## Final Report

## The Battle of Neighborhood :

### **Introduction: A description of the problem and a discussion of the background**

The purpose of this project is to help people to know about the neighbourhood so that to select the neighbourhood having better facilities. It will help people to take smart decision on selecting the better neighbourhood out of numbers of other neighbourhood in North York, Canada. Now a day’s lot’s of people are migrating to various cities of Canada in search of jobs or business, The project is for those people who are looking for better neighbourhood in terms of accessibility, Schools, Market, Theatres, hospitals, opportunities etc. The aim of project is to analyse the features to do the comparative analysis of neighbourhoods. The attributes includes crime rates, housing prices, school’s according to rating, connectivity, weather conditions, hospitals, water resources, waste water treatment facilities. The project will be helpful for people to get aware about the locality before moving to a new city or else to decide the location which is more suitable for him/her.

### Data Attributes: A description of the data and how it will be used to solve the problem.

We are going to use the dataset consisting of postal codes, Borough and Neighborhood of North York which is scrapped from the Wikipedia page mentioned below:  
Data Link: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>,

Most time consuming part of any data science project is collection of data and data cleaning, Some of the data can be gathered with the help of web-scrapping and some of the data can be obtained from the data providers but it adds some cost to the project. By considering the limitations related to the cost, we are using API’s and other data sources which are available for free, as free membership offers limitation of usage, we are limiting the scope of project.

The data gathered from Foursquare API for the North York Borough will be used to cluster the neighborhood according to the similarities by using K-means clustering algorithm, Foursquare provides the location data within the area of interest such as venue name, locations, business etc.

The Data obtained from web-scraping is analyzed and processed and further meta data is added to the original data to get insights of the data. Some of the attributes of the data are:

1. Postal codes
2. Neighborhood’s
3. Neighborhood’s latitude
4. Neighborhood’s longitude
5. Venue
6. Venue’s latitude
7. Venue’s longitude
8. Venue Category
9. Nearby Schools
10. Nearby Hospitals

Some of the future scope of data attributes are mentioned below:

1) Crime rate: The crime rate of locality is a good way to judge about the ease of life in the city, low crime rate indicates good locality and high crime rate indicates safety issues.

2) Housing prices: Median of housing prices can be used to find out the price range of the houses in the neighbourhood for comparison.

3) School’s: This attribute shows the average rating of School’s for the given city in terms of quality, fees and ratings, Highest rating indicates better Schools and educational facilities.

4) Connectivity: It indicates the connectivity of the locality, good connectivity indicates ease of travel.

5) Weather condition: It shows the weather conditions of the cities.

6) Hospitals: It tell about the availability medical facilities

7) Water resources: It is very important to have a good amount of water resources nearby the city for ease of living.

8) Waste water treatment: Poor waste management implies poor life style and poor management.

### **2) Problem which we are trying to solve:**

The main question we are trying to answer is “Can we suggest the better neighbourhood to the person who is about to migrate to North York, Canada by considering various essential facilities ?”

### **Approach:**

The data gathered for this project belongs to unlabelled data i.e. without any target column, the approach is basically is to use unsupervised machine learning algorithm K-Means clustering to compare the similarities of the two cities, K-means can be used to explore the neighbourhood and cluster the cities according to their similarities.

Libraries and API’s needed: Some of the libraries which will be used in this project are mentioned below: 1) Foursquare API: Foursquare offers best results in terms of data gathering. It has a database of millions places and it’s API has the ability to perform location search, business search etc. Foursquares developers credentials are used for this project and due to the limited access to the number of calls we will limit the number of neighbourhood and radius to minimum.

2) Pandas: It will be used to handle the data

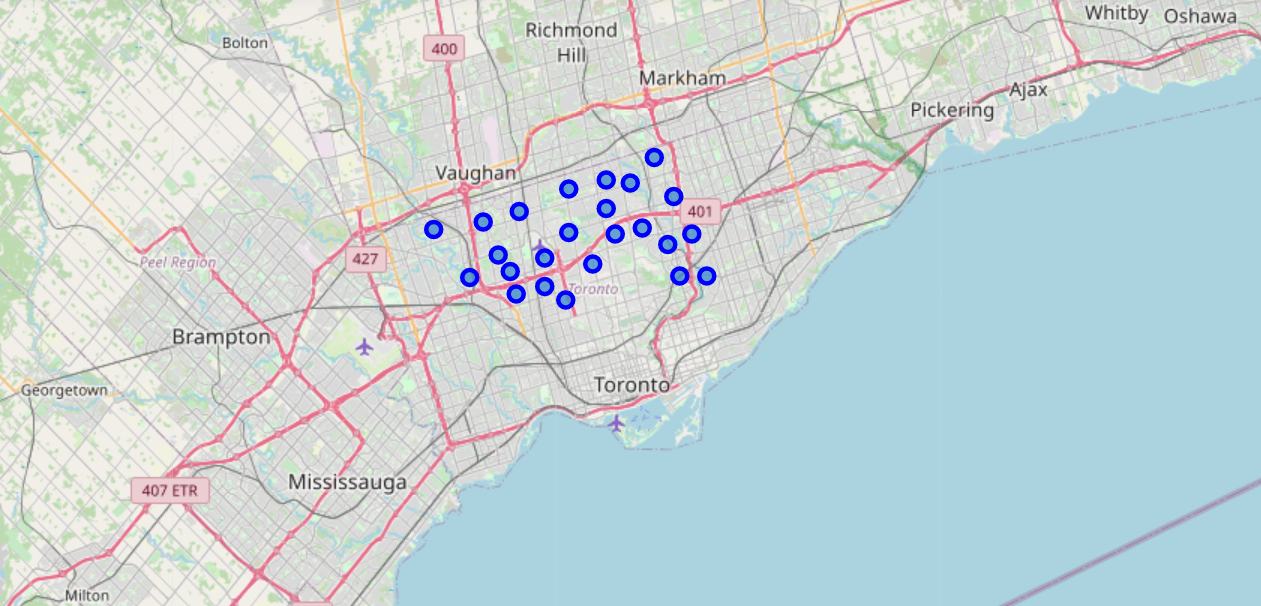
3) Matplotlib: It will be used to plot the graphs and visualise the data

4) Scikit-learn: This machine learning library is used to import k-means clustering algorithm.

5) Folium: This library will be used to visualise the neighbourhoods.

6) Geocoder: It will be used to retrieve the location data.

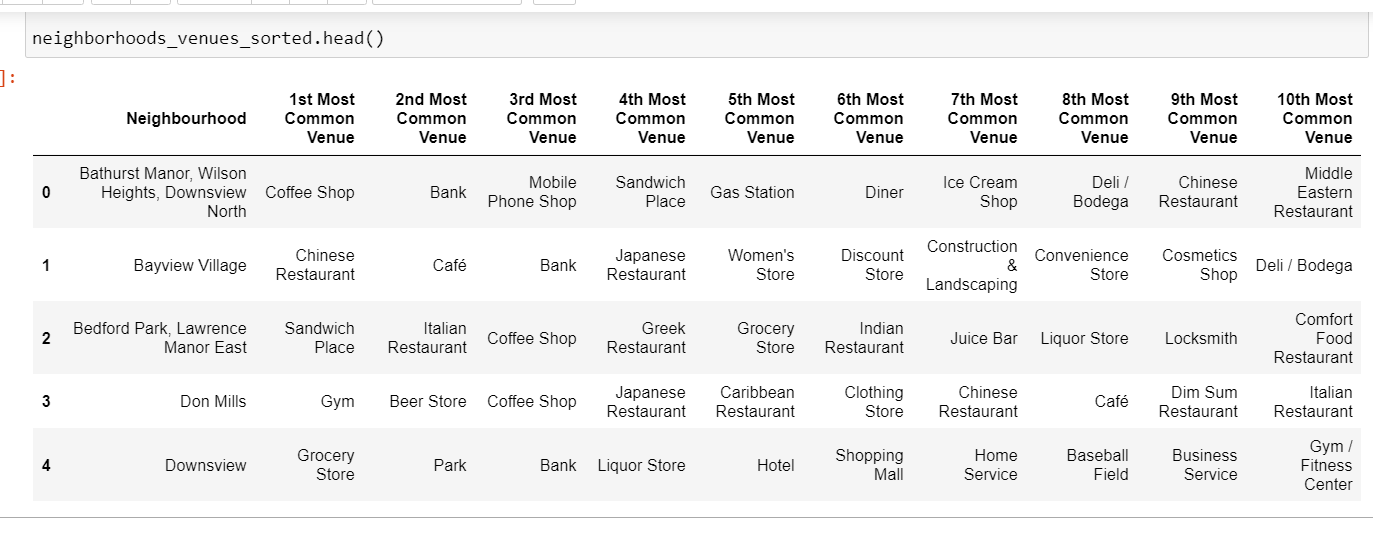
**Map of North-York**



### **3)Methodology Section**

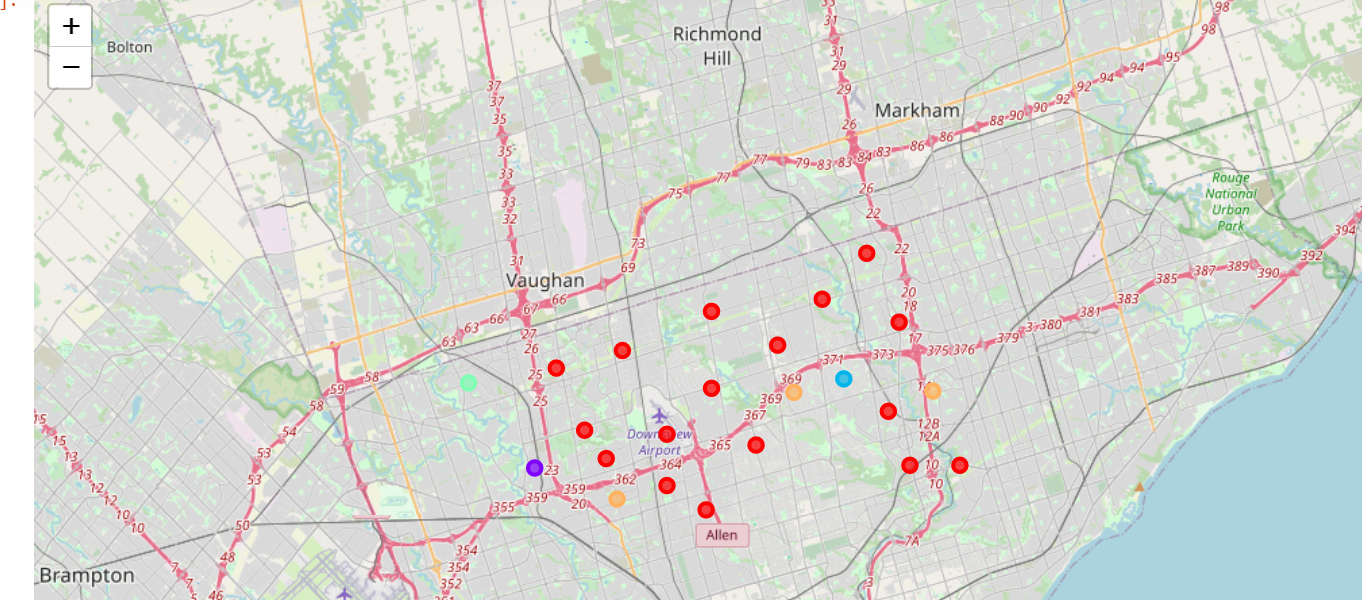
To compare the similarities of two cities, we decided to explore neighborhoods, segment them, and group them into clusters to find similarities. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

**Using K-Means Clustering Approach** | Most Common Venue

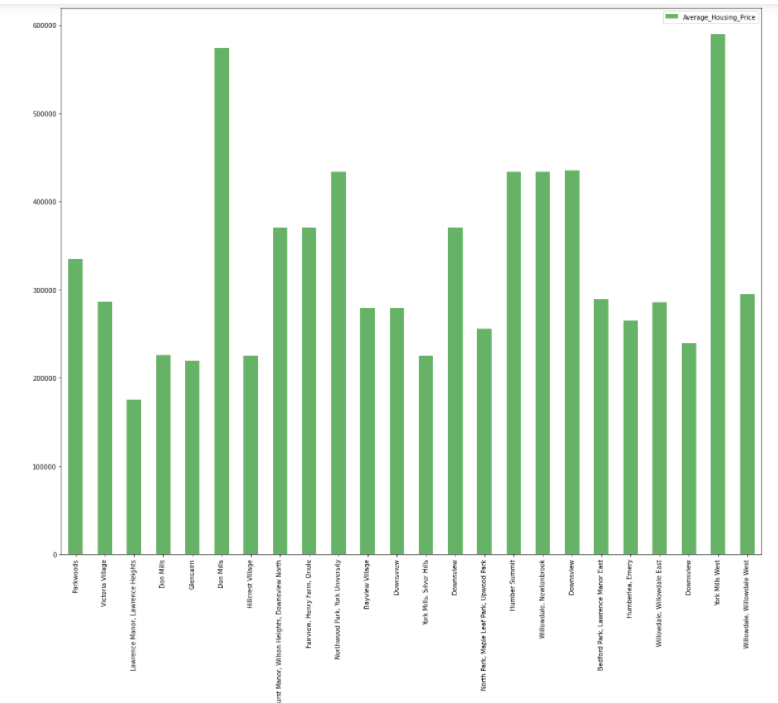


### **4. Results Section**

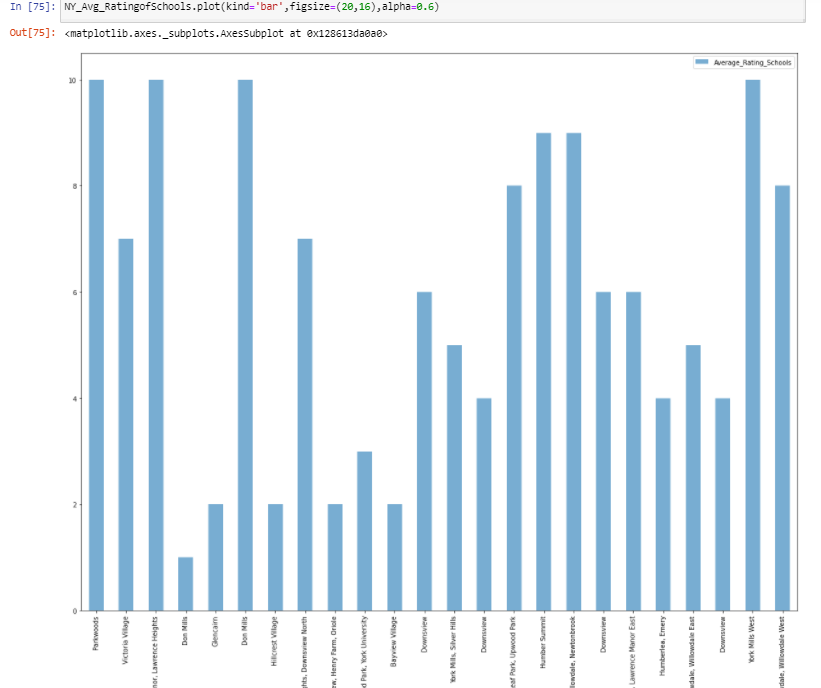
**Map of Clusters for North York**



**Average Housing Price by Clusters in North York**



**School Ratings by Clusters in North York**



### **5) Conclusion Section**

In this Capstone project, using k-means cluster algorithm we separated the neighborhood into five different clusters, which have very-similar neighborhoods around them. Using the charts above results presented to a neighborhood based on average house prices and school rating have been made.

It has been a great journey with the Coursera and IBM, I learned a lot in this course and will try to implement this in real applications.

**…Thank You!!!**