported on holidays can be explained by the lower general activity in the city during these days when many services are closed or operating at reduced capacity, and fewer citizens are active in the city to report problems.

Table 13 - Holiday Indicator

Holiday Indicator	Number of Days	Number of Reports	Average Reports per Day
Not Holiday	2 107	727 297	345
Holiday	84	7 525	90

To conclude the temporal analysis, the distribution of the number of reports by period of the day was also analysed. The use of a variable that segments the day into different periods allows for a more detailed view of when citizens are most active in reporting occurrences.

The graph shown in *Figure 20* illustrates the total number of reports accumulated over the period from 2018 to 2023, broken down into the following periods: *dawn*, *morning*, *lunchtime*, *afternoon* and *evening*. It can be seen that the *morning* is the period with the highest number of accumulated reports, totalling 270 608 occurrences, followed by the *afternoon* with 236 429, then the *lunchtime* period with 114 213 occurrences. *Dawn*, as expected, saw the lowest number of reports, with just 11 463.

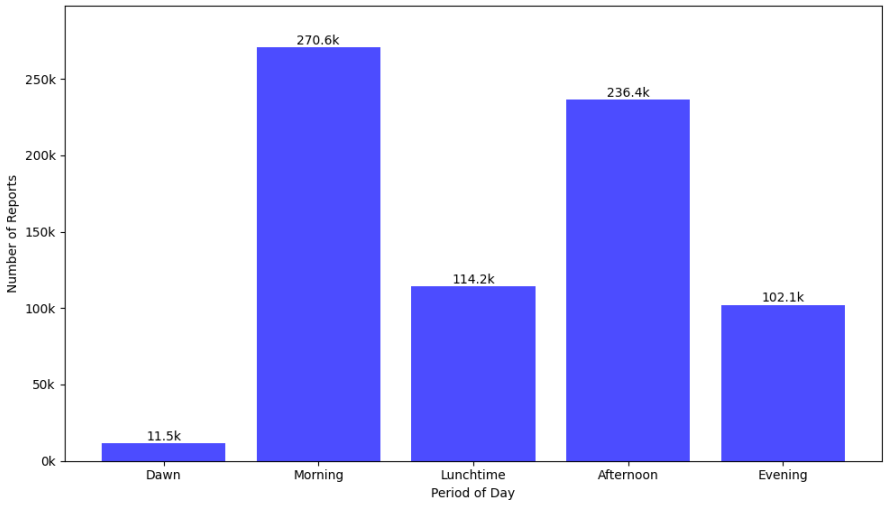

Figure 20 - Number of reports per period of day

Below in *Figure 21* is a second graph showing the average number of reports per hour for each period of the day. This graph adjusts the perspective by taking into account the length of the five periods, for example, lunchtime comprises fewer hours than the morning or afternoon. Here, it can be seen that the *afternoon* and *lunchtime* have the highest averages of reports per hour, with 59 107 and 57 107 respectively, standing out as the periods of the day with the highest intensity of records.

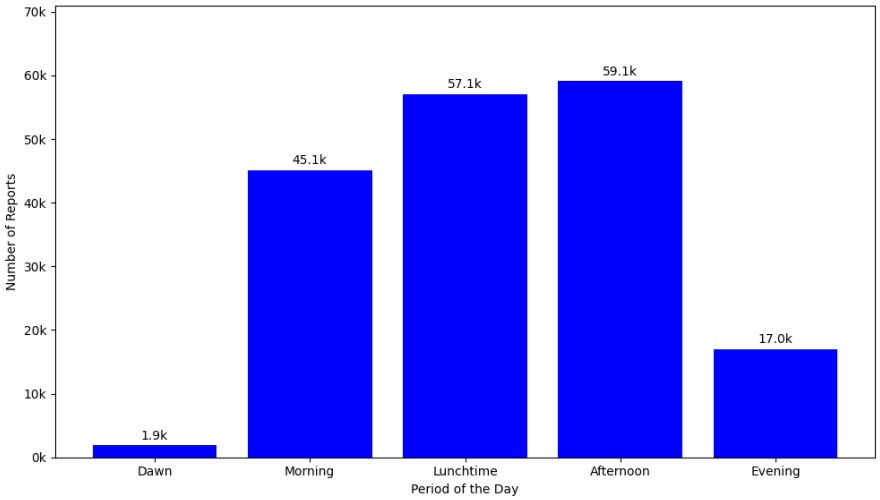

Figure 21 - Average number of reports per hour per period of the day

This is an important detail for urban management, as it indicates the periods of the day when citizens are most active in submitting occurrences.

4.2. Spatial Analysis of Reported Occurrences by parish

Now that we have concluded the chronological analysis of occurrences, we move on to a spatial analysis, focusing on specific geographical areas within the city of Lisbon. This step is crucial to understanding how the city's different parishes vary in terms of the number of occurrences reported. Through this analysis, we will be able to identify the parishes with the highest and lowest number of reports, as well as understand the patterns of spatial distribution of occurrences.

Figure 23 shows the number of occurrences reported by parish in the city of Lisbon. It can be observed that *Alvalade* is the parish with the highest number of occurrences, totalling 61 408 reports, followed by *Arroios* with 60 442. On the other hand, the parishes of *Carnide* and *Beato* recorded a much lower number of occurrences, 9 761 and 8 750 respectively.

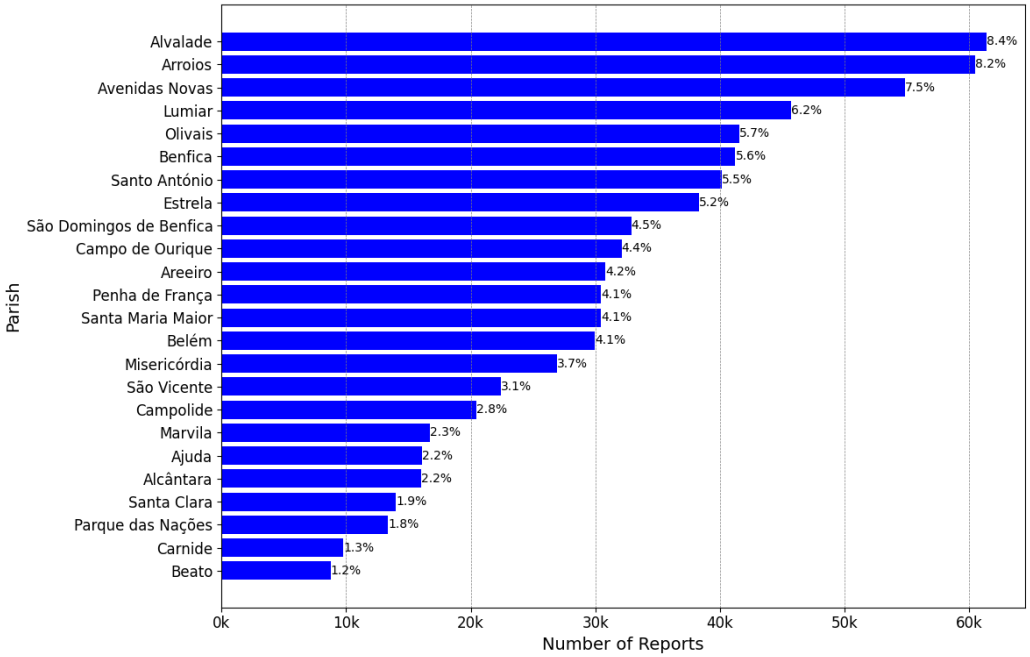

Figure 22 - Number of reports per parish

The shape map presented in *Figure 23* below shows that the more central parishes of the city, such as *Alvalade*, Arroios and *Avenidas Novas*, are the ones that show a greater predominance of occurrences than the surrounding parishes.

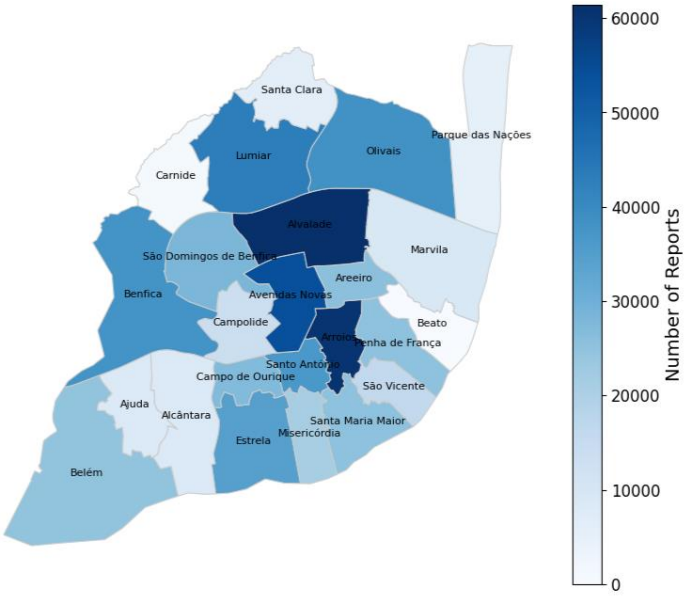

Figure 23 - Number of reports per parish (shape map)

After analysing the total number of reports per parish between 2018 and 2023, it is important to understand how the distribution of occurrences has evolved over these years. *Figure 24* below shows six shape maps illustrating the evolution of reports in different Lisbon parishes during the period under analysis.

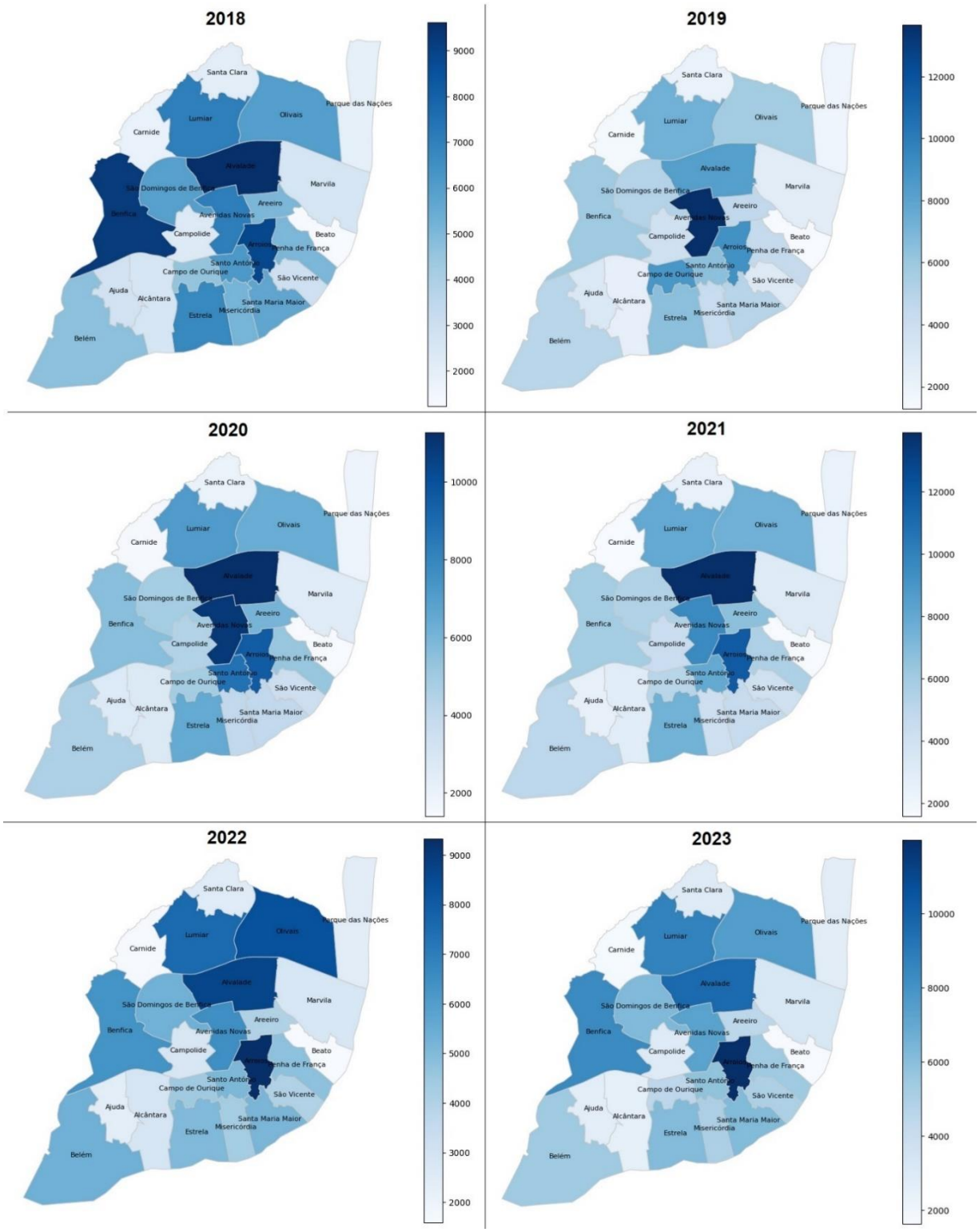

Figure 24 - Evolution of the number of reports per parish between 2018 and 2023 (shape map)

Analysing the reports over the years revealed some interesting trends and patterns, both in terms of parishes and categories of occurrences. Regarding the parishes that showed the highest incidence of reports:

- *Alvalade*: Consistently one of the parishes with the highest number of reports, standing out in 2021 with 13 892 occurrences. The main categories are *"Removal - Large Waste Collection Request"* and "*Rubble and objects abandoned on the public highway"*.

- *Avenidas Novas*: In 2019, it registered the highest number of reports, with 13 662 occurrences, dominated by requests for *"Removal - Large Waste Collection Request"*, which accounted for 67% of the total.

- *Arroios*: Shows a significant increase in occurrences, especially in 2021 and 2023. In 2023, it recorded 11 983 occurrences, with *"Removal - Large Waste Collection Request"* accounting for 28% and "*Rubble and objects abandoned on the public highway"* 11%.

- *Benfica* and *Olivais*: These parishes generally show higher than average occurrence values. In 2022, *Olivais* had 8 344 occurrences, with *"Removal - Large Waste Collection Request"* (22%) and "*Rubble and objects abandoned on the public highway"* (17%). The parish of *Benfica*, on the other hand, stood out in 2018 as the second parish with the most reports that year (9 202), with "*Rubble and objects abandoned on the public highway"* accounting for 25% of occurrences

Switching the focus to the main type of occurrences by parish, we have the following:

- *"Removal - Large Waste Collection Request"*: This is the dominant category each year in most parishes. In 2019, the parish of Avenidas Novas had the highest number of reports for this type of occurrence, accounting for 67% of the total.

- "*Rubble and objects abandoned on the public highway"*: This category is often the second most reported category. However, in some years it has even been the most reported category in the main parishes. This is the case in parishes such as *Alvalade* and *Benfica*. In 2021, *Alvalade* had 42% of occurrences in this category, and *Benfica* in 2018 with 25% of occurrences, surpassing *"Removal - Large Waste Collection Request"*.

Finally, some patterns and changes felt over the years are as follows:

- In *2018*, *Alvalade* and *Benfica* led the way with a significant number of occurrences related to both collection requests and rubble.

- In *2019*, *Avenidas Novas* stood out with a clear predominance of requests for *"Removal - Large Waste Collection Request"*.

- In *2020*, both *Alvalade* and *Avenidas Novas* showed a more even split between "*Rubble and objects abandoned on the public highway"* and *"Removal - Large Waste Collection Request"*, with a slight predominance of the first one.

- In *2021*, *Alvalade* registered the highest number of occurrences, with "*Rubble and objects abandoned on the public highway"* representing the majority of reports.

- In *2022* and *2023*, *Arroios* took the lead, with a significant mix of *"Removal - Large Waste Collection Request"* and "*Rubble and objects abandoned on the public highway"*, although collection requests were slightly more predominant.

4.3. Spatial Analysis of Reported Occurrences detailed by Statistical Section

After analysing and visualizing the distribution of occurrences by parish, it is now important to detail some of the situations observed. To do this, we used the "*Statistical Section*", which offers a more detailed breakdown of the parishes. This analysis allows for a more granular observation, which can help identify specific areas within a parish where there is a higher concentration of reported occurrences.

Looking at *Figure 25*, where we place the shape map of occurrences by parish side by side and compare it with the shape map of occurrences by section, we can identify certain variations.

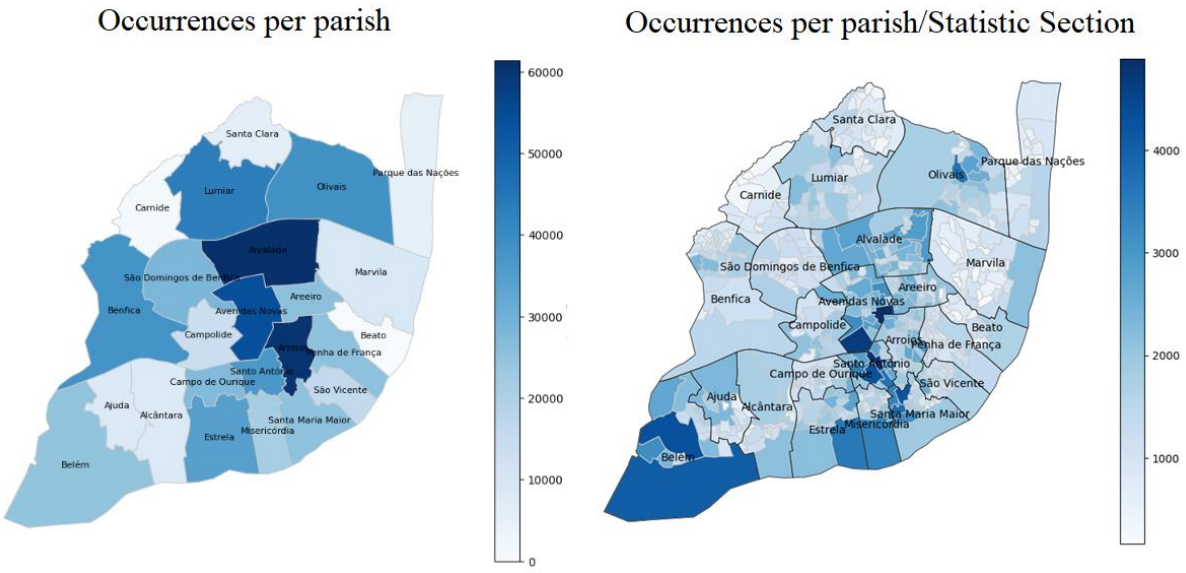

Figure 25 - Shape map of occurrences by parish and occurrences by Statistical Section

The two sections that stand out most in terms of the number of occurrences are the "*110657005"* section, belonging to the Avenidas Novas parish, more specifically the area of Saldanha, with 4 887 occurrences. Next, section "*110666012"*, in Santo António, more precisely at the beginning of Avenida da Liberdade with 4 811 occurrences. However, when we analyse the shape map of the sections, we notice that the areas with the highest incidence change. The more centralised areas lose prominence to the riverside regions. This visual difference between the two maps is strongly influenced by the number of sections per parish and the area that each section covers. Parishes like Alvalade, which has 31 sections, have a more detailed distribution, while Belém, with only 14 sections, covers spatially wider areas, which may explain some of these variations.

It is also important to show and understand which types of occurrences predominate in the sections with the highest incidence of reports, as displayed in *Figure 26*. This graph shows the sections with more than 4 000 occurrences and the four most frequent types of occurrences, making it possible to compare them.

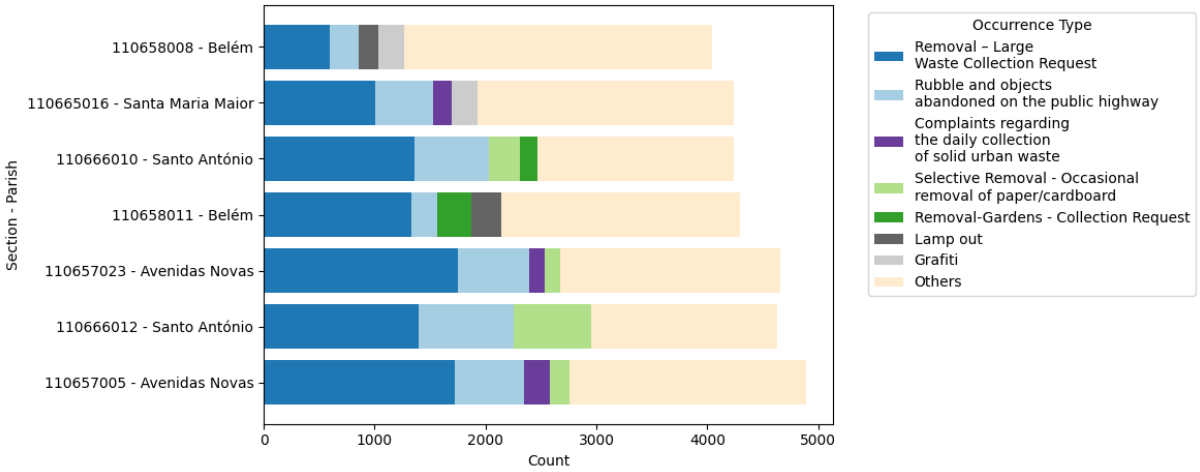

Figure 26 - Top typologies in the sections with the most occurrences

The graph shows the sections with the highest number of reports, highlighting the four most frequent types of occurrences in each section, allowing a direct comparison between them. In practically all sections, the *"Removal - Large Waste Collection Request"* (in dark blue) and "*Rubble and objects abandoned on the public highway"* (in light blue) are the most prevalent types of occurrences, registering considerable occurrences in all sections. However, when analysing this graph, it is possible to see significant differences in the distribution of these occurrences between the selected sections.

Now that we analysed the main sections of the city, let us focus again on the parishes with the highest number of reports, which are Alvalade and Arroios. Using the statistical sections, a more detailed analysis of the distribution of occurrences in these parishes can be performed. By making this more refined analysis, the aim is to identify the specific areas within these parishes that have the highest incidence of occurrences.

In the specific case of the parish of Alvalade, which stood out as the parish with the highest number of reports over the period analysed, it becomes relevant to understand which areas within this parish concentrate the majority of incidents. *Figure 27* illustrates the distribution of occurrences by statistical section within this parish.

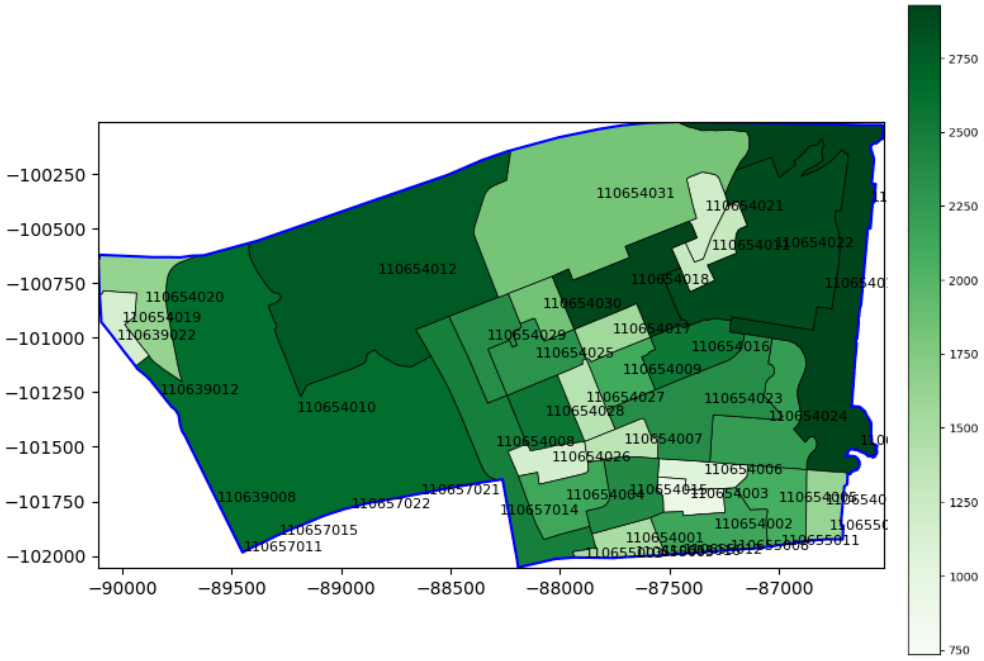

Figure 27 - Statistical Sections of Alvalade

In *Figure 27*, it can be seen that several sections of the parish of Alvalade have a high number of incidents. In particular, the sections near Avenida Almirante Gago Coutinho and Praça do Aeroporto, located in the top right-hand corner, where the incidence of reports is significantly higher. In addition, the Campo Grande area, located on the left-hand side of the image, also has a high number of occurrences, highlighting the importance of these areas in the borough's urban management.

Another parish that registered a high number of reports is the parish of Arroios, with the second highest number of occurrences. It is therefore important to deepen our analysis in this part of the city. *Figure 28* shows the occurrences in this parish, broken down by statistical section.

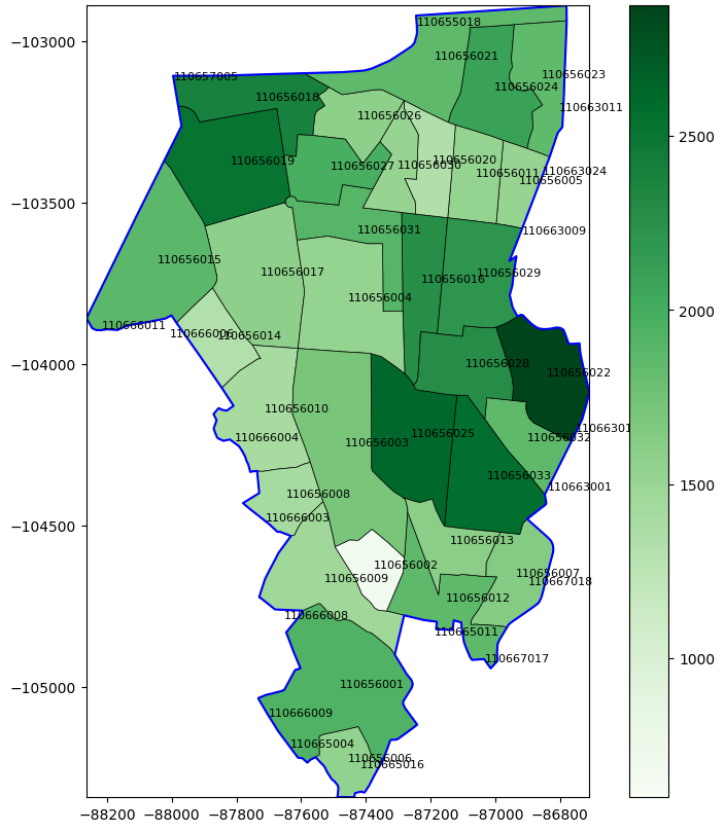

Figure 28 - Statistical Sections of Arroios

The shape map shows that the Anjos area stands out as the area with the highest incidence of reports. In particular, Monte Agudo viewpoint, located in this area, has a high concentration of occurrences, as indicated by the darker section in the image.

4.4. Analysis of Occurrence Classifications and Typology

After analysing the distribution of occurrences by parish, it is essential to look at another equally important topic, the classification of reported occurrences. This analysis allows us to better understand the types of problems most frequently reported by citizens and to identify the thematic areas that require greater attention from urban services.

Figure 29 shows the distribution of the number of occurrences in eight different thematic areas. "*Urban Hygiene"* stands out as the area with the highest number of occurrences, reflecting a very significant number of problems related to the cleaning and maintenance of urban areas. Other important areas include "*Public Lighting"* and "*Walkways and Accessibility"*, which also register a considerable number of reports, although much lower than the "*Urban Hygiene"* occurrences.

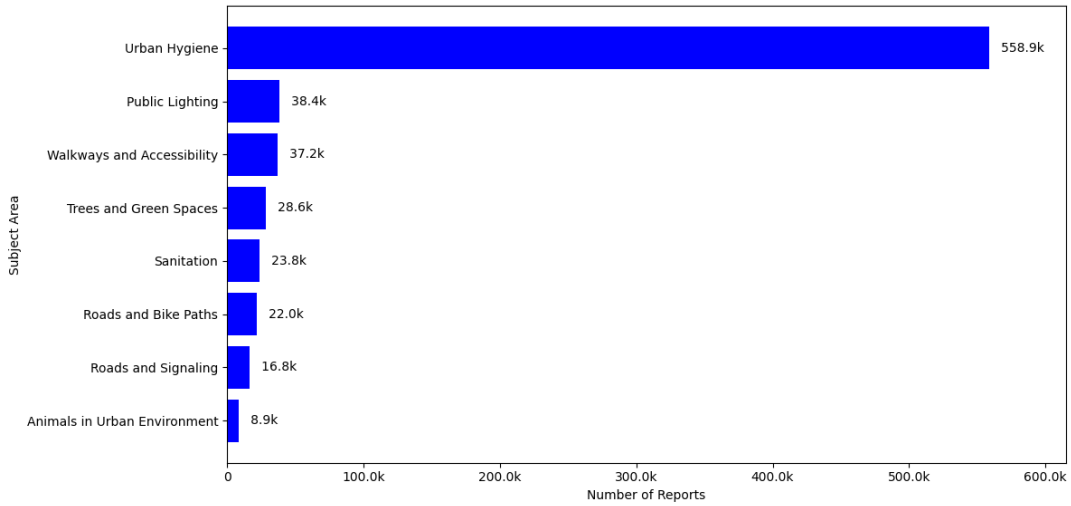

Figure 29 - Number of reports per subject area

As the "*Urban Hygiene"* area represents around 76% of the dataset, it is important to detail the occurrences within this area. The bar chart in *Figure 30* shows the 10 most frequent occurrence types in the "*Urban Hygiene*" area.

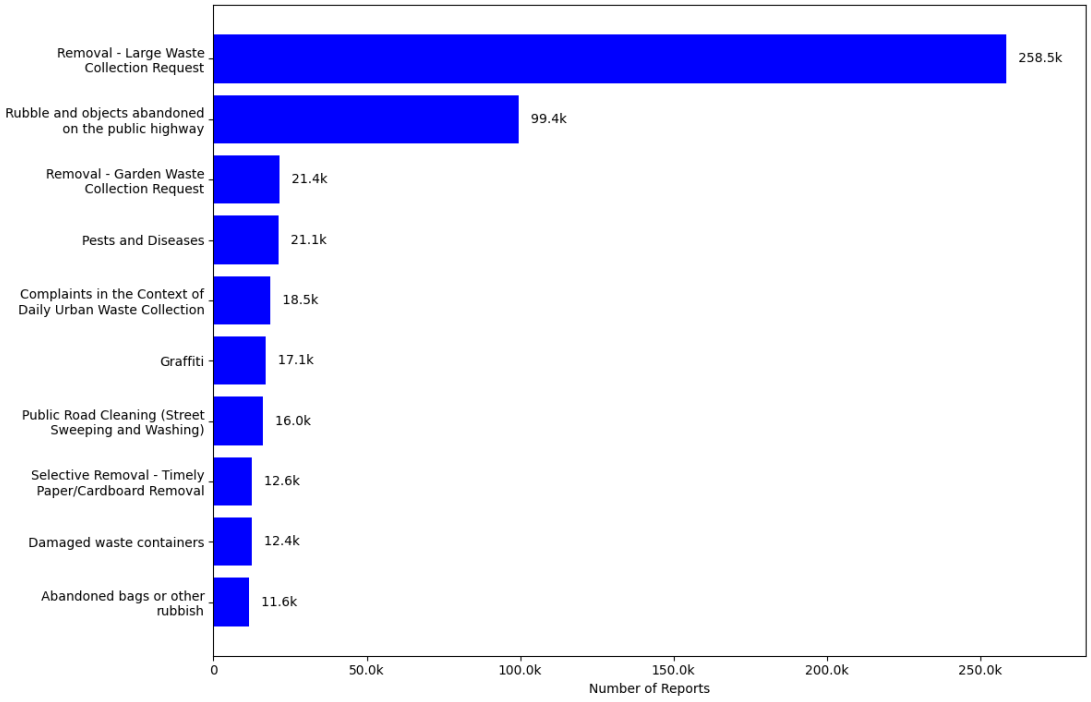

Figure 30 - Top 10 types of occurrences in Urban Hygiene

It can be observed that "*Removal-Large Waste-Collection Request*" is the most reported type, representing around 46% of the reports in the "*Urban Hygiene"* Area and 35% of the total dataset. This category concerns requests from citizens to collect bulky items, indicating a significant need for large waste removal services. Below in *Figure 31* and *Figure 32* we can see which parishes in the city have the highest incidence of this type of occurrence.

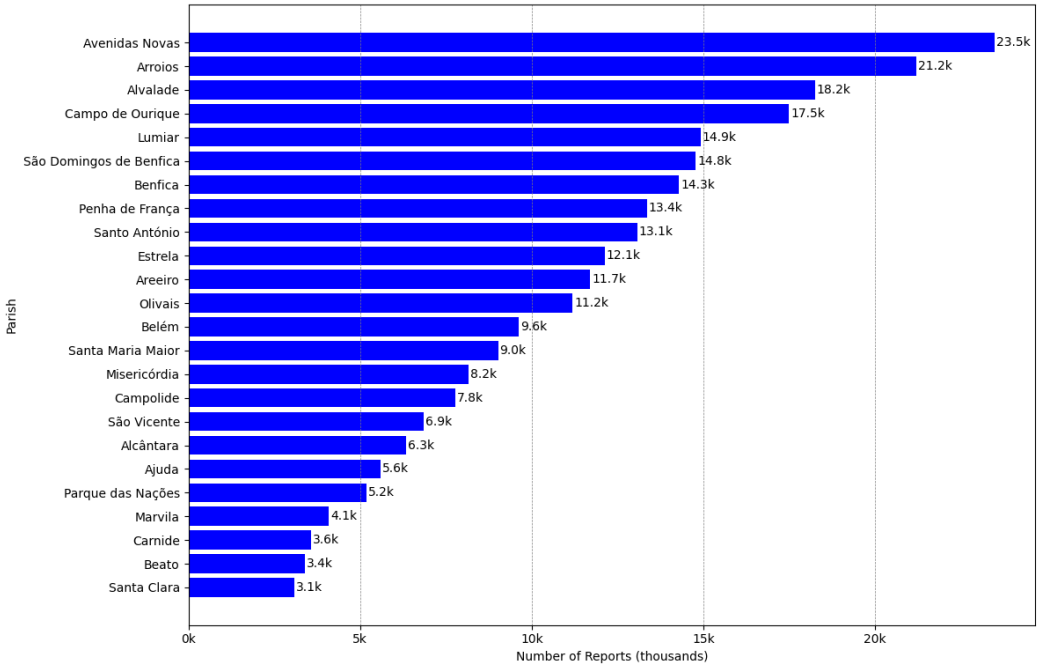

Figure 31 - Number of reports per parish for "Removal-Large Waste-Collection Request"

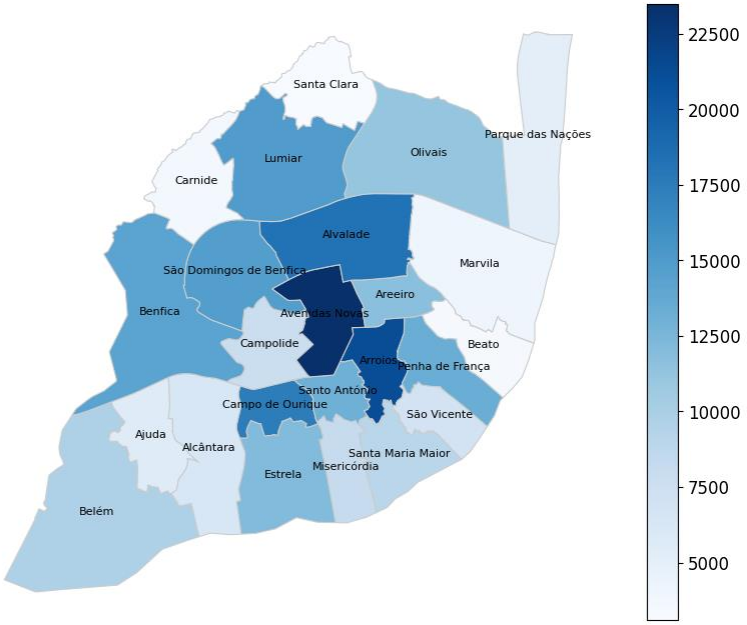

Figure 32 - Number of reports per parish for "Removal-Large Waste-Collection Request" (shape map)

The visualizations above illustrate the distribution of the number of reports by parish for the category "*Removal - Large Waste Collection Request*". It can be noted that *Avenidas Novas* has the highest number with 23 481 occurrences, followed by *Arroios* with 21 187 and *Alvalade* with 18 239 reports. Parishes such as *Campo de Ourique*, *Lumiar* and *São Domingos de Benfica* also had high numbers, indicating a greater need for bulky item removal services in these areas. In contrast, parishes such as *Santa Clara*, *Beato* and *Carnide* have significantly fewer occurrences reported. This distribution shows that more centralized parishes tend to have more occurrences of this type.

The second most frequent type is "*Rubble and objects abandoned on the public highway*", representing 18% of the "*Urban Hygiene"* Area and 15% of the total data. This type of occurrence reflects problems with the improper dumping of different types of waste on the public highway, presenting a challenge in urban waste management. For this reason, it is important to understand where in the city this type of occurrence is prevalent, and *Figure 33* and *Figure 34* provides us with that information.

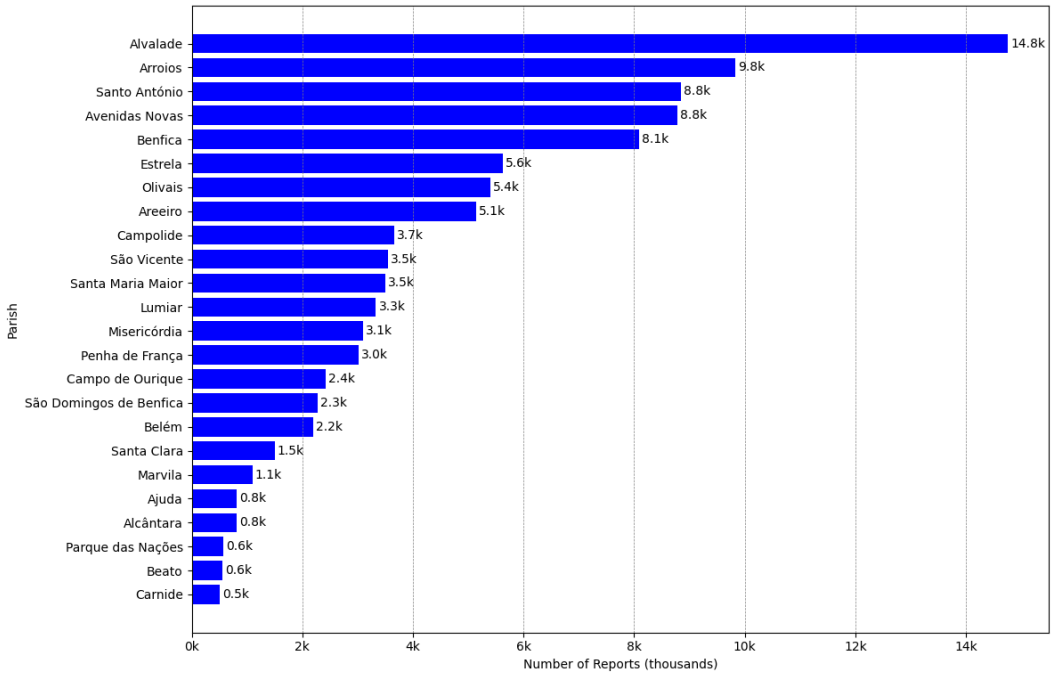

Figure 33 - Number of reports per parish for "Rubble and objects abandoned on the public highway"

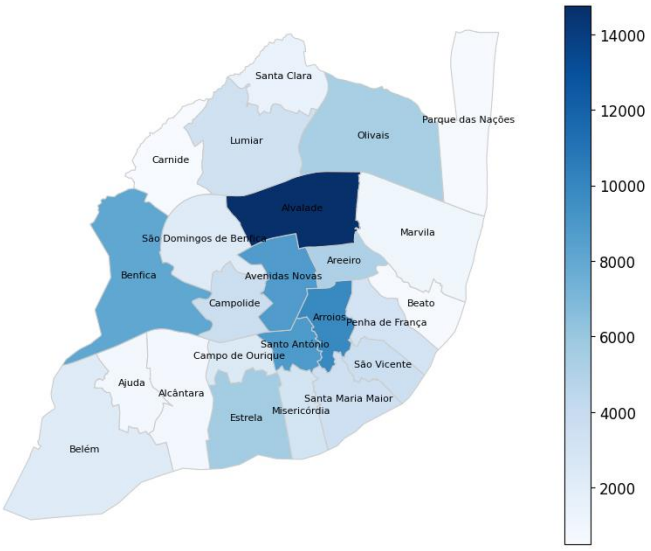

Figure 34 - Number of reports per parish for "Rubble and objects abandoned on the public highway" (shape map)

The images above show the distribution of the number of reports by parish for the category "*Rubble and objects abandoned on the public highway*". *Alvalade* stands out with 14 757 reports, much higher than the rest. This was followed by *Arroios* with 9 825 and *Santo António* and *Avenidas Novas*, both with 8 849 and 8 784 reports respectively. Parishes such as *Benfica*, *Estrela* and *Olivais* also have significant figures, indicating frequent problems with improper waste disposal in these areas. In contrast, parishes such as *Carnide*, *Beato* and *Parque das Nações* have a much smaller number of reports, suggesting a lower incidence of this type of problem. This distribution may reflect the variations in population density, waste disposal habits and the urban characteristics of each parish.

4.5. Analysis of Time Efficiency in the Resolution of Reported Occurrences

In addition to analysing reported occurrences and their temporal and spatial distribution, it is also crucial to assess the efficiency of urban services in resolving these occurrences. Analysing the average time it takes for events to be reported as resolved, offers a valuable insight into the effectiveness of the responses of the services responsible.

This analysis is made by using the "*Time Elapsed"* variable, which represents the time interval from when an occurrence is registered on the "*Na Minha Rua*" portal until it is classified as "*Resolved*". This indicator helps to assess the efficiency and effectiveness of the services responsible for resolving reported occurrences. In order to proceed with this analysis, only records with a "*Resolved*" status were selected, representing 96,1% of the dataset.

The graph presented in *Figure 35* illustrates the average resolution time per area of occurrence, measured in days. This graph makes it possible to identify which areas, on average, have the shortest resolution times and which take the longest to resolve.

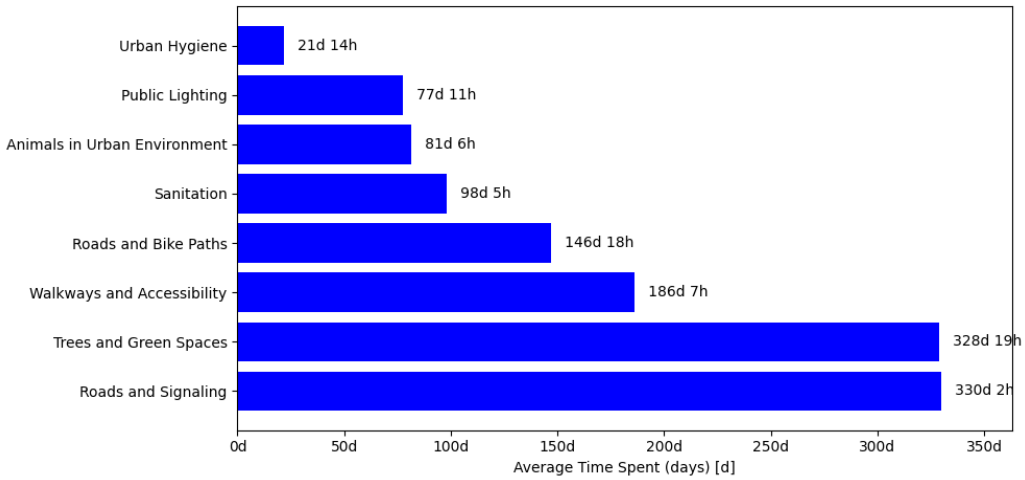

Figure 35 - Average resolution time per classification area

The "*Urban Hygiene*" class has a much lower average resolution time than the others, as can be seen in the previous graph. It is therefore important to understand whether this average resolution time is maintained across the various parishes, or whether there is a more marked variation in any area of the city. The *Figure 36* below helps us to better visualize this variation.

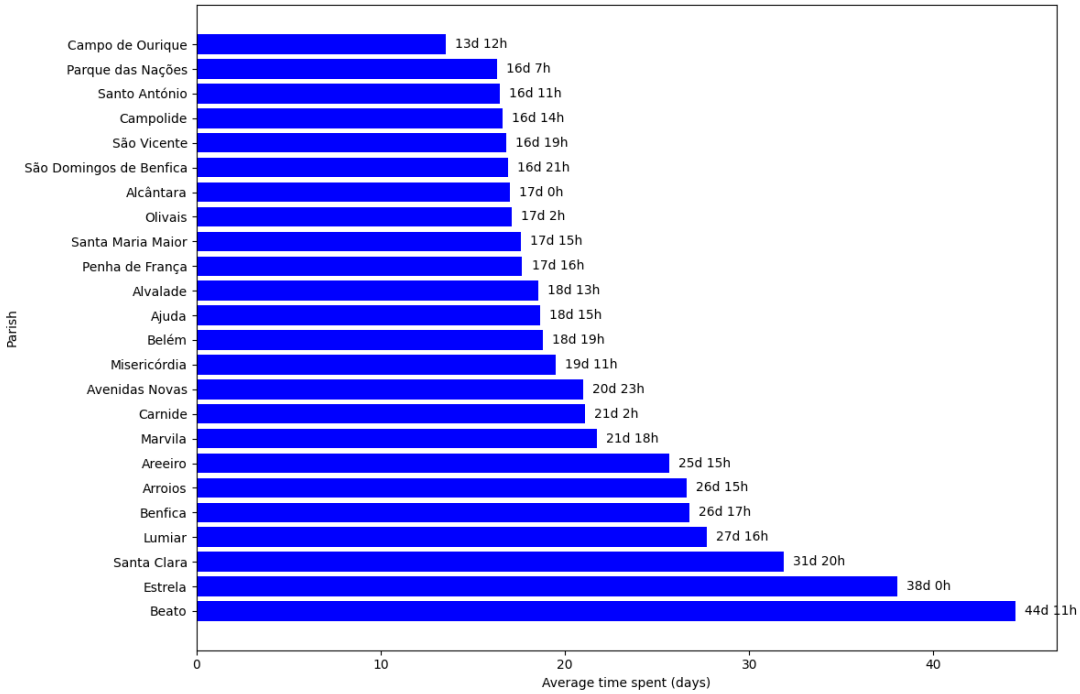

Figure 36 - Average resolution time per parish for Urban Hygiene

The graph shows the average time taken to resolve occurrences related to "*Urban Hygiene*" broken down by parish. There is a significant variation between the different parishes, indicating differences in the efficiency of the services responsible for resolving these problems. *Campo de Ourique* stands out as the parish with the shortest average resolution time, registering just 13 days and 12 hours. This may indicate a more efficient and rapid response from urban cleaning services in this specific area.

On the other hand, the parish of *Beato* has the longest average resolution time, with 44 days and 11 hours. This resolution time is substantially longer compared to *Campo de Ourique*, suggesting that urban hygiene problems in this parish take considerably longer to resolve. This variation can be attributed to various factors, including the availability of resources, the efficiency of local services and the complexity of the problems reported.

4.6. Correlation Between Socioeconomic Factors and Reported Occurrences

Finally, after analysing in detail the occurrences reported on the "*Na Minha Rua*" platform from different temporal and spatial perspectives, it is important to deepen our understanding of other variables that can influence the reporting of occurrences in the city. To this end, we will correlate the number of reports per parish with the socioeconomic indicators provided by the INE. From this correlation, we obtained the matrix shown in the *Figure 37* below, which illustrates the relationship between the various socioeconomic characteristics and the total number of occurrences reported by parish.

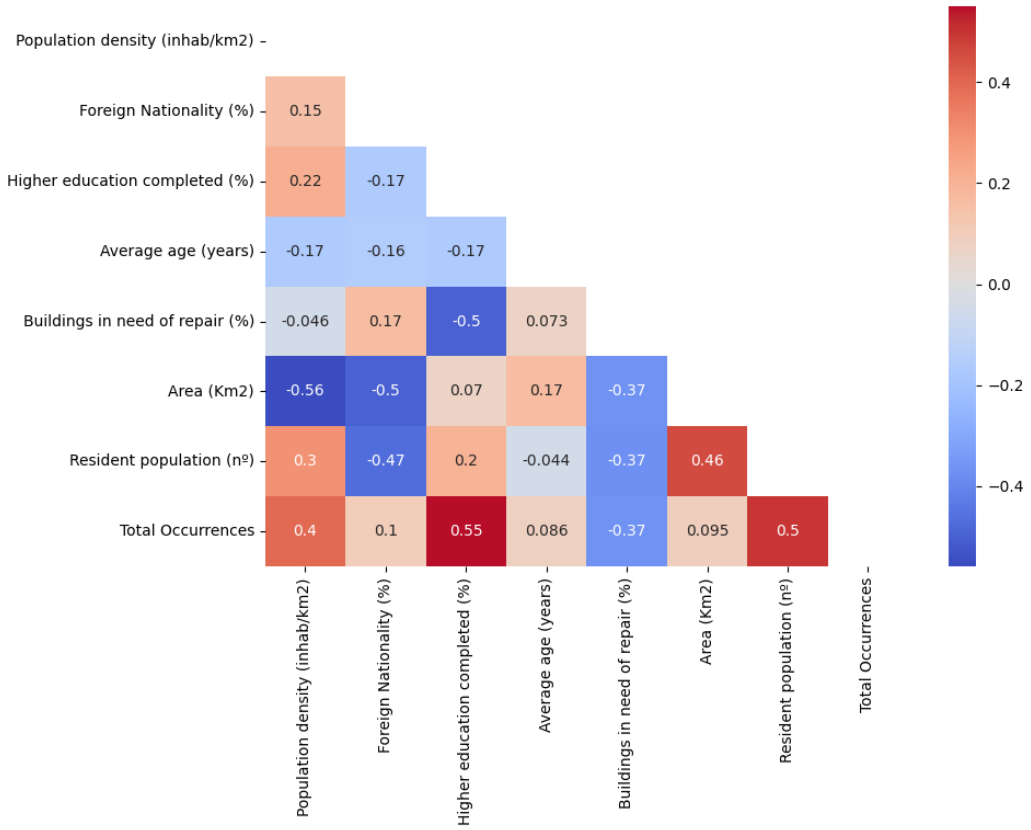

Figure 37 - Socioeconomic indicators correlation Matrix

Firstly, the total number of occurrences per parish was calculated, and then these figures were combined with the socio-economic data provided by INE. The variables considered in the analysis include population density, percentage of inhabitants with foreign nationality, percentage of inhabitants with complete higher education, average age, percentage of buildings in need of repair, area (km2) and the number of the resident population.

When analysing the correlation matrix, a positive correlation of 0,55 was found between the *percentage of the population with completed higher education* and the *total number of occurrences reported*. This result may indicate that citizens with a higher level of education are more likely to use this type of platform. There was also a positive correlation of 0,5 between the *number of inhabitants* and the *total number of occurrences*. This suggests that areas with more inhabitants tend to report more occurrences.

The scatterplots in *Figure 38* below help to visualize these correlations in more detail. A trend line, calculated using a simple linear regression model, has been added to the graphs to help visualize the relationship between the socioeconomic variable and the total number of occurrences.

The first graph shows that parishes like *Avenidas Novas* and *Alvalade*, with a high percentage of the population having completed higher education, have a high number of reported occurrences. In the second graph, parishes like *Arroios* and *Alvalade*, which have a higher number of inhabitants, also have a high number of occurrences.

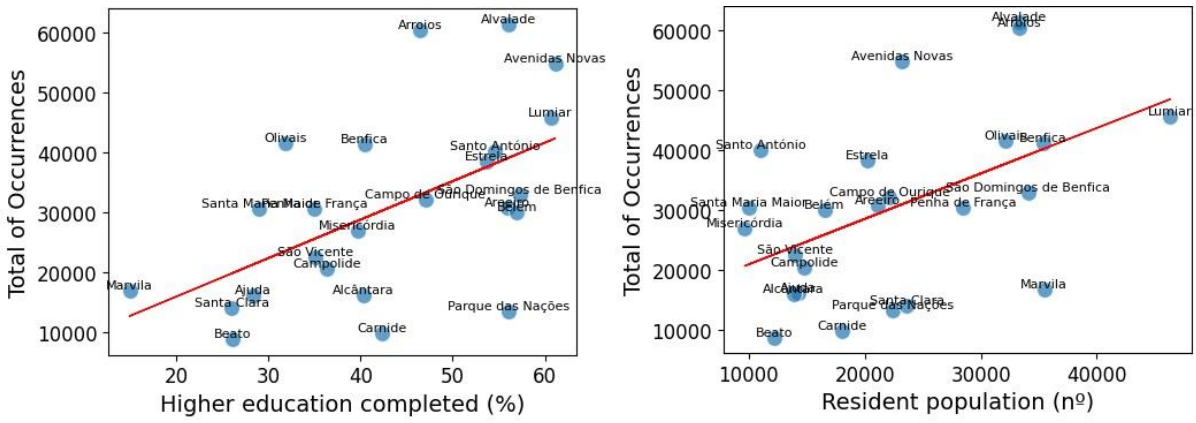

Figure 38 - 1- Scatterplot of Higher education completed per number of reports. 2- Scatterplot of Resident population per number of reports

The heatmap in *Figure 37* also showed a negative correlation of -0.37 between the *percentage of buildings in need of repair* and the *total number of occurrences* reported by parish. This pattern is visually reinforced by the scatterplot below in *Figure 39*.

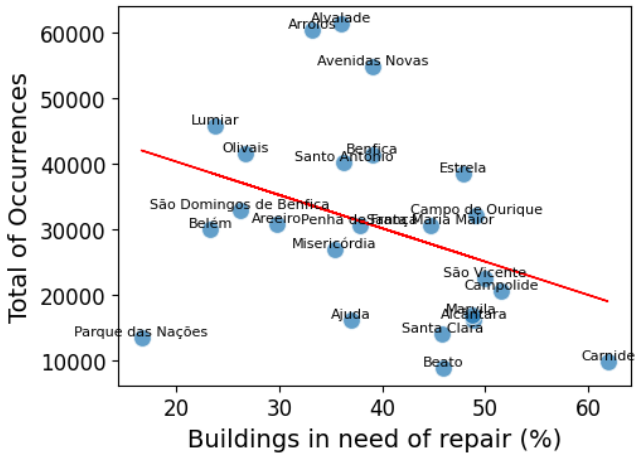

Figure 39 - Scatterplot of Buildings in need of repair per number of reports

It can be observed that parishes with a higher percentage of buildings in need of repair tend to report fewer occurrences. For example, parishes like *Carnide* and *Campolide*, which have a high percentage of buildings in need of repair, have a lower number of reported occurrences. On the other hand, parishes like *Alvalade* and *Arroios*, which have a lower percentage of buildings in need of repair, show a much higher number of reported occurrences.

Given the high correlation between the number of inhabitants per parish and the number of reported occurrences, it was important to find the average number of reports per inhabitants. *Figure 40* shows the number of reports per 1 000 residents in each Lisbon parish in 2023, ordered in descending order. This visualization shows the density of reported occurrences in relation to the population of each parish. The red dashed line on the graph represents the average of 257,84 reports per 1 000 residents, which serves as a benchmark for comparing the performance of each parish in relation to the city average.

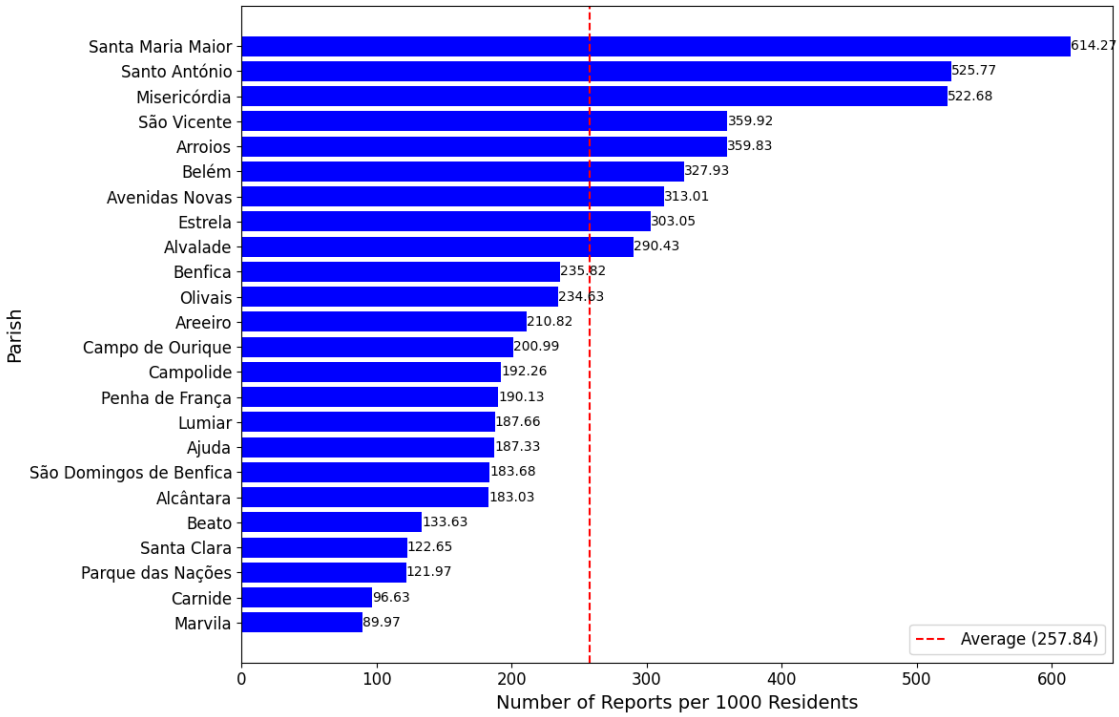

Figure 40 - Number of reports per 1 000 residents by parish

Santa Maria Maior stands out with the highest number of reports per capita, totalling 614,27 reports per 1 000 residents. This is followed by *Santo António* with 525,77 and *Misericórdia* with 522,68. These parishes, located in the historic centre of Lisbon, have a higher density of urban problems, probably due to the intense movement of people and activities. On the other hand, parishes such as *Marvila*, *Carnide* and *Parque das Nações* have lower occurrences per capita, with 89,97, 96,63 and 121,97 reports per 1 000 residents, respectively. This can be explained by the fact that these parishes are far from the city's historic centre, and therefore not as busy as the city's central areas.

This analysis shows a transition in the incidence of reports from the more central areas to the more historic parts of the city. Contrary to the previous analysis when we looked only at accumulated values over the years. The two images in *Figure 41* highlight this differentiation, showing on the left the incidence of reports per parish, and on the right the incidence of reports per 1 000 inhabitants per parish.

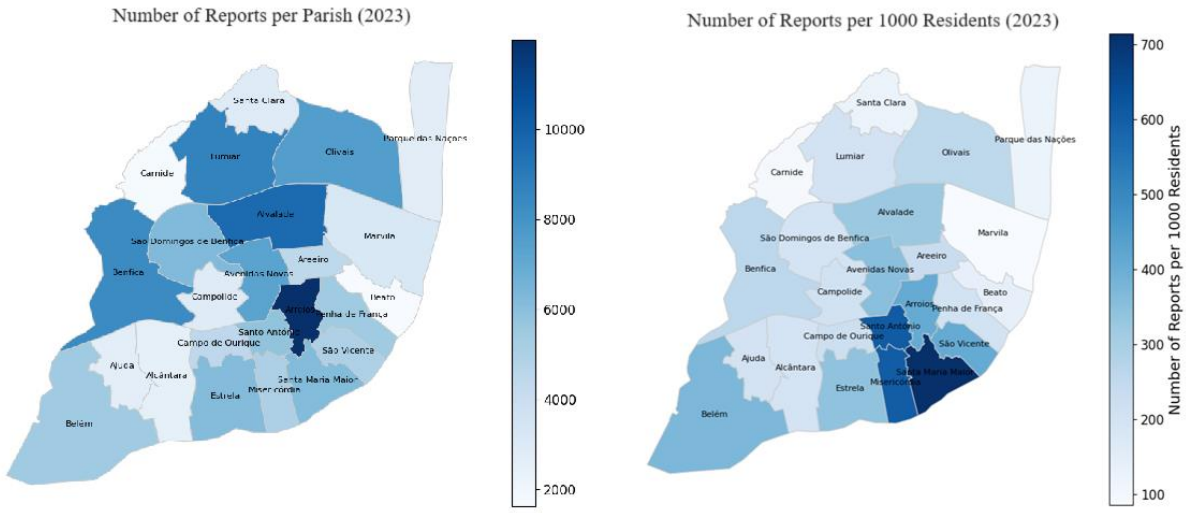

Figure 41 - Difference between the total number of reports by parish in 2023 and the number of reports per 1 000 residents per parish in 2023

Figure 42 shows the five most frequent types of reports in the parishes of *Santo António*, *Santa Maria Maior* and *Misericórdia*, which are the parishes with the highest number of reported occurrences per 1 000 inhabitants.

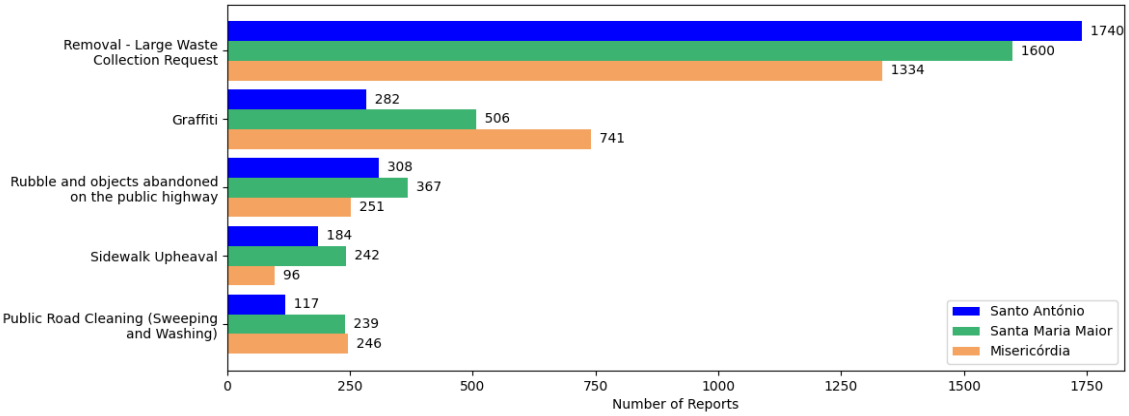

Figure 42 - Top 5 occurrence types in Santo António, Santa Maria Maior and Misericórdia

The "*Removal-Bulky items - Collection Request*" is clearly the most reported occurrence in the tree parishes analysed, with Santo António having the highest number of reports, registering 1 740 occurrences, followed by *Santa Maria Maior* and *Misericórdia* with 1 600 and 1 334 occurrences respectively, which highlights a common concern among these urban areas. "*Graffiti*", on the other hand, shows a contrary spatial incidence, being more frequent in the parish of *Misericórdia* with 741 records, *Santa Maria Maior* with 506 and the parish of *Santo António* with only 282 occurrences, showing a strong contrast with the previous typology. Another relevant type of occurrence is "*Rubble and objects abandoned on the public highway*", with this type of occurrence being more balanced between the three parishes. This type of occurrence, together with waste removal, suggests an ongoing need for urban cleaning and waste management services in these parishes.

Analysing these occurrences makes it possible to identify local priorities, helping to allocate resources more efficiently to solve the most prevalent urban problems in each parish.

4.7. A Proposal for a Dashboard for Exploratory Analysis

We considered it was important to include a data visualization tool on the "*Na Minha Rua*" platform to facilitate the monitoring and analysis of reported incidents. To this end, we developed the dashboard shown in *Figure 43* as a suggestion for those responsible for the platform, providing an overview of occurrences.

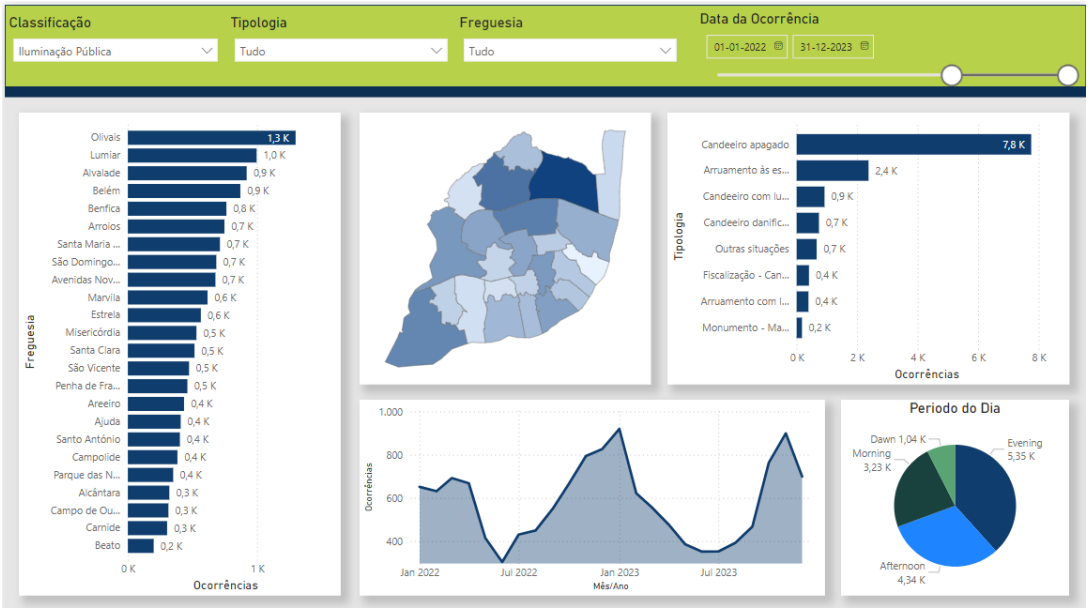

Figure 43 – Dashboard example

At the top of the dashboard is a filter bar that allows the user to customise the data displayed. The filters include the classification of the occurrence, the type, the parish and the time interval, enabling a targeted analysis according to the user's needs. Below, the first graph on the left shows the distribution of occurrences by parish, offering a spatial view of the areas with the most reports. This graphic helps identify the neighbourhoods that require the most attention and resources to resolve problems. On the right, a bar chart details the most frequent types of occurrences. This graph highlights the most common types of problems, allowing managers to prioritise the resolution of the most recurrent ones. At the bottom, the line graph shows the evolution of occurrences over time, allowing trends and monthly variations to be observed, which can be useful for identifying seasonal patterns. The circular graph on the right divides occurrences by time of day, indicating the times with the highest volume of reports, making it easier to manage resources throughout the day.

This dashboard suggestion aims to make accessing and analysing data more intuitive, improving efficiency in decision-making and in managing occurrences reported on the platform.

CHAPTER 5

Conclusions

5.1. Contributions

The objectives defined at the beginning of this work are to investigate the geographical distribution of occurrences, identify and analyse temporal patterns, and explore the nature and typology of the most frequent occurrences in order to provide a comprehensive understanding of the issues affecting the city of Lisbon. This study achieved its objectives by offering a detailed analysis of the occurrences reported through the "*Na Minha Rua*" application. It enabled not only the identification of spatial and temporal patterns of occurrences but also provided a deeper understanding of the nature and typology of the most frequent problems impacting the city.

An important aspect of this work is the inclusion of socioeconomic data from INE, which was not part of the original dataset provided by *LxDataLab*. By integrating these additional data, the study is able to offer more insights into how socioeconomic factors may influence the distribution and nature of urban occurrences, enhancing the overall comprehension of the problems faced by Lisbon's inhabitants. Furthermore, another significant contribution of this work is the methodological approach followed in the dissertation. The process of data preparation and enrichment serves as a replicable framework that can be applied to similar challenges, providing a methodology for future studies in other cities or contexts.

To fulfil its objectives, this work aimed to answer two research questions. To answer the first research question, "*How are these occurrences spatially distributed across different areas of Lisbon*?", a spatial analysis of the reported occurrences was carried out, considering the different parishes and statistical sections of Lisbon. Through data visualization and the use of shape maps, it was possible to identify the spatial distribution of occurrences over time. The results show significant variation in the number and type of incidents between parishes and statistical sections, indicating that certain areas of the city are more prone to specific urban problems.

For example, it was noted that parishes with higher population density, such as Arroios and Areeiro, report a greater number of incidents, especially related to "*Urban Hygiene*" and infrastructure maintenance. This suggests that population pressure directly influences the occurrence of certain issues, providing valuable data for resource allocation by municipal authorities. These spatial patterns reveal that the distribution of incidents is not uniform across the city but rather shaped by factors such as population density, sociodemographic characteristics, and infrastructure conditions.

In response to the second research question, "*What perceptible patterns emerge when analysing occurrence reports on a spatial and temporal level?*", the study identified both spatial and temporal patterns in the reported occurrences. The temporal analysis revealed seasonal and daily variations, highlighting that certain periods, such as the summer, experience an increase in "*Pest and Diseases*" related occurrences. This is likely due to the favourable weather conditions for the development of these issues. Additionally, weekdays recorded a higher number of incidents compared to weekends, and most reports were logged during working hours, reflecting the relationship between daily urban activity and the frequency of reported problems.

Regarding the typology of occurrences, this study revealed that issues related to "Urban Hygiene", such as the request to collect large amounts of waste, dominate the number of reports by a wide margin. This information is important for city managers, as it points to the need for a continuous focus and provision of adequate resources for maintaining urban cleanliness, one of the most visible and significant aspects of the quality of life in the city.

Overall, by achieving the proposed objectives, this work confirms that the analysis carried out provide a solid basis for more informed and effective decision-making by Lisbon's municipal authorities. By mapping and analysing in detail the occurrences reported by citizens, this study contributes to a clearer understanding of the urban challenges faced by the city. The ability to anticipate problems and allocate resources more efficiently can be significantly improved based on the insights generated, thus promoting a cleaner, safer and more efficiently managed city.

5.2. Research limitations

During the course of this study, some limitations were identified that impacted on the analysis carried out and the interpretation of the results. The most significant of these limitations was the lack of detail in the available data, which restricted the ability to carry out a more comprehensive and accurate analysis.

One of the main flaws in the detail of the data was the lack of clear specification of the organisations responsible for resolving the incidents. Although some incidents were attributed to external organisations, it was not clear who these organisations were and exactly what their role was in resolving the problems. In addition, it was not made clear whether the conclusion of the events was reported by the users themselves or by the entities responsible at the time of resolution, which could greatly affect the temporal recording of the final status of the events.

Another relevant limitation is related to the quality of the data, particularly in terms of redundancy and inconsistency in the types of occurrences. During the analysis, several situations were identified in which similar or duplicate typologies complicated the categorisation and interpretation of the data. Additionally, when exploring data with Business Intelligence (BI) tools, it is crucial to have well-defined hierarchies to facilitate analysis. For example, in this context, a clear hierarchy between categories, types and subtypes of occurrences would be necessary to ensure proper data curation and organisation. These limitations largely stem from the nature of the data provided, which, in its current form, lacks the necessary structure and consistency for seamless BI exploration. As a result, part of the challenges faced during this work can be attributed to the quality and structure of the data supplied.

The lack of a more precise and detailed data structure limited the ability to draw more accurate conclusions about the management of urban occurrences. For future research, it would be beneficial to improve the collection and categorisation of data, including more detailed information on the entities responsible, clarifying the process of resolving incidents and eliminating redundancies in the occurrence typologies.

5.3. Future Work

Although this study has achieved its main objectives and provided insights into urban occurrences in Lisbon, there are several opportunities to improve and expand this work in future research and practical implementations.

One of the main directions for future work is the integration of more external databases and indicators that can enrich the analysis of occurrences. For example, the inclusion of meteorological data from IPMA would make it possible to investigate how different weather conditions may influence the type and frequency of reported incidents. Another approach would be to explore the various variables[5] present in the *Geopackage* file of the statistical sections provided by INE. By integrating these variables, it would be possible to establish correlations between data from the "Na Minha Rua" platform and a much wider range of socio-economic indicators, significantly extending the analysis beyond the variables used in this study. This could provide a more detailed understanding of the external factors that contribute to urban problems.

In addition, a proposed dashboard has already been developed and is presented in *Figure 43* of the previews chapter. The next step would be to work in collaboration with *LxDatalab* to build and refine this dashboard, ensuring it is optimized for integration into the "Na Minha Rua" platform. This dashboard would offer city managers and the general public a visual and intuitive tool for monitoring occurrences in real time. With this tool, it would be possible to visualize the spatial distribution of occurrences, analyse temporal patterns, monitor the status of resolutions, and even issue alerts about areas with a high concentration of reports.

Alongside these possible improvements, the integration of machine learning techniques could be explored to predict the occurrence of certain types of problems in specific areas, based on historical patterns and other contextual variables, providing an additional tool to the urban management process.

[5] Examples of variables: Number of classical buildings, number of students, number of vacant dwellings, etc.

References

[1] M. N. I. Sarker, M. N. Khatun, G. M. M. Alam, and M. S. Islam, 'Big Data Driven Smart City: Way to Smart City Governance', in *Int. Conf. Comput. Inf. Technol., ICCIT*, Institute of Electrical and Electronics Engineers Inc., 2020.

[2] U. Nations, '2018 Revision of World Urbanization Prospects', United Nations. Accessed: May 04, 2024. [Online]. Available: https://www.un.org/en/desa/2018-revision-world-urbanization-prospects

[3] Britannica, 'Urbanization - Industrial Revolution, Population, Infrastructure'. Accessed: May 04, 2024. [Online]. Available: https://www.britannica.com/topic/urbanization/Impact-of-the-Industrial-Revolution

[4] I. Turok and G. McGranahan, 'Urbanization and economic growth: the arguments and evidence for Africa and Asia', *Environ. Urban.*, vol. 25, no. 2, pp. 465–482, 2013.

[5] A. Bousios, D. Gavalas, and L. Lambrinos, 'CityCare: Crowdsourcing daily life issue reports in smart cities', in *Proc. IEEE Symp. Comput. Commun.*, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 266–271. doi: 10.1109/ISCC.2017.8024540.

[6] H. Madduri, 'A Smart Road Maintenance System for Cities-An Evolutionary Approach', in *Innovative Technologies in Management and Science*, Springer, 2015, pp. 43–56.

[7] G. Pereira, G. Eibl, P. Parycek, and ACM, 'The Role of Digital Technologies in Promoting Smart City Governance: The Case of SmartGov', presented at the COMPANION PROCEEDINGS OF THE WORLD WIDE WEB CONFERENCE 2018 (WWW 2018), 2018, pp. 911–914. doi: 10.1145/3184558.3191517.

[8] M. do R. M. Bernardo, 'Smart governance em cidades inteligentes europeias', in *CISTI 2019. 14th Iberian Conference on Information Systems and Technologies*, Institute of Electrical and Electronics Engineers, 2019. Accessed: May 01, 2024. [Online]. Available: https://repositorioaberto.uab.pt/handle/10400.2/8943

[9] J. Wang, D. Q. Nguyen, T. Bonkalo, and O. Grebennikov, 'Smart governance of urban data', in *E3S Web Conf.*, Torre A., Martinat S., Kumar V., Lavrikova Y., and Kuzmin E., Eds., EDP Sciences, 2021. doi: 10.1051/e3sconf/202130105005.

[10] World Bank, 'Overview', World Bank. Accessed: May 05, 2024. [Online]. Available: https://www.worldbank.org/en/topic/urbandevelopment/overview

[11] Câmara Municipal de Lisboa, 'naminharuaLX GOPI Gestão e Pedidos de Intervenção de Ocorrências em Lisboa'. Accessed: May 04, 2024. [Online]. Available: https://naminharualx.cm-lisboa.pt/

[12] W. Y. Ayele, 'Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset', *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020, doi: 10.14569/IJACSA.2020.0110603.

[13] R. Wirth and J. Hipp, 'CRISP-DM: Towards a standard process model for data mining', *Proc. 4th Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, Jan. 2000.

[14] D. Rodríguez-García, V. García-Díaz, and C. González García, 'Crowdsl: Platform for incidents management in a smart city context', *Big Data Cogn. Comput.*, vol. 5, no. 3, 2021, doi: 10.3390/bdcc5030044.

[15] J. Zhang and D. Wang, 'Duplicate report detection in urban crowdsensing applications for smart city', presented at the Proceedings - 2015 IEEE International Conference on Smart City, SmartCity 2015, Held Jointly with 8th IEEE International Conference on Social Computing and Networking, SocialCom 2015, 5th IEEE International Conference on Sustainable Computing and Communications, SustainCom 2015, 2015 International Conference on Big Data Intelligence and Computing, DataCom 2015, 5th International Symposium on Cloud and Service Computing, SC2 2015, 2015, pp. 101–107. doi: 10.1109/SmartCity.2015.54.

[16] P. F. Tehrani, S. Pfennigschmidt, U. Kriegel, A. Billig, F. Fuchs-Kittowski, and U. Meissen, 'Multidimensional report analysis in urban incident management', presented at the Proceedings of the 2017 4th International Conference on Information and Communication Technologies for Disaster Management, ICT-DM 2017, 2017, pp. 1–8. doi: 10.1109/ICT-DM.2017.8275689.
[17] J. Tadili and H. Fasly, 'Citizen participation in smart cities: A survey', presented at the ACM International Conference Proceeding Series, 2019. doi: 10.1145/3368756.3368976.
[18] I. M. Ariya Sanjaya, S. H. Supangkat, and J. Sembiring, 'Citizen Reporting Through Mobile Crowdsensing: A Smart City Case of Bekasi', presented at the Proceeding - 2018 International Conference on ICT for Smart Society: Innovation Toward Smart Society and Society 5.0, ICISS 2018, 2018. doi: 10.1109/ICTSS.2018.8549976.
[19] A. Dhini, I. B. N. S. Hardaya, and I. Surjandari, 'Clustering and visualization of community complaints and proposals using text mining and geographic information system', presented at the Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017, 2017, pp. 132–137. doi: 10.1109/ICSITech.2017.8257098.
[20] M. L. Rethlefsen *et al.*, 'PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews', *Syst. Rev.*, vol. 10, pp. 1–19, 2021.
[21] S. Elsevier, 'Scopus', *Rev. Áudio E Base Dados*, vol. 1, pp. ID8–ID8, 2024.
[22] Clarivate, 'Web of Science: Citing Web of Science data'. Accessed: May 26, 2024. [Online]. Available: https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Citing-Web-of-Science-data?language=en_US
[23] BMJ, 'PRISMA 2020', PRISMA statement. Accessed: May 04, 2024. [Online]. Available: https://www.prisma-statement.org/prisma-2020
[24] A. Boumchich, J. Picaut, and E. Bocher, 'Using a Clustering Method to Detect Spatial Events in a Smartphone-Based Crowd-Sourced Database for Environmental Noise Assessment', *Sensors*, vol. 22, no. 22, 2022, doi: 10.3390/s22228832.
[25] J. Brus, J. Vrkoč, and M. Kubásek, 'Design of decision support tools for the quality assessment of illegal dumping notifications based on crowd-sourced data', presented at the Environmental Modelling and Software for Supporting a Sustainable Future, Proceedings - 8th International Congress on Environmental Modelling and Software, iEMSs 2016, 2016, p. 877. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065050319&partnerID=40&md5=a0e64d88cda02732baf78cfae883b2d7
[26] T. Xanthopoulos, T. Anagnostopoulos, C. Kytagias, and Y. Psaromiligkos, 'A smartphone-enabled crowdsensing and crowdsourcing system for predicting municipality resource allocation stochastic requirements', presented at the ACM International Conference Proceeding Series, 2020, pp. 305–310. doi: 10.1145/3437120.3437330.
[27] S. Yosua Grandy Ara and A. Suzianti, 'Analysis of technology adoption for real-time aspiration delivery system', presented at the ACM International Conference Proceeding Series, 2017, pp. 516–520. doi: 10.1145/3162957.3162983.
[28] S. H. Supangkat, R. Ragajaya, and A. B. Setyadji, 'Implementation of Digital Geotwin-Based Mobile Crowdsensing to Support Monitoring System in Smart City', *Sustain. Switz.*, vol. 15, no. 5, 2023, doi: 10.3390/su15053942.
[29] Y. Kaluarachchi, 'Implementing Data-Driven Smart City Applications for Future Cities', *Smart Cities*, vol. 5, no. 2, pp. 455–474, 2022, doi: 10.3390/smartcities5020025.
[30] F. De Filippi *et al.*, 'MiraMap: A We-Government Tool for Smart Peripheries in Smart Cities', *IEEE Access*, vol. 4, pp. 3824–3843, 2016, doi: 10.1109/ACCESS.2016.2548558.
[31] A. C. Pistolato and W. C. Brandão, 'Connectcity: A collaborative e-government approach to report city incidents', presented at the Proceedings of the 15th International Conference WWW/Internet 2016, 2016, pp. 233–237. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85010991317&partnerID=40&md5=814f885482794a837de469eff0b4da50

[32] S. Ruiz-Correa *et al.*, 'SenseCityVity: Mobile Crowdsourcing, Urban Awareness, and Collective Action in Mexico', *IEEE Pervasive Comput.*, vol. 16, no. 2, pp. 44–53, 2017, doi: 10.1109/MPRV.2017.32.
[33] Leiden University, 'VOSviewer - Visualizing scientific landscapes', VOSviewer. Accessed: Apr. 21, 2024. [Online]. Available: https://www.vosviewer.com//
[34] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, 'Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature', *PloS One*, vol. 12, no. 3, p. e0172778, 2017.
[35] R. Wirth and J. Hipp, 'CRISP-DM: Towards a standard process model for data mining',
[36] Instituto Nacional de Estatistica, 'Portal do INE'. Accessed: May 12, 2024. [Online]. Available: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE&xlang=pt
[37] 'Densidade populacional'. Accessed: May 11, 2024. [Online]. Available: https://www.pordata.pt/municipios/densidade+populacional-452
[38] L. Aberta, 'LxDataLab', LISBOA ABERTA. Accessed: Apr. 28, 2024. [Online]. Available: https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/apresentacao
[39] 'Censos 2021'. Accessed: Aug. 27, 2024. [Online]. Available: https://censos.ine.pt/xportal/xmain?xpgid=censos21_main&xpid=CENSOS21&xlang=pt
[40] '4. Built-in Types — Python 3.4.10 documentation'. Accessed: Aug. 27, 2024. [Online]. Available: https://docs.python.org/3.4/library/stdtypes.html
[41] 'Welcome to Python.org', Python.org. Accessed: Aug. 27, 2024. [Online]. Available: https://www.python.org/
[42] M. Melo, 'Estações do ano em Portugal: Conheça o clima no país', IE Intercâmbio no Exterior. Accessed: Aug. 27, 2024. [Online]. Available: https://www.ie.com.br/intercambio/estacoes-ano-portugal/
[43] 'Python strftime() - datetime to string'. Accessed: Aug. 27, 2024. [Online]. Available: https://www.programiz.com/python-programming/datetime/strftime
[44] 'pandas - Python Data Analysis Library'. Accessed: Aug. 27, 2024. [Online]. Available: https://pandas.pydata.org/
[45] 'Python | datetime.timedelta() function', GeeksforGeeks. Accessed: Aug. 27, 2024. [Online]. Available: https://www.geeksforgeeks.org/python-datetime-timedelta-function/
[46] 'Shapefiles—ArcGIS Online Help | Documentation'. Accessed: Aug. 27, 2024. [Online]. Available: https://doc.arcgis.com/en/arcgis-online/reference/shapefiles.htm
[47] 'colab.google', colab.google. Accessed: Aug. 27, 2024. [Online]. Available: http://0.0.0.0:8080/
[48] 'GeoPandas 1.0.1 — GeoPandas 1.0.1+0.g747d66e.dirty documentation'. Accessed: Aug. 27, 2024. [Online]. Available: https://geopandas.org/en/stable/
[49] 'The Shapely User Manual — Shapely 2.0.6 documentation'. Accessed: Aug. 27, 2024. [Online]. Available: https://shapely.readthedocs.io/en/stable/manual.html
[50] 'Geographic Coordinate Systems 101: A Primer for Software Generalists', 8th Light. Accessed: Aug. 27, 2024. [Online]. Available: https://8thlight.com/insights/geographic-coordinate-systems-101
[51] 'shapely.within — Shapely 2.0.6 documentation'. Accessed: Aug. 27, 2024. [Online]. Available: https://shapely.readthedocs.io/en/stable/reference/shapely.within.html
[52] 'shapely.distance — Shapely 2.0.6 documentation'. Accessed: Aug. 27, 2024. [Online]. Available: https://shapely.readthedocs.io/en/stable/reference/shapely.distance.html
[53] 'Instituto Nacional de Estatistica, Censos 2011'. Accessed: Sep. 26, 2024. [Online]. Available: https://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=censos_base_cartogr
[54] 'OGC GeoPackage'. Accessed: Sep. 26, 2024. [Online]. Available: https://www.geopackage.org/
[55] 'pandas.Series.dt.dayofweek — pandas 2.2.3 documentation'. Accessed: Sep. 26, 2024. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.Series.dt.dayofweek.html
[56] *holidays: Generate and work with holidays in Python*. Python.

[57] ‘COVID-19’, SNS24. Accessed: Aug. 27, 2024. [Online]. Available: https://www.sns24.gov.pt/tema/doencas-infecciosas/covid-19/
[58] B. Corp, ‘Cronologia: Covid-19/Dois anos: Principais acontecimentos da pandemia em Portugal’. Accessed: Aug. 27, 2024. [Online]. Available: https://www.antenalivre.pt/covid-19/covid-19-dois-anos-principais-acontecimentos-da-pandemia-em-portugal/

Appendix

Appendix 1 - Categorisation of classification areas in need of review

Typology	Classification Area
Abatimentos - Pesquisas urgentes	Estradas e Ciclovias
	Estradas e Sinalização
Abatimentos superficiais	Estradas e Ciclovias
	Estradas e Sinalização
	Passeios e Acessibilidades
Betão	Estradas e Ciclovias
	Estradas e Sinalização
Betuminoso	Estradas e Ciclovias
	Estradas e Sinalização
	Passeios e Acessibilidades
Buraco envolvente à tampa de saneamento - Betuminoso	Estradas e Ciclovias
	Estradas e Sinalização
Buraco na faixa de rodagem - Betão	Estradas e Ciclovias
	Estradas e Sinalização
Buraco na faixa de rodagem - Betuminoso	Estradas e Ciclovias
	Estradas e Sinalização
Buraco na faixa de rodagem - Cubos	Estradas e Ciclovias
	Estradas e Sinalização
Buracos em Betuminoso	Estradas e Ciclovias
	Estradas e Sinalização
Caixas técnicas SLAT - Manutenção	Estradas e Sinalização
	Passeios e Acessibilidades
Ciclovia - Construção	Estradas e Ciclovias
	Estradas e Sinalização
Ciclovia - Manutenção	Estradas e Ciclovias
	Estradas e Sinalização
Ciclovias - Construção	Estradas e Ciclovias
	Estradas e Sinalização
Ciclovias - Manutenção	Árvores e Espaços Verdes
	Passeios e Acessibilidades
Colocação de novo mobiliário urbano	Árvores e Espaços Verdes
	Passeios e Acessibilidades
Concessionárias - Danos e obras	Estradas e Ciclovias
	Estradas e Sinalização
	Passeios e Acessibilidades
Construção	Estradas e Ciclovias
	Estradas e Sinalização

Typology	Classification Area
Cubos	Estradas e Ciclovias
	Estradas e Sinalização
Dispositivos complementares (Balizadores)	Estradas e Ciclovias
	Estradas e Sinalização
Espelho parabólico	Estradas e Ciclovias
	Estradas e Sinalização
Estabilização de obras de arte	Estradas e Ciclovias
	Estradas e Sinalização
Fresagem de Betuminoso	Estradas e Ciclovias
	Estradas e Sinalização
Grelhas de saneamento	Estradas e Ciclovias
	Estradas e Sinalização
Irregularidades (lombas) - Betuminoso	Estradas e Ciclovias
	Estradas e Sinalização
Lancil	Estradas e Ciclovias
	Estradas e Sinalização
Lancil de segurança danificado ou em falta	Estradas e Ciclovias
	Estradas e Sinalização
Mesas, bancos ou outro mobiliário urbano - Manutenção	Árvores e Espaços Verdes
	Passeios e Acessibilidades
Passadeira sobrelevada	Estradas e Ciclovias
	Estradas e Sinalização
Pavimentação	Estradas e Ciclovias
	Estradas e Sinalização
Pilaretes	Estradas e Ciclovias
	Estradas e Sinalização
	Passeios e Acessibilidades
Recolha de patas de elefante	Estradas e Ciclovias
	Estradas e Sinalização
Reconstrução	Estradas e Ciclovias
	Estradas e Sinalização
Repavimentação	Estradas e Ciclovias
	Estradas e Sinalização
Requalificação	Estradas e Ciclovias
	Estradas e Sinalização
Rotura e/ou desentupimento	Árvores e Espaços Verdes
	Higiene Urbana
Sarjetas	Estradas e Ciclovias
	Estradas e Sinalização
Separadores e proteções de faixa de rodagem	Estradas e Ciclovias
	Estradas e Sinalização
Sinalização horizontal	Estradas e Ciclovias
	Estradas e Sinalização

Typology	Classification Area
Sinalização horizontal	Estradas e Ciclovias
	Estradas e Sinalização
Sinalização informativa e/ou turística	Estradas e Ciclovias
	Estradas e Sinalização
Sinalização vertical	Estradas e Ciclovias
	Estradas e Sinalização
Tampas de saneamento	Estradas e Ciclovias
	Estradas e Sinalização
Viadutos, passagens superiores e inferiores	Estradas e Ciclovias
	Estradas e Sinalização
Viadutos, passagens superiores e inferiores - Estrutura	Estradas e Ciclovias
	Estradas e Sinalização
Viadutos, passagens superiores e inferiores - Manutenção	Estradas e Ciclovias
	Estradas e Sinalização