APPLICATION OF DATA MINING TO THE ANALYSIS OF CRIME DATA

by

Vanessa Audel Afolabi, Hon Bsc, University of Toronto, 2013
Vanessa Audel Afolabi, MMA, Queen's University, 2020

A Major Research Paper
presented to Toronto Metropolitan University

in partial fulfillment of the requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2025

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PAPER (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions, as accepted by my examiners.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Vanessa Audel Afolabi

APPLICATION OF DATA MINING TO THE ANALYSIS OF CRIME DATA

Vanessa Audel Afolabi

Master of Science 2025

Toronto Metropolitan University

ABSTRACT

In this paper, both descriptive and predictive analytics were applied to crime and arrest data from two major metropolitans, namely New York City[1] and Toronto[2], to better understand the when, where, who and what of criminal behaviour. Using both exploratory data analysis and PowerBI, advanced visualizations were created to better understand when and where crimes tend to occur and the demographics of perpetrators who commit these crimes. The second part of my analysis focused on using the NYC Arrests [1] dataset to build a classification model to predict the gender of a perpetrator. After several experiments, the best model was used to generate a feature importance graph to determine which crime features are most important in predicting the gender of a perpetrator. Crime analytics is important when learning about criminal behaviour. It can also help discover trends, mitigate criminal activity and improve public safety and quality of life.

Key words:

NYC Arrests, Toronto Crimes, PowerBI, Random Forest Classifier, SVM Classifier, Performance Metrics

ACKNOWLEDGEMENTS

I will forever be grateful to **Professor Mucahit Cevik** for agreeing to be my research supervisor. His support, guidance and feedback played an immense part in making this project a reality. The knowledge I gained from taking his Advanced Visualization course was used throughout every part of my research project. I would also like to thank **Professor Ceni Babaoglu,** for providing support during each stage of the research project and during the completion of this masters program. She truly embodies what it means to be a reliable resource and sounding board.

Thank you, Professor Cevik and Professor Babaoglu.

TABLE OF CONTENTS

1. INTRODUCTION

A. Overview of MRP Stages

This document provides an in depth description of all the components and phases of my Major Research Project. The first component is an introduction to the project which includes the rationale behind the topic of choice, the research question to be answered and the choice of datasets. This is followed by the data description component which highlights the data sources, data dictionary, any exploratory data analysis performed on the datasets and any feature engineering and data curation that has been done to facilitate the generation of insights. The next step includes a literature review, a thorough breakdown of methodology, a description of all experiments conducted, the results obtained, a presentation of analytical findings, recommendations for future work and an appendix.

B. Problem Definition

The occurrence of crime is an inevitable part of society. Different parts of society see varying rates of occurrence of different types of crimes at different points in time. There exists an immense amount of historical crime data that can be gleaned to extract meaningful information about the nature of different crimes, crime hotspots, criminal personas, criminal demographics and when crimes tend to occur. Data science and data mining are not only great tools for highlighting the what, when, who and where of crimes but they can be used as both predictive and prescriptive tools to predict things such as crime types and the gender, race and age groups of criminals. With such tools, law enforcement can help mitigate criminal activity, solve cases and discover trends ultimately leading to the provision of resources to communities that will help improve overall public safety and quality of life.

C. Research Question

The goal of my Major Research Project is to perform a deep dive analysis into the historical crimes and arrests that have taken place in two major metropolitan cities namely, New York City[1] and Toronto[2]. The goal is to answer questions related to the types of crimes that tend to take place, where and when they usually occur within each city and the persona and demographics of the criminals linked to different types of crimes. Additionally the project aims to discover some similarities and differences in the nature of crimes and criminal behavior between these two major metropolitans. An additional goal is to use details about crimes and arrests to predict the gender (male/female) of criminals and highlight which features are the most important when building such predictive models. By using data science tools techniques such as exploratory data analysis, advanced visualization and predictive analytics one can understand the nature of crimes and predict who is likely to commit different crimes.

D. Datasets

For the purposes of this research project, I will be working with two datasets, namely NYPD Arrests and Toronto Major Crime Indicators (MCI). The first dataset, obtained from the NYC Open Data website[1], contains NYPD arrests including a breakdown of every arrest made in NYC by the NYPD from 2006 to 2024. The second dataset obtained from the City of Toronto Open Data website [2], contains all Major Crime Indicators (MCI) occurrences at the offence and/or victim level.

2. LITERATURE REVIEW

A. Summary of Literature Review Choices

The four research projects that I critically analysed focused on a wide range of data science and analytics techniques to both learn about previous crimes and predict crimes. Some of these techniques were as simple as exploratory data analysis, descriptive statistics and advanced visualization and as complex as building crime predictive models using neural networks, support vector machines and natural language processing techniques. All of these are in line with the goals of my project where I aim to understand the crimes that occur in major metropolitans such as Toronto and New York by looking at the types of crimes that tend to take place, where and when they usually occur within each city and the persona and demographics of the criminals involved.

B. Literature Review 1

This first review was titled, "*Twitter as a Lens for Crime Analysis: A Comprehensive 4W Model for Identifying Crime Patterns and Insights*" [3]. There is a growing body of research on leveraging social media data for crime analysis, crime investigation, crime predictions and crime hotspot analysis. In addition to using historical crime, geographic and demographic data, scientists are realising the importance of utilizing real time, geo-tagged social media data such as Twitter posts to quickly analyse, forecast and solve crimes. Law enforcement, policymakers and community organisations will benefit greatly from such a tool when developing more effective crime prevention and response strategies.

This paper first touches on how machine learning can be used to extract crime-related data from unstructured text data such as Tweets. Secondly, the paper presents the 4W (What, Where, When, and Who) model based on Twitter crowdsourcing to characterise a crime incident. The What portion uses a tweet to determine the type of crime such as Burglary or Homicide. The Where portion of the model uses the geo-location metadata related to tweets to determine the location of a crime. The When portion of the model derives the time of a crime from the date metadata associated with tweets. The Who portion of the model uses the contents of the tweet to identify people associated with the crime such as witnesses, perpetrators and victims.

Using crime-related keywords and the Twitter Search API, a relevant set of crime related tweets were compiled. These tweets were labelled as crime and non-crime and then cleansed using NLP techniques such as stop word removal and lemmatization. The text was transformed using word embeddings (BERT, GLoVe) and significant features were extracted. Then machine learning models such as SVM and LSTM were used to identify tweets most related to crimes. This encompassed the data collection and preparation phase. Then each part of the 4W model was developed separately. The What model used pre-labelled datasets containing different crime types and their corresponding text descriptions to train and test classifiers such as LogisticRegression, RandomForest and MultinomialNB. The Where model used spaCy to extract location related terms and the LocationIQ API to determine the countries based on these location terms. The When model extracted date and time metadata from the crime-related tweets using the LexNLP library. The Who model used The Hugging Face Transformers library to develop a NER model with the ability to identify person entities

The LSTM + BERT model for extracting crime related tweets, obtained an average accuracy of 0.9383 on the training dataset and 0.941 on the test dataset. There were significant challenges with developing the location model because of location abbreviations and too many mentions of different locations making it difficult to extract the correct data. A field study involving experts showed that the 4W model could correctly identify the 4 elements in most cases.

The models developed and their corresponding results demonstrated the potential for this crime analysis framework to provide valuable insights into crime patterns and trends that can inform the development of targeted crime prevention strategies and interventions. It proves that crime related tweets can be gleaned for useful information which can reveal what type of crimes were committed, where these crimes took place, when the crimes happened and who was associated with the crimes in any way.

C. Literature Review 2

The second review was titled, "*Predicting Crime and Other Uses of Neural Networks in Police Decision Making*" [4]. Given the increase in population and rise in urbanization, there has been a drastic increase in crimes. An increase in police presence is not enough to maintain personal and public safety hence the need to plan for and respond effectively to different types of crimes. Extensive research has shown that Neural Networks are ideal for predicting crime hotspots and increasingly used in Spatial Crime Forecasting.

The goal of this project was to gain an in-depth understanding of how law enforcement can utilize Neural Networks in predicting a specific type of crime and in predicting the general location for a specific type of crime hence aiding law enforcement in automated decision making such as crime-preparedness and response by police officers. Discussion of research that demonstrates how Neural Networks can be used for data mining and crime analysis, using old crimes to solve current crimes, to determine the likelihood for repeat offenses and to identify guns based on the markings of a bullet which falls under forensic analysis.

Using crime incidents data from the City of Detroit Michigan from 2016 to 2020, several backpropagation NN models were built. The first models built were used to predict the most likely crime given location and time of day features. The second set of models built were used to predict crime locations given the type of crime and time of day features. The crime data was clustered into 38 different crimes, which were spread across 30 zip code regions of the city. Any small cluster of crimes for minor crimes were removed from the analysis leaving only 27 remaining crime clusters. Several time and location features were included such as latitude and longitude. Related to the crime type NN prediction models, the two-hidden layer MLP had 12 nodes in its first layer and four nodes in its second layer. The single-hidden layer MLP had 12 nodes in its hidden layer. Lastly the RBF NN had an association layer of 54 nodes and a hidden layer of 12 nodes. Related to the NN for predicting location, the two-hidden layer MLPNN had 27 nodes in its first layer and six nodes in its second layer, while the single-hidden layer MLP NN had 36 nodes in its hidden layer.

Backpropagation trained NNs (16.4%) outperformed the RBF NN models (12.9%) when building the model to predict the type of crime. The two hidden layer MLP (8.2%) outperformed the single layer MLP (7.6%) when predicting the crime

location. Both models were trained again focusing on just the top 6 different crime types with a 7th cluster encompassing all other crimes. The models performed a lot better in terms of their model accuracy.
When crime calls come in, these models can be useful in inferring the nature of a crime based on location and time. In other cases, the location of a crime can be inferred based on the type of crime and the time of day. These are all advisory tools that can be used by law enforcement. Future research is needed to further investigate the use of NNs, possibly using additional temporal or different sized geospatial cues to improve near-term (immediate) crime predictions.

D. Literature Review 3

The third review was titled, "*Crime analysis using Data Analytics Background*" [5]. How can gaining a deeper understanding of the nature of past crimes in a given area and during a certain timeframe benefit the city's law enforcement in optimizing the proper allocation of resources towards increased patrolling and building new stations and in devising prevention strategies such as community outreach programs.

The goal was to utilize descriptive and predictive analytics to gain a deeper understanding of the nature of crimes by revealing hidden correlations and patterns and identifying influential features in order to predict the arrest status of serious crimes.

Used exploratory data analysis, correlation analysis, clustering and data visualization tools to develop a descriptive model that highlighted the impact that time and location has on the types of crimes in Chicago from 2017 to 2019. Used naive bayes and decision trees among others to build a classification model for predicting the category of serious crimes such as battery and assault based on input variables such as beat, block and time features. Also identified the most influential features in predicting crime types.

The descriptive model showed that theft, battery and criminal damage were the most frequently occurring category of crimes. It also showed that the highest number of crimes occurred from 8am to 12pm with 12pm being the peak time. Districts 1 and 18 had the most crimes because of their high population and high income per capita. Using the classification models and SMOTE for imbalanced data, the models performed fairly well at predicting the categories of crimes. This could be improved by increasing the number of features and the relevancy of the features.

Features such as beat or block which specify the location of a crime in addition to time ranges, are influential features when predicting crime categories. Ensuring that data fed into the classification algorithms are not imbalanced in terms of arrest status and crime categories are very important in ensuring model accuracy. This modelling attempt highlights the importance of predicting the locations and times of various crimes. This can be used in the decision making process within the city of Chicago when determining when and where policing and community services can be allocated.

E. Literature Review 4

The fourth review was titled, "*Application Of Data Analytics Techniques In Analyzing Crimes Background*"[6]. With the abundance of crime related data, there is immense benefit in gaining a deeper understanding of the impact of socioeconomic indicators like poverty, education and unemployment on crime rates. Crime data can also be used to gain more information and insights on why, when and how criminal activities are carried out in order to help Law enforcement agencies speed up the process of solving crimes and predict the possibility of crimes in the future in certain areas and at certain times.
For the purposes of this research, the goal was to extract meaningful information related to the occurrence of crimes such as identifying hotspots, peak crime periods, rate of arrests in certain locations and the socioeconomic factors (% House Crowding, % below poverty rate, % without high school diploma, hardship index) that contribute the most to crime rates.

Chicago crime data from records of crime cases by the Chicago Police Department from 2008 to 2012 was used to identify dangerous places (crime hotspots), time periods (months of the year) with high and low criminal activity and rate of arrests in different community areas using visualization techniques. A second dataset containing socioeconomic indicators of public health and a hardship index from 2008 to 2012 was used to show the relationship between these socioeconomic factors and the rate of occurrence of crimes (crime rate) by fitting a linear regression model.

The visualizations revealed that in Chicago from 2008 to 2012, more crimes took place during warm weather in the months of May, June, July and August whereas less crimes took place during the colder months of November, December, January and February. The linear regression model showed that there is a positive correlation between crime rate and the hardship index meaning that when people fall on hard times crime rates tend to increase. A negative correlation between % aged 25+ without a high school diploma and the crime rate shows that the more educated the population the less crime there is. A negative correlation between PAUO and crime rate showed that most crimes are committed by people between the ages of 16 and 64. Older (>64) and younger (< 18) groups are less prone to crime. The analysis also revealed that most crimes occur from 6pm to midnight and on January 1st of every year.

Analysis of historical crime data coupled with socioeconomic information, can be used to better police certain areas with higher poverty rates, lower education levels, at certain times of the day and during certain times of the year. This information can also be used to inform decisions related to business development and community outreach projects.

3. EXPLORATORY DATA ANALYSIS

A. Rationale for Data Choice

For the purposes of this research project, I chose to work with two datasets, namely NYPD Arrests[1] and Toronto Major Crime Indicators (MCI)[2]. These datasets were chosen because NYC and Toronto are two major metropolitans where many crimes across a wide variety of offense types are likely to occur. The data went as far back as 2006 for the NYPD Arrests and as far back as 2000 for the Toronto MCIs. Combined, these two datasets contain at least six million crime arrests data points, thereby providing more than enough data points for data mining purposes. Whereas these two datasets both contained information related to the dates of crimes and the types of crimes, only the NYPD Arrests dataset contained information about the perpetrators' age group, gender and race. On the other hand, only the Toronto MCI dataset contained information about the location and premises where a crime occurred. Since no one dataset contained all the desired fields, both were used to create a master dataset for use in both the descriptive and prescriptive analytics portions of this Major Research Project.

B. Data Sources and Background Information

The first dataset, obtained from the NYC Open Data website[1], contains NYPD Arrests including a breakdown of every arrest made in NYC by the NYPD from 2006 to 2024. According to this website, arrests which involve multiple charges are classified according to the top charge. In addition, arrests occurring in open areas such as parks or beaches may be geo-coded as occurring on streets or intersections bordering the area and arrests occurring on a moving train or transit system were geo-coded as occurring at the train's next stop (street intersection).

The Toronto Major Crime Indicators (MCI) dataset[2] contains all crime occurrences by reported date from 2000 to 2024. Some of the Major Crime Indicator categories noted are Assault, Break and Enter, Auto Theft, Robbery and Theft Over (Excludes Sexual Assaults). This data is provided at the offence and/or victim level, therefore one occurrence number may have several rows of data associated with various MCIs used to categorize the occurrence. This data does not include occurrences that have been deemed unfounded. The definition of unfounded according to Statistics Canada[7] is: "It has been determined through police investigation that the offence reported did not occur, nor was it attempted". This data includes all MCI occurrences reported to the Toronto Police Service, including those where the location was not able to be verified. As a result, coordinate fields may appear blank. Likewise, this includes occurrences where the coordinate location is outside the City of Toronto. It is important to note that the fields have been included for both the old 140 City of Toronto Neighbourhoods structure as well as the new 158 City of Toronto Neighbourhoods structure. According to the Toronto open data website[2] this dataset has an overall score of 98% and a grade of gold for its data quality rating.

C. Data Description and Data Dictionary

The NYPD Arrests dataset contains 19 columns and 5,986,025 rows of data. The following is a table containing all its fields/columns and their corresponding descriptions.

Field Name	Description
ARREST_KEY	Randomly generated persistent ID for each arrest
ARREST_DATE	Exact date of arrest for the reported event
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
KY_CD	Three digit internal classification code (more general category than PD code)
OFNS_DESC	Description of internal classification corresponding with KY code (more general category than PD description)
LAW_CODE	Law code charges corresponding to the NYS Penal Law, VTL and other various local laws
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
ARREST_BORO	Borough of arrest. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens)
ARREST_PRECINCT	Precinct where the arrest occurred
JURISDICTION_CODE	Jurisdiction responsible for arrest. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions
AGE_GROUP	Perpetrator’s age within a category
PERP_SEX	Perpetrator’s sex description
PERP_RACE	Perpetrator’s race description
X_COORD_CD	Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

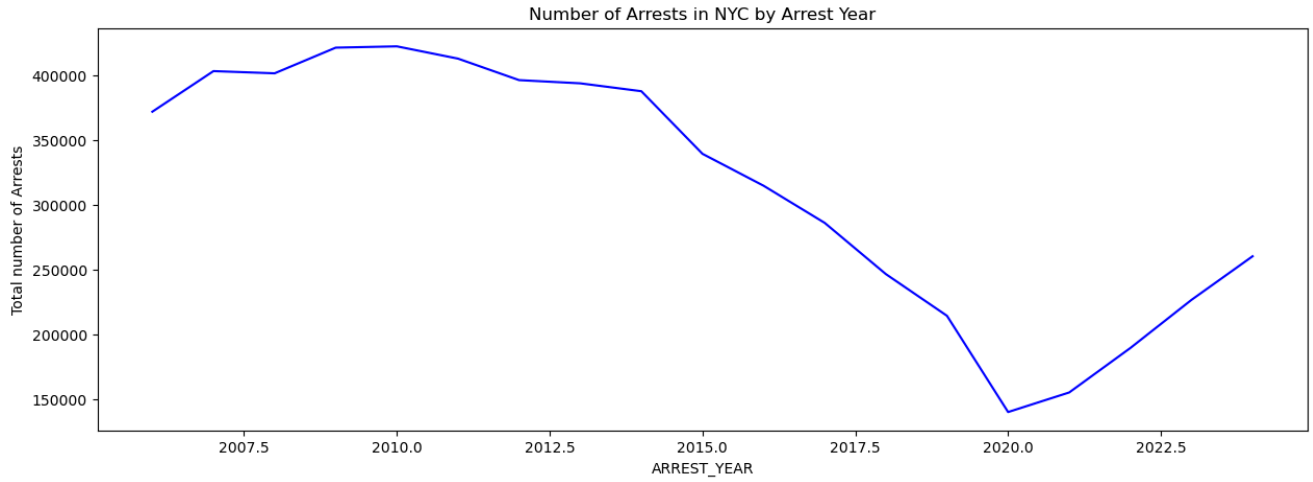
The Toronto Major Crime Indicators (MCI) dataset contains 29 columns and 420,200 rows of data. The following is a table containing all its fields/columns and their corresponding descriptions.

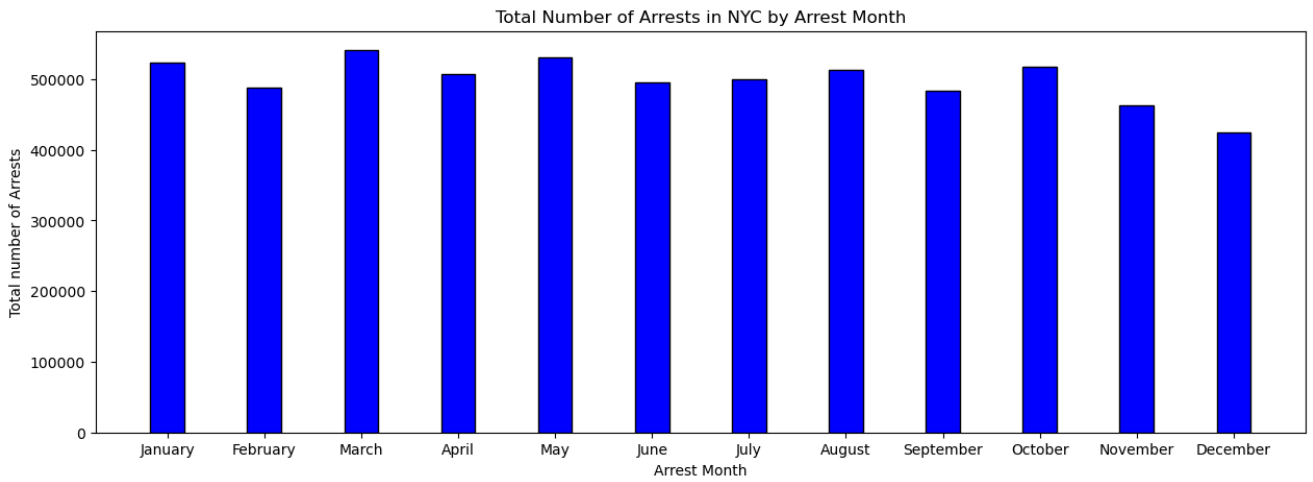
Field Name	Description
EVENT_UNIQUE_ID	Offence Number
REPORT_DATE	Date Offence was Reported
OCC_DATE	Date of Offence
REPORT_YEAR	Year Offence was Reported
REPORT_MONTH	Month Offence was Reported
REPORT_DAY	Day of the Month Offence was Reported
REPORT_DOY	Day of the Year Offence was Reported
REPORT_DOW	Day of the Week Offence was Reported
REPORT_HOUR	Hour Offence was Reported
OCC_YEAR	Year Offence Occurred
OCC_MONTH	Month Offence Occurred
OCC_DAY	Day of the Month Offence Occurred
OCC_DOY	Day of the Year Offence Occurred
OCC_DOW	Day of the Week Offence Occurred
OCC_HOUR	Hour Offence Occurred
DIVISION	Police Division where Offence Occurred
LOCATION_TYPE	Location Type of Offence
PREMISES_TYPE	Premises Type of Offence
UCR_CODE	UCR Code for Offence
UCR_EXT	UCR Extension for Offence
OFFENCE	Title of Offence
MCI_CATEGORY	MCI Category of Occurrence
HOOD_158	Identifier of Neighbourhood using City of Toronto's new 158 neighbourhood structure
NEIGHBOURHOOD_158	Name of Neighbourhood using City of Toronto's new 158 neighbourhood structure
HOOD_140	Identifier of Neighbourhood using City of Toronto's old 140 neighbourhood structure

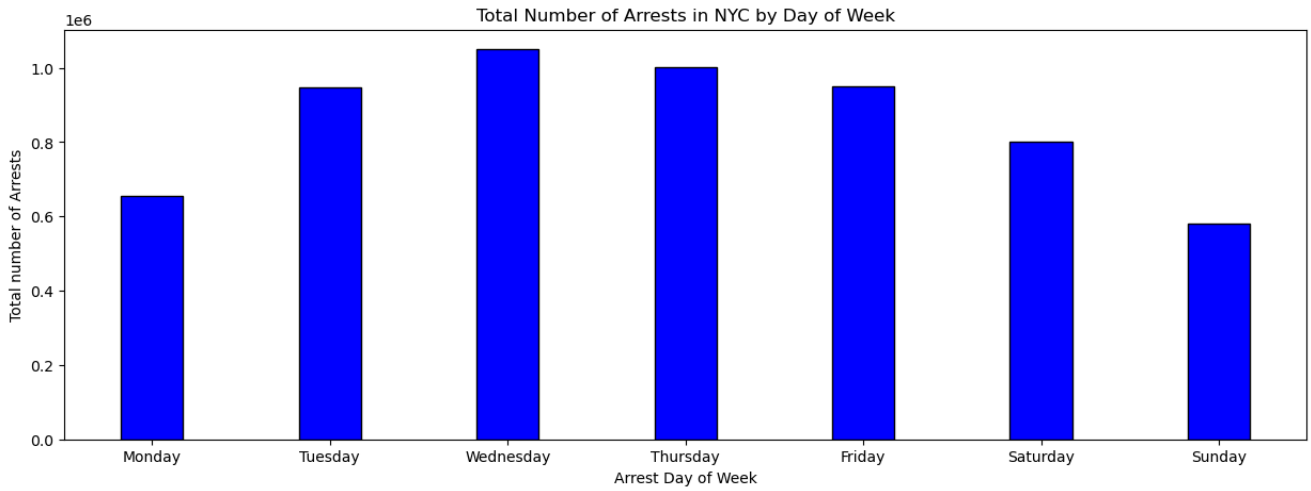
NEIGHBOURHOOD_140	Name of Neighbourhood using City of Toronto's old 140 neighbourhood structure
LONG_WGS84	Longitude coordinate
LAT_WGS84	Latitude coordinate

D. Exploratory Data Analysis and Observations - NYC Arrests

Using the ARREST_DATE field and first converting it to datetime format, the year, month and day of week features were extracted. The following graphs show the number of arrests across the values of ARREST_YEAR, ARREST_MONTH and ARREST_DOW. There was a sharp decline in the number of arrests from about 2014 to 2020. Across all the months the number of arrests fall between 400,000 and 500,000 with the most crimes occurring in January, March, May, August and October. In addition, most arrests tend to occur on Wednesday, Thursday and Friday.

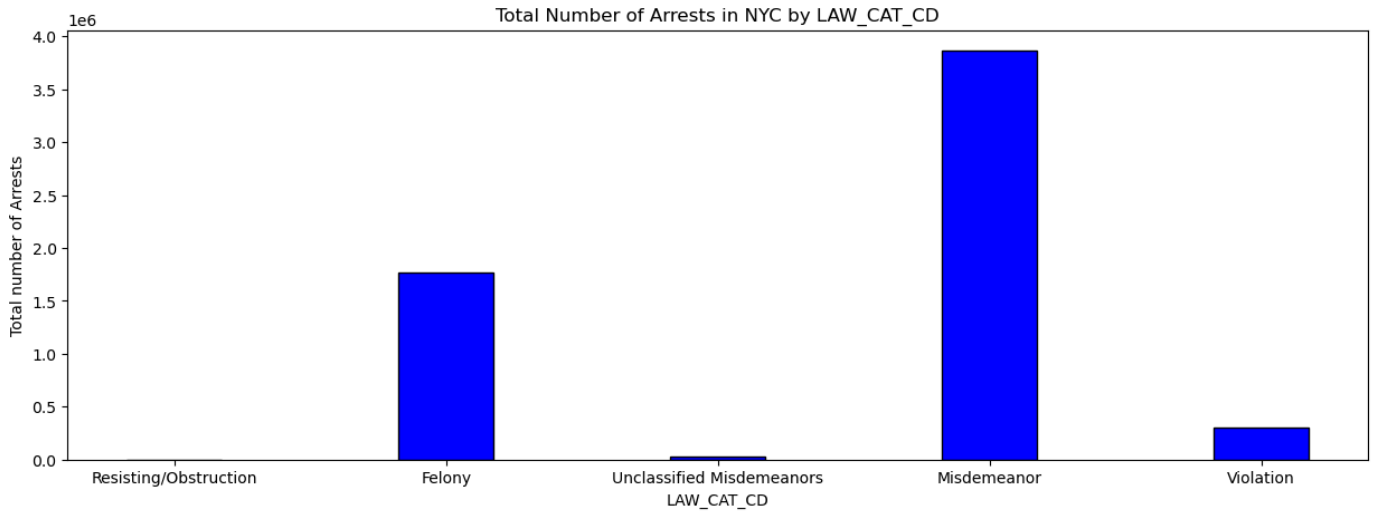






Using the LAW_CAT_CD field, the original distribution across values was obtained in table format. Next, all null values were dropped. Research revealed that a LAW_CAT_CD value of 9 means Resisting/Obstruction so it was labelled as such. The following histogram was then generated.This histogram reveals that most arrests tend to be for misdemeanours followed by felonies. According to wikipedia, a misdemeanor is a nonindictable offense less serious than a felony.

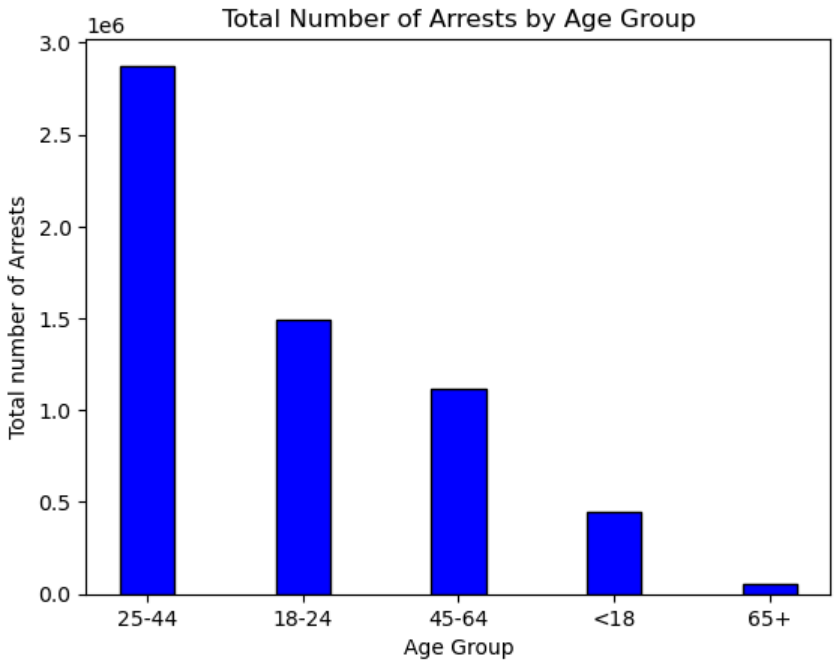
	LAW_CAT_CD	count
6	(null)	10
5	9	1801
1	F	1767834
3	I	27200
0	M	3866398
2	V	297792
4	NaN	24990



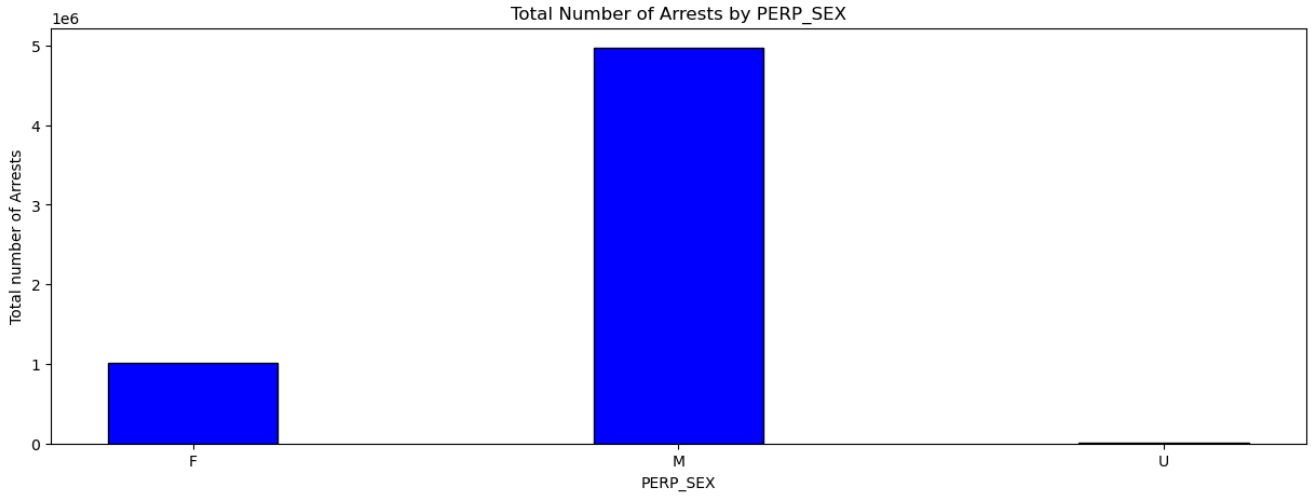
The field OFNS_DESC contains 91 different values. Before any text cleansing and preprocessing, the following are the top common types of OFNS_DESC values. Later in this report, a description of how the values of this field will be cleaned up and binned will be discussed. As an initial assessment, it appears that most offense types are assaults, theft/robbery/larceny and drug related. This will be used to guide the binning process and create more defined offense types.

```
OFNS_DESC
DANGEROUS DRUGS                          1144059
ASSAULT 3 & RELATED OFFENSES              645227
OTHER OFFENSES RELATED TO THEFT           316812
PETIT LARCENY                             306203
FELONY ASSAULT                            288434
OTHER STATE LAWS                          247627
VEHICLE AND TRAFFIC LAWS                  243577
MISCELLANEOUS PENAL LAW                   240181
DANGEROUS WEAPONS                         233174
CRIMINAL TRESPASS                         208608
ROBBERY                                   202169
OTHER TRAFFIC INFRACTION                  183344
GRAND LARCENY                             164147
OFFENSES AGAINST PUBLIC ADMINISTRATION    156544
POSSESSION OF STOLEN PROPERTY 5           149877
CRIMINAL MISCHIEF & RELATED OFFENSES      146039
INTOXICATED & IMPAIRED DRIVING            111592
FORGERY                                    99658
BURGLARY                                   96064
OTHER STATE LAWS (NON PENAL LAW)           79681
Name: count, dtype: int64
```

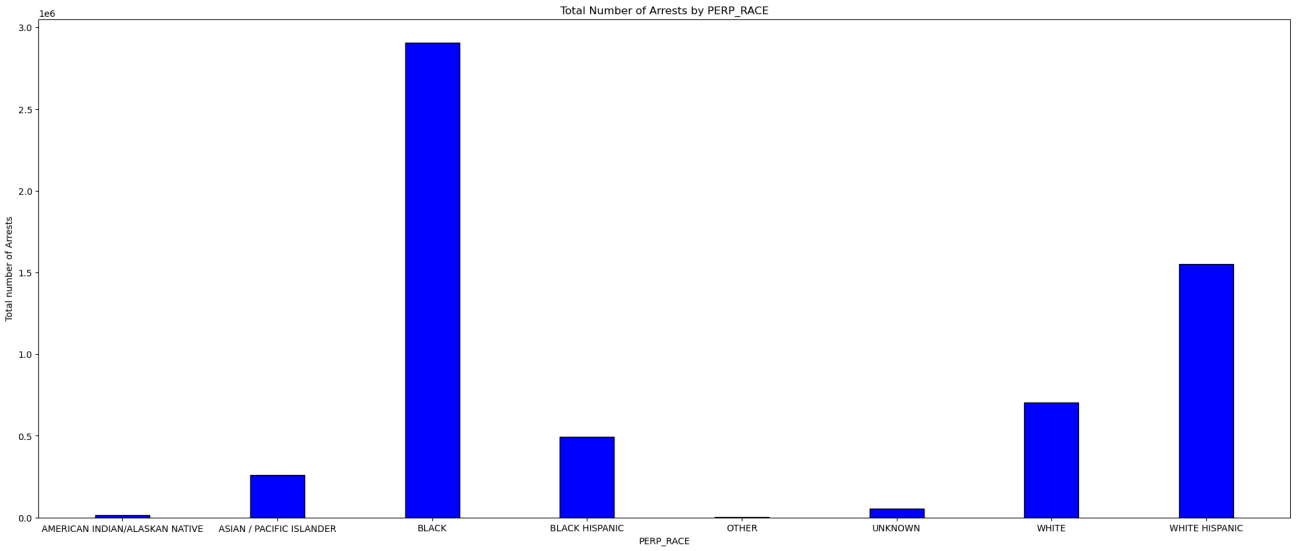
Any AGE_GROUP values with counts less than 8 were removed from the dataset to reveal the following distribution. Most crimes are committed by individuals aged 25 to 44 with people 65 and over committing the least crimes.



The field PERP_SEX contains the following three values. For the predictive and PowerBI portions of this analytics report, the U value which stands for undisclosed will be deleted from the dataset.

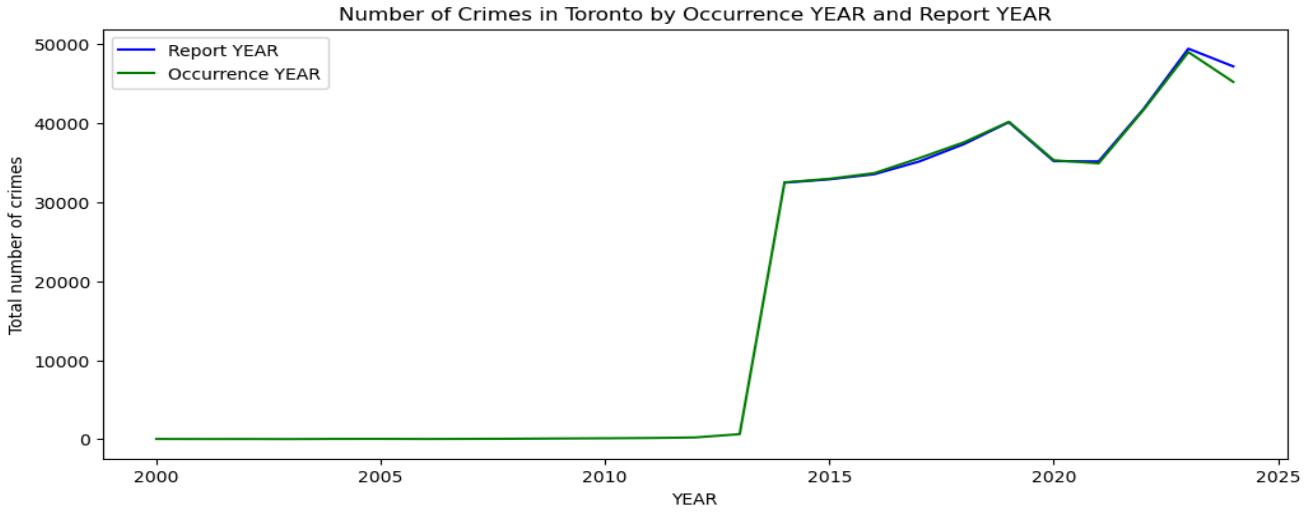


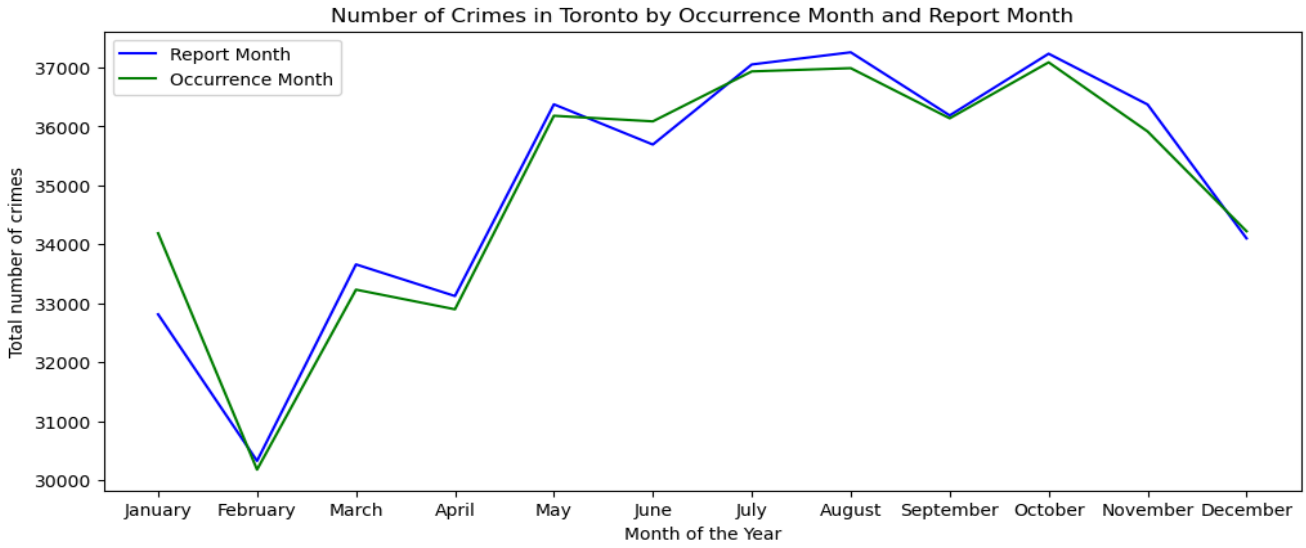
The following histogram is a distribution of the values of the PERP_RACE field. The graph reveals that most crimes are committed by blacks and white hispanics.



E. Exploratory Data Analysis and Observations - Toronto Major Crime Indicators (MCI)

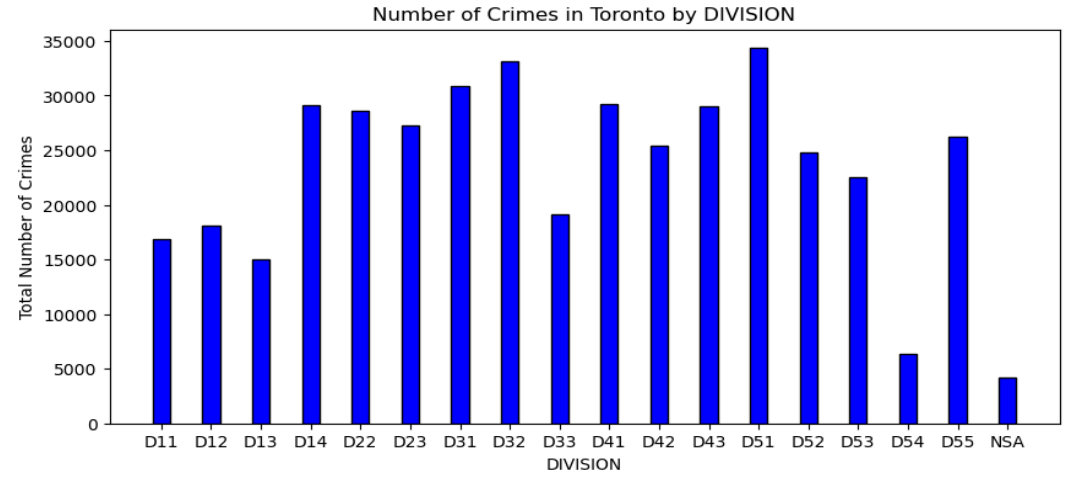
The graphs below show that for the Toronto crimes dataset, crimes tend to occur from 2000 to 2024, whereas these same crimes tend to be reported from 2014 to 2024. It also shows that in any given year, the number of crimes increase from February to August and decrease from October to February. Statistical analysis also reveals that the mean number of days between when a crime occurs and when it is reported is 27 days with a maximum time lapse of 20,011 days.

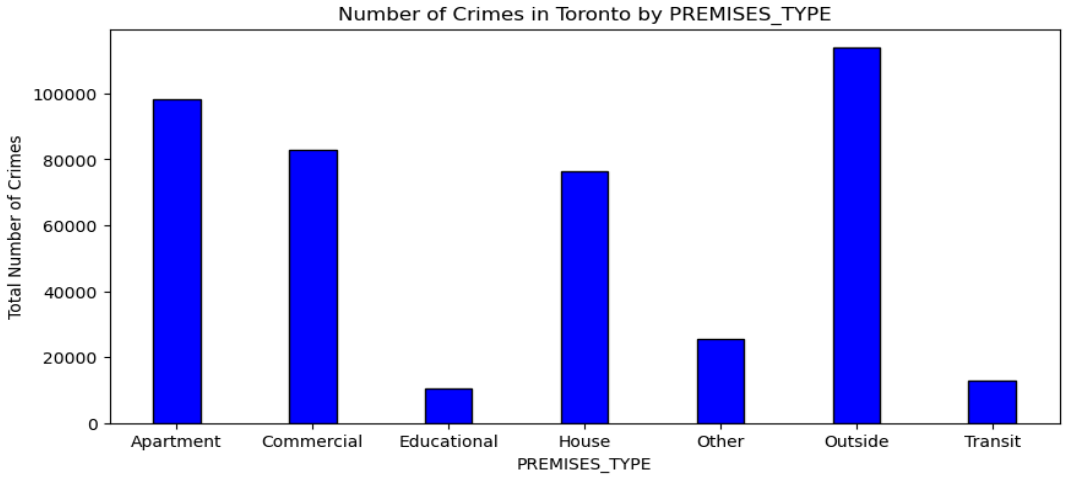





```
count    420200.000000
mean         27.919357
std         303.972336
min           0.000000
25%           0.000000
50%           0.000000
75%           1.000000
max       20011.000000
Name: TIME_LAPSE_OCC_REPORT_DAYS, dtype: float64
```

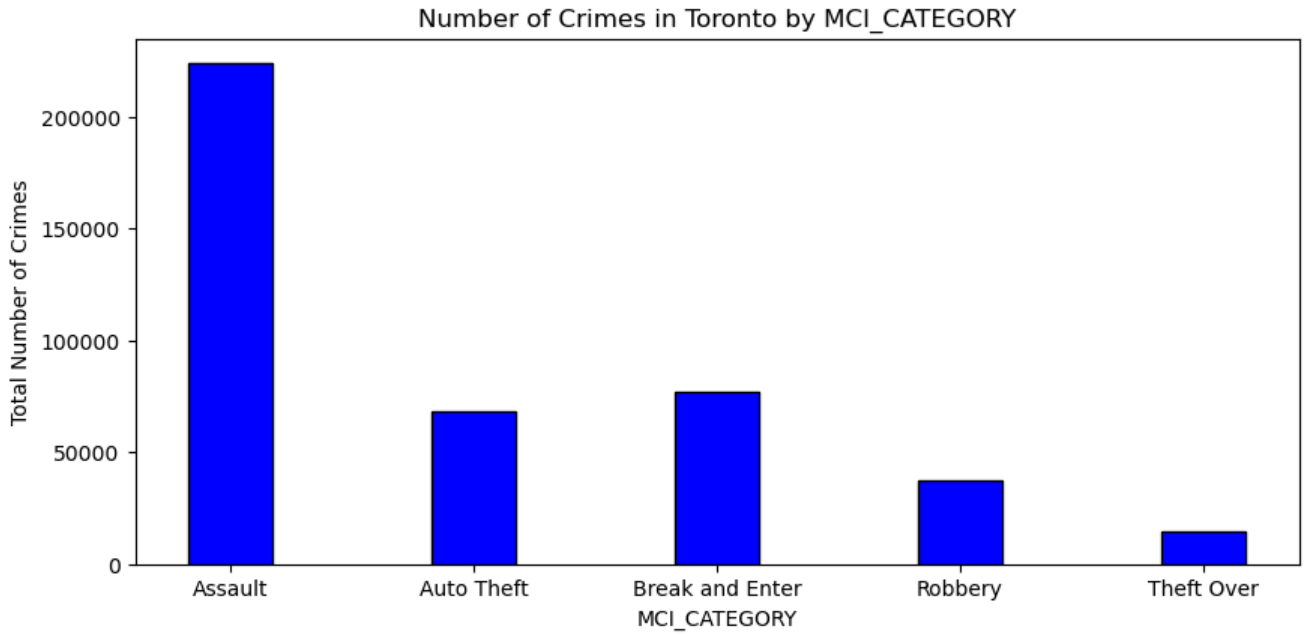
The following Toronto crimes graphs reveal that most crimes occur in Divisions D51 and D32. The PREMISES_TYPE field reveals that most crimes also occur outside, in apartments, in commercial buildings and in houses. The LOCATION_TYPE field shows that most crimes occur in apartments, homes, streets and highways and commercial spaces. This field will not be used in any descriptive and predictive analytics because it requires a lot of text cleansing and preprocessing. The PREMISES_TYPE field will be used instead.





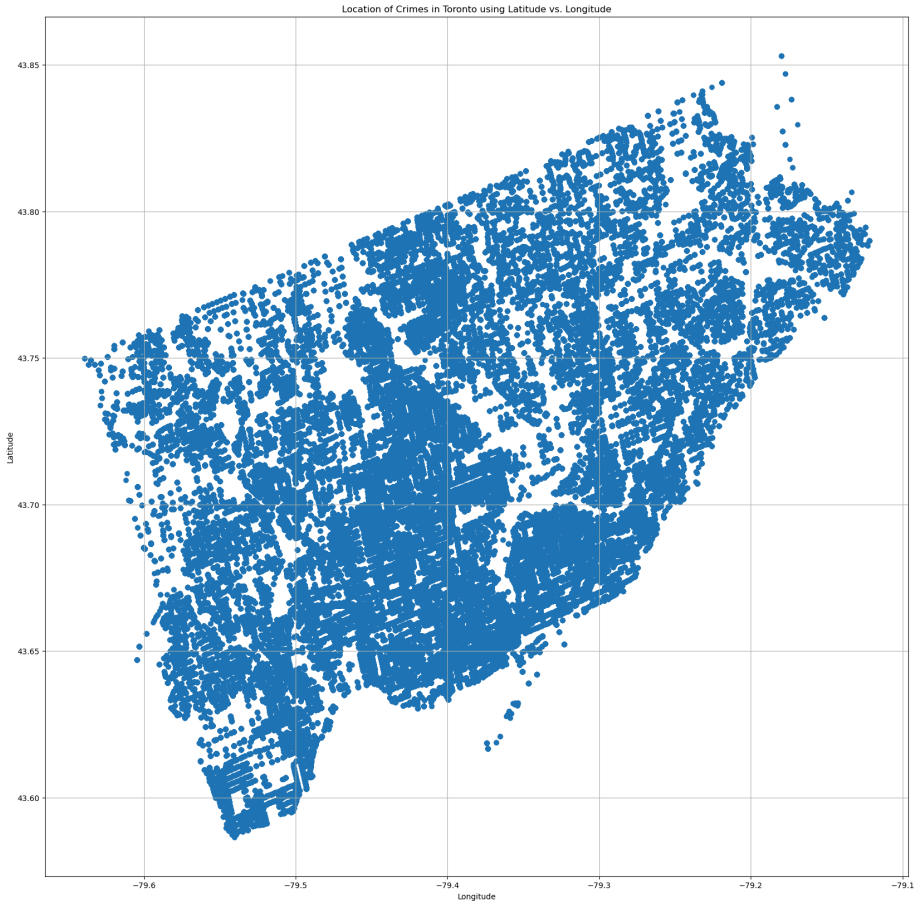
	LOCATION_TYPE	count
0	Apartment (Rooming House, Condo)	98358
1	Single Home, House (Attach Garage, Cottage, Mobile)	76280
2	Streets, Roads, Highways (Bicycle Path, Private Road)	65189
3	Other Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg	47984
4	Parking Lots (Apt., Commercial Or Non-Commercial)	39134
5	Bar / Restaurant	15824
6	Open Areas (Lakes, Parks, Rivers)	8334
7	Schools During Supervised Activity	6175
8	Ttc Subway Station	5544
9	Convenience Stores	5346

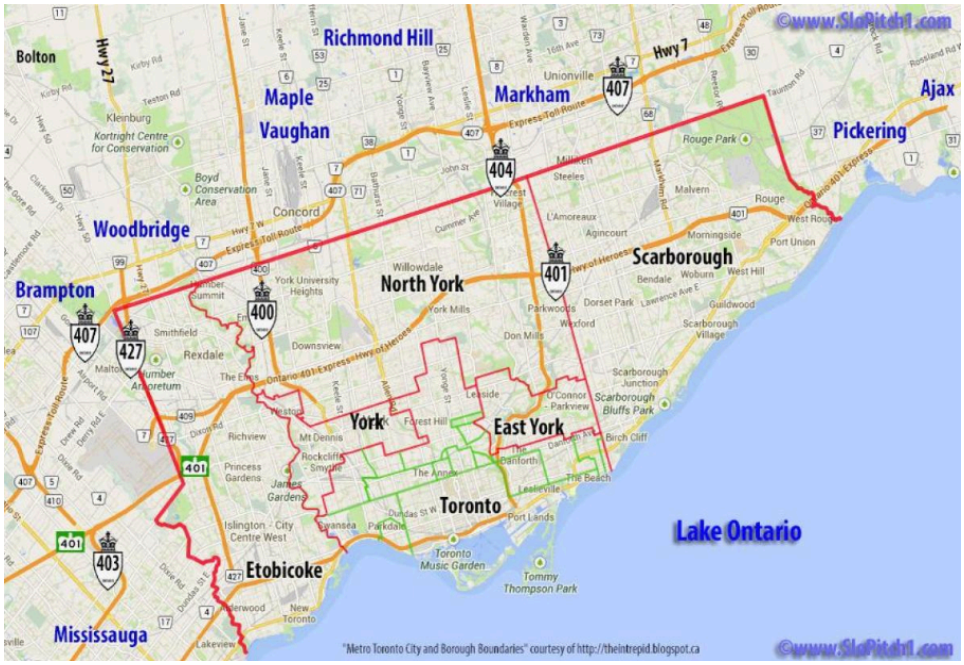
The following Toronto crimes dataset contains both an OFFENCE and MCI_CATEGORY field that provides insight into the types of crimes. MCI_CATEGORY is very clean and doesn't require a lot of text preprocessing and will be used in all the descriptive and predictive analytics. Most of the crimes that occur are usually Assault and Break and Enter.



MCI_CATEGORY	OFFENCE
Assault	{Assault With Weapon, Aggravated Aslt Peace Officer, Assault - Force/Thrt/Impede, Assault, Assault Peace Officer Wpn/Cbh, Discharge Firearm - Recklessly, Pointing A Firearm, Use Firearm / Immit Commit Off, Assault Peace Officer, Discharge Firearm With Intent, Disarming Peace/Public Officer, Assault - Resist/ Prevent Seiz, Hoax Terrorism Causing Bodily, Administering Noxious Thing, Assault Bodily Harm, Aggravated Assault Avails Pros, Crim Negligence Bodily Harm, Aggravated Assault, Traps Likely Cause Bodily Harm, Set/Place Trap/Intend Death/Bh, Air Gun Or Pistol: Bodily Harm, Unlawfully Causing Bodily Harm}
Auto Theft	{Theft Of Motor Vehicle}
Break and Enter	{B&E - M/Veh To Steal Firearm, B&E - To Steal Firearm, Unlawfully In Dwelling-House, B&E W'Intent, B&E, B&E Out}
Robbery	{Robbery - Atm, Robbery - Purse Snatch, Robbery - Business, Robbery - Vehicle Jacking, Robbery - Other, Robbery - Mugging, Robbery - Delivery Person, Robbery - Swarming, Robbery With Weapon, Robbery - Home Invasion, Robbery To Steal Firearm, Robbery - Financial Institute, Robbery - Armoured Car, Robbery - Taxi}
Theft Over	{Theft - Misapprop Funds Over, Theft Over - Shoplifting, Theft Of Utilities Over, Theft Over - Bicycle, Theft Over, Theft From Mail / Bag / Key, Theft Over - Distraction, Theft From Motor Vehicle Over}

The following map shows the location of crimes in Toronto by plotting latitude and longitude. Compared to an actual map of Toronto[8], the data points indicate that most crimes happen within North York, York and East York.





4. METHODOLOGY AND EXPERIMENTS

A. Descriptive Analytics using PowerBI

The goal of this methodology is to study the arrest and crimes data collected over time from two major metropolitans, namely New York City[1] and Toronto[2] to learn about when crimes/arrests usually take place, the nature of these crimes, the characteristics of the individuals who tend to commit these crimes and the location and premises with the highest crime occurrences.

The Business Intelligence tool PowerBI will be used to design a dashboard containing advanced visualizations, trends and patterns that highlight where, when, who and why crimes tend to occur. These visualizations will also be used to highlight the differences and similarities between the crimes that occur in both of these major metropolitans.

NYC arrests and Toronto crimes were picked because of the wealth of data available. For the purposes of this study, only data from 2014 to 2024 will be used in both datasets. Both datasets contain date/time and offence type related information about each crime or arrest. The New York arrests dataset also contains information related to the age group, gender and race of the perpetrators. This type of information is not available in the Toronto crimes dataset.Only the Toronto dataset contains the location and premises type of the crimes.

By analysing busy, congested and populated areas one can get a better understanding of human behaviour. This type of research and data analysis is very useful when providing insights that can help mitigate and predict criminal activity, solve cases, discover trends and provide resources to communities that will help improve overall public safety.

Before any visualizations were created in PowerBI, the underlying dataset was curated by engineering features for both the NYC Arrests and Toronto Crimes datasets and then combining them. Using the NYC Arrests dataset, the CRIME_YEAR, CRIME_MONTH and CRIME_DOW were extracted from the ARREST_DATE field. The numeric values for CRIME_MONTH were mapped to actual month names. The same was done for the CRIME_DOW with the numeric values being mapped to the actual name of each day of the week. The undisclosed value was removed from the PERP_SEX field. The abbreviations for LAW_CAT_CD were replaced with full names and some values were dropped. An extensive amount of data cleansing was done to the OFNS_DESC field by binning very similar crimes into categories such as Assault, Robbery, Murder, Fraud and Child Crimes ultimately creating the TYPE_OF_OFFENSE field. The much cleaner OFFENCE field in the Toronto Crimes dataset was used to guide the binning and regrouping of OFNS_DESC. The Metropolitan field was hardcoded to NYC.

Within the Toronto Crimes dataset the fields CRIME_YEAR, CRIME_MONTH, CRIME_DOW, CRIME_HOUR, LONG_WGS84, LAT_WGS84 and PREMISES_TYPE already existed. The dataset was filtered to ensure that only crimes from 2014 to 2024 were included. This was done to match the NYC Arrests dataset above. TIME_LAPSE_OCC_REPORT_DAYS was computed by subtracting the OCC_DATE from the REPORT_DATE and

converting that to the number of days. The OFFENCE field was also cleaned and binned to create the TYPE_OF_OFFENSE field. The Metropolitan field was hardcoded to Toronto. Once both the NYC Arrests dataset and the Toronto Crimes dataset were curated they were combined to create the final PowerBI dataset that was fed into PowerBI.

The following table shows the fields from each dataset that were included in the descriptive analysis portion of this MRP.

NYC Arrests Columns/Fields	Toronto Crimes Columns/Fields
Metropolitan	Metropolitan
CRIME_YEAR	CRIME_YEAR
CRIME_MONTH	CRIME_MONTH
CRIME_DOW	CRIME_DOW
AGE_GROUP	CRIME_HOUR
PERP_SEX	TIME_LAPSE_OCC_REPORT_DAYS
PERP_RACE	PREMISES_TYPE
LAW_CAT_CD	LONG_WGS84
TYPE_OF_OFFENSE	LAT_WGS84
	TYPE_OF_OFFENSE

This section outlines how the descriptive analytics will be broken down to answer questions related to the when, what, who and where of crimes in major metropolitans (NYC and Toronto).

WHEN CRIMES OCCUR
● Time series that shows the number of crimes that occur every year. The New York Police Department Arrests data runs from 2006 to 2024. The Toronto Crime data runs from 2014 to 2024. This analysis will focus on the 10 years from 2014 to 2024 for both datasets. Comparative analysis will also be done to show how the crime and arrest data compared every year between NYC and Toronto. This is a historical dive to understanding trends of criminal behaviour over time.
● Knowing which months tend to have the highest occurrence of arrests/crimes is an important piece of information. This will be done for both the NYC and Toronto datasets in a comparative side by side fashion with a focus on the 10 years from 2014 to 2024.
● A graph showing the number of arrests that occur on each day of the week between NYC and Toronto from 2014 to 2024 will provide great insights for mitigating criminal behaviour.
● Crimes/arrests occur at different hours of the day and it would be beneficial to understand the hours when crimes occur the most and the least, while comparing that for both NYC and Toronto from 2014 to 2024.

● The Toronto Arrests dataset contains both an occurrence date and a report date for each arrest. The difference between these two dates will be computed to analyze the time lapse between when a crime occurs and when it is reported. Knowing how long it takes victims to report crimes is a piece of information that can help influence legislature and community programs geared at improving the trust between law officials and the community.
WHAT TYPES OF CRIMES OCCUR
● Histograms will be generated to show the number of arrests that occur for each TYPE_OF_OFFENSE such as Assault, Robbery and Murder among others for both NYC and Toronto.
● The NYC Arrests dataset contains a LAW_CAT_CD field which categorizes crimes as felony, misdemeanor or violation. A histogram will be created to show the total number of crimes at each level of offense.
● This field is probably the most important field and it will be layered into other visualizations focused on showing the types of crimes that occur yearly, monthly and for each day of the week. This will also be used to determine the types of crimes committed by gender, race, age group and premise type. This type of information would be useful to law officials when they develop programs for mitigating and preventing crimes in a region throughout the year.
WHO COMMITS THE CRIMES
● Only the NYC Arrests dataset contains age group, sex and race information. Visualizations will be used to show the total number of arrests by age group, sex and race.
● TYPE_OF_OFFENSE will also be compared against age group, sex and race to understand the characteristics of people who tend to commit certain crimes.
WHERE CRIMES OCCUR
● The PREMISE_TYPE field will be used to better understand which locations and premises have the most occurrences of crimes. This is a very useful piece of information as it will reveal if crimes occur the most at home or in public places.

B. Predictive analytics using Machine Learning

A. Aim of Study

The goal of this experiment is to predict the sex of a perpetrator by using historical arrest data from the New York City arrests dataset. This dataset is ideal because it contains information about when arrests occur, where they occur, the characteristics of people who tend to be arrested and different offense types. Because it contains all these key pieces of information unlike the Toronto Crimes dataset, it is ideal for building this classification model. The dataset will be curated, several models will be trained and the best model will be picked based on classification performance metrics. This model will then be enhanced and used to learn about which features are most important when predicting the sex of a perpetrator.

B. Curation of Machine Learning Dataset

a) Feature Engineering

Using the NYC Arrests dataset[1], the CRIME_YEAR, CRIME_MONTH and CRIME_DOW were extracted from the ARREST_DATE field. The numeric values for CRIME_MONTH were mapped to actual month names. The same was done for the CRIME_DOW with the numeric values being mapped to the actual name of each day of the week. The abbreviations for LAW_CAT_CD were replaced with full names. Research revealed that a LAW_CAT_CD value of 9 meant resisting or obstruction and this was used to map that LAW_CAT_CD value. An extensive amount of data cleansing was done to the OFNS_DESC field by binning very similar crimes into categories such as Assault, Robbery, Murder, Fraud and Child Crimes ultimately creating the TYPE_OF_OFFENSE field. The much cleaner OFFENCE field in the Toronto Crimes dataset was used to guide the binning and regrouping of OFNS_DESC.

b) Filtering

Only arrests from 2014 to 2024 were included in the data. The undisclosed value was removed from the PERP_SEX field which meant that only male and female values were considered. All unknown values for PERP_RACE were removed from the data. All null values for LAW_CAT_CD were removed from the dataset. The only AGE_GROUP values considered were <18, 18-24, 25-44, 45-64 and 65+.

C. Response (Dependent) and Independent Variable(s)

To build this classification model the input variables to be used are CRIME_YEAR, CRIME_MONTH, CRIME_DOW, ARREST_BORO, ARREST_PRECINCT, JURISDICTION_CODE, AGE_GROUP, PERP_RACE, TYPE_OF_OFFENSE and LAW_CAT_CD. The response variable is PERP_SEX.

D. Factors and Levels

In this experiment, the factors are the different classification algorithms being tested. In addition, the models will be trained on varying amounts of training data selected by applying stratified sampling of the PERP_SEX feature.

E. Experimental Design

a) Label Encoder

Label encoder will be applied to all the categorical variables to convert the categories from string values to numeric values given that many classification algorithms work best with numeric data. The features that to be label encoded were AGE_GROUP, PERP_SEX, LAW_CAT_CD, TYPE_OF_OFFENSE, PERP_RACE and ARREST_BORO.

b) Randomization (Train/Test Split)

The dataset was divided randomly into two sets, one for training containing 70% of the data and another for testing comprising 30% of the data.

c) Imbalance Data

Given that there was an imbalance in the number of male and female perpetrator data points, this balance was handled by using a Random Over Sampler which essentially creates new data points for the PERP_SEX value with the least data points.

F. Experiment Performance and Revisions

a) Experiment with different train size using Stratified Sampling

The entire dataset has a size of 2,397,604. Training multiple classification models using such a big dataset would take a lot of time. Stratified sampling will be used to train the models on smaller subsets of the data while ensuring an equal representation of data across the values for PERP_SEX.

b) Experiment with different ML Classification Algorithms

The classification model to predict PERP_SEX will be trained using several different classification algorithms including Random Forest Classifier, Ada Boost, XGBoost, Naive Bayes, Decision Trees, KNN Classifier, Logistic Regression and SVM Classifier.

c) Experiment with different Performance Metrics

After training each classification model mentioned above, these models will then be tested and assessed using many different performance metrics including F1 Score, Accuracy Score, Precision, Recall, Log Loss, AUC Score and Cohen Kappa.

G. Insights from final model

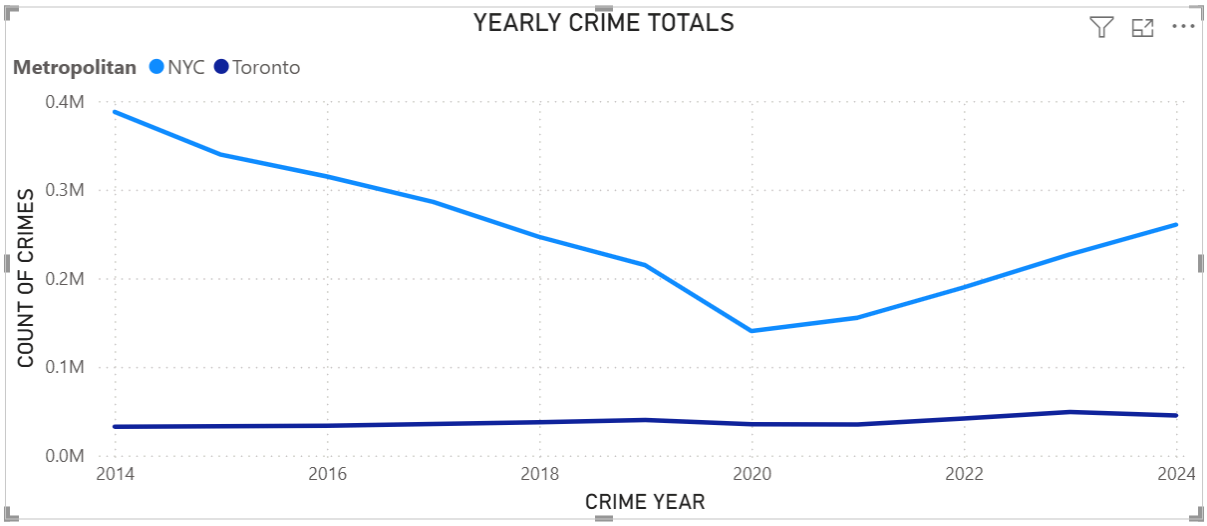
a) Select Best Model and Hyperparameter Tuning

The classification model with the best performance metrics for predicting PERP_SEX is finally picked and its parameters are tuned using hyperparameter tuning to get the most optimal version of that model. This optimal model is then further assessed by creating a classification report including a confusion matrix. The ROC curve is also drawn and a feature importance plot is created to show which input variables are most influential when predicting the gender of a perpetrator.

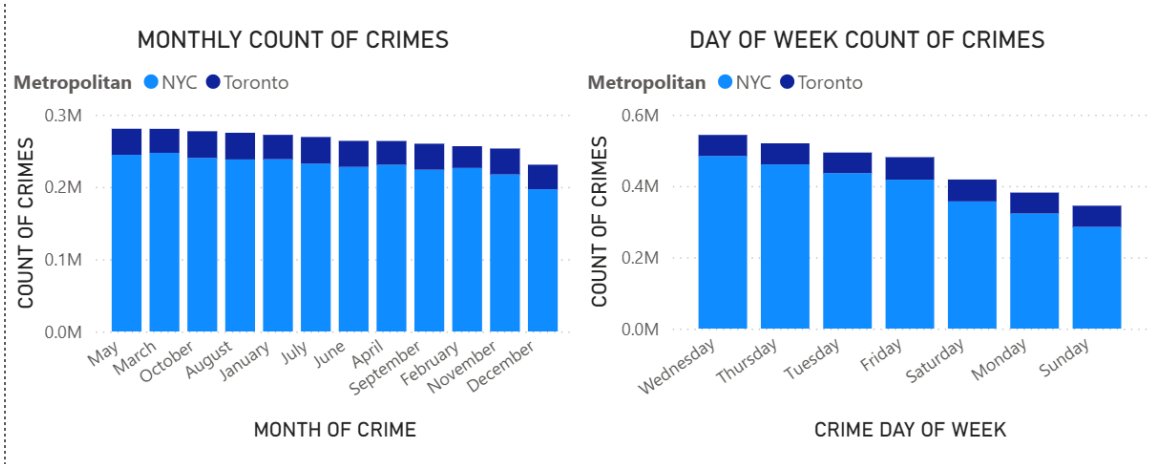
5. RESULTS AND DISCUSSION

A. Descriptive Analytics using PowerBI

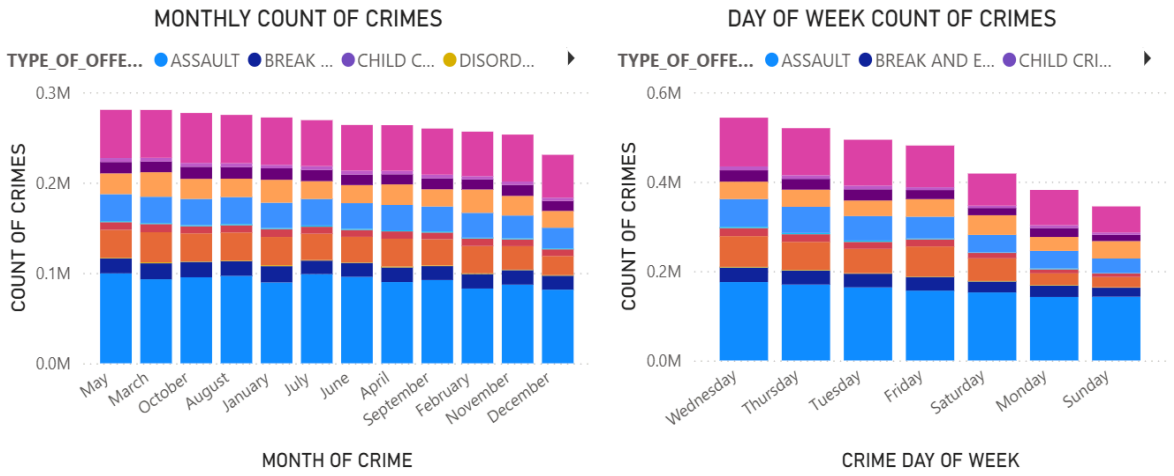
WHEN CRIMES OCCUR: A historical comparison of the total number of crimes occurring each year from 2014 to 2024 in both NYC and Toronto, shows that in this timeframe, New York City had five to ten times the number of crimes compared to Toronto. This could be due in part to the fact that New York City's population is four times that of Toronto. While Toronto's total number of crimes did not see any significant increases or decreases from 2014 to 2024, the opposite was true for NYC. NYC saw a sharp decline in the number of crimes from 2014 to 2020, the year that COVID started. Then from 2020, crimes started increasing sharply again.



To get a deeper understanding of when crimes occur the most, visualizations were generated to show the total number of crimes on both a monthly and daily basis. This type of information is very important for law officials to determine the days and months that require the most surveillance to curtail and control the crimes occurring in society thereby providing safety for communities at large.

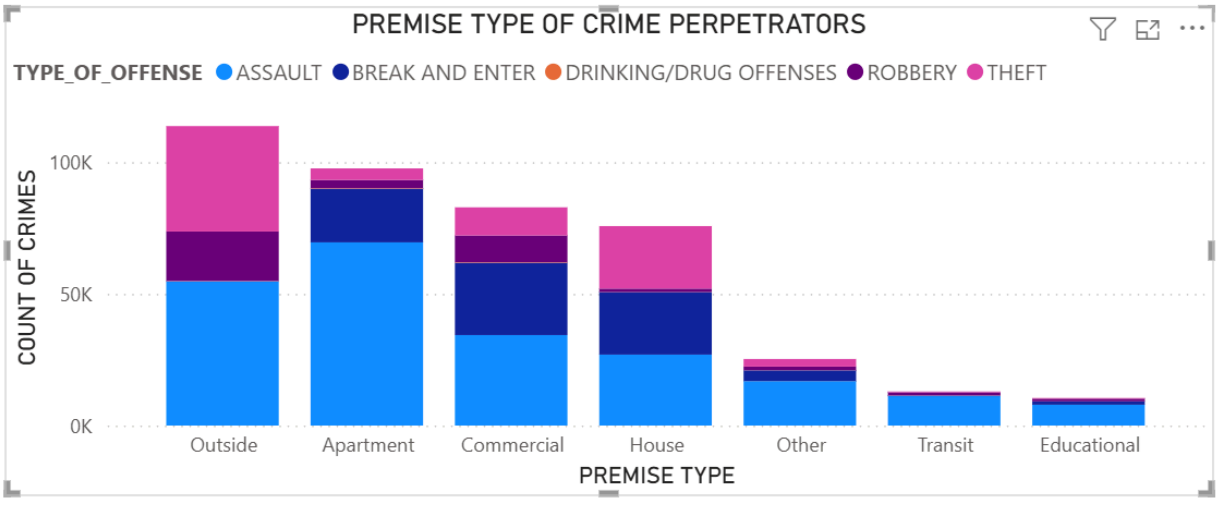


The graphs show that most crimes occur midweek from Wednesday to Friday. Sunday and Monday appear to have the least amount of crimes. While crimes occur consistently throughout the year, spring time appears to have the most crimes with the most crimes occurring in March and May. Oddly enough the data shows that the least crimes occur in December which could be as a result of the festivities and happiness of the Christmas season.

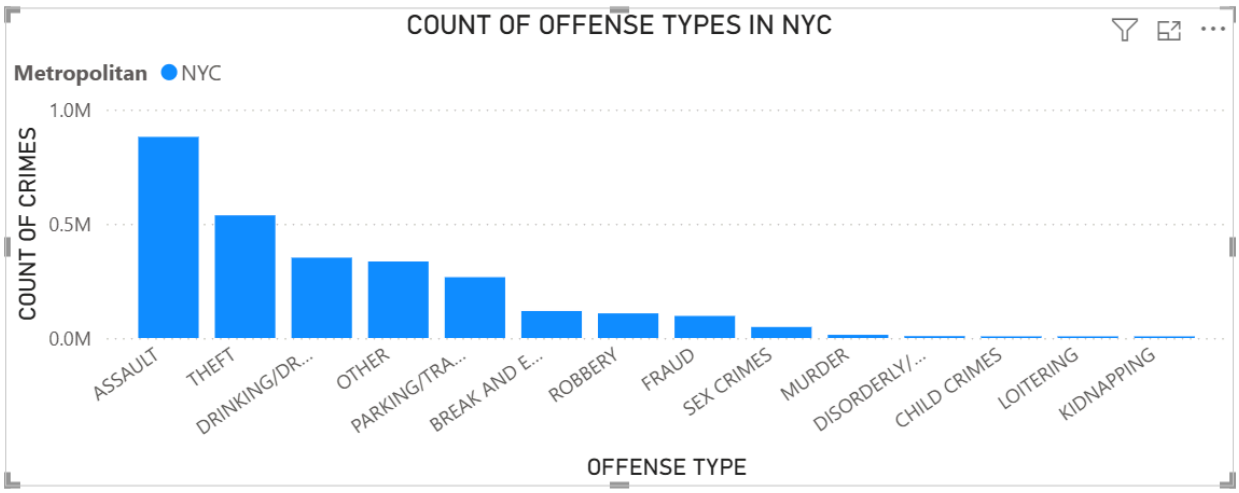


WHERE CRIMES OCCUR: Using the premise type field found only in the Toronto Crimes dataset, the data and visualizations show that most crimes tend to occur outside followed by occurrences in apartments, commercial buildings and houses. The least number of crimes occur in transit and educational locations.

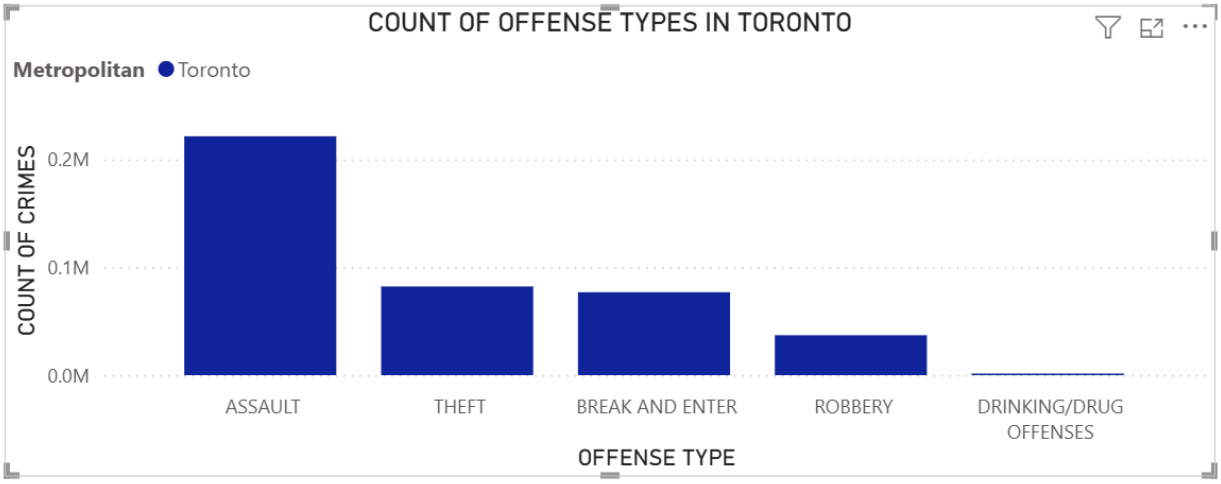
Drilling down to the types of crimes that occur at these premise types shows that assaults, theft and robbery tend to occur the most outside. In buildings such as apartments, houses and commercial spaces two types of crimes occur, namely assaults and break and entry. Transit and educational spaces tend to have mostly assaults. Wikipedia describes an assault as the act of causing physical harm or unwanted physical contact to another person, or the threat or attempt to do so.



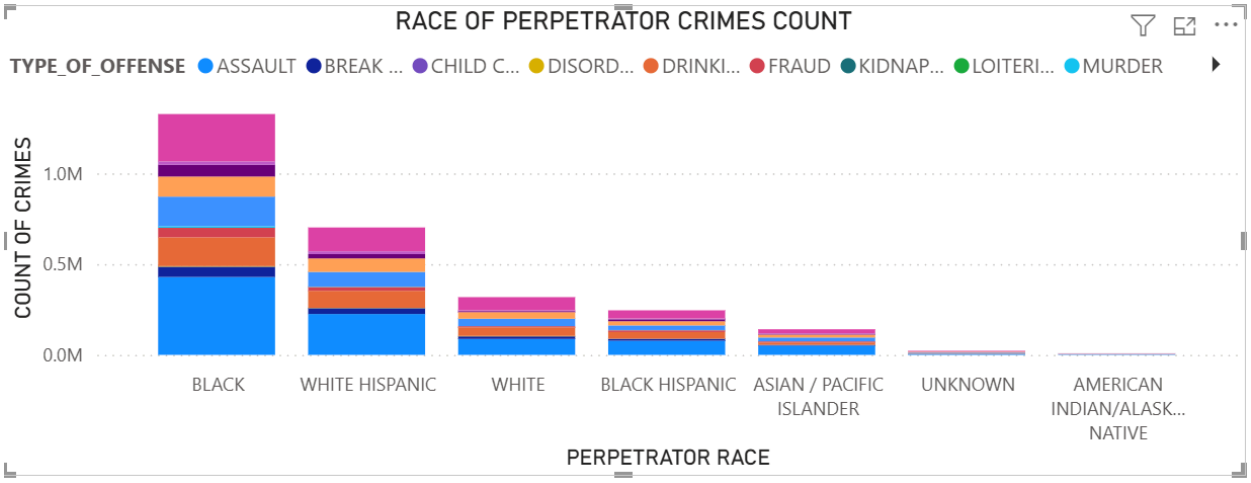
TYPE OF CRIMES: Assault and theft are the two main types of crimes that occur the most in both New York City and Toronto. This information is very useful in ensuring the correct types of resources are allocated across these major metropolitans to increase safety and decrease crimes. NYC also has lots of cases linked to drinking/drugs and parking/traffic. The least number of crimes are linked to muder, disorderly conduct, child crimes, loitering and kidnapping.



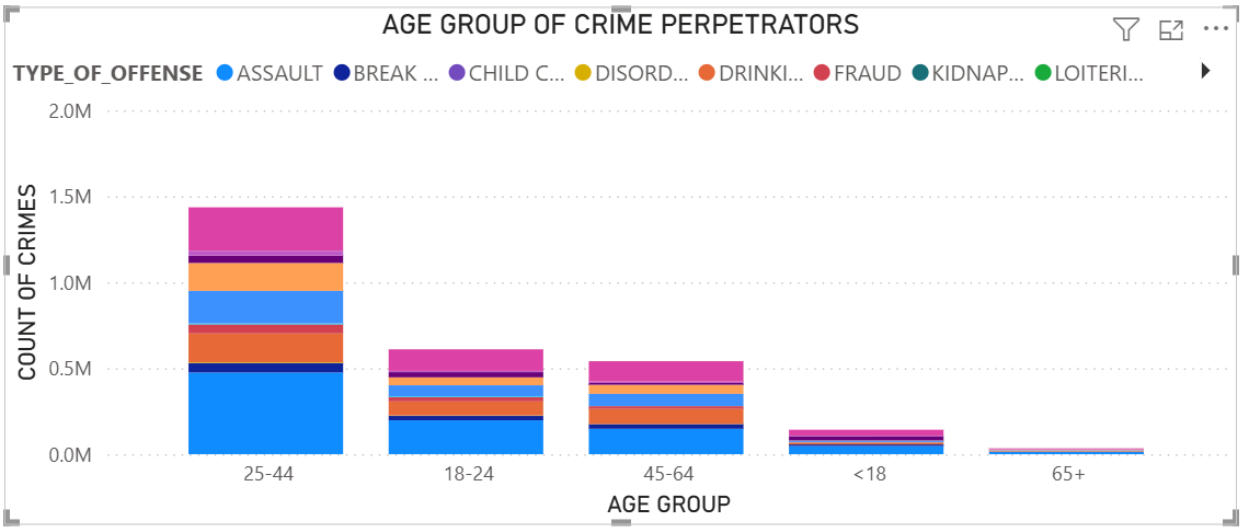
Given that theft, break and enter and robbery are very similar crimes, it is safe to say that crimes associated with stealing or attempts to steal are very common in Toronto. This could also explain why assaults occur the most. The least number of crimes tend to be linked to drinking and drugs.



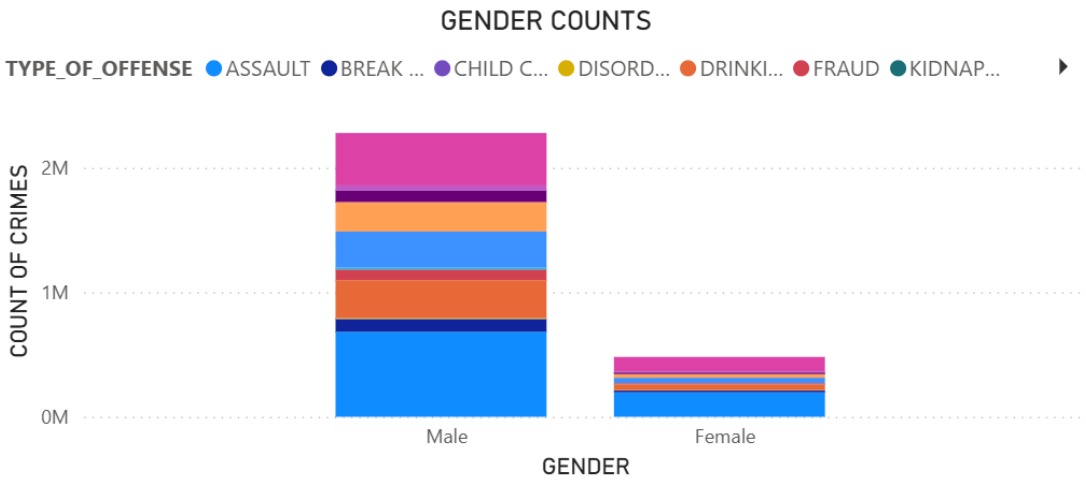
WHO COMMITS CRIMES: The NYC dataset contains information about the race, age group and gender of individuals committing crimes. In relation to race, most crimes are committed by blacks and white hispanics. The least amount of crimes are committed by american indians/alaskan natives and their crimes tend to be assault and theft. The crimes most associated with both blacks and white hispanics tend to be assault, drinking/drug offenses and theft. In decreasing order, whites, black hispanics and asian/pacific islanders are the next three races of people committing a lot of crimes.



Twice the amount of crimes are committed by individuals aged 25 to 44 compared to individuals aged 18 to 24 and those aged 45 to 64. Individuals aged less than 18 and 65 and above commit significantly less crimes compared to all the other age groups.



Males commit almost four times the amount of crimes compared to females. This information can be used to study why males tend to be more problematic and fuel conversions around community programs that can help uplift this group of people.



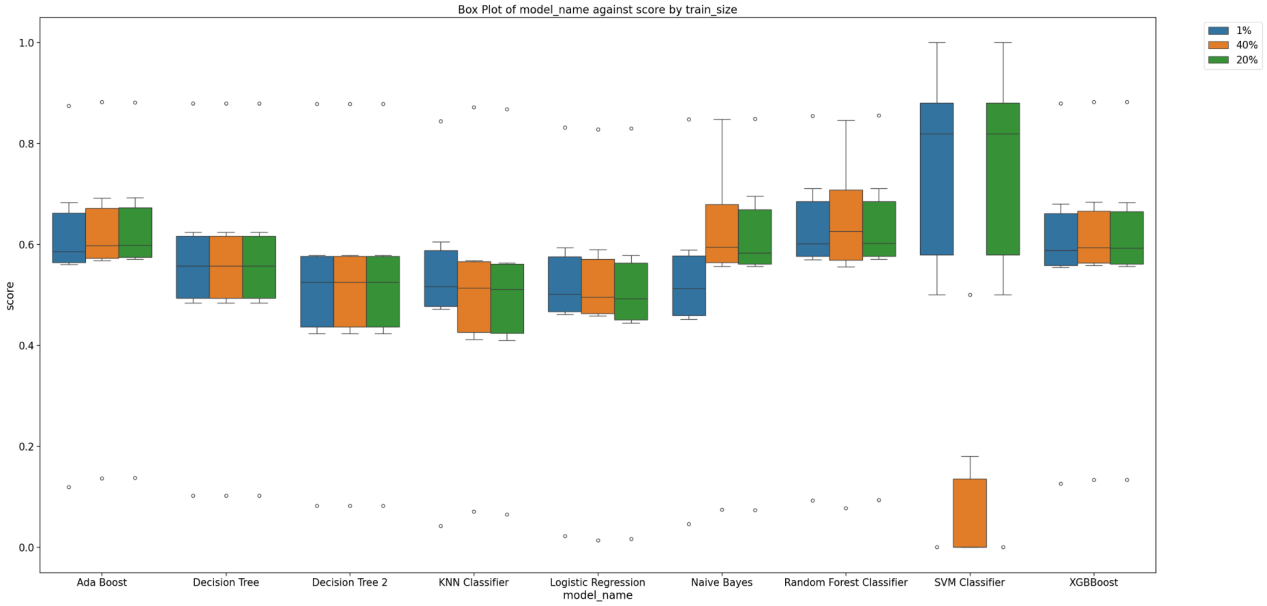
B. Predictive analytics using Machine Learning

A. Discussion of Experiments

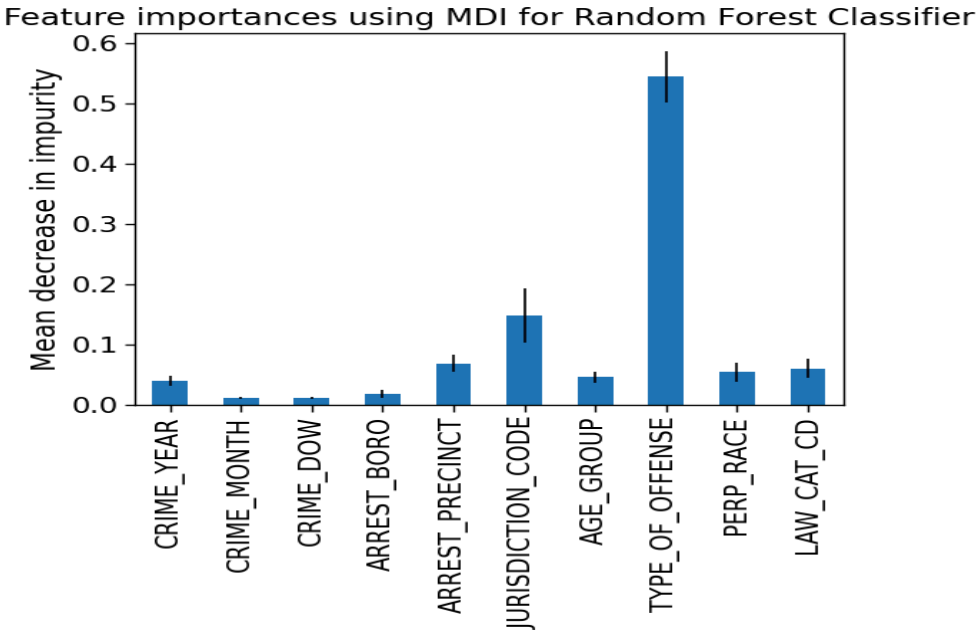
The following table contains all the classification performance metrics obtained as a result of all the experiments which focused on training different classification algorithms while also varying the percentage of the data from the training dataset (1%, 20%, 40%) used to train these models. At first glance it appears that the SVM Classifier and the Random Forest Classifier are the top performing algorithms for predicting the PERP_SEX of a crime data point.

	model_name	f1_score	accuracy_score	cohen_kappa_score	log_loss	precision_score	recall_score	roc_auc_score	train_size
7	Ada Boost	0.683172	0.573990	0.119733	14.713901	0.874674	0.560464	0.597933	1%
7	Ada Boost	0.692832	0.585355	0.137190	14.321396	0.881651	0.570624	0.611428	20%
7	Ada Boost	0.691352	0.584042	0.137005	14.366725	0.882012	0.568470	0.611607	40%
0	Decision Tree	0.624318	0.522848	0.102370	16.480297	0.879882	0.483798	0.591969	1%
0	Decision Tree	0.624318	0.522848	0.102370	16.480297	0.879882	0.483798	0.591969	20%
0	Decision Tree	0.624318	0.522848	0.102370	16.480297	0.879882	0.483798	0.591969	40%
1	Decision Tree 2	0.571022	0.479232	0.082236	17.986725	0.878645	0.422944	0.578865	1%
1	Decision Tree 2	0.571022	0.479232	0.082236	17.986725	0.878645	0.422944	0.578865	20%
1	Decision Tree 2	0.571022	0.479232	0.082236	17.986725	0.878645	0.422944	0.578865	40%
3	KNN Classifier	0.605320	0.495658	0.042294	17.419407	0.843795	0.471939	0.537643	1%
3	KNN Classifier	0.556983	0.465304	0.065094	18.467777	0.867543	0.410156	0.562920	20%
3	KNN Classifier	0.559724	0.468496	0.070401	18.357525	0.871410	0.412265	0.568030	40%
5	Logistic Regression	0.593987	0.482665	0.021969	17.868191	0.832278	0.461775	0.519640	1%
5	Logistic Regression	0.578959	0.470065	0.016625	18.303380	0.829700	0.444598	0.515141	20%
5	Logistic Regression	0.589994	0.477906	0.013928	18.032559	0.827623	0.458382	0.512463	40%
2	Naive Bayes	0.589303	0.484187	0.045870	17.815603	0.847908	0.451576	0.541911	1%
2	Naive Bayes	0.695476	0.577303	0.073424	14.599489	0.848956	0.588994	0.556611	20%
2	Naive Bayes	0.705131	0.586426	0.074612	14.284395	0.848094	0.603414	0.556357	40%
6	Random Forest Classifier	0.710879	0.594526	0.092909	14.004636	0.855121	0.608274	0.570191	1%
6	Random Forest Classifier	0.711276	0.595056	0.093973	13.986341	0.855519	0.608654	0.570985	20%
6	Random Forest Classifier	0.729774	0.610587	0.077914	13.449937	0.845968	0.641645	0.555612	40%
4	SVM Classifier	0.900796	0.819499	0.000000	6.234422	0.819499	1.000000	0.500000	1%
4	SVM Classifier	0.900796	0.819499	0.000000	6.234422	0.819499	1.000000	0.500000	20%
4	SVM Classifier	0.000000	0.180501	0.000000	28.304498	0.000000	0.000000	0.500000	40%
8	XGBBoost	0.679925	0.572360	0.126223	14.770224	0.879294	0.554255	0.604407	1%
8	XGBBoost	0.682842	0.576111	0.133365	14.640669	0.882583	0.556825	0.610248	20%
8	XGBBoost	0.684198	0.577352	0.133856	14.597789	0.882445	0.558686	0.610393	40%

Box plots of the performance metrics for each classification algorithm across each percent of the training set were plotted below. This visualization shows that SVM Classifier, XG Boost and the Random Forest Classifier are the top performers for predicting PERP_SEX. Oddly, the SVM Classifier performs poorly when trained on 40% of the train set. All the other classification algorithms perform similarly across all the train sizes (1%, 20%, 40%).



The Random Forest Classifier was picked as the best model and hyperparameter tuning was done to get the optimal combination of hyperparameter values. Using this optimal model, the feature importance graph was generated to determine which input features are most important when predicting the PERP_SEX of a given crime data point. The graph shows that when predicting PERP_SEX, the most important features in decreasing order are the type of offense, the jurisdiction and precinct of the offense and the race of the perpetrator.



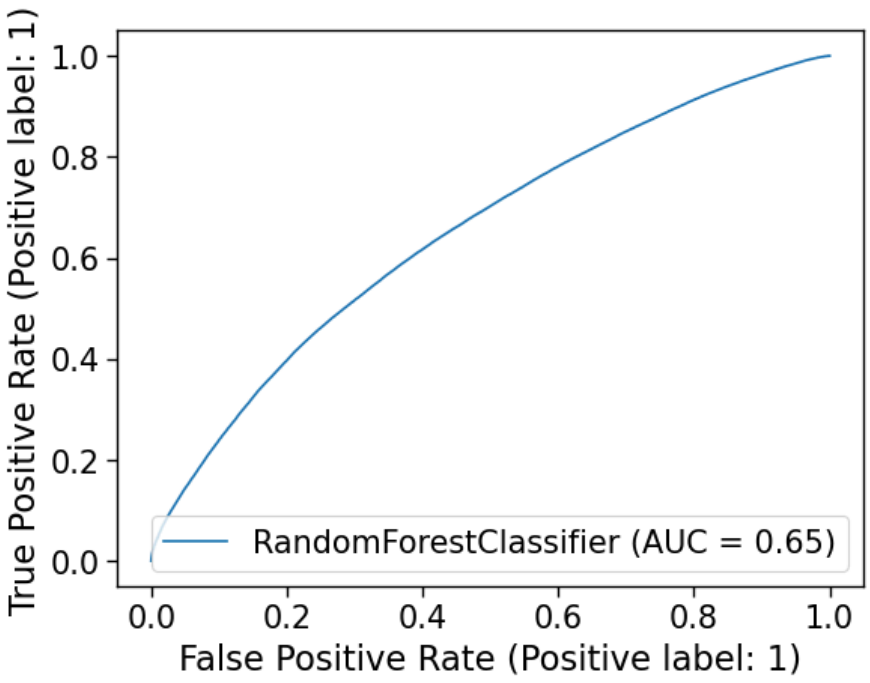
```
Classification Report - Random Forest Classifier:

              precision    recall  f1-score   support

           0       0.24      0.68      0.36    129831
           1       0.88      0.54      0.67    589451

    accuracy                           0.56    719282
   macro avg       0.56      0.61      0.51    719282
weighted avg       0.77      0.56      0.61    719282


 Confusion Matrix - Random Forest Classifier:
 [[ 88233  41598]
 [272450 317001]]
```



6. CONCLUSION AND FUTURE WORKS

In summary, the data shows that black and white hispanic males aged 25 to 44 commit the most crimes, with their crimes of choice usually being assault, drinking/drug offenses and theft. These crimes tend to occur outside and in apartments, commercial buildings and houses during the middle of the week from Wednesday to Friday, all throughout the year but mostly during spring (March and May), August and October. In addition, when crimes occur people take on average 27 days to report that crime when it occurs.

The results also show that when building a classification model to predict the sex of a perpetrator the most important features, in decreasing order are the type of offense, the location of the crime(precinct and jurisdiction) and the race of the perpetrator.

The final model chosen was a Random Forest Classifier which had a low AUC score of 0.65. This means that there is a lot of room for improving this model. Given more time, more meaningful features could be engineered and the model could be trained on a bigger dataset in order to increase its accuracy. After perfecting the model, the learnings could be applied to building a classification model to predict other things like PERP_RACE and TYPE_OF_OFFENSE which are both multi-class classification models.

7. APPENDIX - A | REFERENCES

[1] Police Department (NYPD), “NYPD Arrests Data (Historic),” *https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/about_data*, 2025

[2] Toronto Police Services, “Toronto Major Crime Indicators,” *https://open.toronto.ca/dataset/major-crime-indicators/*, 2025

[3] Banujan Kuhaneswaran, Chamith Sandagiri, Kumara B. T. G. S., Zhenni Li, “Twitter as a Lens for Crime Analysis: A Comprehensive 4W Model for Identifying Crime Patterns and Insights,” *ResearchGate*, 2023

[4] Steven Walczak, “Predicting Crime and Other Uses of Neural Networks in Police Decision Making,” *ResearchGate*, 2021.

[5] Thanu Dayara, Fadi Thabtah, Hussein Abdel-Jaber, Susan Zeidan: “Crime Analysis Using Data Analytics,” *ResearchGate*, 2022.

[6] Christian S Nwankwo, Majeed Kayode, Etinosa S Oghogho, “Application Of Data Analytics Techniques In Analyzing Crimes,” *ResearchGate*, 2018.

[7] Statistics Canada, "Unfounded criminal incidents Statistics Canada’s path to new data collection," *Social https://www150.statcan.gc.ca/n1/en/pub/11-627-m/11-627-m2018023-eng.pdf?st=dg3TASm5,* 2017

[8] Statistics Canada, "Municipalities Toronto Map”, *https://map-of-toronto.com/municipalities-maps/municipalities-toronto-map,* 2017

8. APPENDIX - B | LIST OF FIGURES

9. APPENDIX - C | LIST OF TABLES

10. APPENDIX - D | GITHUB LINK

A. Github Link

https://github.com/Vanaudel/TMU-Major-Research-Project