Segmenting and Clustering Neighborhoods in Toronto

Scrape the Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe

The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood

Import Libraries

In [1]:

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

In [2]:

```
page = requests.get("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M")
soup = BeautifulSoup(page.content, 'html.parser')
```

In [3]:

```
table = soup.find('tbody')
rows = table.select('tr')
row = [r.get_text() for r in rows]
```

In [4]:

```
df = pd.DataFrame(row)
df.head()
```

Out[4]:

	0
0	\nPostcode\nBorough\nNeighbourhood\n
1	\nM1A\nNot assigned\nNot assigned\n
2	\nM2A\nNot assigned\nNot assigned\n
3	\nM3A\nNorth York\nParkwoods\n
4	\nM4A\nNorth York\nVictoria Village\n

In [5]:

```
df1= df[0].str.split('\n', expand=True)
df1
```

Out[5]:

	0	1	2	3	4
0		Postcode	Borough	Neighbourhood	
1		M1A	Not assigned	Not assigned	
2		M2A	Not assigned	Not assigned	
3		M3A	North York	Parkwoods	
4		M4A	North York	Victoria Village	
5		M5A	Downtown Toronto	Harbourfront	
6		M5A	Downtown Toronto	Regent Park	
7		M6A	North York	Lawrence Heights	
8		M6A	North York	Lawrence Manor	
9		M7A	Queen's Park	Not assigned	
10		M8A	Not assigned	Not assigned	
11		M9A	Etobicoke	Islington Avenue	
12		M1B	Scarborough	Rouge	
13		M1B	Scarborough	Malvern	
14		M2B	Not assigned	Not assigned	
15		M3B	North York	Don Mills North	
16		M4B	East York	Woodbine Gardens	
17		M4B	East York	Parkview Hill	
18		M5B	Downtown Toronto	Ryerson	
19		M5B	Downtown Toronto	Garden District	
20		M6B	North York	Glencairn	
21		M7B	Not assigned	Not assigned	
22		M8B	Not assigned	Not assigned	
23		M9B	Etobicoke	Cloverdale	
24		M9B	Etobicoke	Islington	
25		M9B	Etobicoke	Martin Grove	
26		M9B	Etobicoke	Princess Gardens	
27		M9B	Etobicoke	West Deane Park	
28		M1C	Scarborough	Highland Creek	
29		M1C	Scarborough	Rouge Hill	
...	...	...	...	...	...
259		M9X	Not assigned	Not assigned	
260		M1Y	Not assigned	Not assigned	
261		M2Y	Not assigned	Not assigned	
262		M3Y	Not assigned	Not assigned	
263		M4Y	Downtown Toronto	Church and Wellesley	
264		M5Y	Not assigned	Not assigned	

	0	1	2	3	4
265		M6Y	Not assigned	Not assigned	
266		M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	
267		M8Y	Etobicoke	Humber Bay	
268		M8Y	Etobicoke	King's Mill Park	
269		M8Y	Etobicoke	Kingsway Park South East	
270		M8Y	Etobicoke	Mimico NE	
271		M8Y	Etobicoke	Old Mill South	
272		M8Y	Etobicoke	The Queensway East	
273		M8Y	Etobicoke	Royal York South East	
274		M8Y	Etobicoke	Sunnylea	
275		M9Y	Not assigned	Not assigned	
276		M1Z	Not assigned	Not assigned	
277		M2Z	Not assigned	Not assigned	
278		M3Z	Not assigned	Not assigned	
279		M4Z	Not assigned	Not assigned	
280		M5Z	Not assigned	Not assigned	
281		M6Z	Not assigned	Not assigned	
282		M7Z	Not assigned	Not assigned	
283		M8Z	Etobicoke	Kingsway Park South West	
284		M8Z	Etobicoke	Mimico NW	
285		M8Z	Etobicoke	The Queensway West	
286		M8Z	Etobicoke	Royal York South West	
287		M8Z	Etobicoke	South of Bloor	
288		M9Z	Not assigned	Not assigned	

289 rows × 5 columns

In [6]:

```
df2 = df1.rename(columns=df1.iloc[0])
df2.head()
```

Out[6]:

	Postcode	Borough	Neighbourhood
0	Postcode	Borough	Neighbourhood
1	M1A	Not assigned	Not assigned
2	M2A	Not assigned	Not assigned
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village

In [7]:

```
df1.iloc[0]
```

Out[7]:

```
0
1         Postcode
2          Borough
3    Neighbourhood
4
Name: 0, dtype: object
```

In [8]:

```
df3 = df2.drop(df2.index[0])
df3.head()
```

Out[8]:

	Postcode	Borough	Neighbourhood
1	M1A	Not assigned	Not assigned
2	M2A	Not assigned	Not assigned
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Harbourfront

Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.

More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma.

If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough. So for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighborhood columns will be Queen's Park.

In [9]:

```
df4 = df3[df3.Borough != 'Not assigned']
df4.head(11)
```

Out[9]:

	Postcode	Borough	Neighbourhood
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Harbourfront
6	M5A	Downtown Toronto	Regent Park
7	M6A	North York	Lawrence Heights
8	M6A	North York	Lawrence Manor
9	M7A	Queen's Park	Not assigned
11	M9A	Etobicoke	Islington Avenue
12	M1B	Scarborough	Rouge
13	M1B	Scarborough	Malvern
15	M3B	North York	Don Mills North

In [10]:

```
df5 = df4.groupby(['Postcode', 'Borough'], sort = False).agg(','.join)
df5.reset_index(inplace = True)
df5.head()
```

Out[10]:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront,Regent Park
3	M6A	North York	Lawrence Heights,Lawrence Manor
4	M7A	Queen's Park	Not assigned

In [11]:

```
for index, row in df5.iterrows():
    if row["Neighbourhood"] == "Not assigned":
        row["Neighbourhood"] = row["Borough"]

df5.head()
```

Out[11]:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront,Regent Park
3	M6A	North York	Lawrence Heights,Lawrence Manor
4	M7A	Queen's Park	Queen's Park

.shape method to print the number of rows of your dataframe.

In [12]:

```
df5.shape
```

Out[12]:

```
(103, 3)
```

we need to get the latitude and the longitude coordinates of each neighborhood.

In [13]:

```
coordinates = pd.read_csv("Geospatial_Coordinates.csv")
coordinates.rename(columns={"Postal Code": "Postcode"}, inplace=True)
df6 = df5.merge(coordinates, on="Postcode", how="left")
df6.head()
```

Out[13]:

	Postcode	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	Harbourfront,Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

Explore and cluster the neighborhoods in Toronto, will work only boroughs that contain the word Toronto

In [14]:

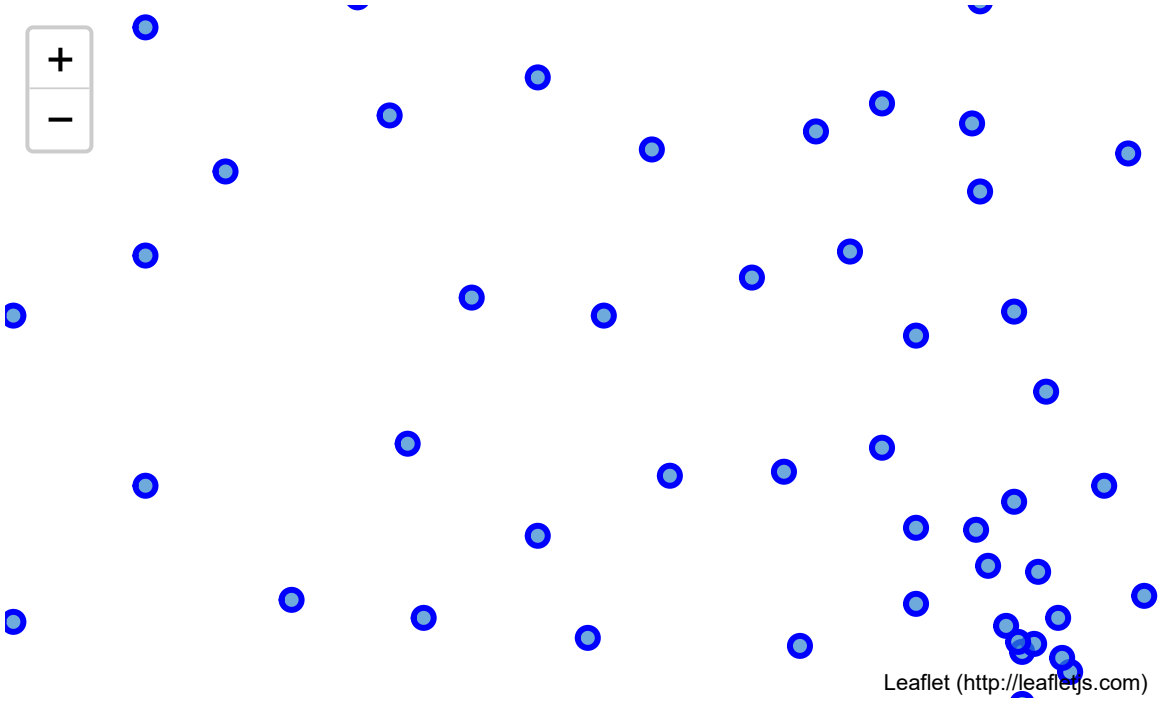
```
#!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim

import matplotlib.cm as cm
import matplotlib.colors as colors

from sklearn.cluster import KMeans

#!conda install -c conda-forge folium=0.5.0 --yes
import folium
```

In [15]:

```
address = 'Toronto'

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


# create map of Toronto using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=12)

# add markers to map
for lat, lng, borough, neighborhood in zip(df6['Latitude'], df6['Longitude'], df6['Boro
ugh'], df6['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).add_to(map_toronto)

map_toronto
```

```
The geograpical coordinate of Toronto are 43.653963, -79.387207.
```

Out[15]:



In [16]:

```
# filter borough names that contain the word Toronto
borough_names = list(df6.Borough.unique())

borough_with_toronto = []

for x in borough_names:
    if "toronto" in x.lower():
        borough_with_toronto.append(x)

borough_with_toronto
```

Out[16]:

```
['Downtown Toronto', 'East Toronto', 'West Toronto', 'Central Toronto']
```

In [17]:

```
df7 = df6[df6['Borough'].isin(borough_with_toronto)].reset_index(drop=True)
df7.head()
```

Out[17]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M5A	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636
1	M5B	Downtown Toronto	Ryerson,Garden District	43.657162	-79.378937
2	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
3	M4E	East Toronto	The Beaches	43.676357	-79.293031
4	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306

In [ ]:

```

```