Rishabh Agarwal

Applied Data Science Capstone Final Report

IBM Data Science

Introduction

An interesting business problem for many businesses being opened in Toronto is location. A certain problem with opening a business or a residential location in a certain area is that that specific area may be filled with crime. Any business or residency would want to minimize the amount of criminal activity going on in their neighborhood. Thus, comes the problem, what neighborhoods in Toronto have the lowest amount of crime in many different categories. Specifically, this can be differentiated into three types of crime: Commercial crime, residential crime, and traffic crimes. Each one of these will be dissected and understood. Using data, our group can understand which neighborhoods are prone to certain types of crime and can cluster these neighborhoods together into desirable, non-desirable, and medium desirable neighborhoods.

Data Used

As said before, our group will use data to understand the crime prevalence in these neighborhoods. The data used to gain this knowledge is the Toronto DataSet given by the City of Toronto. Through this data, we may call SQL statements to see the crimes in certain neighborhoods and which crimes are happening in certain neighborhoods. Each of these data-sets has a premise clause in which it specifies if a crime is in a commercial domain or a residential domain. Using this information, our group can understand which neighborhoods are going through certain crimes more. After this, the use of Four-Square and a Folium map can be used to visualize the neighborhoods with certain problems and see if there is a pattern in the data.

In this data, the number of crimes per year from 2014-2019 in each category is given. The categories are Assault, Auto-Theft, Breaking and Entering, Homicides, Robberies, and Theft Overs. These numbers will then be used to cluster the neighborhoods to see which neighborhoods are the worst in terms of which neighborhoods have the most crime.

Methodology

The methodology used in this study is fairly simple. What will be done is that the data will be manipulated to ascertain the average amount of crime in a certain category in a year. For example, the data set will show the average number of robberies in a certain neighborhood in a certain year. Then, the numbers in each year in each neighborhood will be used to find boxplots and numbers of the number of crimes in each category. In this, we will find the 25% percentile, minimum, maximum, etc. of the crimes in each category. Then, simply, using these numbers, we can find the worst neighborhoods by clustering the neighborhoods by their numbers of crime.

Results

In our data, we found the boxplots of crime in Toronto in each of the major categories

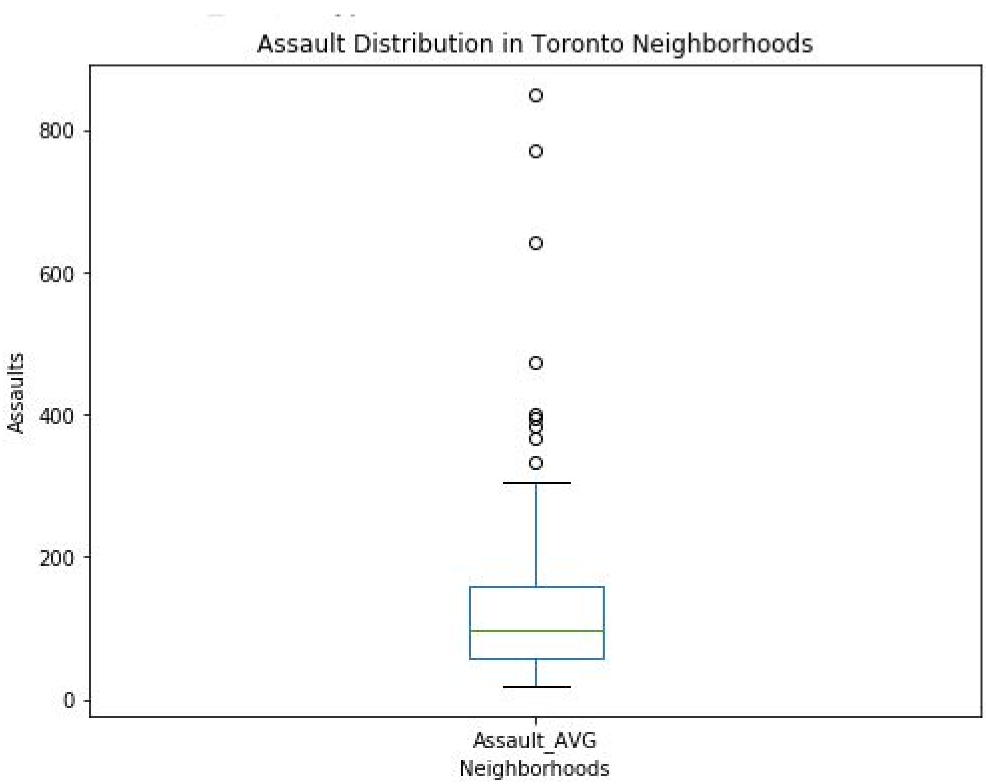

Figure 1 shows the assault distribution in Toronto Neighborhoods

```
count    140.000000
mean     132.646429
std      128.977375
min       18.500000
25%       59.425000
50%       96.500000
75%      160.200000
max      851.800000
```

Figure 2 shows the numerical summary of the number of assaults in Toronto

Then, this is done for Auto-Theft

```
count    140.000000
mean      27.835000
std       35.047468
min        2.700000
25%       13.275000
50%       18.800000
75%       30.975000
max      366.700000
Name: AutoTheft_AVG, dtype: float64
```

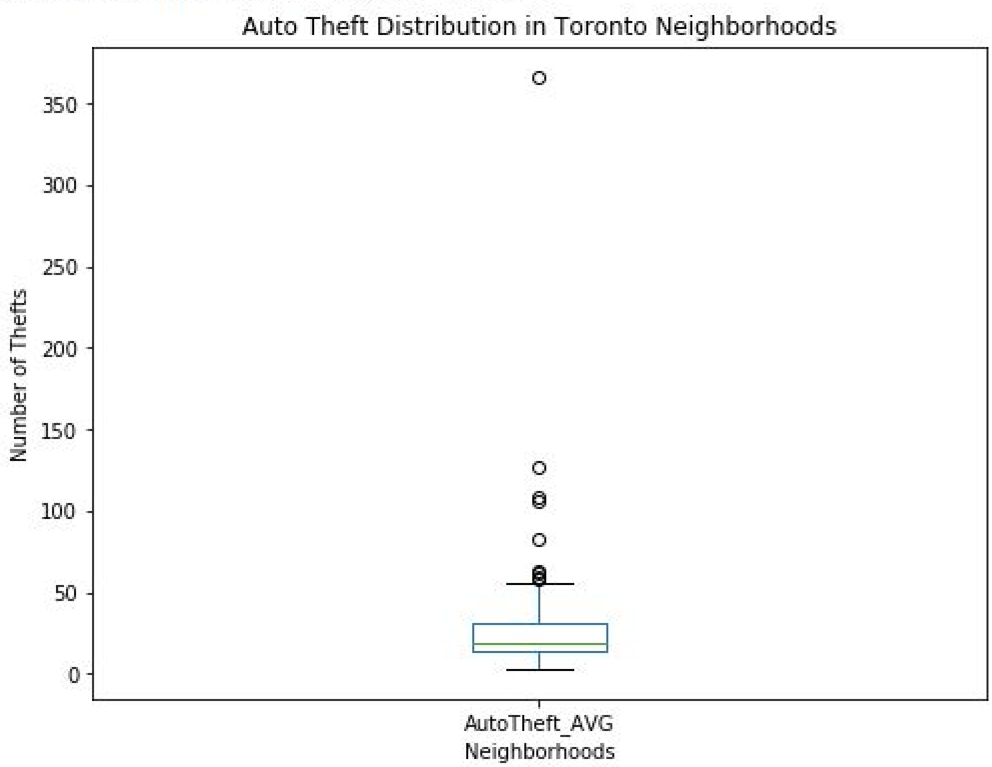

Figure 3 shows Auto-Theft boxplot and numerical summary.

```
count    140.000000
mean      51.548571
std       36.760413
min       10.500000
25%       28.000000
50%       40.750000
75%       64.450000
max      247.300000
Name: BreakandEnter_AVG, dtype: float64
```

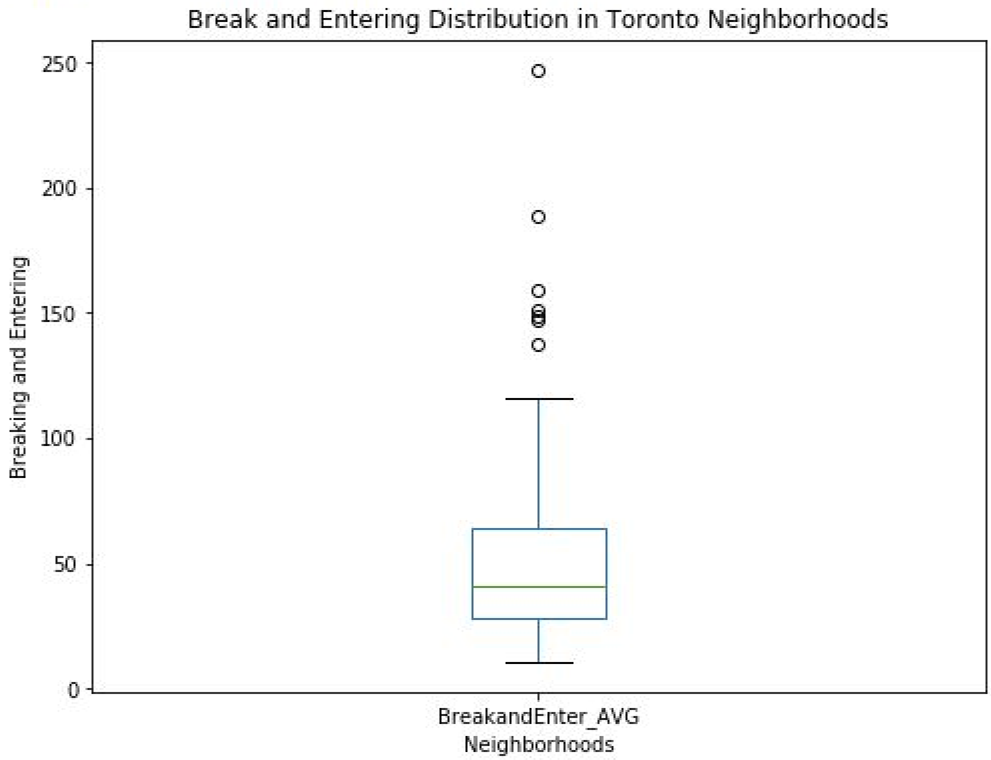

Figure 4 shows the Breaking and Entering distribution in Toronto

```
count    140.000000
mean       0.513571
std        0.517911
min        0.000000
25%        0.200000
50%        0.300000
75%        0.725000
max        2.500000
Name: Homicide_AVG, dtype: float64
```

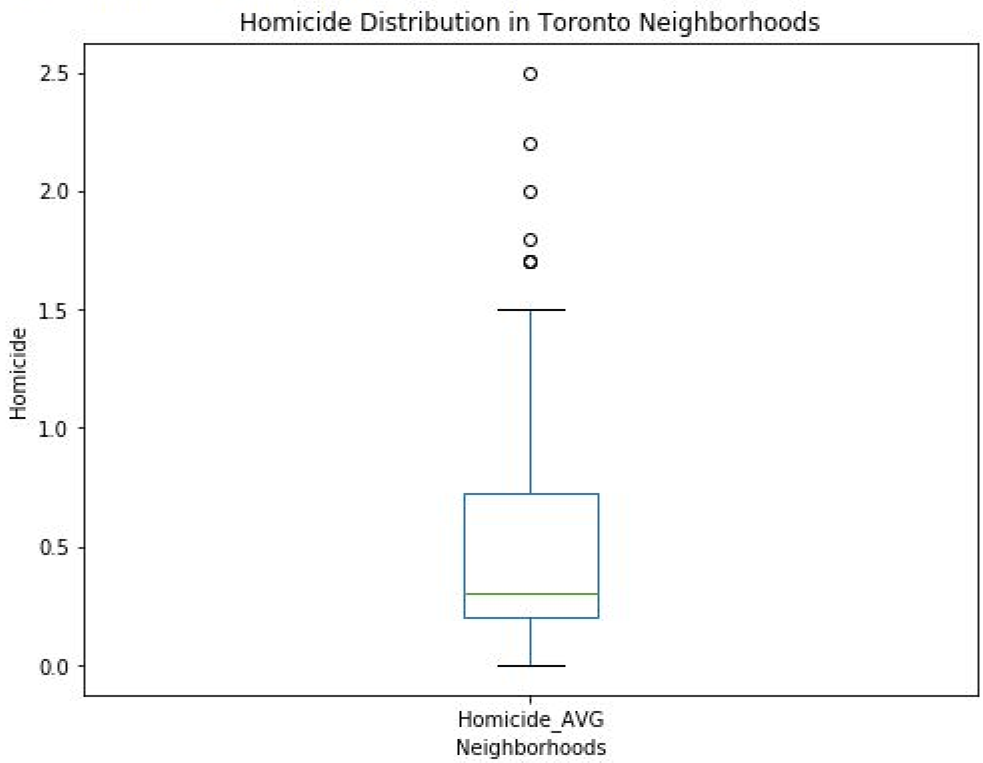

Figure 5 shows the homicide distribution in Toronto

```
count    140.000000
mean      25.647143
std       23.220601
min        3.300000
25%       11.675000
50%       20.100000
75%       30.400000
max      135.700000
Name: Robbery_AVG, dtype: float64
```

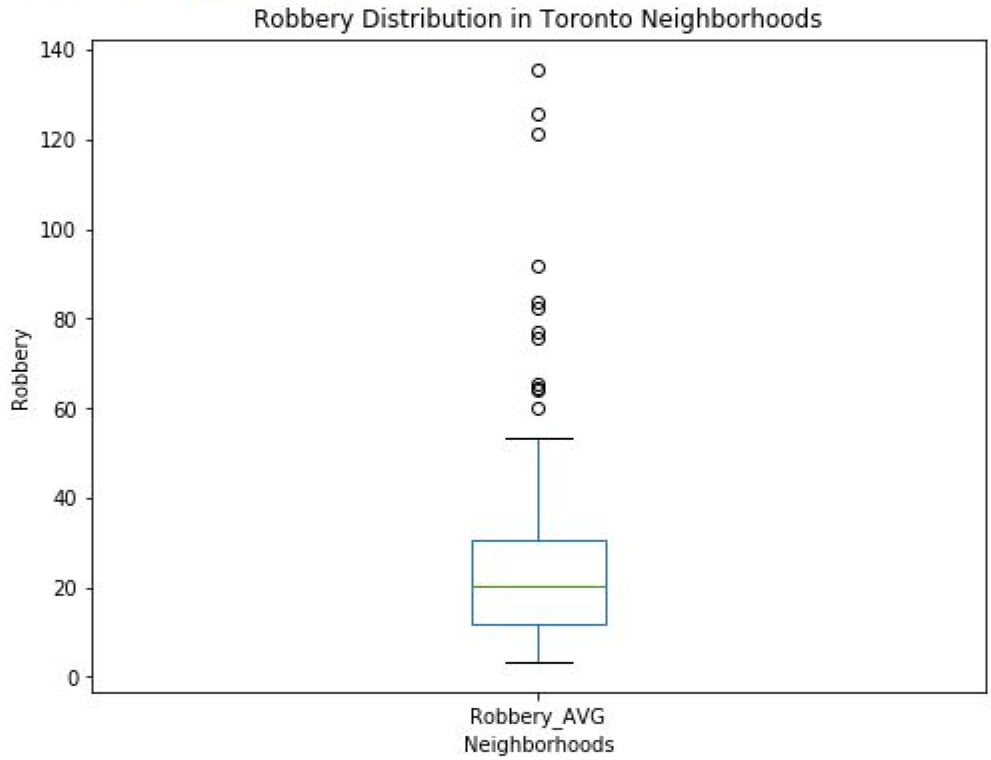

Figure 6 shows the robbery distribution in Toronto

```
count    140.000000
mean       8.082857
std        9.427947
min        1.200000
25%        3.500000
50%        5.200000
75%        8.350000
max       56.200000
Name: TheftOver_AVG, dtype: float64
```

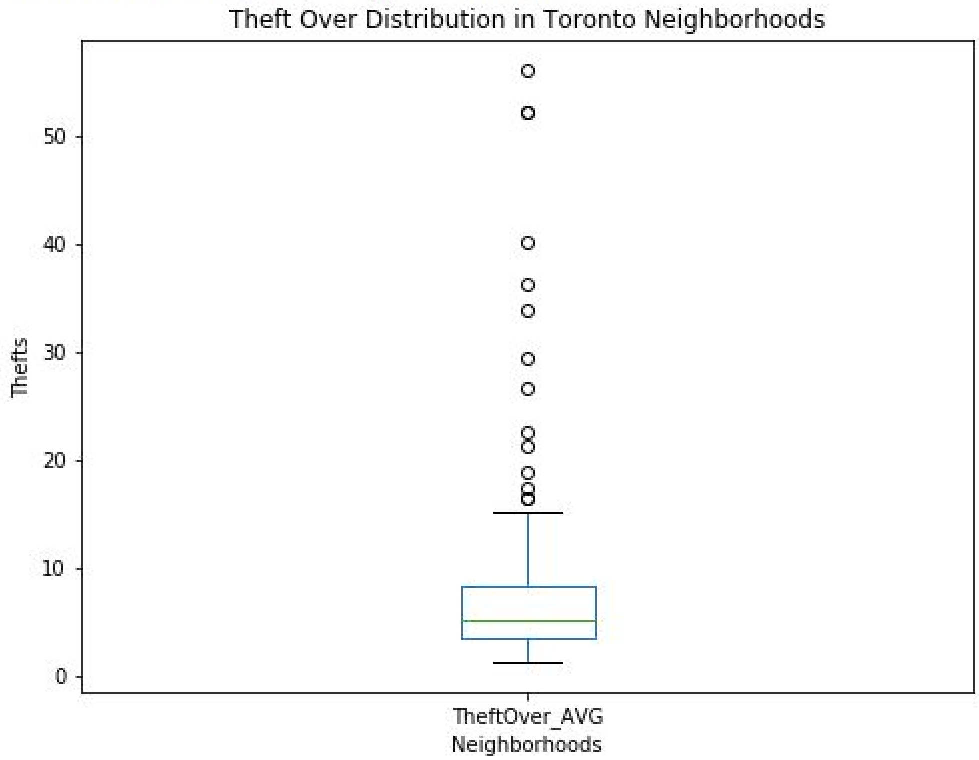

Figure 6 shows Theft Over Distribution in Toronto

Then, using this data, we placed it into a k-means clustering method to find the bad, medium, and good neighborhoods in terms of crime.

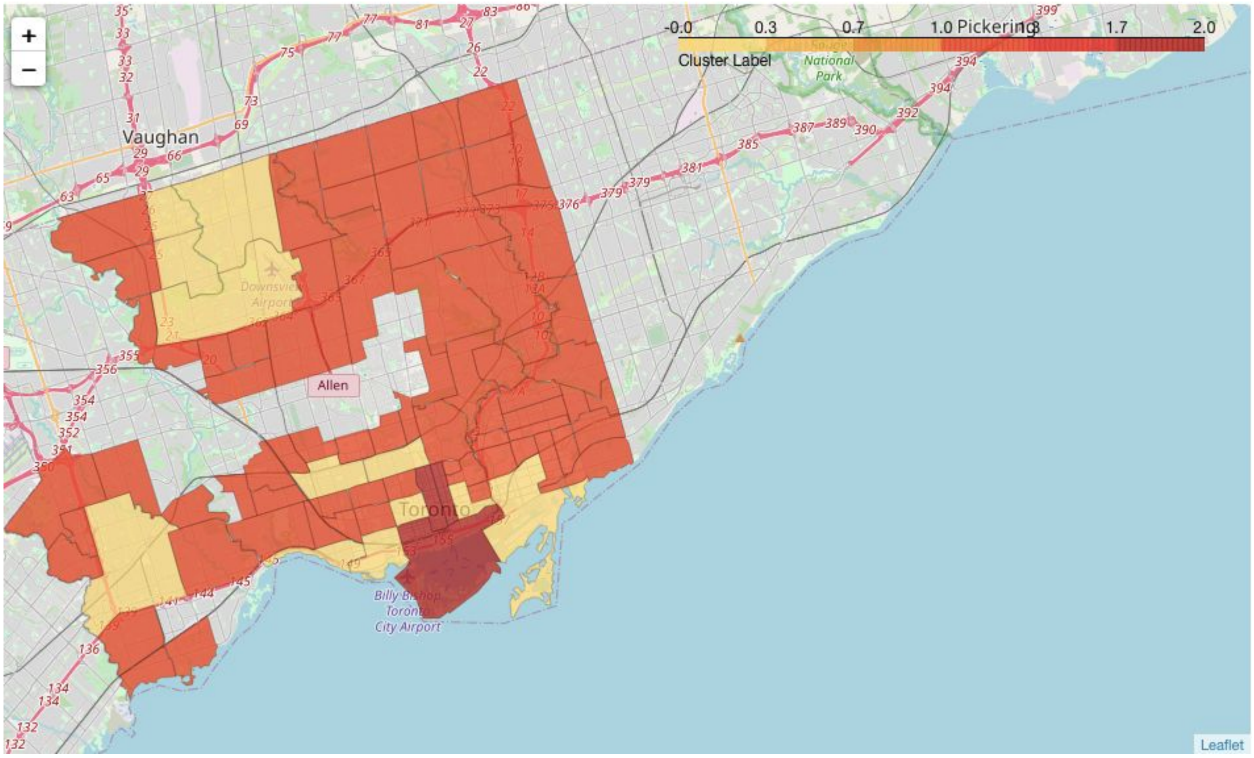

Figure 7 shows the clustering distribution of the Toronto Neighborhoods. In this clustering method, the yellow ones marked as cluster 0 are the ones with medium crime levels. The ones with red marking which are named as cluster 1 are low levels of crime and the ones marked in red which is named as cluster 2 are the ones with high levels of crime.

	Neighbourhood	Cluster Labels	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	Same Column
1	York University Heights	0	333.2	106.3	113.2	0.8	75.8	36.3	True
11	Islington-City Centre West	0	223.0	126.5	116.3	0.8	41.8	40.2	True
15	South Parkdale	0	226.5	18.7	65.3	0.3	33.0	10.0	True
16	South Riverdale	0	244.3	30.8	108.8	1.8	49.0	21.3	True
34	Glenfield-Jane Heights	0	304.8	59.2	36.7	0.8	53.2	8.8	True
51	Kensington-Chinatown	0	368.2	27.5	150.8	1.5	64.0	26.7	True
56	Annex	0	246.3	22.0	147.5	0.5	40.8	29.5	True
65	Dovercourt-Wallace Emerson-Junction	0	240.8	32.0	106.5	1.7	51.0	9.5	True
67	Niagara	0	263.7	24.7	85.5	0.8	20.5	16.5	True
91	Downsview-Roding-CFB	0	395.8	107.8	78.8	1.3	64.7	15.2	True
94	Black Creek	0	218.8	48.8	28.8	0.8	39.2	9.2	True
131	Moss Park	0	474.7	30.2	148.5	2.5	125.5	18.8	True

Figure 8 shows the display for the neighbourhoods with medium amount of crime.

	Neighbourhood	Cluster Labels	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	Same Column
0	Yonge-St.Clair	1	31.0	4.3	23.3	0.0	5.7	4.3	True
2	Lansing-Westgate	1	70.7	23.7	38.8	1.7	14.7	7.0	True
3	Yorkdale-Glen Park	1	160.2	55.5	63.3	1.2	31.5	22.5	True
4	Stonegate-Queensway	1	83.2	28.7	52.8	0.0	20.7	6.0	True
6	The Beaches	1	93.8	16.3	49.3	0.0	20.3	6.2	True
...	...	...	...	...	...	...	...	...	...
133	Woodbine Corridor	1	86.0	9.2	32.5	0.5	14.5	4.0	True
134	Newtonbrook East	1	66.5	11.7	49.8	0.3	9.0	5.2	True
136	Pleasant View	1	46.0	13.5	19.8	0.2	11.8	3.8	True
137	Wychwood	1	70.2	13.2	34.0	0.3	13.8	2.3	True
138	Leaside-Bennington	1	32.8	18.2	33.3	0.2	7.5	5.2	True

Figure 9 shows the neighborhoods with low amounts of crime.

	Neighbourhood	Cluster Labels	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	Same Column
22	Church-Yonge Corridor	2	642.8	37.8	188.5	2.0	135.7	33.8	True
39	Waterfront Communities-The Island	2	851.8	53.7	247.3	1.0	82.2	56.2	True
93	Bay Street Corridor	2	771.0	32.8	158.7	1.5	121.3	52.3	True

Figure 10 shows the neighborhoods with extremely high amounts of crime

After this data analysis, what I realized is that using each crime as an axis in k-means might not be the best way. So, I also decided to conduct a normalized sum of all data and use that to cluster the dataset as well.

To find this normalization, I found the Z-score of each part and summed all the z-scores for each neighborhood together. The dataset after doing this looked like

Neighbourhood	Cluster Labels 2.0	Cluster Labels	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG	TheftOver_AVG	Same Column	Sum of All Crime
Yonge-St.Clair	2	1	31.0	4.3	23.3	0.0	5.7	4.3	True	-4.589174
York University Heights	1	0	333.2	106.3	113.2	0.8	75.8	36.3	True	12.241419
Lansing-Westgate	0	1	70.7	23.7	38.8	1.7	14.7	7.0	True	0.845412
Yorkdale-Glen Park	0	1	160.2	55.5	63.3	1.2	31.5	22.5	True	4.956986
Stonegate-Queensway	2	1	83.2	28.7	52.8	0.0	20.7	6.0	True	-1.603224
The Beaches	2	1	93.8	16.3	49.3	0.0	20.3	6.2	True	-2.173639
Thorncliffe Park	2	1	97.5	9.3	25.5	1.5	11.0	7.2	True	-0.458807
Danforth East York	2	1	65.8	9.3	27.2	0.0	5.8	2.8	True	-4.189957
Islington-City Centre West	1	0	223.0	126.5	116.3	0.8	41.8	40.2	True	11.537141
Danforth	2	1	72.3	6.2	37.3	0.8	20.7	3.7	True	-1.813350

Figure 11 shows the entire dataframe after manipulation.

Then, using this other dataset, another clustering algorithm was used to find the clusters of neighborhoods with high-medium-low crime.

This is the Data for the three clusters

```
Average Assaults in 1st Cluster: 176.94444444444443
Average Auto Thefts in 1st Cluster: 37.32777777777777
Average Break and Enters in 1st Cluster: 69.78333333333335
Average Homicides in 1st Cluster: 0.7388888888888889
Average Robbery in 1st Cluster: 32.455555555555556
Average Theft Over in 1st Cluster: 12.488888888888887
```

```
Average Assaults in 2nd Cluster: 507.5625
Average Auto Thefts in 2nd Cluster: 65.325
Average Break and Enters in 2nd Cluster: 150.2625
Average Homicides in 2nd Cluster: 1.4249999999999998
Average Robbery in 2nd Cluster: 88.875
Average Theft Over in 2nd Cluster: 34.9375
```

```
Average Assaults in 3rd Cluster: 75.35555555555553
Average Auto Thefts in 3rd Cluster: 15.6952380952381
Average Break and Enters in 3rd Cluster: 36.52857142857143
Average Homicides in 3rd Cluster: 0.29523809523809524
Average Robbery in 3rd Cluster: 14.37301587301587
Average Theft Over in 3rd Cluster: 4.7
```

Figure 12 shows the averages for each crime in each cluster.

As we can see, the order of lowest to highest crime goes from the 3rd Cluster to the 1st Cluster to the 2nd Cluster.

These clusters are shown below in a map of Toronto

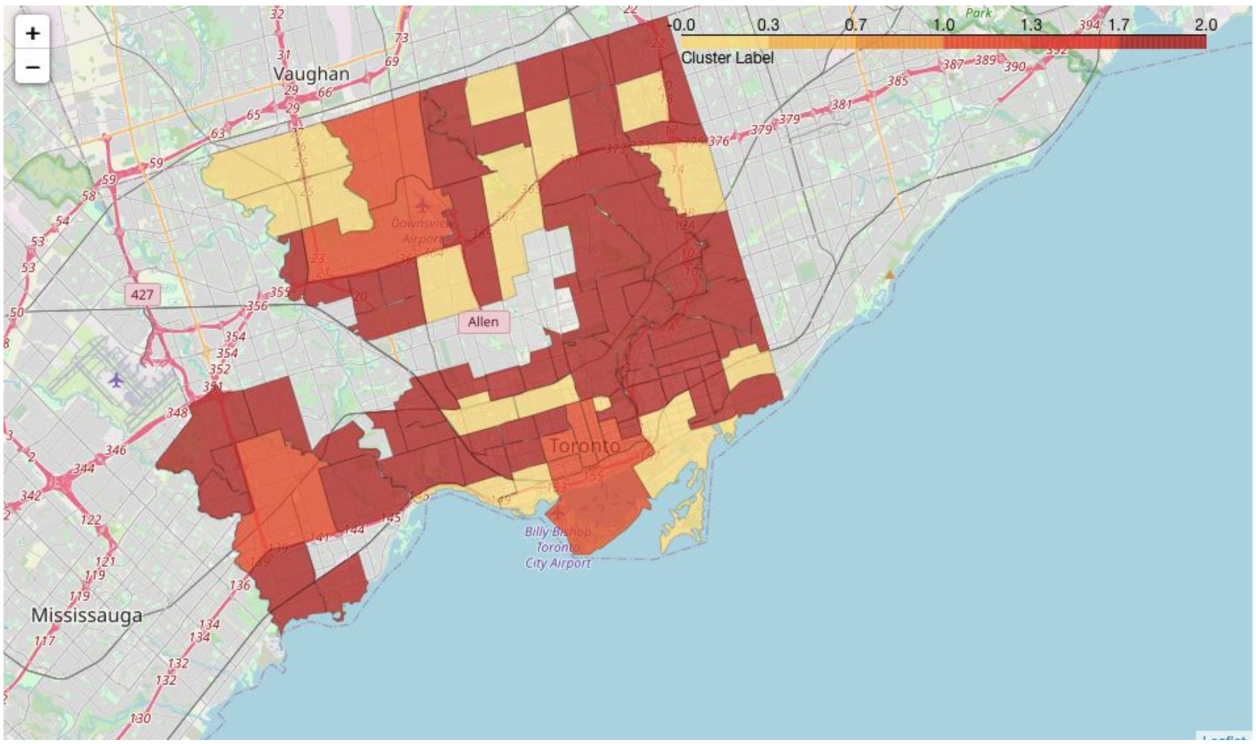

Figure 13 shows the distribution of clusters in the Toronto Neighborhoods

The 1st Cluster is symbolized by 0 (Yellow), the 2nd Cluster is symbolized by 1 (Light Red), and the 3rd Cluster is symbolized by 2 (Dark Red)

In this, the 1st and the 2nd Neighborhoods seem more crime filled. Here is the list of these neighborhoods

	Neighbourhood	Cluster Labels 2.0	Sum of All Crime
1	York University Heights	1	12.241419
2	Lansing-Westgate	0	0.845412
3	Yorkdale-Glen Park	0	4.956986
8	Islington-City Centre West	1	11.537141
11	South Parkdale	0	0.746515
12	South Riverdale	0	7.243345
15	Humber Summit	0	4.620477
16	Humbermede	0	0.157241
17	Church-Yonge Corridor	1	17.528770
25	Glenfield-Jane Heights	0	4.059758
29	Waterfront Communities-The Island	1	19.552487
36	Kensington-Chinatown	1	9.609661
40	Annex	0	5.910805
44	Dovercourt-Wallace Emerson-Junction	0	5.795886
45	Newtonbrook West	0	0.608929
46	Niagara	0	2.938970
49	North St.James Town	0	0.538221
58	Downsview-Roding-CFB	1	10.053878
59	Bay Street Corridor	1	17.975590
60	Black Creek	0	2.280945
61	Willowdale East	0	1.566674
66	East End-Danforth	0	1.342042
69	Parkwoods-Donalda	0	0.302695
74	Bedford Park-Nortown	0	0.155217
76	Don Valley Village	0	0.077357

Figure 14 shows a list of neighborhoods to avoid with their cluster label and sum of all crime

Keep in mind that the Sum of All Crime is a sum of all z-scores

Discussion

In this results section, the data is shown and the logical path from data to cluster is shown. In this path, we see that there are three different clusters of crime: high levels of crime, medium levels of crime, and low levels of crime. In this clustering system, the neighborhoods are categorized by their number of crimes from the years 2014-2019 and used to understand which neighborhoods to avoid. To first understand the distribution of crime in Toronto, boxplots were made. In the Assault distribution, we see that 50% of the data lied in the range of 59-160 assaults per year per neighborhood which also has a maximum of 851.8 and a minimum of 18.5 assaults per year per neighborhood. For Auto-Thefts, 50% of the data was in the range of 13-30 Auto Thefts per year with a minimum of 2.7 and a maximum of 30.975. All of these numerical summaries are shown above in the Results Section. Then, a cluster set was used to decipher which neighborhoods were the worst in terms of crime. In this clustering set, we can see that the ones with high and medium amounts of crime are located together and the neighborhoods with low levels of crime are located closely as well. With this cluster, we can see that the neighborhoods to avoid are the ones shown in Figure 8 and Figure 10. But, even with these neighborhoods, the number of Toronto neighborhoods with low amounts of crime are plentiful.

Then, after considering the fact that each crime is not normalized, we decided to create a k-means cluster based on the sum of all of the z-scores of the crimes in each neighborhood. This is because the z-score also gives an accurate representation of how the crime in this neighborhood relates to other neighborhoods. Since higher than average crime in a neighborhood would lead to a positive z-score and vice versa, a large sum of z-scores would indicate that the neighborhood is high in every single aspect of crime. Using this, we were able to get a second cluster and a second set of neighborhoods that strongly overlapped with the previous set of neighborhoods but differed in smaller ways. This second list can be seen in Figure 14

Conclusion

As shown in Figure 8, Figure 9, and Figure 10, the neighborhoods with medium, low, and high levels of crime are shown respectively. Using this data, a family or business can find a place of residence or business respectively if they are new to Toronto and do not know which neighborhoods are better or worse for crime. Using the data set of number of crimes in Toronto in each neighborhood per year, an understanding of which neighborhoods are bad or good in terms of crime could be ascertained. Using Figure 14 as well, a list can be made of which neighborhoods to specifically avoid in Toronto