Buying a home in Toronto

1. Introduction

1.1 Background

Since I have gotten well acquainted with Toronto in our last couple of assignments, I would like to stay there for this one too. Toronto is a large city with over 5mil of residents. In such a large city like Toronto it might be difficult to figure out what neighborhoods to concentrate on in your search to buy a new home. When searching for a house, one might want to concentrate on neighborhoods with low crime rates, average density of population, access to parks and playgrounds, as well as restaurants and coffee shops. All those criteria are usually considered with a certain budget in mind. I have visited Toronto 2 times and I have stayed with my cousins in the Forest Hill South neighborhood, so I will take that as a point of reference for a comfortable, safe, enjoyable living, and try to find other neighborhoods in Toronto that are similar to Forest Hill South, but are more affordable (have an average home price of less than $800,000).

1.2 Problem

There are about 140 officially recognized neighborhoods in Toronto. To determine neighborhoods that are like Forest Hill, these neighborhoods need to be clustered based on their safety scores, population density, access to parks, restaurants, and coffee shops. Once I determine which neighborhoods properties of interest are like those of Forest Hill South, I will be able to find neighborhoods that satisfy the budget restriction imposed in the last paragraph.

1.3 Interest

This project would be interesting to anyone who is looking to buy a house anywhere, as these methods can be transferred to finding a new house in New York, London or Tokyo. It is always good to do your own research before meeting with a real estate agent, so you can specify what exactly you are looking for. Anyone would be interested to learn more about the neighborhoods and to quickly eliminate neighborhoods that he/she is not interested in.

2. Data

2.1 Sources, cleaning and feature selection

Toronto Police has an open data access to records about various crime ratings, such as assault, robbery, homicide, auto theft and break-and-enter from 2014 to 2018. I have located the geojson file from their website (here), and extracted the crime data for the year 2018 for each neighborhood in Toronto. The file also included the information about population and size of the neighborhood as well as the coordinates of the boundaries of each neighborhood.

Features used:

1. Neighborhood
2. Assault Rate 2018
3. Auto Theft Rate 2018
4. Break and Enter Rate 2018
5. Robbery Rate 2018
6. Homicide Rate 2018
7. Population
8. Size of Hood Area
9. Population density (population/size)
10. Latitude
11. Longitude

2.2 Venues: sources, cleaning, and feature selection

I queried Foursquare API in order find venues that are most popular in each neighborhood of Toronto. The query returned 1475 venues, in 255 unique venue categories. The venue categories were converted to binary variable, using one hot encoding, and grouped by neighborhood. The resulting data set had 134 rows and 255 columns (plus the neighborhood index).

2.3 Average home price: sources, cleaning, and feature selection

The home prices data were scraped from a blog post on Toronto home prices by neighborhood for 2017, which luckily has the same format for the neighborhood names and assignment (link). There were, however, 6 neighborhoods missing from the dataset, and the missing values in those cases were replaced with average home price for all neighborhoods. In future plots, those neighborhoods can be easily spotted since their home prices have decimal points.

The average home prices were listed as strings with commas and dollar signs. Dollar signs and commas were removed, and the prices were converted to floats.

3. Methodology

3.1 Tools

The *Folium* package for Python was used to plot Neighborhoods on the map. The *DivIcon* module for Folium was used to insert text boxes onto maps. The *json* package was used to open and read geojson file. The *.read_html* method from *pandas* was used to scrape table data from a website. The *requests* library for Python was use to make a query request to the Foursquare website. The *KMeans* module from the *sklearn* package was used to cluster the data. The *matplotlib.pylab* was used to make exploratory and summary plot of data. Toronto Police Open Data Portal was used to obtain the city safety and population density data, and Foursquare API was used to obtain venue data.

3.2 Exploratory data analysis

The size of each neighborhood varies, therefore it was important to covert Population variable (Figure 1) into a Population Density (Figure 2) to be used later for clustering of the data. The data on Population, however was still used to calculate per capita crime rates. As shown in Figure 2, Population Density of neighborhoods varies, therefore it was used as one of the features for the clustering analysis.

Plotting the crime rate as, for example “Robbery per capita” (Figure 3) or “Assault per capita (Figure 4), as the size of the corresponding Neighborhood circles shows that the crime rates vary dramatically. A similar distribution of sizes was also observed when other 3 measures of crime rate were plotted (*i.e.*, Homicide per capita, AutoTheft per capita and BreakAndEnter per capita, data not shown).

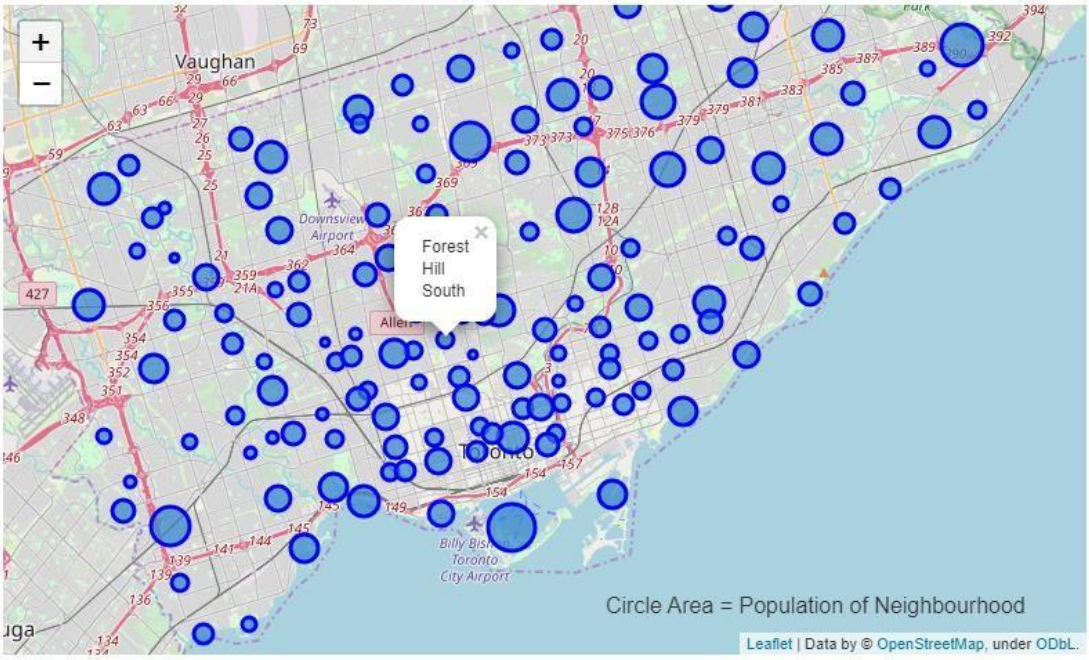

Figure 1. Toronto Neighborhoods represented as circles, where the area of the circle represents Population of the Neighborhood.

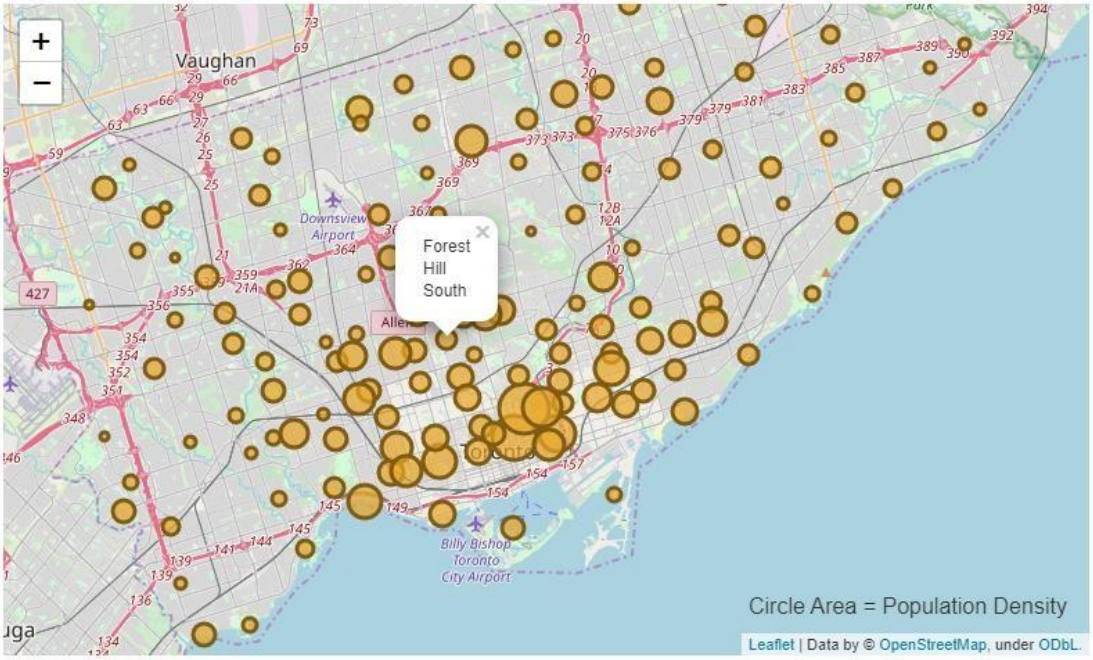

Figure 2. Toronto Neighborhoods represented as circles, where the area of the circle represents Population Density of the Neighborhood.

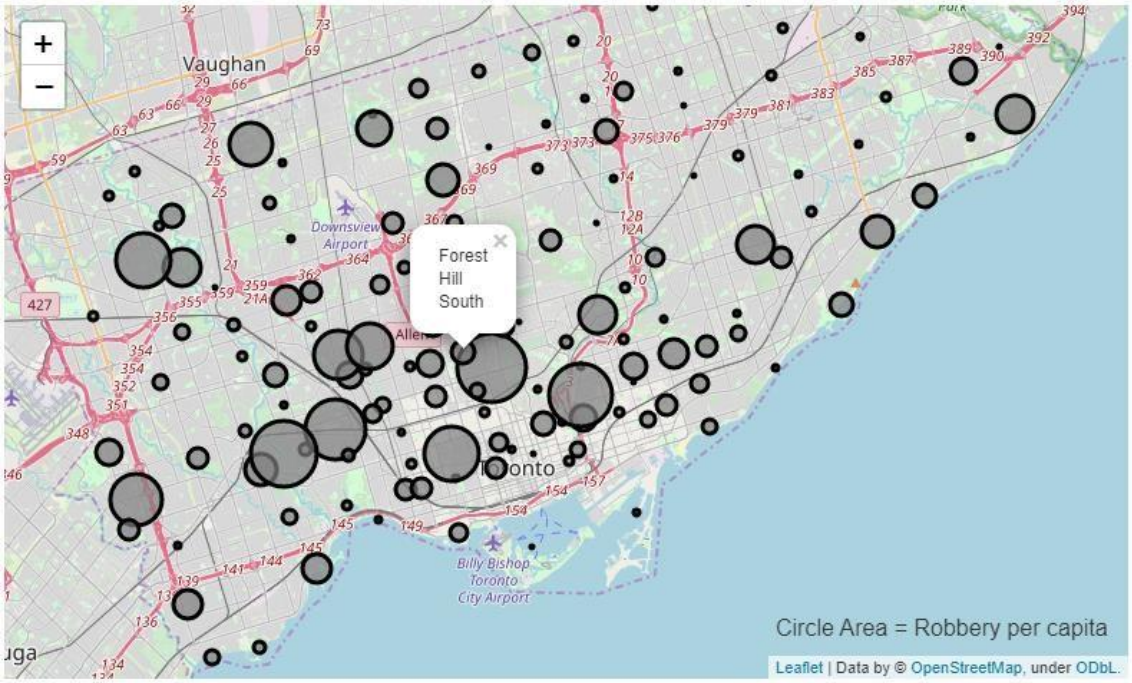

Figure 3. Toronto Neighborhoods represented as circles, where the area of the circle represents Robbery Per Capita (X100) in the Neighborhood.

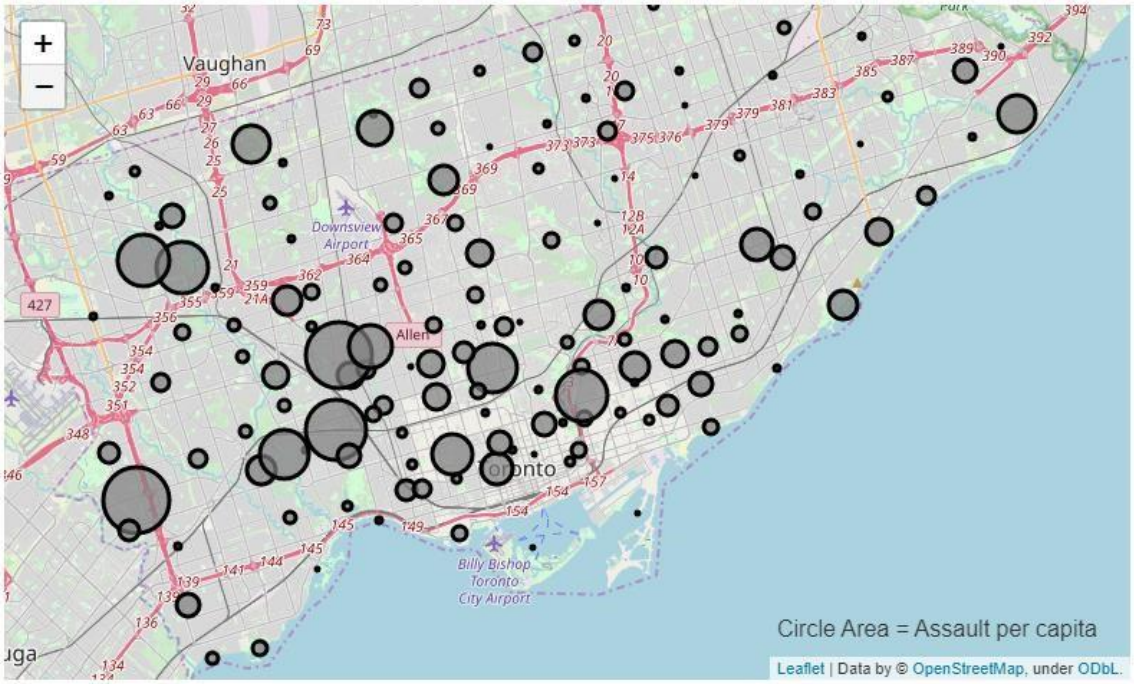

Figure 4. Toronto Neighborhoods represented as circles, where the area of the circle represents Assault Per Capita (X100) in the Neighborhood.

There is no correlation between population density and crime (correlation coefficient is −0.226 when “Assault_per_capita" was used as an example in Figure 5), and therefore both of those features will be useful features to use for clustering.

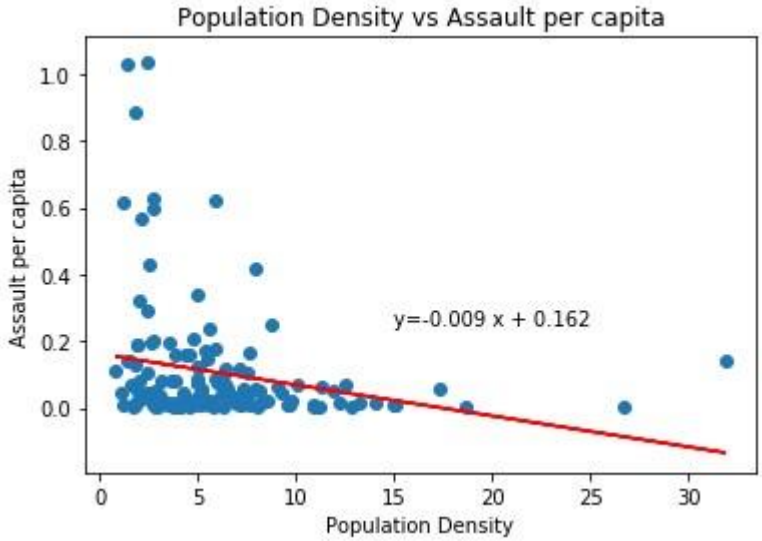

Figure 5. A scatter plot of Population Density and Assault Per Capita rate for Toronto Neighborhoods.

3.3 Selection, normalization and weighting of features.

There are three main factors that I wanted to consider when selecting the right neighborhood to live in: safety (measured as crime), population density, and access to venues. I have collected 5 measures of crime (assault, robbery, homicide, auto theft, and break-and-enter), 1 measure

of population density (population density), and 347 venue category features. I would want the three categories (safety, population density, and venues) to have equal weights in the clustering analysis, and therefore, after normalization of the data to a mean of 0 and variance of 1, features were adjusted by their corresponding weight. Specifically, *StandardScaler* module from *sklearn.preprocessing* package with its default settings was used to normalize all 353 features to a mean of 0 and variance of 1, and then the 5 crime features (assault, robbery, homicide, auto theft, and break-and-enter) were divided by 5, while the 347 venue category features were divided by 347. The population density feature remained as was after normalization.

Following k-means clustering, a new variable called Normalized Crime Rate was calculated by taking the average of the normalized crime features (i.e., assault, robbery, homicide, auto theft, and break-and-enter per capita), to be used for data presentation.

3.4 Clustering

K-Means clustering was used to segment the neighborhood data. First, the elbow method was used to determine the best number of clusters to use. To that end, the number of clusters from 1 to 9 were tested by plotting the distortion function (i.e., the sum of distances for each point to its cluster center), obtained using the *inertia_* data from the kmeans fit, was plotted against the number of clusters. The rate of distortion reduction reduced at 5 clusters (Figure 6), and therefore that's the number of clusters I used in my k-means analysis.

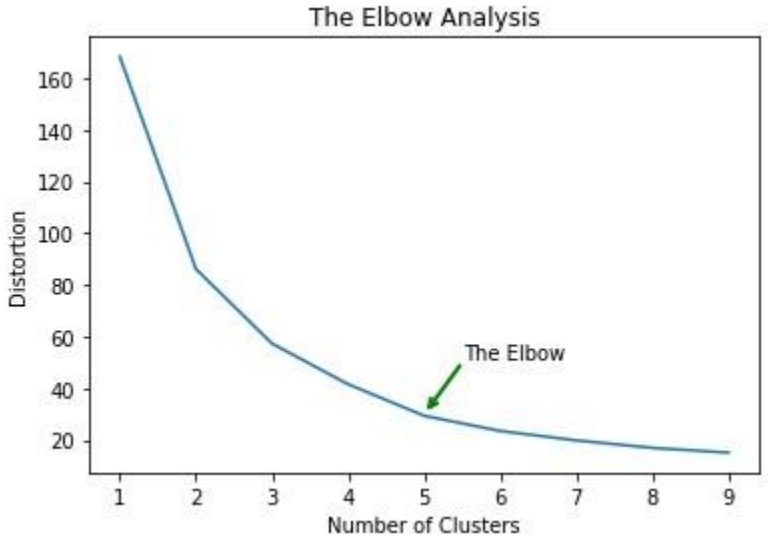

Figure 6. The sum of distance between each point and its cluster (i.e., distortion) as a function of the number of clusters.

4. Results

4.1 Clusters

K-Means clustering analysis segmented the Toronto neighborhoods into five clusters represented by different clusters in Figure 7. Our neighborhood of interest, "Forest Hill South", is part of Cluster 0. In the figure, the size of the circles represents the normalized crime rate for

each neighborhood (calculated as described in *Methodology*). Cluster 0 is the largest cluster, with 72 members. The number of neighborhoods in the other clusters is listed in Table 1.

Even though Cluster 0 has about 50% of the neighborhoods, it does appear that the clustering is reflective of the data. Figures 7 shows that the normalized crime rate is in fact associated with the cluster assignment: clusters of the same color have similar sizes. Figure 8 shows that average normalized crime rates do in fact vary from cluster to cluster. Similarly, Figure 9 shows the association of population density with cluster assignment, and Figure 10 shows that clusters do vary in their average population density.

Cluster Labels	Number of Neighbourhoods
0	72
1	14
2	2
3	8
4	44

Table 1. Number of neighborhoods in each Cluster

Since we have over 300 features representing venue categories, and we adjusted their weights accordingly, it would be difficult to obtain similar representative map plots and bar graphs like those shown in Figures 7-10 showing the dependence of cluster assignment on any particular venue. Figure 11 shoes the average count of three venue categories in each cluster: “Park”, “Coffee Shop”, and “Yoga Studio”. Only the “Yoga Studio” count appear to very significantly between clusters. Such lack of evident dependence of cluster assignment on venue category is not surprising, firstly, considering I set up small weights to each of these features individually, since I cared more about safety and population density, and secondly, considering we observed little heterogeneity in venue categories in Toronto Neighborhoods in Module 3 of this course.

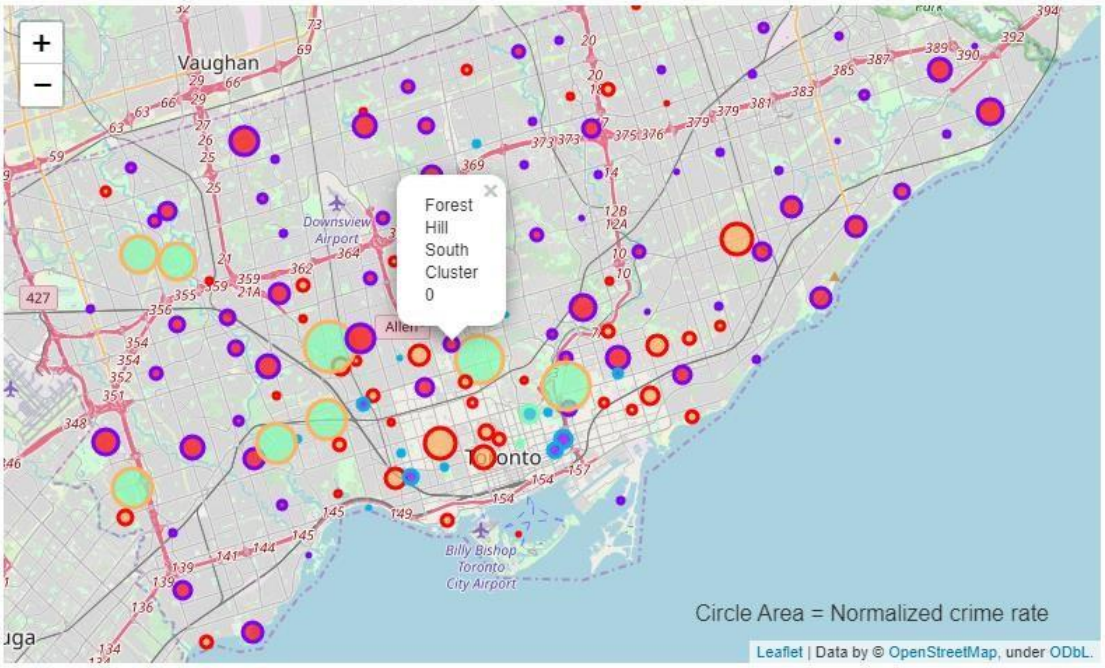

Figure 7. The five clusters of Toronto neighborhoods shown in different colours, where the size of the circle represents the Normalized Crime Rate.

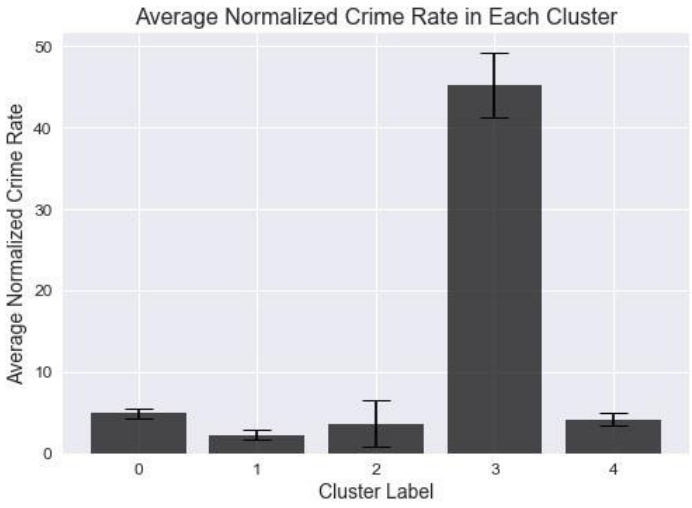

Figure 8. Average normalized crime rate (X 100) by cluster.

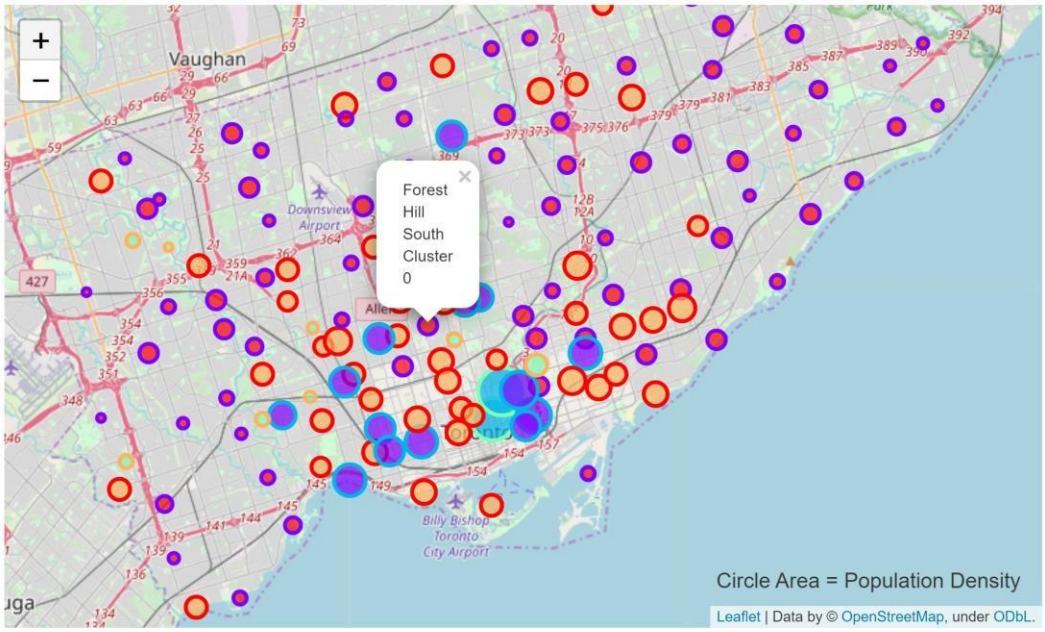

Figure 9. The five clusters of Toronto neighborhoods shown in different colours, where the size of the circle represents the Population Density.

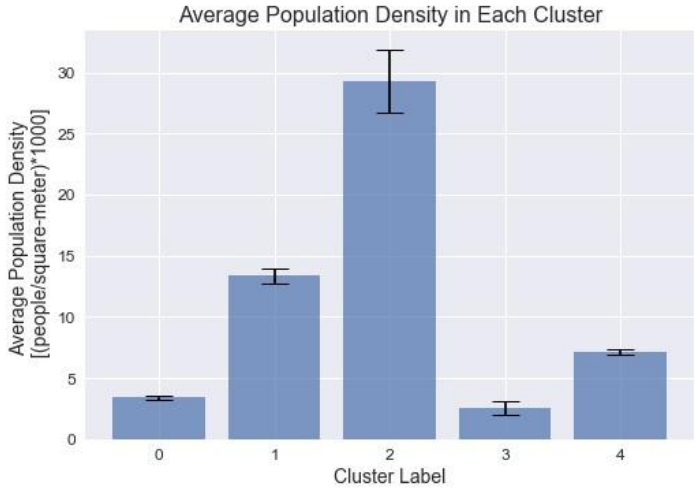

Figure 10. Average Population Density in each Cluster

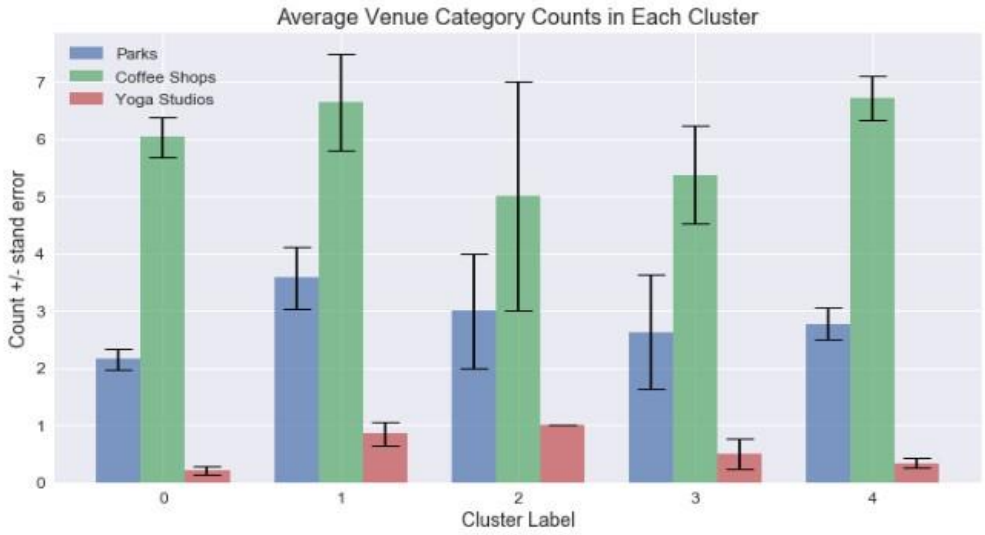

Figure 11. Average counts of parks, coffee shops and yoga studios in each cluster.

4.2 Extracting a short list of Neighborhoods that are affordable and similar to Forest Hill South using clustering results.

When the average home prices for each Neighborhood were used as the circle area representing each Neighborhood, no evident pattern in circle sizes and cluster assignment was observed (Figure 12). Similarly, when plotted as a bar graph, the mean prices for each cluster were very similar (Figure 13). This is not surprising, and it is in fact desirable, since we didn't want home prices to affect clustering. Now we can have a closer look at our Cluster of interest, Cluster 0, which contains the Neighborhood with all my desirable properties, except for home price.

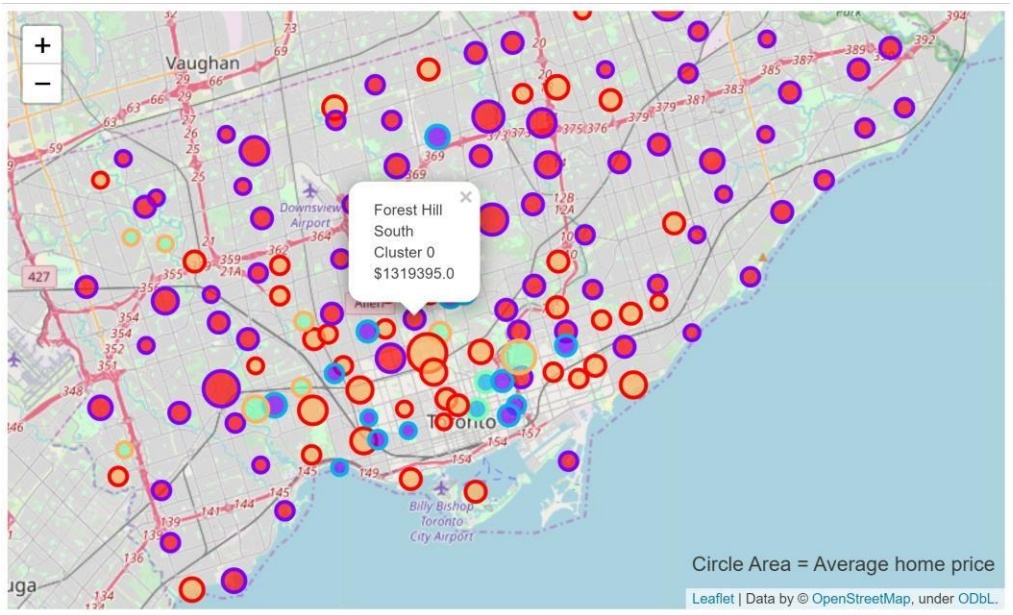

Figure 12. Five clusters of Toronto neighborhoods shown in different colours, where the size of each circle represents the average home price for 2017 in each Neighborhood.

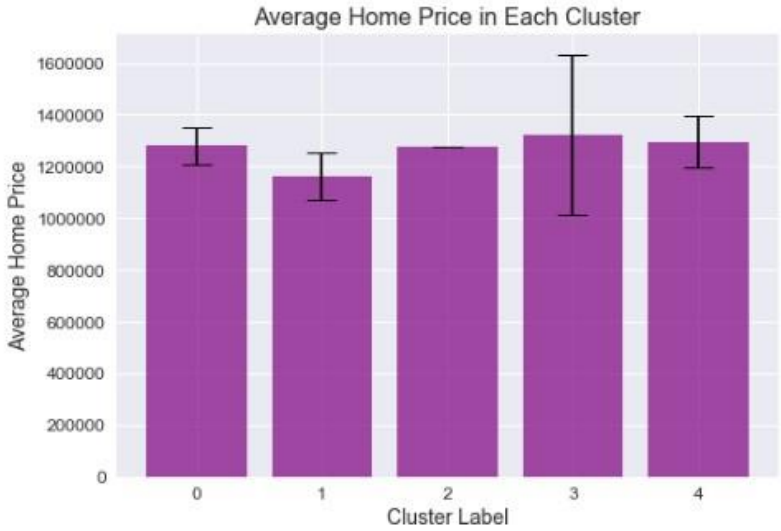

Figure 13. Average home prices in 2017 in each of five clusters of Toronto Neighborhoods.

Figure 14 shows the normalized average crime rate in Neighborhoods in Cluster zero that have an average home price less than $800,000. There are 11 neighborhoods that satisfy that requirement. Out of those 11 Neighborhoods, I chose to filter out neighborhoods with crime rate higher than 4, since that's a criterion very important to me.

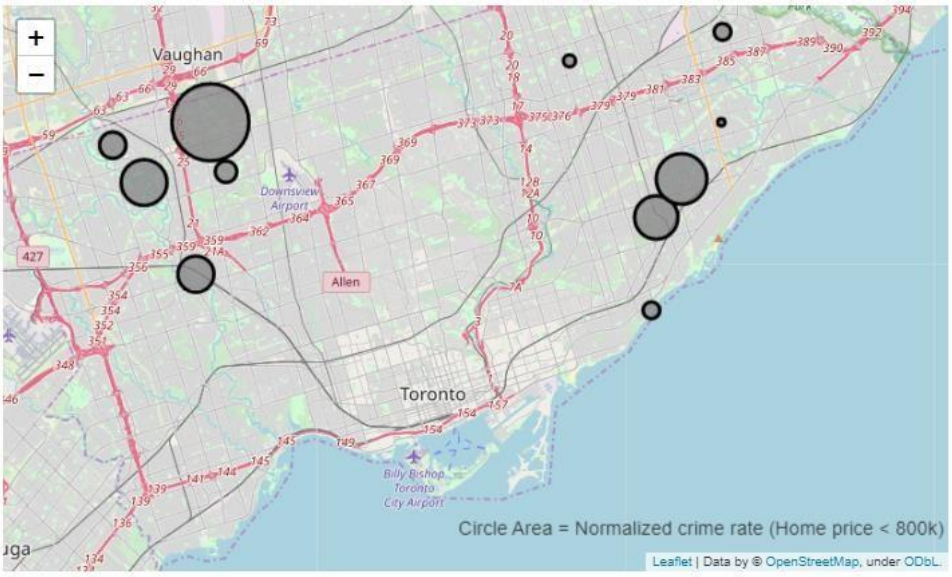

Figure 15. Neighborhoods in Cluster "0" with an average home price of < $800,000, where the size of the circles represents the normalized average crime rate.

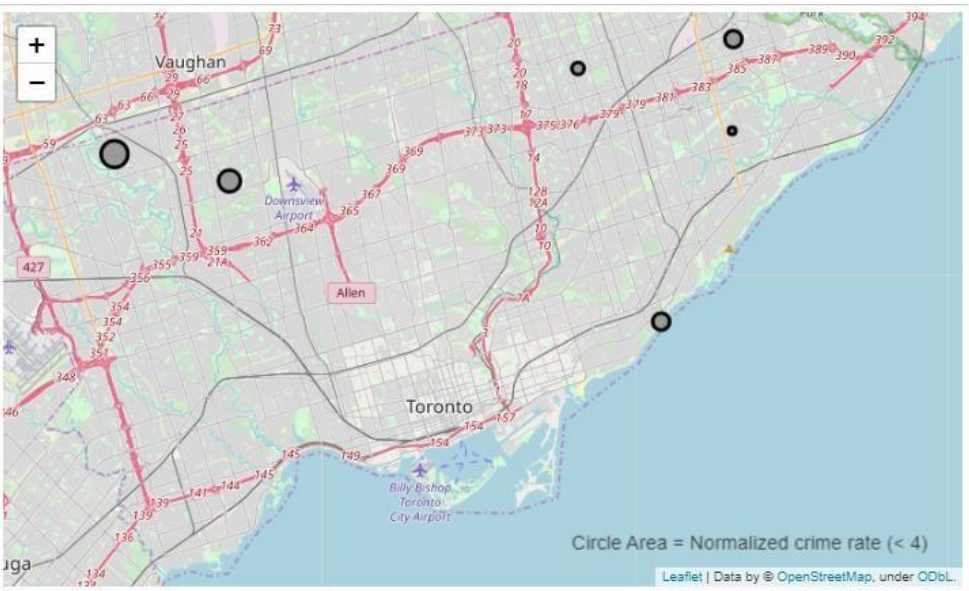

Figure 16. Neighborhoods in Cluster "0" with an average home price of less than $800,000, and an average crime rate of less than 4. The size of the circles represents the normalized average crime rate.

Following this filtering, I ended up with 6 neighborhoods (Humber Summit, Glenfield-Jane Heights, L'Amoreaux, Birchcliffe-Cliffside, Woburn, and Malvern) shown in Figure 16 and Table 2, which are the neighborhoods where I will be looking to buy a home.

	Neighbourhood	Population_Density	Average Crime	Average home price (2017)	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
19	Humber Summit	1.780867	2.251131	706722.0	Coffee Shop	Bank	Park	Asian Restaurant	Hardware Store
34	Glenfield-Jane Heights	5.074424	1.547167	745701.0	Pizza Place	Grocery Store	Vietnamese Restaurant	Fast Food Restaurant	Coffee Shop
83	L'Amoreaux	4.154920	0.529240	784794.0	Fast Food Restaurant	Chinese Restaurant	Coffee Shop	Pharmacy	Sandwich Place
89	Birchcliffe-Cliffside	4.332918	0.846891	725980.0	Coffee Shop	Thai Restaurant	Golf Course	Liquor Store	Beer Store
111	Woburn	2.830533	0.206713	746787.0	Fast Food Restaurant	Coffee Shop	Chinese Restaurant	Pizza Place	Bank
122	Malvern	4.042411	0.872694	692097.0	Fast Food Restaurant	Pharmacy	Pizza Place	Grocery Store	Sandwich Place

Table 2. Six neighborhoods in Toronto, where I will be looking for my new home, and their properties.

5. Discussion

In this case study I used the data about crime, population density, access to different venues categories in Toronto Neighborhood in order segment those neighborhoods, and find neighborhoods similar to Forest Hill South, but more affordable for home purchase. Properties of Forest Hill South that are appealing to me are low crime rates, medium population density, and reasonable access to coffee shops and parks. In other words, Forest Hill South is perfect!

But I need to able to find a neighborhood just like it, but where the house prices are less than $800,000. K-means clustering was used to segment the data, and using the Elbow test, I found that 5 clusters were appropriate for the City of Toronto.

Even though the neighborhoods were not evenly distributed among clusters, cluster assignment did seem to be representative of the data, at least with respect to crime rates and population density (Figures 7-10), since the averages of those two features were different between clusters with a small standard error (Figures 8 and 9). That did not seem to be the case with venue category features, however (Figure 11), though some small differences between clusters were observed. Such a behavior of venue category features is expected for 2 reasons. Firstly, there were over 300 of those features, and a small weight [1/(number of venue category features)] was assigned to each of them individually relative to the weight of crime rates and population density. Secondly, Toronto is a rather homogeneous City in terms of access to different categories of features. In the future analysis, I would further categorize venue types in order to reduce the number of features and increase their weight.

The Cluster that contained Forest Hill South, Cluster 0, has 72 members, but only 11 of those members had an average home price of smaller than $800,000. Furthermore, out of those 11 neighborhoods, I chose to concentrate on the ones that had an average crime rate of 4, and ended up with 6 neighborhoods that satisfied my home search requirements: Humber Summit, Glenfield-Jane Heights, L'Amoreaux, Birchcliffe-Cliffside, Woburn, and Malvern

6. Conclusions

In this case study I identified 6 neighborhoods in Toronto that are similar to Forest Hill South, my favorite neighborhood, but have an average home price of less than $800,000 (which all I can afford). This analysis will be very useful in my search for a new home. This analysis can be used by anyone to find neighborhoods similar to their favorite one, in terms of safety, population density and access to venues.