The Battle of the Neighborhoods

Paulo Fernandes

June 2, 2020

1.Introduction

Moving to a different city is a decision that a lot of people must deal with throughout their lives especially for young people looking for new and better jobs or going to college. Trying to decide about something new and unknown is always stressful but the best decisions usually have strong support with data. I'm going to use this task to try to answer a problem that can be useful for these people including myself: "If I get a job offer in a different city, what would be the best places to considerer moving in?"

Probably there will be more than one place that will fulfill our requirements, but in general, this place is going to be as cheap as possible while complaining about our essential demands.

The city that I will analyze will be Porto, Portugal. It is the second city in Portugal and a place that I'm considering moving in, so I will use this work to get more knowledge about the city and the renting business. The house pricing and rents in Portugal are increasing year over the year especially in Lisbon and Porto due to the concentration of opportunities in these areas.

2.Data

To solve this problem, I build two datasets with the help of Google maps location [1] that will help me solve the problem:

- A dataset of 50 rooms and flats to rent in Porto with the information I obtained from a web page of renting business [2];
- A dataset of the subway of Porto's subway stations because I didn't find any on the web and me considerer the distance from my place to a subway station an important measure;

Both these datasets will be added to the GitHub repository. To complete my work, I'm going to use the Foursquare data to get venues around the places that I select to try to classify them. The goal is to try to get the best deals from my perspective. Having a gym close by is a must be in my opinion, because after a day of work I love to go there and have my workout hour. As I have said already, the answer to this problem will be biased to my demands. But that doesn't mean that the methodology that I'm going to try to develop cannot be used by other people with similar goals but different tasks. Maybe having a gym and a subway station close by isn't that important for you because you rather workout outside and having your car. So, you probably would considerer more important a place with a garage or parking lot and close to a park. Features I will take into consideration:

- Price;

- Number of venues close by;
- Subway station close by;
- Gym close by;

Table 1-Header of the subway stations dataset

Name	Line	Latitude	Longitude

Table 2-Header of the renting places dataset

Type	Furniture included	Bills included	Latitude	Longitude	Price €/month

The subway of Porto has a total of 83 stations right now. On the next map we can see them being displayed:

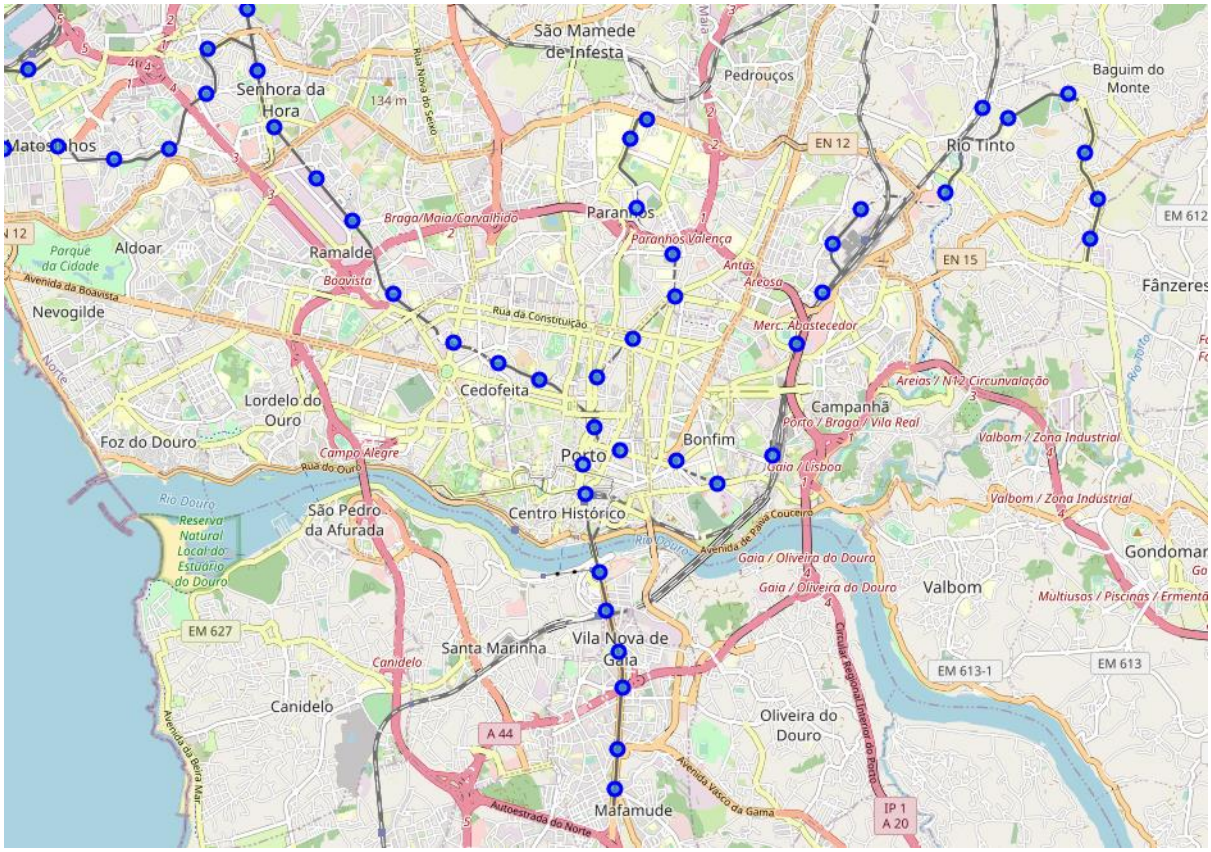

Figure 1-Location of Porto's subway stations

I have selected a total of 50 places between rooms, studios and t1 for renting in Porto. We can see their location on the next map:

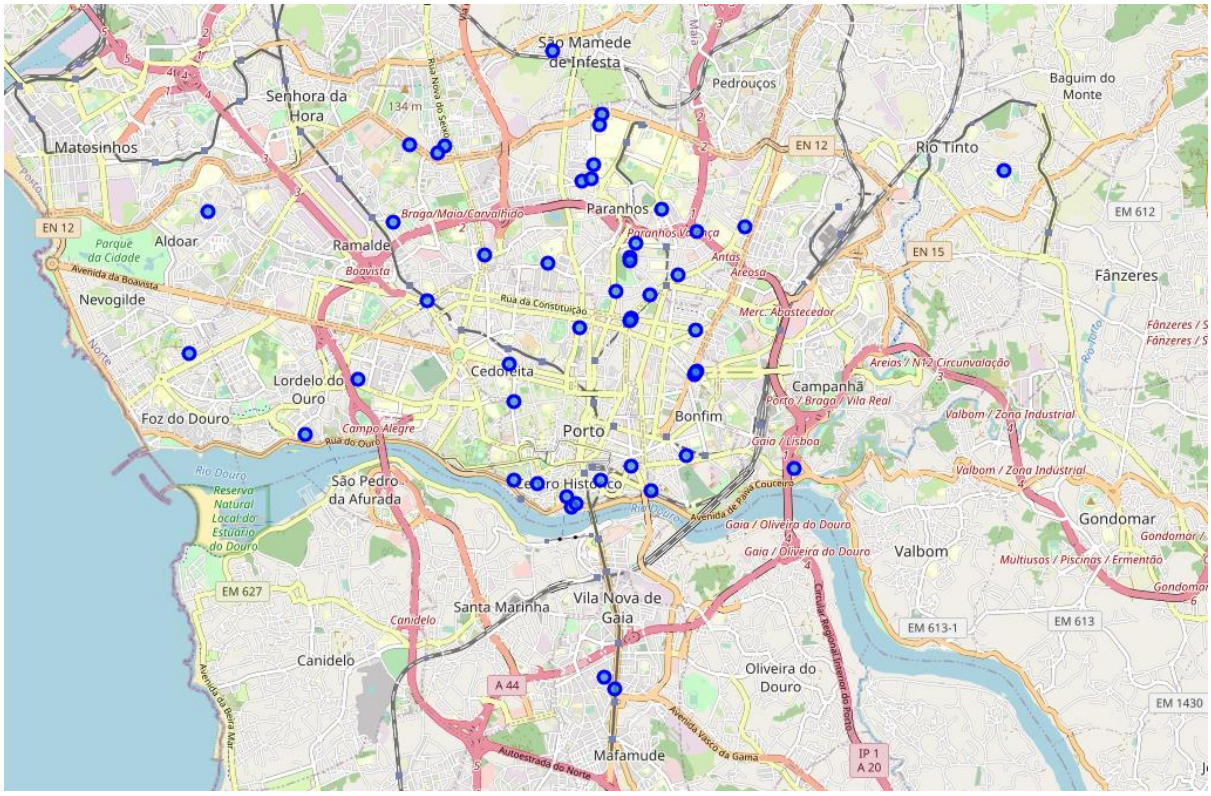

Figure 2-Renting lpaces location

3.Methodology

To detect the best rooms/flats that will comply with my requirements. The place should be less than 1km to a gym and subway station. On top of that, the price and the number of venues will also have an impact on my decision. Because my dataset is relatively small and not diversified enough, the clustering of locations cannot be useful, but I will apply it anyway and try to see if the number of venues and the closeness to a subway station impact the price of the room/flat. To solve my problem, I must use the location information and information from Foursquare to obtain the best places to considerer the renting. So, my goal is to clean and process the renting places dataset.

I removed all the places without furniture because is an expense that I don't want to deal with it. I added 100 €/month to places without bills included. To facilitate my analysis, I created classes of pricing. For rooms:

- Low: <300 €/month
- Medium: (300-400) €/month
- High: >400 €/ month

For studios and t1:

- Low: <350 €/month
- Medium: (350-450) €/month
- High: >450 €/ month

From this first processing I have reduced the total number of options from 50 to 38. The next map shows the location of this places:

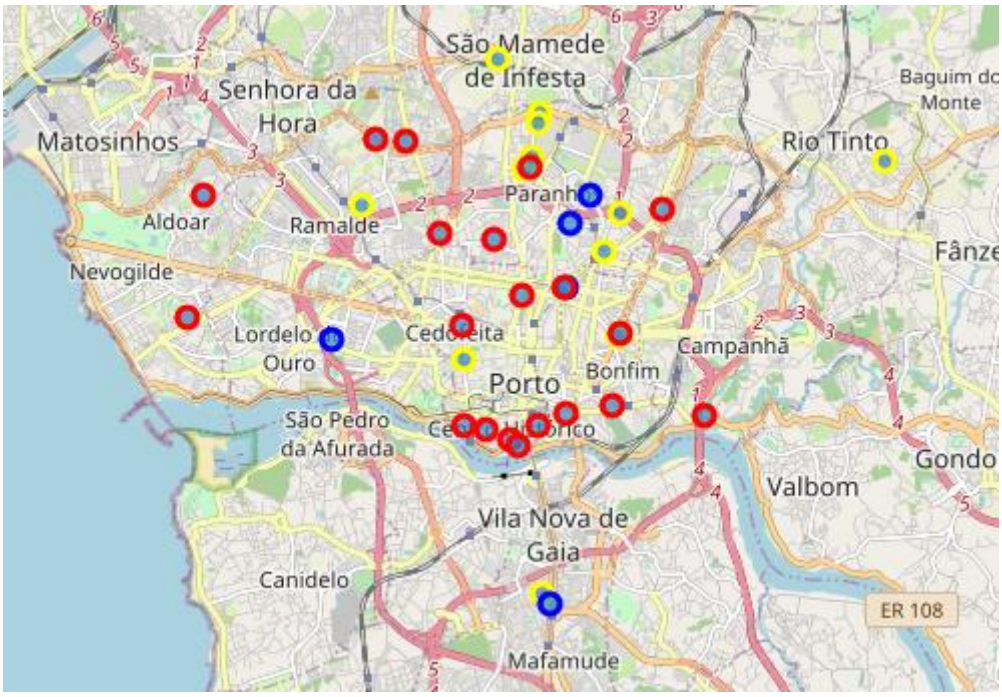

Figure 3-places for renting (red: high price, yellow: medium price, blue: low price)

Then I added a new column to my renting dataset when I put a Boolean value for the condition "subway station closer than 1km". I used the latitude and longitude values from subway stations and renting places and a mathematical algorithm to obtain the distance between two points.

Then I used the Foursquare API to obtain the total number of gyms closer than 1km from each place and added this info to my dataset. I repeat the procedure for the total number of venues, but I reduced the radius to 500m.

Finally, I removed the places that were classified as high price and my final and cleaned dataset was the next:

Type	Furniture	Bills included	Latitude	Longitude	Price €/month	Class price	Subway station <1km	Number gyms <1km	Number venues <500m
room	yes	no	41.166586	-8.597017	380	Medium	True	8	13
room	yes	yes	41.122323	-8.607831	350	Medium	True	4	34
room	yes	no	41.171456	-8.594277	400	Medium	True	7	5
room	yes	yes	41.121041	-8.606290	200	Low	True	3	24
room	yes	no	41.184385	-8.608058	350	Medium	True	3	16
room	yes	no	41.183139	-8.608470	325	Medium	True	3	22
room	yes	no	41.170091	-8.603134	275	Low	True	8	16
room	yes	no	41.172422	-8.638649	360	Medium	True	1	12
room	yes	no	41.173829	-8.599431	290	Low	True	7	4
room	yes	yes	41.161801	-8.603759	285	Low	True	7	36
room	yes	no	41.178752	-8.609370	300	Medium	True	2	24

We started with 50 places and were able to reduce the final amount of viable options to 11.

After solving my main problem, I tried to understand how the number of venues close by could impact the price of the places. To do that I turned my dataset of 50 locations viable to use a clustering method - kmeans. I applied dummy variables on the 'Class Price' and the 'Type'.

	Type	Furniture	Bills included	Latitude	Longitude	Price €/month	Class price	Subway station <1km	Number gyms <1km	Number venues <500m	High	Low	Medium	room	studio	t1
0	room	yes	yes	41.171966	-8.587178	410	High	False	3	17	1	0	0	1	0	0
1	room	yes	yes	41.191279	-8.615421	300	Medium	False	0	10	0	0	1	1	0	0
2	room	yes	no	41.166586	-8.597017	380	Medium	True	8	13	0	0	1	1	0	0
3	room	yes	yes	41.122323	-8.607831	350	Medium	True	4	34	0	0	1	1	0	0
4	room	yes	yes	41.152645	-8.621085	380	Medium	False	7	40	0	0	1	1	0	0

Figure 4-Frame of the dataset using during the clustering.

Using the Elbow method, the best number of clusters is 3:

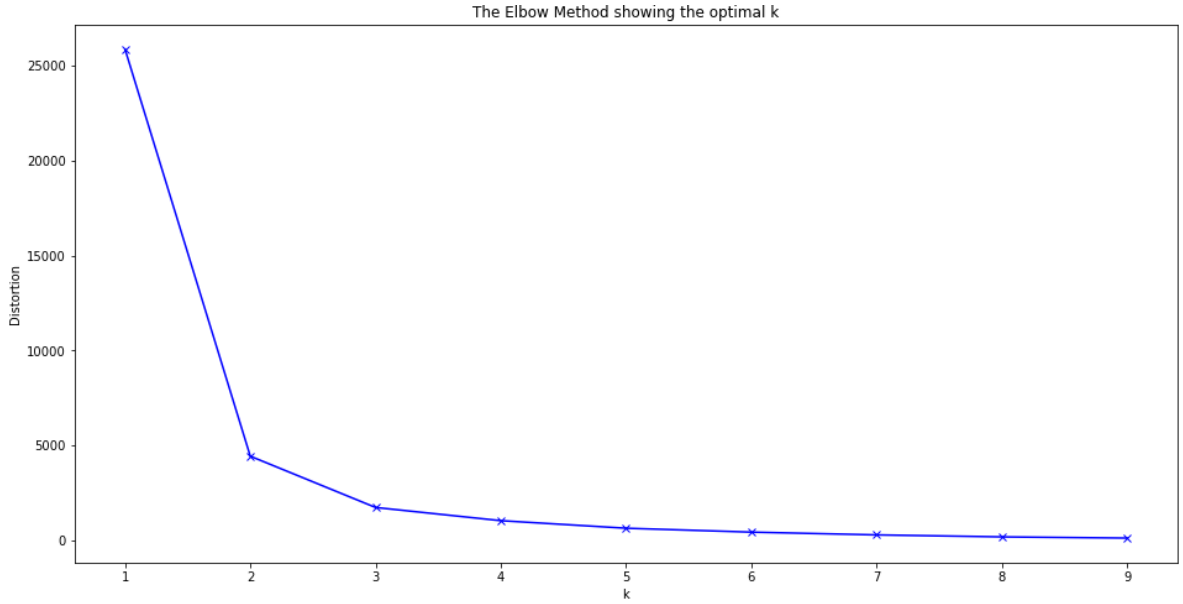

Figure 5-Elbow method showing the optimal number of clusters

Finally, the places were split into the 3 clusters:

First Cluster:

	Type	Longitude	Price €/month	Class price	Subway station <1km	Number gyms <1km	Number venues <500m
3	room	-8.607831	350	Medium	True	4	34
4	room	-8.621085	380	Medium	False	7	40
7	room	-8.606290	200	Low	True	3	24
14	room	-8.603759	285	Low	True	7	36
17	room	-8.603845	415	High	True	7	34
18	room	-8.609370	300	Medium	True	2	24
22	room	-8.611381	600	High	True	9	29
32	studio	-8.621070	620	High	False	5	44

Second cluster:

	Type	Longitude	Price €/month	Class price	Subway station <1km	Number gyms <1km	Number venues <500m
23	room	-8.617487	600	High	False	5	100
24	t1	-8.608352	1090	High	True	4	70
29	studio	-8.613247	550	High	True	5	88
35	t1	-8.603818	700	High	False	4	97
36	t1	-8.611848	850	High	True	5	83

Third cluster:

	Type	Longitude	Price €/month	Class price	Subway station <1km	Number gyms <1km	Number venues <500m
0	room	-8.587178	410	High	False	3	17
1	room	-8.615421	300	Medium	False	0	10
2	room	-8.597017	380	Medium	True	8	13
5	room	-8.611005	400	Medium	False	3	12
6	room	-8.594277	400	Medium	True	7	5
8	room	-8.549249	350	Medium	False	0	4
9	room	-8.608058	350	Medium	True	3	16
10	room	-8.608470	325	Medium	True	3	22
11	room	-8.603134	275	Low	True	8	16
12	room	-8.638649	360	Medium	True	1	12
13	room	-8.599431	290	Low	True	7	4
15	room	-8.594642	350	Medium	False	6	12
16	room	-8.643837	290	Low	False	4	9
19	room	-8.580053	850	High	False	0	5
20	room	-8.621628	700	High	True	11	20
21	room	-8.668642	650	High	False	2	10
25	studio	-8.615969	490	High	False	4	19
26	studio	-8.609709	520	High	False	4	12
27	studio	-8.595686	540	High	True	0	18
28	studio	-8.665756	540	High	False	2	3
30	studio	-8.625298	575	High	False	4	20
31	t1	-8.636309	600	High	False	1	4
33	t1	-8.631125	600	High	False	1	11
34	t1	-8.594397	850	High	False	6	12

4.Results and Discussion

The results of processing the dataset were a reducing from 50 to 11 viable places to considerer renting. This reducing of 1/5 was expectable and making this gape the biggest possible (reducing the

viable options to help on the choice) was the goal of this project. When we try to answer the question that I made at the beginning "If I get a job offer in a different city, what would be the best places to considerer moving in?" we ended up doing a similar analysis. We check the location, we check the neighborhood review, there is a supermarket close? Can I park my car on a safe parking lot? I used my personal preferences on this work, but I still feel that a lot of more could be considered. But the goal of getting all the places close to a subway station and with a gym close by where fulfilled.

While applying the k-means algorithm, we can see that the number of venues close by have an influence on the price of the place. This is especially notorious on the second cluster which corresponds to the overall higher prices and the most umber of venues close by. Obtain this information is important. If 2 places have the same price, but one of them has more venues around, this could be that the place with less venues can be overpriced. But more details are required before assuming this but is an interesting information to start by in a future work.

5. Conclusion

Overall the goal of my project was fulfilled, I used computational tools to solve my problem using data science steps, with a specially focus on the data analysis and preparation. Solving a Data Science problem doesn't mean working with Machine Learning or complicated algorithms as we learned through this course. Sometimes the cleaning of the dataset and adding more info can perfectly fulfill your main goal like I did here. I tried to don't overcomplicated this project because we were supposed to spent around 30 hours on creating a problem to solve, get the dataset and producing a report, notebook and presentation. To increase the quality of this work, as any work in data science, a bigger and more detailed dataset would be a huge improvement. Also, a more detailed analysis of the venues and districts within the city of Porto will certain find interesting conclusion that can help during this filtering process of finding a place to renting.

[1] https://www.google.com/maps

[2] https://www.idealista.pt/