**1.Introduction**

**1.1 Business Problem:**

We are going to compare the Neighborhoods of New York, US and Toronto, Canada to see how similar are these two cities. We will generally be exploring the attributes provided to us by the Four-Square API.

**1.2 Interest:**

In this project we will give analysis which would help suggest the best possible & suitable home location for people across various walks of life. Like Bachelors, Families, Old People etc.

This Analysis is very useful for anyone who is trying to analyze the two cities in terms of the diversity of avenues these two cities offer. This is one of the problem we can try to research through this.

However, this Project will provide the generic Comparison Attributes based on Four-Square data, its Users discretion where ever they want to use it.

**1.3 Assumption:**

We are assuming that the neighborhoods which have the maximum number of venues would be the most visited and most popular place to live around for people making them more expensive and favored.

**2. Data acquisition and cleaning**

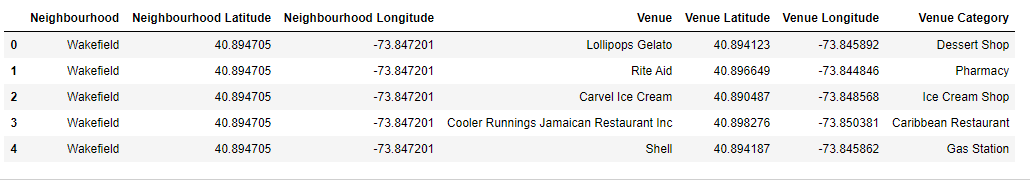
**2.1 Data Sources:**

The Data for the two cities will be taken from a New York Data set [***here***](https://geo.nyu.edu/catalog/nyu_2451_34572)and We will scrapping the Wikipedia Page [https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M,](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) for Toronto and will be combining the Geo-spacial data for Toronto from the link [*here*](http://cocl.us/Geospatial_data).

In addition to having the neighborhoods and geo-spacial data for Toronto and New York, we are going to use data from Four-Square which is going to give us the details of the venues

Some of the Ideas to use the data above is as below:

1. We will take and combine the above data sets and will get the venue categories in and around the neighborhoods of two cities, and we can check the most happening Neighborhood (where there are variety of venues present) with help of #no of venues present.



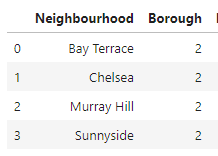
1. Borough can be used as another Grouping criterion to see how densely a Borough is populated with different types of Venues. This is important in case a person has a predominant preference of city and he/she wants to have some data to choose the best place to be so that minimum compromises are made.

**2.2 Data Cleaning**

The data was fetched from various data sources like .json file, CSV file and Toronto data was scrapped from Wikipedia webpage using Beautiful Soup library.

The Data had some anomalies which were needed to be fixed as stated below:

1. The neighborhood data for New York had some cities which were common across Boroughs which could create problems while comparing the neighborhoods, this problem was handled by appending Borough names to the neighborhood for those neighborhoods only. Below is the list of those neighborhoods

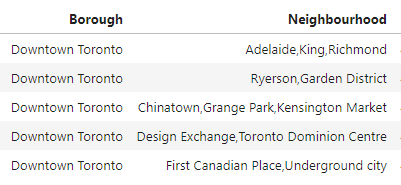


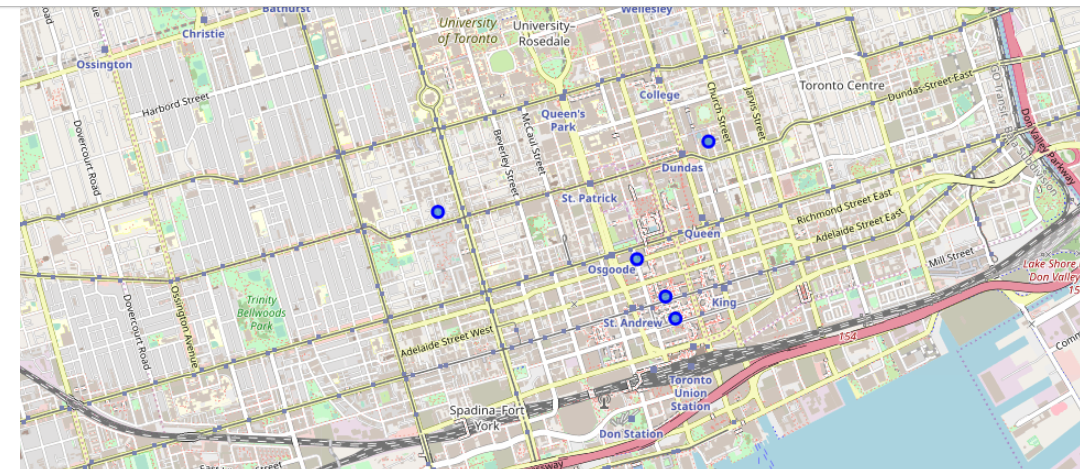
1. Theneighborhood number difference for Toronto and New York was very large, Toronto has some 103 whereas there are 306 neighborhoods for New York. So, we only be comparing just 100 neighborhoods out of both. The selection of these 100 neighborhoods was done according to the maximum number of venues fetched from Foursquare database. The first hundred are sorted out from each. So, we will take first 100 densely populated neighborhoods according to the number of venues available for them.
2. While trying to map first ten most populated neighborhoods in each city, New York had some 35 neighborhoods which had 100 venues each against them, it was due to LIMIT =100 we applied while fetching Data from Four-Square. So, it was difficult to determine which ones are the 10 most popular neighborhoods. To solve this, we extracted these 35 neighborhoods and reran them through Four-Square API to get the number of results for them, this time we removed LIMIT parameter. we got the numbers and then first 10 were sorted out.

**3. Exploratory Data Analysis**

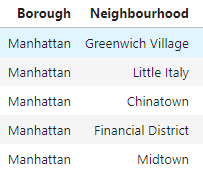
1) We have extracted first 10 neighborhoods for each city and visualized them on maps of Toronto and New York.

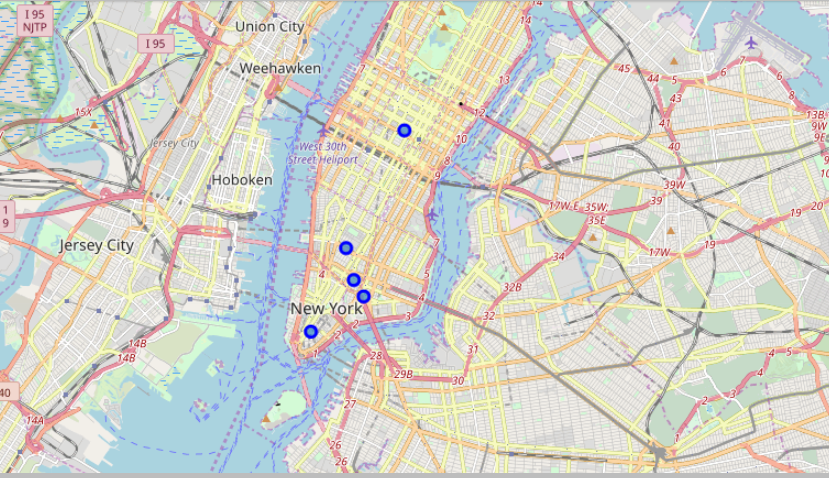
For Toronto, we found that most popular borough is Downtown Toronto. Find the five famous neighborhoods there. These are





For New York, we found that most popular borough is Manhattan. Find the five famous neighborhoods there. These are





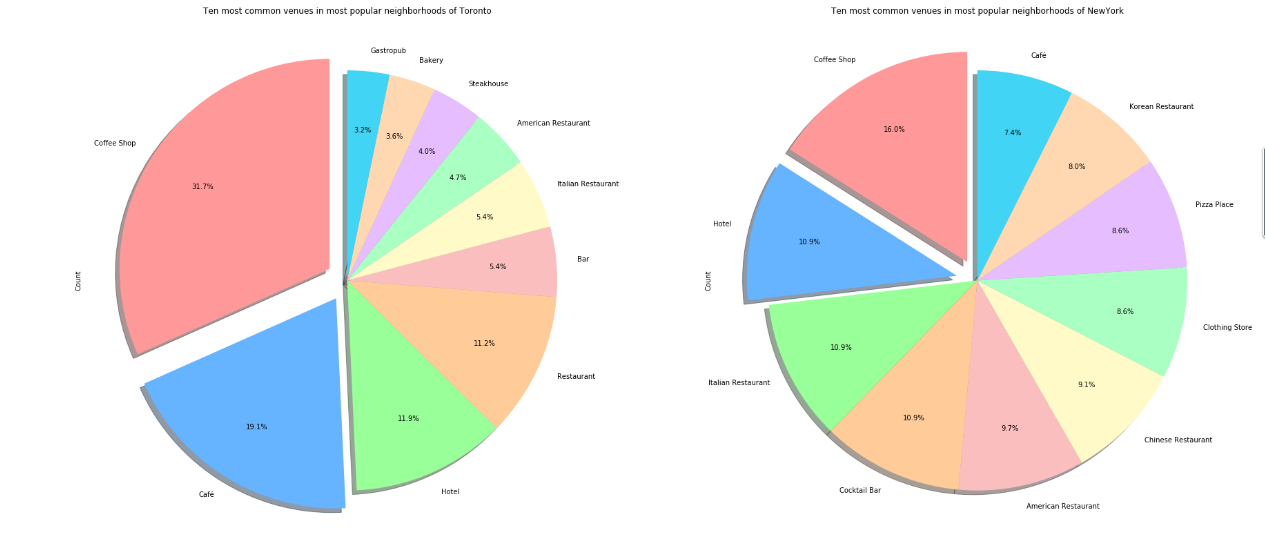
1. New York has some venues which are present in most numbers. Five of them are:



Similarly, Toronto stats are:



And the Chart below shows 10 most densely present venues across 10 most popular neighborhoods in Toronto and New York.



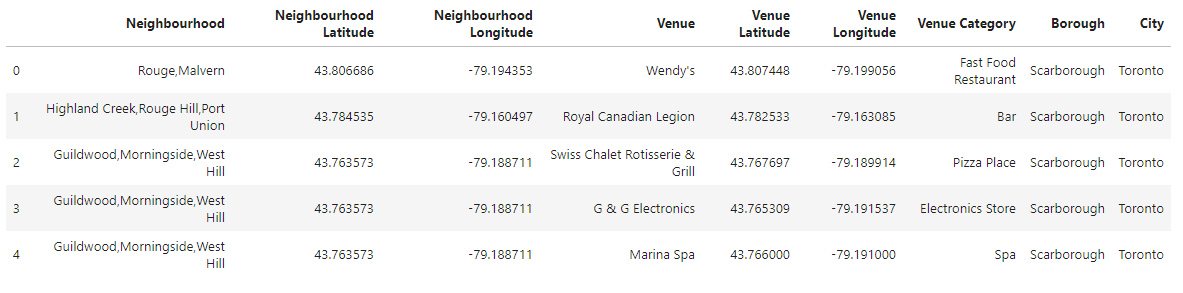
**4. Predictive Modeling**

We will be using classification in our problem to divide the dataset (containing venue data for both Toronto and New York) into Eleven Clusters so that we can have a direct comparison of Boroughs in both the cities, this is done so as to find the Neighborhoods which are common among both the cities and club them together.

This can help the people find similar neighborhoods in Toronto and Canada and find the place which suits them more.

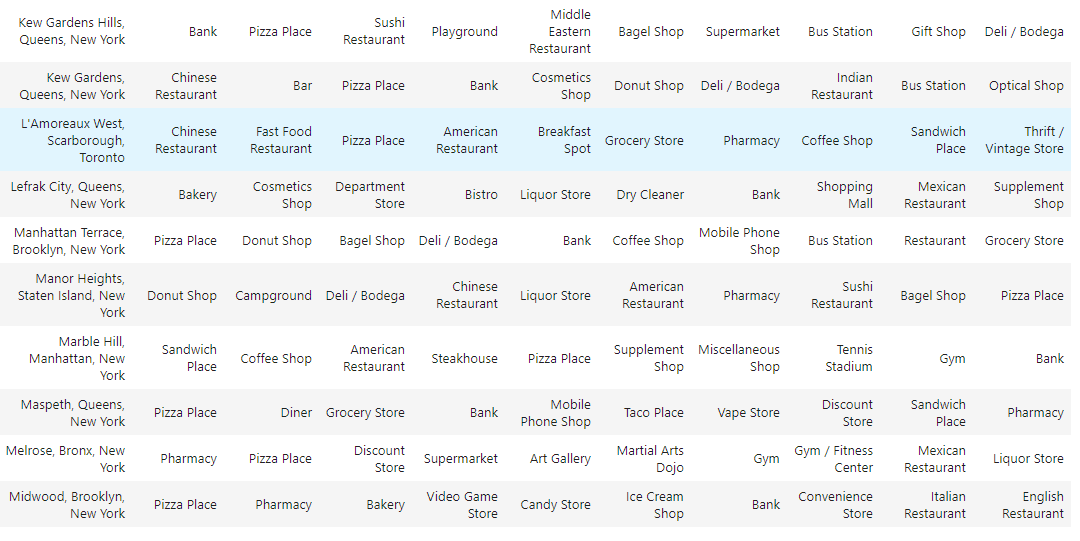
**4.1 Classification using k-means clustering**

**The data** provided to the algorithm had Neighborhoods and venue Category as the different columns and no of venues are marked as data against them. See below sample:



So, basically K- means clustering is grouping them together according to the similarity of Venues and the number of them present in the data set.

Let’s explore Cluster number 1 here, see the sample data below:

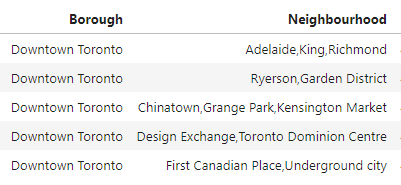


We can see above that **L’Amoreaux West** in Scarborough, Toronto is similar to various neighborhoods of New York like Marbel Hill, Maspeth, Lefrak City.

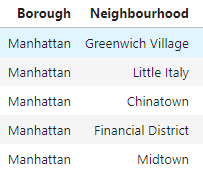
On a different note, if is living in Toronto and wants to move to New York, they can find that Neighborhood in our dataset, and check the neighborhoods for its Cluster label in New York.

**5. Conclusion**

In this study, I have acquired the data for Toronto and New York and with help of Four Square data. I would say that the five most popular neighborhoods in Toronto are as below



And that in New York are:



Manhattan is most venue rich borough in New York and Downtown in Toronto.

We can see that there are a lot of Coffee Shops and Café’s in Toronto, so according to initial analysis coffee business can very rarely go wrong in Toronto (I would not be claiming this; further analysis has to be made)

**6. Future Improvements**

We can improve the above Analysis, if we have data for more neighborhoods in Toronto, the difference in numbers is huge, where dataset has 30 odd number for Toronto where it is staggering 300+ for New York.