# Compare the similarities of financial world cities (New York and Toronto)

## 2. Data
### 2.1. Description

The datasets needed to complete this project are what I used in previous work. They include the data from New York and Toronto, respectively. For the New York dataset, the Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, I essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. The dataset exists for free on the web and the link to the dataset is shown as: https://geo.nyu.edu/catalog/nyu_2451_34572. For the Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto. In general, it is important to first scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format like the New York dataset. The link to the Wikipedia page of the dataset is shown as:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

## 2.2. Usage Method

These two datasets are first faced with some cleaning work. After downloading the data, the first work is to transform the data into a panda dataframe. Next, we need to use some geopy libraries to get the latitude and longitude values of the two cities and create two maps of them with neighborhoods superimposed on top. Here, we need to use a library called Folium, which is a great visualization library. We could zoom into the above map, and click on each circle mark to reveal the name of the neighborhood and its respective borough. Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them. After completing the above steps, we will face the most important part, analysis work. We need group rows by neighbor and by taking the mean of the frequency of occurrence of each category. Finally, we use the k-means method for clustering neighborhoods. This includes the work of comparison and analysis. This project would exchange test datasets to determine how similar the

two models are. The exchanged test datasets are compared with the non-exchanged test datasets. At the same time, the two models are compared intuitively.