Capstone Project - Battle of Neighborhoods

Ramya Sai Bhagavatula

July 26, 2020

1. Introduction

1.1 Background

New York City is one the most populous cities in the United States. It is diverse, multicultural and is the global hub of business and commerce. The city is a major center of financial, and media capital of the world, significantly influencing commerce,entertainment, research, technology, education, politics, tourism, art, fashion, and sports. It also highly promotes business opportunities for anyone who wants to start a new startup.

Toronto is the provincial capital of Ontario and is the most populous city in Canada. It is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. It is considered as the financial capital of Canada and has a high concentration of banks and brokerage firms on Bay Street, in the Financial District. Toronto’s stock exchange is considered to be the world's seventh-largest stock exchange by market capitalization and the Big Five companies having national offices in Toronto.

Considering how perfect the locations are in terms of starting a new business and the markets being highly competitive, cost of doing business will also be very high. Thus, any new business venture or expansion needs to be analysed very carefully. The insights derived from this analysis will give us a good understanding of the business environment which will help us in strategically targeting the market and also help us reduce the risk factor associated.

1.2 Problem Description

Suppose there is a person ‘A’ who is the Managing Director of a Company wants to expand his business and has selected 2 prime locations i.e. Toronto, Canada and New York City, USA. ‘A’ wants to understand the neighborhoods and local businesses in those cities so that he can understand how the cities will be in terms of diversity, food cuisines, living standards and the quality of life for the employees. This project will explore the neighborhoods of both the cities and determine which will be a best fit to their company to expand their business and which area to work upon.

I have been appointed as the Data Analyst to help A's company. After having a clear discussion and understanding the companies concerns, I have understood there are various factors we should be considering before we start analyzing the problem i.e. we need to have an understanding on

- Population Statistics of both the cities
- The Demographics
- Competitors in those locations. and so on. After understanding we can start working on how similar/dissimilar the cities and which would be a good fit.

Fig 1. NY City Map

Fig 2. Toronto City Map

2. Data and Description

2.1 Data Sources

For this project we will be using the following datasets:

To analyze the New York City:

I have used the Beautiful Soup module from the bs4 package to extract the data from the Cities Wikipedia pages. In order to segment the neighborhoods and explore them we need to collect data of the neighborhoods and their boroughs.

1. New York Neighborhoods and their Lat & Long Coordinates - https://geo.nyu.edu/catalog/nyu_2451_34572

2. New York Population and its Demographics - https://en.wikipedia.org/wiki/New_York_City
3. Competitors in a particular business - Can be found by Foursquare API
4. The Geographical Coordinates of the boroughs - can be found using Python Geolibrary

To analyze the City of Toronto:

The following websites are used to analyze the City of Toronto.

1. Toronto-Neighborhoods - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. Toronto Geographical Coordinates - http://cocl.us/geospatial_data

All the above datasets will provide us information on the Postal Codes, Geographical Coordinates, Borough and Neighborhood Names of both the cities.

Foursquare API will help us in finding more information about venues such as best restaurants, coffee shops, must visit places etc. which will help us understand both the locations.

We will also be using Folium and various other Python Packages to create Maps representing both the locations as well so that we can visualize the similarity/dissimilarity of both the regions and use machine learning algorithms such as K- Means Algorithm to cluster the data into various segments.

More understanding about Foursquare and its usage can be found at https://foursquare.com/

3. Methodology

The goal of this project is to group similar neighborhoods in the cities of New York, and Toronto. Since our data is unlabeled i.e. unsupervised we will use k-means clustering technique to segment the neighborhoods.

We have made use of the Beautiful Soup Package to draw http requests to extract the data from the web pages. The data is preprocessed and is joined with the respective geographical coordinates.Later we created a map of the respective cities to have a geographical understanding using Folium Package. Now in order to explore the various venues of a particular neighborhood we had to interact with the Foursquare API by giving our respective credentials.such as Client ID and Client Secret. HTTP requests would be made to this Foursquare API using zip codes of the city neighborhoods.

The API's search feature would be enabled to collect the nearby places of the neighborhoods. Due to certain limitations of the requests we have set the neighborhoods radius parameter to 700. The frequency occurrence of all the venues in that location are sorted from highest to lowest. Folium package is used for the visualizations of all the maps to visualize the neighborhood cluster distribution. For the scope of this project I have selected Scarborough from Toronto City and Long Island City from New York. Desirable insights were drawn using the various packages such as scikit, pandas, numpy etc. Later an unsupervised algorithm - K mean clustering is used to cluster the neighborhoods with the different categories. These neighborhoods can be analyzed individually to derive the results.

4. Results

<u>Visualizing the clusters on the Map:</u>

Scarborough Borough in Toronto, Canada

I used K-means clustering to cluster the neighborhood into 3 clusters. In the figures below we can see the various ways in which the neighborhoods were defined and visualized. In Fig 3 we can see the overall representation of Toronto City with all its neighborhoods imposed on it. We can understand it as a whole city map. In Fig 4 I have zoomed into only the neighborhoods of Scarborough Borough of Toronto. We can see where the neighborhood is located. Later, I have tried to cluster the neighborhoods of the borough with k=3. Fig 5 represents the clusters as follows: Cluster_0 has 15 neighborhoods and has the most common venues as Fast Food Restaurant, Bar, Coffee Shops, breakfast stops, Skating Rinks etc.Cluster_1 has 1 neighborhood with the most common venue as General Entertainment centre and Skating Rink. Cluster_2 had the most common venues as an Accessories Store and Middle Eastern Restaurant.

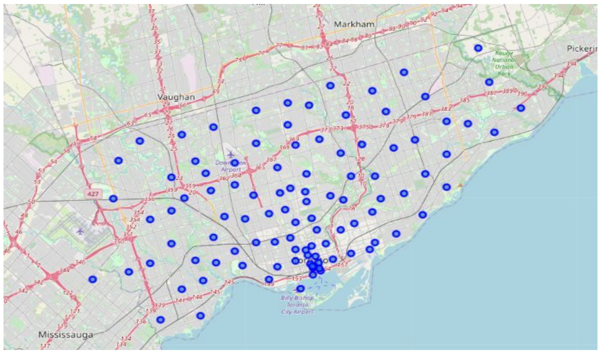

Fig 3. Toronto City with all its neighborhoods superimposed.

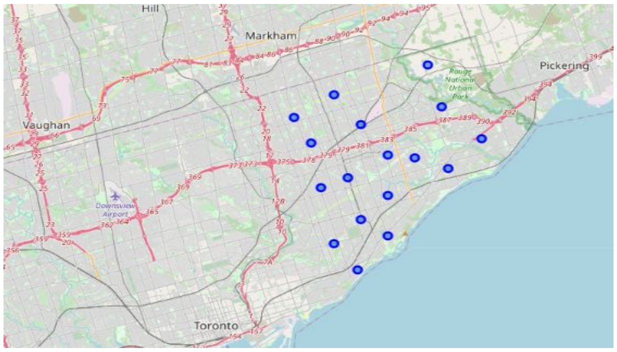

Fig 4. Neighborhoods of Scarborough Borough

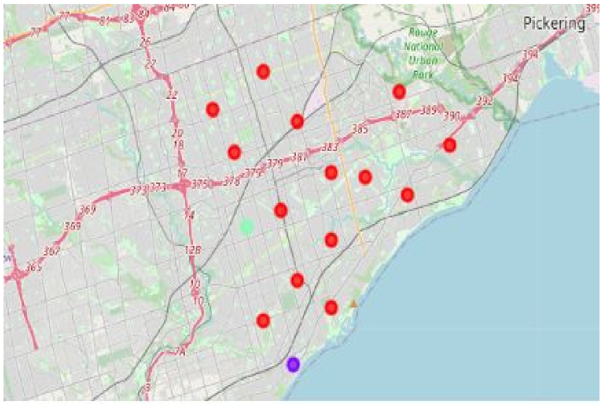

Fig 5. How the venues in neighborhoods of Scarborough are clustered.

Long Island City from Queens Borough

I used k-means clustering to group the Queens into 5 groups. The first cluster had 81 neighborhoods and mostly had the common venues as Pizza Place, Bus station, Beaches etc. The second cluster had 1 neighborhood with the most common venue as a Gym/Fitness Center. The third had 1 neighborhood with the most common value being a Brewery. The fourth had 3 neighborhoods with a Donut Shop, a mobile phone shop and a grocery store as the most common venues. The final cluster had 1 neighborhood having the most common venue as a Deli Store, a Spa, a pizza place etc. Below figures represent the map visualization of various areas of new york. Fig 6 represents the whole new york city with the neighborhoods imposed on top. Fig 7 shows the various neighborhoods of manhattan on the map. Fig 8 shows how the various venues were classified as different clusters on the map of long island city.

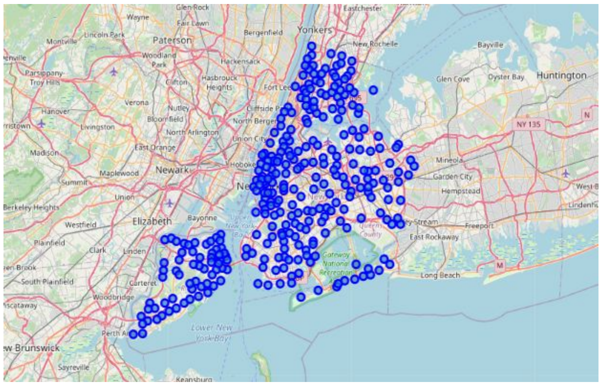

Fig 6. Newyork City with all its neighborhoods superimposed

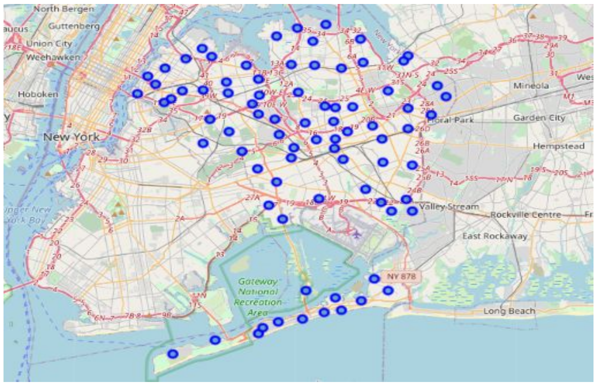

Fig 7. Neighborhoods of Manhattan Borough

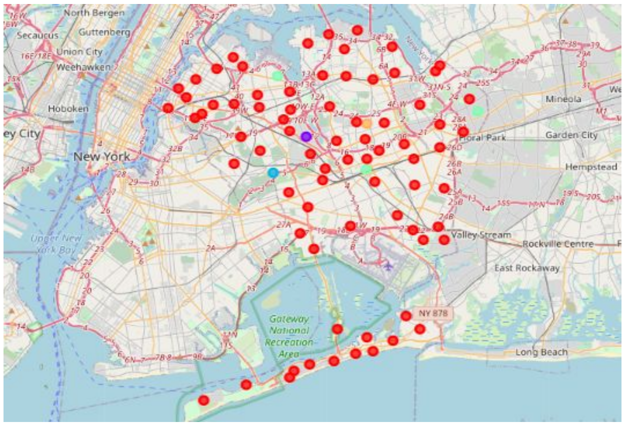

Fig 8. How the venues in neighborhoods of LICity are clustered.

5. Discussion

Since this project is an unsupervised classification, different approaches may tend to lead different results. This project has been done by working on 2 cities Toronto, Canada and New York, USA with their respective zip codes. Toronto has 11 boroughs and 103 neighborhoods with the geographical locations as 43.71 and -79.419. We have worked on the Scarborough Borough of Toronto having 90 venues in 17 neighborhoods. New York City on the other hand had 9 boroughs and 306 neighborhoods with its geographical location as 40.71 and -74.0. We have worked with the Long Island City of Queens Borough and found it to have 2091 venues in 81 neighborhoods. We can see many of the neighborhoods as homogeneous and quite similar to each other. Both cities had a neighborhood cluster with a majority of neighborhoods in one and the other clusters having 1 or 2 or 3 neighborhoods within them. But overall, Queens Borough had more number of venues as compared with Scarborough Borough. Dealing with location details in a deeper way may lead to better dealing of similar data points together.

6. Conclusion

People moving to different cities and having new experiences is ever ending. Companies interested in expanding their businesses to better locations and employees having new challenges are evolving as well. In such cases having such a neighborhood recommendation based on location data would be highly wishful. This can be further used to organize the city's resources in a better way. Overall, based on the quality of venues, it is preferred for Company A to start up its new business at Queens Borough, NY, USA. It offers more choices in terms of restaurants, gyms, indoor and outdoor activities and so on leading to better opportunities for the employees and their families as well.