**Toronto or Newyork?????????**

How one city is compare to another.. Can we do a comparative study between two different cities to see if there is a business exists in one city and what would be the ideal location to open the similar business in the other city.. Can we make this exercise more generic and unify the categories into specific buckets, so that the data can be used for any kind of business.. How about putting the categories into the following buckets for both the cities, so that Kmeans clustering can do a better job in putting similar categories into same bucket...

**Categories which are interested in this study as follows**

* Bar
* Business
* Entertainment
* Fast Food
* Kids
* Parks
* Residential
* Restaurant
* Indian Restaurant
* Services
* Shops
* SightSeeing
* Sports
* Travel

**Data to be used :-**

Newyork data will be downloaded from the following site and cleanedup for this project.

https://cocl.us/new\_york\_dataset

For Toronto, web scrapping will be done to extract the data from the following site

https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

**Methodology:-**

Once we have the data available, the following approach will be used to solve the problem

\* New york data will be first used and cleaned up to get only the columns of our interest and use the foursquare API to access the 100 items within 500 meters and later unify into the categories mentioned above and in the end, sort the list of boroughs with top 10 categories.

\* Similarly do the same assessment for Toronto data and once we have the same kind of data available for Toronto, both the tables will the put into one single table to start the K-Means analysis and analyze the K-Means and put into 2 different categories.

\* Based on the outcome of the clustering, conclusion can be arrived on which cities look similar between Toronto and Newyork.

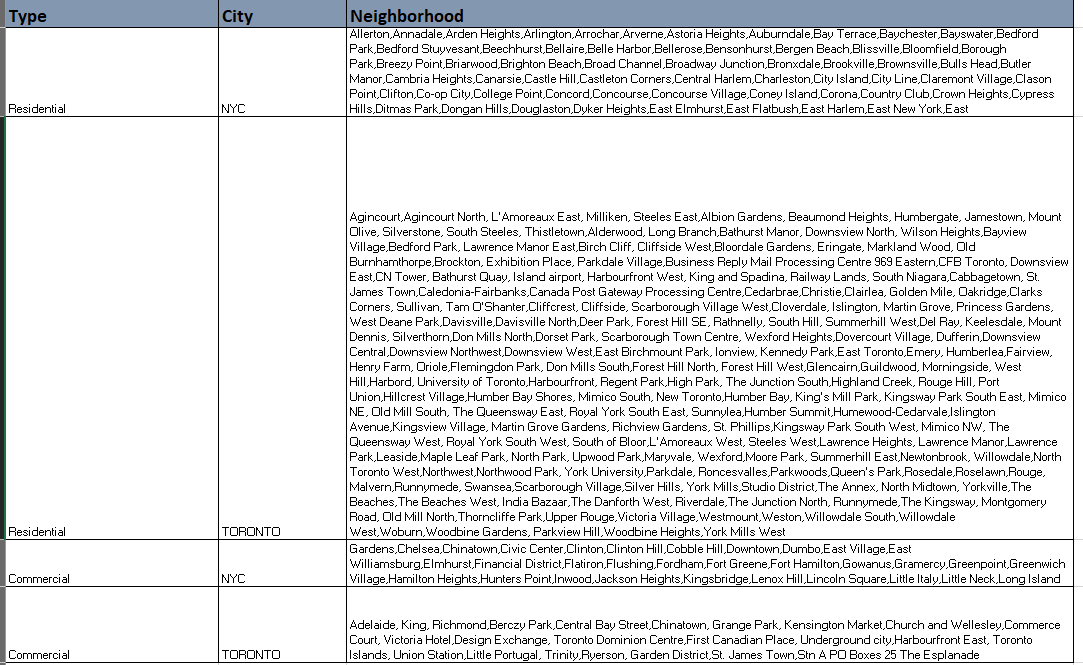
**Results:-**

The main idea of this project to categorize the cities based on the surrounding activities. Since the data is unified for both Toronto and NYC, it become easy to use Kmeans to cluster these into two different groups

Group – 0 – Residential

Group -1 – Commercial

**The following is the results based on the K-Means clustering**



**Discussion:-**

Only K-Means with 2 clusters used in this analysis and the dataset is limited to 100 venues due to free license of Foursquare. This project can further be expanded to multiple analysis like which is the best place to open Indian restaurant or which is the best place to reside with more kids activities etc.. These are beyond the scope of this exercise, but can be expanded based on the data arrived.

**Conclusion:-**

By looking at the data, NYC seems to be more commercial than Toronto Neighborhood as you can see multiple neighborhoods in Toronto are listed as Residential whereas NYC has got very few in this category which is true in nature. The objective of this exercise is achieved by putting these neighborhoods in similar buckets, so that further analysis can be carried out based on this data as mentioned in the discussion session.