COURSERA CAPSTONE PROJECT: BATTLE OF THE NEIGHBORHOODS

An IBM Applied Data Science Course

Analyzing Crime Rate in Toronto Neighborhoods

By: Ani Chidera Priscilla

June 2020.

1 INTRODUCTION/BUSINESS PROBLEM

From the Coursera IBM Data Science Certificate Capstone Project, I learned how to cluster neighborhoods in Toronto using K-Means Clustering and Foursquare API. The Foursquare API has a wide range of abilities which includes providing top venues in a particular location.

A peaceful neighborhood is desired by everyone and reoccurring crime cases would disrupt the peace and serenity of a neighborhood. This project aims to analyze the frequency of crime rates in Toronto neighborhoods from the years 2014 to 2019.

Using data science methodology and machine learning algorithms, this project would iterate through data gathering, data cleaning, data preprocessing, and exploratory data analysis. This capstone project would aim to identify suitable neighborhoods that would be fit for families and individuals to live, business owners to locate their businesses with safety as the major determinant.

2 DATA

To help with this analysis, the following datasets would be required:

- Major Crime Indicators(MCI) of neighborhoods in Toronto, from the years 2014-2019 which I have downloaded from Kaggle(https://www.kaggle.com/kapastor/toronto-police-data-crime-rates-by-neighbourhood/data%22).
- Geographical coordinates data on the latitude and longitude of each neighborhood. Although this was already provided in the MCI data above, I noticed it was inconsistent across each neighborhood so I recompiled it using Python Geocoder library package and stored it in an excel file.
- Toronto 2016 census population data obtained via this website https://open.toronto.ca/dataset/neighbourhood-profiles/.
- Foursquare API to get the associated venues for each neighborhood.

3 METHODOLOGY

3.1 Data Cleaning and Preprocessing

Firstly, we would load the MCI data, drop unnecessary columns, get the crime count for each neighborhood, and drop the duplicate rows. Next, we would load the coordinates data and merge it with the MCI data. For the population data, I had already dropped unnecessary rows and columns before loading it. Then I found the percentage of the population that own private dwellings which is the column we would be needing. I went ahead to merge it the previous dataframe and the resulting dataframe is shown below:

:

	Hood_ID	Neighbourhood	Crime Count	Lat	Long	Percent Occupied
0	79	University	1455	43.728243	-79.377503	0.446037
1	118	Tam O'Shanter-Sullivan	1371	43.781100	-79.298100	0.366647
2	137	Woburn	3798	43.776500	-79.231700	0.344695
3	133	Centennial Scarborough	508	43.781700	-79.148300	0.327720
4	61	Taylor-Massey	1147	43.701668	-79.328331	0.399923
...	...	...	...	...	...	...
135	49	Bayview Woods-Steeles	539	43.794800	-79.382500	0.357154
136	60	Woodbine-Lumsden	377	43.692200	-79.309900	0.438525
137	106	Humewood-Cedarvale	645	43.694500	-79.428100	0.457083
138	58	Old East York	479	43.692000	-79.337800	0.410484
139	29	Maple Leaf	410	43.714800	-79.479400	0.351696

140 rows × 6 columns

3.2 COLLECTING DATA FROM FOURSQUARE API

The next step is to use Foursquare API to get the top 100 venues within a radius of 1.5km. To do this, I have previously registered a Developer Account to have access to a client ID and client secret. After making the API call, the venues will be returned in a JSON format from which we would extract the venue name, category, latitude, and longitude. After this, I would make a

count of the venues in each neighborhood and exclude those with little venues(venues less than 4).

3.3 PREPARING DATA FOR ANALYSIS

To prepare the data for analysis, I would be using one hot encoding to change venue categories from rows to columns and obtain the average frequency occurrence of each category.

3.4 CLUSTERING

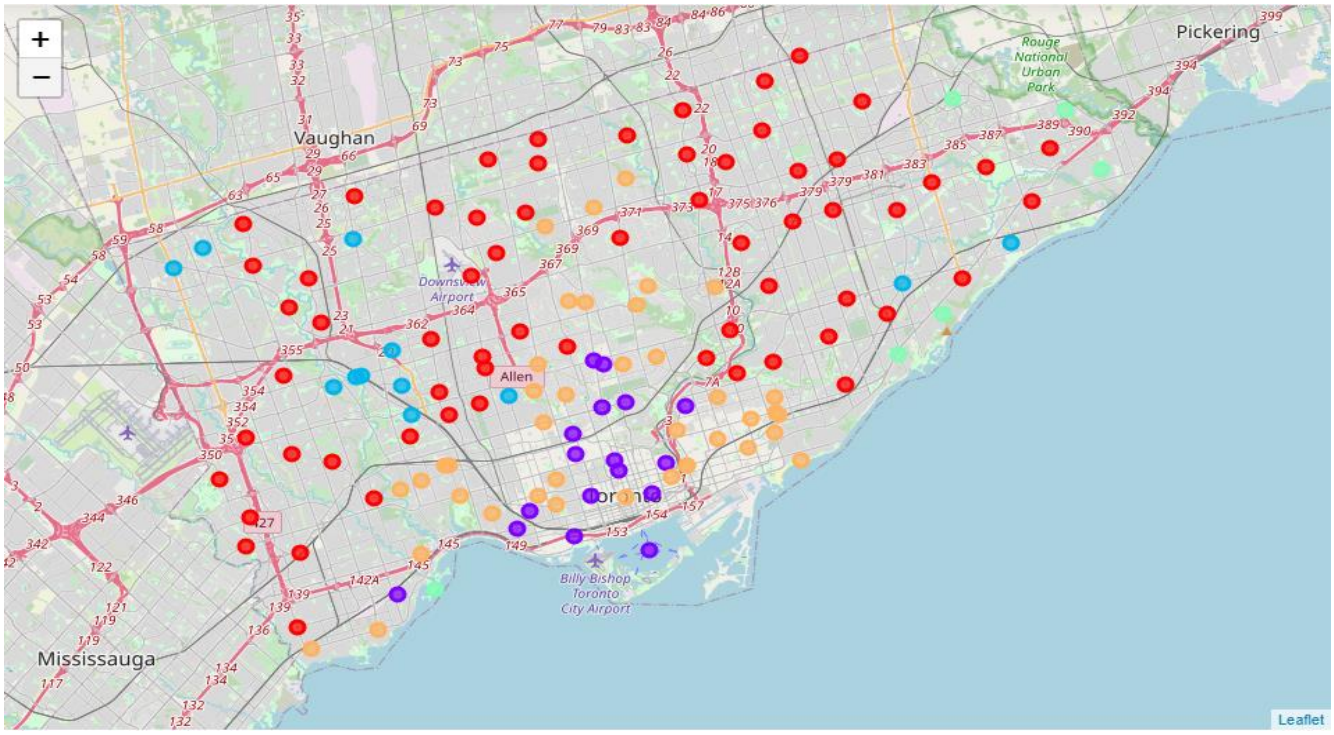
The last step is clustering the data using K-Means Clustering which is an unsupervised machine learning algorithm that divides data into non-overlapping subsets. It identifies k number of centroids and allocates data points to the nearest cluster while minimizing the centroids. The neighborhoods would be segmented into 5 clusters.

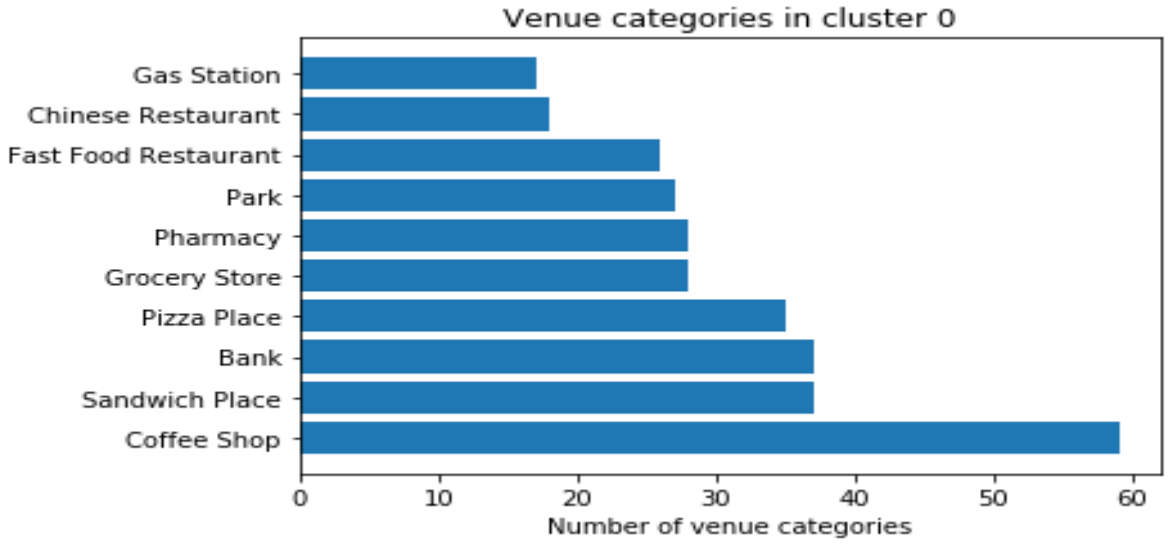
4 RESULTS

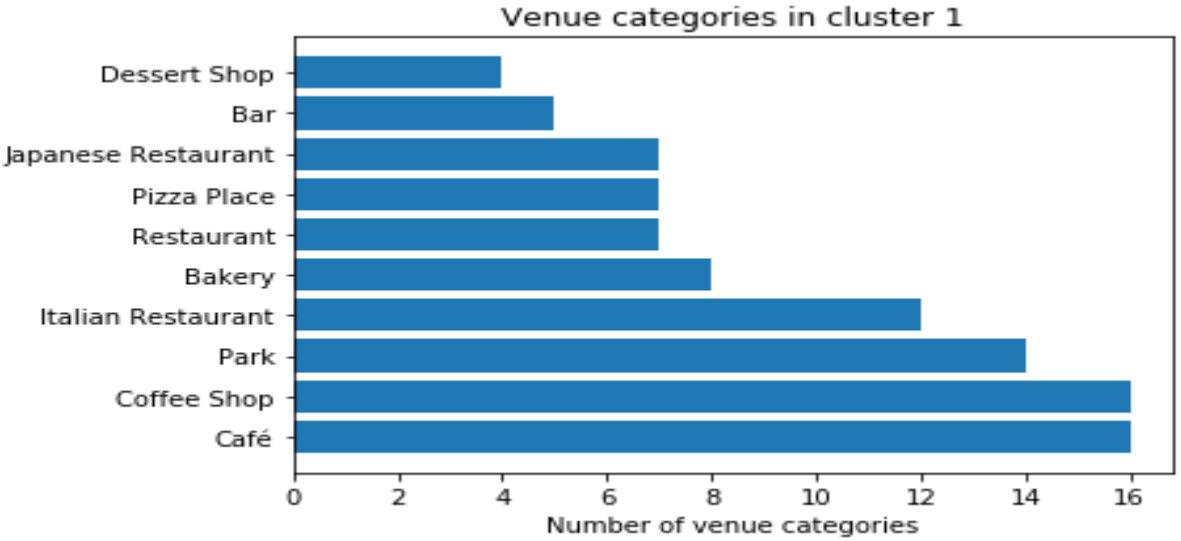
The result from the k-means clustering shows the following clusters:

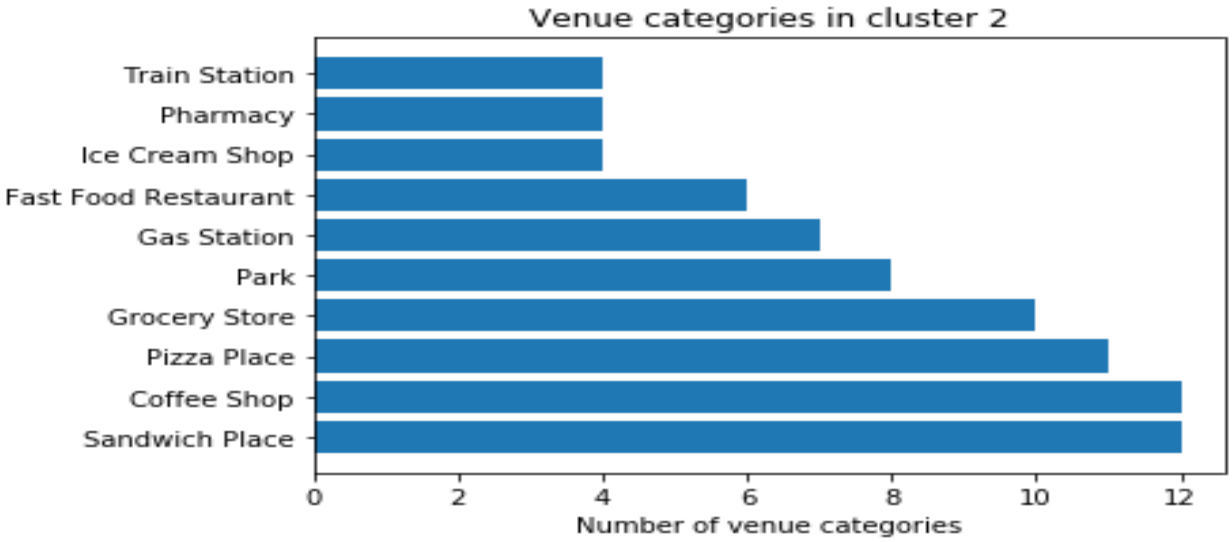
- Cluster 0: Neighborhoods indicated by a red marker
- Cluster 1: Neighborhoods indicated by a purple marker
- Cluster 2: Neighborhoods indicated by a blue marker
- Cluster 3: Neighborhoods indicated by a mint green marker
- Cluster 4: Neighborhoods indicated by an orange marker

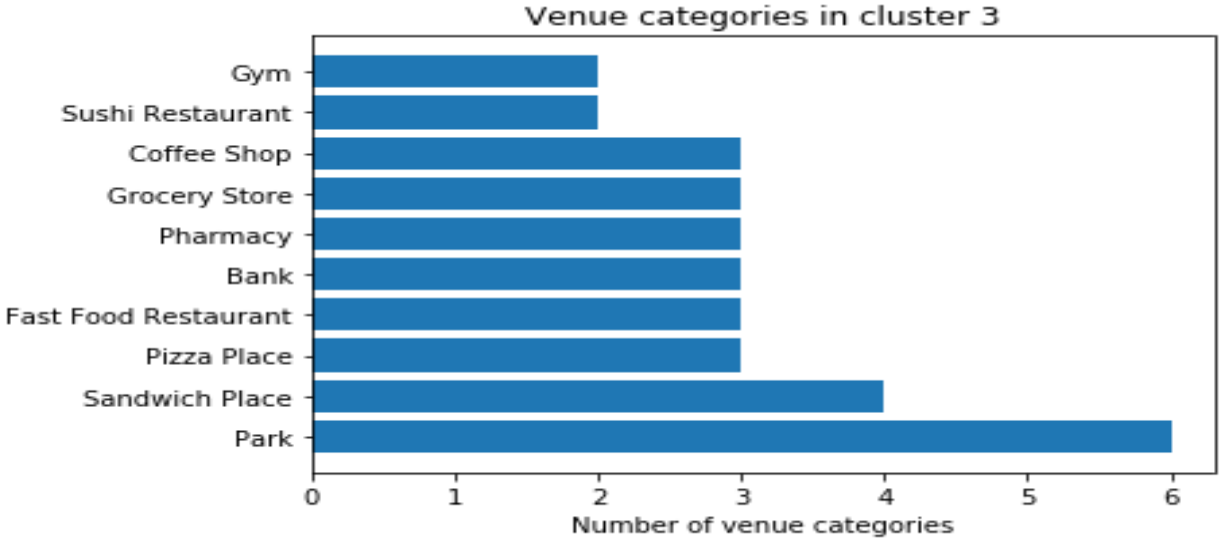
The clustering can be visualized in a map using the folium package as shown below:

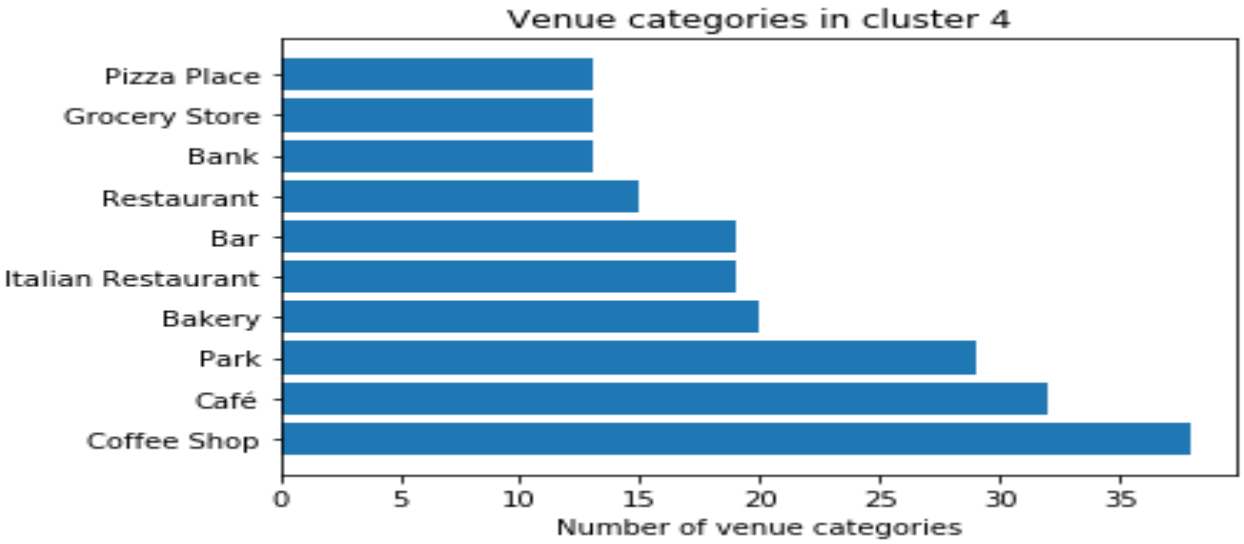





Venue categories in cluster 1
Dessert Shop
Bar
Japanese Restaurant
Pizza Place
Restaurant
Bakery
Italian Restaurant
Park
Coffee Shop
Café
0
2
4
6
8
10
12
14
16
Number of venue categories

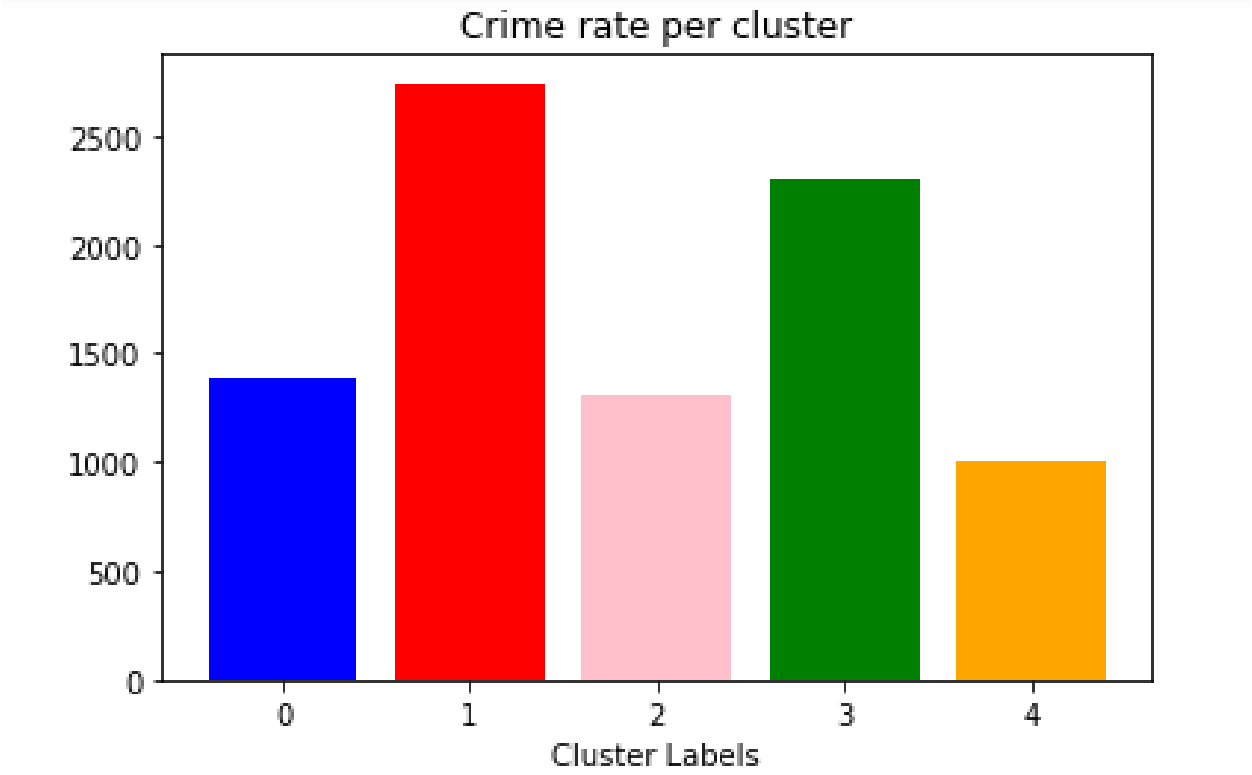

Venue categories in cluster 2
Train Station
Pharmacy
Ice Cream Shop
Fast Food Restaurant
Gas Station
Park
Grocery Store
Pizza Place
Coffee Shop
Sandwich Place
0
2
4
6
8
10
12
Number of venue categories


Venue categories in cluster 3
Gym
Sushi Restaurant
Coffee Shop
Grocery Store
Pharmacy
Bank
Fast Food Restaurant
Pizza Place
Sandwich Place
Park
0
1
2
3
4
5
6
Number of venue categories



5. DISCUSSION

From the clustering above, it can be deduced that cluster 0 consists mainly of coffee shops, cluster 1 consists mainly of cafes and coffee shops, cluster 2 mainly consists of restaurants and eateries, cluster 3 is made up of a large number of parks and cluster 4 constitutes mainly coffee shops as well as other basic amenities required in residential areas.



	Cluster Labels	Hood_ID	Crime Count	Percent Occupied
0	0	65.142857	1383.841270	0.367022
1	1	80.294118	2740.647059	0.539708
2	2	66.750000	1312.083333	0.370302
3	3	107.000000	2303.833333	0.335669
4	4	69.900000	1007.575000	0.425457

Comparing the data with the crime count in each cluster, as shown above, it can be observed that clusters 4, 2, and 0 have the lowest crime rates. Although, it can also be seen from the dataframe above that cluster 1 while having the highest crime count also has the highest percentage of private dwellings. This could be because private dwellings were considered for this analysis and other clusters may be more high profile with fewer buildings. This research undoubtedly has a few limitations.

6. CONCLUSION

As shown by the analysis, neighborhoods in clusters 4, 2, and 0 have the lowest crime rates and would be suitable for those looking for peaceful and safe neighborhoods.

A lot of improvement could be made to this analysis, some of which include:

- Sourcing for more data on the number of residential buildings in Toronto neighborhoods.
- Crowdsourced data providers should make more data readily accessible about Toronto.
- Using a paid Foursquare account instead of a free Sandbox account that comes with a lot of limitations on the number of API calls.

- Further insight into why cluster 1 has the highest crime rate and the highest number of private dwellings.

7. REFERENCES

1. Coursera Capstone Lab: Neighborhoods in New York, hosted on IBM Skills Network Labs, https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DP0701EN/DP0701EN-3-3-2-Neighborhoods-New-York-py-v1.0.ipynb
2. Coursera Capstone Segmenting Neighborhoods in Toronto, https://github.com/chideraani/Coursera_Capstone/blob/master/Segmenting%20and%20Clustering%20Neighborhoods%20in%20Toronto.ipynb
3. Foursquare API, https://developer.foursquare.com/docs/
4. Major Crime Indicators of Neighborhoods in Toronto from Kaggle, https://www.kaggle.com/kapastor/toronto-police-data-crime-rates-by-neighbourhood/data%22
5. Toronto 2016 census population data, https://open.toronto.ca/dataset/neighbourhood-profiles/