**Project Title: The Battle of Neighborhoods, segmenting and clustering**

**Nilu Isakova**

**1) Introduction and Business Problem**

The purpose of this Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods in York, Toronto.

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

Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and reputated schools for their children. This project is for those people who are looking for better neighborhoods. Best schools in the neighborhoods, cafe, super market, medical shops, grocery shops, mall, theatre, hospital etc.

This Project aims to create an analysis of features for a people migrating to York to search a best neighborhood as a comparative analysis between neighborhoods. The features include better school according to ratings and etc.

**2) Data Description**

We will be using Toronto dataset which we scrapped from wikipedia on Week 3. Dataset consisting of latitude and longitude, zip codes. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

The dataset will consist of three columns: PostalCode, Borough, and Neighborhood Only process the cells that have an assigned borough.

To get the latitude and the longitude coordinates of each neighborhood, we will use a link to a csv file that has the geographical coordinates of each postal code: <http://cocl.us/Geospatial_data> to get the latitude and the longitude coordinates of each neighborhood. For schools rating [https://www.greatschools.org](https://www.greatschools.org/)

We will be using the Foursquare API to explore neighborhoods in York, Toronto. Also, will use explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. To clustering we will use the k-means clustering algorithm to complete this task. Finally, we will use the Folium library to visualize the neighborhoods in York and their emerging clusters.

List of all the necessary packages:

* numpy library to handle data in a vectorized manner
* pandas library for data analsysis
* json library to handle JSON files
* geopy.geocoders, Nominatim convert an address into latitude and longitude values
* requests ibrary to handle requests
* pandas.io.json, json\_normalize tranform JSON file into a pandas dataframe
* Matplotlib and associated plotting modules
* matplotlib.pyplot
* k-means from clustering stage
* sklearn.cluster, KMeans
* folium map rendering library
* wordcloud, WordCloud, STOPWORDS for wordcloud plots
* BeautifulSoup parse html data, and create a dataframe.

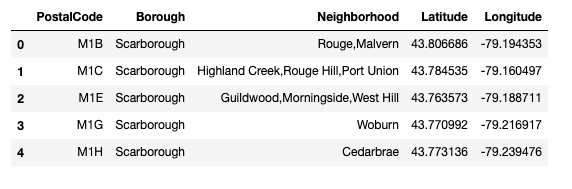
**2.1) Data cleaning**

We used Geospatial Data to get the postal code of each neighborhood along with the borough name and neighborhood name, in order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighborhood.

For the York, Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in York, Toronto. We scraped the Wikipedia page and wrangle the data, clean it, and then read it into a *pandas* dataframe so that it is in a structured format. After downloading the data from wikipedia as html, we used use BeautifySoup package to parse html data, and create a dataframe.

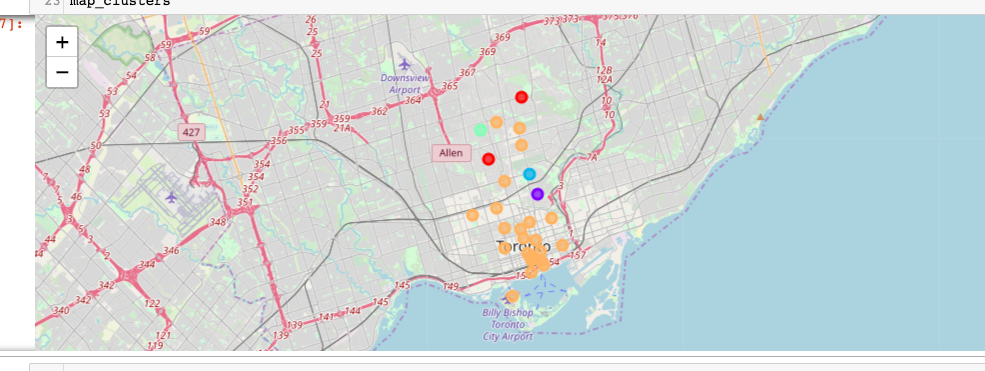
We used a link below to a csv file that has the geographical coordinates of each postal code: <http://cocl.us/Geospatial_data> to get the latitude and the longitude coordinates of each neighborhood.

After structured format the data looks like this:



**3) Methodology**

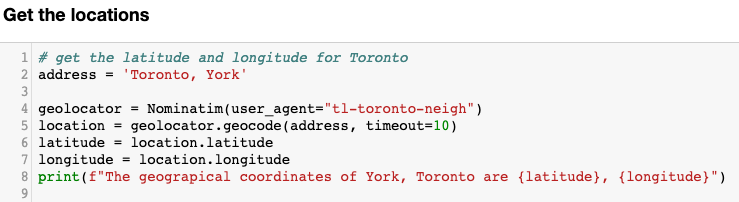
Clustering Approach: To find similar neighborhoods we explore and cluster neighborhoods, segment them, and group them into 5 clusters, and we used k-means clustering algorithm. The below picture shows the five clusters on map.



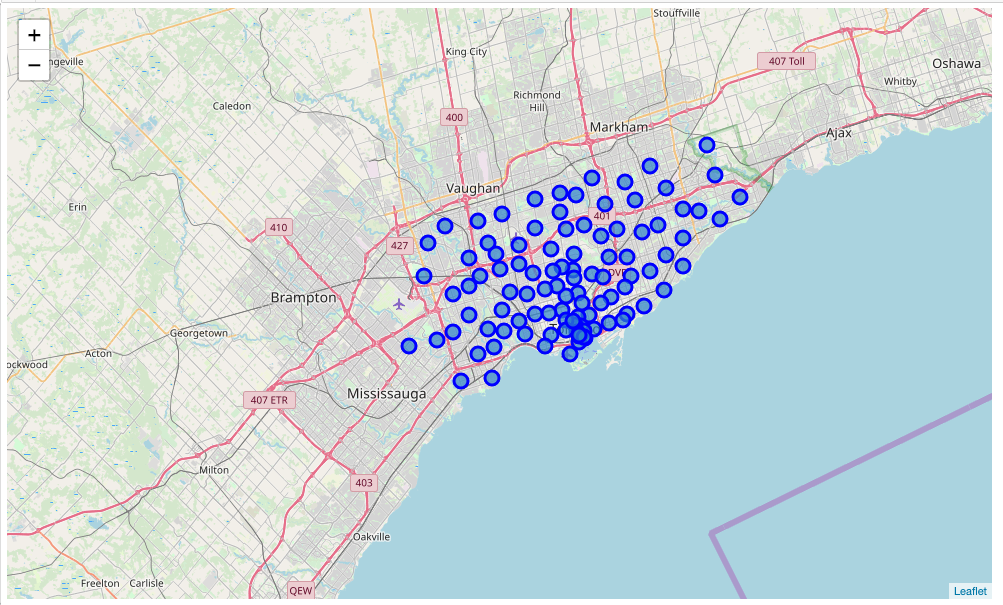
**4) Results**

As a result, York is one of the most diverse and multicultural areas in the Greater Toronto Area. We used Foursquare API to get neighborhoods, values. As a result we found out that 1) Park 2) West 3) North South were the most prefered neighborhoods. The best rating schools were in in the following neighborhoods: Glencairn  
Lawrence Park  
Dorset Park,Scarborough Town Centre  
Cliffcrest,Cliffside,Scarborough Village West

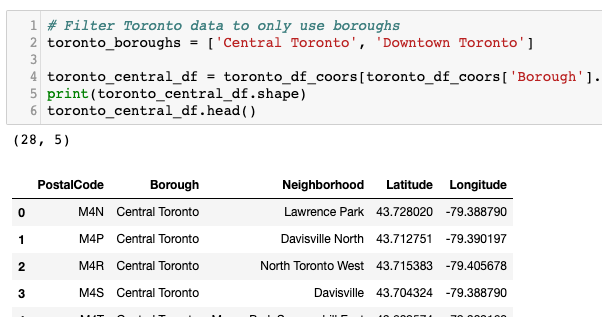
For the latitude and longtitude information we used the York Toronto address.



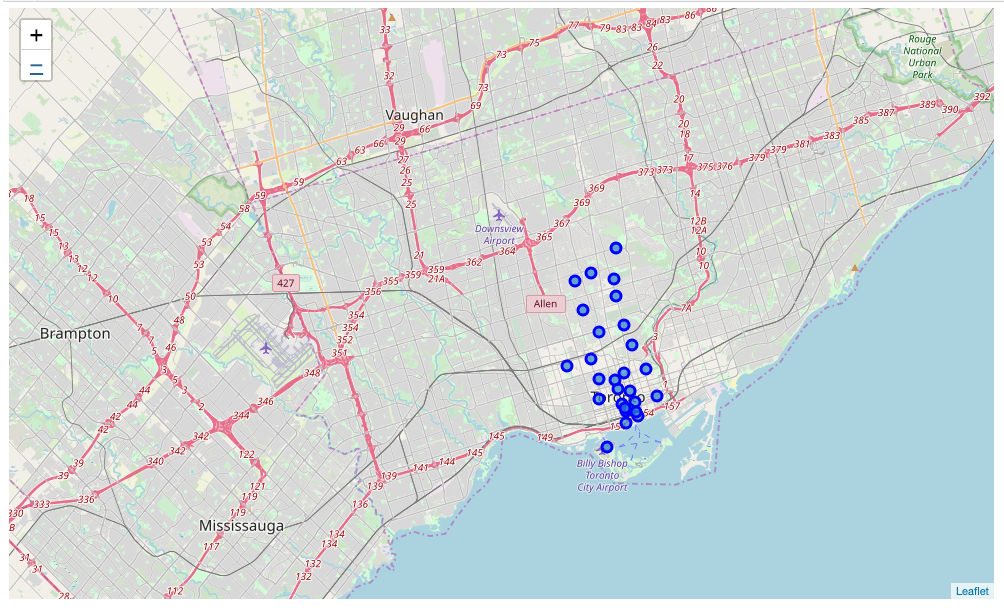
And the map looks like below



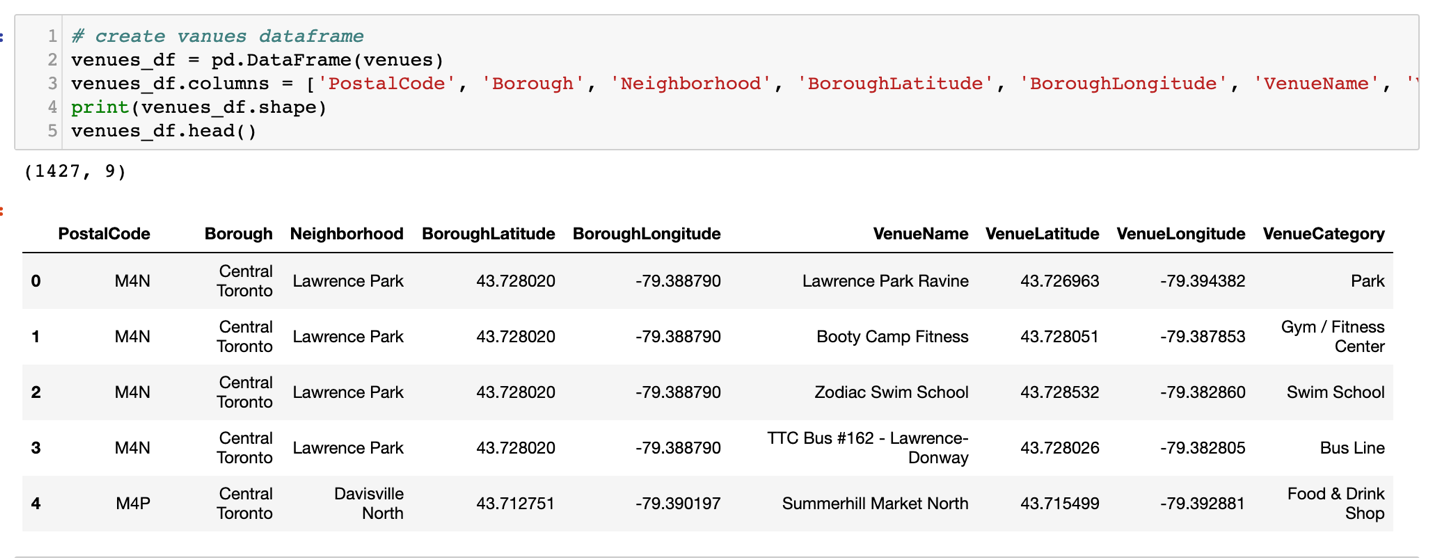
And we narrowed the area for only look Central Toronto and Downtown Toronto, as shown below picture.



And now the map looks like as below



We used the following vanues for the analysis purpose



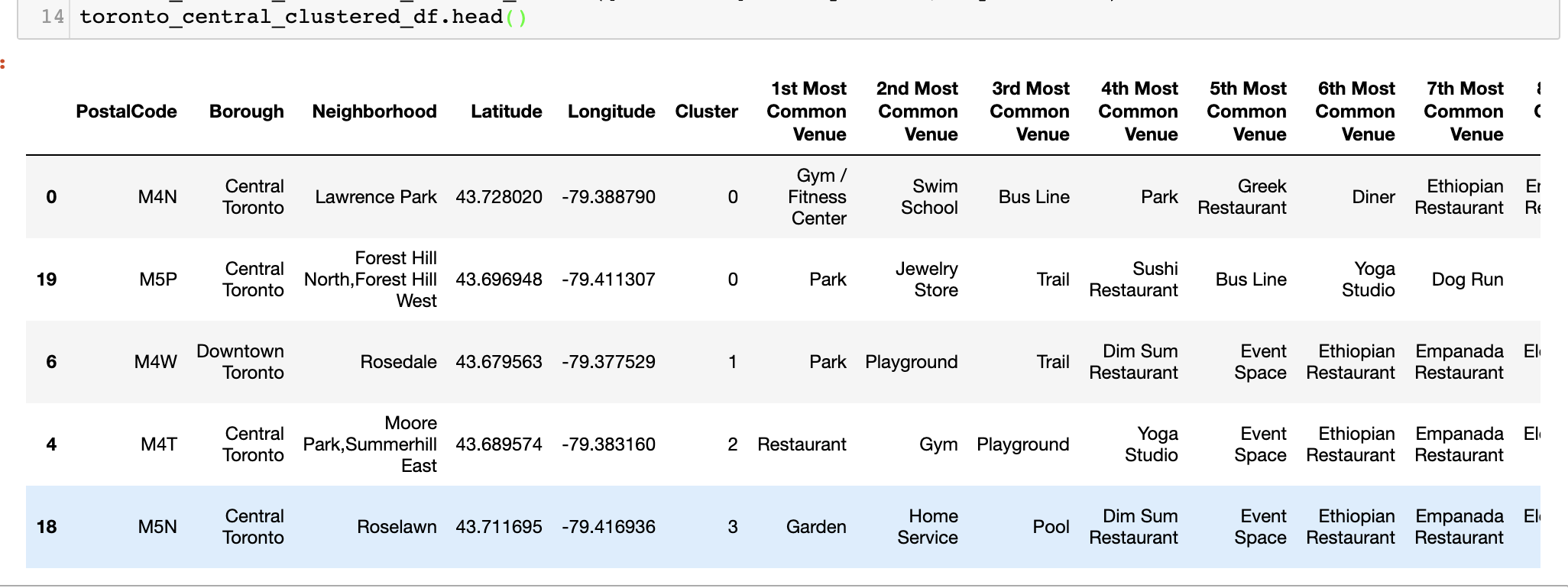
And the venue names:



We used 217 categoriesfor the top 10 vanues:



We divided the dataset into 5 clusters:



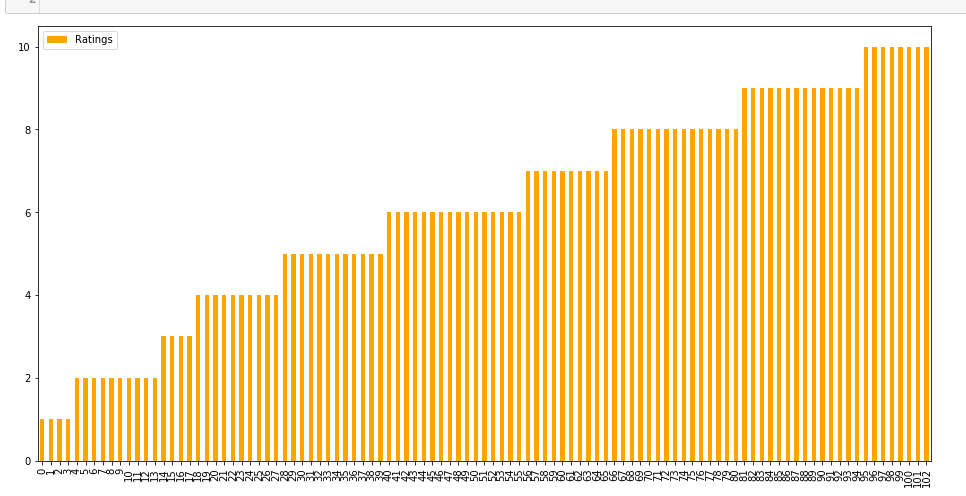
**5) Discussion**

We tried to solve the following two main problems in Toronto York:

1. Sorted list of best neighborhoods and most of them belong to cluster4



1. Sorted list of schools in terms of rating and reviews



**6) Conclusion**

In this project, using k-means cluster algorithm we separated the neighborhood into five different clusters and for 103 different lattitude and logitude from dataset, and analyzed neighborhoods in York city.