Picking a restaurant location in Porto, Portugal

1 Introduction

1.1 Background

Porto is the second biggest city in Portugal, situated in the northern part of the country. In the last few years it has become a very popular touristic destination, due to its antiqueness, beautiful landscapes, its proud and friendly people and the fact that it is cheaper than most European big cities, having won multiple Best European Destination awards.

This has caused a huge boom of restaurants, tourism services and housing prices, presenting good business opportunities but also riskier investments as rent is very high, and competition is fierce.

I have chosen Porto because it's a city I'm familiar with, having lived there for 5 years, so I have some knowledge to perform assumptions when/if key data of some part of the problem is missing. It could also mean that this assignment will have some practical usefulness

1.2 Problem

In this project we will try to find a good location to open a new restaurant in Porto and its adjacent counties. We will make our recommendation based on competition, population density, purchasing power, and expected rent price.

2 Data

2.1 Datasets used

The following datasets will be used to perform this analysis:

- Foursquare API to retrieve venues in the area;
- Population density per County table;
- Price per m^2 of new lease agreements by Borough table;
- Purchase power index by County table;

All of these tables are public and available at http://www.ine.pt; This is the National Institute of Statistics of Portugal webpage which aggregates all kinds of official data for the country, so it should be the most reliable source for our analysis.

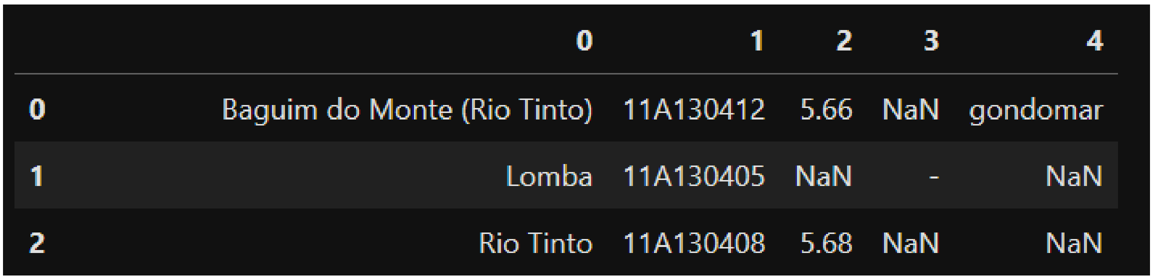
We will analyze a part of the Metropolitan area of Porto, focusing on Porto and its 5 adjacent Counties:

- Gondomar
- Maia
- Matosinhos
- Porto
- Valongo
- Vila Nova de Gaia

3 Methodology

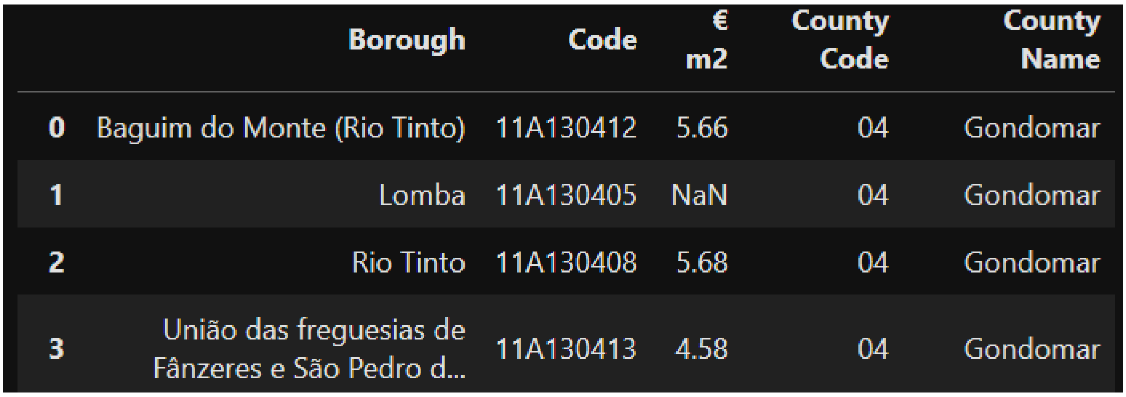
3.1 Data cleaning

We started off by importing the Price per m^2 of new lease agreements by Borough table, which included the Median house rental value per m2 of new lease agreements of dwellings (€) for all the Boroughs in the and applying some simple changes to provide titles and list the corresponding county for each of the Boroughs.



	0	1	2	3	4
0	Baguim do Monte (Rio Tinto)	11A130412	5.66	NaN	gondomar
1	Lomba	11A130405	NaN	-	NaN
2	Rio Tinto	11A130408	5.68	NaN	NaN

Figure 1 - Result of initial data import

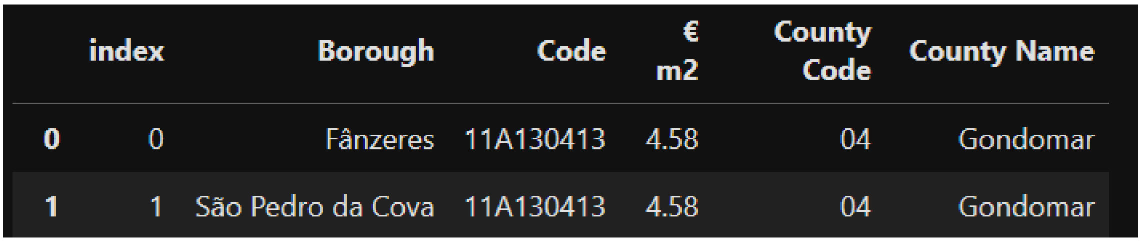


	Borough	Code	€ m2	County Code	County Name
0	Baguim do Monte (Rio Tinto)	11A130412	5.66	04	Gondomar
1	Lomba	11A130405	NaN	04	Gondomar
2	Rio Tinto	11A130408	5.68	04	Gondomar
3	União das freguesias de Fânzeres e São Pedro d...	11A130413	4.58	04	Gondomar

Figure 2 - After some cleaning

Next, we tried to get the centroid location for each of the Boroughs with the Nominatim from the geopy.geocoders. However, when we tried to retrieve them for some of the Boroughs it wasn't possible. To circle around this problem, we had to do some more cleaning to the dataframe. Boroughs in Portugal have undergone a change in 2013, and some of them were grouped, now being named "União de Freguesias de …" (translates to Union of Boroughs of.." and the name of all the former Boroughs they now encompass. We did some operations with string formulas to separate all the grouped Boroughs and then joined all in a single dataframe.

For the '€ m2' value, all the 'new split' had the old value of the "União…" assigned. For example, row 3 of Figure 2 is now the following two separate rows:

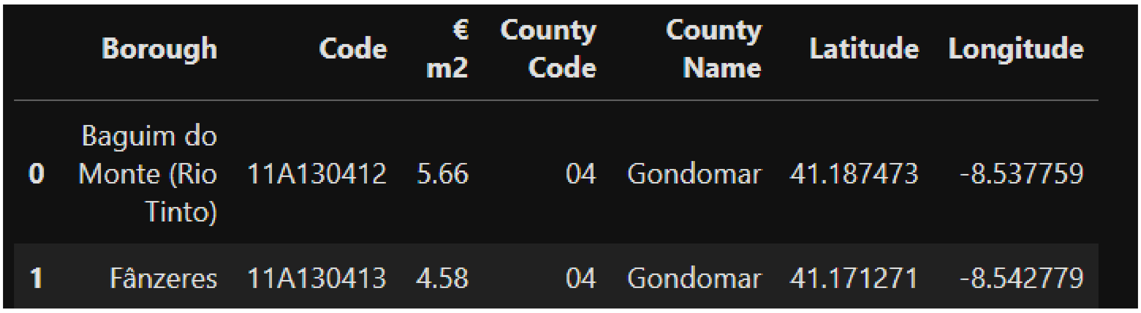


	index	Borough	Code	€ m2	County Code	County Name
0	0	Fânzeres	11A130413	4.58	04	Gondomar
1	1	São Pedro da Cova	11A130413	4.58	04	Gondomar

Figure 3 - Split Borough example

After this we were left with a finished list of Boroughs to work with, in total 76 to consider for the location of our restaurant.

We used the Nominatim, which worked for the split names, to retrieve the coordinates for each of the Boroughs. Next, we replaced missing values for the '€ m2' column with the minimum value of the price of the respective County. Out of personal experience, most of the Boroughs with missing data are poorer, more rural zones of these counties.



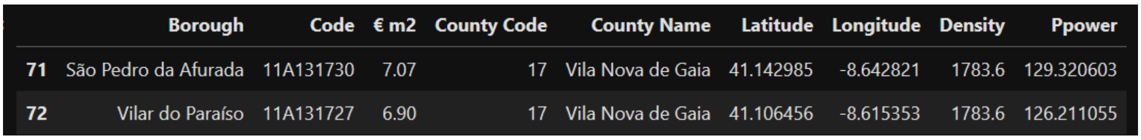
	Borough	Code	€ m2	County Code	County Name	Latitude	Longitude
0	Baguim do Monte (Rio Tinto)	11A130412	5.66	04	Gondomar	41.187473	-8.537759
1	Fânzeres	11A130413	4.58	04	Gondomar	41.171271	-8.542779

Figure 4 - Dataframe after split and NA assignment

The Population density per County file was imported next. Unfortunately, this information was only available by County. The most recent data for the by Borough level was from 2011, before the changes and the grouping occurred, so that data wasn't compatible with our Dataframe.

The information was imported and mapped to the Dataframe from Figure 4, assuming that the County density was an ok approximation for the Borough.

The purchasing power index was imported next. Similarly, this information was only available by county. After mapping similarly to what was done for the 'Density', we assumed that the purchasing power varied similarly to the '€ m2' and so we transformed the column by multiplying it per each row '€ m2' value and then dividing by the mean of the '€ m2' of the whole county. Finally the dataset was ready to analyze!



	Borough	Code	€ m2	County Code	County Name	Latitude	Longitude	Density	Ppower
71	São Pedro da Afurada	11A131730	7.07	17	Vila Nova de Gaia	41.142985	-8.642821	1783.6	129.320603
72	Vilar do Paraíso	11A131727	6.90	17	Vila Nova de Gaia	41.106456	-8.615353	1783.6	126.211055

Figure 5 - Dataset after cleaning

3.2 Exploratory analysis

After all that cleaning, we plotted the Boroughs by using the Follium library:

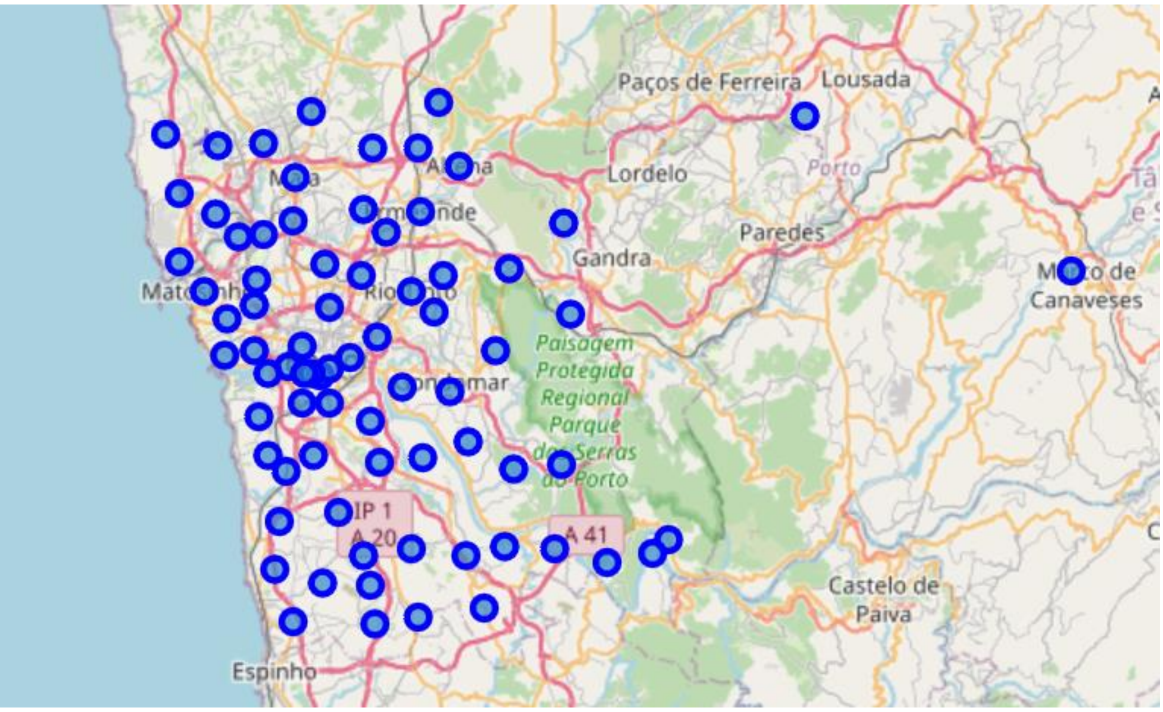


Figure 6 - Initial Borough map

No need to be an expert in Porto to notice right away a couple of Boroughs were assigned the wrong coordinates by the Nominatim. This happened because some other locations in the country had the same name. This was corrected by manually retrieving the coordinates and replacing the value in the dataframe:

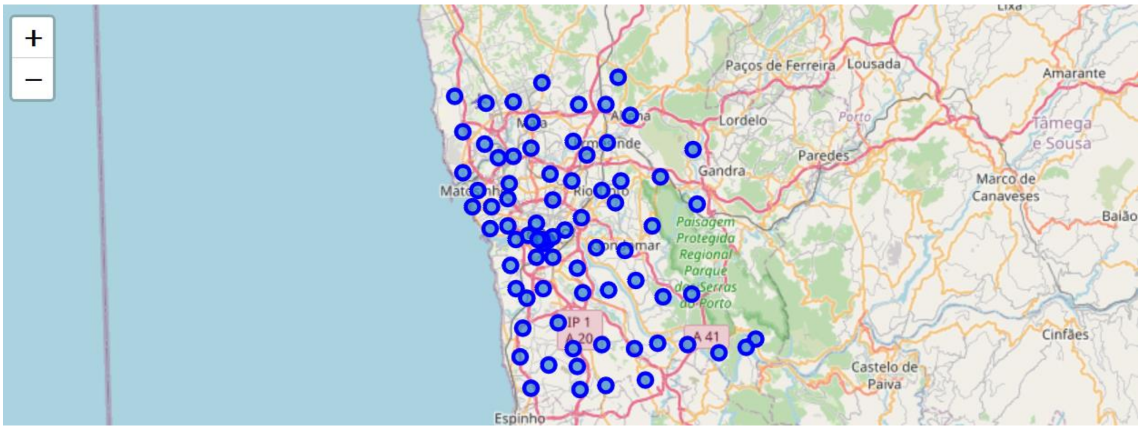


Figure 7 - Corrected map

Next, the Foursquare API was used to retrieve the nearby venues list for each Borough, within a radius of 500m of the centroid. The probable restaurants were classified with a new Boolean variable 'Is a restaurant'. The Nº of venues per Borough and Nº of restaurants per Borough were plotted:

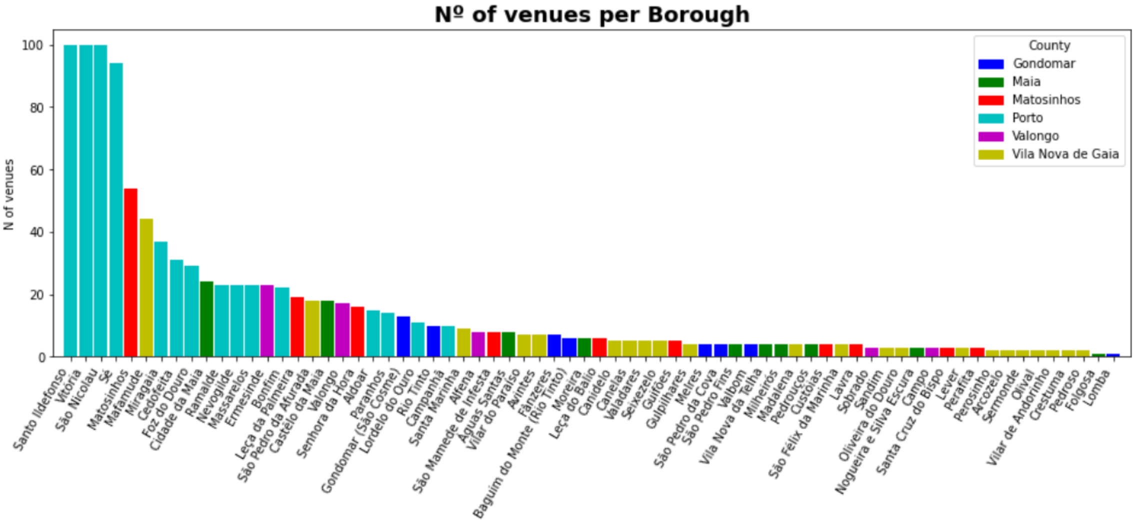


Figure 8 - N of venues per Borough

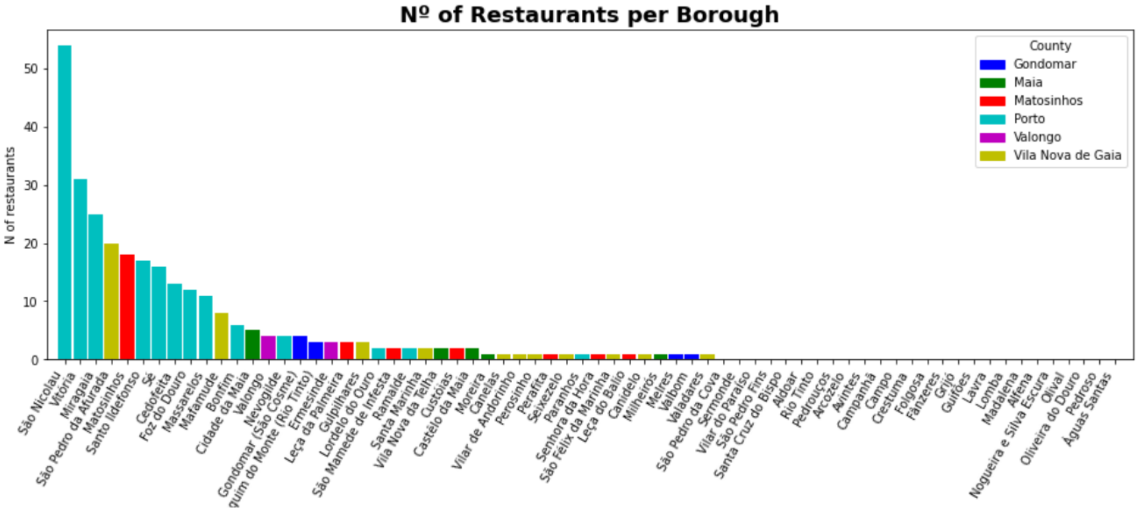
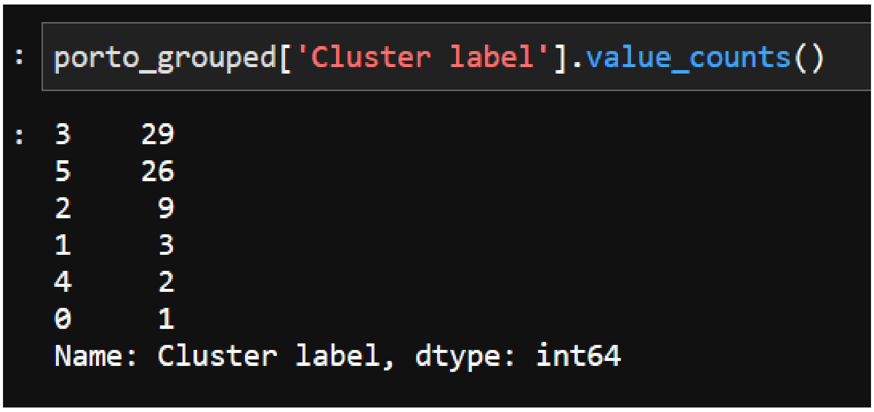


Figure 9 - Nº of restaurants per Borough

Right away we can notice some concentration of restaurants in some specific Boroughs. If we wanted to open a restaurant in the County of Porto, we could probably go with Paranhos or Campanhã, that have 1 and 0 restaurants registered in Foursquare respectively.

We tried clustering the Boroughs based on the percentage of each type of venue they had. The KMeans method was chosen with a K=6 arbitrarily chosen. The results were a bit unbalanced, which was slightly expected since some of the Boroughs have very few venues recorded in Foursquare.

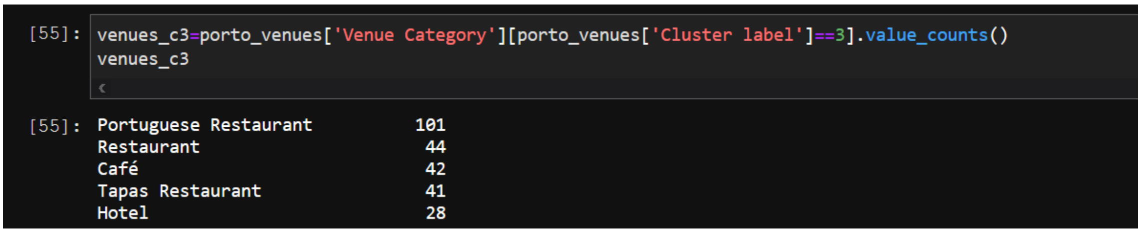


```
: porto_grouped['Cluster label'].value_counts()

: 3    29
  5    26
  2     9
  1     3
  4     2
  0     1
  Name: Cluster label, dtype: int64
```

Figure 10 - Cluster results

Cluster 3 was very restaurant heavy:



```
[55]: venues_c3=porto_venues['Venue Category'][porto_venues['Cluster label']==3].value_counts()
      venues_c3

[55]: Portuguese Restaurant         101
      Restaurant                     44
      Café                           42
      Tapas Restaurant               41
      Hotel                          28
```

Figure 11 - Cluster 3 venues

The results of the cluster were plotted (Cluster 3 are the baby blue colored markers):

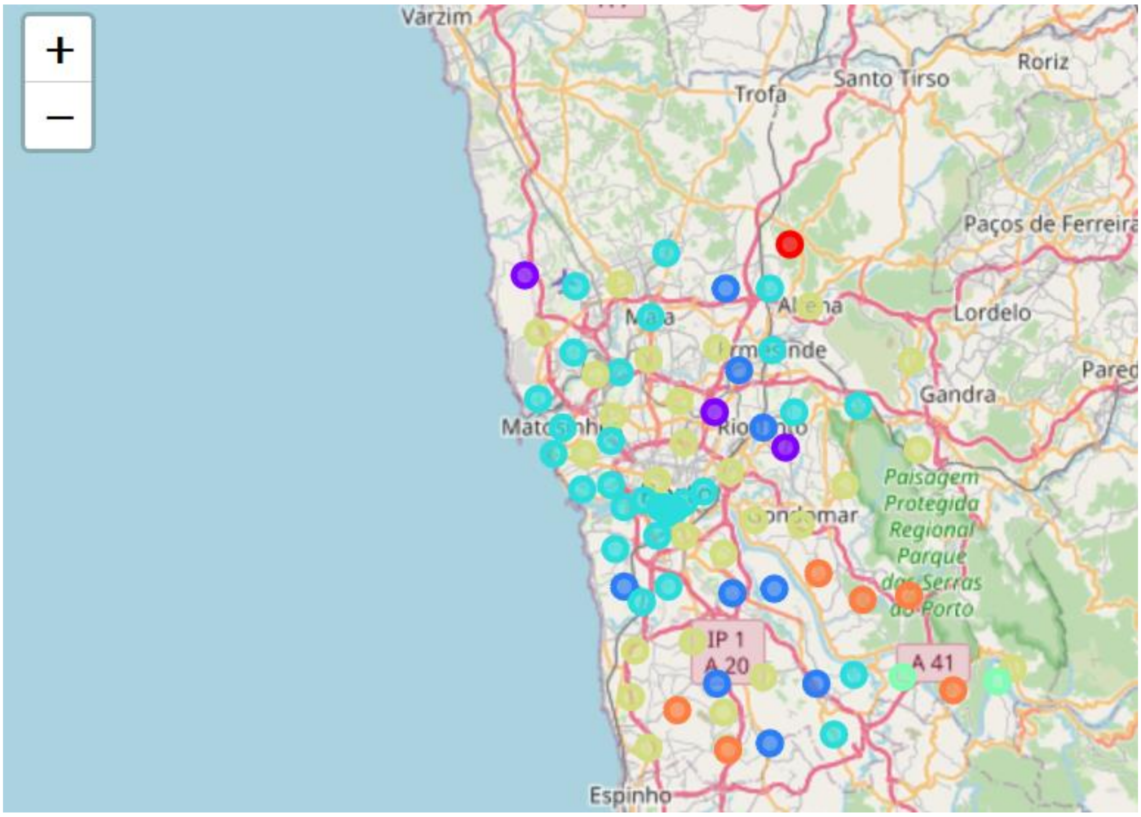


Figure 12 - Boroughs plotted by Cluster

4 Results

We decided to exclude the boroughs of cluster 3 as the competition should be very fierce in those.

By logic, we would want to choose a place with high pop. density, purchasing power (to have a higher number of customers, willing to pay more) and low price per m2 of rent.

A good rule of thumb for restaurants is to have ~1 m2 available per customer. You want your kitchen to be ~40% of the total area. For a 60 people restaurant, a good estimate would be to have 100 m2 of area.

Cluster 3 Boroughs were removed some analysis on the remaining ones was done, estimating a monthly rent and creating a normalized (0 to 1) index for customers*price (expected income). A scatterplot with the inverse of rent price and the expected income index was plotted. Ideally, we would recommend Boroughs in the 1st quadrant, simultaneously low cost and the potential of many wealthy customers:

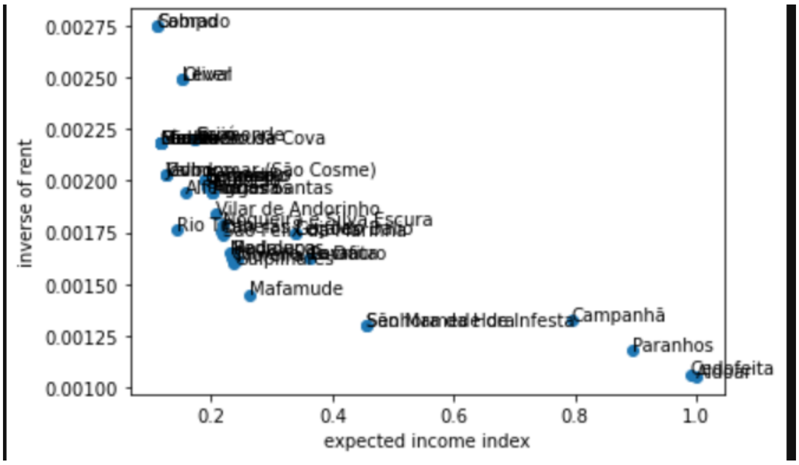
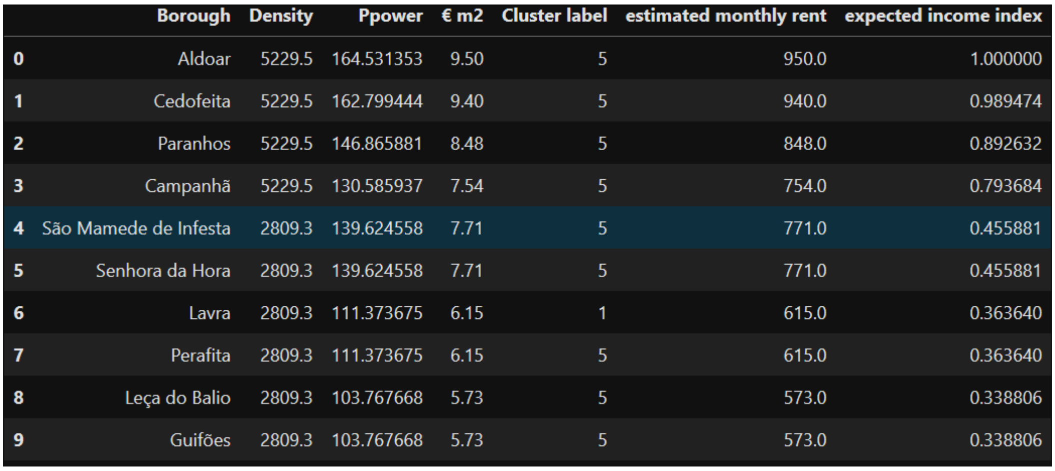


Figure 13 - Boroughs plotted by expected income vs inverse of rent

No clear winners emerged. As the index varied more than the inverse of the rent, we looked to the top 10 in the expected income, ordered and plotted them, and it led us to being ready to this list:



	Borough	Density	Ppower	€ m2	Cluster label	estimated monthly rent	expected income index
0	Aldoar	5229.5	164.531353	9.50	5	950.0	1.000000
1	Cedofeita	5229.5	162.799444	9.40	5	940.0	0.989474
2	Paranhos	5229.5	146.865881	8.48	5	848.0	0.892632
3	Campanhã	5229.5	130.585937	7.54	5	754.0	0.793684
4	São Mamede de Infesta	2809.3	139.624558	7.71	5	771.0	0.455881
5	Senhora da Hora	2809.3	139.624558	7.71	5	771.0	0.455881
6	Lavra	2809.3	111.373675	6.15	1	615.0	0.363640
7	Perafita	2809.3	111.373675	6.15	5	615.0	0.363640
8	Leça do Balio	2809.3	103.767668	5.73	5	573.0	0.338806
9	Guifões	2809.3	103.767668	5.73	5	573.0	0.338806

Figure 14 - Final list

5 Discussion

The best place to situate a restaurant should be dependent on the type of restaurant we are talking about. If we are looking for a more high-end restaurant, Aldoar and Cedofeita are very good candidates as they are very central, very dense and wealthy areas. For ~200€ a month less, Campanhã is a nice balance between density and rent prices. Guifões and Leça do Balio in the Matosinhos county appear as good options for low price restaurants as they are still fairly dense and rent prices are lower.

6 Conclusion

Although a recommendation was provided, we are not very sure of it.

It would be better to have more coherent data (for example density per Borough as it is expected to deviate significantly from the County average) to be surer of the recommendations given. The Foursquare API returned almost no results for the less central Boroughs which probably skewed the results. Tourism influx and nights slept in hotels could also be interesting to analyze.

Having a clearer idea of which type of restaurant would probably make the analysis less guess based and more certain.