Data Science Project

REPORT

Segmenting Neighbourhoods of Toronto on the basis of Livability

ANANTH S G

29/Feb/2020

1. INTRODUCTION

- ## Background

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of about 3 Million as of July 2018. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto is the New York of Canada.

Toronto ranks in the top 5 of the most livable cities in the world according to the Economist Intelligence Unit. Thanks to the booming economy, the inflow of immigrants the real estate market of Toronto is heating up.

For prospective home buyers and businessmen, it will be useful to know which neighborhoods of Toronto are more attractive for investment.

- ## Problem

We know that Toronto is among the top livable cities of the world. But no comprehensive study done to rank the livability of the neighborhoods of Toronto. Livability Score of a neighborhood depends on proximity to Restaurants, Nightlife options, Shops, Schools, Hospitals, Public Transport and Offices. For simplicity we only take into consideration the Restaurants, Bars, Schools and Shops to access the Livability Score.

- ## Interest

Grouping neighborhoods by Livability Scores will be extremely useful for the following:

1. Real Estate agents would want to know how the commercial & housing prices vary with Livability Score
2. Online listing companies such as Realtor.com, Zillow can provide customers the options to browse through livability scores of neighbourhoods
3. Most importantly the customers would like compare the livability score of neighbourhoods with the Housing Prices before making the purchase

2. Data

- ## Data Sources

The list of Toronto Neighborhoods and the postal codes can be found in

wikipedia(web-scraping). The location data of each neighbourhood can be obtained from geocoder api. Foursquare API provides the list of Restaurants, Bars, Schools and Shops in a neighborhood.

- ## Data Cleaning

The neighborhood data was scraped from the website and merged with the location data.

The neighbourhood data had some unassigned boroughs which were removed. If the neighborhood name was unassigned but it had a corresponding borough value, the borough name was used the neighborhood name.

The location data that we got was based on the postal codes. That is why the neighborhood data had to be grouped. The grouping was done on the basis of postal code. If multiple neighbourhood names shared a single postal code, then the names were concatenated with 'comma' separator.

The final merged data set containing the neighborhood & location data had data points for Toronto & its suburbs. For this analysis we only consider the data points for Toronto city.

Since neighbourhoods are grouped w.r.t the postal codes. Whenever neighborhoods are mentioned, it actually means a set of neighborhoods having a common postal code.

This is what the merged data set looks like.

	Postal-Code	Borough	Neighbourhood	Latitude	Longitude
0	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568
1	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
2	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191
3	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558
4	M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420

- ## Feature Selection

The livability score of a Toronto neighbourhood depends on how many venues are present **within 500m (walking distance)** of its geographical location.

For simplicity, we assume that the livability score depends on criteria such as lifestyle and the convenience. The criteria Lifestyle is the number of Bars & Restaurants nearby. The criteria

convenience is the number of schools and shops nearby. High lifestyle score is preferable for young people and a high convenience score is preferred by everybody.

So, **Livability Score = Number of Restaurants,Bars,Schools & Shops**

We can extract the count of all these venue categories using the Foursquare API.

Once we fetch the number of Restaurants,Bars,Schools & Shops for every neighborhood, we merge this data with the original data set.

This is a snapshot of Toronto neighbourhood data along with the feature set needed to group **neighbourhoods by Livability Score**.

	Postal-Code	Borough	Neighbourhood	Latitude	Longitude	Restaurants	Bars	Shops	Schools
0	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568	42.0	50.0	50.0	12.0
1	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	13.0	45.0	24.0	7.0
2	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	4.0	9.0	5.0	3.0
3	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	0.0	3.0	2.0	0.0
4	M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420	0.0	0.0	0.0	1.0

3. Methodology

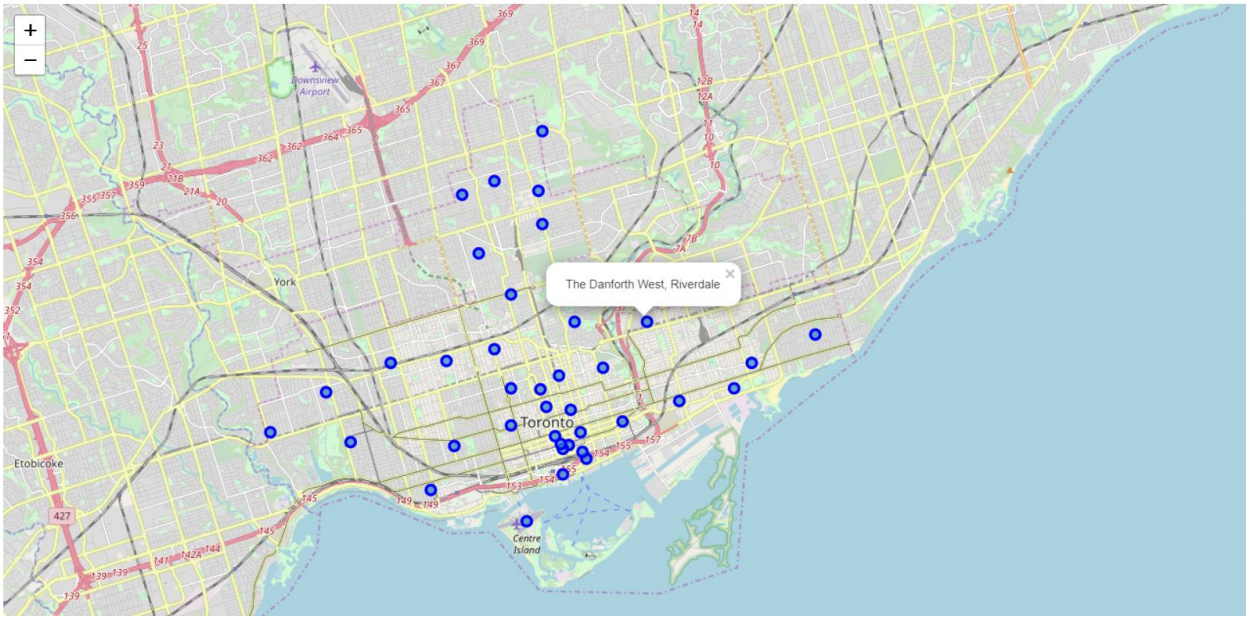
- ### Exploratory Data Analysis

The initial data set, with the data of neighborhoods of Toronto and its suburbs, had 103 unique postal codes and 10 boroughs. For the analysis we only consider the Toronto City data which contains 39 unique postal codes and 4 boroughs.

The details of how many unique postal codes and unique neighborhoods in Toronto city is shown below:

	Postcode	Neighbourhood
Borough		
Central Toronto	9	17
Downtown Toronto	19	36
East Toronto	5	7
West Toronto	6	13

Using the above details let us map the neighbourhoods.

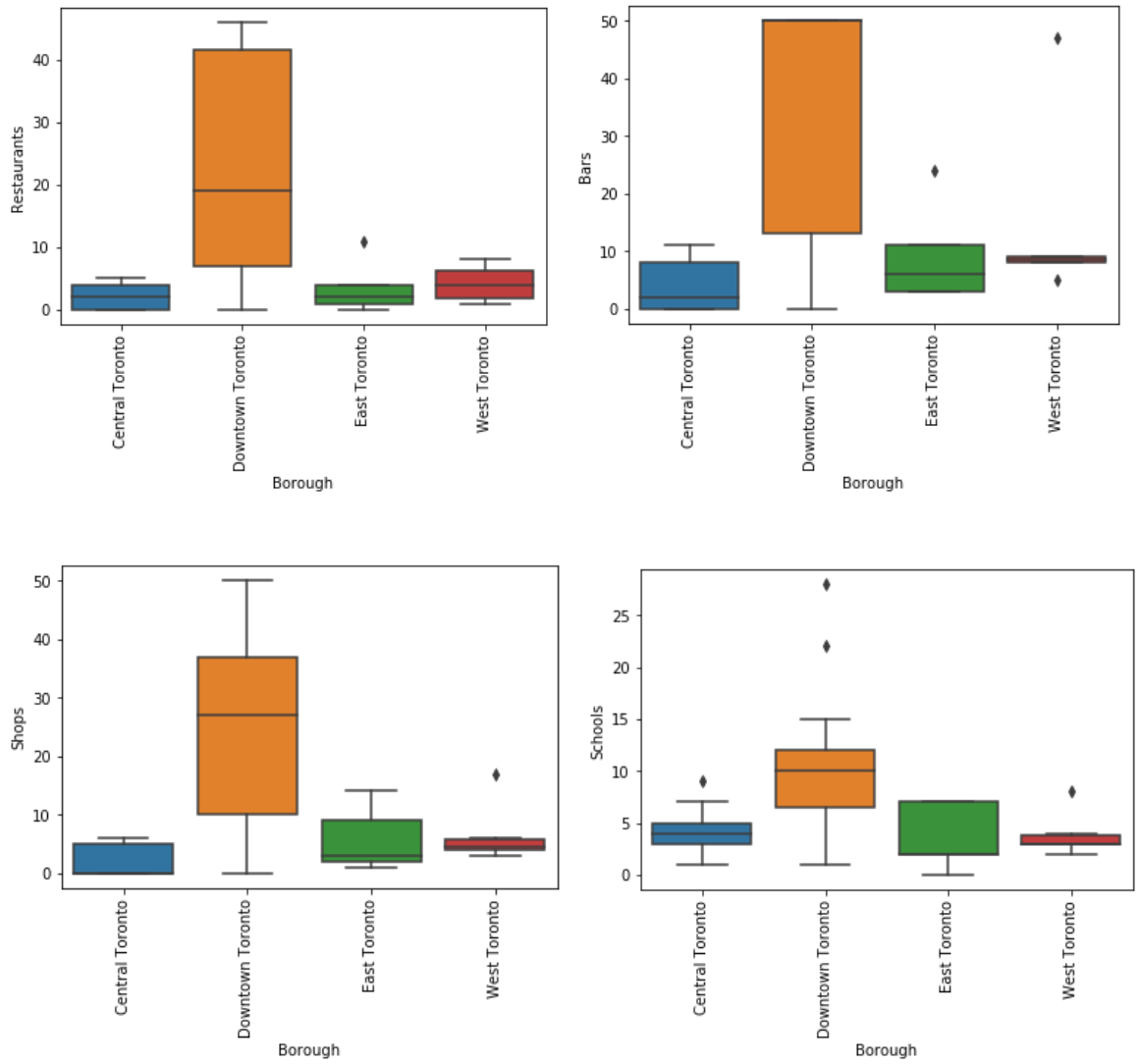


After fetching the number of Restaurants,Bars,Schools & Shops for every neighborhood, we end up with the below data set.

	Postal-Code	Borough	Neighbourhood	Latitude	Longitude	Restaurants	Bars	Shops	Schools
0	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568	42.0	50.0	50.0	12.0
1	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	13.0	45.0	24.0	7.0
2	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	4.0	9.0	5.0	3.0
3	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	0.0	3.0	2.0	0.0
4	M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420	0.0	0.0	0.0	1.0

Let us explore **how each boroughs fare when compared to our feature set data**.

These are the **box plots of boroughs v/s Features.**



Out of all the boroughs Downtown Toronto is the clear winner and has the highest number of Restaurants, Bars, Schools and Shops within a distance of 500m.

Central Toronto & East Toronto have similar numbers. West Toronto is the clear loser with the least number of Restaurants, Bars, Schools and Shops within a distance of 500m.

Let us now see how things look at neighbourhood level using clustering algorithms.

- ## Machine Learning- Clustering Algorithms:

The main problem statement is to segment and cluster the neighborhoods of Toronto on the basis of Livability Score. The final data along with features has been compiled for the 39 postal codes and 4 boroughs of Toronto and it looks like this:

	Postal-Code	Borough	Neighbourhood	Latitude	Longitude	Restaurants	Bars	Shops	Schools
0	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568	42.0	50.0	50.0	12.0
1	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	13.0	45.0	24.0	7.0
2	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	4.0	9.0	5.0	3.0
3	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	0.0	3.0	2.0	0.0
4	M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420	0.0	0.0	0.0	1.0

Livability Score = Number of Restaurants,Bars,Schools & Shops

For the clustering, let us use the K-Means algorithm with 4 Clusters. The results can be interpreted better when we use number of clusters = 4.

The features are the columns Restaurants, Bars, Schools and Shops. The features are normalized and transformed before being fed to the algorithm.

The K-Means algorithm groups every neighborhood into one of the 4 clusters and gives a cluster label.

The output looks like this:

	Postal-Code	Borough	Neighbourhood	Latitude	Longitude	Restaurants	Bars	Shops	Schools	Labels
0	M4R	Central Toronto	North Toronto West	43.715383	-79.405678	2.0	8.0	3.0	3.0	1
1	M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	4.0	11.0	6.0	4.0	1
2	M4P	Central Toronto	Davisville North	43.712751	-79.390197	4.0	2.0	0.0	5.0	1
3	M4S	Central Toronto	Davisville	43.704324	-79.388790	5.0	11.0	5.0	9.0	1
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	0.0	0.0	0.0	4.0	1

When we look at the average number of restaurants, bars, schools and shops for each label, we get more clarity to interpret the clustering.

	Restaurants	Bars	Shops	Schools
Labels				
0	40.571429	47.714286	41.285714	10.857143
1	2.434783	5.130435	3.043478	3.956522
2	12.000000	36.428571	18.714286	8.142857
3	43.500000	50.000000	34.000000	25.000000

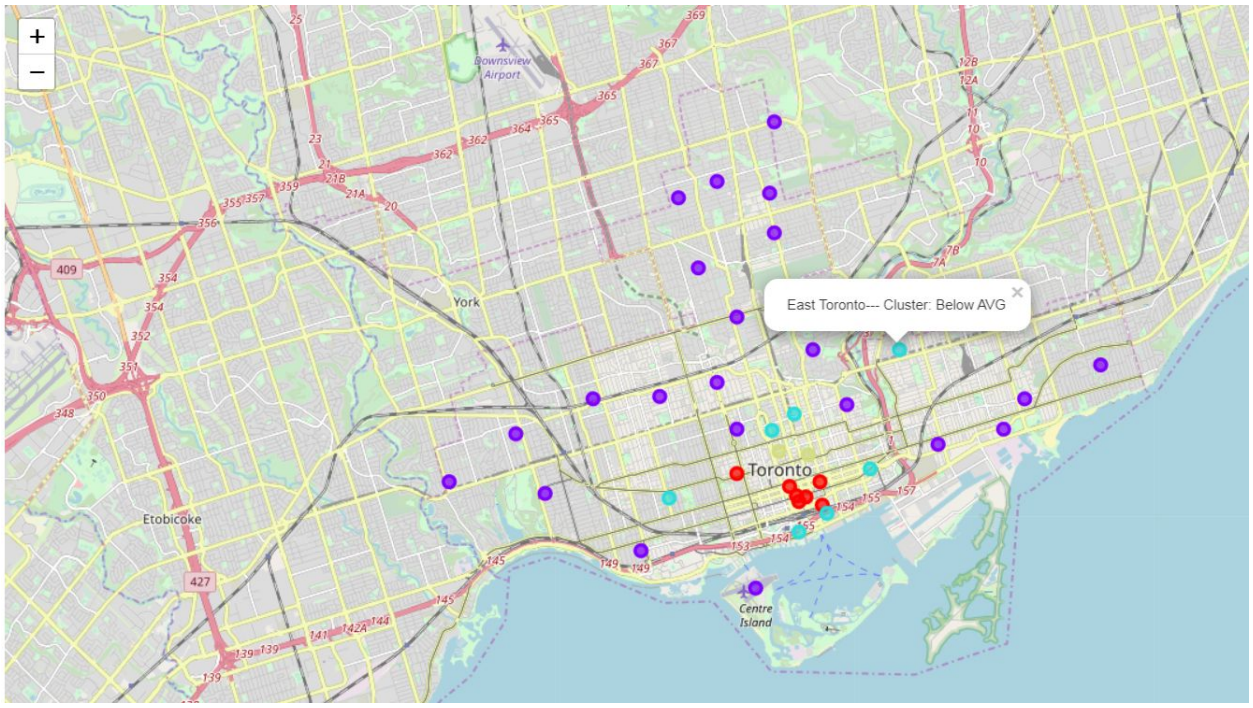
From the above number we **assign meaningful labels** to the above numeric labels. There are 4 clusters giving us the livability score grouping. Since label 3 has the highest number of venues for each category, let us name it 'HIGH'. Then comes label 0 with 'Above AVG' livability score. After that we have label 2 with a 'Below AVG' and finally the label 1 with a 'LOW' livability score.

Once we give meaningful cluster labels, let us append these to the old table with column name "Livability".

	Postal-Code	Borough	Neighbourhood	Latitude	Longitude	Restaurants	Bars	Shops	Schools	Labels	Livability
0	M4R	Central Toronto	North Toronto West	43.715383	-79.405678	2.0	8.0	3.0	3.0	1	LOW
1	M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	4.0	11.0	6.0	4.0	1	LOW
2	M4P	Central Toronto	Davisville North	43.712751	-79.390197	4.0	2.0	0.0	5.0	1	LOW
3	M4S	Central Toronto	Davisville	43.704324	-79.388790	5.0	11.0	5.0	9.0	1	LOW
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	0.0	0.0	0.0	4.0	1	LOW

Using the above details let us map the neighbourhoods in a map.

The below map shows all the 4 clusters on the map of Toronto City and if you click on the colored dots, we get the information of the Cluster Label and the Borough.



Cluster Color	Livability Label
Yellow	HIGH
Red	Above Avg
Light Blue	Below AVG
Purple	LOW

4. Results & Conclusion

Based on the features (Number of restaurants, bars, schools & shops) the neighborhoods of Toronto have been segmented or **clustered** into 4 similar groups. The clustered neighborhoods have similar Livability Scores.

From the above map, the Yellow and Red dots offer a great living experience (High Livability Score) for the customers. These correspond to the neighbourhoods in Downtown Toronto & Central Toronto. The neighbourhoods close to the purple dots don't offer such a great living

experience. These are the neighborhoods in West & East Toronto. This doesn't mean that the people living in these neighborhoods are miserable. It means that compared to Downtown Toronto, they don't offer much facilities.

The **implication** for **real-estate** deals is that the housing prices in neighborhoods with above average Livability Score must be more expensive. We need to include housing prices data in our analysis. If a traveller is looking to book a hotel in Toronto and he/she wants to experience the nightlife of Toronto, then Downtown Toronto is the best choice.

5. Further Discussion & Recommendation

To calculate the livability score we are only considering the number of restaurants, bars, schools and shops in the vicinity. But in actuality we need to consider **more data points** to decide on the livability score. We need to include features such as median housing prices, offices, bus stops, shopping malls, population, resident complaints, parks etc.

Finally there is also a scope for developing a **recommendation system** for a customer to check which neighborhood would be better for him/her. This would be useful for online listing companies such as realtor.com and zillow.com. For example, we need to take the customers information about his income and place of work and input these parameters into our recommendation system. The system then uses clustering and decision trees to suggest the best neighborhoods and best houses for the customer.

6. References

- Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Location data from geocoder from the location http://cocl.us/Geospatial_data
- Features data from Foursquare API venues query