Lisbon Housing Market Study

Diogo Marques (20231428@novaims.unl.pt, 20231428); Diogo Estevens (20231424@novaims.unl.pt, 20231424); João Mineiro (20231426@novaims.unl.pt, 20231426); Renato Poirier (20231427@novaims.unl.pt, 20231427): Sérgio Monteiro (20231430@novaims.unl.pt, 20231430)

**Group.** 3 **TP.** 2

Abstract

In the last couple of years, the housing market prices in Lisbon have suffered a significant increase, being of the most expensive cities in Europe to buy a house. Several factors, such as inflation, foreign investment and market speculation have played their role in this sudden spike in prices. Some consider it temporary; others say it is permanent. Either way, due to the current market conditions, it is important to have accurate predictive models to understand housing prices in the current market.

For the creation of an accurate predictive model, a real estate listing from a prominent housing website, Supercasa, will be used as our primary database. This dataset will have detailed information about the houses for sale in the Lisbon region, where we will have the number of rooms, area, location, and the type of building. Beyond the primary information, extras will also be analyzed; if the house has elevator, garage or even a swimming pool. The impact of these variables in the house pricing will be considered.

Exterior factors, such as the proximity to educational, cultural sites and metro stations will also be considered since these variables can have an impact on the final price of a house.

To answer our main question, machine learning algorithms will be applied. Web-scrapping, data preprocessing, data exploration and regression techniques will be covered in this report to build a predictive model capable of estimating a house price accurately, via predictive modelling, considering different intrinsic and exterior factors.

Keywords: Predictive Modelling; Web-scrapping; Housing Prices

# Introduction

The Lisbon housing market prices has been one of the main topics of discussion in Portugal. Everyone from the average person to policymakers have been discussing the topic since the prices have risen to a level where an average person in Portugal is not able to afford a house in the Lisbon region. The average total salary in Portugal, in 2023, is of 1.505€[[1]](#footnote-1), and with the increase in mortgage interest rates from the ECB, the average Portuguese inhabitant must pay the value of 362€ per month to pay their mortgage[[2]](#footnote-2). These values show the weight of the house mortgage in the average Portuguese person’s budget. So, to avoid the impact of possible additional speculation in the housing market, an accurate predictive model of the housing prices, considering several internal and external factors, is an important tool to understand if the pricing is correct and adapted to the current market.

As shown in the next figure, according to the data obtained through the Supercasa dataset, which will be explained on the next chapters, the median square meter in Lisbon costs, on average, around 6000€, while the Portuguese median is less than 2000€[[3]](#footnote-3), according to the Portuguese Statistics Institute.

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Descrição gerada automaticamente

Figure 1 – Median Square Meter Price by Lisbon Region, Lisbon (red line) and Portugal (green line).

To create such model, a database of the current houses on sale in the Lisbon area is needed. Since there are no official databases available. from the Portuguese State Institutions, regarding the prices of current houses on the market, we need to extract our own database. For the acquisition of such database, web-scrapping techniques will be used to obtain the necessary data of the house listings on the Supercasa website. Main information such as area, number of rooms, price, location, and other factors such as having a garage, elevator or swimming pools will be at our disposal. The same technique will also be applied to extract information on cultural sites, metro stations and schools, to understand if the pricing of a house is related to the proximity to these sites. Since we are applying web-scrapping techniques to obtain our main database and several others, the raw data will not be 100% accurate. Meaning that a lot of data processing and exploration will be needed. Several different techniques will be used to clean and process our data, so it can be analysed correctly, since machine learning algorithm are very sensitive to bad data.

After the data processing and exploration part, which is the most time-consuming part of the report, we are ready to start to apply machine learning algorithms and predict, with a certain level of certainty, which is the fair price of a house in the Lisbon region considering several factors, some usually not accounted when pricing a house. Several regression models will be tested, such as Multiple Linear Regression or Ridge Regression Model, and compared to each other to understand which model is more accurate in predicting the correct price for each house.

# Data and Methods

The main database will be extracted from the Supercasa website, via Web-scrapping. To acquire this dataset a Python program was created to scrape the data directly from the website. The main libraires used for this scrape were the following: Requests, Beautiful Soup, JSON. The Request library allows to create HTTP request, so it can take HTML content from webpages. The Beautiful Soup (bs4) is a library that pulls data out of HTML and XML files. Additionally, it can parse the content obtained from the website. Finally, the JSON library allows to decode JSON string into Python strings, allowing to decode the information obtained in the Supercasa website. So, after running the Python script, we obtain a dataset with 1309 properties for sale in the Lisbon region and with 9 descriptive columns. The following information was extracted for each property: “ID”, “Title”, “price”, “num\_rooms”, “total\_area”, “latitude”, “longitude”, “region” and “extras”.

As expected, the main dataset obtained had several errors and inconsistencies, so data preprocessing and exploration had to be made. Firstly, duplicated houses were excluded from our dataset, since we verified that the same house could be announced on different pages. Afterwards, data quality checks were run in each individual column, and inconsistencies were found. Some of the houses had the area information in the “num\_rooms” column and in the “total\_area” column other types information were presented. We verified that these inconsistencies were found on T0 houses (0 rooms), which had no number of rooms information and distorted the extraction. From here a rule was created: if the house had the string “T0” on its “Title” column, the value of 0 would be created in the “num\_rooms” column and the “total\_area” would be corrected. After this correction, we excluded houses which did not have the correct number of rooms or the correct total area. Afterwards, there were some rows with missing values on the “total\_area” column. One of the methods to handle with missing values is the K Nearest Neighbors Imputer. Firstly, the ideal number of neighbours had to be calculated, in this case was 1, which delivered the higher accuracy. All columns were transformed into the correct type since some integer variables had to be transformed to float and so forth. An additional column, which represents the real total of rooms in the house was created, “Total\_N\_Rooms”, since we had house being sold as having 3 rooms but in the announcement the property had 3+1 rooms, being 4 rooms. Finally, since the main goal of this report is to create a predictive model for the houses price, several dummy variables were created for each region, property type (apartment, studios) and extras (garage, swimming pool). After all these transformations, we ended up with a dataset of 1205 properties and with 94 columns, having six objects variables, two float variables (latitude and longitude) and the rest being integers.

For the other datasets, the web-scrapping technique was also applied. For the Metro dataset, the information was extracted for the Wikipedia page. For the Cultural dataset, information was scraped from museums, theatres, cinemas, auditoriums sites. Afterwards these individual datasets were put together to form an individual dataset. For the Educational dataset, information from pre-schools, 1st/2nd,3rd cycle schools, high schools and universities were extracted and put all together. The data quality check and necessary transformations were applied to all datasets, like in the first dataset.

For all datasets, two columns were very important to obtain and transform, which were the “latitude” and “longitude” columns, since those columns are used to calculate distance between houses in the main dataset to the locations on the other datasets. Transformations were applied in all datasets to assure that the latitude and longitude variables were as float, to apply the Haversine function, which allows to calculate the distance between two points having their coordinates. Finally, after having the distance between the houses and the rows on the other datasets, we applied distance thresholds for each dataset. After joining the datasets, three new columns were added to the main dataset for the regression models: "Stations\_within\_0.5km," "Cultural\_fac\_within\_1.5km," and "Edu\_within\_1.5km," indicating the number of nearby sites for each house, by distance threshold.

Uma imagem com texto, mapa

Descrição gerada automaticamente

Figure 2 – House Location (hue by price) plus Educational and Cultural Sites

# Results and Discussion

Use this section to discuss and report your main findings. Summarize the methods used and describe your main findings. Focus on reporting results that are relevant.

# Conclusions

Use this section to connect the Results with the problem you discussed in the introduction. What have been the main challenges in the development of your project? What would be the next steps?

# Statement of contr2ibution & Acknowledgments

Use this section to specify what was the contributions of each group member. Did some else also helped in the project development, perhaps with useful insights or in the data acquisition? Use this section to acknowledge their role.

# References

References should follow the APA 6th edition, consider using a reference manager such as Zotero or Mendeley. Use academic references. Avoid referencing blog posts.

1. https://www.portugal.gov.pt/pt/gc23/comunicacao/comunicado?i=salario-medio-por-trabalhador-atingiu-1505-euros-em-2023 [↑](#footnote-ref-1)
2. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine\_destaques&DESTAQUESdest\_boui=639464096&DESTAQUESmodo=2 [↑](#footnote-ref-2)
3. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine\_destaques&DESTAQUESdest\_boui=593987855&DESTAQUESmodo=2 [↑](#footnote-ref-3)