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

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:

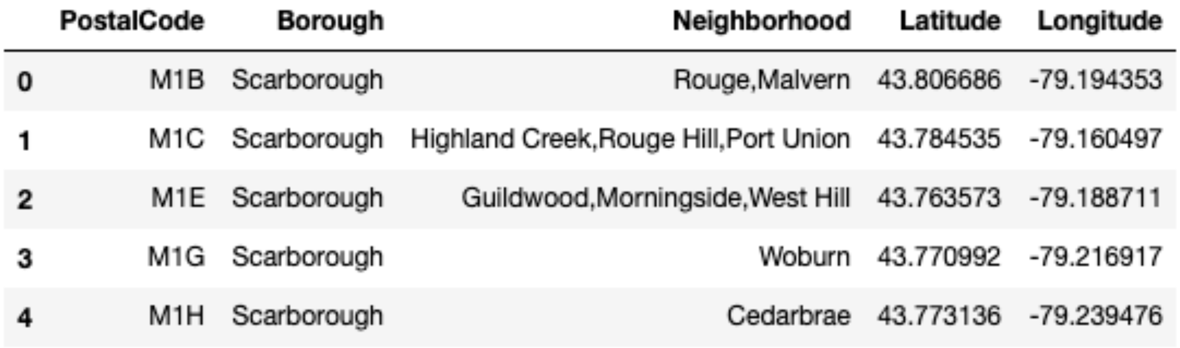
- numpy library to handle data in a vectorized manner
- pandas library for data analsysis
- json library to handle JSON files
- geopy.geocoders, Nominatim convert an address into latitude and longitude values
- requests ibrary to handle requests
- pandas.io.json, json_normalize tranform JSON file into 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.

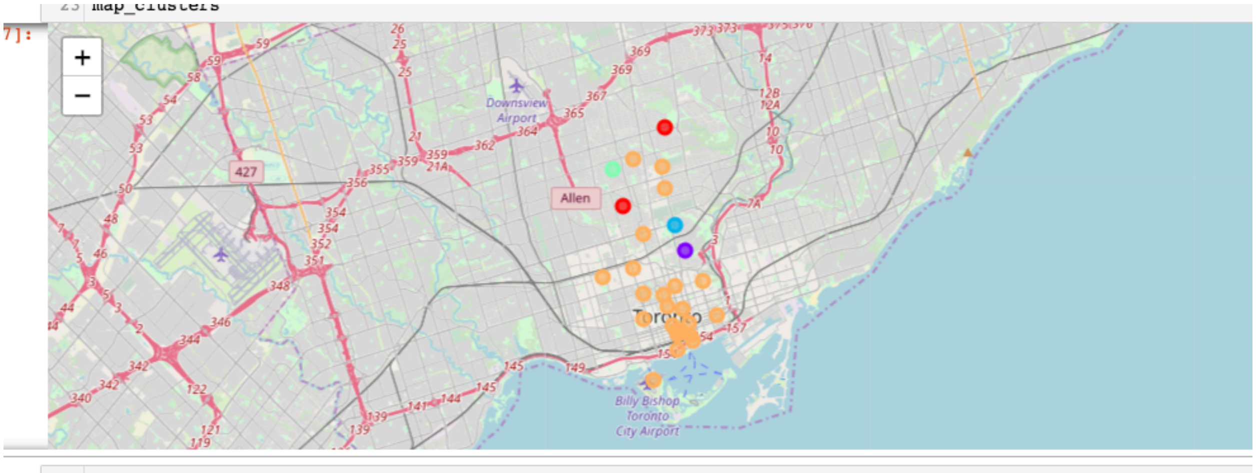
After structured format the data looks like this:



	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

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.

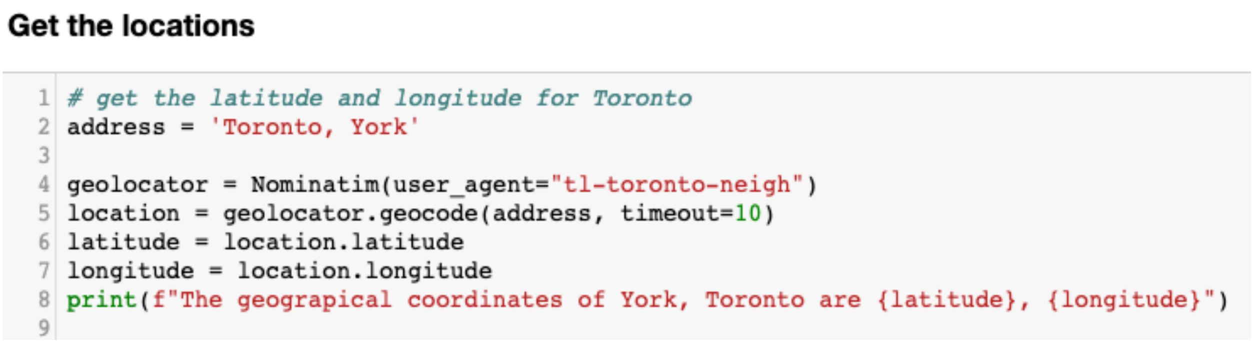


```
23 map_clusters
```

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

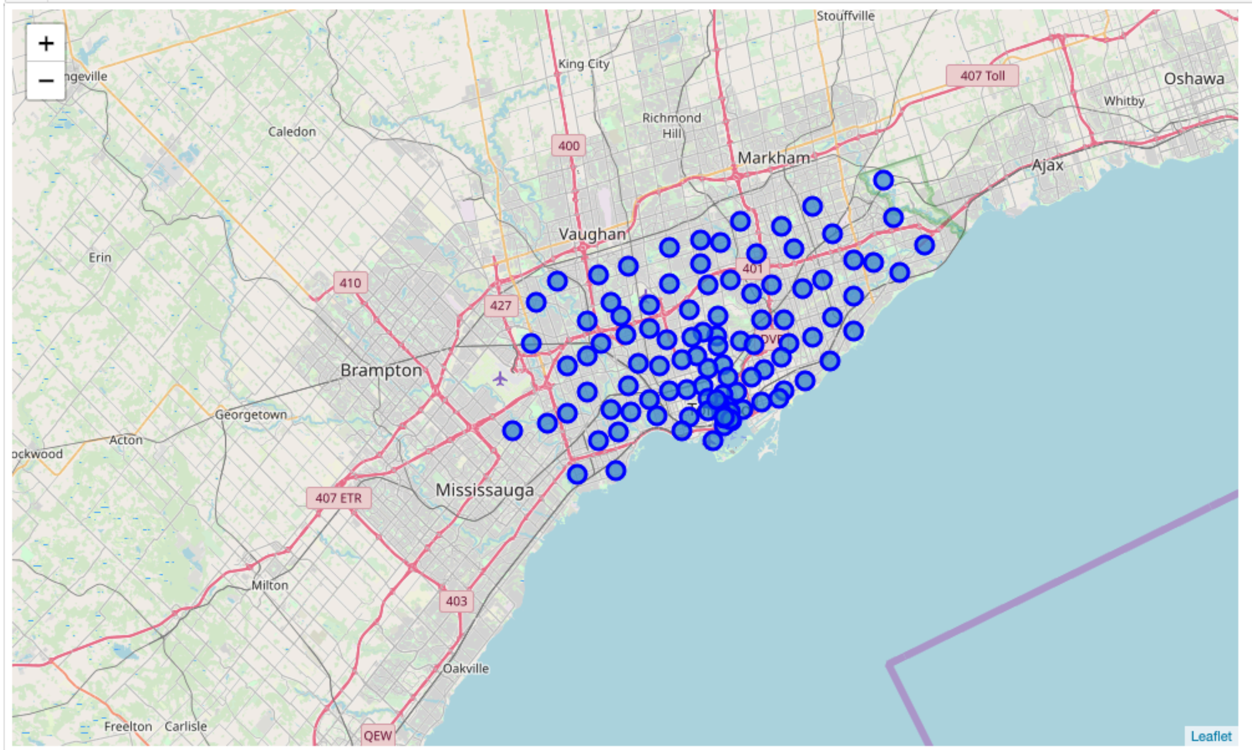
For the latitude and longtitude information we used the York Toronto address.



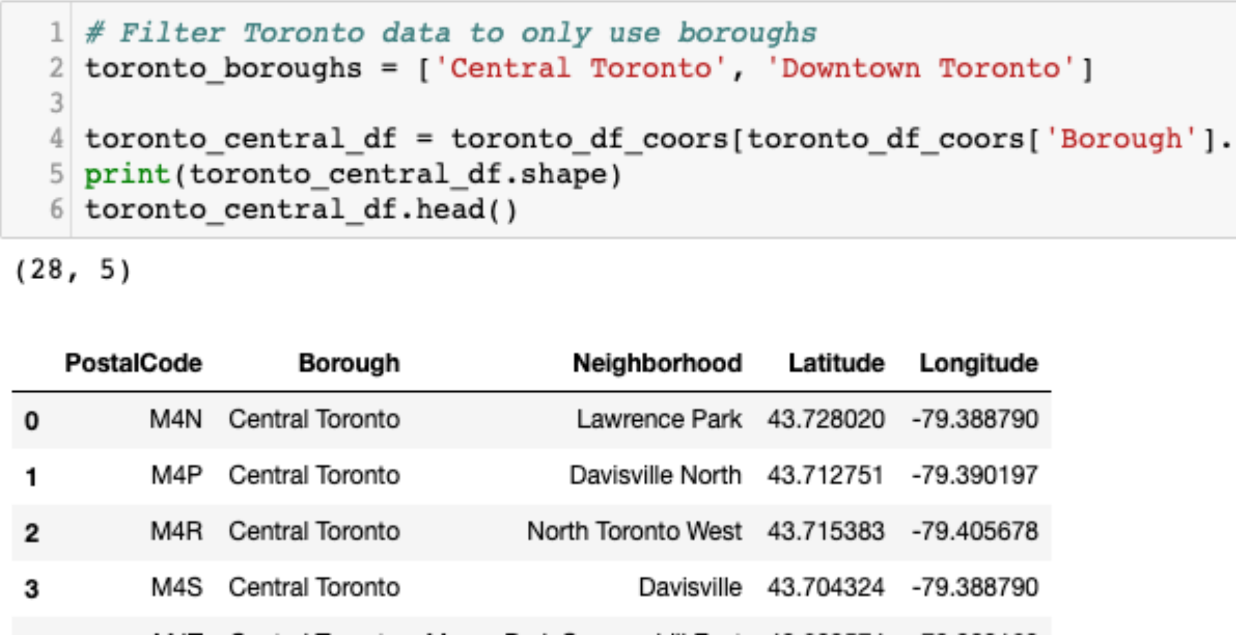
Get the locations

```
# get the latitude and longitude for Toronto
address = 'Toronto, York'

geolocator = Nominatim(user_agent="tl-toronto-neigh")
location = geolocator.geocode(address, timeout=10)
latitude = location.latitude
longitude = location.longitude
print(f"The geograpical coordinates of York, Toronto are {latitude}, {longitude}")

```

And the map looks like below



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

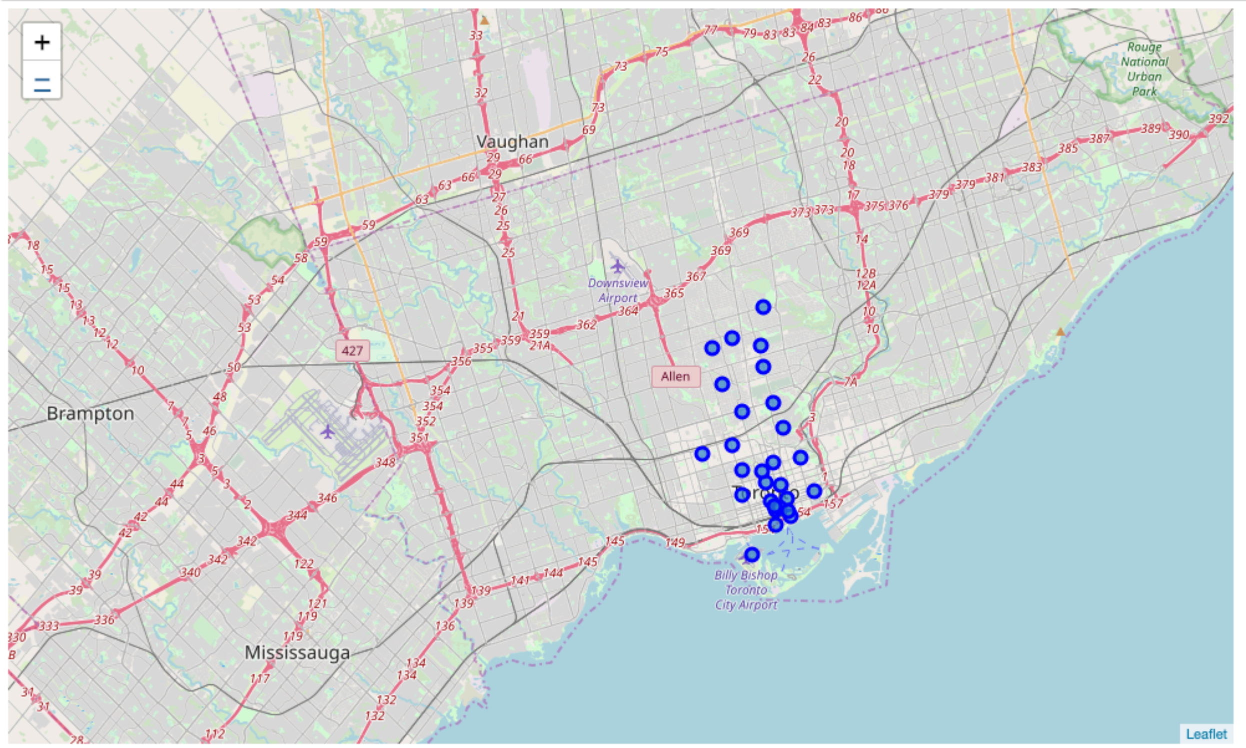


```
# Filter Toronto data to only use boroughs
toronto_boroughs = ['Central Toronto', 'Downtown Toronto']

toronto_central_df = toronto_df_coors[toronto_df_coors['Borough'].
print(toronto_central_df.shape)
toronto_central_df.head()
```

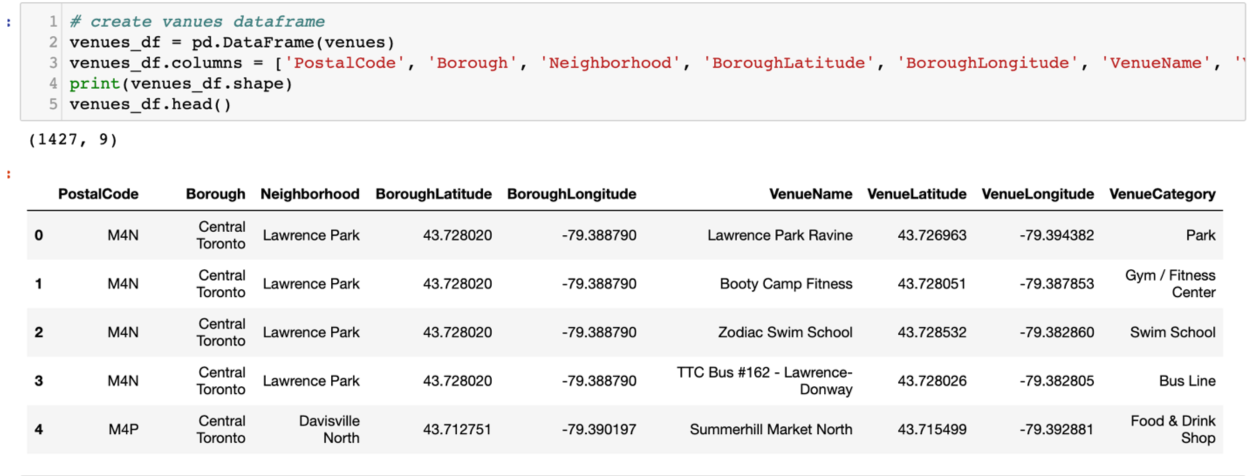
```
(28, 5)
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
1	M4P	Central Toronto	Davisville North	43.712751	-79.390197
2	M4R	Central Toronto	North Toronto West	43.715383	-79.405678
3	M4S	Central Toronto	Davisville	43.704324	-79.388790

And now the map looks like as below



We used the following vanues for the analysis purpose

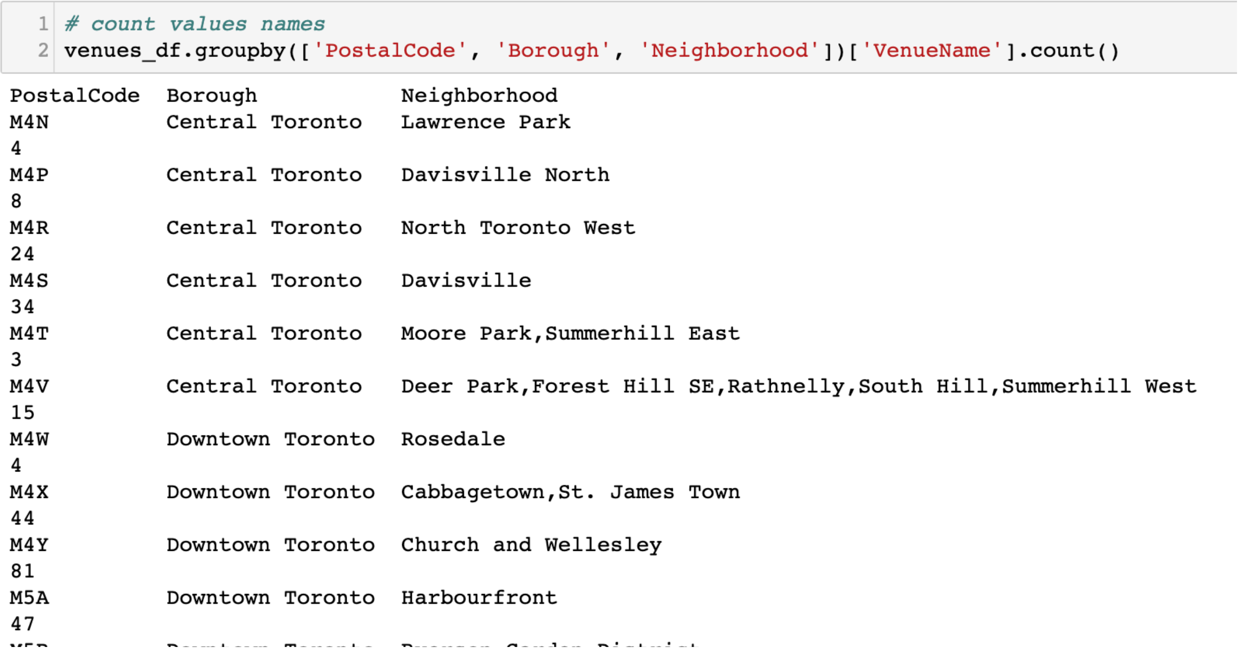


```
# create vanues dataframe
venues_df = pd.DataFrame(venues)
venues_df.columns = ['PostalCode', 'Borough', 'Neighborhood', 'BoroughLatitude', 'BoroughLongitude', 'VenueName', '
print(venues_df.shape)
venues_df.head()
```

```
(1427, 9)
```

	PostalCode	Borough	Neighborhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	Booty Camp Fitness	43.728051	-79.387853	Gym / Fitness Center
2	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
3	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
4	M4P	Central Toronto	Davisville North	43.712751	-79.390197	Summerhill Market North	43.715499	-79.392881	Food & Drink Shop

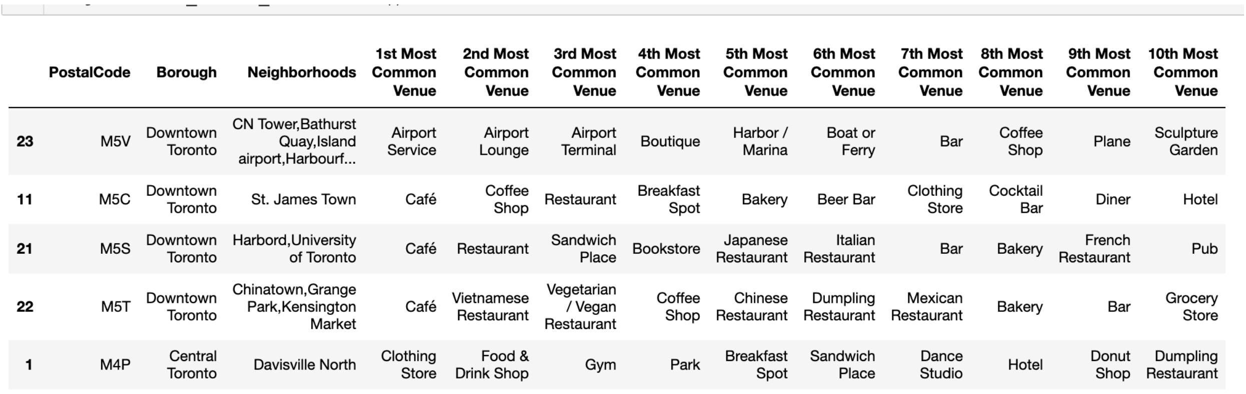
And the venue names:



```
# count values names
venues_df.groupby(['PostalCode', 'Borough', 'Neighborhood'])['VenueName'].count()
```

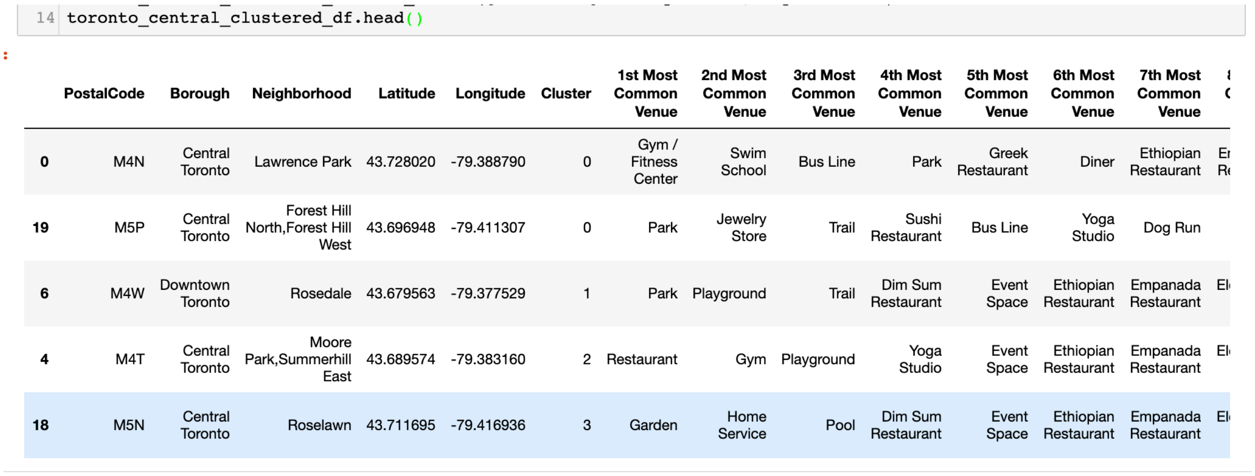
```
PostalCode  Borough           Neighborhood
M4N         Central Toronto   Lawrence Park
4
M4P         Central Toronto   Davisville North
8
M4R         Central Toronto   North Toronto West
24
M4S         Central Toronto   Davisville
34
M4T         Central Toronto   Moore Park,Summerhill East
3
M4V         Central Toronto   Deer Park,Forest Hill SE,Rathnelly,South Hill,Summerhill West
15
M4W         Downtown Toronto  Rosedale
4
M4X         Downtown Toronto  Cabbagetown,St. James Town
44
M4Y         Downtown Toronto  Church and Wellesley
81
M5A         Downtown Toronto  Harbourfront
47
M5B         Downtown Toronto  Ryerson,Garden District
```

We used 217 categoriesfor the top 10 vanues:



	PostalCode	Borough	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	M5V	Downtown Toronto	CN Tower,Bathurst Quay,Island airport,Harbourf...	Airport Service	Airport Lounge	Airport Terminal	Boutique	Harbor / Marina	Boat or Ferry	Bar	Coffee Shop	Plane	Sculpture Garden
11	M5C	Downtown Toronto	St. James Town	Café	Coffee Shop	Restaurant	Breakfast Spot	Bakery	Beer Bar	Clothing Store	Cocktail Bar	Diner	Hotel
21	M5S	Downtown Toronto	Harbord,University of Toronto	Café	Restaurant	Sandwich Place	Bookstore	Japanese Restaurant	Italian Restaurant	Bar	Bakery	French Restaurant	Pub
22	M5T	Downtown Toronto	Chinatown,Grange Park,Kensington Market	Café	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Coffee Shop	Chinese Restaurant	Dumpling Restaurant	Mexican Restaurant	Bakery	Bar	Grocery Store
1	M4P	Central Toronto	Davisville North	Clothing Store	Food & Drink Shop	Gym	Park	Breakfast Spot	Sandwich Place	Dance Studio	Hotel	Donut Shop	Dumpling Restaurant

We divided the dataset into 5 clusters:



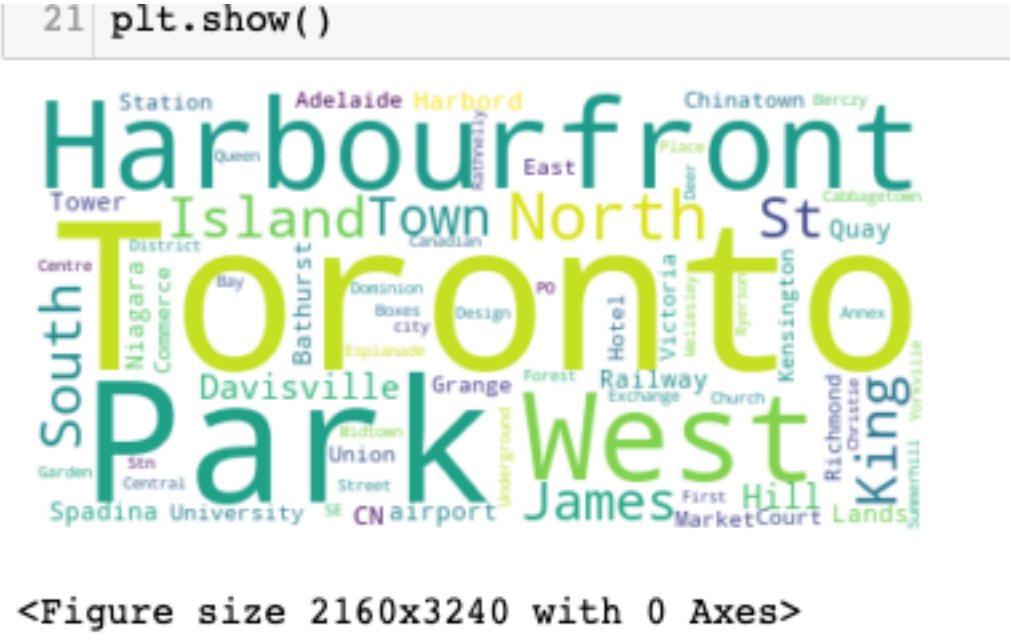
```
toronto_central_clustered_df.head()
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8 C
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	0	Gym / Fitness Center	Swim School	Bus Line	Park	Greek Restaurant	Diner	Ethiopian Restaurant	Er Re
19	M5P	Central Toronto	Forest Hill North,Forest Hill West	43.696948	-79.411307	0	Park	Jewelry Store	Trail	Sushi Restaurant	Bus Line	Yoga Studio	Dog Run	
6	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	1	Park	Playground	Trail	Dim Sum Restaurant	Event Space	Ethiopian Restaurant	Empanada Restaurant	El
4	M4T	Central Toronto	Moore Park,Summerhill East	43.689574	-79.383160	2	Restaurant	Gym	Playground	Yoga Studio	Event Space	Ethiopian Restaurant	Empanada Restaurant	El
18	M5N	Central Toronto	Roselawn	43.711695	-79.416936	3	Garden	Home Service	Pool	Dim Sum Restaurant	Event Space	Ethiopian Restaurant	Empanada Restaurant	El

5) Discussion

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

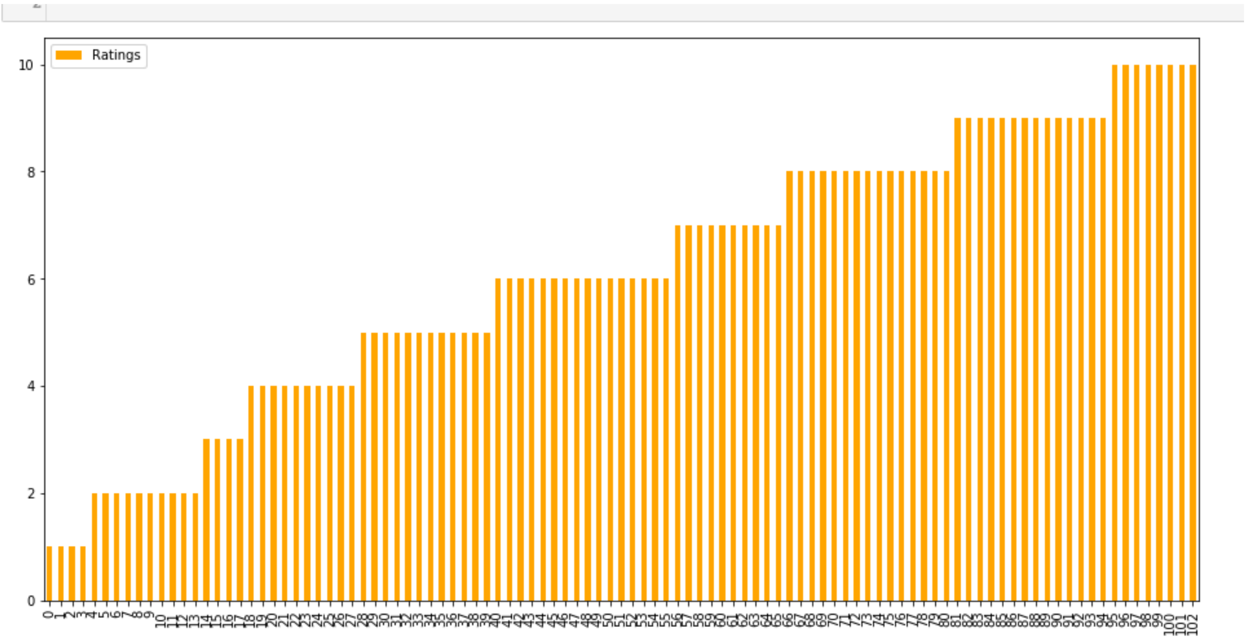
1) Sorted list of best neighborhoods and most of them belong to cluster4



```
plt.show()
```

```
<Figure size 2160x3240 with 0 Axes>
```

2) 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.