Coursera Capstone: Analyzing neighborhoods of Madrid, Spain

Introduction

The city of Madrid is the capital of Spain and one of the largest cities in this country. Being a big city it offers more opportunities to start a new business like a restaurant. But it is a big issue to decide where to locate that new business. It makes sense to locate the restaurant in a neighborhood where there are more eating places because many people are used to going to that location to eat. But also if there is a big amount of restaurants the business will have a higher competence to rise.

With this idea in mind this project explored the neighborhoods in the city of Madrid to know if opening a restaurant is a good choice and where to do it. So anyone interested in starting a new business of this type could make a better choice based on data.

Methodology

- Data:

The data was extracted from Wikipedia. The web page contains a list of all the neighborhoods in Madrid separated by districts. This unstructured data was processed to obtain structured data (for this step BeautifulSoup was used). Only the columns “District name” and “Name” were saved in a data frame. From this data the latitude and longitude of each neighborhood were calculated using Geopy and later on information about nearby venues in each neighborhood was extracted using the Foursquare API. This information about venues was used to cluster the neighborhoods in Madrid and look for the best cluster to open a new restaurant.

The data from Wikipedia was downloaded making a request and retrieving the html file. The file was parsed using BeautifulSoup. The empty rows were skipped and there were two types of rows with data: rows with a district and a neighborhood and rows with only a neighborhood. To differentiate the district from the neighborhood in the first case a parenthesis was searched in the last position (the district always ended in a closing parenthesis). To extract the district the cell containing the name and the district number were split and the district name was selected. To differentiate the neighborhood a string not empty, different from a number and different from “[[ ]]” (the images in the table) was searched. To extract the neighborhood the spaces before and after the name were removed.

The next step was calculating the latitude and longitude of each neighborhood. As said before Geopy was used for this step. The locations were searched as “*District name*, *neighborhood name,* Madrid, España”. After this calculation a null value was generated and that row was deleted. The resulting data frame looked like the following one:

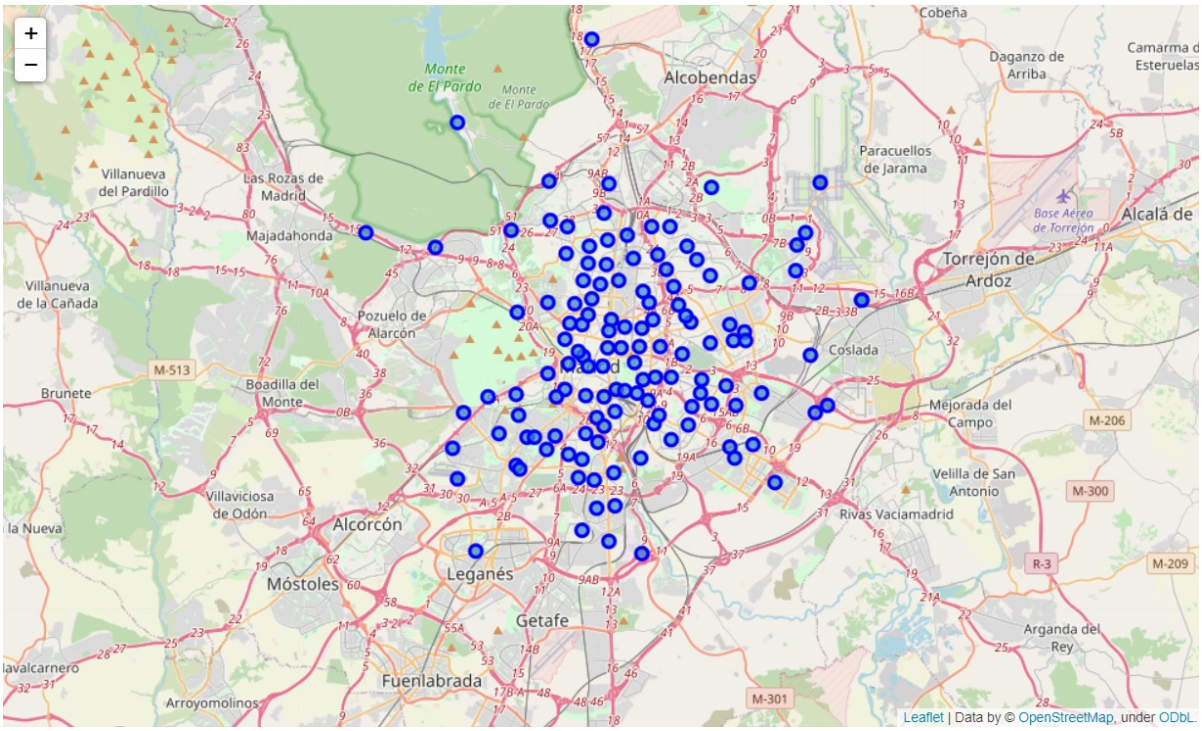
	District	Neighborhood	Latitude	Longitude
0	Centro	Palacio	40.417821	-3.715111
1	Centro	Embajadores	40.421058	-3.707185
2	Centro	Cortes	40.416389	-3.696463
3	Centro	Justicia	40.424246	-3.693508
4	Centro	Universidad	40.422753	-3.709849

The names, latitudes, longitudes and categories of nearby venues of each neighborhood were extracted making a request to Foursquare. The resulting data frame looked like the following one:

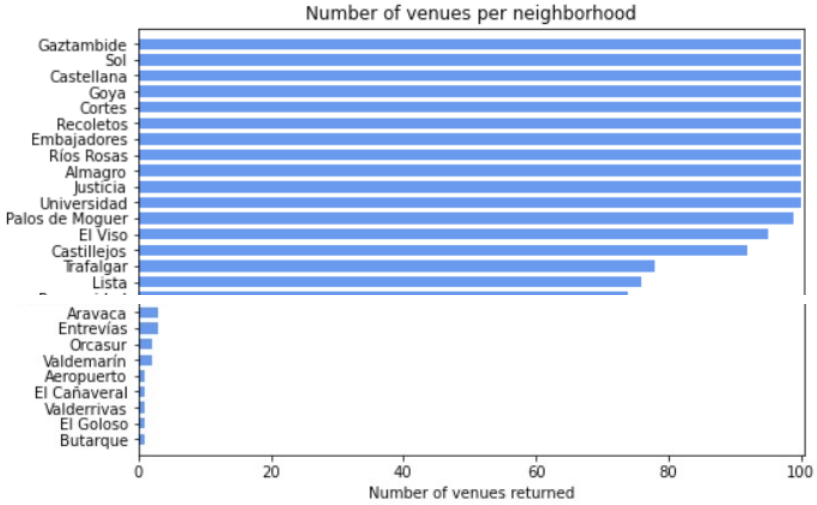
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	40.417821	-3.715111	Palacio Real de Madrid	40.417940	-3.714259	Palace
1	Palacio	40.417821	-3.715111	Plaza de la Almudena	40.416320	-3.713777	Plaza
2	Palacio	40.417821	-3.715111	Santa Iglesia Catedral de Santa María la Real ...	40.415767	-3.714516	Church
3	Palacio	40.417821	-3.715111	Plaza de Oriente	40.418326	-3.712196	Plaza
4	Palacio	40.417821	-3.715111	Jardines de Sabatini	40.419954	-3.713126	Garden

- Exploratory data analysis:

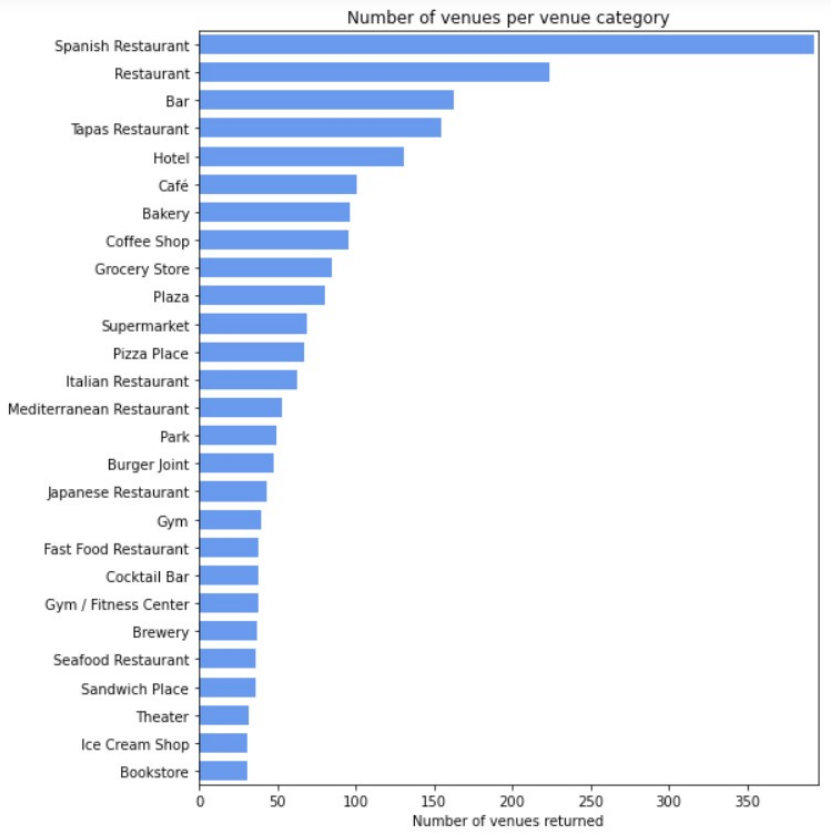
The distribution of the neighborhoods was showed on a map. The points were distributed normally from the city center of Madrid. This did not give much information to open a new restaurant.



In the graph “Number of venues per neighborhood” the total number of venues returned by neighborhood is showed. The limit specified was only reached in a few neighborhoods and some of them had a low number of returned nearby venues. In this graph the size and relevance of each neighborhood can be observed. For example, “El Goloso” is a neighborhood where we can find a military base and farmland and the population is low compared to other neighborhoods. On the other hand “Justicia” is a neighborhood located in the city center and its population is higher.



The total number of venues of each category (just for the most common ones) is displayed in the graph “Number of venues per venue category”. The most common venue category by difference was “Spanish restaurant” and some of the most common categories were also related to restaurants. The reason behind this could be that Spain is a country that has an economy oriented towards the service sector.

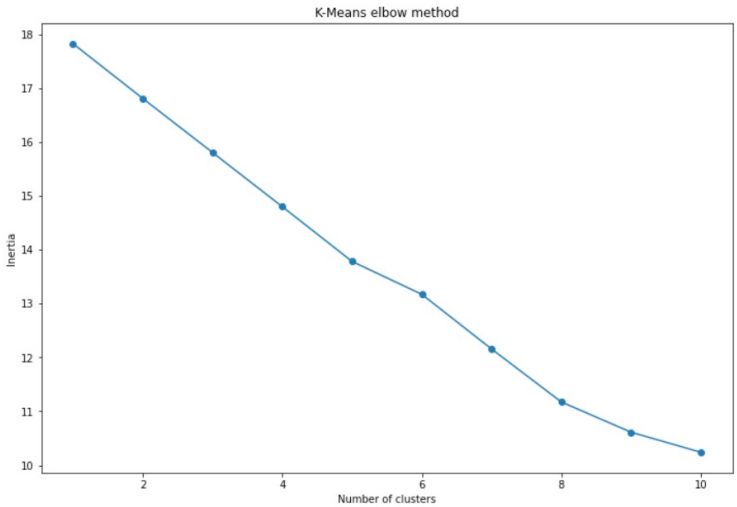


- Machine learning:

The machine learning technique used was clustering to group the locations in Madrid and search for the best one to open a restaurant. The clustering algorithm used was k-means.

The first step to use this algorithm was transform the venue data into a data frame where each column is a venue category and each row is a neighborhood containing the frequency of each category in each location.

Then the optimal number of clusters was searched with the elbow method. As seen in the graph “k-means elbow method” the optimal number of clusters seemed to be 8.

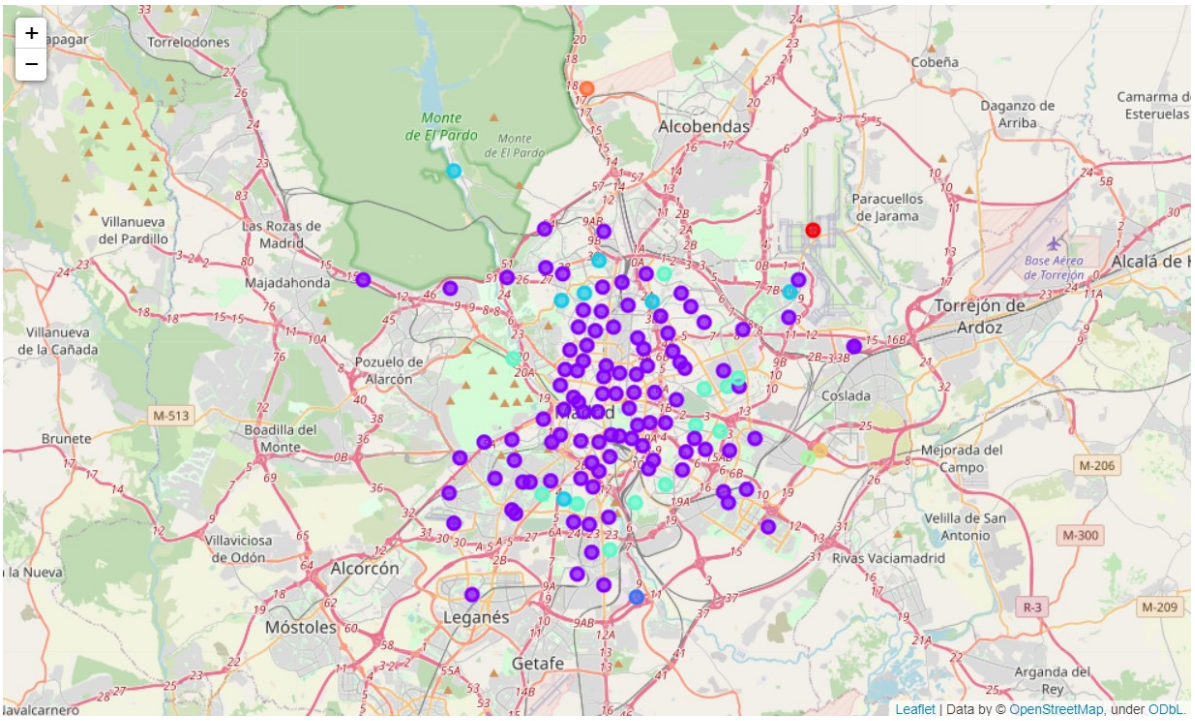


Afterwards the final model was trained with this number of clusters and the results were stored in a new data frame, containing the neighborhood name, the five most common venue categories (in parenthesis the frequency of each one was showed), the latitude and longitude and the label of each cluster. The resulting data frame looked like the following one:

	Neighborhood	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue	Latitude	Longitude	Label
0	Abrantes	Bakery (0.2)	Soccer Field (0.2)	Gym / Fitness Center (0.1)	Pizza Place (0.1)	Park (0.1)	40.380998	-3.727985	3
1	Acacias	Bar (0.09)	Tapas Restaurant (0.07)	Spanish Restaurant (0.07)	Pizza Place (0.07)	Art Gallery (0.05)	40.404075	-3.705957	0
2	Adelfas	Supermarket (0.06)	Grocery Store (0.06)	Tapas Restaurant (0.06)	Bar (0.06)	Fast Food Restaurant (0.06)	40.401903	-3.670958	0
3	Aeropuerto	Business Service (1.0)	Women's Store (0.0)	Escape Room (0.0)	Food Court (0.0)	Food & Drink Shop (0.0)	40.494838	-3.574081	7
4	Alameda de Osuna	Restaurant (0.08)	Tapas Restaurant (0.08)	Hotel (0.08)	Bakery (0.08)	Cocktail Bar (0.04)	40.457581	-3.587975	0

Results

The resulting clusters were displayed in a map by colors. The majority of the neighborhoods were inside one of the clusters (the purple one) and some of the clusters were only formed by one neighborhood.



Also the five most common venue categories of each neighborhood were displayed by clusters to inspect the clusters. Cluster 0 (in purple) is the one with more instances, integrated by a wide variety of venue categories. Cluster 1 (in dark blue), 4 (in green), 5 (in yellow) and 6 (in orange) could be seen as noise or an error since they are clusters formed just by one venue category not really relevant. Cluster 2 (in blue) is formed mainly by restaurants. Cluster 3 (in cyan) seems to be a “sport-like” neighborhood cluster since it has gyms and parks. Cluster 7 (in red) is the airport. The results obtained for each one were:

Cluster 0

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
1	Bar (0.09)	Tapas Restaurant (0.07)	Spanish Restaurant (0.07)	Pizza Place (0.07)	Art Gallery (0.05)
2	Supermarket (0.06)	Grocery Store (0.06)	Tapas Restaurant (0.06)	Bar (0.06)	Fast Food Restaurant (0.06)
4	Restaurant (0.08)	Tapas Restaurant (0.08)	Hotel (0.08)	Bakery (0.08)	Cocktail Bar (0.04)
5	Spanish Restaurant (0.14)	Restaurant (0.13)	Mediterranean Restaurant (0.05)	Bar (0.04)	Japanese Restaurant (0.04)
6	Chinese Restaurant (0.17)	Hotel (0.06)	Library (0.06)	Coffee Shop (0.06)	Restaurant (0.06)
...	...	...	...	...	...
121	Coffee Shop (0.18)	Breakfast Spot (0.12)	Soccer Field (0.06)	Spanish Restaurant (0.06)	Snack Place (0.06)
122	Pizza Place (0.13)	Soccer Field (0.13)	Grocery Store (0.13)	Garden (0.07)	Bar (0.07)
123	Restaurant (0.25)	Pizza Place (0.25)	Train (0.25)	Train Station (0.25)	Farm (0.0)
124	Playground (0.09)	Restaurant (0.09)	Flower Shop (0.09)	Plaza (0.09)	Food & Drink Shop (0.09)
125	Fast Food Restaurant (0.17)	Pizza Place (0.11)	Bar (0.06)	Comedy Club (0.06)	Pharmacy (0.06)

Cluster 1

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
20	Grocery Store (1.0)	Farmers Market (0.0)	Escape Room (0.0)	Event Space (0.0)	Exhibit (0.0)

Cluster 2

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
15	Spanish Restaurant (0.36)	Restaurant (0.09)	Bus Station (0.09)	Paella Restaurant (0.09)	Chinese Restaurant (0.09)
35	Spanish Restaurant (0.31)	Restaurant (0.15)	Fast Food Restaurant (0.08)	Hotel (0.08)	Bar (0.08)
43	Spanish Restaurant (0.38)	Restaurant (0.23)	Plaza (0.15)	Tapas Restaurant (0.08)	Government Building (0.08)
63	Spanish Restaurant (0.25)	Restaurant (0.12)	Food & Drink Shop (0.12)	Park (0.12)	Athletics & Sports (0.12)
115	Spanish Restaurant (0.5)	Breakfast Spot (0.17)	Tapas Restaurant (0.17)	Art Studio (0.17)	Fast Food Restaurant (0.0)
119	Spanish Restaurant (0.33)	Soccer Field (0.33)	Supermarket (0.17)	Asian Restaurant (0.17)	Farmers Market (0.0)
126	Spanish Restaurant (0.5)	Beer Garden (0.17)	Park (0.17)	Athletics & Sports (0.17)	Women's Store (0.0)

Cluster 3

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
0	Bakery (0.2)	Soccer Field (0.2)	Gym / Fitness Center (0.1)	Pizza Place (0.1)	Park (0.1)
9	Park (0.33)	Bakery (0.17)	Bar (0.17)	Other Repair Shop (0.17)	Café (0.17)
10	Restaurant (0.2)	Park (0.2)	Soccer Field (0.2)	Grocery Store (0.2)	Metro Station (0.2)
12	Castle (0.33)	Bus Stop (0.33)	Park (0.33)	Event Space (0.0)	Exhibit (0.0)
49	Park (0.33)	Pizza Place (0.33)	Gym / Fitness Center (0.33)	Women's Store (0.0)	Farm (0.0)
57	Park (0.25)	Music Venue (0.25)	Gym (0.25)	Pharmacy (0.25)	Women's Store (0.0)
59	Spanish Restaurant (0.25)	Pizza Place (0.25)	Soccer Stadium (0.25)	Park (0.25)	Farm (0.0)
69	Pizza Place (0.25)	Restaurant (0.12)	Track Stadium (0.12)	Supermarket (0.12)	Chinese Restaurant (0.12)
72	Park (0.6)	Plaza (0.2)	Brewery (0.2)	Women's Store (0.0)	Farmers Market (0.0)
85	Toy / Game Store (0.2)	Food & Drink Shop (0.2)	Bar (0.2)	Park (0.2)	Optical Shop (0.2)
94	Beer Garden (0.25)	Theater (0.25)	Park (0.25)	Athletics & Sports (0.25)	Women's Store (0.0)
96	Park (0.33)	Metro Station (0.33)	Bar (0.17)	Supermarket (0.17)	Women's Store (0.0)

Cluster 4

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
118	Mediterranean Restaurant (1.0)	Women's Store (0.0)	Farmers Market (0.0)	Event Space (0.0)	Exhibit (0.0)

```
##### Cluster 5 #####
```

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
41	Toll Booth (1.0)	Farmers Market (0.0)	Escape Room (0.0)	Event Space (0.0)	Exhibit (0.0)

```
##### Cluster 6 #####
```

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
42	Shoe Store (1.0)	Women's Store (0.0)	Embassy / Consulate (0.0)	Food & Drink Shop (0.0)	Food (0.0)

```
##### Cluster 7 #####
```

	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
3	Business Service (1.0)	Women's Store (0.0)	Escape Room (0.0)	Food Court (0.0)	Food & Drink Shop (0.0)

Discussion

Regarding the results presented in the previous section and remembering the idea presented in the introduction I believe the best location to open a new restaurant is the cluster number 2 (the names of the neighborhoods in this cluster are listed in the final cell of the notebook). This cluster is formed mainly by restaurants so there is a high chance that opening a restaurant there people will try it. The main disadvantage is that having a high amount of restaurants there will be a higher competence and maybe people refuse trying a new restaurant and only eat where they are used to.

About the results obtained it can be mentioned that the clusterization seemed to work well. As said before there are some clusters that appear to be noise but the other clusters types could be named as: general, restaurants, sport and airport.

Conclusion

This project covered the analysis of neighborhoods in the city of Madrid with the idea of opening a new restaurant. To do it data was extracted from Wikipedia as well as other resources like Foursquare. With this data a clusterization using k-means algorithm was performed and the resulting clusters showed a good segmentation of the neighborhoods. To open a new restaurant I suggest choosing neighborhoods found in the cluster number 2.