Capstone week 3 Assignment

In [2]:
```
#!pip3 install lxml
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
import numpy as np

print ('libraries imported')
```

```
libraries imported
```

PART ONE

1. setting up the url to access the data from wikipedia

2. parsing the data into html

In [3]:
```
url = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
').text
soup=bs(url,'lxml')
table = str(soup.table)
```

3. reading the html into a pandas dataframe

In [12]:
```
toronto_df = pd.read_html(table)
toronto_df=toronto_df[0]
toronto_df
```

Out[12]:

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
...	...	...	...
175	M5Z	Not assigned	Not assigned
176	M6Z	Not assigned	Not assigned
177	M7Z	Not assigned	Not assigned
178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...
179	M9Z	Not assigned	Not assigned

180 rows × 3 columns

4. droping the 'Not assigned' values , combining neighborhoods, and replacing 'Not assigned' neighborhood value with the corresponding value from Borough column

In [13]:
```
# drop the 'Not assigned' values in Borough column
dropValues = toronto_df[ toronto_df['Borough'] == 'Not assigned' ].index
toronto_df.drop(dropValues , inplace=True)

# join neghibours with same postal area code
# please note that the data is already joined from wikipedia so the following code
is
# not necessary
toronto_df = toronto_df.groupby(['Postal Code','Borough'], sort=False).agg(', '.joi
n)
toronto_df.reset_index(inplace=True)

# changin the not assigned value in Neighbors with the corresponding value in Borou
gh
toronto_df['Neighbourhood'] = np.where(toronto_df['Neighbourhood'] == 'Not assigned
',
                                       toronto_df['Borough'], toronto_df['Neighbour
hood'])

toronto_df
```

Out[13]:

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

103 rows × 3 columns

5. the data frame should have no duplicates, test to see if true

In [14]:
```
duplicateVAlues = toronto_df[toronto_df.duplicated()]
print(duplicateVAlues)
```

```
Empty DataFrame
Columns: [Postal Code, Borough, Neighbourhood]
Index: []
```

6. Print data frame shape

In [15]:
```
toronto_df.shape
```

Out[15]:
```
(103, 3)
```

PART TWO:

---

1. Importing CSV file to obtain Lat, Lon

In [16]:
```
lon_lat = pd.read_csv('http://cocl.us/Geospatial_data')
lon_lat.head(10)
```

Out[16]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848

2. join the longitude and latitude with toronto_df

In [17]:
```
toronto_df=pd.merge(toronto_df,lon_lat,on='Postal Code')
```

In [18]:
```
toronto_df
```

Out[18]:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

103 rows × 5 columns

PART THREE:

1. For the analysis i chose to work with borough that contain 'Etobicoke'

In [27]:
```
Etobicoke_df=toronto_df.loc[toronto_df['Borough'] == 'Etobicoke']
Etobicoke_df.reset_index(inplace = True, drop = True)
Etobicoke_df
```

Out[27]:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
1	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650943	-79.554724
2	M9C	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201
3	M9P	Etobicoke	Westmount	43.696319	-79.532242
4	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724
5	M8V	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321
6	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
7	M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484
8	M9W	Etobicoke	Northwest, West Humber - Clairville	43.706748	-79.594054
9	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
10	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
11	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

2. import classes to analyze and visualize the above data

In [32]:
```
#!pip install geopy
from geopy.geocoders import Nominatim
import matplotlib.cm as cm
import matplotlib.colors as colors
#!pip install sklearn
from sklearn.cluster import KMeans
#!pip install folium
import folium
print('Libraries imported.')
```

```
Libraries imported.
```

3. getting the geo data of Etobicoke

In [33]:
```
address = 'Etobicoke, Toronto, Canada'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Etobicoke are {}, {}.'.format(latitude, longit
ude))
```

```
The geograpical coordinate of Etobicoke are 43.6435559, -79.5656326.
```

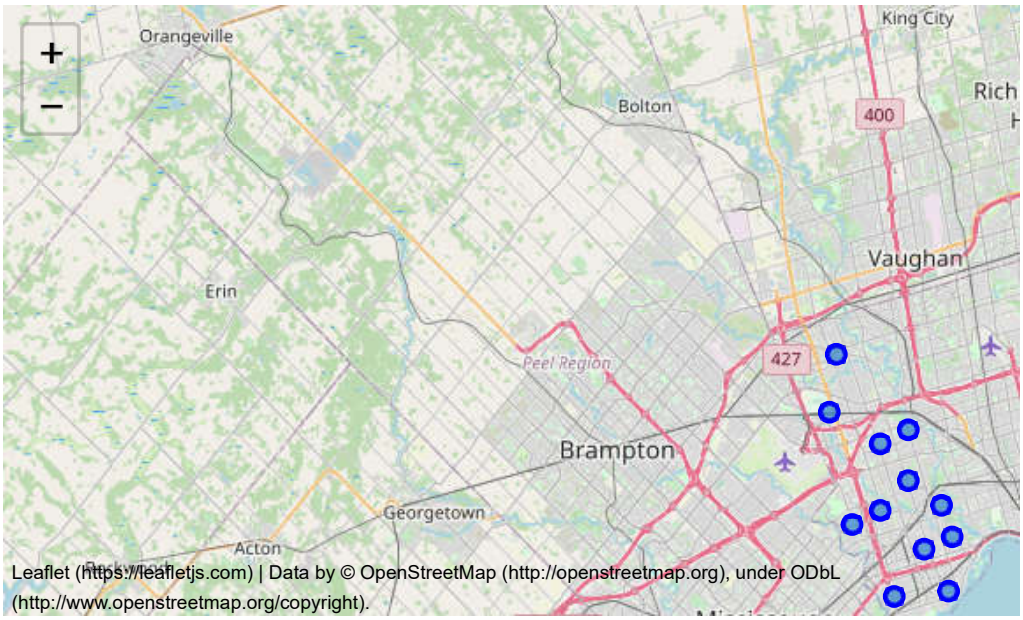
4. create a map of the above geo data

In [52]:
```
map_Etobicoke = folium.Map(location=[latitude, longitude], zoom_start=10)

for lat, lng, borough, neighborhood in zip(Etobicoke_df['Latitude'],
                                           Etobicoke_df['Longitude'],
                                           Etobicoke_df['Borough'],
                                           Etobicoke_df['Neighbourhood']):
    label = '{}, {}'.format(neighborhood, borough)
    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_Etobicoke)

map_Etobicoke
```

Out[52]:



5. Clustering neighborhoods

In [56]:
```
kclusters = 5
Etobicoke_cluster = Etobicoke_df.drop(['Postal Code','Borough','Neighbourhood'], 1)
kmeans = KMeans(n_clusters=kclusters, random_state=0)
kmeans.fit(Etobicoke_cluster)
kmeans.labels_
Etobicoke_df.insert(0, 'Cluster Labels', kmeans.labels_)
```

In [57]:
```
Etobicoke_df
```

Out[57]:

	Cluster Labels	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	3	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
1	0	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650943	-79.554724
2	0	M9C	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe...	43.643515	-79.577201
3	3	M9P	Etobicoke	Westmount	43.696319	-79.532242
4	3	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724
5	4	M8V	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321
6	2	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
7	4	M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484
8	2	M9W	Etobicoke	Northwest, West Humber - Clairville	43.706748	-79.594054
9	1	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
10	1	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
11	1	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

6. visualize clusters

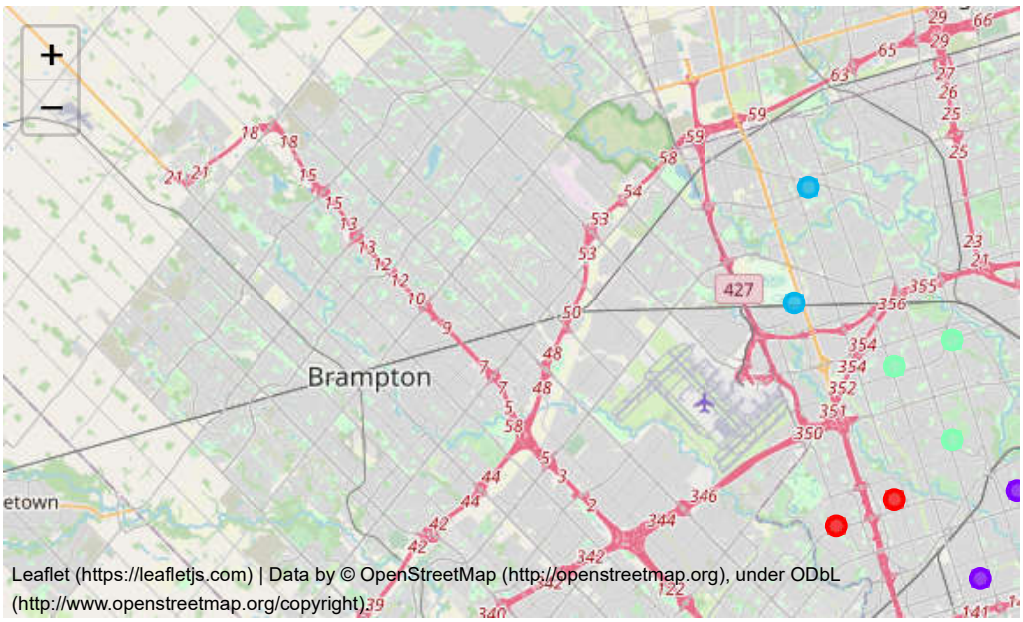
In [60]:
```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)

x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

markers_colors = []
for lat, lon, poi, cluster in zip(Etobicoke_df['Latitude'],
                                  Etobicoke_df['Longitude'],
                                  Etobicoke_df['Neighbourhood'],
                                  Etobicoke_df['Cluster Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

Out[60]:



In [ ]: