Crime Rate Predictive Analysis

CSDA1010-031: Introduction to Big Data

Final Project - Team 2

January 2021

by

Don Sohn

Herby Robinson

Pushpendra Sharma

Rachna Kumari

Suzanne Douglas

Abstract

The use of “Big Data” in predictive analytics of crime rates has been applied across Canada in the last decade due to the increasing availability of both computer resources and volume of data. Ethical concerns regarding bias, as well as questions of the practical benefits of these systems are widely debated, and the specific methods of analysis are therefore extremely socially and politically relevant. This report evaluates the application of big data analytics in geographical and person-based criminology.

It focuses on using point of interest (POI) data within the city of Toronto, and attempting to recreate the predictive utility of using the boundaries of each Toronto neighbourhood, incorporating demographic based data of the relevant areas. The exploratory data analysis shows the importance of different independent variables on the output variable i.e. crime rate. Costs and benefits are evaluated in the context of how big data analytics will impact the police department and public safety.

In this project we will be using a dataset created by combining features of datasets from the Toronto Open Data Catalog and merging them with the Toronto Police MCI dataset. The Catalog datasets used in the merge had features pertaining to Housing, Community Service, Income and Education / Skill set.

The goal is to examine the features to see which combination is the best predictor of or contributor to crime based on the crime data found in the MCI dataset. While all the neighbourhoods were represented in the dataset, the time period for the summary data varied and, in some cases, organizing the features by date was not possible. This shortcoming will be addressed in later sections.

Introduction

Crime is important to study and monitor to help identify trends and relationships to patterns and activities pertaining to crime (Wheeler, 2016). These values have been adapted from police organizations nationally to develop core strategies to policing and further develop a reliable relationship with the public community (Braga, 2016). Thus, it is important to recognize that factors influencing crimes are not solely internal but external as well.

This is why it is important to continue studying the spatial demographic patterns between neighborhood and crime occurrences as population changes continuously. Braithwaite (1975) found that population growth and the increase in population density has resulted in higher crime occurrences due to an increase in opportunities and reduction in social cohesion.

This study focuses on crime patterns within the City of Toronto, Ontario which is the largest city in Canada with a total population of 2,731,571 as of 2016 (Statistic Canada, 2016). However, due to unreadily available census data for 2016, attributes from the 2011 census data were used. In 1998, the municipal government restructured Toronto to combine seven large municipalities and to further improve the municipal and provincial responsibilities (City of Toronto, n.d.). The City of Toronto is located Northwest of Lake Ontario and consists of four city-wide ward boundaries which are: Etobicoke-York, North York, Toronto East York, and Scarborough.

In many areas where the MCI values are highest, without looking at the data, i.e. by casual observation, it has been noted that there are pockets in communities that seem to have certain characteristics. These characteristics are

1) An over representation of single parent homes in which children are left often with no outlet for pent up energy.

2) Homes in which the income is closely aligned with the minimum wage.

3) Homes in which the working age occupants have little certification and are therefore disenfranchised workers

4) Homes in which the individuals have poor ‘life coping’ skills.

On observing these characteristics, Sociologist have postulated that this increase in crime can be stemmed by the funding of more community based programs, better access to housing and family life related services, an increase in the minimum wage and better access to academic and trade programs.[1] While this has been vigorously argue and debated in many circles, the documentation produced by the said interactions and the results of the implemented policies coupled with the still increasing crime rate shows that there is somewhat of a disconnect between the results and the perceived solution.[2]

It should be noted however that while there is a disconnect there is value in implementing the policies that are recommended by the sociologist. Humans are complex beings and our interactions even more so. The challenge is for us to find and implement the right combination that will reduce crime in the GTA. It is also important to note that while our analysis may yield a correlation it does not necessarily prove causation.

Background

In criminology, both people and place perspectives have been important aspects of study for researchers. While statistics have always been a primary tool for the prediction of crime trends, with large compendiums of data having been amassed over the course of more than a hundred years.[1] Stemming from them are decisions for the optimal allocation of police resources with the purpose of predicting and preventing crime. The modern era of “big data” has presented both new opportunities and new challenges for predictive policing. As the sheer volume and variety of data available grows beyond what any human can manually analyze or interpret, computer algorithms have become increasingly prevalent to develop predictive models used to allocate police resources, and such data driven models are already being implemented across Canada[2] and numerous other cities across the world.[3] One of the key aspects of machine learning and data analytics is providing a solution to the new era of everyday data collected by different tools and devices around the world.[4] This has helped with answering questions in regards to commonalities and variability of information and further summarize the results in a comprehensible format. Were there a robust and accurate method of predicting future crime, the potential benefits to society would be immense, both in more efficient use of judicial resources as well as reduced crime, a perennial social ill. Data driven crime prediction has given researchers a chance of prediction of crime probability with a high accuracy unavailable to classical statistical analyses.[4][5] For instance, one of the methods of creating combined data is combining POI data from geographical aggregation and hotspot identification. Geographical aggregation is prepared by dividing the study area into small squares of population, while hotspot identification, is crime counts for particular crime types within a geographical grid. An example of such models is presented by Eck, John et al.[4]

While there are many claims of the benefits of predictive policing, some empirical reviews have also found no tangible benefit once these models were actually applied.[6] There are additional concerns that the machine learning models will recreate the racial and social biases which are present in the data.[7] A machine learning approach, Series Finder, was proposed in a study done by Wang et al.[8] to address the problem of detecting specific correlation between crimes committed by the same offender or group of offenders. It stands to reason that, if a certain demographic is overrepresented in the data being fed to the algorithm, then said algorithm will infer a correlation which, while accurately representing the data, may not accurately represent the real world from which the data was drawn. While crude, the adage “garbage in, garbage out” applies equally to big data. It is therefore of special importance to thoroughly assess the generalizability of a given predictive model across multiple contexts. Though the application of these systems is controversial, they are nonetheless a very active area of research, and numerous new methodologies are being developed each passing year. Science is a process of iteration and refinement, and a failure of one model does not predict the failure of all models.

This work will focus on using machine learning models to predict crime risk in the neighbourhoods of Toronto, using POI data, which may include anything from bus stops to convenience stores, libraries, benches, trees, or really any arbitrary but quantifiable/identifiable feature of the urban space. In a study done by Powet Cichosz, POI data was used to predict urban crime rates. In the mentioned work; predictive power evaluation, dimensionality reduction and model transfer was evaluated using training for different data models using UK police crime records for the time span between 2016 and 2019. Powet used shapefiles and bar plots to coordinate between POI subjects and geographical data. Powet also provided visualized predictions using Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) Models.[4] Of particular note in the paper is this quote: “The experimental results obtained for four UK urban areas suggest that POI attributes have high predictive utility. Classification models using these attributes, without any form of location identification, exhibit good predictive performance when applied to new, previously unseen micro-areas.” This is significant, as many of the critiques of these systems allege that the models do not actually have predictive power, however being able to accurately predict the crime rate of a previously unseen area provides strong evidence that the model is indeed robust.

Publications on the topic have only used US and UK based cities to date. In this work, we will attempt to recreate the methods and results employed by Powet, applied to the City of Toronto. POI data will be combined with demographical data using an estimation of gender, age and income group splits. By combining these data in different algorithms and scopes we hope to create a visual representation of the “Major Crime Indicators”(MCI) tracked by the Toronto Police Department from a span between 2014 and 2019, dividing the area by neighbourhood, which have officially defined boundaries which are consistent between different sources, and constitute 140 distinct areas. The predicted crime rates will be compared against the actual rates to determine the accuracy of the method.

Proposed Plan

Different steps of the approach used in the study are:

Exploratory data analysis: Data is collected from the Toronto police department, spanning from 2014 to 2019 and including all categories under “Major Crime Indicators” which are assault, murder, auto theft, theft over $5000, robbery, break&enter, and sexual violation. Exploratory data analysis is performed to understand the data, to find patterns, to spot anomalies and to develop a hypothesis for the model. A detailed exploratory data analysis is carried out in this study to propose the plan for the model development.

Data Cleaning: Based on the hypothesis (crime rate is dependent on the income, neighbourhood, and education / skill set) developed from the exploratory data analysis, data cleansing is performed in the model development phase.

Feature Engineering: Different features will be selected on the basis of the hypothesis and the exploratory data analysis. Based on the exploratory data analysis from this study, it is observed that the rate of the crime is dependent on the day, month of the year, part of the day and income of a person.

Transformation: Data transformation is the process in which we take data from its raw, siloed and normalized source state and transform it into data that’s joined together, dimensionally modeled, denormalized, and ready for analysis. Data transformation will be performed in order to take the effect of all variables into account in the machine learning model.

Data modelling and model evaluation: sci-kit learn will be used to deploy different machine learning models for the study. Data will be split into training, testing and validation sets to fit and validate the model. Furthermore, k-fold cross validation scheme will be used to avoid any errors from the data split. Finally, the accuracy of the model will be evaluated based on the different matrices such as roc, rmse to name the few.

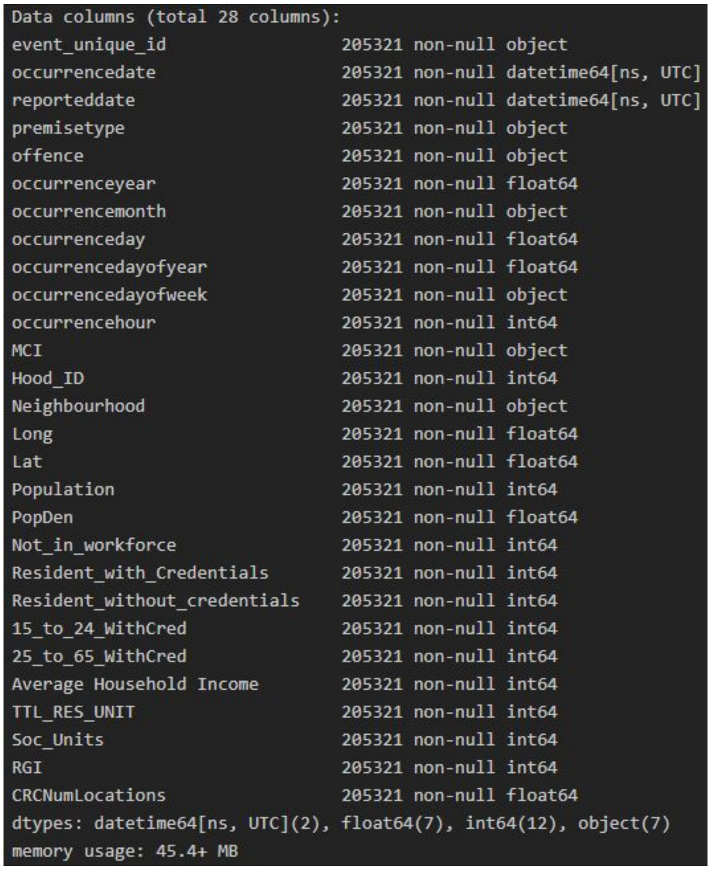
Data Understanding

Fig. 1 (b) describes the quantity and type of data available for the analysis and it indicates a few things. Fig 1 (b) describes the descriptive statistics for the numerical columns present in the data.

1) There were no missing values and there are 28 columns in the file.

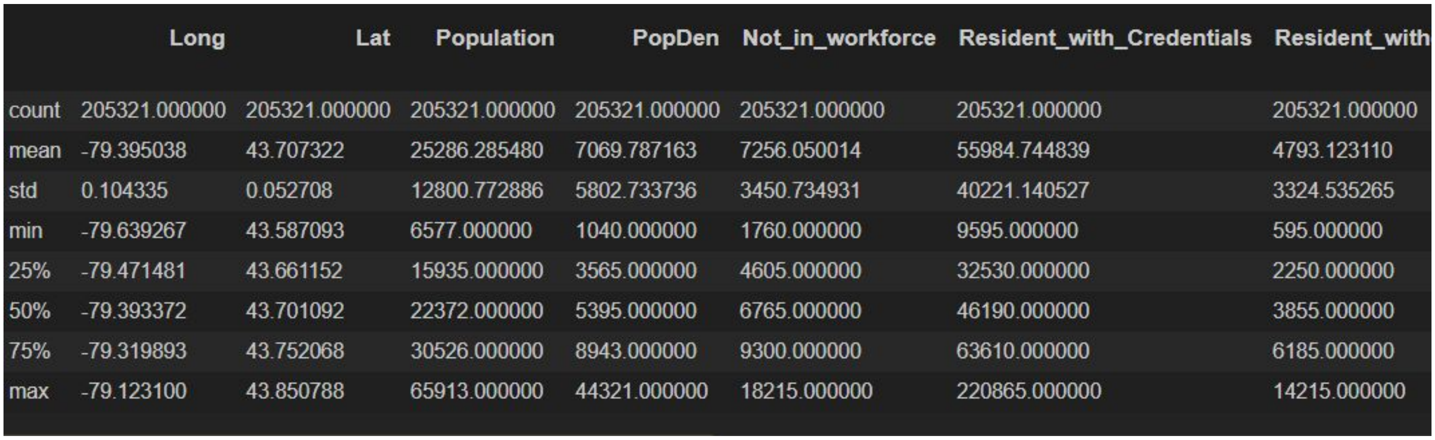
2) The field-names are self explanatory and their types are dates, floats, integers and objects, It was noted that most of the object types could be parsed as meaningful strings based on the fieldname.

3) There are less values for the occurrence of the crime fields as compared to other fields. This was due to the fact that some of the crimes happened before 2014 however, they were reported in the years between 2014 and 2019. Therefore, the values are missing for the occurrence of the offense..



```
Data columns (total 28 columns):
event_unique_id                205321 non-null object
occurrencedate                 205321 non-null datetime64[ns, UTC]
reporteddate                   205321 non-null datetime64[ns, UTC]
premisetype                    205321 non-null object
offence                        205321 non-null object
occurrenceyear                 205321 non-null float64
occurrencemonth                205321 non-null object
occurrenceday                  205321 non-null float64
occurrencedayofyear            205321 non-null float64
occurrencedayofweek            205321 non-null object
occurrencehour                 205321 non-null int64
MCI                            205321 non-null object
Hood_ID                        205321 non-null int64
Neighbourhood                  205321 non-null object
Long                           205321 non-null float64
Lat                            205321 non-null float64
Population                     205321 non-null int64
PopDen                         205321 non-null float64
Not_in_workforce               205321 non-null int64
Resident_with_Credentials      205321 non-null int64
Resident_without_credentials   205321 non-null int64
15_to_24_WithCred              205321 non-null int64
25_to_65_WithCred              205321 non-null int64
Average Household Income       205321 non-null int64
TTL_RES_UNIT                   205321 non-null int64
Soc_Units                      205321 non-null int64
RGI                            205321 non-null int64
CRCNumLocations                205321 non-null float64
dtypes: datetime64[ns, UTC](2), float64(7), int64(12), object(7)
memory usage: 45.4+ MB
```

Fig. 1 (a) Description and quantity of data available for the analysis.



	Long	Lat	Population	PopDen	Not_in_workforce	Resident_with_Credentials	Resident_with
count	205321.000000	205321.000000	205321.000000	205321.000000	205321.000000	205321.000000	205321.000000
mean	-79.395038	43.707322	25286.285480	7069.787163	7256.050014	55984.744839	4793.123110
std	0.104335	0.052708	12800.772886	5802.733736	3450.734931	40221.140527	3324.535265
min	-79.639267	43.587093	6577.000000	1040.000000	1760.000000	9595.000000	595.000000
25%	-79.471481	43.661152	15935.000000	3565.000000	4605.000000	32530.000000	2250.000000
50%	-79.393372	43.701092	22372.000000	5395.000000	6765.000000	46190.000000	3855.000000
75%	-79.319893	43.752068	30526.000000	8943.000000	9300.000000	63610.000000	6185.000000
max	-79.123100	43.850788	65913.000000	44321.000000	18215.000000	220865.000000	14215.000000

Fig 1(b) Descriptive statistics for the numerical type columns used in the modelling

Analysis of Crime facilitated by time based features

To understand the crime rate variation with time and neighbourhood, the crime rate potted with different independent variables. The categories of different crimes in Toronto are shown in Fig. 2. It can be observed that the majority of crimes are assaults followed by entry break, auto theft and theft over.

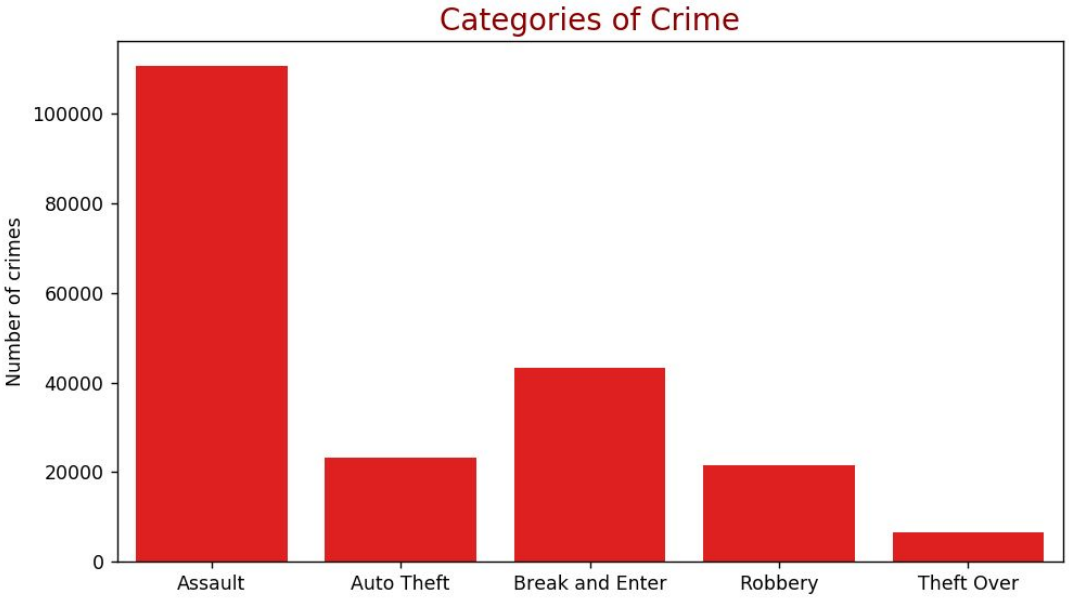


Fig. 2 Number of different crimes in Toronto.

Fig. 3 shows the number of crimes occurring on different days of a week. It can be observed that the number of crimes is largest on Friday followed by Saturday, Sunday, Monday, Thursday, Wednesday and Tuesday. In general, the crime rate was maximum on the weekends as compared to the week days as shown in Fig. 3.

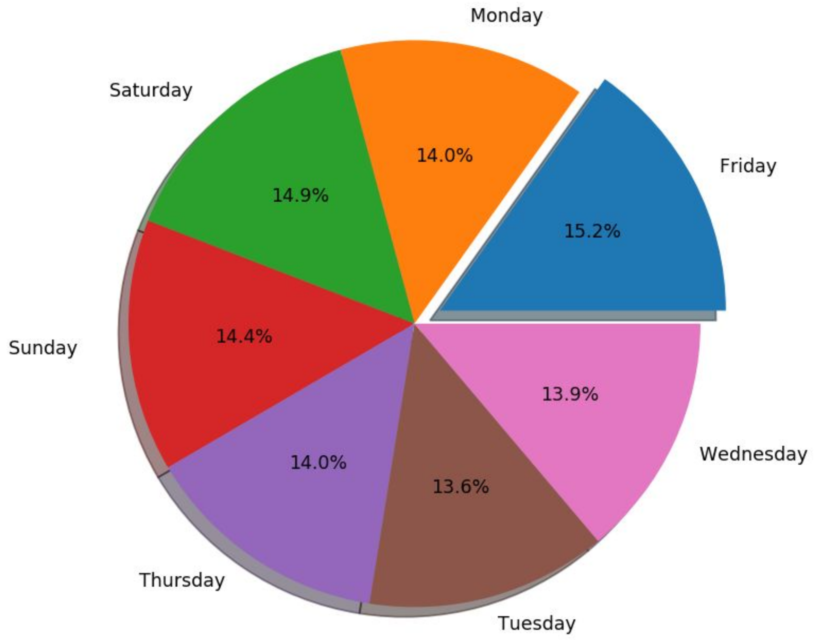


Fig. 3 Number of crimes on different days of the week.

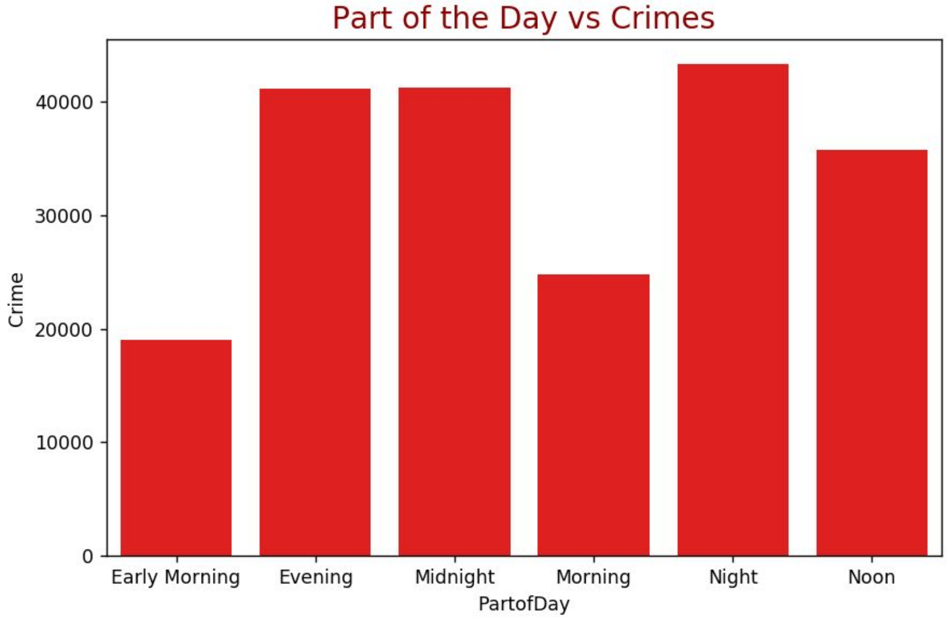


Fig. 4 Number of crimes at different parts of the day.

Furthermore, to understand how crime rates vary within a day, a feature called part of the day is created. This feature of the data tells which part of the day the crime occurred, for instance, morning evening or night. Fig. 4 shows the numbers of crimes for part of the day. It can be observed that the crime rate is higher in the night time followed by midnight, evening, noon, morning and early morning. The crime rate was found to be minimum in the early morning time as shown in Fig. 4. To understand the crime rate with part of the day for a particular day in the week, a 2-D heatmap is plotted as shown in Fig. 5.

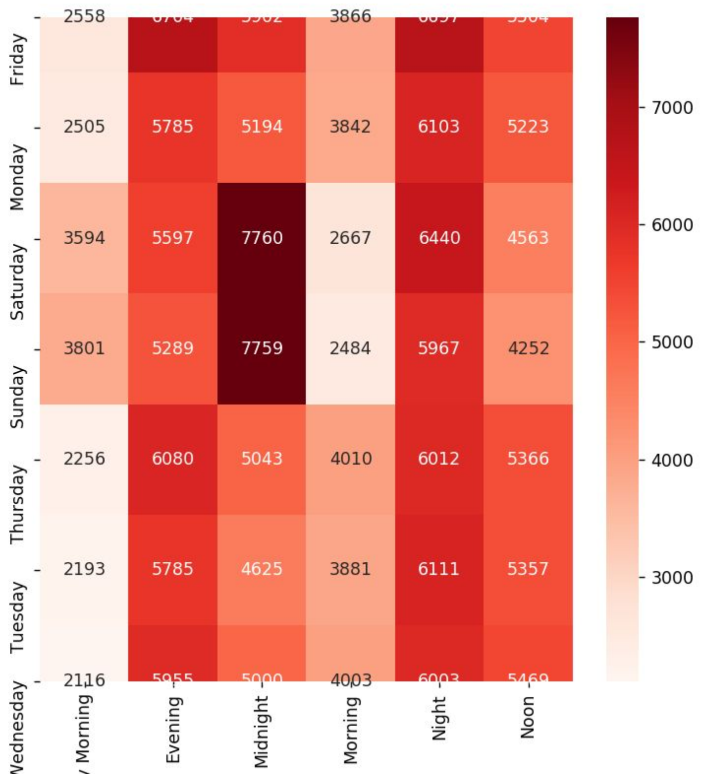


Fig. 5 Number of crimes for a part of day for a particular day.

It can be observed that the crime rate on the weekend midnights is very high. This may indicate that weekend nights are not safe in general in different parts of the Toronto city.

Fig. 6 shows the crime rate in the city of Toronto for different years. It should be noted here that the data is for the years between 2014 to 2019 based on the date of the reporting of the crime. It can be observed that the crime rate was highest in the year 2019. In general, the crime rate is increasing with time.

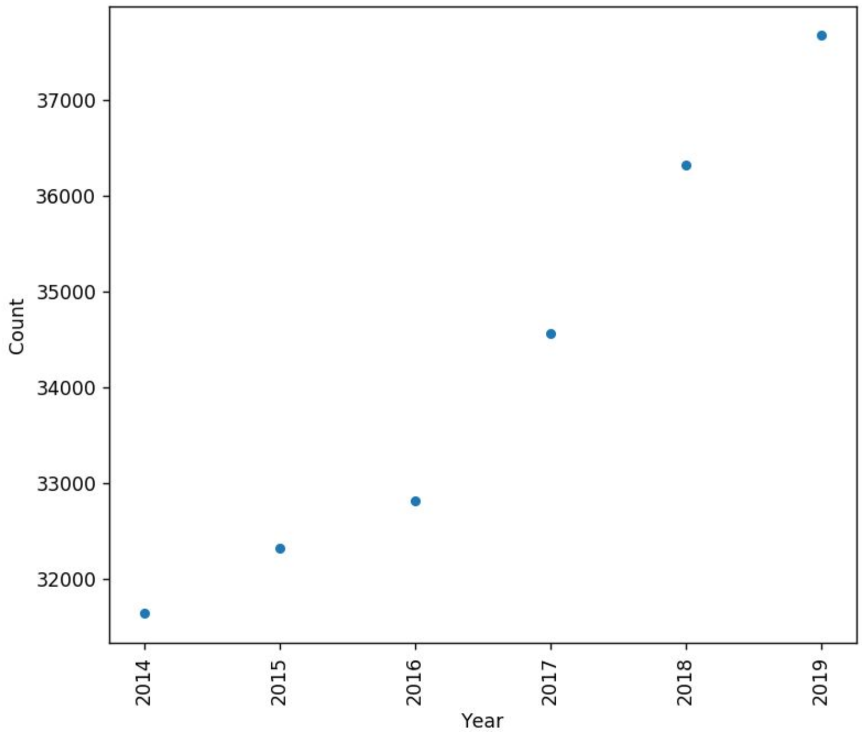


Fig. 6 Crime rate in different years for the city of Toronto.

Fig. 7 shows crime rates for different months within a year. It was observed that there was not much difference between the crime rates for the different months of the year. All the months were found to have approximately similar crime rates. This graph indicates that month may not be a good independent variable for predicting the output for the model.

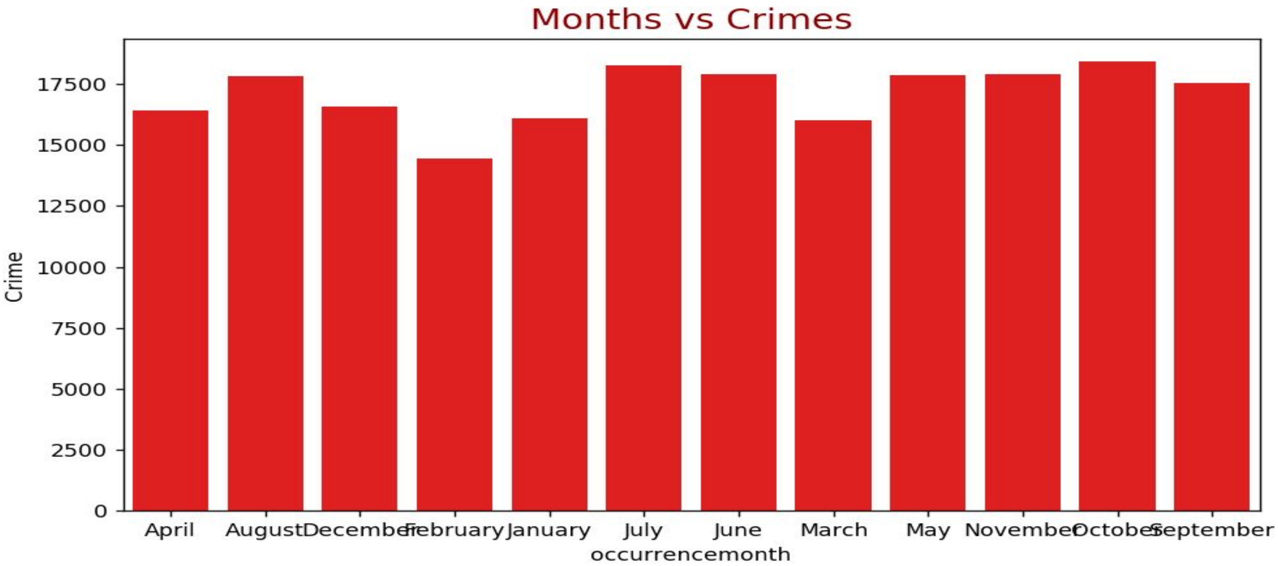


Fig. 7 Crime rate for different months of a year.

To understand the crime rate with days for a particular month in the year, a 2-D heatmap is plotted as shown in Fig. 8. It was observed that there was a major crime event on the day of May 1 of the year as the number of crimes is high on this day as indicated in the Fig. 8. Except May 1, the crime rate for all the days was found to be varied linearly.

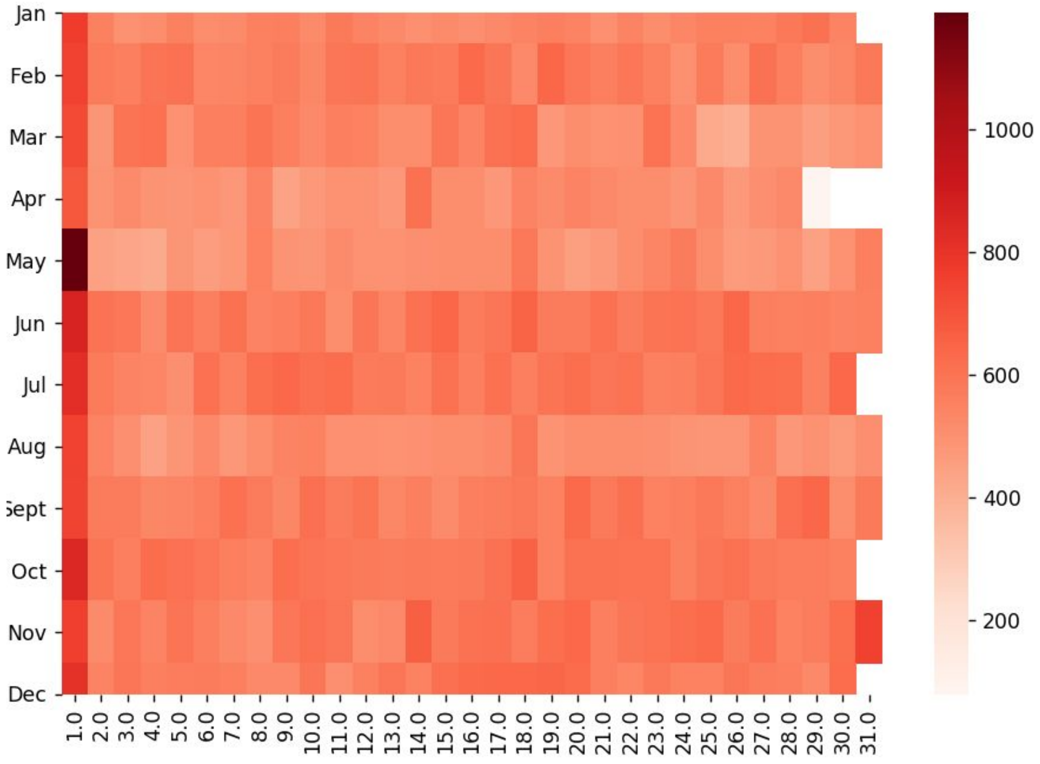


Fig. 8 The crime rate on different days of a particular month.

Fig. 9 shows the crime rate in 140 neighbourhoods of the toronto. The size of the circle is proportional to the number of crimes in the neighbourhood. The highest crime rate was observed for the neighbourhood Waterfront communities- the island.

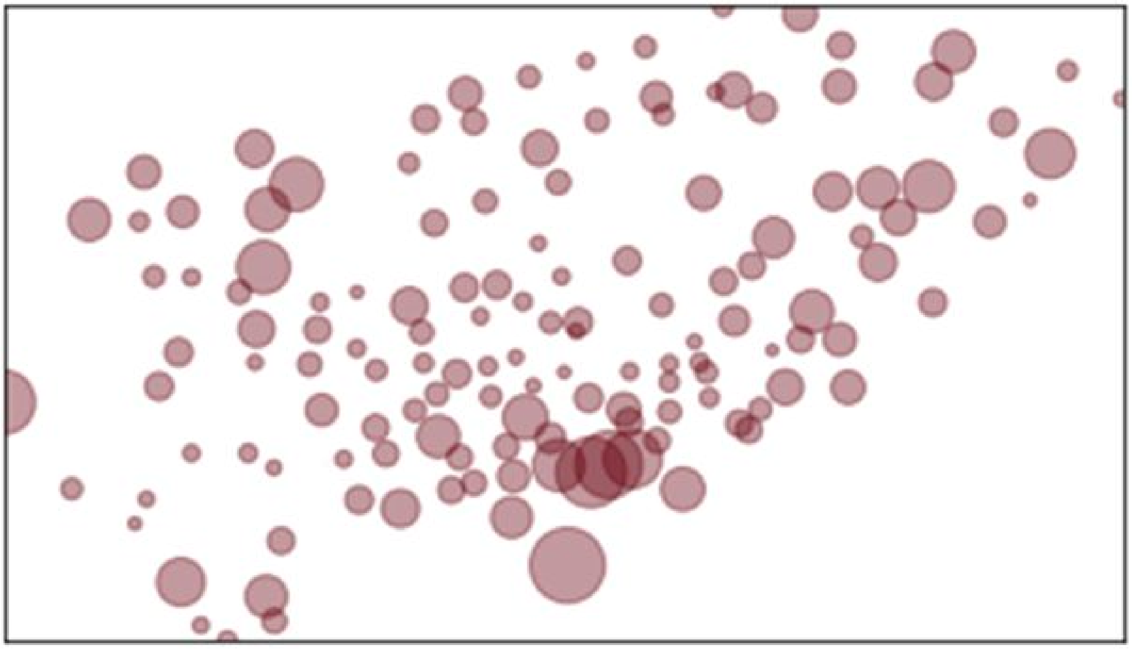


Fig. 9 The crime rate in 140 neighbourhoods of Toronto.

Fig. 9b shows the types of crime Toronto as ranked and displayed via Wordcloud. Sometimes words speak louder than words.

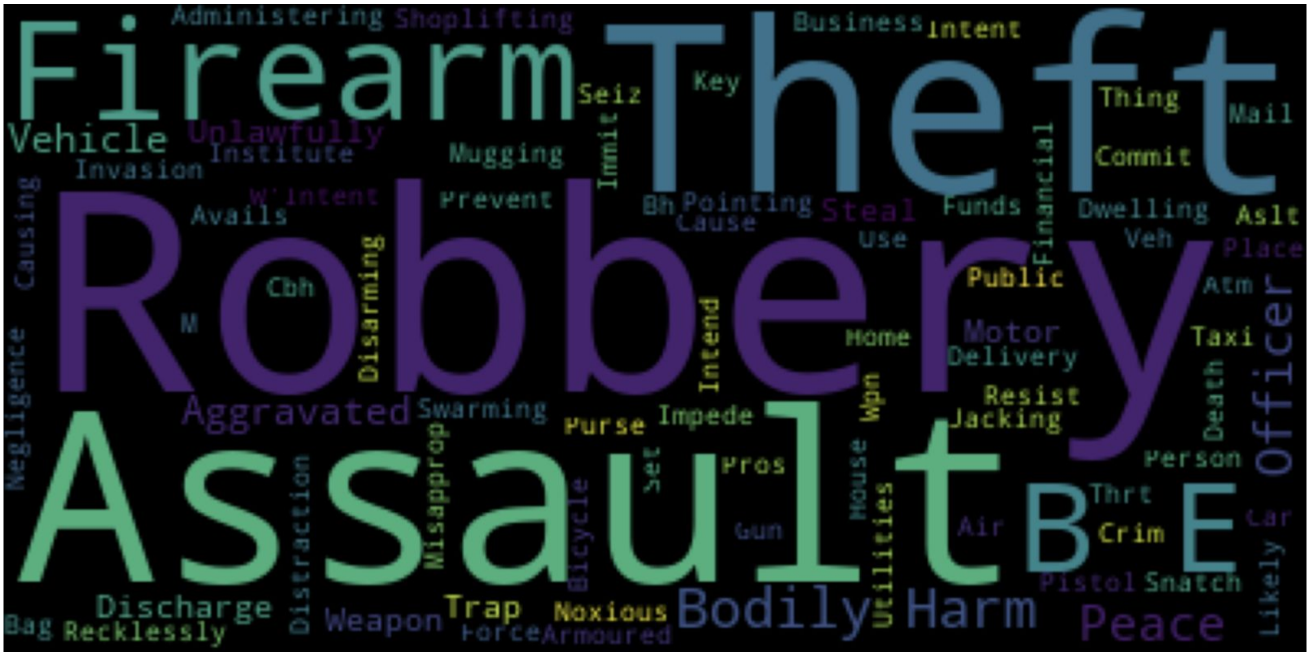


Fig. 9b Types of crime ranked

Data Exploration:

The exploratory portion of this study allows for a visual investigation of the distribution of crime patterns across the city. Inferences can be made about the clustering of crime occurrences across the city to further identify areas that have higher or lower rates of crime.

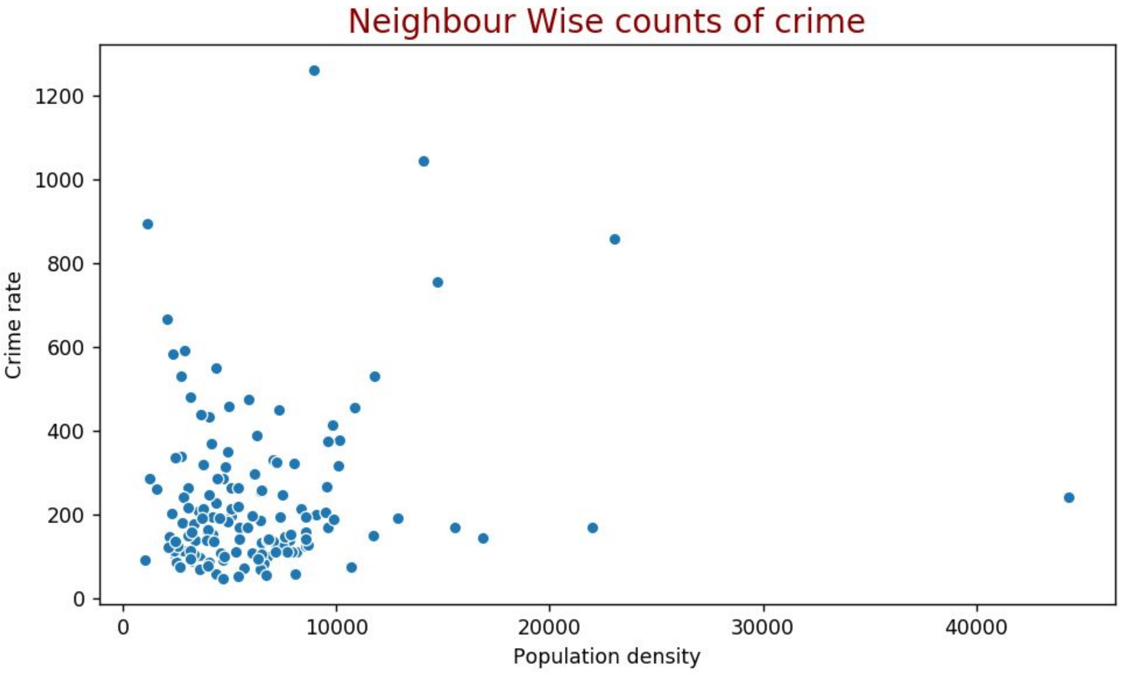


Fig. 10 The crime rate in 140 neighbourhoods of Toronto with population density.

Fig. 10 shows the rate in 140 neighbourhoods of Toronto with population density. There was no general trend between the crime rate and population density.

Based on the exploratory data analysis, it was observed that the crime rates are correlated with the region of crime occurrence, time of the crime, category of the crime, income and education / skill set. Based on the analysis, a hypothesis that the crime rate can be predicted based on the neighbourhood, time, education/ skillset and the income of the citizens living in the neighbourhood. This hypothesis will be tested further during the model development using different model evaluation matrices.

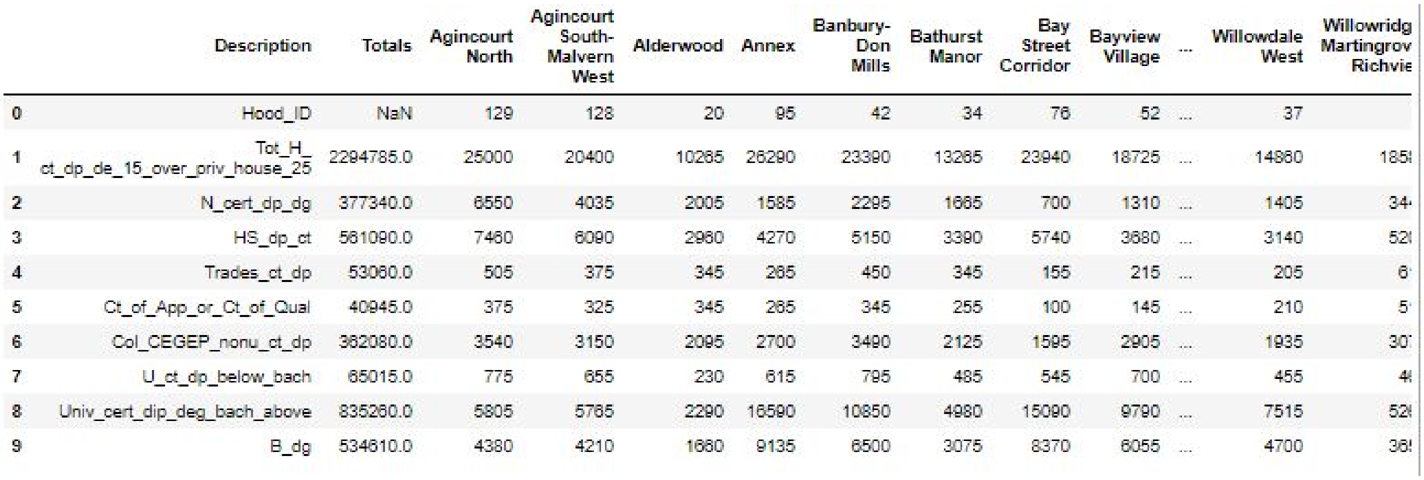
After the exploratory data analysis the next steps for the proposed plan are data cleaning, feature engineering, data transformation to prepare the data for the machine learning model. These steps will be completed in the next phase of the project in the coming week.

Using the developed machine learning model, the crime rates for the year 2020 will be predicted. The predicted values of the crime rate will be compared to the actual crime rates in the year 2020. The latter will be done to understand the effect of coronavirus on the crime rates.

Analysis of Crime based on the Residents Skillset

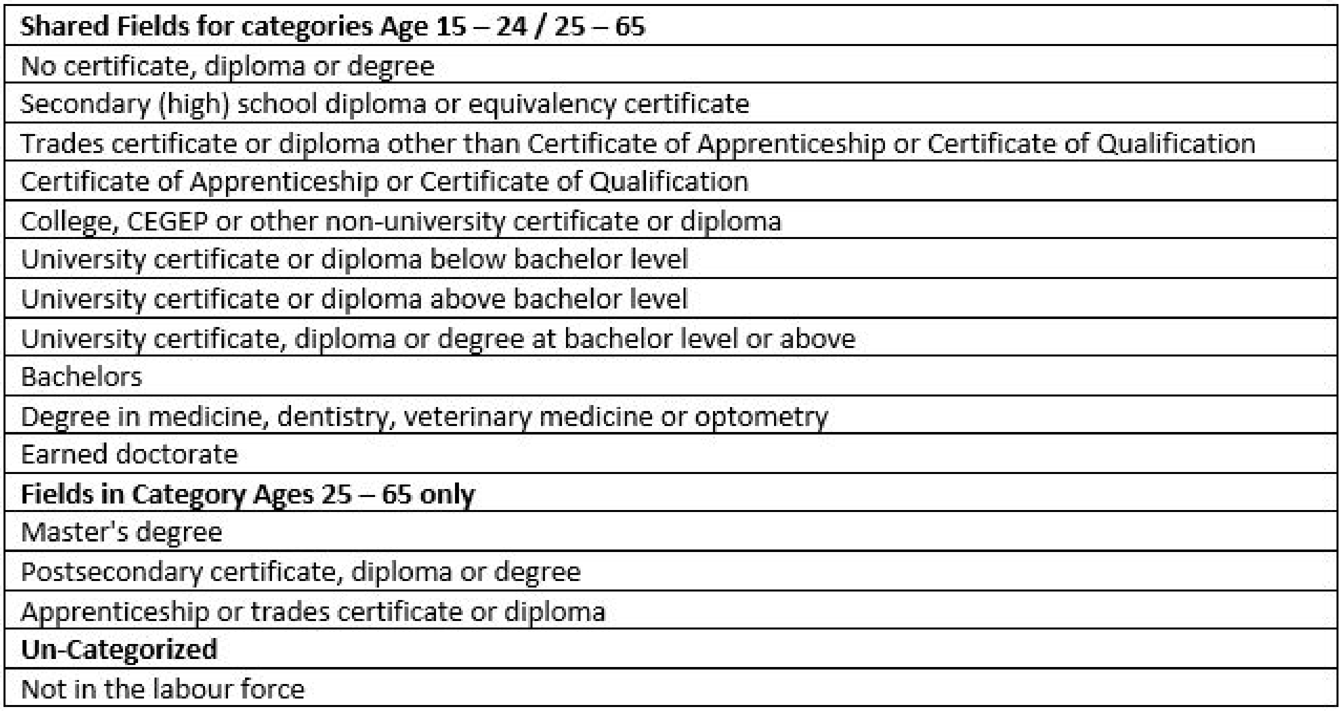
On reading the documentation on the dataset we realized that the Skillset dataset on Toronto was only current as of 2016. One of the criteria for our project was that we should have in excess of 100K records per dataset. However, when a filter was done on the crime dataset it was found that if we were to restrict the MCI data to 2016 then the criteria would not be met. Although the odds were that this dataset was not going to be used we still went ahead and did some analysis with the restricted dataset to see what information could be gleaned.

The file as formatted can be seen below but for our purposes the data must be transposed .



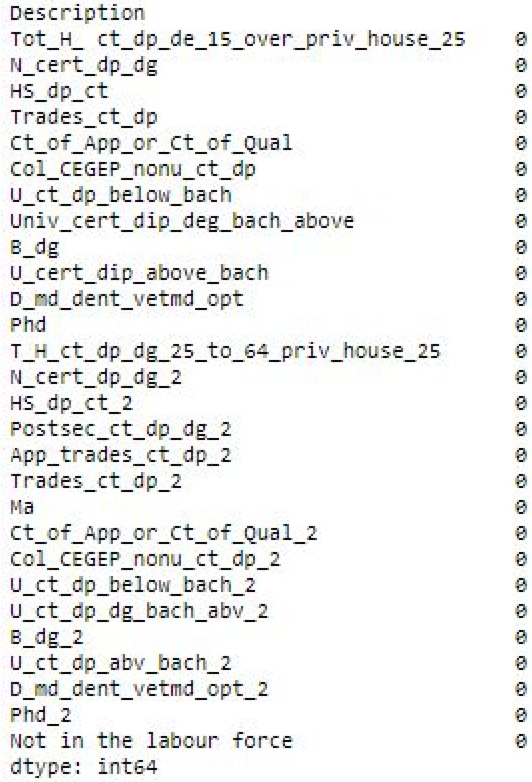
	Description	Totals	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village	...	Willowdale West	Willowridg Martingrov Richvie
0	Hood_ID	NaN	129	128	20	95	42	34	76	52	...	37	
1	Tot_H_ ct_dp_de_15_over_priv_house_25	2294785.0	25000	20400	10265	26290	23390	13265	23940	18725	...	14860	185
2	N_cert_dp_dg	377340.0	6550	4035	2005	1585	2295	1665	700	1310	...	1405	34
3	HS_dp_ct	561090.0	7460	6090	2960	4270	5150	3390	5740	3680	...	3140	52
4	Trades_ct_dp	53060.0	505	375	345	265	450	345	155	215	...	205	6
5	Ct_of_App_or_Ct_of_Qual	40945.0	375	325	345	265	345	255	100	145	...	210	5
6	Col_CEGEP_nonu_ct_dp	362080.0	3540	3150	2095	2700	3490	2125	1595	2905	...	1935	30
7	U_ct_dp_below_bach	65015.0	775	655	230	615	795	485	545	700	...	455	4
8	Univ_cert_dip_deg_bach_above	835260.0	5805	5765	2290	16590	10850	4980	15090	9790	...	7515	52
9	B_dg	534610.0	4380	4210	1660	9135	6500	3075	8370	6055	...	4700	36

The transposed dataset that was used consisted of 26 fields, 11 fields were categorized into the age groups 15 to 24 and then 25 to 60 and , Three fields were unique to the over 25 category and one field was a summary field for persons not in the labour force. Below is a table highlighting the fields.



Shared Fields for categories Age 15 – 24 / 25 – 65
No certificate, diploma or degree
Secondary (high) school diploma or equivalency certificate
Trades certificate or diploma other than Certificate of Apprenticeship or Certificate of Qualification
Certificate of Apprenticeship or Certificate of Qualification
College, CEGEP or other non-university certificate or diploma
University certificate or diploma below bachelor level
University certificate or diploma above bachelor level
University certificate, diploma or degree at bachelor level or above
Bachelors
Degree in medicine, dentistry, veterinary medicine or optometry
Earned doctorate
Fields in Category Ages 25 – 65 only
Master's degree
Postsecondary certificate, diploma or degree
Apprenticeship or trades certificate or diploma
Un-Categorized
Not in the labour force

As seen below the data did not have any null values and all values were of the right type (numeric).



```
Description
Tot_H_ ct_dp_de_15_over_priv_house_25    0
N_cert_dp_dg                             0
HS_dp_ct                                 0
Trades_ct_dp                             0
Ct_of_App_or_Ct_of_Qual                  0
Col_CEGEP_nonu_ct_dp                     0
U_ct_dp_below_bach                       0
Univ_cert_dip_deg_bach_above             0
B_dg                                     0
U_cert_dip_above_bach                    0
D_md_dent_vetmd_opt                      0
Phd                                      0
T_H_ct_dp_dg_25_to_64_priv_house_25      0
N_cert_dp_dg_2                           0
HS_dp_ct_2                               0
Postsec_ct_dp_dg_2                       0
App_trades_ct_dp_2                       0
Trades_ct_dp_2                           0
Ma                                       0
Ct_of_App_or_Ct_of_Qual_2                0
Col_CEGEP_nonu_ct_dp_2                   0
U_ct_dp_below_bach_2                     0
U_ct_dp_dg_bach_abv_2                    0
B_dg_2                                   0
U_ct_dp_abv_bach_2                       0
D_md_dent_vetmd_opt_2                    0
Phd_2                                    0
Not in the labour force                  0
dtype: int64
```

Various plots against crime were created but the data was too granular so to achieve more meaningful plots 7 totals per neighbourhood were created

1)Totals for individuals in the age 15 - 24 with a certified skillset

2)Totals for individuals in the age 15 - 24 with no certified skillset

3)Totals for individuals in the age 25 - 65 with a certified skillset

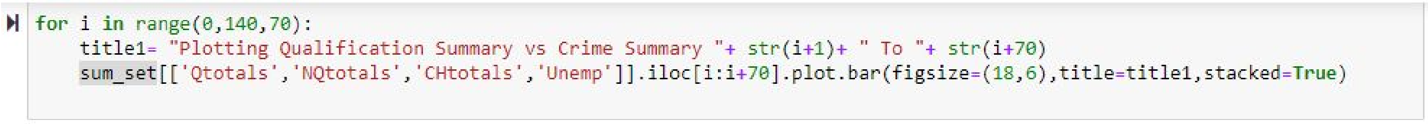
4)Totals for individuals in the age 25 - 65 with no certified skillset

5)Totals for individuals with a certified skillset

6)Totals for individuals with no certified skillset

7)Totals for individuals in the age 25 - 65 with no certified skillset

These totals were then plotted against a crime summary total for each neighbourhood. Below is a subset of



```
for i in range(0,140,70):
    title1= "Plotting Qualification Summary vs Crime Summary "+ str(i+1)+ " To "+ str(i+70)
    sum_set[['Qtotals','NQtotals','CHtotals','Unemp']].iloc[i:i+70].plot.bar(figsize=(18,6),title=title1,stacked=True)
```

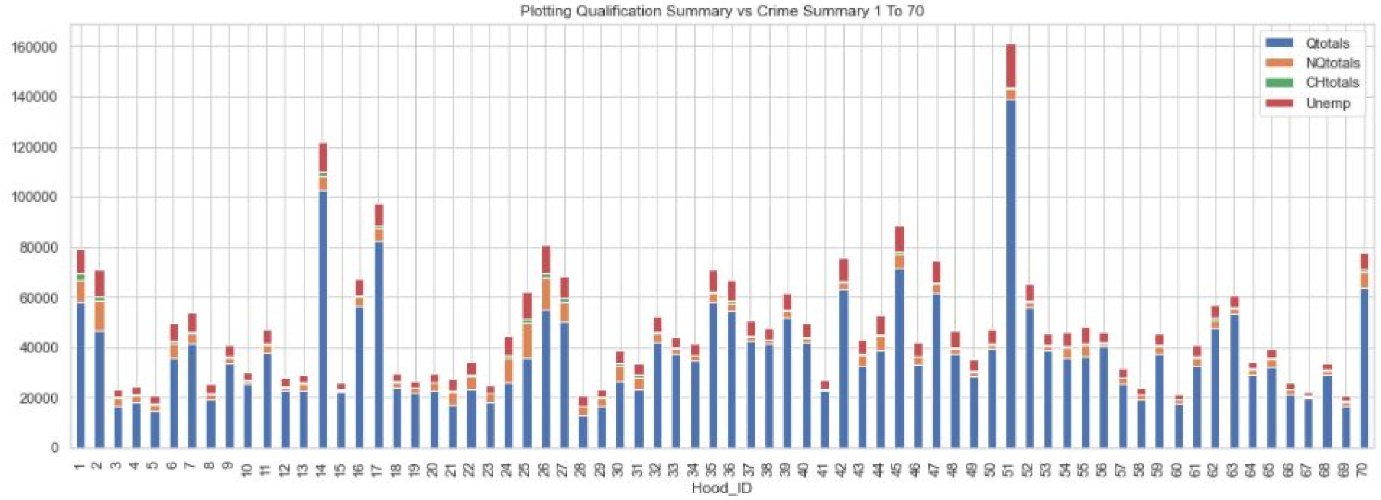


Fig. 11 Plotted against a crime summary total for each neighbourhood

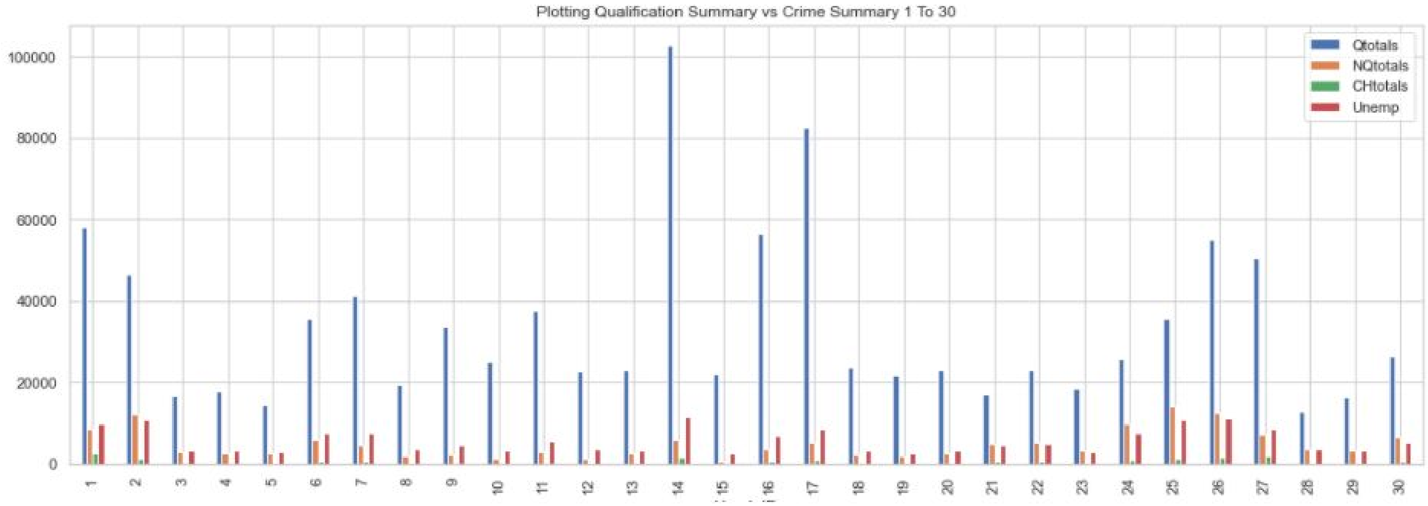
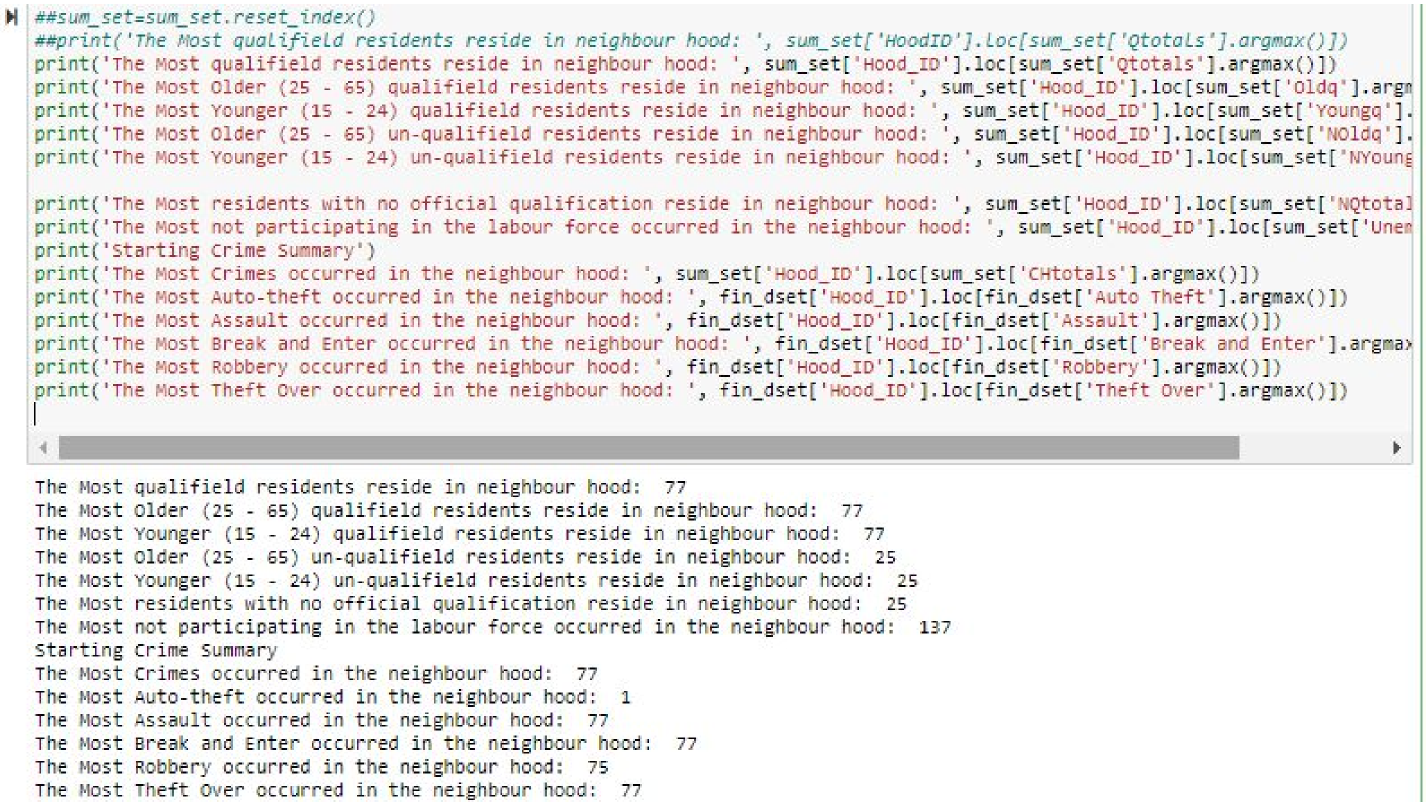


Fig. 12 the graphs and summaries generated.



```
##sum_set=sum_set.reset_index()
##print('The Most qualifield residents reside in neighbour hood: ', sum_set['HoodID'].loc[sum_set['Qtotals'].argmax()])
print('The Most qualifield residents reside in neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['Qtotals'].argmax()])
print('The Most Older (25 - 65) qualifield residents reside in neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['Oldq'].argm
print('The Most Younger (15 - 24) qualifield residents reside in neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['Youngq'].
print('The Most Older (25 - 65) un-qualifield residents reside in neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['NOldq'].
print('The Most Younger (15 - 24) un-qualifield residents reside in neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['NYoung

print('The Most residents with no official qualification reside in neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['NQtotal
print('The Most not participating in the labour force occurred in the neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['Unem
print('Starting Crime Summary')
print('The Most Crimes occurred in the neighbour hood: ', sum_set['Hood_ID'].loc[sum_set['CHtotals'].argmax()])
print('The Most Auto-theft occurred in the neighbour hood: ', fin_dset['Hood_ID'].loc[fin_dset['Auto Theft'].argmax()])
print('The Most Assault occurred in the neighbour hood: ', fin_dset['Hood_ID'].loc[fin_dset['Assault'].argmax()])
print('The Most Break and Enter occurred in the neighbour hood: ', fin_dset['Hood_ID'].loc[fin_dset['Break and Enter'].argmax
print('The Most Robbery occurred in the neighbour hood: ', fin_dset['Hood_ID'].loc[fin_dset['Robbery'].argmax()])
print('The Most Theft Over occurred in the neighbour hood: ', fin_dset['Hood_ID'].loc[fin_dset['Theft Over'].argmax()])
```

```
The Most qualifield residents reside in neighbour hood:  77
The Most Older (25 - 65) qualifield residents reside in neighbour hood:  77
The Most Younger (15 - 24) qualifield residents reside in neighbour hood:  77
The Most Older (25 - 65) un-qualifield residents reside in neighbour hood:  25
The Most Younger (15 - 24) un-qualifield residents reside in neighbour hood:  25
The Most residents with no official qualification reside in neighbour hood:  25
The Most not participating in the labour force occurred in the neighbour hood:  137
Starting Crime Summary
The Most Crimes occurred in the neighbour hood:  77
The Most Auto-theft occurred in the neighbour hood:  1
The Most Assault occurred in the neighbour hood:  77
The Most Break and Enter occurred in the neighbour hood:  77
The Most Robbery occurred in the neighbour hood:  75
The Most Theft Over occurred in the neighbour hood:  77
```

The output generated from the data was contrary to what was expected as it was observed that the highest crime was also in the neighborhood that had the most skilled residents in both age-groups. It was also interesting that the neighbourhood with the most skilled also had the highest crime totals in the categories of assault, breaking and enter and theft over.

The exploratory analysis of the 2016 census Education / skill set data of neighborhoods when combined with the crime data for neighborhoods allowed us to conclude that education / skill set was not a contributing factor in the neighborhoods crime rate. Given the previously stated result and that the record count criteria was not met, the possibility of the education / skill set features being omitted from the final model is a distinct possibility.

Analysis of Crime based on their proximity to Customer Recreation Centres

Crimes reported by TPS include: homicide, assault, burglary, theft etc. Personal crime is defined as homicide and assault whereas property crime were burglary, theft, etc.

Plotted in relation to the number of recreation centres in each respective neighbourhood, yielded the following graphs:

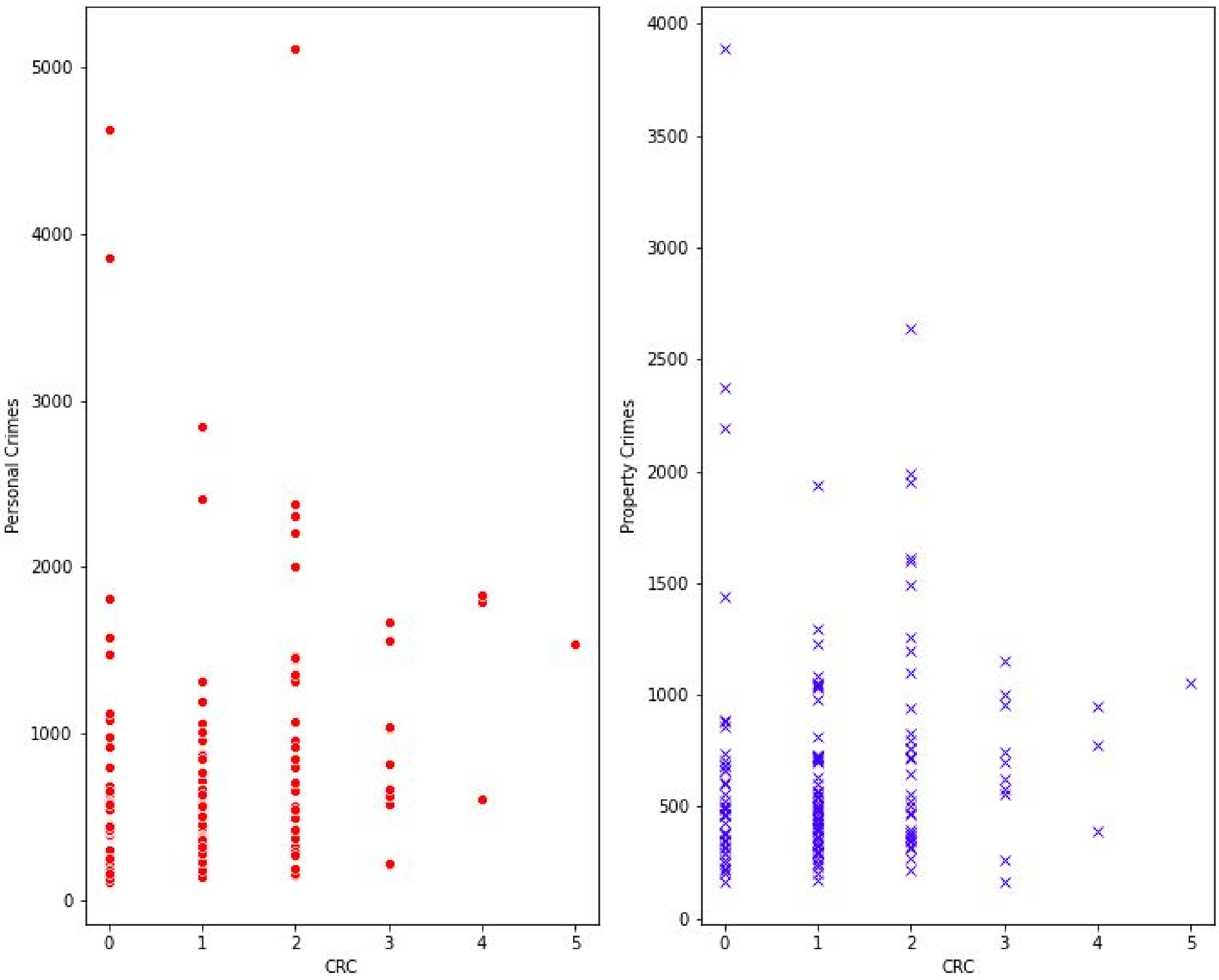
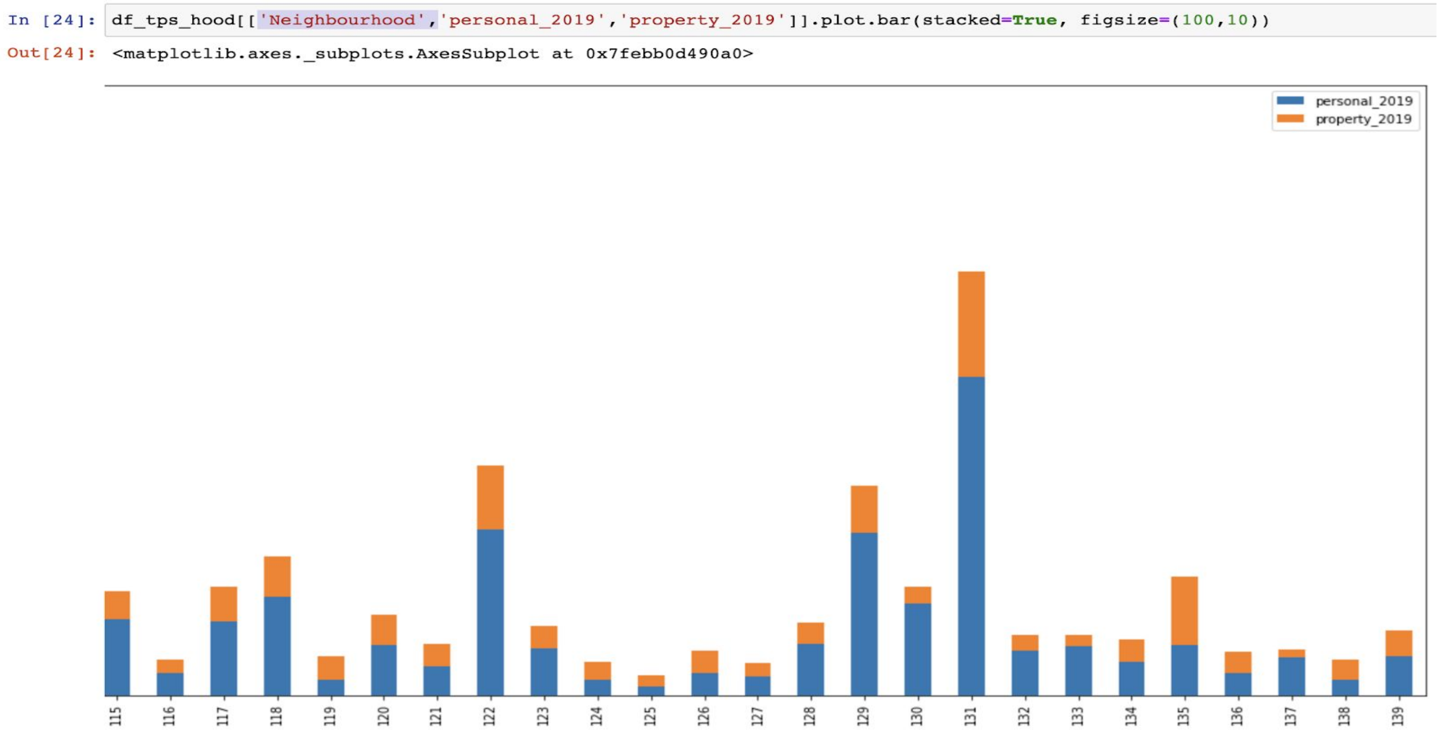


Fig 13: Number of recreation centres

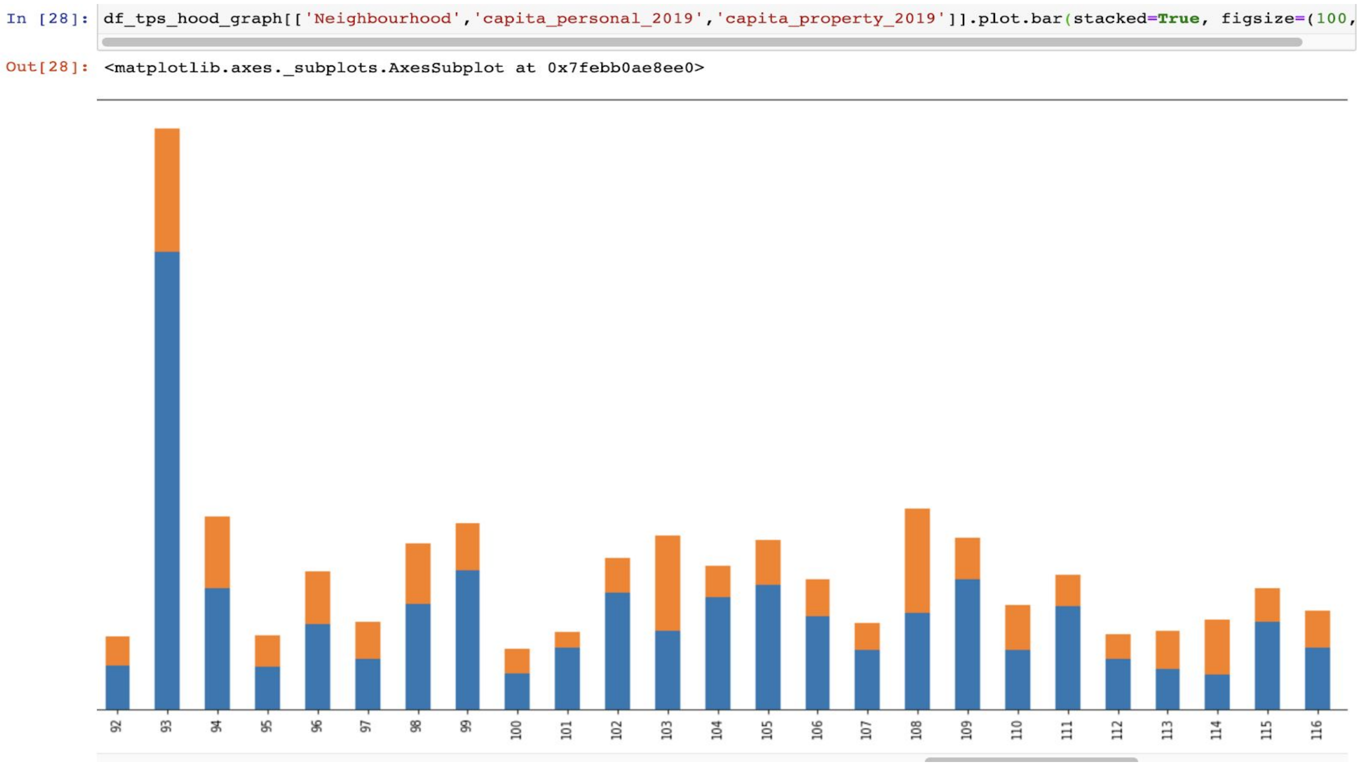
Stack Plotted personal and property crime for each of the 140 neighbourhoods. Surprisingly and consistently there is more personal crime than property crime which is disturbing.



```
In [24]: df_tps_hood[['Neighbourhood','personal_2019','property_2019']].plot.bar(stacked=True, figsize=(100,10))

Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x7febb0d490a0>
```

Fig 14: Personal and Property crime for each of the 140 neighbourhood

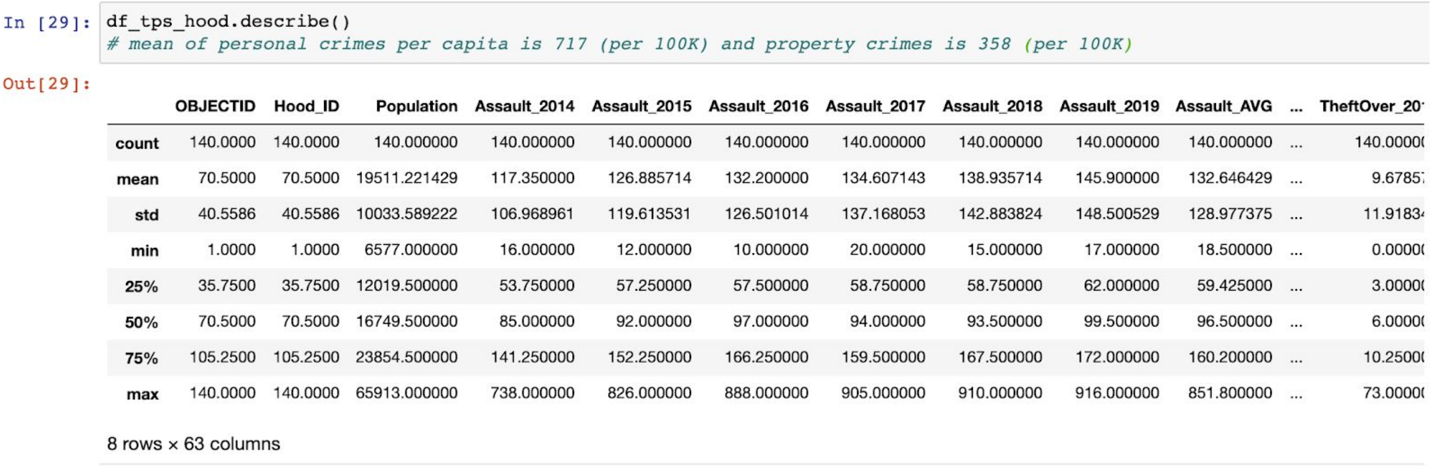
Performed the same plot but per capita to see if population size would have large impact (it did not):



```
In [28]: df_tps_hood_graph[['Neighbourhood','capita_personal_2019','capita_property_2019']].plot.bar(stacked=True, figsize=(100,

Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x7febb0ae8ee0>
```

Fig 15: Per capita to see if population size

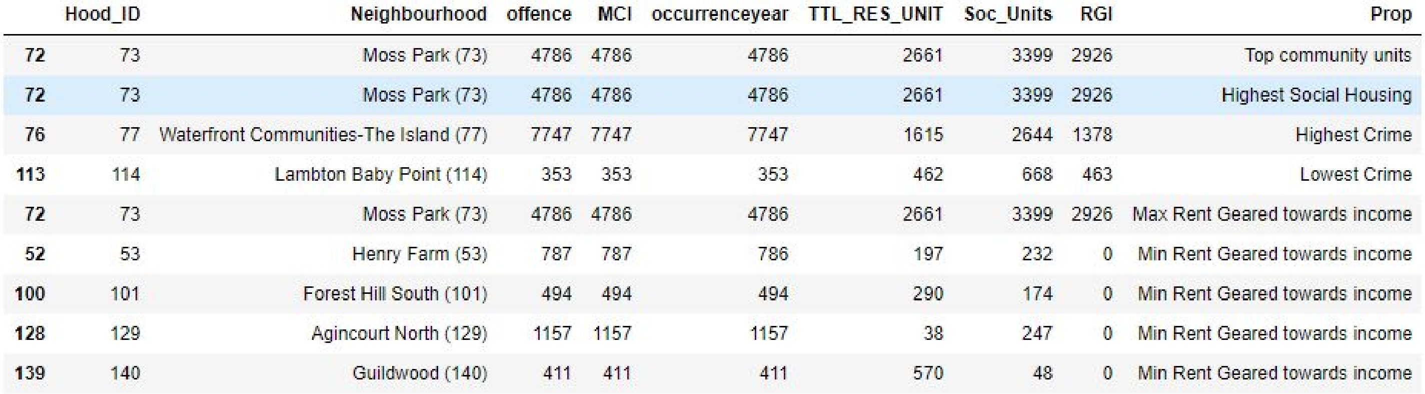
The mean/average of personal crimes per capita is 717 (per 100K) and property crimes is 358 (per 100K)



```
In [29]: df_tps_hood.describe()
         # mean of personal crimes per capita is 717 (per 100K) and property crimes is 358 (per 100K)

Out[29]:
        OBJECTID  Hood_ID      Population  Assault_2014  Assault_2015  Assault_2016  Assault_2017  Assault_2018  Assault_2019  Assault_AVG  ...  TheftOver_20
count   140.0000  140.0000   140.000000    140.000000    140.000000    140.000000    140.000000    140.000000    140.000000   140.000000  ...    140.0000(
mean     70.5000   70.5000 19511.221429    117.350000    126.885714    132.200000    134.607143    138.935714    145.900000   132.646429  ...      9.6785
std      40.5586   40.5586 10033.589222    106.968961    119.613531    126.501014    137.168053    142.883824    148.500529   128.977375  ...     11.9183
min       1.0000    1.0000  6577.000000     16.000000     12.000000     10.000000     20.000000     15.000000     17.000000    18.500000  ...      0.0000(
25%      35.7500   35.7500 12019.500000     53.750000     57.250000     57.500000     58.750000     58.750000     62.000000    59.425000  ...      3.0000(
50%      70.5000   70.5000 16749.500000     85.000000     92.000000     97.000000     94.000000     93.500000     99.500000    96.500000  ...      6.0000(
75%     105.2500  105.2500 23854.500000    141.250000    152.250000    166.250000    159.500000    167.500000    172.000000   160.200000  ...     10.2500(
max     140.0000  140.0000 65913.000000    738.000000    826.000000    888.000000    905.000000    910.000000    916.000000   851.800000  ...     73.0000(

8 rows × 63 columns
```

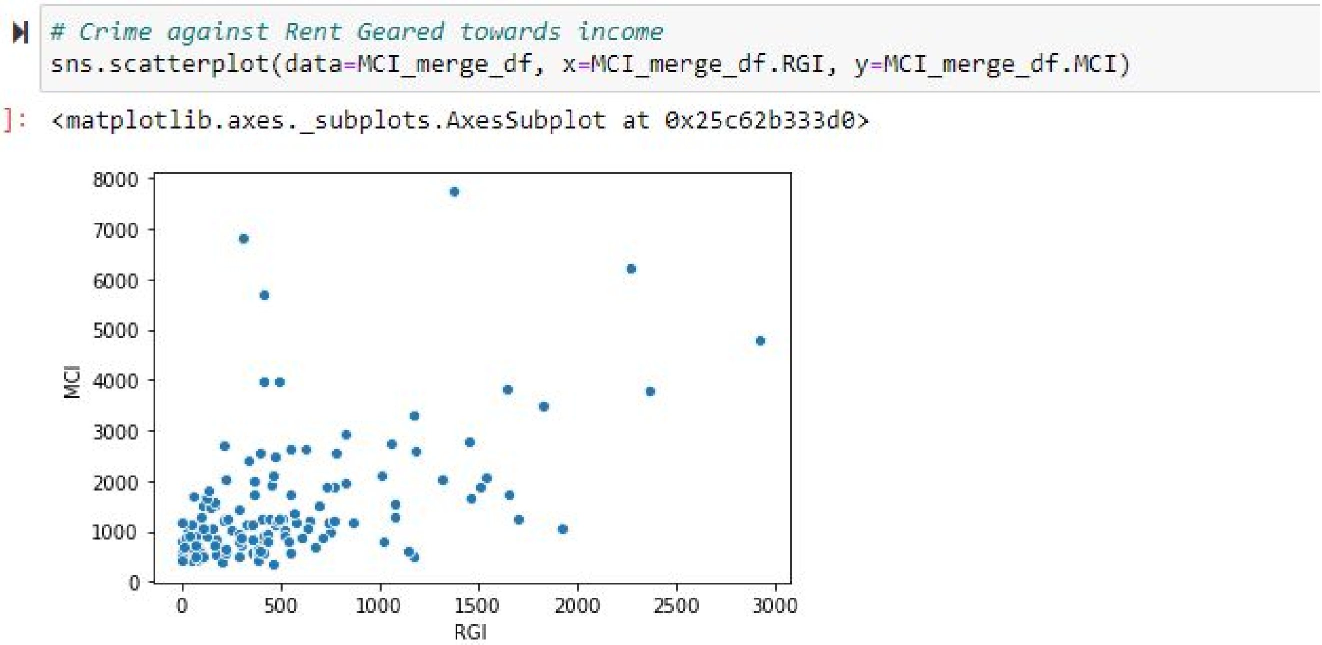
Analysis of Crime based on Rent Geared towards Income

A summary of the rent geared towards income (RGI) , Community Housing and Social Housing noting the low crime levels in areas without RGI's e.g. Forest Hill



	Hood_ID	Neighbourhood	offence	MCI	occurrenceyear	TTL_RES_UNIT	Soc_Units	RGI	Prop
72	73	Moss Park (73)	4786	4786	4786	2661	3399	2926	Top community units
72	73	Moss Park (73)	4786	4786	4786	2661	3399	2926	Highest Social Housing
76	77	Waterfront Communities-The Island (77)	7747	7747	7747	1615	2644	1378	Highest Crime
113	114	Lambton Baby Point (114)	353	353	353	462	668	463	Lowest Crime
72	73	Moss Park (73)	4786	4786	4786	2661	3399	2926	Max Rent Geared towards income
52	53	Henry Farm (53)	787	787	786	197	232	0	Min Rent Geared towards income
100	101	Forest Hill South (101)	494	494	494	290	174	0	Min Rent Geared towards income
128	129	Agincourt North (129)	1157	1157	1157	38	247	0	Min Rent Geared towards income
139	140	Guildwood (140)	411	411	411	570	48	0	Min Rent Geared towards income

The analysis below displayed neighbourhoods with lower rent geared to income units (RGI's) had lower tendencies towards crime. Here MCI - Crime was calculated as a count of the total reported crimes between 2014 and 2019.

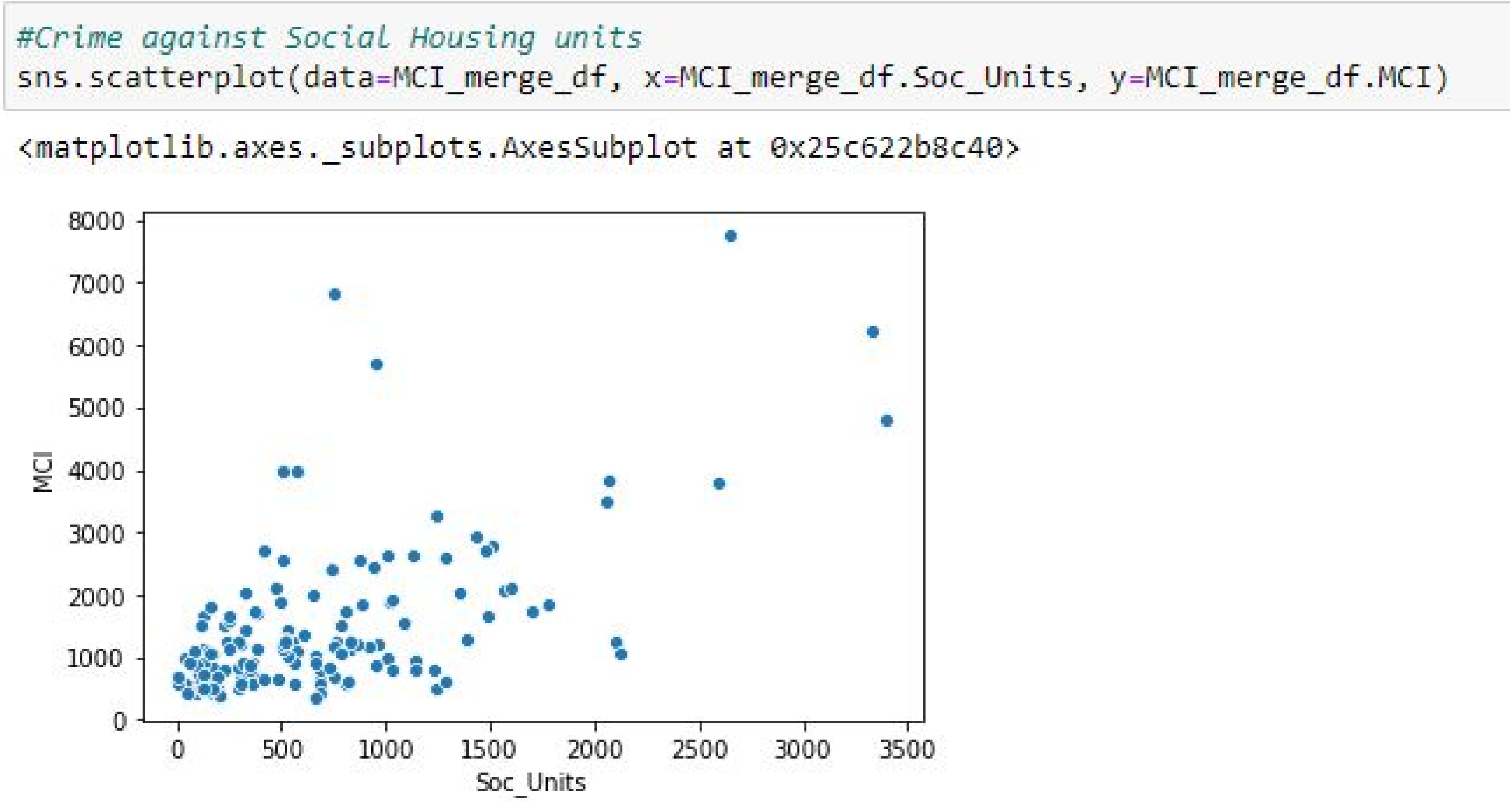


```
# Crime against Rent Geared towards income
sns.scatterplot(data=MCI_merge_df, x=MCI_merge_df.RGI, y=MCI_merge_df.MCI)
```

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x25c62b333d0>
```

Fig 16: Crime against Rent Geared towards Income

Scatterplot below addition shows that Social Housing units (Inclusive of Community Housing, City of Toronto public housing also demonstrated a similar trend in our analysis. We therefore concluded that RGI and Social housing units had some impact on crime.



```
#Crime against Social Housing units
sns.scatterplot(data=MCI_merge_df, x=MCI_merge_df.Soc_Units, y=MCI_merge_df.MCI)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x25c622b8c40>
```

Fig 17: Crime against Social Housing Units

Analysis of Crime based on Household Income

Scatterplot below shows the relationship between Average household Income on Crime Rate of Toronto Neighbourhoods and its shows that whenever we have lower household Income the crime rate is high. Whereas the high Income neighbourhoods have lower crime rates.

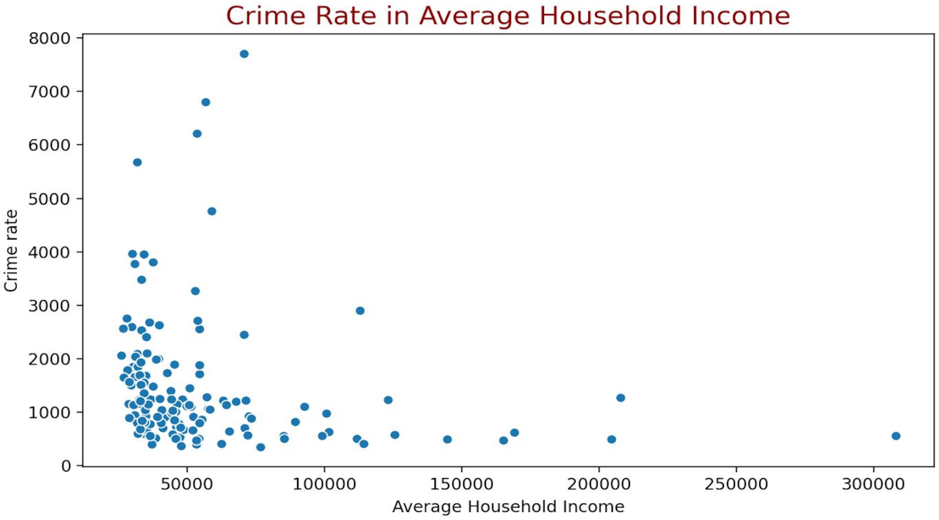


Fig 18: Crime against Average Household Income.

The analysis below displayed lower Average Household Income had a higher tendency towards crime.

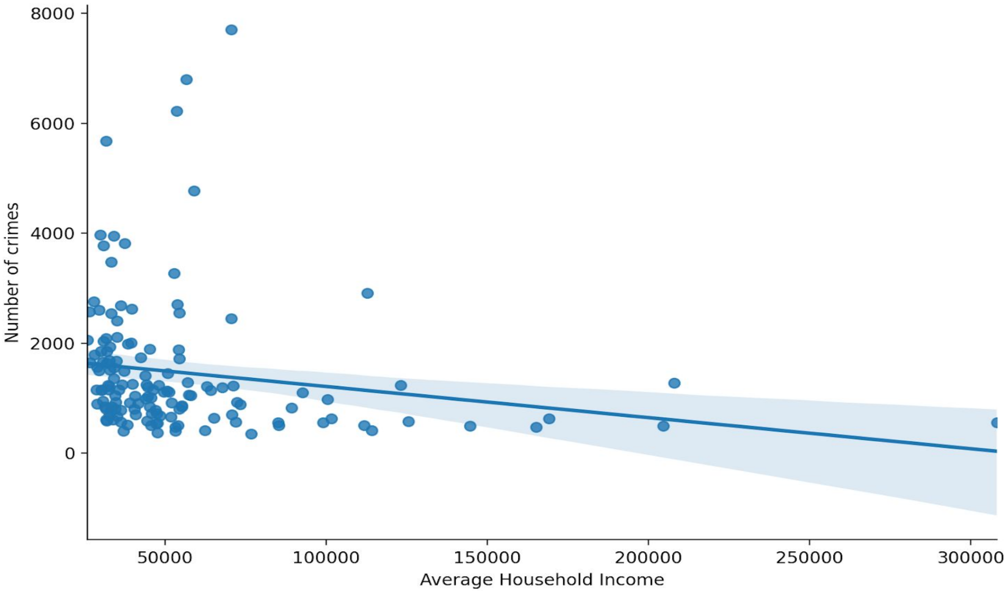


Fig 19: Crime against Average Household Income.

Modeling and Evaluation

The exploratory portion of this study has merely given a general perception of the distribution of crime across the city. However, to explore various types of individual and neighbourhood socioeconomic characteristics that impact property, Income education and violent crime rates. After understanding the effects of individual variables on rates of crimes in the city of Toronto from exploratory data analysis, different methods of modelling were used to predict the rates of crime. The features for the models were selected based on the importance of each feature from the exploratory data analysis and also using the inbuilt functions available in the sk-learn library.

For **data cleaning**, the features with string type and time type (year, day, week) were converted to the category type. All the null and duplicate values were removed from the data sets. The crime category type was set as the target variable to be predicted.

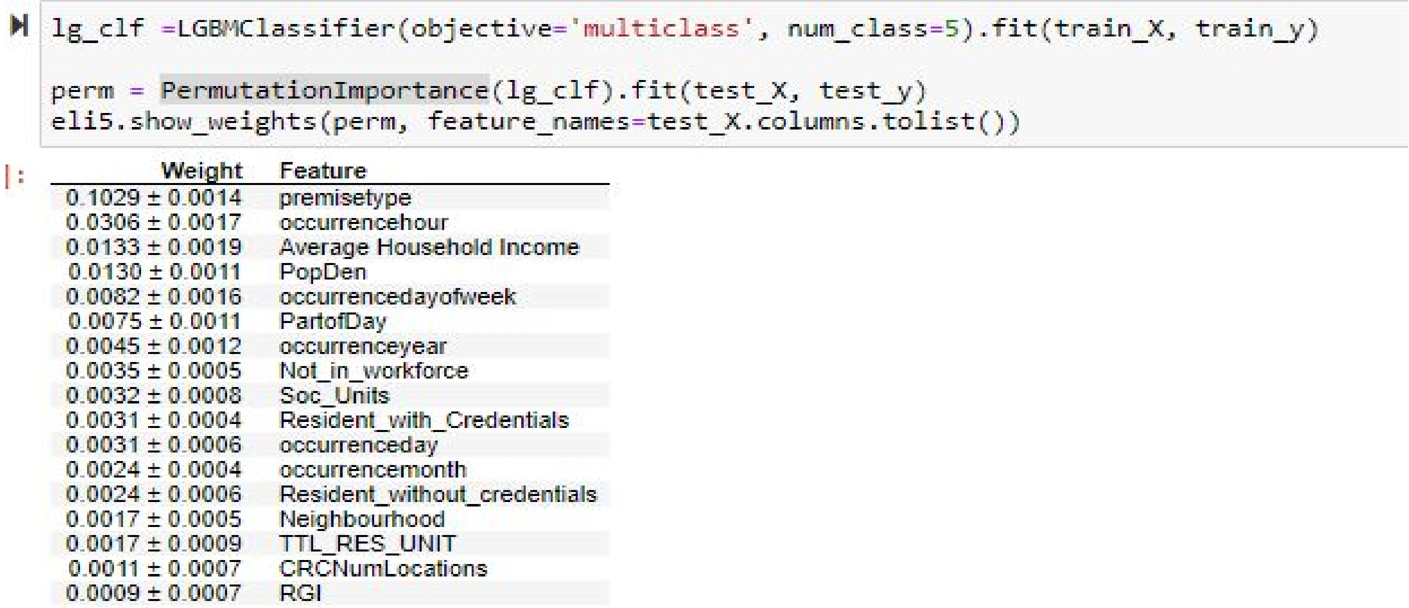
For **feature engineering**, all the categorical variables were encoded with dummy values. A new feature named Part of the Day with six categories was created for the model. This was done to minimize the number of categories for the occurrence hour (24 to 6). Furthermore, the columns with the numerical columns were standardized to have the mean equals to zero and variance equals to 1 using the function **sklearn.preprocessing.StandardScaler().**

Supervised Learning:

The open source library sklearn was used to deploy different supervised learning methods in python. Different supervised learning methods such random forest, knn classifier, logistic and linear regression, lightgbm were used to train the model to be used for the prediction of crime type in the city of Toronto.

Feature selection for the models can be done in a variety of methods but we decided to use the Permutation Importance method for our models. When using Permutation importance it directly measures variable importance by observing the effect on model accuracy by randomly shuffling each predictor variable. This technique is broadly-applicable because it doesn't rely on internal model parameters, such as linear regression coefficients. Based on the exploratory data analysis we used input features ['premisetype', 'occurrenceyear', 'occurrencemonth', 'occurrenceday', 'occurrencedayofweek', 'PopDen', 'Resident_with_Credentials', 'Resident_without_credentials', 'Average Household Income', 'TTL_RES_UNIT', 'Soc_Units', 'CRCNumLocations', 'Long', 'Lat', 'PartofDay'] for the model and used the 'MCI' as the output feature.

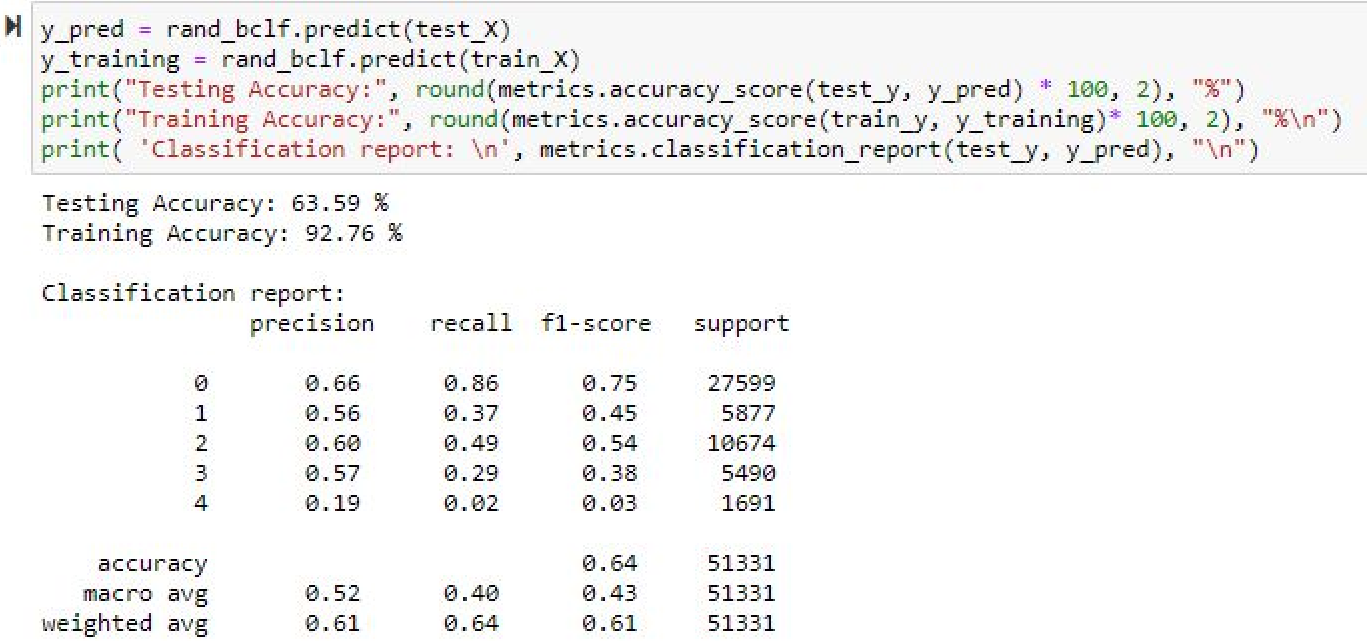
Below is the chart obtain when this was done on our existing features:



```
lg_clf =LGBMClassifier(objective='multiclass', num_class=5).fit(train_X, train_y)

perm = PermutationImportance(lg_clf).fit(test_X, test_y)
eli5.show_weights(perm, feature_names=test_X.columns.tolist())
```

Weight	Feature
0.1029 ± 0.0014	premisetype
0.0306 ± 0.0017	occurrencehour
0.0133 ± 0.0019	Average Household Income
0.0130 ± 0.0011	PopDen
0.0082 ± 0.0016	occurrencedayofweek
0.0075 ± 0.0011	PartofDay
0.0045 ± 0.0012	occurrenceyear
0.0035 ± 0.0005	Not_in_workforce
0.0032 ± 0.0008	Soc_Units
0.0031 ± 0.0004	Resident_with_Credentials
0.0031 ± 0.0006	occurrenceday
0.0024 ± 0.0004	occurrencemonth
0.0024 ± 0.0006	Resident_without_credentials
0.0017 ± 0.0005	Neighbourhood
0.0017 ± 0.0009	TTL_RES_UNIT
0.0011 ± 0.0007	CRCNumLocations
0.0009 ± 0.0007	RGI

After selecting the features we used the random forest classifier to predict the category of the crime. The data set was divided into the training and test set randomly with a train to test set ratio 70 to 30 percent. The random forest classifier had an accuracy of 52% on the test set data when the default parameters were used. In order to hypertune the parameters, we used the **GridSearchCV** method with a dictionary of range of different parameters for the classifier. After hypertuning the parameters, the accuracy on the training set went up to 93% while the accuracy on the test set went up 63%. The lower value of accuracy on the test set shows that the model is overfitting.



```
y_pred = rand_bclf.predict(test_X)
y_training = rand_bclf.predict(train_X)
print("Testing Accuracy:", round(metrics.accuracy_score(test_y, y_pred) * 100, 2), "%")
print("Training Accuracy:", round(metrics.accuracy_score(train_y, y_training)* 100, 2), "%\n")
print( 'Classification report: \n', metrics.classification_report(test_y, y_pred), "\n")
```

```
Testing Accuracy: 63.59 %
Training Accuracy: 92.76 %

Classification report:
               precision    recall  f1-score   support

           0       0.66      0.86      0.75     27599
           1       0.56      0.37      0.45      5877
           2       0.60      0.49      0.54     10674
           3       0.57      0.29      0.38      5490
           4       0.19      0.02      0.03      1691

    accuracy                           0.64     51331
   macro avg       0.52      0.40      0.43     51331
weighted avg       0.61      0.64      0.61     51331
```

We speculated one cause of the overfitting can be the encoding of different categories. The sklearn **LabelEncoder** method was used to encode the categories into numeric values. This method encodes the categories using the numbers 0 to C-1 where C is the total number of categories. This method has a pitfall that it will give more weightage to the categories with numbers higher in values than the categories whose number representations has a lower value. To deal with this problem, we used the **OneHotEncoder** method to convert all the categories into dummy variables with categories one and zeros. After one hot encoding the test set accuracy went up to 65 percent from the 63 percent though the training set accuracy was decreased from 93 percent to 82 percent. Although the one hot encoding method decreased the overfitting but no significant improvement in the accuracy of the test set was observed. Following are the different methods used to predict the category of the crime in the city of Toronto

Method	Test set accuracy
Linear regression	0.0032
Logistic regression	0.58
Gradient boosting decision tree	0.57
K nearest neighbour classifier	0.63
Support vector classification	0.54
Decision tree classifier	0.515
Naïve bayes classifier	0.39
Random forest classifier	0.65
Neural network with 3 hidden layers	0.59

It was observed that the random forest method performed better than the other methods based on the values of the accuracy on the test set.

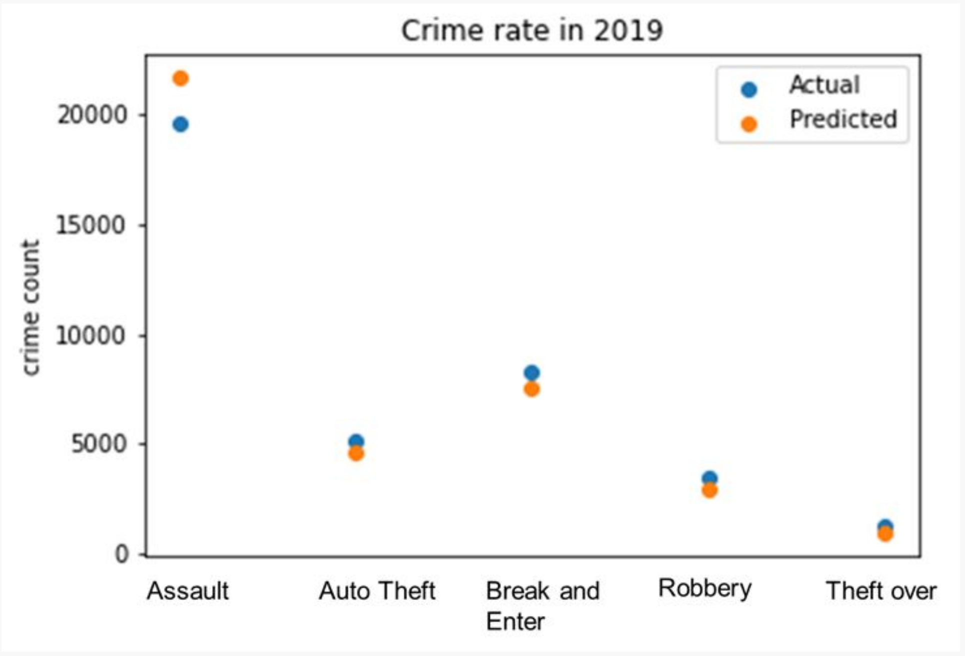
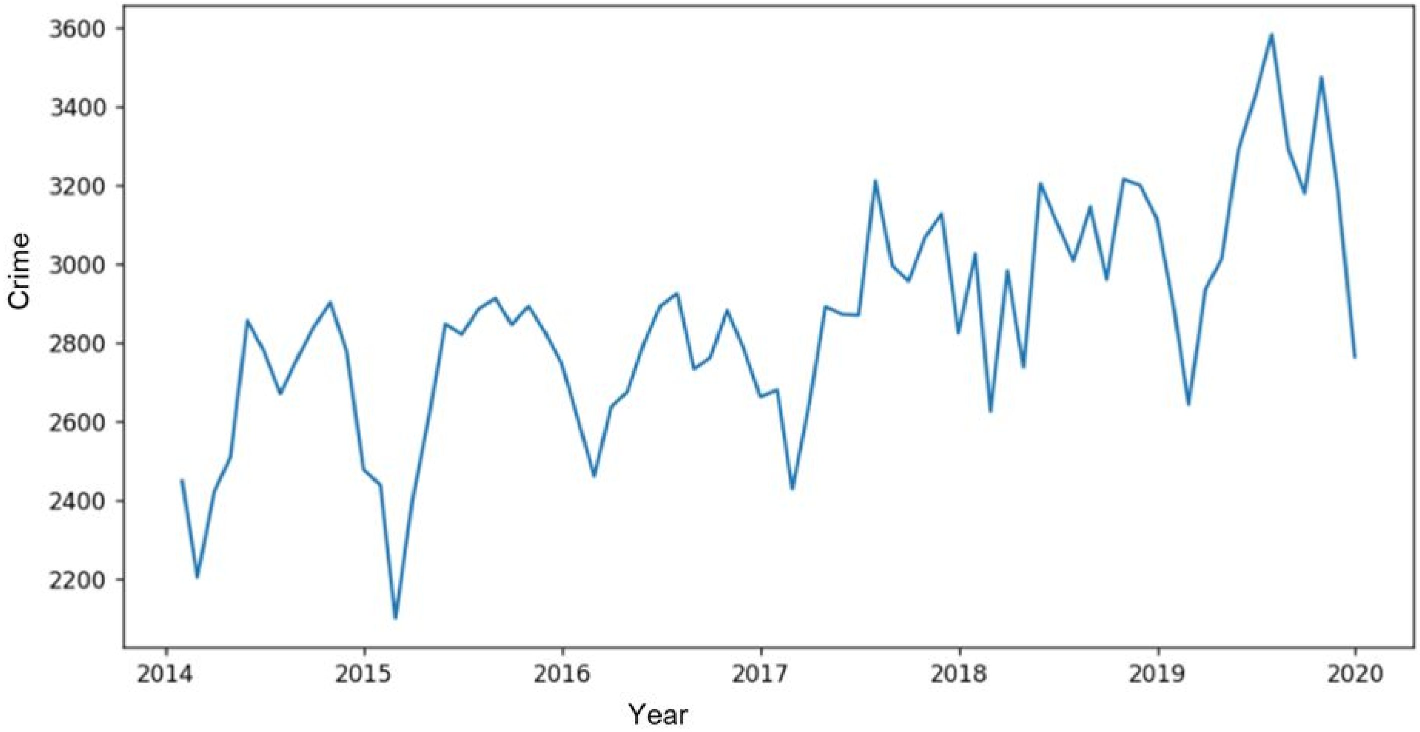


Fig 20: Predicted and Actual crime rates.

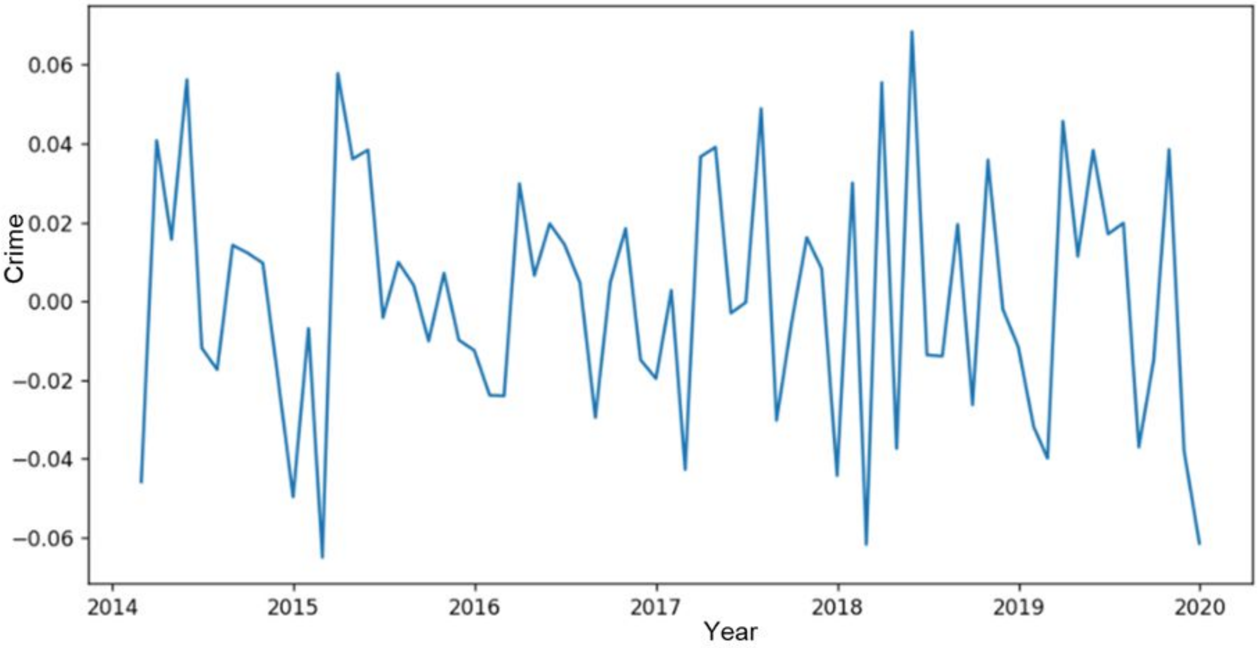
We observed that using a single model to classify the crime into five categories (0,1,2,3,4) yields a test set accuracy of 0.65 and training set accuracy of 0.99. This shows that the model may be overfitting on the data. Also, the random forest classifier was found to have the best performance among all the considered methods of supervised learning for this study. To improve the accuracy of the model on the test set we tried 5 models instead of one model for all the categories. These five models were constructed by considering only one category at a time. For this, we encoded the considered category 1 and all the other categories as 0. It should be noted that this process may result in the class imbalance because the number of 1's in the data is far greater than the number of 0's. To remove the class imbalance we used **Random under sampling** present in the sklearn library. This function randomly selects the data from the original data with the same number of data points for both the classes. Following table shows the test set accuracy for each of the models considered for each of the classes.

Category	Test set accuracy
Auto theft	0.79
Break and enter	0.77
Theft over	0.76
Robbery	0.75
Assault	0.7

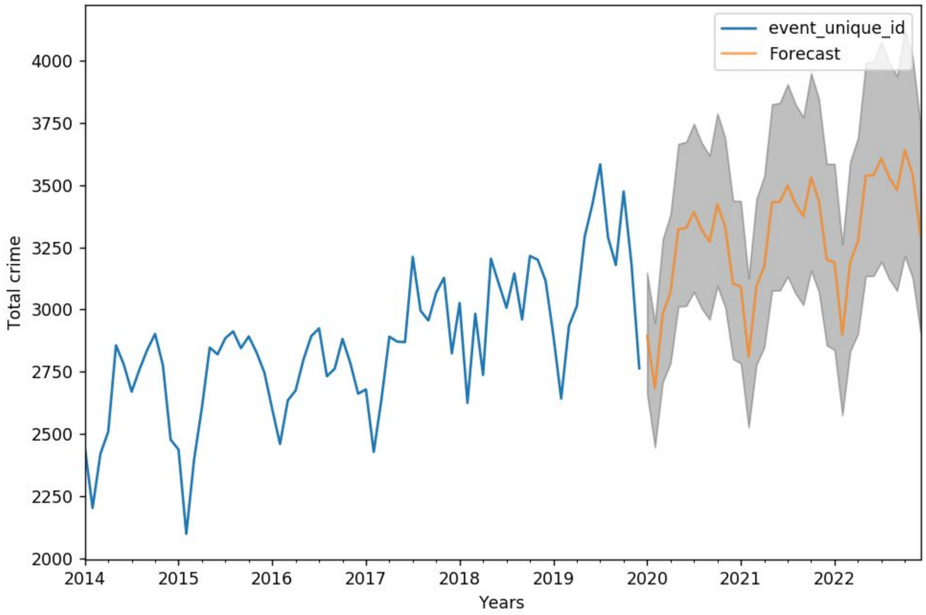
Here we can see that the accuracy on the test set has been increased from 0.65 to 0.79 for the auto theft category and also significantly for all other categories. The accuracy for the test was 0.79 for the new random forest classifier trained on auto theft category while the accuracy on the training set was found to be 0.98. Therefore, we conclude that the new models improve the test set accuracy though the models still have the problem of overfitting. To get further insight into the modelling process and to forecast the crime rate for the future 3 years we also carried out the time series analysis using the ARIMA model.



The above figure shows the crime rate with a monthly frequency. We can see that the crime rate increases with time. In order to use the ARIMA model for crime forecasting, the mean and variance of the time series should be constant with time. To make the mean and variance of the series constant the series was transformed. The series was made stationary by subtracting the previous crime and also by making the crime scale as logarithmic as shown in the following figure.

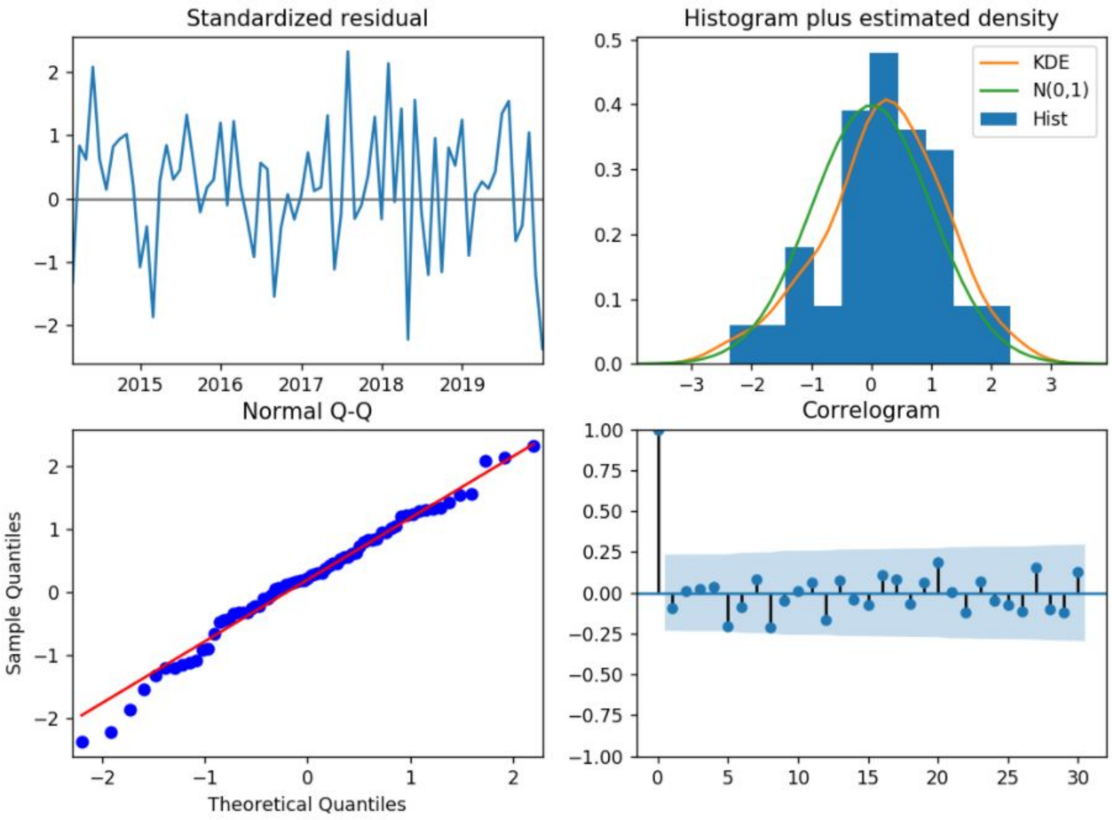


After making the series stationary the ARIMA model was used to forecast the crime in the city of Toronto. The following figure shows the forecasted crime for the city of Toronto.



It can be seen from the time series analysis that the crimes in the city of Toronto are expected to increase with the time.

The following figure shows the different plots for the performance evaluation matrix for the ARIMA model. It can be observed from the standard residual plot that there is no significant correlation between the residuals of the model. Also, the KDE plot shows that the mean is zero and the distribution is normal for the model. This further confirms the accuracy of the model. The bottom left figure shows that all the predicted values line on the redline which is a sign of good accuracy. Finally, the correlogram (bottom right) plot shows that the residuals from the ARIMA model do not have major correlation between them which again is a good sign for the accuracy of the model.



Unsupervised Learning

Cluster Analysis:

Our approach to Cluster Analysis was governed by two questions:

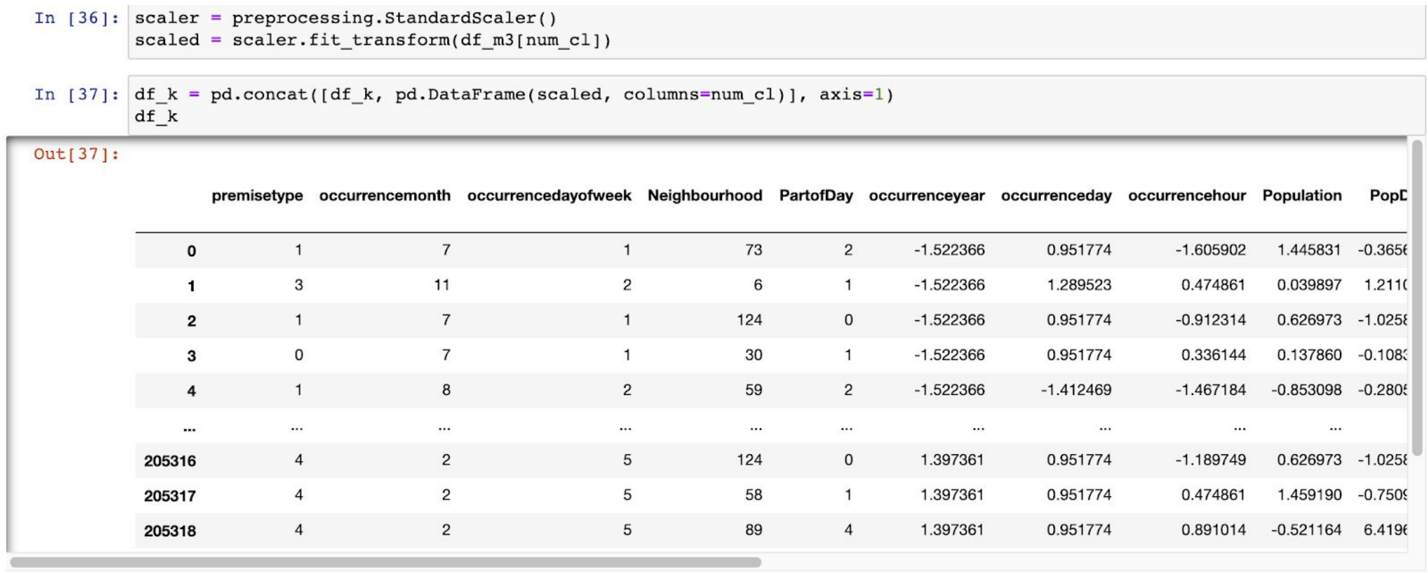
1) Does the presence or absence of social services impact crime in a neighbourhood?

2) What role does demographics play when investigating the crime rate of a neighbourhood?

Although we were governed by those two questions, we also used other tools to do wholesale discovery of the data.

Preparation for cluster analysis

As seen below, in our preparatory steps for Clustering we normalized the data using the StandardScaler function which is found in the sklearn library.



```
In [36]: scaler = preprocessing.StandardScaler()
         scaled = scaler.fit_transform(df_m3[num_cl])
```

```
In [37]: df_k = pd.concat([df_k, pd.DataFrame(scaled, columns=num_cl)], axis=1)
         df_k
```

Out[37]:

	premisetype	occurrencemonth	occurrencedayofweek	Neighbourhood	PartofDay	occurrenceyear	occurrenceday	occurrencehour	Population	PopD
0	1	7	1	73	2	-1.522366	0.951774	-1.605902	1.445831	-0.3656
1	3	11	2	6	1	-1.522366	1.289523	0.474861	0.039897	1.2110
2	1	7	1	124	0	-1.522366	0.951774	-0.912314	0.626973	-1.0258
3	0	7	1	30	1	-1.522366	0.951774	0.336144	0.137860	-0.1083
4	1	8	2	59	2	-1.522366	-1.412469	-1.467184	-0.853098	-0.2805
...	...	...	...	...	...	...	...	...	...	
205316	4	2	5	124	0	1.397361	0.951774	-1.189749	0.626973	-1.0258
205317	4	2	5	58	1	1.397361	0.951774	0.474861	1.459190	-0.7509
205318	4	2	5	89	4	1.397361	0.951774	0.891014	-0.521164	6.4196

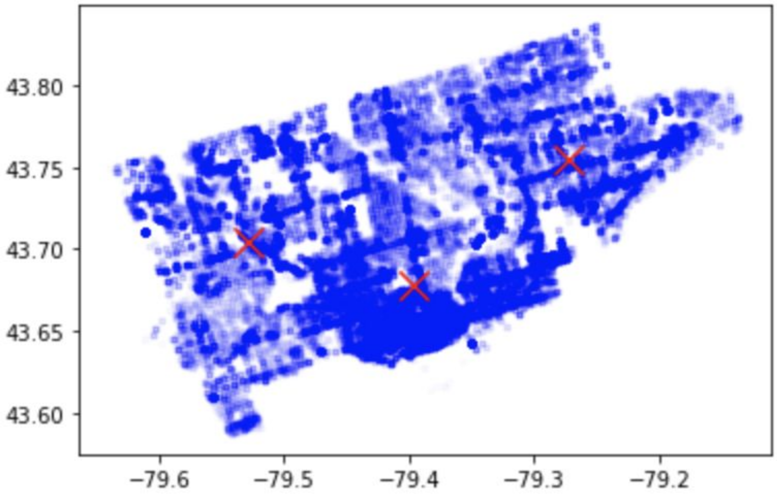
Social services and demographics

After normalization, we started the process of cluster analysis by randomly assigning k, the number of clusters, to 3 and from the attribute featureset we selected two features::

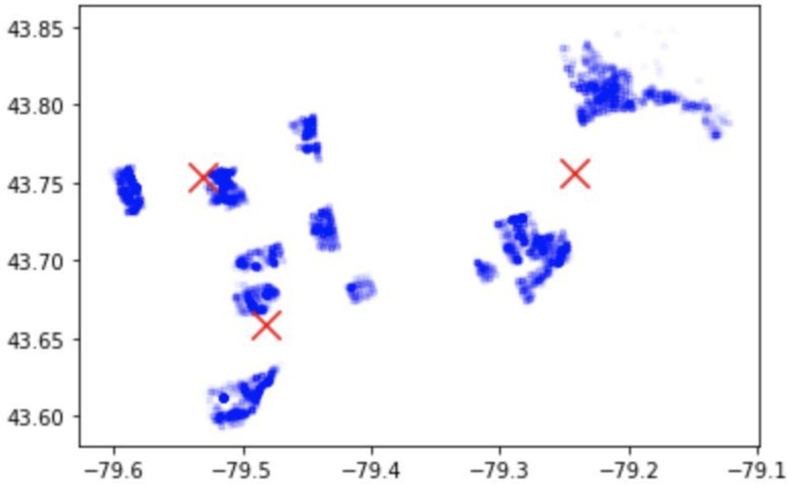
- Number of recreation centers in given neighbourhood
- Population density in given neighbourhood

In an attempt to obtain better clusters we modified the value indicating the number of recreation centers per neighbourhood. In the dataset the value ranged from 0 to 5 but we decided to make it categorial with the governing condition being greater than or equal to 2. Again, based on the data, the boundary of 2 instead of 3 was chosen for better clustering.

The images below show the clustering transition when considering two or less rec centers.



Neighbourhoods with more than 2 rec centers:

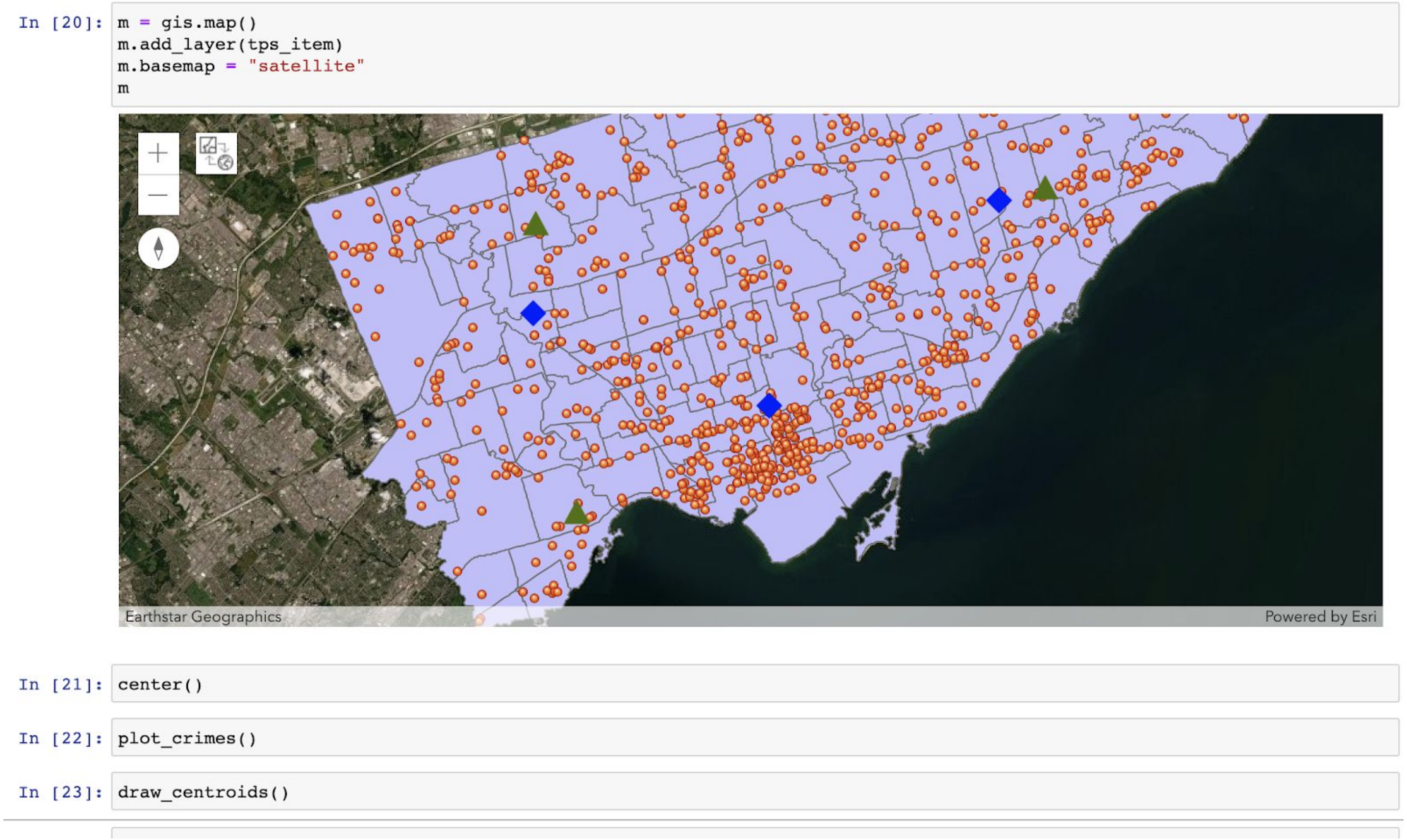


Based on the initial results, we decided to apply evaluation methods such as elbow or silhouette to derive a better estimation of K. However, even in the initial results, the objective to see clustering patterns caused by the impact for the attribute is met i.e. neighbourhoods with more rec centers seem to have less crime.

To further visualize this we utilized ESRI ArcGIS to map layers on top of each other.
First, we took base mapping from Toronto Police in the form of their map layer published via ESRI ArcGIS. The REST endpoint for that layer: https://services.arcgis.com/S9th0jAJ7bqgIRjw/ArcGIS/rest/services/Neighbourhood_MCI/FeatureServer

Second, we overlaid the above image with a plot of crimes between the years 2014 to 2019 to visualize the individual crimes and get a better understanding of the overall densities of crime.

Third, we overlaid the centroids for crime where neighbourhoods had recreation centers less than or equal to 2 (represented in blue) and those with 3 or more recreation centers (represented in green)



```
In [20]: m = gis.map()
         m.add_layer(tps_item)
         m.basemap = "satellite"
         m
```

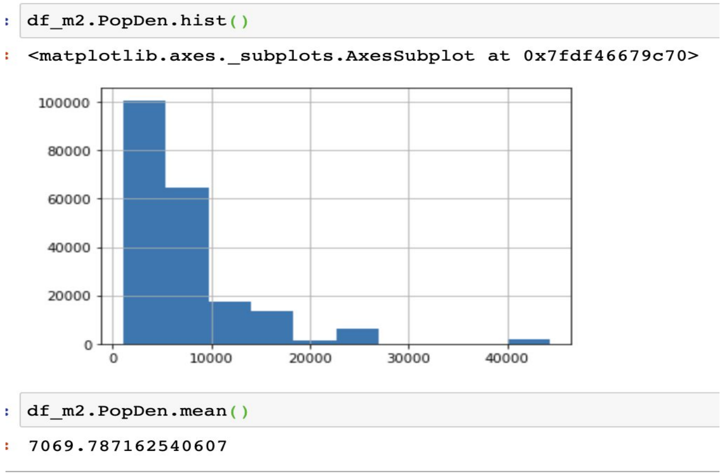
```
In [21]: center()
```

```
In [22]: plot_crimes()
```

```
In [23]: draw_centroids()
```

Interestingly, there are some areas where a neighbourhood high in crime is in close proximity to those in low in crime. Using ArcGIS to geocode the exact addresses of these centroids, it was observed indeed that there were certain neighbourhoods with high crime reputation next to low crime neighbourhoods that could be literally blocks away.

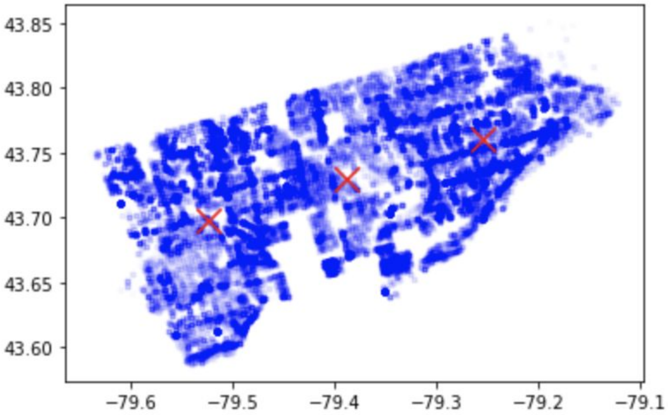
To further explore this, we looked at demographic information such as the population density.
In this exploration, the mean value for all the population densities per neighbourhood and the plots below generated.



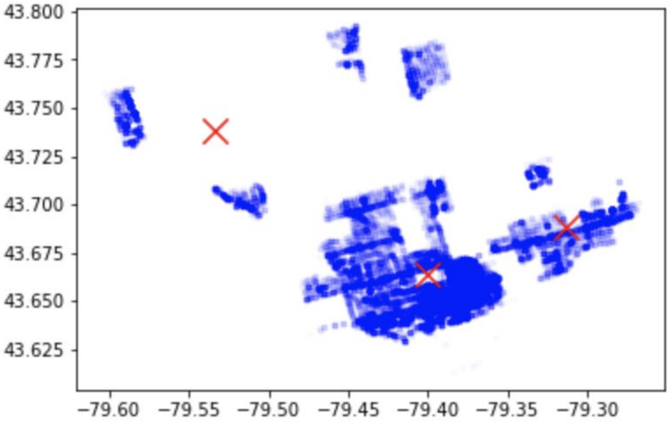
```
: df_m2.PopDen.hist()
: <matplotlib.axes._subplots.AxesSubplot at 0x7fdf46679c70>
```

```
: df_m2.PopDen.mean()
: 7069.787162540607
```

Neighbourhoods with population density greater than the mean:



Neighbourhoods with population density less than the mean:



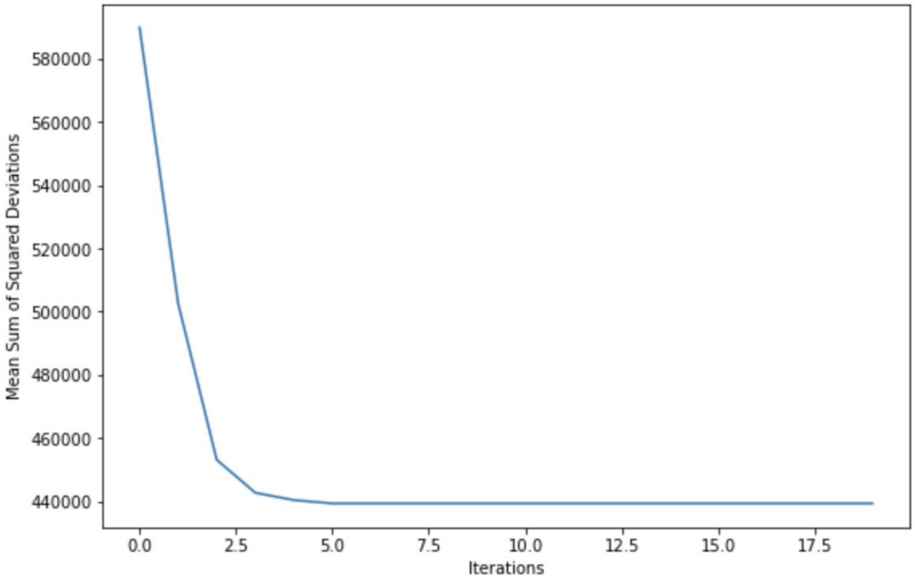
Determine optimal number of clusters

We then turned our attention to clustering on all attributes and the decision was made to use functions from python libraries to assist in the determining of clusters. This was done using the elbow criterion and silhouette coefficient.

When using the elbow criterion, the best k value is obtained by using an approximation value that coincides with the elbow or knee of the curve. This is the point where the curve visibly bends, specifically from high slope to low slope (flat or close to flat), or in the other direction.

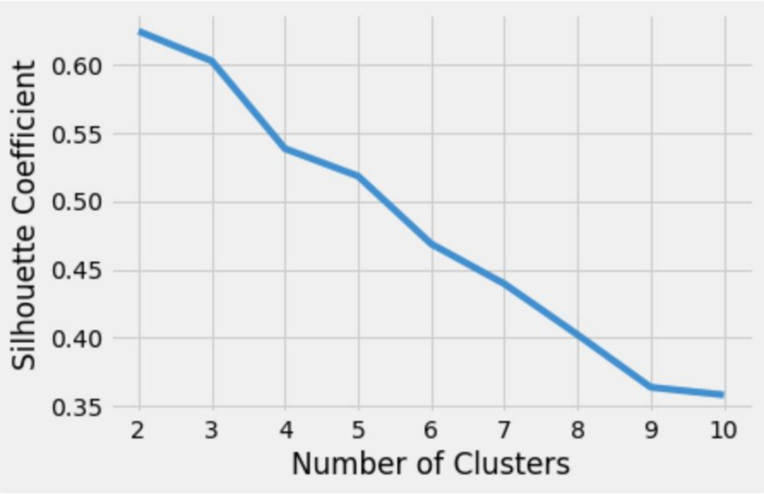
The silhouette coefficient refers to a method of interpretation that gives a graphical representation of how well each object has been classified. Its values ranges from -1 to 1 and tells how similar an object is to other objects in its own cluster when compared to other clusters. A high coefficient value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

The Elbow Curve Plot.



The Elbow curve shows K would be somewhere between 2 and 5 and a best approximation of K=3. This value coincided with our initial guess of k for our cluster plot.

The Silhouette Coefficient Plot

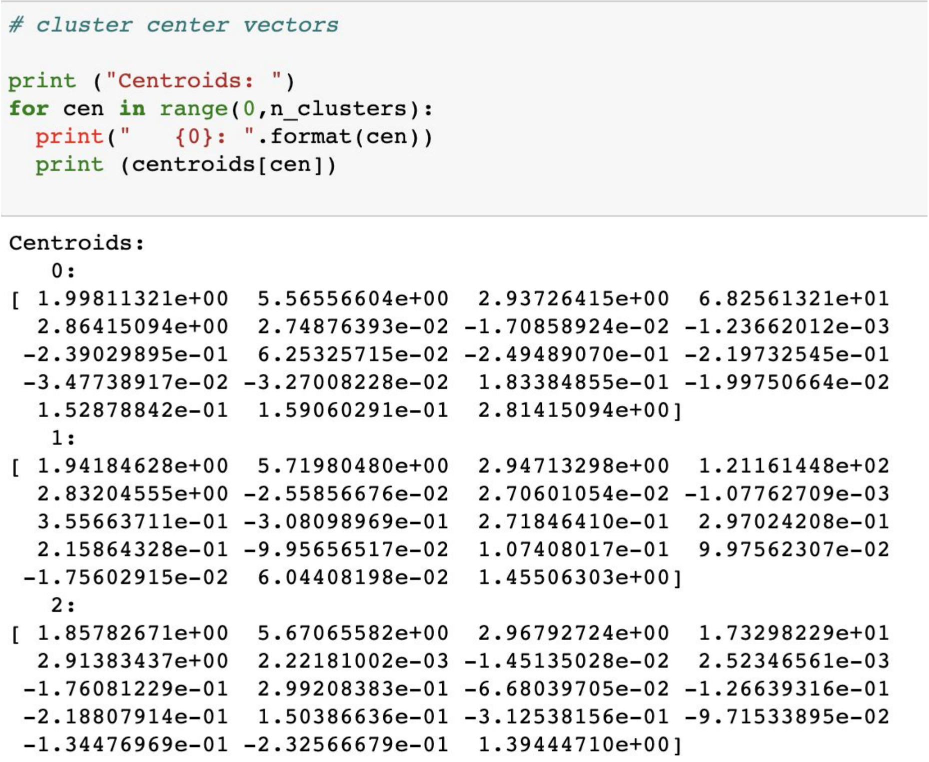


Silhouette coefficient method shows inconclusive result:

Clustering using K-means method

This is done by applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means thereby classifying new data into the existing clusters. This is known as the nearest centroid classifier.

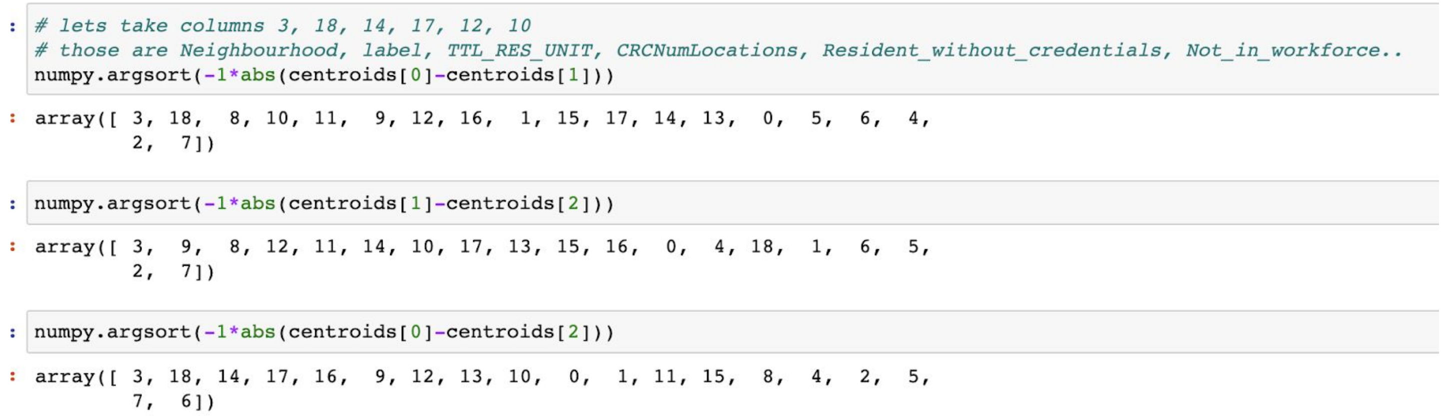
Compared the distance between the centroids to see if clear separation between clusters but also which attributes were most impactful to those distances:



```
# cluster center vectors

print ("Centroids: ")
for cen in range(0,n_clusters):
  print("   {0}: ".format(cen))
  print (centroids[cen])
```

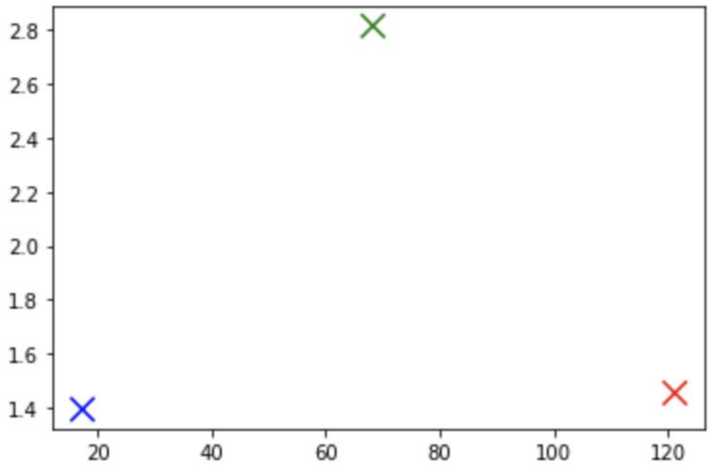
```
Centroids:
   0:
[ 1.99811321e+00  5.56556604e+00  2.93726415e+00  6.82561321e+01
  2.86415094e+00  2.74876393e-02 -1.70858924e-02 -1.23662012e-03
 -2.39029895e-01  6.25325715e-02 -2.49489070e-01 -2.19732545e-01
 -3.47738917e-02 -3.27008228e-02  1.83384855e-01 -1.99750664e-02
  1.52878842e-01  1.59060291e-01  2.81415094e+00]
   1:
[ 1.94184628e+00  5.71980480e+00  2.94713298e+00  1.21161448e+02
  2.83204555e+00 -2.55856676e-02  2.70601054e-02 -1.07762709e-03
  3.55663711e-01 -3.08098969e-01  2.71846410e-01  2.97024208e-01
  2.15864328e-01 -9.95656517e-02  1.07408017e-01  9.97562307e-02
 -1.75602915e-02  6.04408198e-02  1.45506303e+00]
   2:
[ 1.85782671e+00  5.67065582e+00  2.96792724e+00  1.73298229e+01
  2.91383437e+00  2.22181002e-03 -1.45135028e-02  2.52346561e-03
 -1.76081229e-01  2.99208383e-01 -6.68039705e-02 -1.26639316e-01
 -2.18807914e-01  1.50386636e-01 -3.12538156e-01 -9.71533895e-02
 -1.34476969e-01 -2.32566679e-01  1.39444710e+00]
```

It appears that ‘Neighbourhood’ was the attribute most consistently cause for the distance between centroids:



```
: # lets take columns 3, 18, 14, 17, 12, 10
  # those are Neighbourhood, label, TTL_RES_UNIT, CRCNumLocations, Resident_without_credentials, Not_in_workforce..
  numpy.argsort(-1*abs(centroids[0]-centroids[1]))

: array([ 3, 18,  8, 10, 11,  9, 12, 16,  1, 15, 17, 14, 13,  0,  5,  6,  4,
          2,  7])

: numpy.argsort(-1*abs(centroids[1]-centroids[2]))

: array([ 3,  9,  8, 12, 11, 14, 10, 17, 13, 15, 16,  0,  4, 18,  1,  6,  5,
          2,  7])

: numpy.argsort(-1*abs(centroids[0]-centroids[2]))

: array([ 3, 18, 14, 17, 16,  9, 12, 13, 10,  0,  1, 11, 15,  8,  4,  2,  5,
          7,  6])
```

Mapping centroids:

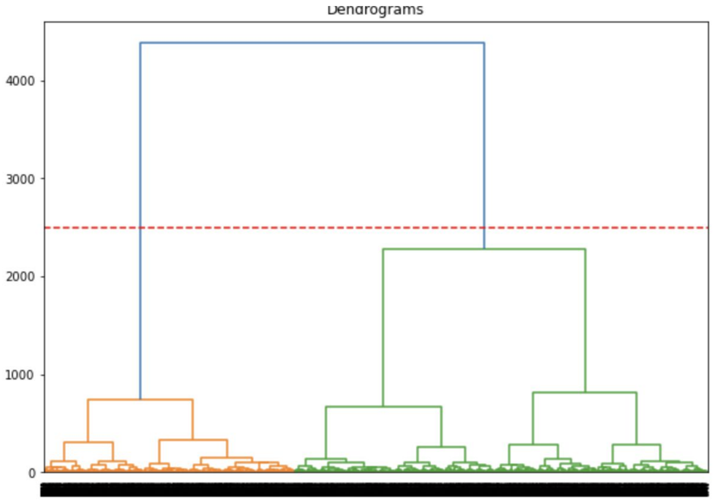


We also did K Means on K= 5 with similar results

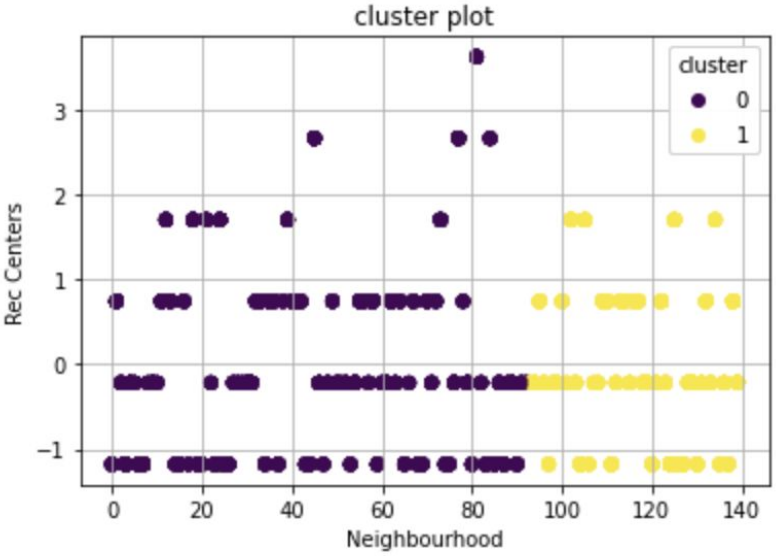
Using hierarchical method (dendrogram):

A Dendrogram arranges clusters in a hierarchical manner in what is called Hierarchical clustering. In this type of clustering and algorithm groups similar objects into clusters and the endpoint is reached when the final set of clusters are distinct from each other and objects in the same cluster similar to each other.

When this was applied to our dataset it appears that there are 2 clusters based on the dendrogram seen below. The biggest separation between clusters is denoted by the two blue lines.

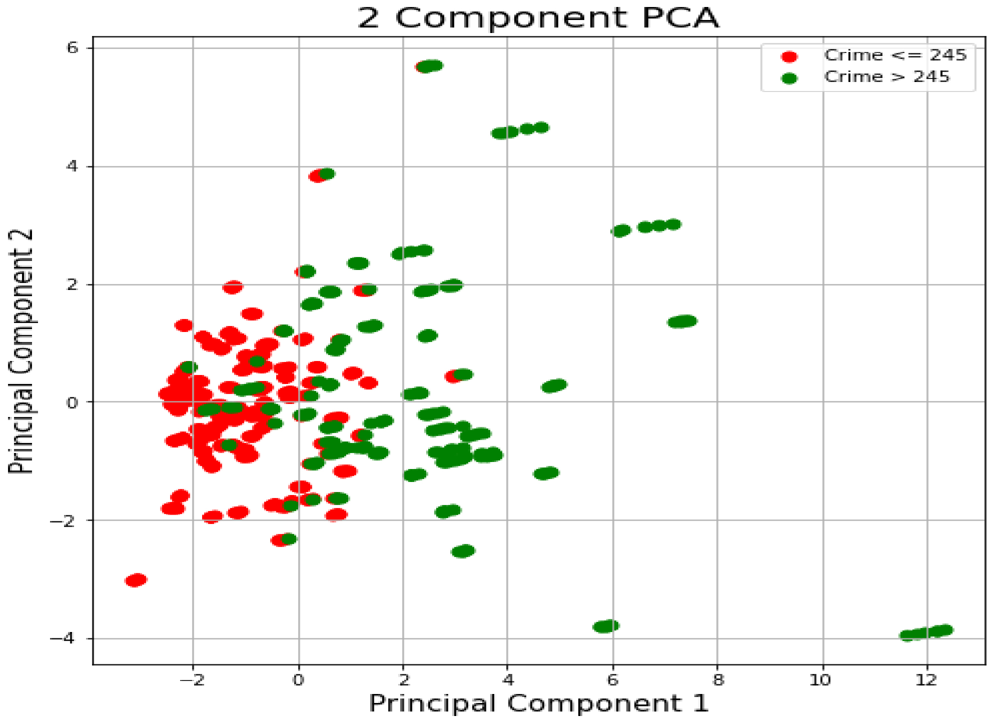


Based on the 2 clusters, plotted by arbitrarily choosing 2 attributes from each cluster to plot and visualize the separation of the clusters:



Principal Component Analysis (PCA)

We looked at Principal Component Analysis (PCA) however our visualization did not show significant separation of our clusters after application of eigenvalues and eigenvectors. We applied two dimensional or two components which yielded the resultant visualization below:



The most separation displayed was with the use of the following target variables:

- *MCI (total crime) < = 245, and*
- *MCI (total crime) > 245*

Features used on our PCA

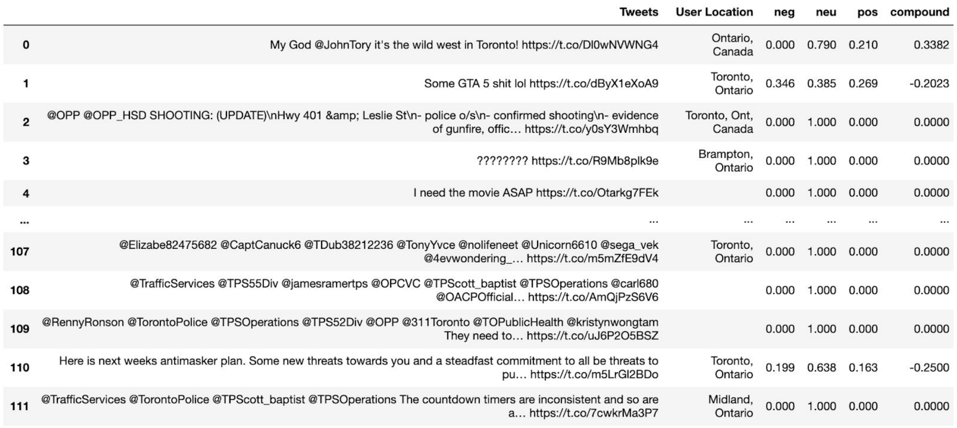
- *Neighbourhood,*
- *Crime (MCI),*
- *Not_in_workforce*
- *Resident_with_Credentials*
- *Resident_without_credentials*
- *15_to_24_WithCred' (Education)*
- *'25_to_65_WithCred',*
- *Average Household Income*
- *TTL_RES_UNIT' (Total residential Units)*
- *Soc_Units - Social Housing Units (includes Toronto housing)*
- *'RGI', MCI - Crime counts*

Real time feeds

As a side tangent, we explored the possibility of incorporating unstructured data in the form of social media to supplement the data. The TPS data is historical and when used for modeling and ultimately predicting, it is based on a specific point in time that has passed. The thought was to add another source which provided real time updates as they happened.

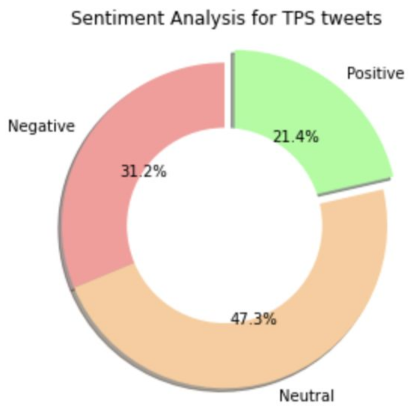
Twitter was identified as a possible source for current crime data as they occurred since it was thought that concerned citizens would tweet about occuring or potential crimes.

Using Twitter API and Tweepy, we were able to extract crimes from Twitter. Applied Natural Language Processing in the form of sentiment analysis to arrive at an overall sentiment, positive or negative (demonstrated in the compound column below, -1 representing negative and +1 representing positive)

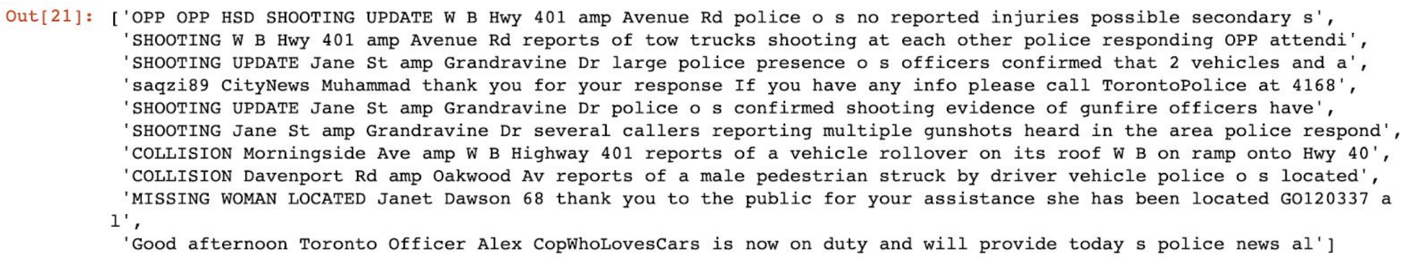


	Tweets	User Location	neg	neu	pos	compound
0	My God @JohnTory it's the wild west in Toronto! https://t.co/Dl0wNVWNG4	Ontario, Canada	0.000	0.790	0.210	0.3382
1	Some GTA 5 shit lol https://t.co/dByX1eXoA9	Toronto, Ontario	0.346	0.385	0.269	-0.2023
2	@OPP @OPP_HSD SHOOTING: (UPDATE)\nHwy 401 &amp; Leslie St\n- police o/s\n- confirmed shooting\n- evidence of gunfire, offic… https://t.co/y0sY3Wmhbq	Toronto, Ont, Canada	0.000	1.000	0.000	0.0000
3	???????? https://t.co/R9Mb8plk9e	Brampton, Ontario	0.000	1.000	0.000	0.0000
4	I need the movie ASAP https://t.co/Otarkg7FEk		0.000	1.000	0.000	0.0000
...	...	...	...	...	...	...
107	@Elizabe82475682 @CaptCanuck6 @TDub38212236 @TonyYvce @nolifeneet @Unicorn6610 @sega_vek @4evwondering_… https://t.co/m5mZfE9dV4	Toronto, Ontario	0.000	1.000	0.000	0.0000
108	@TrafficServices @TPS55Div @jamesramertps @OPCVC @TPScott_baptist @TPSOperations @carl680 @OACPOfficial… https://t.co/AmQjPzS6V6		0.000	1.000	0.000	0.0000
109	@RennyRonson @TorontoPolice @TPSOperations @TPS52Div @OPP @311Toronto @TOPublicHealth @kristynwongtam They need to… https://t.co/uJ6P2O5BSZ		0.000	1.000	0.000	0.0000
110	Here is next weeks antimasker plan. Some new threats towards you and a steadfast commitment to all be threats to pu… https://t.co/m5LrGl2BDo	Toronto, Ontario	0.199	0.638	0.163	-0.2500
111	@TrafficServices @TorontoPolice @TPScott_baptist @TPSOperations The countdown timers are inconsistent and so are a… https://t.co/7cwkrMa3P7	Midland, Ontario	0.000	1.000	0.000	0.0000

Upon further investigation the decision was made to forego real-time analysis as the volume of data was miniscule as not enough citizens reported crime via text so sentiment analysis did not have any real impact.



It was observed that Toronto Police Operations themselves reported crimes with more structure on Twitter, notably that crimes would be reported with the first word(s) in uppercase on the tweets:



```
Out[21]: ['OPP OPP HSD SHOOTING UPDATE W B Hwy 401 amp Avenue Rd police o s no reported injuries possible secondary s',
          'SHOOTING W B Hwy 401 amp Avenue Rd reports of tow trucks shooting at each other police responding OPP attendi',
          'SHOOTING UPDATE Jane St amp Grandravine Dr large police presence o s officers confirmed that 2 vehicles and a',
          'saqzi89 CityNews Muhammad thank you for your response If you have any info please call TorontoPolice at 4168',
          'SHOOTING UPDATE Jane St amp Grandravine Dr police o s confirmed shooting evidence of gunfire officers have',
          'SHOOTING Jane St amp Grandravine Dr several callers reporting multiple gunshots heard in the area police respond',
          'COLLISION Morningside Ave amp W B Highway 401 reports of a vehicle rollover on its roof W B on ramp onto Hwy 40',
          'COLLISION Davenport Rd amp Oakwood Av reports of a male pedestrian struck by driver vehicle police o s located',
          'MISSING WOMAN LOCATED Janet Dawson 68 thank you to the public for your assistance she has been located GO120337 a
         l',
          'Good afternoon Toronto Officer Alex CopWhoLovesCars is now on duty and will provide today s police news al']
```

A future exercise could incorporate inclusion of these tweets to supplement the existing crime data so that the model is based on updated data throughout the day. This would make the model as close to real-time as possible.

Conclusion:

Although the accuracy is not the best, it is our belief that the model should be useful as it will provide clients with at least a starting point for their projections as it is over 50 percent accurate.

The model would be deployed to interested parties as a subscription to a web based service. We envision that our clients would be either individuals or companies and as such we would have both individual and corporate sign-on features.

In addition to the existing data, we would consider any information that is related to patrol data for the neighborhoods as being invaluable. It is our belief that the addition of this other dataset would create a good reference point when analyzing the crime statistics. If we were to consider an example in which the prevailing thought may be that more patrols are needed but if crime is spiking when patrols are ongoing then we need to look at other factors.

While the frequency of the data update depends on the dataset providers the model would be inspected daily to see if the required operating benchmarks are maintained. On a weekly basis we would meet to analyze the daily benchmarks and make the necessary adjustments.

In summary our conclusions are as follows:

- Strategies that balance community services, rent geared towards income and social housing across Toronto neighbourhoods will yield better management of existing police resources.
- Neighbourhoods with lower rent geared to income units (RGI's) showed lower tendencies towards crime.
- Lower average household income neighbourhoods had higher tendencies towards crime.
- Crime most likely to occur was assault.
- **Supervised learning:** Best predictive analysis model on crime categories occurred with random forest classified modelling with a 0.65 accuracy.
- **Unsupervised learning:** Resulted in well distanced clusters, but primarily due to Neighbourhood.
- Clusters were not used for improving the supervised models nor for imputing.
- Future phases of this project could include - using clusters to help train supervised models, incorporation of real time data through the form of social media and Natural Language Processing e.g. Twitter and/or digitized police scanner.
- PCA visualizations did not yield significant separation of data, when using more than 2 target groups or categories and using 2 dimensional components.

References:

[1a]Beware Default Random Forest Importances (explained.ai)

[1b]**Government of Canada supports Indigenous community-based justice**
Government of Canada supports Indigenous community-based justice - Canada.ca

[2][2]https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/rslts-crm-prvntn-12-17/index-en.aspx

[3] Ellis L., Farrington D., Hoskin A. “Handbook of Crime Correlates 2nd Edition”. https://www.elsevier.com/books/handbook-of-crime-correlates/ellis/978-0-12-804417-9.

[4] McQuigge, Michelle. “Canadian police using controversial ‘predictive policing’ tools, report finds”. https://globalnews.ca/news/7309391/canada-police-predictive-tools-report/

[5] Lau, Tim. “Predictive Policing Explained”. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained

[6] Cullen, Julie Berry, and Steven D. Levitt. “Crime, urban flight, and the consequences for cities”. https://www.jstor.org/stable/2646853?seq=1

[7] Ellis L., Farrington D., Hoskin A. “Handbook of Crime Correlates 3rd Edition”. https://www.elsevier.com/books/handbook-of-crime-correlates/ellis/978-0-12-804417-9

[8] Meijer A., Wessels M. “Predictive Policing: Review of Benefits and Drawbacks”. https://doi.org/10.1080/01900692.2019.1575664

[9] Lau, Tim. “Predictive Policing Explained”. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained

[10] Wang, Johnjian; Kifer, Daniel; Graif, Corina; Li, Zhenhui; et al. “Crime Rate Inference with Big Data”. https://dl.acm.org/doi/10.1145/2939672.2939736

[11]