Battle of Neighborhoods

Neighborhood segmentation and clustering in Toronto

Author: Anmol Bansal

Libraries Used and data Acquisition

```
In [1]: import numpy as np
        import pandas as pd
        import json
        import requests
        from pandas.io.json import json_normalize
        import matplotlib.cm as cm
        import matplotlib.colors as colors
        from sklearn.cluster import KMeans
        from bs4 import BeautifulSoup

In [2]: from geopy.geocoders import Nominatim

In [3]: import folium # map rendering library

In [4]: import wikipedia as wd
        html = wd.page("List of postal codes of Canada: M").html().encode("UTF-8")
        df = pd.read_html(html)[0]
        df_drop = df[df.Borough != "Not assigned"].reset_index(drop=True)
        toronto_df_grouped = df_drop.groupby(["Postal Code", "Borough"], as_index=False).agg(lambda x : ",".join(x))
        for index, row in toronto_df_grouped.iterrows():
            if row["Neighborhood"] == "Not assigned":
                row["Neighborhood"] = row["Borough"]
        print(toronto_df_grouped.shape)
        toronto_df_grouped.rename(columns={"Postal Code": "PostalCode"}, inplace=True)
        toronto_df_grouped.head()

        (103, 3)
```

Merged data using geopy library

Creating a pandas dataframe with all three old details along with latitudes and longitudes of neighborhoods for Foursquare API

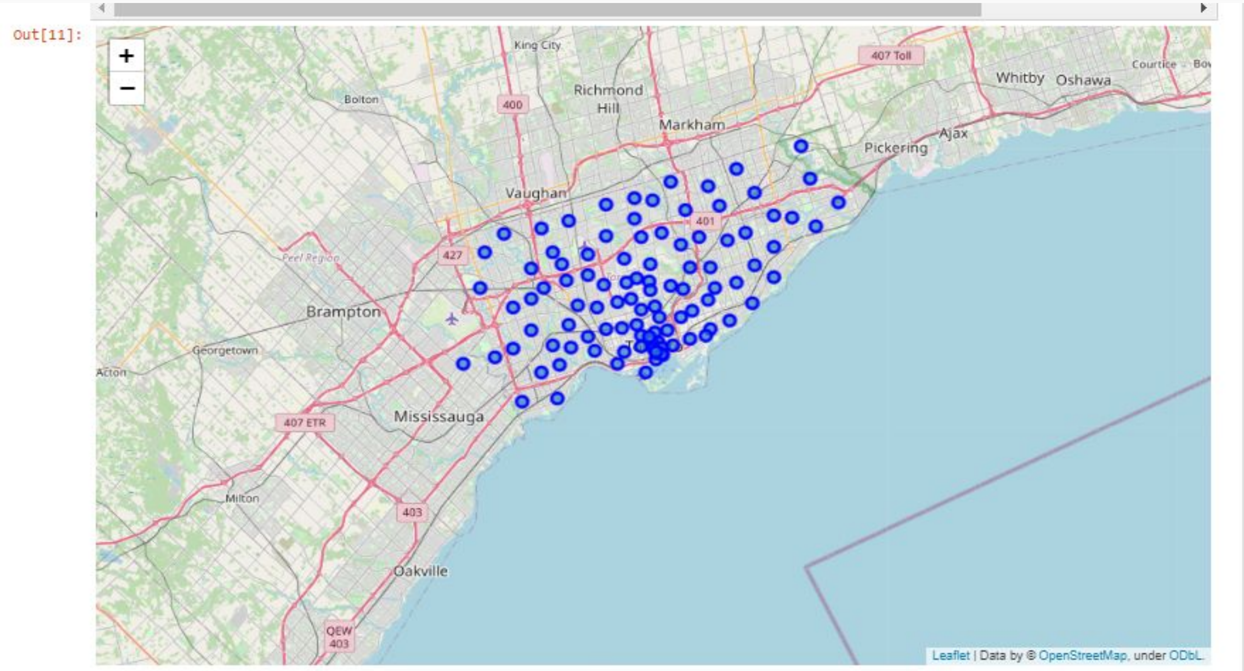
Now merging the data

In [8]:
```
toronto_df_new = toronto_df_grouped.merge(coordinates, on="PostalCode", how="left")
toronto_df_new.head()
```

Out[8]:

	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

Data Visualization Using folium library



Adding Markers on all the grouped neighborhoods based on Boroughs. Folium is a powerful map data visualization tool.

Defining Foursquare Parameters for API calls

Foursquare is a robust API to access geospatial data.

Define Foursquare Credentials and Version

In [15]:

```
CLIENT_ID = 'CVUZFVHSE2ZM2NR40I3DCX4K3MPCEU2G4FCHA4KSZUQSU335' # your Foursquare ID
CLIENT_SECRET = 'PLUVUAQLSYRK5KTQTAWTBTJES0NGPGB2FPLSOH0AALNPBHRX'  # your Foursquare Secret
VERSION = '20180605' # Foursquare API version

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

```
Your credentails:
CLIENT_ID: CVUZFVHSE2ZM2NR40I3DCX4K3MPCEU2G4FCHA4KSZUQSU335
CLIENT_SECRET:PLUVUAQLSYRK5KTQTAWTBTJES0NGPGB2FPLSOH0AALNPBHRX
```

One-Hot encode Categorical Parameters

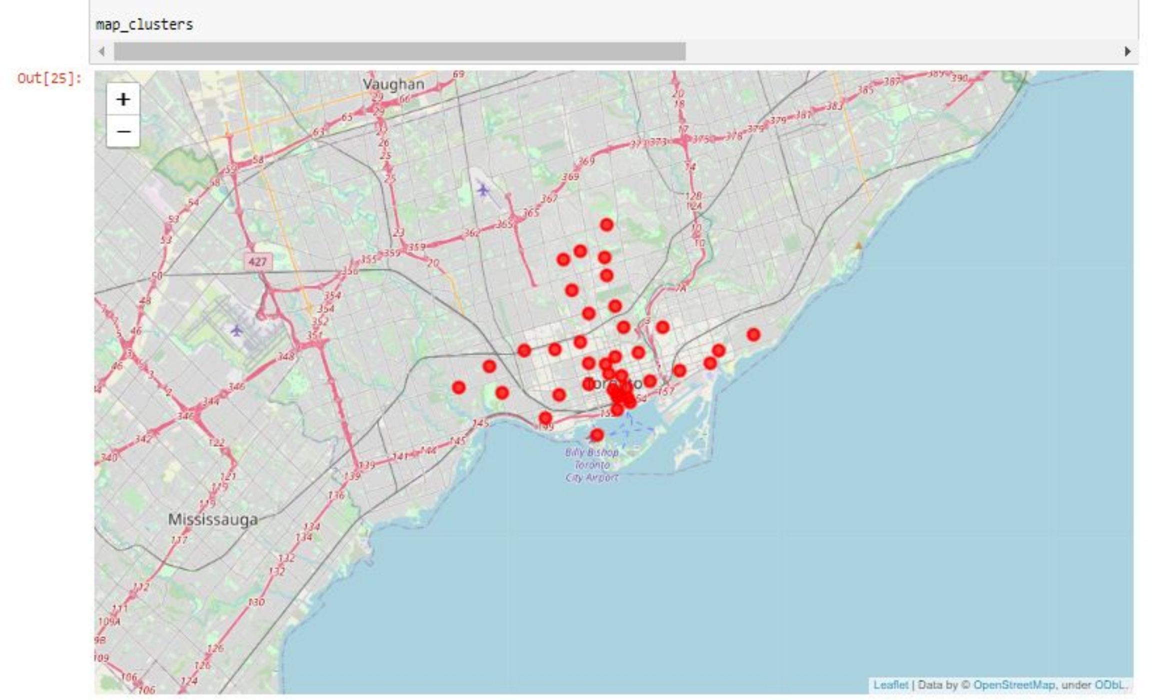
One hot encoding allow us to convert categorical values to numeric values for easy calculations.

Analysing each area

In [19]:
```
# one hot encoding
toronto_onehot = pd.get_dummies(venues_df[['VenueCategory']], prefix="", prefix_sep="")

# add postal, borough and neighborhood column back to dataframe
toronto_onehot['PostalCode'] = venues_df['PostalCode']
toronto_onehot['Borough'] = venues_df['Borough']
toronto_onehot['Neighborhoods'] = venues_df['Neighborhood']

# move postal, borough and neighborhood column to the first column
fixed_columns = list(toronto_onehot.columns[-3:]) + list(toronto_onehot.columns[:-3])
toronto_onehot = toronto_onehot[fixed_columns]

print(toronto_onehot.shape)
toronto_onehot.head()
```

```
(1677, 36)
```

Visualization of clusters via Folium



Clusters are plotted and marked using foliuj, now K-means clustering can be used to separate neighborhoods into k clusters based on similarity.

K-means Clustering of neighborhoods

CLUSTERING

```
In [22]: kclusters = 5

         toronto_grouped_clustering = toronto_grouped.drop(["PostalCode", "Borough", "Neighborhoods"], 1)

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

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

         C:\Users\Anmol\Anaconda3\envs\gammavishwanathan\lib\site-packages\sklearn\cluster\k_means_.py:971: ConvergenceWarning: Number o
         f distinct clusters (1) found smaller than n_clusters (5). Possibly due to duplicate points in X.
           return_n_iter=True)

Out[22]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Clusters

Cluster 2

In [27]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[1] + \
                                                      list(range(5, toronto_merged.shape[1]))]]
```

Out[27]:

Borough	Cluster Labels	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

Cluster 3

In [28]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1] + \
                                                      list(range(5, toronto_merged.shape[1]))]]
```

Out[28]:

Borough	Cluster Labels	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

Cluster 4

In [29]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[1] + \
                                                      list(range(5, toronto_merged.shape[1]))]]
```

Conclusion

Cluster 5

In [30]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4, toronto_merged.columns[[1] + \
                                                                                 list(range(5, toronto_merged.shape[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	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue

Conclusion

Most of the neighborhoods fall into Cluster 1 which are the areas with cafe, restaurants, supermarkets etc

In [ ]:

Thanks!

Neighborhoods in Toronto are segmented based on similarity of their neighborhoods.