Implementation of a Data Science Analysis for the location of a new Hotel in Bogota.

Juan Pablo Mendoza.

January 26, 2020.

1. Introduction

In recent years Bogota, capital city of Colombia, has been subject of a rise in both foreign investment and inner economic development, which is precisely why, when faced with the question of where to install a new hotel, it is very important to know the characteristics of interest in the sector to locate the investment.

2. Data

In order to make a proper analysis over the different sectors in which to establish the new hotel, it will be required to separate the city into different locations and inspect the venues surrounding each of them. For this purpose the data concerning the locations will be scrapped from the official subdivisions of localities that can be found in the following link.

For the venues in each subdivision the data will be taken from the Foursquare database using a developer account.

3. Methodology

Once all the corresponding libraries have been imported, the project will begin by using the read_html function in order to subtract the localities names from the wikipedia page, however, these will also come with unnecesary data, which will be cleaned leaving only the names and the code of each locality.

The coordinates of each locality will then be attained by using the geocoder function and passing in the names of each locality from the already made dataframe. The localities will be mapped to see how they are distributed inside of bogota.

Now the venues will be obtained through the foursquare database providing our cliding data. A function will be created to exploit the "explore" functionality in the foursquare url for it to give us the surrounding localities and their coordinates. All this information will then be turned into a database.

We will explore the different unique venue categories surrounding the localities and provide a onehot encoding function for their future use in the machine learning algorithm. The dataframe will be grouped into neighbors to check the frequency of every kind of venue in each.

We will now select the different kinds of venues that will be of our most interest regarding the establishment of the hotel. The dataframe will be cleaned of the venues that won't make much of a difference to have nearby a hotel.

Now we will execute the machine learning algorithm in order to cluster the data.

Lastly, we will map the different clusters and check what the differences between each are, we will select the cluster that best fits our hotel project.

4. Results.

The clustering algorithm provided three different clusters with different frequencies, most notable changes may be seen in the food and entertainment related vanues.

5. Discussion.

It seems that the most significant variation between the localities comes in terms of food related businesses. As the entertainment ones are not quite a significant percentage of the venues.

6. Conclusion.

The clustering algorithm, with the help of the Foursquare database has provided 3 different clusters involving the 20 localities of Bogota, Colombia. As it can be seen from the segmentation of each cluster, the localities in Cluster No. 0 seem to have slightly more presence of cultural and entertainment facilities in comparison with the other clusters. Cluster No. 1 has a significantly higher presence of restaurants and food-related businesses compared to the other clusters. Lastly, Cluster No. 2 seems to be rather empty.

In conclusion, even though Cluster No. 0 does seem to provide more entertainment facilities, it is a slight difference compared to Cluster No. 1, which in turn has a significant upper-hand in restaurants and similar businesses. Therefore, in regards to the positioning of a new Hotel, Cluster No. 1 is the recmomended option.