Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

In this project we will try to find the relationship of using different type of data for clustering the neighborhoods of Barcelona, Spain. We will try to use socioeconomic data of every neighborhood and we will also use data of the most common venues in every neighborhood. It will be interesting to analyze if the clusters are similar because Barcelona is a multicultural city and also has a lot of tourism. This fact makes that usually the neighborhoods more visited for tourists are very different from other neighborhoods less visited.

I will use my few data science knowledge to analyze if the idea that I have of big difference between neigborhoods in the city is real.

Data

Based on definition of our problem, factors that will influence our analysis are:

- Most common venues in every Neighborhood.
- Socioeconomic data of every Neighborhood in the city of Barcelona.
- Geographical coordinates of every Neighborhood.

The location in the map of every neighborhood will be the coordinates of the center of the hood and I will use the geopy API.

Following data sources will be needed to extract/generate the required information:

- The location in the map of every neighborhood will be the coordinates of the center of the hood and I will use the **Geopy API**.
- number of venues and their type in every neighborhood will be obtained using **Foursquare API**
- coordinate of Barcelona center will be obtained using **Google Maps API geocoding** of well known Barcelona location (Plaça Catalunya)
- The socioeconomic data will be extracted from the local government of Barcelona Website, you can access there with the following link.

https://www.bcn.cat/estadistica/catala/dades/barris/index.html (be careful, the web is only available in Spanish)

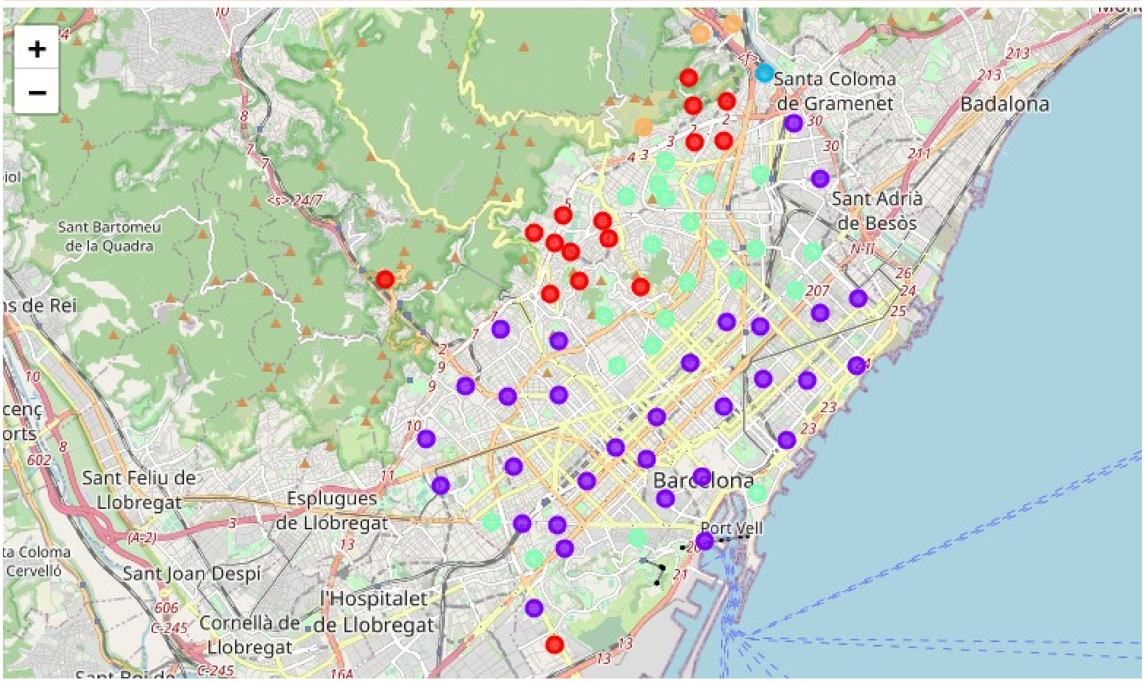
Methodology

Purpose of this project was to identify differences in clustering Barcelona Neighborhoods using two different types of data. By getting the most common venues in the neighborhoods we can start clustering the hoods using the k means clustering algorithm which will classify the more similar neighborhoods in the city in the same cluster.

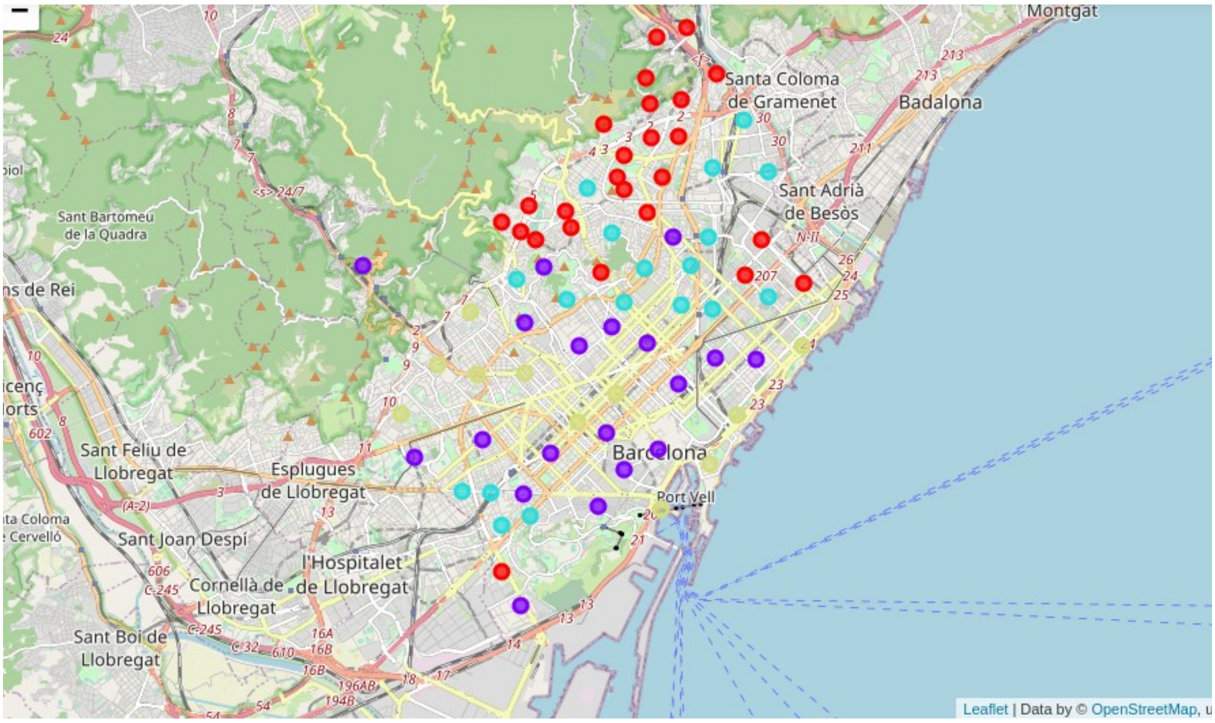
We are going to use the same procedure for the socioeconomic data clustering. In this case the data is properly prepared in a csv file and we will take the columns of data that are relevant for the analysis and the other columns will be dropped from the data frame. Then we can start clustering the hoods using the kmeans clustering algorithm which will classify the more similar neighborhoods in the city in the same cluster.

Results and Discussion

Once we have the two clustered maps of Barcelona we can start to analyze the results. Our analysis shows that the most common venues clustering gets 3 big clusters, the 'purple' cluster has the most touristic hoods the most common venues in this hoods are usually restaurants and hotels an they are located in the town center and the left part of the city. The second big cluster is the 'blue-green' cluster, this cluster contains hoods that are not as touristic as the first cluster and the most common venues are for example supermarkets or other shops more oriented in selling to local people. And finally the third big cluster is the 'red' one which contains the hoods of the outskirts of the city and they don't have a lot of venues and are usually different from the town center neighborhoods.



Secondly we have to analyze the clustering using socioeconomic data from the hoods of Barcelona. We can see that we have 4 big different clusters. The 'brown' cluster contains the most wealthy hoods in the city (Cluster Label, 3) for example if we take a look at the rent house prices we can see that are much higher than in the other clusters. Also if we take a look at the unemployment rate we can see that is very low in comparison of the rest of the town. The 'purple' cluster we can say that is a medium class cluster (Cluster Label,1) for example if we take a look at the rent house prices we can see that are higher than other clusters but lower than the 'brown' clusters. Also if we take a look at the unemployment rate we can see that is low or average in comparison of the rest of the town. Finally the 'blue' and 'red' clusters located in the (Clusters Labels, 0 and 2) We can see that the unemployment ratings and the rent prices are the opposite of the 'brown' cluster so we can say that they are the poorest hoods in town.



Conclusion

The main purpose of this project was to identify different shapes of clustering depending on the data used. We can conclude that despite the clusters don't have the same shape with the venues and the socioeconomic data we can conclude that they are quite similar and show us which are the hoods that are more oriented to the tourists and also which are the richest and the poorest hoods and where they are geographically located in the city.

Finally this project helped me to understand better how my city which is Barcelona has different types of hoods and how tourism has affected the city.