APPLIED DATA SCIENCE CAPSTONE PROJECT:
Statistical solutions on finding optimal places to start a local business in Barcelona

1. Introduction

Barcelona is one of the most densely populated Mediterranean cities which does not stop growing, both in tourism and in new business models. That is why it has become an excellent location to start a new commerce.

However, land expansion is a Barcelona tough problem, given by its topography situation and being surrounded by other big cities like Hospitalet de Llobregat. This circumstances increase the risk of opening a new store due the very competitive market and the high investment required.

For this reason, now it is becoming more important than ever knowing how Barcelona territorial distribution works, the differences between each district and neighborhood, how its population is distributed, and understanding the virtues that enhance every zone. Using all this information will make easier to find bigger opportunities and accomplish the greatest success when opening a new store.

This study expects to answer decisive questions in the private sector, as for example:

- Is my business model too crowned in a specific district? Which alternative districts would be a better option to open a particular store?
- In order to open a second store of the same kind, which would be the most similar districts to the main store district?
- Which neighborhood has the least number of restaurants per square meter? Which of these options has a bigger population density?

But there also solutions for the public sector:

- Analyzing the number of students per square meter, which district is more in need of a library?
- If we want to build a new nursing home, which neighborhoods have the largest amount of old man without one near? ¿And if we want to build a new children's park?

In conclusion, this study main objective is to support the decision of where to start a specific business, using all the information from the city and showing graphically the best options to have in mind.

2. Data description and processing

2.1 Data sources:

- Venues from every Barcelona district

Through Foursquare API (https://developer.foursquare.com/) I will get every relevant venue from each district, processing and grouping them in order to record how many locals of each kind exist in each neighborhood.

- Barcelona population Dataset

Every statistical data from Barcelona city and its neighborhoods is extracted from the statistical department section of Barcelona local government webpage, as:

  - Population density (https://www.bcn.cat/estadistica/castella/dades/barris/terri/sup/sup417.htm)
  - Age distribution (https://www.bcn.cat/estadistica/castella/dades/barris/tpob/pad/ine/a2017/ine08.htm)

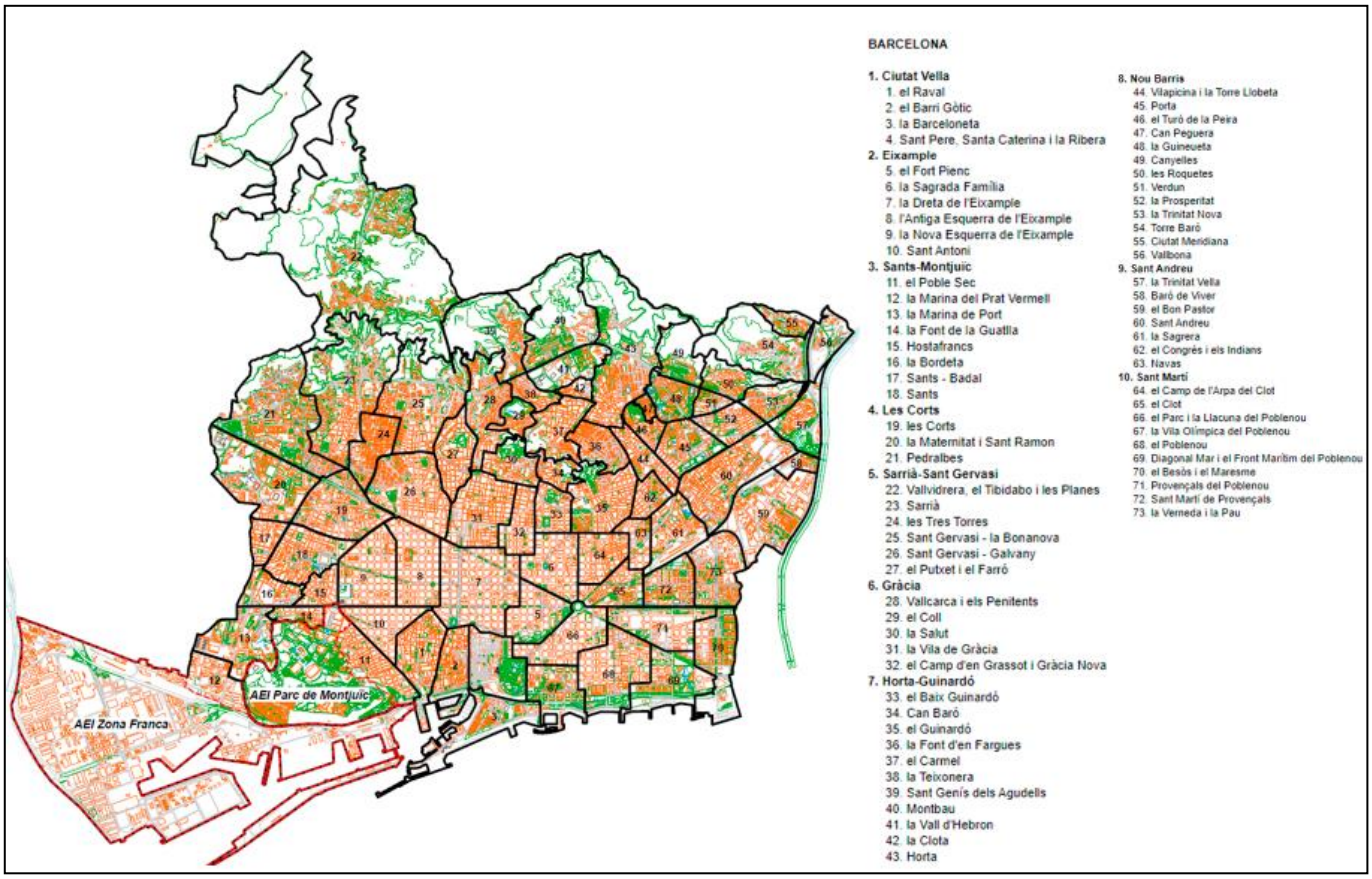
I decided to use 2017 datasets because it is the most actual year which was fully loaded in every section.

- Longitude and latitude from Barcelona neighborhoods and districts:

Unfortunately, I have been unable to find a current dataset which contains the longitude and latitude of all districts. I even contacted the statistical department of Barcelona local government to request this information, but I was told that they did not have this dataset.

So, understanding how important this data for this project is, I decided to collect manually this information for every single district using Google Maps application and save it in a CSV file that will be allocated at my GitHub repository.

To get an idea of how Barcelona is territorially distributed, here attach an images and a table of the 10 borough and 73 Neighborhoods.



2.2 Data Cleaning

These datasets have been loaded as csv from GitHub repository, using semicolon as separator and 'ISO-8859-1" as encoding.

Latitude/longitude file needed some transformations to allow a correct use of it, as changing his column types. In this case, this file was saved using comma as decimal separator, which is a standard in a majority of European countries. In order to change it, commas have been replaced by points in latitude/longitude fields and manually changed its types to float.

Population density file did not have this issue, however, it had a few columns that were not selected as relevant information for the model, and were eliminated. These columns were "AREA" and "DENSITY", which had a similar but less relevant information than "RESIDENTIALAREA" and "NET_DENSITY".

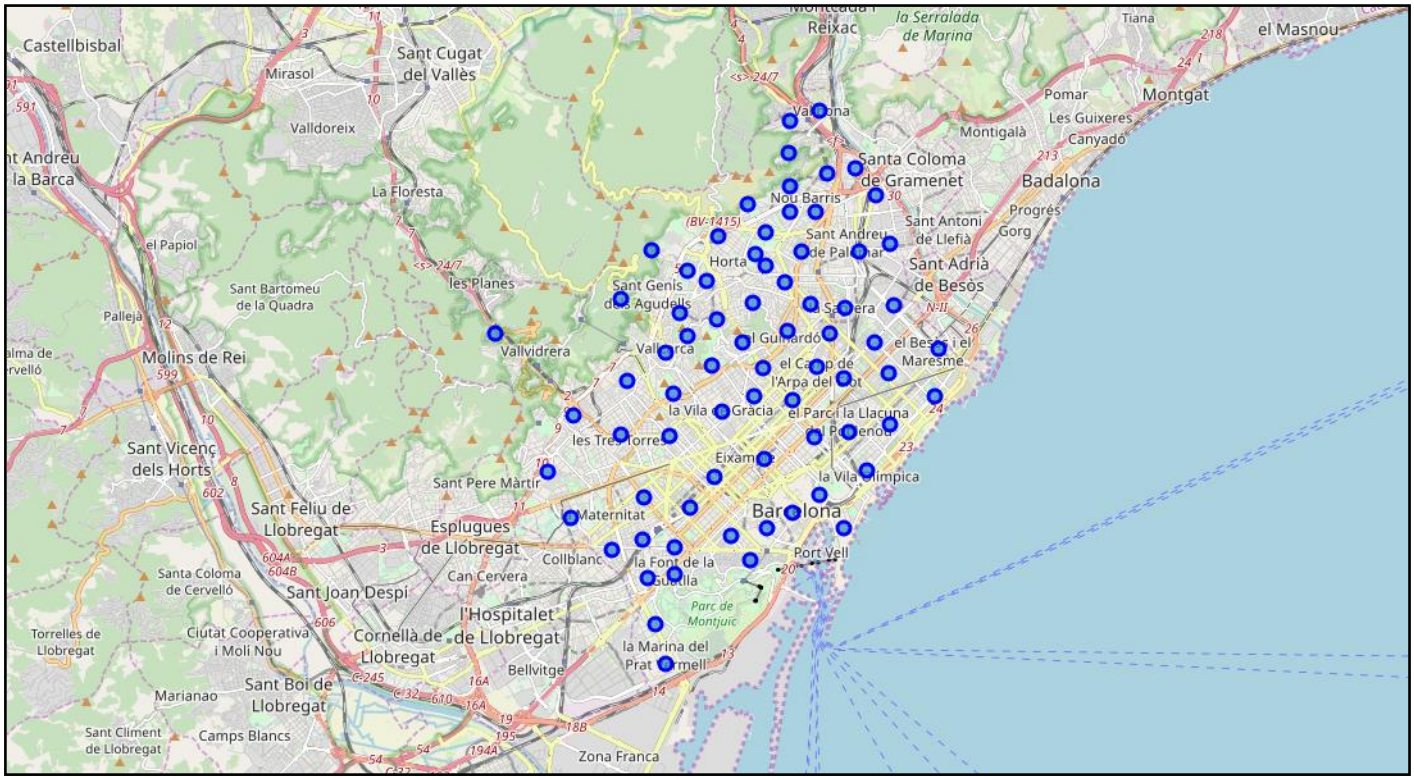
Age of distribution file were pre-processed before loading, so it was decided to not do any cleaning transformation after checking dtypes.

3. Methodology

3.1 Exploring our data

First of all it will be useful to check how our database looks and to help understand how its will be used.

Using folium library and a dataframe with latitude and longitude of each neighborhood is easy to create a map like this one showing the correct distribution of Barcelona city:



Afterward, using Foursquare API we can explore every relevant venue from each neighborhood. To give a value to 'Radius' and 'Limit' variables, I made the following estimations:

Having in mind that the neighborhood with the biggest residential area have a total of 114 hectares, which are 1.14 square kilometers and using this value of a circumference area, the maximum radius should be 600 meters.

Accepting this radius, there are not more than 200 venues in any neighborhood alone, so we used this value as 'Limit'.

At this point we got every venue from Barcelona neighborhoods, but it may be interesting to show which are the venues more repeated among every zone.

	index	Venue Category	Count
0	248	Spanish Restaurant	59
1	224	Restaurant	58
2	51	Café	56
3	261	Tapas Restaurant	54
4	171	Mediterranean Restaurant	49
5	213	Plaza	49
6	201	Park	47
7	210	Pizza Place	45
8	126	Grocery Store	45
9	20	Bakery	44
10	256	Supermarket	43
11	144	Hotel	42
12	45	Burger Joint	40
13	65	Coffee Shop	40
14	21	Bar	39

As before, it shows a pretty expected results, highlighting 'Spanish, Tapas and Mediterranean Restaurants'.

It is time to train our model using k-mean clustering algorithm for unlabeled data.

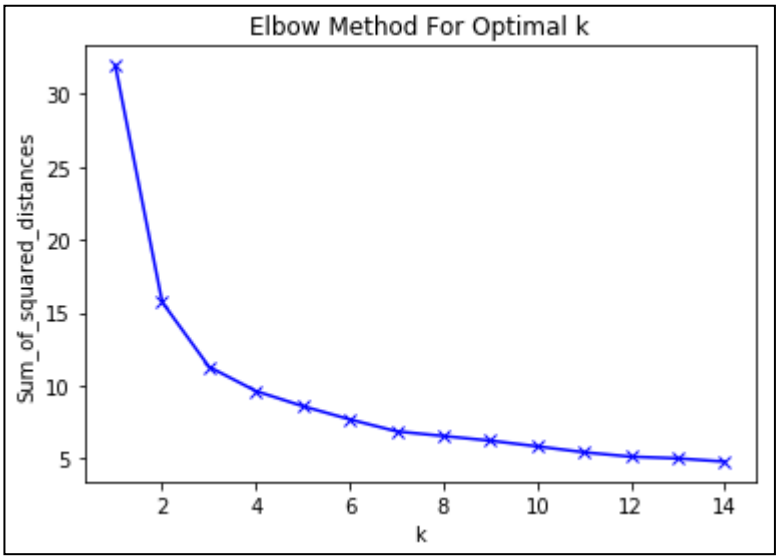
First of all, we create a dataframe grouping the most relevant information for each neighborhood:

- Number of category venues existing on this zone.
- Population, residential_Area and density information
- Population age distribution.

Each of this columns have been normalized in order give the same value to every field.

	Neighborhood_main_field	Accessories Store	African Restaurant	American Restaurant	Amphitheater	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Yoga Studio	POPULATION	AREA	RESIDENCIAL_AREA	DENSITY	NET_DENSITY	0-14 years	15-24 years	25- 64 years	65 years and more
0	Baro de Viver	0.00	0.000000	0.047619	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.033507	0.007766	0.000000	0.187832	0.423309	0.045395	0.044897	0.030907	0.023532
1	Can Baro	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.145631	0.018639	0.094545	0.400343	0.408286	0.128336	0.134880	0.141166	0.148786
2	Can Peguera	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.028852	0.000000	0.020000	0.324250	0.235217	0.027865	0.033767	0.025301	0.031952
3	Canyelles											0.000000	0.108494	0.047515	0.065455	0.146872	0.402044	0.090790	0.099604	0.098927	0.129614
4	Ciutat Meridiana											0.000000	0.169046	0.016591	0.101818	0.497344	0.446495	0.209969	0.214865	0.155975	0.134200
5	Diagonal Mar i el Front Maritim del Poblenou											0.000000	0.226142	0.078862	0.173636	0.187489	0.384072	0.338566	0.166195	0.228161	0.145779
6	Horta											0.000000	0.453448	0.209122	0.540909	0.147215	0.267183	0.422030	0.419166	0.418880	0.502519
7	Hostafrancs											0.000000	0.265659	0.020474	0.168182	0.663068	0.465290	0.222135	0.238068	0.280712	0.223291
8	Montbau											0.000000	0.078027	0.136049	0.068182	0.041302	0.285430	0.063448	0.067912	0.065292	0.110443

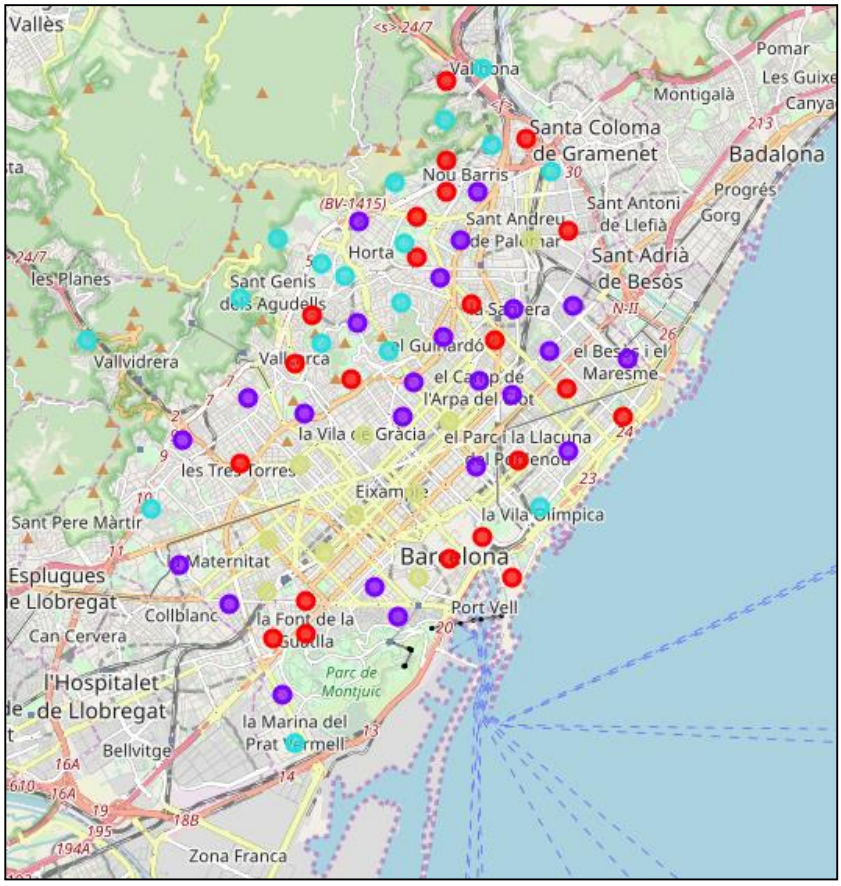
Next, using elbow method we can detect which is the optimal number of kclusters:



Any value from 2 to 4 would be a good option, but in this case I choose 4 to be able to work with more different situations.

4. Results

Using again folium library to draw a map, but this time separating neighborhoods by the recent created clusters. In this image is clearly showed the different zones (outland, east, west…) of each cluster that determine the similarities of each neighborhood.



Once we got our data ready, we could answer questions like:

4.1 Which are the most interesting neighborhoods to start an "Italian restaurant"?

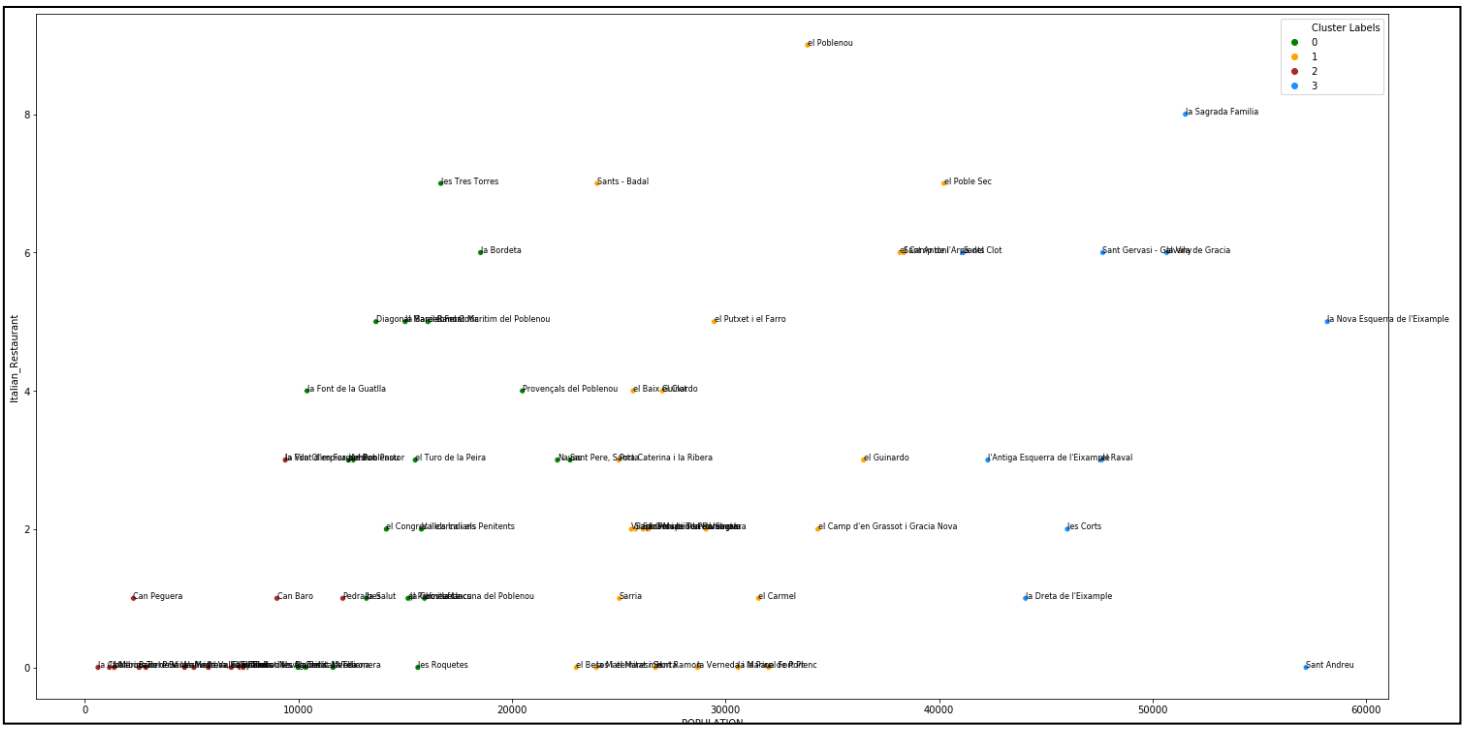
Italian gastronomy is pretty famous around the world, and Barcelona is not an exception. There are a lot of pizza and pasta restaurants around de city, but unfortunately Foursquare label them different ('Italian restaurant' and 'pizza restaurant').

In order to get a bigger data work with I decided to add them and create a new dataframe keeping key information like **Neighborhood**, **Population**, **Cluster labels** and the number of **Italian restaurants**.

It also has been created a new calculated field with the **population per Italian restaurant**.

	Neighborhood_main_field	POPULATION	Italian_Restaurant	Popu/rest	Cluster Labels
0	el Poblenou	33843	9	3760.0	1
1	la Sagrada Familia	51539	8	6442.0	3
2	les Tres Torres	16667	7	2381.0	0
3	el Poble Sec	40228	7	5746.0	1
4	Sants - Badal	23987	7	3426.0	1
5	la Bordeta	18530	6	3088.0	0
6	Sants	41127	6	6854.0	3
7	la Vila de Gracia	50662	6	8443.0	3
8	el Camp de l'Arpa del Clot	38168	6	6361.0	1
9	Sant Antoni	38345	6	6390.0	1
10	Sant Gervasi - Galvany	47666	6	7944.0	3

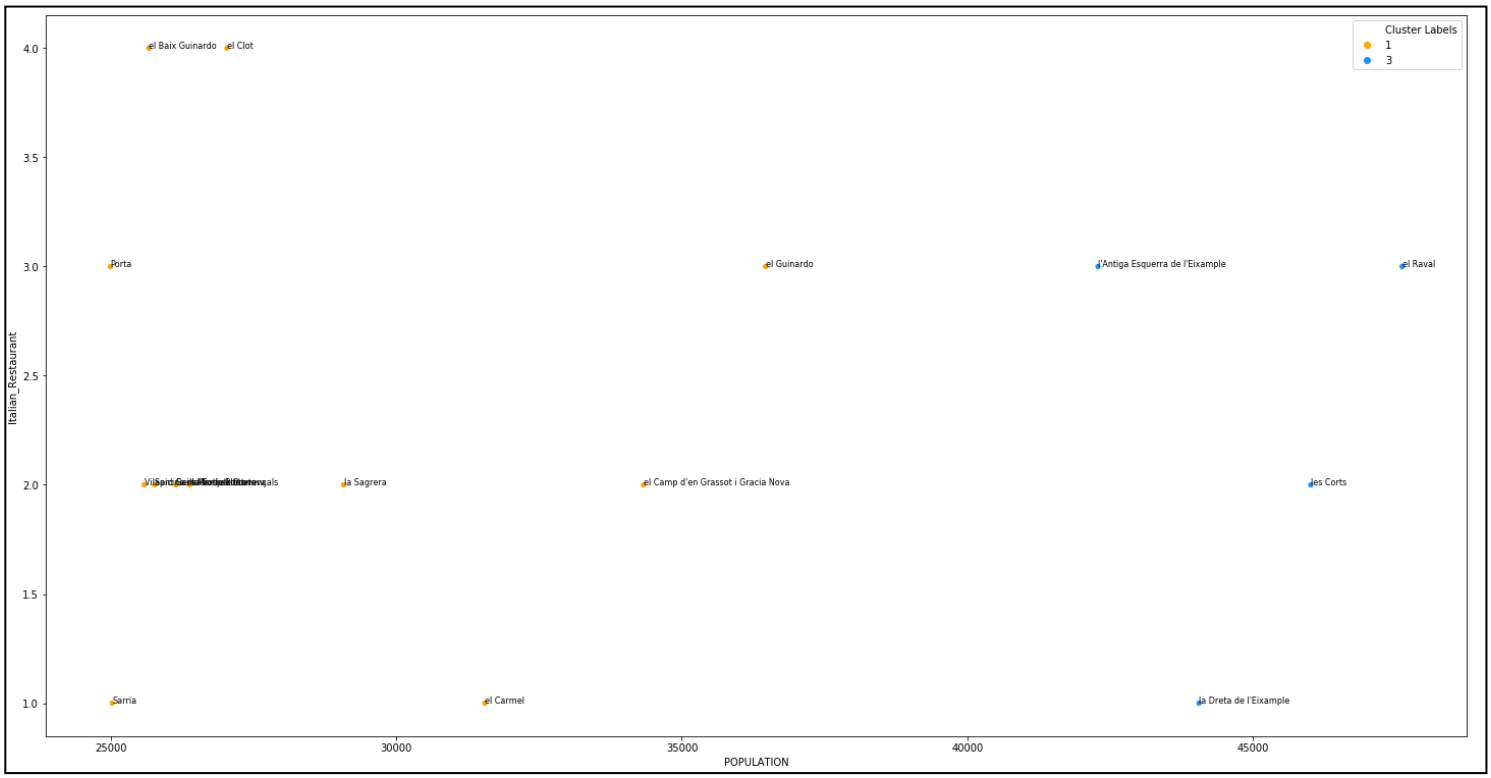
Next I decided to create a scatterplot to get an easier idea of which neighborhoods would be better to open a new restaurant of this kind.



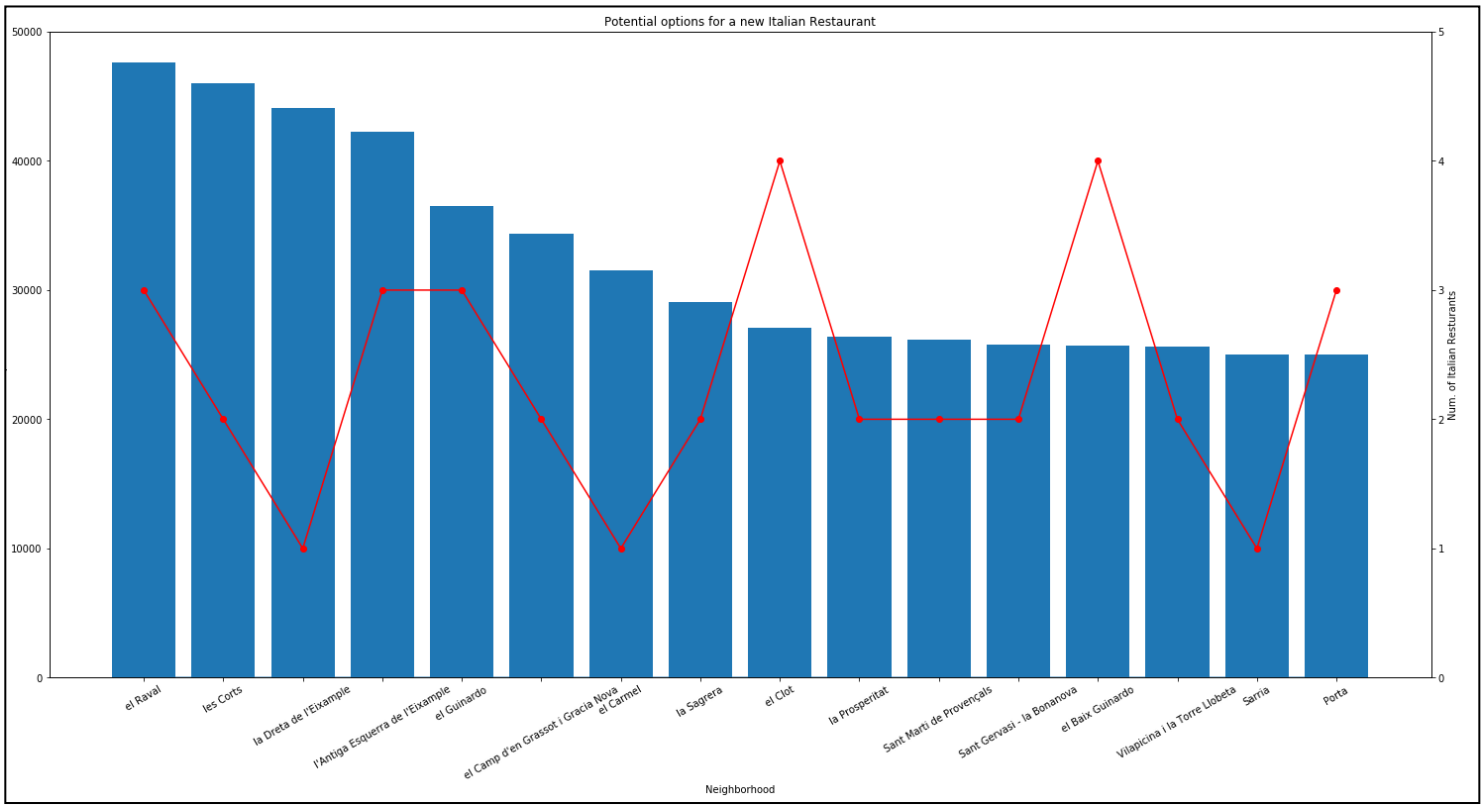
Sadly, as can be seen in the picture, there is too many values to show the information correctly. That is why I decided to filter following the

- Keep only the two cluster with a higher population (1 and 3)
- Drop neighborhoods with more than 4 Italian restaurants, because there is too many unnecessary competence.
- Drop neighborhoods with 0 Italian restaurants, because it is always a safer to start in a zone with an existing market.

The result is the following:



Even an easier way to see the best option for our search is using a bar plot sorted by population and showing too the number of Italian restaurants as a linear plot:

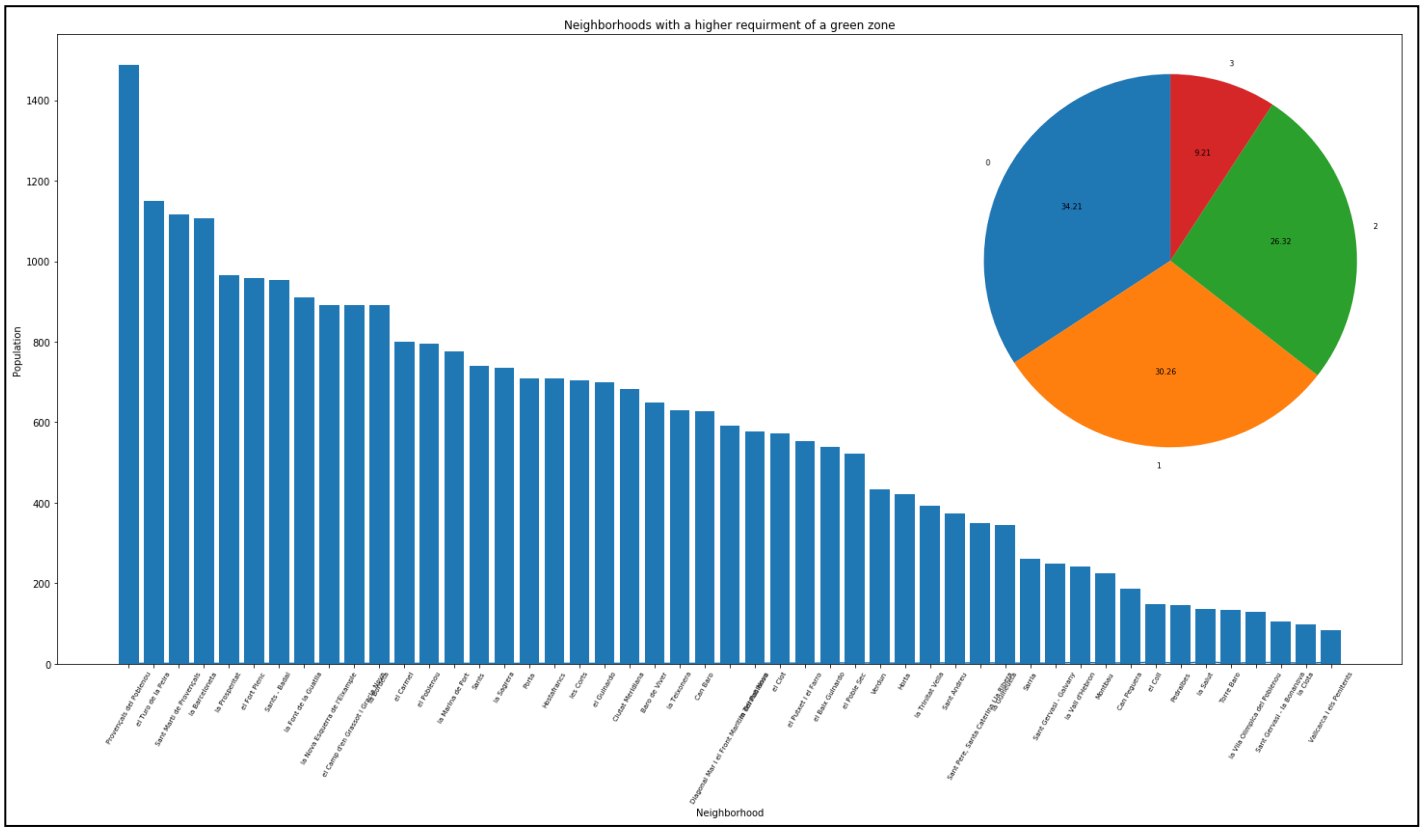


Any one of the eight first neighborhoods would be an excellent selection, but seems like that 'la Dreta de l'Eixample' is the optimal neighborhood to start a new Italian restaurant.

4.2 Which neighborhoods are in greater need of a green area (Park)?

In this case it was necessary compare the amount of **Park** of every neighborhood with the population density of each one. I created a similar dataframe to the previous one:

	Neighborhood_main_field	NET_DENSITY	Park	Net_density/Park	Cluster Labels
0	la Salut	680.1	5	136.0	0
1	el Coll	597.6	4	149.0	2
2	Vallcarca i els Penitents	335.1	4	83.0	0
3	Sant Gervasi - la Bonanova	314.0	3	104.0	1
4	la Vila Olimpica del Poblenou	387.3	3	129.0	2
5	la Vall d'Hebron	724.0	3	241.0	2
6	Sant Andreu	746.2	2	373.0	3
7	el Poble Sec	1043.1	2	521.0	1
8	el Baix Guinardo	1081.9	2	540.0	1
9	Verdun	866.7	2	433.0	0
10	Sant Pere, Santa Caterina i la Ribera	700.1	2	350.0	0



Visualizing the result on this graphs we can get the neighborhoods that need more a new green zone, and as it could be expected, not all of them are the ones which have less, because the density of each neighborhood is a key point to have in mind.

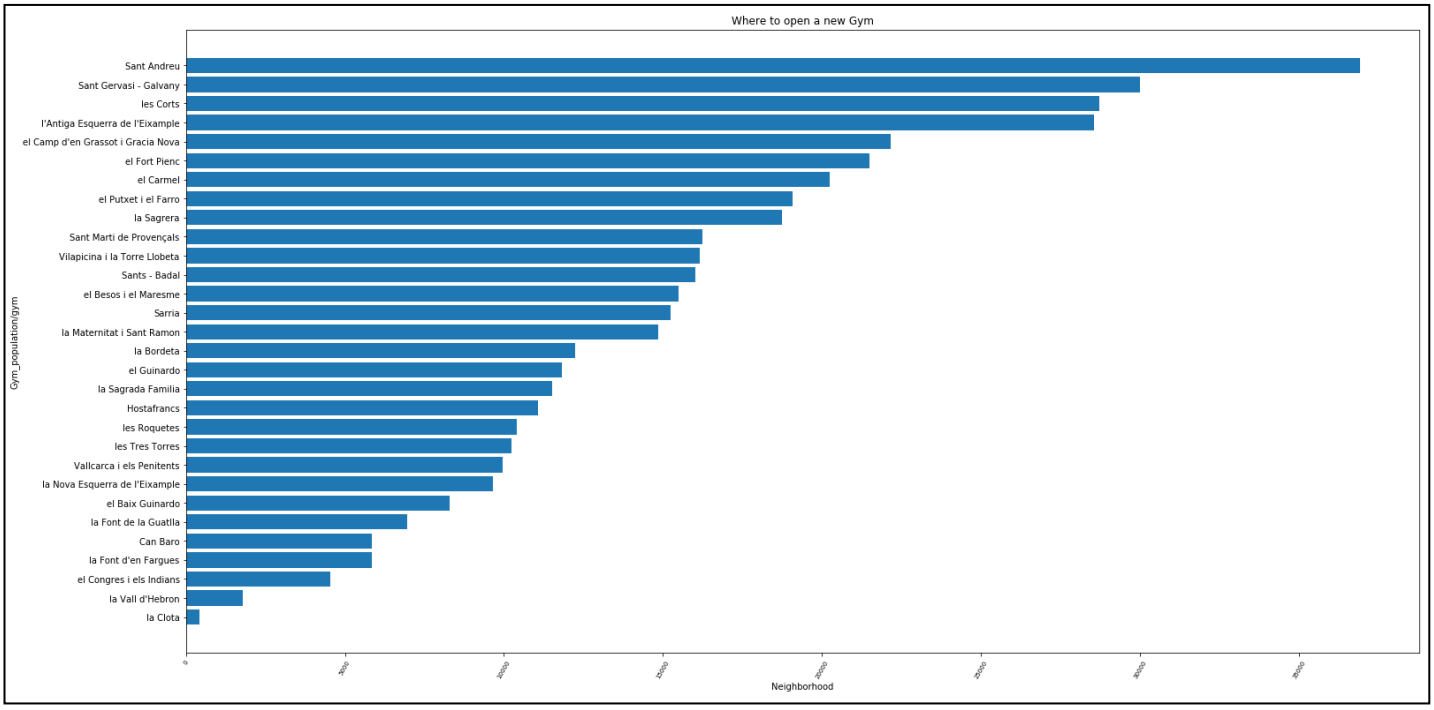
Moreover, I expected a relationship between the requirement of parks and the different neighborhood cluster, but surprisingly there was not. All clusters have a similar percentage, except cluster number 3, which is the one with less neighborhoods.

4.3 Which are the zones with a higher potential clients to open a new Gym?

In this situation, I created a new dataframe including the neighborhoods with a Gym near, and I decided to focus on population age that goes to a gym more frequently. So after a cleaning processes and adding fields like '15 – 24 years' and ' 25 – 64 years' to de dataframe, the results was this one:

	Neighborhood_main_field	15-24 years	25- 64 years	Gym	gym_population/gym	Cluster Labels
19	la Clota	53	378	1	431.0	2
4	la Vall d'Hebron	510	3063	2	1786.0	2
3	el Congres i els Indians	1270	7797	2	4533.0	0
10	la Font d'en Fargues	885	4949	1	5834.0	2
15	Can Baro	768	5087	1	5855.0	2
8	la Font de la Guatlla	887	6083	1	6970.0	0
5	el Baix Guinardo	2083	14484	2	8283.0	1
0	la Nova Esquerra de l'Eixample	4868	33736	4	9651.0	3
7	Vallcarca i els Penitents	1326	8632	1	9958.0	0
11	les Tres Torres	2099	8146	1	10245.0	0
23	les Roquetes	1616	8784	1	10400.0	0
25	Hostafrancs	1315	9742	1	11057.0	0
1	la Sagrada Familia	4331	30249	3	11526.0	3

Again, a bar plot can simply show which are the neighborhoods that we are finding to open a new Gym in the city. The first four options are the optimal selections, and even we could like to find which of this options more similar to a neighborhood where we already have an own Gym, and then use the cluster labels to choose the perfect one.



5. Discussion

As mentioned at the beginning of this report, the main objective here was to gather the highest amount of information about every neighborhood in Barcelona, in order to cluster them correctly and to use its data resolve business questions.

However, there were some key factors in the study that are worthy of note and to discuss:

- The neighborhood datasets needed a worthy amount of work to search, process and clean and even create them manually. Now that these new datasets have been published, I hope it helps in future studies around the city.

- The k-means was analyzed using elbow method, which worked but did not show an exact result, and the number of clusters could vary between 2 to 5.

- Foursquare API is a great option to obtain the location of the most relevant venues in the city. But the results are strongly influenced by the activity of his community on every zone, and in Spain the support is not as big as in United States or other countries. That is why there were a lot of locals missing, some of them not correctly labeled, and overcrowded by the same kind of restaurants. In the end that is what made that the model looked much more affected by population and density than by venues values.

6. Conclusion

In my opinion, this kind of studies can end up helping people willing to open new locals more that we could expect. It is undeniable that the number of local openings will keep increasing, but at the same time the new building space is limited.

That is why new investors will need to gather a lot of information before start a new business in the city, and one of them should be the information showed in this study. Furthermore, even city managers could benefit from this results, as showed in the examples.

However, if this kind of information must be used by final clients as investors or city manager, there are a few improves that in my opinion should be approached, like:

- **Find a better source of venues information:** As mentioned before, the venues results showed from every application depends on the community support from each country, and in this situation I think that maybe some alternatives to Foursquare could achieve better results.
- **Add local ratings to the model:** I think it would be interesting to include the ratting of each local to the model, in order to be able to separate the best rated to the worst, and find new opportunities around 'bad-rated' business.
- **Automatize results:** Finally one of my initial ideas was to create a simply program to enable users to get all the information from a zone only by writing a few lines, like the venue category, district or potential clients age. This functionality was dismissed because the amount of work required, but I think it would be an excellent way to introduce this solution to a higher user market.