FINDING MY NEW HOME IN TORONTO

by Dar’ya Redka

November 12th, 2019

1. Introduction

1.1 Background

In such a large and heterogenous city like Toronto it might be difficult to figure out what neighbourhoods to concentrate on in one's search to buy a new home. Particularly, considering the current high house prices in certain areas of the city. When searching for a house with family in mind, one might want to concentrate on neighbourhoods 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. This report is a case study that's takes Forest Hill South neighbourhood as a point of reference for a comfortable, safe, enjoyable living (our case subject, myself, is currently renting an apartment here), and tries to find other neighbourhoods in Toronto that are similar to Forest Hill South, but are more affordable (i.e., have an average home price of less than $800,000) for our subject to be able to go from renting to owning a house (average home price in Forest Hill South in 2017 was $1.32 million in 2017, see below).

1.2 Problem

There are about 100-140 neighbourhoods in Toronto (depending on assignment). In order to determine neighbourhoods that are similar to Forest Hill, these neighbourhoods need to segmented based of their safety scores, population density, access to parks, restaurants, and coffee shops using a clustering technique. Once I determine which neighbourhoods properties of interest similar to those of Forest Hill South (i.e., the neighbourhood my subject knows she likes living in), out of those neighbourhoods I'll be able to find neighbourhoods that satisfy my budget restriction.

1.3 Interest

Whenever anyone meets with a real estate agent or visits a real estate website, it would be great to have a very good about what areas to limit your search to. Otherwise, the amount of data would be overwhelming and one might miss a home of their dreams due to the sheer volume of houses for sales around the city. Anyone buying a house would be interested in learning what neighbourhoods are similar to the one they already love, and for many people, crime rates and population density (i.e., traffic and social like in the neighbourhood) would be or more important than venues available in the area.

2. Data

2.1 Crime rate, population density, latitude and longitude of each neighbourhood: sources, cleaning, feature selection

In order to determine how safe each neighbourhood is, I need to obtain some measure of the crime `rates`. Toronto Police has an open data access to records about various crime ratings, such as assault, robbery, homicide, autotheft 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 neighbourhood in Toronto. The file also included the information about population and size of the neighbourhood (area in square-meters), as well as the coordinates of the boundaries of each neighbourhood.

I extracted the following properties from the geojson file: 'Neighbourhood', 'Assault_Rate_2018','AutoTheft_Rate_2018','BreakandEnter_Rate_2018', 'Robbery_Rate_2018','Homicide_Rate_2018','Population','Size_of_hood_area'. Then I used the information about each neighbourhood's population and size (area in square-meters) to calculate the population density ("Population"/ "Size_of_hood_area" I also used the neighbourhoods' population to calculate the crime rate per capita (times 100), as opposed to using counts as is. From those features I created my own data set with the following features: 'Neighbourhood','Population_Density','Assault_per_capita','AutoTheft_per_capita','BreakandEnter_per_capita','Robbery_per_capita','Homicide_per_capita'.

I kept different types of crimes separate for my analysis, and combined them in the end of the study for making conclusions.

There were 140 neighbourhoods in the geojson file from Toronto Police, and the format of their names and assignment differed from the neighbourhoods we used earlier in Module 3. Therefore I had to obtain another measure of the coordinates for the centers of the neighbourhoods that would match my data. To do so, I used the coordinates of the boundaries to estimate the center point of each area (i.e., I took averages of all latitudes and all longitutes provided for each neighbourhood). Those central coordinates were needed in order to carry out queries later in the report and for plotting neighbourhood clusters. I added the "Latitude' and "Longitude" features to my data set as well.

2.2 Venues: sources, cleaning, and feature selection

I queried Foursquare API in order find venues that are most popular in each neighbourhood of Toronto, using the central neighbourhood coordinates are I extracted from geojson file shared by Toronto Police website. The query returned 10922 venues, in 347 unique venue categories. The venue categories were converted to binary variable, using onehot encoding, and grouped by neighbourhood. The resulting data set had 140 rows and 347 columns (plus the neighbourhood index), which means that the query returned the data for all 140 neighbourhoods. When this data set was later merged with the data on crime.

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

After segmentation of the data I wanted to overlay it with the information about home prices for each neighbourhood. It was not included in the segmentation process because I didn't want to be one of the criteria for similarity between neighbourhoods. The home prices data were scraped from a blog post on Toronto home prices by neighbourhood for 2017, which luckily has the same format for the neighbourhood names and assignment (link). There were, however, 6 neighbourhoods missing from the dataset, and the missing values in those cases were replaced with average home price for all neighbourhoods. In future plots, those neighbourhoods 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.

The resulting dataframe was merged with the entire dataset, to be used for plotting maps.

3. Methodology

3.1 Tools

The *Folium* package for Python was used to plot Neighbourhoods 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 neighbourhood 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 neighbourhoods 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 Neighbourhood 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).

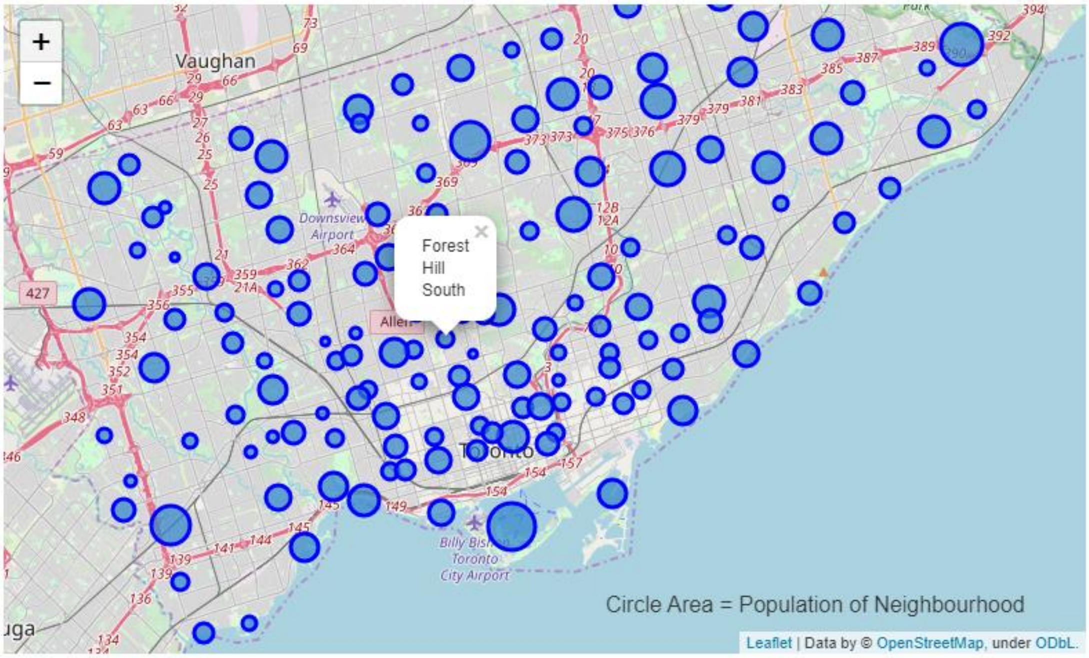

+
−
Vaughan
Downsview Airport
Forest Hill South
Allen
Toronto
Billy Bishop Toronto City Airport
Circle Area = Population of Neighbourhood
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 1. Toronto Neighbourhoods represented as circles, where the area of the circle represents Population of the Neighbourhood.

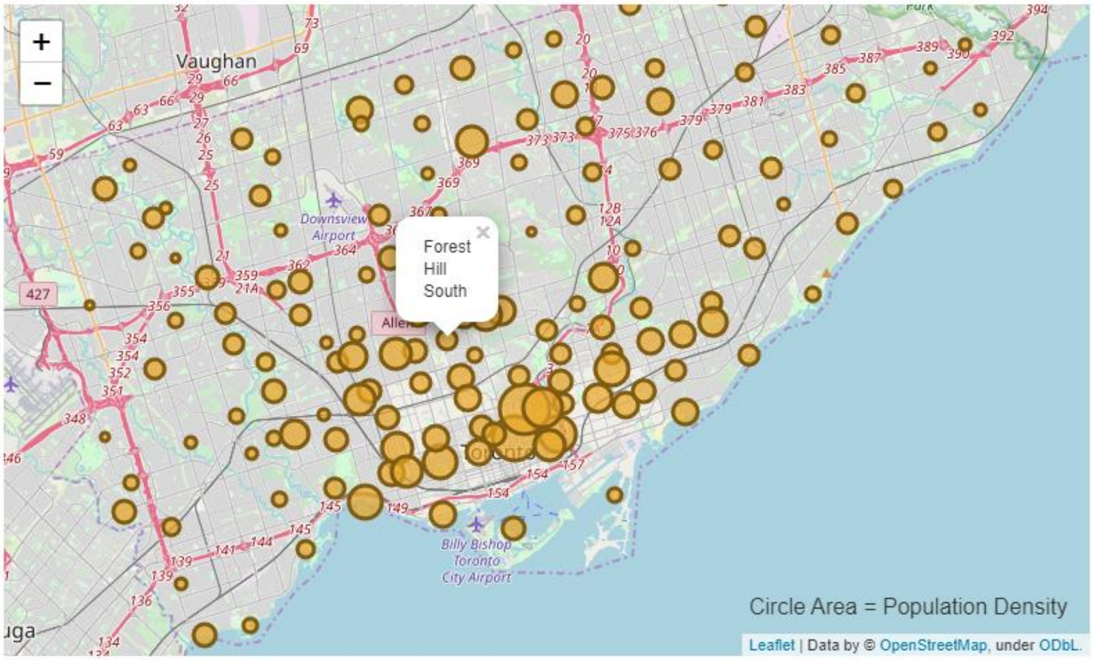

+
−
Vaughan
Downsview Airport
Forest Hill South
Allen
Toronto
Billy Bishop Toronto City Airport
Circle Area = Population Density
Leaflet | Data by © OpenStreetMap, under ODbL.

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

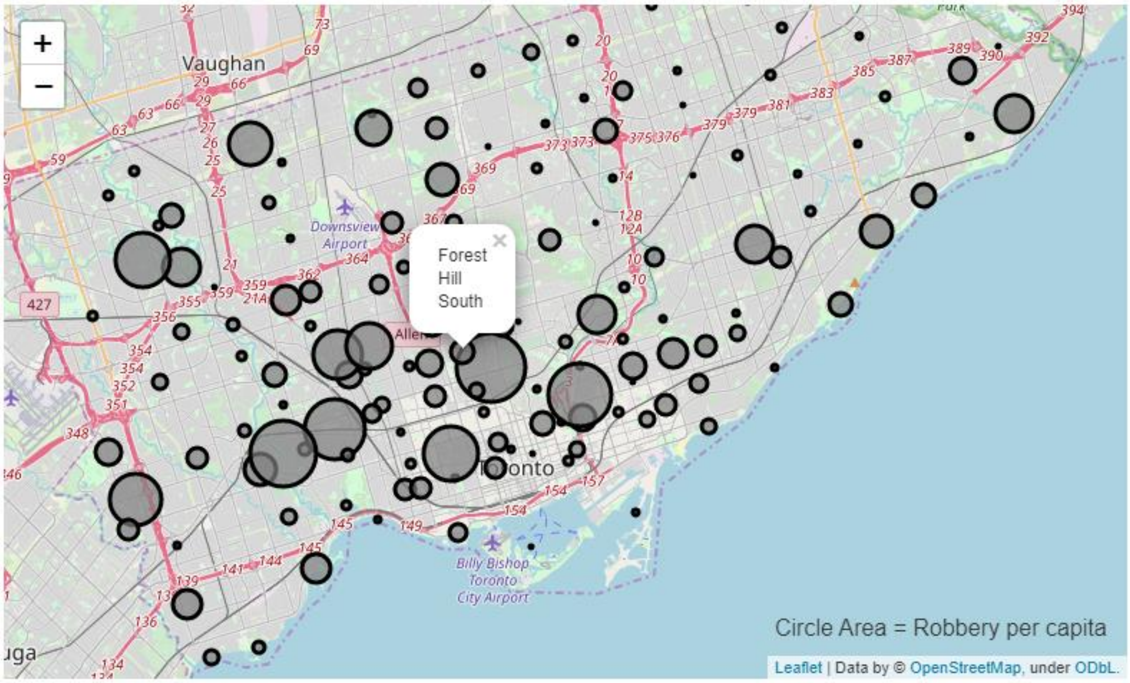

+
−
Vaughan
Downsview Airport
Forest
Hill
South
Allen
Toronto
Billy Bishop Toronto City Airport
427
Circle Area = Robbery per capita
Leaflet | Data by © OpenStreetMap, under ODbL.

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

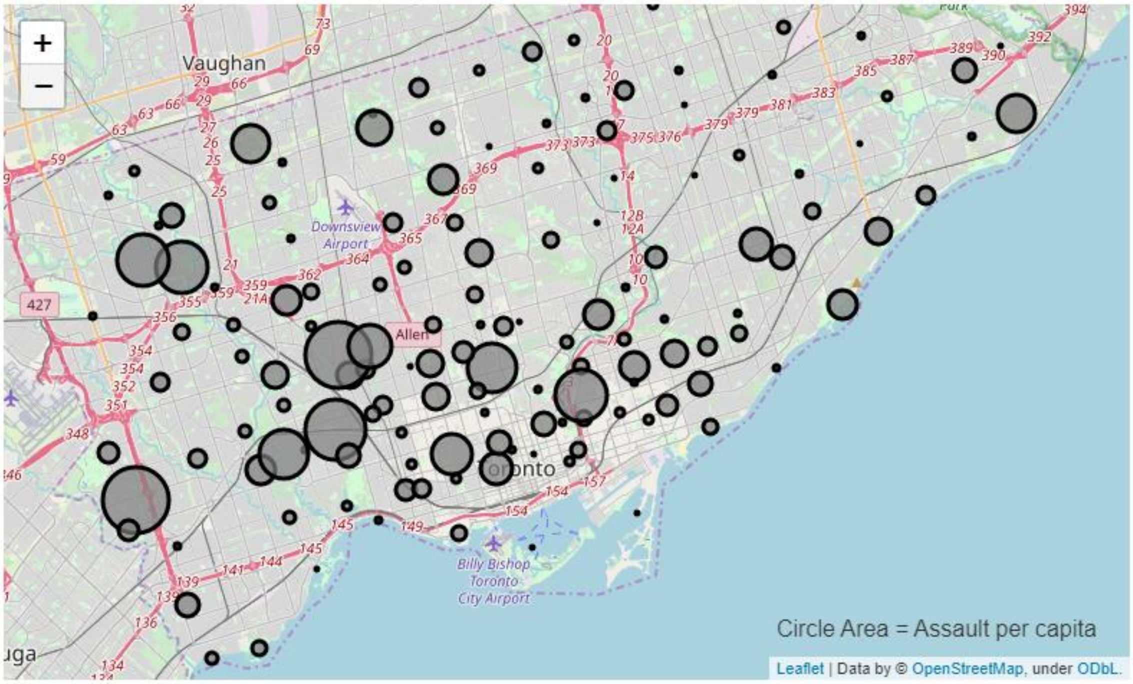

+
−
Vaughan
Downsview Airport
Allen
Toronto
Billy Bishop Toronto City Airport
427
Circle Area = Assault per capita
Leaflet | Data by © OpenStreetMap, under ODbL.

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

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.

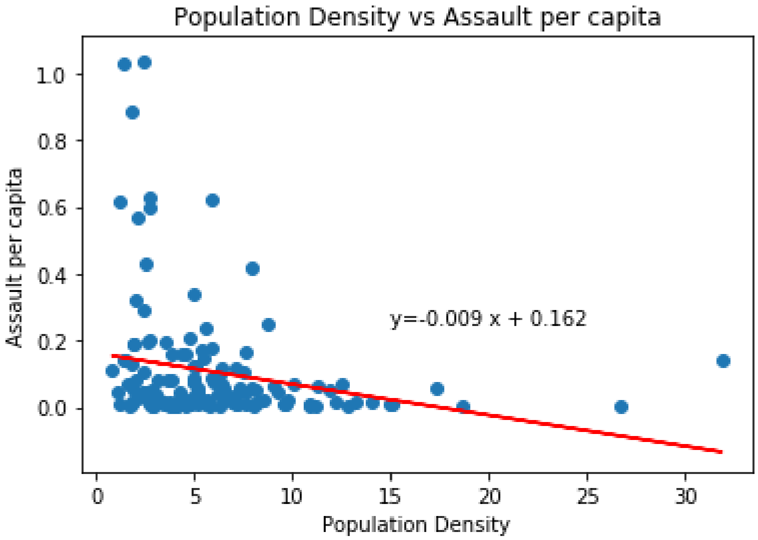


Figure 5. A scatter plot of Population Density and Assault Per Capita rate for Toronto Neighbourhoods.

3.3 Selection, normalization and weighting of features.

There are three main factors that I wanted to consider when selecting the right neighbourhood 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 neighbourhood 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.

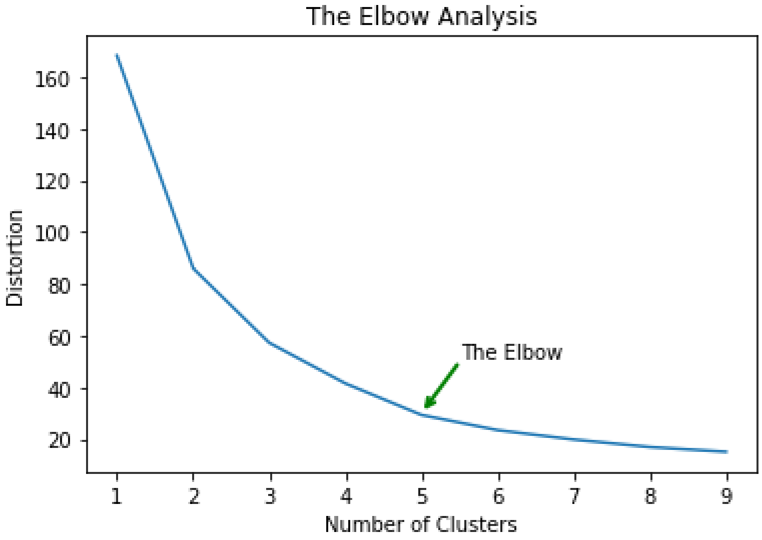


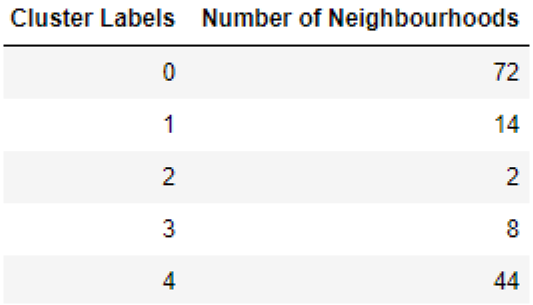
Figure 6. The sum of distance between each point and its cluster (i.e., distortion) as a function of the number of clusters.

4. Results

4.1 Clusters

K-Means clustering analysis segmented the Toronto neighbourhoods into five clusters represented by different clusters in Figure 7. Our neighbourhood 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 neighbourhood (calculated as described in *Methodology*). Cluster 0 is the largest cluster, with 72 members. The number of neghbourhoods in the other clusters is listed in Table 1.

Even though Cluster 0 has about 50% of the neighbouhoods, 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 colour 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 neighbourhoods 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 Neighbourhoods in Module 3 of this course.

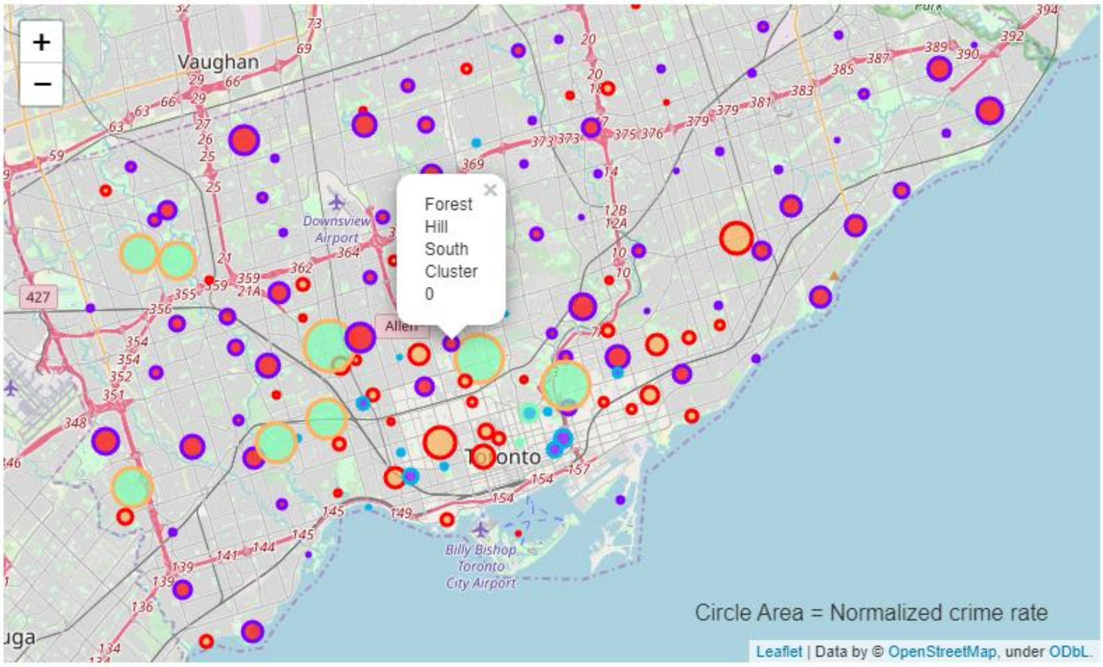

+
−
Vaughan
Downsview Airport
Forest Hill South Cluster 0
Allen
Toronto
Billy Bishop Toronto City Airport
427
Circle Area = Normalized crime rate
Leaflet | Data by © OpenStreetMap, under ODbL.

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

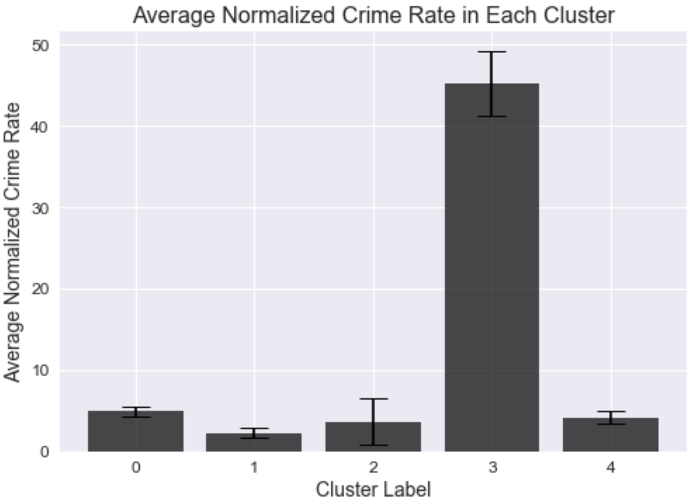


Figure 8. Average normalized crime rate (X 100) by cluster.

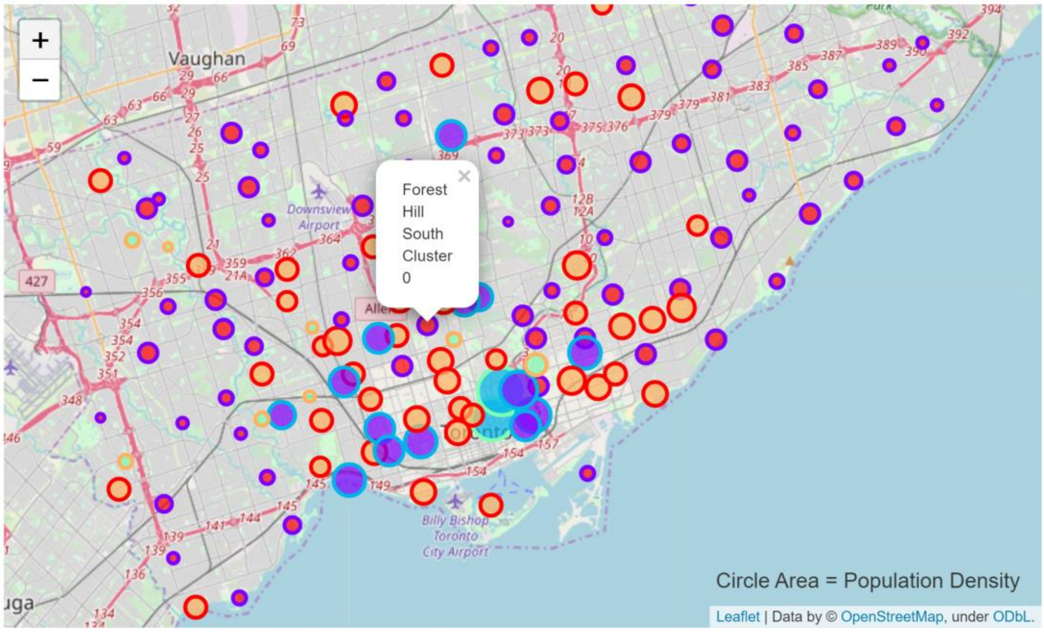

+
−
Vaughan
Downsview Airport
Billy Bishop Toronto City Airport
Toronto
Forest Hill South Cluster 0
×
Circle Area = Population Density
Leaflet | Data by © OpenStreetMap, under ODbL.

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

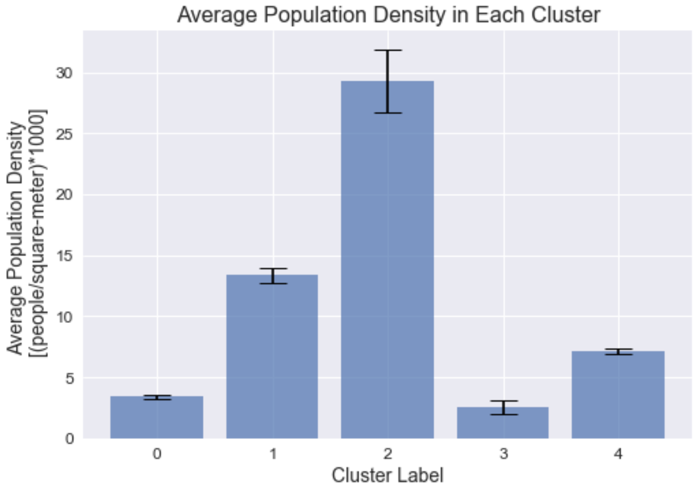


Figure 10. Average Population Density in each Cluster

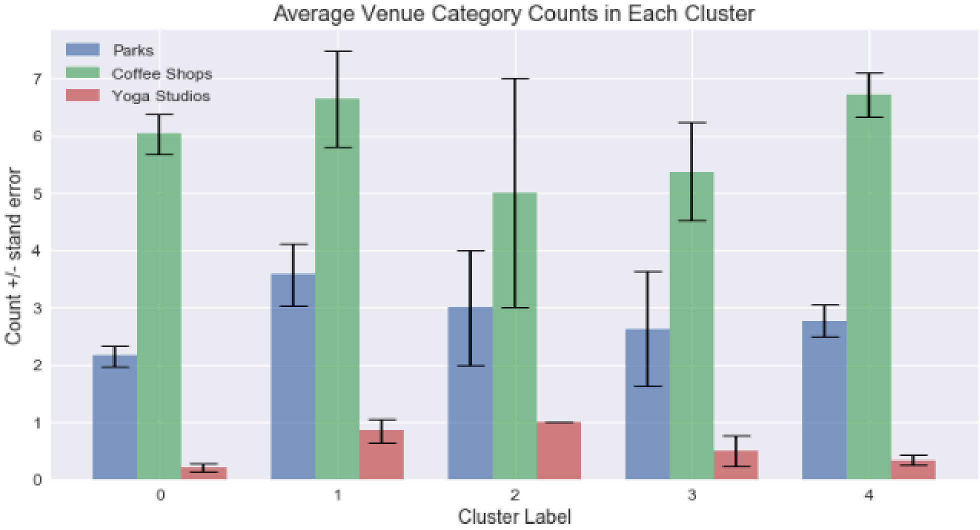


Figure 11. Average counts of parks, coffee shops and yoga studios in each cluster.

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

When the average home prices for each Neighbourhood were used as the circle area representing each Neighbourhood, 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 Neighbourhood with all my desirable properties, except for home price.

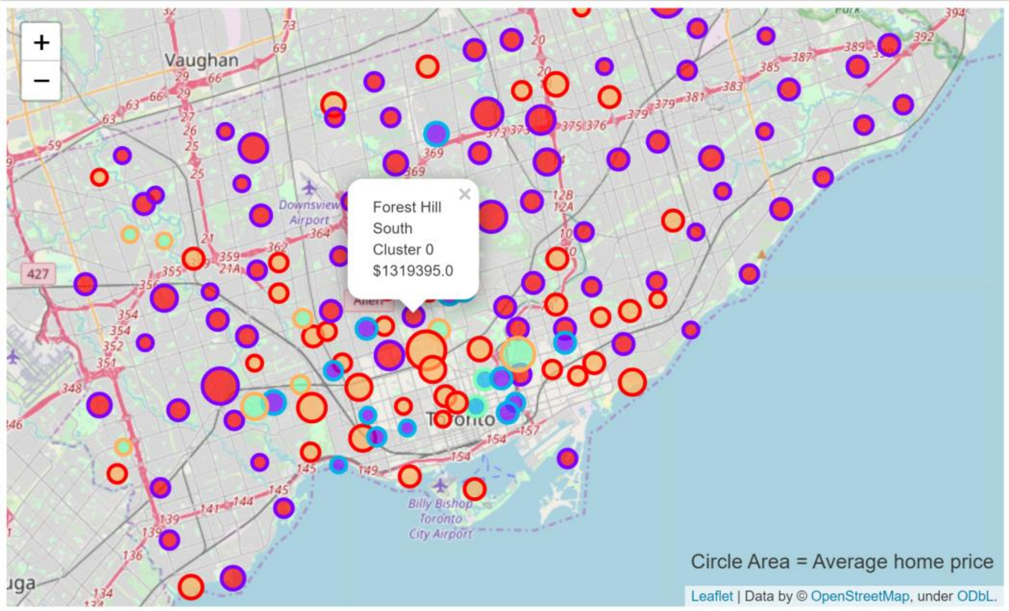

Forest Hill
South
Cluster 0
$1319395.0
Circle Area = Average home price
Leaflet | Data by © OpenStreetMap, under ODbL.
Vaughan
Downsview Airport
Toronto
Billy Bishop Toronto City Airport
+
−

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

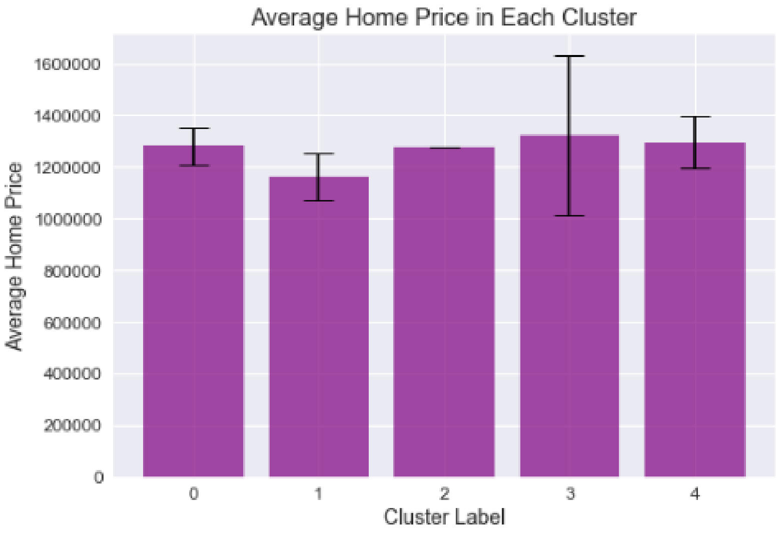


Figure 13. Average home prices in 2017 in each of five clusters of Toronto Neighbourhoods.

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

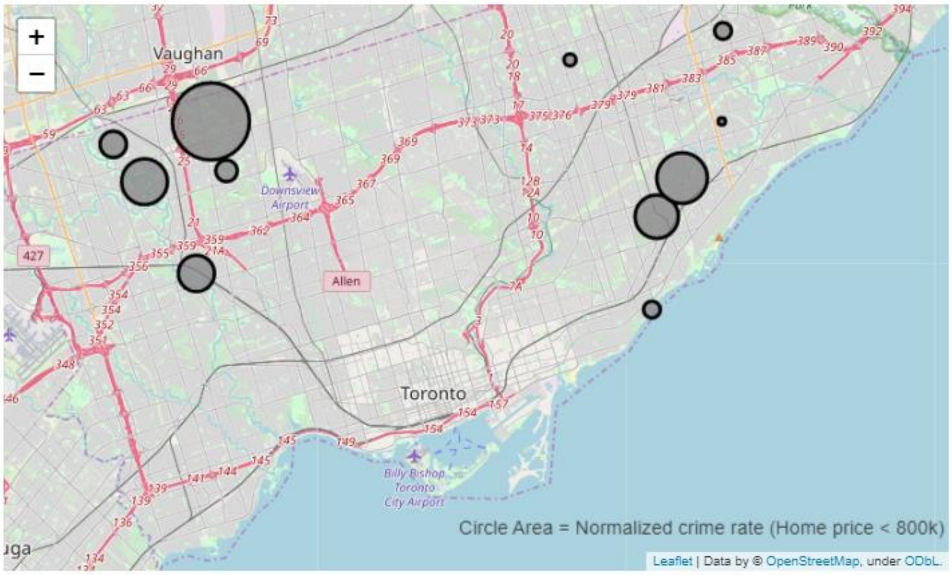

Vaughan
Downsview Airport
Allen
427
Toronto
Billy Bishop Toronto City Airport
Circle Area = Normalized crime rate (Home price < 800k)
Leaflet | Data by © OpenStreetMap, under ODbL.

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

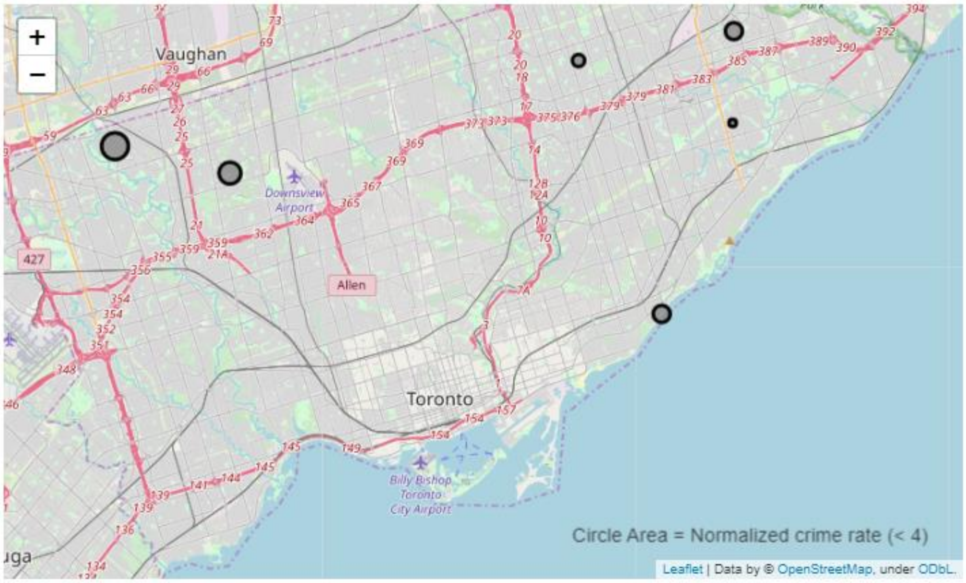
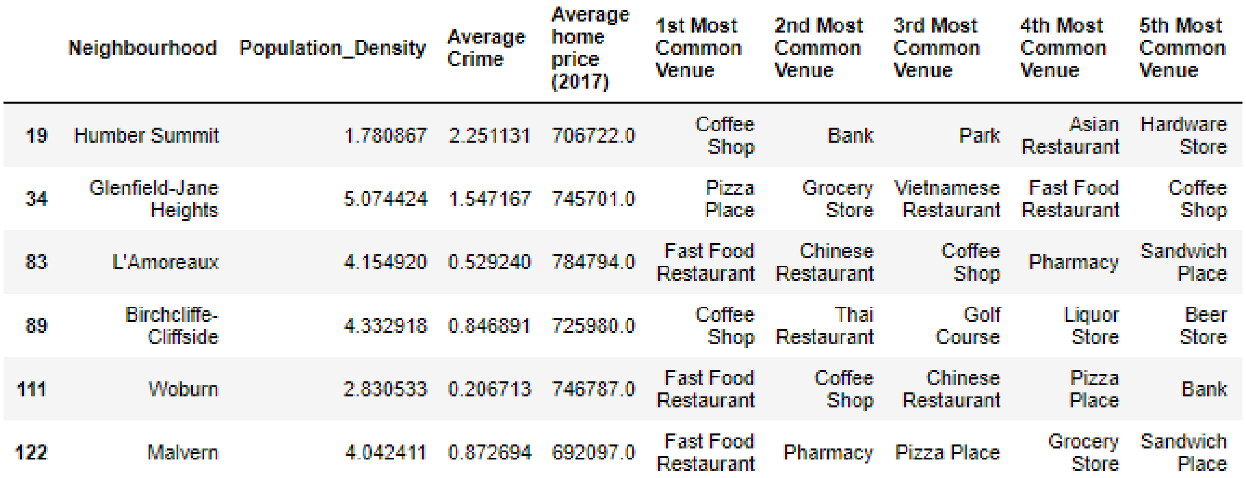

Vaughan
Downsview Airport
Allen
427
Toronto
Billy Bishop Toronto City Airport
Circle Area = Normalized crime rate (< 4)
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 16. Neighbourhoods 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 neighbourhoods (Humber Summit, Glenfield-Jane Heights, L'Amoreaux, Birchcliffe-Cliffside, Woburn, and Malvern) shown in Figure 16 and Table 2, which are the neighbourhoods 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 neighbourhoods 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 Neighbourhood in order segment those neighbourhoods, and find neighbourhoods 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 neighbourhood 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 neighbourhoods 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 neighbourhoods, I chose to concentrate on the ones that had an average crime rate of 4, and ended up with 6 neighbourhoods 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 neighbourhoods in Toronto that are similar to Forest Hill South, my favourite neighbourhood, 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 neighbourhoods similar to their favourite one, in terms of safety, population density and access to venues.