MADRID LIFESTYLE

José Javier Rueda

May 2020

Introduction

Background

I am a 24 year old engineer who has recently completed its master's degree. I live in Madrid with my family and since I have studied in Madrid, I have not had the opportunity to live myself. Currently, I am willing to let go and find an apartment to live alone or with friends.

Problem

Madrid is composed of 21 districts which have several neighborhoods each one. I am trying to study the neighborhoods according to their location and to the businesses in them. My actual neighborhood is very fun and it has a lot of restaurants and shops and I would like to find a neighborhood to move in similar to this one.

I am going to analyze the different neighborhoods of the city I live in, Madrid. I will cluster the neighborhoods according to the characteristics of their venues as restaurants, social places, parks or residential areas.

Interest

This project has a personal interest to me but it also can be interesting to real state companies which can study the characteristics of the neighborhoods and offer personalized offers to their clients.

Data Acquisition

Sources

The data will be obtained from several sources. First of all, data about Madrid neighborhoods will be obtained by scraping the Wikipedia webpage in which this information is shown. Then, the coordinates of each neighborhood will be uploaded manually from a website that provides the coordinates of any Google Maps site (https://geocode.localfocus.nl/). Finally, with the Foursquare API, the neighborhoods will be explored so they can be clustered according to the venues sited in each of the districts.

Cleaning

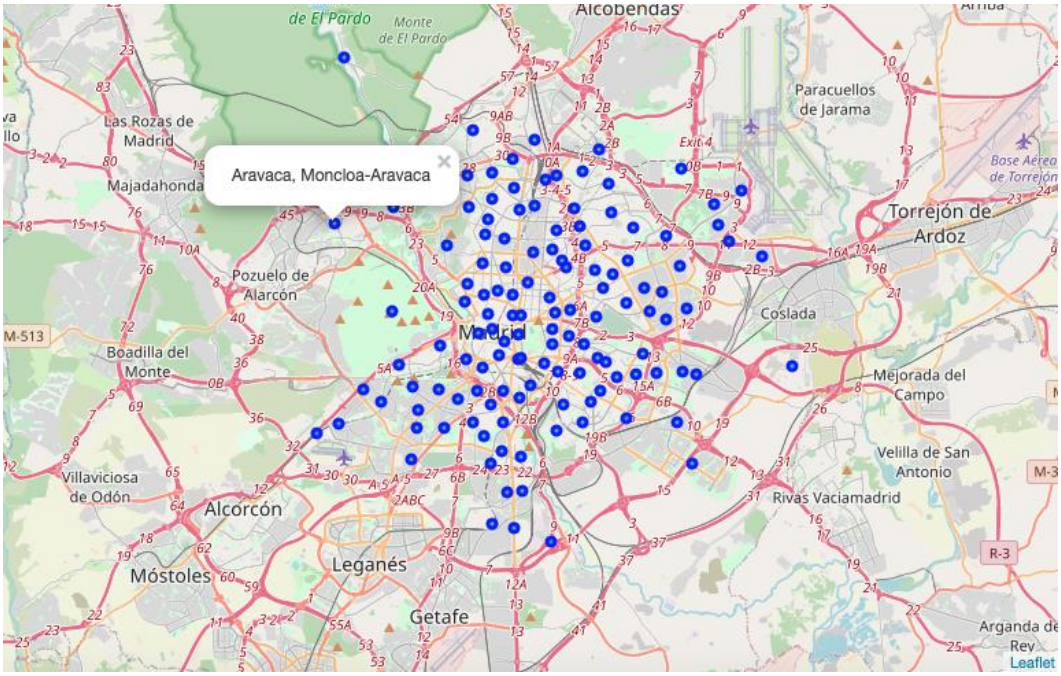
The task of cleaning the data consists in merging the two tables by the neighborhood name. The first table includes the district name, the neighborhood name and the neighborhood code. The second table has the neighborhood name alongside the latitude and longitude coordinates. The final data frame will show the neighborhood name and code, the district name and the coordinates.

Since the decimal separator in Spain is the comma, it is very important to change all the commas to dots in the data frame. Also, the data type is changed from object to integer for the neighborhood number code and from object to float for the coordinates.

The size of the data frame is checked with a simple command and the result retrieved is right: Madrid has 21 districts and 131 neighborhoods.

The coordinates of Madrid city are obtained from a quick search on the internet: (40.416775, -3.70379).

Madrid neighborhoods are depicted in the map shown below.



Exploratory Data Analysis

Exploring the data shapes

I have decided to limit the amount of venues per neighborhood to 100. I believe this quantity if more than enough to describe an area. Besides, the great majority do not reach this limit and there are 18 neighborhoods who do not even retrieve ten venues. For further development and adjustment, these neighborhoods could be dropped from the data frame.

With a radius of 700 meters, the final quantity of venues retrieved from the Foursquare database has been 5535. Each venue is characterized with a category, that might be park, Italian restaurant, gym or hotel for example. Within the database of venues, a total of 296 unique categories can be found, which represents a very rich selection.

Obtaining the most common venues

The data is prepared for clustering by obtaining the frequency of each category venue per neighborhood. This is done by encoding the data frame and then grouping by neighborhood. The data is ordered from the most common to the least category in each neighborhood.

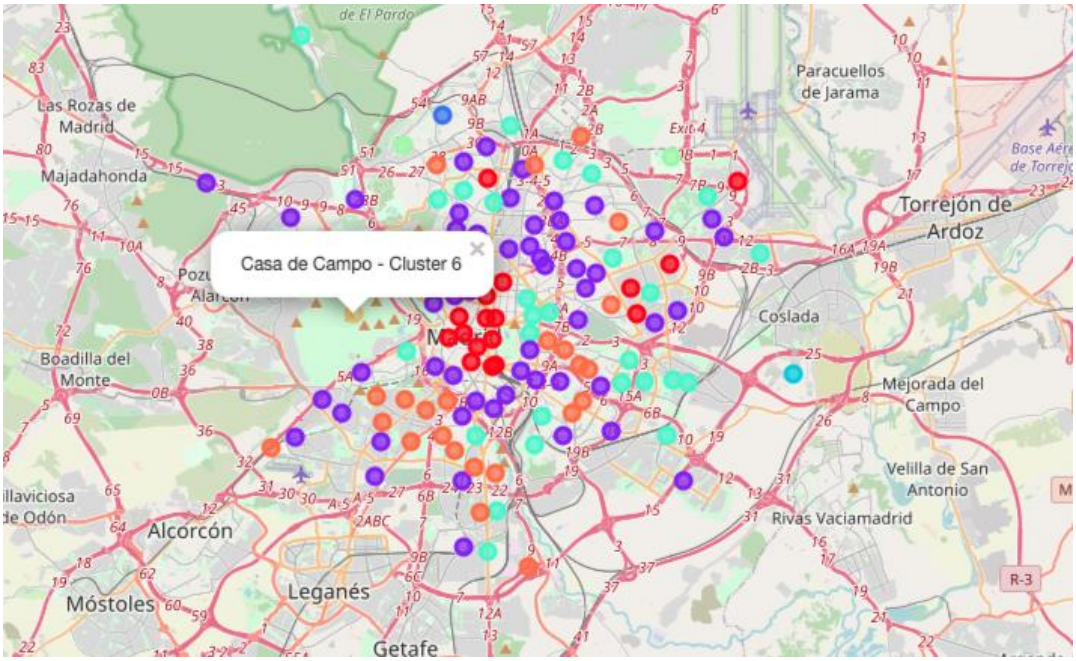
It is decided that a neighborhood is well defined by its 5 most common venues. A smaller value could lead to miss information while a greater value could introduce noise in the model.

Clustering

Model

The model developed follows the k-means methodology. The number of clusters has been chosen by trial and error. After studying the model it has been concluded that there are a few areas in Madrid that are very characteristics and cannot be clustered, such as Casa de Campo or Mirasierra. These neighborhoods can be considered outliers. The optimal k for the model is 8, which leads to five main clusters and three outliers. A greater value would increase the complexity of the model and a smaller value would reduce the accuracy of the diverse areas that can be found in Madrid.

The following map shows the clusters.



Results

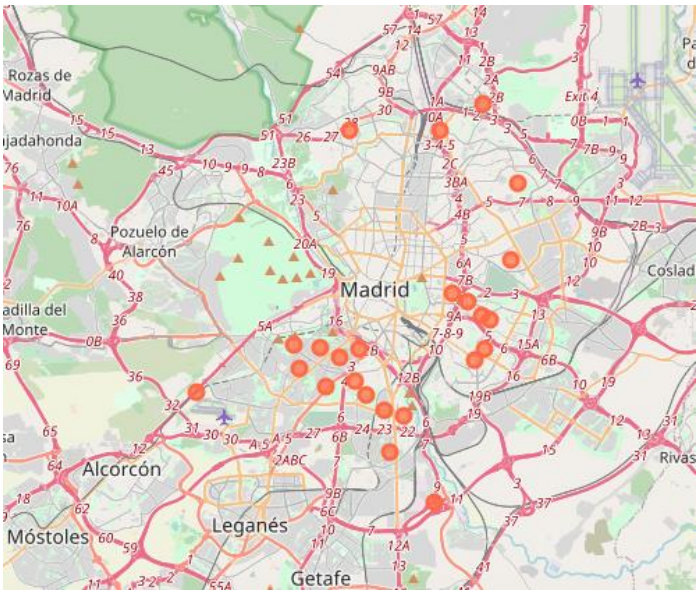
The following table shows the clusters labels and their relationship with the most common venues found in each of them.

Label	**Category**
Cluster 0	Tourists (hotels and coffee shops)
Cluster 1	Food (restaurants and bars)
Cluster 2	Mirasierra (outlier)
Cluster 3	El Cañaveral (outlier)
Cluster 4	Spanish restaurants
Cluster 5	Sports facilities
Cluster 6	Casa de Campo (main park)
Cluster 7	Residential (plazas and green areas)

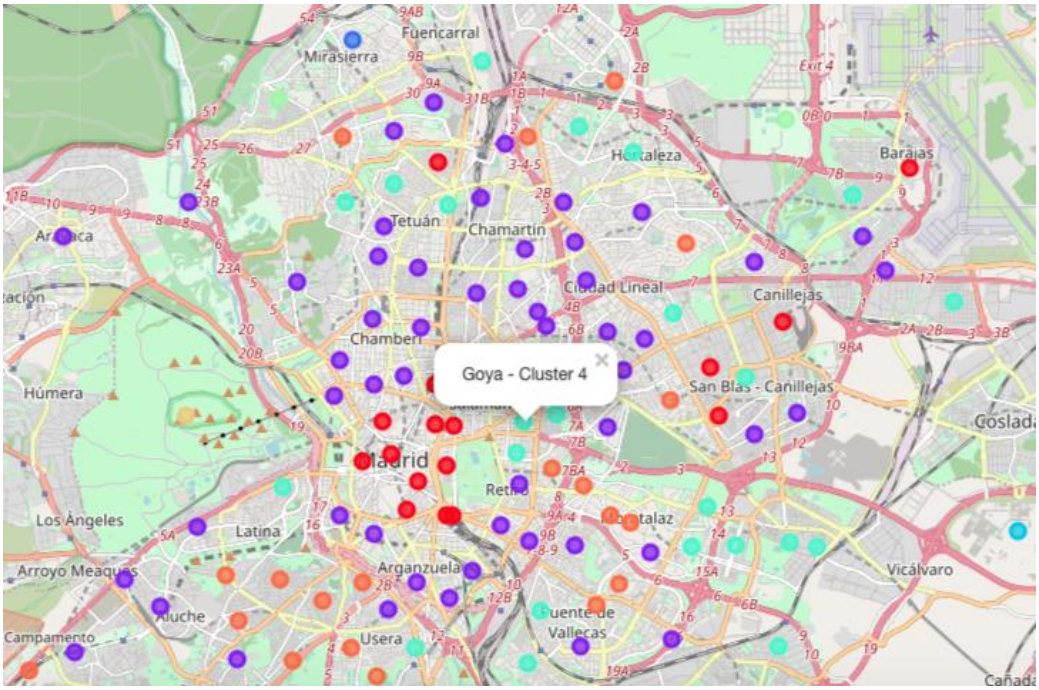
Conclusions

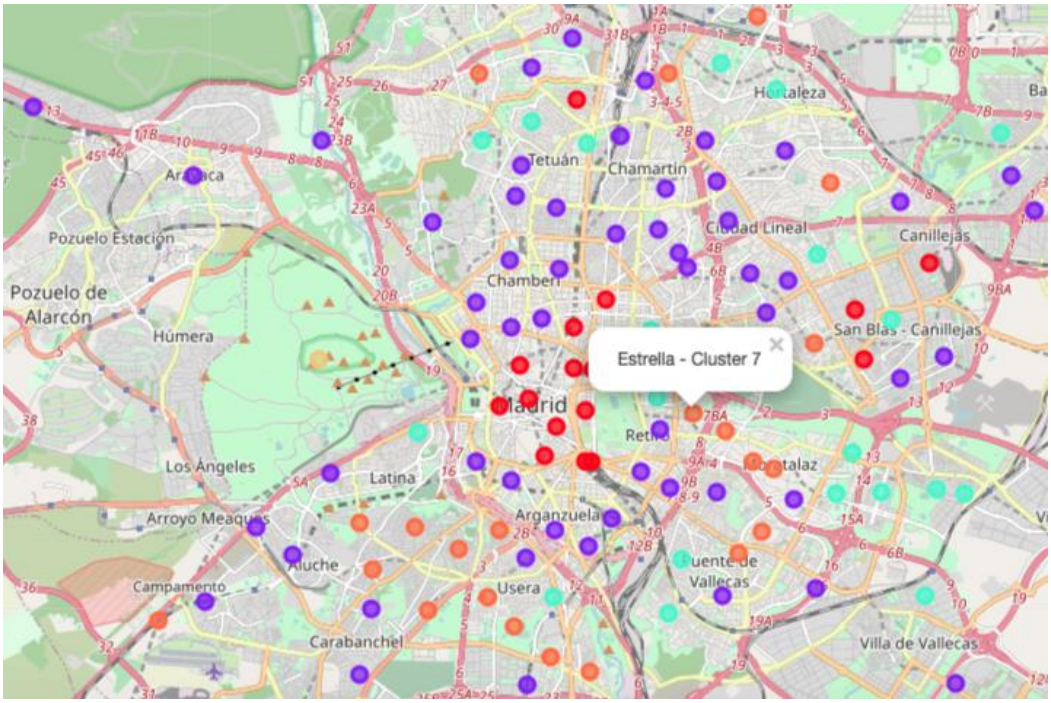
As I have said before, I currently live in Goya neighborhood, which according to the analysis, belongs to Cluster 4 – Spanish restaurants. I am happy with this lifestyle but I would be willing to live in a more quiet area, such as Cluster 7, with parks and plazas to go on relaxing walks.

The following map shows the neighborhoods belonging to Cluster 7.



Since I am also interested in living close to my family to be able to visit them regularly, I would finally choose Estrella neighborhood.





Further directions

As I have mentioned earlier in this report, the further work for this project would be to drop the neighborhoods with a small number of venues retrieved so they do not generate noise in the clustering model.

Also, outliers should be detected earlier to avoid complications while modelling the problem and designing a solution.

These issues have not been tackled in this project since the main goal was to study the raw data provided by Foursquare about Madrid city.