Housing Sales Prices & Venues Data Analysis of Lisbon

1. Introduction

1.1 Background

Lisbon is the capital and the largest city of Portugal, with an estimated population of 505,526 [1] within its administrative limits in an area of 100.05 km2 [2]. Lisbon's urban area extends beyond the city's administrative limits with a population of around 2.8 million people, being the 11th-most populous urban area in the European Union [3]. About 3 million people live in the Lisbon metropolitan area, including the Portuguese Riviera (which represents approximately 27% of the country's population) [4].

Tourism has been growing steadily since 2011, with hundreds of new hotels and thousands of refurbished apartments for tourists opening across Portugal, promoting the country to one of the top 3 hottest travel destinations worldwide for 2018, according to the Lonely Plant travel book publishers. The number of foreign tourists visiting Portugal rose nearly 12 percent last year to a record 12.7 million people, contributing to the once-bailed out country's strongest economic growth since 2000, official data showed on Wednesday [5]. Tourism and all travel-related revenues account for about 10 percent of Portugal's gross domestic product, which expanded 2.7 percent last year. The tourism sector is also a key source of employment and big component of exports of services [5].

1.2 Problem

Lisbon is becoming more and more popular, which is driving major changes in the city. With the appearance of Airbnb and other lodging companies, long-term renting homes were re-converted to short-term renting since it's more profitable for landlords. This has led to a significant rise in house prices, either for buying or for long-term renting. However, prices haven't rise evenly across the city, with some neighborhoods showing a significantly higher rise of prices when comparing to others [6]. Why is there a higher demand for housing in certain neighborhoods? Are the neighborhood venues the main driver of house prices?

1.3 Interest

Anyone who might want to move to Lisbon, or relocate within the city might have interest in this study. In the end, the reader should be able to understand which neighborhoods have more expensive housing, and how they relate to each other, based on their venues.

2. Data Acquisition and Cleaning

2.1 Data Sources

The Data used to solve this problem was:

 CSV File from the Portuguese Nacional Statistics Institute (INE) website, with the average house price of Portugal, Lisbon, and Lisbon neighborhoods over the last few years [7].
 This will allow to analyze and understand how house prices have risen over the last years in Lisbon.

	Location	2019	2018	2017	2016
0	PT: Portugal	1011	950	881	830
1	1701106: Lisboa	3111	2581	2143	1875
2	170110601: Ajuda	3085	2259	1847	1535
3	170110602: Alcântara	2951	2271	1856	1512
4	170110654: Alvalade	3304	2957	2394	1979

Table 1 - Top Rows of Housing Prices Dataframe

 Geojson File provided by the Lisbon City Hall, with 3rd level Administrative Divisions of Lisbon [8]. This will allow us to visualize the neighborhoods more clearly.

	OBJECTID	COD_SIG	NOME	IDTIPO	PERIMETRO	AREA_M2	FREGUESIAS53	GlobalID	Shape_Area	Shape_Length	geometry
0	1	117	Olivais	4015	12584.00	8088293.95	Santa Maria dos Olivais	a230fa77-aa34-4608-a0cc-9b8c9d27dae9	1.332305e+07	16147.938402	POLYGON ((-1016824.6818 4690909.7691, -1016783
1	2	115	Marvila	4015	12216.54	7122521.90	Marvila	d8017559-b785-49fe-bb47-195d94bbba0c	1.172476e+07	15674.400256	POLYGON ((-1015864.8618 4687711.8074, -1015825
2	3	108	Belém	4015	16050.78	10426948.88	Santa Maria de Belém + São Francisco Xavier	a604614f-ded1-4444-b4e4-34ae93a272b2	1.713895e+07	20577.231501	POLYGON ((-1024796.8455 4680967.0274, -1024829
3	4	110	Campo de Ourique	4015	6605.93	1651405.17	Santa Isabel + Santo Condestável	c76364c0-c54a-49eb-82d5-dfa19c8a89f8	2.716167e+06	8472.132735	POLYGON ((-1019423.5055 4681993.6429, -1019391
4	5	112	Carnide	4015	10445.19	3688827.32	Carnide	c978f503-3fcb-446b-bec4-46f736191aa1	6.075093e+06	13404.425191	POLYGON ((-1021380.2693 4690245.2015, -1021380

Table 2 - Top Rows of Lisbon Administrative Divisions Geodataframe

• Foursquare API to get the most common venues of Lisbon neighborhoods [9]. This will helps us characterizing the neighborhoods based on their main venues.

2.2. Data Cleaning and Feature Analysis

2.2.1 Housing Prices

Looking at the original dataframe, we can see that we have some information in the "Location" column strings. We will remove the Location_ID since we won't need it. For our study we won't need also the first two rows since they those rows are not Lisbon neighborhoods, so we'll also remove those rows. Finally, we'll add ", Lisboa" to the Location strings to guarantee that all locations retrieved correspond to Lisbon neighborhoods.

NOME	2016	2017	2018	2019
Ajuda, Lisboa	1535	1847	2259	3085
Alcântara, Lisboa	1512	1856	2271	2951
Alvalade, Lisboa	1979	2394	2957	3304
Areeiro, Lisboa	1803	2193	2550	3019
Arroios, Lisboa	1763	1997	2587	3194
	Ajuda, Lisboa Alcântara, Lisboa Alvalade, Lisboa Areeiro, Lisboa	Ajuda, Lisboa 1535 Alcântara, Lisboa 1512 Alvalade, Lisboa 1979 Areeiro, Lisboa 1803	Ajuda, Lisboa 1535 1847 Alcântara, Lisboa 1512 1856 Alvalade, Lisboa 1979 2394 Areeiro, Lisboa 1803 2193	Alcântara, Lisboa 1512 1856 2271 Alvalade, Lisboa 1979 2394 2957 Areeiro, Lisboa 1803 2193 2550

Table 3 - Clean Housing Prices Dataframe

We now have a Dataframe with Neighborhood Name, and the average price of m² for each year (2016-2019) per neighborhood.

Doing a quick analysis, we can confirm that housing prices in Lisbon have grew much more than on the rest of the country. We calculated the Rate of Price Growth from 2016 to 2019 and plotted it into a bar chart:

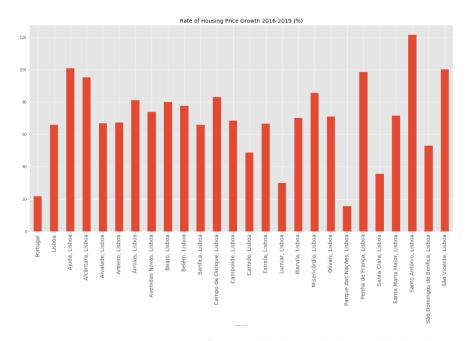


Figure 1 - Average Housing Prices for Portugal, Lisbon and Lisbon Neighborhoods

We can see that, while prices grew 20% in the whole country, in Lisbon we had a growth over 60%, meaning prices grew 3 times more in Lisbon.

2.2.2 Lisbon Administrative Divisions

There are many columns in this dataframe that are unnecessary. Let's drop those columns to reduce the dataframe size. Also, analyzing the geometry column we can see that the polygons coordinates are not is epsg4269, and because of that the folium library will not be able to plot the geometries. Let's convert the coordinate system to epsg4269, so we will be able to build a choropleth map further ahead in this study. After that, we'll use Nominatim to get coordinate pairs for each neighborhood.

	NOME	geometry	longitude	latitude
0	Olivais, Lisboa	POLYGON ((-9.13429152930909 38.78491931477943,	-9.125823	38.770324
1	Marvila, Lisboa	POLYGON ((-9.125669319549052 38.76252241258399	-9.112754	38.748259
2	Belém, Lisboa	POLYGON ((-9.205906694301216 38.71526233954682	-9.209432	38.697769
3	Campo de Ourique, Lisboa	POLYGON ((-9.157637159813509 38.7224577607051,	-9.165223	38.718213
4	Carnide, Lisboa	POLYGON ((-9.175215068103029 38.78026559915881	-9.192649	38.759206

Table 4 - Clean GeoDataframe of Lisbon Administrative Divisions

The resulting Dataframe contains only the Neighborhood Name, the geometry to plot the neighborhood area on a folium map, and Latitude and Longitude of the neighborhood center.

2.2.3 Merging Data

In this step we will merge the geodataframe with coordinates with the housing prices dataframe. The resulting dataframe will be a Geodataframe, not losing its geographical characteristics.

	NOME	geometry	longitude	latitude	2016	2017	2018	2019
0	Olivais, Lisboa	POLYGON ((-9.13429152930909 38.78491931477943,	-9.125823	38.770324	1323	1444	1750	2263
1	Marvila, Lisboa	POLYGON ((-9.125669319549052 38.76252241258399	-9.112754	38.748259	1637	1692	1483	2786
2	Belém, Lisboa	POLYGON ((-9.205906694301216 38.71526233954682	-9.209432	38.697769	1928	2288	2736	3426
3	Campo de Ourique, Lisboa	POLYGON ((-9.157637159813509 38.7224577607051,	-9.165223	38.718213	1957	2273	2965	3583
4	Carnide, Lisboa	POLYGON ((-9.175215068103029 38.78026559915881	-9.192649	38.759206	2050	2192	2572	3049

Table 5 - Merged GeoDataframe

3. Exploring neighborhoods using Foursquare API

3.1 Retrieving data

In this step we will use the Foursquare API to retrieve Venue information. We will try to get 100 venues per neighborhood, in a radius of 2000 meters.

	NOME	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Olivais, Lisboa	38.770324	-9.125823	Tryp Lisboa Aeroporto	38.770242	-9.125251	Hotel
1	Olivais, Lisboa	38.770324	-9.125823	Star Inn Hotel Lisboa	38.770619	-9.125421	Hotel
2	Olivais, Lisboa	38.770324	-9.125823	Fnac	38.771350	-9.130824	Electronics Store
3	Olivais, Lisboa	38.770324	-9.125823	Mundo Fantástico da Sardinha Portuguesa	38.773705	-9.130853	Gourmet Shop
4	Olivais, Lisboa	38.770324	-9.125823	Victoria's Secret	38.771349	-9.130834	Lingerie Store

Table 6 - Dataframe from Foursquare API

3.2 Data Understanding and Cleaning

Let's analyze the venues retrieved by the Foursquare API. First, let's see how many venues were returned by the API:

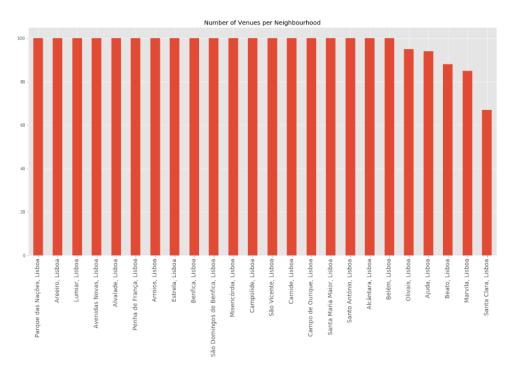


Figure 2 - Venues retrieved by Foursquare API per Neighborhood

We have a fair amount of venues for each neighborhood. Let's now analyze the Venues Type.

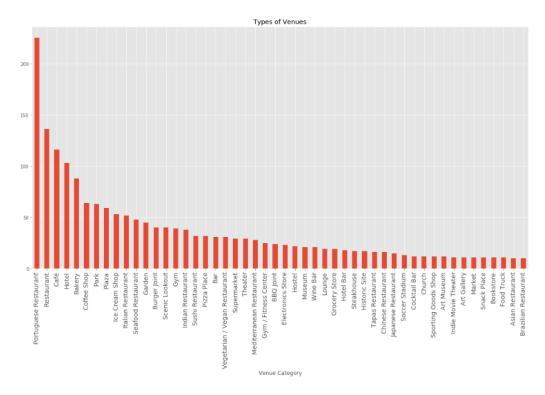


Figure 3 - Types of Venues Retrieved by Foursquare

The top 4 venues in Lisbon are Restaurants, Cafés (coffee shops) and hotels, which clearly suggests that Lisbon truly is a very touristic city. We can learn from analyzing the graph that we have some categories that should be grouped. Doesn't make much sense separate restaurants from portuguese or italian restaurants, for example. Also, cafés and coffee shops are the same thing.

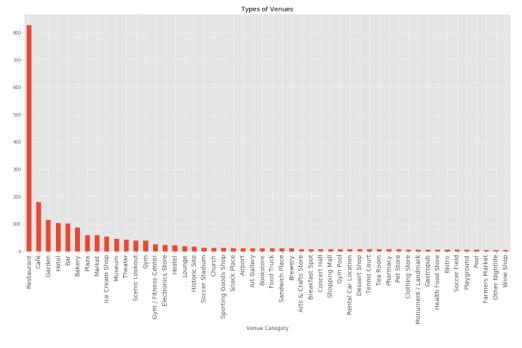


Figure 4 - Types of Venues after Grouping

3.3 Frequency of Venues per Neighborhood via one-hot encoding

Let's do one hot encoding, and group venues by neighborhood calculating the frequency mean.



Table 7 - One-Hot Encoding of Venues

Now that we know the frequency, we are able to list the 10 most common types of venue per neighborhood.



Table 8 - Top 10 of most frequent Venue per Neighborhood

4. Clustering neighborhoods based on their typical venues

To accomplish the task of clustering the neighborhoods, we will use a very common unsupervised technique of Machine Learning called K-Means, used for segmenting data based on its features.

4.1 Finding the right number of clusters

In this step we will use the Elbow Method to find out which number of clusters is the least errorprone, so that our results may be as close as possible to the truth.

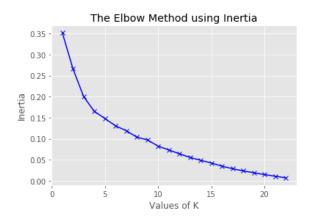


Figure 5 - Elbow Method for K-Means Clustering

4.2 Creating Clusters

Using the results obtained with the Elbow Method, we'll define the optimal number of clusters as 4.

After fitting the model, we'll merge Cluster information with our resulting Dataframe obtained in chapter 2.



Table 9 - Dataframe with Cluster Label

4.3 Cluster Analysis

In this step we will analyze the resulting clusters, and name them accordingly to their characteristics.

	NOME	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Belém, Lisboa	Restaurant	Café	Museum	Garden	Bakery
4	Carnide, Lisboa	Restaurant	Café	Garden	Theater	Gym
5	Lumiar, Lisboa	Restaurant	Bakery	Café	Garden	Market
6	Ajuda, Lisboa	Restaurant	Garden	Café	Museum	Bakery
11	São Domingos de Benfica, Lisboa	Restaurant	Garden	Café	Gym	Hotel
13	Parque das Nações, Lisboa	Restaurant	Café	Garden	Hotel	Market
21	Benfica, Lisboa	Restaurant	Café	Garden	Sporting Goods Shop	Gym

Table 10 - Cluster 1

	NOME	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Marvila, Lisboa	Restaurant	Market	Ice Cream Shop	Motorcycle Shop	Art Gallery
3	Campo de Ourique, Lisboa	Restaurant	Hotel	Café	Garden	Bar
7	Alvalade, Lisboa	Restaurant	Bakery	Bar	Gym / Fitness Center	Plaza
8	Areeiro, Lisboa	Restaurant	Bakery	Hotel	Bar	Gym / Fitness Center
10	Alcântara, Lisboa	Restaurant	Garden	Café	Bakery	Museum
12	Beato, Lisboa	Restaurant	Café	Hotel	Market	Electronics Store
15	Avenidas Novas, Lisboa	Restaurant	Hotel	Café	Bakery	Garden
16	Estrela, Lisboa	Restaurant	Café	Ice Cream Shop	Lounge	Bar
19	Penha de França, Lisboa	Restaurant	Bar	Café	Scenic Lookout	Garden
20	Misericórdia, Lisboa	Restaurant	Bar	Café	Plaza	Ice Cream Shop
23	Campolide, Lisboa	Restaurant	Hotel	Garden	Bar	Gym

Table 11 - Cluster 2

	NOME	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	São Vicente, Lisboa	Restaurant	Café	Bar	Hotel	Scenic Lookout
17	Santa Maria Maior, Lisboa	Restaurant	Hotel	Café	Plaza	Bar
18	Santo António, Lisboa	Restaurant	Bar	Hotel	Garden	Plaza
22	Arroios, Lisboa	Restaurant	Hotel	Bar	Café	Plaza

Table 12 - Cluster 3

	NOME	1st wost Common venue	Znd Wost Common Venue	3rd Most Common venue	4th Most Common venue	our most common venue
(Olivais, Lisboa	Restaurant	Café	Airport	Bakery	Market
9	Santa Clara, Lisboa	Restaurant	Café	Bakery	Market	Garden

Table 13 - Cluster 4

Looking at the resulting clusters we can now label them:

- Cluster 1 Gardens and Cafés
- Cluster 2 High diversity of venues
- Cluster 3 Accomodation and Social Areas
- Cluster 4 Residential area, low offer of services

4.4 Setting Labels

In this step we will change the Cluster Labels accordingly to their characteristics to identify the area. We will also create new labels to identify the neighborhoods accordingly to its average housing price by plotting it on a histogram with 4 bins, to match the number of clusters:

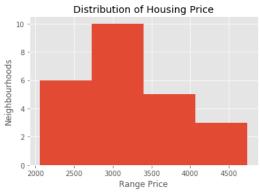


Figure 6 - Histogram of Housing Prices

We'll indentify the bins as:

- 2058€ 2729€: Low Price Areas
- 2729€ 3400€: Low-to-Mid Price Areas
- 3400€ 4071€: Mid-to-High Price Areas
- 4071€ 4742€: High Price Areas

Adding this information to the dataframe, and editing Cluster Labels, we'll end up with the following dataframe:



Table 14 - Final Dataframe

5. Results

Now that we have a consolidated dataframe with all the necessary information, we will plot all of this data.

We'll first draw a choropleth map identifying the neighbourhoods by its average housing price. We will then superimpose markers identifying the cluster. The label will show us the name of the area, the type of neighbourhood and the class of housing prices of that neighbourhood.

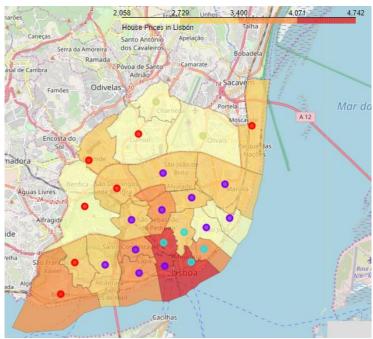


Figure 7 - Choropleth Map with superimposed Clusters

6. Discussion

Looking at the map, it's clear that the housing price doesn't really play a role in determining the type of existing venues in a neighborhood. The opposite is also true, with the existing venues not influencing the house prices.

Addressing the segmentation, we can see some logic on how it happened, being circular around downtown Lisbon (the light blue cluster).

The light blue cluster corresponds to the touristic area, which falls on the oldest part of the city and therefore, has more touristic interest. It's also an area that has a lot of restaurants, hotels and nightlife. The purple cluster corresponds to the business area, and it's located around downtown. This is where we can find many businesses, and that's why we can find such a high diversity of venues in this cluster. The red cluster corresponds to business/residential hybrid area, with a lot of parks and cafés to drink coffee, a cold drink or just spend some time during the day. The light green cluster corresponds to residential areas, with a reduced number of venues, since these areas provide mainly housing.

Addressing the housing prices, we can't really see a geographical correlation between price and location (like proximity to the city center). Although the historical center of the city is the most expensive area of the city, the prices are not evenly distributed around these areas. This can be explained by the fact that some neighborhoods are, by tradition, more fancy and with richer

households. These neighborhoods are usually called "classic neighborhoods" in Lisbon. This explains why Belém, São Vicente, Estrela and Campo de Ourique appear in the Mid-to-High Housing Price range. Avenidas Novas is one of the most modern neighborhoods in the city, and that's why it also belongs to this class.

7. Conclusions and Next Steps

Lisbon is a very popular city right now, as we could confirm by the analysis of the growth in housing prices in the city. This study can be helpful to anyone interested in moving into the city to work or start a business. We could also see that Lisbon has a very high quantity of restaurant venues, showing how gastronomy is such an import part of the Portuguese identity. This study might also be useful for any city planner that wishes to observe how the city is developing, either in price and in type of services. We can see that distance to the city center doesn't play a big role in housing price, and that we can find similar neighborhoods with different ranges of housing prices, which may help people that are on a tighter budget, with minimal loss of quality of life.

The suggested next steps would be trying to relate public transportation services, which is probably a big player in housing prices. Areas with a lot of public transportation are usually more expensive. We could also explore the evolution of the city, applying this study to the data of previous years (we only used 2019 data, but we have data available since 2016).

8. References

- [1] https://www.pordata.pt/MicroPage.aspx?DatabaseName=Municipios&MicroN ame=Popula%C3%A7%C3%A3o+residente+total+e+por+grandes+grupos+et%C3%A1 triox&MicroURL=390&
- [4] https://dre.pt/application/dir/pdf1s/2013/09/17600/0568805724.pdf
- [3] http://www.demographia.com/db-worldua.pdf
- [5] https://www.reuters.com/article/us-portugal-tourism/portugal-tourist-arrivals-spike-12-percent-to-new-record-in-2017-idUSKCN1FY1S9
- [2] https://web.archive.org/web/20181105172426/http://www.dgterritorio.pt/ca
 rtografia e geodesia/cartografia/carta administrativa oficial de portugal caop /caop download /carta administrativa oficial de portugal versao 2017
 em vigor /
- [6] https://www.publico.pt/2018/07/31/economia/noticia/preco-das-casas-em-lisboa-e-porto-dispara-20-1839651
- [7] https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrod=0009484&contexto=bd&selTab=tab2&xlang=pt
- [8] http://dados.cm-lisboa.pt/dataset/freguesias-2012
- [9] https://developer.foursquare.com/