## Newspaper Route Franchise

## OCT 22, 2019

**1. Introduction**

My brother wants to buy paper routes franchise in North America. The newspaper company pays by the number of neighborhood in each borough. The two major cities that has a franchise availability are New York and Toronto. My brother wants to explore and compare both cities to figure out which city will bring more cash flow.

We need the neighborhood data for Toronto and New York

We have to create data frame for neighborhoods

We have to create folium maps and scatter plots for comparison

We have to find the number of neighborhoods and compare

We don't know what insights we may uncover. We will dive deep into the location data and data from Foursquare.

**2. Data Description**

**2.1 New York Location Dataset**

The data is taken from https://cocl.us/new\_york\_dataset. The original format of this dataset is a GeoJson file (newyork\_data.json), which is then converted to a csv.

The CSV File has the following features:

Borough

Neighborhood

Latitude

Longitude

There are a total of 307 records available in this dataset. However cleaning needs to be done.

**2.2 Toronto Location Dataset**

The dataset is scraped from https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M. The dataset is then cleaned and saved as CSV file. We are using Jupyter labs for scraping the notebook.  
This notebook will be updated with Markdown for easy understanding. The final CSV File will have the following features:

Postal Code

Borough

Neighborhood

Latitude

Longitude

The above mentioned datasets are used in conjunction with the Foursquare API.

**3. Methodology**

This notebook follows the following Methodology:

Collecting the data

Cleaning the data

Visualization

Machine Learning

**3.1 Collecting Data**

The datasets used in this notebook are collected from two main sources namely, Wikipedia and https://cocl.us/new\_york\_dataset. Please refer to the Data Description section for more information about the data.  
The Foursquare data, will be added within the notebook and no external file will be created for it. The API calls will be made within the code cell. Once the necessary data is collected, cleaning the data is done.

**3.2 Cleaning Data**

This step consumes a lot of time. The cleaning of data in this notebook is done in two phases, one for the Toronto Data and other for the New York City Data.

**3.3 Visualization**

A few simple visualizations are made in order to get a better understanding of the data. Folium is used for this purpose. Folium generates interactive and beautiful maps, using the Latitude and Longitude we provide. The folium visualizations will be made available in the comments section.  
Further to enhance the understanding of the data, count plots are used to better understand the final result.

**3.4 Machine Learning**

KMeans clustering algorithm is used on the most common data frame to identify the probable clusters for different types of boroughs. In both the datasets, we have initialized 5 clusters.

**4. Results**

New York has 306 postal codes in the available franchise area while as Toronto has 103 postal codes. We can conclude that New York dominates Toronto. In Brooklyn, 1st Most Common Restaurant have American Restaurant as top followed by Italian and Caribbean Restaurant. For 2nd Most Common Restaurant, we have rather a different picture. It is topped by Chinese followed by American, Falafel and Vietnamese Restaurants.

**5. Discussions**

There is a small discrepancy in the datasets. There are some entries in the dataset that belong to the condo buildings which means we have more vertical numbers than horizontal numbers which could be both good and bad. We may be right or we may be wrong. There are a few methods by which we can solve this discrepancy. One such method is to consider this a majority in that area. The problem with this method is we might not always be correct. Another concrete way of solving this is rather tiresome approach, which is not feasible. We look for the corresponding coordinates and use google to find what type of structure is that.

**6. Conclusion**

Hence, this notebook serves as a reference to identify the comparison between Toronto and New York City. This notebook can be extended to include not only the data related to postal codes but also other aspects like restaurants, entertainment venues, etc.

Data Description

New York Location Dataset

The data is taken from [NYU Spatial Data Repository](https://geo.nyu.edu/catalog/nyu_2451_34572). The original format of this dataset is a GeoJson file, which is then converted to a csv. The raw data can be found [here](https://github.com/KrishnaChaitanya1/Coursera_Capstone/blob/master/Final%20Capstone%20Project/Data%20Files/nyc_geojson.json). The CSV File has the following features:

Borough

Neighborhood

Latitude

Longitude

There are a total of 307 records available in this dataset. However cleaning needs to be done. The cleaned dataset is found [here](https://github.com/KrishnaChaitanya1/Coursera_Capstone/blob/master/Final%20Capstone%20Project/Data%20Files/New%20York%20City%20Boroughs.csv).

Toronto Location Dataset

The dataset is scraped from [this](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) link. The dataset is then cleaned and saved as CSV file. The notebook for scraping is available [here](https://github.com/KrishnaChaitanya1/Coursera_Capstone/blob/master/Final%20Capstone%20Project/Scraping%20for%20Toronto%20Postal%20Codes.ipynb).  
This notebook will be updated with Markdown for easy understanding. The [final CSV File](https://github.com/KrishnaChaitanya1/Coursera_Capstone/blob/master/Final%20Capstone%20Project/Data%20Files/Toronto%20Postal%20Codes.csv) will have the following features:

Postal Code

Borough

Neighborhood

Latitude

Longitude

The above mentioned datasets are used in conjunction with the [Foursquare API](https://foursquare.com/).