CAPSTONE PROJECT
BATTLE OF NEIGHBORHOODS

-ANALYSIS OF “NEIGHBORHOOD” ON “CRIME” IN THE CITY OF TORONTO.

SRIHARI G K

Table of Contents

- Business Problem Introduction
- Description of Data set
- Methodology of Problem Solving
  - Data Wrangling
  - Data Analysis
  - Explore Neighborhood using Four Square API.
  - Clustering of Neighborhood
- Conclusion

Business Problem Introduction

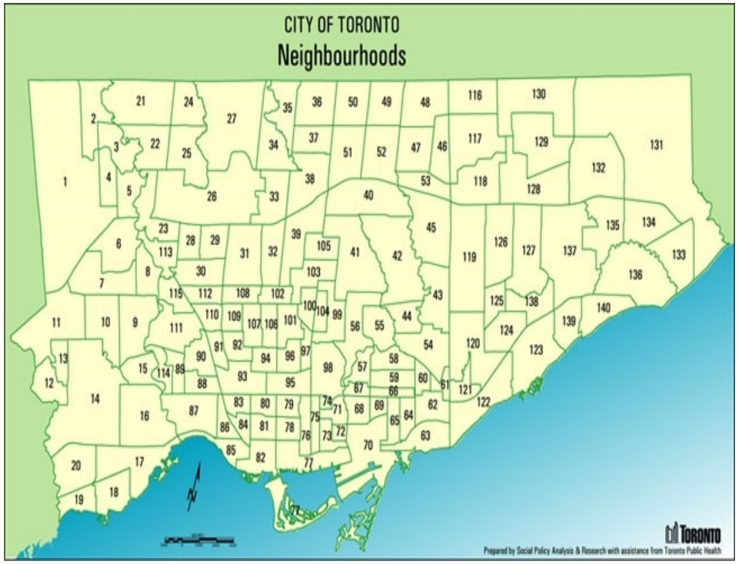
Have you any time felt that if the place is busy, like lot of restaurants etc, there would be lot of people and the place would be relatively safer than the quieter places?
Why take a Guess! Lets take a data driven decision!

Business Problem

This presentation explores the crime data in the Toronto neighborhood using Crime Statistics published by "Toronto Police"

In the same neighborhood the number of venues are explored using Four Square API.

Correlation between Number of Venues Vs Number of Crimes is correlated


CITY OF TORONTO
Neighbourhoods
TORONTO
Prepared by Social Policy Analysis & Research with assistance from Toronto Public Health

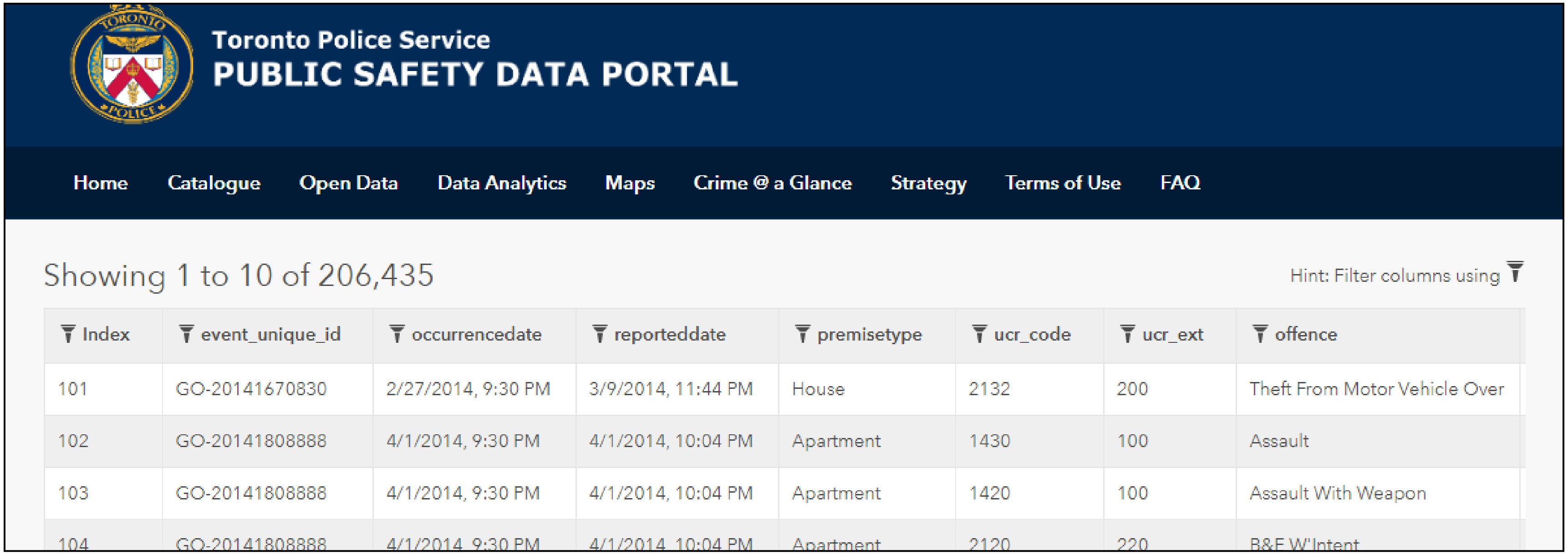
Results/Conclusion

Use of Solution

People visiting the places can take precautionary steps before visiting.

Source : https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

Description of Data set



Showing 1 to 10 of 206,435

Hint: Filter columns using

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

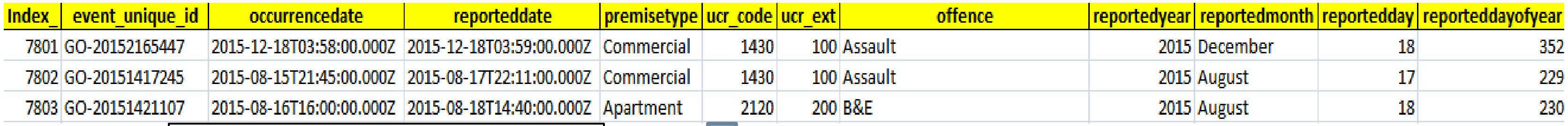
The data for analysis is taken from "**Public Safety Data Portal**"[1] is available is .csv, geojson format for public use. .CSV file format will be used here in the analysis.

In the portal, information about Type ,Occurrence of Crime Day/Week/Year, Latitude and Longitude data of the crime are published

[1] http://data.torontopolice.on.ca/datasets/mci-2014-to-2019/data

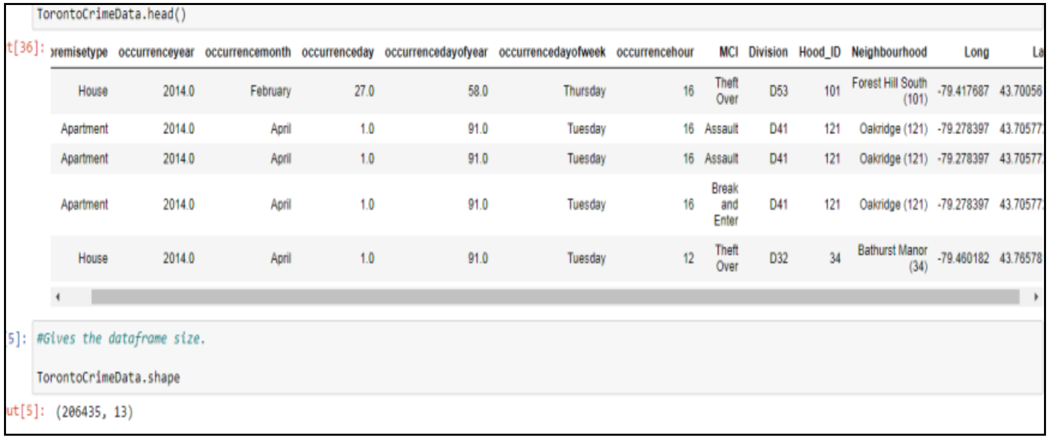
Methodology of Problem Solving

Data Wrangling : .CSV format with required columns are loaded into a data frame.



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

.CSV format of data



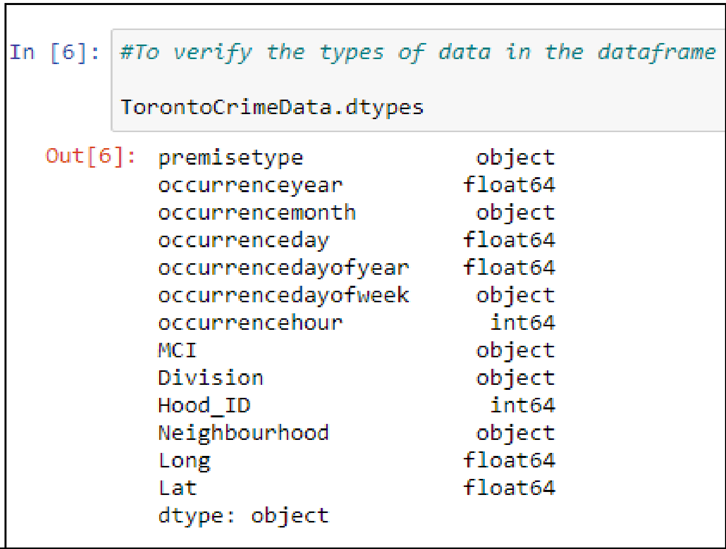
```
TorontoCrimeData.head()
```

t[36]:

remisetype	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)
```

Data Frame loaded into Pandas

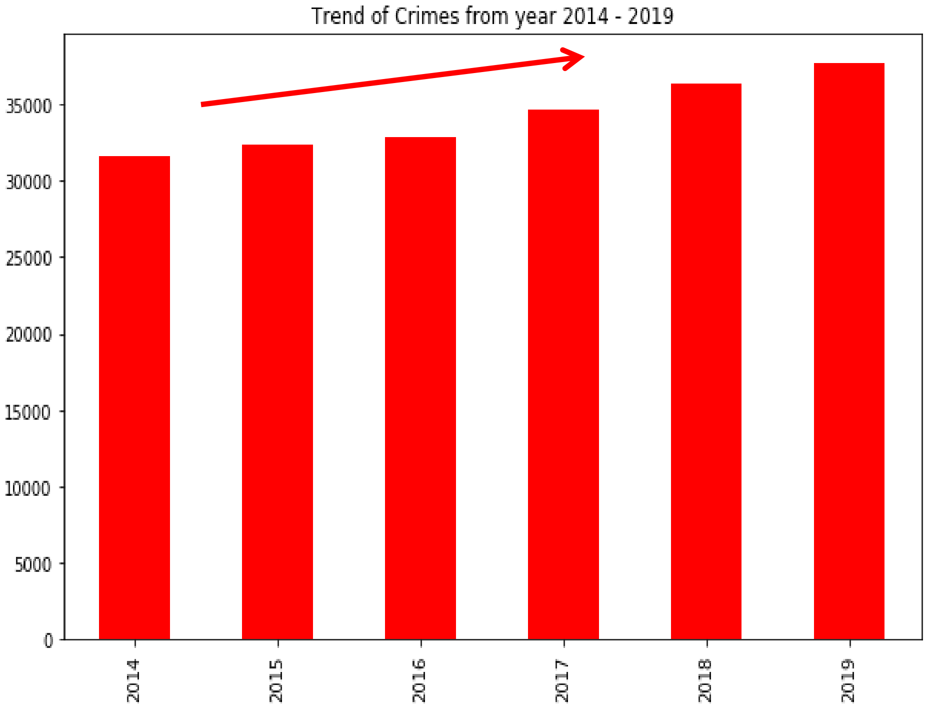


```
In [6]:  #To verify the types of data in the dataframe

         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
```

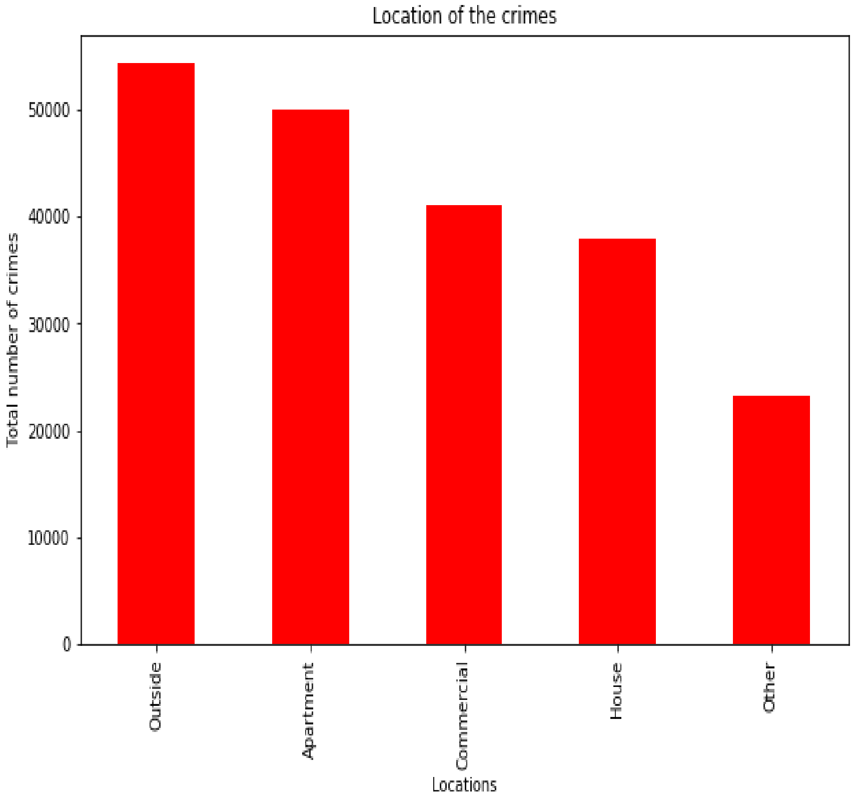
Data Frame is Analyzed for data types

Methodology of Problem Solving

Data Analysis- Crime trend year on year and location of Crime



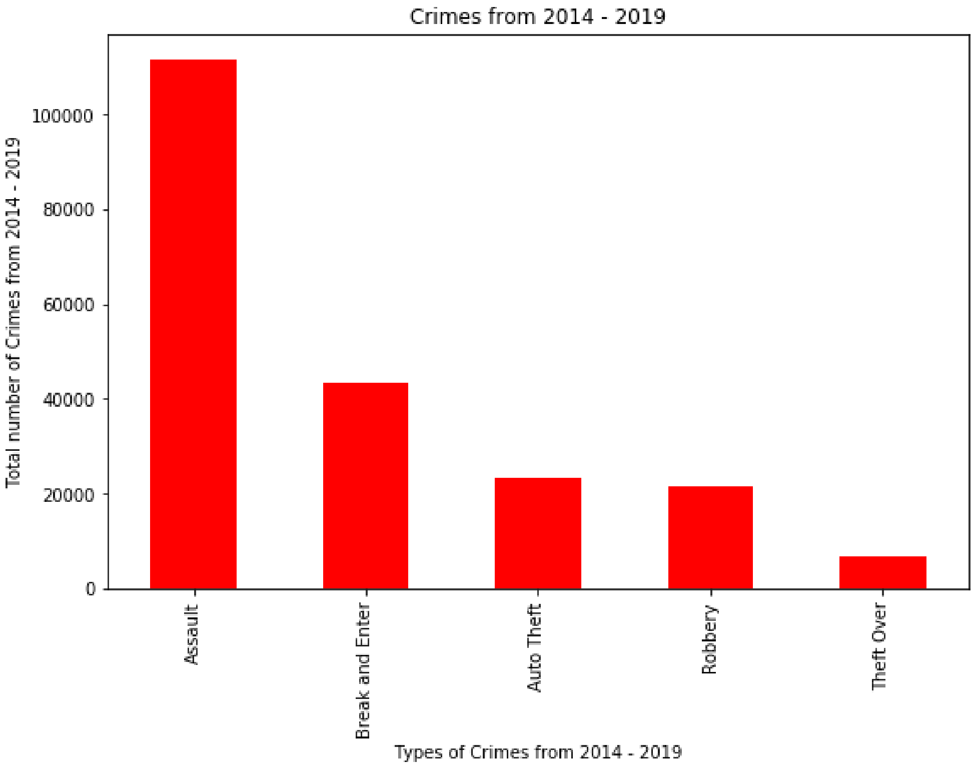
Crime Trend based on year 2014-2019 : **Crimes are increasing!!**

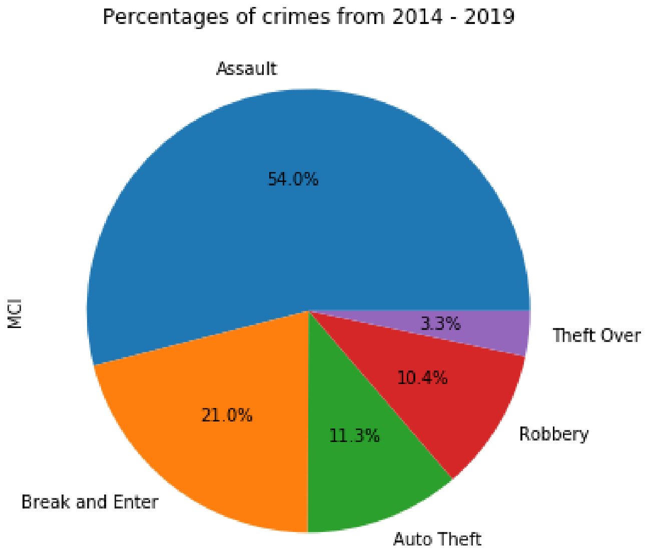


Location of Crimes : **Most Number of Crimes happens Outside followed by Apartments**

Methodology of Problem Solving

Data Analysis- Categories of Crime and their percentages

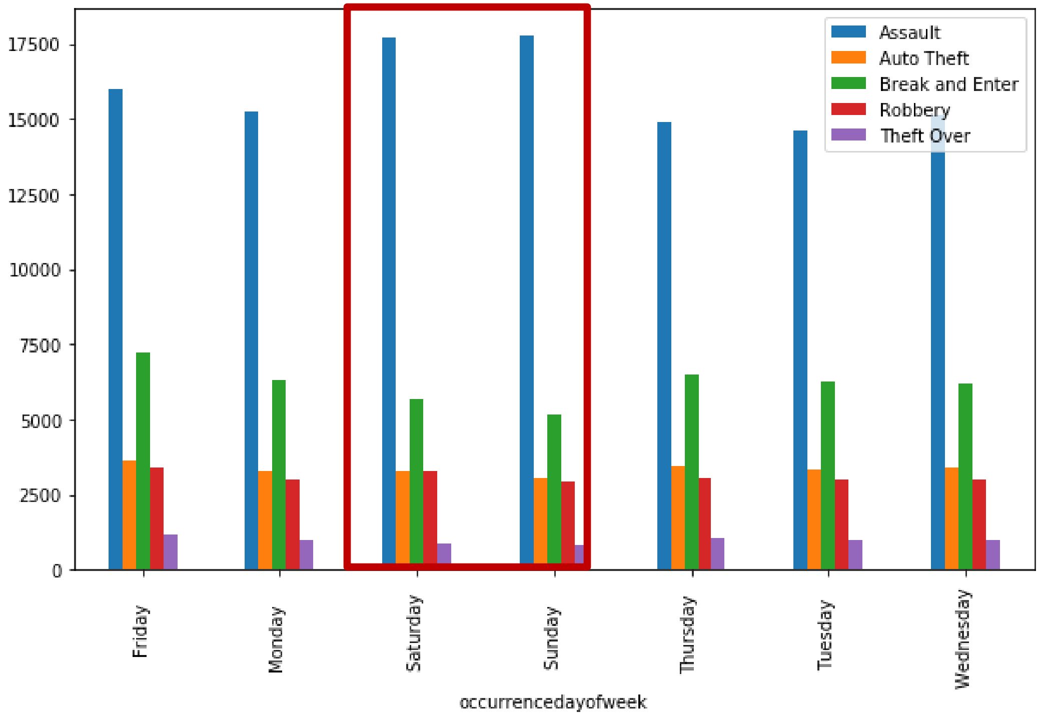




Major type of crime which has been committed in the city of Toronto is **<u>"Assault"</u>** amounting to about **54%** and **<u>"Break and Enter"</u>** amounts to **21%** of the Total crimes

Methodology of Problem Solving

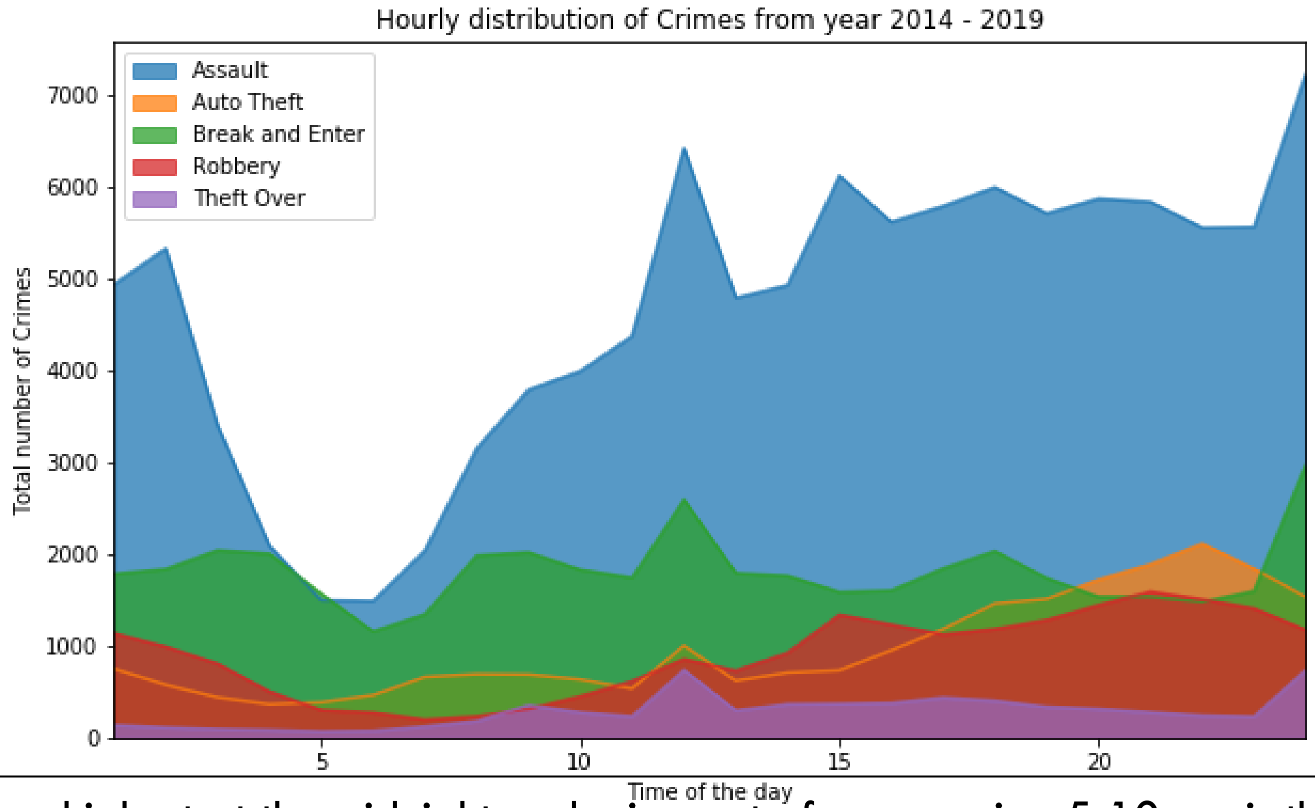
Data Analysis-Weekly distribution of Crime



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%**

Methodology of Problem Solving

Data Analysis-Hourly distribution of Crime

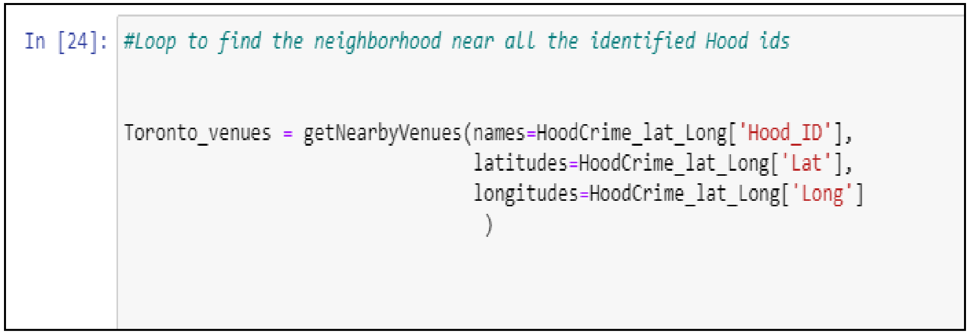


All Crimes are highest at the midnight and crime rate from morning 5-10 am is the least. Although a sudden spike in the crime can be seen in mid afternoon!! Robbery and Auto Theft crimes are highest at around 22 hours.

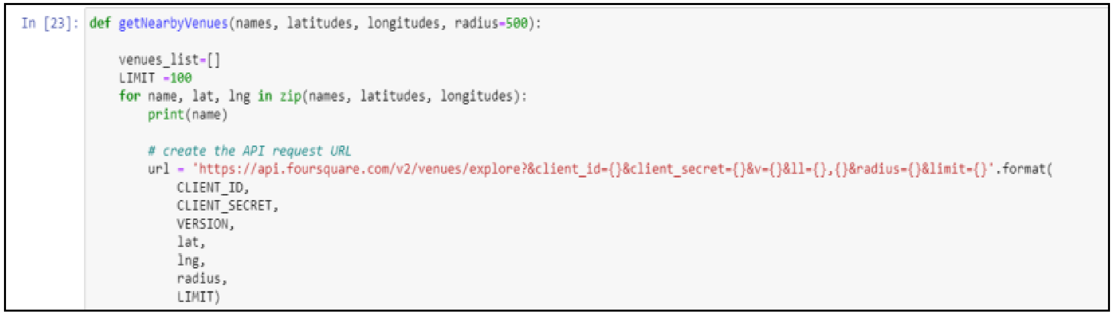
Methodology of Problem Solving

Explore Neighborhood using Four Square API.

"getNearbyVenues" Function

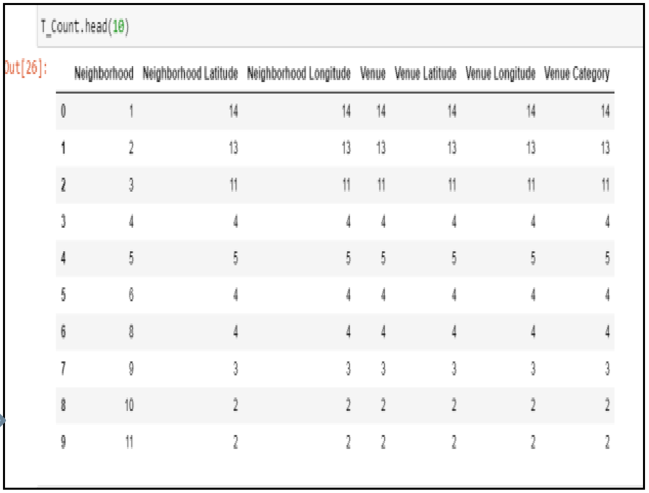


```
In [24]: #Loop to find the neighborhood near all the identified Hood ids


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



```
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)
```

Four Square API



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

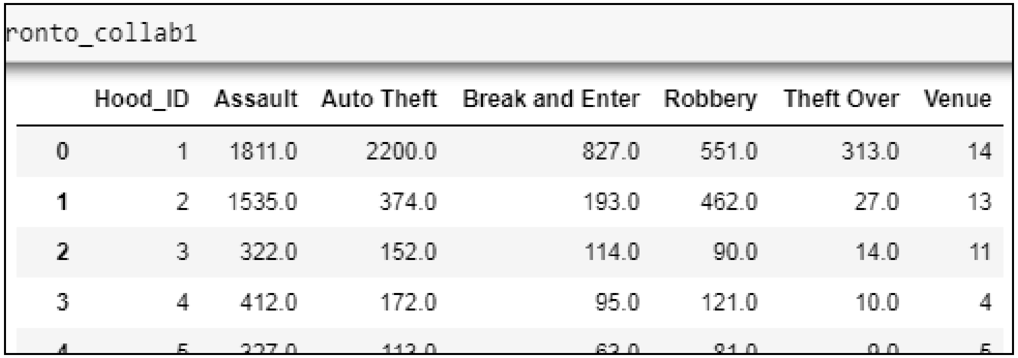
Out[26]:

	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

Count of Venues
mapped to
Neighborhood

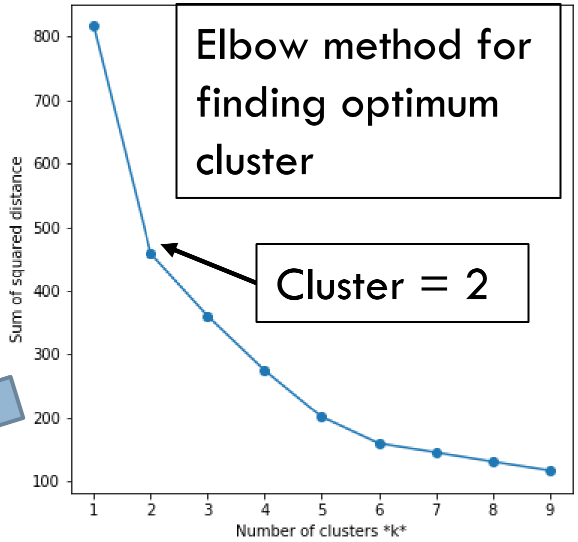
Methodology of Problem Solving

Clustering of Neighborhood – Using K means Clustering



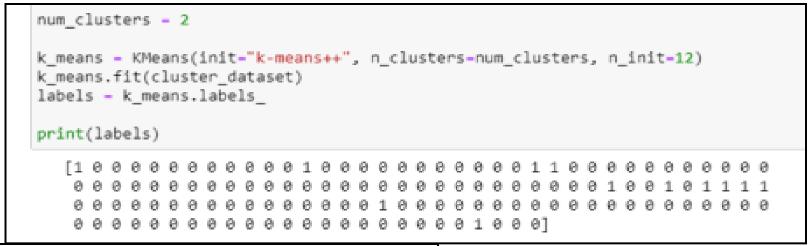
ronto_collab1

	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



Elbow method for finding optimum cluster

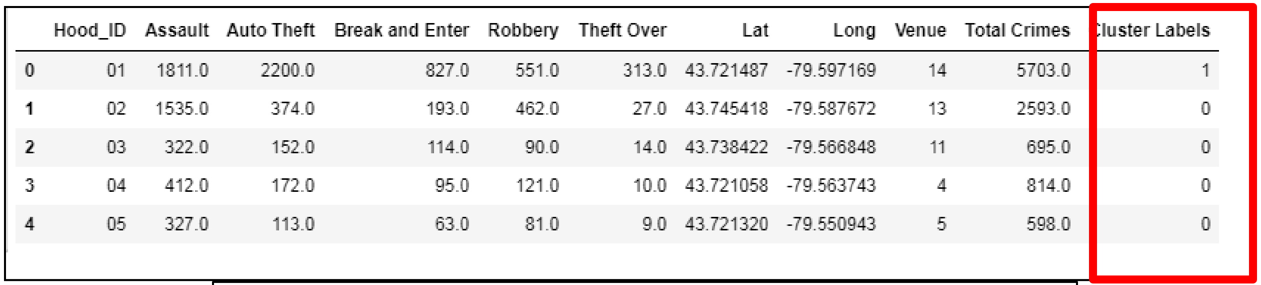
Cluster = 2



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

K- means Clustering Algorithm



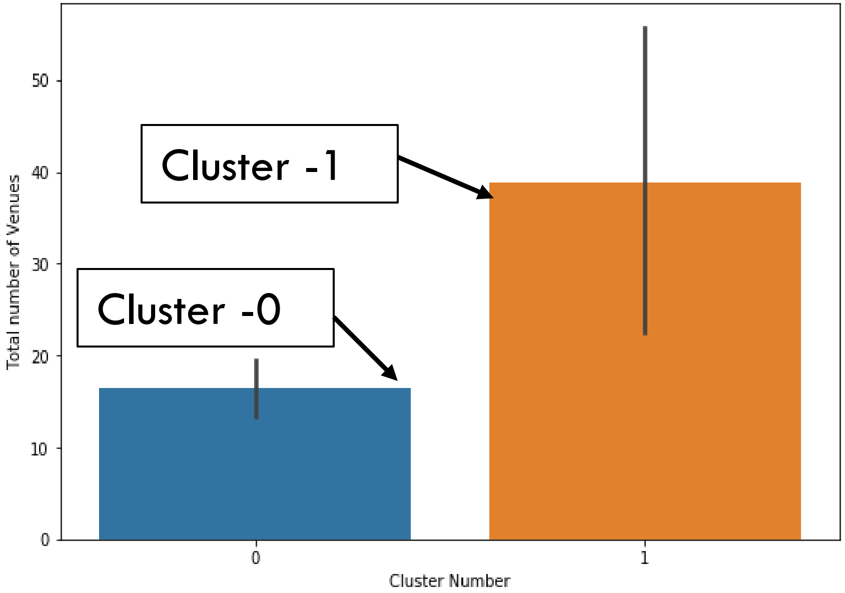
	Hood_ID	Assault	Auto Theft	Break and Enter	Robbery	Theft Over	Lat	Long	Venue	Total Crimes	Cluster Labels
0	01	1811.0	2200.0	827.0	551.0	313.0	43.721487	-79.597169	14	5703.0	1
1	02	1535.0	374.0	193.0	462.0	27.0	43.745418	-79.587672	13	2593.0	0
2	03	322.0	152.0	114.0	90.0	14.0	43.738422	-79.566848	11	695.0	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

Hood data frame divided into two clusters

Conclusion

Analysis of clustered Data frame using Bar Graph and Box Plot

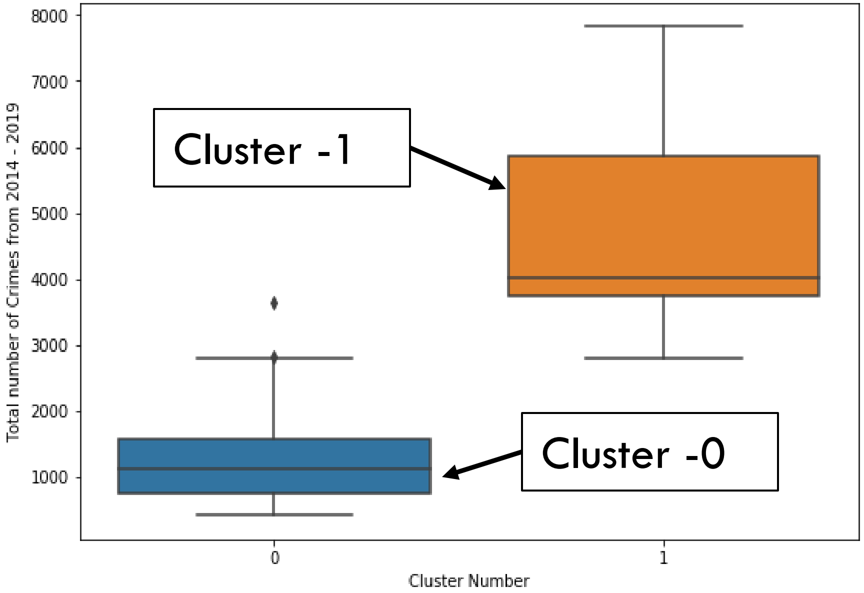
Venues in each cluster



Cluster -1

Cluster -0

Crimes in each cluster



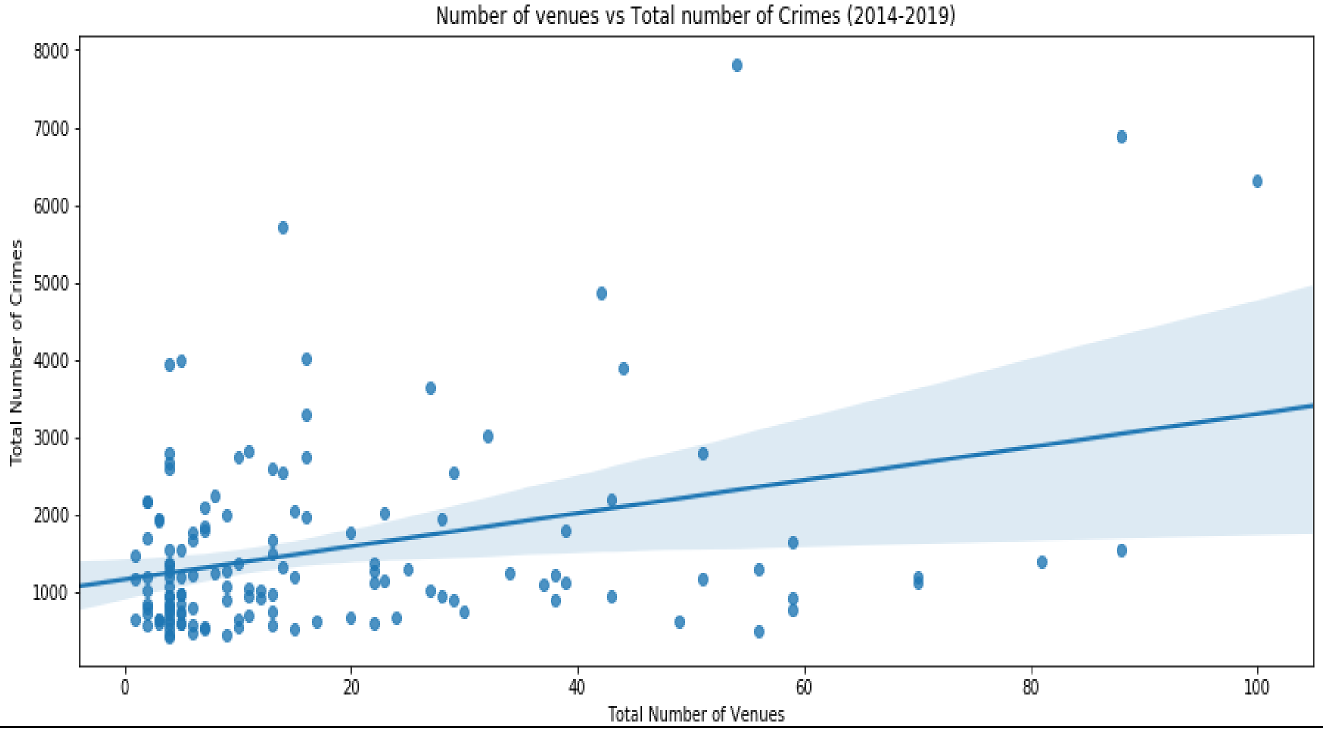
Cluster -1

Cluster -0

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”.

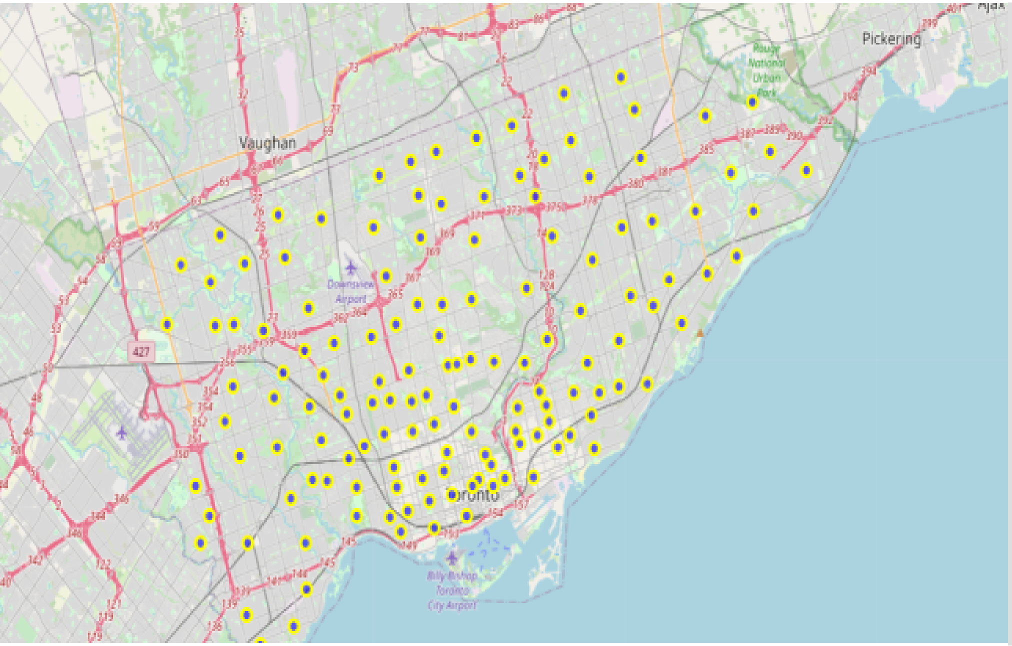
Conclusion

Regression plot of Number of Venues to Total Number of Crimes

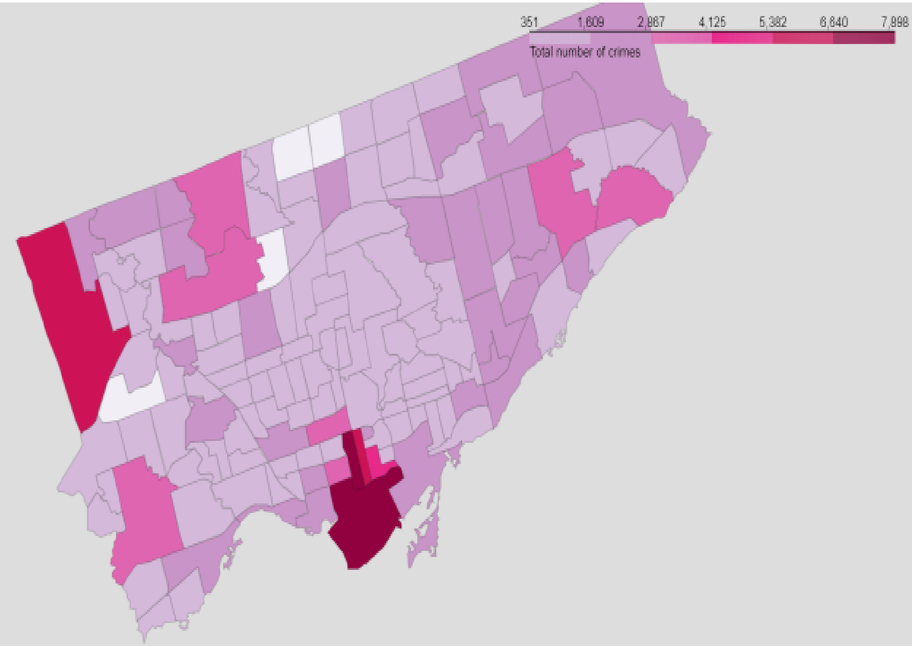


As we can see there is definitely a positive correlation which was further strengthened by the taking the Pearson's correlation and p value. The Pearson Correlation Coefficient came up as **0.3503682188715298** with a P-value of **2.8948696915466882e-05.**

Conclusion



Crime location plotted on the Toronto Map



Intensity of Crime plotted on the Toronto Map

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>."