**Determining similarity between NEW YORK CITY and TORONTO**

**Problem statement**

In this project I am going to determine how NEW YORK CITY and TORONTO are similar or dissimilar by comparing the most common venue categories in their neighbourhoods.

**Background**

New York and Toronto, two major cities of the world. Lots of businesses and companies have their head offices in and around these cities, which urges lots of people from around the world from different professional background to come and settle here. Additionally, many people have to move among these two cities too, i.e., New York to Toronto or vice versa for job requirements or business needs. Suppose if a resident of New York has to permanently move to Toronto and he/she is searching for a place to live there. That person would like to move to similar area as their current area in New York based on nearby venues including coffee shops or schools or parks or types of restaurants, etc. And also for new comers as both cities are very diverse and are the financial capitals of their respective countries it will be helpful for someone wanting to choose between these two places for having an ideal place to live where he or she can compare things like community culture, good educational institution, hangout places etc.

**Data Acquisition**

URLs used to collect the neighbourhood data for New York city and Toronto are respectively(<https://cocl.us/new_york_dataset>)and(<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>). The Toronto data needed web scraping to get the required data frame consisting boroughs and neighbourhoods. Here **Foursquare API**  will be used to get the location data to explore neighbourhoods in New York City and Toronto

* Firstly the data will be collected for all the neighbourhood areas in NEW YORK AND TORONTO with the help of the URLs shared with us in this course .
* Then the latitude and longitude data for all those neighbourhoods will be collected , here also we have access to those data in this course.

Libraries used in this project: I will use the **k-means clustering algorithm** to complete this task. Finally, I will use the **Folium library** to visualize the neighbourhoods in New York City and Toronto city and their emerging clusters. Along with this other libraries like **Sklearn, Pandas, Numpy, Matplotlib** will be used for the analysis, clustering algorithm , easy data manipulation and data visualization work.

**Data Cleaning**

For simplification in analysis the New York data is converted to only Manhattan data by extracting only Manhattan boroughs from this dataframe. The Toronto data cleaning needed more consideration as there were NULL datas in the table scraped from the web given in above. Those rows with NULL were dropped, same postal code neighborhoods were grouped in one row and then postal code column was dropped. Then these two city’s dataframes were concatenated and the **Foursquare API** calls were made to get top 100 most important venues of each neighbourhood. The venues were grouped by their categories and finally for each neighbourhood top 10 most common venue categories were selected and the final refined cleaned and ready for future analysis dataset is formed.

**Methodology**

In this project for finding similar neighbourhoods between New York and Toronto, K means clustering algorithm is used.

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster is a collection of data points aggregated together because of certain similarities and k is the number of centroids we need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the error. In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

Here in total 8 clusters were taken for analysis and using pyhton’s sklearn package k means algorithm was done on dataset containg features about neighbourhoods in New York and Toronto city.

**Results and Discussion**

After the completion of the analysis part, data visualization packages like folium was used to have a visual understanding of the different clusters in the neighbourhoods of New York and Toronto. It was observed that all New York / Manhattan neighbourhoods were clustered into only one group, although by increasing the number of clusters in the Kmeans algorithm we will find different clusters in the New York neighbourhoods too. Here for K=8, cluster 6 includes all the neighbourhoods in New York and some neighbourhoods from Toronto. The 1st common venues in this cluster 6 neighbourhoods are cafes, fine dining continental restaurants, grocery stores etc. **So if someone wanting to switch places between New York and Toronto ,they should look for places in the neighbourhoods from cluster 6 to get a similar feeling with their current location after shifting.**  
And also people who are searching for some kind of events in these two city's, they can narrow down their search area by looking at these 8 different clusters depending on their interest. If they are looking for a Pizza place, they should look for neighbourhoods in cluster 1. Cluster 2 neighbourhoods have mostly breakfast places, bakery shops etc. For drugstore , women's store cluster 3 neighbourhoods are ideal. Cluster 4, cluster 5 and cluster 7 have very less number of neighbourhoods , so these places are outliers in clustering distribution. Cluster 6 includes the most happening places in these two cities , here one can find amazing nightlives with excellent Chinese, Italian, Mexican restaurants and bars, nice coffee shops, supermarkets etc. Cluster 8 mainly group places with playground, many kind of sports arena ,parks etc.

**Conclusion**

The objective of this project was to find similar neighbourhoods in New York and Toronto by applying K-means clustering algorithm in their neighbourhoods and grouping then based on some discriminating category. From the visualisation of the result we can see that we weren't able to cluster places within New York, this is because of very large number of input data and less number of the assumed K value in this algorithm. Although we were able to find the most similar places in these two cities, that is neighbourhoods from cluster 6.