**Business Problem:**

* Compare the neighbourhoods in Toronto & New York city.
* Given that these cities are the financial capitals of the respective countries, how much do they differ in terms of neighbourhood clusters.

**Data for addressing the problem:**

* ***Four square API*** is used for location services that can provide information regarding venues around each neighbourhood. <https://developer.foursquare.com/>.
* Data with postal codes in Canada is pulled from Wikipedia using requests module
* The data is massaged & filtered down to "Central Toronto" before applying the machine learning technique.
* New York data is downloaded in the form of json file (provided in the lab)
* The latitude & longitude information for each neighbourhood is derived & transformed into a dataframe.
* Foursquare API calls are made to retrieve ‘***venue categories’*** within each neighbourhood.

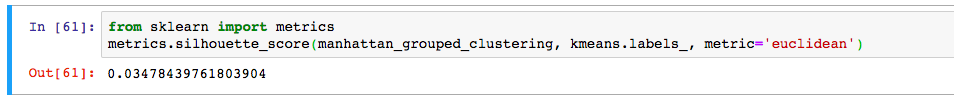
**Methodology:**

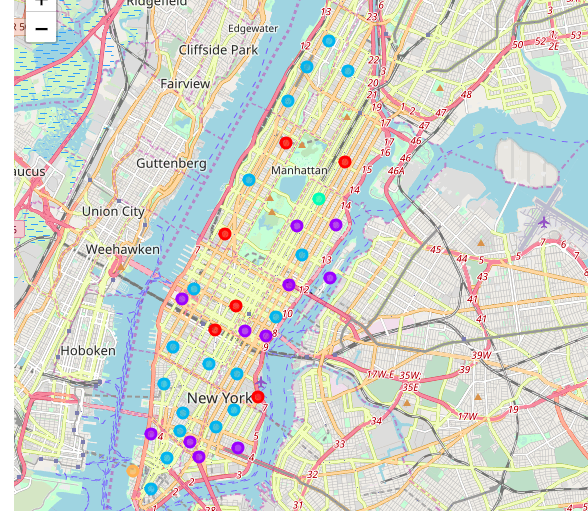
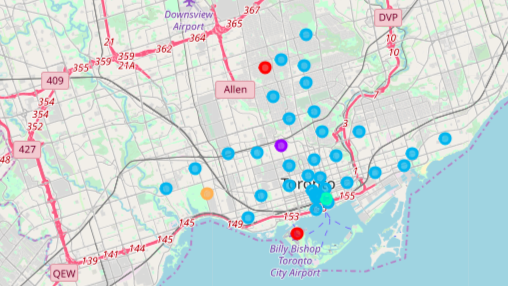
* Top 5 venues in each neighbourhood are analysed through defined functions.
* The Top 10 most common venues for each of the venue are analysed.
* By grouping the dataframe based on ‘neighbourhood’, the dataset is categorized into 5 clusters using sklearn’s Kmeans model.
* Since we do not know the ‘ground truth labels’, we will use Silhouette metric for comparing the cluster results.
* The clusters generated for New York & Toronto dataset are analysed using ‘silhouette’ metric.
* The Silhouette metric provides a score between -1 to 1 with -1 representing incorrect clusters & 1 indicative of highly dense clustering.

**Results:**

* The silhouette metric for New York dataset is 0.034 suggesting that the clusters are sparsely dense.
* The metric for Toronto dataset is 0.35 indicating densely populated clusters.





**Discussion:**

The clusters in Toronto neighbourhood are densely populated with well separated clusters.

The clusters in Newyork are sparsely populated with a low silhouette score.

**Conclusion:**

The neighbourhoods in Toronto & New York are dissimilar with varying cluster density & centroids.