Analyzing neighborhoods of Madrid for a possible coffee shop opening

Eniko Kuris

January 20, 2021

Final project link:

https://github.com/eniko22/Coursera_Capstone/blob/main/Capstone_Final%20Project_Eniko%20Kuris.ipynb

1. Introduction

This is the Final Project for the IBM Data Science Professional Certificate Course.

For the final project I chose and worked on the following scenario: A client contracted me to run a data analyzation for the best possible neighborhood in Madrid for opening a coffee shop. The stakeholder is the client, this project is made directly to him so I can provide detailed information and he can make the right business decisions.

Madrid, as the capital of Spain attracts millions of tourists every year. However native Spanish people prefer having a *desayuno/almuerzo* (breakfast/brunch) in a local bar, the tourists prefer a fancy coffee shop where there is huge variety of coffees. My task is to detect the most popular neighborhood.

2. Data acquisition and cleaning

In this paragraph I will describe the data that was used to solve the problem and the source of the data.

2.1. Data sources

My data was retrieved from wikipedia, from the following link, where we can see list of neighborhoods in Madrid: https://en.wikipedia.org/wiki/Districts_of_Madrid

I chose my source carefully – wanted to make sure that the data is correct, so I cross-checked with several webpages.

2.2. Data Cleaning

The data was raw, it had to be cleaned and simplified. I wanted to have a clean tab sheet only with the name of the neighborhoods and the longitude, latitude which was needed for further analyzation.

3. Methodology

After importing all the libraries to make sure that we have all the necessary information to work with I started cleaning the data. This meant deleting some irrelevant columns, checking the data type and making sure that I only leave the relevant information so it would be easier to work with.

Once the data was easy to understand I used Geocoder library to get the coordinates (latitude and longitude) for each neighborhood.

After this step I was ready to connect to Foursquare API to further analyze the retrieved data.

First here I wanted to show on the map the neighborhoods so I decided to use Folium for that. With the help of Foursquare API I got the most common venues for the neighborhoods so I could determine the most common locals in each area. This helped a lot for further investigating. I also used the *group by* function to see where are the most common venues and as I supposed before, the most popular venues are in the Center area and near to Retiro, which is a beautiful park in the heart of Madrid.
Once I had these information I decided to use the *cluster* function, because I wanted to group the similar neighborhoods so I could have a better understanding. I made a formula to calculate the "k", and based on that I made four clusters – so I divided the neighborhoods to four different group.

4. Results

My hypothesis was that the most beneficial area for opening the coffee shop is the center but after the clustering function it became clear that the center ineed is the best neighborhood to open the said local. The first cluster brought the most accurate results.

5. Discussion

With this report my Client will have a better understanding about the most profiting area for his coffee shop in Madrid. It is very import to know, that my report is mostly based on popularity, as I used the most common venues for detecting it. However the popularity of the neighborhood is very important, my Client also need to take into account the transportation possibilities, the rent fees, the parking places available, etc.

6. Conclusion

With this data analyzation I managed to detect the most popular neighborhood in Madrid for opening a coffee shop. The center is the best area to do that as it is full of nice venues and as thousands of tourist spend their days in the center, the most beneficial are for the coffee is there too.