IBM APPLIED DATA SCIENCE CAPSTONE

THE BATTLE OF NEIGHBORHOODS

1. BUSINESS PROBLEM

Berlin is the capital and largest city of Germany by area and population. It's 3.748.148 inhabitants making it the second most populous city of the European union after London. Berlin is also one of the most visited cities in the continent thanks to its huge cultural, economic, political and scientific offer. It makes the city the third most visited destination.

Moreover, most Berliners have got used to rising rents because of the high demand of housing among different other reasons. In recent years, the capital has seen rents rise faster than any other city in the country. Online portals of the so – called "sharing economy", such as Airbnb, are marketing to tourists and have proven to be a profitable business. Currently there are 11.701 listings in Berlin being equal to 0.4% of all the city flats and 34.418 Airbnb offered beds equivalent to an average of 2.9 beds per flat.

As a tourist and business man who has to come often to the city it is helpful to know the relationship between the location and kind of nearby venues and spots with the price for a room per night. Being Airbnb the most popular low-cost offer to afford a vocational flat or room in the city, this analysis will identify the different possible existent cluster in order to understand and select the best accommodation according to the needs to the tourist.

2. DATA COLLETION

To consider the above problem, the data were collected as followed:

- Airbnb listing data set of Berlin. Including key features as the location coordinates, neighborhood and price per night of the room.
  URL: http://insideairbnb.com/get-the-data.html

- Foursquare API to get the most common venues of given neighborhood of Berlin.

3. METHODOLOGY

The obtained data set from the webpage was transformed into a data frame using pandas. With the purpose of having a clean data set, all unnecessary columns were drop keeping just the information required for the analysis, being price, latitude, longitude and name of the neighborhood the selected features. It was also decided to normalize the neighborhoods according to the 12 zones which the city is divided in. In order to get the centralized zone location coordinates, a group operation by the neighborhood was executed.

The Foursquare API was utilized to explore the different zones and segment them. The limit and radius were set to 100 venues and 4.5 Kilometers for each neighborhood respectively. Taking into account the first created data frame, a new one was created having the information of the 10 most visited venues in each zone. All the used future for the statistical model were normalized.

4. MACHINE LEARNING ALGORITHM

There are some common venue categories along the neighborhoods. For this reason, the unsupervised K-means algorithm was used in this project. The K-means algorithm is one of the most recognized and used clustering methods.

First, the algorithm was run with different k values to select the best one for the model according to the elbow method. In this case, a k value of 4 clusters was selected. Afterwards, the algorithm was executed again to segment the 12 neighborhoods in 4 different clusters.

5. RESULTS

As result, it was obtained 4 apparently different clusters. After analyzing them, it was observed that types of venues are relative equally distributed along the clusters, being the restaurants and cultural spots the most commons. However, the room princes change according to the location of the neighborhood as followed:

- **HIGH PRICE ZONES (CLUSTER 2):** Average prices between 97 and 110 euros per night. This cluster includes the following zones: Charlottenburg-Wilmersdorf, Mitte and Tempelhof-Schöneberg. These are the most central areas in the city.
- **HIGH – INTERMEDIATE PRICE ZONES (CLUSTER 0):** Average prices between 59 and 70 euros per night. This cluster includes the following zones: Friedrichschein-Kreuzberg, Lichtenberg, Neuköln, Pankow, Spandau, Steglitz-Zehlendorf and Trettow-Köpenick. They are areas which are a bit further from the center of the city.
- **LOW – INTERMEDIATE PRICE ZONE (CLUSTER 4):** Average prices of 59 euros per night. This cluster includes the following zone: Marzahn-Hellensdorf. It is located at the eastern side of the city (peripheric area).
- **LOW PRICE ZONE (CLUSTER 1):** Average prices of 49 euros per night. This cluster includes the following zone: Reinickendorf. It is a peripheric area located in the north of the city.

6. CONCLUSION

The location selection has a big influence in the price of an Airbnb room in the city. The central areas are the most expensive and the price decreases according to the distance from the center of the city. The further the location, the lower the price will be taking into account the trend of getting the lowest prices in the eastern and northern side of the city. All the zones have relatively good transport connection and different kind of venues.