CLUSTERING TORONTO NEIGHBORHOODS BASED ON SAFETY, HOUSING PRICES, SCHOOLS & ENTERTAINMENT

by

Maryam Momodu Bassey

INTRODUCTION

Canada opened its borders to skilled workers from all over the world. People with families are among this group of emigrating skilled workers.

Finding the right neighborhood to reside is an issue for people new to a city. There are several factors to consider and to pick a few essentials are; housing, safety, schools for children and places of relaxation and entertainment. There are also less general issues like proximity to work, job opportunities.

This project aims to use data to help guide Toronto immigrants with choosing neighborhoods to reside in with regards to safety, housing prices, availability of schools, entertainment and relaxation activities. The project will categorize similar neighborhoods which will simplify the process of choosing a neighborhood to reside in.

DATA

Requirements

- Toronto crime incidents data including incident type and location data
- Data on average housing prices with regards to neighborhoods in Toronto
- Location data for Schools, Arts and Entertainment centers in Toronto neighborhoods
- Toronto neighborhood latitude and longitude data

Sources

- Average Housing Sale prices
  Sourced from- https://www.zolo.ca/toronto-real-estate/neighbourhoods, 29/04/19 6.17pm
  The data was scraped from the website and it contains the average selling price of houses for neighborhoods in Toronto for 28days prior to the day it was scraped.

- Toronto crime data
  Sourced from- http://data.torontopolice.on.ca/datasets/98f7dde610b54b9081dfca80be453ac9_0, 28/04/19
  The downloaded dataset contains information on crimes committed in Toronto and their location data from 2014 to 2018. The 2018 incidents data is of interest for this project.

- Data for schools and entertainment centers in Toronto was sourced from the Four Square Location data.

- Neighborhood CDN- https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto

- Toronto neighborhood Geojson data
  https://github.com/jasonicarter/toronto-geojson/blob/master/toronto_topo.json
  https://portal0.cf.opendata.inter.sandbox-toronto.ca/dataset/neighbourhoods/

Cleaning

Data downloaded and scraped were individually cleaned.

- For the average housing dataset, redundant values and unnecessary data were dropped. The characters for dollar sign($), thousand(k) and million(M) were also removed to allow me work with the values as integers. The values in millions were converted to thousands and the CDN for neighborhoods manually added in a text editor.

- The Toronto crime dataset downloaded contained only 2018 data. Columns containing the Neighborhood ID, Neighborhood name, Latitude, Longitude and Major Crime Indicator(MCI) were selected to create a new dataset. The new dataset was organized to show the total crime committed in each neighborhood and the remaining columns retained.

- Four Square Location data was utilized to get data on venues within a 500 meters radius of each neighborhood latitude and longitude. The resulting dataset of 292 venues was the filtered to get only data in two venue categories; Arts and entertainment and Schools.
  - Art and Entertainment category was made up of 47 venue types;
    Amphitheater, 'Aquarium', 'Arcade', 'Art Gallery','Bowling Alley', 'Casino', 'Circus', 'Comedy Club', 'Concert Hall','Country Dance Club', 'Disc Golf', 'Exhibit', 'General Entertainment', 'Go Kart Track', 'Historic Site', 'Karaoke Box', 'Laser Tag', 'Memorial Site', 'Mini Golf', 'Drive-in Theater', 'Indie Movie Theater', 'Multiplex', 'Art Museum', 'History Museum', 'Planetarium', 'Science Museum', 'Jazz Club', 'Piano Bar', 'Rock Club', 'Dance Studio', 'Indie Theater', 'Opera House', 'Theater', 'Pool Hall', 'Racecourse', 'Racetrack', 'Roller Rink', 'Salsa Club', 'Samba School', 'Stadium', 'Theme Park', 'Water Park', 'Zoo'

  - Schools category was made up of 7 venue types;
    'Elementary School', 'High School', 'Middle School', 'Preschool', 'Private School', 'Driving School', 'Nursery School'

  The dataset was further modified to contain 3 features, Neighborhood, Arts and Entertainment(sum of all venues under this category) and Schools(sum of all venues under this category), with 139 rows and 3 columns. Two Neighborhoods were missing, they were replaced manually and assigned values zero.

The final dataset contained Neighborhood names, Latitude and longitude data, 3 features; Arts and Entertainment, Average Housing price in thousands and Crime rate. The Schools feature was dropped as it did not contain much information.

METHODOLOGY

As the aim of the project which is to discover similarity of Toronto neighborhoods with regards to specified criteria required the use of clustering algorithms to generate clusters.
There are several algorithms available for clustering with differing strengths and weakness. Three were employed; K means, DB-SCAN and agglomerative algorithms.

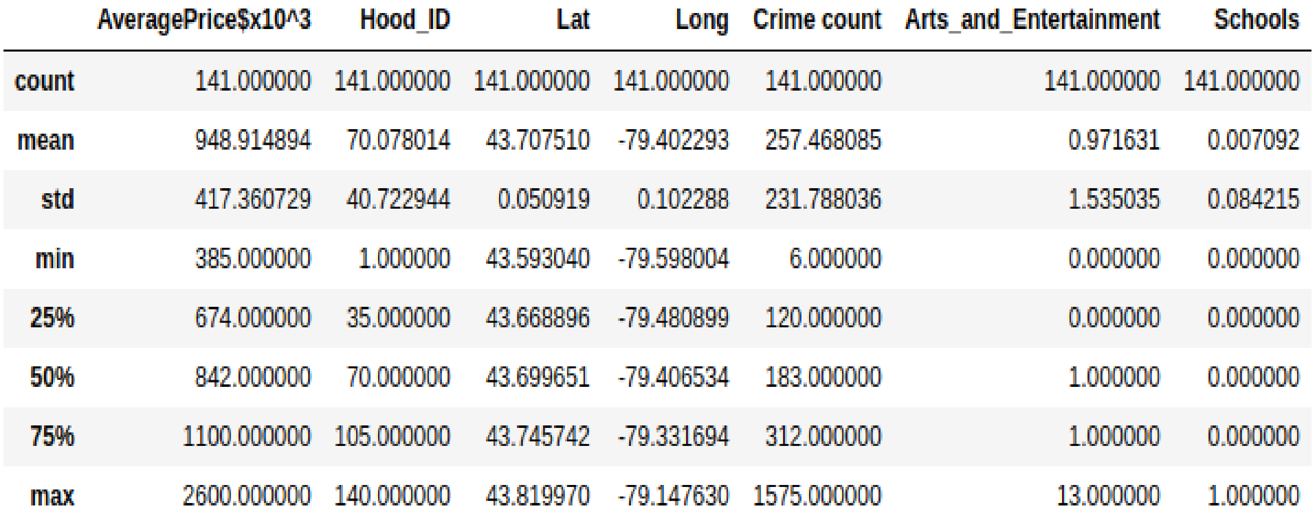
- K-means was employed to partition the neighborhoods into groups with similar characteristics. It is known that when applied to tasks with arbitrary shape clusters, or clusters within cluster, the traditional techniques might be unable to achieve good results. This was taken into consideration, as *Yellowbrick* visualizer was employed to plot a graph of different iterations of K, with the dataset, and the elbow of the line plot was visually estimated. This was accepted as the optimal K to run the model. The resulting neighborhood clusters were visualized using *folium*.

- The Density Based Spatial Clustering of Application with Noise (DB-SCAN) has the ability to locate and separate high density regions from low density regions; and it doesn't require the number of clusters to be specified. While its drawbacks is minimized in this project as the data used is not highly dimensional. The minimum number of points and epsilon were selected using the defined functions and information from- https://stackoverflow.com/questions/43160240/how-to-plot-a-k-distance-graph-in-python. The resulting neighborhood clusters and outliers were visualized using *folium*.

- Agglomerative hierarchical clustering employs a bottom up approach to clustering.. It work well with small datasets; which is the case in this project. Selection of the appropriate method for computing distances was done using a code from- https://stackoverflow.com/questions/21638130/tutorial-for-scipy-cluster-hierarchy. The proposed number of clusters from the method used was then compared to what was gotten by visually ascertaining the maximum distance in the plotted dendogram. The clusters of neighborhoods was visualized using folium.

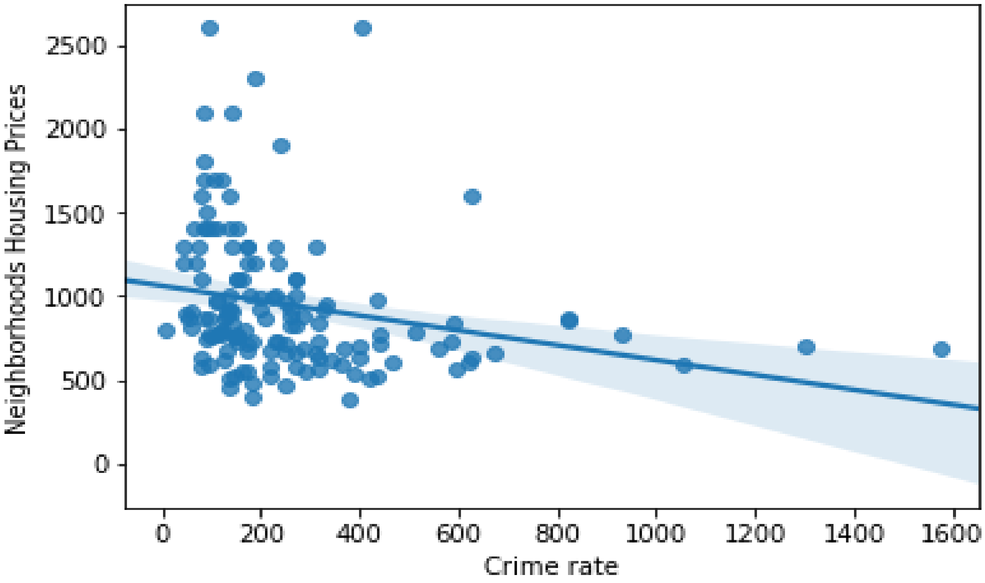
RESULT

Descriptive analysis of data from each feature was done to evaluate the spread of the data.



	AveragePrice$x10^3	Hood_ID	Lat	Long	Crime count	Arts_and_Entertainment	Schools
count	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000
mean	948.914894	70.078014	43.707510	-79.402293	257.468085	0.971631	0.007092
std	417.360729	40.722944	0.050919	0.102288	231.788036	1.535035	0.084215
min	385.000000	1.000000	43.593040	-79.598004	6.000000	0.000000	0.000000
25%	674.000000	35.000000	43.668896	-79.480899	120.000000	0.000000	0.000000
50%	842.000000	70.000000	43.699651	-79.406534	183.000000	1.000000	0.000000
75%	1100.000000	105.000000	43.745742	-79.331694	312.000000	1.000000	0.000000
max	2600.000000	140.000000	43.819970	-79.147630	1575.000000	13.000000	1.000000

Table 1 Descriptive analysis of the dataset



Scatter plot and regression line of Crime rate vs Housing Price.

The plot above shows that the relationship between Crime rate and Housing price is non-linear.

Visualization of data on average housing price, Crime rate and Art and entertainment opportunities was done using the box plot. This was essential in creating the grading -using the percentiles- for each feature and assessing outliers.

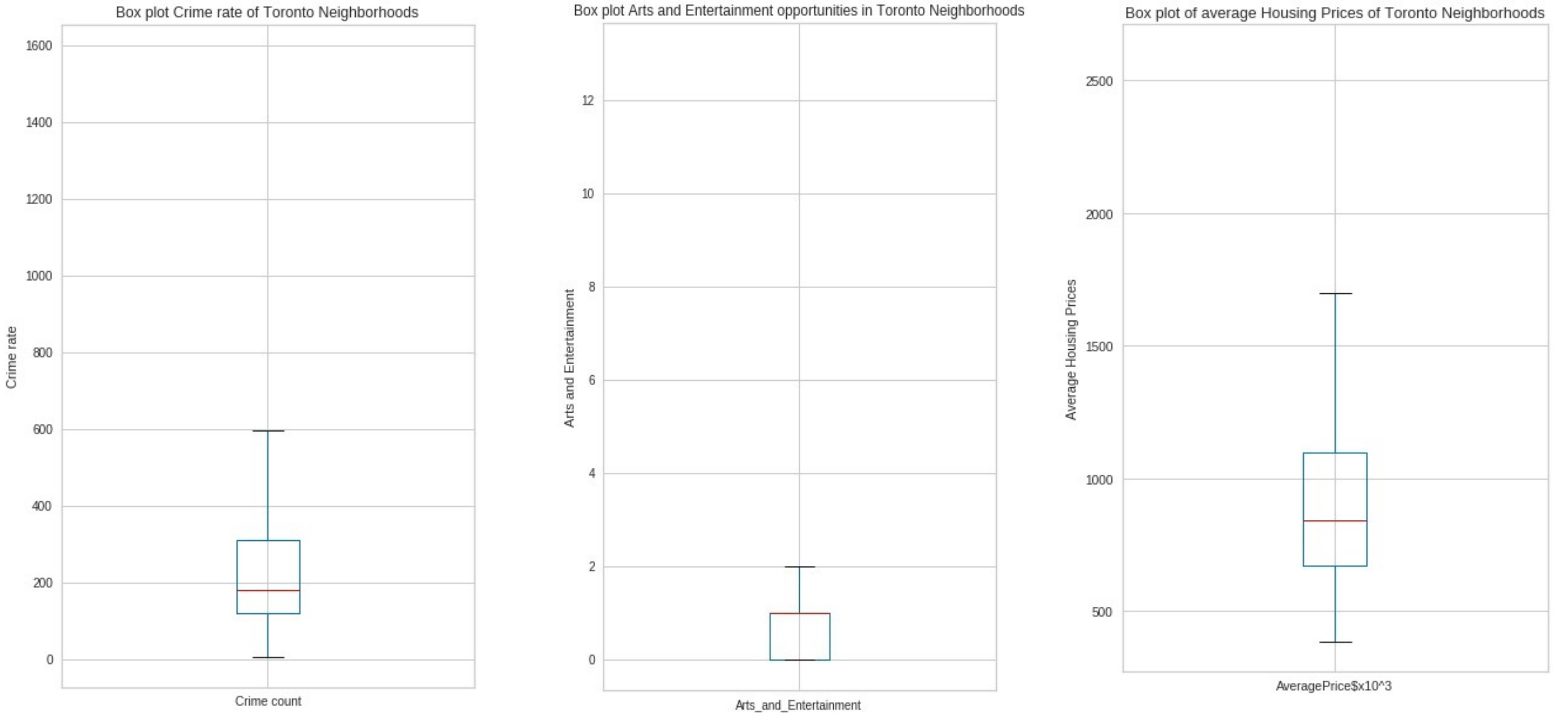


Fig 1   Box plots

After examining the distribution of the data in each feature as shown in the table and box plots above, I came up with the grading shown in the below. This grading was necessary in labeling the clusters generated.

GRADING	PERCENTILES	HOUSING PRICE($ x10^3)	CRIME RATE
LOW	Below 25	385 - 674	6- 120
AVERAGE	25 to 50	675- 842	121- 183
ABOVE AVERAGE	50 to 75	843 - 1100	184- 312
HIGH	75 to maximum	1101- 1750	313- 600
VERY HIGH	outliers	1751- 2600	601-1575

Table 2 Housing prices and Crime rate grading * maximum is Q3 + 1.5*IQR

GRADING	ARTS AND ENTERTAINMENT
NO	0
FEW	1- 2
FAIR NUMBER	3- 5
LARGE NUMBER	6- 13

Table 3 Grading for Arts and Entertainment opportunities

K-MEANS

The *KElbow Visualizer* was used to visualize the elbow point shown in the figure below.

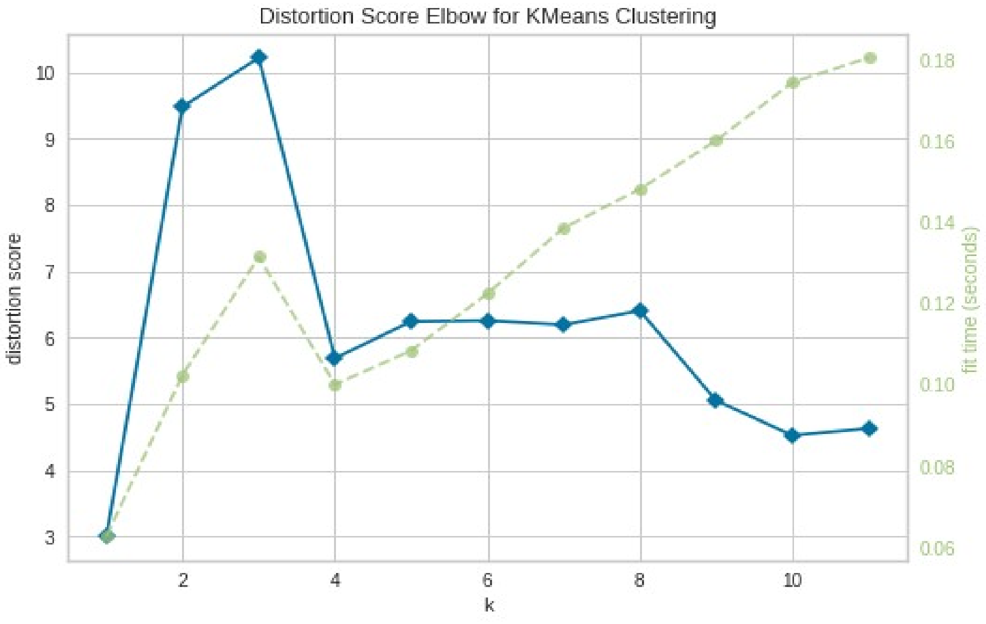
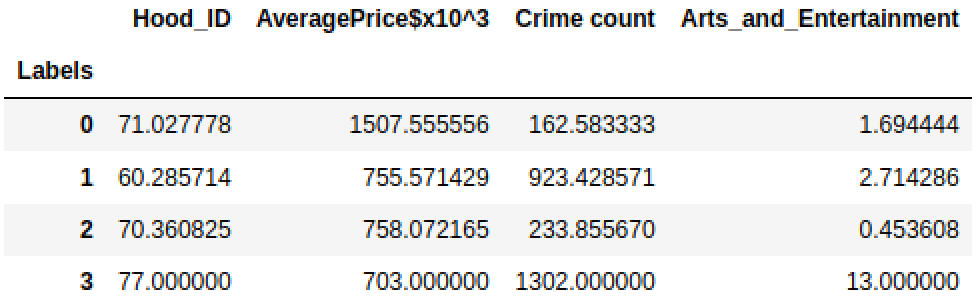


Fig 2

In Fig 3, the elbow point is at k=4, this was chosen as the optimal k value and 4 clusters were generated using the K-means algorithm. Their characteristics are shown in the table below.



Labels	Hood_ID	AveragePrice$x10^3	Crime count	Arts_and_Entertainment
0	71.027778	1507.555556	162.583333	1.694444
1	60.285714	755.571429	923.428571	2.714286
2	70.360825	758.072165	233.855670	0.453608
3	77.000000	703.000000	1302.000000	13.000000

Table 4 K-means clusters

A profile was created for each cluster, considering the characteristics of its features. The 4 clusters are:

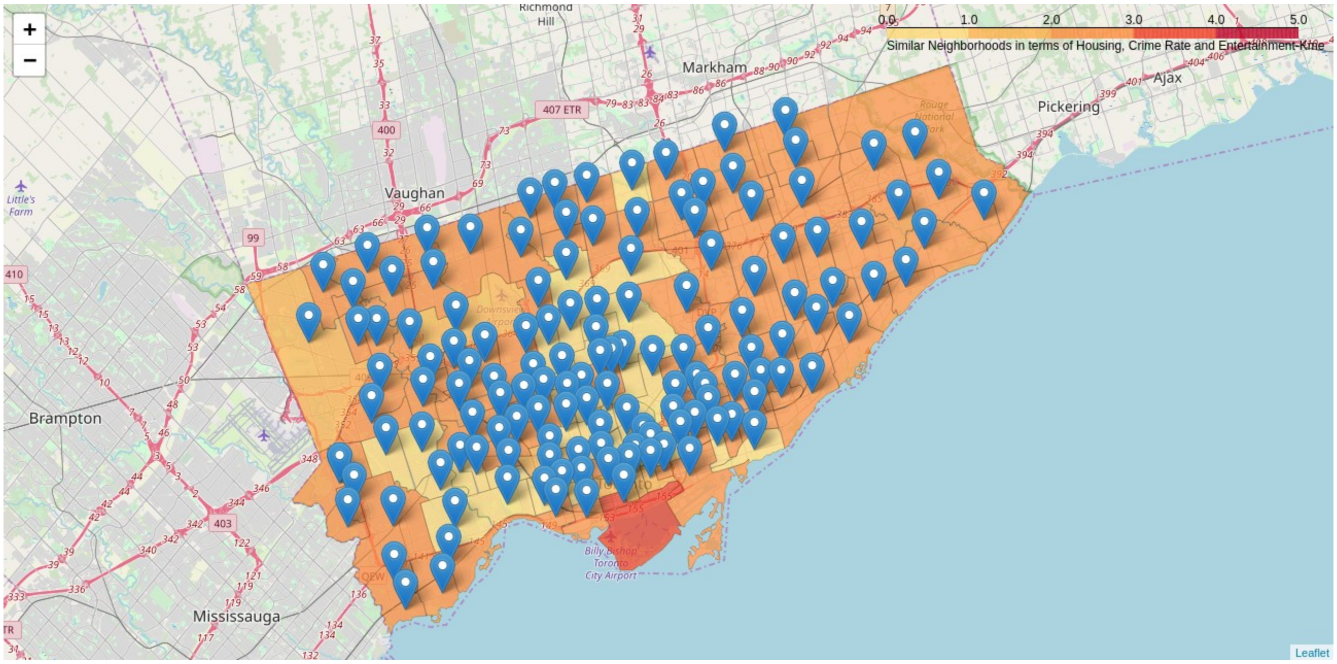
LABEL 0- High Housing Price, Average Crime Rate, Few Entertainment Opportunities, **35 neighborhoods**

LABEL 1- Average Housing Price, Very High Crime Rate, Fair Number of Entertainment Opportunities, **7 neighborhoods**

LABEL 2- Average Housing Price, Above Average Crime Rate, Very Few Entertainment Opportunities, **97 neighborhoods**

LABEL 3- Average Housing Price, Very High Crime Rate, Large Number of Entertainment Opportunities, **1 neighborhood**

A *Choropleth* map was generated -using *folium* -to visualize the neighborhood clusters according to their labels.


+
−
0.0
1.0
2.0
3.0
4.0
5.0
Similar Neighborhoods in terms of Housing, Crime Rate and Entertainment Kme
Richmond Hill
Markham
Pickering
Ajax
Vaughan
Brampton
Mississauga
Billy Bishop Toronto City Airport
Leaflet

DB-SCAN

The minimum number of points was taken as 4, 2* dimensionality of datasets. The optimal epsilon value was calculated using Knearest neighbor distance plots and histogram.

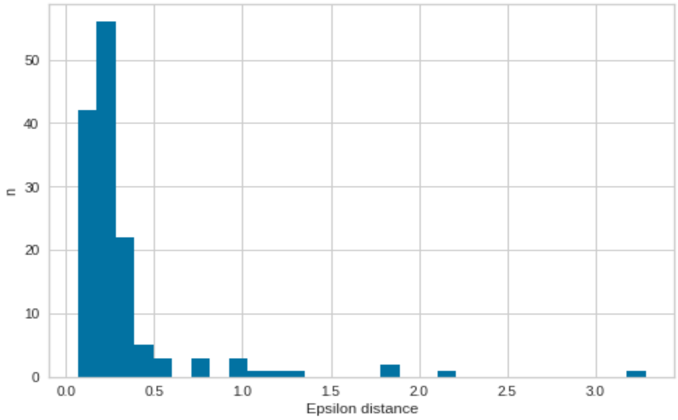


Fig 4 Epsilon value determination - Knn distance histogram with using a defined function

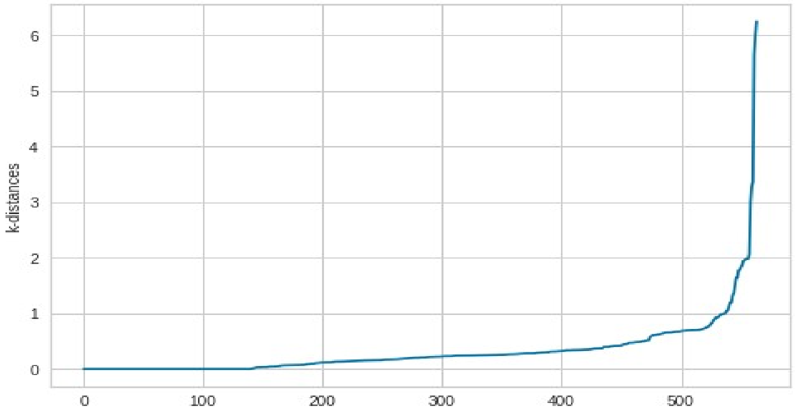
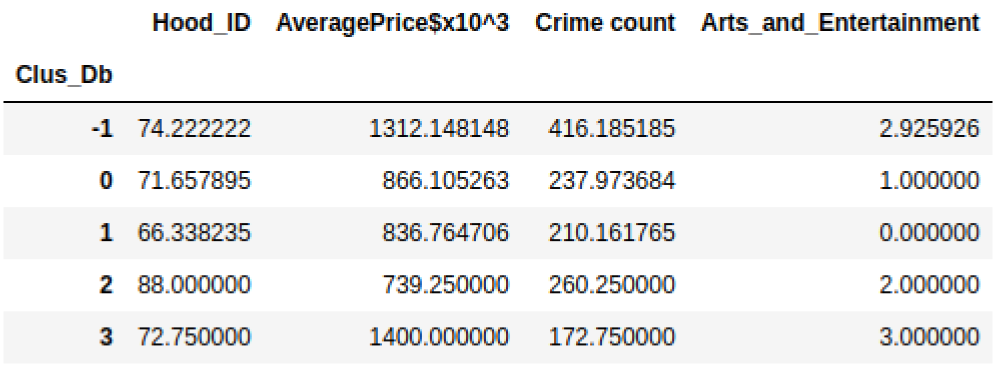


Fig 5 Epsilon value determination - Knn distance histogram with using a second defined function

The knee point in both figures above can be seen at 0.6, this was accepted as the epsilon value.



Clus_Db	Hood_ID	AveragePrice$x10^3	Crime count	Arts_and_Entertainment
-1	74.222222	1312.148148	416.185185	2.925926
0	71.657895	866.105263	237.973684	1.000000
1	66.338235	836.764706	210.161765	0.000000
2	88.000000	739.250000	260.250000	2.000000
3	72.750000	1400.000000	172.750000	3.000000

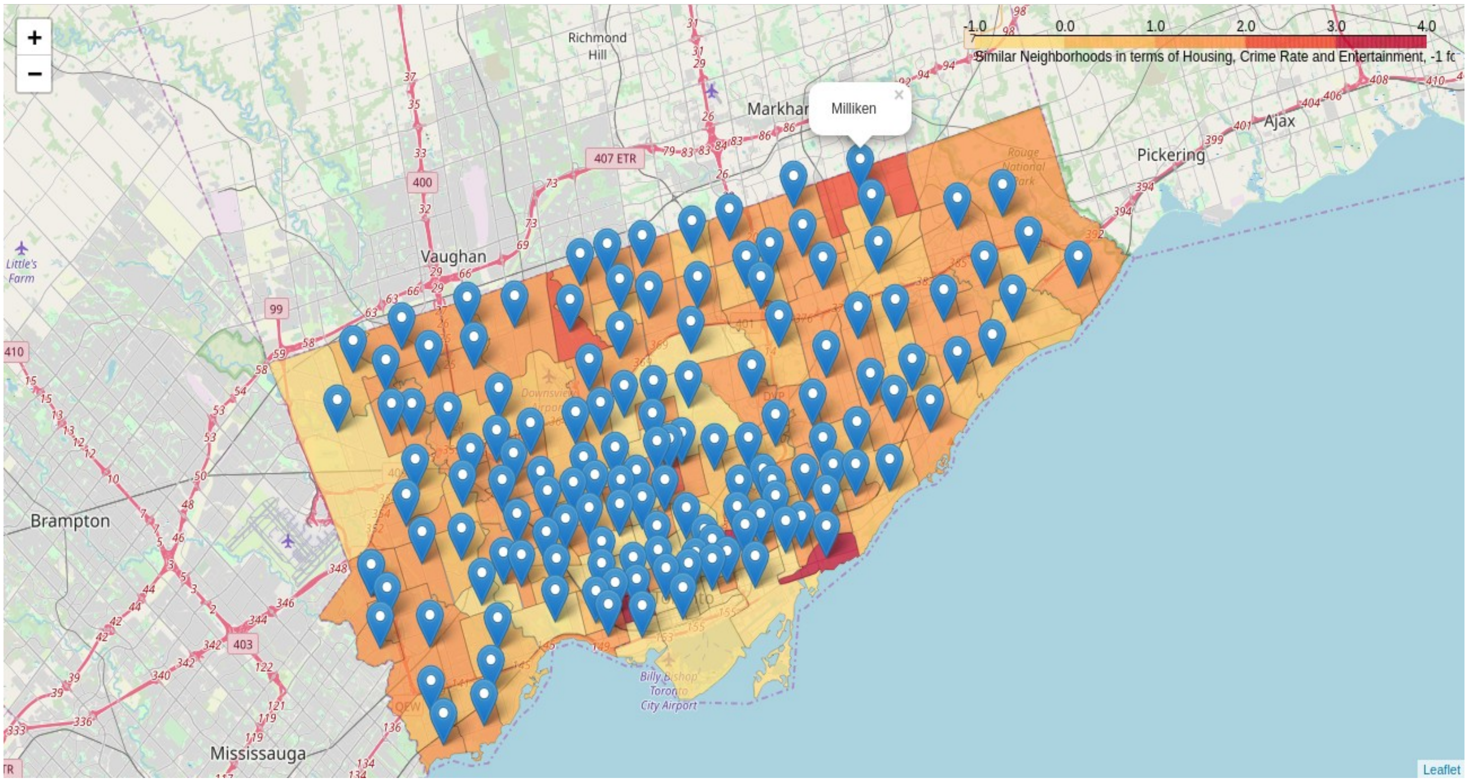
Table 5 DB-SCAN clusters

The DB-SCAN algorithm produced 5 cluster made up of 4 actual clusters and a outlier category. Out of 140 neighborhoods, 27 neighborhoods were located in the outlier category. This can be seen in the table above.

A profile was created for each cluster, considering the characteristics of its features. The 4 clusters are:

LABEL 0- Above Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities
LABEL 1- Average Housing Price, Above Average Crime Rate, No Entertainment Opportunities
LABEL 2- Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities
LABEL 3- High Housing Price, Average Crime Rate, Fair Amount of Entertainment Opportunities
LABEL -1 Outliers

A *Choropleth* map was generated -using f*olium* -to visualize the neighborhood clusters according to their labels.



AGGLOMERATIVE CLUSTERING

A comparison of complete and average linkage method was done showing the possible clusters of the data. This was done to get a tentative idea of the optimal method to employ with the algorithm. This is evident in the figure below.

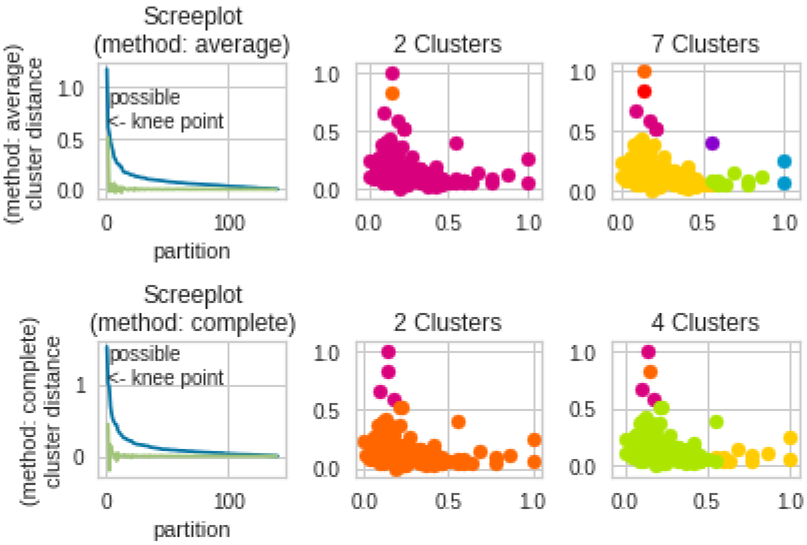


Fig 6 Comparing Hierarchical algorithm methods

Examining the clustering in the figure above, it can be seen that the **complete** method, with **4** clusters produces more distinct clusters.

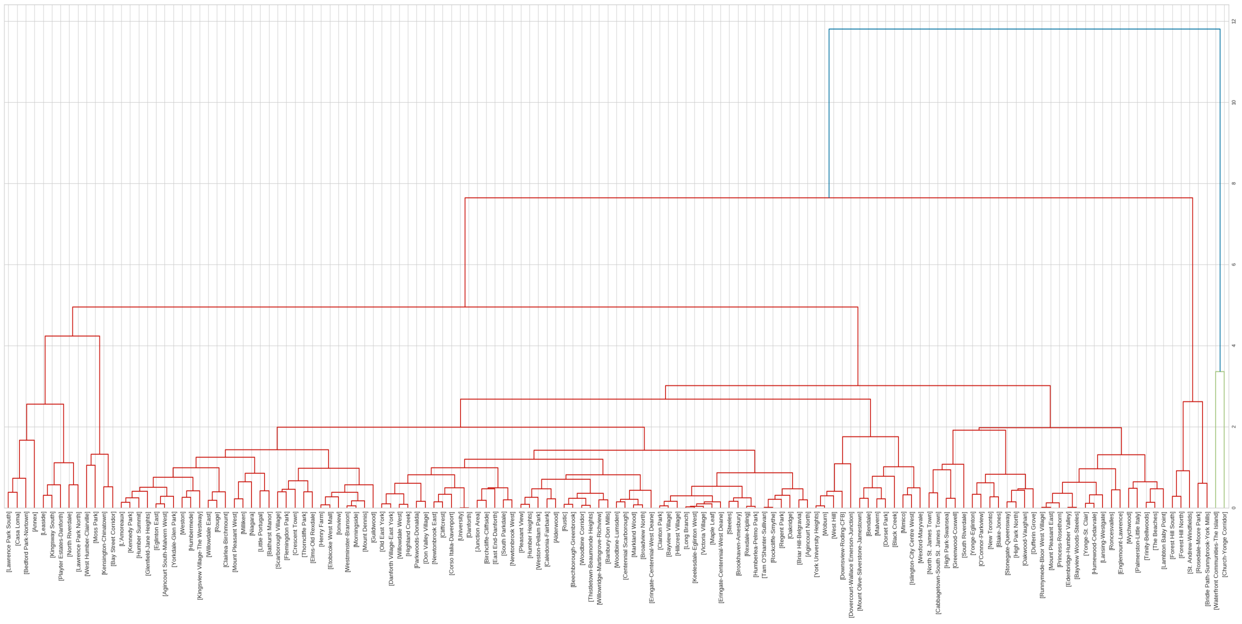
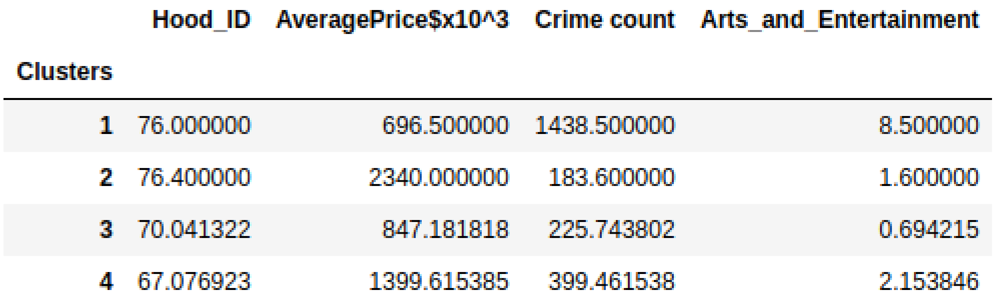


Fig 6 Dendrogram of Toronto Neighborhoods

Examining the dendrogram above, a jump is seen close to a distance of 4, if we place a straight line through approximately 3.75 distance tick, we will get 5 clusters, if the line is placed a little above the the 4 distance tick, we will have 4 clusters. Since the choice is between 5 or 4 clusters, a choice of 4 clusters was made as the **elbow method** also suggests 4 clusters.



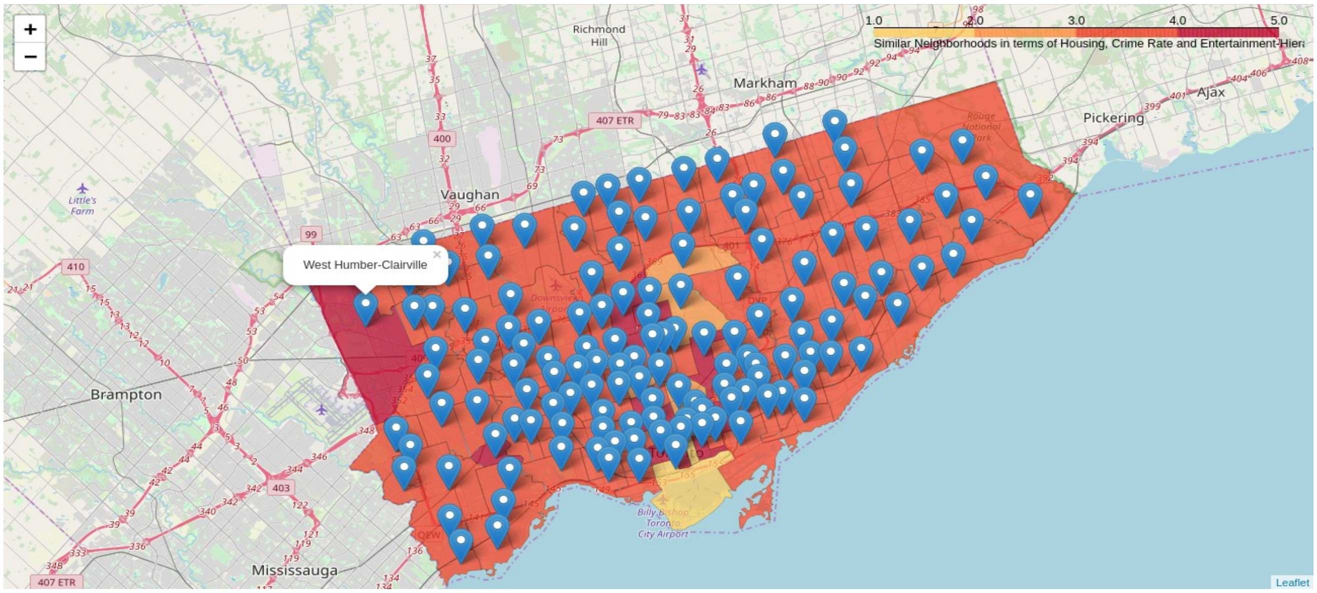
	Hood_ID	AveragePrice$x10^3	Crime count	Arts_and_Entertainment
Clusters				
1	76.000000	696.500000	1438.500000	8.500000
2	76.400000	2340.000000	183.600000	1.600000
3	70.041322	847.181818	225.743802	0.694215
4	67.076923	1399.615385	399.461538	2.153846

Table 6 Agglomerative algorithm clusters

A profile was created for each cluster, considering the characteristics of its features. The 4 clusters are:

LABEL 1- Average Housing Price, Very High Crime Rate, Large Number of Entertainment Opportunities, **2 neighborhoods**
LABEL 2- Very High Housing Price, Average Crime Rate, Few Entertainment Opportunities, **3 neighborhoods**
LABEL 3- Above Average Housing Price, Above Average Crime Rate, Very Few Entertainment Opportunities, **122 neighborhoods**
LABEL 4- High Housing Price, High Crime Rate, Fair Number of Entertainment Opportunities, **13 neighborhoods**

A *Choropleth* map was generated -using f*olium* -to visualize the neighborhood clusters according to their labels.



DISCUSSION

The decision to use 3 different clustering algorithms on the dataset was just to see what kind of clusters could be gotten, this has no bearing on the aim of the project.

The Db-Scan algorithm was the choice algorithm because it has the ability to detect outliers. One hundred and forty (140) Toronto neighborhoods were divided into 4 clusters, of these

27 neighborhoods were outliers (Very High Housing prices, Very High Crime rate or Very High Entertainment opportunities);

4 neighborhoods were in the High Housing Price, Average Crime Rate and Fair Amount of Entertainment Opportunities cluster;

Another 4 neighborhoods were in the Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities cluster;

68 of the neighborhoods were in the Average Housing Price, Above Average Crime Rate, No Entertainment Opportunities cluster;

The final cluster with the characteristics Above Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities had 37 neighborhoods.

In comparing the characteristics of the neighborhoods to that of the clusters they were assigned to, I discovered that some neighborhood features were somewhat different from that of their assigned cluster.

CONCLUSION

The purpose of this project was to place Toronto neighborhoods in categories depending on crime rate, average housing prices, number of schools and arts and entertainment opportunities; this was not totally achievable as data for schools from Foursquare Location data was sparse.

By employing the DB-SCAN I was able to get 4 labeled clusters. This will enable anyone migrating to Toronto to filter out neighborhood to reside in based on their preferences. This project provides a starting point in considering potential neighborhoods for residence and reduces the time and energy needed for such research.

APPENDIX

DB-SCAN NEIGHBORHOOD CLUSTERS

LABEL 0	LABEL 1	LABEL 2	LABEL 3	OUTLIERS
Agincourt North Agincourt South-Malvern West Alderwood Bayview Woods-Steeles Bendale Birchcliffe-Cliffside Briar Hill-Belgravia Caledonia-Fairbank Clairlea-Birchmount Danforth Village-East York Don Valley Village Downsview-Roding-CFB Old East York Edenbridge-Humber Valley Eglinton East Englemount-Lawrence Flemingdon Park Guildwood High Park North Highland Creek Humber Heights Junction Area Kingsview Village-The Westway Leaside Mimico Mount Olive-Silverstone-Jamestown North Riverdale Oakwood-Vaughan Parkwoods-Donalda Pleasant View	THE REMAINING 68 NEIGHB0RHOODS	Bathurst Manor Little Portugal Milliken Mount Pleasant West	Palmerston-Little Italy Playter Estates-Danforth The Beaches Trinity-Bellwoods	Annex Bay Street Corridor Bedford Park-Nortown Bridle Path-Sunnybrook-York Mills Cabbagetown-South St. James Town Casa Loma Church-Yonge Corridor Crescent Town Dovercourt-Wallace Emerson-Junction Dufferin Grove Forest Hill North Forest Hill South Greenwood-Coxwell High Park-Swansea Kensington-Chinatown Kingsway South Lambton Baby Point Lawrence Park South Moss Park Niagara North St. James Town Rosedale-Moore Park South Riverdale St. Andrew-Windfields Waterfront Communities-The Island

Roncesvalles Scarborough Village Stonegate-Queensway Thorncliffe Park West Hill Weston-Pellam Park Willowdale West Yorkdale-Glen Park				West Humber-Clairville Wychwood