Coursera Capstone

IBM Applied Data Science Capstone

Finding Best Neighbourhoods for a New Home in Toronto, Canada

By: Suhail Ur Rahman

February 2020

Introduction

Problem Statement

In a diverse and huge city like Toronto, it might be intimidating to figure out which neighbourhoods are best in one's quest to buy a new home, considering the exorbitant prices of homes in certain neighbourhoods. Not only that, one needs to look into different factors like population density, low crime rates, access to parks and playgrounds, and restaurants and cafes while looking at the budget in mind. Here the neighbourhood of Forest Hill South is taken as a reference by the customer due to its high scores in various factors like low crime rate, access to restaurants etc. The only downside is that the homes in Forest Hill South are well above the customer's budget which is CAD 800,000. The average price of homes in Forest Hill South in 2017 was CAD 1.329 million (as shown in www.moneysense.ca).

There are well over 120 neighbourhoods in Toronto. So in order to find the neighbourhoods which are similar to Forest Hill South, one needs to segment the neighbourhoods based on their low crime rates, the density of population, close proximity to restaurants, cafes, parks etc. The segmentation will be done using clustering technique. After determining which neighbourhoods are similar to Forest Hill South in terms of the features mentioned above, I will be able to find neighbourhoods according to the customers budget and liking.

Target Audience

When a customer visits a real estate agent or a website dealing with real estate, it would be great to have a good idea about the areas to limit the search to. Otherwise, it will be overwhelming and quite intimidating for the customer to choose from over 120 neighbourhoods. Also, the customer will be interested in neighbourhoods similar to the one they like. The customer would also consider moving to neighbourhoods with low crime rates, population etc even though there are fewer venues like restaurants etc.

Data

To solve the problem, we need to find the following data:

- List of Neighbourhoods and their properties.
- Latitude and Longitude of the neighbourhoods.
- Venue data of the neighbourhoods.
- Average Prices of homes in each neighbourhood

Sources of data and methods to extract the data.
To find the list of neighbourhoods and their crime rates in order to determine the safety of the neighbourhood, I accessed the website of Toronto Police. They have open data access to various crimes' ratings from 2014 to 2018. They had a geojson file(https://opendata.arcgis.com/datasets/af500b5abb7240399853b35a2362d0c0_0.geojson) which helped in determining the crime according to the neighbourhoods. The following information would be extracted from the dataset: "Neighbourhood", "Assault_Rate_2018", "AutoTheft_Rate_2018", "BreakandEnter_Rate_2018", Robbery_Rate 2018", "Homicide_Rate_2018", "Population", "Size_of_hood_area". The Population Density would be calculated by dividing "Population" with "Size_of_hood_area".The crime rates were calculated according to the population i.e. per capita.

They were 140 neighbourhoods in the geojson file. Since the coordinates were of boundaries, the centre coordinates of each neighbourhood have to be calculated according to the average of all latitudes and longitudes. This would be needed in order to plot the neighbourhood clusters.

To find the venue data, Foursquare API was used. It would show the most popular venues in each neighbourhood of Toronto while using the central coordinates of the respective neighbourhood.

After Segmenting the data, it would be combined with the average home prices of each neighbourhood. The data will be obtained from the blog post from the following link https://www.moneysense.ca/spend/real-estate/where-to-buy-2019-toronto/. There are some neighbourhoods which are missing from the dataset, and the missing prices will be replaced with the mean price of homes across all neighbourhoods.

Methodology

Tools
The following tools were used:

- The *.read_html* from *pandas* library to scrape the table data from website.
- The *json* package was used to open and read the geojson file.
- The *KMeans* module from *sklearn* package was used to cluster the data.
- The *Foursquare API* was used to obtain venue data.
- The *Folium* package was used to plot Neighbourhoods on the map of Toronto.

Exploratory Data Analysis

If we see the size of neighbourhoods, it varies. Therefore we need to convert the population (Figure 1) into population density(Figure 2) so that we can use it for clustering of the data. However, the population data needs to be retained, so we can convert crime rates into per capita crime rates.

When we plot the crime rates per capita, for example, “Assault per capita”(Figure 3) or “Robbery per capita”(Figure 4) as the circle size for a neighbourhood, we can see that the crime rate vary very differently from other neighbourhoods. Same is the case with other crimes.

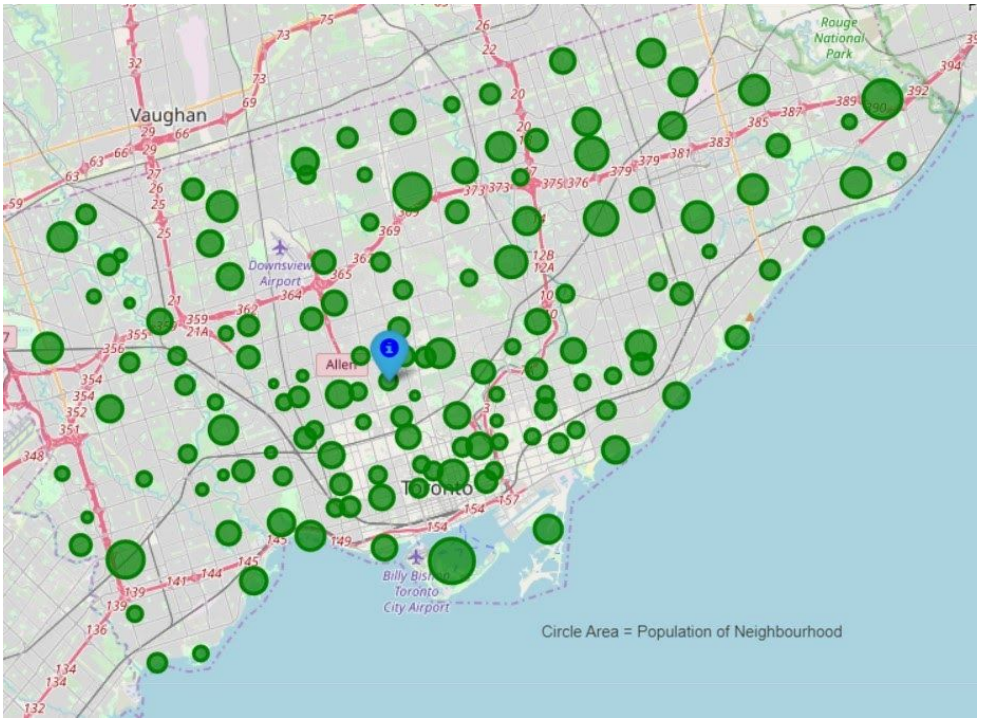

Figure 1. Neighbourhoods represented as circles, where the circle size represents the population in the respective neighbourhood.

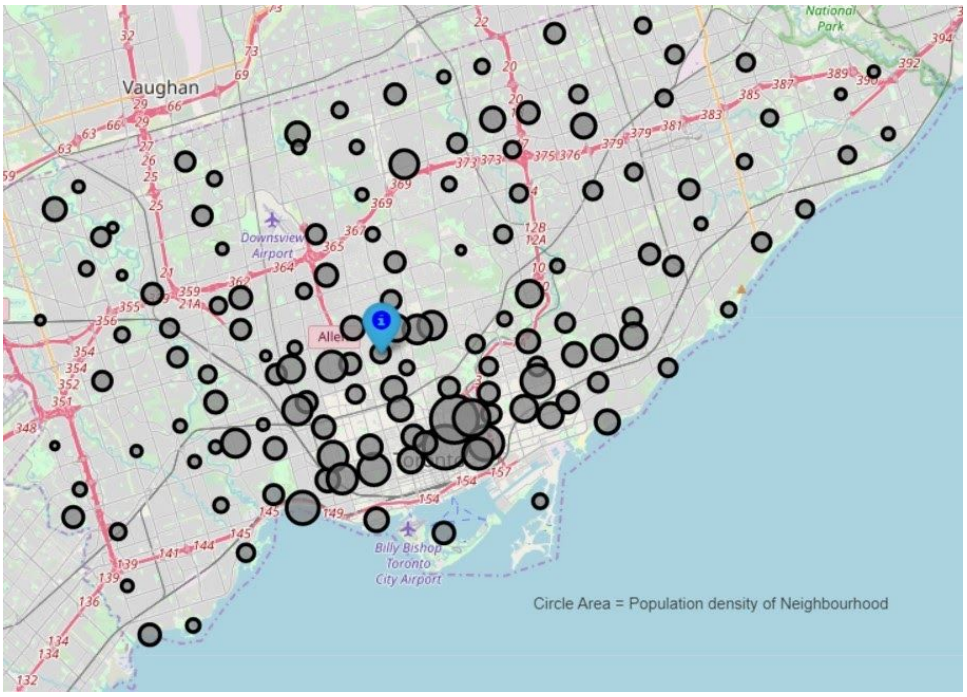

Figure 2. Neighbourhoods represented as circles, where the circle size represents the population density in the respective neighbourhood.

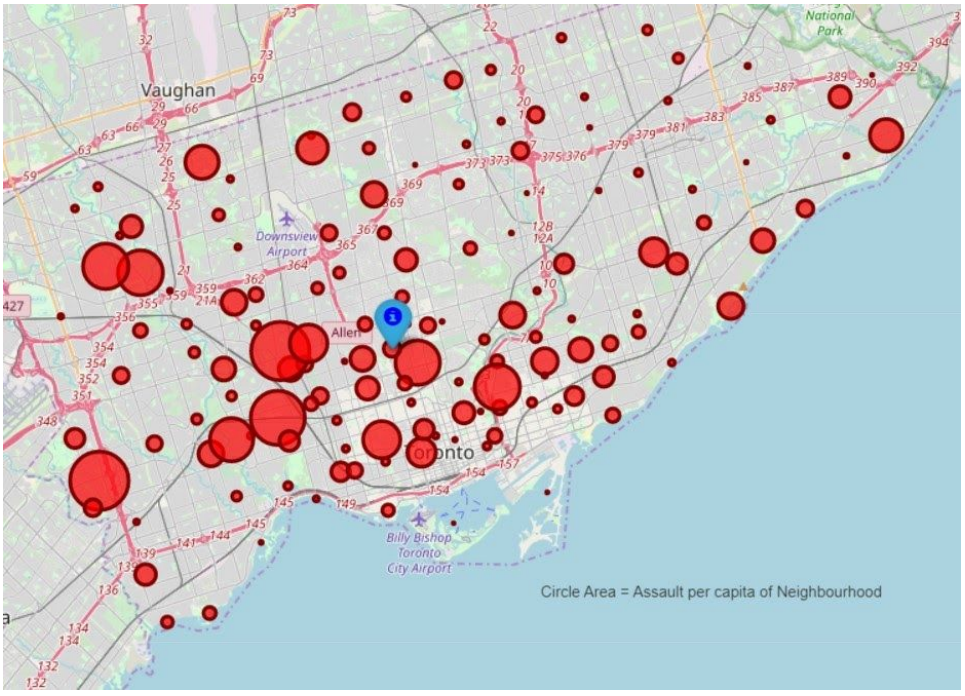

Figure 3. Neighbourhoods represented as circles, where the circle size represents the Assault per capita(x100) in the respective neighbourhood.

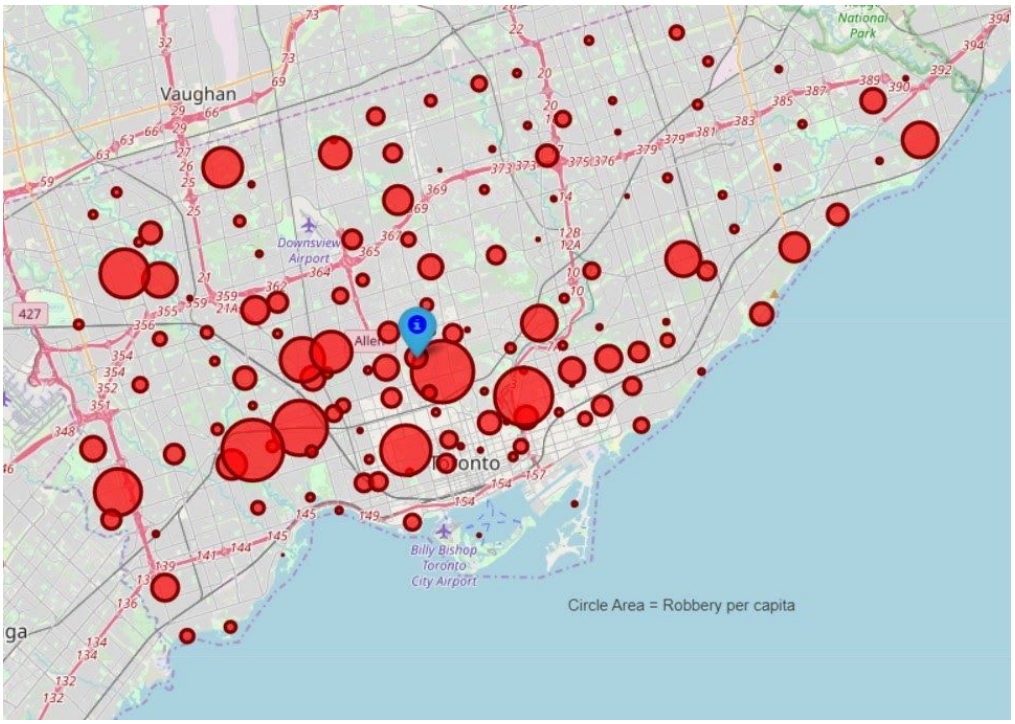

Figure 4. Neighbourhoods represented as circles, where the circle size represents the Robbery per capita(x100) in the respective neighbourhood.

The next step was to find whether we can use both population density and crime(for example, Assault per capita) for clustering. The correlation coefficient was calculated between population density and Assault per capita. The value was -0.226, hence it was obvious there was no correlation between the two(as shown in Figure 5). Hence both features could be used for clustering.

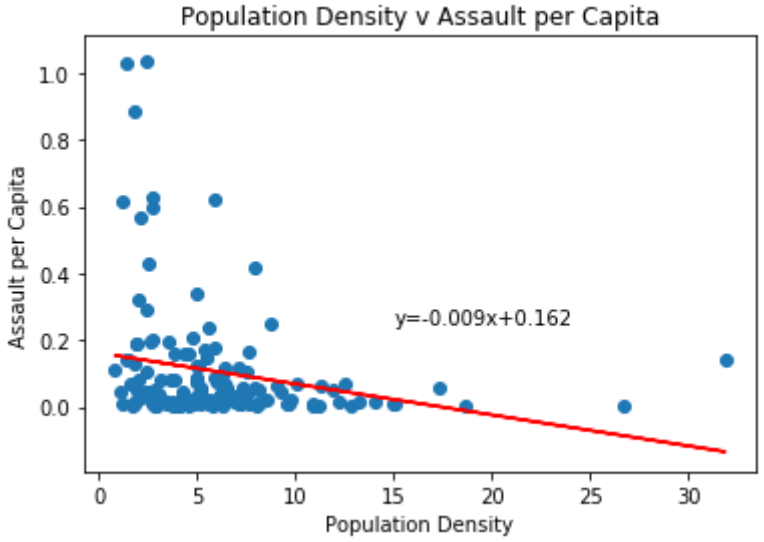

Figure 5. A scatter plot of Population Density vs Assault per capita for neighbourhoods in Toronto

Selection, Normalization and Weighting of features of the dataset

Three factors were the most important when selecting a neighbourhood for a new home: Population Density, Low Crime Rate and Access to Venues. For the crime rates, five measure of crimes were taken, namely Assault, Auto Theft, Break and Enter, Homicide, and Robbery, a measure of population density and 334 venue category features. The three factors should be normalized and adjusted to their corresponding weights so that we can use those factors for clustering analysis. Therefore, *StandardScaler* module from *sklearn.preprocessing* was used to normalize all the 340 features to mean of 0 and variance of 1. Then we need to divide the 5 crime features by 5, while the 334 venue categories were divided by 335. The population density will remain the same.

A new variable called Normalized Crime Rate was calculated following the k-means clustering. The Normalized Crime Rate was calculated by taking the average of the normalized crime features namely Assault, Auto Theft, Break and Enter, Homicide, and Robbery.

Clustering

Here, K-Means clustering was used to segment the neighbourhood data. Initially, we need to find the best number of clusters to be used. First, the elbow method was used to determine the best value k to be used. To that end, the number of clusters between 1 to 9 was tested by plotting the distortion function. Here the distortion function is the sum of distance for each point to its cluster centre. This is obtained by using *inertia_* data from the k-means fit and plotting it against the number of clusters. The rate of distortion was reduced at k=5 as shown in Figure 6. Hence the number of clusters for the k-means analysis will be 5.

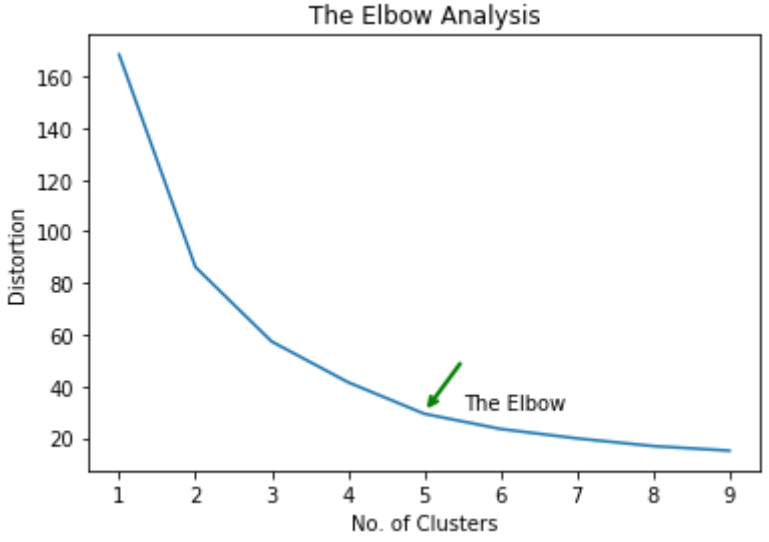

Figure 6. Plot of Distortion vs Number of clusters.

Results

Clusters

The K-Means clustering analysis segmented the neighbourhood data into 5 clusters, each cluster represented from 0 to 4. Our initial neighbourhood of interest, Forest Hill South, belongs to cluster. The following figure shows the map of Toronto with the size of circles representing normalized crime rate for each neighbourhood. Cluster 0 is the largest with 72 members. The number of neighbourhoods in other clusters are listed in Table 1.

	Cluster Labels	No. of Neighbourhoods
0	0	72
1	1	14
2	2	2
3	3	8
4	4	44

Even though cluster 0 has around 50% of neighbourhoods, it appears the clustering is reflective of the data. The figure below also confirms this, as the clusters of the same colour have similar sizes. Figure 8 shows that the average normalized crime rate varies from cluster to cluster. Figure 9 shows the association of population density with that of the cluster assignment. Figure 10 shows that clusters vary in their average population density.

Since we have more than 300 venue features with their adjusted weights, it would be difficult to obtain similar map plots and bar graphs as shown in Figures 7-10, showing the dependence of cluster assignment on any particular venue. Figure 11 shows the average count of three venue categories in each cluster namely Coffee Shop, Park, and Yoga Studio. Only Yoga Studio appears to vary between the clusters. The lack of dependence on venue categories is unsurprising considering the fact the weights were set up for each feature individually, were low crime rate and population density were more important.

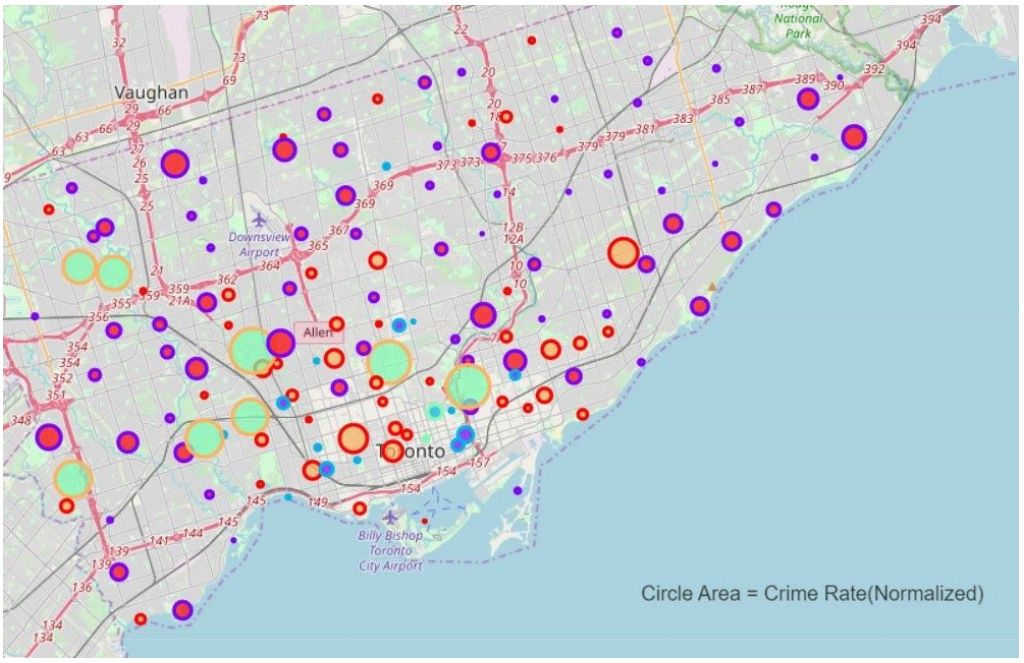

Figure 7. The 5 clusters of neighbourhoods represented as circles of different colours, where the circle size represents the Crime Rate(Normalized) in the respective neighbourhood.

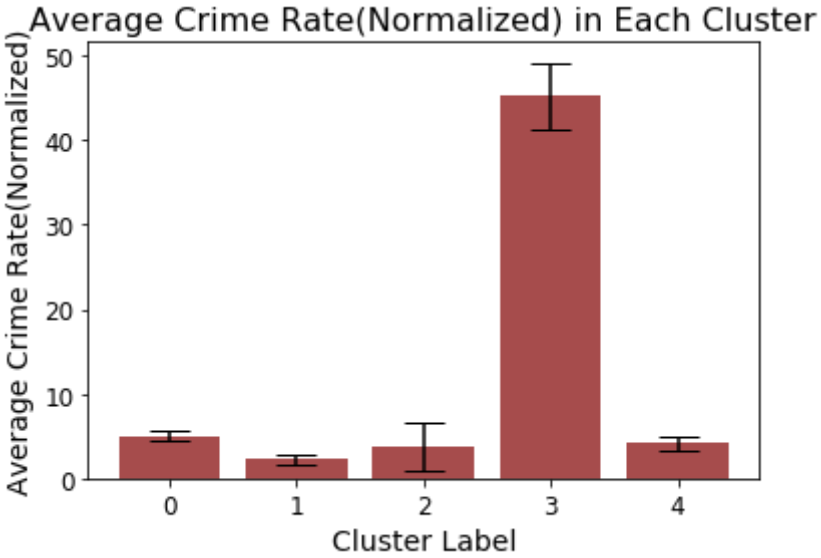

Figure 9. Average Normalized Crime Rate(x100) in each cluster.

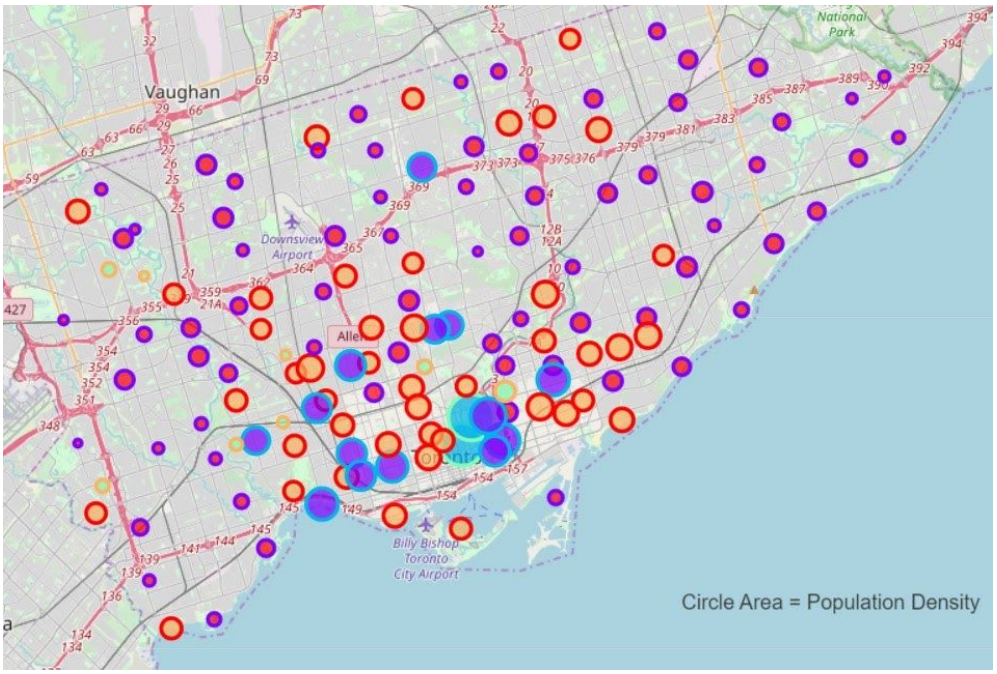

Figure 9. The 5 clusters of neighbourhoods represented as circles of different colours, where the circle size represents the population density in the respective neighbourhood.

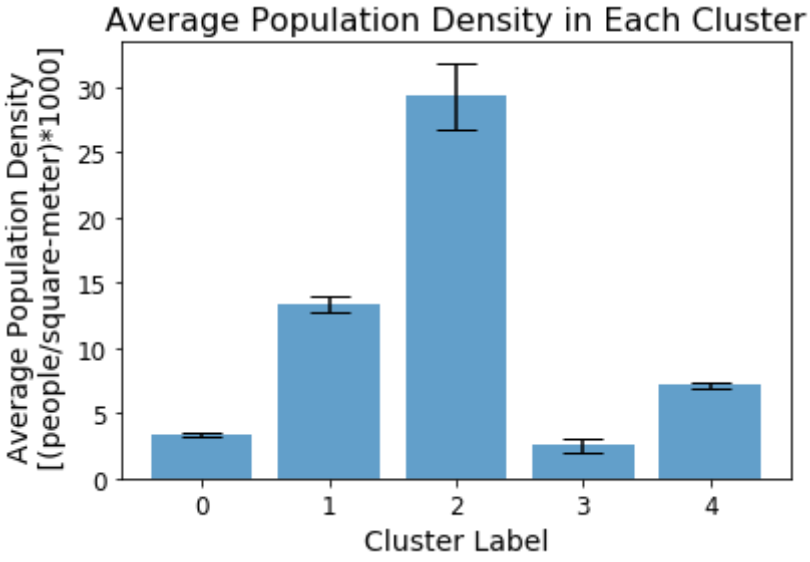

Figure 10. Average Population Density in each cluster.

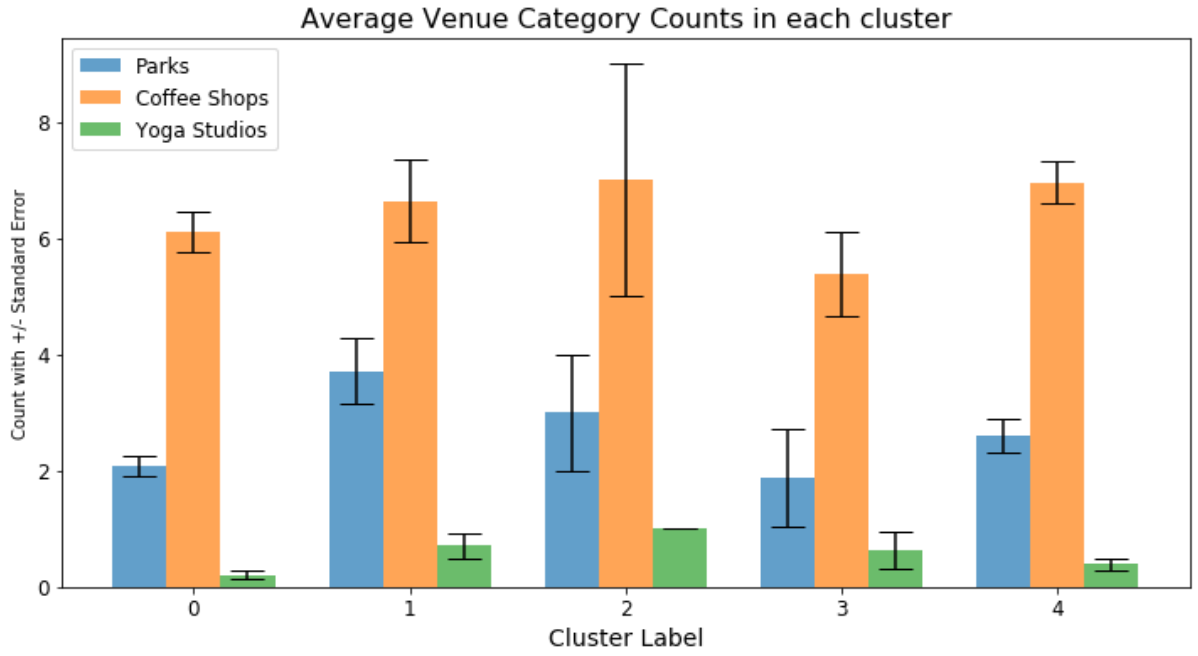

Figure 11. Average counts of parks, coffee shops and yoga studios in each cluster.

Extracting a list of neighbourhoods that are similar to Forest Hill South but affordable

When the average home price for each neighbourhood was used as the circle size representing each neighbourhood(Figure 12), there was no pattern in circle size and cluster assignment. The mean price of every cluster was also similar(Figure 13). This is unsurprising considering the fact that we did not want home prices to affect clustering. Now we can have look at the cluster 0, which contains the neighbourhood of our liking, with only the price being the exception.

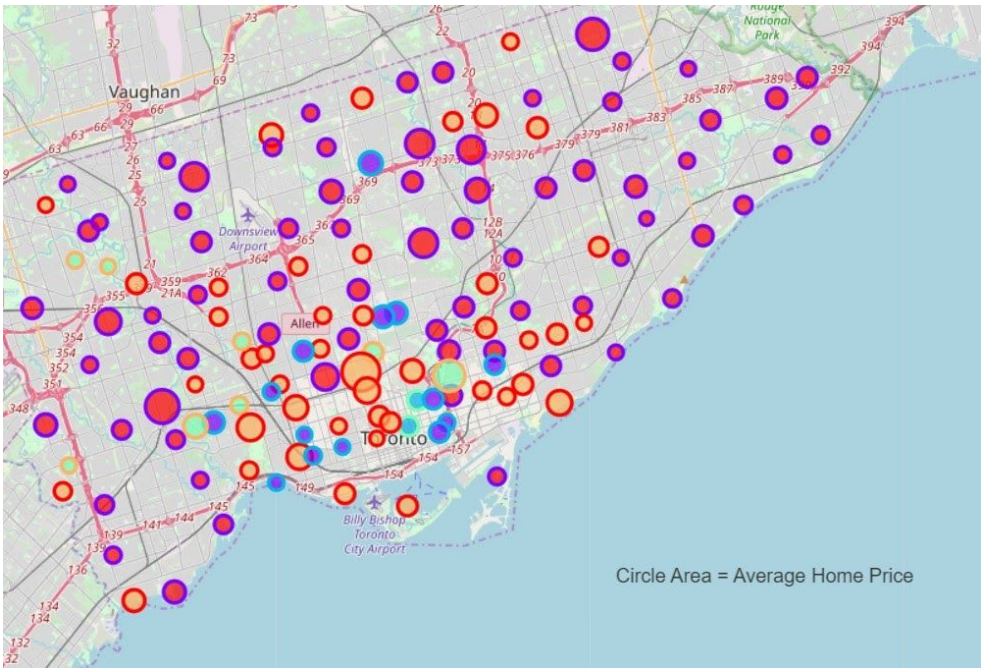

Figure 12. The 5 clusters of neighbourhoods represented as circles of different colours, where the circle size represents the Average Home Price in the respective neighbourhood.

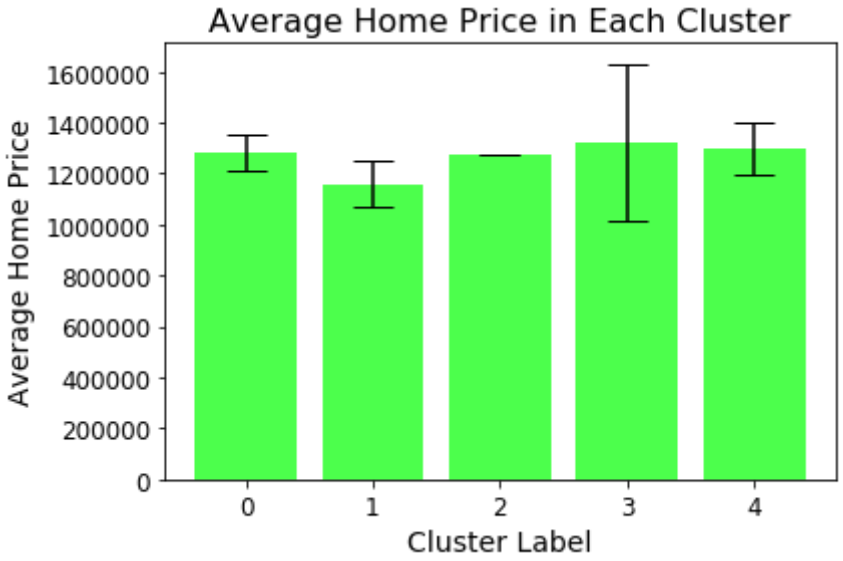

Figure 13. Average Home Prices in each cluster

Figure 14 shows the normalized average crime rate in Neighbourhoods in cluster 0 that have an average home price less than 800,000 CAD. There are around 11 neighbourhoods that satisfy the requirement. Since the low crime rate is also important, additional filtering was used to get the neighbourhoods whose crime rates were less than 4. The result was 6 neighbourhoods with the price less than 800,000 CAD and crime rate less than 4(Figure 15). These are the neighbourhoods to look at for a new home(Table 2).

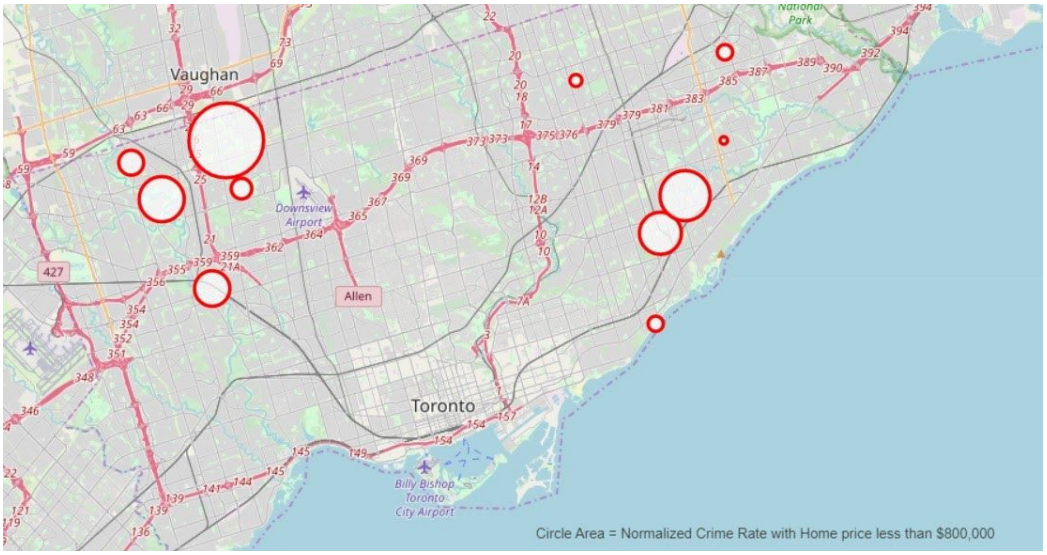

Figure 14. Neighbourhoods in Cluster 0, where the circle size represents the Normalized Crime rates with Average Home prices less than 800k CAD in the respective neighbourhood.

Figure 14. Neighbourhoods in Cluster 0, where the circle size represents the Normalized Crime rates<4 along with Average Home prices less than 800k CAD in the respective neighbourhood.

	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	Electronics Store	Bank	Coffee Shop	Asian Restaurant	Park
34	Glenfield-Jane Heights	5.074424	1.547167	745701.0	Pizza Place	Coffee Shop	Grocery Store	Gas Station	Park
83	L'Amoreaux	4.154920	0.529240	784794.0	Fast Food Restaurant	Chinese Restaurant	Coffee Shop	Pharmacy	Park
89	Birchcliffe-Cliffside	4.332918	0.846891	725980.0	Coffee Shop	Restaurant	Grocery Store	Gym	Diner
111	Woburn	2.830533	0.206713	746787.0	Coffee Shop	Fast Food Restaurant	Pizza Place	Pharmacy	Sandwich Place
122	Malvern	4.042411	0.872694	692097.0	Fast Food Restaurant	Grocery Store	Pizza Place	Pharmacy	Sandwich Place

Table 2. Six Neighbourhoods along with their features, which are the best fit for a new home.

Discussion

In this case study, the data about crime, population density, access to different venues were considered in order to segment the neighbourhoods similar to that of Forest Hill South, but more affordable. Since Forest Hill has low crime rates, medium population density and reasonable access to coffee shops, they were chosen. The only problem was the average price range exceeded the budget. So we needed to find neighbourhoods comparable to Forest Hill South but with house prices less than 800,000 CAD. K-Means clustering was used to segment the data, and the number of clusters k was found using the elbow test. Here the appropriate value of k was found to be 5.

Even though the neighbourhoods were not evenly distributed among the clusters, it seemed to be representative of data, with respect to crime rates and population density as the averages of those two features were different between clusters with small standard error. However, with venue category features, it wasn't the case although small differences between clusters were observed. This was to be expected as there were over 300 venue features and a small weight (1/(number of venue category features)) was assigned to each of them individually relative to the weight of crime rate and population density. In a future analysis, we could further categorize venue types, hence reducing the number of features and improving their weights as a result.

The Cluster that contained Forest Hill South has 72 members. However, only 6 neighbourhoods fulfilled the criteria of crime rate less than 4 and average price lower than 800,000 CAD. The neighbourhoods which satisfied the requirements are as follows: Humber Summit, Glenfield-Jane Heights, L'Amoreaux, Birchcliffe-Cliffside, Woburn, and Malvern.

Conclusions

In this case study, 6 neighbourhoods were identified in Toronto with similarity to Forest Hill South but with an average price less than 800,000 CAD. This analysis will help the customer in their search to find a new home. This analysis also can be used by other customers to find neighbourhoods similar to their favourite neighbourhood, in terms of population density, low crime rate and access to venues.

References

Toronto Police Open Data Portal
http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-

Moneysense Blogpost
https://www.moneysense.ca/spend/real-estate/where-to-buy-2019-toronto/

Foursquare Developers Documentation
https://developer.foursquare.com/docs