Project: *Exploring and Comparing Neighbourhood Clusters in Toronto and Manhattan to Select the Most Similar Neighbourhood*

1.0 Introduction/Business Problem

One issue that people must face in a globalized world is moving far away from home. Whether they are pursuing a dream career, moving closer to family and friends, or just wanting a change of environment, it is inevitable that some people will make big changes in their life and move elsewhere. People would rather move to areas where they feel the most comfortable so they can adjust to their new surroundings a little easier; some may want to move somewhere that has similar food culture, shops, recreational spots, etc. However, when you're moving somewhere new, it can be overwhelming to search through the whole city and find these areas.

Both New York City (NYC) and Toronto are major metropolitan cities renown for being financial capitals of their respective countries with incredible multiculturalism. However, they are in different countries and have different histories; their environments, though similar in some ways, are vastly different in other ways. If someone was moving from Toronto to Manhattan, it would be ideal to move to a neighbourhood that had the greatest similarity to their home city to ease the adjustment process.

This Capstone project aims to analyze neighbourhoods in Toronto and NYC and determine their similarity (this will mostly be based on similarity in venues using Foursquare API), using these insights as a reference for recommending which neighbourhood one should choose in NYC when moving there from Toronto. Following data science methodology, machine learning techniques such as k-means clustering will be used to provide answers to the business question: What similar neighbourhood in Manhattan would you recommend an individual move to if they were moving from Toronto?

2.0 Data Acquisition and Cleaning

The following data will be needed to answer the business question:

- List of neighbourhoods in Toronto and their latitude and longitude coordinates
- List of neighbourhoods in NYC and their latitude and longitude coordinates
- Information of the venues in each of the neighbourhoods

2.1 Data sources

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) "List of postal codes of Canada: M" will be used to get all the names of each neighbourhood, their boroughs, and postal codes. Web scraping techniques will be used to extract the data, using BeautifulSoup and other Python libraries.

To get Manhattan neighbourhood data, New York City data was needed. Luckily, this data could be found here: https://geo.nyu.edu/catalog/nyu_2451_34572 in the form a JSON file, which was used to get all the names of each neighbourhood, their boroughs, and their latitude and longitude coordinates.

The geographical coordinates (latitude, longitude) of each neighbourhood in Toronto was retrieved using Google's ARCGIS package.

Foursquare's API will be used to get the nearby venue data for each neighbourhood.

The machine learning technique k-Means clustering was used to cluster the neighbourhoods, and the Python folium library was used to visualize these clusters on a map. Then, the clusters of neighbourhoods that have the most similarity in terms of types of venues were compared.

2.2 Data Pre-Processing

Data that was downloaded or scraped were combined into one data frame table. There were no missing or duplicated values for either dataset.

I used Google's ARCGIS package to iterate through each address in the Toronto dataset in order to get their respective latitude and longitude values. I then appended this to the main data frame.

Because I was only interested in one specific borough for each city, I filtered both the Toronto and NYC data frames so that they would only display neighbourhoods for Downtown Toronto and Manhattan, respectively.

Finally, I merged both data frames into one new working dataset, as I would be clustering both Downtown Toronto and Manhattan neighbourhoods to determine their similarities.

2.3 Foursquare API

After obtaining the working dataset for both Downtown Toronto and Manhattan, I used the Foursquare API to explore all the nearby venues within 1km of each borough. The limit value for the API was set to 100, meaning that a maximum of 100 nearby venues will be returned for each neighbourhood.

2.4 Exploratory Data Analysis

I first looked at how many venues were returned for each neighbourhood, this was done via grouping by counts for each neighbourhood. I also noticed that some neighbourhoods had less than 100 nearby venues, which could skew the results of the clustering. To address this, I decided to only use the top 10 venues for each neighbourhood for clustering. I used one hot encoding to make grouping venues simpler. This way, I can take the mean of the frequency of occurrence of each category, then sort these into a new data frame that displays the top 10 venues of each neighbourhood.

Once that was done, I used k-means clustering on the data set to determine similarity of each neighbourhood based on their top 10 common venues. I then appended the cluster label of each neighbourhood to the data frame.

3.0 Results

After adding the cluster labels to the dataset, I created folium maps to visualize each neighbourhood and their clusters. The coloured markers on the map represents a neighbourhood, and markers in the same colour represent the same cluster.

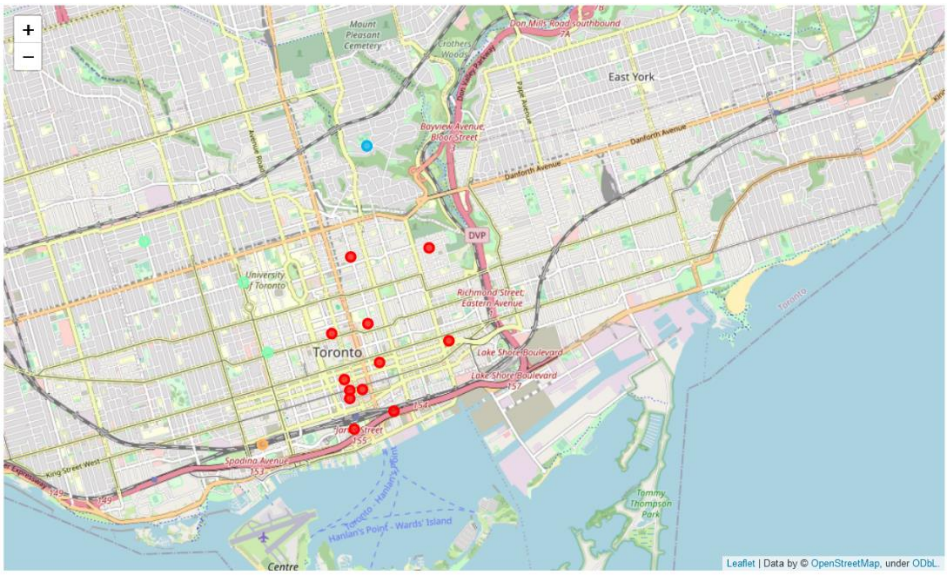

Figure 1. A map of Toronto marked with clusters of each neighbourhood

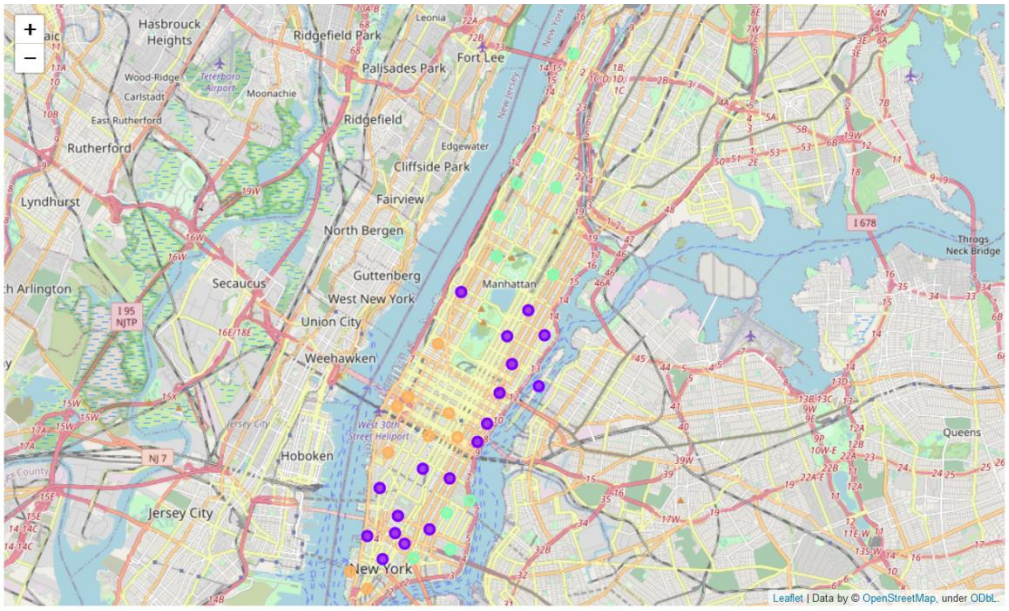

Figure 2. A map of Manhattan marked with clusters of each neighbourhood.

As you can see, Manhattan is much bigger than Toronto, and has more venues and neighbourhoods to explore compared to Toronto. It would therefore be appropriate and more time efficient to select 5 neighbourhoods that are in the Toronto dataset (representing where the individual is moving from) and see which clusters in Manhattan are the most similar to those.

I randomly selected 5 neighbourhoods from the Toronto dataset. The 5 neighbourhoods that were picked were:

1. University of Toronto, Harbord (cluster 3)
2. Harbourfront East, Union Station, Toronto Islands (*This will be shorted to Harbourfront East*) (cluster 0)
3. Kensington Market, Chinatown, Grange Park (*This will be shortened to Kensington Market*) (cluster 3)
4. Garden District, Ryerson (cluster 0)
5. Toronto Dominion Centre, Design Exchange (T*his will be shortened to Toronto Dominion Centre*) (cluster 0)

I parsed through the Manhattan dataset to compare which neighbourhoods were similar to these neighbourhoods. This was done by comparing which neighbourhoods fell into the same clusters (cluster 3 and 0).

It was found that if you are moving to Manhattan from University of Toronto, Harbord or Kensington Market (cluster 3), the similar neighbourhoods in Manhattan you could choose are:

- Marble Hill
- Chinatown
- Washington Heights
- Inwood
- Hamilton Heights
- Manhattanville
- Central Harlem
- East Harlem
- East Village
- Lower East Side
- Manhattan Valley
- Morningside Heights

- Stuyvesant Town

However, if you are moving from Harbourfront East, Garden District, Ryerson, or Toronto Dominion Centre (cluster 0), there are no similar neighbourhoods in Manhattan.

4.0 Conclusions

From the clusters, we can see that Downtown Toronto had 4 clusters in total (0, 2, 3, and 4) while Manhattan only had 3 clusters (0, 1, and 4).

The biggest cluster in Toronto was cluster 0, which is why randomly selecting values from the data frame had returned cluster 0 neighbourhoods as the majority. Disappointingly, Manhattan does not have any neighbourhoods that fall in this cluster. This could indicate that Toronto and Manhattan are not that similar despite being major multicultural cities.

However, some neighbourhoods are similar to Manhattan, such as those that fell into cluster 3. When we take a closer look at the top 10 venues of the neighbourhoods in cluster 3, we can see that these are neighbourhoods that have a lot of focus on food, coffee shops; food culture is important to both cities, so it would stand to reason that the most similar neighbourhoods would have a lot of food and cafe venues.

In this project, I was comparing neighbourhoods in Downtown Toronto and Manhattan to determine their similarities in an attempt to provide insight to those who were interested in moving from Toronto to Manhattan.

While there are some neighbourhoods that share similarities (i.e. similar venues concerning food), it appears that the majority of neighbourhoods do not share much similarity between Toronto and Manhattan.

It should be noted that this analysis was done only taking into account nearby venues within 1km of each neighbourhood. Similarities in culture and environment can be determined in many ways, so do not solely rely on factors like venues to make your decision.