ISEL

Study and Modeling of CTT Locker Operation

MARTA DE FARIA RIJO FERREIRA

(Licenciada)

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Doutor António Serrador
Doutora Matilde Pós-de-Mina Pato

Júri:

Presidente: Doutor Nuno Miguel Machado Cruz
Vogais: Doutor João Moura Pires
Doutora Matilde Pós-de-Mina Pato

outubro de 2024

Study and Modeling of CTT Locker Operation

MARTA DE FARIA RIJO FERREIRA

(Licenciada)

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Doutor António Serrador, ISEL/IPL
Doutora Matilde Pós-de-Mina Pato, ISEL/IPL

Júri:

Presidente: Doutor Nuno Miguel Machado Cruz, ISEL/IPL
Vogais: Doutor João Moura Pires, FCT/UNL
Doutora Matilde Pós-de-Mina Pato, ISEL/IPL

outubro de 2024

To my grandpa.

Acknowledgements

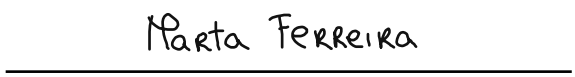
I would like to express my gratitude to my supervisors Matilde Pato and António Serrador, for their support and effort in helping me produce my best work throughout this year. Their advice and encouragement have been crucial to finishing this thesis successfully. I would also like to thank everyone involved with this project, who helped make it better, as well as express my gratitude to Instituto Superior de Engenharia de Lisboa – Instituto Politécnico de Lisboa for the opportunity and resources they gave me over these past years and for the friendships it forged.

Lastly, I owe a huge debt of appreciation to my family and close friends, whose constant support and patience – both in good times and bad – have guided me through this journey.

Statement of integrity

I declare that this dissertation is the result of my personal and independent research. Its content is original, and all sources listed in the bibliographic references were consulted and are duly mentioned in the text. I further declare that all scientific and technical references relevant to the development of the work are duly cited and included in the bibliographic references.

The author Marta Ferreira



Lisbon, October, 23

Abstract

The rise of e-commerce, greatly accelerated by the COVID-19 pandemic, has created a need for more efficient and dependable delivery methods. As a result, alternative delivery points, such as parcel lockers, are being explored as effective solutions for distributing e-commerce products. As e-commerce continues to grow, so does the study of last-mile delivery solutions.

Examining the social context, distribution, and density of parcel lockers is crucial in highlighting their importance. Previous research has identified the factors that influence the selection of delivery locations, assessed environmental risks and compared delivery methods. Few studies have been done regarding the social climate and population characteristics without surveys and questionnaires. This thesis studies and understands the social environment of parcel locker locations and density. It also suggests a method to optimise locations for installing new lockers.

This study confidently analyses locker usage patterns, by studying 750 Portuguese lockers and more than 200,000 parcels, as well as load and turnover rates, by delving into the demographic characteristics of Portuguese parishes, focusing on age, education, and employment status, and Portuguese municipalities, focusing on gross income. Real-world data from a prominent Portuguese parcel locker provider reveals that education and employment status significantly impact the selection of parcel locker locations. A comparison of the top ten municipalities by gross income, number of boxes available and the number of parcels deposited shows that gross income and locker utilization are strongly correlated. Based on the models built, it can be concluded that the optimal algorithm to use is GBM with a complete dataset. By employing this algorithm, it becomes possible to forecast locker usage by examining the correlation between PL usage, population income, and other pertinent factors.

Keywords: Visual data-analytics, locker service case study, decision-making, locker load, sociodemographic factors, machine learning

Resumo

O aumento do mercado de comércio eletrónico (recentemente impulsionado pela pandemia de Covid-19) levou à necessidade de formas mais rápidas e melhores de entrega de mercadorias. Como resultado, pontos alternativos de entrega, como os cacifos, estão a ser explorados como soluções eficazes para a entrega de produtos de comércio eletrónico. À medida que o comércio eletrónico continua a crescer, também cresce o estudo de soluções de entrega de *last-mile deliveries*.

Examinar o contexto social, a distribuição e a densidade dos cacifos de encomendas é crucial para destacar a sua importância. Investigações anteriores identificaram os fatores que influenciam a seleção dos locais de entrega, avaliaram os riscos ambientais e compararam os métodos de entrega. Poucos estudos foram feitos sobre o ambiente social e as características da população sem usar questionários. Esta tese tem como objetivo estudar e compreender o ambiente social da localização e densidade dos cacifos de encomendas. Também sugere uma forma de otimizar locais para instalação de novos cacifos.

Este estudo analisa com segurança os padrões de utilização dos cacifos, estudando 750 cacifos e mais de 200.000 encomendas, bem como as taxas de carga e rotatividade, investigando as características demográficas das freguesias portuguesas, com foco na idade, escolaridade e situação profissional, e nos municípios portugueses, com foco no rendimento bruto. Este estudo utiliza dados reais de um grande operador de cacifos português e os dados revelam que a educação e a situação profissional impactam significativamente a seleção de locais para cacifos de encomendas. Uma comparação dos dez principais municípios por rendimento bruto, número de caixas disponíveis e número de encomendas depositadas mostra que o rendimento bruto e a utilização de cacifos estão fortemente correlacionados. Com base nos modelos construídos, pode-se concluir que o algoritmo ideal a ser utilizado é o GBM com um conjunto de dados completo. Ao usar este algoritmo, torna-se possível prever o uso de armários examinando a correlação entre o uso de cacifos de encomendas, o rendimento da população e outros fatores pertinentes.

Palavras-chave: Análise de dados; Visualização; Tomada de decisão; Lockers, Cargas; Estudos demográficos; Tableau

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

DRF	Distributed Random Forest 20, 46
GBM	Gradient Boosting Machine 20, 45, 46
GLM	Generalized Linear Models 20, 46
INE	Portuguese National Statistics Institute 2, 13, 15, 16, 30, 47, 48
LMD	Last-Mile Delivery 5, 8
MAE	Mean Absolute Error 45, 46
MSE	Mean Squared Error 45, 46
PL	Parcel Locker xxi, 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 15, 18, 23, 25, 26, 27, 29, 30, 43, 45, 47, 48, 49
R^2	R Squared 45
RMSE	Root Mean Squared Error xxi, 20, 45, 46
RMSLE	Root Mean Squared Logarithmic Error 45, 46

1 Introduction

In this chapter, the thesis context is briefly presented, emphasising its significance, and outlining the thesis objectives. The chapter ends by describing the structure of the document.

E-commerce traffic is increasing, bringing challenges for urban freight transportation [32]. According to [14], in Europe (2023), estimated values, 92% of the population accessed the Internet and 78% of those bought goods or services online. In Portugal, those numbers are 85% and 62%, respectively. Since 2021, those 62% have been almost fixed, however, the revenue for the e-commerce market is forecast to increase in the next coming years continuously [30]. In 2023, according to CTT e-Commerce Report 2023 [19], e-commerce in Portugal was estimated to exceed 10.000 million euros, a global increase of 4.3% compared to 2022.

As the e-commerce market increases, so does the demand for efficient and fast delivery of goods. The most common methods of getting items to a retail client are picking them up in-store (BOPIS, or Buy Online, Pick Up In-Store) or via other Pick Up Drop Off (PUDO) locations; courier services; logistics companies; and online retailers. One solution to increase last-mile efficiency is parcel lockers (PLs), making it unnecessary for both the deliverer and the receiver to be in the same place and moment. Thus, there is no need for second or third delivery attempts. Furthermore, PLs also add value as they can work in reverse, meaning the return of goods when required, and contribute to environmental sustainability by reducing pollution, congestion, and accidents [10, 23].

Across Europe, many countries have successfully incorporated PLs into e-commerce delivery strategies. Examples include InPost in Poland, PostNord in Sweden, Denmark and Norway, PostNL in the Netherlands, La Poste in France, and Bpost in Belgium [13].

In Portugal, the largest network of PLs is operated by Locky, with a network of over 1,000 lockers installed throughout the mainland and islands. The lockers are strategically placed in high-traffic areas such as supermarkets, shopping centres, and car parks. Plenty of Locky's locations are accessible 24 hours a day, providing customers with a convenient and reliable option [19]. In addition, Amazon Lockers and DPD Lockers can also be found in use in Portugal, for example in gas and subway stations.

In light of this, it is essential to understand the PL users as a way to determine optimal locations

for PLs. When one mentions an increase or decrease in lockers, it could be either a locker nearby or more boxes in that area. It is important to consider that various factors can impact the use of PLs, such as availability, accessibility, security, the age of the users, and the distance to a locker, among others. One aim is to determine if certain population characteristics influence the use of PLs as a method of goods distribution, without conducting surveys and directly contacting users.

1.1 Objectives and approach

The current thesis focuses on characterising the PL Portuguese users by identifying social indicators that might impact the use of PL and analysing different social key features or characteristics, to determine the optimal locations for installing PLs, adding new ones or changing previous locations, by maintaining detailed records of their usage and identifying overcrowded lockers or areas with excessive PL density through the analysis of the Portuguese population.

In summary, the following objectives are pursued:

O1 Determine the optimal locations for installing PLs, while also maintaining detailed records of their usage and identifying any overcrowded lockers or areas with excessive PL density; and

O2 Analyse key social factors to understand the impact of these factors on the use of PLs.

To achieve these goals, we want to answer the following research question:

Q1 What are the main characteristics of Portuguese locker users in the most used locations?

The datasets, used in this work, are gathered from real data provided by one PL operator regarding deliveries in Portugal from January to September 2023. The datasets regarding the population are gathered from Portuguese National Statistics Institute (INE). Employing visualisation methods enabled us to establish relevant relations among locker usage and sociodemographic variables, offering valuable insights for operators to optimise revenue and enhance service quality.

Machine learning algorithms make it possible to analyse the relationship between demographic features and PL usage more effectively, producing more insightful and practical results.

However, since the data about the lockers is private, neither the datasets nor the dashboards will be made public, instead, examples of the dashboards will be shown in the results part of this thesis.

1.2 Publications and presentations

During this thesis, we published a paper in the 28 International Conference Information Visualisation, titled "Understanding Portuguese Users of Parcel Locker Services". This article is published in the IV 2024 Conference Proceedings, available here. This paper focuses on some

of the characteristics discussed in this thesis and examines and compares them between two groups of parishes.

Additionally, the work done for the previous paper was adapted to focus only on the city of Lisbon and presented at the 2º Encontro do Laboratório de Dados Urbanos de Lisboa, titled "*Caraterização do utilizador de uma rede de cacifos*". The characteristics discussed in this presentation were the same as in the paper above but focused on the Lisbon region.

1.3 Document structure

This document is structured in 5 chapters: Chapter 1 introduces the problem. In Chapter 2 is the literature review on different strategies for PLs. Chapter 3 reflects how we collect and analyse data. Chapter 4 presents and discuss results. Finally, Chapter 5 concludes and suggests some future research directions.

2 Related Work

This chapter provides a context to the PL solution and the different areas of study that have already been conducted. From the literature review, common topics emerged. The last section also discusses the database used.

2.1 Lockers research

Last-mile delivery The final phase of a delivery process, known as the Last-Mile Delivery (LMD), plays an important role in the logistics supply chain. It is often the most costly and time-consuming aspect of the delivery journey, and it has the potential to influence customer satisfaction and loyalty [6, 8, 24]. Different relevant areas related to LMD appear, such as carbon emissions and the environmental impact of this step [2, 10, 20, 25], the delivery costs [8, 26], and service quality and factors influencing this process, discussed in [3, 8, 10, 12].

Despite the challenges associated with this phase, PLs offer a solution to mitigate these issues by streamlining the delivery process and reducing costs. However, the implementation of PLs requires careful consideration of several factors. Research indicates that seven key variables significantly influence the selection of PL locations [13]. This highlights the importance of strategic planning to optimise the use of PLs. The variables included are described in Table 2.1.

Availability is the possibility of delivery and collection of parcels 24/7, while accessibility is the degree of connection with the different infrastructure and transport modes. Security is the state of being free from danger or threat. Another variable considered is how PLs impact the environment in terms of emissions and land occupation, and also the price of the installation and PLs maintenance. Method of use is the procedure to implement the PLs usage, whilst regulations are the official rules from different countries to regulate PLs activity.

The article by Molin et al. [15] examines consumers' preferences for receiving parcels through a stated choice experiment between home delivery, service point and PL, varying the price, delivery moment and distance. They conclude that price variations have a significant impact on usage. The research conducted by Jagoda et al. [11] results in a cross-generational perspective of e-customer preferences on LMD revealing that the parcel price and the possibility of free return are still the most important factors in choosing the delivery method. Based on a

Table 2.1: Variables influencing the selection of PL locations - (a) Availability, (b) Accessibility, (c) Security, (d) Environmental impact, (e) Costs, (f) Method of use, (g) Regulations.

Authors & year	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Iwan et al. (2016) [10]	✓	✓	✓	✓			
Huong and Thiet (2020) [8]	✓		✓		✓		✓
Pinchasik et al. (2023) [20]				✓			
Russo and Comi (2023) [25]				✓			
Carotenuto et al. (2022) [2]				✓			
Sawik et al. (2022) [26]					✓		
Lachapelle et al. (2018) [12]	✓	✓	✓			✓	
Cieśla (2023) [3]	✓	✓	✓				
Iannaccone et al. (2021) [9]	✓	✓					
Moslem and Pilla (2023) [16]		✓					

Kano questionnaire, Cie'sla [3] classifies attributes into must-be, one-dimensional, attractive and indifferent. According to users, the essential attributes that should become mandatory for service providers include parcel stations, improvements for the disabled, adjusting the size of the parcel to the size of the box, proper placement of the parcel in the box, and a well-functioning dedicated application.

The study performed by Stašys et al. [29] analyses average satisfaction with PL services revealing four primary reasons for their satisfaction: not having to wait at home for a parcel, the ease of use of PLs, the service being available at any time, and the convenient location of the PLs. A second group of less important factors included the speed of delivery, quality of service, lack of human interaction, reliability of delivery, and the balance between price and quality. Additionally, the study identified three factors that were even less significant for customer satisfaction: technical novelty, quick response to customer concerns, and company reputation.

The cost and emissions implications are evaluated by Dong et al. [4]. Results suggest that using PLs is beneficial and improves service efficiency as it creates economic and environmental savings compared to home deliveries.

According to Iwan et al. [10], the most important factor of efficiency regarding PLs is the proper location of the lockers used for deliveries in the city area. The other most important reasons for PLs' utilisation are their availability and localisation concerning the user, close to home, on the way to work, among others.

Locker size The market offers a variety of PLs, each tailored to specific installation environments (indoors or outdoors) and intended uses (such as package delivery or refrigerated storage). The choice of PL also depends on the user base, which may be public/private, individual/collective, or mobile/stationary. Regardless of their size and configuration, PLs accommodate packages of varying sizes, with small packages fitting into medium and large lockers and medium packages fitting into large lockers.

Ranjbari et al. [24] identified two key metrics to evaluate locker performance, which they consider paramount in assessing the effectiveness and efficiency of PLs. The metrics are:

(1) Locker occupancy rate – the percentage of occupied locker boxes; (2) Number of excess (or rejected) packages – the number of delivered packages that exceed the locker capacity and require rescheduling.

They also highlight the potential impact of slight variations in locker cell sizes on the number of excess packages, underscoring the need for careful planning and consideration before deciding on a locker's size and configuration. To make informed decisions, locker providers must gather data on online shopping behaviours and population density for the targeted location. This data is crucial for understanding the demand patterns and selecting the most appropriate PL size and configuration to efficiently meet customers' needs. The authors attempt to find the right size and configuration for a residential building locker.

In Viana's thesis [31], a mathematical model is proposed to aid decision-making, depending on a specific demand, the model returns the number of towers of a maximum of three configurations that will best help reduce the breakage with the least number of towers possible.

Location problem Installing PLs systems in car-dominant cities is quite challenging considering the transportation infrastructures and sustainability of the city, for instance. In Australia, Lachapelle et al. [12] state all PLs are in easily accessible locations, with disability access, and all with extra lighting outside the lights of the locker. The locations are: (1) urban areas with high potential for pedestrian traffic; (2) suburban areas with more diversified access features like parking and high street visibility; (3) suburban post locations with adjoining post office locations, indoor lockers, and located in small malls; (4) shopping centres, typically identified as the least safe locations but that have otherwise varying features and mostly abundant parking. It should be emphasised that, unlike many European systems, no PL has been installed next to bus stations or ferry terminals. They conclude that PLs are predominantly situated in areas with higher population density, balancing working and resident populations. Additionally, areas with a significant proportion of households having internet access are more likely to have PLs.

The research presented by Iannaccone et al. [9] highlights a robust demand for PLs among young e-commerce consumers. This emphasises the importance of a dense PL network (with approximately four PLs/km^2) and ensuring round-the-clock accessibility. The optimal number of PLs for Rome (1,285 km^2) would be approximately 5,140. In Poland, according to Cie'sla [3], in 2023, 59% of the population can access a PL within a walking distance of less than seven minutes. This value is even higher in urban areas, where 85% of residents can reach a PL within the same or shorter distance.

A study from Wang et al. [33] proposes a robust optimisation model considering the demand uncertainties, including the large and small parcels to be sent and received. It can optimally determine the location, by the constraints of acceptable walking distance, and the number of large and small PLs for each location under various robust levels.

The work presented by Moslem and Pilla [16] strives to identify the ideal locations for PLs, in Dublin, by applying a unique decision-making model, using the Euclidean Distance-Based Aggregation Method (EDBAM) for the selection of PL locations. The first step in selecting a location typically involves identifying the demand for usage and any additional capacity needs. Following this, an analysis is conducted to determine the optimal location.

Lastly, the work done by Ottaviani et al. [18] introduces an innovative approach to solving the location problem by integrating mixed-integer linear programming with greedy heuristics algorithms. The proposed solution was tested using real customer demand data from Turin, Italy. The results indicate that, on average, 10 to 11 PLs are needed to cover 90% of the estimated potential demand.

Environment Pinchasik et al. [20] study the effects of using PLs as an alternative to home deliveries, combined with different locker network expansions in Oslo. They assess the cost, traffic, environmental, and societal damage cost effects when home deliveries are instead delivered to PLs. This study shows that more usage of PL can potentially lower last-mile transportation emissions, traffic, and logistical costs, providing insights into the relationships between various network development tactics, receivers' walking distances, LMD efficiency issues, and the implications of recipients' parcel collection visits. Increased usage of PL has the potential to significantly reduce traffic, emissions, and other negative social effects in urban areas, particularly when pick-up excursions can be conducted spontaneously in conjunction with other errands and without using an exhaust vehicle (ideally by walking or cycling). The benefits of rerouting home deliveries to PLs, in terms of reducing emissions, decrease, but the reductions in traffic and societal damage remain significant.

Sustainability Cano et al. [1] offer a bibliometric analysis and literature review to identify recent articles, important subjects, and trends regarding the sustainability of logistics operations in e-commerce contexts. This study analyses the main management topics surrounding sustainable e-commerce logistics, with a focus on LMD, as well as urban and city logistics' environmental impact. From a technological standpoint, the major subjects are vehicle routing supported by multi-objective optimisation, optimisation methods, and planning and decision-making strategies to increase cost and energy efficiency. Similarly, the literature review revealed research trends in freight transportation systems using electronic vehicles, drones, and delivery technologies. The literature review establishes that sustainable e-commerce logistics is approached from economic, social, and environmental perspectives.

Finally, Silva et al. [28] characterise the sustainable urban last-mile logistics research field through a systematic review, into six thematic clusters that identify the main concepts addressed in the different areas of the last-mile research and group the solutions in vehicular, operational and organisational solutions. The clusters are supply chain and channels, delivery methods and attributes, innovative vehicles, logistics infrastructures and schemes, operational optimisation and emerging business models.

Reduce delivery times The study presented by Ranjbari et al. [23] builds a nonequivalent groups pre-test/post-test control experiment to estimate the causal effects of a PL on delivery times in a residential building in Seattle, WA. The causal impacts are measured using the difference-in-difference method using a nearby residential building as a control. This includes vehicle dwell time and delivery courier time spent inside the building. The findings revealed a statistically significant decrease in time spent inside the building, as well as a minor but insignificant reduction in delivery vehicle stay time at the curb. In this study, building residents

were automatically registered to receive their parcels through the locker, removing the possibility of alternative delivery. The success of lockers in lowering delivery times is dependent on user uptake. If only a few people utilise the locker, there will be no delivery consolidation and no noticeable improvement in delivery times. This could be a more serious issue in the case of neighbourhood public lockers or lockers in commercial buildings where people choose to become users. Reduced dwell time at the curb enhances parking turnover, which is especially beneficial in urban core regions, and expands network capacity without the need for new infrastructure. Furthermore, by eliminating failed delivery efforts, lockers cut delivery truck mileage, traffic congestion, and emissions.

Forecast demand Lastly, Sethuraman et al. [27] propose a combination of machine learning techniques to predict locker demand and package dwell time with linear programming to maximise throughput in lockers. With rising demand, capacity management has become crucial for Amazon Locker operations.

One summarises the three steps in capacity management as predicting the number of packages expected in each locker, for each shipping option, and on each day over the next week; estimating the probability that a package will remain in the locker for up to six days, depending on the shipping speed and delivery day; and determining the optimal capacity reservations for different shipping options, aiming to maximise throughput using linear programming.

While most studies have concentrated on the location, usage, cost, and sustainability of PLs, there are more research topics to be investigated, for instance, PL user's characteristics like age and education level. This study addresses this gap by examining how these factors influence the use of PLs, by studying census data instead of surveys. One also attempts to guide decision-making on where and if to install or remove lockers based on the population characteristics. With the help of a machine learning tool, one is also building models to predict locker usage.

2.2 Database

Since we need to operate on parcels and locker information, the fastest way to manipulate these data is to use a database.

MySQL MySQL [17] is an open-source relational database management system that stores data in tables with rows and columns and, because of its ACID compliance, it ensures that all committed transactions are permanently saved. It is reliable, performs well, can scale and is easy to use. MySQL supports a wide range of data types, as well as extensive built-in functions for data manipulation and analysis. Also, since MySQL is a more mature and full-featured database, with extensive documentation, it offers more resources. Initially, this database is used because it connects to Tableau. However, DuckDB is chosen due to its columnar format.

DuckDB DuckDB [5] is a free analytical database that stores data in a columnar format, enabling fast query processing and compression. It has no external dependencies, neither for compilation nor during run-time, making it extremely portable. DuckDB does not run as a separate process but completely embedded within a host process, having the additional

advantage of high-speed data transfer to and from the database. A key aspect is that the DuckDB Python package can run queries directly on Pandas data without ever importing or copying any data. DuckDB provides transactional guarantees (ACID properties) and data can be stored in persistent, single-file databases.

DuckDB is engineered to facilitate online analytical processing (OLAP), sometimes referred to as analytical query workloads. To do this, just-in-time query execution engines or vectorized data management state-of-the-art are used. DuckDB has a query execution engine that is columnar-vectorized. This means that while queries are still parsed, a big batch of values (referred to as a "vector") are processed all at once. This significantly lowers costs associated with conventional systems that process each row sequentially, such as PostgreSQL, MySQL, or SQLite. For OLAP queries, vectorized query execution produces far better results.

Despite not being one of the database types supported by Tableau, there is a way to connect to DuckDB using a JDBC connector. Because the generic JDBC connection doesn't use the numerous connection-specific attributes that a named Tableau connector uses to maximise speed, you'll notice noticeable differences in performance between this "generic" JDBC connector and a named Tableau connector. However, since we are dealing with large amounts of data and want fast query processing, DuckDB is used as the database management system.

3 Methodology

This chapter provides a detailed explanation of the analysis of the data and the interconnections between different variables. It outlines the steps involved in data understanding and pre-processing using the PL operator. Furthermore, it presents the rationale behind the choice of characteristics to be studied within the Portuguese population and territory.

It is divided into sections regarding the data's origin and use, so we have lockers, population and territory sections. Sections regarding the dataset processing. Plus sections describing the knowledge extraction and the overall design of this study.

3.1 Lockers

From the PL operator, three datasets were initially provided, two regarding the lockers and another about the parcels going through those lockers. The data concerning lockers is static and reflects the period from January to September 2023. Later, the PL operator load dataset was also provided. Table 3.1 describes succinctly the size and conditions of the datasets, while Table 3.2 and 3.3 describe their attributes. The `Locker Loads` dataset only has three attributes: `provider_terminal_id`, `date` and `load`.

Concerning Table 3.1, it is worth noticing that the `Parcels` dataset has missing values, however, the PL operator has considered that and as mentioned in Table 3.3, the attribute `Valid locker event date` acts as a safeguard for those missing values. These missing values exist due to synchronisation problems in the lockers. As for the `Lockers`, despite there being 750 lockers, in reality, we are looking at 719 lockers, after cleaning and pre-processing the `Lockers` datasets. The `Locker Loads` dataset has null values with some representing the dates where a locker hasn't been installed yet.

The `Lockers` data was stored in a DuckDB database. Since all records are historical and data accumulates over time, these tables are efficiently updated using Python, which was also used to compute the locker load. Once the comprehensive dataset was assembled, Tableau Prep Builder (Version 2023.3) was used to manipulate and aggregate the data, and Tableau (Version 2023.2) for creating clear visualisations to enable enhanced data comprehension and build dashboards. Using Tableau Prep Builder allowed for easy aggregation of the `Lockers`

Table 3.1: Statistic of the Datasets

Dataset	# of Tuples	Observations
Lockers	750	No null values.
Lockers info	768	No null values.
Parcels	211,064	1,696 tuples w/o deposit date and 240 tuples w/o exit date.
Loads	275,940	45,739 tuples without load.

Table 3.2: Lockers' attributes and description.

Attribute	Description	Original
Provider terminal id	PL operator locker identification.	Yes
Customer face name	Name of the locker.	
City	City where the locker is.	
Region	Region where the locker is.	
Latitude	Latitude of the position of the locker.	
Longitude	Longitude of the position of the locker.	
Number of boxes	Number of boxes the locker has.	
Municipality	Municipality where the locker is.	No
Parish	Parish where the locker is.	
Code INE	Administrative region code, defined by INE.	
Installation date	Date of installation of the locker.	
Description	Type of place where the locker can be found.	
Closest locker	Name of the locker that is closest.	
Closest locker distance	Distance, in kilometres, to the closest locker.	

Table 3.3: Parcels' attributes and description.

Attribute	Description	Original
Partner parcel label	PL operator partner identification.	Yes
Locker id	PL operator locker identification.	
In demand	Timestamp of the locker reservation.	
Deposited	Timestamp of when the courier deposited the parcel.	
Customer collected	Timestamp of when the customer picked up the parcel.	
Customer collect error	Timestamp of when an error occurred.	
Customer collect finished	Timestamp of when the locker is released after a pick-up (10 minutes after `Customer collected`).	
Valid locker event date	Older timestamp between the previous four, in case the deposited event is missing.	
Courier collected	Timestamp of when a courier picked up a missed delivery.	

datasets by joining them through locker identification.

Note that in Table 3.2, there is an additional column "Original". This column distinguishes between the original attributes and those added to this dataset. As we want to cross-reference lockers with population data, we need something to link them and, as neither `City` nor `Region` are geographical levels considered by INE, we need to add this information to each locker. Python was employed to access geoapi.pt, a RESTful API that offers information on administrative regions, georeferencing, censuses, and postal codes. This API was instrumental in obtaining the parish for each PL, complementing the locker data by filling in the missing parish-level information. However, the free API requests are currently restricted to a number considerably lower than the number of lockers. Considering this limitation, if one adds more than 100 lockers at once to the database, `municipality`, `parish` and `code INE` are going to be empty, unless the PL operator starts to use those attributes as well.

Parameters like `Installation date` and `zone` were later added to the lockers dataset. They were also provided by the PL operator but, since they did not come with the initial datasets, we are categorising them as not original.

The remaining two attributes provide additional, useful information for each locker. To calculate the closest locker, one uses the haversine formula, which is available through a Python module. This formula determines the distance between two points on a sphere (Earth), using their latitude and longitude [22].

Some errors in lockers' `city`, `region` and `number of boxes` were also corrected before adding them to the database.

To evaluate and compare the performance of different lockers, two key indicators are considered, as follows:

(a) locker load (lo_l), and

(b) locker turnover (lo_t).

The locker's load, adapted from [23], defined by eq. (3.1), provides insights into the occupancy rate of the lockers at any given time, offering a clear picture of the ratio between occupied boxes and the total number of boxes installed. Meanwhile, daily locker turnover, eq. (3.2), allows for the identification of which lockers are utilised more frequently, thereby highlighting their operational efficiency:

$$lo_l = \frac{p}{b} \tag{3.1}$$

$$lo_t = \frac{d}{b} \tag{3.2}$$

where p is the the no. of parcels, b is the no. of boxes, d represents the number of deliveries.

The lifespan of a parcel within a specific locker is defined by two key dates: (1) its entry date, which is the date on which the parcel is deposited into a box, and (2) its exit date, the latest date on which the box is emptied.

The calculation of parcels occupying a locker involves the addition of a parcel upon its deposit and its subtraction upon retrieval. In reality, this calculation involves two algorithms. Initially, the assumption, in regards to the load calculation was that all the lockers existed from January to September 2023, and, as such, there was no distinction between a locker having zero load and not yet having been installed. From the results viewpoint, both produced zero load and, from a visualisation standpoint, that is not helpful. As a result, Algorithm 1 obtains the date of when the box is liberated and Algorithm 2 calculates the load, when a locker has not been installed yet, the load is null. The entry date of a parcel is considered to be the `deposited` date, however, if that attribute has no value and `in_demand` has one, we consider that as the entry date. Also, in order not to over-complicate, the load is calculated for every hour of every day between January and September 2023. Ideally, it would be advantageous to be connected live with the lockers, instead of dealing with previous records.

Algorithm 1 Algorithm to obtain the earliest date of when a locker is liberated.

Require: `customer_collect_finished`, `courier_collected`, `customer_collect_error`, `customer_collected` **return** earliest of these dates or 1/1/2000 ▷ Arbitrary date that will not be found in the dataset

Algorithm 2 Algorithm to calculate locker load.

```
for all Lockers do
    for all Parcels do
        if parcel provider_terminal_id matches locker provider_terminal_id then
            get parcel entry date
            get parcel exit date
            while entry date < exit date do
                count 1
            end while
        end if
    end for
    if locker already installed then
        count/number_of_boxes
    end if
    write to the database
end for
```

This study calculates locker load and turnover indicators for each locker and parish. To calculate the load or turnover of a parish, it's necessary to aggregate all lockers within that parish. In short, for parishes, one is summing the number of boxes each locker in that parish has, the number of boxes that are occupied and the number of parcels that are deposited, for that hour or day. These operations are being made in Tableau Prep Builder.

These indicators can be interpreted in the context of parish distribution, population density, age group, education, and employment status, providing a comprehensive understanding of their usage patterns.

To increase the performance of the models defined in section 4.6, new variables were derived from the load data set. They are `day_of_week` – which indicates which day of the week (in words) the load is being calculated for, `special_day_of_week` – which shows if it is the weekend or a holiday (yes/no), and `holiday` – which indicates, in case it is a holiday, the name of the holiday. To accomplish this, a small dataset containing the Portuguese holidays of 2023 was also added to Tableau Prep Builder.

An additional insightful indicator is the ratio of boxes to population density, eq. (3.3), which helps identify whether parishes are under-serviced, over-serviced, or if the ratio is standard:

$$b_p = \frac{bo}{po} \tag{3.3}$$

where bo is the the number of boxes and po is the population density for each parish. This indicator was calculated inside Tableau and applied directly to the dashboards.

3.2 Population characteristics

To complement the locker usage data, secondary quantitative data collected by INE is utilised to investigate the factors influencing PL usage. This approach allows for the analysis of datasets that encompass information about the lockers themselves, the parcels processed through these lockers, and various demographic attributes of Portugal's population. The INE data specifically references the most recent census, Censos 2021, providing a comprehensive view of the population. The geographic level of analysis may vary depending on the factor under investigation, ranging from municipality (NUTS III) to parish.

Data from INE was consulted online to obtain a dataset with population information for each parish in Portugal, focusing on minimum education levels and employment status. This data was organised by age group, offering a detailed demographic profile that complements the analysis of PL usage.

In short, each parish is characterised by:

(i) Population density (No/km^2);

(ii) Age Group (0-14, 15-24, 25-64, 65-older (Count));

(iii) Proportion of the resident population with completed higher education, with at least the 3rd cycle of basic education completed, and with at least complete high school education; and,

(iv) Employment status of the resident population aged 15 and above (Count).

To characterise the population concerning income, we have to look at municipalities. Table 3.4 provides a comprehensive overview of the datasets used in this study from INE, regarding the population characteristics. Datasets referent to the parishes have 3,092 tuples representing one for each parish of Portugal, plus *Corvo* (that is not considered a parish). When the granularity is municipalities, the datasets have 308 tuples since that is how many municipalities there are in Portugal.

Table 3.4: Population datasets' short description, from INE.

Datasets	# Tuples	Observations
Population's density (No/km^2) by Place of residence at Census date [2021] (NUTS - 2013) and Sex	3,092	No null values.
Resident population (No) by Locality (Census), Sex and Age group (By life cycles)	3,092	No null values.
Resident population with 15 and more years old (No) by Place of residence at Census date [2021] (NUTS - 2013), Sex, Activity status and Source of income	3,092	No null values.
Proportion of resident population with higher education completed (%) by Place of residence at Census date [2021] (NUTS - 2013) and Sex	3,092	No null values.
Proportion of resident population with at least the lower secondary education 3rd cycle completed (%) by Place of residence at Census date [2021] (NUTS - 2013) and Sex	3,092	No null values.
Proportion of resident population with at least upper secondary education completed (%) by Place of residence at Census date [2021] (NUTS - 2013) and Sex	3,092	No null values.
Gross reported income per inhabitant (€) by Geographic localisation (NUTS - 2013); Annual	308	10 tuples do not have median values.

3.3 Territory

From the INE's data, a key detail is missing necessary to represent the data on a map view, that is information about the geographic coordinates of each location. To best represent the parishes in visualisations, we obtained Portugal's shapefiles from Carta Administrativa Oficial de Portugal, available on Direção-Geral do Território. The shapefiles are organised by main geographic area, namely *Continente*, *Açores* and *Madeira*, producing three sets of files. However, to better understand and read these data in Tableau it is useful if we can join these files. So far Tableau does not have a way to join or union different data sources into a single table. The workaround for this problem is easy. We simply have to put the unzipped files into a single folder, regardless of the area, and now we can union the data from these areas into one single table with the geographic information of Portugal using a single data source.

About the administrative division of Portugal in regards to the use of administrative units in the context of INE, each administrative region is assigned a numerical code (referenced as `code INE` in this document), according to a set of criteria [21]. Hence, a three-level structure:

- level 1: *distrito* – district in the case of the mainland, and island in the case of the Autonomous Region of Açores and Madeira, identified through a two-digit numerical code
- level 2: *município* – municipality, identified through a four-digit numerical code; two of these digits correspond to the municipality within the district
- level 3: *freguesia* – commune (parish), identified through a six-digit numerical code; two

of these digits correspond to the common within the municipality.

Table 3.5: Territory datasets' short description, from INE.

Datasets	# Tuples	Observations
Freguesias classified by Degree of urbanisation (Eurostat), 2016, organized by NUTS 2013	3,092	No null values
Freguesias classified by Classification of urban areas (CAOP2013) level 1 and 2 for NUTS 2013	3,092	No null values

Furthermore, to better understand the territory and the population distribution, Table 3.5 provides an overview of two datasets concerning the degree of urbanisation and classification of urban areas of each parish.

The degrees of urbanisation vary from densely populated areas (ADP) to intermediate areas (AMP) and thinly populated areas (APP). Table 3.6 shows the classification of Portugal's parishes according to the degree of urbanisation based on population density. The classification of urban areas falls into three categories: predominantly urban area (APU), medium urban area (AMU), and predominantly rural area (APR), and Table 3.7 shows the classification of the parishes by these categories, also by population density.

Figure 3.1 depicts the distribution, by percentage, of the classification of urban areas and the degree of urbanisation.

Table 3.6: Classification of parishes based on degree of urbanisation according to population density.

	ADP		AMP		APP	
Population density	**Nº of parishes**	**%**	**Nº of parishes**	**%**	**Nº of parishes**	**%**
100 - 500 hab/km^2	121	34.28	325	58.45	392	17.96
< 100 hab/km^2	27	7.65	36	6.48	1,787	81.86
> 500 hab/km^2	205	58.07	195	35.07	4	0.18

Table 3.7: Classification of parishes based on classification of urban areas according to population density.

	AMU		APR		APU	
Population density	**Nº of parishes**	**%**	**Nº of parishes**	**%**	**Nº of parishes**	**%**
100 - 500 hab/km^2	498	67.57	59	3.64	281	38.39
< 100 hab/km^2	238	32.29	1,564	96.36	48	6.56
> 500 hab/km^2	1	0.14			403	55.05

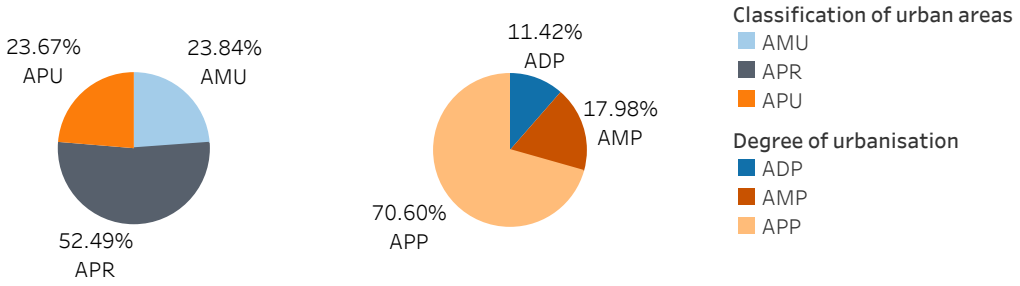


Figure 3.1: *On the left is the distribution of the classification of urban areas, and on the right is the distribution of the degree of urbanisation.*

Table 3.8: Duration of main database operations in MySQL and DuckDB.

Database operation	**MySQL**	**DuckDB**
Add lockers	1 second	0 seconds
Add parcels	2 minutes	0 seconds
Calculate load	$\approx$ 4 hours	40 minutes

3.4 Database

One of the reasons to use DuckDB, in addition to it being an analytics database, is the speed with which it processes the datasets involved. Essentially, there are three operations involving the `Lockers` datasets: (1) add lockers to the lockers table, (2) add parcels to the parcels table, (3) calculate the load and add it to the loads table.

Table 3.8 describes the times involved in each of these operations for a MySQL and DuckDB database. The fact that DuckDB is so fast at processing this data, especially the calculation of locker load, makes it an optimal choice for this work.

3.5 Tableau pre-processing

DuckDB can read files into tables but only CSV, Parquet and JSON. They offer an Excel extension, but unlike what its name may suggest, it does not provide support for reading Excel files. Since we need to pre-process the data we obtained from the PL operator, we are using Tableau Prep Builder to help us. This tool is also useful since we need to guarantee the parish identification, `code INE`, is always read as a string, and one of the ways we can output a flow in Tableau Prep Builder is in CSV, which DuckDB reads. Tableau Prep Builder is also being used to build the datasets described in the next section for knowledge extraction.

All the different data sources mentioned above are then imputed into Tableau. A Tableau Data Source acts as a bridge between the original data and Tableau and it encompasses the data, whether live or extracted, connection details, table or sheet names. When the data spans multiple tables or databases, as is our case, we need to combine it to create a view.

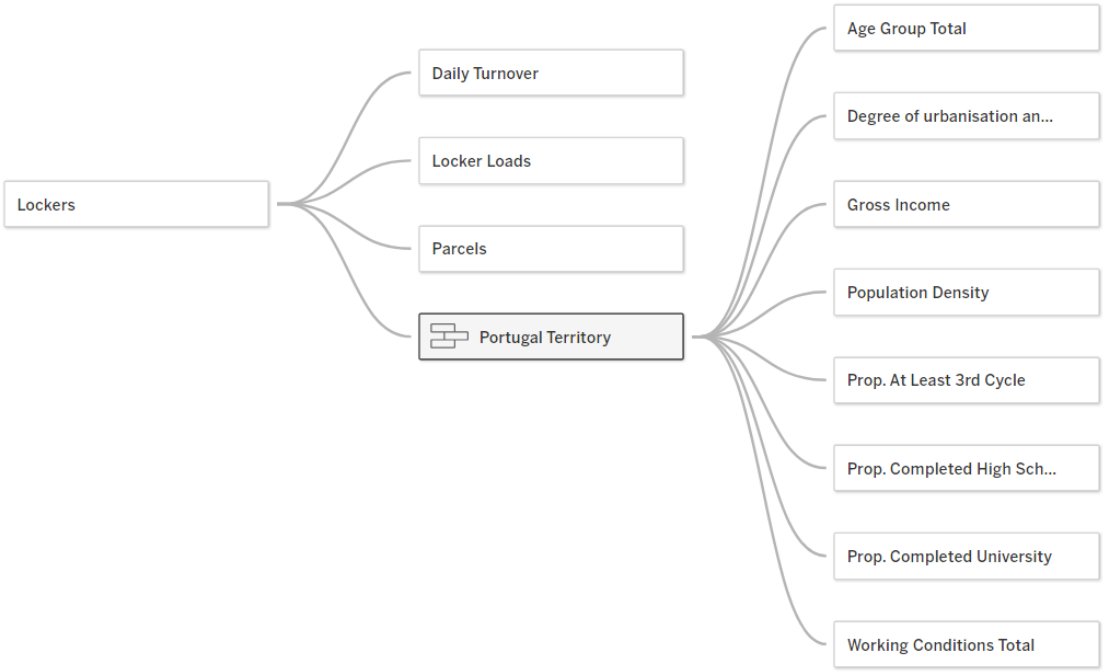


Figure 3.2: *Overall Tableau Data Source view.*

In Tableau, relationships are the primary method for combining data. They offer a dynamic and flexible way to merge data from various tables for analysis. If needed, tables can also be manipulated. Relationships link data from multiple tables based on shared fields, guiding Tableau on how to connect rows across tables. Unlike hard-coded joins, relationships don't immediately merge rows. Instead, when we create a visualisation, Tableau traces the fields involved and generates the necessary joins to retrieve the correct data. This approach is beneficial when dealing with data at different levels of detail or granularity.

As seen in Figure 3.2, our Tableau is built using 17 tables and databases. `Lockers`, `Parcels` and `Locker Loads` correspond to tables in the DuckDB database, `Portugal Territory` is the union of the shapefiles of *Açores*, *Madeira* and *Continente*, and the remainder are excel tables with information about the Portuguese population.

3.6 Knowledge extraction

Three versions of the dataset, as described in Table 3.9, were created, with 4,789,513 tuples, each used to generate predictive models.

Table 3.9: Description of the datasets for H2O.

Dataset	Description
A	Load dataset plus locker dataset, with no added or derived variables.
B	Previous dataset where the extra variables of lockers were added, plus the derived variables of load.
C	Previous dataset plus the population characteristics.

The models were built using the H2O tool (Version 3.46.0.1), specifically, AutoML. This tool parses the datasets easily. Given data from January to September 2023, one splits the frames (0.75/0.25 ratio) to use for training and validation. By running AutoML, one can choose different configurations but the easiest is to select the training and validation frames and the response column, which in this case, is the load of a locker. This will build models using supervised algorithms: Generalized Linear Models (GLM), Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), deep learning, and XGBoost [7]. AutoML orders the resulting models according to Root Mean Squared Error (RMSE), best first. In the end, one chooses the best one for each dataset and algorithm and compares them to see which is possibly the best model to be used to predict the load.

The algorithms, in summary, are as follows:

- GLM estimates regression models for outcomes following exponential distributions.
- DRF is a classification and regression tool.
- GBM, for regression and classification, is a forward learning ensemble method.
- H2O's deep learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation.
- XGBoost is a supervised learning algorithm that implements a process called boosting to yield accurate models.

To build these models with H2O, we needed to allocate at least 10G. The command to run this tool is: `java -Xmx10g -jar h2o.jar`.

3.7 Python module considerations

The Python module presents a menu with several options, including adding lockers, and parcels and calculating the load. It is important to note that when adding lockers and parcels, the datasets to be added cannot have tuples that have already been added. One of the advantages of DuckDB is being able to read a CSV dataset directly into a table quickly and therefore that is what it is doing. When you try to insert a tuple that is already included in the database table, an error message is displayed and you need to find out which tuple already exists in the database. The exception is one of the methods for adding lockers which is for incomplete datasets, in this case for lockers that do not have `parish`, `municipality`, `code INE`, and information on the nearest locker. Given the limitation of the geoapi.pt mentioned, two methods were created for adding lockers. By keeping this database up to date, it is easy to manually enter missing data in order to use the complete locker addition method. As the `Parcels` dataset represents a history, the parcels added need to have parcels relating to different periods. When calculating the load, there are a couple of considerations:

(i) insert the dates of the parcels' interval to be processed, ensuring that they have not yet been processed, otherwise, it will give an error because the pair `<locker id, date>` already exists in the database.

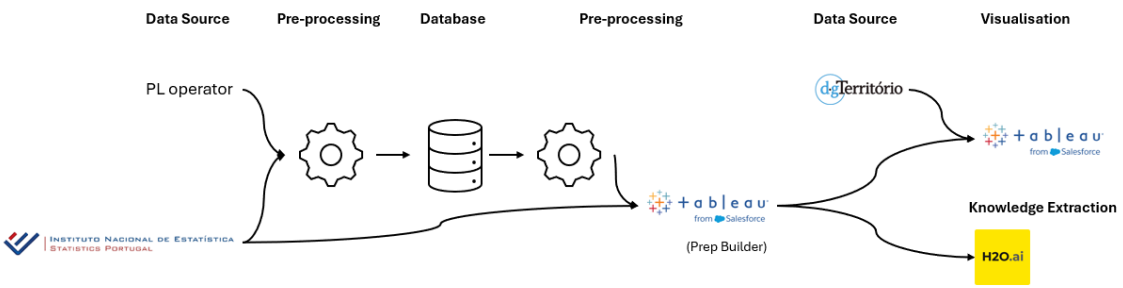


Figure 3.3: *Overall architecture.*

(ii) to process different years, separate the calculation request, that is, if you want to calculate loads from October 2023 to February 2024, make two requests, one from October to December 2023 and the other from January to February 2024.

3.8 Overall architecture

The solution entails creating a system designed to gather and store data from multiple information sources. The primary goal is to extract valuable insights using visualisations and predictive models.

The solution is presented in Figure 3.3 and it is divided, as such:

(i) Data Source, where the data sourced to be used are represented;

(ii) Pre-processing involves processing the data to ensure it can be integrated into the graphical visualisation tool and the knowledge extraction tool;

(iii) Database, where lockers data will be stored, namely `Lockers`, `Parcels` and `Locker Loads`;

(iv) Visualisation allows for the graphical visualisation of the data;

(v) Knowledge Extraction enables the development and implementation of predictive models.

4 Results

This chapter analyses the data collected and the population characteristics to investigate and identify the PLs acceptance factors, based on the datasets detailed in the previous chapter. As mentioned, we are using Tableau to create visualisations and build dashboards.

It is divided into six sections. The first section visually describes the variables regarding Portugal as a unit, and the second details some parishes. The third focuses on the turnover trend, while the fourth regards the visualisation of loads. Section five discusses the calculated load and the PL operator load. The last section considers the application of machine learning algorithms to the datasets.

4.1 Portugal

The first visualisations are maps of Portugal, Figure 4.1 and 4.2, showing the parishes that have lockers and how many they have. These images allow us to identify which areas of Portugal have or don't have lockers, but more details are required.

To better understand how the lockers are distributed regarding the number of boxes and the population of each parish, one creates three dashboards showing the boxplot distribution for each variable.

Figure 4.3 shows a positively (right) skewed distribution, where the median number of boxes is 33, the lower whisker is 16 and the upper is 115. The upper hinge is 56 and the lower is 16. As seen, there are a lot of outliers, reinforcing that some parishes have more boxes available than others.

Figure 4.4 also shows a positively (right) skewed distribution, where the median number of residents is 11,912, the lower whisker is 397 and the upper is 50,806. The upper hinge is 23,261 and the lower is 4,747. The higher outliers represent the more heavily populated parishes, such as *Algueirão-Mem Martins*, *União das freguesias de Cascais e Estoril*, *Odivelas*, *São Domingos de Rana*, to name a few.

And finally, Figure 4.5 shows a positively (right) skewed distribution, with a medium of 696 residents/km^2. The lower whisker is 4, the higher is 5,023, the upper hinge is 2,135 and

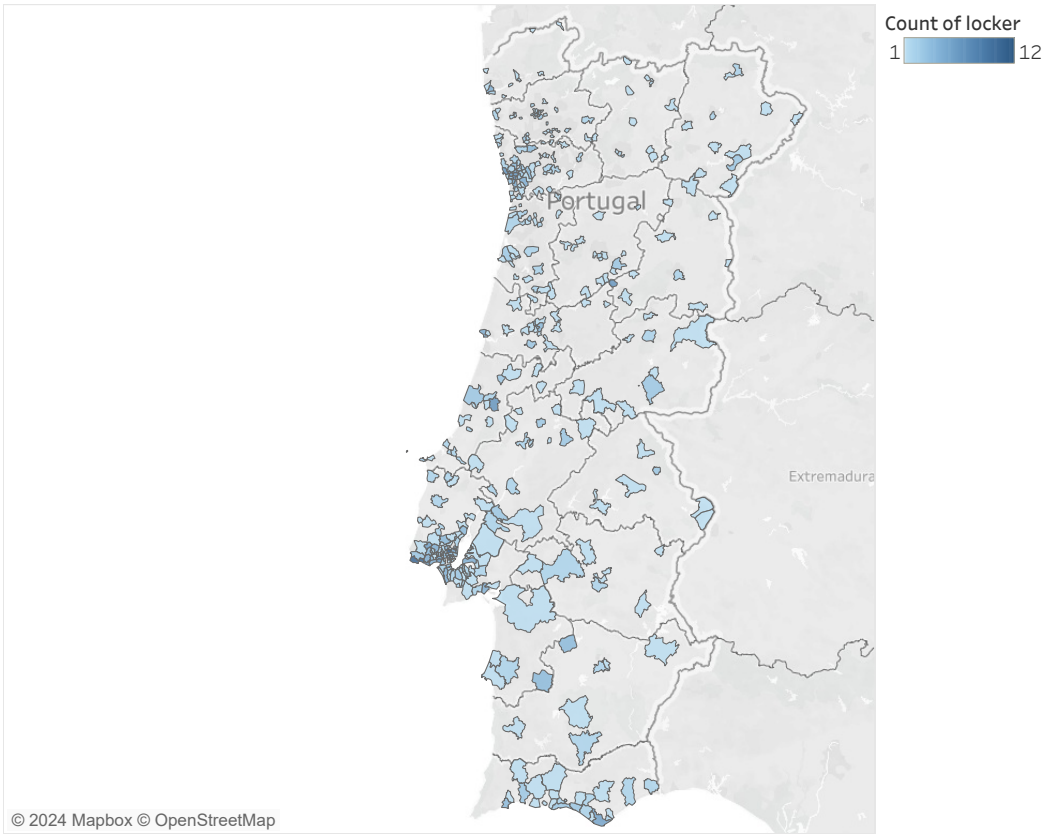


Figure 4.1: *Map view of the number of lockers distribution throughout Continental Portugal.*

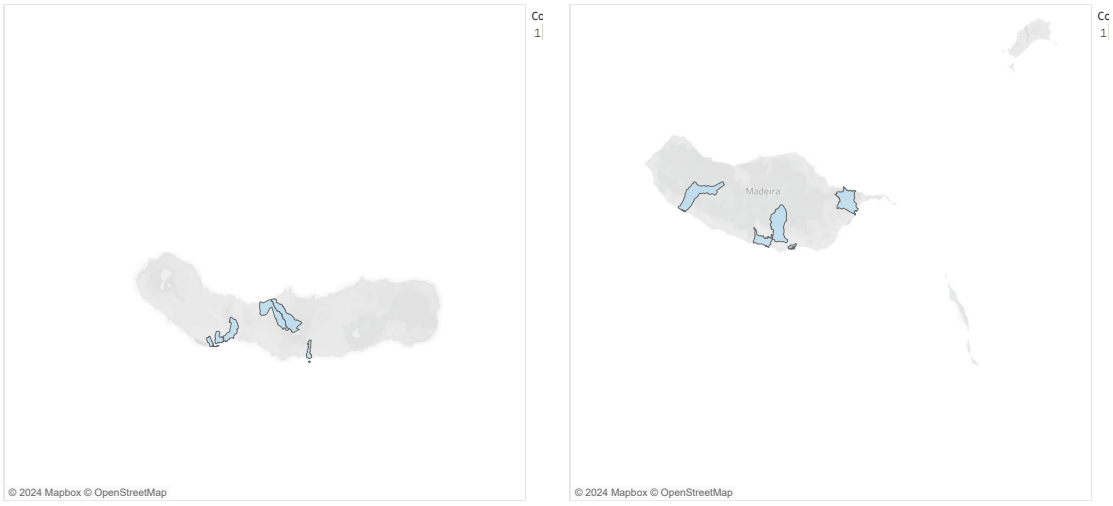


Figure 4.2: *On the left, a map view of the number of lockers distributed throughout the Autonomous Region of Açores, showing only Ilha de São Miguel; on the right, the Autonomous Region of Madeira. Both use the same scale of Figure 4.1.*

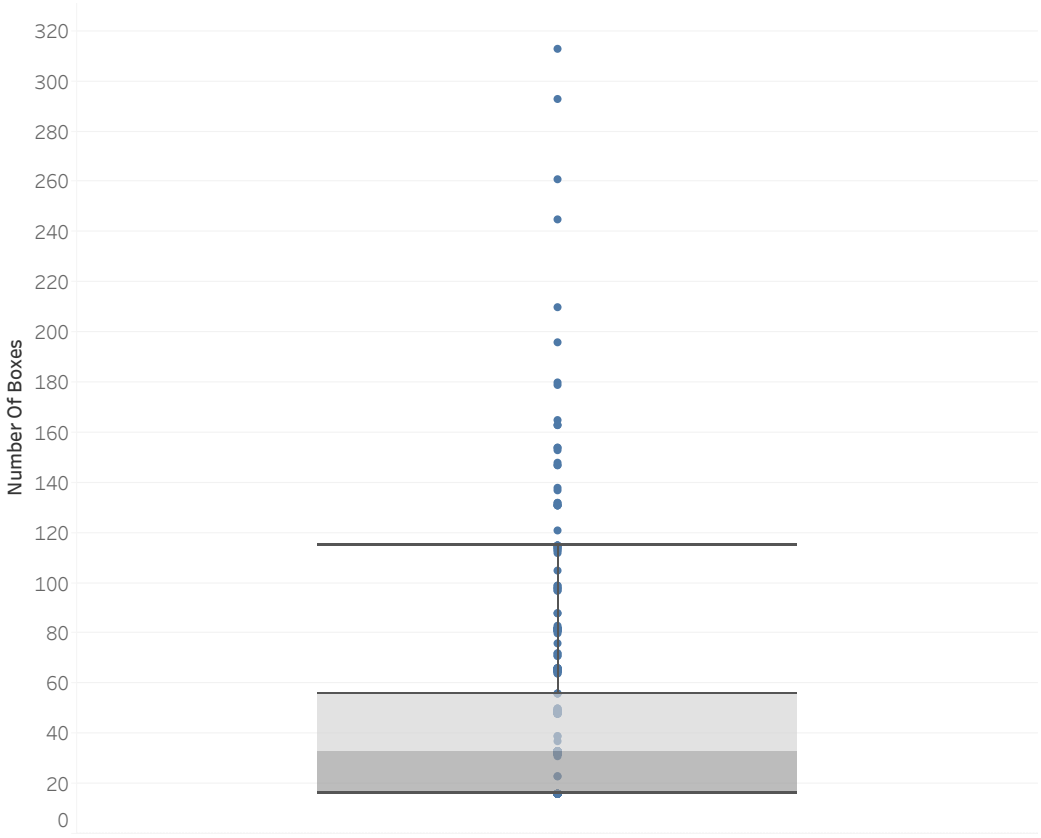


Figure 4.3: *Boxplot distribution of the number of boxes of each parish.*

the lower is 207. Here, we also have outliers different from the parishes in Figure 4.4. This indicates that we have different population distributions regardless of how many residents live in a parish.

Another interesting way of seeing these distributions is through a scatter plot, to see the relation between the number of boxes and the population. In Figure 4.6, one can see that in the parishes with fewer residents, the number of boxes available is more levelled, whereas where there are more residents we start to see a more scattered distribution. This dashboard allows seeing which parishes have more boxes, such as *União das freguesias de Cascais e Estoril*. In Figure 4.7, one notices the same pattern where when the population density is lower, the distribution of boxes is more levelled. What is interesting here is that one can visually see which parishes have more boxes available considering the number of residents and the area of that parish.

These visualisations effectively highlight the disparities in PL availability across different regions, with remote areas exhibiting a significantly higher density of boxes. These serve as a tool for identifying areas that may be under-serviced or over-serviced with PLs. They underscore the importance of considering geographic location and population characteristics when evaluating PL acceptance and utilisation, offering valuable insights into the distribution and accessibility of PLs across Portugal.

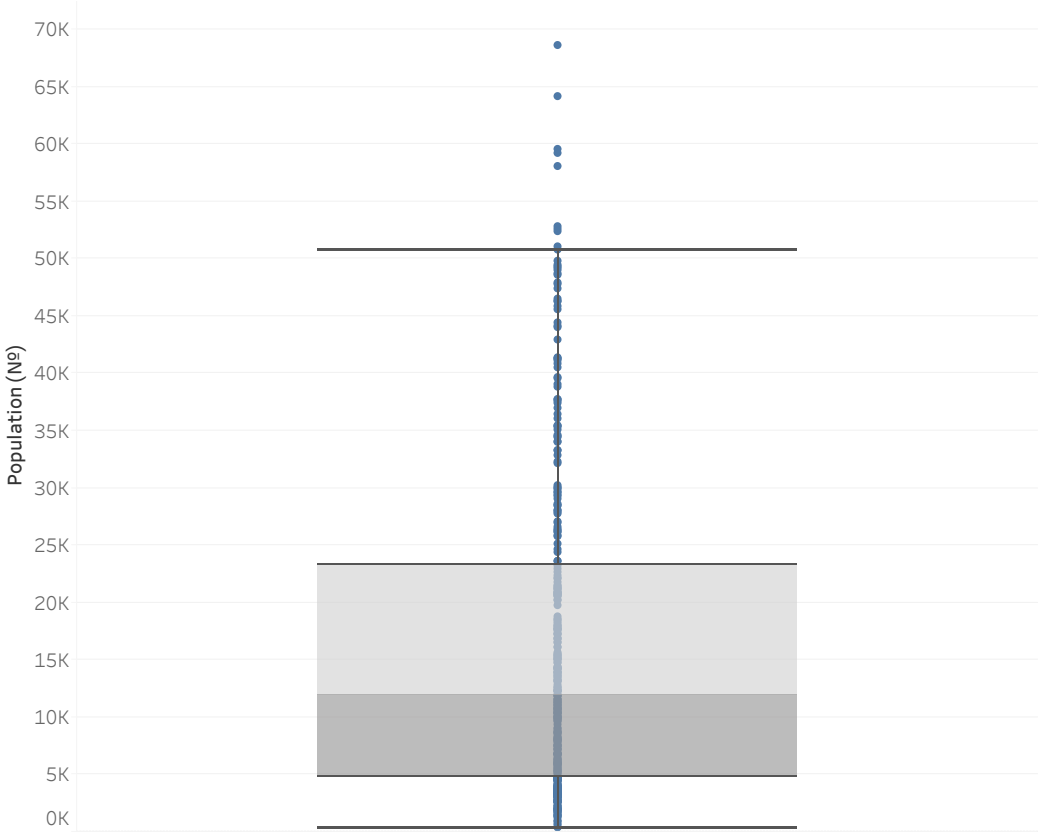


Figure 4.4: *Boxplot distribution of the population of each parish.*

Table 4.1 shows summary statistics on parcel deliveries in Portugal differentiating between weekdays and weekends, during the period studied. The main indicator in this study is the load of a locker, adapted from Ranjbari et al. [24] as previously mentioned. However, the study conducted by Ranjbari et al. also measured the number of rejected packages and its study is contained to a single residential building in a highly populated city. As such, this statistics summary also focuses on the degree of urbanisation of the parish the locker is located in, which can go from densely populated areas (ADP) to intermediate areas (AMP) and thinly populated areas (APP). On average, couriers delivered 20,952.2 parcels daily. Depending on how densely populated a parish is, the number of parcels delivered varies, as well as the peak delivery and withdrawal hours. The peak delivery hours for ADP were between 10:00 and 11:00, and AMP and APP are between 9:00 and 10:00, on weekdays. On the weekends, those times are 10:00 for ADP, 9:00 for AMP and 16:00 for APP. The peak withdrawal hours are 17:00 for weekdays and 11:00 for weekends, regardless of how densely populated a parish is.

Crossing this information with that on Table 4.2 allows for a better understanding of the type of place the lockers have been positioned so far and how that relates to the population of that area. In densely populated and intermediate areas there are more lockers in supermarkets, shopping centres and PL operator stores. Yet, in thinly populated areas, there are more lockers

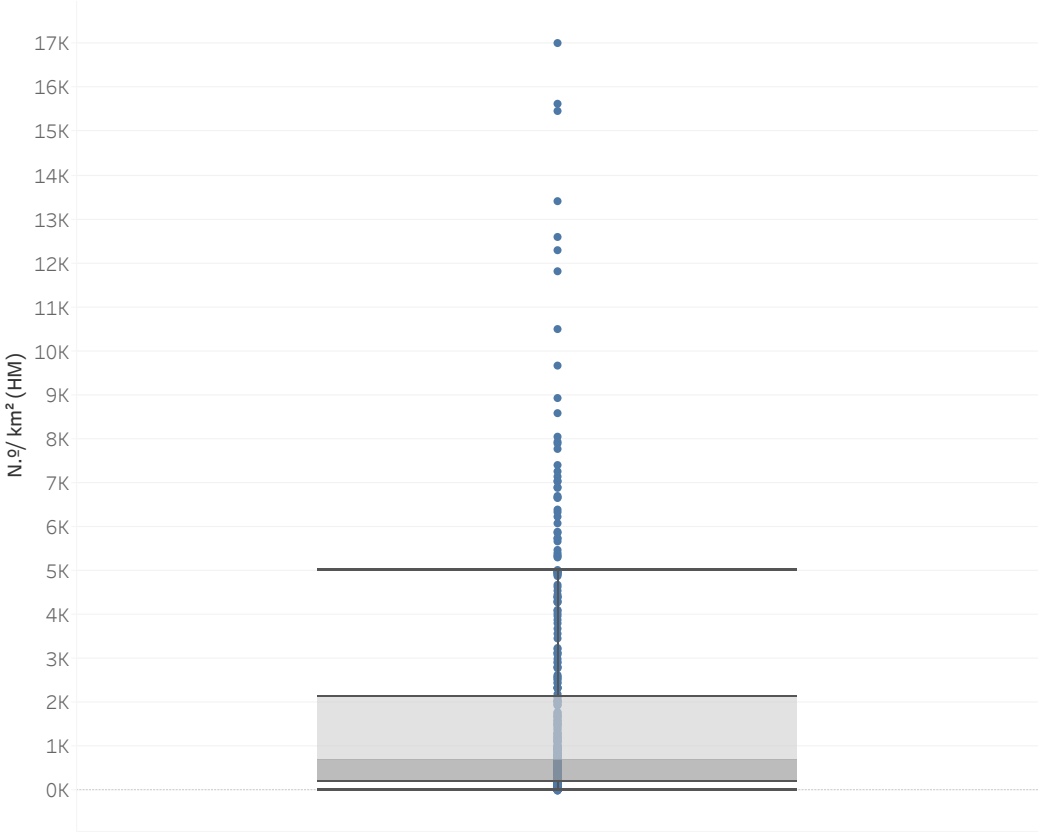


Figure 4.5: *Boxplot distribution of the population density of each parish.*

in PL operator stores, supermarkets and service stations.

The same table is used to relate with the variables used in the selection of PLs locations, discussed in 2.1. The `zone` attribute does not give complete insight into the availability of each locker, however, since most supermarkets and shopping centres do not function 24/7, it is safe to assume that those lockers are also unavailable while those establishments are closed. Assuming the PL operator stores lockers are accessible from outside the store, those are available all the time. The same can be said for the lockers placed in a service station since those are usually also placed to the side of the service station. Even in more thinly populated areas, the lockers are placed in accessible places that supposedly have more traffic. Regarding the remainder variables, there is not enough information in the datasets to discuss them.

4.2 Population characteristics and lockers

Instead of displaying the population characteristics we are studying (age group, level of education, employment status) for each parish, we will show those for a few selected parishes. We have to make this cut because, since Portugal has 3091 parishes (plus Corvo) and there are more than 700 lockers, it is impossible to look at every single one of them individually.

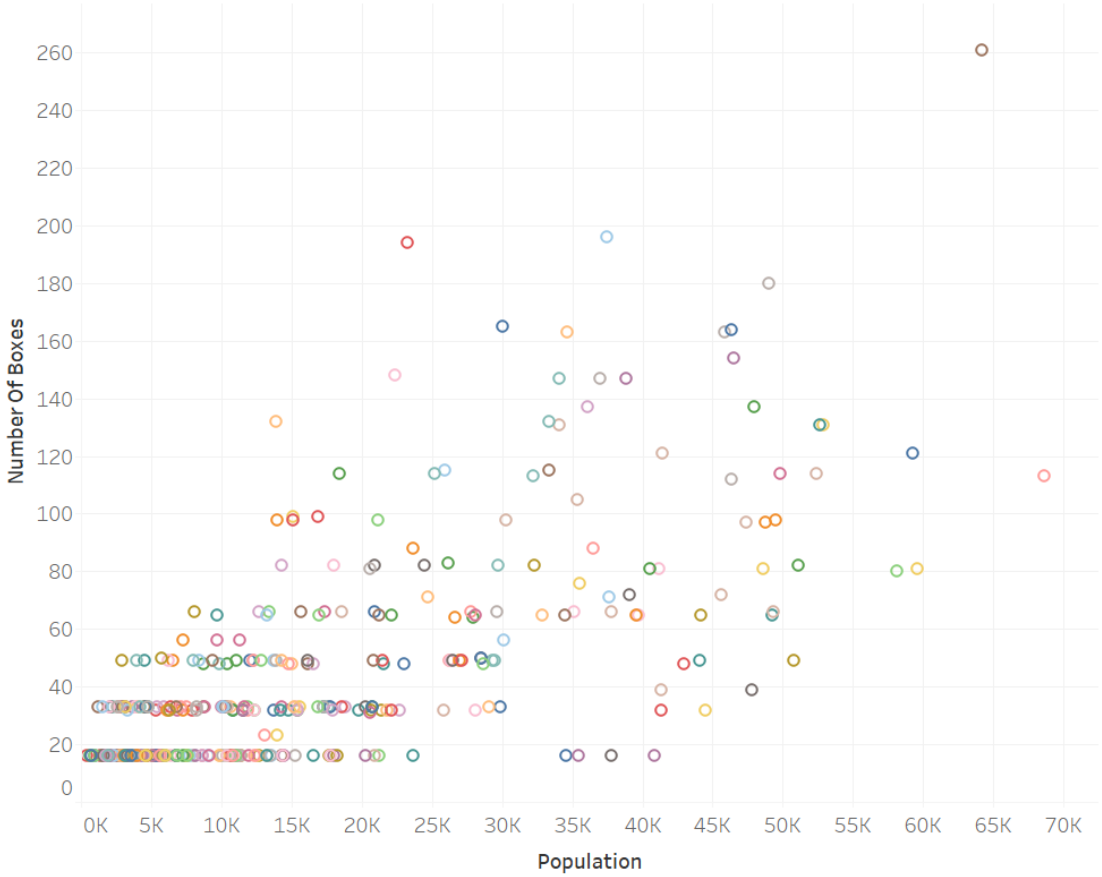


Figure 4.6: *Scatter plot distribution of the population of each parish regarding the number of boxes.*

Table 4.1: Summary statistics on parcel deliveries, in Portugal, during September-October 2023, filtered by degree of urbanisation.

	Weekdays			**Weekends**			**Overall**
	ADP	AMP	APP	ADP	AMP	APP	
Total parcels delivered	130,091	65,891	6,598	2,034	728	16	205,358
% of all deliveries	63.35	32.09	3.21	0.99	0.36	≈ 0	100
Peak entry hour	10:00	9:00		10:00	9:00	16:00	9:00
Peak exit hour	17:00			11:00			17:00
Mean entries per day	26,018.2	13,178.2	1,319.6	1,017	364	8	20,952.2
Mean exits per day	24,204.6	12,364.4	1,241.4	5,551	2,398.5	203.5	22,981.7
Mean max. load per day	0.13	0.11	0.05	0.14	0.12	0.06	0.11

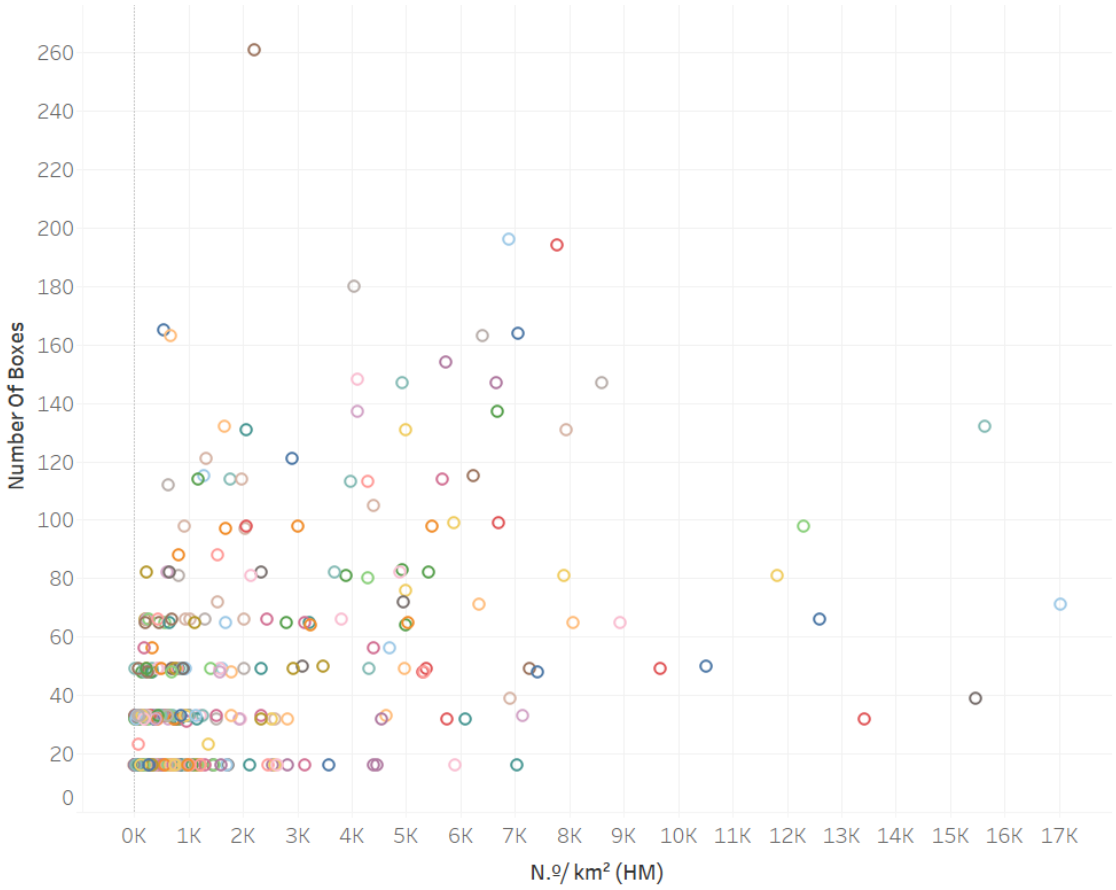


Figure 4.7: *Scatter plot distribution of the population density of each parish regarding the number of boxes.*

Table 4.2: Summary of locker zone distribution, in Portugal, during September-October 2023, according to the degree of urbanisation.

Zone	**Degree of urbanisation (%)**		
	ADP	**AMP**	**APP**
Supermarket	30.56%	32.15%	26.26%
Shopping Centre	19.04%	27.02%	4.91%
PL Operator Store	17.68%	13.79%	35.18%
Service Station	10.92%	12.65%	18.17%
Commercial Park	4.14%	4.21%	3.22%
Laundry	3.71%	0.15%	
Car Park	3.70%	1.46%	
Hospital/Health Centers	2.71%	0.15%	
University	2.47%	0.17%	
Transport	2.07%	2.62%	4.25%
Municipal Space	2.06%	3.98%	7.43%
Office	0.56%		
Pharmacy	0.37%		0.05%
Auto Workshop		0.54%	
Industrial Zone		0.48%	
PL Operator Point		0.28%	
Condominium/Residential		0.21%	
Operational Center		0.13%	0.54%

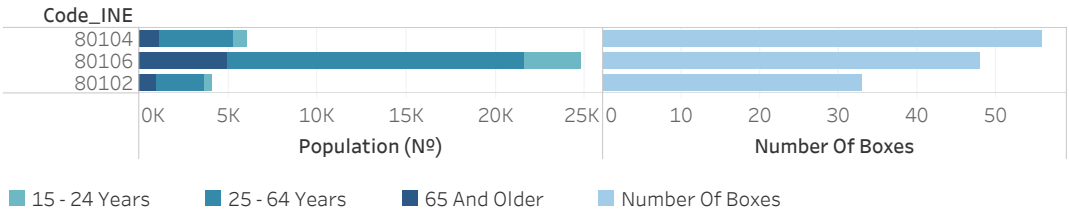


Figure 4.8: *Distribution of the population by age group of each parish for Albufeira, ordered by descending the number of boxes.*

We could decide which parishes to analyse with the previously calculated parish locker load, eq. (3.1) and in a dashboard displaying the loads of each locker per parish, filter them to display the parishes where the load of the parish is higher than 0.8. This limit is an arbitrary threshold that the PL operator also uses to monitor the lockers. However, these exclude interesting parishes resulting in a dashboard with less used lockers since it is looking at a parish. One could use the locker load of individual lockers and pick the parishes that reach that threshold. The problem with that approach is that a big list of parishes is obtained. Alternatively, only a few municipalities are selected, where the number of residents is known and also have more displaced students because of local universities. By filtering through municipalities, one still gets parishes with fewer residents and from more remote areas. Thus, the selected regions are the following: *Albufeira*, *Castelo Branco*, *Coimbra*, *Lisboa* and *Porto*. For each, a dashboard shows the lockers' load in that parish, stacked age group (with the number of lockers), stacked education level and employment status. For better reading of the dashboards, the dashboards show the `code INE` instead of the parish name - some parishes have long names making the visualisations unreadable, as such, to know which parish is which, any of the datasets, regarding the population, from INE contain both the code and the parish name. Furthermore, only locker load is analysed in this section, but it is possible also to look at locker turnover.

4.2.1 Albufeira

According to Figure 4.8, this municipality has six lockers across three parishes. The residents in these parishes are mainly aged between 25 and 64 years old, followed by 65 and older. The parish with more residents is *Albufeira e Olhos de Água*, code 080106, but despite that it is not the parish with more boxes available, which is *Ferreiras*, code 080104.

Despite differences in the age group, those parishes show similar percentages by education level, where between 65% to 70% have, at least, the 3rd cycle of basic education, 45% and 50% of the residents have, at least, completed high school, and between 15% and 20% have finished university, as can be seen in Figure 4.9. As for employment status, most residents are employed or retired, Figure 4.10.

Figure 4.11 displays the dashboard showing the load of each locker per parish, calculated with eq. (3.1). Notably, only one locker surpasses the 0.8 load threshold in *Guia*, code 080102.

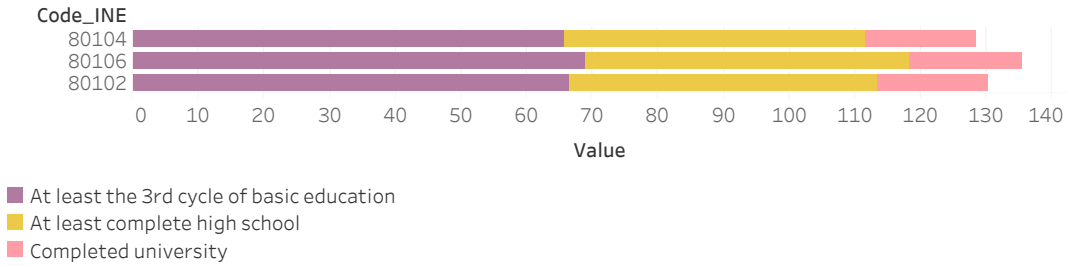


Figure 4.9: *Distribution of the population by the education level of each parish for Albufeira, ordered by descending the number of boxes.*

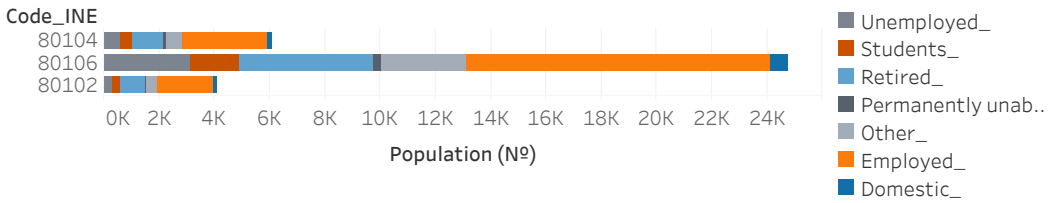


Figure 4.10: *Distribution of the population by employment status of each parish for Albufeira, ordered by descending the number of boxes.*

Note that *Guia* is neither the parish with more residents nor more boxes available. However, it is well placed in a shopping centre, as is the other locker that shows heavier use in a supermarket. Barely noticeable in the figure is one of the lockers in parish 80104 that started to have load towards the end of September, and that is because of when the locker was installed. Each line in the load figures, with a different colour, represents a different locker.

4.2.2 Castelo Branco

Figure 4.12 displays the age of the parishes where lockers in this municipality exist (four lockers in two parishes). We can see that *Castelo Branco*, 50205, has much more residents than *Alcains*, 50201, both with more residents aged between 25 and 64 years old.

Regarding the education level, Figure 4.13, between 55% to 75%, at least, the 3rd cycle of basic education, 40% and 55% of the residents have, at least, completed high school, and between 15% and 30% have finished university.

As for employment status, Figure 4.14, these follow the same pattern where the majority of residents are employed or retired, followed by students.

Figure 4.15 displays the dashboard showing the load of each locker per parish, calculated with eq. (3.1). Notably, only one locker surpasses the 0.8 load threshold in *Alcains*. This is the parish with fewer residents but it also has fewer boxes available.

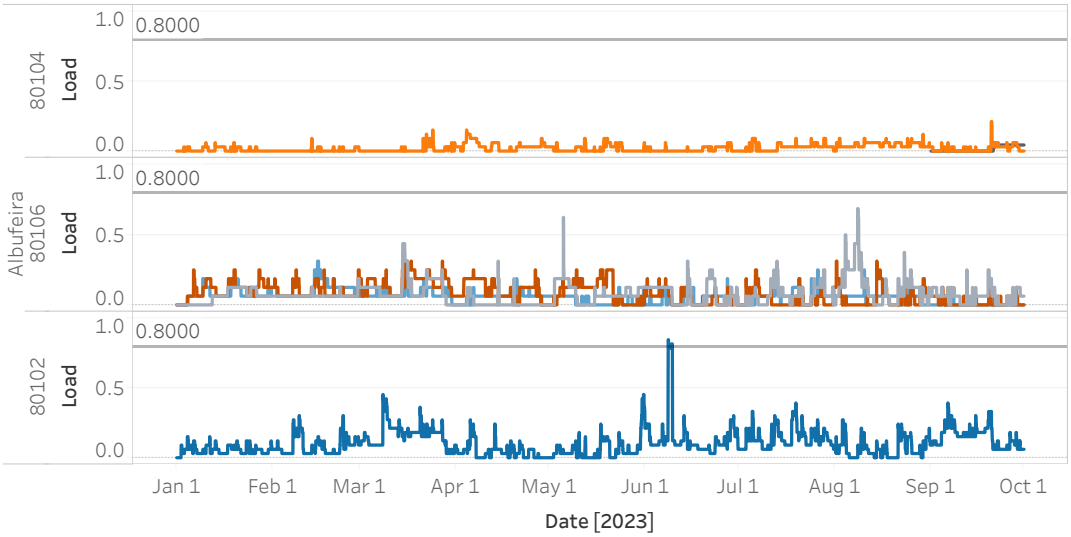


Figure 4.11: *Locker load, ordered by parishes with the highest number of boxes for Albufeira.*

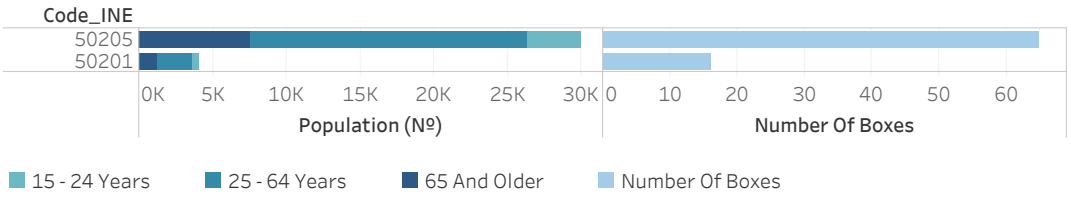


Figure 4.12: *Distribution of the population by age group of each parish for Castelo Branco, ordered by descending the number of boxes.*

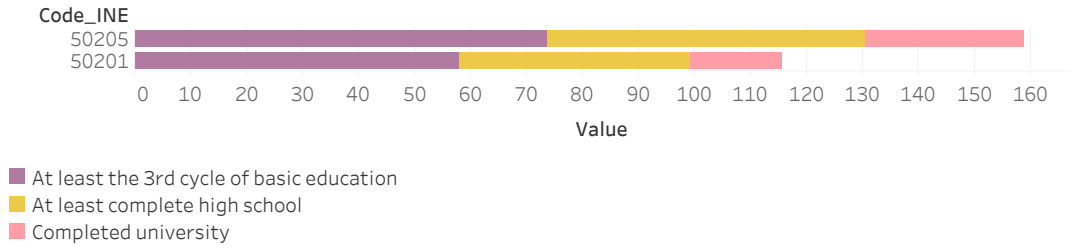


Figure 4.13: *Distribution of the population by the education level of each parish for Castelo Branco, ordered by descending the number of boxes.*

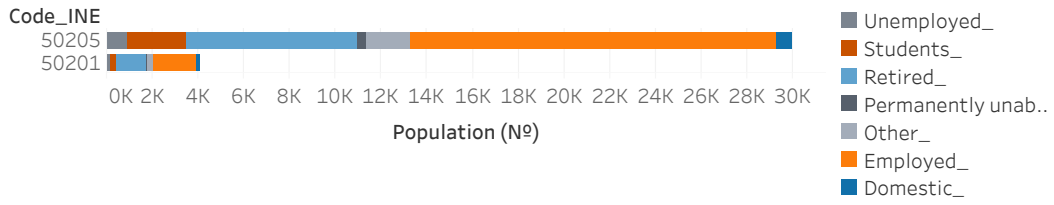


Figure 4.14: *Distribution of the population by employment status of each parish for Castelo Branco, ordered by descending the number of boxes.*

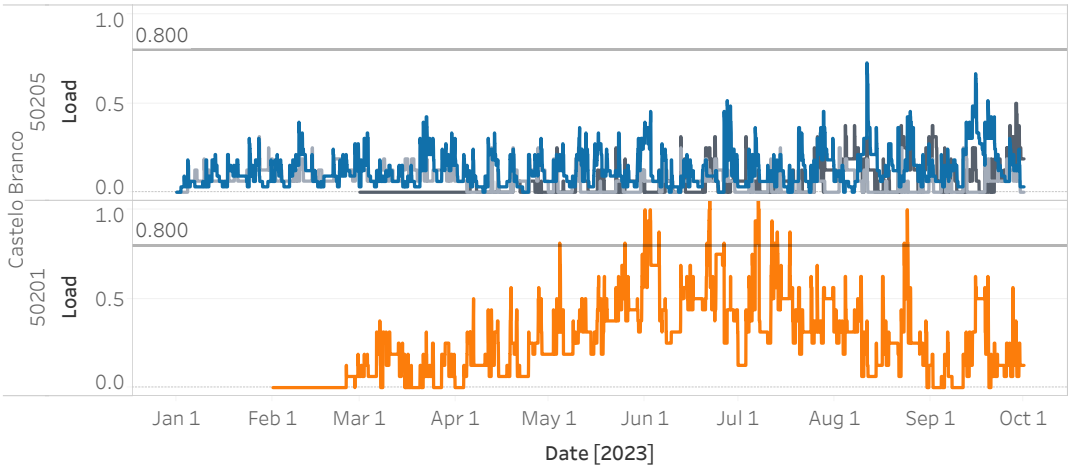


Figure 4.15: *Locker load, ordered by parishes with the highest number of boxes for Castelo Branco.*

4.2.3 Coimbra

According to Figure 4.16, this municipality has twelve lockers across six parishes. The residents in these parishes are mainly aged between 25 and 64 years old, followed by 65 and older. The parish with more residents is *Santo António dos Olivais*, code 060318, but despite that it is not the parish with more boxes available, which is *União das freguesias de Coimbra (Sé Nova, Santa Cruz, Almedina e São Bartolomeu)*, code 060334.

Despite differences in the number of residents, these parishes show similar percentages by education level, where between 60% to 85% have, at least, the 3rd cycle of basic education, 40% and 75% of the residents have, at least, completed high school, and between 20% and 55% have finished university, as can be seen in Figure 4.17.

Figure 4.18 shows the employment status of this municipality's residents. These follow the same pattern where the majority of residents are employed or retired, followed by students.

Figure 4.19 displays the dashboard showing the load of each locker per parish, calculated with eq. (3.1). Only one locker surpasses the 0.8 load threshold in *Santo António dos Olivais* and the locker is placed in a shopping centre.

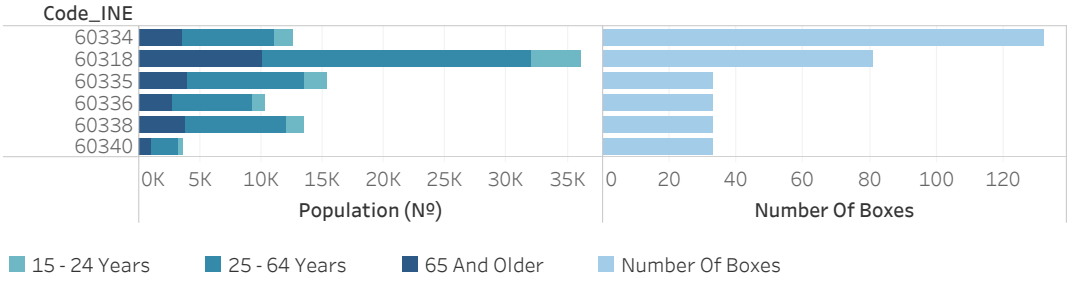


Figure 4.16: *Distribution of the population by age group of each parish for Coimbra, ordered by descending the number of boxes.*

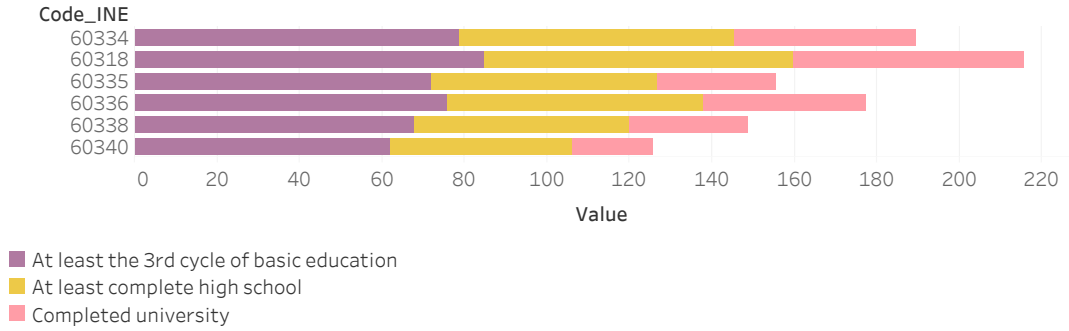


Figure 4.17: *Distribution of the population by the education level of each parish for Coimbra, ordered by descending the number of boxes.*

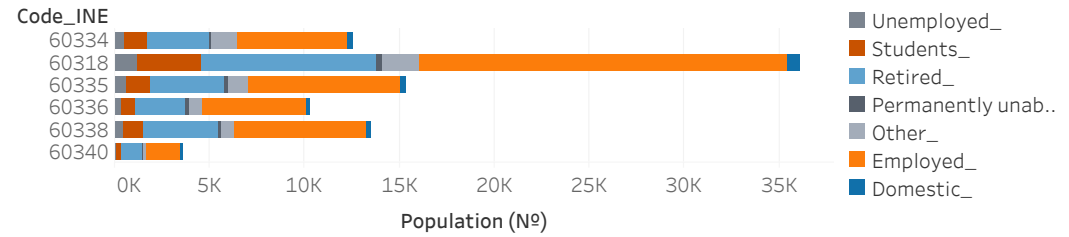


Figure 4.18: *Distribution of the population by employment status of each parish for Coimbra, ordered by descending the number of boxes.*

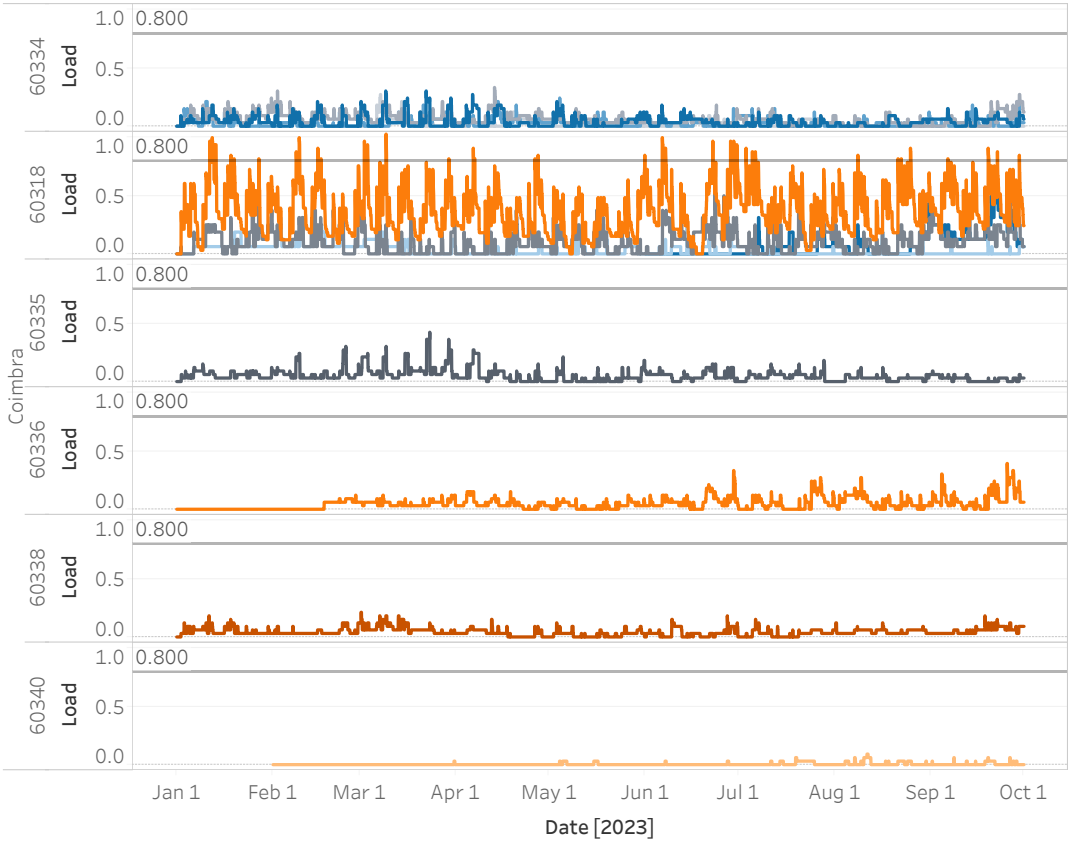


Figure 4.19: *Locker load, ordered by parishes with the highest number of boxes for Coimbra.*

4.2.4 Lisboa

As for Figure 4.20, this municipality has seventy-five lockers across twenty-one parishes. The residents in these parishes are mainly aged between 25 and 64 years old, followed by 65 and older. The parish with more residents is *Avenidas Novas*, code 110618, and also the parish with more boxes.

Regarding the education level, Figure 4.21, shows that between 55% to 90%, at least, the 3rd cycle of basic education, 35% and 80% of the residents have, at least, completed high school, and between 15% and 60% have finished university.

As for employment status, Figure 4.22, these follow the same pattern where the majority of residents are employed or retired, followed by students.

Figures 4.23, 4.24, 4.25 displays the dashboard showing the load of each locker per parish, calculated with eq. (3.1). This dashboard shows that eleven parishes have lockers reaching the 0.8 threshold. One of these parishes is *Parque das Naçoes*, code 110662. This is an interesting parish because, despite there being lockers near each other, we can see that some lockers have more use than others. This can possibly be justified by the location of these lockers since the lockers with more use are placed near shopping centres and supermarkets.

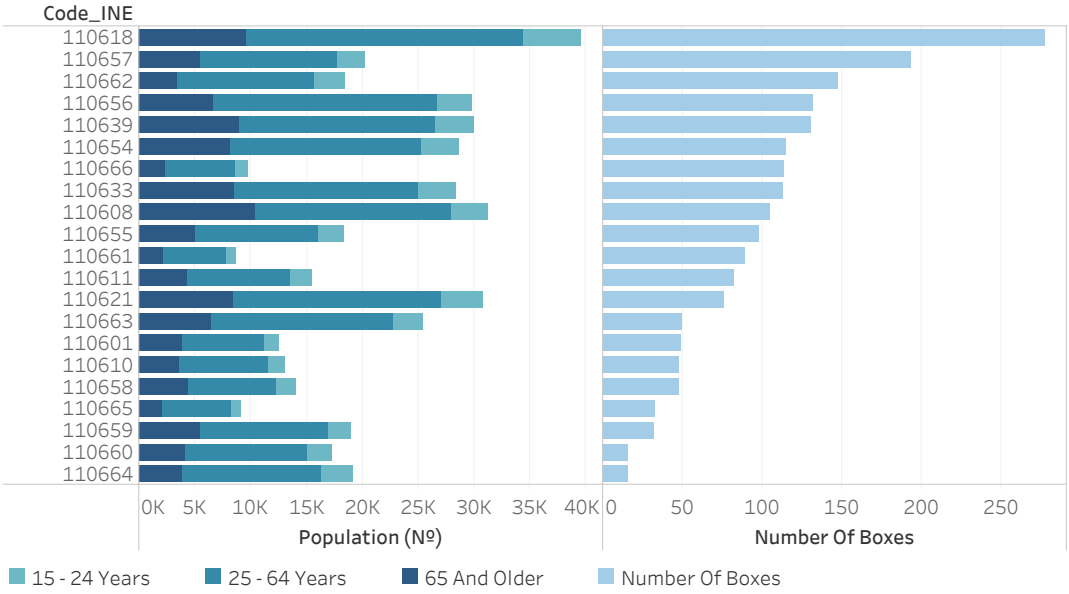


Figure 4.20: *Distribution of the population by age group of each parish for Lisboa, ordered by descending the number of boxes.*

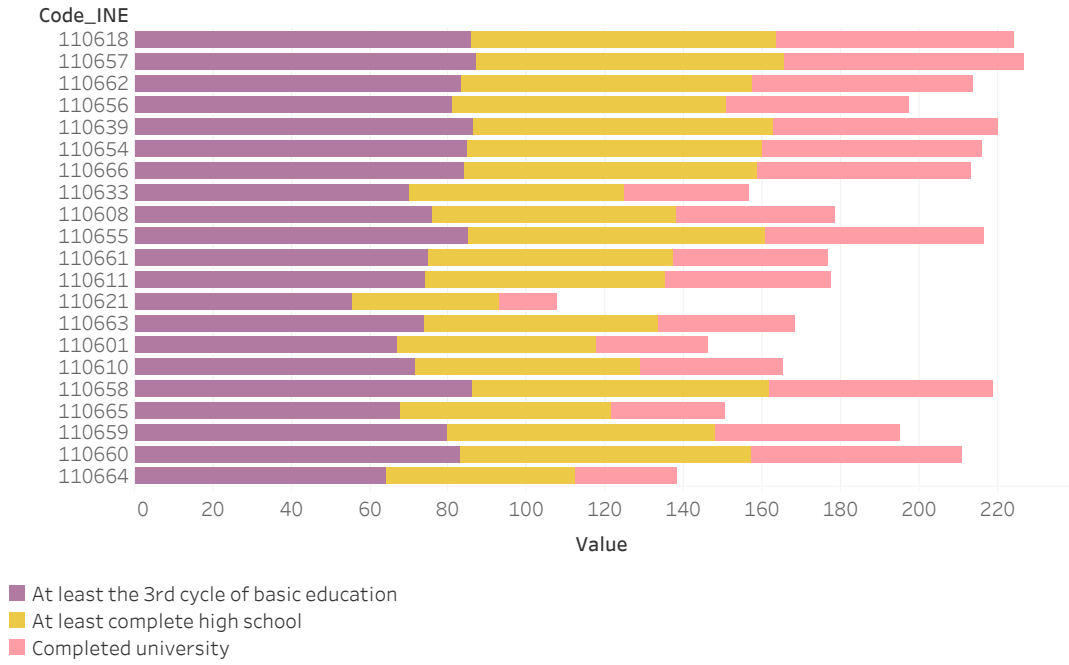


Figure 4.21: *Distribution of the population by the education level of each parish for Lisboa, ordered by descending the number of boxes.*

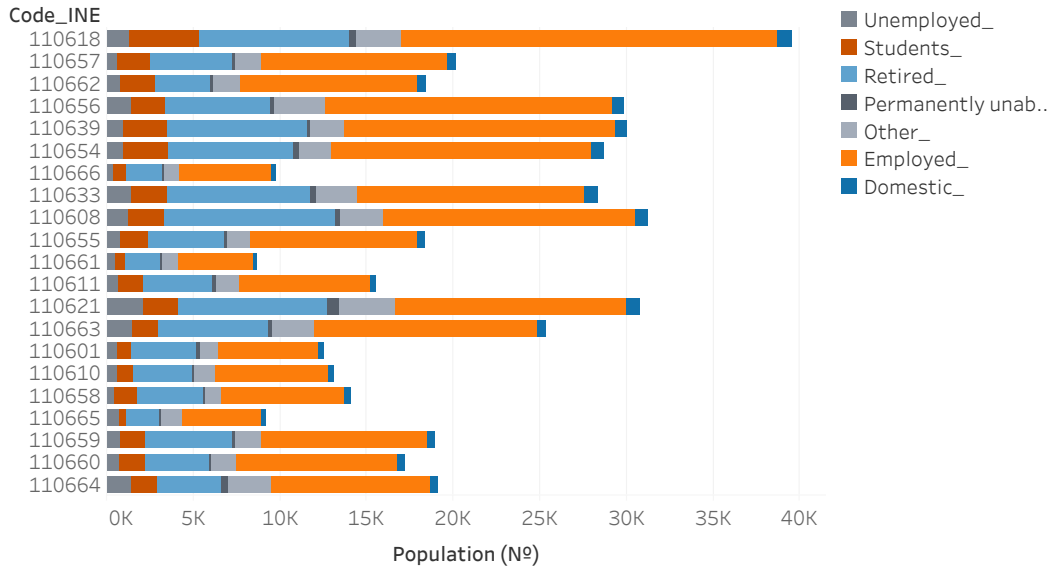


Figure 4.22: *Distribution of the population by employment status of each parish for Lisboa, ordered by descending the number of boxes.*

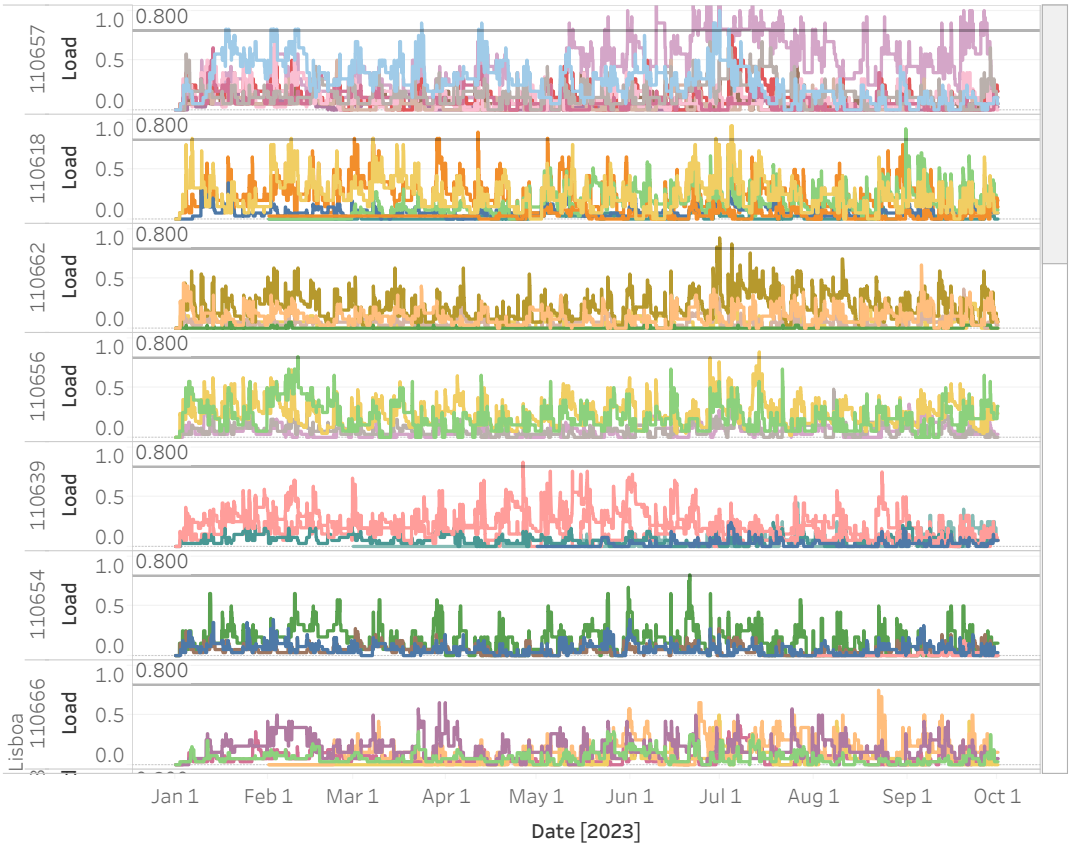


Figure 4.23: *Locker load, ordered by parishes with the highest number of boxes for Lisboa, part one.*

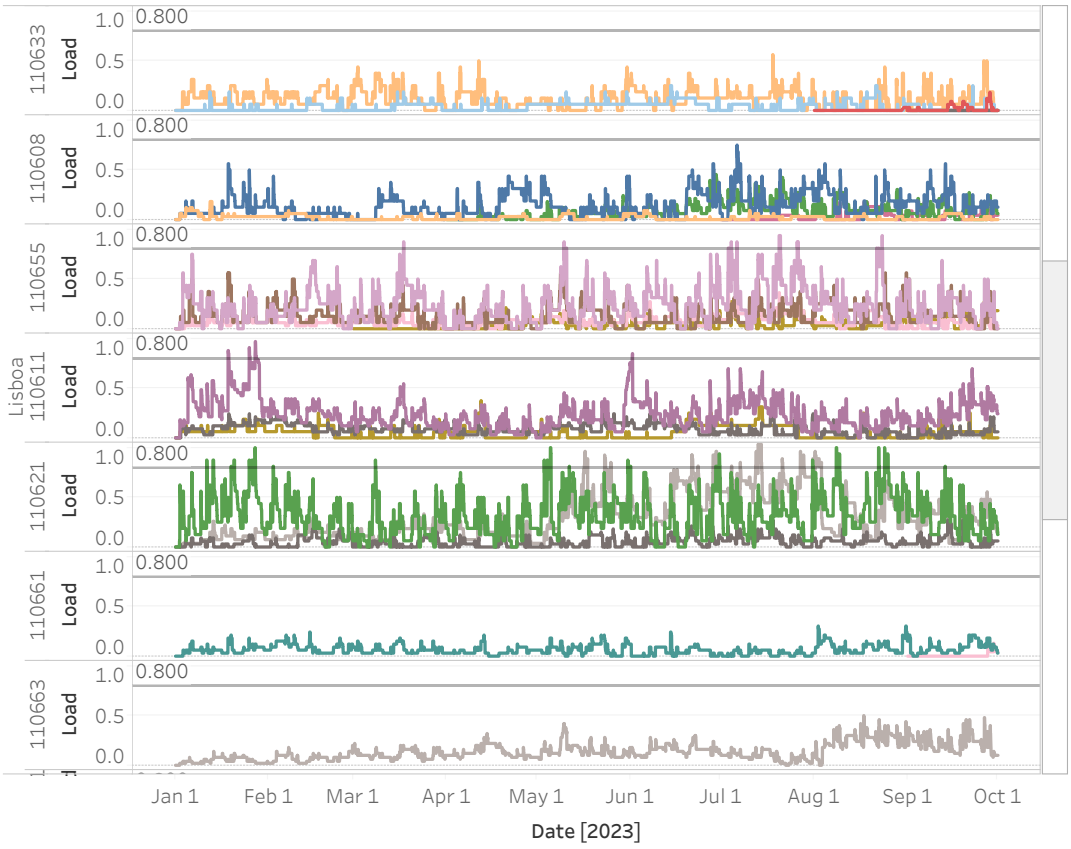


Figure 4.24: *Locker load, ordered by parishes with the highest number of boxes for Lisboa, part two.*

4.2.5 Porto

As observed in Figure 4.26, these parishes have distinct age group distributions and the parish with more boxes is not the heavier populated. The majority of the residents in these parishes are between 25 and 64 years old or older, followed by 65 and older. This municipality has thirty-three lockers across seven parishes.

Regarding the education level, Figure 4.27, shows that between 55% to 80%, at least, the 3rd cycle of basic education, 35% and 70% of the residents have, at least, completed high school, and between 15% and 50% have finished university.

The employment status, in Figure 4.28, shows that the majority of residents are employed or retired, followed by students.

Figure 4.29 displays the dashboard showing the load of each locker per parish, calculated with eq. (3.1). This dashboard shows that four parishes have lockers reaching the 0.8 threshold. Almost all of these lockers are situated in a supermarket, except one in a post office.

In summary, all these parishes have more residents aged between 25 and 64 years old, which is pre-retirement age, indicating a higher number of employed residents when compared to the retired. The employed residents outnumber the retired by more than double, with a significant

Figure 4.25: *Locker load, ordered by parishes with the highest number of boxes for Lisboa, part three.*

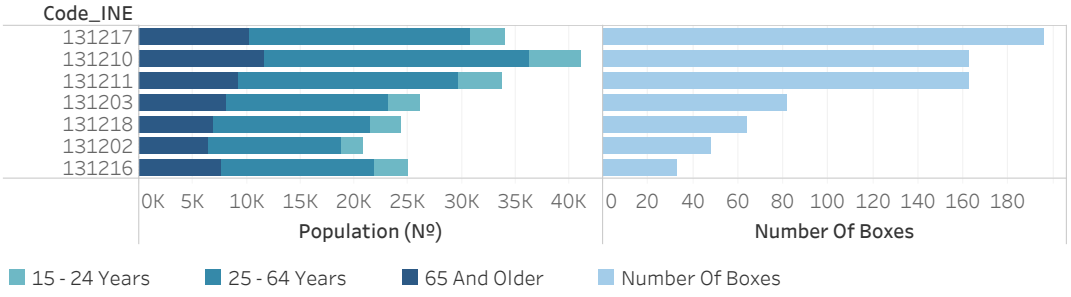


Figure 4.26: *Distribution of the population by age group of each parish for Porto, ordered by descending the number of boxes.*

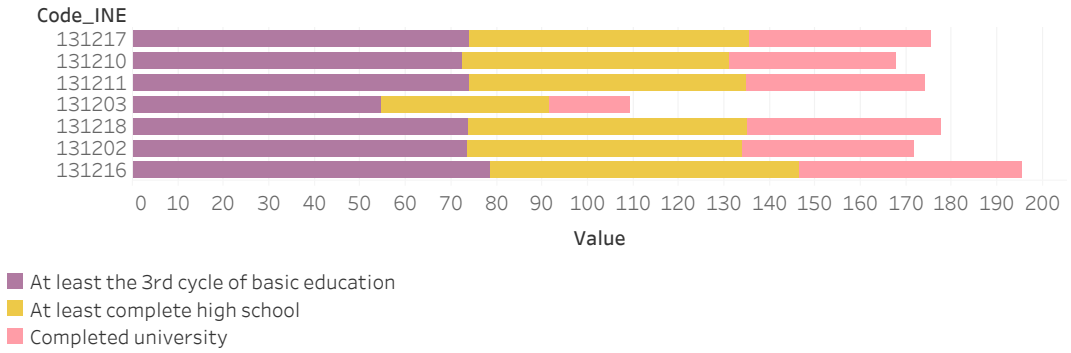


Figure 4.27: *Distribution of the population by the education level of each parish for Porto, ordered by descending the number of boxes.*

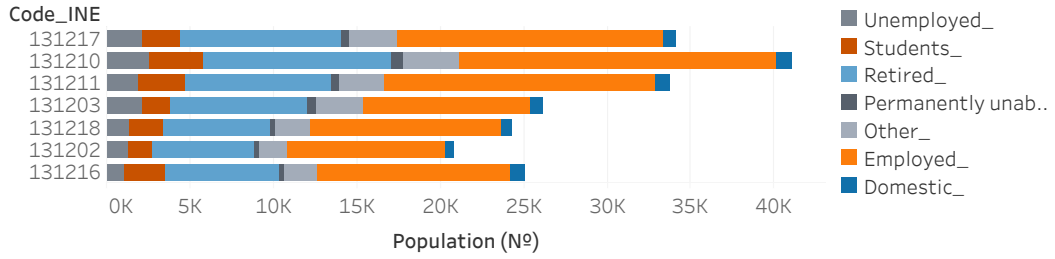


Figure 4.28: *Distribution of the population by employment status of each parish for Porto, ordered by descending the number of boxes.*

portion of the population being students. Parishes such as *Paranhos* in *Porto*, *Avenidas Novas* in *Lisboa*, *Santo António dos Olivais* in *Coimbra*, *Castelo Branco* in *Castelo Branco*, *Albufeira e Olhos de Água* in *Albufeira* have more residents and also show heavier use of lockers. Concerning the education in these parishes, they tend to have more residents completing each education level. On the other hand, some parishes with less use are also parishes with fewer residents and less education.

In B, it was added a dashboard for each of the previous municipalities showing the behaviour in deposited parcels for each month. These dashboards show that most of these parishes behave similarly monthly. One can also have dashboards showing these distributions according to the day of the week, or hour, the parcels are collected, for example.

Some parishes, where the lockers reach the 0.8 threshold multiple times, could benefit from increasing the number of boxes, either by adjusting the number of boxes in specific lockers or by adding new locations, such as in *Avenidas Novas*, *Santo Antonio dos Olivais*, and *Santa Maria Maior*.

One factor that can impact locker usage is, assuming that students use this service more, particularly, students who go to universities outside their home parish. As such, in A, there are more dashboards to complement the ones above, namely of *Braga*, *Évora* and *Faro*,

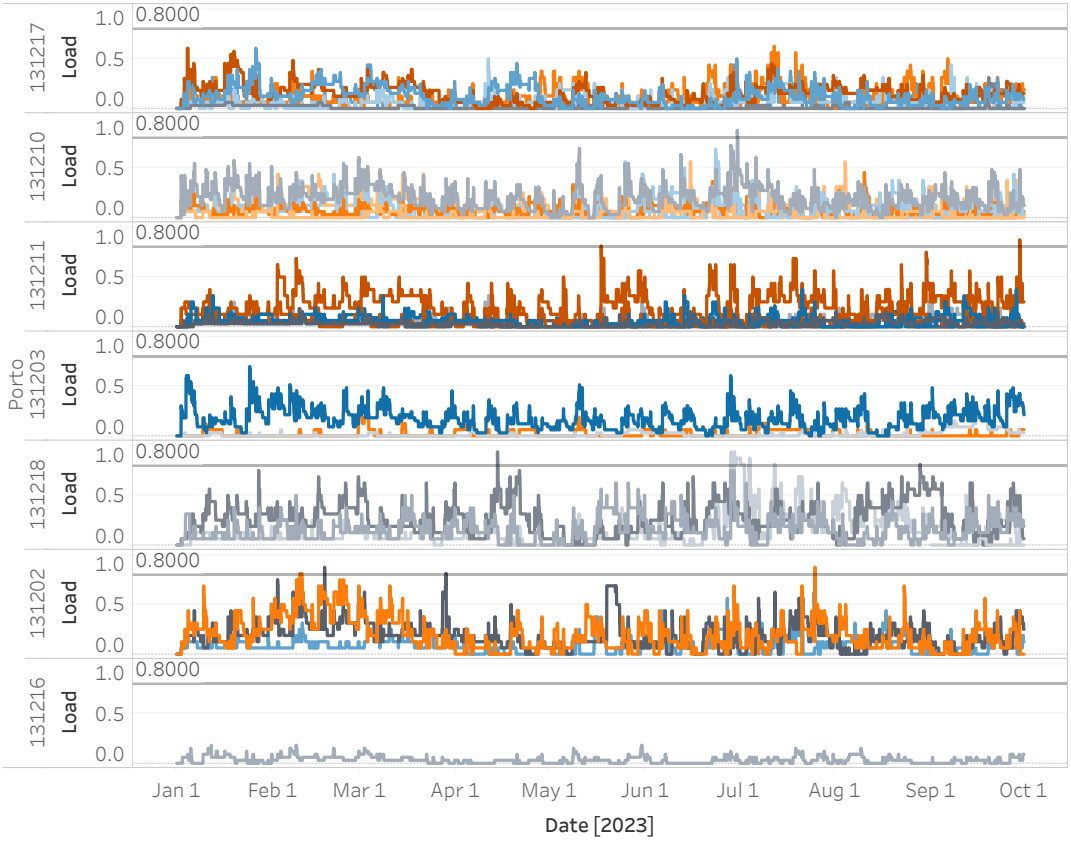


Figure 4.29: *Locker load, ordered by parishes with the highest number of boxes for Porto.*

more municipalities beyond the ones described that have universities. Another factor that can influence the use of each locker is the people who work outside of their home parish and can either choose a locker closer to work or home, for this study there are no data available to distinguish.

4.2.6 Population Income

One of the datasets regarding the population characteristics is the gross income per inhabitant. However, the lowest geographic granularity here is by municipality. As such, the number of boxes in each municipality can be related to the income, as seen in Figure 4.30. Except for *Albufeira*, the gross income seems to follow the number of boxes available and the number of deposited parcels. The top ten municipalities with more parcels is *Lisboa*, *Porto*, *Cascais*, *Sintra*, *Oeiras*, *Vila Nova de Gaia*, *Braga*, *Almada*, *Maia* and *Amadora*. If one looks at the top ten by gross income per inhabitant then, it is *Lisboa*, *Oeiras*, *Porto*, *Cascais*, *Coimbra*, *Alcochete*, *Aveiro*, *Évora*, *Matosinhos* and *Almada*. The municipalities present in these three shortlists are *Lisboa*, *Porto*, *Cascais*, *Oeiras*, and *Almada* indicating a strong relationship between gross income and locker usage. Looking at the lower end of income is trickier, however, more than half of the municipalities have less than a sixth of the deposited parcels in *Lisboa*.

Based on the gross income, a starting point when deciding where to install new lockers can be

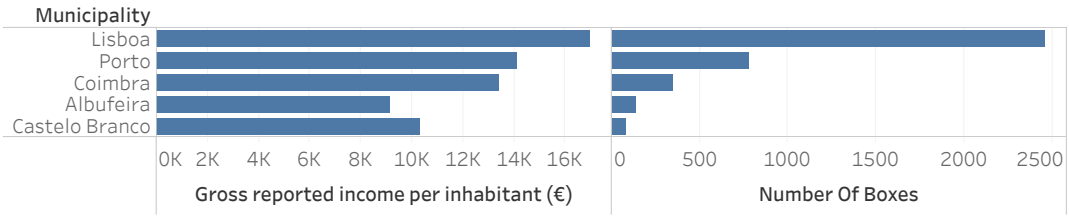


Figure 4.30: *Gross reported income per inhabitant (€) and number of boxes per municipality, ordered by number of boxes and filtered to the municipalities discussed in this chapter.*

by choosing similar municipalities. The top five municipalities that do not have lockers, at least during this study, are *Porto Santo*, *Vila do Porto*, *Arruda dos Vinhos*, *Corvo* and *Horta*, with gross reported incomes per inhabitant varying from 13,000 (€/annual) to 10,000(€/annual), approximately. From here, we could look at the characteristics of municipalities with similar gross incomes, and analyse the characteristics of the parishes, to support the decision of where to install the new lockers. For example, the closest municipality to *Arruda dos Vinhos*' income is *Sines* which only has one locker in a post office, though it was installed in August 2023, in around one month it has reached the 0.8 threshold once. This locker is placed in the parish with more residents, *Sines*, code 151301. In *Arruda dos Vinhos*, the parish with more residents is *Arruda dos Vinhos*, code 010202. Regarding the education level, parish *Arruda dos Vinhos* also has the highest values for this municipality. With this in mind, this parish might be a good fit for a locker.

4.3 Turnover trend and forecast

In Tableau, one can add trend lines to visualisations and that is what is being done in Figure 4.31. This figure shows the turnover, calculated with eq. (3.2) for the parishes of the municipalities *Albufeira* and *Castelo Branco*. In the dashboard, the user can pick any municipality and parish, however, to make reading this figure easier, only these two are shown. Looking at the locker in parish *Guia*, code 080102, the trend has a p-value of 0.0568005. Now, the locker in parish *Alcains*, code 050201, the trend has a p-value of 0.0247698. The p-value serves as an indicator of the significance of the trend line. Typically, a p-value of 0.05 or lower is deemed significant. In other words, the smaller the p-value, the more significant the model becomes, meaning that both these trends are significant. Both trends are increasing but the locker in *Guia* is growing faster.

Tableau also has a forecast option, however, since our sample period is so small, we cannot use this technique.

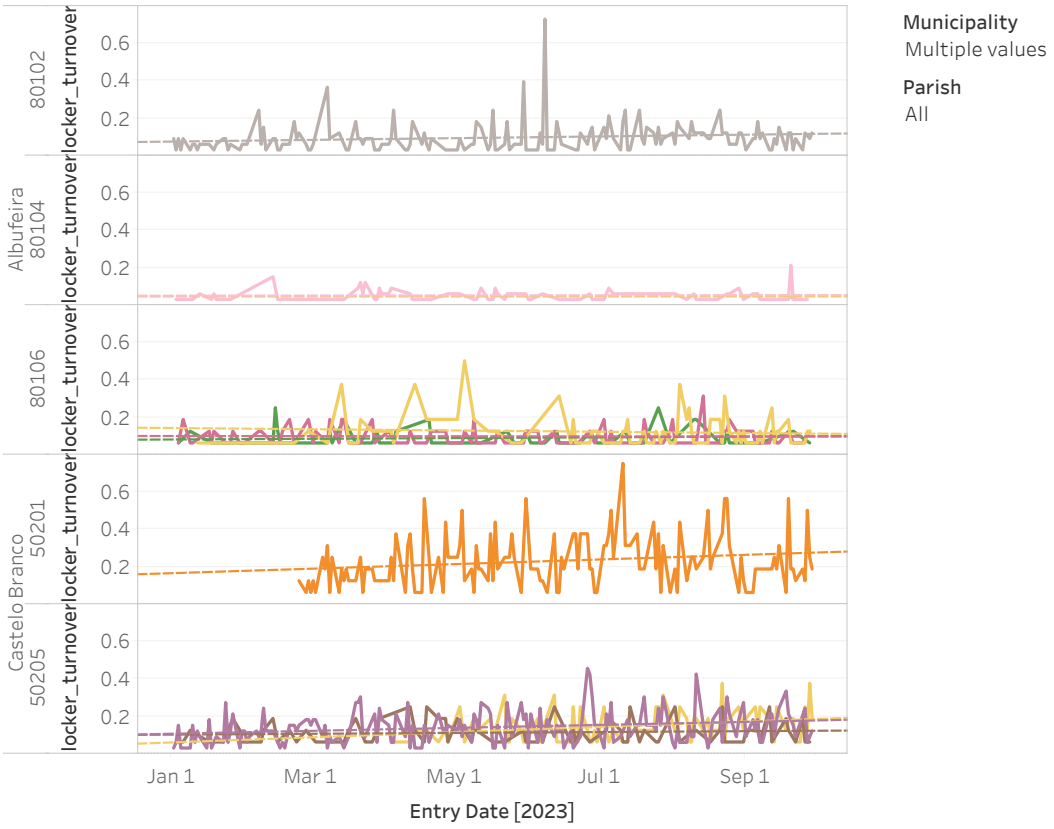


Figure 4.31: *Turnover evolution and trend from January to September 2023, filtered to Albufeira and Castelo Branco.*

4.4 Load visualisation

Another helpful visualisation is the story in Figure 4.32. This dashboard has some filters to centre and guide the analysis:

- Maximum load which can be `above 0.8`, notice that this is the same threshold used by the PL operator, `below 0.1` and in between those values.
- Table name to focus on *Continente*, *Açores* or *Madeira*.
- Municipality to filter the analyses to a specific zone, only shows municipalities where lockers exist.
- Parish to focus on all or just one parish of the chosen municipality, only shows parishes where lockers exist.

This visualisation can support the decision to install or remove lockers in a location. Depending on the PL operator guidelines and the interval they choose to study, through this story, one can see which lockers are at risk of always being almost full or which ones are rarely used. Note that this visualisation does not consider when a locker was installed, meaning that from the moment the locker exists, its load is calculated and shown here. Therefore, for the lockers

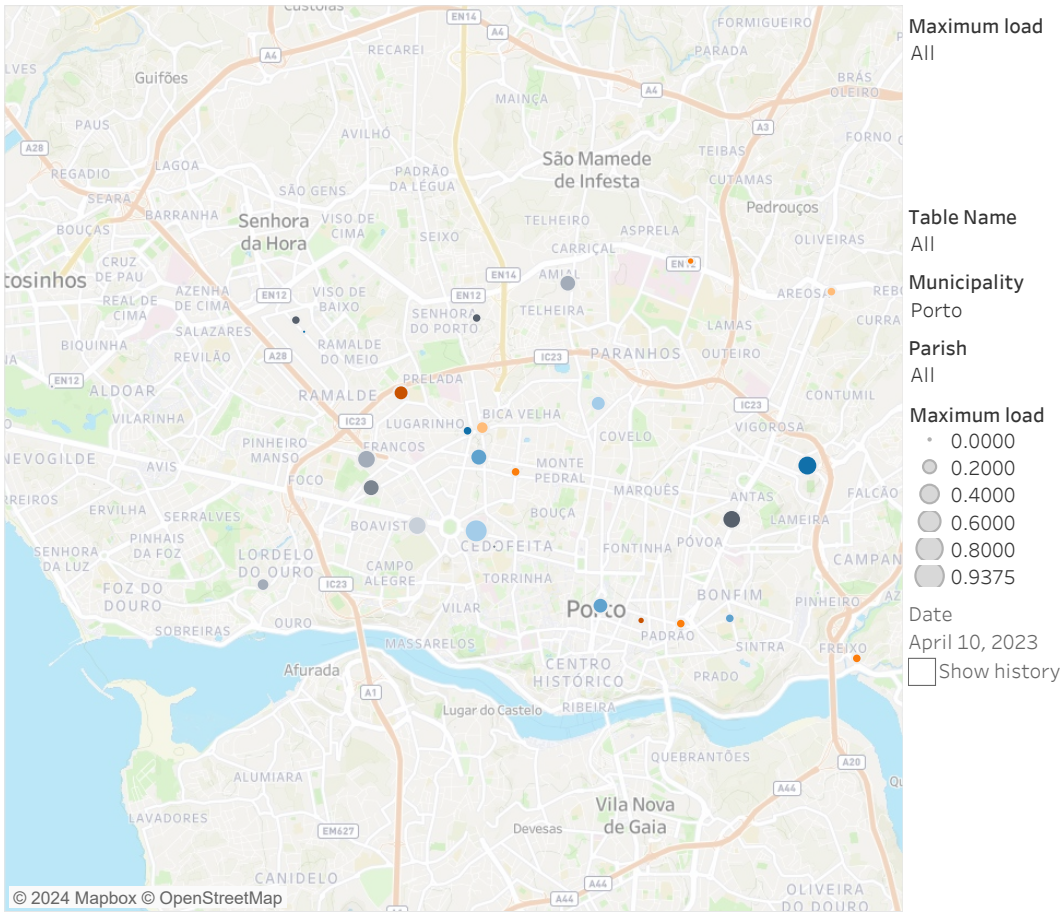


Figure 4.32: *Story describing the locker loads, in this case, filtered to the Porto municipality. All the lockers in this municipality are represented with the maximum load of April 10, 2023.*

that have loads below 0.1, several factors can justify that number, one of them being its short time in operation and the possibility that the locker is yet unknown. The same can be said for the lockers with loads above 0.8. It is easy to think first of installing more lockers or boxes, instead of removing them, since when the load is lower, the locker can always be used. The same cannot be said for when a locker is full. However, there can also be factors influencing the higher usage of lockers. Even so, with such a small sample of data, as we have here, we cannot make sound justifications for high peaks. A couple of reasons can be vacation periods or back-to-school shopping, but with only data from January to September of 2023, one cannot say for certain. Nonetheless, Figure 4.32 can be a useful tool to guide decision-making.

Alternatively, another dashboard was created to visualise this information but in table format for a more concise reading. It has the same filters as the previous dashboard, the main difference is that here it does not show the locker position on a map and there is an additional filter to only look at the last three months of data.

4.5 Load verification

The PL operator provided a dataset with locker loads at noon to evaluate our load results. Our load dataset has 4,710,888 tuples, 719 lockers x 273 days (January to September) x 24 hours, while the PL operator load dataset, consisting of two tables, has 229,824 tuples, 756 lockers x 304 days (January to October), after removing the duplicated tuples. After joining the two datasets to keep the tuples with the same locker identification and date, at noon, we keep 196,287 tuples, 719 lockers x 273 days (January to September). For these tuples, we are checking if our load is the same as the PL operator. 80% show the same result, either the same load or both are null, and 20% do not. This discrepancy is probably a result of the PL operator saving a snapshot of the actual loads of the lockers every day at noon, without fail, providing a real load, while our method computes the load based on records that are not always complete, making us dependent on various times in a parcel lifespan. From the full parcels dataset, 187 parcels are discarded because they do not have a `deposited` or `in_demand` date, meaning they do not have an entry date, and 8 parcels are discarded because they do not have an exit date. Additionally, older calculated loads do not include parcels that were deposited before January 2023.

4.6 Machine learning

The models generated with H2O were evaluated and compared by R Squared (R^2), RMSE, Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Error (MAE) and Mean Squared Error (MSE). R^2 represents the degree that the predicted value and the actual value move in unison and vary between zero and one, where zero represents no correlation and one complete correlation. The higher the better. RMSE evaluates how well a model can predict a continuous value, the fact that the RMSE units are the same as the predicted target is useful for understanding if the error size is problematic. The smaller the RMSE, the better. RMSLE is a metric that measures the ratio between actual and predicted values and takes the log of the predictions and actual values. The smaller this value is, the better. MAE is an average of the absolute errors, again the units are the same as the predicted target and the smaller the MAE the better the model. MSE measures the average of the squares of the errors or deviations. The smaller the MSE, the better the model's performance. RMSE and MSE are sensitive to outliers, and MAE is robust to outliers.

Table 4.3 shows the values obtained for the evaluation metrics for the built models, for each model we selected the best RMSE for each algorithm. For the three datasets used to generate the models, the algorithm with the best results is GBM, particularly in datasets B and C. Focusing on the results produced by this algorithm, the worst is A noting that the basic information about a locker and its load harms model performance.

Based on the analysis and comparison of all these measures between the models developed, the model built with dataset B and algorithm GBM was selected as the best to forecast, among these. However, dataset C is not very far behind, and by applying the GBM algorithm to dataset C, we could effectively analyse the relationship between PL usage, population characteristics,

Table 4.3: Evaluation model metrics for the built models, by best RMSE.

Dataset	Algorithm	R^2	RMSE	RMSLE	MAE	MSE
A	GLM	0.0287	0.1092	0.0917	0.0775	0.0119
	DRF	0.3894	0.0866	0.0725	0.0580	0.0075
	GBM	**0.8086**	**0.0485**	**0.0402**	**0.0299**	**0.0024**
	DeepLearning	0.5551	0.0739	0.0619	0.0489	0.0055
B	GLM	0.0368	0.1087	0.0912	0.0770	0.0118
	DRF	0.4251	0.0840	0.0702	0.0559	0.0071
	GBM	**0.8234**	**0.0466**	**0.0386**	**0.0285**	**0.0022**
	DeepLearning	0.4634	0.0811	0.0684	0.0542	0.0066
C	GLM	0.0576	0.1075	0.0902	0.0755	0.0116
	DRF	0.4250	0.0840	0.0702	0.0559	0.0071
	GBM	**0.8224**	**0.0467**	**0.0387**	**0.0287**	**0.0022**
	DeepLearning	0.4631	0.0811	0.0681	0.0535	0.0066

and other factors that could help identify trends and patterns within the data.

5 Conclusion

This last chapter goes over the main considerations and findings that emerged during the thesis development and results analysis. It also discusses the challenges encountered and provides recommendations for future work.

The major points and conclusions that came to light during the thesis development are covered in this last chapter. It also addresses the difficulties that occurred and offers suggestions for further research.

5.1 Main considerations and findings

In this study, the Portuguese population is characterised by cross-referencing parishes with PLs, using datasets from INE and a PL operator. The dashboards presented facilitate the description of the population based on age, minimum education, and employment status. One concludes that there is a correlation between municipalities with PLs and users' gross income.

Since the smallest geographic granularity studied is a parish, five municipalities were selected that have more residents, and the population characteristics and locker usage among their parishes were compared.

Based on the population characteristics studied, the analysis reveals that:

Age – All parishes follow a similar curve throughout their parishes, with the difference that the parish with more residents in *Albufeira* has 25k residents and in *Lisboa* it has 40k. The majority of the population falls within the 25 to 64-year age range, followed by 65 and older.

Education – All parishes exhibit a similar educational curve. However, *Paranhos* in *Porto*, *Avenidas Novas* in *Lisboa*, *Santo António dos Olivais* in *Coimbra*, *Castelo Branco* in *Castelo Branco*, *Albufeira e Olhos de Água* in *Albufeira* show a higher proportion of individuals who have completed each level of education. For each municipality, these are also the parishes with more residents and that also show heavier lockers' usage.

Employment Status – Since the majority of residents are 25-64 years old, which is below the

pre-retirement age, there are more employed residents when compared to the remaining population. The employed residents outnumber the retired by more than double, with a significant portion of the population being students.

These findings suggest that PL usage can be influenced by consumers' employment status and education level. In the parishes with fewer residents, leading to different distributions in education and employment status, the minimum level of education is lower than in the parishes with more residents, which have higher load values, indicating a potential link between education level and PL usage. The parishes with more PL usage also have more employed residents, suggesting higher income levels, which may encourage online shopping. To consider income as a characteristic to study, we looked at municipalities as the smallest granularity. Generally, the gross income seems to follow the number of boxes available and the number of parcels deposited. Comparing the top ten municipalities for the last three variables indicates a strong relationship between gross income and locker usage.

As for determining the optimal locations for installing PLs and identifying overcrowded lockers, one successfully built a dashboard that can help with this decision. For instance, in *Lisboa* municipality, for the past three months of records, there are lockers in *Avenidas Novas*, *Arroios*, *Campolide Areeiro*, and *Marvila* that show multiple days when the locker usage is critical. The dashboard also answers the other end of the spectrum, meaning locker with little usage. However, to make a decision based on these values, more information is required.

Incorporating machine learning algorithms provides a powerful tool for identifying trends and patterns within the data. By using H2O, new models were built with different datasets and algorithms. From the models' result analysis, one concludes that the best algorithm to use is GBM with a full dataset. This algorithm could then be used to predict locker usage based on the relationship between PL usage, population income, and other relevant factors, potentially leading to more nuanced and actionable insights.

5.2 Limitations and difficulties

During the implementation of the work described in this thesis, limitations and difficulties were identified. Some of the obstacles were possible to overcome; others had an impact on the conclusions that can be extracted from this work.

The first difficulty appeared when the population characteristics with locker data were related because they referred to different geographic levels. Lockers were described through city and region, and population characteristics through municipalities and parishes. This was overcome by accessing geoapi.pt to complement locker data with municipality, parish and identifier in INE. Now, these datasets can be related. The first approach was by municipalities for all characteristics, but upon discussion with the PL operator, it was decided to, when possible, look at parishes. Ideally, one would characterise the population by postal code area, but given the population characteristics datasets that was not possible for this study.

The fact that the parcels dataset only covers the period of January to September 2023, is quite limiting for a couple of reasons. The first is that the oldest the records, the more incomplete

the dataset is, meaning that parcels that were deposited before January are not considered since we have no way of knowing how many are already in a locker at the start of the analysis. Furthermore, due to synchronism issues, not all the key moments of a parcel lifespan are filled. This does not mean that these records are ignored since the PL operator has an attribute that potentially solves this problem. Nonetheless, the observed dates as the box is occupied and the box is free, differ and might refer to different points in the lifespan of a parcel. Preferably, these dates would not have to be calculated. Finally, since the dataset does not cover more than a year of data, one can not make definitive conclusions regarding locker usage. One can not detect patterns over time about seasonality and special occasions.

5.3 Future work

For future research, we could further complement the datasets with information regarding traffic and pedestrian traffic inside cities and see how that relates to locker usage. One could describe the population by postal code area or neighbourhoods. Another aspect that could be considered is the residents that regularly move outside their residence area, for work or school, for instance. Another dimension that could be studied is e-commerce trends and see how those are reflected in locker usage.

Second, keep adding more data about parcels and lockers, producing more in-depth analyses of the datasets. Also, add another level to load analysis and calculate it minute-by-minute.

Last, continuing to incorporate machine learning techniques to more effectively analyse the relationship between PL usage and other relevant factors, possibly yielding more complex and useful insights.

Bibliography

[1] J. A. Cano, A. Londoño-Pineda, and C. Rodas. “Sustainable logistics for e-commerce: A literature review and bibliometric analysis”. In: *Sustainability* 14.19 (2022), p. 12247 (cit. on p. 8).

[2] P. Carotenuto, R. Ceccato, M. Gastaldi, S. Giordani, R. Rossi, and A. Salvatore. “Comparing home and parcel lockers’ delivery systems: a math-heuristic approach”. In: *Transportation Research Procedia* 62 (2022), pp. 91–98 (cit. on pp. 5, 6).

[3] M. Cieśla. “Perceived Importance and Quality Attributes of Automated Parcel Locker Services in Urban Areas”. In: *Smart Cities* 6.5 (2023), pp. 2661–2679 (cit. on pp. 5–7).

[4] B. Dong, I. B. Hovi, and D. R. Pinchasik. “Analysis of service efficiency of parcel locker in last-mile delivery: A case study in Norway”. In: *Transportation Research Procedia* 69 (2023), pp. 918–925 (cit. on p. 6).

[5] DuckDB. *Duck DB*. https://duckdb.org/. (Online; accessed 08.09.2024). 2024 (cit. on p. 9).

[6] J. R. van Duin, B. W. Wiegmans, B. van Arem, and Y. van Amstel. “From home delivery to parcel lockers: A case study in Amsterdam”. In: *Transportation Research Procedia* 46 (2020), pp. 37–44 (cit. on p. 5).

[7] H2O.ai. *H2O Algorithms*. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html. (Online; accessed 08.09.2024). 2024 (cit. on p. 20).

[8] T. T. Huong and B. N. Thiet. “Smart Locker-a Sustainable Urban Last-Mile Delivery Solution: Benefits and Challenges in Implementing in Vietnam”. In: *12th NEU-KKU International Conference Socio-Economic and Environmental Issues in Development*. 2020 (cit. on pp. 5, 6).

[9] G. Iannaccone, E. Marcucci, and V. Gatta. “What young e-consumers want? Forecasting parcel lockers choice in Rome”. In: *Logistics* 5.3 (2021), p. 57 (cit. on pp. 6, 7).

[10] S. Iwan, K. Kijewska, and J. Lemke. “Analysis of parcel lockers’ efficiency as the last mile delivery solution–the results of the research in Poland”. In: *Transportation Research Procedia* 12 (2016), pp. 644–655 (cit. on pp. 1, 5, 6).

[11] A. Jagoda, T. Kolakowski, J. Marcinkowski, K. Cheba, and M. Hajdas. “E-customer preferences on sustainable last mile deliveries in the e-commerce market: A cross-generational perspective”. In: *Equilibrium. Quarterly Journal of Economics and Economic Policy* 18.3 (2023), pp. 853–882 (cit. on p. 5).

[12] U. Lachapelle, M. Burke, A. Brotherton, and A. Leung. “Parcel locker systems in a car dominant city: Location, characterisation and potential impacts on city planning and consumer travel access”. In: *Journal of Transport Geography* 71 (2018), pp. 1–14 (cit. on pp. 5–7).

[13] A. Lagorio and R. Pinto. “The parcel locker location issues: An overview of factors affecting their location”. In: *International Conference on Information Systems, Logistics and Supply*. 2020. URL: https://api.semanticscholar.org/CorpusID:229219841 (cit. on pp. 1, 5).

[14] S. Lone and J. Weltevreden. *2023 European E-commerce Report*. Amsterdam/Brussels: Amsterdam University of Applied Sciences & Ecommerce Europe. 2023 (cit. on p. 1).

[15] E. Molin, M. Kosicki, and R. Van Duin. “Consumer preferences for parcel delivery methods: The potential of parcel locker use in the Netherlands”. In: *European Journal of Transport and Infrastructure Research* 22.2 (2022), pp. 183–200 (cit. on p. 5).

[16] S. Moslem and F. Pilla. “Addressing last-mile delivery challenges by using euclidean distance-based aggregation within spherical Fuzzy group decision-making”. In: *Transportation Engineering* 14 (2023), p. 100212 (cit. on pp. 6, 7).

[17] MySQL. *MySQL*. https://www.mysql.com/. (Online; accessed 08.09.2024). 2024 (cit. on p. 9).

[18] F. M. Ottaviani, G. Zenezini, A. De Marco, and A. Carlin. “Locating Automated Parcel Lockers (APL) with known customers’ demand: a mixed approach proposal”. In: *European Journal of Transport and Infrastructure Research* 23.2 (2023), pp. 24–45 (cit. on p. 8).

[19] A. Pimenta. *e-Commerce Report 2023*. https://www.ctt.pt/application/themes/pdfs/empresas/report_ecommerce_ctt2023.pdf. (Online; accessed 15.01.2024). 2023 (cit. on p. 1).

[20] D. R. Pinchasik, I. B. Hovi, and B. Dong. “Replacing home deliveries by deliveries to parcel lockers: cost, traffic, emissions, and societal cost effects of locker network expansions in greater Oslo”. In: *International Journal of Logistics Research and Applications* (2023), pp. 1–26 (cit. on pp. 5, 6, 8).

[21] S. Portugal. *Administrative division*. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_cont_inst&INST=6251013&xlang=en. (Online; accessed 18.08.2024). 2024 (cit. on p. 16).

[22] PyPi. *Haversine*. https://pypi.org/project/haversine/. (Online; accessed 18.08.2024). Jan. 2024 (cit. on p. 13).

[23] A. Ranjbari, C. Diehl, G. Dalla Chiara, and A. Goodchild. “Do parcel lockers reduce delivery times? Evidence from the field”. In: *Transportation Research Part E: Logistics and Transportation Review* 172 (2023), p. 103070 (cit. on pp. 1, 8, 13).

[24] A. Ranjbari, C. Diehl, G. Dalla Chiara, and A. Goodchild. “What is the Right Size for a Residential Building Parcel Locker?” In: *Transportation Research Record* 2677.3 (2023), pp. 1397–1407 (cit. on pp. 5, 6, 26).

[25] F. Russo and A. Comi. “Urban Courier Delivery in a Smart City: The User Learning Process of Travel Costs Enhanced by Emerging Technologies”. In: *Sustainability* 15.23 (2023), p. 16253 (cit. on pp. 5, 6).

[26] B. Sawik, J. Faulin, A. Serrano-Hernandez, and A. Ballano. “A simulation-optimization model for automated parcel lockers network design in urban scenarios in Pamplona (Spain), Zakopane, and Krakow (Poland)”. In: *2022 Winter Simulation Conference (WSC)*. IEEE. 2022, pp. 1648–1659 (cit. on pp. 5, 6).

[27] S. Sethuraman, A. Bansal, S. Mardan, M. G. Resende, and T. L. Jacobs. “Amazon locker capacity management”. In: *INFORMS Journal on Applied Analytics* (2024) (cit. on p. 9).

[28] V. Silva, A. Amaral, and T. Fontes. “Sustainable urban last-mile logistics: A systematic literature review”. In: *Sustainability* 15.3 (2023), p. 2285 (cit. on p. 8).

[29] R. Stašys, D. Švažė, and E. Klimas. “The main reasons for customer satisfaction with parcel locker services: the case of Lithuania”. In: *Regional formation and development studies: journal of social sciences.* 37 (2022), pp. 175–187 (cit. on p. 6).

[30] Statista. *Revenue of the e-commerce market in Portugal from 2019 to 2028*. https://www.statista.com/forecasts/1263657/portugal-retail-e-commerce-sales. (Online; accessed 03.01.2024). 2024 (cit. on p. 1).

[31] J. R. d. S. N. Viana. “Algoritmo de optimização para cacifos modulares”. PhD thesis. Instituto Politecnico do Porto (Portugal), 2022 (cit. on p. 7).

[32] M Viu-Roig and E. Alvarez-Palau. “The impact of E-Commerce-related last-mile logistics on cities: A systematic literature review.” In: *Sustainability, 12 (16), 6492* (2020) (cit. on p. 1).

[33] Y. Wang, Y. Zhang, M. Bi, J. Lai, and Y. Chen. “A robust optimization method for location selection of parcel lockers under uncertain demands”. In: *Mathematics* 10.22 (2022), p. 4289 (cit. on p. 7).

A Appendix 1 - More parishes details

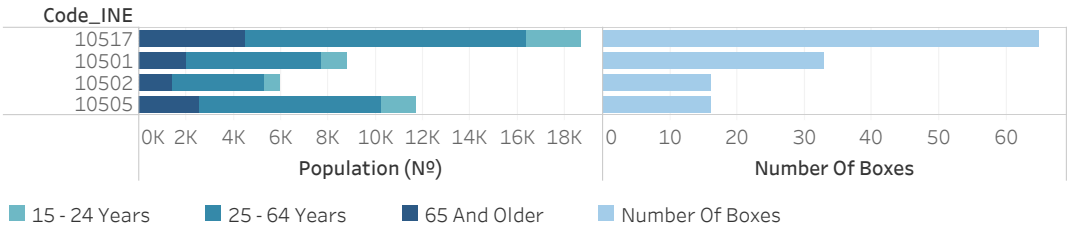


Figure A.1: *Distribution of the population by age group of each parish for Aveiro, ordered by descending the number of boxes.*

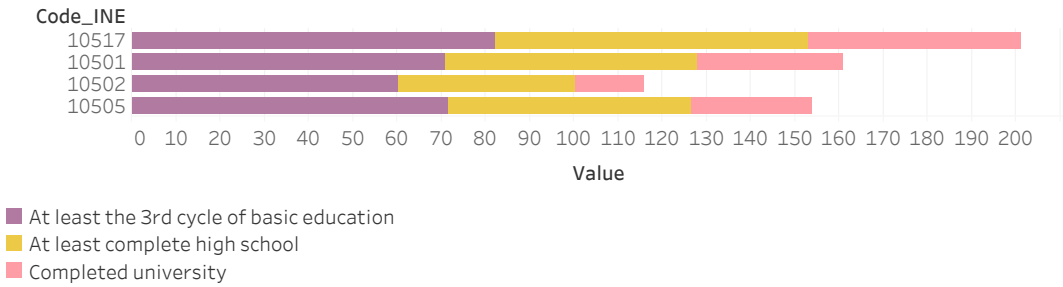


Figure A.2: *Distribution of the population by the education level of each parish for Aveiro, ordered by descending the number of boxes.*

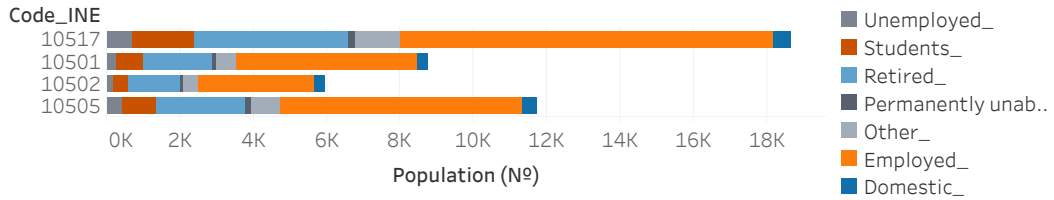


Figure A.3: *Distribution of the population by employment status of each parish for Aveiro, ordered by descending the number of boxes.*

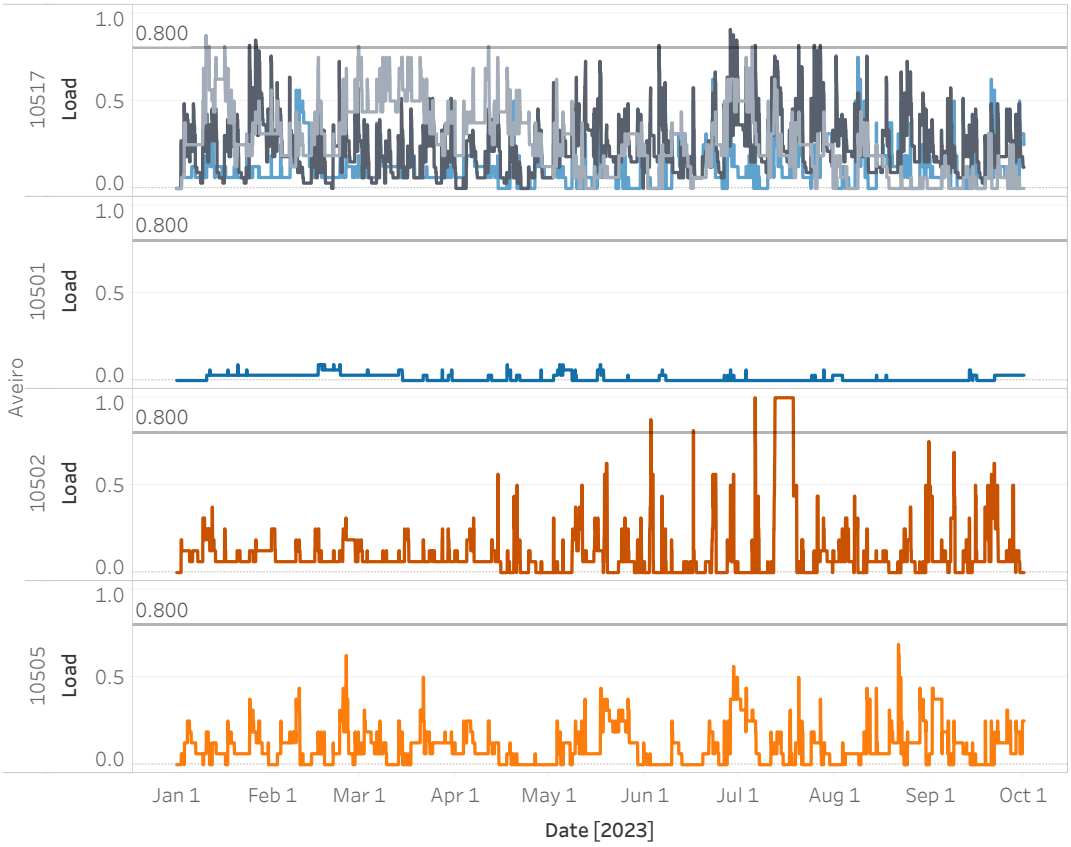


Figure A.4: *Locker load, ordered by parishes with the highest number of boxes for Aveiro.*

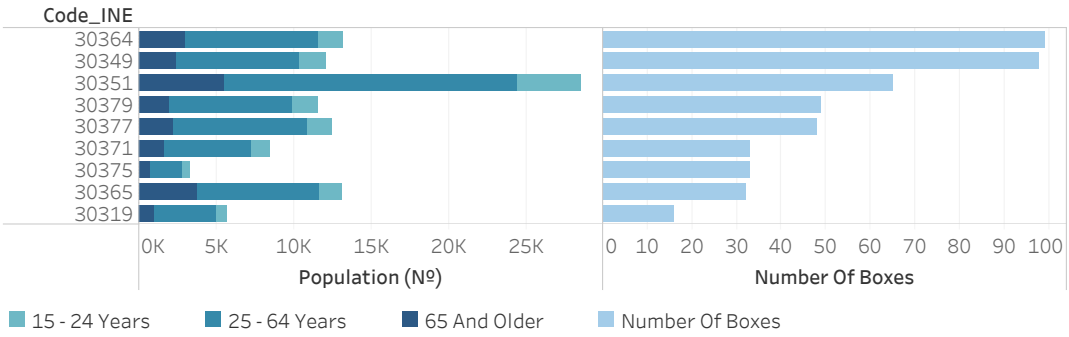


Figure A.5: *Distribution of the population by age group of each parish for Braga, ordered by descending the number of boxes.*

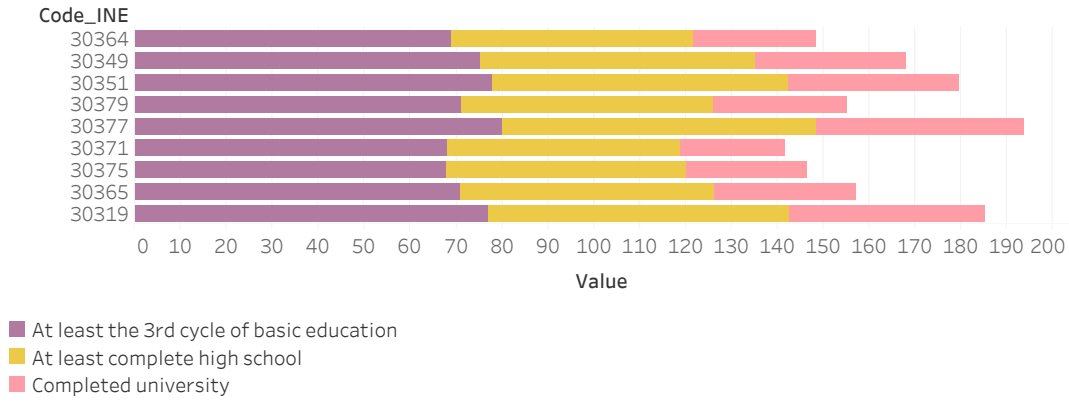


Figure A.6: *Distribution of the population by the education level of each parish for Braga, ordered by descending the number of boxes.*

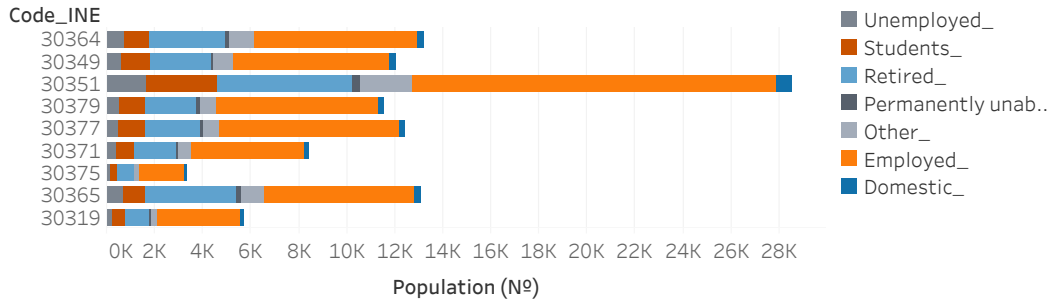


Figure A.7: *Distribution of the population by employment status of each parish for Braga, ordered by descending the number of boxes.*

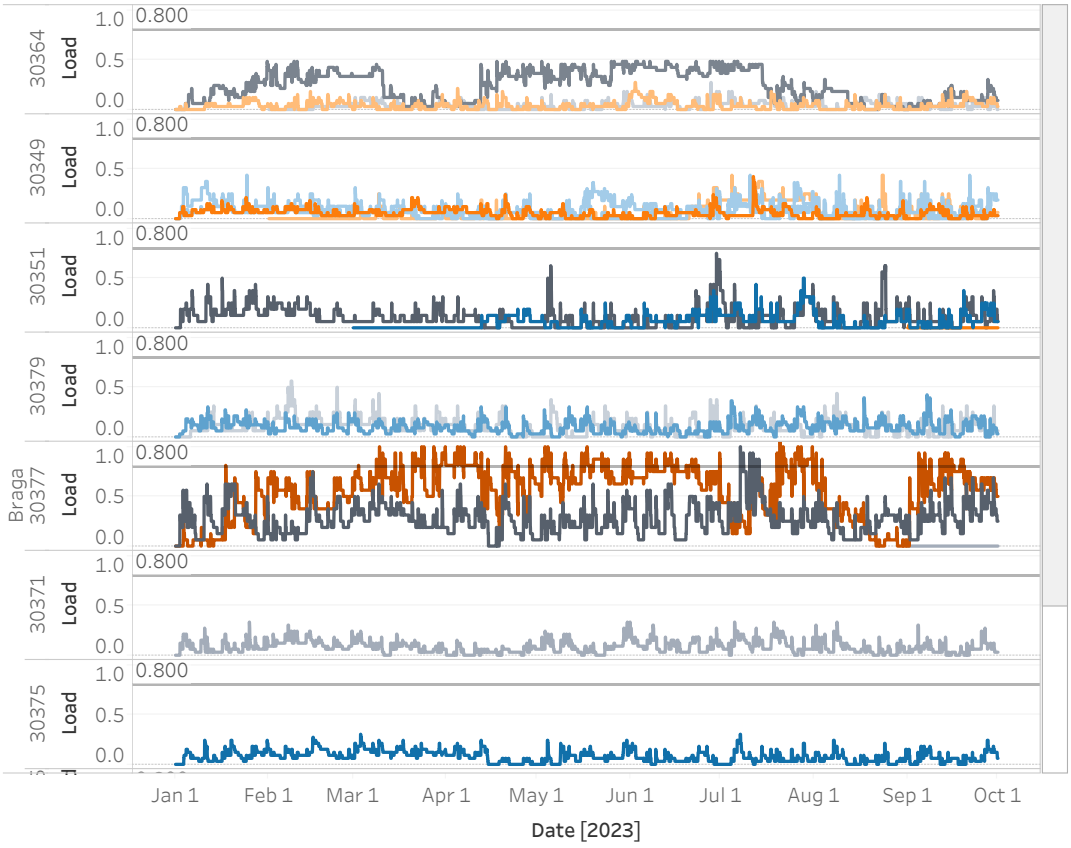


Figure A.8: *Locker load, ordered by parishes with the highest number of boxes for Braga, part one.*

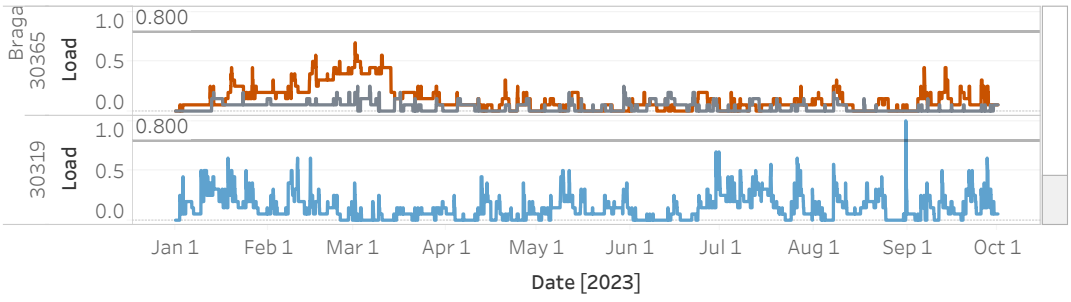


Figure A.9: *Locker load, ordered by parishes with the highest number of boxes for Braga, part two.*

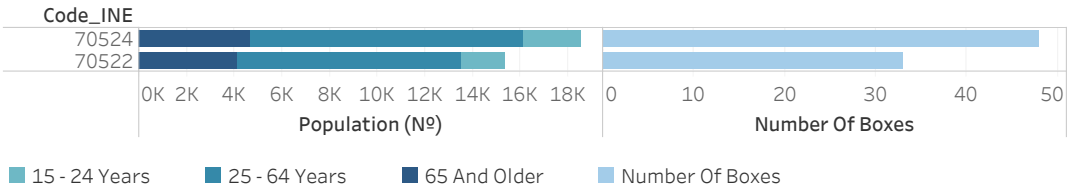


Figure A.10: *Distribution of the population by age group of each parish for Évora, ordered by descending the number of boxes.*

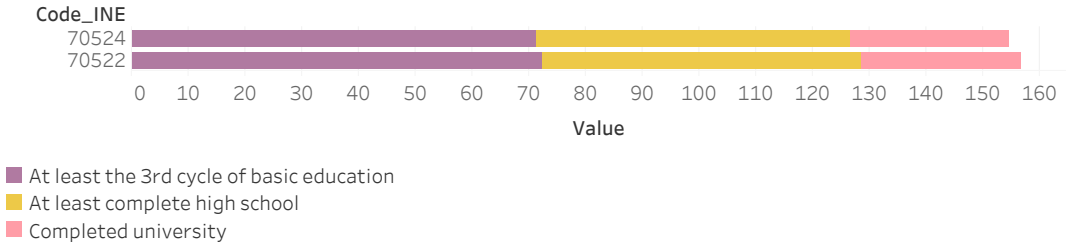


Figure A.11: *Distribution of the population by the education level of each parish for Évora, ordered by descending the number of boxes.*

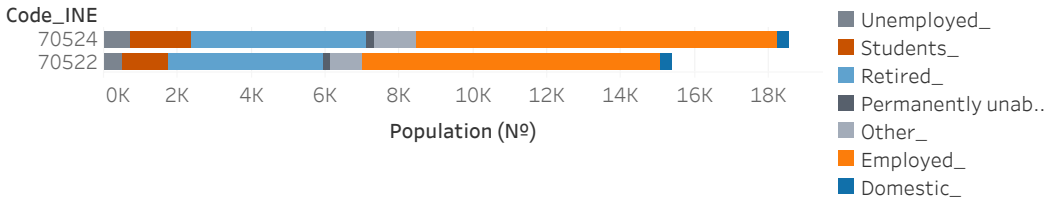


Figure A.12: *Distribution of the population by employment status of each parish for Évora, ordered by descending the number of boxes.*

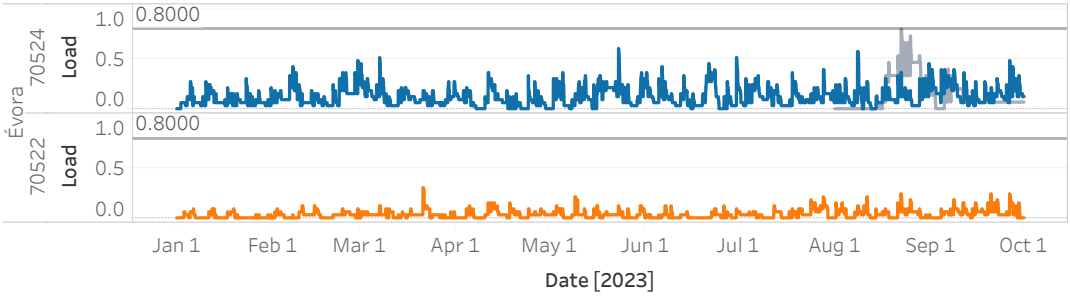


Figure A.13: *Locker load, ordered by parishes with the highest number of boxes for Évora.*

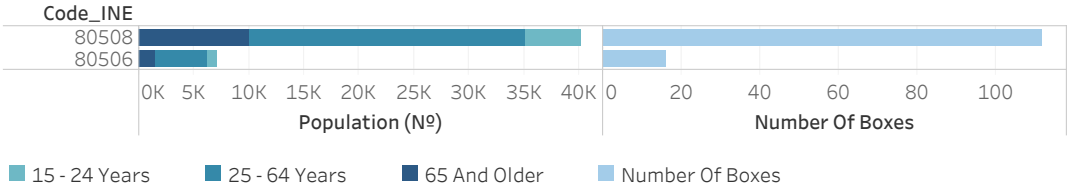


Figure A.14: *Distribution of the population by age group of each parish for Faro, ordered by descending the number of boxes.*

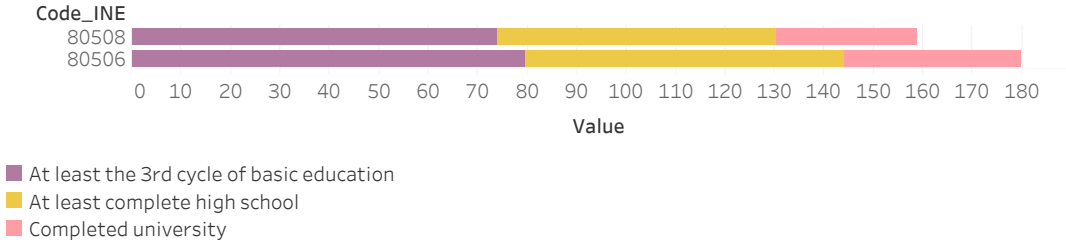


Figure A.15: *Distribution of the population by the education level of each parish for Faro, ordered by descending the number of boxes.*

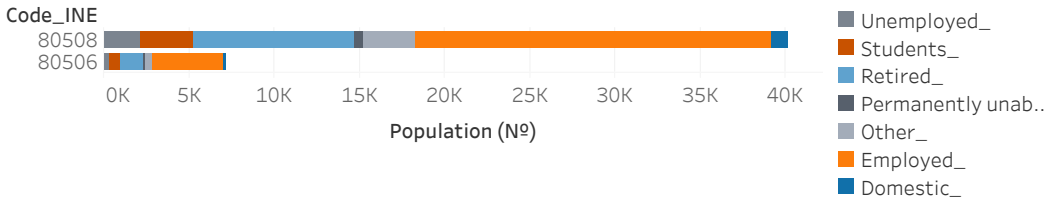


Figure A.16: *Distribution of the population by employment status of each parish for Faro, ordered by descending the number of boxes.*

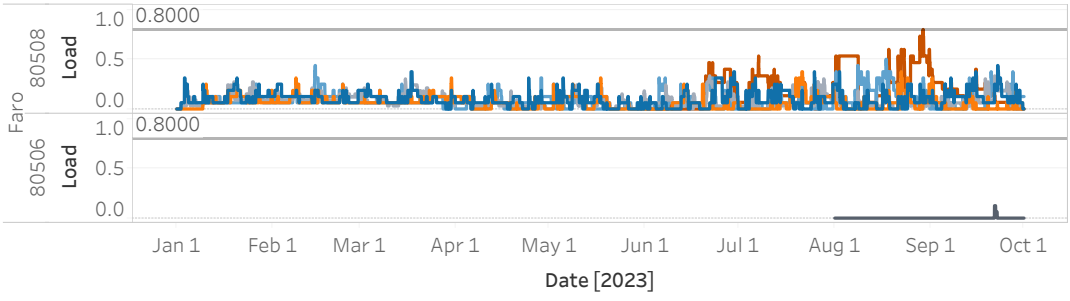


Figure A.17: *Locker load, ordered by parishes with the highest number of boxes for Faro.*

B Appendix 2 - Parcels data

Distribution of deposited parcels for the municipalities/parishes discussed in Chapter 4 from January to September 2023.

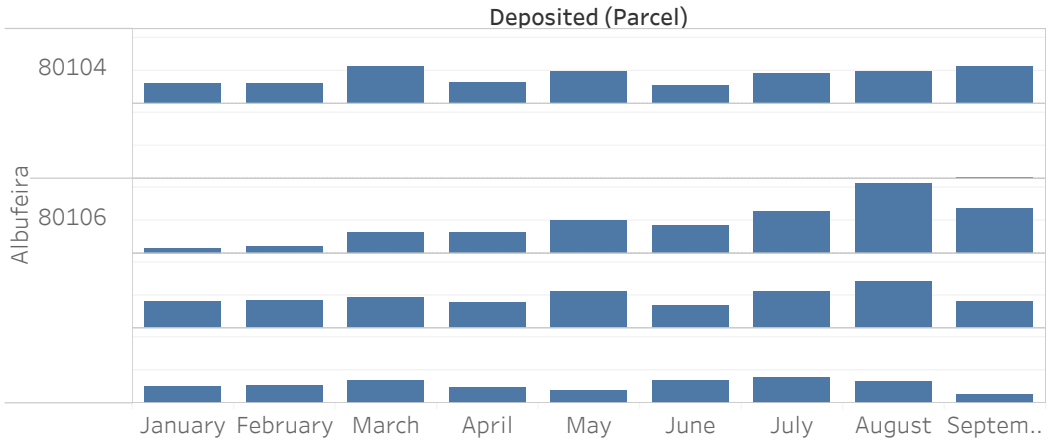


Figure B.1: *Distribution of deposited parcels in* `Albufeira`*'s parishes.*

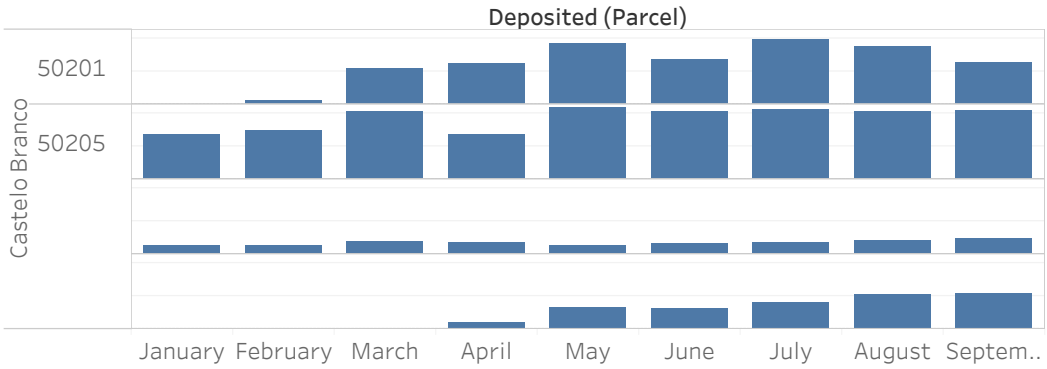


Figure B.2: *Distribution of deposited parcels in `Castelo Branco`'s parishes.*

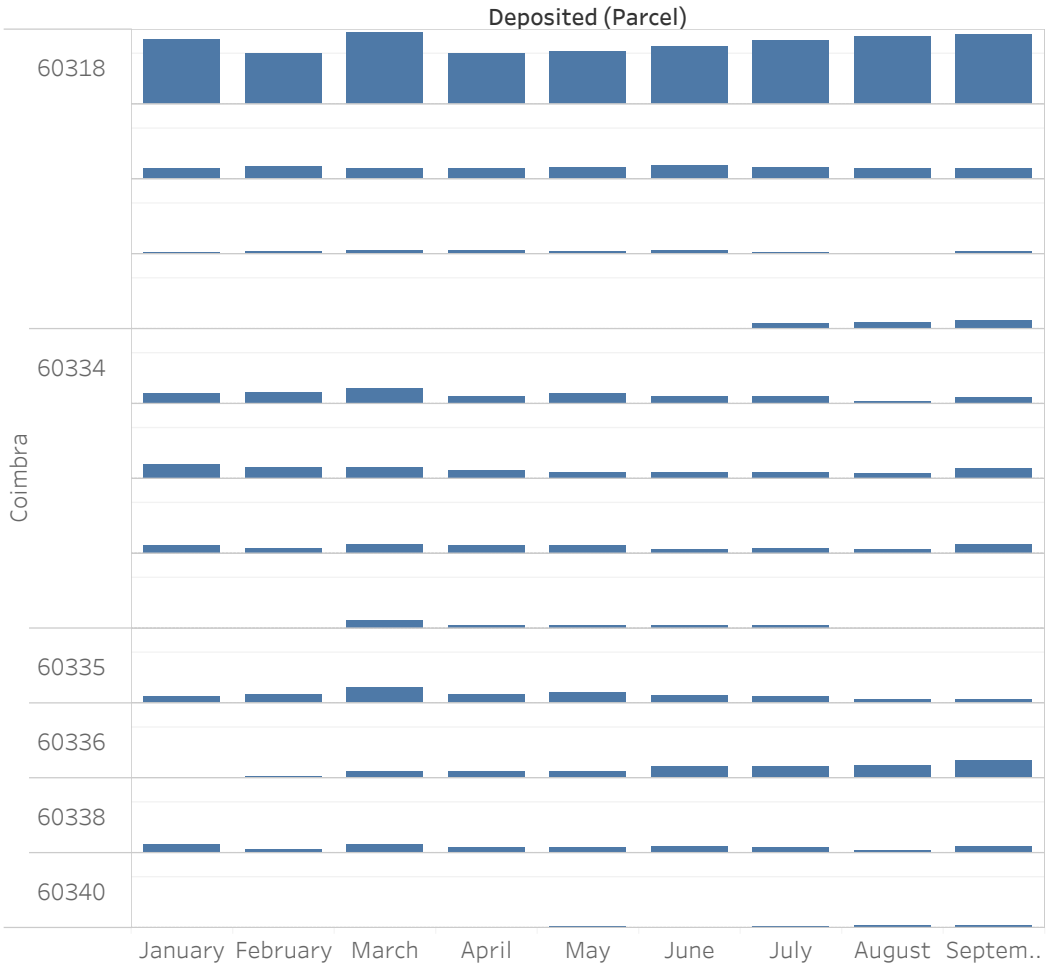


Figure B.3: *Distribution of deposited parcels in Coimbra's parishes.*

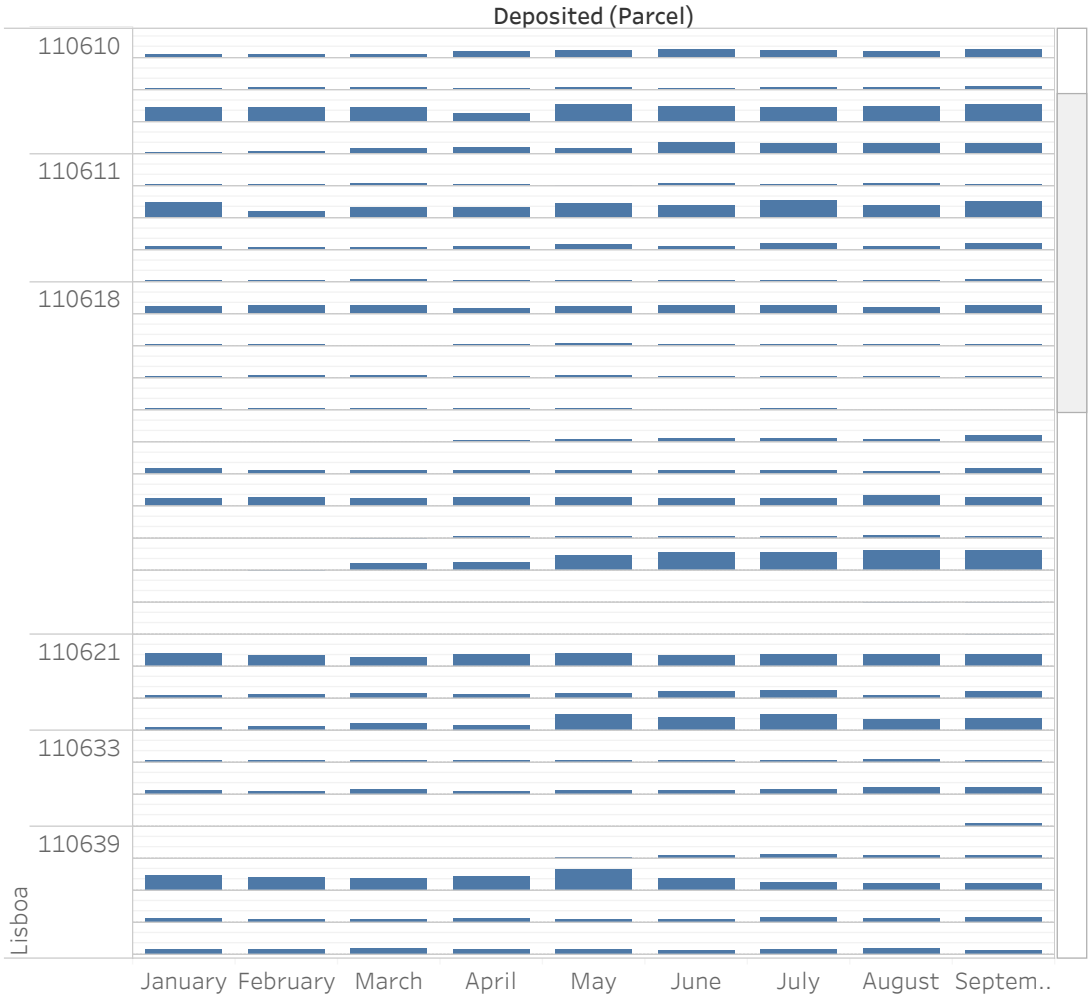


Figure B.4: *Distribution of deposited parcels in some parishes of* `Lisboa`.

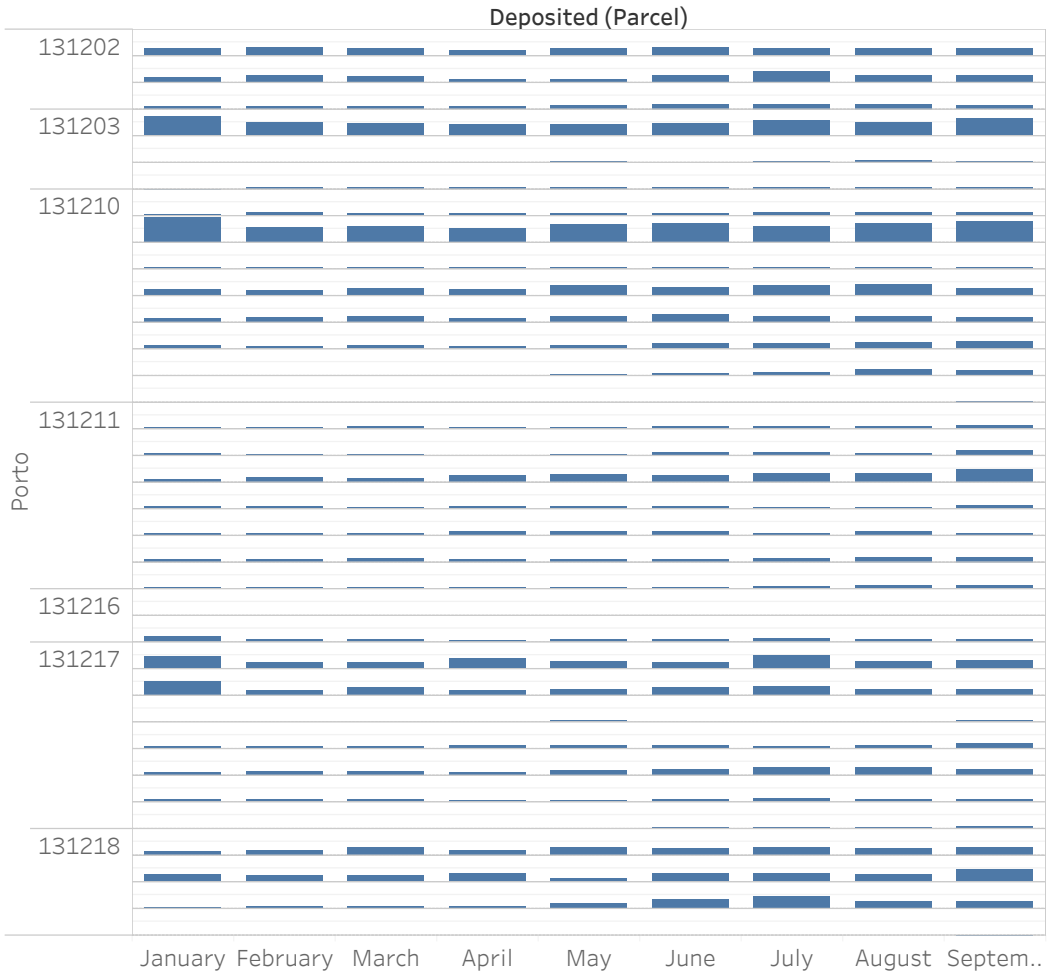


Figure B.5: *Distribution of deposited parcels in `Porto`'s parishes.*