Exploring Crime Occurrences and Neighbourhood Venues in Toronto

1. Introduction

1.1 Background

When I was planning a move to the city of Edmonton in the Canadian province of Alberta, I had difficulties determining the safety of its neighbourhoods. The online advice from locals appeared subjective and anecdotal. One contributor posited that any neighbourhood with a Walmart was unsafe.

To find a safe and comfortable neighbourhood to live in, I needed to visit the city and explore a number of different areas. I looked at homes and observed the surrounding neighbourhoods. I drove the back alleys behind the homes of interest. I even visited the area during at nighttime to observe any changes.

Building on the Walmart contributor's claim, I want to address a key question: can we predict the safety of a neighbourhood by the type of venues in it?

1.2 Problem

Is there a correlation between the number of crime occurrences and the types of venue in a neighbourhood?

If we knew that the neighbourhoods with certain venues or combinations of venues correlated to crime rates, it would give us a rule of thumb to judge the safety of neighbourhoods.

1.3 Usefulness

This information would very likely be useful to people who are moving to a new city. They could narrow their search for neighbourhoods to live in, and unlike me, could save money and use their time more efficiently.

This information may also help tourists to avoid areas that are less safe.

City officials may find it useful, as well. When planning zones for a neighbourhood, they could avoid certain combinations of venues that correlate with higher crime, or inversely, plan for venues that correlate with lower crime.

2. Data Acquisition and Cleaning

2.1 Data Sources

I used a .csv file from the Toronto Police Service Public Safety Data Portal site (which contains information licensed under the Open Government Licence – Ontario) for crime data by neighbourhood. It provides data on serious crime (assault, auto theft, break and enter, robbery, theft over $5,000 and homicide) by neighbourhood for the years 2014 to 2018.

To collect venue data on each of the neighbourhoods, I used the coordinates of Toronto neighbourhoods and the Foursquare API to find a maximum of 30 venues within each neighbourhood.

2.2 Crime Data Cleaning

From the Toronto Police Service crime data, I extracted the serious crime occurrences for each neighbourhood from the year 2018. For columns, I use the crime types (assault, auto theft, break and enter, robbery, theft over $5,000 and homicide), and each row represents a neighbourhood. One data entry was missing in one of the neighbourhoods, and so I dropped that row (neighbourhood) from the data.

The column representing theft over $5,000 was uploaded as a python float type, and needed to be changed it to type int. I also created a new column "Crime_Occurrences" that summed all the serious crime numbers (see Figure 1). The resulting data consisted of 139 neighbourhoods (rows) and 8 columns (Figure 1).

	Neighbourhood	Assault_2018	Auto_Theft_2018	BreakandEnter_2018	Robbery_2018	Theft_Over_2018	Homicide_2018	Crime_Occurrence
0	Yonge-St.Clair	61	69	23	19	3	0	175
1	York University Heights	138	23	52	15	2	1	231
2	Lansing-Westgate	197	22	52	41	6	0	318
3	Yorkdale-Glen Park	127	28	56	35	13	2	261
4	Stonegate-Queensway	128	41	41	36	4	0	250
5	Tam O'Shanter-Sullivan	56	46	18	11	2	1	134
6	The Beaches	457	22	236	78	30	0	823
7	Thistletown-Beaumond Heights	30	25	25	8	7	2	97
8	Thorncliffe Park	135	114	60	42	15	0	366
9	Danforth East York	227	156	115	33	54	0	585

Figure 1. Toronto Neighbourhoods and Crime Occurrences (1st ten rows)

I then used the geopy library and geocoders attribute to find the latitudes and longitudes of the neighbourhoods and added them to the data. Geocoders could not provide coordinates for 34 neighbourhoods, and these were removed.

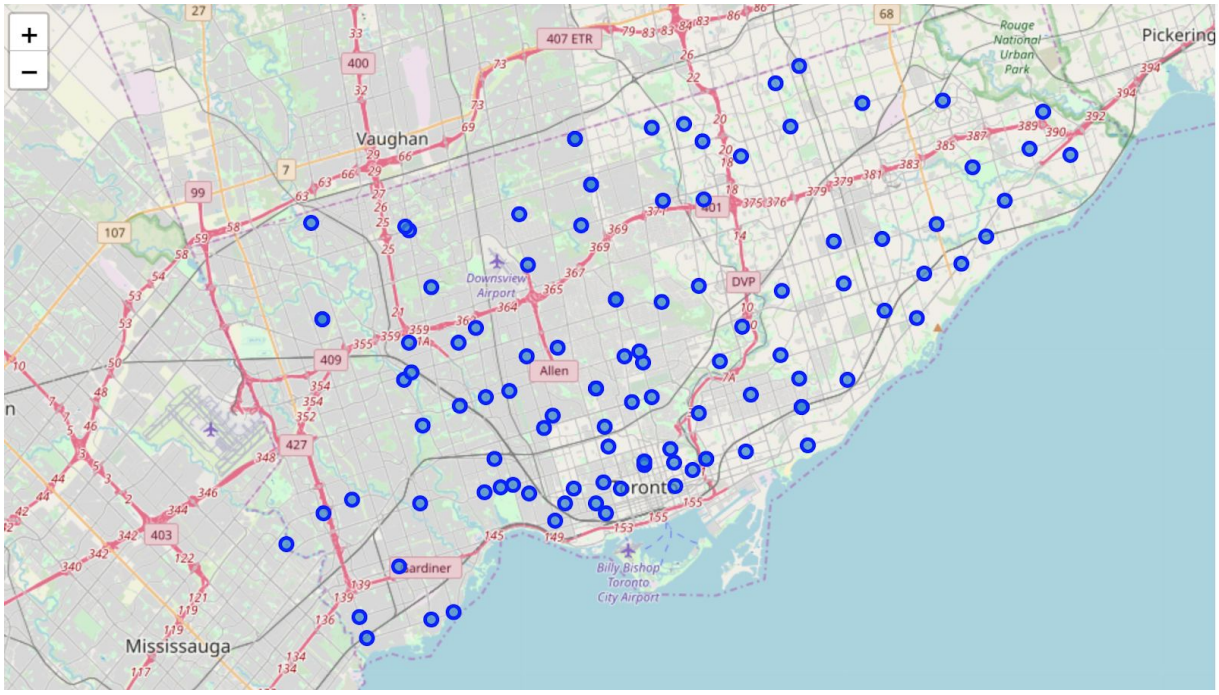
I sorted the data in descending order of crime occurrences so that the neighbourhood with the highest numbers appears at the top and the neighbourhood with the lowest crime is at the bottom. Lastly, I added a “Crime_Rank” column to reflect the relative ranking of crime for neighbourhoods in Toronto. The first row, with “Crime_Rank” of 1, represents the neighbourhood with the highest crime occurrences in 2018; the last row, with “Crime_Rank” of 105, represents the neighbourhood with the lowest crime occurrences in 2018.

The resulting data consists of 105 rows, or neighbourhoods, and 10 columns (see Figure 2).

	Neighbourhood	Assault_2018	Auto_Theft_2018	BreakandEnter_2018	Robbery_2018	Theft_Over_2018	Homicide_2018	Crime_Occurrence	Latitude	Longitude	Crime_Rank
94	Black Creek	1005	79	221	224	46	4	1579	43.734634	-79.505355	1
24	Cliffcrest	787	58	314	93	50	1	1303	43.721939	-79.236232	2
27	Ionview	284	495	154	69	50	0	1052	43.735990	-79.276515	3
106	Palmerston-Little Italy	547	40	145	159	37	0	928	43.655854	-79.410116	4
6	The Beaches	457	22	236	78	30	0	823	43.671024	-79.296712	5
85	Bathurst Manor	404	109	83	64	14	0	674	43.763893	-79.456367	6
21	Centennial Scarborough	385	60	98	72	13	0	628	43.787491	-79.150768	7
103	Rexdale-Kipling	295	34	189	54	53	0	625	43.721362	-79.565513	8
14	Scarborough Village	340	77	96	85	22	0	620	43.743742	-79.211632	9
43	Kennedy Park	411	26	103	51	5	2	598	43.724878	-79.253969	10

Figure 2. Cleaned Data on Toronto Crime by Neighbourhood

Folium was used to visualize a map of the neighbourhoods of Toronto (see Map 1).



Map 1. The Neighbourhoods of Toronto

2.3 Venue Data Cleaning

With the Foursquare API and the coordinates from each neighbourhood, I collected a maximum of 30 neighbourhoods for each venue (see Figure 3).

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Black Creek	43.734634	-79.505355	KTX Insurance Brokers	43.735430	-79.501780	Insurance Office
1	Cliffcrest	43.721939	-79.236232	Dairy Queen	43.722758	-79.235625	Fast Food Restaurant
2	Cliffcrest	43.721939	-79.236232	Dairy Queen Ltd Brazier	43.722637	-79.235452	Fast Food Restaurant
3	Cliffcrest	43.721939	-79.236232	Canadian Tire	43.721005	-79.237790	Furniture / Home Store
4	Cliffcrest	43.721939	-79.236232	Pizza Pizza	43.722491	-79.235277	Pizza Place
5	Cliffcrest	43.721939	-79.236232	Big Boy's Burger	43.721751	-79.236301	Burger Joint
6	Cliffcrest	43.721939	-79.236232	Wild Wing	43.721102	-79.236665	Wings Joint
7	Cliffcrest	43.721939	-79.236232	I.D.A. - St. Clair & Brimley Pharmacy	43.721212	-79.242254	Pharmacy
8	Cliffcrest	43.721939	-79.236232	LCBO	43.725183	-79.232039	Liquor Store
9	Cliffcrest	43.721939	-79.236232	Dollarama	43.725136	-79.231929	Discount Store

Figure 3. Neighbourhoods and Venues (first 10 rows)

I then one-hot encoded the venues by neighbourhood and took the mean frequency of occurrence of each category to use for k-means clustering. The resulting data was 105 rows and 236 columns. Figure 4 below shows the first 5 neighbourhoods and 15 venues.

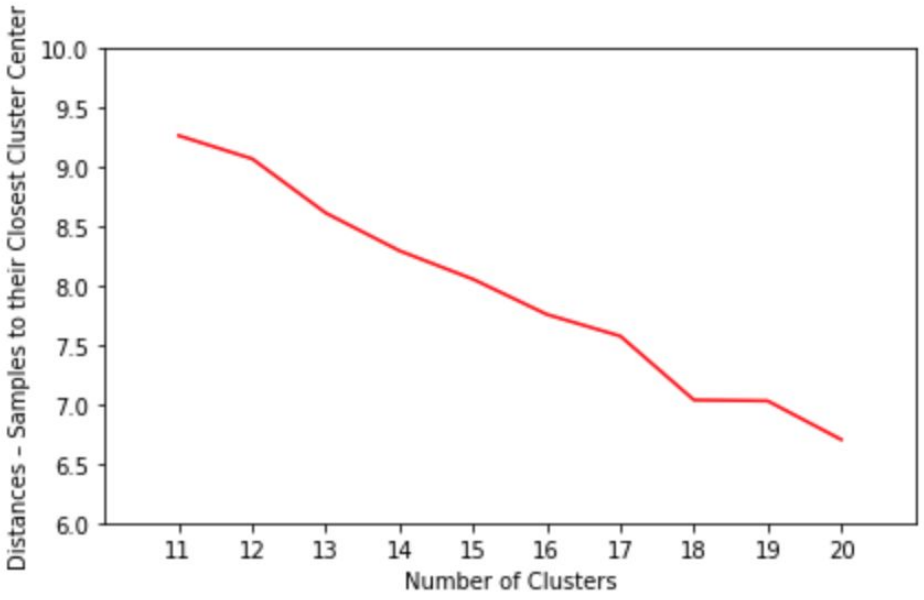
Neighbourhood	Adult Boutique	Afghan Restaurant	American Restaurant	Animal Shelter	Antique Shop	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	BBQ Joint	Bagel Shop	Bakery	Bank	Bar
Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.066667	0.033333	0.0
Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
Annex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.033333	0.000000	0.0
Banbury-Don Mills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
Bathurst Manor	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0

Figure 4. Neighbourhoods and Mean Frequency of Venues

3. Methodology

I ran k-means clustering on various cluster sizes to determine the most accurate cluster size to use with my data.

With 105 neighbourhood entries, I chose neighbourhood cluster sizes from 11 to 20, ran k-means clustering on the data and calculated the sum of squared distances of the samples to their closest cluster center (the error). I plotted the error and found the “elbow point”, the most accurate k-size, to be 18 (Graph 1).



Graph 1. The Elbow Point

Then I ran k-means with a cluster size of 18 to place the neighbourhoods in 18 different groups. In order to view the neighbourhood crime rankings and cluster groups, I added the cluster labels to the data (Figure 5).

hbourhood	Assault_2018	Auto_Theft_2018	BreakandEnter_2018	Robbery_2018	Theft_Over_2018	Homicide_2018	Crime_Occurrence	Latitude	Longitude	Crime_Rank	Cluster Labels
Black Creek	1005	79	221	224	46	4	1579	43.734634	-79.505355	1	1
Cliffcrest	787	58	314	93	50	1	1303	43.721939	-79.236232	2	8
Ionview	284	495	154	69	50	0	1052	43.735990	-79.276515	3	11
Palmerston-Little Italy	547	40	145	159	37	0	928	43.655854	-79.410116	4	12
he Beaches	457	22	236	78	30	0	823	43.671024	-79.296712	5	14
hurst Manor	404	109	83	64	14	0	674	43.763893	-79.456367	6	11
Centennial carborough	385	60	98	72	13	0	628	43.787491	-79.150768	7	1
dale-Kipling	295	34	189	54	53	0	625	43.721362	-79.565513	8	5
carborough Village	340	77	96	85	22	0	620	43.743742	-79.211632	9	1
nnedy Park	411	26	103	51	5	2	598	43.724878	-79.253969	10	4

Figure 5. Neighbourhood Crime Data and Cluster Groups

4. Results

The neighbourhoods are not evenly distributed within the cluster groupings. 63.8% of the neighbourhoods (67/105) fall into two clusters: cluster 1 and cluster 11 (see Figure 6). Half of the clusterings, 9/18, consist of only one neighbourhood (Figure 6). This may reflect the lack of venue data in some cases. 31 of our neighbourhoods consist of 5 or fewer venues, which likely makes it difficult to cluster them. In turn, this may hamper our ability to find correlations between crime occurrences and venue types. However, let's look for patterns among the neighbourhoods with the highest and the lowest crime occurrences.

Cluster Labels	Neighbourhood
0	1
1	25
2	1
3	1
4	1
5	1
6	1
7	7
8	6
9	5
10	2
11	42
12	3
13	1
14	4
15	1
16	1
17	2

Figure 6. Number of Neighbourhoods in each Cluster

Of the 10 neighbourhoods with the highest crime occurrences, 3 of them fall into cluster group 1, and 2 of them fall into cluster group 11 (see Figure 5). However, the 10 neighbourhoods with the fewest crime occurrences all fall into the same 2 groups: 7 of them are in cluster 1 and 3 in cluster 11 (see Figure 7). In fact, the two neighbourhoods with the highest and lowest crime numbers fall both fall into the same cluster group (cluster=1). This rules out the notion that venues in neighbourhood clusters 1 and 11 are correlated to crime.

Crime_Rank	Cluster Labels
96	1
97	1
98	11
99	1
100	1
101	1
102	11
103	11
104	1
105	1

Figure 7. Crime and Cluster Groups for the Lowest-Crime Neighbourhoods

Examining the other the top 10 crime neighbourhoods, we can see they fall into five other cluster groups, each represented once: clusters 8, 12, 14, 5 and 4 (Figure 7). Clusters 4 and 5 consist of only one neighbourhood each, which is not enough to predict a pattern, and so I ruled them out.

Next, let's look more closely at the crime rankings of neighbourhoods in cluster groups 8, 12 and 14. Figure 8 (below) shows that only two neighbourhoods are among the highest 40 in crime occurrences, ranking 2nd and 39th. However, one of the neighbourhoods has low crime numbers – it ranks 91 out 105 Toronto neighbourhoods examined. This ambiguity makes it difficult to conclude that the neighbourhoods in cluster 8 have any correlation to crime occurrences.

Crime_Rank	Cluster Labels
2	8
39	8
41	8
44	8
56	8
91	8

Figure 8. Crime Rank and Cluster Group 8

In Figure 9 (below), we can see three neighbourhoods that rank 4, 36 and 77 in crime occurrences. This wide range suggests that the neighbourhoods in cluster group 12 are not particularly associated with crime, either.

Crime_Rank	Cluster Labels
4	12
36	12
77	12

Figure 9. Crime Rank and Cluster Group 12

Lastly, we can see that cluster group 5 has three neighbourhoods that rank in the top third of crime occurrences; however, it also has a neighbourhood that ranks 82/105 (see Figure 10). Thus, it is difficult to say decisively that the neighbourhoods in this cluster have venues that correspond to higher crime rates.

Crime_Rank	Cluster Labels
5	14
32	14
33	14
82	14

Figure 10. Crime Rank and Cluster Group 14

5. Discussion

I relied on geopy's geocoder to find the coordinates of each neighbourhood, which may not be strictly accurate. In Map 1 above, we can see neighbourhoods that are nearly overlapping.

A more nuanced exploration with better neighbourhood coordinates may yield greater clarity on the degree of correlation – or lack of correlation – that exists between crime rates and venue types.

Also, the k-means clustering method placed 63.8% (67/105) of the neighbourhoods into only 2 of 18 clusters (Figure 6). This may be due to the sparsity of venue data for some neighbourhoods.

An intra-city comparison of neighbourhoods with more venue types may help shed light on this issue.

6. Conclusion

In this study, I explored the relationship between crime and venue types in Toronto neighbourhoods. I used data for a variety of serious crimes (assault, auto theft, break and enter, robbery, theft over $5,000 and homicide numbers), along with the type of venues that are in each neighbourhood. I then used k-means to cluster the neighbourhoods and examine similarities between neighbourhoods and crime rates.

If a correlation existed, it would be useful for people moving to or travelling in a new city, as well as to city planners.

However, in conclusion, I found that neighbourhoods with similar venues do not share similar crime rates. I feel relatively safe to conclude that there is not any correlation between neighbourhood venues and crime occurrences.