Capstone Project: The Battle of Neighborhoods

Naiara Marcos Gonzalez

May 27, 2020

1 Introduction

1.1 Description of the problem

Previous weeks, we studied two cities, New York and Toronto, which are two big financial cities from North America and their population is made up of diverse nationalities. In this project, we will focus on comparing London to these other two cities and discuss if it is more similar to New York or Toronto or none.

1.2 Discussion of the background

Finding similarities or dissimilarities between cities can be very interesting from different points of views. For example, expanding business, that is to say, if there is a business or infrastructure which is working well in one city, maybe it would be a good idea reproduce it in the city which looks more alike. Another possible example, it might be finding accommodation for people who move from another city and they want to live in similar area than they used to, and so on.

In this study, we chose London because it shares the same characteristics as New York and Toronto but in Europe, it is a big financial centre with high and heterogenous population.

1.3 Target audience

The potential audience of this project is a person or company which already has a business in London and wants to move/expand to North America, but he has doubts about which of the two cities (New York and Toronto) is a better fit. For that reason,

we should study the similarity between these three cities: London, New York and Toronto.

2 Data

In order to compare these three cities, we will use the studies of New York and Toronto which we made the previous weeks in Coursera. We add new data from the city of London. The data sources are:

New York:https://geo.nyu.edu/catalog/nyu_2451_34572
Toronto:https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto
https://open.toronto.ca/dataset/neighbourhoods/
London:https://www.doogal.co.uk/london_postcodes.php

This data provides us the Postcode, Borough's Name, Neighbourhood, the respective Latitude and Longitude for each Neighbourhood that every city contains. From this data, we can use the Foursquare API to get information about restaurants, coffee shops, parks, theatres, and so on for each neighbourhood. Then, with this data we plan to apply the k-means algorithm to create a cluster analysis of the neighbourhoods from each cities. With these analysis we plan to tell which city is London more alike and to find a correspondence between each city neighbourhood.

3 Methodology

We compare the three cities: London, New York and Toronto. Firstly, we analyse the neighbourhoods' venues which got from the Foursquare API, using different plots such as, the amount of neighbourhoods are for each borough, the 80 most common venues for each city followed by a word cloud with all places. Moreover, there are two tables where are represented the number of venues per neighbourhood /borough, and the 5 most common places in each neighbourhood.

After that, we select the 10 most common spots for each neighbourhood. Then, we apply k-means algorithm from *scikit-learn* library to an unsupervised data, in order to create clusters between the neighbourhoods from every city. With the aim of finding an optimal k, number of cluster, we plot the inertia model, which is defined as sum of squared distances of samples to their closest cluster centre vs number of cluster k. Thus, choosing k when the plot is observed an elbow or location of a bend. Note,

sometimes the second elbow is the k which has been selected, because we decide to take the second if this is closer to the k observed in other cities.

4 Results

Observing the neighbourhoods distribution around the boroughs, Toronto's data contains less neighbourhoods than the others cities, and London's data has more neighbourhoods and boroughs with respect to the others cities. Therefore, from those data, a dataframe is created with the locations belongs to these areas. Bellow, we can observe the total results for each city:

- New York: 10147 venues distributed in 425 unique categories.
- Toronto: 2015 venues distributed in 278 unique categories.
- London: 12236 venues distributed in 400 unique categories.

There are significant differences between the number of venues from each city. However, we still compare the cities because we only consider the category venue. Furthermore, observing the plot of the most common sites, the categories are practically the same.

Applying K-mean algorithm in the 10 most common places for each neighbourhood, we decide to use different k's, number of clusters for every city, for New York there are 8 clusters, London there are 9 clusters and Toronto there are 6 clusters. All clusters are visualised in every city via folium map.

5 Discussion

Observing the clustering results, we can find similarities between neighbourhoods from different cities, because they are formed by common places such as coffee shop, pizza place, bus stop, and so on. However, there is a table in every notebook where is included the number of venues for each neighbourhood and borough and the values are not homogeneous. Therefore, the data is quite limited and poor in order to determine likeness between cities. It would be interesting to get more data, not only 10th common category venues, but also data related to people who live in the neighbourhoods, the incomes by family, housing prices, etc.

6 Conclusion

Summing up the capstone, three cities have been chosen, New York, London and Toronto, in order to compare the neighbourhoods of each city. The data comes from different Websites and it is composed by the following features: Boroughs, Neighbourhoods, Longitude and Latitude. Using the Foursquare API, we acquire the venues from each neighbourhoods, we note that the number of venues per neighbourhood is quite unbalanced. After that, we select the 10 most common venue category, and then we apply an unsupervised data k-mean algorithm in order to find patterns, that is to find the similarity between these areas. Following this, we try to compare the 10 most common venues which belong to different clusters with the three cities.

We realise that this data is not enough to determine whether the cities are similar or not, since the obtained data is very limited. Therefore, it would be attractive to have features such as the amount of the population, housing prices, even venues from another application, etc. and larger dataset, not only 10 categories for each neighbourhood.