**Introduction**

Through this exercise, a study on neighbourhoods of Toronto and New York (particularly Manhattan) has been carried out. Foursquare API is a great tool to explore around any geographical location on the planet and thus locations have been fed into the API and each location has been further explored.

Further, through Folium the neighbourhoods have been displayed on the map. The neighbourhoods of each city have been clustered using K-Means algorithm by using venues in each neighbourhood as the features. It has been shown that Toronto and New York share 193 common venues, which represents a large correlation between these two cities. Further, population density data between these 2 cities has also been compared to understand the relationship between cities better.

This can be an exciting avenue for people choosing a particular neighbourhood to start a business or for tourists looking to explore areas in the city based on their interest. For people interested in any kind of cuisine can also use the methods in this study to arrive to similar results.

**Data Sources**

1. Toronto Neighbourhoods data:
   * https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M
2. New York Neighbourhoods data:
   * https://cocl.us/new\_york\_dataset
3. Population Density Data:
   * https://en.wikipedia.org/wiki/Boroughs\_of\_New\_York\_City
4. Population density data for Toronto:
   * https://en.wikipedia.org/wiki/Demographics\_of\_Toronto\_neighbourhoods

**Methodology**

Analytic Approach:

In order to study the neighbourhoods of these two cities, common venues between Toronto and New York will be extracted. A large number of common venues will suggest similarity between these two cities. Further income and population density of these two cities will be compared in order to highlight similarities and differences.

Data Requirements:

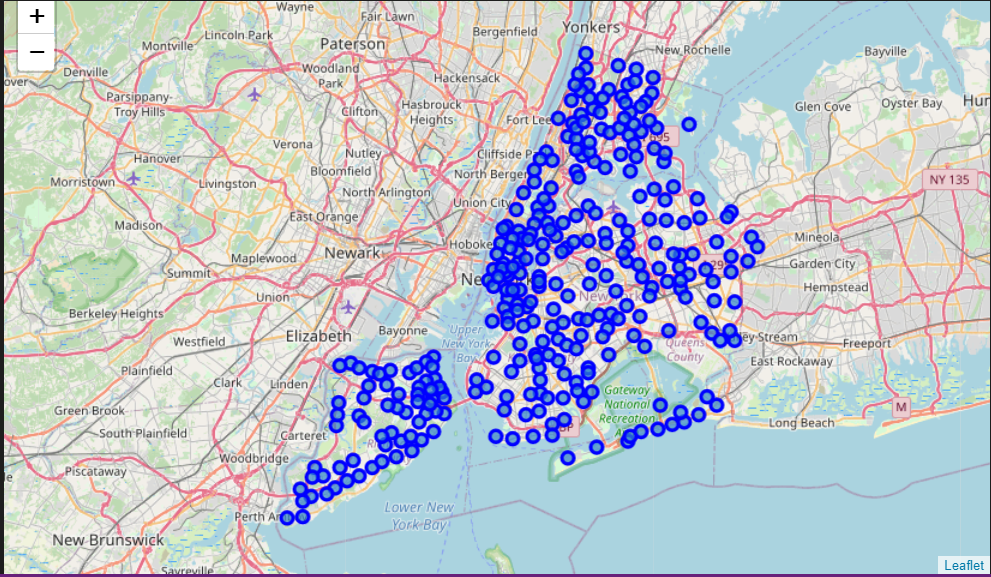
A list of all neighbourhoods in Toronto has been extracted through reference [1] and for Manhattan, New York has been extracted through reference [2]. Further, venues for each neighbourhood are extracted through Foursquare API. Web-scraping has been used for extracting data on income and population density.

Data Understanding and Preparation:

Missing entries of neighbourhood data or missing borough data have been removed. Dummy variables have been created for each associated venue for neighbourhoods.

Modelling:

Through web-scraping and using folium visualization tool, neighbourhoods in Toronto and Manhattan have been obtained on a map [see figure 1 and figure 2].



**Figure 1: Neighbourhoods of New York City**



**Fig 2: Neighbourhoods in Toronto**

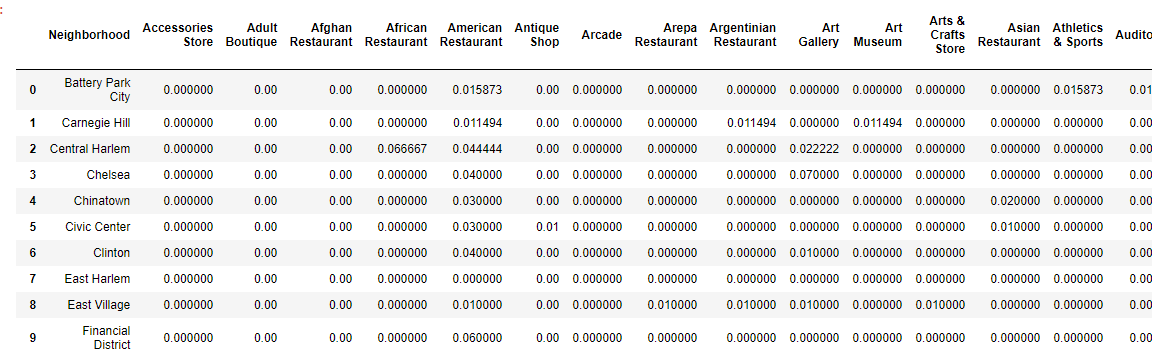
Clusters have been recognized in Manhattan which are similar to each other using K-Means Clustering algorithm. K-Means Clustering algorithm is an unsupervised learning algorithm which tries to assemble examples into k different clusters based on the co-ordinate proximity of the example w.r.t. the centroid (mean position) of different clusters. The algorithm initializes the centroid of each cluster at random positions and for each example the allocation is made to one of the clusters. With the addition of a new element to the cluster, the mean position/centroid is re-calculated and the algorithm is repeated on the next example.

The cost function (also referred to as the distortion function) is defined as the sum of the square distance between each example x(i) and the cluster centroid c(i) of the cluster to which the example has been assigned.

Although the cost function is a non-convex function, and is thus susceptible to finding local minima using gradient descent but one can run k-means clustering algorithm several times with different initialization centroids to choose the one with the lowest cost.

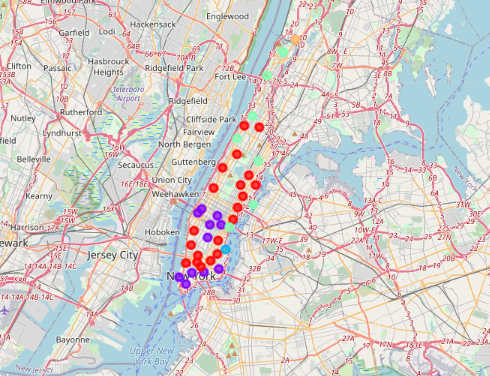
**Results**

For clustering of neighbourhoods, venues in a neighbourhood have been used as features. In order to do this, ‘Nominatim’ library from Geopy has been used to generate a list of latitude and longitude values from a list of names of neighbourhood. Venues have been obtained using Foursquare API by feeding in latitudes and longitudes of neighbourhoods. Dummy variable has been used to give a weight to each venue based on number of such venues within a neighbourhood [shown in Figure 3].



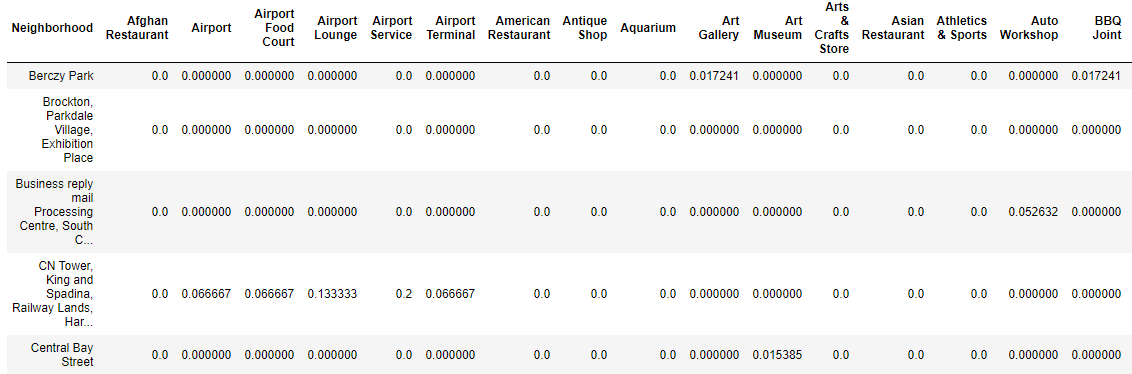
**Figure 3: List of Venues for each neighbourhood in Manhattan**

Since now the features for each neighbourhood have numeric values, it is possible to compute clustering for these neighbourhoods. Using K-Means algorithm, neighbourhoods in Manhattan have been clustered into 5 different clusters. With different colour-coding for each cluster, a map of clustered neighbourhoods in Manhattan can be seen in figure 4.



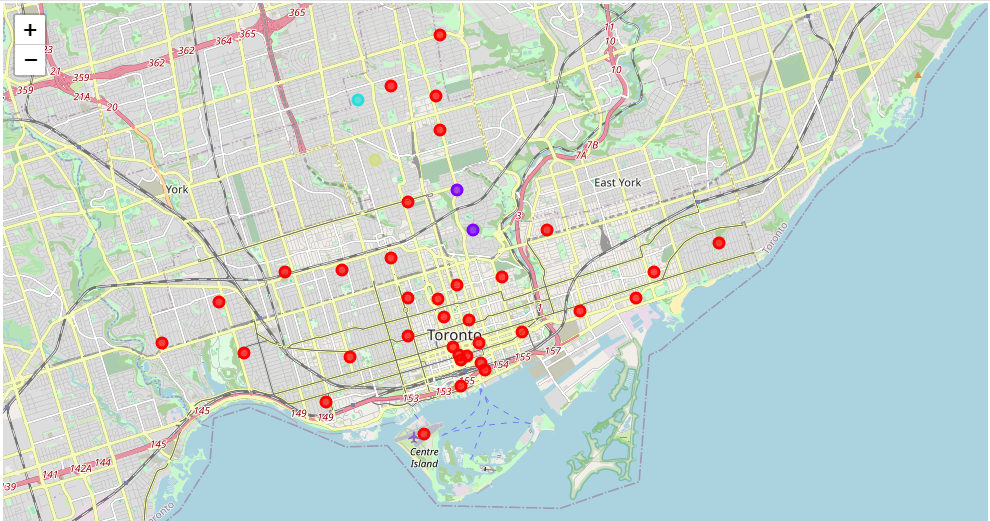
**Fig 4: K-Means Clustering of Neighbourhoods in Manhattan, New York**

Using same approach as for Manhattan, locations of neighbourhoods in Toronto are used to obtain a list of venues in each neighbourhood. Based on number of venues in any neighbourhood, a weighted numeric value is given to each venue type within a neighbourhood.



**Fig 5: List of Venues in each neighbourhood on Toronto**

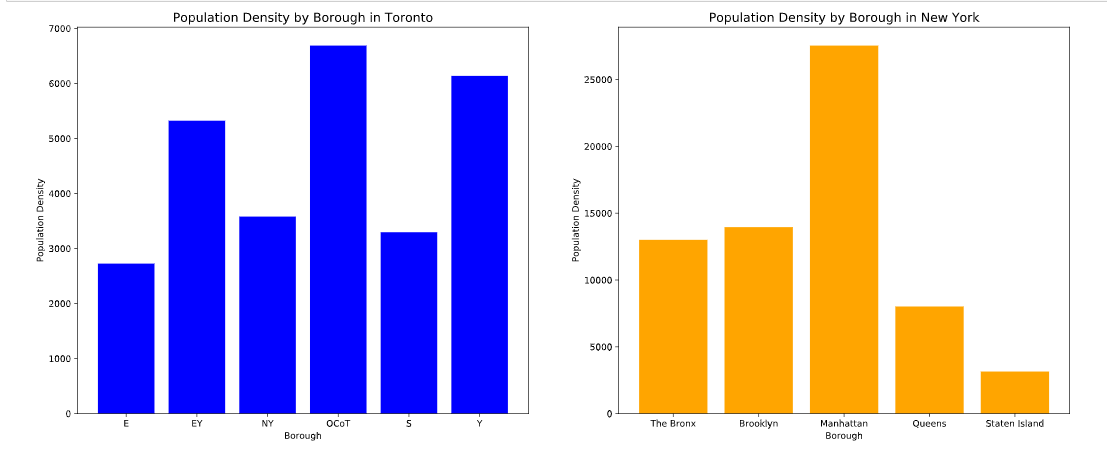
Two neighbourhoods can be similar to each other if they have similar weights for venues. Using this approach, k-means algorithm is used to cluster these neighbourhoods into 4 different clusters. The clustered map can be observed in figure 6.



**Fig 6: K-Means Clustering of Neighbourhoods in Toronto**

Also, there are 193 common venues between Toronto (total 234 venues) and Manhattan, New York (total 327 venues). It can be said that while New York is more diverse in terms of venues but these two cities have similar character when it comes to locations to explore.

A comparison of population density of these two cities, however, exhibits a difference between these two cities as New York has almost twice the population density of Toronto [data taken from reference 3 and reference 4].



**Fig 7: Population density comparison between New York and Toronto**

**Conclusion**

In this Capstone project, geospatial data has been studied using several python libraries such as Nominatim, Geopy, Folium etc. to obtain top venues for two different cities in North America viz. Toronto and New York City. With the use of data refinery techniques, web-scraped data has been neatly imported into Pandas dataframes. Categorical features were converted into dummy numerics for clustering by use of python libraries.

Clustering of neighbourhoods based on venues was conducted using K-Means algorithm for each city. While the neighbourhoods of Manhattan in New York were clustered into five different clusters, Toronto neighbourhoods were clustered into four different clusters. For clustering, dummy weights of types of venues were used to calculate similarity and dissimilarity between neighbourhoods.

Further, the two cities were also compared to each other based on the commonality of the venues and their population density data. It was observed that out of 234 venues for Toronto, 193 venues were common with those of Manhattan, New York. This exhibits a large degree of similarity between these two cities.