IBM Applied Data Science Capstone

Coursera

Recommending Italian Restaurants in Bogotá D.C., Colombia

By: Diego Ferreira

June 2020

Introduction

Bogotá D.C. is the capital and largest city of Colombia. With a population of 7.4 million, Bogotá is considered the core of Colombia economy.

One of the tourist attractions of Bogotá is the diversity of the cuisine. Colombian, Italian, Mexican, Peruvian, among other types of restaurants can be found in Bogotá and they are all very popular. Bogotá offers an increasingly wide range of flavors encompassing Colombian and International cuisine, which has converted the city into one of Latin America's major emerging culinary hotspot. Particularly, in this project we will provide some insights about the Italian restaurants that are currently available in Bogotá.

Problem Description

The objective of this capstone project is to propose some locations in the city of Bogotá to recommend to someone that has not been in the city but also to spot potential neighborhoods where someone can open a new Italian restaurant. Using data science and machine learning concepts and techniques, this project aims to provide answers to the following questions:

1. If someone who likes Italian food wants to open a new restaurant, where should that person build that restaurant?
2. If I like Italian food, in which neighborhoods I would find this cuisine?

Target Audience

This project is useful for investors looking to open new restaurants in Bogotá but also for people who would like to eat Italian food in the city. The target audience of this project is everyone who is exploring Italian restaurants in Bogotá.

Data

To solve this problem, we will need the following data:

- List of neighborhoods in Bogotá D.C.
- Latitude and longitude coordinates of the neighborhoods. This is necessary to locate each neighborhood in the map using Python.
- Venue data, we will use this data for our clustering algorithm.

Sources of the data and methods to extract them

The geolocation data will be acquired through geopostcodes (http://www.geopostcodes.com/). This database contains the list of regions and cities, as well as their neighborhoods and the coordinates for all of them.

After that, to get the venues of the neighborhoods that are of interest for us, we will use Foursquare API. Foursquare has one of the largest databases, it contains 105+ million places and it is used by over 125,000 developers. This tool will provide us many categories of venue data and we are particularly interested in the restaurants category to solve the problem description. This project will put in practice many of the skills that we have learned through the IBM data science course, such as working with API (Foursquare), data cleaning, data wrangling, machine learning (k-means clustering), and map visualization (folium library).

Methodology

Initially, we will get the list of neighborhoods and its coordinates from geopostcodes. This database already comes with the information that we need; hence we will save some time and we will not perform a web scraping using Python. We will use the pandas library from Python and put this database into a dataframe and visualize the neighborhoods in a map using the folium library.

We will use Foursquare API afterwards to get the top 100 venues that are within a 500-meter radius. Since we already used a Foursquare developer account through the past labs of the course, it means we already have a client ID and key. We then make API calls to Foursquare, passing the geolocation of each neighborhood to obtain the venues. From the Foursquare data that we receive, we will use the venue name, venue category, venue latitude and longitude. We will check how many venues were returned for each neighborhood and how many unique categories were returned for each neighborhood. We will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are preparing the data for use in clustering.

Finally, we will perform clustering on the data using the k-means algorithm. We will identify the number of clusters needed according to the elbow methodology and allocate each neighborhood to its corresponding cluster. This algorithm is one of the simplest unsupervised learning machine learning algorithms that will be very useful for this project. The results we will help us to visualize the locations of Italian restaurants in Bogotá so we can recommend them to someone interested in tasting Italian dishes but also will help us to identify location prospects where investors may want to open a new Italian restaurant.

Results

The results from the elbow method in the k-means shows that we need to use 4 clusters to group the neighborhoods according to their venues.

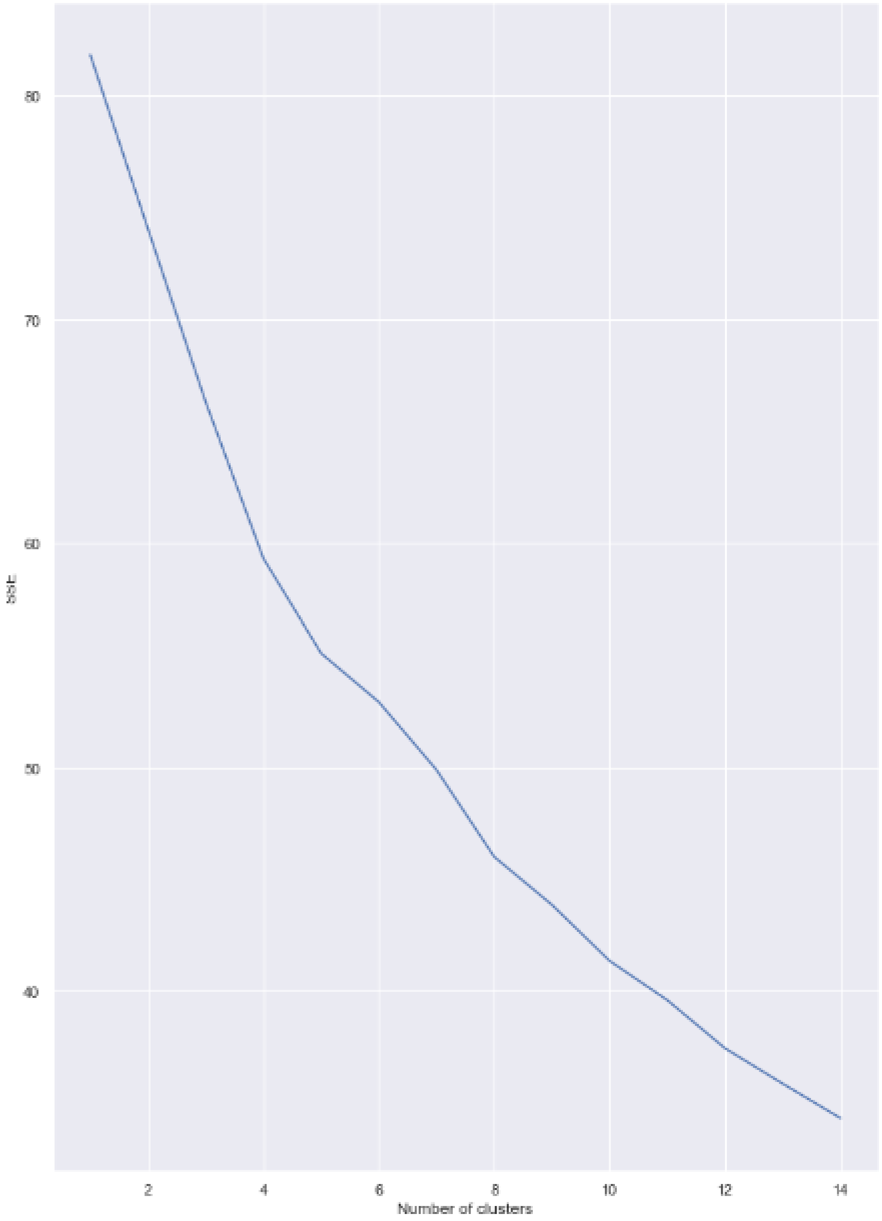

Figure 1. Elbow method to define the number of clusters

The results of the k-means clustering are shown in figure 2. The algorithm classified the neighborhood into 4 clusters. The cluster number 2 (blue circles) are the ones who have more venues in common, mostly restaurants in general.

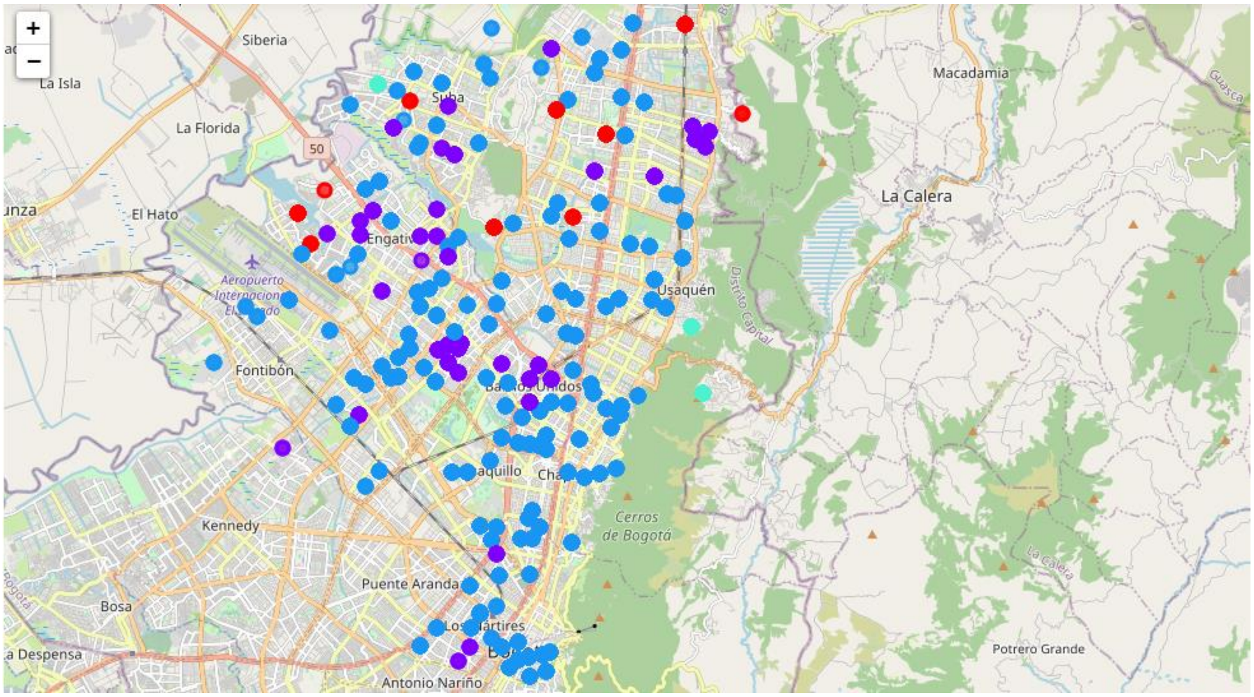

Figure 2. Clustering visualization of Bogotá’s neighborhoods.

The Italian restaurants are mainly located in the downtown, north and northeast parts of the city, as shown in the below figure.

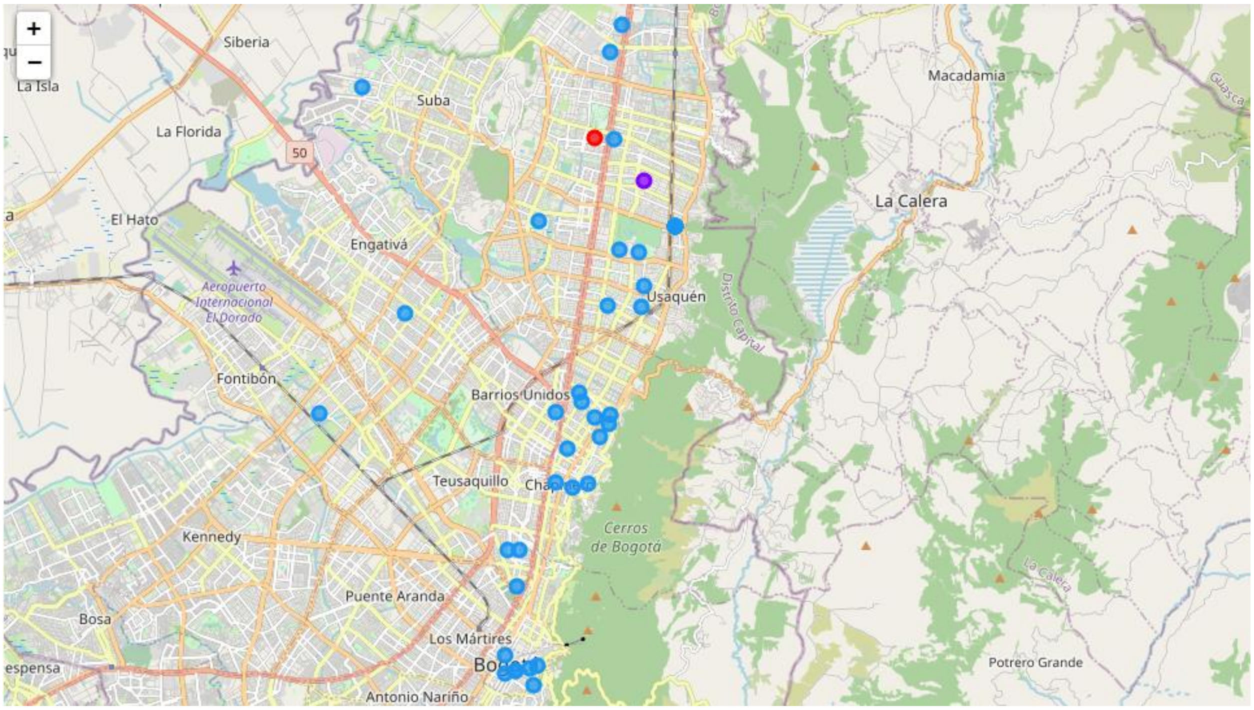

Figure 3. Italian restaurants in Bogotá.

Discussion

As seen in figure 3, most of the Italian restaurants are in the downtown, north and northeast parts of the city, with the highest number in the cluster 2. We can see that there are great opportunities for investment, as there are only two Italian restaurants in the west and 1 Italian restaurant in the borough called Suba. This represents potential areas to open new Italian restaurants in the city.

It is also possible to recommend people who like Italian food to concentrate in these zones when they want to try new Italian restaurants. Moreover, the owners of the restaurants could be suffering from intense competition due to oversupply and high concentration of Italian restaurants in these zones.

Limitations and suggestions for future research

In this project we only considered the frequency of occurrence of Italian restaurants as the only factor for our clustering algorithm. There are other factors such as population and income of the neighborhoods residents that may influence the location decision of new Italian restaurants. However, this data was not available at the time of the project, but it would be interesting to see the outcome when having such data. Moreover, due to the limitation of the requests that can be sent to Foursquare API, it was not possible to consider neighborhoods from the south zones of the city.

Conclusion

In this project, we were able to identify a business problem, specify the data required, propose a methodology to solve the business problem, apply a machine learning algorithm, and provide recommendations to potential stakeholders, such as investors and Italian food lovers, where they can find some locations where they can open new restaurants and also locations where they can go to try Italian food respectively. We identified that the cluster number 2 is where most of the Italian restaurants are concentrated and potential zones where new Italian restaurants can be opened.

References

- Foursquare Developers Documentation. *Foursquare.* Retrieved from https://developer.foursquare.com/docs
- Geolocation database. *Geopostcodes.* Retrieved from http://www.geopostcodes.com/
- Bogotá neighborhoods information. *Wikipedia.* Retrieved from https://es.wikipedia.org/wiki/Anexo:Barrios_de_Bogot%C3%A1