# 1. Introduction

**Background**

New York City (NYC) and Toronto are located in North America and are major financial hubs in the world. They are made up of different skyscrapers and business centers. Both are very cosmopolitan and have dynamic life style. Apart from the commercial perspective, they also build with many high-rise residential buildings. Many Global organization around the world have office located in these 2 countries. Many people often relocate from other countries to these 2 cities and working in the central business district (CBD) areas. They may not be aware of the similarities or differences in these 2 cities. One of the examples is related to the ethnic makeups in NYC and Toronto. NYC has a much larger Black and Latino population, whereas Toronto has proportionally more Asians and Indians. Hence the likelihood of NYC having more America or south America Restaurant than Toronto is higher.

The target audience for this project is the expatriate who will move to either cities and will work on the CBD areas. Hence the scopes will focus on the Manhattan New York and East, downtown, central and West Toronto areas

**Problem and Interests**

Given the diversity of the culture, this project will compare the following neighbourhoods of these two cities and determine how similar or dissimilar they are. In total,

- Manhattan consists of 40 neighbourhoods
- East, downtown, central and West Toronto consists of 39 neighbourhoods

It will focus on 3 topics

- Difference of the venue category between these 2 cities.
- Difference between the food culture based on the type of restaurant.
- Both cities will be independently split into clusters by neighbourhood. And then comparison between clusters will be done and identify similarity based on the venue category

It meant to provide the information for expatriates who plan to live in the neighbourhoods around the CBD areas so that they can choose the neighbourhoods best suit to their life style and needs.

# 2. Data

**Source of Data & Data**

Two data sets, one for Manhattan, one for Toronto, created from the previous labs or projects of the training course will be used as the source of data. These datasets have already populated with the information of the boroughs and neighbourhoods of NYC and Toronto as well as the respective latitudes and longitudes.

Before the data analysis, the neighbourhood candidates (NC) need to be filtered from the source of datasets. The outcome will have 2 datasets.

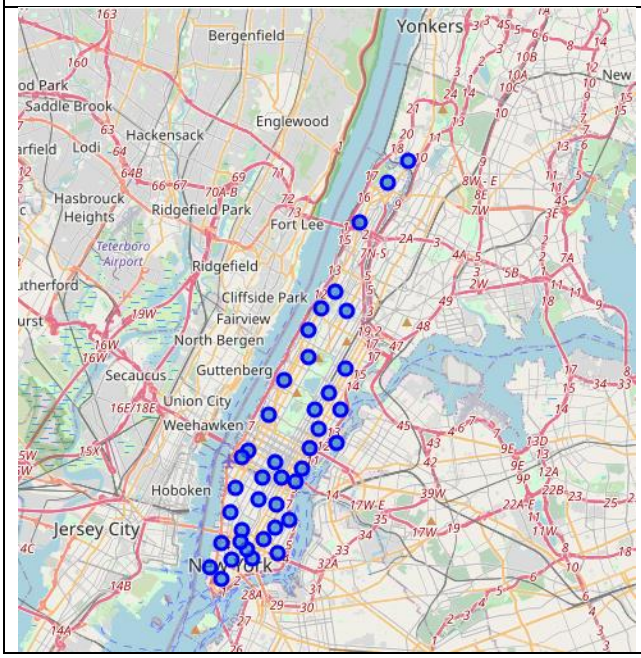
- Neighbourhood Candidates Set A - represent the  40 neighbourhoods of Manhattan. The sample data is as follows

| | number | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| **6** | 6 | Manhattan | Marble Hill | 40.876551 | -73.910660 |
| **100** | 100 | Manhattan | Chinatown | 40.715618 | -73.994279 |
| **101** | 101 | Manhattan | Washington Heights | 40.851903 | -73.936900 |
| **102** | 102 | Manhattan | Inwood | 40.867684 | -73.921210 |
| **103** | 103 | Manhattan | Hamilton Heights | 40.823604 | -73.949688 |

- Neighbourhood Candidates Set B - represent the 39 neighbourhoods of East, downtown, central and West Toronto. . The sample data is as follows

| | number | Postal Code | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **38** | 37.0 | M4E | East Toronto | The Beaches | 43.67635739999999 | -79.2930312 |
| **42** | 41.0 | M4K | East Toronto | The Danforth West, Riverdale | 43.6795571 | -79.352188 |
| **43** | 42.0 | M4L | East Toronto | India Bazaar, The Beaches West | 43.6689985 | -79.31557159999998 |
| **44** | 43.0 | M4M | East Toronto | Studio District | 43.6595255 | -79.340923 |
| **45** | 44.0 | M4N | Central Toronto | Lawrence Park | 43.7280205 | -79.3887901 |

Then, the geographical locations of the neighbourhoods will be reviewed to ensure the neighbourhoods are next to each other to ensure they are not scattered too far apart.

| Manhattan | Toronto City Area |
|---|---|
|  |  |

**Features selection**

The venues and venue categories will be the key features for the analysis. Hence, Foursquare API will be used to extract the revenues and revenue categories of all the neighbourhoods for these 2 cities. These data will combine with the datasets Set A and Set B to create new datasets that have the neighbourhoods and the revenue categories.

**Approach**

After the data source have been loaded into the data frame with data cleansing and filtering, Foursquare API will be used to collect the venues, latitudes, longitudes and venue categories for the neighbourhoods of Manhattan and Toronto City area.

To address the 1st audience interest, multiple datasets will be created to store venue categories followed by using **"SET"** operations to identify

- The common venue categories for both cities.
- The venue categories existed in Manhattan but not in Toronto City Area.
- The venue categories existed in Toronto City but not in Manhattan.

Difference between the food culture based on the type of restaurant will be the 2nd part of interest in this project. The "Restaurant" will be the key word to extract the records from the previous datasets and conduct an analysis or comparison.

Finally, the similarity of neighbourhood based on the venue category will be assessed. To do that, **one hot encoding** will be use to split the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Both cities will be independently split into clusters by neighbourhood using cluster algorithm "**kmeans**"; and the comparing the clusters and surface out the similarity based on the venue category