Capstone Project : The Battle of Neighborhood

TORONTO NEIGHBORHOOD COMPARISION

Applied Data Science Capstone | Jan-07-2020

Introduction

This report has been prepared as part of Applied Data Science Capstone Project on Coursera. This document will cover the following key topics

1. Business problem attempted to be resolved using data science
2. Data description and source
3. Methodologies used for data cleansing and machine learning
4. Inferences and conclusion
5. Future direction

Business Problem Statement

One of the key issues anybody moving to Toronto faces is "what is the best place to stay?". Now this question can be answered in many ways like interviewing friends and colleagues, advise from the realtors, job location etc. However, any such process mentioned above with be qualitative and does not take into account all the signification information available about the city of Toronto neighborhood based on which a quantitative analysis and decision can be formed. In addition, everybody's tastes are different i.e. somebody want to leave in a quiet neighborhood while another person might want to leave with more nightlife spots. So whatever qualitative analysis is done, it should also take into account the individuals taste and should allow the individual to make adjustments.

This project will take into account various data available about Toronto neighborhoods including all the different kind of categories of venues like schools, restaurants, entertainment etc. to build a cluster of neighborhoods based on the weightage on the various categories and allow user to modify the weightage and find what neighborhood to best suitable for their needs.

Data Sources

The project is leveraging the data already made available through the project guidelines. Ideally there are a number of other dimensions can be leveraged as part of the project like commute options, safety etc. but to keep the scope of the project limited only the following sources of data has been used :

1. Toronto Postal codes, Neighborhoods and Borough names from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. Toronto Postal code based coordinates i.e. latitudes and longitudes from http://cocl.us/Geospatial_data/Geospatial_Coordinates.csv
3. Category based venue data from Four Square. The following categories have been used :

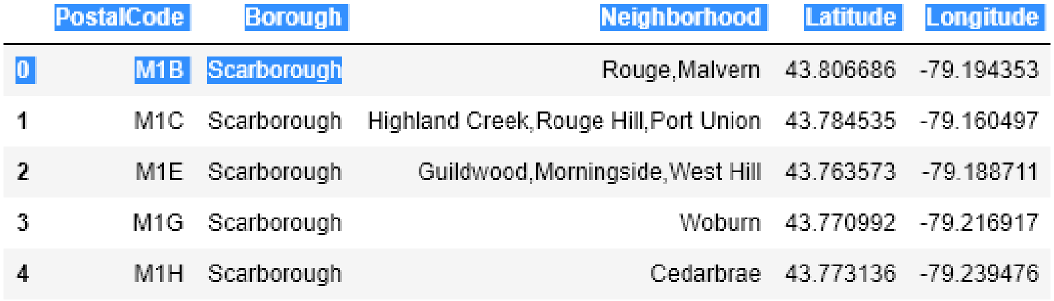
    a. Arts&Entertainment
    b. School
    c. Restaurant
    d. Nightlife Spot
    e. Parks&Recreation
    f. ProfessionalServices
    g. Office
    h. Place of Worship
    i. Shopping
    j. Grocery

4. Map data using folium

Methodology

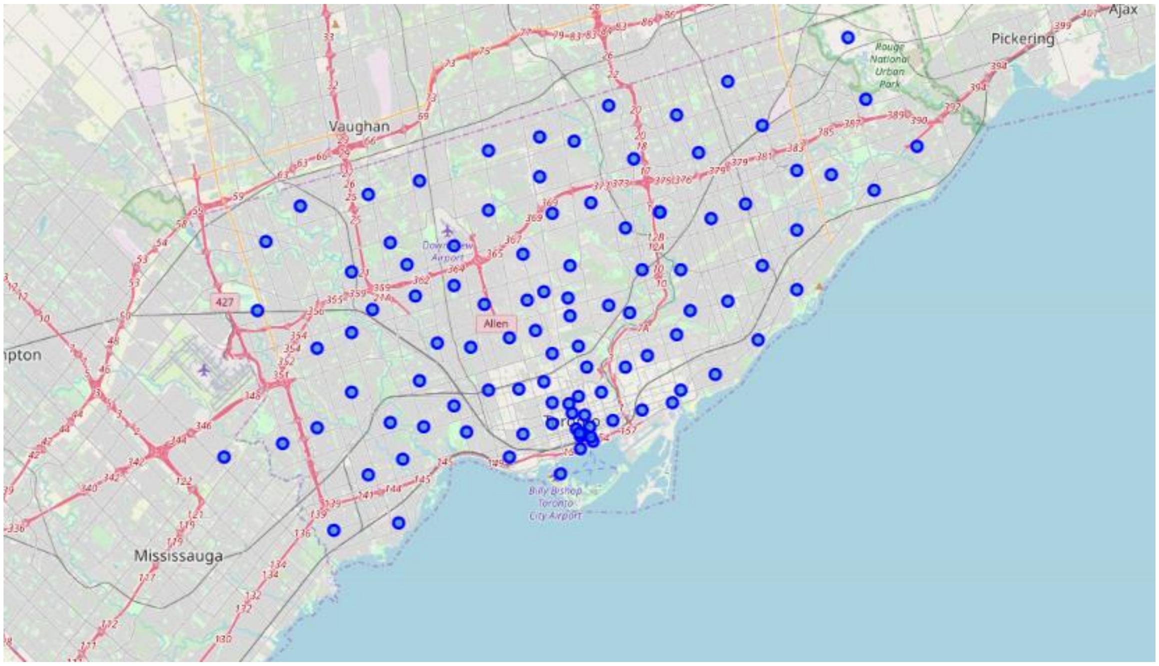
The following steps have been used to combine the data sources and build the neighborhood cluster

1. Read the Wiki Page for Toronto based postal code data and parse it using BeautifulSoup to build a dataframe that contains Toronto Postal Code, Borough and Neighborhood. (Any missing data for Borough i.e. Not assigned is ignored)
2. Combine the data from the wiki with the project provided data of coordinates. Ideally this could be built using geocoder but due to unstable code of geocoder previously provided file is used. The data looks like as follows :

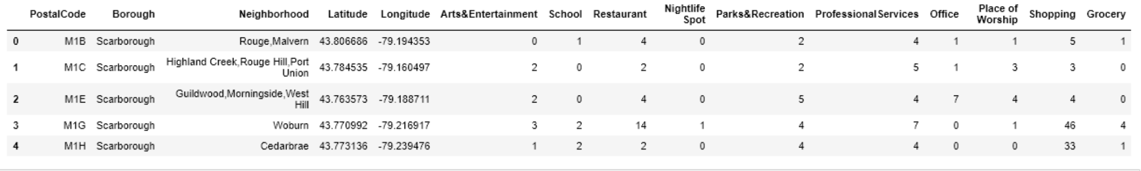


	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

3. The data above is now represented on a folium map of Toronto to understand how the neighborhoods look like.



4. The foursquare API is used recursively to pull count of venues of different categories and added to the dataframe.



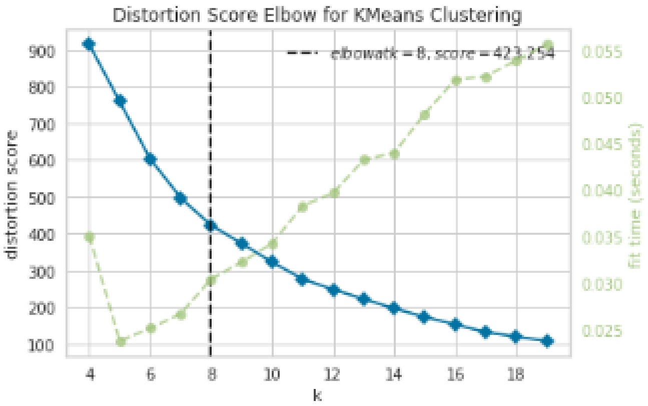
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Arts&Entertainment	School	Restaurant	Nightlife Spot	Parks&Recreation	ProfessionalServices	Office	Place of Worship	Shopping	Grocery
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	0	1	4	0	2	4	1	1	5	1
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	2	0	2	0	2	5	1	3	3	0
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	2	0	4	0	5	4	7	4	4	0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	3	2	14	1	4	7	0	1	46	4
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	1	2	2	0	4	4	0	0	33	1

5. Now we use a min max scaler to normalize the data of each of the categories and the weightage of each category is multiplied to the normalized data.

6. The weightage used for the different categories was as follows :

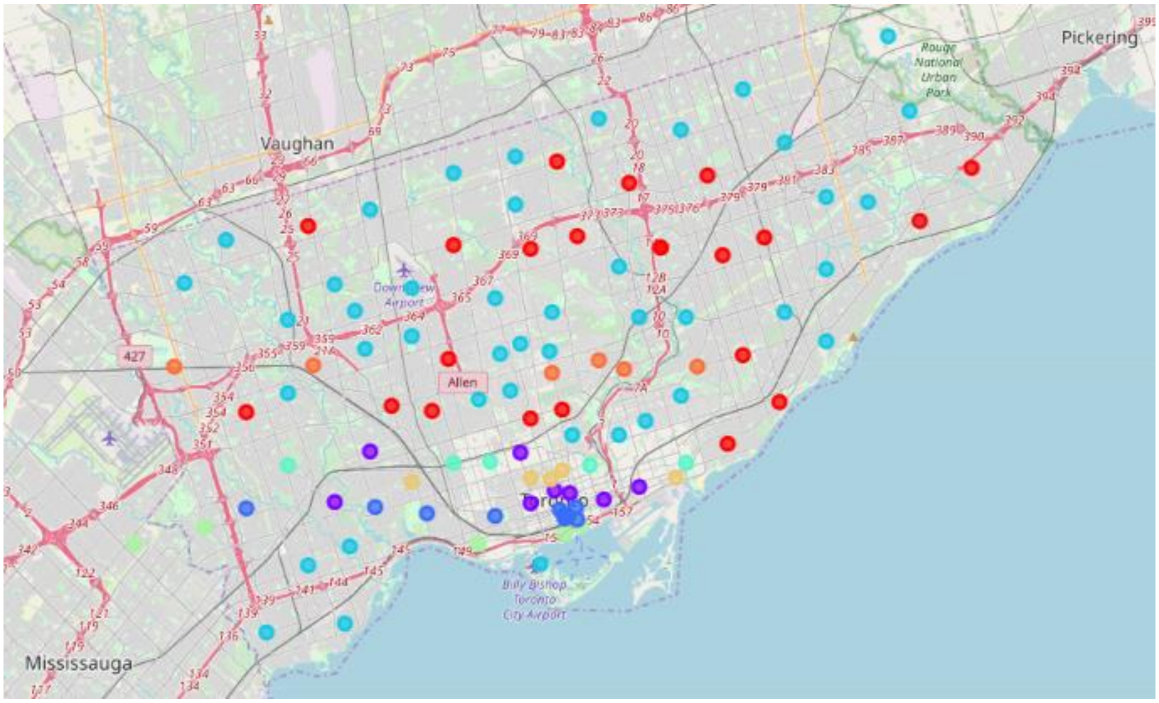
   a. Arts & Entertainment : 10
   b. School : 10
   c. Restaurant : 10
   d. Nightlife Spot : 5
   e. Parks & Recreation : 10
   f. Professional Services : 5
   g. Office : 5
   h. Place of Worship : 10
   i. Shopping : 5
   j. Grocery : 10

7. With the normalized and weighed data, K elbow visualizer is implemented to identify the best K Value for the K Means clustering.



8. The elbow value determined is 8 so the KMeans clustering is now implemented for 8 clusters thus dividing the neighborhoods in Toronto to 8 different clusters.

9. Now the folium map is displayed again with different colors for each clusters to show how different neighborhoods in Toronto share the same characteristics

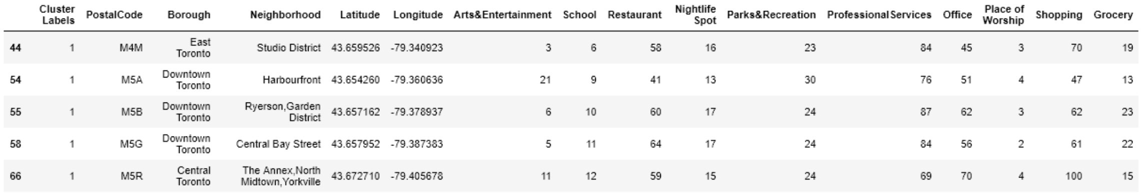


10. Now the cluster data is displayed for sample clusters to see what kind of neighborhoods have same venues

Conclusion

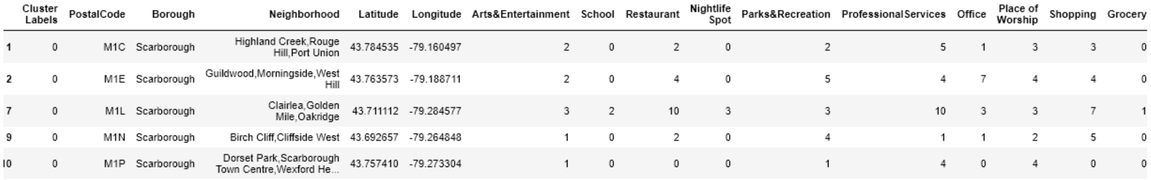
Based on the 10 criteria defined above, we were able to cluster all the 103 neighborhoods in Toronto into 8 different clusters.

Looking at the data it is identified that the borough of Central Toronto and neighborhood of The Annex, North Midtown and Yorkville has the most number of schools and it falls under Cluster 1 so if we want to get similar neighborhoods to that then filtering on Cluster 1 gives us the neighborhoods of East Toronto Studio District , Downtown Toronto Harbourfront etc. These neighborhoods will be good choice for people who share similar weightage and are primarily looking for more schools in the neighborhood.



	Cluster Labels	PostalCode	Borough	Neighborhood	Latitude	Longitude	Arts&Entertainment	School	Restaurant	Nightlife Spot	Parks&Recreation	ProfessionalServices	Office	Place of Worship	Shopping	Grocery
44	1	M4M	East Toronto	Studio District	43.659526	-79.340923	3	6	58	16	23	84	45	3	70	19
54	1	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	21	9	41	13	30	76	51	4	47	13
55	1	M5B	Downtown Toronto	Ryerson,Garden District	43.657162	-79.378937	6	10	60	17	24	87	62	3	62	23
58	1	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	5	11	64	17	24	84	56	2	61	22
66	1	M5R	Central Toronto	The Annex,North Midtown,Yorkville	43.672710	-79.405678	11	12	59	15	24	69	70	4	100	15

Similarly the neighborhood of Silver Hills and York Mills in North York have 0 schools in the neighborhood so if want to identify the neighborhoods to absolutely avoid based on above criteria then we have to look for neighborhoods in the same cluster as Silver Hills and York Mills which is cluster 0. This gives us neighborhoods like Highland Creek, Rouge Hill, Port Union, Guildwood, Morningside etc. in Scarborough.



	Cluster Labels	PostalCode	Borough	Neighborhood	Latitude	Longitude	Arts&Entertainment	School	Restaurant	Nightlife Spot	Parks&Recreation	ProfessionalServices	Office	Place of Worship	Shopping	Grocery
1	0	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	2	0	2	0	2	5	1	3	3	0
2	0	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	2	0	4	0	5	4	7	4	4	0
7	0	M1L	Scarborough	Clairlea,Golden Mile,Oakridge	43.711112	-79.284577	3	2	10	3	3	10	3	3	7	1
9	0	M1N	Scarborough	Birch Cliff,Cliffside West	43.692657	-79.264848	1	0	2	0	4	1	1	2	5	0
10	0	M1P	Scarborough	Dorset Park,Scarborough Town Centre,Wexford He...	43.757410	-79.273304	1	0	0	0	1	4	0	4	0	0

Future direction

Currently this project has only used venue data from Four Square but there are many other datasets available to add more dimensions for the comparison. One example is the crime data for Toronto that is available from Toronto police that can be leveraged to determine the safety index of the neighborhood. The data was downloaded but it had different neighborhood names and although it had latitude and longitude we needed a mechanism to get the postal code or Borough names hence it has been left for future enhancement.