This notebook will be mainly used for the capstone project.

I am happy to share my project in data science professional certificate IBM

In [1]:

```
import pandas as pd
import numpy as np
from geopy.geocoders import Nominatim
import matplotlib.cm as cm
import matplotlib.colors as colors
import folium
import requests
import json
from pandas.io.json import json_normalize
from sklearn.cluster import KMeans
```

Building dataframe for Toronto

In [2]:

```
df_PostalCode = pd.read_csv('Postal_Code.csv')
df_PostalCode.head()
```

Out[2]:

	Postal Code	Borough	Neighborhood
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

In [3]:

```
df_PostalCode.shape
```

Out[3]:

```
(103, 3)
```

In [4]:

```
df_geo = pd.read_csv('Geospatial_Coordinates.csv')
df_geo.head()
```

Out[4]:

	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

In [5]:

```
neighborhoods = df_PostalCode.groupby(by = 'Postal Code')
neighborhoods = df_PostalCode.merge(df_geo, how ='left')
neighborhoods
```

Out[5]:

	Postal Code	Borough	Neighborhood	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 969 East...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South,King's Mill Park ,Sunnylea,Humb...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW,The Queensway West,South of Bloor,Ki...	43.628841	-79.520999

103 rows × 5 columns

In [6]:

```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(neighborhoods['Borough'].unique()),
        neighborhoods.shape[0]
    )
)
```

```
The dataframe has 11 boroughs and 103 neighborhoods.
```

Visualization neighborhood in Toronto

In [7]:

```
address = 'Toronto, Ontario'

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

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

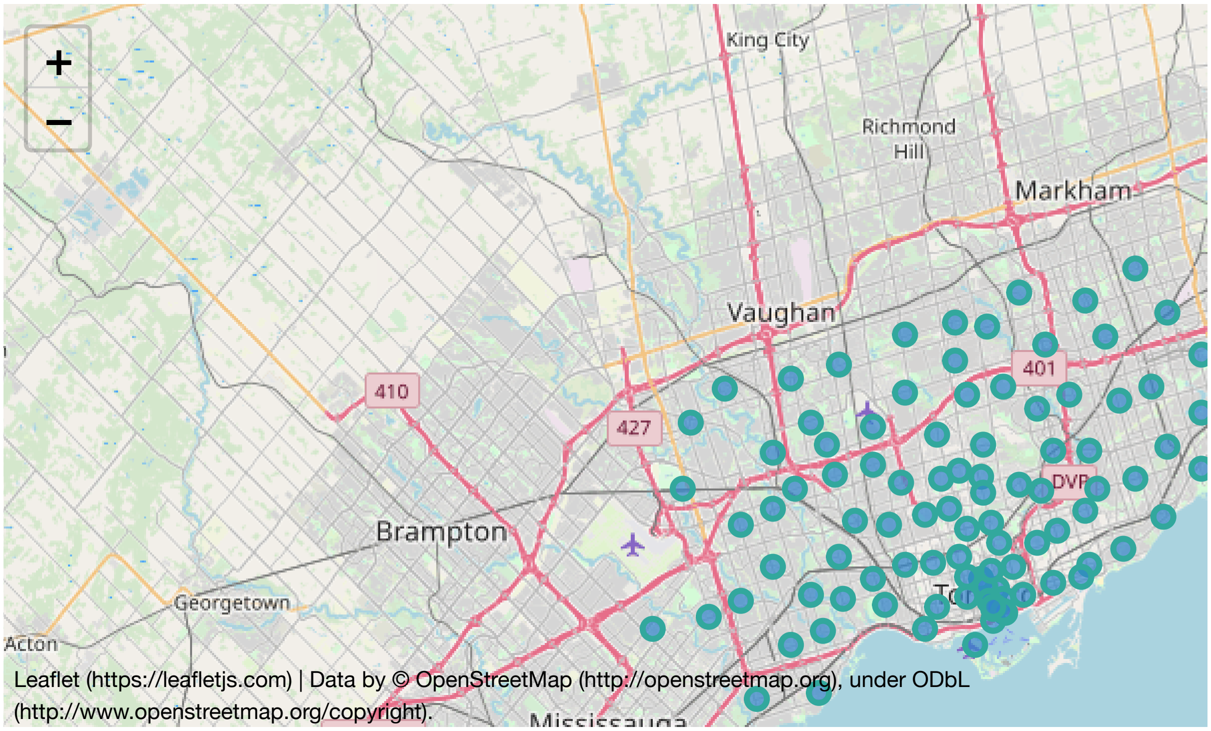
In [8]:

```
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)


for lat, lng, borough, neighborhood in zip(neighborhoods['Latitude'], neighborhoods['Longitude'], neighborhoods['Borough'], neighborhoods['Neighborhood']):
    label = '{}, {}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='#32a89d',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Out[8]:



Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

Explore neighborhoods in North York

In [9]:

```python
NY_df = neighborhoods[neighborhoods['Borough'] == 'North York'].reset_index(drop=True)
NY_df.head()
```

Out[9]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M6A	North York	Lawrence Manor,Lawrence Heights	43.718518	-79.464763
3	M3B	North York	Don Mills,North	43.745906	-79.352188
4	M6B	North York	Glencairn	43.709577	-79.445073

In [10]:

```python
NY_df.shape
```

Out[10]:

```
(24, 5)
```

In [11]:

```python
address = 'North York,Toronto'
geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geographical coordinate of North York Toronto are {}, {}.'.format(latitude, longitude))
```

```
The geographical coordinate of North York Toronto are 43.7543263, -79.44911696639593.
```

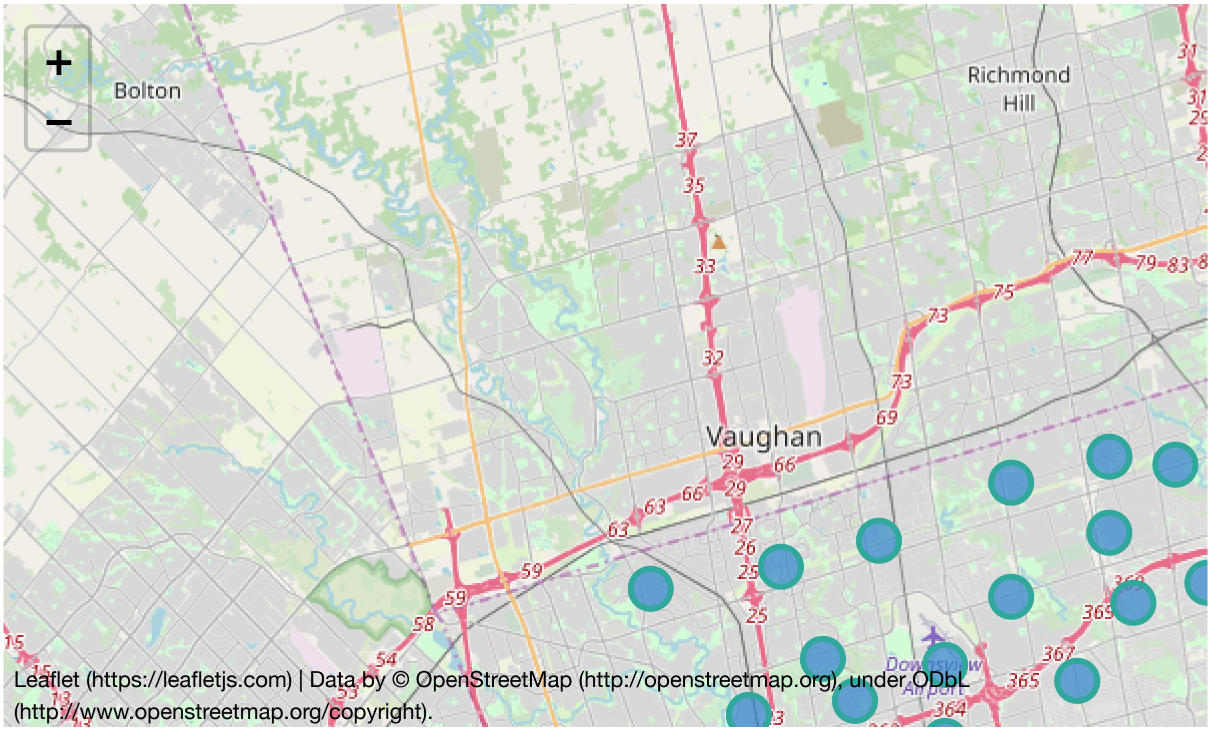
Visualizing neighborhood in North York

In [12]:
```
# create map of North York using latitude and longitude values
map_NY = folium.Map(location=[latitude, longitude], zoom_start=11)

# add markers to map
for lat, lng, label in zip(NY_df['Latitude'], NY_df['Longitude'], NY_df[
'Neighborhood']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        rradius=5,
        popup=label,
        color='#32a89d',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_NY)

map_NY
```

Out[12]:


+
−
Bolton
Richmond Hill
Vaughan
Downsview Airport
Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

In [13]:
```
CLIENT_ID = 'MBBKD0HFFWF5PCLNLKDDZ03TXY23UBCJWUUIGHIXG2VLSMDZ' # your Fo
ursquare ID
CLIENT_SECRET = 'JZWLUUEQYWUQZARLKV2WDPFN0J2IW3GAYQJOQTBDN2CHVQH1' # you
r Foursquare Secret
VERSION = '20203030' # Foursquare API version

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: MBBKD0HFFWF5PCLNLKDDZ03TXY23UBCJWUUIGHIXG2VLSMDZ
CLIENT_SECRET:JZWLUUEQYWUQZARLKV2WDPFN0J2IW3GAYQJOQTBDN2CHVQH1
```

Neighborhoods in Norht York within a radius of 500 meters.

In [14]:

```
LIMIT = 100
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id=
{}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['ite
ms']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for
item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

In [15]:

```
NY_venues = getNearbyVenues(names=NY_df['Neighborhood'],
                            latitudes=NY_df['Latitude'],
                            longitudes=NY_df['Longitude']
                            )
```

```
Parkwoods
Victoria Village
Lawrence Manor,Lawrence Heights
Don Mills,North
Glencairn
Don Mills South,Flemingdon Park
Hillcrest Village
Bathurst Manor,Wilson Heights,Downsview North
Fairview,Henry Farm,Oriole
Northwood Park,York University
Bayview Village
Downsview,CFB Toronto
York Mills,Silver Hills
Downsview West
North Park,Maple Leaf Park,Upwood Park
Humber Summit
Willowdale,Newtonbrook
Downsview,Central
Bedford Park,Lawrence Manor East
Humberlea,Emery
Willowdale South
Downsview Northwest
York Mills West
Willowdale West
```

Venues that are in Norht York

In [16]:

```
print(NY_venues.shape)
NY_venues.head()
```

```
(239, 7)
```

Out[16]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	649 Variety	43.754513	-79.331942	Convenience Store
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

In [17]:

```
print('There are {} uniques categories.'.format(len(NY_venues['Venue Category'].unique())))
```

```
There are 104 uniques categories.
```

In [18]:

```
# one hot encoding
NY_onehot = pd.get_dummies(NY_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
NY_onehot['Neighborhood'] = NY_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [NY_onehot.columns[-1]] + list(NY_onehot.columns[:-1])
NY_onehot = NY_onehot[fixed_columns]

NY_onehot.head()
```

Out[18]:

	Neighborhood	Accessories Store	Airport	American Restaurant	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Bakery	Bank
0	Parkwoods	0	0	0	0	0	0	0	0
1	Parkwoods	0	0	0	0	0	0	0	0
2	Parkwoods	0	0	0	0	0	0	0	0
3	Victoria Village	0	0	0	0	0	0	0	0
4	Victoria Village	0	0	0	0	0	0	0	0

5 rows × 105 columns

In [19]:

```
NY_onehot.shape
```

Out[19]:

```
(239, 105)
```

In [20]:

```
NY_grouped = NY_onehot.groupby('Neighborhood').mean().reset_index()
NY_grouped
```

Out[20]:

	Neighborhood	Accessories Store	Airport	American Restaurant	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Baker
0	Bathurst Manor,Wilson Heights,Downsview North	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
1	Bayview Village	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
2	Bedford Park,Lawrence Manor East	0.000000	0.0	0.041667	0.000000	0.000000	0.000000	0.00000
3	Don Mills South,Flemingdon Park	0.000000	0.0	0.000000	0.000000	0.050000	0.000000	0.00000
4	Don Mills,North	0.000000	0.0	0.000000	0.000000	0.000000	0.166667	0.00000
5	Downsview Northwest	0.000000	0.0	0.000000	0.000000	0.000000	0.250000	0.00000
6	Downsview West	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
7	Downsview,CFB Toronto	0.000000	0.5	0.000000	0.000000	0.000000	0.000000	0.00000
8	Downsview,Central	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
9	Fairview,Henry Farm,Oriole	0.000000	0.0	0.015385	0.000000	0.015385	0.000000	0.03076
10	Glencairn	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
11	Hillcrest Village	0.000000	0.0	0.000000	0.000000	0.000000	0.200000	0.00000
12	Humber Summit	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
13	Humberlea,Emery	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
14	Lawrence Manor,Lawrence Heights	0.083333	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
15	North Park,Maple Leaf Park,Upwood Park	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.25000
16	Northwood Park,York University	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
17	Parkwoods	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
18	Victoria Village	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
19	Willowdale South	0.000000	0.0	0.000000	0.029412	0.000000	0.000000	0.00000
20	Willowdale West	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
21	York Mills West	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000
22	York Mills,Silver Hills	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00000

23 rows × 105 columns

The top10 categories for venues in each neighborhood.

In [21]:

```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

In [22]:

```
num_top_venues = 5

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

# create columns according to number of top venues
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))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = NY_grouped['Neighborhood']

for ind in np.arange(NY_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venue
s(NY_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

Out[22]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bathurst Manor,Wilson Heights,Downsview North	Coffee Shop	Bank	Middle Eastern Restaurant	Shopping Mall	Pizza Place
1	Bayview Village	Japanese Restaurant	Chinese Restaurant	Café	Bank	Women's Store
2	Bedford Park,Lawrence Manor East	Coffee Shop	Sandwich Place	Italian Restaurant	Restaurant	Comfort Food Restaurant
3	Don Mills South,Flemingdon Park	Restaurant	Gym	Beer Store	Coffee Shop	Discount Store
4	Don Mills,North	Japanese Restaurant	Gym / Fitness Center	Athletics & Sports	Caribbean Restaurant	Café

Cluster Neighborhoods

Run *k*-means to cluster the neighborhood into 5 clusters.

In [23]:

```
kclusters = 5

NY_grouped_clustering = NY_grouped.drop('Neighborhood', 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(NY_grouped_clu
stering)

kmeans.labels_[0:10]
```

Out[23]:

```
array([1, 1, 1, 1, 1, 1, 1, 3, 1, 1], dtype=int32)
```

In [24]:

```
neighborhoods_venues_sorted.insert(0, 'Cluster_Labels', kmeans.labels_)

NY_merged =NY_df

NY_merged = NY_merged.join(neighborhoods_venues_sorted.set_index('Neighb
orhood'), on='Neighborhood')

NY_merged.head()
```

Out[24]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	3.0	Park	Food & Drink Shop
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0	Coffee Shop	Hockey Arena
2	M6A	North York	Lawrence Manor,Lawrence Heights	43.718518	-79.464763	1.0	Clothing Store	Furniture / Home Store
3	M3B	North York	Don Mills,North York	43.745906	-79.352188	1.0	Japanese Restaurant	Gym / Fitness Center
4	M6B	North York	Glencairn	43.709577	-79.445073	1.0	Park	Pizza Place

In [25]:

```
NY_merged = NY_merged.fillna(0)
NY_merged.Cluster_Labels.astype(int)
NY_merged.dtypes
```

Out[25]:

```
Postal Code                 object
Borough                     object
Neighborhood                object
Latitude                   float64
Longitude                  float64
Cluster_Labels             float64
1st Most Common Venue       object
2nd Most Common Venue       object
3rd Most Common Venue       object
4th Most Common Venue       object
5th Most Common Venue       object
dtype: object
```

Visualizing the resulting clusters

In [26]:

```
# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11,
control_scale=True)
```

In [27]:

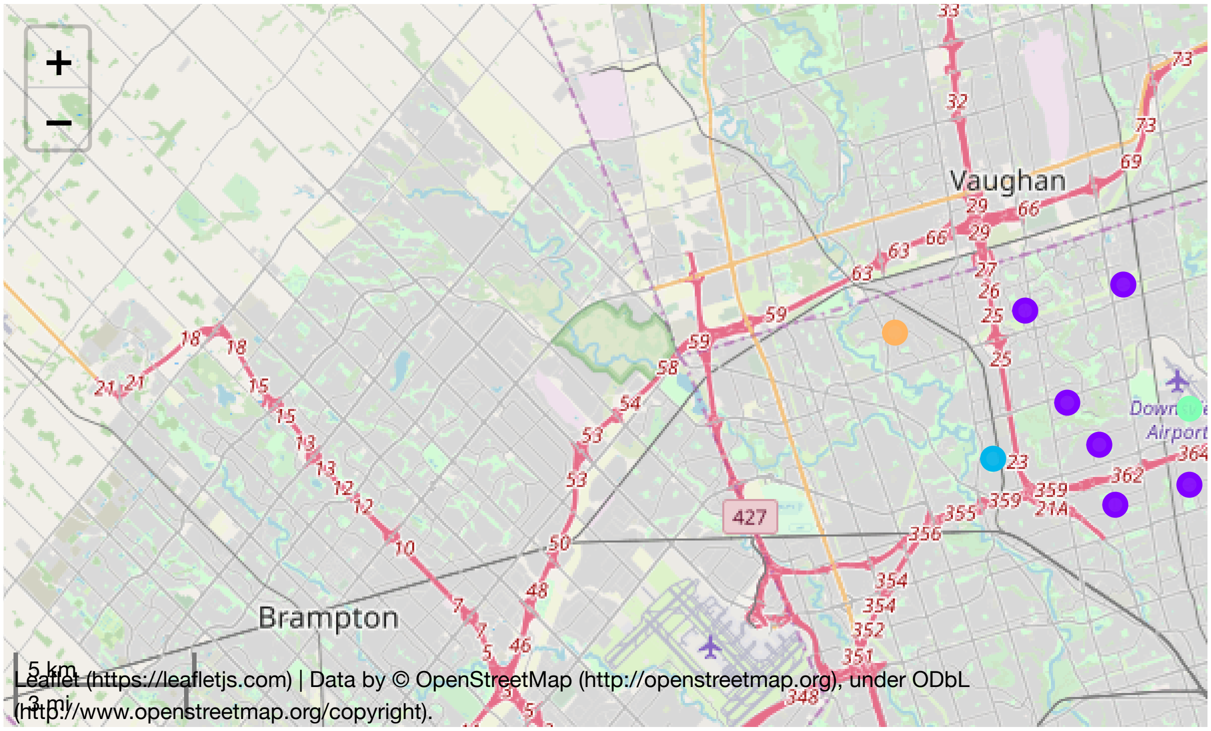
```
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]
```

In [28]:

```
# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(NY_merged['Latitude'], NY_merged['Long
itude'], NY_merged['Neighborhood'], NY_merged['Cluster_Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_ht
ml=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
      # popup=label,
        color = rainbow[int(cluster)-1],
      # color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[int(cluster)-1],
        #fill_color=rainbow[cluster-1],
        fill_opacity=0.9,
        line_opacity=0.2
      ).add_to(map_clusters)

map_clusters
```

Out[28]:


+
−
Vaughan
Brampton
Downsview Airport
427
5 km
3 mi
Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

Cluster 1

In [29]:

```
NY_merged.loc[NY_merged['Cluster_Labels'] == 0,
                         NY_merged.columns[[1] + list(range(5, NY_merge
d.shape[1]))]]
```

Out[29]:

	Borough	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	North York	0.0	Cafeteria	Martial Arts Dojo	Women's Store	Coffee Shop	Concert Hall
16	North York	0.0	0	0	0	0	0

Cluster 2

In [30]:

```
NY_merged.loc[NY_merged['Cluster_Labels'] == 1,
                     NY_merged.columns[[1] + list(range(5, NY_merged.sha
pe[1]))]]
```

Out[30]:

	Borough	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	North York	1.0	Coffee Shop	Hockey Arena	Portuguese Restaurant	Intersection	Diner
2	North York	1.0	Clothing Store	Furniture / Home Store	Women's Store	Miscellaneous Shop	Boutique
3	North York	1.0	Japanese Restaurant	Gym / Fitness Center	Athletics & Sports	Caribbean Restaurant	Café
4	North York	1.0	Park	Pizza Place	Pub	Japanese Restaurant	Women's Store
5	North York	1.0	Restaurant	Gym	Beer Store	Coffee Shop	Discount Store
6	North York	1.0	Golf Course	Mediterranean Restaurant	Athletics & Sports	Pool	Dog Run
7	North York	1.0	Coffee Shop	Bank	Middle Eastern Restaurant	Shopping Mall	Pizza Place
8	North York	1.0	Clothing Store	Coffee Shop	Fast Food Restaurant	Japanese Restaurant	Food Court
9	North York	1.0	Miscellaneous Shop	Massage Studio	Caribbean Restaurant	Bar	Coffee Shop
10	North York	1.0	Japanese Restaurant	Chinese Restaurant	Café	Bank	Women's Store
13	North York	1.0	Park	Grocery Store	Bank	Hotel	Shopping Mall
14	North York	1.0	Park	Construction & Landscaping	Bakery	Basketball Court	Women's Store
17	North York	1.0	Food Truck	Home Service	Baseball Field	Korean Restaurant	Women's Store
18	North York	1.0	Coffee Shop	Sandwich Place	Italian Restaurant	Restaurant	Comfort Food Restaurant
20	North York	1.0	Ramen Restaurant	Coffee Shop	Restaurant	Café	Sandwich Place
21	North York	1.0	Grocery Store	Gym / Fitness Center	Athletics & Sports	Discount Store	Women's Store
23	North York	1.0	Discount Store	Pharmacy	Pizza Place	Bank	Butcher

Cluster 3

In [31]:

```
NY_merged.loc[NY_merged['Cluster_Labels'] == 2,
                     NY_merged.columns[[1] + list(range(5, NY_merged.sha
pe[1]))]]
```

Out[31]:

	Borough	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
19	North York	2.0	Baseball Field	Women's Store	Distribution Center	Concert Hall	Construction & Landscaping

Cluster 4

In [32]:

```
NY_merged.loc[NY_merged['Cluster_Labels'] == 3,
                     NY_merged.columns[[1] + list(range(5, NY_merged.sha
pe[1]))]]
```

Out[32]:

	Borough	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	North York	3.0	Park	Food & Drink Shop	Convenience Store	Diner	Coffee Shop
11	North York	3.0	Airport	Park	Women's Store	Discount Store	Comfort Food Restaurant
22	North York	3.0	Park	Convenience Store	Bank	Women's Store	Discount Store

Cluster 5

In [33]:

```
NY_merged.loc[NY_merged['Cluster_Labels'] == 4,
                     NY_merged.columns[[1] + list(range(5, NY_merged.sha
pe[1]))]]
```

Out[33]:

	Borough	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
15	North York	4.0	Pizza Place	Empanada Restaurant	Women's Store	Diner	Comfort Food Restaurant