**Neighborhood comparison between Madrid and Barcelona**

14th June 2020

Author: Claudio Pravia Silvera

# Introduction

The business problem we want to address is to provide franchise operators with information that makes it easier to identify the probability of success when opening a franchise in either Madrid or Barcelona.

To do this we will compare the two cities to discover the neighborhoods for similarities related to the type and number of venues that they have. We will cluster the neighborhoods using k-means and will then examine the clusters to find similarities between the two cities. We will analyse what makes them more attractive to tourists.

The target audience is therefore the franchise operators that focus on tourism in these two cities and needs to know where to open the franchise.

# Data acquisition and cleaning

First, we will extract the list of neighborhoods of the two cities. After that, we will use Nominatim to obtain the geographical coordinates for the neighborhoods by using the neighborhoods’ name, and finally using Foursquare to obtain the different venues that exist in each of the neighborhoods.

## Data sources

The list of neighborhoods’ data was obtained using the following Wikipedia links:

<https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid>

The neighborhood names are the column "Nombre".

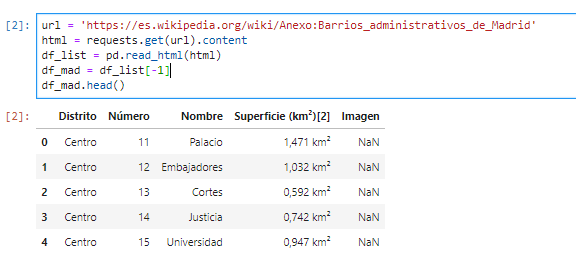
<https://es.wikipedia.org/wiki/Distritos_de_Barcelona>

The neighborhood names are the column "Barrios" that are a set that will be separated.

We will use a ‘get’ request using the Foursquare API, to obtain the following venue data: name, latitude, longitude and category.

### Madrid.

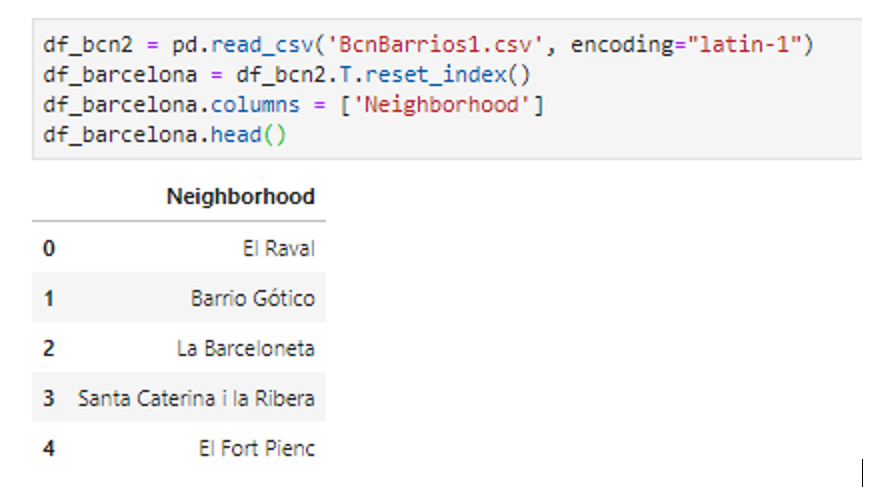
The first step was to get the Names of the neighborhoods.



### Barcelona.

In this case the Names of the neighborhoods came as a set in one field by district and though they were labeled with a number, some of them were separated by commas and other by “and”: sometimes the “and” was in Spanish (“y”) and other times in Catalan (“i”). The decision was made to save them to a csv file and do the cleansing manually.



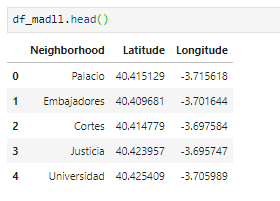


## Data cleansing

When getting the coordinates, an exception was programmed to discard the neighborhoods without data.

**2.2.1 Madrid.**

We got the coordinates using Nominatim:

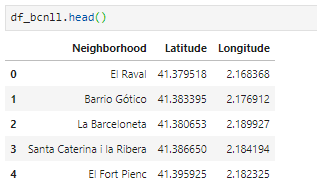


The list below shows the neighborhoods names for which we could not get the coordinates.



**2.2.2 Barcelona.**

This is the list with coordinates for Barcelona.



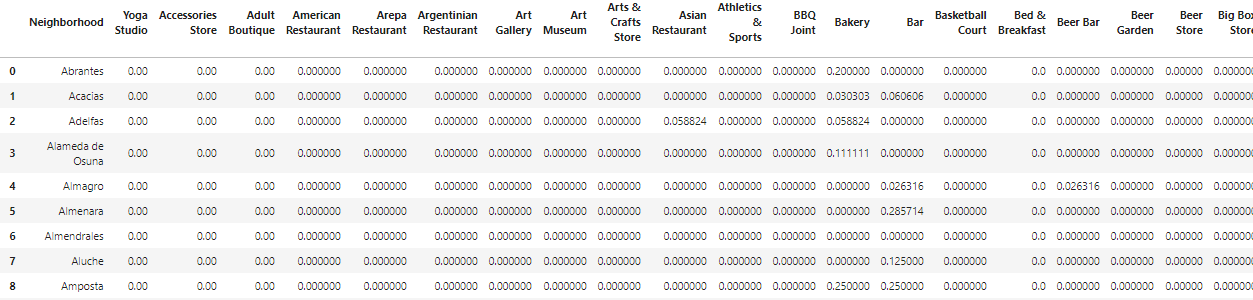
The list below shows the neighborhoods names for which we could not get the coordinates.



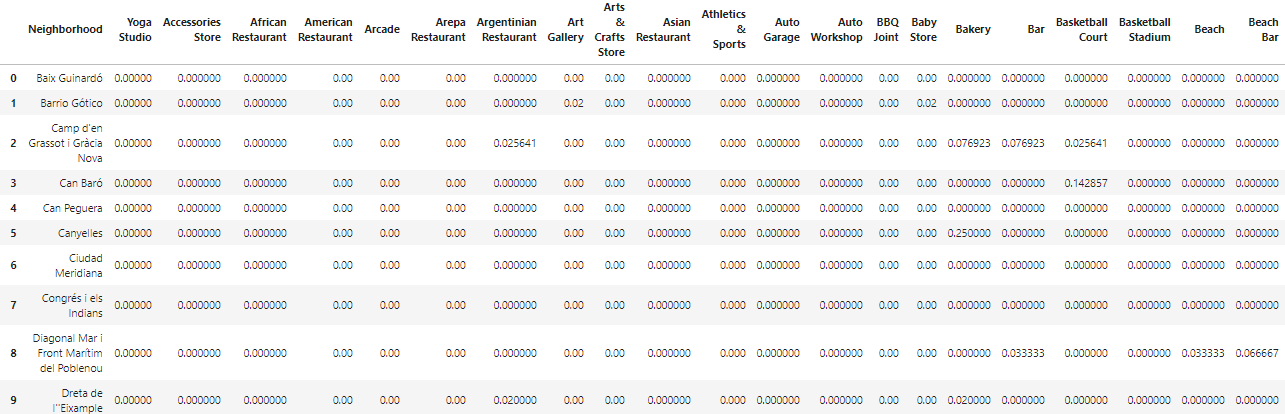
## Feature selection

Using Foursquare data, we identified the top 50 venues for each neighborhood that are within a 200-meter radius from the center of the neighborhood. We encoded and normalized the venues per neighborhood.

### Madrid.

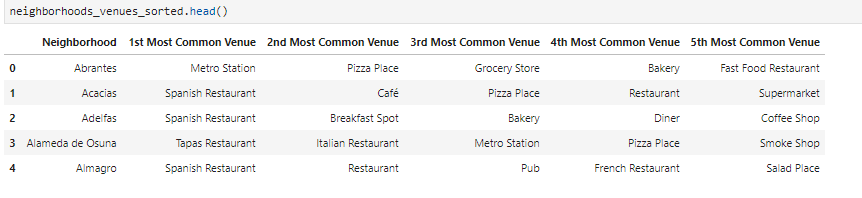


### Barcelona.



We got the 5 most common venues for each neighborhood.

### Madrid venues.



### Barcelona venues.



# Methodology

We wanted to classify the different neighborhoods by the amount of common venues that they have.

## Machine learning used

We used K-Means clustering to cluster the neighborhoods in different groups to find patterns that can help us get the findings that will help to solve our business problem.

## Exploratory Data Analysis

We used different values of k and found k=6 to be the optimal. Less than that will confound the clusters and more will include clusters that are not relevant.

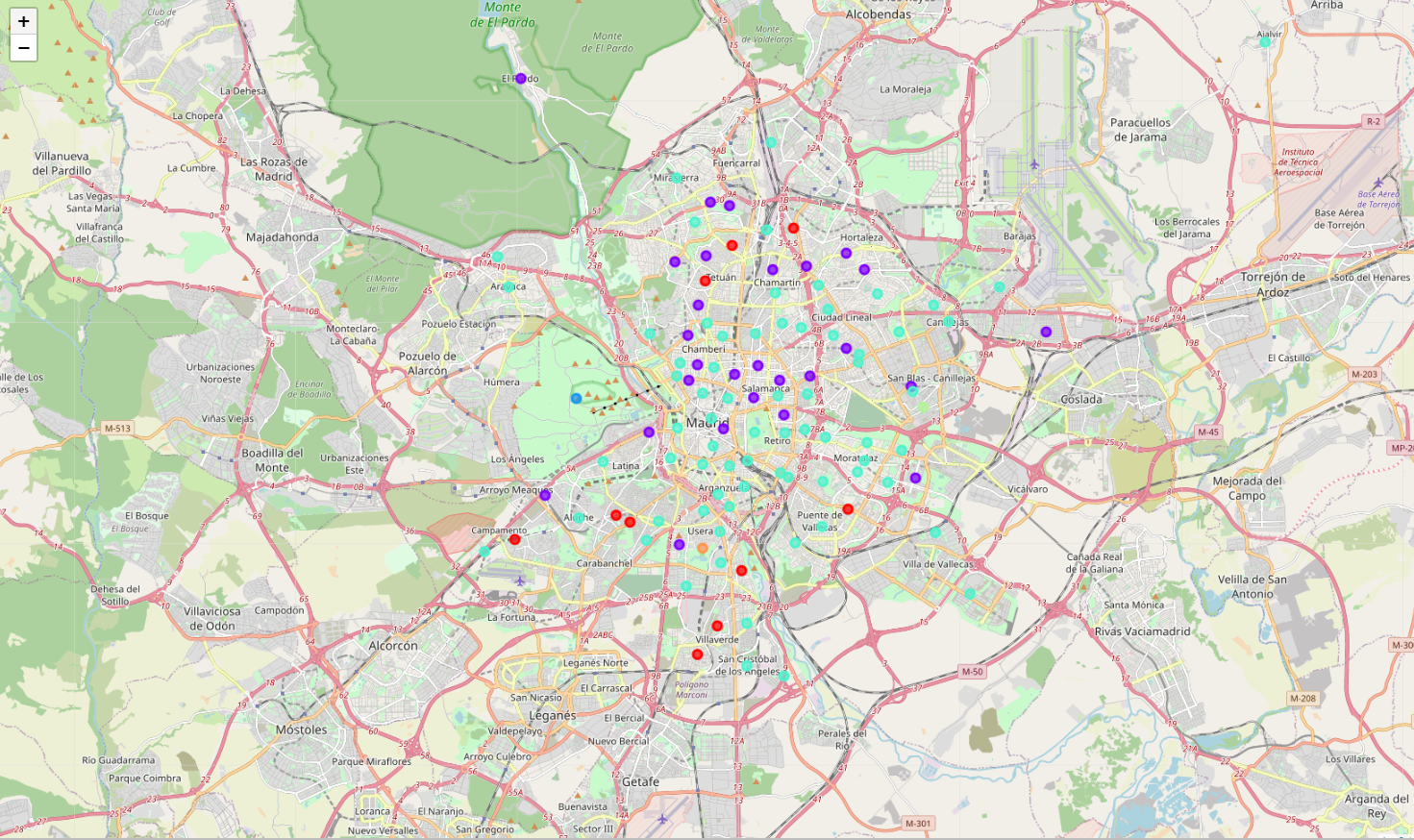
We also changed the radius of the circles we used to identify venues in each neighborhood. We started with 500 meters, and then reduced to 300 meters and finally to 200 meters. We realized that with the larger radius some venues will appear to be part of more than one neighborhood and that was to be avoided.

# Results

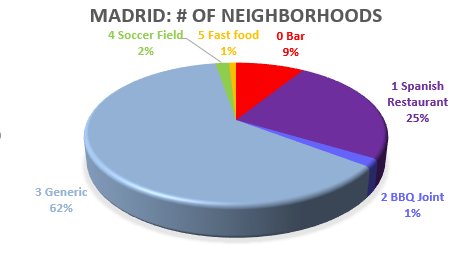
## 

## Clustering neighborhoods in Madrid.

The following figure shows a map of Madrid with the neighborhoods shown in different colors for each cluster.

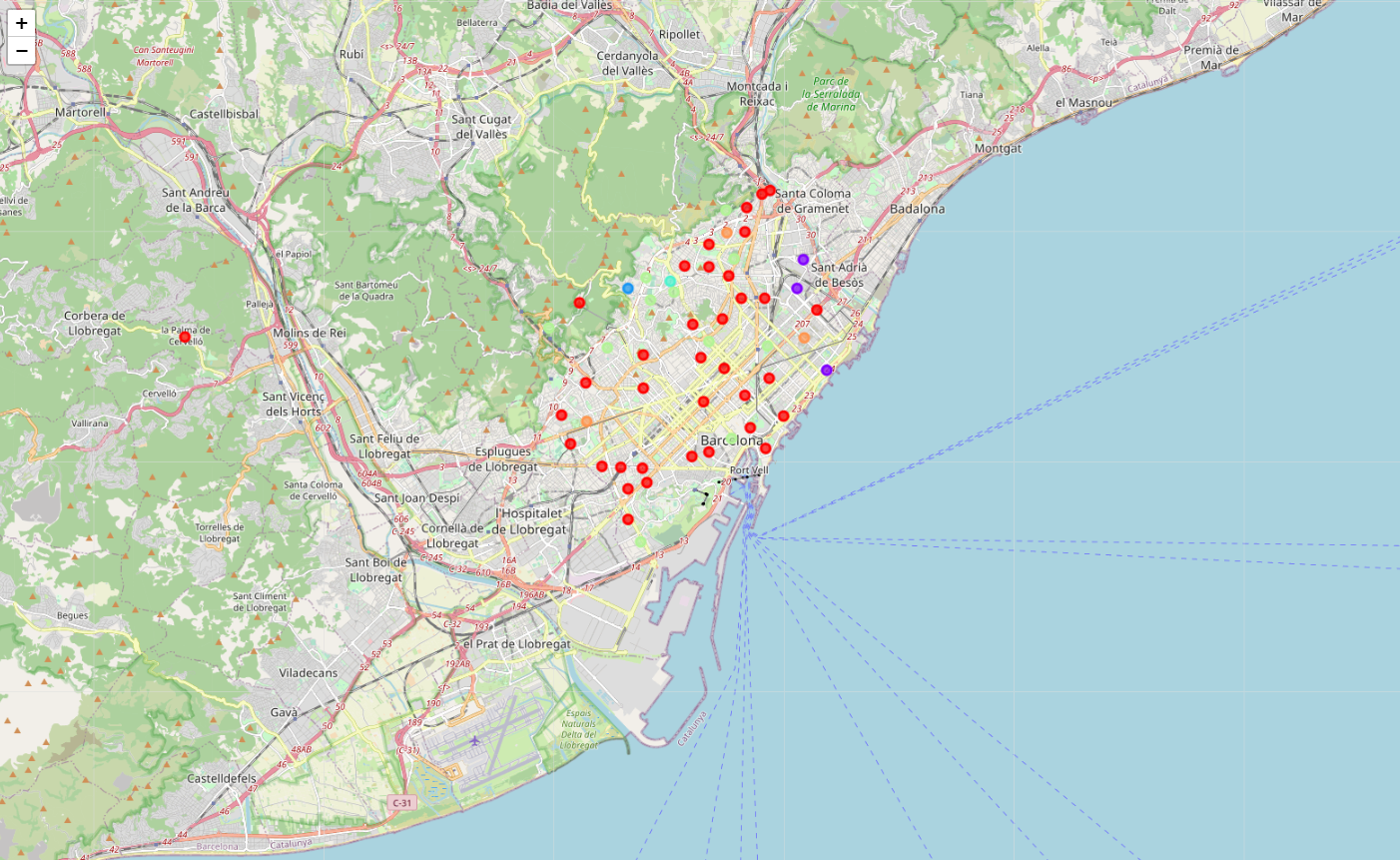


This one shows a chart with the percentage of the number of neighborhoods that fall into each cluster in Madrid.

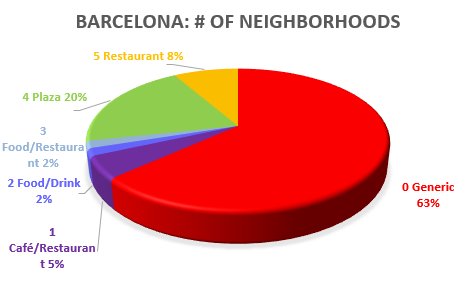


## Clustering neighborhoods in Barcelona

This figure shows a map of Barcelona with the neighborhoods shown in different colors for each cluster.



And this one shows a chart with the percentage of the number of neighborhoods that fall into each cluster for Barcelona.



# Discussion

We see in both cases a predominant cluster of more than 60% that contains venues that we labeled as Generic. The venues in these clusters have many different categories that cannot be summarized.

In Madrid, the second largest cluster has a clear bias for the category “Spanish Restaurant” while the third has the category “Bar”.

In the case of Barcelona, the second largest cluster has the category “Plaza” (meaning “park”) as predominant while the third has the category “Restaurant”.

# Conclusion

We can recommend, for a franchise in Spanish food that would like to open in Madrid, choose one of the neighborhoods in the Madrid cluster 1.

In Barcelona, that is in Catalonia, a province that wants to be independent, there is no cluster related to Spanish Restaurant. The Barcelona cluster 5 though, shows neighborhoods with the main category as Restaurants. We can therefore recommend a more general food franchise to be opened in these neighborhoods.

# Future directions

We could deepen this study by concentrating only in restaurants and then try to find more specific relationships between the type of cuisine and the neighborhoods in Madrid and Barcelona.