**Battle of Neighbourhoods**

**Vidya Sankar.S**

**Jul 31, 2019**

1. **Introduction**
   1. **Background**

Immigration to US and Canada is happening at a very great rate these days. Both the countries support the immigrants in many ways. For anyone from a developing country, settling off in one among these countries is a big dream come true. In today’s world, leaving one’s homeland and settling off in any of the foreign countries is a very common thing.

As part of this project, I am trying to identify the best suited place among two famous cities such as New York and Toronto for a common man to move into. Both the cities have their own pros and cons.

When New York city is well known for its diversity, energy, landmarks, excitement, opportunities and convenience, Toronto is well suited for a much more peaceful life in terms of cost of living, commutation, travel and connectivity to nature.

The purpose of moving into these cities will vary for each and every person. The decision of which city to chose when it comes to a living completely depends upon the purpose.

* 1. **Problem**

The purpose of this project is to help people understand various factors and identify which city would be preferable and what factors to consider when planning to move in.

The internet has tons of raw data already about these two famous cities. However, it would be best when a statistical analysis is performed with the data available. It would be much more better when the findings are portrayed through user friendly charts and plots, so that one can easily understand the facts.

In this project, let us consider a restauranteur from a developing country planning to move into one among these two cities. The factors considered would be one that would help them establish their business over these places.

* 1. **Interest**

Obviously, this would help anyone in future who has such plans of starting off in either of the places.

1. **Data Acquisition and Cleaning**
   1. **Data sources**

I needed data corresponding to the New York city and Toronto to proceed further. If had to collect the Postcodes along with the Borough and Neighbourhood information, for these corresponding cities. I was able to get the data for US from the Kaggle dataset [here](https://cocl.us/new_york_dataset).

For Toronto, I had to scrape the data from the webpage <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>.

* 1. **Data Cleaning**

The NY data downloaded was perfect and required no cleaning.

However, the Toronto data scraped had to be cleaned in order to ensure that only correct data is being used for analysis.

First, there were Postcodes which did not have Borough information. I had to drop those columns as further processing would be possible only if the borough information is available.

Second, few Boroughs did not have neighbourhood values populated. For those cases, I copied the corresponding Borough as the neighbourhood as so would be the case most of the times.

Third, the Toronto data did not have the Latitude and Longitude co-ordinates for the respective Postal codes. The co-ordinates where available in the dataset [here](http://cocl.us/Geospatial_data). This data was fetched and merged to the scraped data.

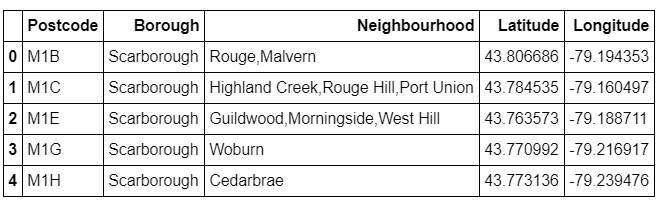
The data for the cities where stored in the below format in respective dataframes.

* 1. **Feature selection**

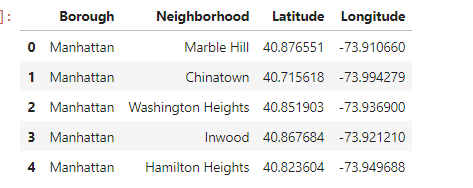
Once the data cleaning was done for the Toronto data, I observed that there were multiple neighbourhoods for the same Post codes. I grouped all those postcodes and joined the corresponding neighbourhoods in the same column. This reduced the size of data from (288,3) to (103,3).

Once the feature selection was done, the data looked in the below format:

**Toronto Data**



**New York Data**



Once data was formatted in the above way for both the cities, next task was to consider the best suited Borough among both the cities which can be used for further analysis.

Factors considered to decide the same were as follows:

|  |
| --- |
| Lifestyle |
| Proximity to tourist attraction |
| Weather |
| Value of money |
| Cost of Living |
| Processes involved in starting a new restaurant |
| Education |

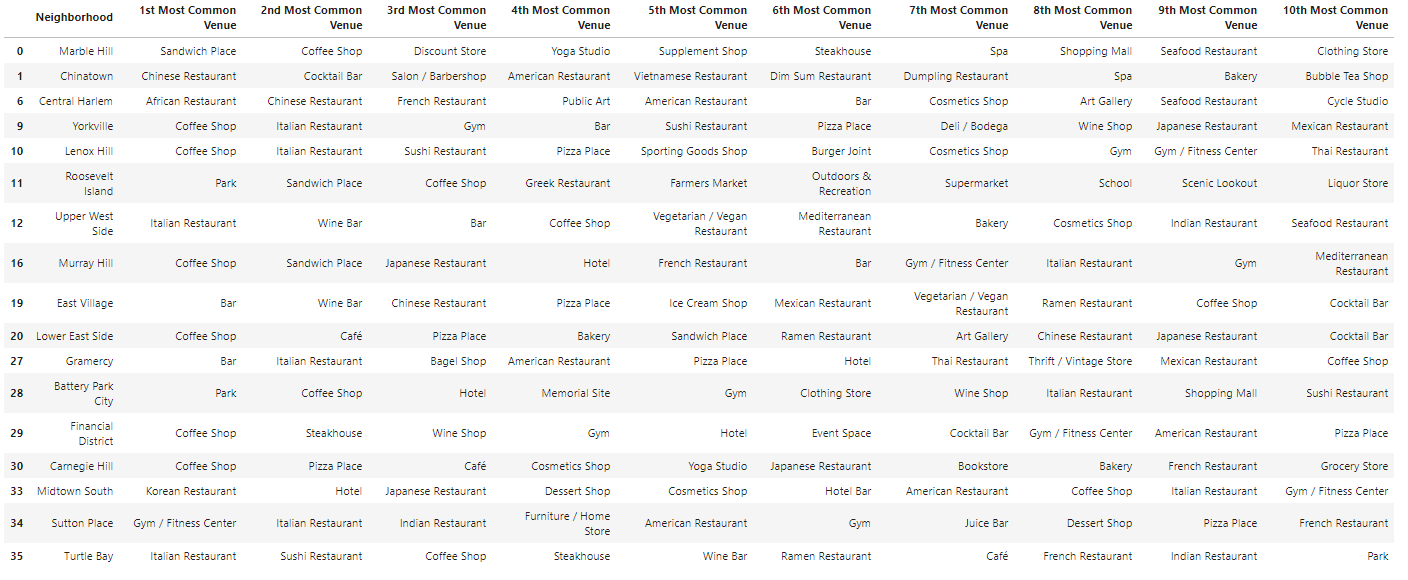
Various web pages were browsed (links can be found in **Appendix I**)to decide on the Boroughs and the following boroughs were considered for the respective cities.

|  |  |
| --- | --- |
| **Borough** | **City** |
| Manhattan | New York |
| Scarborough | Toronto |

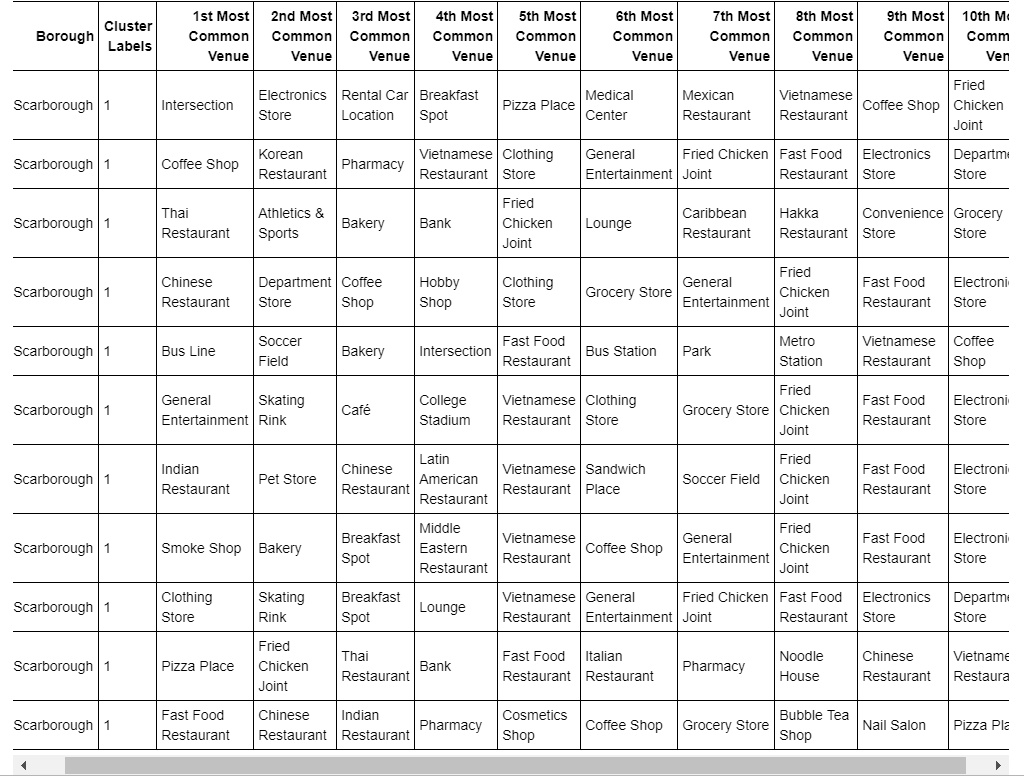
Once the borough was decided, Foursquare API was used to obtain furthermore information of the two boroughs considered.

Explore API was used to collect information and land on the top 10 venues for each borough. The result was optimised using K mean clustering.

**Manhattan Data:**



**Scarborough data:**



These data were further used to perform the ERD and come up with a solution.

# **Appendix I**

<https://www.timeout.com/newyork/things-to-do/reasons-manhattan-is-the-best-borough>

<https://www.quora.com/What-is-Manhattan-NYC-borough-known-for>

<https://en.wikipedia.org/wiki/Scarborough,_Toronto>

<https://moving2canada.com/where-to-live-in-toronto-neighbourhoods/>