The Battle of Neighborhoods: Toronto, CA

Author: Sandip Gupta **Date: 16/08/2020**

1. Introduction

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America.

1.1 Problem

The city's population grew by 4 per cent (96,073 residents) between 1996 and 2001. Goal is to understand and divide the neighborhoods of the city based on the population and features (such as Hospitals, Schools, Parks, Banks and Shopping Malls). This would also help us categorize neighborhoods into commercial and residential zones.

1.2 Interest

Exercise could help municipal governments understand the features of residential and commercial areas. Also how features affect population of neighborhoods and what areas require certain features. Based on distinction a clear strategy can be taken up to provide services as per the neighborhood type.

2. Data acquisition, cleaning and pre-processing

1.3 Data Source

By scraping Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M we will get all names of neighborhoods in Toronto along with their Postal Code. We will get the coordinates of each neighborhood from http://cocl.us/Geospatial_data. We will use Foursquare API (https://foursquare.com) to get the data on nearby Hospitals, Schools, Shopping Malls, Parks and Banks in each Neighborhood based on their coordinates. We will source neighborhood population data of census 2016 from https://www12.statcan.gc.ca segregated by Postal Code.

1.4 Data cleaning & pre-processing

i. Dropping rows where Postal Code as not been assigned to a Borough from table scrapped from Wikipedia link.

ii. Removing '\n' from all table data as it has appeared by default in every cell.

iii. If a cell has a borough but a 'Not assigned' neighborhood, then assigning neighborhood same as borough.

iv. Neighborhood dataframe is merged with Coordinates dataframe with Postal Code as reference feature.

	PostalCode	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

v. Using Foursquare API search query to fetch data of various features and all feature are merged into a single data frame.

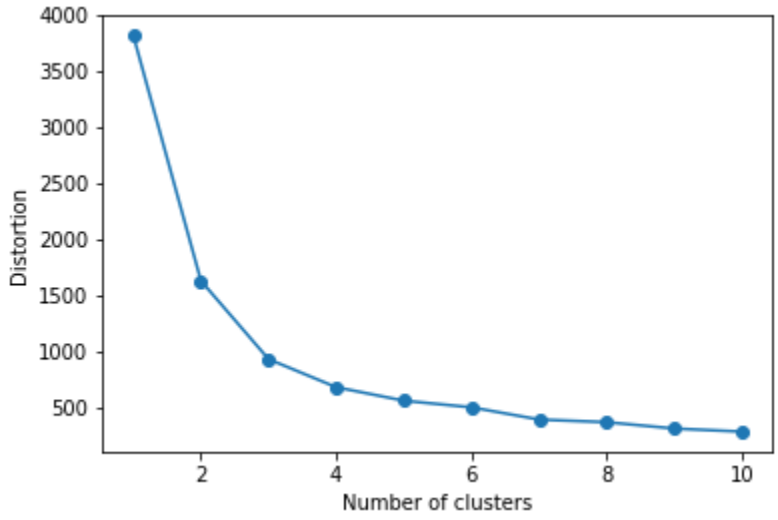
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Place Name	Place Latitude	Place Longitude	Category
0	Regent Park, Harbourfront	43.654260	-79.360636	Bay Cat Hospital	43.655393	-79.358540	Hospital
1	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Women's College Hospital	43.661491	-79.387602	Hospital
2	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Toronto General Hospital	43.658762	-79.388292	Hospital
3	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Mount Sinai Hospital Women's and Infants' Depa...	43.659612	-79.390761	Hospital
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Mount Sinai Hospital, Joseph and Wolf Lebovic ...	43.658247	-79.391473	Hospital

vi. Finally, dataframe is one hot encoded 'Categories' feature and grouped by 'Neighborhood' feature.

	Neighborhood	Bank	Hospital	Park	School	Shopping Mall
0	Agincourt	0	0	0	0	2
1	Alderwood, Long Branch	2	0	0	1	1
2	Bathurst Manor, Wilson Heights, Downsview North	2	0	3	1	0
3	Bayview Village	0	0	1	0	1
4	Bedford Park, Lawrence Manor East	0	1	0	2	0
5	Berczy Park	9	0	5	0	1

3. Methodology

As we need to perform unsupervised clustering, we selected the **K-means Clustering** as the clustering algorithm. To get accurate result from K-means clustering we need to determine the best value for 'K', which stands for Number of clusters by the elbow method.



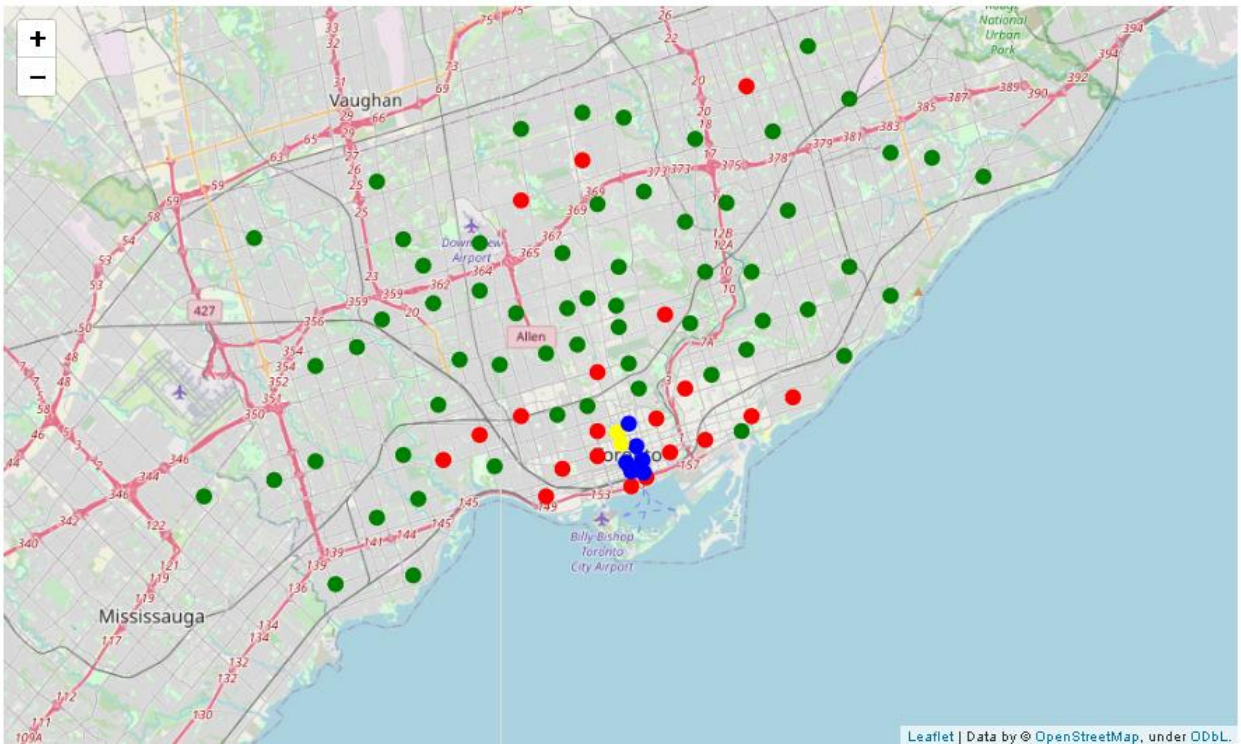
Thus, a line plot was generated with value of ‘K’ ranging from 1 to 10. From the above plot we can confirm 4 as the ideal number clusters.

Cluster labels were generated and added to the dataframe.

	Cluster Labels	Neighborhood	Bank	Hospital	Park	School	Shopping Mall
0	2	Agincourt	0	0	0	0	2
1	2	Alderwood, Long Branch	2	0	0	1	1
2	0	Bathurst Manor, Wilson Heights, Downsview North	2	0	3	1	0
3	2	Bayview Village	0	0	1	0	1
4	2	Bedford Park, Lawrence Manor East	0	1	0	2	0

Main dataframe and cluster label table were merged and neighborhood map with color coded clusters was generated.

	PostalCode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Bank	Hospital	Park	School	Shopping Mall
0	M3A	North York	Parkwoods	43.753259	-79.329656	2	0.0	0.0	1.0	0.0	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	2	0.0	0.0	0.0	1.0	0.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	4.0	1.0	4.0	0.0	0.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	2	0.0	0.0	0.0	2.0	0.0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	3	5.0	22.0	5.0	1.0	1.0



From the above cluster map we can infer that clusters have a pattern. Neighborhoods closer to the city center are clustered together and neighborhoods in the outskirts have similar features. This could be because city center would have more commercial establishments like banks and hospitals.

On further examination of each cluster we can list out the following findings:

i. Cluster 0 & 2 (red & green) neighborhoods have more schools & parks which indicates residential areas.
ii. Cluster 1 (blue) neighborhoods have more Banks nearby which indicates that these neighborhoods are mostly commercial.
iii. Cluster 3 (yellow) neighborhoods have more hospitals. High numbers also may indicate hospital in this neighborhood may have multiple wings and other medical facilities.

4. Results

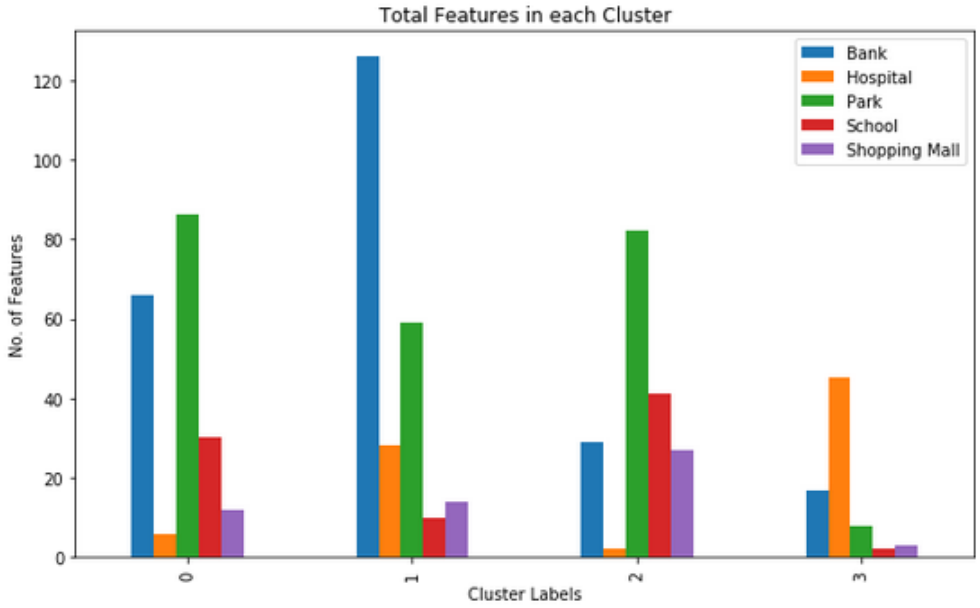
To confirm above finding we downloaded population data from Statcan website segregated Postal Code wise into a dataframe. Features 'Population 2016' & 'Total dwellings' were merged to the main dataframe by referencing Postal Code.

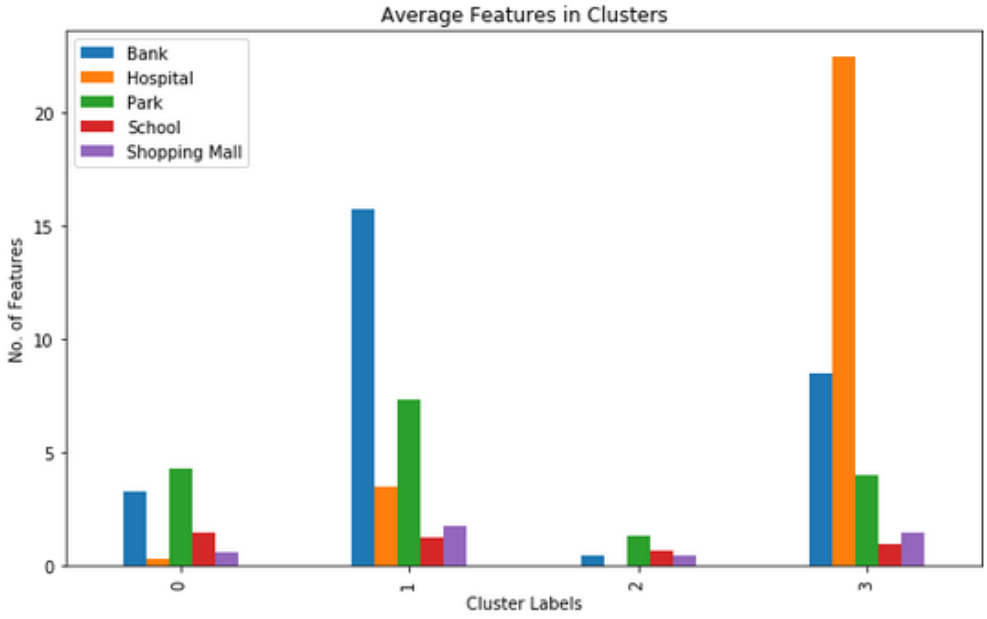
	PostalCode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Bank	Hospital	Park	School	Shopping Mall	Population 2016	Total Dwellings
0	M3A	North York	Parkwoods	43.753259	-79.329656	2	0.0	0.0	1.0	0.0	0.0	34615.0	13847.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	2	0.0	0.0	0.0	1.0	0.0	14443.0	6299.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	4.0	1.0	4.0	0.0	0.0	41078.0	24186.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	2	0.0	0.0	0.0	2.0	0.0	21048.0	8751.0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	3	5.0	22.0	5.0	1.0	1.0	10.0	6.0

Then, we grouped the main Dataframe by Clusters.

	Cluster Labels	Bank	Hospital	Park	School	Shopping Mall	Population 2016	Total Dwellings
0	0	3.300000	0.300000	4.300000	1.500000	0.60	31213.000000	15200.050000
1	1	15.750000	3.500000	7.375000	1.250000	1.75	6029.750000	4229.875000
2	2	0.483333	0.033333	1.366667	0.683333	0.45	27374.183333	11215.633333
3	3	8.500000	22.500000	4.000000	1.000000	1.50	4216.500000	2941.000000

We generated bar plots with the total and average **features** in the neighborhood.



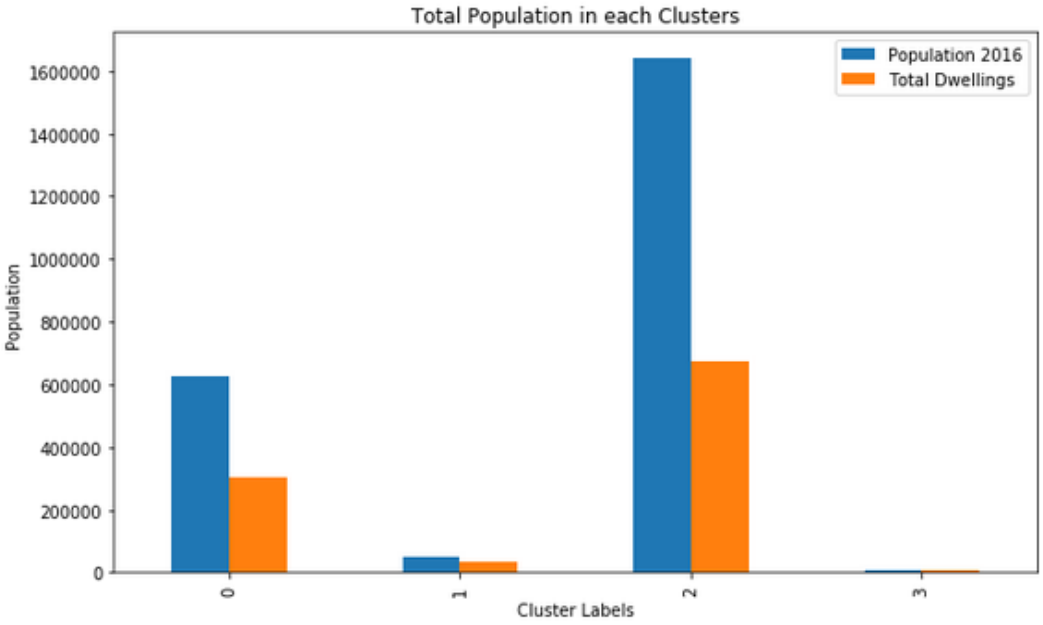


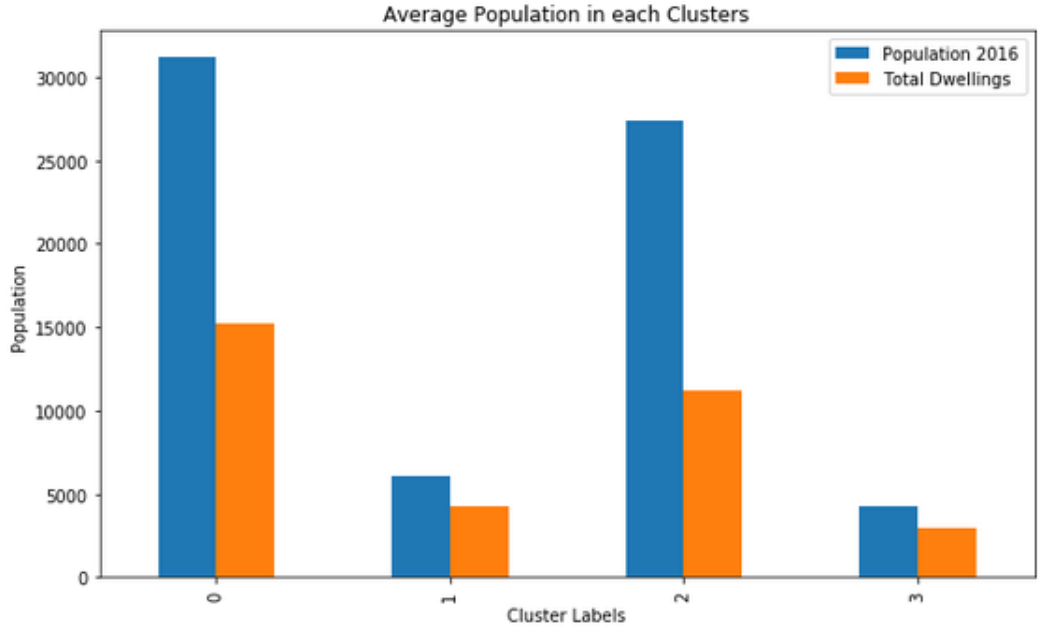
From the above bar charts we can infer that clusters have patterns.

1. Neighborhoods Cluster 1 and 3 have more number of average and total commercial establishments such as Banks and Hospitals.

2. In the rendered map we also saw that these clusters are closer to city center.

We generated bar plots with the total and average **population** in the neighborhoods.





From the above bar charts we can establish that clusters have patterns.

1. Neighborhoods Cluster 0 and 2 have more average and total population, also lesser commercial establishments thus are residential neighborhoods.

2. In the rendered map we saw that these clusters are away from city center.

5. Conclusion

Based on the observations we can conclude the following points:

1. Clusters created by the model have a clear relationship with population in those neighborhoods.
2. Maximum population stays away from city center despite of higher number of facilities available.
3. We can clearly segregate Residential and Commercial neighborhoods based on available data.