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

by

Maryam Momodu Bassey

Outline

- Problem description
- Aim of the Project
- Data Sources
- Data Cleaning
- Analytical Approach
- Results

Problem Description

- 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.
- Some important considerations are; housing, safety, schools for children and places of relaxation and entertainment.

Aim of the Project

- This project aimed 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.

- Creating clusters of similar neighborhoods to simplify the process of choosing a neighborhood to reside in.

Data Sources

- Average Housing Sale prices from- https://www.zolo.ca/toronto-real-estate/neighbourhoods, 29/04/19 6.17pm

- 2018 Toronto crime data from- http://data.torontopolice.on.ca/datasets/98f7dde610b54b9081dfca80be453ac9_0, 28/04/19

- Data for schools and entertainment centers in Toronto neighborhoods 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://portal0.cf.opendata.inter.sandbox-toronto.ca/dataset/neighbourhoods/

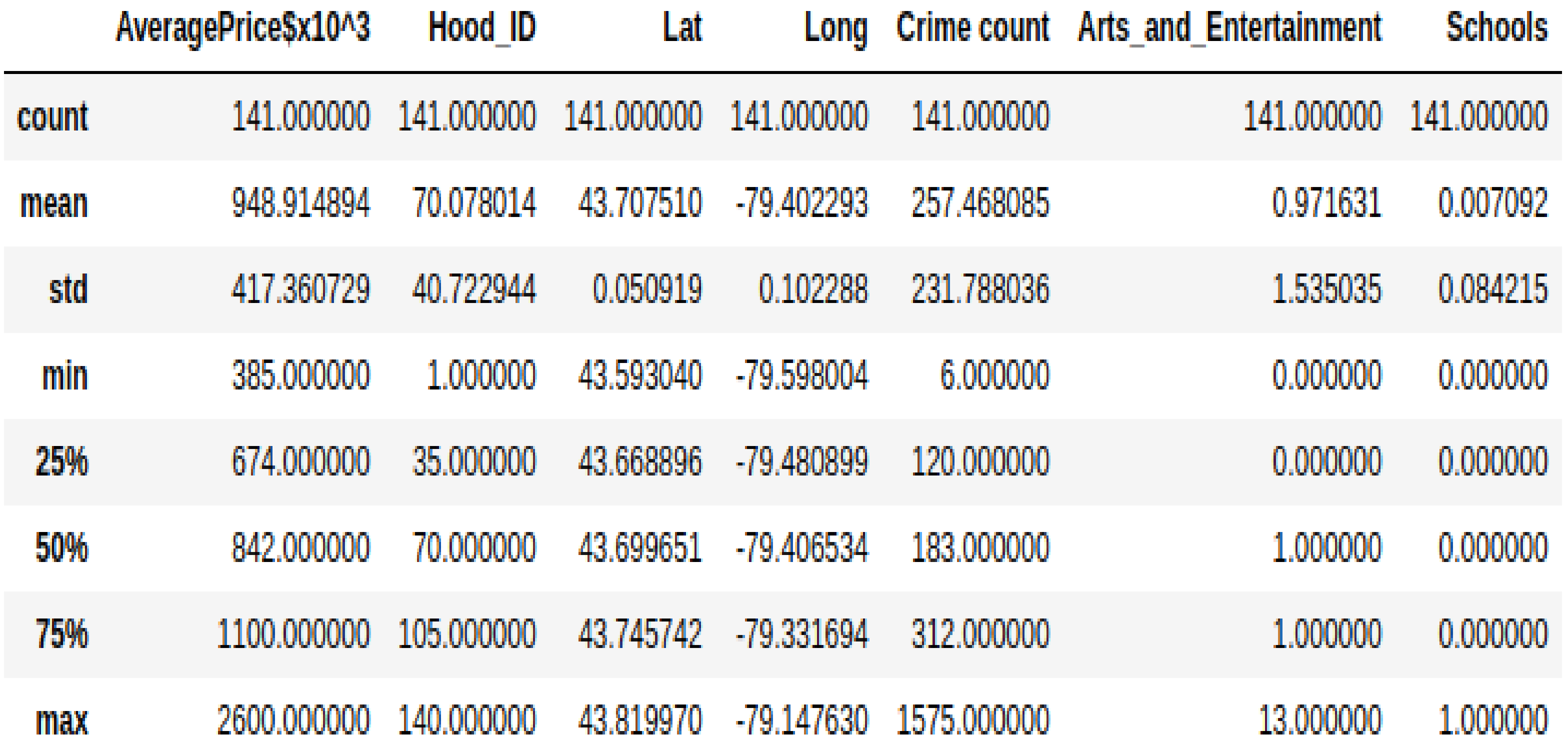
Data Cleaning

- Data downloaded and scraped were individually cleaned.
- Art and Entertainment category was made up of 47 venue types while the Schools category was made up of 7 venue types.
- The final dataset contained 140 Neighborhoods, their 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.

Analytical Approach

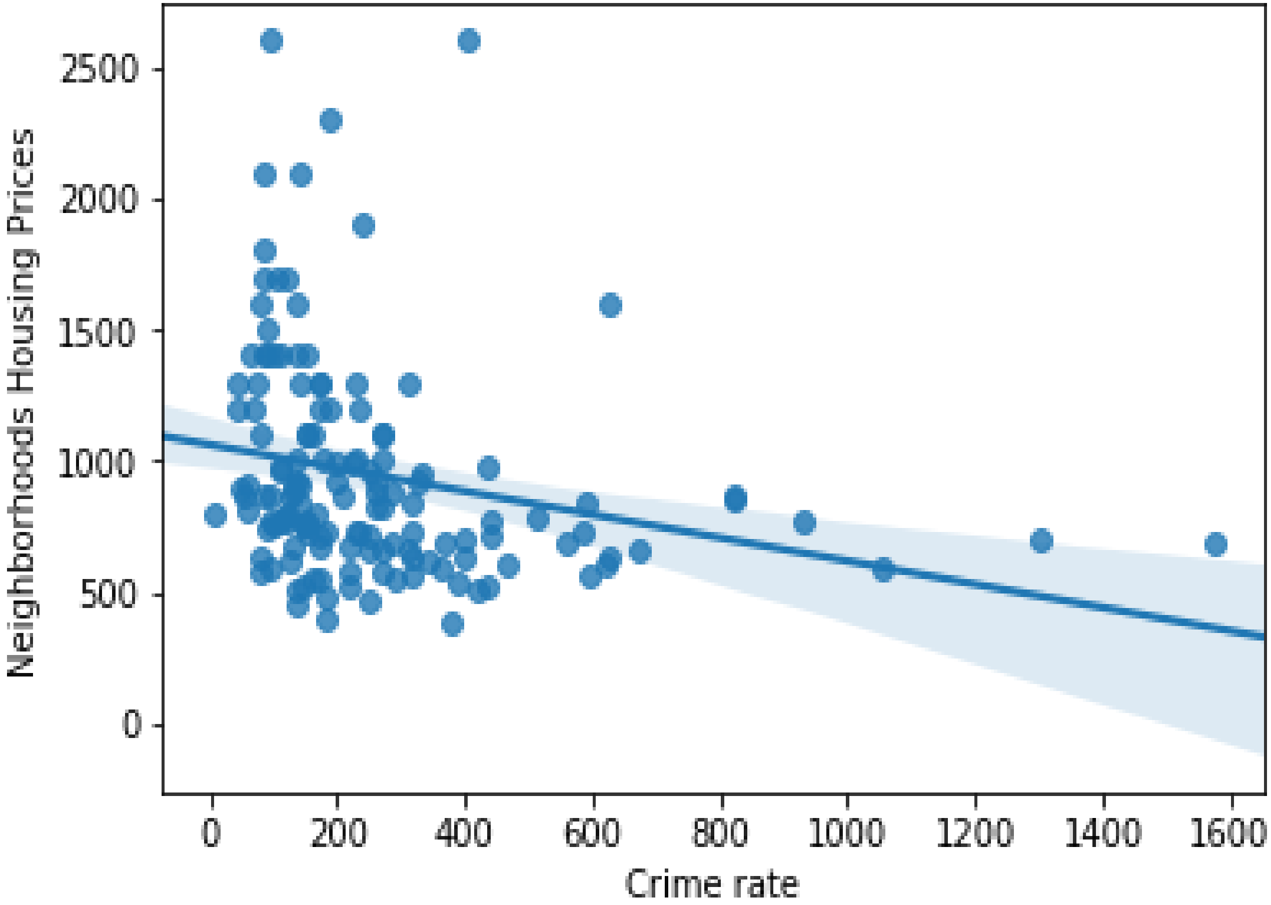
- Although DB-SCAN, K-means, Hierarchical Clustering algorithms were employed on the dataset, DB-SCAN was the algorithm of choice.
- DB-SCAN's ability to locate and separate high density regions from low density regions and locating outliers was of interest.

Descriptive analysis of the dataset



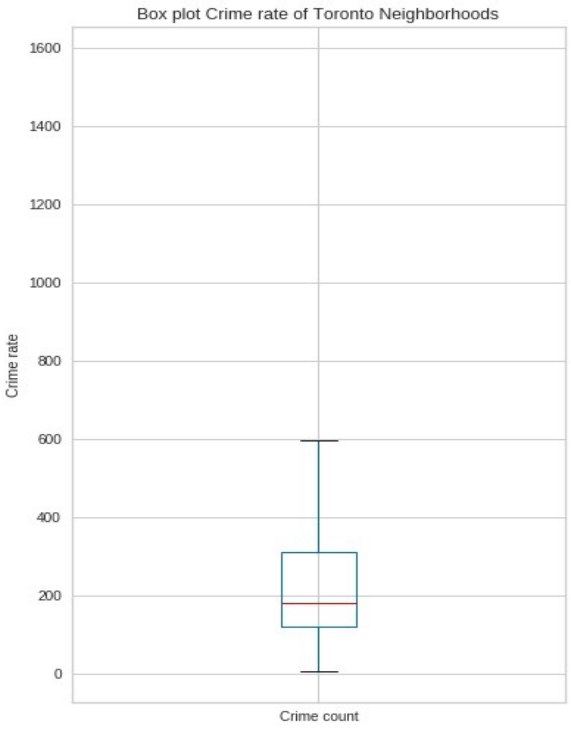
	AveragePrice$x10^3	Hood_ID	Lat	Long	Crime count	Arts_and_Entertainment	Schools
count	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000
mean	948.914894	70.078014	43.707510	-79.402293	257.468085	0.971631	0.007092
std	417.360729	40.722944	0.050919	0.102288	231.788036	1.535035	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

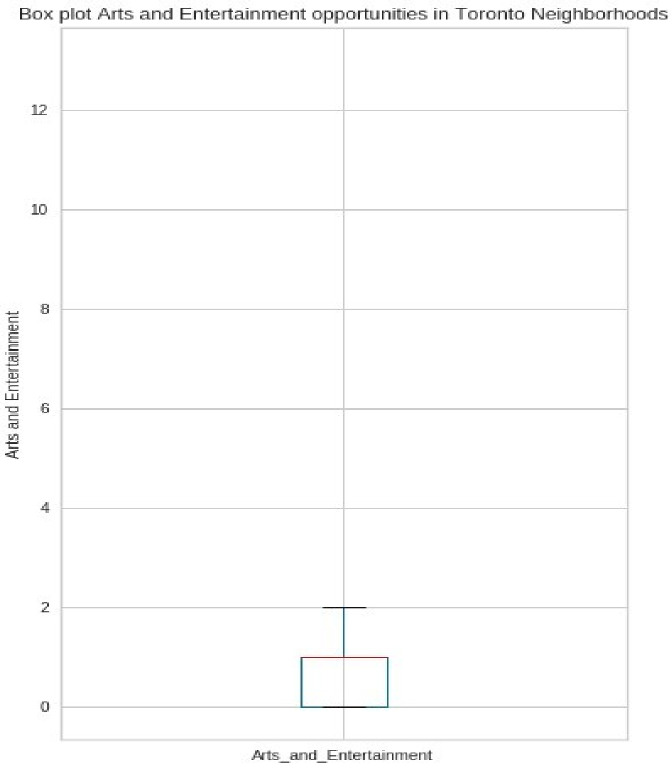
Scatter plot and regression line of Crime rate vs Housing Price

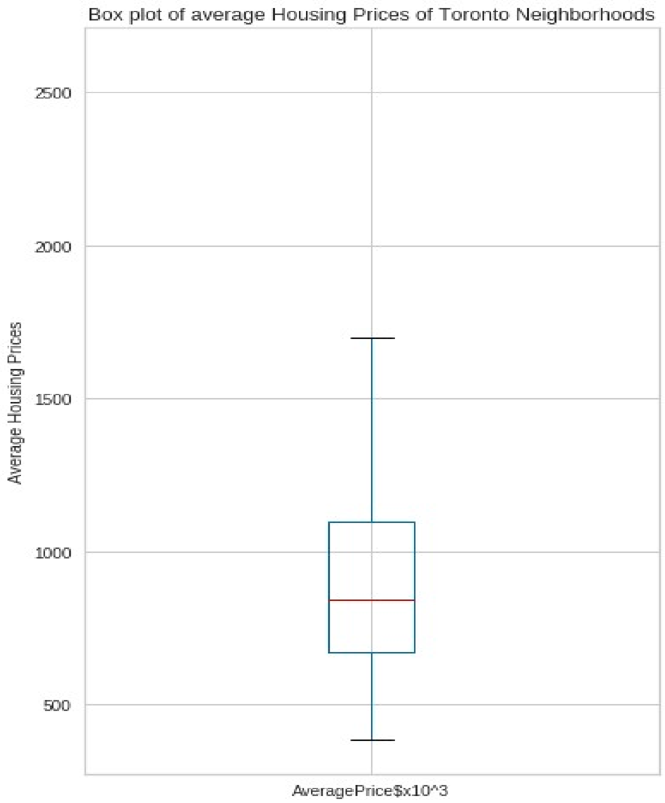


The plot above shows that the relationship between Crime rate and Housing price is non-linear.

Box Plots


Box plot Crime rate of Toronto Neighborhoods


Box plot Arts and Entertainment opportunities in Toronto Neighborhoods


Box plot of average Housing Prices of Toronto Neighborhoods

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

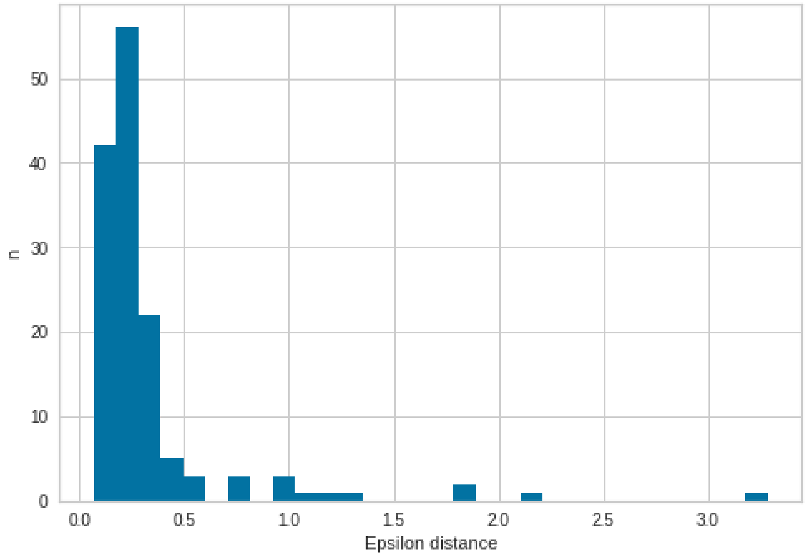
Creating Grades

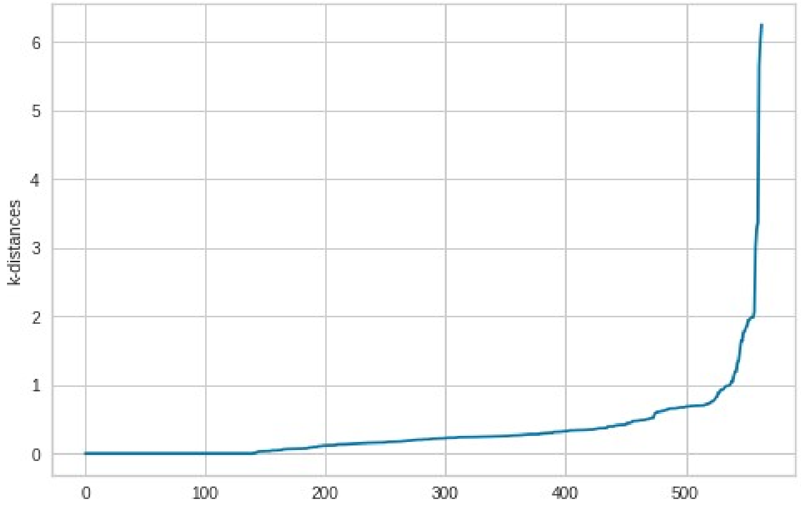
GRADING	PERCENTILES	HOUSING PRICE($ x10^3)	CRIME RATE
LOW	Below 25	385 - 674	6- 120
AVERAGE	25 to 50	675- 842	121- 183
ABOVE AVERAGE	50 to 75	843 - 1100	184- 312
HIGH	75 to maximum	1101- 1750	313- 600
VERY HIGH	outliers	1751- 2600	601-1575

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

* maximum is Q3 + 1.5*IQR

DB- SCAN- episilon determination

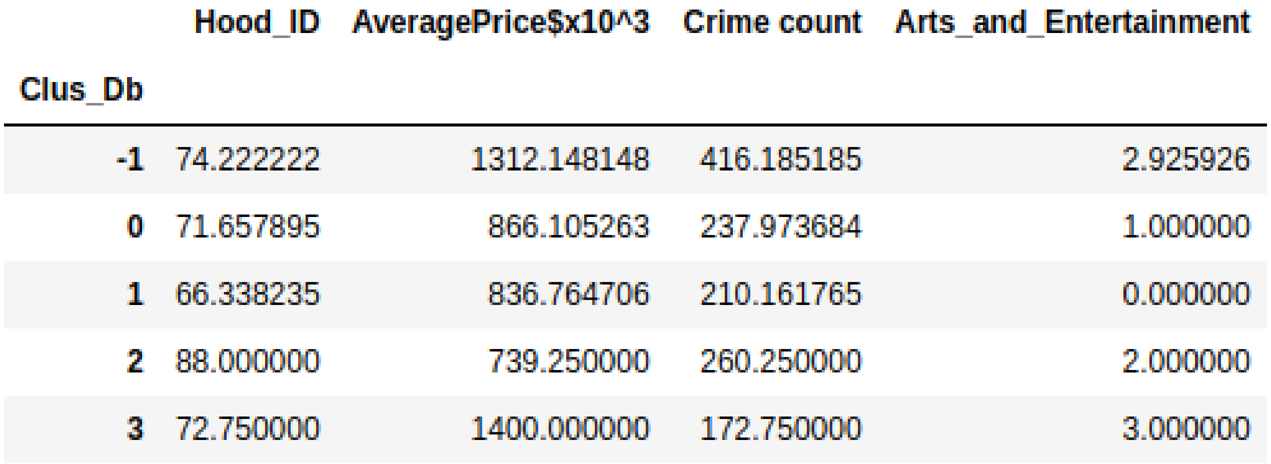




The minimum number of points was taken as 4, 2* dimensionality of datasets. The optimal epsilon value was calculated using Knearest neighbor distance plots and histogram.

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

DB-SCAN- Clusters



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

A profile was created for each cluster, considering the characteristics of its features;

LABEL 0- Above Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities

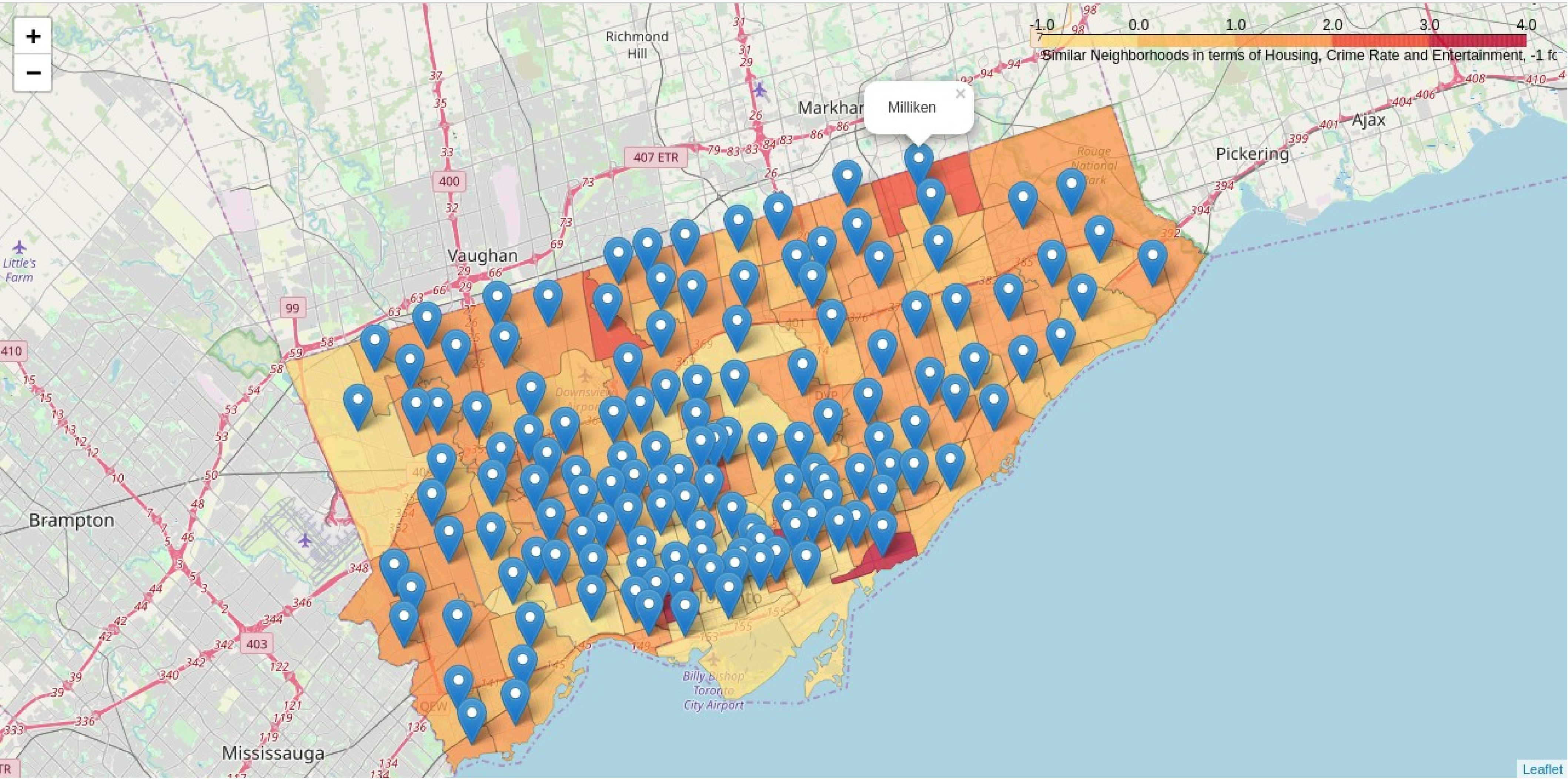
LABEL 1- Average Housing Price, Above Average Crime Rate, No Entertainment Opportunities

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

DB-SCAN- Map



This interactive map is better experienced on a browser.

Neighborhood Clusters

LABEL 0- Above Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities, **37 neighborhoods**

LABEL 1- Average Housing Price, Above Average Crime Rate, No Entertainment Opportunities, **68 neighborhoods**

LABEL 2- Average Housing Price, Above Average Crime Rate, Few Entertainment Opportunities, **4 neighborhoods**

LABEL 3- High Housing Price, Average Crime Rate, Fair Amount of Entertainment Opportunities, **4 neighborhoods**

LABEL -1 Outliers, **27 neighborhoods**

Conclusion

This project provides a starting point in considering potential neighborhoods for residence and reduces the time and energy needed for such research.

Thus enabling anyone migrating to Toronto filter out neighborhoods to reside in based on their preferences.