David López Companys

FINDING THE OPTIMAL LOCATION FOR A NEW STORE

Content:

1. Problem description

In this project, the problem attempted to solve will be to find the best possible location or the most optimal, for a luxury supermarket in the city of Barcelona (Catalunya).
To achieve this task, an analytical approach will be used, techniques used will be focused on clustering to have a group of optimal neighborhoods where index of immigration and rental costs fit the supermarket needs and capabilities.
During the process of analysis, several data transformations will be performed, in order the find the best possible data format for the machine learning model to ingest.
Once the data is set up and prepared, a modeling process will be carried out, and this statistical analysis will provide the best possible places to open this new supermarket.

2. Data presentation

The data used for this analytical process will come from dispare open data sources, which are:

1. Foursquare API: This data will be accessed via Python and used to obtain the most common venues per neighborhood in the city of Barcelona. This way, it is possible to have a taste of how the city's venues are distributed, what are the most common places for leisure, and in general, it will provide an idea of what people's likes are.
2. % of immigration by nationality and neighbourhood: This information is provided as an open data source by the Barcelona's city hall, the information provided includes:
    a. Year
    b. Month
    c. Neighbourhood
    d. Country / Region of immigrant
    e. Number of immigrants

    This information can be downloaded as csv file in the following url: https://opendata-ajuntament.barcelona.cat/data/es/dataset/est-demo-immigrants-lloc-procedencia
3. Average rental price per neighbourhood: This information is also provided by the Barcelona's city hall as statistical data, the information contained in this dataset is as follows:
    a. Neighbourhood
    b. Average price per sqm / per year

    This dataset can be acceced via the following url: https://www.bcn.cat/estadistica/angles/dades/timm/ipreus/hab2mave/evo/t2mab.htm

This three data sources will be used to cluster the different neighbourhoods of barcelona in different groups depending on the:

- %immigration by precedence living in every single neighbourhood and venue.
- rental price

With this information, first the optimal number of clusters will be found via elbow method (or others, to be defined which fits better in development), after the optimal number of clusters is found, different clustering methods will be applied:

- Partitioning clustering: K-means
- Hierarchical clustering: Bottom up

With that, venues will be clustered in how they perform similar in the different variables analysed, having that information and optimal cluster will be chosen to place the new supermarket depending on budget available and immigration.

3. Methodology

The methodology used to find the best solution for this statistical problem consists on using the data of immigration and venues comming from Foursquare to cluster the distinct district of Barcelona to fins the cluster of district that best fit to our supermarket.

To be more concrete, the algorithm used is the K-means algorithm with the Elbow method to find the optimal number of clusters.

Here comes and example of the data of population downloaded from the different data sources:

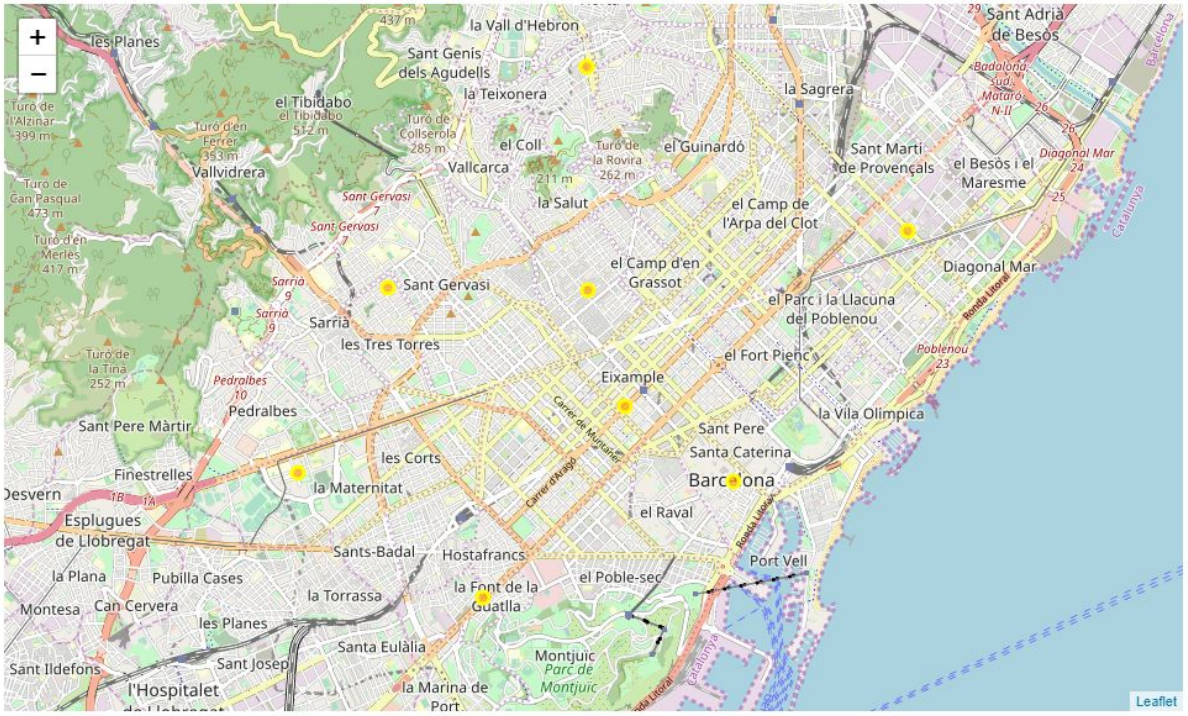
	Lloc_de_procedencia	Nom_districte	Nombre
2	Andalusia	Ciutat Vella	223
3	Andalusia	Eixample	476
4	Andalusia	Gracia	187
5	Andalusia	Horta-Guinardó	180
6	Andalusia	Les Corts	89
7	Andalusia	Nou Barris	204
8	Andalusia	Sant Andreu	125
9	Andalusia	Sant Marti	302
10	Andalusia	Sants-Montjuic	290
11	Andalusia	Sarria-Sant Gerv	125
12	Aragó	Ciutat Vella	69
13	Aragó	Eixample	167
14	Aragó	Gracia	49
15	Aragó	Horta-Guinardó	64
16	Aragó	Les Corts	54
17	Aragó	Nou Barris	57
18	Aragó	Sant Andreu	52
19	Aragó	Sant Marti	84
20	Aragó	Sants-Montjuic	103
21	Aragó	Sarria-Sant Gerv	43
22	Canaries	Ciutat Vella	106
23	Canaries	Eixample	181
24	Canaries	Gracia	60
25	Canaries	Horta-Guinardó	65
26	Canaries	Les Corts	28
27	Canaries	Nou Barris	69
28	Canaries	Sant Andreu	38
29	Canaries	Sant Marti	103
30	Canaries	Sants-Montjuic	104
31	Canaries	Sarria-Sant Gerv	70
32	Cantabria	Ciutat Vella	19

The first step done is to manipulate that dataset and group it by district and immigration origin:

:

	Unnamed: 0	Ciutat Vella	Eixample	Gracia	Horta-Guinardó	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi
0	Andalusia	223	476	187	180	89	204	125	302	290	125
1	Aragó	69	167	49	64	54	57	52	84	103	43
2	Canaries	106	181	60	65	28	69	38	103	104	70
3	Cantabria	19	32	17	12	5	11	22	13	27	8
4	Castella - la Manxa	49	77	26	36	20	39	40	71	43	20

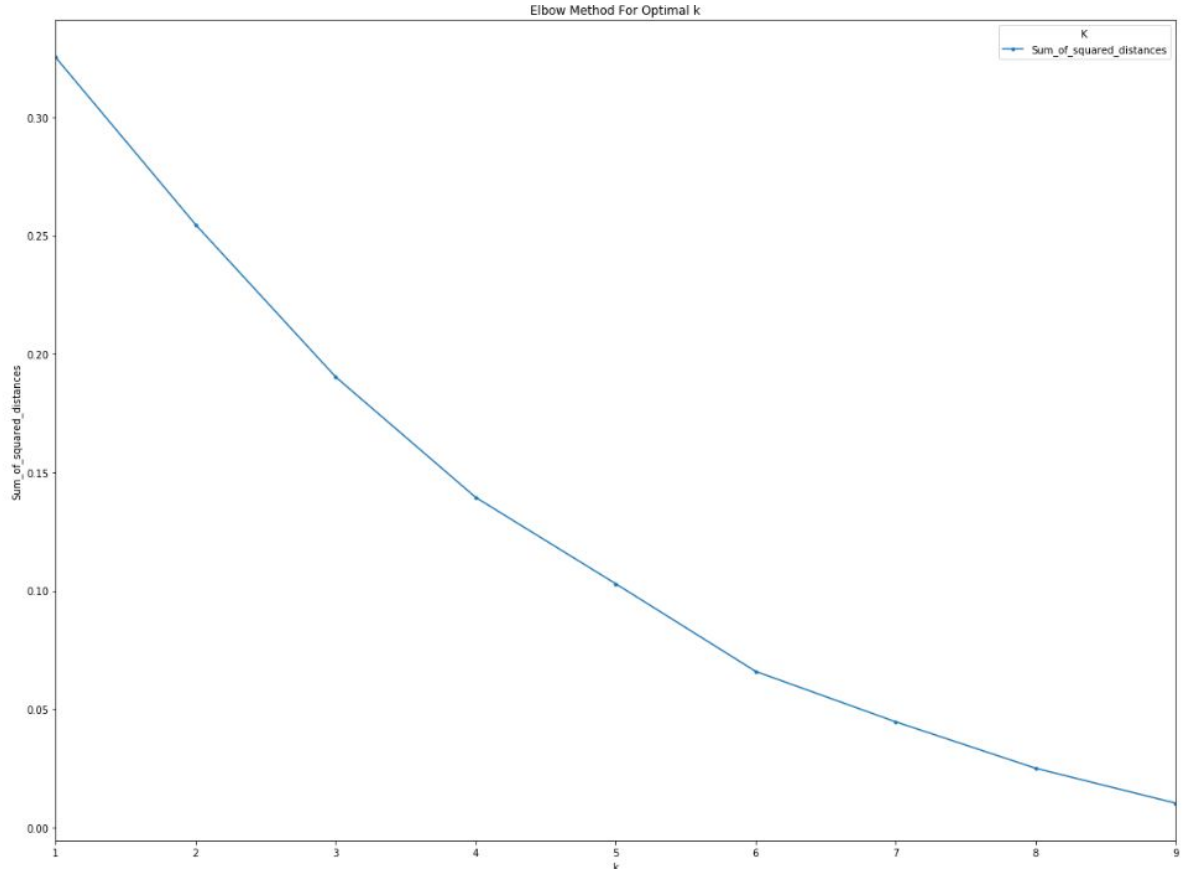
Once our data is ready to be analysed we made a geolocation of the distinct district of Barcelona and plot it in a map to view the distribution of them:



The next step has been to find the nearby venues of each district via API calls to the Foursquare developer API, the results from that part after being transformed look like follows:

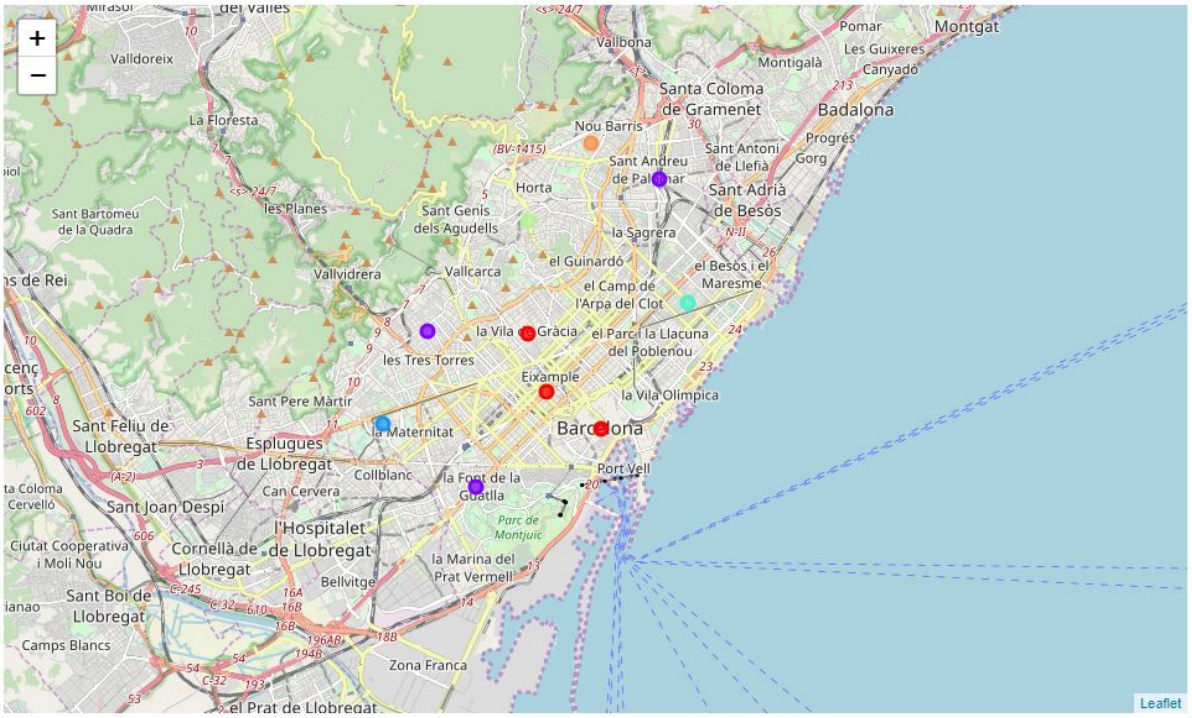
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ciutat Vella	Tapas Restaurant	Plaza	Bar	Spanish Restaurant	Wine Bar	Hotel	Cocktail Bar	Vegetarian / Vegan Restaurant	Ice Cream Shop	Tea Room
1	Eixample	Hotel	Tapas Restaurant	Boutique	Clothing Store	Mediterranean Restaurant	Hostel	Japanese Restaurant	Cocktail Bar	Gym	Bookstore
2	Gracia	Tapas Restaurant	Cocktail Bar	Bar	Café	Plaza	Pizza Place	Ice Cream Shop	Restaurant	Mediterranean Restaurant	Arts & Crafts Store
3	Horta-Guinardó	Supermarket	Food & Drink Shop	Plaza	Gym	Grocery Store	Food Court	Café	Coffee Shop	Metro Station	Chinese Restaurant
4	Les Corts	Tram Station	Soccer Field	Nightclub	Gym	Garden	Print Shop	College Cafeteria	Donut Shop	Museum	Science Museum
5	Nou Barris	Spanish Restaurant	Grocery Store	Plaza	Park	Diner	Building	Seafood Restaurant	Metro Station	Brewery	Skate Park
6	Sant Andreu	Spanish Restaurant	Café	Supermarket	Restaurant	Tapas Restaurant	Park	Brewery	Plaza	Coffee Shop	Bar
7	Sant Marti	Hotel	Pizza Place	Supermarket	Park	Convenience Store	Office	Falafel Restaurant	Mediterranean Restaurant	Café	Kebab Restaurant
8	Sants-Montjuic	Spanish Restaurant	Restaurant	Grocery Store	Plaza	Hotel	Tapas Restaurant	Breakfast Spot	Burger Joint	Supermarket	Bakery
9	Sarria-Sant Gervasi	Bakery	Restaurant	Coffee Shop	Spanish Restaurant	Hotel	Seafood Restaurant	Bar	Café	Mediterranean Restaurant	Japanese Restaurant

After that point, we are ready to apply our k-means algorithm to find the different clusters of districts with those two variables.
First, we need to find the optimal number of clusters using the elbow method:



With that, we identified 6 as the number of clusters needed, with that, we are now ready to apply the K-means algorithm.
Once the results are transformed and plotted in one single table with all the information, the results look like as follow:

Map:



Dataset:

	Unnamed: 0	Ciutat Vella	Eixample	Gracia	Horta-Guinardó	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3r Co
0	Andalusia	223	476	187	180	89	204	125	302	290	125	Ciutat Vella	0	Tapas Restaurant	Plaza	
1	Aragó	69	167	49	64	54	57	52	84	103	43	Eixample	0	Hotel	Tapas Restaurant	Bc
2	Canaries	106	181	60	65	28	69	38	103	104	70	Gracia	0	Tapas Restaurant	Cocktail Bar	
3	Cantabria	19	32	17	12	5	11	22	13	27	8	Horta-Guinardó	4	Supermarket	Food & Drink Shop	
4	Castella - la Manxa	49	77	26	36	20	39	40	71	43	20	Les Corts	2	Tram Station	Soccer Field	Ni
5	Castella i Lleó	67	146	52	75	43	54	60	103	105	39	Nou Barris	5	Spanish Restaurant	Grocery Store	
6	Catalunya	2619	4541	1960	2799	1579	3265	2639	3889	4214	2297	Sant Andreu	1	Spanish Restaurant	Café	Super
7	Ceuta	5	2	0	5	1	4	0	5	5	4	Sant Marti	3	Hotel	Pizza Place	Super
8	Comunitat Foral de Navarra	15	59	18	20	12	16	20	25	24	29	Sants-Montjuic	1	Spanish Restaurant	Restaurant	C
9	Comunitat Valenciana	218	432	175	190	96	144	167	309	297	141	Sarria-Sant Gervasi	1	Bakery	Restaurant	Coffe

Now, we have all the information available to be able to make a final decision on which cluster or districts are the best place to put our supermarket.

4. Result

Cluster 1:

	Unnamed: 0	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	Canaries	28	69	38	103	104	70	Gracia	0	Tapas Restaurant	Cocktail Bar	Bar
1	Aragó	54	57	52	84	103	43	Eixample	0	Hotel	Tapas Restaurant	Boutique
0	Andalusia	89	204	125	302	290	125	Ciutat Vella	0	Tapas Restaurant	Plaza	Bar

Cluster 2:

	Unnamed: 0	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9	Comunitat Valenciana	96	144	167	309	297	141	Sarria-Sant Gervasi	1	Bakery	Restaurant	Coffee Shop
8	Comunitat Foral de Navarra	12	16	20	25	24	29	Sants-Montjuic	1	Spanish Restaurant	Restaurant	Grocery Store
6	Catalunya	1579	3265	2639	3889	4214	2297	Sant Andreu	1	Spanish Restaurant	Café	Supermarket

Cluster 3:

	Unnamed: 0	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
4	Castella - la Manxa	20	39	40	71	43	20	Les Corts	2	Tram Station	Soccer Field	Nightclub

Cluster 4:

	Unnamed: 0	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
7	Ceuta	1	4	0	5	5	4	Sant Marti	3	Hotel	Pizza Place	Supermarket

Cluster 5:

	Unnamed: 0	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
3	Cantabria	5	11	22	13	27	8	Horta-Guinardó	4	Supermarket	Food & Drink Shop	Plaza

Cluster 6:

	Unnamed: 0	Les Corts	Nou Barris	Sant Andreu	Sant Marti	Sants-Montjuic	Sarria-Sant Gervasi	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
5	Castella i Lleó	43	54	60	103	105	39	Nou Barris	5	Spanish Restaurant	Grocery Store	Plaza

5. Discussion

It is interesting how the venues and people from different countries varies to one cluster to another. The main differentiation is located on these two variables.
Each cluster has its own characteristics, but also common spots with other clusters. If we examine with more detail these results, some conclusions can be made.
As a recommendation, it must be said in a study of this size, to make good predictions about where to open a certain business or shop, more data is needed. For example, socio-demographic data about the population, like their income level, if they have children or not, the education level, what kind of job do they make a living from, etc.…
Also, one of the most important data to examine carefully are the data related to the people's likes and tastes about how they prefer to spend their leisure time, what kinds of food do they like, or what are their hobbies. With all these data gathered, a more in depth analysis could be performed, and the segmentations would be more accurate. For this project, these data weren't available, and was also out of the project's scope.

6. Conclusion

Analysing the data that we have been working in with the main goal to find the best districts to place a luxury product supermarket, the conclusions are as follow:
- The district where service similar to our luxury supermarket exist is the district of Sarria, where we can find that there are supermarkets similar.
- With that, we can assume that this district, the current population of it, likes this type of supermarkets.

- Then, the next step is to find similar districts where population is quite similar to Sarria, but no services like this exist.

Conclusion: We can figure out that district cluster 2 is the most suitable to place our supermarket, so districts of Sants-Monjuic and Sant Andreu should be a good place to start our new supermarket.