New opening location decision using machine learning

Abstract

This Capstone Project for IBM-DataScience course consists of comparing and analysing the metropolitan areas of the two biggest cities of Spain (Madrid and Barcelona) by exploring the places around the subway stations. The target is to find the best areas in Barcelona to open a new gym of a very successful gym firm from Madrid based on neighborhood similarity. The analysis will perform using the unsupervised machine learning K-means clustering method with public transport station locations. The differences between clusters are based on the venue's categories around the station area. The result of this study will help the gym company to decide the places that are suitable for the next opening.

1 Introduction/Business problem

1.1 Background

Barcelona is the capital and the largest city of Catalonia, which is in the northeastern of Spain in the Mediterranean Sea coast. Barcelona has a population of 1.6 million within the city limits and 5.2 million considering the metropolitan area. These numbers convert the city to the sixth most populated area in the European Union. The population density is 15.9 k inhabitants per square kilometer, but the city center has a population density of 40.3 k inhabitants per km^2. Inhabitants between 25 and 44 years old, which are the primary target, represent the 30.6 percent of the population. All these facts make the city a perfect place for business expansion.

The historical attractions, the sports, the good climate and the night entertainment, are keeping the tourist income growing. It is the 4^{th} preferred place to visit among tourists in Europe. Last year Barcelona hosted 9 million of tourists. Nevertheless, this is not the unique motor of the city, the best universities of Spain, industries and many international businesses are placed here, making Barcelona attractive for investments.

On the other hand, Madrid is the capital of Spain and the biggest city situated in the strategic center of the country. The capital has a population of 3.3 million and 6.5 million, considering the metropolitan area. It is regarded as the major financial center and the leading economic hub of the Iberian Peninsula. The population density is 5 k inhabitants per square kilometer and the city center has a population density 26 k inhabitants per km^2. And the middle age structure is 38% of the population. We could find similar features between two cities apart from tourist impact.

1.2 Problem definition

A successful gym club chain from Madrid is planning to open a new gym in Barcelona. The locations of the gyms are usually placed in residential areas, near to downtown and close to a subway or train station. They are looking for locations in Barcelona with similar features to place the new gym. This study will cluster the subway and train stations of both cities, taking into account the kind of venues around. To concrete the best places, we will also make a comparison between the population density in the neighborhoods.

2 Data

2.1 Data requirements

- List of subway and train stations and coordinates within Madrid metropolitan area
- List of subway and train stations and coordinates within Barcelona metropolitan area
- List neighborhoods and its demographic data of Madrid and Barcelona.
- List of the venues close to the stations. We will use o Foursquare data for this. This data will be the key to identify similar areas.

2.2 Data sources

The list of all stations, names, type and coordinates within the metropolitan area of Barcelona can be easily downloaded from the Open Data service of the city: https://opendata-ajuntament.barcelona.cat/data/es/dataset/transports/resource/e07dec0d-4aeb-40f3-b987-e1f35e088ce2

For stations data of the Madrid metropolitan area, there is also an open data service, but in this case, the data is divided. The subway (metro) stations coordinates can be downloaded from here: https://datos.madrid.es/egob/catalogo/200073-12-puntos-transporte-navegadores.gpx and the train (cercanias) stations can be found here: https://datos.madrid.es/egob/catalogo/200073-1-puntos-transporte-navegadores.gpx. The links may be temporally, so here there is the general link for all kind of transports: https://datos.madrid.es/sites/v/index.jsp?vgnextoid=08055cde99be2410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

The administrative division of the Madrid neighborhoods can be downloaded as shapefile from https://datos.madrid.es/egob/catalogo/200078-10-distritos-barrios.zip

The administrative division of the Barcelona neighbourhoods can be downloaded as JSON here: https://github.com/martgnz/bcn-geodata/tree/master/barris or from https://opendata-ajuntament.barcelona.cat/data/es/dataset/20170706-districtes-barris

The demographic data of Barcelona neighborhoods can be scrapped from here: https://www.bcn.cat/estadistica/castella/dades/inf/lecpadro/lec19/t13.htm

The demographic data of Madrid neighborhoods can be downloaded with a tool on this page: http://www-2.munimadrid.es/TSE6/control/seleccionDatosBarrio

2.3 Data cleaning

Madrid Data

For Madrid stations, we have downloaded train stations XML file which is easily converted to a dataframe using *XML.etree.ElementTree* library for getting the clean data:

Table 1. Train stations data

	station	line	Latitude	Longitude
0	AEROPUERTO T4	C-1	40.49176762416515	-3.593336761776833
1	ATOCHA	C-1	40.40658651228024	-3.6893092868843853
2	CHAMARTIN	C-1	40.47209555413991	-3.6824760485471546
3	DELICIAS	C-1	40.400367595318635	-3.6927692985734324
4	FUENTE DE LA MORA	C-1	40.4847413558896	-3.662829237134829

And we do the same for the Madrid subway data:

	station	line	Latitude	Longitude
0	ALTO DEL ARENAL	1	40.389768700352434	-3.6452252835565764
1	ALVARADO	1	40.45033105413335	-3.7033178070796913
2	ANTON MARTIN	1	40.41246302818812	-3.6993757416732405
3	ATOCHA	1	40.408846792956474	-3.692490883566612
4	ATOCHA RENFE	1	40.406586090726144	-3.689379993715528

The data preparation will consist of:

1. Merging both datasets
2. Names without coordinates will be removed
3. Several stations are repeated because they can belong to different lines or exists in both tables. We will only keep one

For the Foursquare query, the coordinates are the important data for venues exploring, but we will keep the station and line categories and use later as representation labels.

Madrid Neighborhood limits have been downloaded as a shapefile which is read directly in python using the library geopandas, an extension for pandas for geometry datasets:

	OBJECTID	geodb_oid	CODDIS	NOMDIS	CODBAR	CODDISTRIT	CODBARRIO	NOMBRE	ORIG_FID	geometry
0	108	108	17	Villaverde	172	17	17-2	San Cristobal	107	POLYGON ((441930.8668000005 4466853.1887, 4419...
1	109	109	17	Villaverde	173	17	17-3	Butarque	108	POLYGON ((444144.8566044134 4464473.210504748,...
2	111	111	17	Villaverde	175	17	17-5	Los Angeles	110	POLYGON ((441147.7280000008 4466374.483400001,...

Barcelona Data

In this case, the metro and train stations are in a csv file. The file includes one point per station entrance. The data preparation will consist of:

1. Removing the NaN values
2. Separating the station names into two columns ‘line’ and ‘station’ as we have done with Madrid data.
3. Removing the duplicate stations.

	EQUIPAMENT	NOM_BARRI	Longitude	Latitude	line	station
0	METRO (L3, L5) - VALL D'HEBRON (C. de les Bass...	la Vall d'Hebron	2.142987	41.424923	METRO (L3, L5)	VALL D'HEBRON
1	FGC - PROVENÇA (C. de Provença)-	l'Antiga Esquerra de l'Eixample	2.158326	41.392331	FGC	PROVENÇA
2	FGC (L6) - REINA ELISENDA (Sortida Duquesa d'O...	Sarrià	2.119370	41.399203	FGC (L6)	REINA ELISENDA
3	FGC (L6) - LA BONANOVA-	Sant Gervasi - Galvany	2.135427	41.397791	FGC (L6)	LA BONANOVA
4	METRO (L11) - CASA DE L'AIGUA (C. Vila-Real)-	la Trinitat Nova	2.185391	41.451492	METRO (L11)	CASA DE L'AIGUA

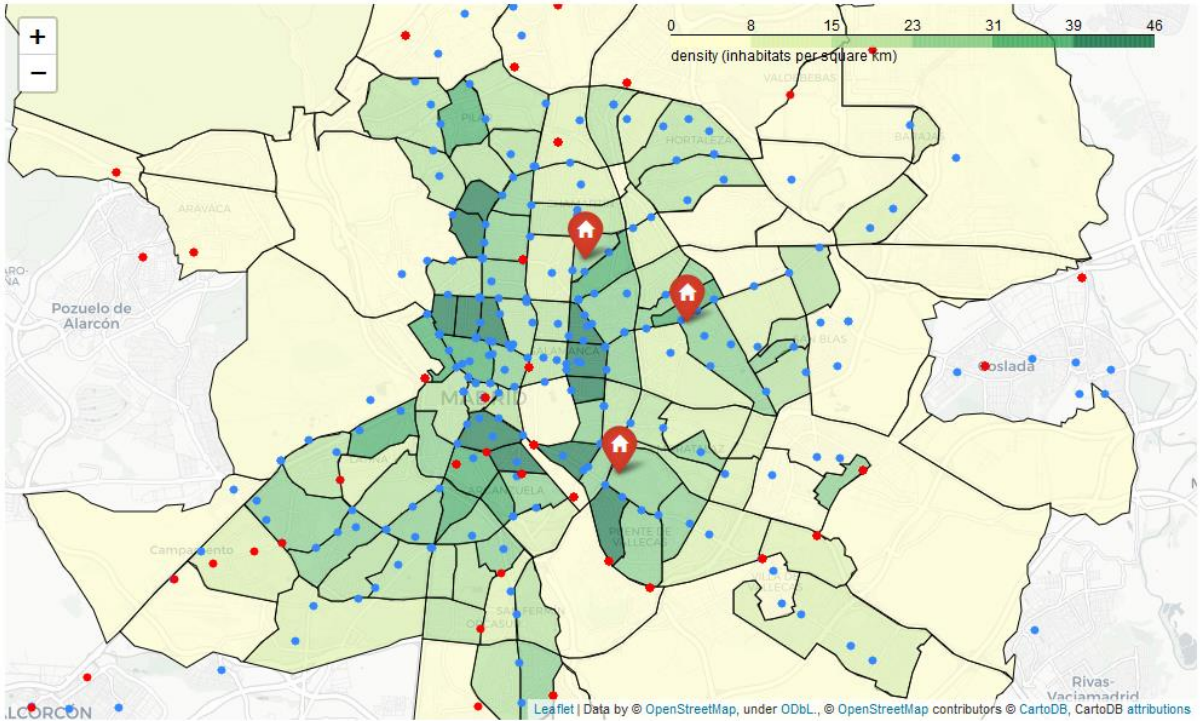
For Barcelona neighborhoods limits, the data is a shapefile that can be used directly for the analysis.

3 Methodology

I used GitHub repository in this study. I will put all the data and the Jupyter notebooks I used for this capstone project. The most important libraries I am going to use are *folium*, *pandas*, *geopandas*, *matplotlib*, *seaborn* and *sklearn*.

3.1 Madrid

First, we can visualise the neighborhoods of Madrid and the situation of the gyms on a map. I used python folium library to visualise the Madrid locations and its position. The metro stations are in blue and train stations in red. The choropleth shapes represent the neighborhoods limits and the color scale represents the population density.



The gyms are in three different places represented by a red icon, we can join the neighborhoods shapes, and the gyms coordinates using *sjoin* tool from *geopandas* to see the density of each gym's neighborhood:

	name	NOMBRE	Barrio	density
0	place1	Pueblo Nuevo	PUEBLO NUEVO	27.506699
1	place2	Numancia	NUMANCIA	26.335952
2	place3	Ciudad Jardín	CIUDAD JARDIN	24.884911

The table shows that the gyms are in neighborhoods within a population density of around 26 in/km2. So, this information will be used to decide the suitable locations for the new aperture in Barcelona.

To segment and identify the special features of the gym's areas, we will analyse all the metro and train stations in the city of Madrid and the gym locations.

3.1.1 Foursquare data

I utilised the Foursquare API to explore the venues around the locations and segment them. The venues will be limited to 100 venues in 500 meters radius from each location. We will explore 320 locations in Madrid.

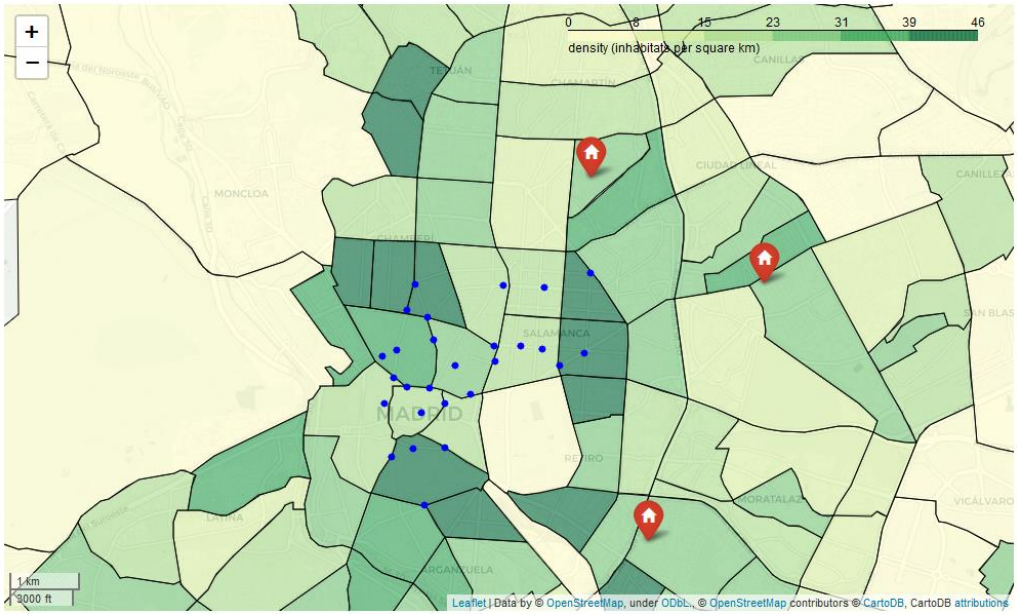
The resulted venue Dataframe consists of a total of 7685 venues. In the table, we see the venues dataset head including the location of the station.

```
In [33]: venues.head()
```

Out[33]:

	stations	stations Latitude	stations Latitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ALTO DEL ARENAL/1	40.389769	40.389769	Stones Rock Bar	40.391972	-3.649905	Bar
1	ALTO DEL ARENAL/1	40.389769	40.389769	Calle Pedro Laborde	40.388294	-3.648227	Building
2	ALTO DEL ARENAL/1	40.389769	40.389769	AhorraMas	40.391022	-3.643129	Grocery Store
3	ALTO DEL ARENAL/1	40.389769	40.389769	Restaurante chino Jardín de bambú	40.388884	-3.642610	Chinese Restaurant
4	ALTO DEL ARENAL/1	40.389769	40.389769	Jardin Bambu	40.388602	-3.641056	Asian Restaurant

All kind of station category in the venues list are removed, I do not want to count the stations for the clustering analysis, the final size is 7587 venues. In the figure, we can see the stations with more venues; they are in the city center as expected.



In the table is shown the top 10 categories I have found:

	Venue Category	n
0	Spanish Restaurant	779
1	Restaurant	487
2	Bar	343
3	Tapas Restaurant	340
4	Hotel	296
5	Café	226
6	Coffee Shop	189
7	Plaza	184
8	Italian Restaurant	159
9	Bakery	158

It is possible to apply some data cleaning before the clustering. The most common venues are restaurants and hotels. One of the things we can do is to combine 'Spanish restaurant' and "Restaurant" because both are the same. "Tapas restaurant" can be considered the same as 'Bar', however, in this case, Tapas Restaurant could be more linked to touristic places, so it is important to keep it for segmentation. We combine 'Café' and 'Coffe Shop' and 'Gym' and Fitness Center'. Finally, we will exclude stations with less than five venues.

For the input data preparation required for k-means clustering, we will use the one-hot encoding for venue categories column. Then we have reduced the dataframe to one row per station with a normalisation based on the frequency of appearance of each category. In the end, we will have a dataframe with all stations and, in columns, the frequencies of each category. The final Dataframe contains 299 columns and 202 rows, and it is exactly the input for k-means. I put an example of this normalisation in the table:

	station	Restaurant	Coffe Shop	Bar	Tapas Restaurant	Grocery Store	Gym	Pizza Place	Supermarket	Hotel	...	Cajun / Creole Restaurant	Lebanese Restaurant	College Classroom
0	ABRANTES/11	0.166667	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.0	0.000000	...	0.0	0.0	0.0
1	ACACIAS/5	0.058824	0.029412	0.058824	0.088235	0.029412	0.058824	0.058824	0.0	0.000000	...	0.0	0.0	0.0
2	AEROPUERTO T1 T2 T3/8	0.000000	0.117647	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.0	0.0
3	AEROPUERTO T4/8	0.090909	0.090909	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.030303	...	0.0	0.0	0.0
4	AGUILAS, LAS/C-5	0.235294	0.058824	0.176471	0.058824	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.0	0.0

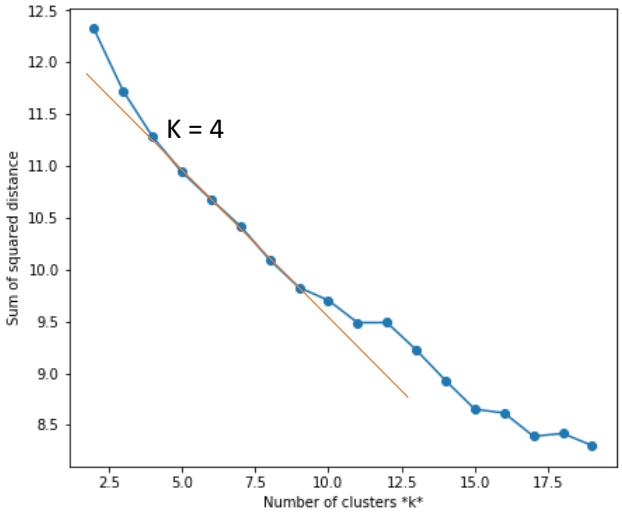
To make the comparison between stations easier, we will create another dataframe with only ten columns showing the ten most common venues per each station:

	station	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	ABRANTES/11	Restaurant	Pizza Place	Soccer Field	Fast Food Restaurant	Athletics & Sports	Park	Lake	Gym Pool	Arcade	Office
1	ACACIAS/5	Tapas Restaurant	Art Gallery	Restaurant	Bar	Gym	Pizza Place	Theater	Event Space	Sporting Goods Shop	Bookstore
2	AEROPUERTO T1 T2 T3/8	Coffe Shop	Airport Service	Airport Terminal	Duty-free Shop	Rental Car Location	Airport Lounge	Airport Gate	Bus Station	Diner	Frozen Yogurt Shop
3	AEROPUERTO T4/8	Airport Service	Restaurant	Coffe Shop	Airport Lounge	Deli / Bodega	Fast Food Restaurant	Bus Station	Accessories Store	Rental Car Location	Police Station
4	AGUILAS, LAS/C-5	Restaurant	Bar	Shopping Mall	Gym Pool	Convenience Store	Seafood Restaurant	Park	Smoke Shop	Sporting Goods Shop	Tapas Restaurant

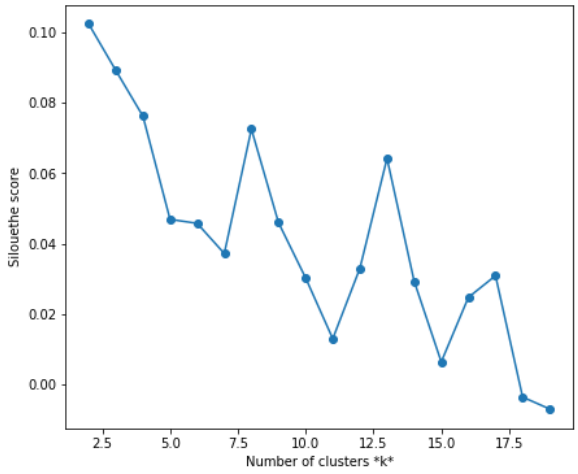
3.1.2 Clustering

3.1.2.1 Number of clusters determination

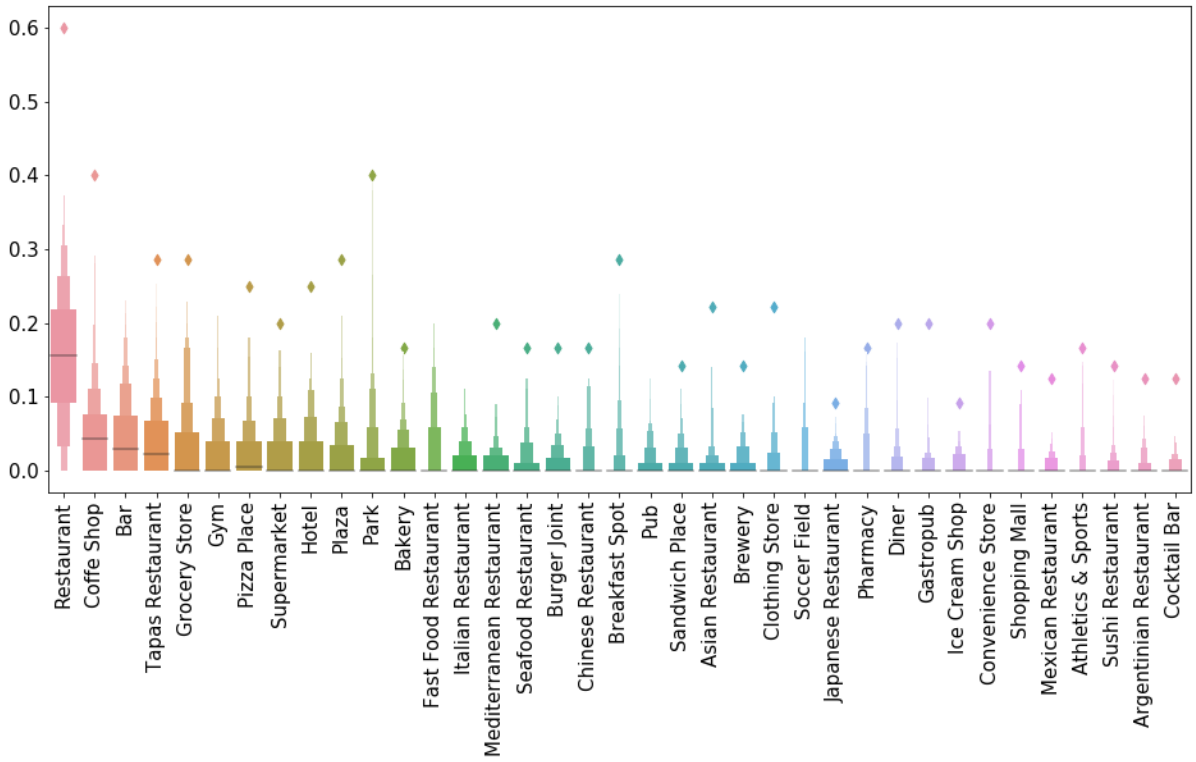
We can use the elbow method to determine the number of clusters. From the graph Sum of squared distances vs Number of clusters (also called inertia), the point of inflation on the curve is the indication that the model fits good at this number of clusters.



In this case, there is a minimum inflation change for k = 4. However, there is not a clear elbow in the curve. We can also use the silhouette score to find the optimum number of clusters:

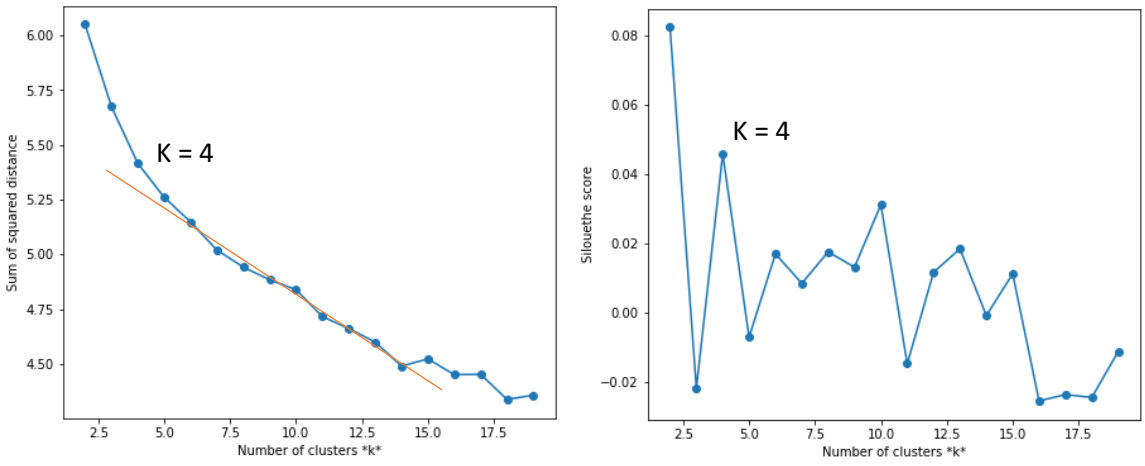


But again, the silhouette score does not give us additional information; we need to investigate the data to see if we can find any bias that might affect the data.



From the boxplot showed above, we see how the restaurants are in every location; this can produce a bias in clustering. To deal with it, we can try to make a k-means weighted, taking the weight as the total number of categories we have found per each station.

For the k-means weighted method, we obtain these curves for inertia and silhouette score:



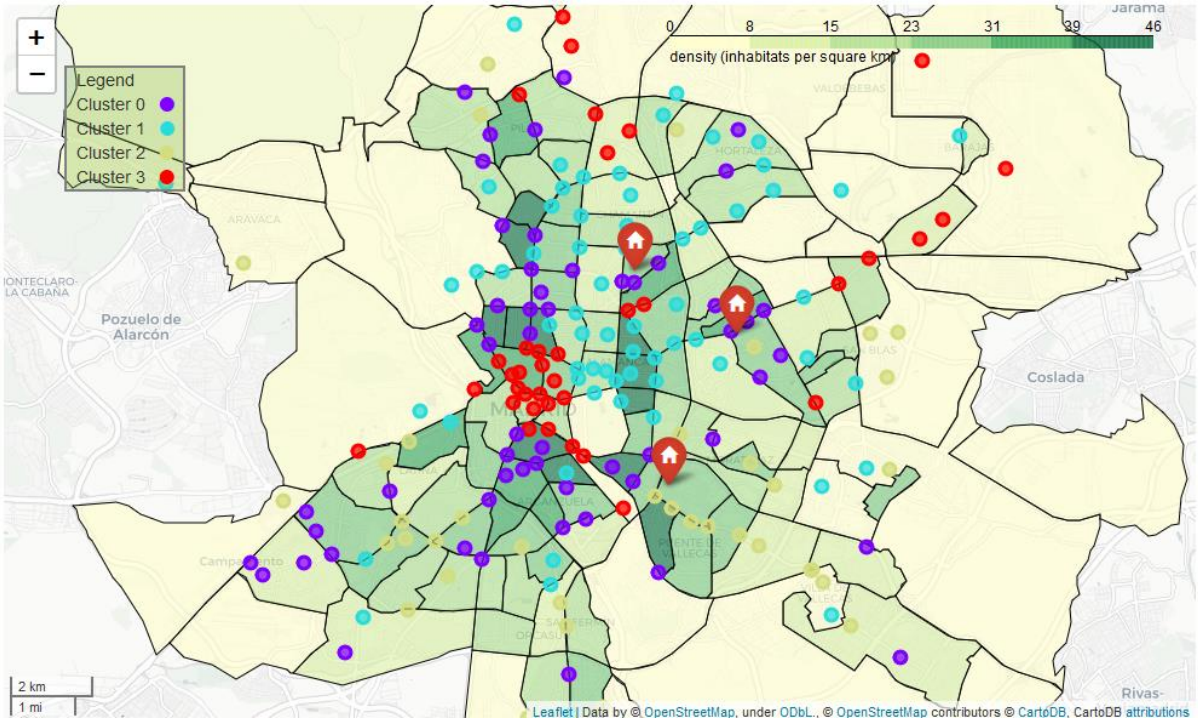
The scores are better than with the unweighted case, now the curves are better but it is hard to determine the optimum number of clusters from the graphs. Nevertheless, from the curves and after tries with a different number, we have seen that k = 4 is a good option.

3.1.3 Results

Finally, K-Means weighted was carried out with k = 4. The distribution of locations by clusters are:

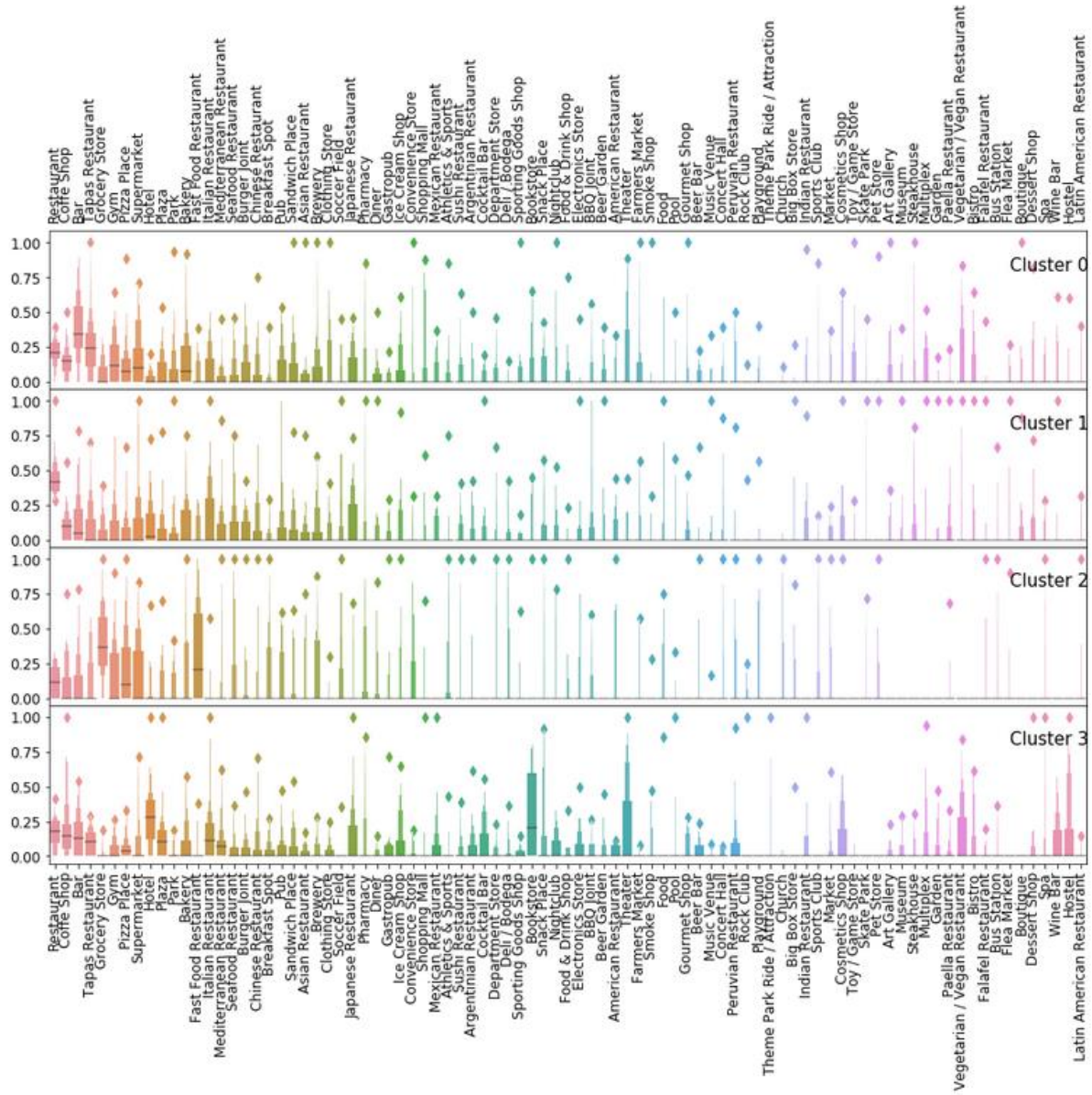
```
Locations by Cluster (202,)

1    66
0    58
2    40
3    38
dtype: int64
```

The locations are well distributed among the clusters. In the figure below, the visualisation of the clusters on a choropleth map is showed:



We can identify the main features of the clusters using a normalised boxplot representation and looking which are the most common venues for each cluster directly on the table:

1. Cluster 0 (purple): They are the locations mostly present in high density areas which are closed to the center. The most common categories in this group are all kind of restaurants, Supermarkets, Grocery Stores, Gyms, and Clothing Store. Supermarkets and Grocery Stores indicates that they may be residential areas.
2. Cluster 1 (cyan): They are the locations mostly present in high density areas covering the north east of the city. The most common categories in this group are normal restaurants, Italian restaurants, supermarkets, Hotels and Gyms. These areas are a combination of residential and business places.
3. Cluster 2 (green): They are located mainly in the suburbs. The most common categories in this group are Groceries, Fast food restaurants, Supermarket and miscellaneous shop. They are residential locations.
4. Cluster 3 (red): They are the locations with more numbers of venues, most of them are in the city center. The most common categories in this group are all kind of restaurants, Coffee shops, Museum, Hotels, night clubs and tapas bars. Typical venue configuration of a touristic and centric place.

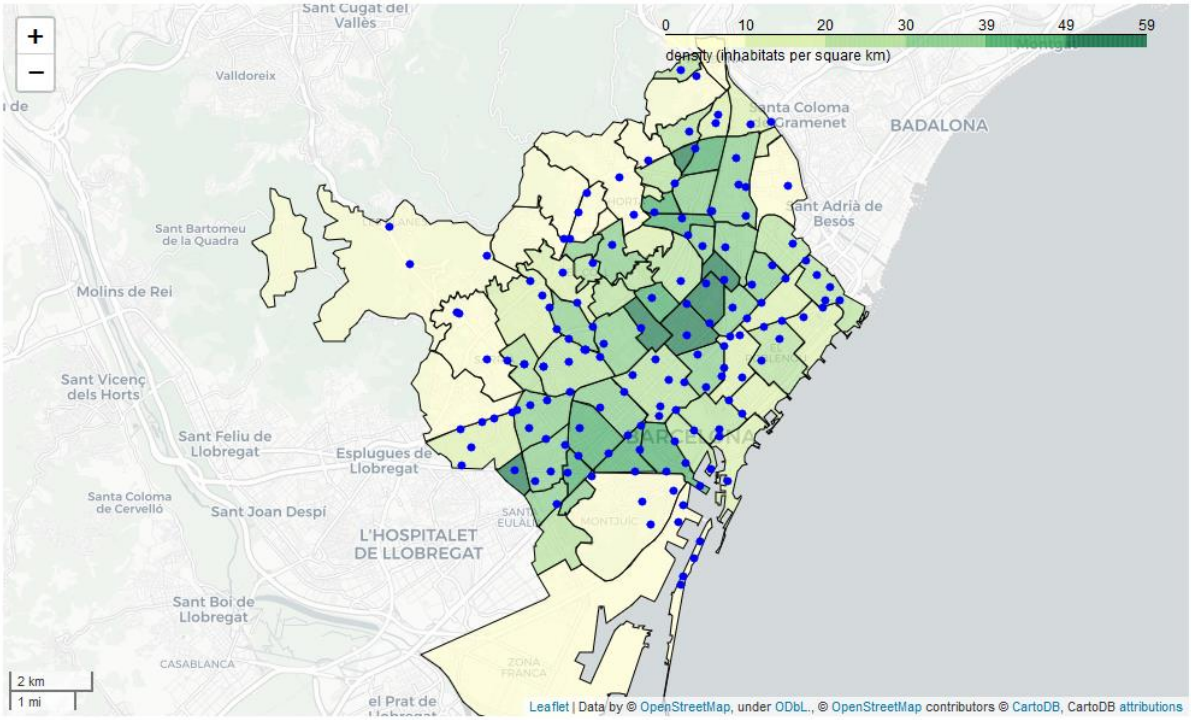


Boxplots diagrams can help to visualise fast the main changes among the clusters. We can see how the restaurants have a mean always higher than other categories.

3.2 Barcelona

We will do the clustering process for Barcelona as well. In the following map, the Barcelona map with the stations is represented. The metro and train stations are in blue. The choropleth shapes represent the neighborhoods' limits, and the color scale represents the population density. In Barcelona, we will analyse 144 locations in total.

In the map, we can see that most populated are corresponds to the city center of Barcelona (La Sagrada Familia Neighborhood).



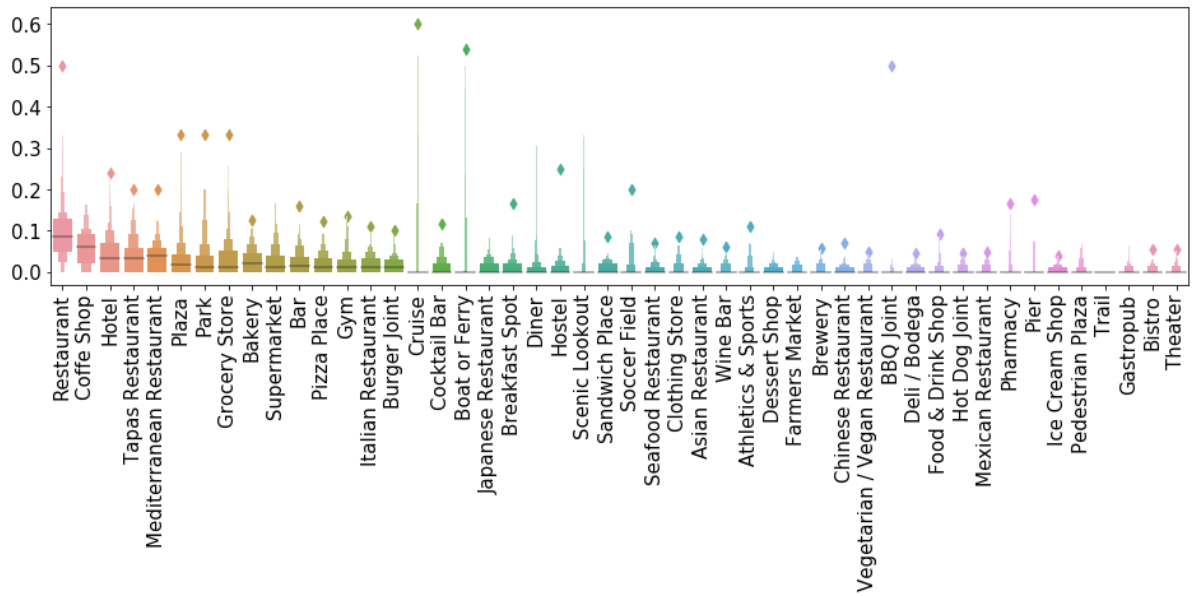
3.2.1 Foursquare data

After data cleaning following the same directives used for Madrid. In the case of Barcelona, we have 134 locations to explore with Foursquare API.

The resulted venue Dataframe consists of a total of 7360 venues. In the table, we see the venues dataset head including the location of the station.

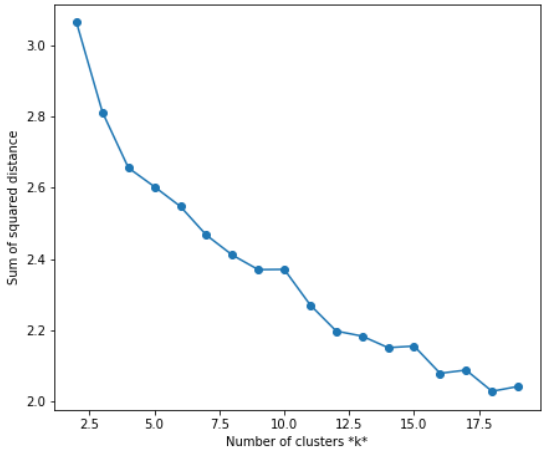
	stations	stations Latitude	stations Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	geometry
0	VALL D'HEBRON/METRO (L3, L5)	41.424923	2.142987	Camp Hebron-Teixonera	41.425532	2.146341	Soccer Field	POINT (2.142987 41.424923)
1	VALL D'HEBRON/METRO (L3, L5)	41.424923	2.142987	Bar Plaza	41.421942	2.141679	Spanish Restaurant	POINT (2.142987 41.424923)
2	VALL D'HEBRON/METRO (L3, L5)	41.424923	2.142987	Bar Smith	41.424204	2.147051	Bar	POINT (2.142987 41.424923)
3	VALL D'HEBRON/METRO (L3, L5)	41.424923	2.142987	Caprabo	41.424942	2.140979	Grocery Store	POINT (2.142987 41.424923)
4	VALL D'HEBRON/METRO (L3, L5)	41.424923	2.142987	Mercat de Vall d'Hebron	41.424229	2.142355	Farmers Market	POINT (2.142987 41.424923)

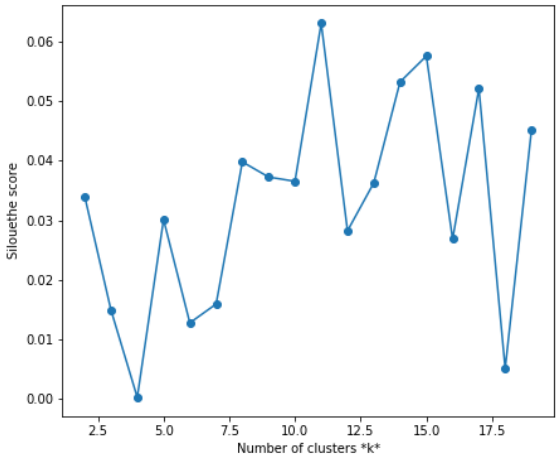
All kind of station category in the venues list are removed, as well as in Madrid; the final size is 7294 venues. There are 317 unique categories in the dataframe. We applied the same data preparation for the clustering and we will remove two stations: the harbor station and the station of the top of the funicular, because there are no venues in this area or the venues are too different in comparison with the other locations. In the figure we can see a boxplot with the most common categories. As in Madrid, the most common category is the Restaurant an Coffee Shop.



3.2.2 Clustering

We have chosen 10 clusters in this case for a k-means weighted because there is an abrupt change in the inertia curve for k= 10 and we see also a maximum for a silhouette score. We also tried with different cluster and we have obtained decent results with this number.



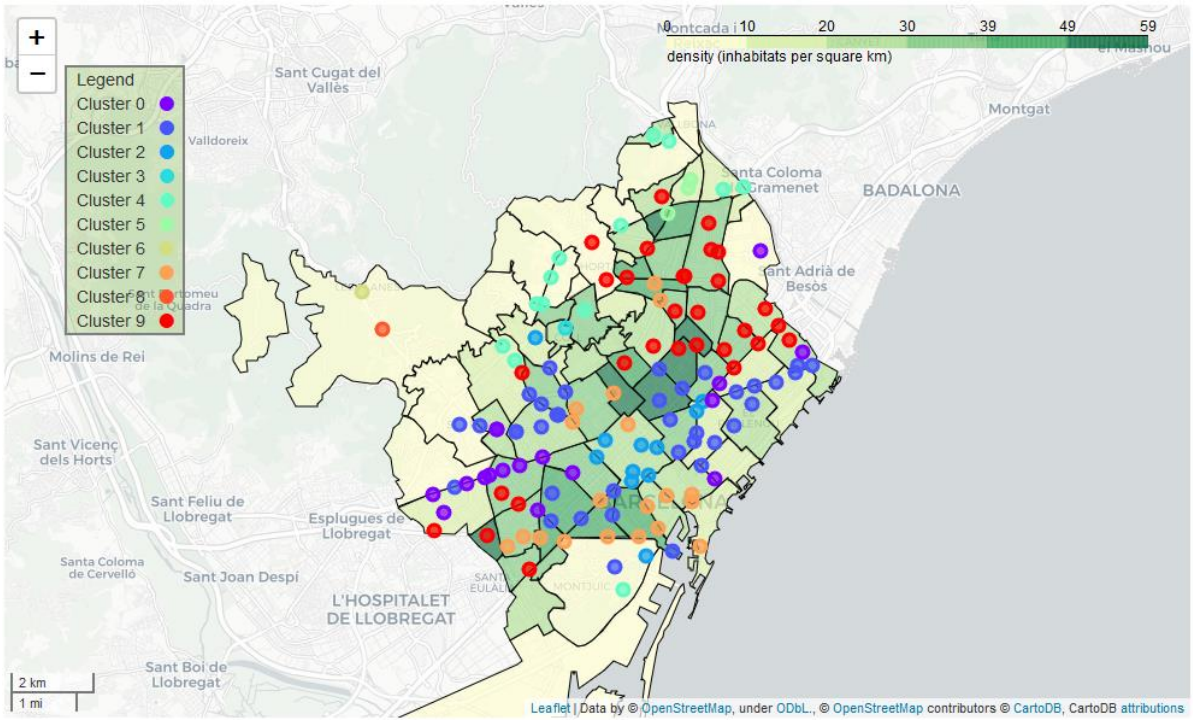


The distribution of the clusters is not as homogeneous; we have got three clusters with only one location. We are not going to take these clusters for evaluation; these locations corresponds to stations in the suburbs areas with few venues.

```
Location by Cluster

1    38
9    30
7    19
0    17
4    13
2    11
5     3
8     1
6     1
3     1
dtype: int64
```

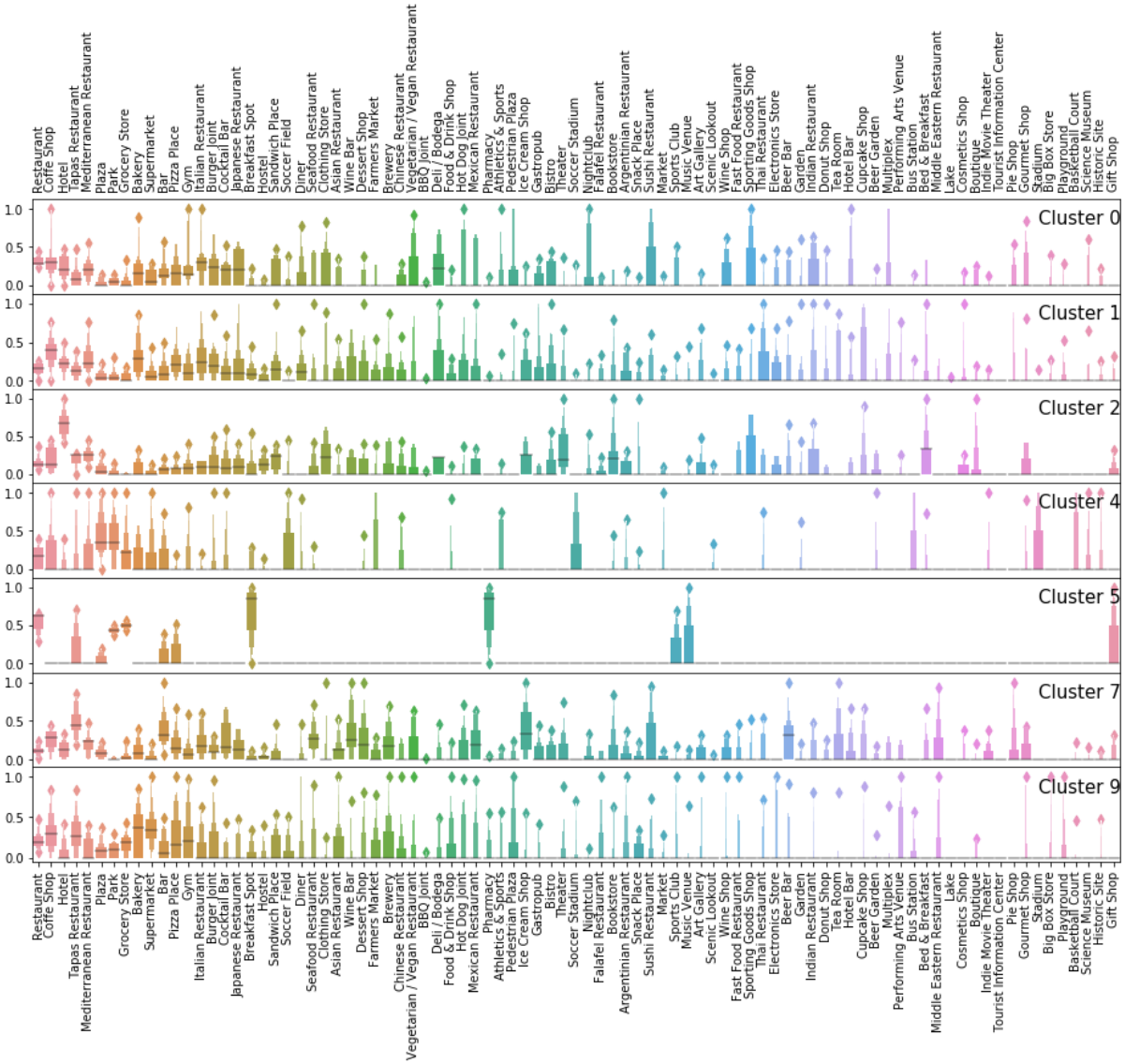
3.2.3 Results

As we can see in the map, the empty clusters are in areas with low density in the west border of the city and the *El Coll* station.



We can identify the main features of the clusters using a normalised boxplot representation and looking which are the most common venues for each cluster directly:

1. Cluster 0 (purple): They are the locations mostly present in the diagonal street on the west side. The most common categories in this group are Restaurants, Hotels, Gyms, Coffee Shops, and many restaurants. This location looks more similar to business places and **residential**.
2. Cluster 1 (navy blue): They are the locations mostly present in high-density areas in the historical center and some residential areas. The most common categories in this group are normal Restaurants, Coffee Shops, other restaurants, Gyms and Supermarkets. These areas are a combination of **residential** and touristic place.
3. Cluster 2 (blue): They are located mainly in the center. The most common categories in this group are Hotels, Restaurants, Coffee Shops. Touristic place in the city center.
4. Cluster 4 (cyan): They are the locations in residential zones in the north suburbs. The most common categories in this group are Parks, Plaza, Coffee shops, restaurants, supermarkets and gyms. A venue configuration of **residential** location.
5. Cluster 5 (green-cyan): Cluster with only three locations. They are similar to cluster 4, **residential** areas in the north of the city, but these locations have fewer venues and fewer Gyms and Coffee shops. (It could be a right place for the new opening)
6. Cluster 7 (orange): This kind of location is in the center and the south of the city. They are like cluster 1. The most common venues for these locations are Tapas Restaurants, Cocktail Bar, Coffee Shop, all kind of restaurants and Hotels.
7. Cluster 9 (red): Locations situated in the north of the city. The most popular venues are Restaurants, Coffee Shops, Supermarkets, Gyms, Parks and Grocery Store. These locations are also a **residential** area.



4 Result - Discussion

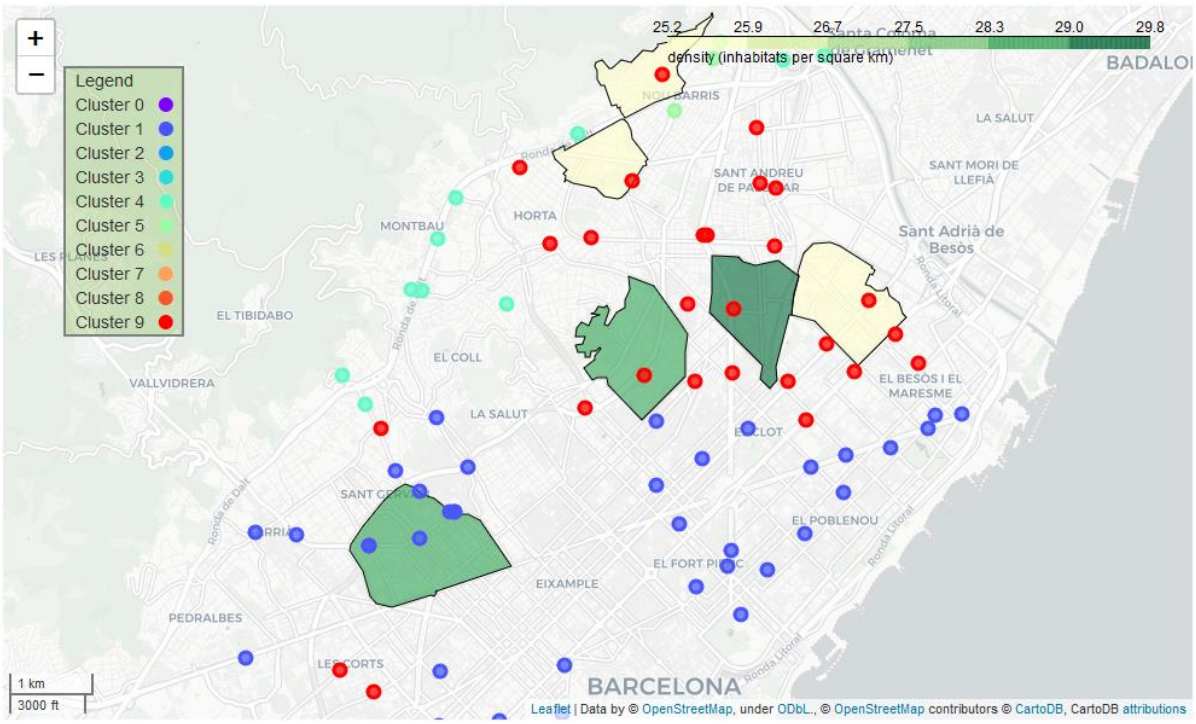
The purpose of the study is to find the best place, close to a public transport station, for a new opening based on the similarities between locations.

First, we can identify the type of location in which the firm has the actual gyms in Madrid. From the Madrid map, we can presume that gyms are placed in locations identified as clusters 0 (purple). The cluster 0 has been identified as a residential area having in one of the top categories **Supermarkets** and **Grocery stores**.

The target neighborhoods are the ones between **25 and 30 inhabitants per square kilometer** and close to the city center.

The clusters which represent the residential areas in Barcelona with similar features to gyms locations are **clusters 1, 4, 5 and 9**. They can be considered residential and also has Supermarket and groceries as an essential category.

We can see only the clusters and the neighborhoods which accomplish the requirements on a map:



The stations which best fit the conditions are the MUNTANER STATION, HOSPITAL DE SAN PAU, LA SAGRERA and LA PAU.

Analysis has been done using the Foursquare API. Both cities have a lot of Restaurants per location in comparison with other venues, I believe this may make the analysis harder, and it could be the reason why is so difficult to find the best cluster number. However, K-means weighted method has produced a decent clustering.

Further development on this study to increase the performance might be including more information to the clustering analysis, for example, neighborhoods information as density or housing prices.

5 Conclusion

As a result, in this Capstone study, I have tried to simulate a job for a gym firm who wants to open a new gym in a different city. They might want to find the locations with the same features that work for them before, and it is possible to do it using machine learning.

6 Repository

The repository of this project can be found in
https://github.com/IsmaelGSerrano/Coursera_Capstone

The jupyter notebooks used for this project are included in the repository:

- Madrid analysis:
  - Capstone_MadridNeigbours_notebook1: Data downloading and preparation
  - Capstone_MadridNeigbours_notebook2: Foursquare API data query
  - Capstone_MadridNeigbours_notebook3: clustering
- Barcelona analysis:
  - Capstone_BarcelonaNeigbours_notebook1: Data downloading and preparation
  - Capstone_BarcelonaNeigbours_notebook2: Foursquare API data query
  - Capstone_BarcelonaNeigbours_notebook3: clustering

All the data used in this project is also included in the repository.