Capstone Project - The Battle of Neighborhoods (Week 2)

1. Description of the Problem & Discussion of the Background (Introduction Section)

IBM data science professional program's Capstone Project - The Battle of Neighborhoods is a project in which we apply data science and machine learning algorithm to find the best neighbourhood with in Scarborough,Toronto. This project make use of the FourSquare API to get the data of the school , housing prices and other important reviews of the neighborhood. Making use of the python web scrapping and k means clustering we cluster the data and find the best location to fit the requirements.
The main purpose of this project is to identify the best suitable neighborhood. This project will make the businesses or user to easily decide the best neighborhood.Many people are migrating to Canada and needs lots to search for good housing prices and schools for their children. This project is for those people who are looking for better neighborhoods with low cost. For ease of accessing to Cafe, School, Super market, medical shops, grocery shops, marts etc.The main objective of the project is to define a cluster where the facilities are readily available.

In this project, we acquire data by using web scraping method of python and then clean the data,populate the data then use foursquare API to collect the data of the all the neighborhood places then use k-means clustering method to find the best fit.

First settled by Europeans in the 1790s since then Scarborough has grown from a collection of small rural villages and farms to become fully urbanized with a diverse cultural community. Incorporated in 1850 as a township, Scarborough became part of Metropolitan Toronto in 1953 and was reconstituted as a borough in 1967. Scarborough rapidly developed as a suburb of Old Toronto over the next decade and became a city in 1983. In 1998, Scarborough and the rest of Metropolitan Toronto were amalgamated into the present city of Toronto. Scarborough still exists in name and as a borough of Toronto. The Scarborough Civic Centre, the former city‘s last place of government, is occupied by City of Toronto government offices. Scarborough is a popular destination for new immigrants in Canada to reside. As a result, it is one of the most diverse and multicultural areas in the Greater Toronto Area, being home to various religious groups and places of worship. Although immigration has become a hot topic over the past few years with more governments seeking more restrictions on immigrants and refugees, the general trend of immigration into Canada has been one of on the rise.

2. Methodology

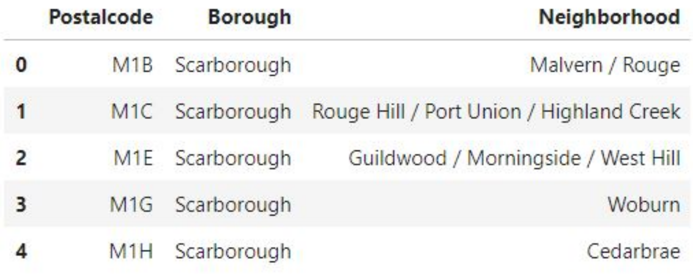
i. Data Acquisition and Cleaning

a) the data I am using is from the wiki portal : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M . THis data has all the information of the neighbourhood data of the Toronto.

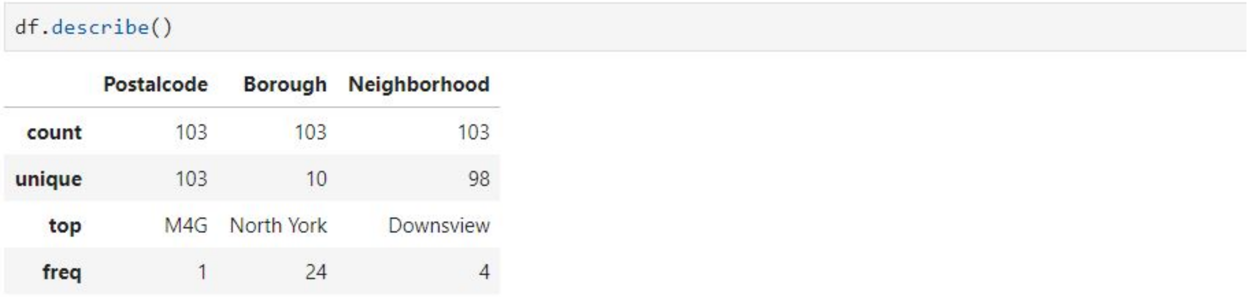
b) Then to get the geo graphical coordinates of the neighbor hood by using the foursquare api and geopy library
c) Other venue location are also fetched using the foursquare API

The data from the wikipedia source is scarped and imported directly to the dataframe in pandas. The dataframe consisted of three columns namely PostalCode, Borough, and Neighborhood. The dataframe whose value is not null in the Borough field is considered rest data are ignored, and if there are two postal Code we conbine those two rows to one.

Wiki Scraped Data:



	Postalcode	Borough	Neighborhood
0	M1B	Scarborough	Malvern / Rouge
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek
2	M1E	Scarborough	Guildwood / Morningside / West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

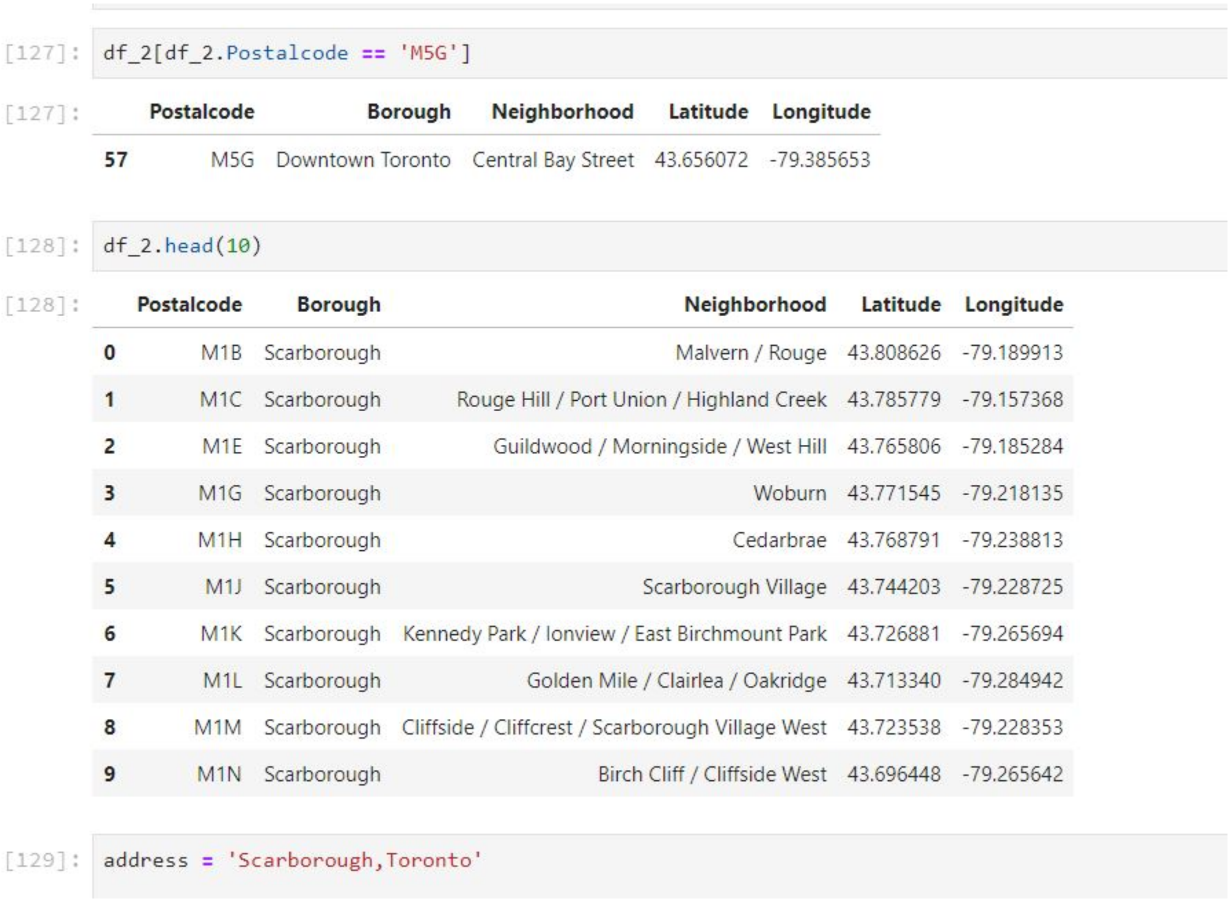


```
df.describe()
```

	Postalcode	Borough	Neighborhood
count	103	103	103
unique	103	10	98
top	M4G	North York	Downsview
freq	1	24	4

ii. Getting Latlong Data using get_latilong Function

get_latilong function uses arcgis tools to get the latitude and longitude of the neighborhood using just the postal code of the neighborhood.



```
[127]: df_2[df_2.Postalcode == 'M5G']
```

[127]:

	Postalcode	Borough	Neighborhood	Latitude	Longitude
57	M5G	Downtown Toronto	Central Bay Street	43.656072	-79.385653

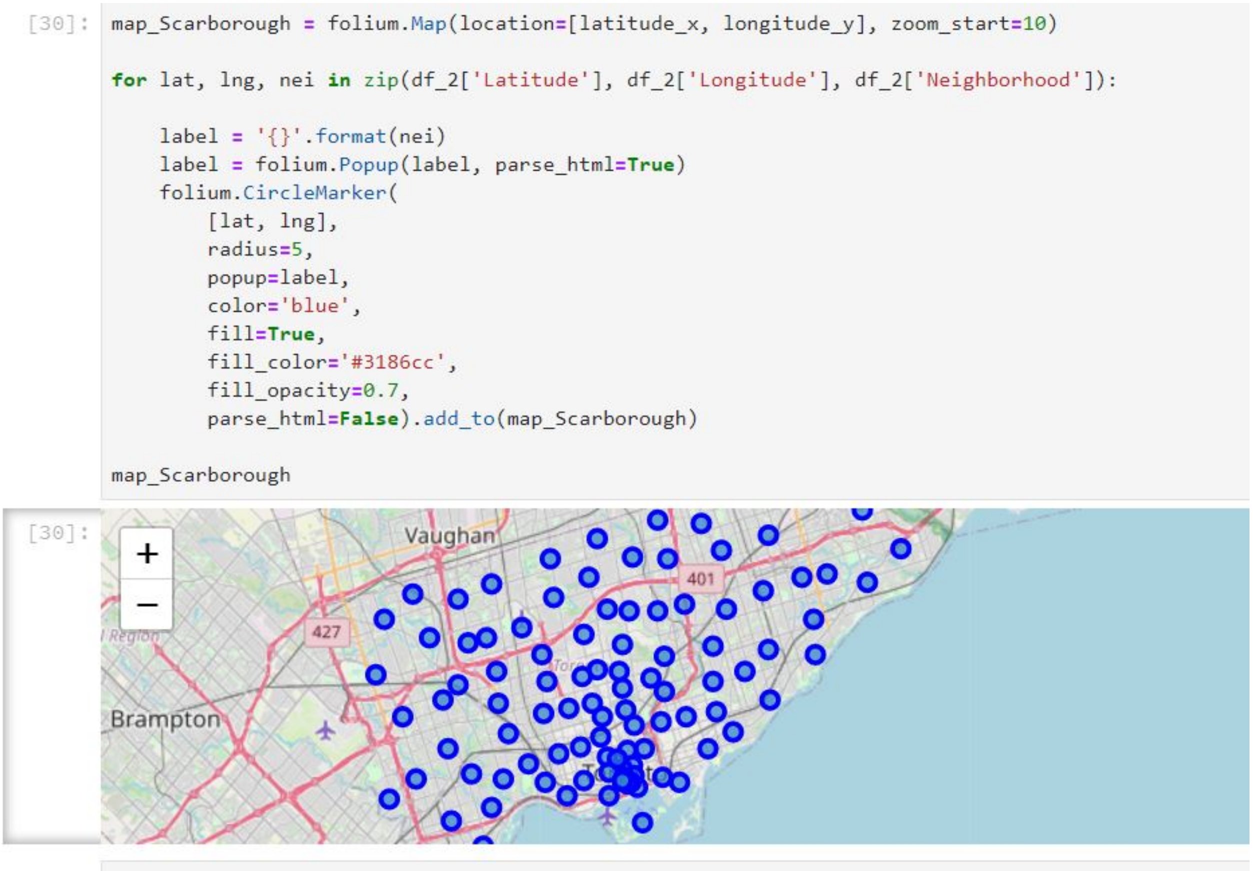
```
[128]: df_2.head(10)
```

[128]:

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern / Rouge	43.808626	-79.189913
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	43.785779	-79.157368
2	M1E	Scarborough	Guildwood / Morningside / West Hill	43.765806	-79.185284
3	M1G	Scarborough	Woburn	43.771545	-79.218135
4	M1H	Scarborough	Cedarbrae	43.768791	-79.238813
5	M1J	Scarborough	Scarborough Village	43.744203	-79.228725
6	M1K	Scarborough	Kennedy Park / Ionview / East Birchmount Park	43.726881	-79.265694
7	M1L	Scarborough	Golden Mile / Clairlea / Oakridge	43.713340	-79.284942
8	M1M	Scarborough	Cliffside / Cliffcrest / Scarborough Village West	43.723538	-79.228353
9	M1N	Scarborough	Birch Cliff / Cliffside West	43.696448	-79.265642

```
[129]: address = 'Scarborough,Toronto'
```

iii. Plotting the map of Scarborough using the folium library



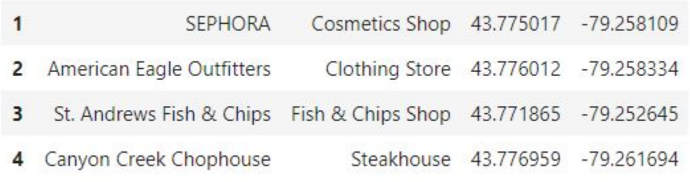
```
[30]: map_Scarborough = folium.Map(location=[latitude_x, longitude_y], zoom_start=10)

      for lat, lng, nei in zip(df_2['Latitude'], df_2['Longitude'], df_2['Neighborhood']):

          label = '{}'.format(nei)
          label = folium.Popup(label, parse_html=True)
          folium.CircleMarker(
              [lat, lng],
              radius=5,
              popup=label,
              color='blue',
              fill=True,
              fill_color='#3186cc',
              fill_opacity=0.7,
              parse_html=False).add_to(map_Scarborough)

      map_Scarborough

[30]:
```

iv. iv iv) Connecting to Foursquare API to get data

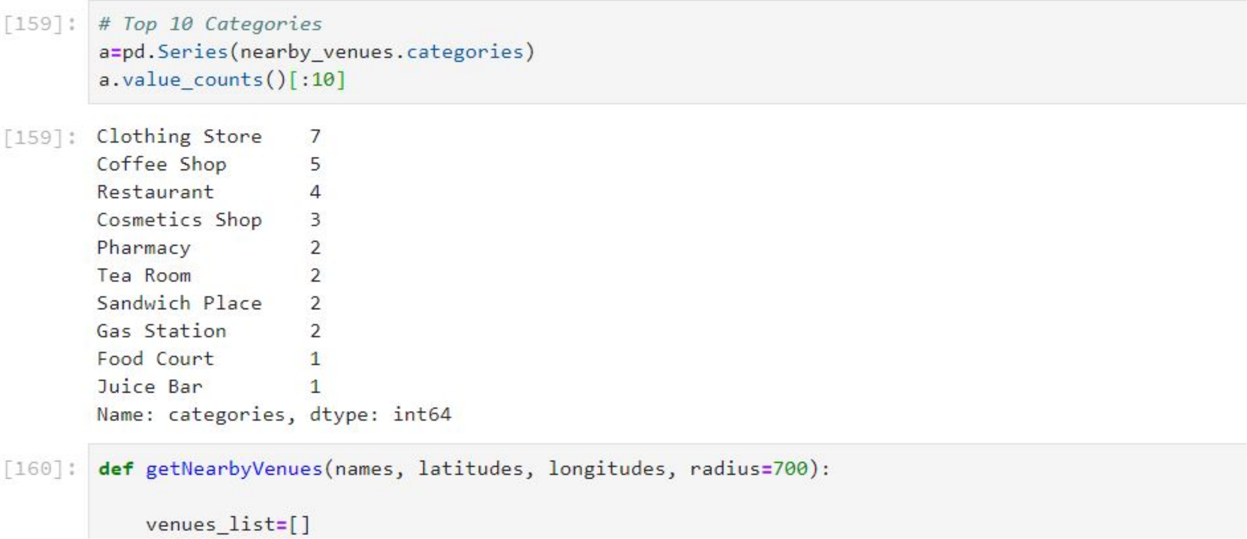
The foursquare developers client id and client secret is used to fetch the data's needed data about different venues in different neighborhoods of that specific borough. The fetched data included are venue names, locations, menus photos etc. All the information about the prices and school near 100 meters of radius are chosen for this project. The In-formations gathered from the foursquare API are

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

v) Getting Near By Venus, Schools and locations Category



1	SEPHORA	Cosmetics Shop	43.775017	-79.258109
2	American Eagle Outfitters	Clothing Store	43.776012	-79.258334
3	St. Andrews Fish & Chips	Fish & Chips Shop	43.771865	-79.252645
4	Canyon Creek Chophouse	Steakhouse	43.776959	-79.261694

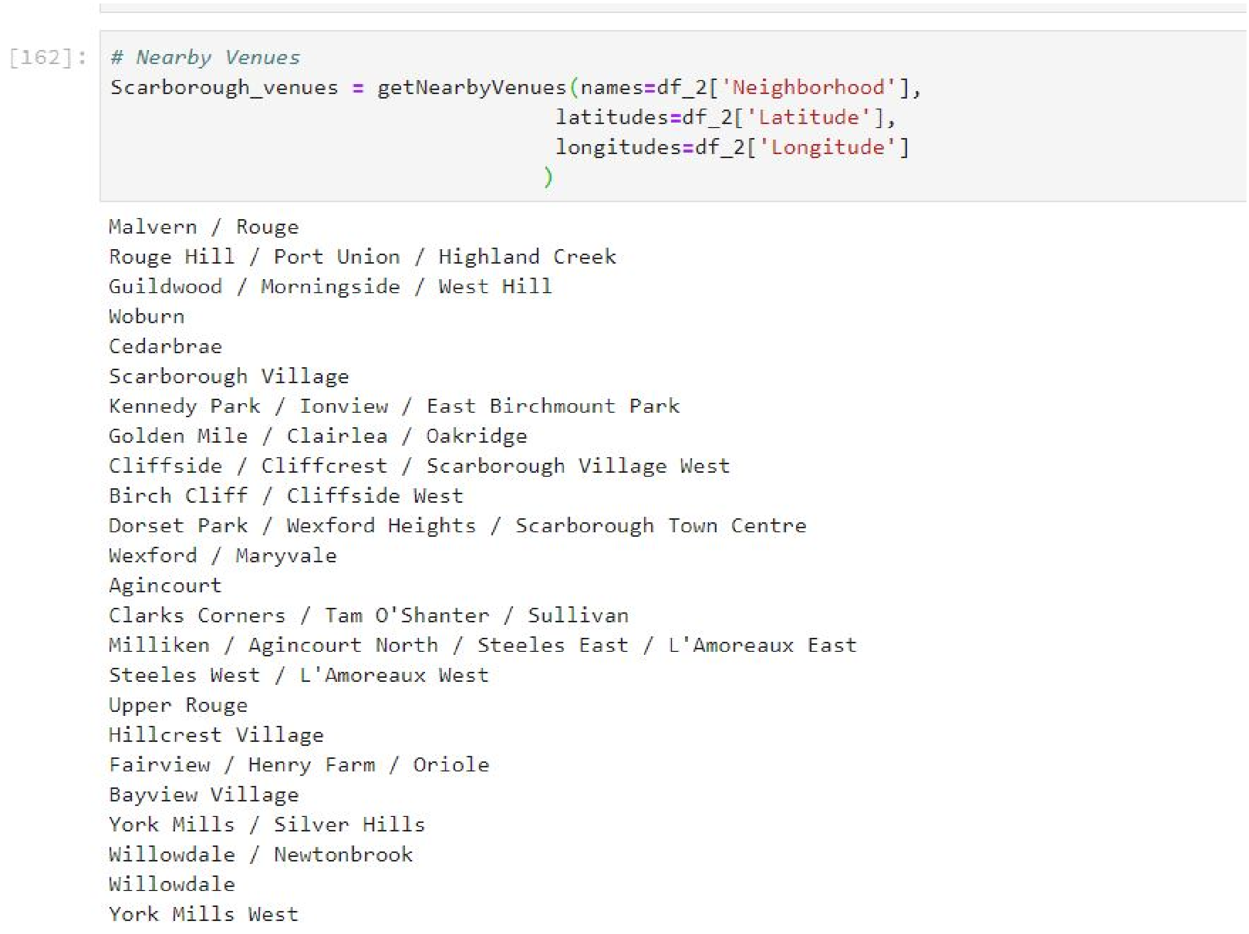


```
[159]: # Top 10 Categories
       a=pd.Series(nearby_venues.categories)
       a.value_counts()[:10]
```

```
[159]: Clothing Store    7
       Coffee Shop       5
       Restaurant        4
       Cosmetics Shop    3
       Pharmacy          2
       Tea Room          2
       Sandwich Place    2
       Gas Station       2
       Food Court        1
       Juice Bar         1
       Name: categories, dtype: int64
```

```
[160]: def getNearbyVenues(names, latitudes, longitudes, radius=700):

           venues_list=[]
```

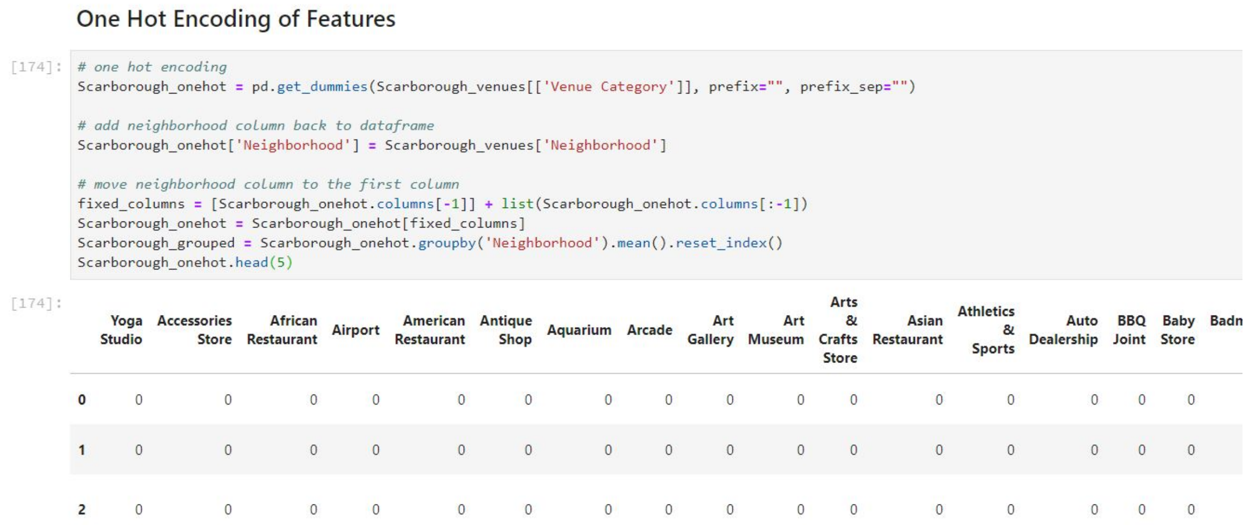
vi) Getting Near By Venus, Schools and locations Names



```
[162]: # Nearby Venues
       Scarborough_venues = getNearbyVenues(names=df_2['Neighborhood'],
                                           latitudes=df_2['Latitude'],
                                           longitudes=df_2['Longitude']
                                          )
```

```
Malvern / Rouge
Rouge Hill / Port Union / Highland Creek
Guildwood / Morningside / West Hill
Woburn
Cedarbrae
Scarborough Village
Kennedy Park / Ionview / East Birchmount Park
Golden Mile / Clairlea / Oakridge
Cliffside / Cliffcrest / Scarborough Village West
Birch Cliff / Cliffside West
Dorset Park / Wexford Heights / Scarborough Town Centre
Wexford / Maryvale
Agincourt
Clarks Corners / Tam O'Shanter / Sullivan
Milliken / Agincourt North / Steeles East / L'Amoreaux East
Steeles West / L'Amoreaux West
Upper Rouge
Hillcrest Village
Fairview / Henry Farm / Oriole
Bayview Village
York Mills / Silver Hills
Willowdale / Newtonbrook
Willowdale
York Mills West
```

Vii) One Hot Encoding of features



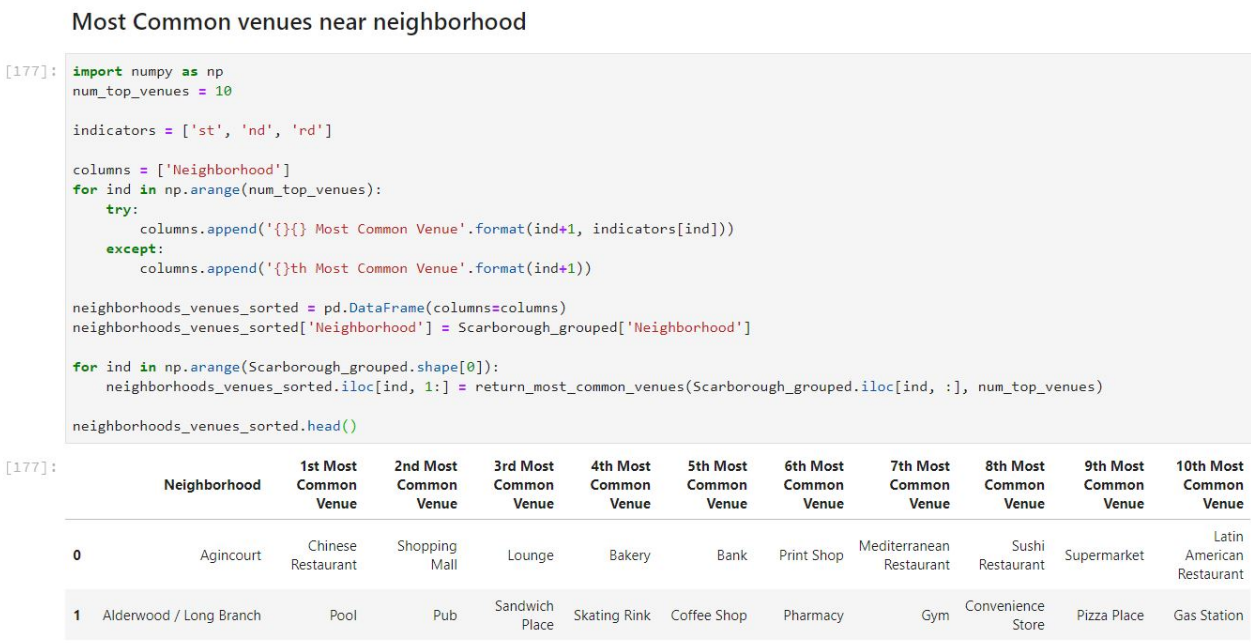
One Hot Encoding of Features

```
[174]: # one hot encoding
       Scarborough_onehot = pd.get_dummies(Scarborough_venues[['Venue Category']], prefix="", prefix_sep="")

       # add neighborhood column back to dataframe
       Scarborough_onehot['Neighborhood'] = Scarborough_venues['Neighborhood']

       # move neighborhood column to the first column
       fixed_columns = [Scarborough_onehot.columns[-1]] + list(Scarborough_onehot.columns[:-1])
       Scarborough_onehot = Scarborough_onehot[fixed_columns]
       Scarborough_grouped = Scarborough_onehot.groupby('Neighborhood').mean().reset_index()
       Scarborough_onehot.head(5)
```

[174]:

	Yoga Studio	Accessories Store	African Restaurant	Airport	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	BBQ Joint	Baby Store	Badn
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Viii) Most Common venues near neighborhood



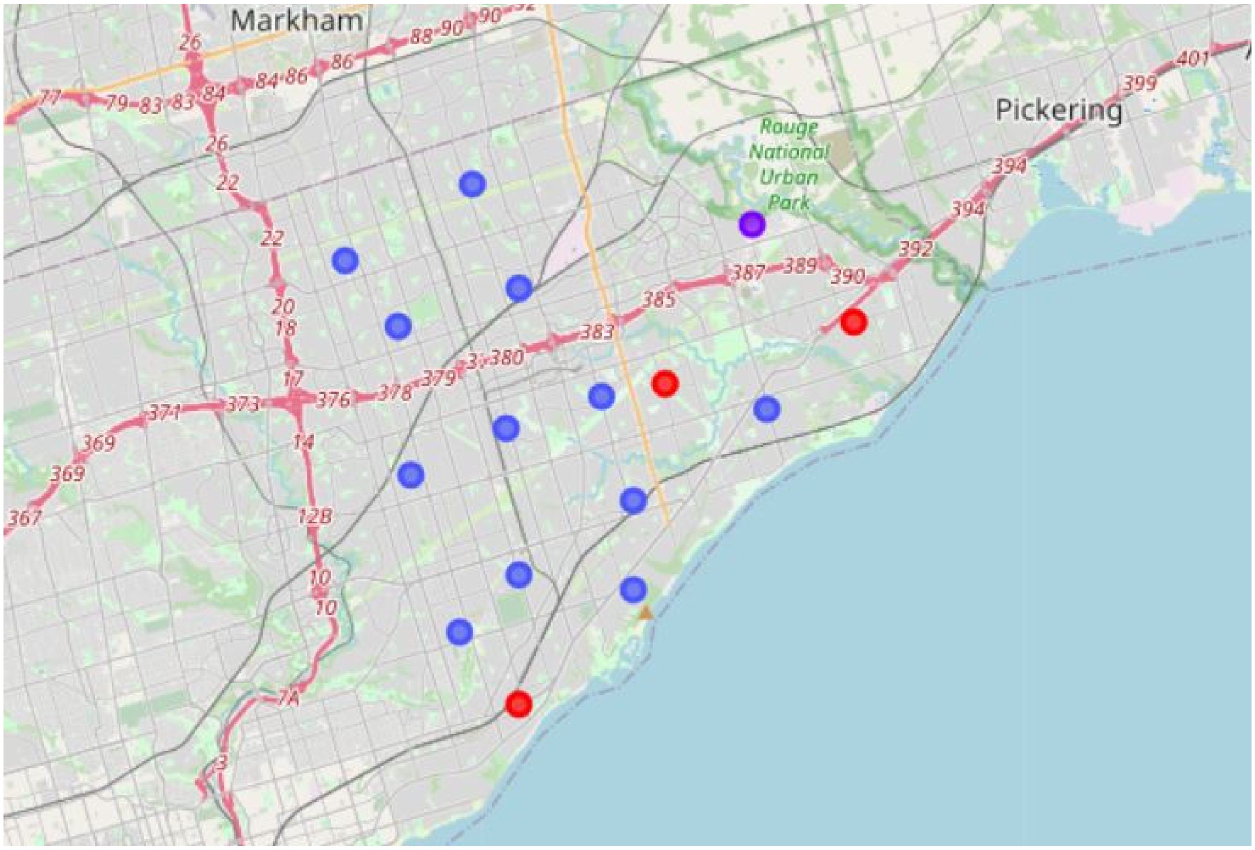
Most Common venues near neighborhood

[177]:
```python
import numpy as np
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = Scarborough_grouped['Neighborhood']

for ind in np.arange(Scarborough_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(Scarborough_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

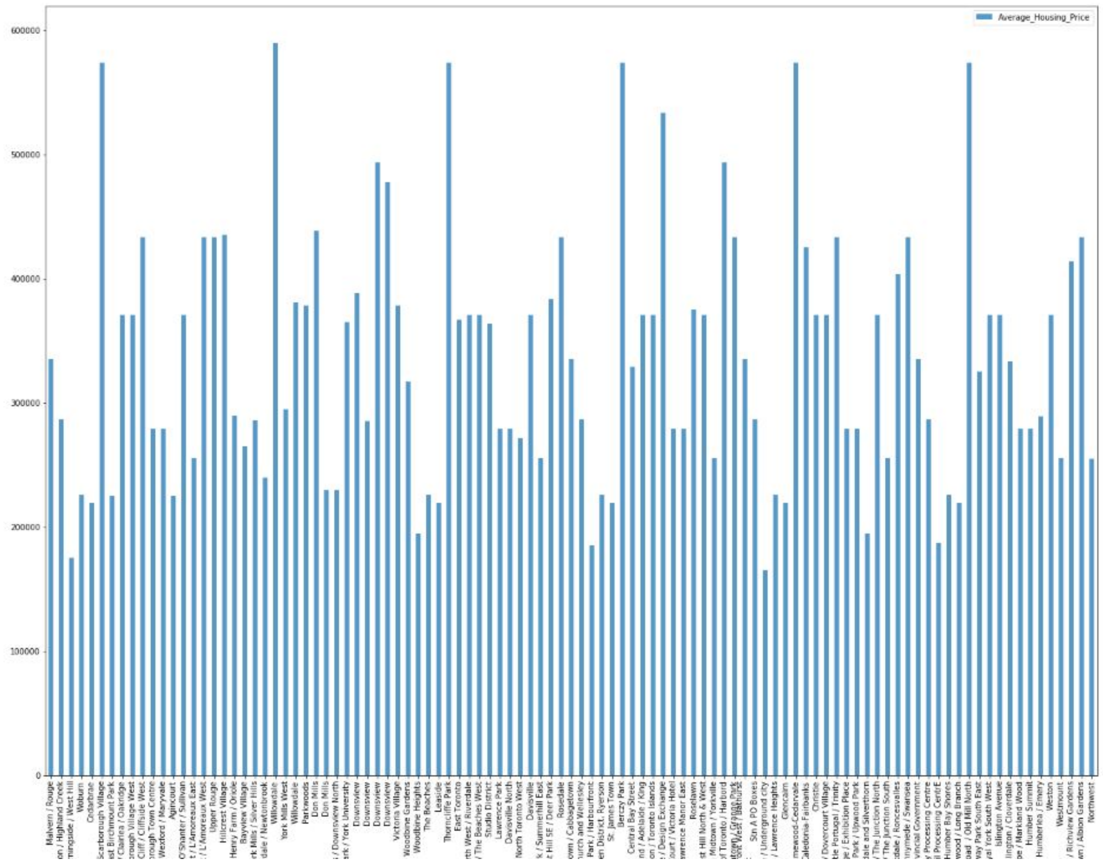
[177]:

	Neighborhood	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
0	Agincourt	Chinese Restaurant	Shopping Mall	Lounge	Bakery	Bank	Print Shop	Mediterranean Restaurant	Sushi Restaurant	Supermarket	Latin American Restaurant
1	Alderwood / Long Branch	Pool	Pub	Sandwich Place	Skating Rink	Coffee Shop	Pharmacy	Gym	Convenience Store	Pizza Place	Gas Station

iX) K - Means Clustering Approach:



To compare the similarities of two cities, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods 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.



Work Flow: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.

3. Results and Discussion Section

The major purpose of this project, is to suggest a better neighborhood in a new city for the person who are shiffting there. Social presence in society in terms of like minded people. Connectivity to the airport, bus stand, city center, markets and other daily needs things nearby.

1. Sorted list of house in terms of housing prices in a ascending or descending order
2. Sorted list of schools in terms of location, fees, rating and reviews

4. Conclusion Section

In this project, using k-means cluster algorithm I separated the neighborhood into 10(Ten) different clusters and for 103 different lattitude and logitude from dataset, which have very-similar neighborhoods around them. Using the charts above results presented to a particular neighborhood based on average house prices and school rating have been made.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation.
This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools.
The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

5. Future Directions

This project can be continued for making it more precise in terms to find best house in Scarborough. Best means on the basis of all required things(daily needs or things we need to live a better life) around and also in terms of cost effective.