FIND ME A HOME

An Analytics approach to segment Toronto's neighbourhoods

Abstract

When a person plans to migrate to a new place, the first thing they search for is a House. However, their decision to rent/buy a house is based on various aspects including crime rate, local employment rate, child care services and nearby venues. Using Data Science and Machine Learning, we will group different neighbourhoods of Toronto into various segments and provide suggestions on which segment could be better for users.

Nitish Bhardwaj

Contents

Table of Figures

1. Business Problem

Whenever somebody migrates to a new place, they instantly start their search for an ideal house. It is generally noticed that a decision to buy/rent a house is highly supported by the location [1][2] like nearby venues of the house or the crime rate, and local employment rate. The married people having children give preference to places which have nearby childcare spaces. It becomes difficult for a user to find all these attributes. They end up doing their research either by visiting the location or by visiting various websites.

Toronto is a highly populated city in the Ontario province of Canada. It is also considered as the Silicon Valley of Canada [3]. Every month a lot of people do migrate to Toronto in search of better opportunities and a world-class lifestyle.

This project addresses the business problem of the users who are looking for a house in Toronto. The solution will suggest the most common venues near the neighbourhoods as well as the crime rate, local employment rate and several childcare spaces in the area.

2. Data Collection

The data for this research problem is based on various sources. Firstly the data of Toronto neighbourhoods having geolocation details like the Latitude and Longitude is fetched from the "Open Canada Neighbourhood" website [4]. Secondly, data for the home prices, childcare spaces and local employment is fetched from a URL [5] of "Open Canada". Thirdly, the data for crime-related cases in each neighbourhood is fetched from another URL [6]. At this step, the data contains 140 unique neighbourhoods of Toronto and the other fetched details.

Once we have a single dataset having the geolocation details of each neighbourhood in Toronto, the details of nearby venues are fetched for each location using Foursquare API [7].

The data is then cleaned and processed to have the final dataset ready to feed for the clustering algorithm.

3. Data Understanding

To understand the data, a thorough Exploratory Data Analysis(EDA) is performed on the created dataset. No missing values were found in the data fetched from Open Toronto Data websites[4-6]. Please refer to figure 1 of the captured missing values information on the Dataframe of the Open Toronto Neighbourhood details dataset.

Verify the number of missing values in the dataset:

```
Neighbourhood                 0
Neighbourhood_Id              0
LONGITUDE                     0
LATITUDE                      0
Home Prices                   0
Child Care Spaces             0
Local Employment              0
Arsons                        0
Assaults                      0
Break & Enters                0
Drug Arrests                  0
Fire Medical Calls            0
Fire Vehicle Incidents        0
Fires & Fire Alarms           0
Hazardous Incidents           0
Murders                       0
Robberies                     0
Sexual Assaults               0
Thefts                        0
Total Major Crime Incidents   0
Vehicle Thefts                0
dtype: int64
```

Figure 1: Snapshot of Missing values in the Open Data based neighbourhood details

Next, the venue data fetched from the Foursquare API was analyzed. Please refer to figure 2 for the venue details fetched using FourSquare API.

	No. of Venues per Neighbourhood
Church-Yonge Corridor	100
Mount Pleasant West	70
Bay Street Corridor	64
Junction Area	62
Dufferin Grove	59
...	...
Newtonbrook West	1
Princess-Rosethorn	1
Rustic	1
Willowdale East	1
Willowridge-Martingrove-Richview	1

139 rows × 1 columns

View the total unique categories of venues

```
print('There are {} uniques categories.'.format(len(toronto_venues['Venue Category'].unique())))
```

```
There are 282 uniques categories.
```

Figure 2: A screenshot of the details of venues fetched using FourSquare API

As shown in figure 2, it was found that 282 unique categories of venues were fetched. When the venue details were grouped by neighbourhood, it was uncovered that 139 neighbourhood details were fetched

from FourSquare API. Figure 3 shows a screenshot of the missing venue details for one of the neighbourhood.

```
#Printing the rows for the neighbourhoods for which no data is fetched from FourSquare API
toronto_grouped[toronto_grouped.isna().any(axis=1)]
```

	Neighbourhood	Neighbourhood_Id	LONGITUDE	LATITUDE	Home Prices	Child Care Spaces	Local Employment	Assaults	Break & Enters	Drug Arrests	...	Video Store	Vietnamese Restaurant	Warehouse Store
50	St.Andrew-Windfields	40	-79.379037	43.756246	1363202	124	13023	63	112	5	...	NaN	NaN	NaN

1 rows × 297 columns

Figure 3: A screenshot of missing venue details for four neighbourhoods

Though only one neighbourhood out of 140 has missing venue details, however, every neighbourhood was important in this analysis. So, the missing venue details were handled in the data preparation step.

Using visualization of Crime cases recorded in each neighbourhood, it was found that Fire Medical Cases occur most in Toronto city and the least number of crime-related cases are Murders. Please refer to figure 4 showing a bar chart of the crime-related cases in Toronto arranged in descending order.

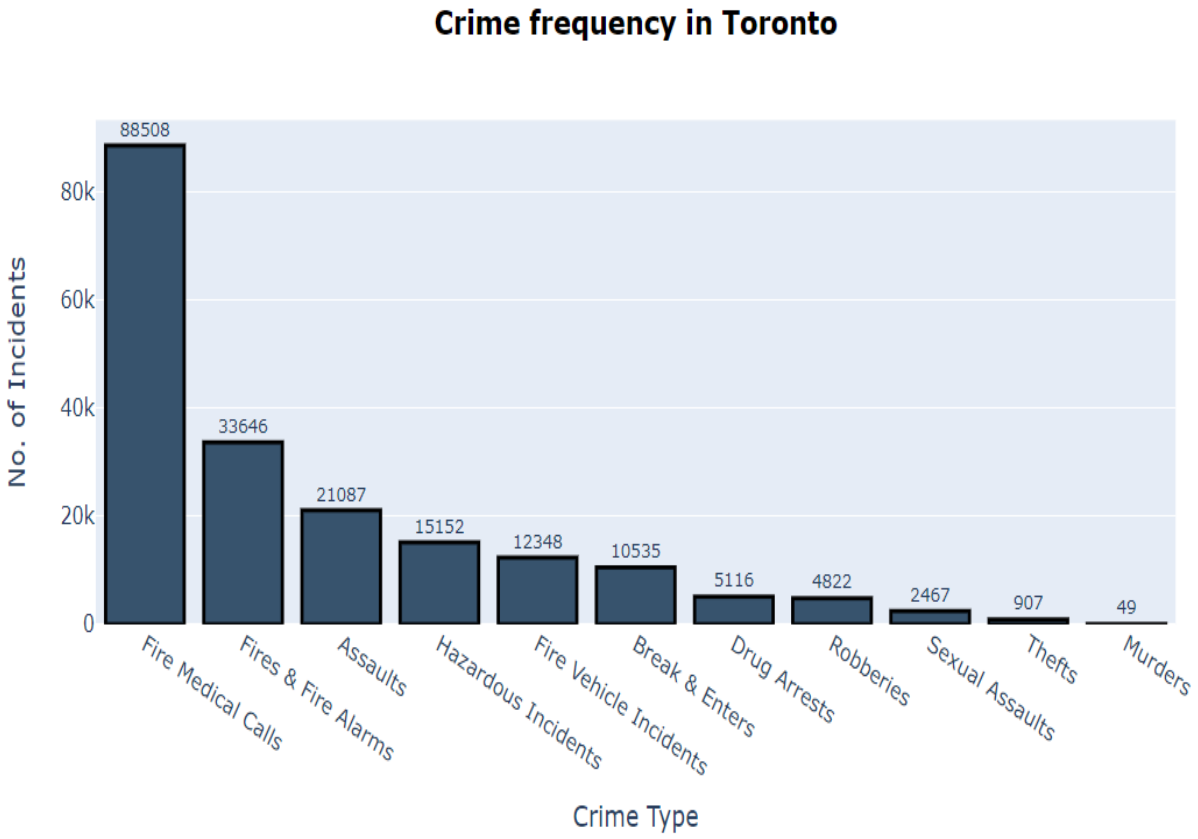

Figure 4: Bar chart for the Crime related cases arranged by descending order

Considering the business problem, as fire-related cases are not considered a huge aspect in deciding a house or a neighbourhood, these variables will be removed from the Dataframe.

The neighbourhoods are also plotted on a Folium map to visualize and validate the geocode details fetched. To increase the readability of the map and validation of the location details, clusters are created based on the geocode details using the MarkerCluster() method of Folium [8]. Please refer to figure 5 for a screenshot of the Toronto's neighbourhoods plotted on a Folium map.

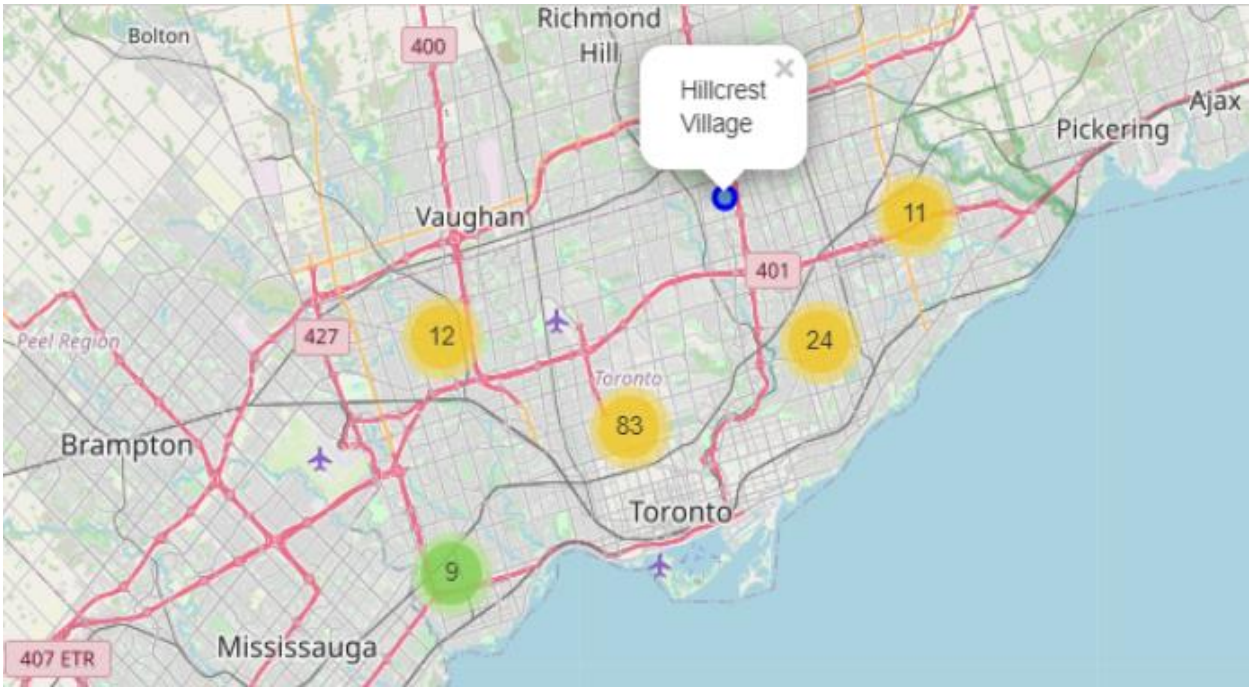

Figure 5: A screenshot of Toronto's neighbourhoods clustered based on geolocation details

Please note that the shape of the Dataframe created based on the Open Toronto website [4-6] was found as 140 where each row corresponds to a unique neighbourhood. To validate that the geolocation details present in the Dataframe are correct, a simple math validation was done i.e. add the total number of points plotted on the map, if the points plotted in the Toronto area are similar to the shape of the Dataframe then no outlier is present in the Dataset.

The shape of the Dataframe = SUM(Number of points plotted in Toronto area. Please refer figure 5.)

140 = SUM(12+9+83+24+1+11)

140 = 140

Later, the distribution of home prices in the dataset is analyzed. Please refer to figure 6 for a screenshot of the histogram plotted showing the home price range and the number of neighbourhoods in that range.

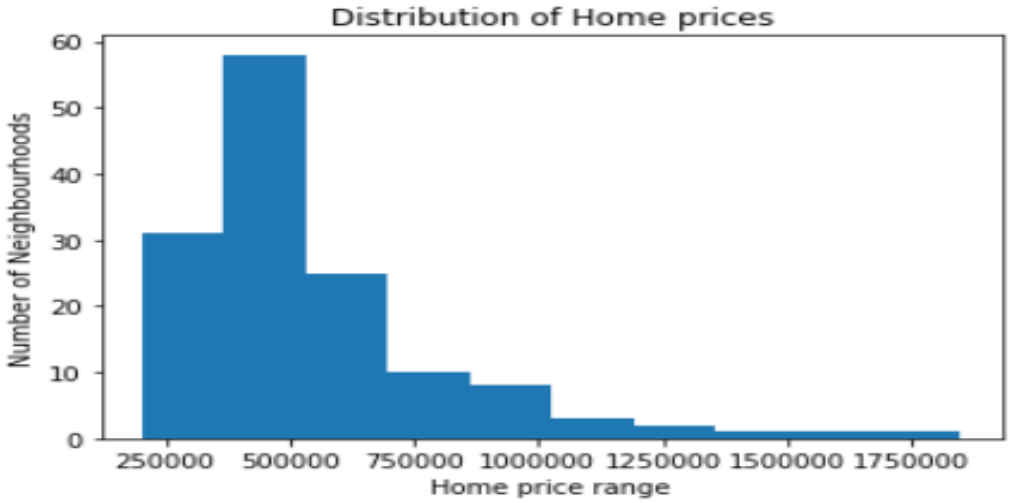

Figure 6: A screenshot of the Home price range and the number of neighbourhoods in that range

Based on figure 6, it is clear that most of the neighbourhoods have a price range of 250K to 700K units. It is also noticeable that only less than 5 neighbourhoods have a home price range of 1.7M units. Interestingly, as shown in figure 6, the home price data is right-skewed. The understanding from figure 6, further leads us to visualize and create quick histograms of all the useful numeric variables from the dataset. Please refer to figure 7 for a screenshot of the histograms of these variables.

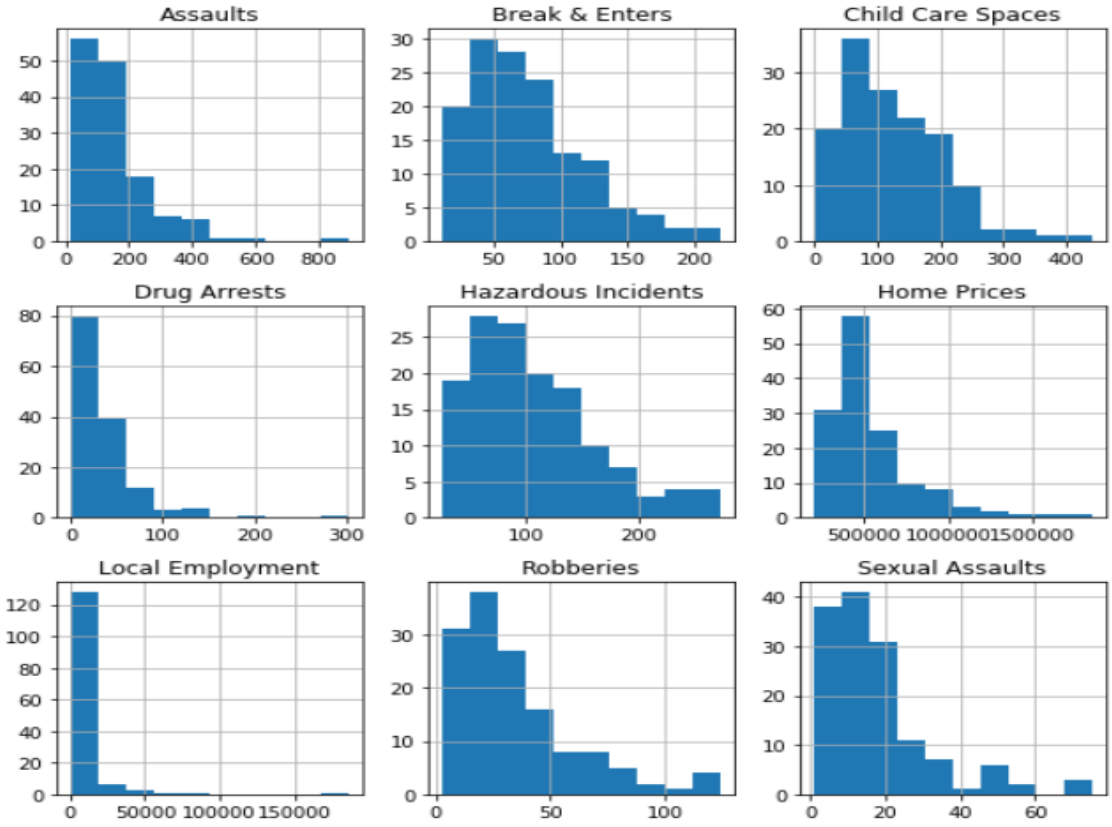

Figure 7: Histograms of useful numeric variables in the dataset

The histograms shown in figure 7 have a tail on the right side which conveys that most of these numeric variables are right-skewed.

4. Data Preparation

In this section, we will find any anomalies in the dataset and fix them to get a Dataset ready to be fed to the Machine Learning model. Based on figure 7, we got a hint of skewness in the Dataset. Next, we used statistical methods to verify the skewness in the variables. Please refer to figure 8, for a screenshot of the unbiased skewness in the variables.

```
Local Employment       6.556823
Vehicle Thefts         4.499711
Thefts                 3.469959
Drug Arrests           3.179012
Assaults               2.459699
Home Prices            1.995973
Sexual Assaults        1.774144
Robberies              1.484055
Break & Enters         1.060041
Hazardous Incidents    0.922938
Child Care Spaces      0.886435
dtype: float64
```

Figure 8: Screenshot of skewness in each relevant numeric variable in the dataset

A lot of variables in the dataset contains a mixture of positive and negative values. To reduce the skewness, we transformed the data using the “Yeo-johnson” method of the Scikit-learn Python library. After the transformation, most of the variables started exhibiting the normal distribution as shown in figure 9.

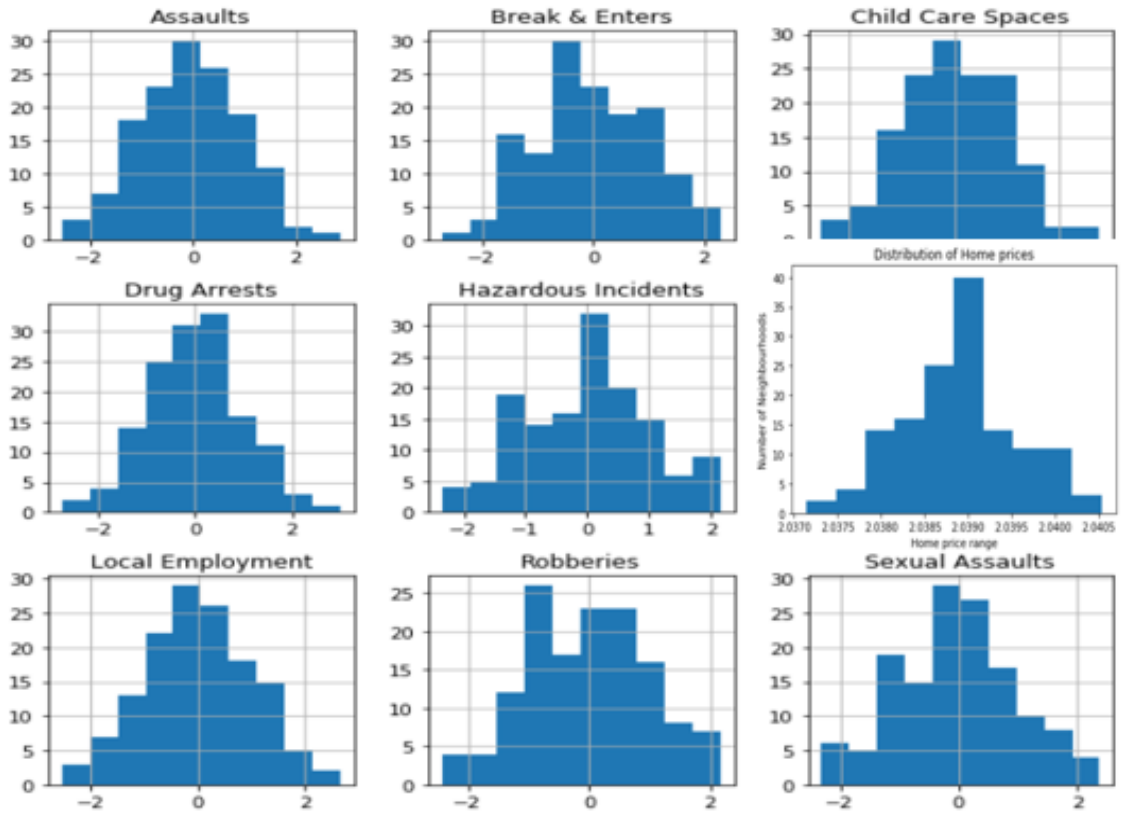

Figure 9: A screenshot of data after Yeo-Johnson transformation

For a few variables like Home prices, the Yeo-Johnson method did not yield the desired result. Hence, the Box-cox method is applied to transform the data into a normal distribution. The transformed Home price graph can be referred to in figure 9.

The venue details fetched using FourSquare API was one hot encoded so that they can be processed by the clustering algorithm. After the one-hot encoding, data was arranged based on the mean values for each neighbourhood. To address the missing venue data for one of the neighbourhood as shown in figure 3, the missing values were replaced by the mean of each column i.e. the category of venue. Post this processing, no variables had any missing values as shown in figure 10.

```
toronto_grouped[toronto_grouped.isna().any(axis=1)]
```

Neighbourhood	Neighbourhood_Id	LONGITUDE	LATITUDE	Home Prices	Child Care Spaces	Local Employment	Assaults	Break & Enters	Drug Arrests	...	Video Store	Vietnamese Restaurant	Warehouse Store	Wir Bi

0 rows × 297 columns

Figure 10: A screenshot showing no missing values after processing the data with the mean values

To make the data more presentable after clustering, venue data was arranged based on the top 10 famous venues near each neighbourhood. This arrangement is based on computing the mean of venues for each neighbourhood and then sorting the values in descending order. The highest mean values are the most famous venues in that location. Figure 11 shows a few neighbourhoods and their top 10 common venues.

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Wychwood	Convenience Store	Event Space	Park	Farmers Market	Electronics Store	Falafel Restaurant	Farm	Fast Food Restaurant	Field	Filipino Restaurant
Yonge-Eglinton	Coffee Shop	Fast Food Restaurant	Gym	Restaurant	Movie Theater	Supermarket	Breakfast Spot	Buffet	Skating Rink	Shopping Mall
Yonge-St.Clair	Coffee Shop	Italian Restaurant	Restaurant	Sushi Restaurant	Bagel Shop	Pizza Place	Thai Restaurant	Café	Grocery Store	Gym
York University Heights	Bar	Japanese Restaurant	Miscellaneous Shop	Bank	Fast Food Restaurant	Caribbean Restaurant	Massage Studio	Coffee Shop	Farmers Market	Field
Yorkdale-Glen Park	Restaurant	Fast Food Restaurant	Furniture / Home Store	Fried Chicken Joint	Rental Car Location	Café	Seafood Restaurant	Bowling Alley	Bookstore	Sandwich Place

Figure 11: Top 10 most common venues of neighbourhoods

5. Modeling and Evaluation

K means is used to cluster the data based on home price, child care spaces, local employment, primary crime categories and venue details. Three clusters were created based on the analysis done using the Elbow method. Due to a high number of features fed to Kmeans, the Elbow method resulted in a little smoother curve as shown in figure 12.

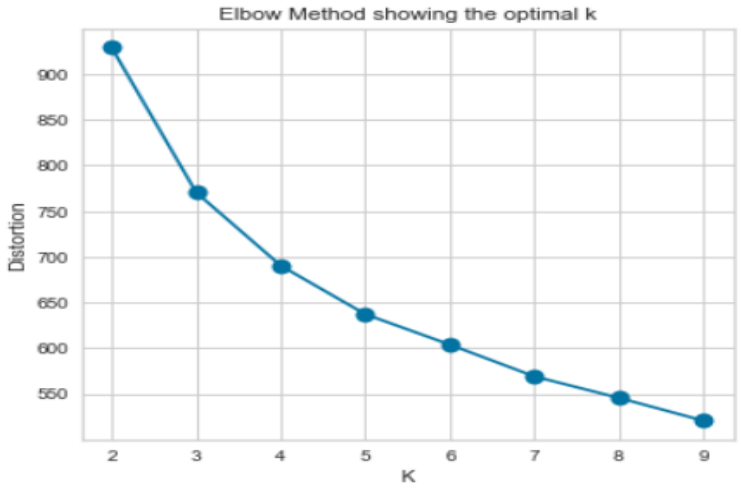

Figure 12: Elbow method to find an optimum value of K for K-means

As shown in figure 12, three or four clusters both could be good in this analysis. But to get more confidence in the selection of K parameter i.e. the number of clusters for K means, the Silhouette score was computed for each K and values were plotted on a graph as shown in figure 13.

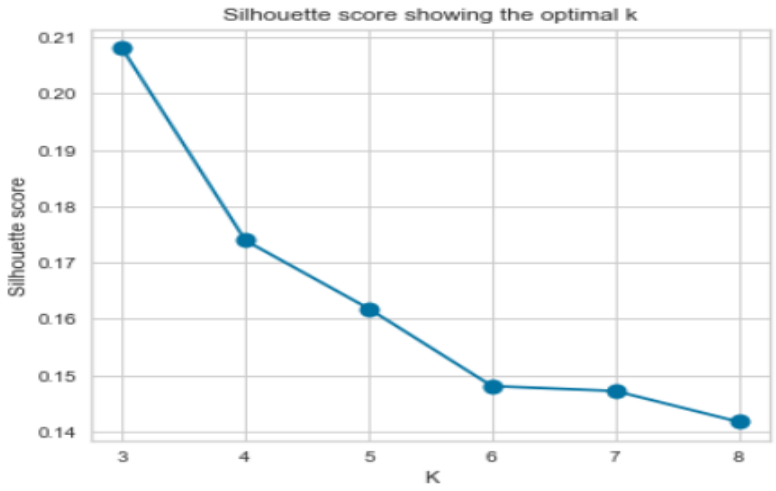

Figure 13: Silhouette score for different K values of K-means

From figure 13, it is visible that for cluster 3, the silhouette score is the highest. Hence, three clusters were created on the data using the K-means algorithm based on the K-means++ initialization method to speed up the convergence. The three clusters created are shown in figure 14. They are represented with three different colours and a click on each point displays the attributes of the neighbourhoods.

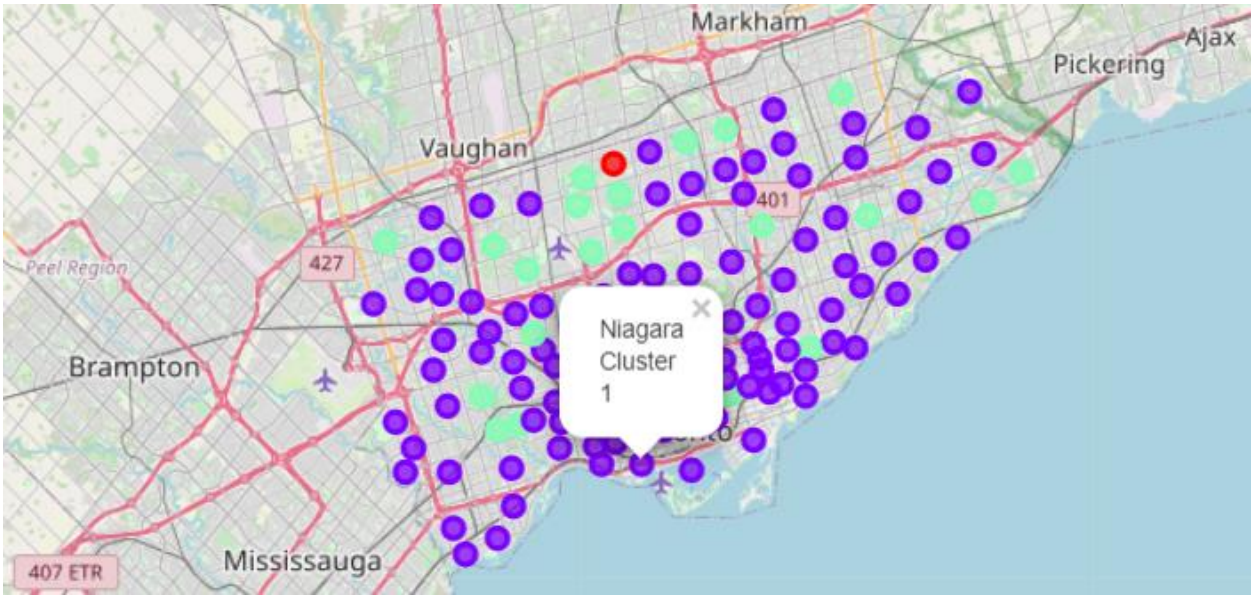

Figure 14: Clusters created based on K-means

Please note that the clusters are not created based on geolocation details. Hence, a neighbourhood in the extreme west and another neighbourhood in the extreme east maybe both in the same cluster.

6. Insights

Cluster/ Segments	Cluster/Segment Details	Number of Neighbourhoods
Cluster 1	**House Price Range:** ~550K **Childcare Spaces:** Low to Medium **Crime Rate:** High. Common cases of Assaults, Break & Enter, and Hazardous Incidents. **Nearby Venues:** Less restaurants, photography studio and Farms, Farmers market **Local Employment:** Medium	1
Cluster 2	**House Price Range:** 250K to 600K **Childcare Spaces:** Medium to High. Some neighbourhoods don't have childcare spaces **Crime Rate:** Medium to High. Common cases of Assaults, Break & Enter. **Nearby Venues:** Restaunrants, coffee shops, pubs and bars **Local Employment:** Medium to High	114
Cluster 3	**House Price Range:** 250K to 800K **Childcare Spaces:** Medium to High (Each nighbourhood has a childcare space) **Crime Rate:** Low to High. Common cases of Assaults, Hazardous Incidents, and Thefts. **Nearby Venues:** Playgrounds or parks, restaurants, pubs/bars, gym and clothing store **Local Employment:** Medium to High	25

Figure 15: Cluster details

7. Conclusion

Before moving to a new place or a city, people do their research on nearby venues, crime rates, local employment rates and childcare spaces. In this project, Toronto's neighbourhoods are clustered based on various attributes.

With the help of the insights provided in section 6, a user can choose the cluster based on their preference. As future work, a dashboard can be created to show the neighbourhoods in each cluster and display visualizations for each feature.

References

[1] "10 Important Features to Consider When Buying a House | HOMEiA.com", *HOMEiA*, 2020. [Online]. Available: https://homeia.com/10-important-features-to-consider-when-buying-a-house/. [Accessed: 30- Apr- 2020].

[2] "The 5 Factors of a 'Good' Location", *Investopedia*, 2020. [Online]. Available: https://www.investopedia.com/financial-edge/0410/the-5-factors-of-a-good-location.aspx. [Accessed: 30- Apr- 2020].

[3] V. Monga, "Silicon Valley Looks North as Tech Giants Expand in Toronto", *WSJ*, 2020. [Online]. Available: https://www.wsj.com/articles/silicon-valley-looks-north-as-tech-giants-expand-in-toronto-11566054001. [Accessed: 30- Apr- 2020].

[4] "Open Data Dataset", *Open.toronto.ca*, 2020. [Online]. Available: https://open.toronto.ca/dataset/neighbourhoods/. [Accessed: 14- May- 2020].

[5] "Open Data Dataset", *Open.toronto.ca*, 2020. [Online]. Available: https://open.toronto.ca/dataset/wellbeing-toronto-economics/. [Accessed: 14- May- 2020].

[6] "Open Data Dataset", *Open.toronto.ca*, 2020. [Online]. Available: https://open.toronto.ca/dataset/wellbeing-toronto-safety/. [Accessed: 14- May- 2020].

[7] "Endpoints | Places API", *Developer.foursquare.com*, 2020. [Online]. Available: https://developer.foursquare.com/docs/places-api/endpoints/. [Accessed: 30- Apr- 2020].

[8] "folium — Folium 0.11.0 documentation", *Python-visualization.github.io*, 2020. [Online]. Available: https://python-visualization.github.io/folium/modules.html. [Accessed: 14- May- 2020].