Criminal Statistics in Toronto

JSC370 Final Project Report

Chan Yu

April 28, 2023

Contents

Introduction

Background

Crime is a common social problem that affects communities worldwide. In Canada, Toronto has been the city with the highest crime rate, which includes violent and property crimes. The Toronto Police Service is the agency responsible for enforcing the law in the city, and it publishes criminal data on the Public Safety Data Portal. This data includes information on the type of crime, location, and date of occurrence.

The inflation and unemployment rates are two economic indicators that reflect the financial well-being of a community. Inflation refers to the rate at which the general level of prices for goods and services is rising, and unemployment refers to the percentage of people who are not employed but are actively seeking work. Both indicators can have a significant impact on the daily lives of people, including their safety and security.

Research Questions

The purpose of this project is to analyze the crime statistics in the city of Toronto from 2014 to 2021, and investigate whether there is a correlation between the number of crimes in Toronto and changes in the inflation and unemployment rates over time. To achieve this, we will explore the crime data provided by the Toronto Police Service and compare it to the monthly inflation rate and unemployment rates obtained from the website of City of Toronto between January 2018 and January 2023. This analysis aims to shed light on the relationship between economic conditions and crime in Toronto, which can inform policies and strategies to improve public safety in the city.

Overall, we will provide an overview of the crime situation in the city of Toronto, including the types of crimes and their frequency. We will also discuss the inflation and unemployment rates in the city and how they have changed over time. Finally, we will present the results of our analysis and discuss the implications of the findings.

Please be aware that in this report, tables or figures may be truncated, and interactive visualizations can only be accessed by visiting the website.

Method

The initial stage of this statistical project is to collect data that is to be examined and analyzed, then explore and clean the data to fit the requirements.

Data Collection & Cleaning

1. Crime Data

Most of the data utilized in this project can be accessed through the Public Safety Data Portal. The Police Service identifies Assault, Break and Enter, Auto Theft, Robbery, and Theft Over as the categories of Major Crime Indicators, while this project focuses on analyzing four specific crime categories that occurred in Toronto, as listed below. The data is available in CSV format on the Data Portal and can be imported into R for analysis.

- Shootings and Firearm Discharges: shooting occurrence refers to any incident in which a projectile is discharged from a firearm and injures a person, excluding events like suicide and police involved firearm discharges.
- Break and Enter: the type of premises were reported break and enter includes commercial, house, apartment, educational, transit, and others. One occurrence may have several rows of records since the data is provided at both offense and/or victim levels.
- Auto Theft: similarly, one occurrence may have several rows of record.
- Robbery: includes robbery for mugging, purse snatching, vehicle jacking, robbing ATMs, armoured cars, taxis, etc.

The datasets contain columns of information such as a unique event ID, the date of occurrence, the neighbourhood, as well as the longitude and latitude coordinates of each event. For the purpose of this project, the occurrence date will be used instead of the report date for each record. Additional information, such as the type of premises and locations, may be included in some datasets.

In order to facilitate the merging of the datasets into one for analysis, relevant variables are selected and standardized to ensure that variable names and types (numberical/categorical) match across all categories.

Meanwhile, a new variable named `crime_cat` is created to store the name of their category for each record, then use `bind_rows` to merge all categories into one dataset. A `NA` value will be filled into columns that some categories do not have but others do.

Following the initial EDA checks, the following updates were made to serve the project's objectives:

- Columns were renamed to enable easy access.
- New columns were created to represent `month` and `day_of_week` in numeric format.
- Suspicious values in `year`, `day`, `latitude`, and `longitude` columns were replaced with `NA`.
- Although all datasets from the Public Safety Data Portal include a date column, some may lack `year/month/day/day_of_week` information. To address this, the `date` column was reformatted using `as.POSIXct()`. This enabled the `date` column to fill in the missing values in `year/month/day/day_of_week` correspondingly.
- Finally, the project retained only those observations with `latitude` and `longitude` values within or around city of Toronto, and limited the analysis to years between 2014 and 2021, as there were no records of `Break and Enter` incidents in 2022.

The complete dataset, comprising the downloaded CSV files for shooting and firearm discharges, break and enter, auto theft, and robbery, consists of 122,157 observations and 17 feature variables. Below lists the variables in the dataset.

Variable	**Description**
event_unique_id	Identifier of the offence occurrence
date	Occurrence date of the crime, in the format of "YYYY/MM/DD HH:mm:ss"
year	Occurrence year
month	Occurrence month, e.g. April, May
month_num	Occurrence month in numeric type, e.g. 4, 5
day	Occurrence day of the month, e.g. 26, 27
dayofweek	Day of week of occurrence, e.g. Friday, Saturday
dayofweek_num	Day of week in numeric type starting on Monday, e.g. 5, 6
division	Police division where offence occurred
hood_id	Identifier of neighbourhood where offence occurred
neighbourhood	Name of neighbourhood where offence occurred
longitude	Longitude coordinates
latitude	Latitude coordinates
crime_cat	Category of the criminal offence, e.g. Auto Theft, Robbery
location_type	Location type of offence, e.g. Go Station, Bar/Restaurant
premises_type	Premises type of offence, e.g. Apartment, House
offence	Title of offence, e.g. Robbery-Atm, Theft Of Motor Vehicle

Displayed below is Table 1, which illustrates a random sample of 6 rows from the cleaned data.

Table 2: Sample Data from Public Safety Data Portal

event_unique_id	date	year	month	month_num	day	dayofweek	dayofweek_num	divis
GO-20141807710	2014-04-01 04:00:00	2014	April	4	1	Tuesday	2	D55
GO-20141608353	2014-02-27 05:00:00	2014	February	2	27	Thursday	4	D42
GO-2020646005	2020-04-01 04:00:00	2020	April	4	1	Wednesday	3	D55
GO-2015548747	2015-04-02 04:00:00	2015	April	4	2	Thursday	4	D41
GO-20211858657	2021-09-27 04:00:00	2021	September	9	27	Monday	1	D32
GO-2016633608	2016-04-08 04:00:00	2016	April	4	8	Friday	5	D55

Table 3: Data from the City of Toronto for Year 2021

month_date	year	month	inflation	unemployment
2021-01-01	2021	1	0.8	10.8
2021-02-01	2021	2	0.9	11.0
2021-03-01	2021	3	1.7	9.1
2021-04-01	2021	4	2.4	9.4
2021-05-01	2021	5	2.9	10.5
2021-06-01	2021	6	2.5	9.8
2021-07-01	2021	7	2.8	9.7
2021-08-01	2021	8	3.3	8.6
2021-09-01	2021	9	3.8	8.1
2021-10-01	2021	10	4.0	7.8
2021-11-01	2021	11	4.3	7.9
2021-12-01	2021	12	4.7	8.0

2. Financial Indicators

Additionally, for the two datasets from the City of Toronto, which record the monthly inflation rate and unemployment rates between January 2018 and February 2023, are also provided in CSV format.

The monthly inflation and unemployment rates were downloaded and imported into R as separate datasets, each containing two columns named `month` and `rate.` However, since the `month` columns in both datasets have the same format and range, they can be merged into a single dataset using the `merge()` function.

Moreover, the original `month` column is in the format `YYYY-mm-dd`, with the day section set to 01. To facilitate future manipulation, the `month` column is renamed to `month_date`, and new columns named `year` and `month` are generated based on the `month_date` column.

Displayed below as Table 2 is the data for the monthly inflation and unemployment rates in the year 2021 as example.

3. Neighbourhood & Demographics

The Open Data Portal of the City of Toronto contains information on neighbourhoods and demographics. For this project, we will utilize the `neighbourhoods` package from the `opendatatoronto` library, which provides geometry data containing the coordinates of the boundaries of 158 neighbourhoods in Toronto that will be used to create map visualizations.

Table 3 displays an example dataset of neighbourhood information with selected variables. It can be observed that the identifier for each neighbourhood is represented by the variable `AREA_SHORT_CODE`, which is equivalent to the `hood_id` variable in the crime dataset. To facilitate future usage, a new column has been generated by converting the data type of `AREA_SHORT_CODE` to match it with other datasets.

Table 4: Example of Neighbourhood Information

AREA_SHORT_CODE	AREA_NAME	geometry	hood_id
001	West Humber-Clairville (1)	POLYGON ((-79.55236 43.7094...	1
002	Mount Olive-Silverstone-Jamestown (2)	POLYGON ((-79.60338 43.7578...	2
003	Thistletown-Beaumond Heights (3)	POLYGON ((-79.57751 43.7338...	3
004	Rexdale-Kipling (4)	POLYGON ((-79.55512 43.7151...	4
005	Elms-Old Rexdale (5)	POLYGON ((-79.55512 43.7151...	5
006	Kingsview Village-The Westway (6)	POLYGON ((-79.55236 43.7094...	6

Table 5: Example of Neighbourhood Population in 2016

hood_id	neighbourhood	pop_2016
1	West Humber-Clairville	33312
2	Mount Olive-Silverstone-Jamestown	32954
3	Thistletown-Beaumond Heights	10360
4	Rexdale-Kipling	10529
5	Elms-Old Rexdale	9456
6	Kingsview Village-The Westway	22000

Additionally, the portal contains open data on the demographics of Toronto. The project will extract the population data for each neighbourhood in 2016 to compute crime occurrence density over population.

The Table 4 below displays the example data of population in 2016 for the first six neighbourhoods. Only useful variables were extracted from the dataset and were renamed properly.

Data Grouping

To examine how different types of crimes vary across various neighbourhoods in Toronto, the data needs to be grouped by the `neighborhood` and `crime_cat` columns. As noted earlier, a single occurrence may have multiple rows of records. Therefore, instead of counting the number of observations, we count the number of unique values in `event_unique_id` column for each group.

Displayed below in Table 5 are the top three neighbourhoods with the highest number of crime occurrences for each crime category, with their average latitude and longitude in each neighbourhood.

To analyze how the number of crime occurrences in Toronto changes over time, the data needs to be grouped by `year` and `month_num`, as well as by `crime_cat` for deeper insights.

Displayed below as Table 6 is an example of the data, which shows the number of crime occurrences for each category in the first quarter of 2021.

Furthermore, Table 7 provides a breakdown of the total number of criminal offences committed for the four categories. It is worth noting that the highest number of occurrences between 2014 and 2021 in Toronto belongs to break and enter, with over 55,000 incidents, while auto theft and robbery have fewer incidents in comparison. On the other hand, firearm discharge has the lowest number of incidents, with around 3,000 occurrences.

Data Merging

To analyze the correlation between the inflation rate and unemployment rate, we can merge the data that is grouped by month and crime category with the rates data.

Table 6: Example Data Grouped by Neighbourhood and Crime

neighbourhood	crime_cat	count	latitude	longitude
West Humber-Clairville	Auto Theft	2783	43.71811	-79.59669
Islington-City Centre West	Auto Theft	969	43.63100	-79.54140
York University Heights	Auto Theft	909	43.76371	-79.48771
Waterfront Communities-The Island	Break & Enter	2186	43.64506	-79.38673
Church-Yonge Corridor	Break & Enter	1655	43.65970	-79.37912
Bay Street Corridor	Break & Enter	1293	43.65783	-79.38413
Church-Yonge Corridor	Robbery	998	43.65918	-79.37991
Moss Park	Robbery	902	43.65709	-79.37059
Bay Street Corridor	Robbery	645	43.65719	-79.38345
Glenfield-Jane Heights	Shooting & Firearm Discharge	172	43.74767	-79.51256
Black Creek	Shooting & Firearm Discharge	102	43.76699	-79.51804
Waterfront Communities-The Island	Shooting & Firearm Discharge	84	43.64523	-79.38839

Table 7: Example Data Grouped by Month and Crime

year	month_num	crime_cat	count
2021	1	Break & Enter	517
2021	1	Auto Theft	383
2021	1	Robbery	142
2021	1	Shooting & Firearm Discharge	27
2021	2	Break & Enter	412
2021	2	Auto Theft	349
2021	2	Robbery	95
2021	2	Shooting & Firearm Discharge	17
2021	3	Auto Theft	426
2021	3	Break & Enter	423
2021	3	Robbery	125
2021	3	Shooting & Firearm Discharge	30

Table 8: Number of Crime Occurrences for each Category

crime_cat	count
Break & Enter	55797
Auto Theft	32906
Robbery	21619
Shooting & Firearm Discharge	3054

Table 9: Example Data Grouped by Month and Crime

year	month_num	crime_cat	count	month	inflation	unemployment
2021	1	Break & Enter	517	2021-01-01	0.8	10.8
2021	1	Auto Theft	383	2021-01-01	0.8	10.8
2021	1	Robbery	142	2021-01-01	0.8	10.8
2021	1	Shooting & Firearm Discharge	27	2021-01-01	0.8	10.8

Table 10: Example of Merged Neighbourhood Information

AREA_SHORT_CODE	AREA_NAME	hood_id	pop_2016	geometry
001	West Humber-Clairville (1)	1	33312	POLYGON ((-79.55236 43.7
002	Mount Olive-Silverstone-Jamestown (2)	2	32954	POLYGON ((-79.60338 43.7
003	Thistletown-Beaumond Heights (3)	3	10360	POLYGON ((-79.57751 43.7
004	Rexdale-Kipling (4)	4	10529	POLYGON ((-79.55512 43.7
005	Elms-Old Rexdale (5)	5	9456	POLYGON ((-79.55512 43.7
006	Kingsview Village-The Westway (6)	6	22000	POLYGON ((-79.55236 43.7

Table 8 displayed below is an example of the data, illustrating the number of crime occurrences for each category in January 2021.

We merge the population number for each neighbourhood to neighbourhood info data to facilitate future usage. Table 9 below displays the example results.

Modeling

As previously mentioned, the project goal is to investigate whether there is a relationship between changes in the inflation and unemployment rates and the number of crimes in Toronto. To achieve this, I plan to train three different linear models using the variables from Table 8, which include inflation rate, unemployment rate, crime category, and occurrence date, to predict the number of crime occurrences.

Since the dataset only contains approximately 190 observations, I will split it into training and testing sets using an 80/20 split.

The first model I will train is a generalized linear model (GLM). This model extends the linear model to handle non-normal response variables, such as counts of occurrences. I will use the Poisson distribution assuming that the mean and variance of the count are equal, and that the counts are independent of each other.

Next, I will use a generalized additive model (GAM) as an extension of the GLM. GAM allows for non-linear relationships between response and predictors, and I plan to apply smoothing functions to the date variable since I suspect a non-linear relationship between date and counts.

Lastly, I will train a linear mixed-effects model, which was covered in STA303. Since the data can be clustered by crime category, a linear mixed-effects model allows for random effects on categorical variables, which can account for the variability within and between groups.

Although the dataset is small, I plan to use all available predictors to create pruned regression trees and a random forest model for predicting crime count. Ultimately, I aim to compare the mean squared errors of all these methods.

Tool

Data cleaning and wrangling were completed with `tidyverse`, `dplyr`, `data.table`, `dplyr`, `lubridate`, `sf`, and `opendatatoronto`.

Plots and figures were created with `ggplot2`, interactive visualizations were created using `plotly`, `mapboxapi`, `leaflet`.

Tables were created with `knitr`, `kable`, `kableExtra`, and `broom`.

Packages used for modelling include `mgcv`, `rpart`, `rpart.plot`, `randomForest`, and `lme4.`

Result

1. Occurrences correlation with financial indicators

Through Visualizations

To examine how the frequency of crime has evolved over time, we have plotted the occurrences of different crime categories between 2014 and 2022.

- The data indicates that auto theft has consistently increased over time, while robbery has experienced a decreasing trend.
- The category of shooting and firearm discharge shows a relatively steady frequency over time, albeit with fewer observations compared to the other categories.
- Break and Enter has the largest number of observations in the overall dataset, almost equivalent to the sum of occurrences of the other three categories before 2018, but it has experienced a downward trend since then.

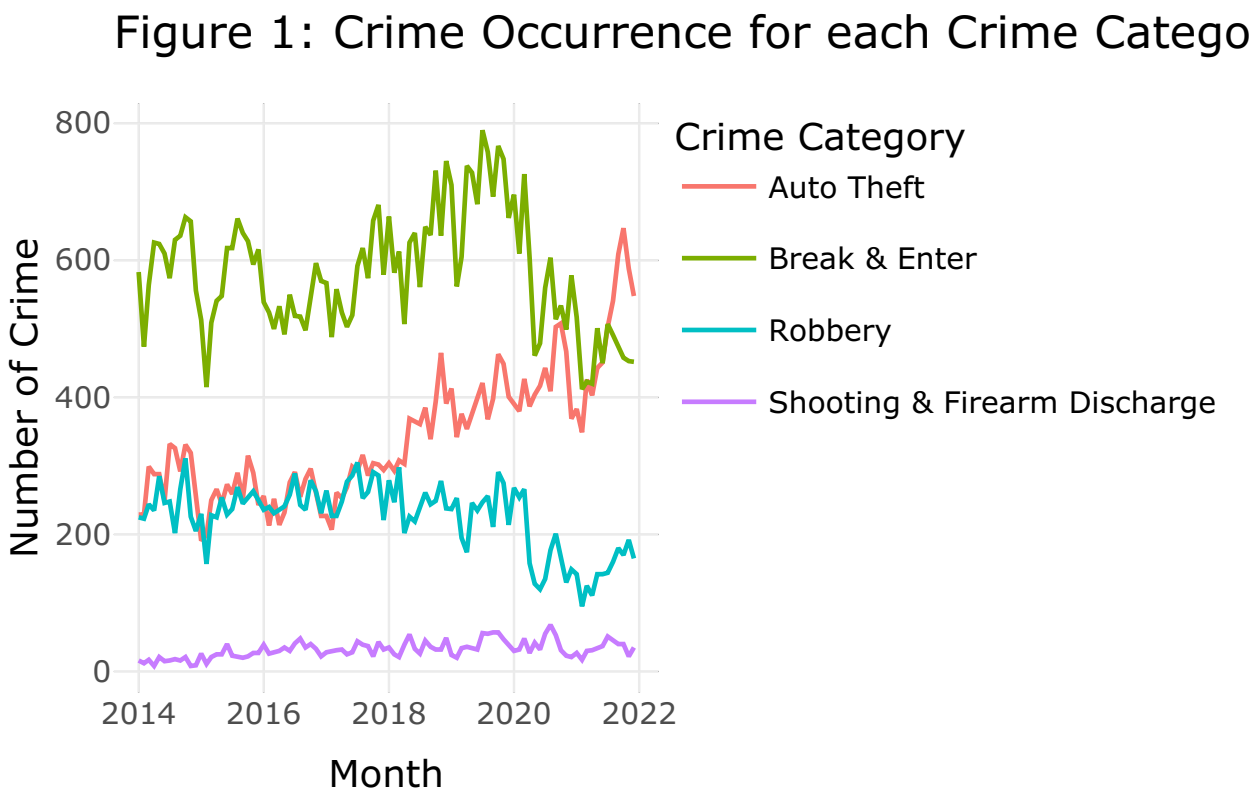
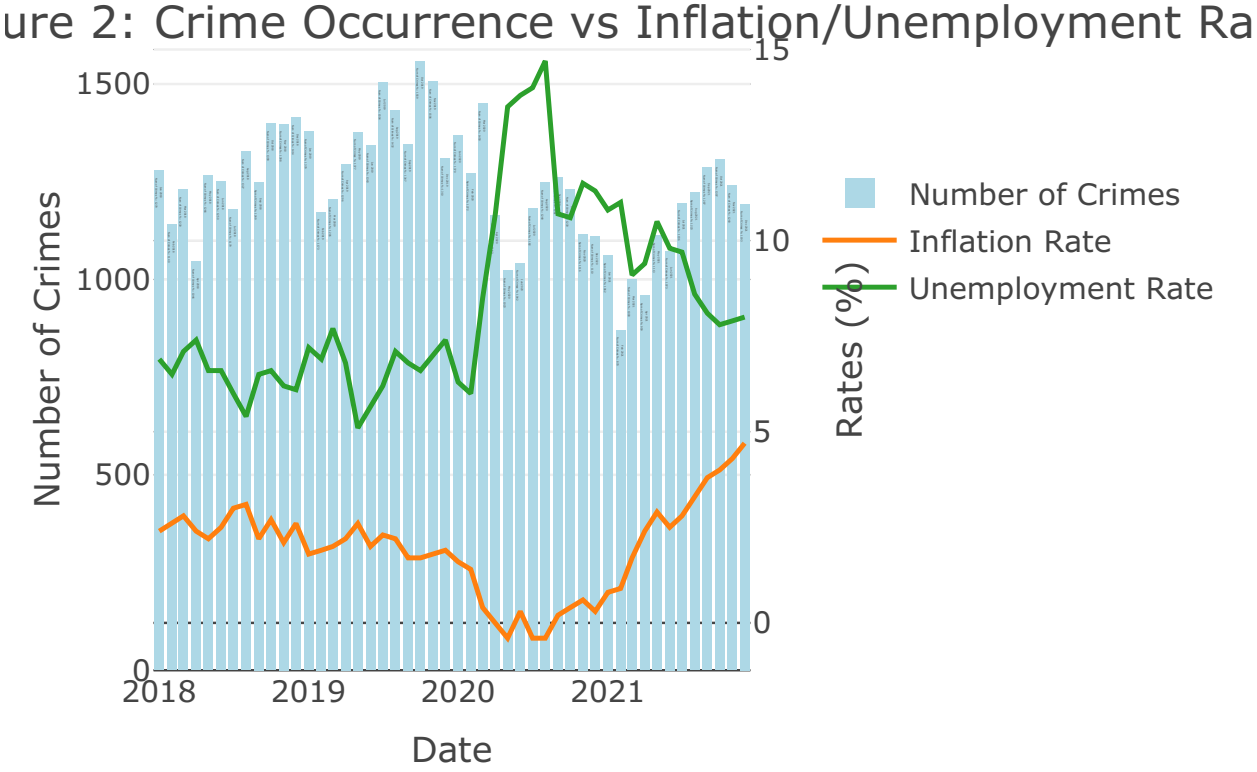


Figure 2 below displays the relationship between crime occurrences and financial indicators. Here are some observations:

- The occurrence of crime in Toronto peaked at over 1500 occurrences per month in October 2019 between 2018 and 2022, but then decreased in the following two years to below 1300 occurrences per month. There appears to be a seasonal trend with lower crime rates in winter.
- Historically, inflation and unemployment have shown an inverse relationship, where inflation tends to decrease when unemployment rises. In other words, increased employment leads to higher spending power and demand. This pattern can be observed from the rate lines in the visualizations.
- According to the barplot depicting the number of crime occurrences, there appears to be a correlation between decreasing inflation rates around 2020 and a decrease in the number of crimes. Additionally, starting from 2021, as the inflation rate continues to increase, the number of crimes also decreases.
- Despite the sharp increase in the unemployment rate at the beginning of 2020 due to COVID-19, it appears to have had no discernible effect on the number of crimes committed. Moreover, in later 2021, despite a decrease in the unemployment rate, the number of crime occurrences continued to increase. This suggests that the unemployment rate may not be a significant factor in the occurrence of crimes.



The animated graph confirms the previous observation of a trend between the inflation rate and the number of crime occurrences, visit on the website.

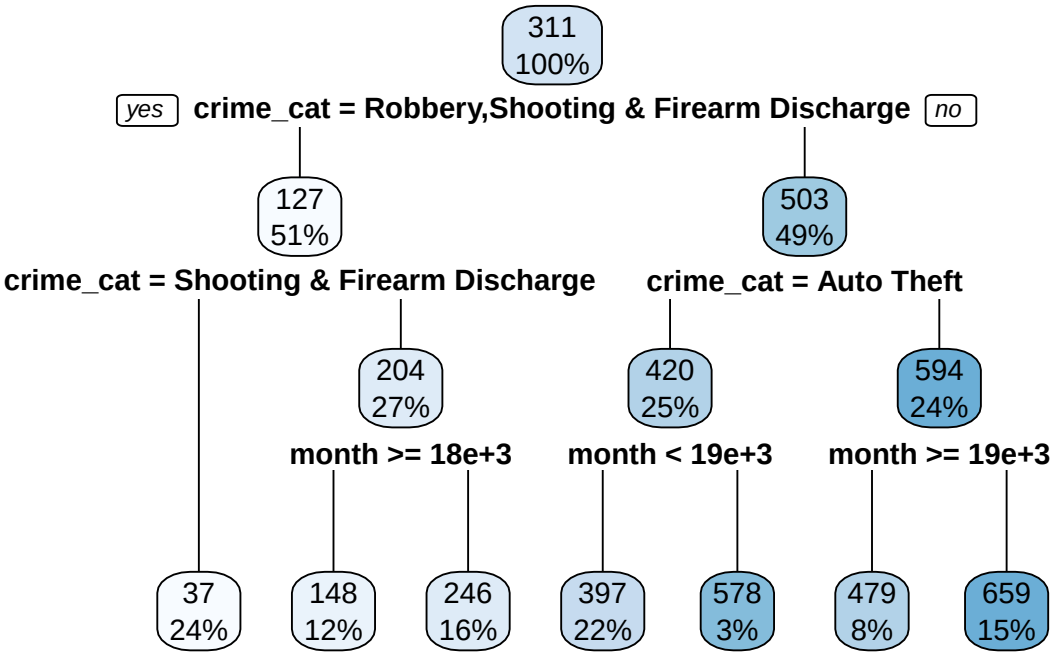
Model Prediction

According to the Poisson GLM model output, the response variable is negatively associated with both inflation rate (approximately -0.029) and unemployment rate (around -0.039), and the latter has a greater effect on changes in crime occurrence than expected. Additionally, the crime category is a significant predictor with a clear impact. The testing set's MSE for this model is approximately 4287.

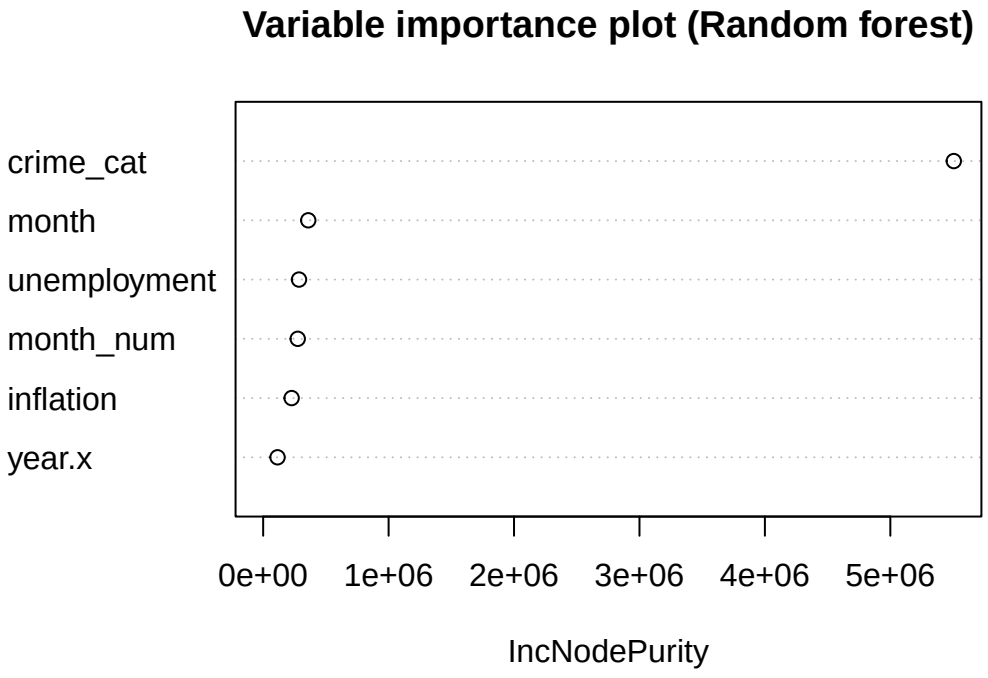
The output of the GAM model shows a different pattern. Specifically, the estimated coefficients reveal that, when controlling for other variables, the count of crimes is significantly greater for the categories 'Break & Enter', 'Robbery', and 'Shooting & Firearm Discharge', relative to 'Auto Theft'. The coefficients for 'inflation' and 'unemployment' are negative, but not statistically significant, which suggests that there is no linear relationship between the financial indicators and the occurrence of crimes. The adjusted R-squared value of 0.914 indicates that the model accounts for 91.4% of the variability in the data, which highly possible caused by overfitting. The MSE of the testing set for this model is approximately 5829, which is higher than that of the first model.

In the linear mixed-effects model, I assigned a random slope to the unemployment rate, as it appeared to be relatively more significant compared to the other two models. According to the model summary, the month and inflation rate did not have a significant impact on the number of crimes. The variance of the random effects indicated that there was considerable variation in the number of crimes between different crime categories. This model had the lowest MSE of approximately 3700 on the testing set.

The variable importance plot, which was created using the pruned regression tree with an optimal complexity parameter value, reveals that the crime category is a significant predictor of offense counts, while both the inflation rate and unemployment rate are not. The testing set MSE for this model was approximately 4300.

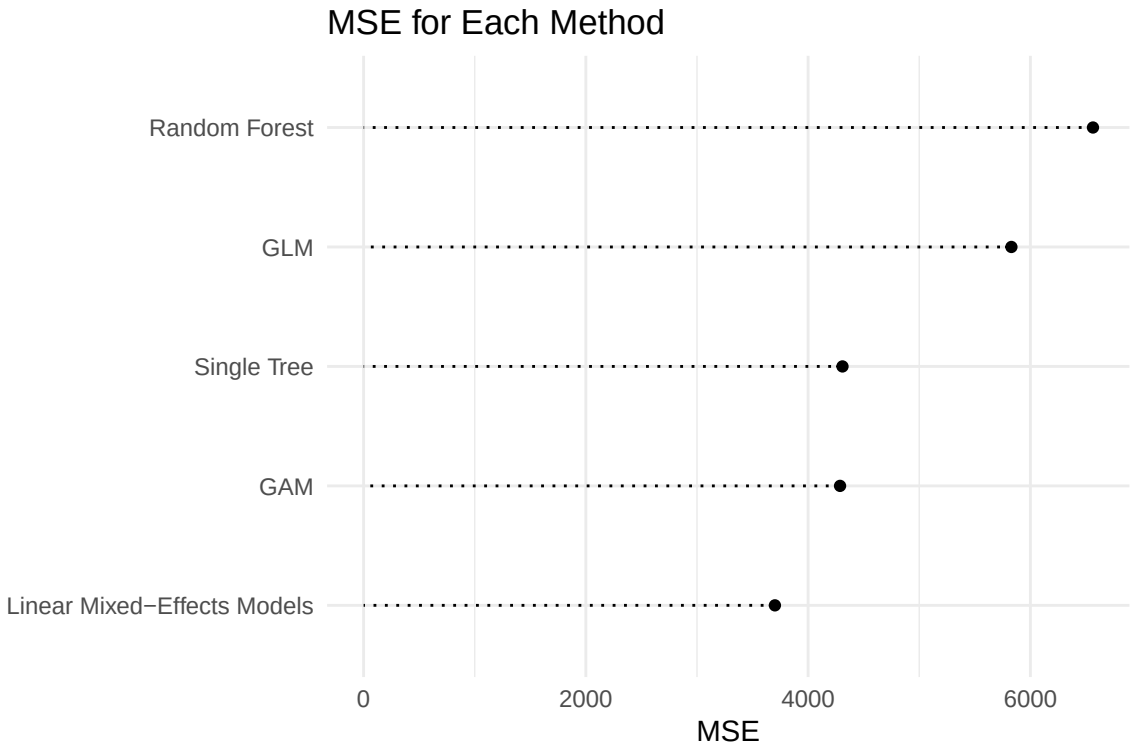


The importance plot generated by the random forest model also confirms that the crime category is a crucial predictor, while the unemployment rate has a relatively higher importance compared to the inflation rate. However, the MSE obtained for the testing set is approximately 6563, which is higher than the other models.



The plot below displays the MSEs for all methods. The Linear Mixed-Effects model has the smallest MSE,

followed by GAM and single regression tree, which have similar MSE values. Despite the limitations of a small dataset and limited predictors, if I had to choose a model from these options for predicting crime statistics in Toronto, I would select the linear mixed-effect model.



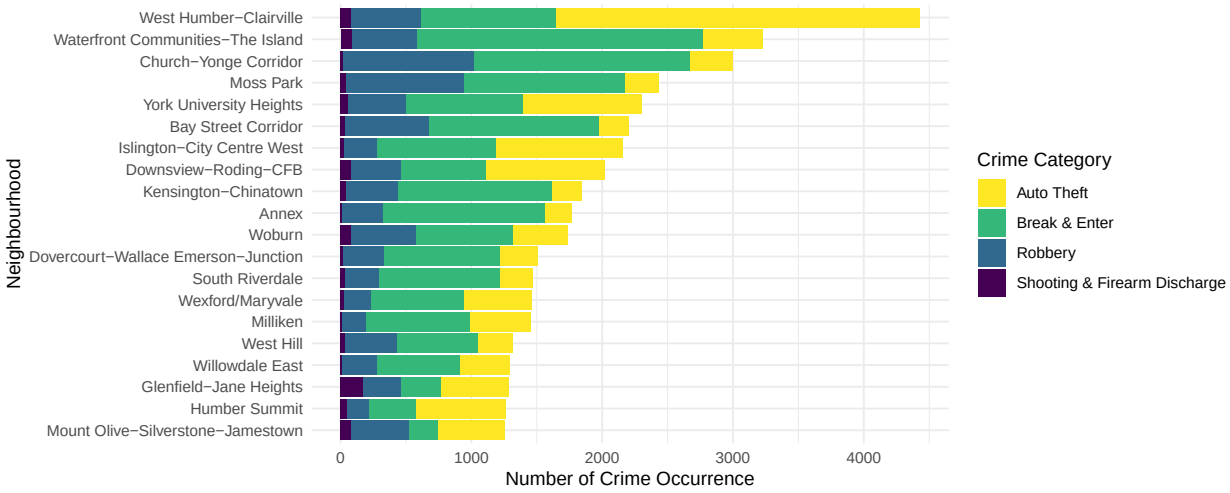
2. Occurrences correlation with neighbourhood profiles

In Figure 4, we plot the top 20 neighbourhoods with the highest number of crime occurrences between 2014 and 2021.

The highest number of crime occurrences took place in **West Humber-Clairville**, with approximately 4300 recorded instances. **Waterfront Communities** and **Church-Yonge Corridor** also had over 3000 recorded occurrences each.

Figure 4: Barplot of Crime Occurrence by Neighbourhood

Focus on top 20 neighbours between 2014−2021

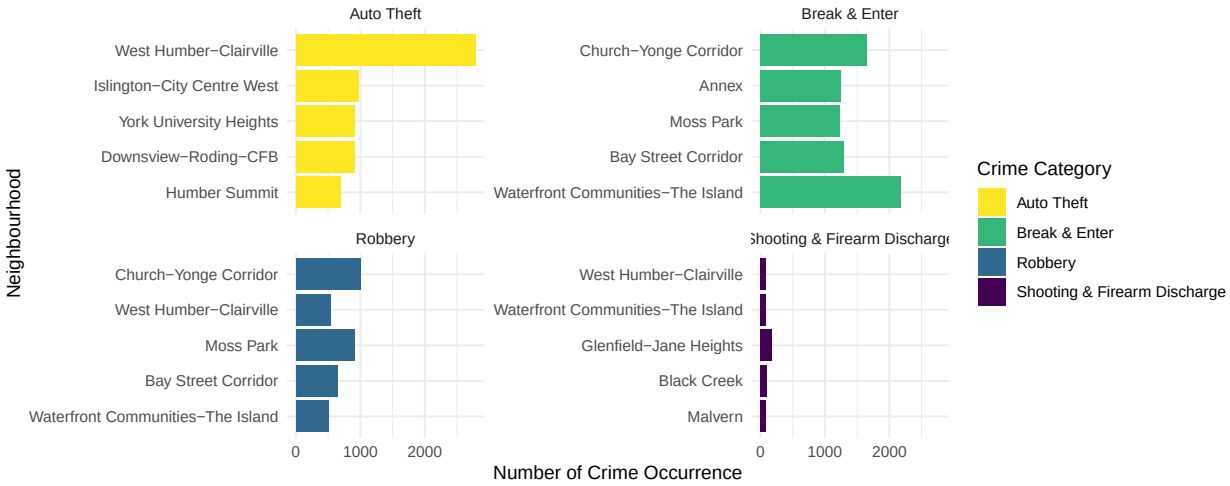


We further investigate the data by visualizing the top five neighbourhoods with the highest number of occurrences in each category in Figure 5. The names of these neighbourhoods are also listed below:

- For auto theft, **West Humber-Clairville** has the highest number of instances.
- For break and enter, **Waterfront Communities** has the highest number of occurrences and **Church-Yonge Corridor** is the second highest.
- **West Humber-Clairville** and **Waterfront Communities** also appear as two of the top five neighbourhoods with the highest number of robbery and firearm discharge occurrences.
- **Church-Yonge Corridor** has the highest number of robbery occurrences.

Figure 5: Barplot of Crime Occurrence by Neighbourhood

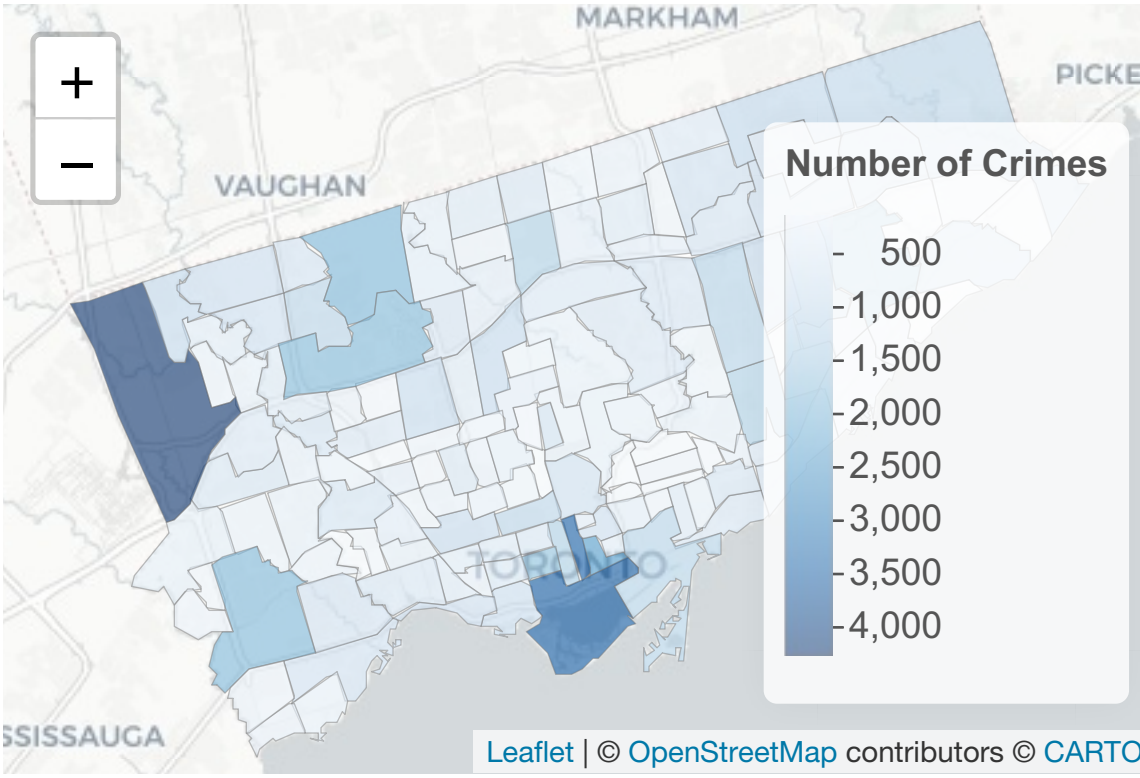
Focus on top 5 neighbours in each crime category



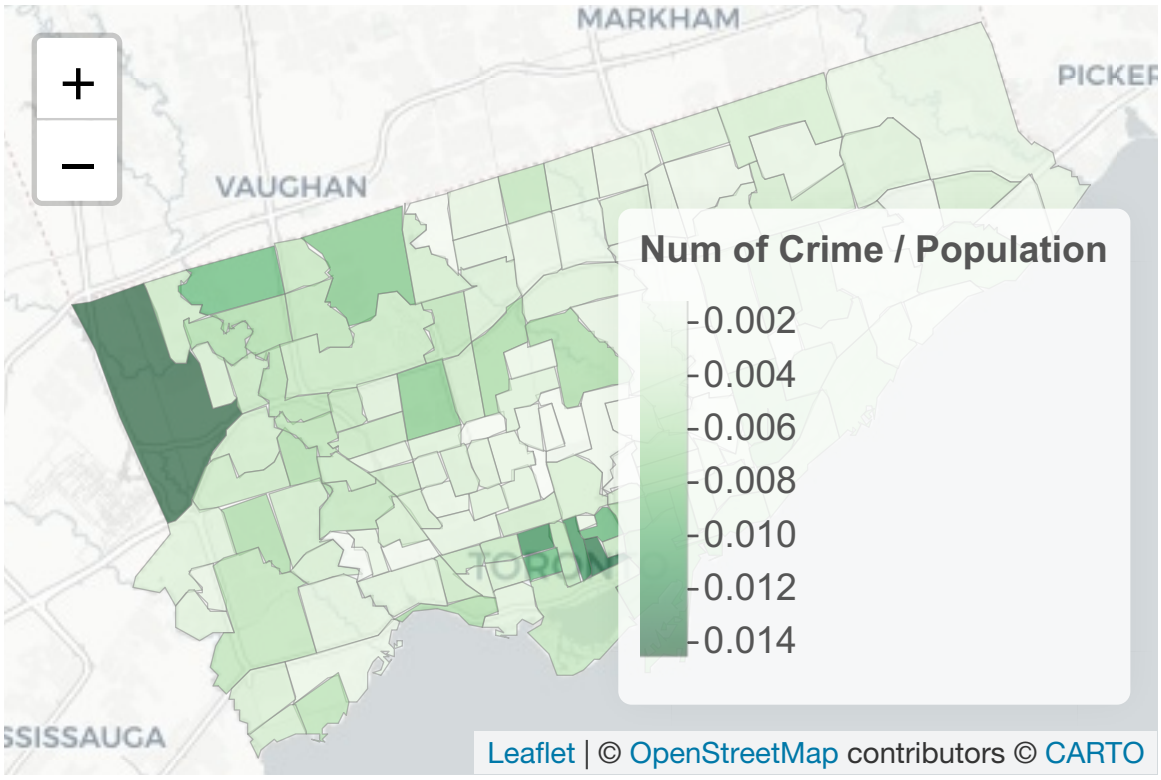
Based on this information, we can conclude that shooting and firearm discharge incidents are mostly concentrated around the edges of the city, whereas auto theft occurs throughout Toronto but is more common in the northwest region. Break and enter, and robbery incidents, on the other hand, are predominantly concentrated in the downtown area of Toronto, where population density is relatively high.

The following map provides a better representation of the location of each neighbourhood.

Occurrence Polygons



Density Polygons



- The map view with polygons proved what we obtained from the previous observations. With the darkest shade indicating the highest number of crimes in the northwest region due to auto theft, and

a relatively higher number of crimes in the downtown area of Toronto. In contrast, areas with low population density show lighter shade on the map indicating fewer crimes.

- The scatter plots (on website) provide a more accurate representation of the average location of all crimes that occurred in each neighborhood area. The size of the scatter plots corresponds to the number of crimes, with larger scatter plots indicating higher crime counts. Similarly, larger scatter plots are observed in the northwest region and downtown area, indicating higher crime levels in these areas.
- Regarding the density polygon, it displays the ratio of crime occurrences in 2016 to the population of each neighbourhood in the same year. The darker shades observed in the northwest region and downtown area correspond to the findings we obtained.

Conclusions & Summary

Conclusions

The crime rate in the City of Toronto has varied over time. According to the dataset, the most common types of crimes reported in the city are "Break & Enter" and "Auto Theft" more than "Robbery" and "Firearm Discharge". Moreover, there has been a significant increase in the total number of reported crimes in the last decade, with a peak in 2019. By observing the location of crime offenses in multiple charts, it concludes that the highest number of crimes in the northwest region, and a relatively higher number of crimes in the downtown area of Toronto. In contrast, areas with low population density show fewer crimes.

In general, inflation and unemployment are important economic indicators that can have an impact on crime rates. In Toronto, inflation has remained relatively stable over the years before 2021, while the unemployment rate has seen some fluctuation, with a peak in 2020 due to the COVID-19 pandemic. In this project, we applied five different models to explore the association between these financial indicators and the number of crimes. The results from the models showed that the crime category is a significant predictor in the city, but neither of the rates is significant in better-fitted models, which is not under expectation.

In conclusion, our findings suggest that the crime category is the most important predictor of crime rates in the city of Toronto, which could be considered while allocating resources and developing targeted strategies to reduce crime in the city.

Limitation & Improvement

There are several limitations and potential improvements that should be considered for this project.

Firstly, the dataset used in this project is relatively small and has a limited number of predictors. To enhance the accuracy of our models, it would be beneficial to collect and merge additional data on various variables that could influence crime rates, such as income level, education level, age, gender, ethnicity, community policing efforts, and drug and alcohol use.

Secondly, even though the majority of the data in this project was provided by the Toronto Police Service, there may be discrepancies between the features of each dataset for some specifications, such as the relationship between different `premises_type` or `location_type`. This aspect was not explored in this project, although it may be significant for some types of crime.

Additionally, during the exploratory data analysis phase, it might be worth investigating whether there are any seasonal patterns in crime incidents in Toronto, considering that weather conditions can play a significant role in Canada.