Capstone Project – Battle of Neighborhoods

May 3

2020

Srihari G K

Table of Contents

Table of Figures

Capstone Project
Battle of Neighborhoods

Problem Description

Analysis of "Neighborhood" on "Crime" in the city of Toronto.

Compare Neighborhoods in the city of Toronto with the help of Crime statistics to find the best place which has lower or higher crime rates and to understand the effect of other venues on crime rate using Foursquare API.

Introduction and business problem selection

Toronto is the provincial capital of Ontario and has become the most populated city in Canada with a population of more than 3 million as per 2019[1]. By far Toronto is the fourth largest city in North America and has the lowest homicide rate which fluctuated between 2.1 to 3.1 per 100,000 people over 2010's decade. Although the crime rates are pretty less compared to other major North American cities, there are large criminal organizations which are operating in the Toronto region since at least the mid-19th century. Crime is Toronto has mostly been the domain of international crime syndicates. [2]

The crime data in Toronto has been published by the "Toronto Police Service" in a "**Public Safety Data Portal**" [3]. And all the crimes in Toronto from 2014 – 2019 has been published in the portal here. [4] The venue details can be obtained from the **"Foursquare API"**[5]. First the crime data is overlapped on the Toronto map using **Folium Library** to see the occurrences of the crimes and using Foursquare API data the number of venues are mapped. And later both the maps are compared for correlation.

This project tries to analyze the patterns of crime, like what are categories of crimes, their occurrence based on time of the day, day of the week, or month of the year and get more insights into the data. This also explores a correlation between the presence of venues to number of crimes. This analysis tries to solve the common notion which exists that the presence of number of venues in the neighborhood makes the area busy and due to presence of lot of traffic and people around the venues, the crimes may be less.

The results of this analysis can help the residents and the visitors of Toronto about the crime locations near the venues and take appropriate precautions.

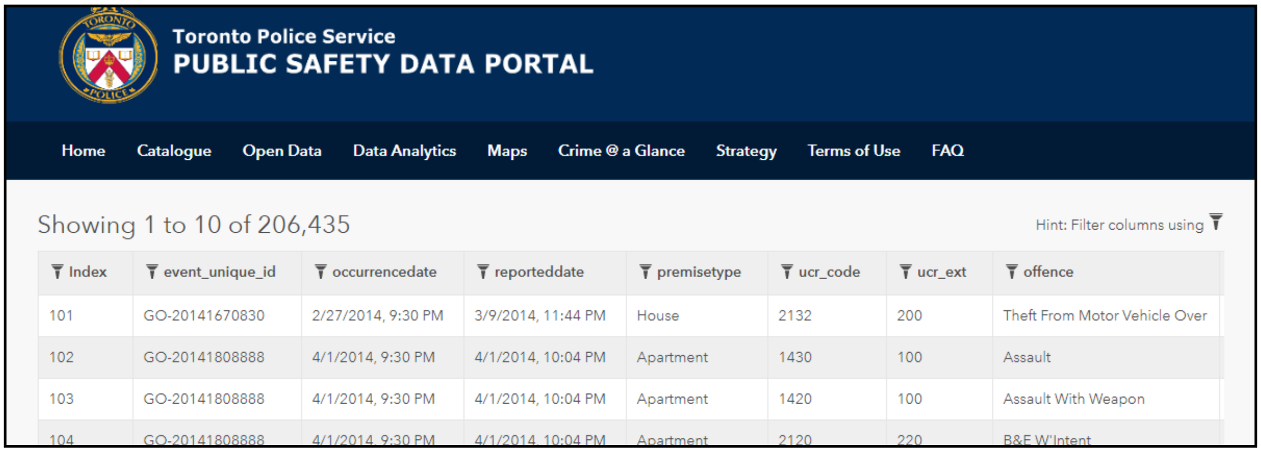
---

1 https://en.wikipedia.org/wiki/Toronto
2 https://en.wikipedia.org/wiki/Crime_in_Toronto
3 http://data.torontopolice.on.ca/
4 http://data.torontopolice.on.ca/datasets/mci-2014-to-2019/data
5 https://foursquare.com/user

DATA

Data Description

The data in "**Public Safety Data Portal"** is available is .csv, geojson format for public use. .CSV file format will be used here in the analysis as it represents the real-time data and will be able to make this data analysis viable for a long time. The typical data set in the Police portal looks like the one in **Figure 1 : Toronto city- Crime data set**.



Index	event_unique_id	occurrencedate	reporteddate	premisetype	ucr_code	ucr_ext	offence
101	GO-20141670830	2/27/2014, 9:30 PM	3/9/2014, 11:44 PM	House	2132	200	Theft From Motor Vehicle Over
102	GO-20141808888	4/1/2014, 9:30 PM	4/1/2014, 10:04 PM	Apartment	1430	100	Assault
103	GO-20141808888	4/1/2014, 9:30 PM	4/1/2014, 10:04 PM	Apartment	1420	100	Assault With Weapon
104	GO-20141808888	4/1/2014, 9:30 PM	4/1/2014, 10:04 PM	Apartment	2120	220	B&E W'Intent

Figure 1 : Toronto city- Crime data set


```
Out[12]: [{'type': 'Feature',
  'properties': {'Index_': 7801,
   'event_unique_id': 'GO-20152165447',
   'occurrencedate': '2015-12-18T03:58:00.000Z',
   'reporteddate': '2015-12-18T03:59:00.000Z',
   'premisetype': 'Commercial',
   'ucr_code': 1430,
   'ucr_ext': 100,
   'offence': 'Assault',
   'reportedyear': 2015,
   'reportedmonth': 'December',
   'reportedday': 18,
   'reporteddayofyear': 352,
   'reporteddayofweek': 'Friday    ',
   'reportedhour': 3,
   'occurrenceyear': 2015,
   'occurrencemonth': 'December',
   'occurrenceday': 18,
   'occurrencedayofyear': 352,
   'occurrencedayofweek': 'Friday    ',
   'occurrencehour': 3,
   'MCI': 'Assault',
   'Division': 'D14',
   'Hood_ID': 79,
   'Neighbourhood': 'University (79)',
   'Long': -79.4052277,
   'Lat': 43.6569824,
   'ObjectId': 7001},
  'geometry': {'type': 'Point', 'coordinates': [-79.4052277, 43.6569824]}}]
```


Figure 2 : Geojson data of crimes in Toronto

The data set when imported from the portal as Geojson file and explored, the output looks like the one in **Figure 2 : Geojson data of crimes in Toronto**. The data represents the type of crime, reported and occurrence date/month/time, Longitude and Latitude of the crime location.

Index_	event_unique_id	occurrencedate	reporteddate	premisetype	ucr_code	ucr_ext	offence	reportedyear	reportedmonth	reportedday	reporteddayofyear
7801	GO-20152165447	2015-12-18T03:58:00.000Z	2015-12-18T03:59:00.000Z	Commercial	1430	100	Assault	2015	December	18	352
7802	GO-20151417245	2015-08-15T21:45:00.000Z	2015-08-17T22:11:00.000Z	Commercial	1430	100	Assault	2015	August	17	229
7803	GO-20151421107	2015-08-16T16:00:00.000Z	2015-08-18T14:40:00.000Z	Apartment	2120	200	B&E	2015	August	18	230

occurrencedayofweek	occurrencehour	MCI	Division	Hood_ID	Neighbourhood	Long	Lat
Friday	3	Assault	D14	79	University (79)	-79.4052277	43.6569824
Saturday	21	Assault	D42	118	Tam O'Shanter-Sullivan (118)	-79.3079071	43.7787323
Sunday	16	Break and Enter	D43	137	Woburn (137)	-79.225029	43.7659416

Figure 3 : CSV data of crimes in Toronto

Figure 3 : CSV data of crimes in Toronto shows the .CSV file format which consists of the same data sets used in the JSON file

The longitude and latitude fields of the crime dataset will be used to see the number of venues within 500 meters of the locations using Foursquare API.

```
results = requests.get(url).json()
results
```


```
{'meta': {'code': 200, 'requestId': '5e915efeed78b8001b03b8d6'},
 'response': {'headerLocation': 'Corktown',
  'headerFullLocation': 'Corktown, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 45,
  'suggestedBounds': {'ne': {'lat': 43.6587599045, 'lng': -79.3544279001486},
   'sw': {'lat': 43.6497598955, 'lng': -79.36684389985142}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '54ea41ad498e9a11e9e13308',
       'name': 'Roselle Desserts',
       'location': {'address': '362 King St E',
        'crossStreet': 'Trinity St',
        'lat': 43.653446723052674,
        'lng': -79.3620167174383,
        'labeledLatLngs': [{'label': 'display',
          'lat': 43.653446723052674,
          'lng': -79.3620167174383}],
        'distance': 143,
```


Figure 4: Four square API data for identified location data

And the foursquare API data looks like the one in **Figure 4: Four square API data for identified location data**. This data gives the venue name, address, latitude and longitude and distance from the requested locations. The data gathered from the Foursquare API will be compared with crime data and correlation analysis will be carried out.

```
In [13]: ColumnNames = ['offence','reportedyear','reportedmonth','reportedday','reporteddayofyear', 'reporteddayofweek','reportedhour','o
ccurrenceyear','occurrencemonth','occurrenceday','occurrencedayofyear','occurrencedayofweek','occurrencehour','MCI','Divisio
n','Hood_ID','Neighbourhood','Longitude','Latitude']
```

Figure 5 : Relevant columns extracted for Analysis

Out of all the data obtained from the portal, only few fields will be input into the data frame for further analysis as shown in **Figure 5 : Relevant columns extracted for Analysis.**

Though initially all the column names shown in Figure 5 were selected for analysis, some information were deemed to be redundant and only the column names mentioned in the **Table 1** were selected.

Column/ Data Names	Data Description
'premisetype'	Describes the location of the Crime like house, apartment, outside etc.
'occurrenceyear’	Consists the year data when the crime had occurred
'occurrencemonth’	Consists the month data when the crime had occurred
‘occurrenceday’	Consists the day data when the crime had occurred
'occurrencedayofyear'	Consists the day data in terms of 365 days when the crime occurred
'occurrencedayofweek'	Consists the day data in terms of 55 weeks when the crime occurred
'occurrencehour'	Consists the hour data when the crime occurred
'MCI'	These are 5 categories of Crime recorded : Assault, Break and Enter, Auto Theft, Robbery and Theft
'Division'	Toronto map is divided into 55 divisions and this data covers the division where the crime had occurred.
'Hood_ID'	Toronto map is divided into 140 hoods and this data covers the division where the crime had occurred.
'Neighbourhood'	It is the name corresponding to Hood ID
‘Long’	Longitude location of the crime
‘Lat’	Latitude location of the crime

Table 1: Selected data for Analysis

Methodology

Data Wrangling

Downloading the dataset and converting into a pandas dataframe

Geojson files are very difficult to convert into a dataframe because of more than 20000 rows of data and limited computer capabilities. That's the reason .CSV file has been utilized to frame a data set. All the column names shown in **Table 1** is extracted into a dataframe as shown in Figure 1**Figure 6.**

```
TorontoCrimeData.head()
```

t[36]:

premisetype	occurrenceyear	occurrencemonth	occurrenceday	occurrencedayofyear	occurrencedayofweek	occurrencehour	MCI	Division	Hood_ID	Neighbourhood	Long	La
House	2014.0	February	27.0	58.0	Thursday	16	Theft Over	D53	101	Forest Hill South (101)	-79.417687	43.70056
Apartment	2014.0	April	1.0	91.0	Tuesday	16	Assault	D41	121	Oakridge (121)	-79.278397	43.70577
Apartment	2014.0	April	1.0	91.0	Tuesday	16	Assault	D41	121	Oakridge (121)	-79.278397	43.70577
Apartment	2014.0	April	1.0	91.0	Tuesday	16	Break and Enter	D41	121	Oakridge (121)	-79.278397	43.70577
House	2014.0	April	1.0	91.0	Tuesday	12	Theft Over	D32	34	Bathurst Manor (34)	-79.460182	43.76578

```
[5]: #Gives the dataframe size.

TorontoCrimeData.shape
ut[5]: (206435, 13)
```

```
In [6]: #To verify the types of data in the dataframe

TorontoCrimeData.dtypes

Out[6]: premisetype              object
        occurrenceyear          float64
        occurrencemonth          object
        occurrenceday           float64
        occurrencedayofyear     float64
        occurrencedayofweek      object
        occurrencehour            int64
        MCI                      object
        Division                 object
        Hood_ID                   int64
        Neighbourhood            object
        Long                    float64
        Lat                     float64
        dtype: object
```

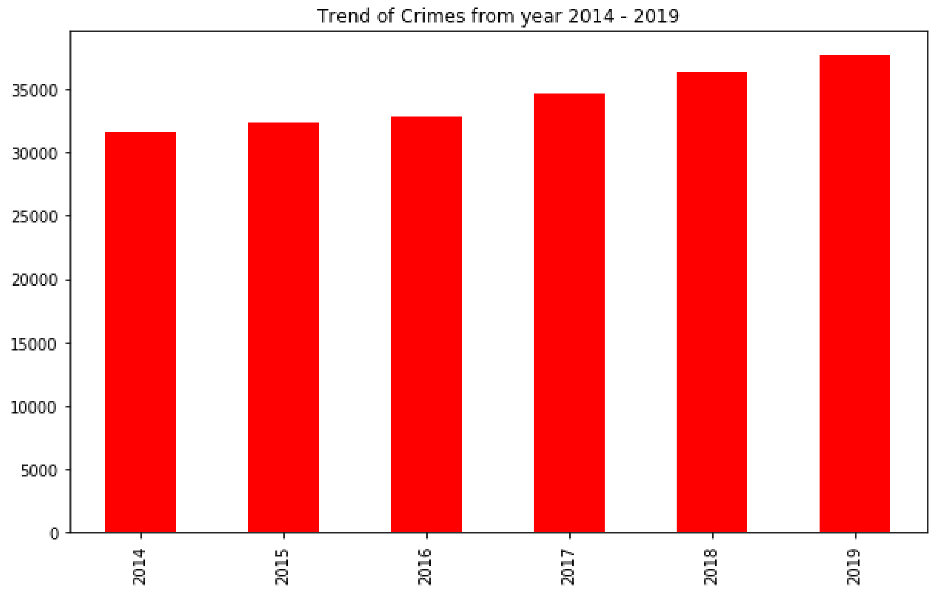
Figure 6: CSV data downloaded into a dataframe

As seen in the figure above the dataframe consists of **206435 Rows and 13 Columns.** Figure 6 also shows the types of data types in the data set. The "occurrence year", "occurrenceday" column data are type casted into "integer" from "float" variable type.

Data Analysis

Crime Trend from year 2014-2019

The data is analyzed for the increase in total number of crimes from the year 2014 to 2019. Although the dataset has crimes from 2006, it is ignored because they are very less compared to other years.



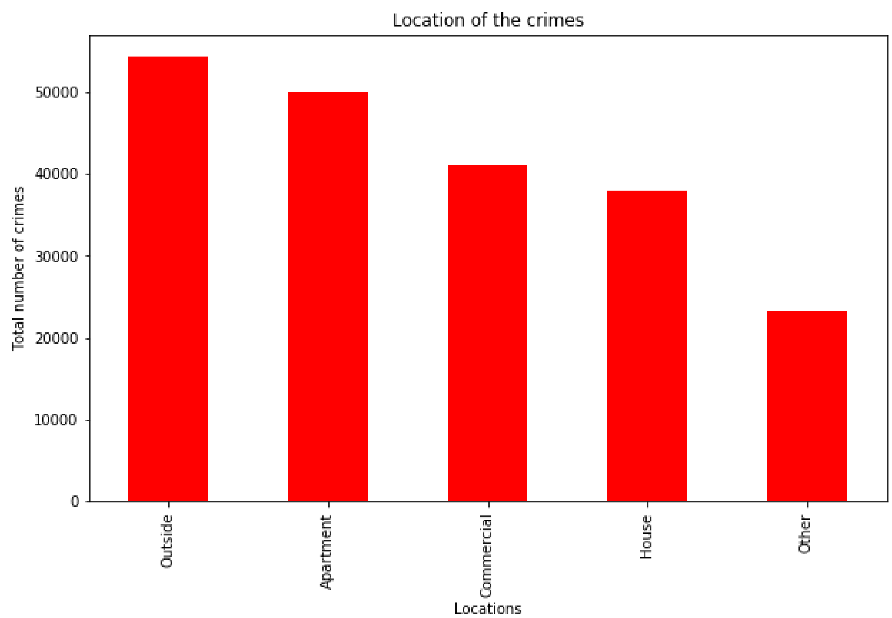


Figure 7 : Trend of Crimes from year 2014-2019

From above graph we can see that the crime rate has been increasing on a yearly basis and most of the crime happens outside followed by apartments and commercial establishments. This does not give a clear picture of the nature of the crime. Hence, the crimes are divided into nature of crimes in the following section.

Types of Crimes

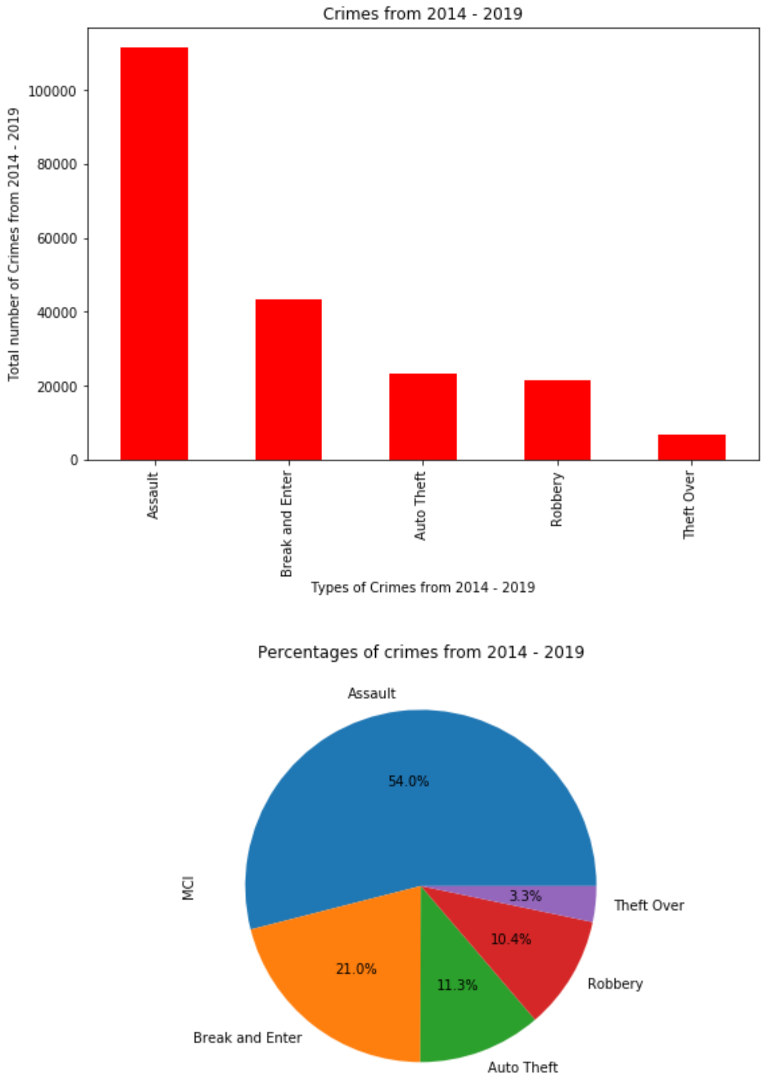


Figure 8 : Types of Crime from 2014-2019

From Figure 8, we can see that major type of crime which has been committed in the city of Toronto is “Assault” amounting to about 54% and “Break and Enter” amounts to 21% of the Total crimes. This is achieved by converting the “MCI” column in the main data frame into a dummy variable dataframe and then number of counts data is extracted and plotted using Matplotlib libraries.

Crime distribution as per week.

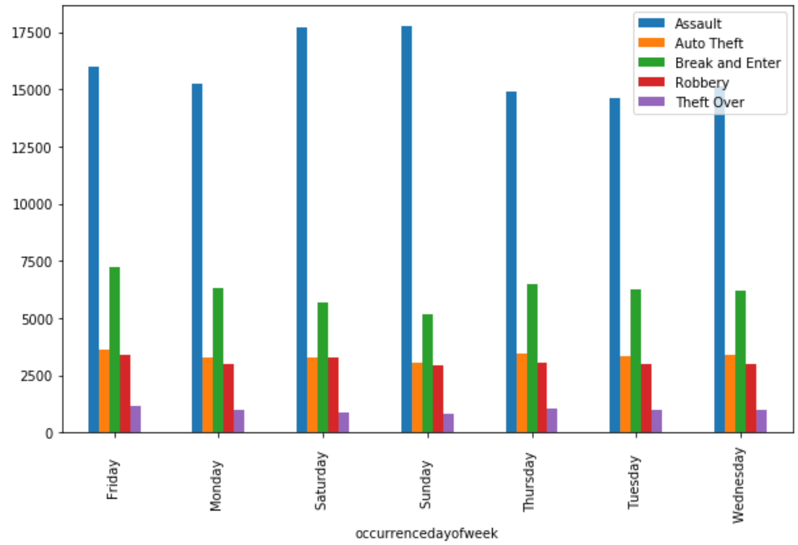


Figure 9 : Weekly distribution of Crime

The above is graph is obtained by merging the "MCI" dummy variable data frame with "occurrencedayofweek" . This new data frame is plotted with occurrence day of the week on the X axis and Total number of Crimes on the Y-axis. As shown in Figure 9. The crime rate drastically increases on weekends particularly on Sunday. From the above graph we can see that **"Assault"** which is most committed crime in Toronto is highest on Weekends approximately **22% increase** particularly on Sundays. But interesting fact is, the second biggest crime which is **"Break and Enter"** decreases on weekends by approximately **40%.**

Crime distribution as per time of the day

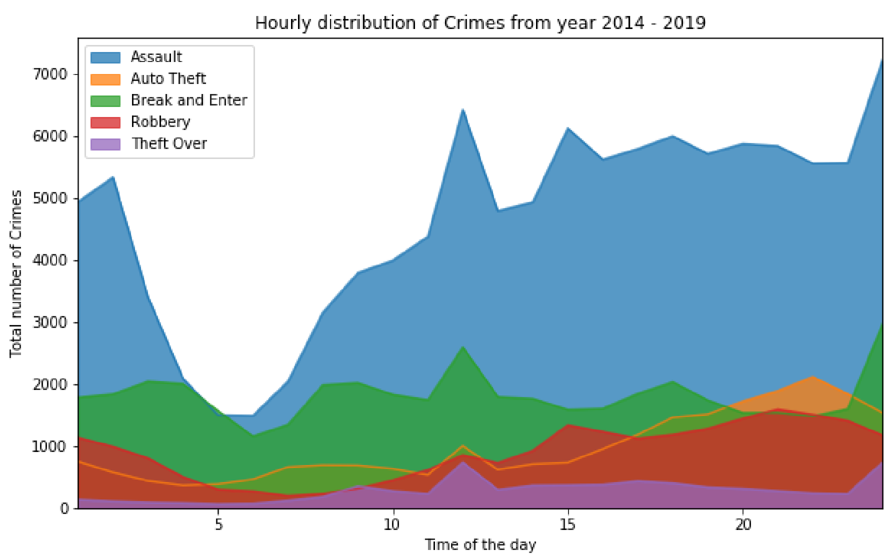


Figure 10: Hourly Distribution of Crime

As seen in the Figure 10, all the Crime are highest at the midnight and crime rate from morning 5-10 am is the least. Although a sudden spike in the crime can be seen in mid afternoon.

Explore Neighborhoods using Four Square API

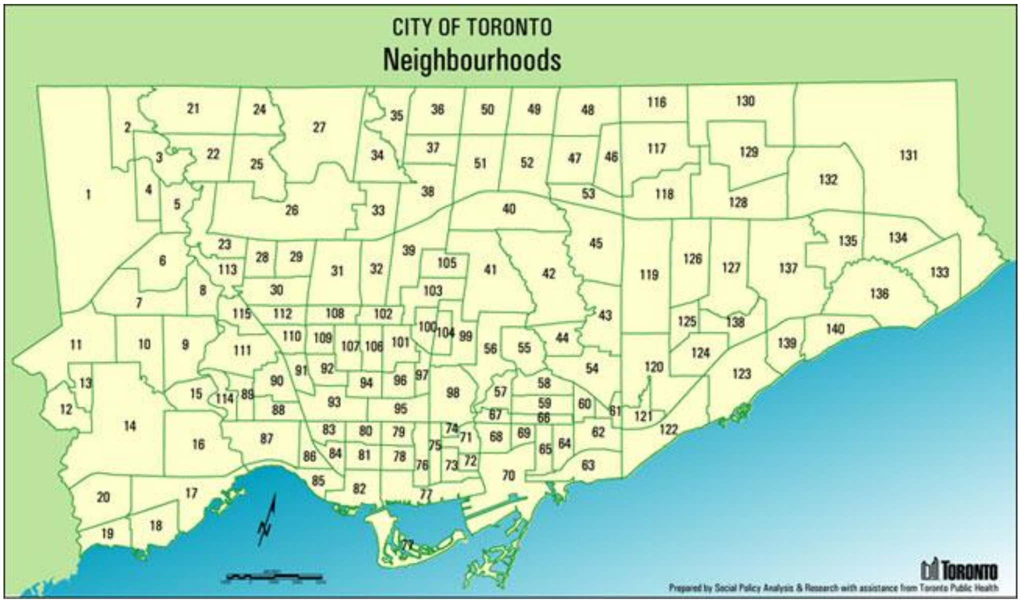


Figure 11: City of Toronto Neighborhoods

As shown in the above figure, city of Toronto is divided into 140 hoods and the same has been mentioned in the data set as “Hood_ID”. Each Hood_ID has location data mapped in the data set as latitude and longitude. Same data is mapped to the 140 divisions .

```
print(HoodCrime_lat_Long.shape)
HoodCrime_lat_Long.head()
```

```
(140, 8)
```

57]:

	Hood_ID	Assault	Auto Theft	Break and Enter	Robbery	Theft Over	Lat	Long
0	1	1811.0	2200.0	827.0	551.0	313.0	43.721487	-79.597169
1	2	1535.0	374.0	193.0	462.0	27.0	43.745418	-79.587672
2	3	322.0	152.0	114.0	90.0	14.0	43.738422	-79.566848
3	4	412.0	172.0	95.0	121.0	10.0	43.721058	-79.563743
4	5	327.0	113.0	63.0	81.0	9.0	43.721320	-79.550943

Figure 12: Aggregated data of crime as per Hoods

A consolidated data frame is developed having the sum of all crimes per crime category and per hoods. The location data in this frame is fed into the "Four Square API" as a URL and corresponding 100 venues in the radius of 500 meters are obtained.

```
In [23]: def getNearbyVenues(names, latitudes, longitudes, radius=500):

             venues_list=[]
             LIMIT =100
             for name, lat, lng in zip(names, latitudes, longitudes):
                 print(name)

                 # create the API request URL
                 url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
                     CLIENT_ID,
                     CLIENT_SECRET,
                     VERSION,
                     lat,
                     lng,
                     radius,
                     LIMIT)
```

```
In [24]: #Loop to find the neighborhood near all the identified Hood ids


         Toronto_venues = getNearbyVenues(names=HoodCrime_lat_Long['Hood_ID'],
                                          latitudes=HoodCrime_lat_Long['Lat'],
                                          longitudes=HoodCrime_lat_Long['Long']
                                           )
```

```
T_Count.head(10)
```

Out[26]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1	14	14	14	14	14	14
1	2	13	13	13	13	13	13
2	3	11	11	11	11	11	11
3	4	4	4	4	4	4	4
4	5	5	5	5	5	5	5
5	6	4	4	4	4	4	4
6	8	4	4	4	4	4	4
7	9	3	3	3	3	3	3
8	10	2	2	2	2	2	2
9	11	2	2	2	2	2	2

Figure 13: Data Frame consisting of Four Square API data

As shown in the Figure 13, credentials are sent to access the API . Function "getNearbyVenues' used in the course has been used to fetch the counts of the neighborhood. The venue data has been incorporated into the data frame. As shown in the above picture count of venues are mapped to corresponding Hood ID and fed into a Clustering Algorithm.

Clustering of Neighborhood

Initial Clustering using K-means algorithm

The hoods are clustered using **K-means clustering** to examine the effect of neighborhood on the Crime. It is expected from the algorithm that it divides the dataset into clusters depending on the total number of crimes. The data frame shown in Figure 14 is fed into the algorithm after normalizing. Normalization is a statistical method that helps mathematical-based algorithms interpret features with different magnitudes and distributions equally. **StandardScaler()** is used to normalize our dataset.

```
Toronto_collab1
```

Out[29]:

	Hood_ID	Assault	Auto Theft	Break and Enter	Robbery	Theft Over	Venue
0	1	1811.0	2200.0	827.0	551.0	313.0	14
1	2	1535.0	374.0	193.0	462.0	27.0	13
2	3	322.0	152.0	114.0	90.0	14.0	11
3	4	412.0	172.0	95.0	121.0	10.0	4
4	5	327.0	113.0	63.0	81.0	9.0	5

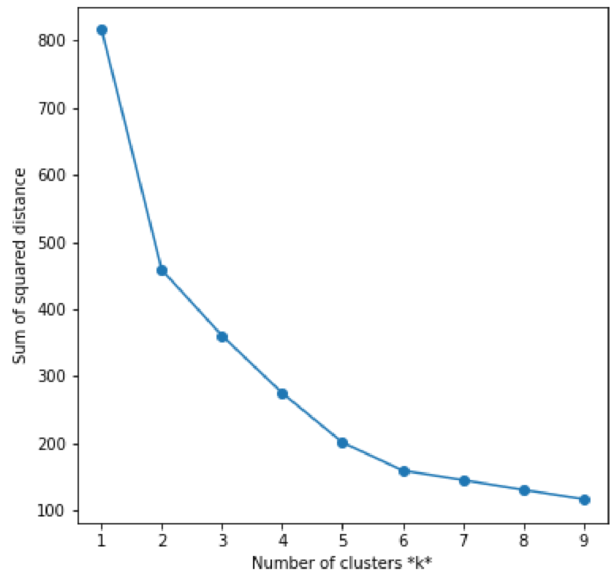
```
#As there are 5 types of crimes in Toronto, first iteration will be with 5 clusters.

num_clusters = 5

k_means = KMeans(init="k-means++", n_clusters=num_clusters, n_init=12)
k_means.fit(cluster_dataset)
labels = k_means.labels_

print(labels)
```

```
[4 2 0 0 0 0 0 0 0 0 0 0 2 0 0 2 0 0 0 2 0 0 2 2 2 2 0 0 0 2 0 0 0 0 0 0
 0 3 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 2 3 0 3 0 3 3 0 2 3 0 1 3 1 1 1 2
 3 3 3 2 3 3 3 0 0 0 0 3 0 0 2 3 2 0 3 0 0 3 0 0 0 3 3 0 0 0 0 0 0 0 2 0 0
 0 2 0 2 2 0 0 0 2 0 2 2 2 0 2 2 2 0 0 0 2 2 0 0 0]
```

Figure 14: Data frame used for K-means Clustering

Initially number of clusters was decided to be "5" corresponding to the categories of crime. But owing to incorrect clustering by the algorithm, it was decided to change the number of clusters. Elbow method is used to arrive at optimum number of clusters, which is explained in the further section.

Finding the right number of cluster with Elbow method




```
num_clusters = 2

k_means = KMeans(init="k-means++", n_clusters=num_clusters, n_init=12)
k_means.fit(cluster_dataset)
labels = k_means.labels_

print(labels)
```

```
[1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 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 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 1 1 1
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 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 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]
```

	Hood_ID	Assault	Auto Theft	Break and Enter	Robbery	Theft Over	Lat	Long	Venue	Total Crimes	Cluster Labels
0	01	1811.0	2200.0	827.0	551.0	313.0	43.721487	-79.597169	14	5703.0	1
1	02	1535.0	374.0	193.0	462.0	27.0	43.745418	-79.587672	13	2593.0	0
2	03	322.0	152.0	114.0	90.0	14.0	43.738422	-79.566848	11	695.0	0
3	04	412.0	172.0	95.0	121.0	10.0	43.721058	-79.563743	4	814.0	0
4	05	327.0	113.0	63.0	81.0	9.0	43.721320	-79.550943	5	598.0	0

Figure 15: Finding right Clusters using Elbow method

The right number of clusters was decided as per Elbow method. As seen in Figure 15, right number of clusters was decided as 2 and corresponding clusters were mapped to corresponding hoods. This data frame was used for correlation analysis.

Conclusion

The clusters identified from K-means algorithm has been evaluated in terms of Bar graph, Box plot and Regression plots. As shown in **Figure 16**, Bar graph shows the number of venues present in each cluster. From the graph it can be noted that Cluster number “1” has highest number of the venues than Cluster “0”, correspondingly in the box plot we can see that the total number of the crimes in the cluster “1” is way ahead of cluster “0”.

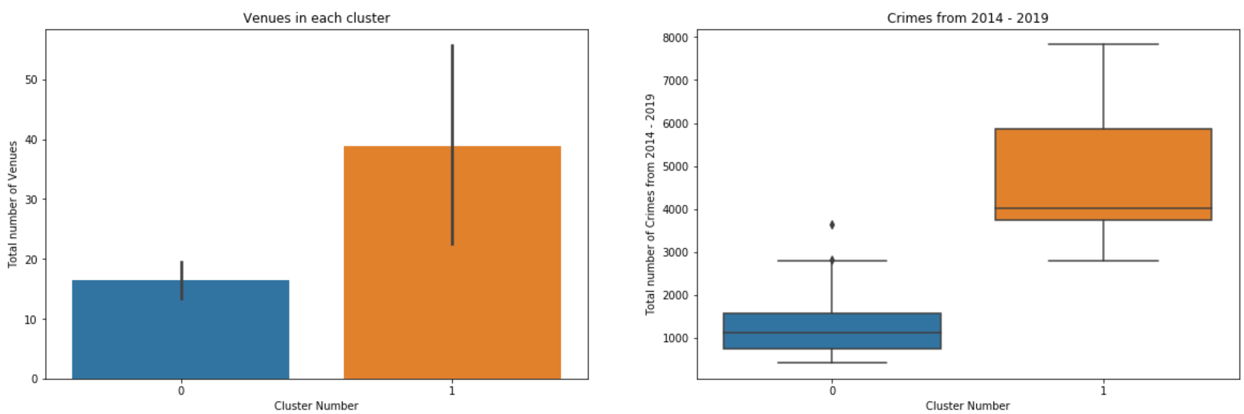


Figure 16: Bar graph and Box plots of cluster

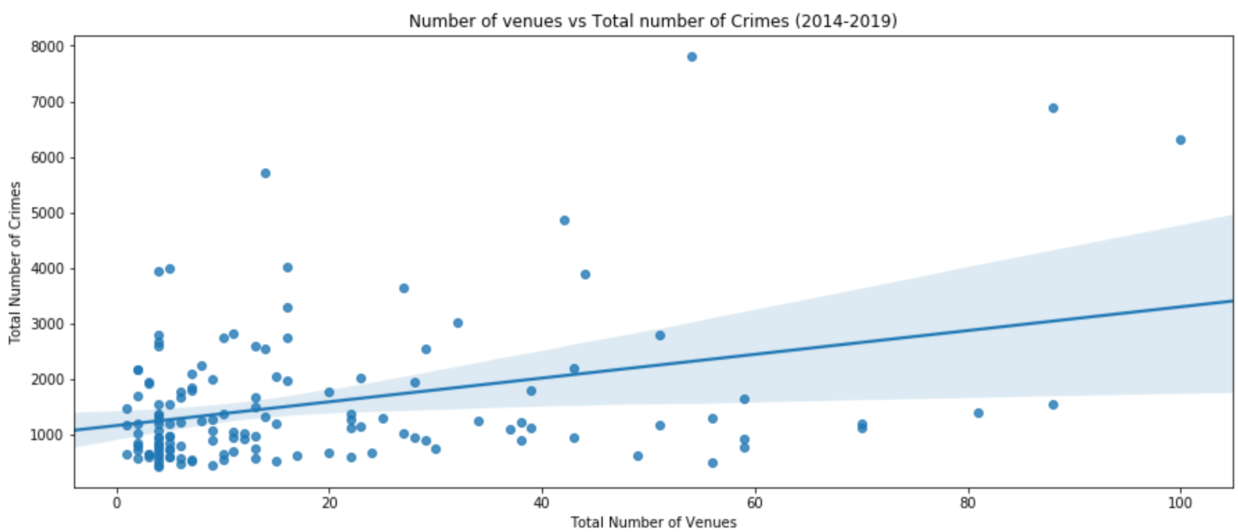


Figure 17: Effect of Venues on the Crimes

Taking clues from cluster “1” from the bar chart and box plot , regression graph was plotted using Seaborn library, where “Total Number of Venues” are plotted on the X –axis and “Total Number of Crimes are plotted on Y-axis. As we can see there is definitely a positive correlation which was further strengthened by the taking the Pearson’s correlation and p value.

The Pearson Correlation Coefficient came up as **0.3503682188715298** with a P-value of **2.8948696915466882e-05.**

From the Box plot, Regression plots, Pearson Correlation and P-value it can be confidently concluded that the **"<u>Total number of Crime increases as the number of venues present at the location</u>."**

However, the model can be further filtered based on the category of crimes near the venues. The analysis will act as a definite guide for people visiting the location to take appropriate precautionary actions.

Figure 18: Number of Crimes plotted on Toronto map

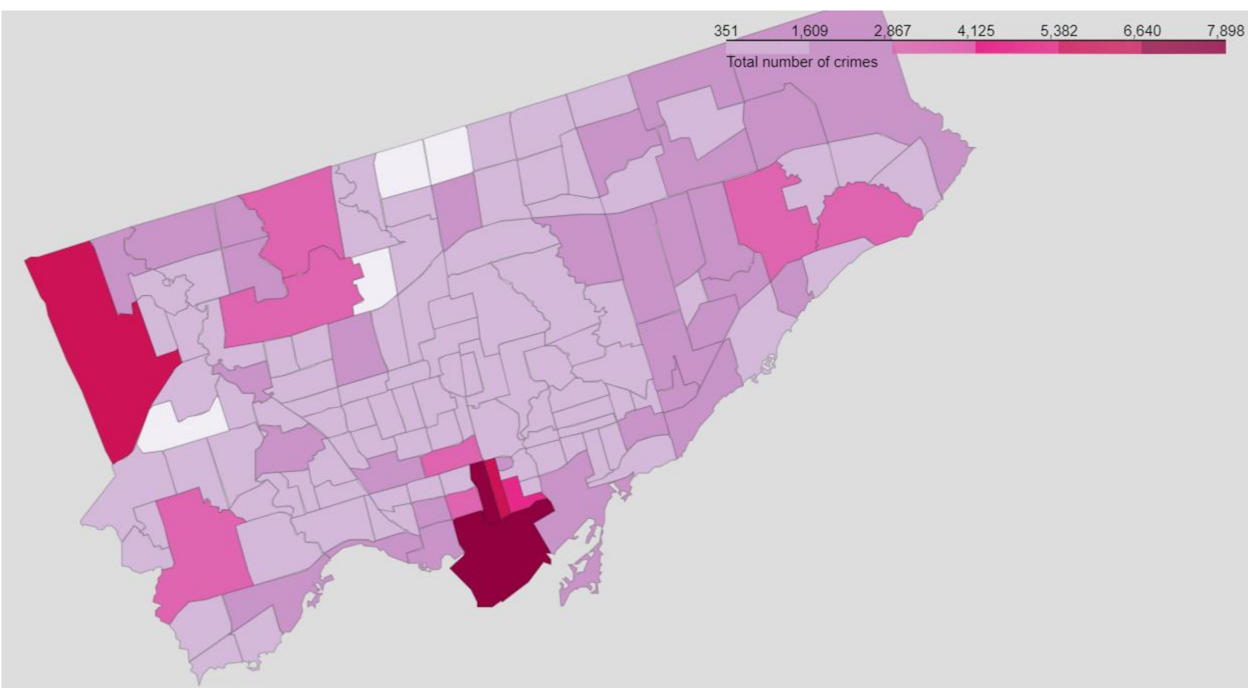


Figure 19: Intensity of Crime plotted on Toronto Map