The Battle of Neighborhoods

May 31, 2020

1. Introduction:-

1.1 Background:-

The purpose of this project is to recommend places for people who are traveling too much and need to know what is the best place to go and visit.

For now, the project will be just for Toronto. Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario’s northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower. Toronto also has many green spaces, from the orderly oval of Queen’s Park to 400-acre High Park and its trails, sports facilities, and zoo.

1.2 The Problem:-

The problem is that we want to get all places in Toronto and arrange them so when anyone wants to visit the city we easily give him the best places including coffee shops, restaurants, hotels, etc.

1.3 Foursquare API:-

As I mention I want to get the data for all places in Toronto and one of the best places I can get the data from is Foursquare API so I will use it to grab all the data I need.

1.4 Clustering:-

I should use clustering to be able to divide places into clusters.

1.5 Other important things:-

There are other important things I will use such as Python to get and handle the data and other libraries. The other libraries are:

Pandas: For creating data frames.

Folium, Seaborn, and Matplotlib: for visualization and showing the map.

Beautiful Soup and requests: for scraping data and tables from websites and handle the requests and the data.

Scikit Learn: to import K-Means for clustering.

2.The Data:-

2.1 Data Description:-

Here I will use these datasets:-

First: (List of postal codes of Canada: M) From Wikipedia. This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Data Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Second: this is a link to a CSV file that has the geographical coordinates of each postal code: **Data Link:** http://cocl.us/Geospatial_data

2.2 Foursquare API Data:-

And for sure I'll need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus, and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

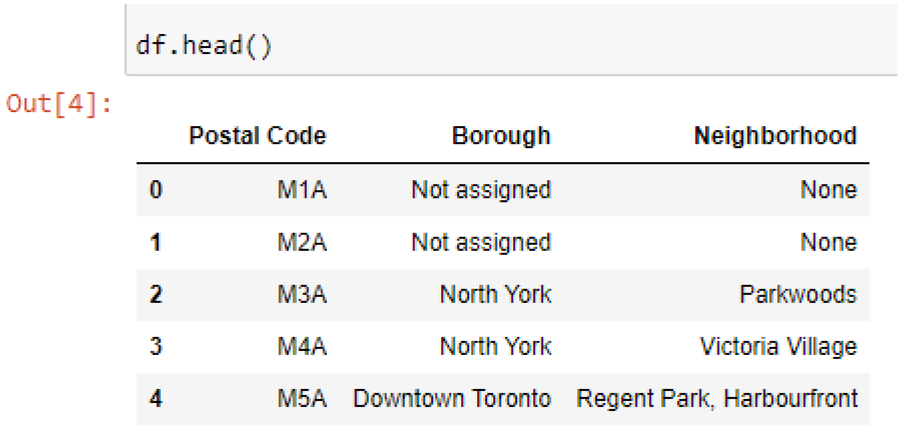
After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

3.Exploratory Data Analysis:-

This is the raw data:-



```
df.head()
```

Out[4]:

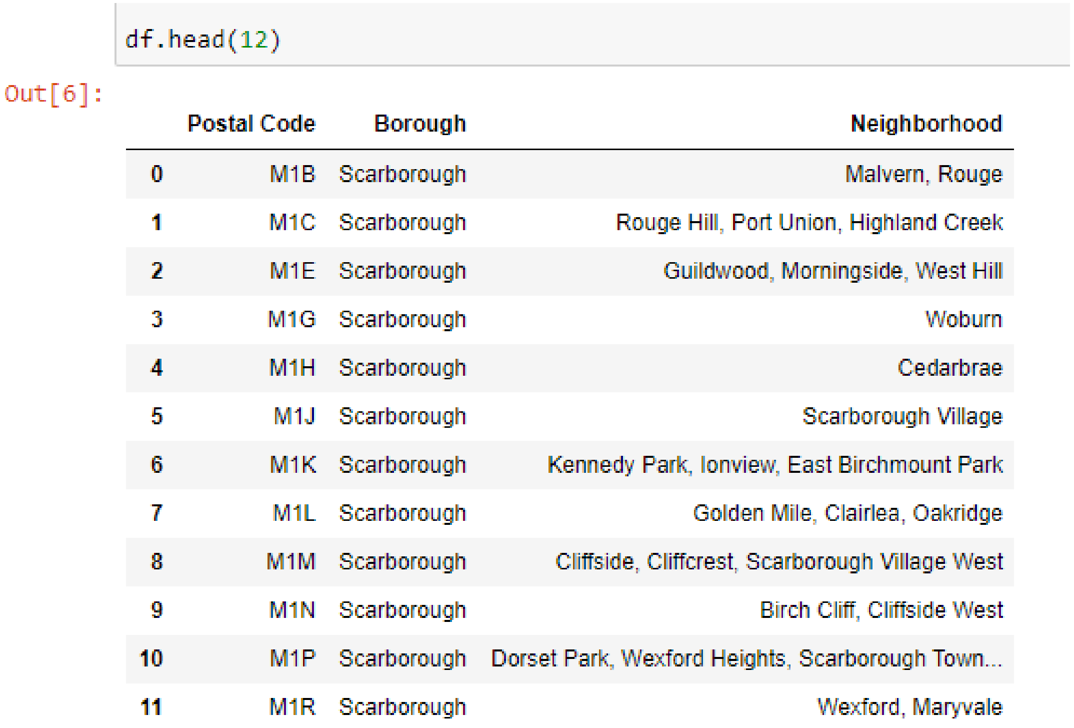
	Postal Code	Borough	Neighborhood
0	M1A	Not assigned	None
1	M2A	Not assigned	None
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

3.1 Cleaning the data:-

First, I ignored cells with a borough that isn't assigned (Null values)

Second, I grouped the data by [Postal Code, Borough] to be cleaner and not to be duplicated values.

And this is the data:-



```
df.head(12)
```

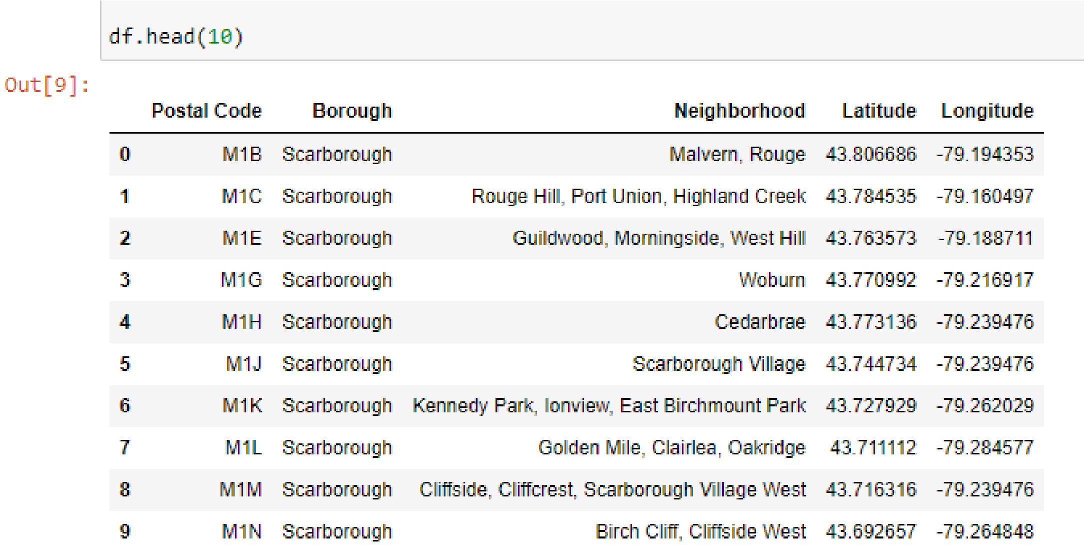
Out[6]:

	Postal Code	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 Town...
11	M1R	Scarborough	Wexford, Maryvale

The shape of the data is (103, 3)

Now I need the coordinates (Latitude and Longitude) to be able to draw the map so is imported the data which have the coordinates. Then I merged the raw data with the coordinates data.

And this is the data now:-



```
df.head(10)
```

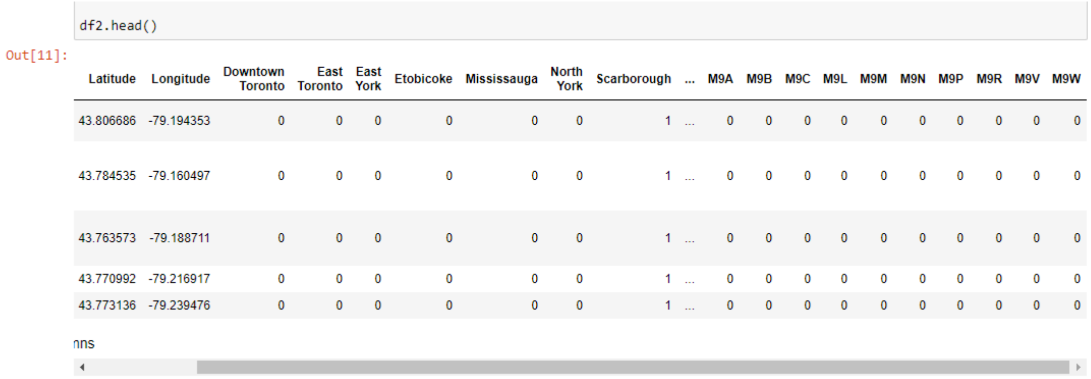
Out[9]:

	Postal Code	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
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848

I should now deal with categorical columns [Postal Code and Borough] so I convert it to be one hot encoder.

I don't deal with Neighborhood columns because I'll drop it.

And this is the data after all these things:-



```
df2.head()
```

Out[11]:

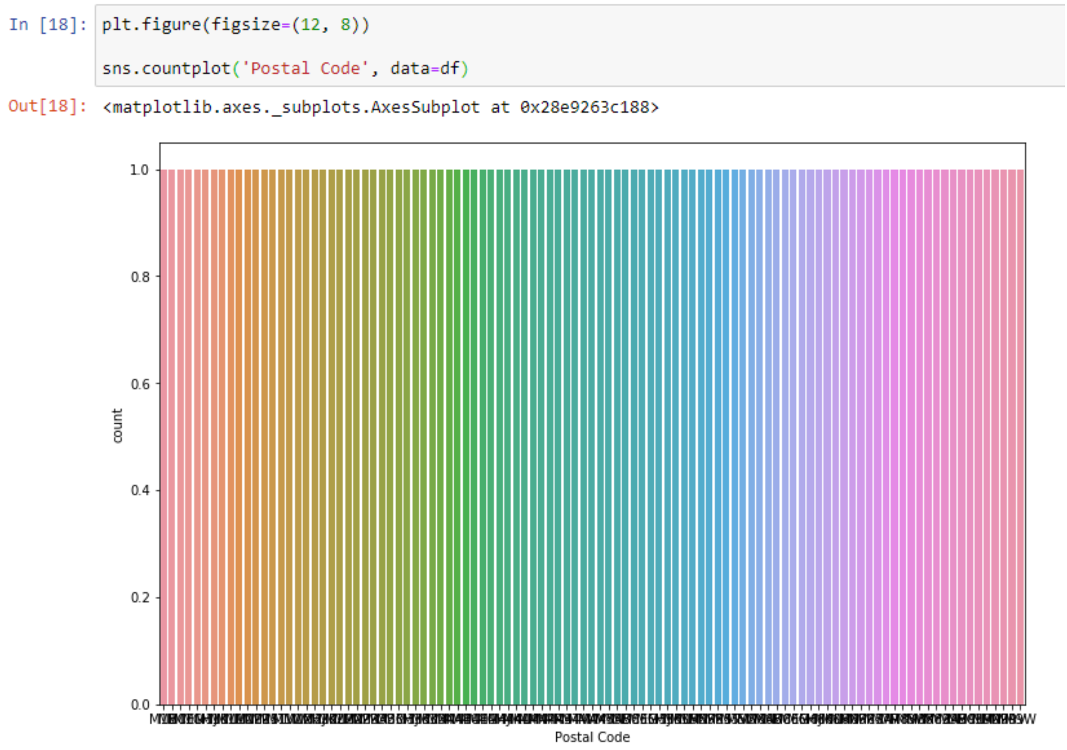
Latitude	Longitude	Downtown Toronto	East Toronto	East York	Etobicoke	Mississauga	North York	Scarborough	...	M9A	M9B	M9C	M9L	M9M	M9N	M9P	M9R	M9V	M9W
43.806686	-79.194353	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
43.784535	-79.160497	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
43.763573	-79.188711	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
43.770992	-79.216917	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
43.773136	-79.239476	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0

nns

4.Data Visualization:-

This visualize for Postal Code:-

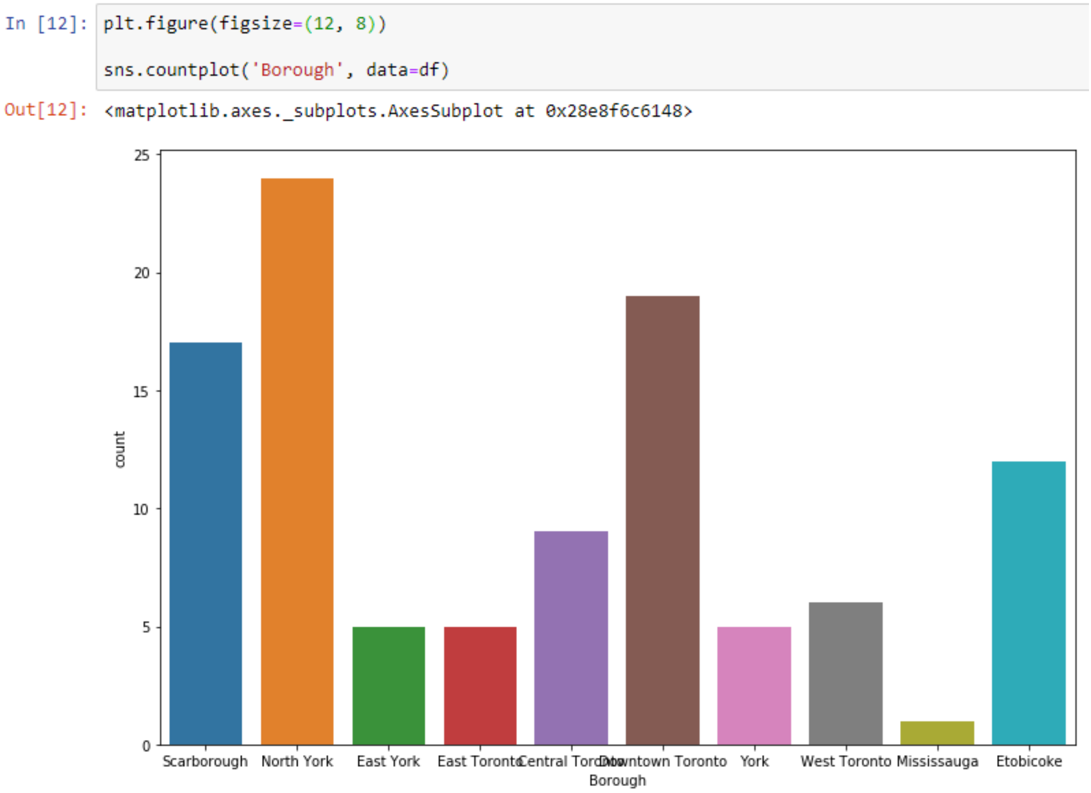
After seeing it we will notice that there is a lot of Postal Code.



```
In [18]: plt.figure(figsize=(12, 8))

         sns.countplot('Postal Code', data=df)

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x28e9263c188>
```

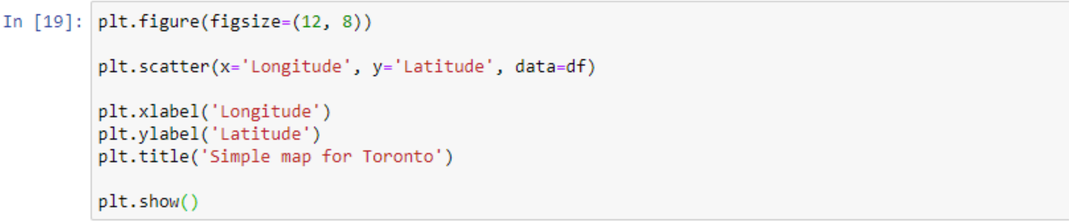
This visualize for Borough:-

And this is more clear and clean than Postal code.

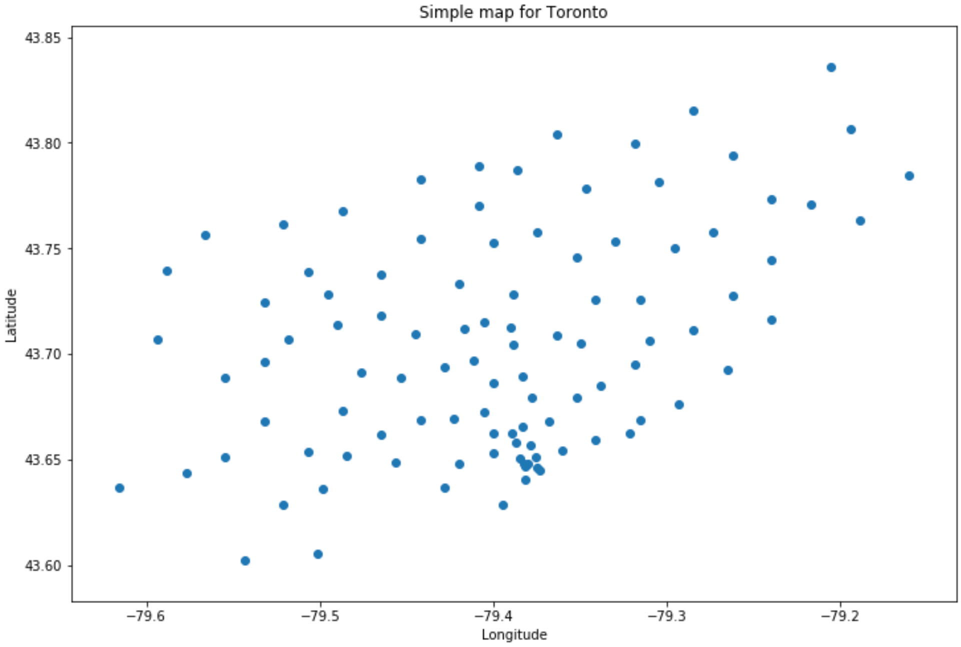


```
In [12]: plt.figure(figsize=(12, 8))

         sns.countplot('Borough', data=df)

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x28e8f6c6148>
```

And for last I made a simple map for Toronto by using the longitude and latitude columns:-



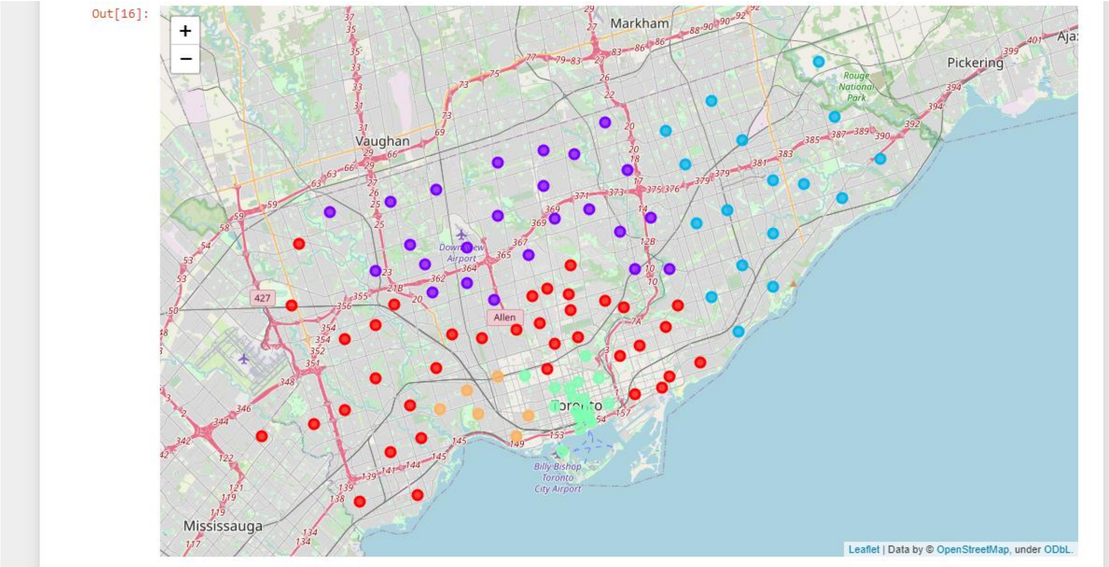
```
In [19]: plt.figure(figsize=(12, 8))

         plt.scatter(x='Longitude', y='Latitude', data=df)

         plt.xlabel('Longitude')
         plt.ylabel('Latitude')
         plt.title('Simple map for Toronto')

         plt.show()
```



5. Building the model:-

Now I built the model and visualize the clear map for Toronto with the cluster on it.

And this is the result:-



Out[16]:

BY: Mohamed Ashraf Gaber