Analysis on Toronto Neighborhoods

Clustering Toronto Neighborhoods into Clusters of Neighborhoods That are Located at the City Center, Middle between Center and Outer Part of the City, and Outer Border.

1. Introduction

Toronto is the largest city in Canada in terms of number of population and diversity of cultural backgrounds. As of the 2021 statistic, it has around 2.6 millions of population, followed by Montreal with around 1.6 millions. It is also the financial capital in Canada sharing the similar function as New York. Due to this fact it has appealed not only home buyer within the country to own a house in Canada as well foreign real estate investor to invest in.

In this report, we will cluster neighborhoods in Toronto based on degree of occurrence of top two types of retail store in a neighborhood into three clusters, cluster of neighborhoods that are located in the city center, in between city center, outer border of the city of Toronto. Specifically, this report will be targeting to home buyers or investors whether be in local and foreign wishes to own a house in a neighborhood based on the location.

2. Data

2.1 Data Description

In light of the issue stated above, the following data is required:

- Information of number of neighborhoods in Toronto and their corresponding coordinate and borough in which it belongs to. Fortunately, Wikipedia does provide the neighborhoods list with borough it belongs to and postal code, and their respective latitude and longitude are provided from the readily available Geocode package.

- Foursquare API that contains the information of retail stores in a neighborhood. Types of retails store, name and coordinate. In this analysis, I am only interested in five types of retail store, supermarket, plaza, shopping mall, department store, and big box store.

2.2 Data Preparation

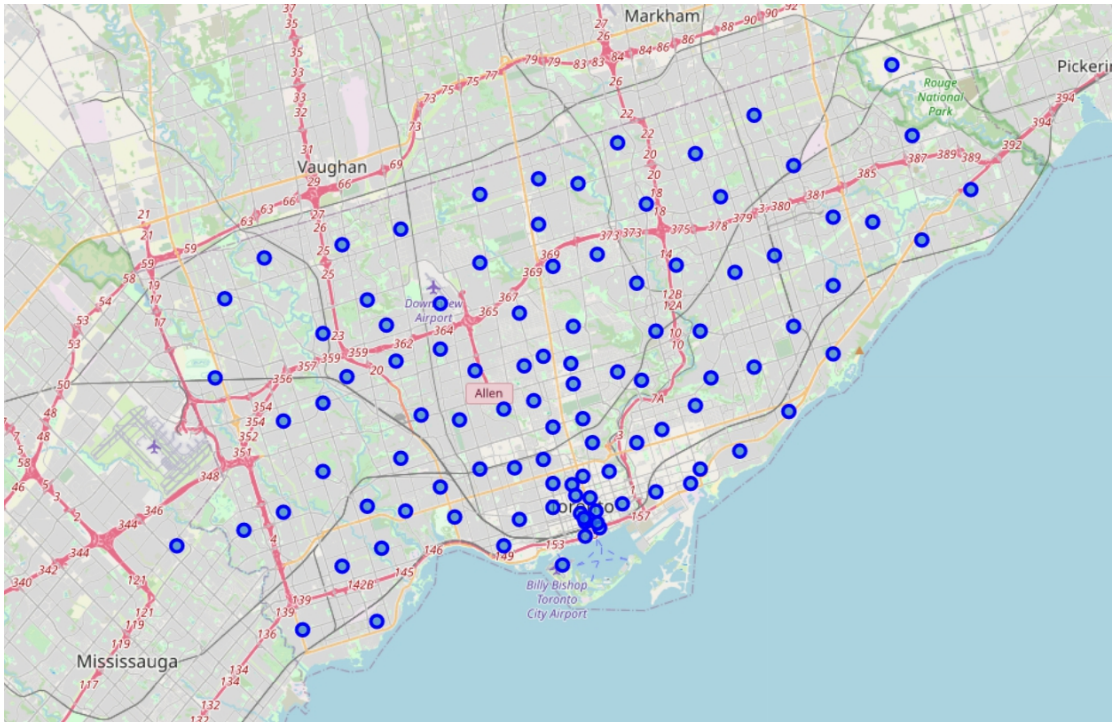
The data obtained initially are from three different sources. The list of neighborhoods is scrapped from Wikipedia using the library of BeautifulSoup, coordinate of each neighborhood from readily available Geocode list, and retail stores information located in each neighborhood extracted through API call on Foursquare. The primary task is to merge these information into a single dataset.

The first task is to combine the two datasets from Wikipedia and readily available Geocode. Once the two datasets have been merged, it has 103 samples or neighborhoods with five columns of *PostalCode*, *Borough*, *Neighborhood*, *Latitude*, and *Longitude*.

	PostalCode	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	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	Enclave of M4L	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

103 rows × 5 columns

The 103 neighborhoods can be visualized in the following map plot created by folium library.



In the Foursquare part, I have set the radius to 8km from the center grid of each neighborhood. The category ID for the API call is 4bf58dd8d48988d1fd941735, which is the category ID assigned to shopping mall in Foursquare. As mentioned earlier that we are only interested in shopping mall, plaza, supermarket, department store, and big box store, other non relevant retail store types that are called from Foursquare based on the category ID will be excluded from the analysis. Here is the summary of the dataset when the dataset obtained from Foursquare and the previous merged dataset.

```
Toronto_Shop = Toronto_venues1[Toronto_venues1['Venue Category'].isin(['Shopping Mall','Plaza','Supermarket','Department Store','Big Box Store'])]
Toronto_Shop
```

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Parkwoods	43.753259	-79.329656	CF Shops at Don Mills	43.735011	-79.344787	Shopping Mall
1	North York	Parkwoods	43.753259	-79.329656	CF Fairview Mall	43.777981	-79.344397	Shopping Mall
2	North York	Parkwoods	43.753259	-79.329656	Bayview Village	43.768527	-79.385494	Shopping Mall
3	North York	Parkwoods	43.753259	-79.329656	Scarborough Town Centre	43.775277	-79.257413	Shopping Mall
4	North York	Parkwoods	43.753259	-79.329656	Leaside Village	43.705682	-79.360777	Shopping Mall
...	...	...	...	...	...	...	...	...
4997	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Galleria Shopping Centre	43.667592	-79.442053	Shopping Mall
4998	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Rockwood Mall	43.625328	-79.604702	Shopping Mall
4999	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Sino Mall	43.597794	-79.594803	Shopping Mall
5001	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Crossways Mall	43.656775	-79.452498	Shopping Mall
5003	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Walmart	43.656229	-79.435530	Big Box Store

4275 rows × 8 columns

```
# Number of stores under each retail store type
Toronto_Shop['Venue Category'].value_counts()
```

```
Shopping Mall       3999
Plaza                101
Supermarket           72
Department Store      53
Big Box Store         50
Name: Venue Category, dtype: int64
```

Together there are 4275 rows with only 5 types of retail store, that are shopping mall, plaza, supermarket, department store, and big box store.

The next part is to identify the top two retail store types in each neighborhood based on the frequency of occurrence. The process required that the data is in integer or floating type. So, I have to create a dummy variable for each retail store type. The mean of occurrence will be the frequency of occurrence to identify the top retail store types.

	Borough	Neighborhood	Big Box Store	Department Store	Plaza	Shopping Mall	Supermarket
0	Central Toronto	Davisville	0.016393	0.016393	0.032787	0.918033	0.016393
1	Central Toronto	Davisville North	0.015873	0.015873	0.031746	0.920635	0.015873
2	Central Toronto	Forest Hill North & West	0.018182	0.018182	0.036364	0.909091	0.018182
3	Central Toronto	Lawrence Park	0.000000	0.000000	0.018182	0.963636	0.018182
4	Central Toronto	Moore Park, Summerhill East	0.017857	0.017857	0.035714	0.928571	0.000000
...	...	...	...	...	...	...	...
98	York	Caledonia-Fairbanks	0.018868	0.018868	0.037736	0.905660	0.018868
99	York	Del Ray, Mount Dennis, Keelsdale and Silverthorn	0.027027	0.000000	0.027027	0.918919	0.027027
100	York	Humewood-Cedarvale	0.018519	0.018519	0.037037	0.907407	0.018519
101	York	Runnymede, The Junction North	0.029412	0.000000	0.000000	0.970588	0.000000
102	York	Weston	0.000000	0.000000	0.000000	1.000000	0.000000

103 rows × 7 columns

Now, I can use the information above to identify the top two types of retail store for each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue
0	Davisville	Shopping Mall	Plaza
1	Davisville North	Shopping Mall	Plaza
2	Forest Hill North & West	Shopping Mall	Plaza
3	Lawrence Park	Shopping Mall	Supermarket
4	Moore Park, Summerhill East	Shopping Mall	Plaza
...	...	...	...
98	Caledonia-Fairbanks	Shopping Mall	Plaza
99	Del Ray, Mount Dennis, Keelsdale and Silverthorn	Shopping Mall	Supermarket
100	Humewood-Cedarvale	Shopping Mall	Plaza
101	Runnymede, The Junction North	Shopping Mall	Big Box Store
102	Weston	Shopping Mall	Supermarket

103 rows × 3 columns

The dataset is ready for clustering using K-Nearest Neighbors algorithm.

3. Methodology

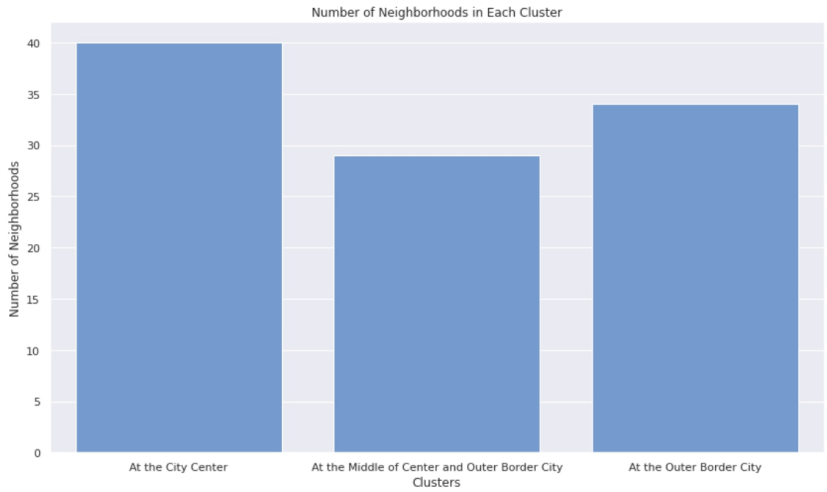
3.1 K-Nearest Neighbor

The main function of the K-Nearest Neighbors algorithm is to cluster neighborhoods that have similar characteristic into clusters of neighborhoods. As mentioned in the beginning that we will use the top two retail store types based on the frequency of occurrence as the characteristic for clustering.

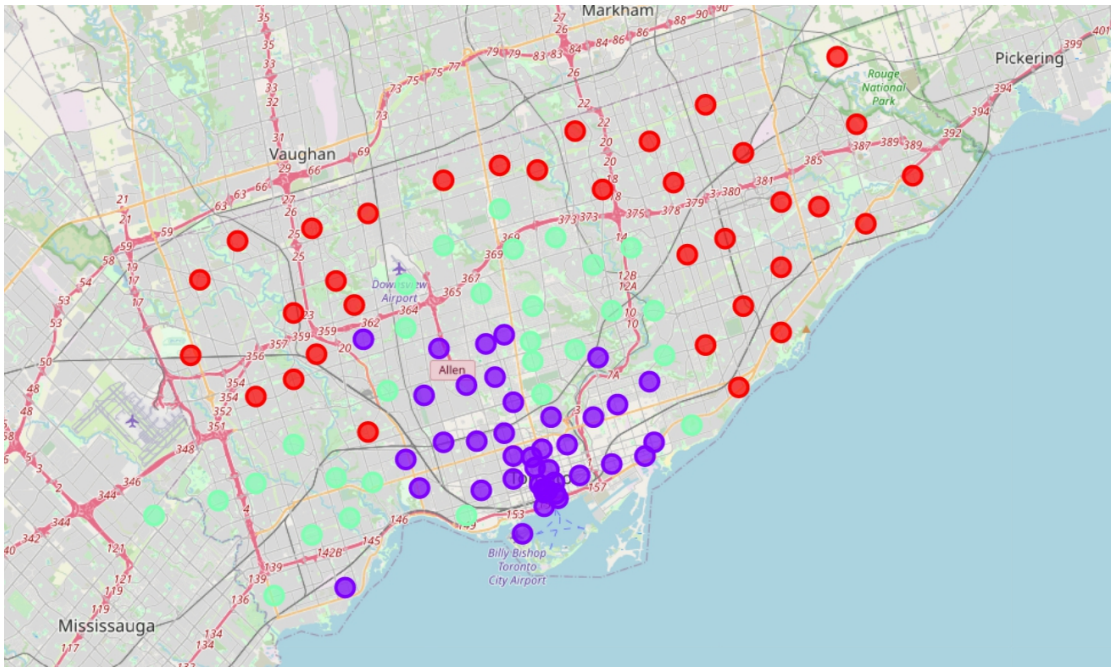
Neighborhoods in Toronto will be clustered into three categories, neighborhoods in the center of the city, in between center and outer part of the city, and outer border.

4. Finding

The three clusters have been plotted on the map. The purple cluster is located at the center of the city with 40 neighborhoods in it, followed by light green with 29 neighborhoods and 34 neighborhoods located in the red cluster which is the outermost of the Toronto city.



We can visualize the neighborhoods on the map plot using folium library.



5. Discussion

With the help of K-Nearest Neighbors, we are able to cluster neighborhoods into three clusters. The first cluster which is purple colour is located at the city center, light green colour is located in between the city center and the outer border of the city of Toronto, and cluster located at the outer border of the Toronto city which is in red colour.

The analysis could help buyers and investors to decide which neighborhoods they are interested based on the location whether be in center or outer part of the Toronto city. Apart of that it is as well could act as a proxy or reference on measuring the relative average housing price between the

clusters. On average, the neighborhoods in the purple cluster will will have relative higher average housing prices as compared to clusters which are in red and light green colour.

6. Conclusion

The analysis report has clustered neighborhoods into three clusters of neighborhoods, which are neighborhoods that are in the city center, cluster that is located at the middle between the city center and outer border, and outer border of Toronto city. Now, is all down to home buyers or investors both local and foreign to decided based on their preference on location in which cluster of neighborhoods they are interested in, be in city center or outer part.