**THE BATTLE OF NEIGHBOURHOODS**

SHREYAS BALIGA

1. **Introduction**

The goal of the project is to create an analysis of the neighbourhoods in Toronto which will aid people to explore facilities in and around their neighbourhood.

This analysis also helps people who are moving into Toronto make an efficient decision on choosing the right neighbourhood to live in based on their personal requirement and needs.

A comparative analysis will be conducted between different neighbourhoods so as to choose the best neighbourhood for the end user.

This Capstone Project aims to create an analysis of features for a people migrating to Scarborough to search a best neighbourhood as a comparative analysis between neighbourhoods. The features include median housing price and better school according to ratings, crime rates of that particular area, road connectivity, weather conditions, good management for emergency, water resources both fresh and waste water and excrement conveyed in sewers and recreational facilities.

It will help people to get awareness of the area and neighbourhood before moving to a new city, state, country or place for their work or to start a new fresh life.

1. **Data cleaning**

Link to the database: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

I have used the same dataset that derived from Wikipedia during Week 3.

Please consider the data in the given link as an example of how the data is represented.

**Foursquare API Data:**

We require data about different venues in different neighbourhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, we have chosen the radius to be 200 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighbourhood
2. Neighbourhood Latitude
3. Neighbourhood Longitude
4. Venue
5. Name of the venue
6. Venue Latitude
7. Venue Longitude
8. Venue Category

This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business. Using credentials of Foursquare API features of near-by places of the neighbourhoods would be mined. Due to http request limitations the number of places per neighbourhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

1. **Exploratory Data Analysis**

Map of Toronto:

The map of Toronto was plotted using the folium library.

This map comprises of all the neighbourhoods in and around the given area.



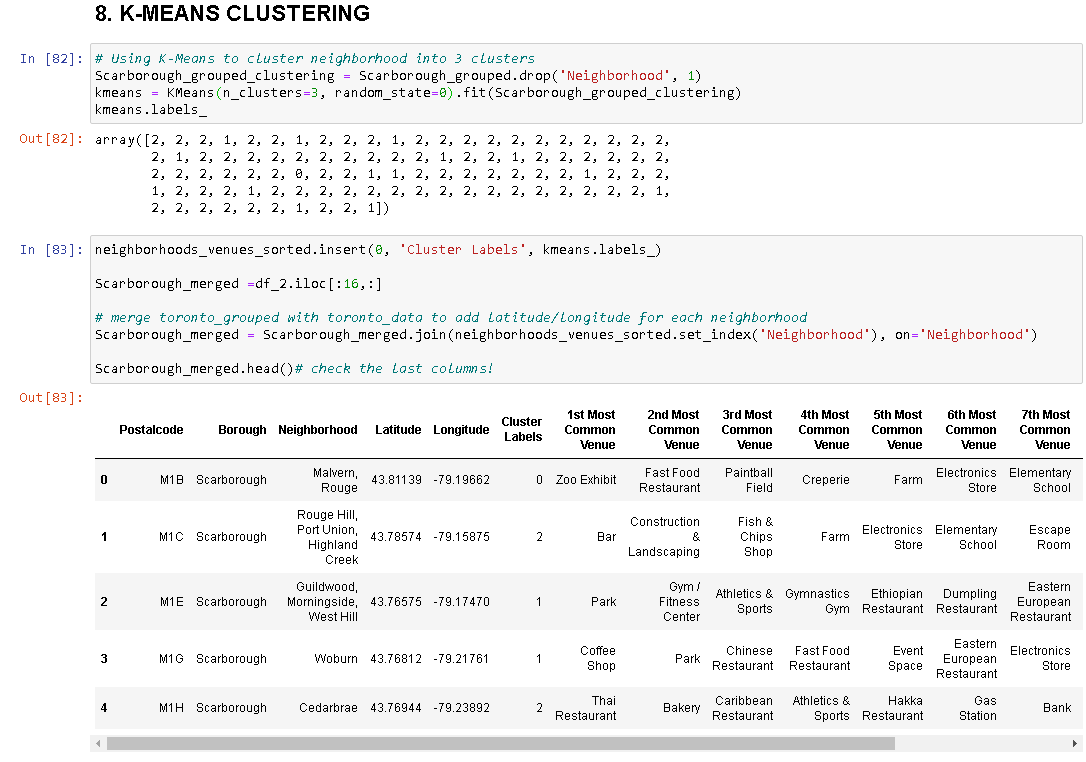
1. **Modelling**

**The clustering model that is employed is k-means clustering.**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

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

**Implementation:**

****

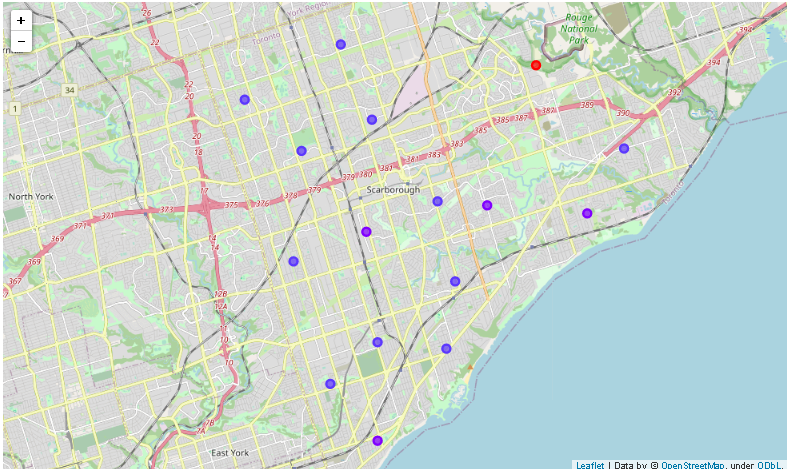
Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

1. **Conclusions:**

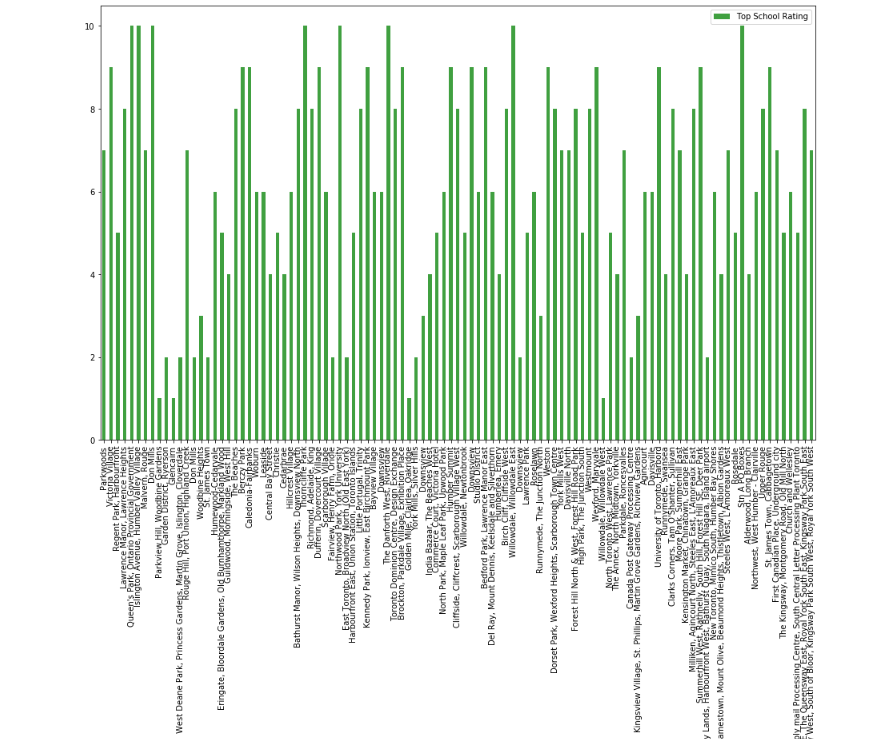
Using k-means cluster algorithm I was able to separate the neighbourhood into 15 different clusters and for 108 different latitude and longitude from dataset, which have very-similar neighbourhoods around them. Using the charts below results presented to a particular neighbourhood based on average house prices and school rating have been made.

I was able to gain a fair amount of knowledge about Folium and other plotting techniques. Applying this knowledge in a practical approach has taught me a lot.

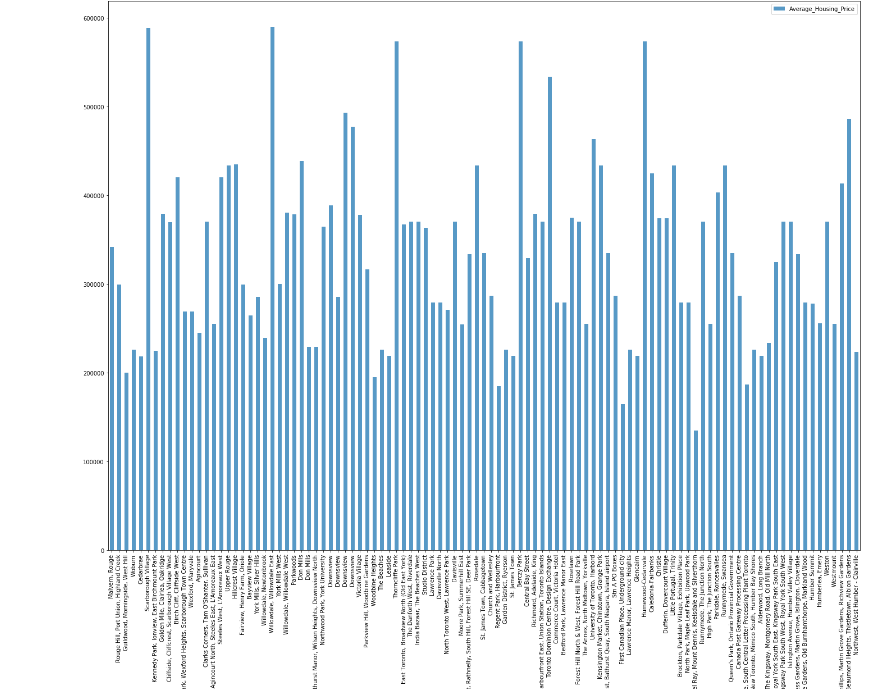
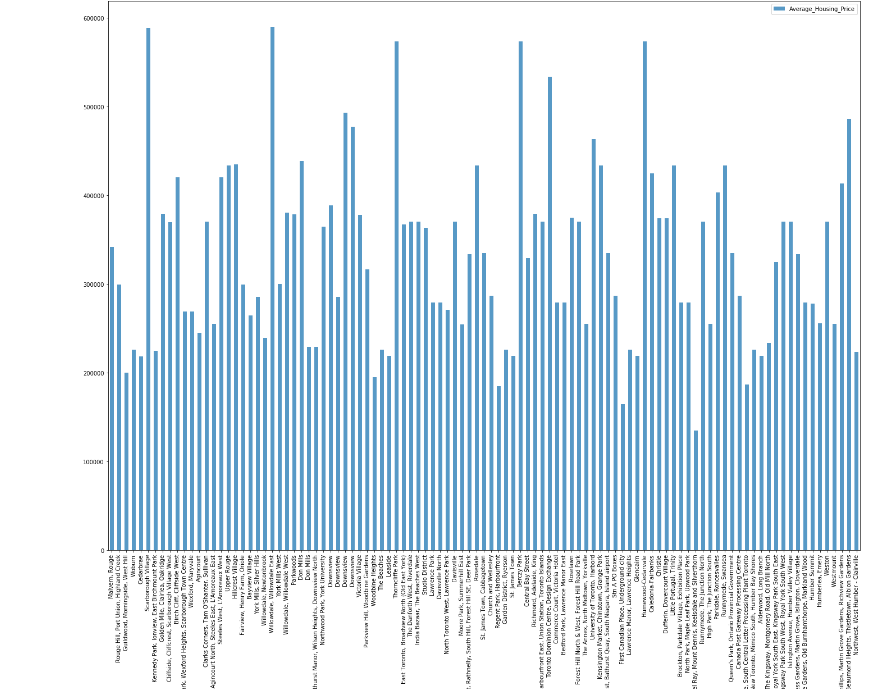
**Map of Clusters in Scarborough**

****

**Average Housing Price by Clusters in Scarborough**

****

**School Ratings by Clusters in Scarborough**

****

1. **Future Scope**

This project can also be implemented in different cities provided there exists a generous dataset.

This project can also be more precise by implementing other various clustering techniques to find the best neighbourhood around Toronto.

Libraries Which are Used to Develop the Project:

Pandas: For creating and manipulating data frames.

Folium: Python visualization library would be used to visualize the neighbourhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.