Segmenting and Clustering Neighborhoods of the City of Toronto

Introduction:

Toronto is a big city with a quite good number of neighborhoods. The neighborhoods will be segmented and clustered to group venues. We will use Foursquare API to explore the neighborhoods. To get most common venue categories in each neighborhood, we will use "explore" function. The neighborhoods can be grouped into clusters by using k-means clustering algorithm. Then using Folium library, the neighborhoods and its clusters can be visualized.

Steps to complete the task:

A. Import following libraries.
   a. Pandas as pd for data analysis
   b. Numpy as np to handle data in a vectorized manner.
   c. Json to handle JSON files.
   d. Nominatim from geopy.geocoders for coverting address to latitude and longitude.
   e. Request to handle request
   f. Matplot library for ploting
   g. KMeans from sklearn.cluster
   h. Folium for rendering maps

B. Downloading and Exploring Dataset.
   In order to segmenting and exploring the neighborhoods we need a dataset that contains all the boroughs of Toronto and the neighborhoods that exists in each borough including their latitude and longitude.

C. Scraping the web page for necessary tables.

D. Cleaning and transform the data to usable format.

E. Two dataset- one contains Postal code, borough and neighborhoods and the other contains latitude and longitude. We merged these two sets of data to get a single data frame.

F. Using KMeans clustering algorithm the data set segmented and clustered.

After scraping the web page we got the table below which shows each group of postal code, borough and neighborhoods e.g all M1 group in column-0, M2 group in column-1 and so on.

	0	1	2	3	4	5	6	7	8
0	M1ANot assigned	M2ANot assigned	M3ANorth York(Parkwoods)	M4ANorth York(Victoria Village)	M5ADowntown Toronto(Regent Park / Harbourfront)	M6ANorth York(Lawrence Manor / Lawrence Heights)	M7AQueen's Park(Ontario Provincial Government)	M8ANot assigned	M9AEtobicoke(Islington Avenue)
1	M1BScarborough(Malvern / Rouge)	M2BNot assigned	M3BNorth York(Don Mills)North	M4BEast York(Parkview Hill / Woodbine Gardens)	M5BDowntown Toronto(Garden District, Ryerson)	M6BNorth York(Glencairn)	M7BNot assigned	M8BNot assigned	M9BEtobicoke(West Deane Park / Princess Garden...
2	M1CScarborough(Rouge Hill / Port Union / Highl...	M2CNot assigned	M3CNorth York(Don Mills)South(Flemingdon Park)	M4CEast York(Woodbine Heights)	M5CDowntown Toronto(St. James Town)	M6CYork(Humewood-Cedarvale)	M7CNot assigned	M8CNot assigned	M9CEtobicoke(Eringate / Bloordale Gardens / Ol...
3	M1EScarborough(Guildwood / Morningside / West ...	M2ENot assigned	M3ENot assigned	M4EEast Toronto(The Beaches)	M5EDowntown Toronto(Berczy Park)	M6EYork(Caledonia-Fairbanks)	M7ENot assigned	M8ENot assigned	M9ENot assigned
4	M1GScarborough(Woburn)	M2GNot assigned	M3GNot assigned	M4GEast York(Leaside)	M5GDowntown Toronto(Central Bay Street)	M6GDowntown Toronto(Christie)	M7GNot assigned	M8GNot assigned	M9GNot assigned

We appended all columns in a single column. The code and data's screen shot is given below;

```
number_of_col = new_table.shape[1]
for i in range(number_of_col):
    new_table.rename(columns={i: 'test'},inplace = True)
    df1=pd.DataFrame(new_table['test'])
    df = df.append(df1)
    new_table.rename(columns={'test': i}, inplace = True)
df
```

	test
0	M1ANot assigned
1	M1BScarborough(Malvern / Rouge)
2	M1CScarborough(Rouge Hill / Port Union / Highl...
3	M1EScarborough(Guildwood / Morningside / West ...
4	M1GScarborough(Woburn)
5	M1HScarborough(Cedarbrae)
6	M1JScarborough(Scarborough Village)
7	M1KScarborough(Kennedy Park / Ionview / East B...
8	M1LScarborough(Golden Mile / Clairlea / Oakridge)
9	M1MScarborough(Cliffside / Cliffcrest / Scarbo...
10	M1NScarborough(Birch Cliff / Cliffside West)

We then split each row into three columns; PostalCode, Borough and Neighborhoods. After proper cleaning we the dataset as below.

	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
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	Kennedy Park , Ionview , East Birchmount Park
7	M1L	Scarborough	Golden Mile , Clairlea , Oakridge
8	M1M	Scarborough	Cliffside , Cliffcrest , Scarborough Village West
9	M1N	Scarborough	Birch Cliff , Cliffside West
10	M1P	Scarborough	Dorset Park , Wexford Heights , Scarborough To...

We get the latitude and longitude from different dataset and merged with the above data sets to have the tidy data for creating map and final clustering. The code snippet and the data frame is given below;

86]:
```
location = pd.read_csv(path_coordinates)
location.rename(columns ={'Postal Code': 'PostalCode'}, inplace = True)
location.head(3)
```

86]:

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711

Merging two dataframe on column 'PostalCode'

87]:
```
Toronto_Postal_Location = pd.merge(newdf, location)
Toronto_Postal_Location.head()
```

87]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern , Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill , Port Union , Highland Creek	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

Then we render the map of Toronto with itts venue points as below:

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

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

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

```
92]: # create map of New York using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, borough, neighborhood in zip(Toronto_Postal_Location['Latitude'], Toronto_Postal_Location['Longitude'], Toronto_Pos
    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_toronto)

map_toronto
```

And got the map rendered as below;

Let us see Central Toronto neighborhood

[94]:
```
central_toronto_data = Toronto_Postal_Location[Toronto_Postal_Location['Borough'] == 'Central Toronto'].reset_index(drop=True)
central_toronto_data.head()
```

[94]:

	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
4	M4T	Central Toronto	Moore Park , Summerhill East	43.689574	-79.383160

Let us find the restaurant in the neighborhood.

	Neighborhood	Venue	Venue Category
18	North Toronto West	Sushi Shop	Restaurant
20	North Toronto West	Tio's Urban Mexican	Mexican Restaurant
24	North Toronto West	C'est Bon	Chinese Restaurant
25	North Toronto West	A&W	Fast Food Restaurant
31	North Toronto West	Roberto's	Italian Restaurant
37	Davisville	Marigold Indian Bistro	Indian Restaurant
40	Davisville	Zee Grill	Seafood Restaurant
43	Davisville	Sakae Sushi	Sushi Restaurant
44	Davisville	Florentia Ristorante	Italian Restaurant
45	Davisville	Positano	Italian Restaurant
47	Davisville	Thai Spicy House	Thai Restaurant
52	Davisville	Hokkaido Sushi	Sushi Restaurant
55	Davisville	souvlaki express	Greek Restaurant
59	Davisville	Starving Artist	Restaurant
66	Davisville	Chef Upstairs	American Restaurant
73	Summerhill West , Rathnelly , South Hill , For...	Daeco Sushi	Sushi Restaurant
74	Summerhill West , Rathnelly , South Hill , For...	Mary Be Kitchen	Restaurant
75	Summerhill West , Rathnelly , South Hill , For...	Union Social Eatery	American Restaurant

Number of Venues in each neighborhoods

Let's check how many venues are present in each neighborhood

```
central_toronto_venues.groupby('Neighborhood').count()
```

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue
Davisville	34	34	34
Davisville North	12	12	12
Forest Hill North & West	4	4	4
Lawrence Park	4	4	4
Moore Park , Summerhill East	2	2	2
North Toronto West	19	19	19
Roselawn	4	4	4
Summerhill West , Rathnelly , South Hill , Forest Hill SE , Deer Park	14	14	14
The Annex , North Midtown , Yorkville	19	19	19

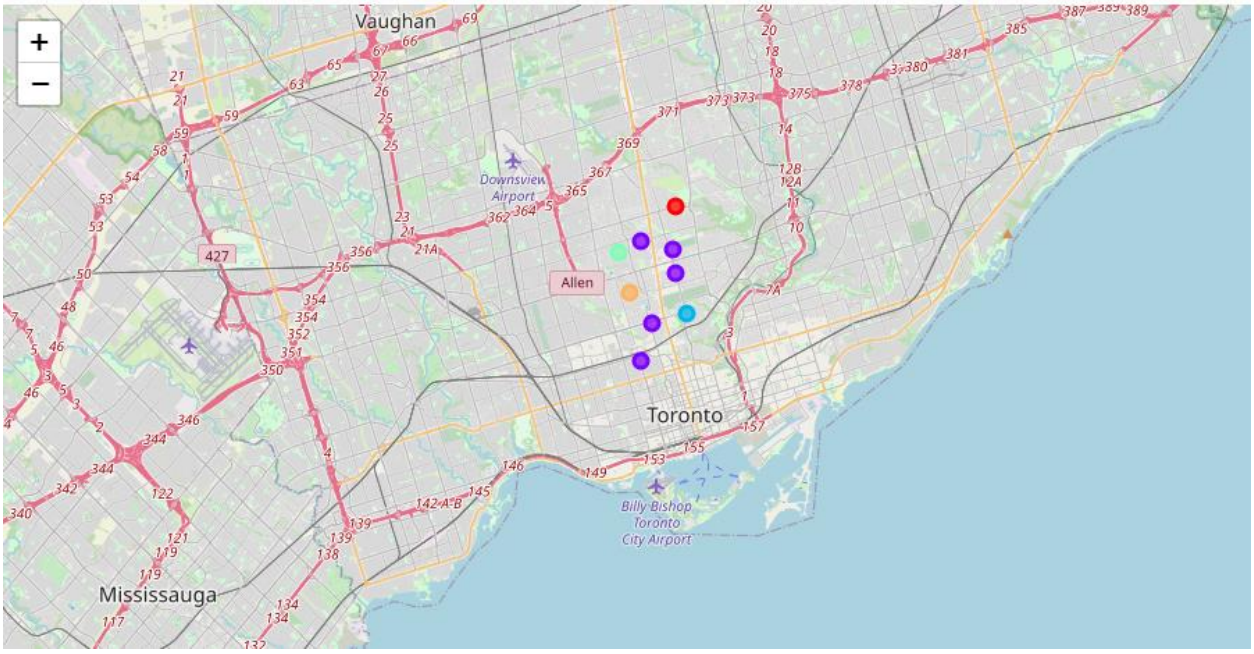
Cluster Neighborhoods

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

8]:
```
# set number of clusters
kclusters = 5

central_toronto_grouped_clustering = central_toronto_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(central_toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

8]:
```
array([1, 1, 4, 0, 2, 1, 3, 1, 1])
```

Map of the cluster



Summary : We collected data from website, then scrapped to get the relevant table and using Foursquare API and Folium library got the exact location of restaurant including rendering maps. Then using KMeans clustering algorithm venues are clustered into five groups.