Assessment of Toronto Crime Data through Exploratory Data Analysis and Classification Methods

Katherine Ault
501092397
CIND 820: Big Data Analytics Project

Supervisor: Dr. C. Babaoglu

June 27th, 2022

1. INTRODUCTION

While the concept of crime forecasting can be traced back a century, it was through the adoption of geographic information systems (GIS) to map crime data during the 1990s which led researchers to recognize the potential for predictive analytics to forecast crime (Hvistendahl, 2016). The use of machine learning methods in the field of crime analysis to identify crime patterns and predict criminal activity is of considerable interest to law enforcement agencies with the results used to support evidence-based decisions in addition to informing choices regarding resource allocation, deployment, divisional staffing, and patrol plans (Lau, 2020).

Through exploration and analysis of the recently released Toronto Major Crime Indicators (MCI) dataset (Toronto Police Service, 2022), several research questions will be investigated through the application of various analytical tools and machine learning methods. Research questions include, but are not limited to:

- Can crime type(s) be predicted based on neighbourhood attributes (e.g., population density, unemployment rate, average income, average education level)?
- Which neighborhoods are the most violent and which the least violent?
- Which neighborhood has the highest incidence of crime and which neighbourhood the lowest?
- Applying the crime severity index weights (StatCan, 2021) to incidents, which neighbourhood has the highest overall crime weighting and which the lowest?
- What are the general crime trends within the City of Toronto?
- Are there recognizable temporal trends?
- Are specific crime types concentrated within certain geographical areas?

The data used for this project includes the Toronto MCI dataset noted above, in addition to Crime Severity Index weights for Canada (Statcan, 2021), geographic feature files (shapefiles) of Toronto Police patrol zones and Toronto neighbourhoods, and Toronto neighbourhood profiles (all from open.Toronto.ca, 2022).

The techniques employed will be data cleaning and exploratory data analysis of the Toronto MCI and Toronto neighbourhood datasets, merging of datasets, identification of relevant factors associated with crime, and application of k-NN, Naïve Bayes and logistic regression classification algorithms for crime prediction; all methods to be conducted using R.

The source code for this project can be found on GitHub:
https://github.com/kmault/CIND820_Capstone.

2. Literature Review

The literature search and review focused on studies where machine learning methods, preferably comparative studies, were applied to crime data to determine their effectiveness in crime pattern recognition and crime prediction. While several papers were found where clustering and classification methods were conducted through machine learning methods, they often simply described the application of the techniques(s) and failed to provide outcomes in terms of model performance or qualitative comparisons and subsequently were not included in the literature summaries.

The prediction of crime in San Francisco using k-NN, support vector machines, random forest and Naïve Bayes models was completed by Palanivinayagam et al., (2021). The authors generated 3 models for each algorithm with Naïve Bayes typically outperforming each method for every iteration with average accuracies of 94.8%, 93.8%, 93.7% and 93.7% for Naïve Bayes, random forest, k-NN, and SVM, respectively.

Wibowo and Oesman (2020) conducted a comparative analysis of the k-NN, Naïve Bayes, and decision tree algorithms to detect and predict crime within the Sleman Regency of Indonesia. The data focused on the crimes of theft, fraud, and embezzlement over a three-year period with the dataset containing a total of 1,735 incidents and 15 variables (i.e., day, time, season, victim gender, occupation, location etc.). Using accuracy to evaluate the comparative performance of the algorithms, Naïve Bayes achieved the highest level of accuracy (65.6%), followed by the k-NN model with accuracies ranging from 57.9% to 61.6%, and the decision tree model with an accuracy of 60.3%.

Zhang et al., (2020) compared several machine learning algorithms to predict property crime hotspots within a large coastal city in Southeast China using 2015 to 2018 public crime data. Through the application and comparison of 6 algorithms in predicting crime hotspots, the long-short term memory (LSTM) neural network model consistently outperformed those generated by k-NN, random forest, support vector machine, Naïve Bayes, and convolutional neural networks.

Alves et al., (2018) proposed that statistical learning methods were the best way to predict crime in relation to urban factors and applied a random forest algorithm to forecast crime and evaluate the importance of various urban indicators (e.g., literacy rates, employment etc.). Using the number of homicides as the dependent variable, the authors were able to obtain a 97% accuracy on the prediction of homicides in Brazilian cities. The authors also observed that factors such as employment rate and literacy influenced crime.

The use of k-means, decision tree, and tree-based algorithms to analyze and predict crimes were applied to the 2011 to 2012 UK crime data by Akila and Mohana (2017). The attributes of crime type, month, year, and location were assessed and predicted with the results showing that the k-means model provided higher accuracies compared to the decision tree and tree-based algorithms; 90.3% compared to 80.8% and 88.7%, respectively.

Classification of crime data from the Nigerian Prisons Services was conducted using decision tree (J48), Naïve Bayes, and ZeroR rule induction algorithms by Obuandike et al., (2015). Comparison of the 3 techniques revealed that the decision tree classifier returned the highest accuracy; 59.15% compared to 56.78% from Naïve Bayes and 56.78% from ZeroR.

Multivariate linear regression was used to predict crime counts using historical crime data from Chicago by Sengupta et al., (2014). The authors thoroughly described the process from data extraction to modeling and performance evaluation. The model was considered useful for predicting the probability of crime counts as the p-value was less than 0.05.

Shojaee et al., (2013) compared the performance of 5 machine learning algorithms in the classification of crime using the American Communities and Crime Unnormalized dataset. Utilizing crime status as the dependent variable, the authors determined that the k-NN algorithm returned the highest accuracy (87.5%) compared to Naïve Bayes (84.6%), neural network (85.3%), support vector machine (85.6%), and J48 decision tree (84.9%) models. The authors also noted that model performance was enhanced through feature selection. Iqbal et al., (2013), classified the dataset used by Shojaee et al., (2013) by implementing Naïve Bayes and decision tree algorithms using 10-fold cross validation to predict crime category. The performance measures of accuracy, precision and recall for the decision tree classifier consistently exceeded those for Naïve Bayes by at least 10% for each measure; 83.9%, 83.5% and 84% for decision tree compared to 70.8%, 66.4% and 70.8% for Naïve Bayes.

Using a dataset comprised of a combination of publicly available datasets from various American cities, Yu et al., (2011) assessed the performance of support vector machine (SVM), decision tree (J48), neural network and k-NN (k=1) classification models on crime forecasting. The authors attempted to predict residential burglaries using the attributes such as the number of arrests, commercial burglaries, foreclosure, street robberies etc. The model results indicated that the k-NN method consistently underperformed compared to, from highest to lowest, neural network, decision tree, and SVM algorithms.

Kim et al., (2018), using 15 years of Vancouver crime data, generated predictive models by KNN and boosted decision tree methods. Both models returned low accuracies, 39% for KNN and 44% for boosted decision tree and the authors noted that while the predictive accuracy was poor, they could be used as a framework for developing and executing future models.

Several studies were found that specifically investigated Toronto crime data (e.g., Oliveira, 2021., Stodulka, 2021., Sundar, 2020., Uwoghiren, 2020., Li, 2017., Vempala, 2016., Taneja et al., n.d.). Oliveira (2021) performed an analysis of 2019 Toronto major crime indicator data to identify neighbourhoods with the highest and lowest number of crimes as well as any spatial and temporal trend in crimes. Based on his analysis, the Church-Yonge, and Bay Steet Corridors were identified as the most dangerous neighbourhoods and that crimes were most likely to occur at noon and between the 11 pm and 3 am. Stodulka (2021) analysed Toronto major crime indicators between 2014 and 2018 to determine neighbourhoods with the most crime in addition to evaluating the different assault types. The results of his study showed that most crime in Toronto was comprised of assaults which represented over 50% of incidents in the dataset and that the Church-Yonge

Corridor was the most dangerous. Sundar (2020) applied decision tree (J48), k-NN, Naïve Bayes, and random forest classifiers to predict Toronto crime categories (i.e., assault, auto theft, B&E, robbery, and theft over $5000). Following feature selection and applying the SMOTE technique to rectify the imbalanced nature of the dataset, the author found that while none of the algorithms had particularly good precision and recall values for the theft over category, the random forest model outperformed the other methods, returning the highest precision and recall values for all crime categories, 52% compared to 40%, 49% and 43% for Naïve Bayes, KNN, and decision tree models, respectively. Review of incident data revealed that the Waterfront, Bay Street Corridor, and the Yonge-Church neighbourhoods were the most dangerous while Lambton Baby Point, Woodbine Lumsden, and Maple Leaf were the safest. Uwoghiren (2020) sought to group Toronto neighbourhoods via k-means clustering in order of desirability based on factors such as location of venues, crime rate, employment rate etc. The most desirable neighbourhoods were identified as Mount Pleasant West, Church-Yonge Corridor, Yonge-St Clair, and Bay Street Corridor with most neighbourhoods in the north-west region of the city (e.g., Etobicoke) considered least desirable. Using the 2008 to 2011 Toronto crime data, Vempala (2016) determined via linear regression and random forest models that the percentage of males, number of businesses and social assistance recipients were the most important crime predictors for 2011 although neither model was considered to perform particularly well, and the author noted that the features did not allow for the generation of a model that could make practical predictions. Li (2017) explored the 2016 Toronto MCI crime dataset, mapped crimes, and applied k-means clustering to group neighbourhoods based on crime categories. The results of her data exploration, mapping, and clustering indicated that the Waterfront was the most dangerous neighbourhood followed by the Church-Yonge Corridor (based on incident counts) with the safest neighbourhoods identified as Richmond Hill and Leaside Bennington. Cluster analysis indicated that 2 clusters were sufficient to group neighbourhoods: one cluster representing neighbourhoods with low crime (all categories) and a second cluster for those neighbourhoods with high crime (all categories). Taneja et al., (n.d.) applied, post PCA feature selection, the cluster techniques of k-means, agglomerative and DBSCAN to the 2000 to 2017 Toronto MCI dataset to identify violent versus non-violent neighbourhoods. Using internal validation measures such as silhouette, Dunn Index, C-H score, and D-B score to evaluate performance, agglomerative clustering provided the most suitable result followed by DBSCAN and k-means.

Following review of available literature pertinent to classification of crime data, the decision was taken to use the k-NN, Naïve Bayes and logistic regression algorithms to classify crime in Toronto. Several research questions were answered during review of previous studies involving crime data including:

- *Which Toronto neighborhoods are the most violent and which the least violent (generally assessed as high crime versus low crime)?* The most and least dangerous neighbourhoods were consistently identified as the Waterfront, Yonge-Church and Bays Street Corridors, and Lambton Baby Point.
- *What are the general crime trends within the City of Toronto?* In general, most crime in Toronto was comprised primarily of assaults which typically represented over 50% of incidents, that the Waterfront and Church-Yonge Corridors were consistently ranked among the most dangerous neighbourhoods and most crimes occurred outside, followed by apartments and commercial establishments.

- *Are there recognizable temporal trends?* Most incidents were noted to occur between May and October, on Friday and Saturday, at noon and between 11 pm and 3 am.
- *Are specific crime types concentrated within certain geographical areas?* While assaults were the main crime type for each neighbourhood, Waterfront and the Yonge-Church Corridor had the most break and enters while the most vehicle thefts occurred in West Humber-Clairville.

The current project comprises several similar elements to that of Sundar (2020); however, this study will use a more expansive and up to date dataset (2014 to 2021, compared to 2014 to 2019) and a few different predictive algorithms will also be executed. The outcome and performance measures will then be evaluated and compared.

This project will build upon the knowledge and framework produced during previous predictive studies of crime data, both for Toronto and other regions. Given that crime is, and will always be, an ongoing societal issue, studies that seek to provide any predictive capability would be of great value to law enforcement agencies.

3. Data & Data Description

3.1 Data & Data Cleaning

3.1.1 Toronto MCI Dataset

The primary dataset used for this project was the 2014 to 2021 Toronto Major Crime Indicator (MCI) data released by the Toronto Police Service and available through the Toronto Police Service Public Safety Data Portal (data.torontopolice.on.ca). The dataset contains a summary of founded incidents that fall within the categories of assault, break and enter, auto theft, robbery, and theft over $5000. Initial assessment of the MCI dataset showed that it contained 281,692 observations and 30 variables; variable data types were comprised of character (14), integer (12), and numeric (4). The dataset contained no NA values, 20,233 duplicated rows (based on event ID and offense), and 1372 entries for occurrences prior to 2014. Several redundant columns were identified during the initial evaluation based on the similarity of information: columns X, Y and Long, Lat both listed geographic coordinates, the dates of interest were for occurrences between 2014 and 2021 and not reports generated between those dates, and premises type and location type were quite similar with preference given to the premises type which contained a higher-level categorization of physical location.

Following preliminary review of the Toronto MCI dataset, initial feature reduction was performed by removing the following columns: X, Y, index, reported date, location type, reported year, reported month, reported day, reported day of year, reported day of week, reported hour and object ID. All rows with duplicated incident IDs and offence types, and those with an occurrence date prior to 2014 were removed. Two new columns were added: weight based on the Crime Severity Index weights for Canada dataset and a season column based on the occurrence month.

Additionally, several variable columns (e.g., day, month, crime time, premises type) were converted to factors to avoid potential complications during subsequent analysis.

The cleaned Toronto MCI dataset contained 260175 observations and 17 variables (Figure 1).

```
'data.frame':   260175 obs. of  17 variables:
 $ event_unique_id     : chr  "GO-20141625305" "GO-20141297201" "GO-20141302953" "GO-20141304312" ...
 $ Division            : Factor w/ 17 levels "D11","D12","D13",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ occurrencedate      : chr  "2014/03/02 05:00:00+00" "2014/01/03 05:00:00+00" "2014/01/08 05:00:00+00" "2014/01/08 05:00:00+00" ...
 $ premises_type       : Factor w/ 7 levels "Apartment","Commercial",..: 4 2 2 4 2 4 5 1 6 1 ...
 $ ucr_code            : int  1430 2120 2130 2120 2130 2120 1430 1430 1450 1430 ...
 $ weight              : num  26.4 211 144.4 211 144.4 ...
 $ offence             : Factor w/ 50 levels "Administering Noxious Thing",..: 6 13 43 13 43 13 6 6 21 6 ...
 $ occurrenceyear      : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ occurrencemonth     : Factor w/ 12 levels "April","August",..: 8 5 5 5 5 5 5 5 5 5 ...
 $ occurrencedayofweek: Factor w/ 7 levels "Friday    ","Monday    ",..: 4 1 7 7 2 1 4 4 4 2 ...
 $ occurrencehour      : int  8 10 2 11 0 18 10 21 22 9 ...
 $ MCI                 : Factor w/ 5 levels "Assault","Auto Theft",..: 1 3 5 3 5 3 1 1 1 1 ...
 $ Hood_ID             : chr  "1" "1" "1" "1" ...
 $ Neighbourhood       : Factor w/ 141 levels "Agincourt North",..: 126 126 126 126 126 126 126 126 126 126 ...
 $ Long                : num  -79.6 -79.6 -79.6 -79.6 -79.6 ...
 $ Lat                 : num  43.7 43.7 43.7 43.7 43.7 ...
 $ season              : Factor w/ 4 levels "Autumn","Spring",..: 2 4 4 4 4 4 4 4 4 4 ...
```

Figure 1. Structure of the cleaned Toronto MCI dataset.

3.1.2 Neighbourhood Profile Dataset

The second dataset reviewed was the Toronto Neighbourhood Profiles data released by the Division of Social Development, Finance & Administration and available from the City of Toronto Open Data Portal (https://open.toronto.ca/dataset/neighbourhood-profiles/). The neighbourhood profiles dataset contained 2,383 entries and 146 attribute columns; variable datatypes were integer (1) and character (145). The data frame was structured in such a way that individual neighbourhoods comprised the columns while each row contained information obtained primarily from 2016 census profile data; preliminary review revealed the dataset contained several redundant attribute columns. The dataset contained 7847 NA values and no duplicated rows. The dataset was transposed resulting in each row representing a unique neighbourhood and each column the various census-derived attributes.

Following cursory examination of the Neighbourhood Profiles dataset, initial feature reduction was carried out through the removal of rows such as: X_id, Category, Data.Source, and Characteristic as well as those with information pertaining to the 2011 population, population change 2011-2016, private dwellings occupied by usual residents, languages spoken, income statistics other than average income, mobility status, and anything that did not pertain to general neighbourhood details (e.g., detailed population breakdown rather than general population for the area).

The cleaned Neighbourhood dataset contained 141 observations and 24 variables (Figure 2).

```
'data.frame':   141 obs. of  24 variables:
 $ Neighbourhood           : chr  "City of Toronto" "Agincourt North" "Agincourt South-Malvern West" "Alderwood" ...
 $ Neighbourhood.Number    : int  NA 129 128 20 95 42 34 76 52 49 ...
 $ Population              : chr  "2,731,571" "29,113" "23,757" "12,054" ...
 $ TotalPrivateDwellings   : chr  "1,179,057" "9,371" "8,535" "4,732" ...
 $ PopulationDensity       : chr  "4,334" "3,929" "3,034" "2,435" ...
 $ MalePop                 : chr  "1,264,670" "13,200" "11,150" "5,680" ...
 $ FemalePop               : chr  "1,417,995" "15,200" "12,145" "6,140" ...
 $ AvgHouseholdSize        : num  2.42 3.16 2.88 2.6 1.8 2.23 2.56 1.7 2.22 2.7 ...
 $ EnglishOnly             : chr  "2,323,235" "21,645" "19,225" "10,840" ...
 $ NeitherOfficialLanguage: chr  "132,765" "5,945" "3,430" "295" ...
 $ Avg_AfterTaxIncome      : chr  "81,495" "427,037" "278,390" "168,602" ...
 $ Citizen                 : chr  "2,296,365" "23,550" "19,015" "11,425" ...
 $ NotCitizen              : chr  "395,300" "5,270" "4,465" "600" ...
 $ VisibleMinority         : chr  "1,385,855" "26,365" "20,155" "2,490" ...
 $ NotVisibleMinority      : chr  "1,305,815" "2,465" "3,320" "9,535" ...
 $ NoEducation             : int  377340 6550 4035 2005 1585 2295 1665 700 1310 1295 ...
 $ Education_HighSchool    : int  561090 7460 6090 2960 4270 5150 3390 5740 3680 2385 ...
 $ HigherEducation         : int  1356360 11005 10275 5305 20435 15940 8200 17495 13760 7480 ...
 $ EmploymentRate_Total    : num  59.3 50 53.2 62.4 65.8 55.6 60.3 56.2 58.5 51.3 ...
 $ UnemploymentRate_Total : num  8.2 9.8 9.8 6.1 6.7 7.2 7.2 10.2 7.7 8 ...
 $ EmploymentRate_Male     : num  63.8 54.5 58.2 67.1 68.9 62 64.6 61.2 63.7 56.6 ...
 $ UnemploymentRate_Male   : num  8 9 8.8 5.4 6.9 6.4 6.5 9.5 6.8 6.3 ...
 $ EmploymentRate_Female   : num  55.2 46 48.7 57.9 63.1 50.2 56.4 51.7 54 46.8 ...
 $ UnemploymentRate_Female: num  8.5 10.6 10.8 6.8 6.7 8.2 7.7 11.1 8.6 9.7 ...
```

Figure 2. Structure of the cleaned Neighbourhood dataset.

3.1.3 CSI Weights Dataset

The Crime Severity Index weights for Canada dataset, provided via email request to StatCan, contained 284 entries and 3 attributes. The attributes were comprised of the Uniform Crime Reporting (UCR) violation codes (e.g., 1110), description of violation (e.g., Murder 1st Degree) and violation weighting (e.g., UCR 1110 has an associated weight of 7656.16). The dataset contained no duplicates or NA values. No changes were made to the file.

3.1.4 Shape Files

3.1.4.1 TPS Patrol Zones.

The shapefile for the Toronto Police Patrol Zones was provided by the Toronto Police Service and available from the City of Toronto Open Data Portal (https://open.toronto.ca/dataset/patrol-zones/). The file contained information such as the number of polygon features representing each patrol zone (n=17), divisional address, geographic coordinate system, zone, and datum (i.e., UTM zone 17N/NAD27), and extent of each patrol zone. No changes were made to the patrol zones file. Plots showing the extent of police patrol zones are shown on Figure 3.

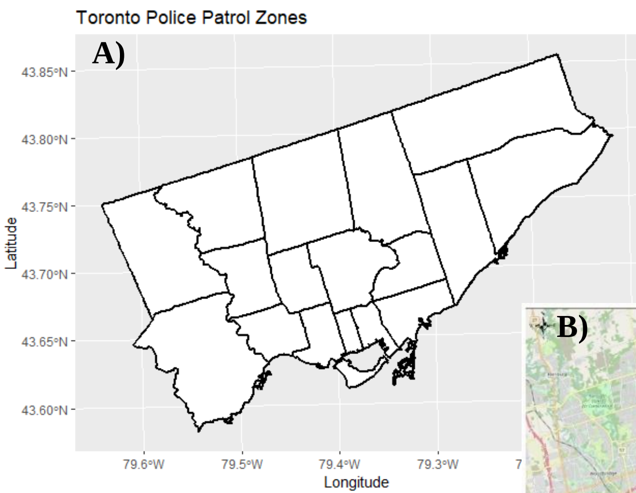
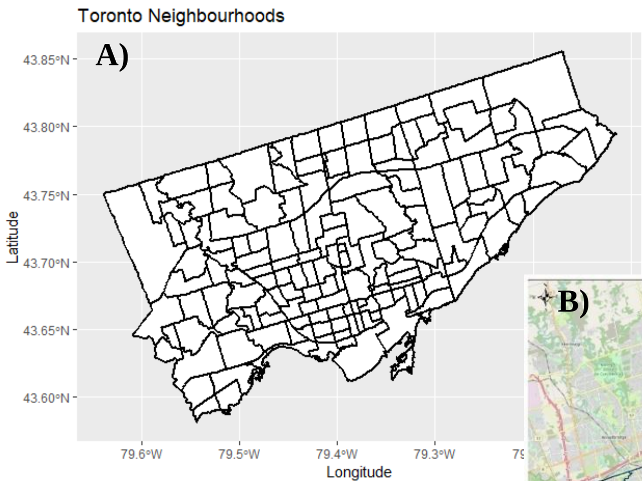


Figure 3. General plot of TPS patrol zones (A) and a more detailed figure with OpenStreetMap as the base map (B).

3.1.4.2 Neighbourhoods.

The Neighbourhoods shapefile was provided by the Division of Social Development, Finance & Administration and available from the City of Toronto Open Data Portal (https://open.toronto.ca/dataset/neighbourhoods/). The file contained information such as the number of polygon features representing each neighbourhood, the neighbourhood designation (e.g., neighbourhood improvement area - NIA), geographic bounds as well as geographic coordinate system, zone, and datum (i.e., Lat/Long, WGS84), and neighbourhood name and number (e.g., West Humber-Clairville: 1). This file was not changed, and a simple plot of Toronto neighbourhoods is presented in Figure 4.



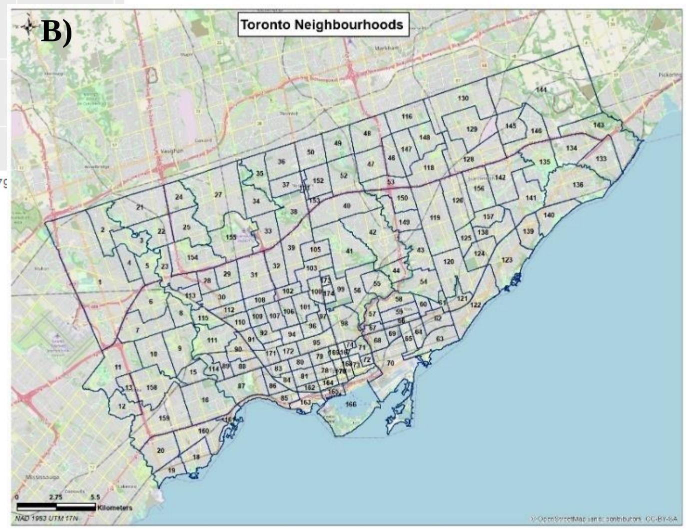


Figure 4. General plot of Toronto neighbourhoods (A) and a more detailed figure with OpenStreetMap as the base map (B).

4. Exploratory Data Analysis & Descriptive Statistics

The Toronto MCI dataset is the primary dataset under review, as such it is the only data that underwent EDA. The number of founded criminal incidents in Toronto exhibited a gradual increase from 2014 to 2019 after which time incident numbers started to decrease (Figure 5); annual crime counts ranged from 30197 to 37024 and a median value of 31849 (Figure 6). The monthly trend in crime counts showed minor variation, with higher values observed between May and October (Figure 7); the number of incidents ranged from 18736 to 23248 with a median value of 22180 (Figure 8). The number of incidents per day was consistent (Figure 9) with the highest counts noted over the weekend (Friday to Sunday); the number of daily incidents ranged from 35569 to 39552 and exhibited a median value of 36734 (Figure 10) MCT. There was minor variation in the number of crimes committed per season (Figure 11) with a gradual increase from winter to summer; seasonal numbers ranged from 60783 (Winter) to 68512 (Summer). The number of incidents per premises type ranged from 6607 for educational to 67550 for outside (Table 1, Figure 12).

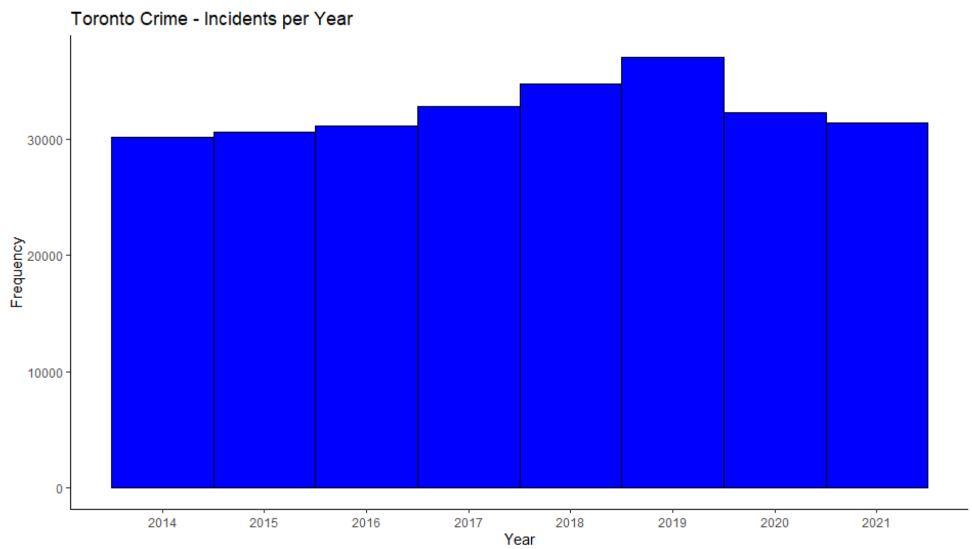


Figure 5. Trend in Toronto crime counts per year.

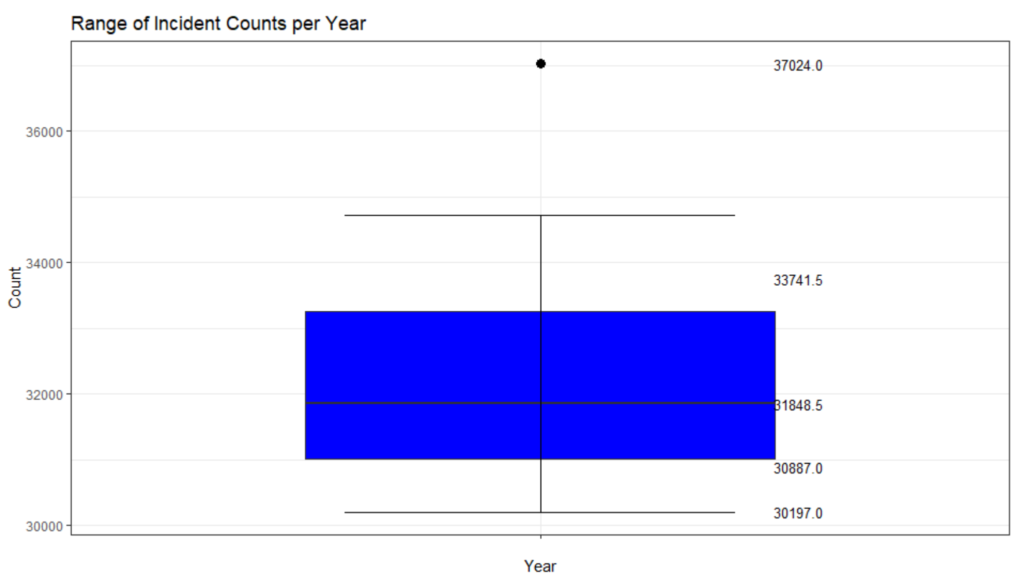


Figure 6. Range in Toronto crime counts per year.

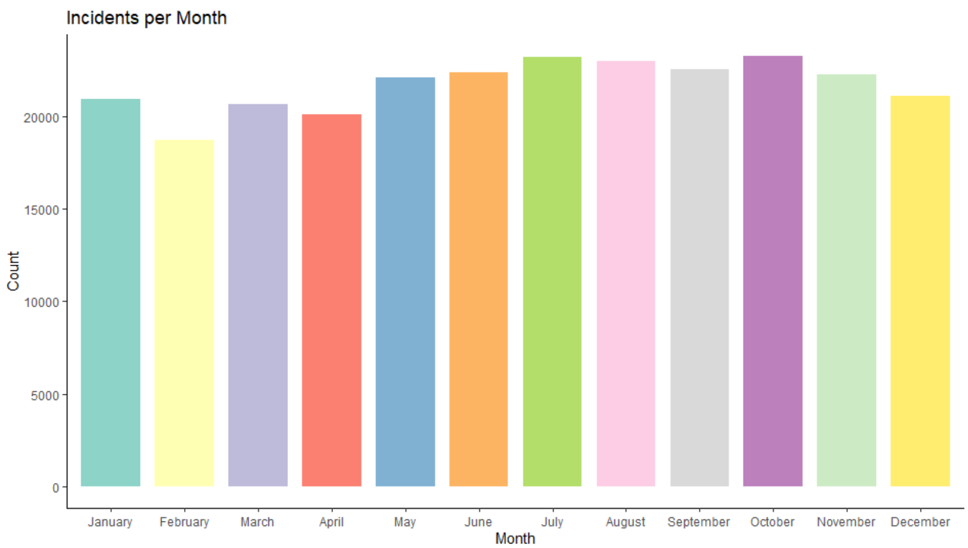


Figure 7. Trend in Toronto crime counts per month.

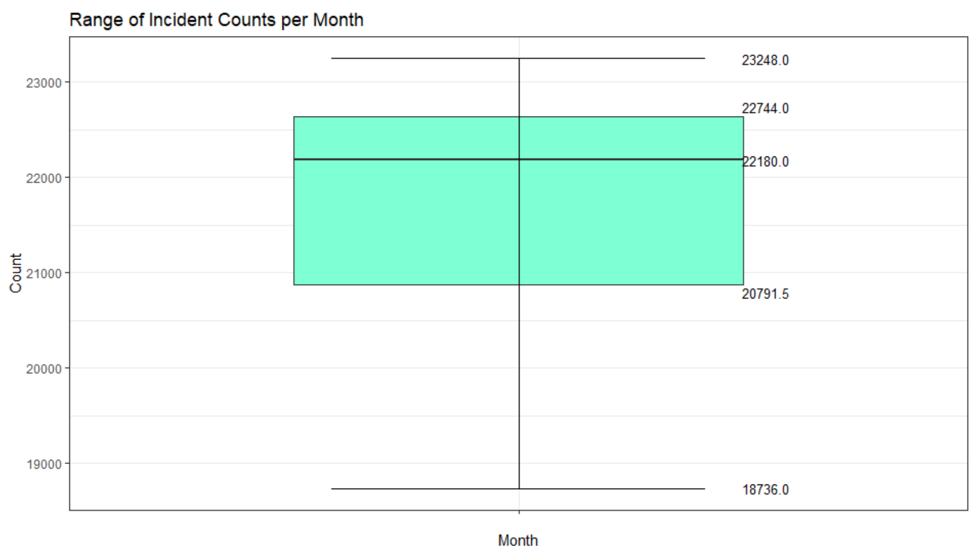


Figure 8. Range in Toronto crime counts per month.

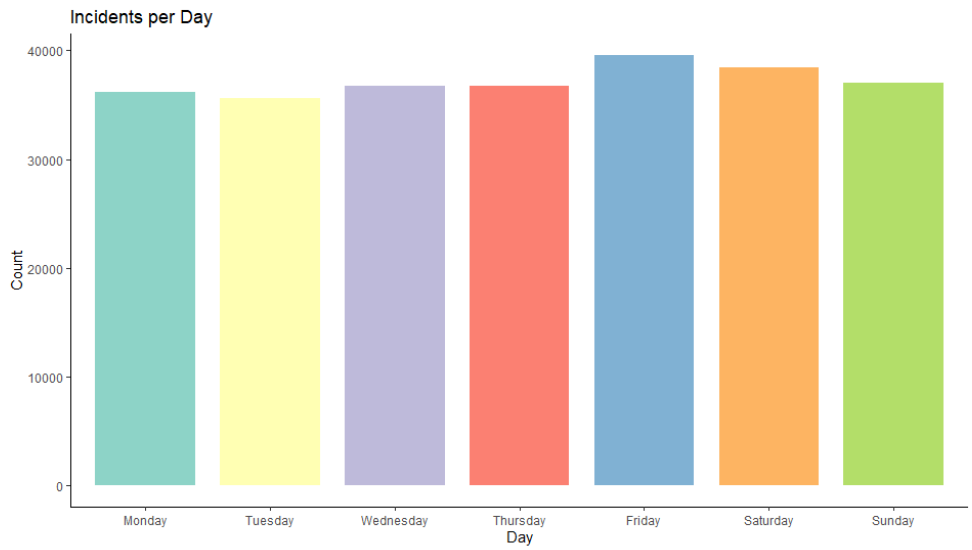


Figure 9. Trend in Toronto crime counts per day of week.

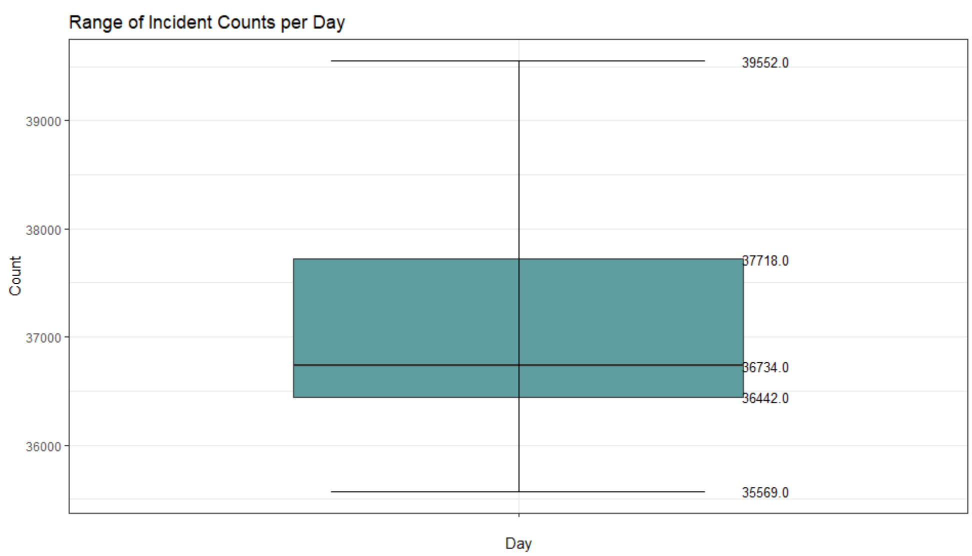


Figure 10. Range in Toronto crime counts per day of week.

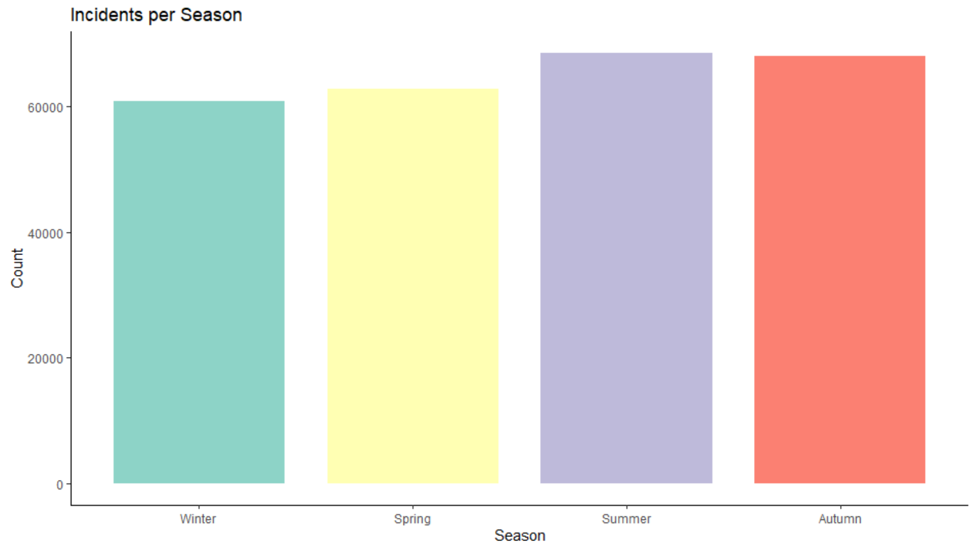


Figure 11. Trend in Toronto crime counts per season.

Premises	Incident Count
Apartment	63513
Commercial	52657
Educational	6607
House	47694
Other	14920
Outside	67550
Transit	7234

Table 1. Incident count per premises type.

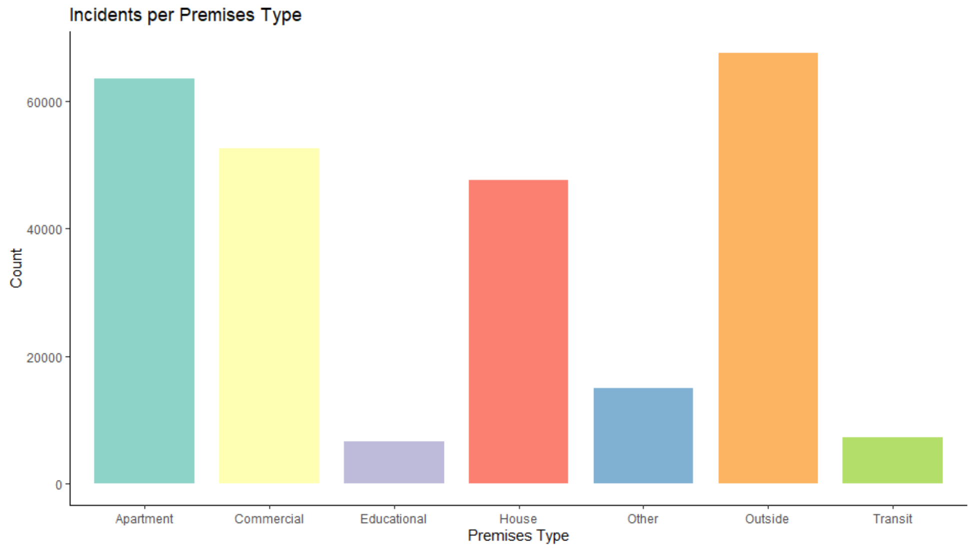


Figure 12. Trend in Toronto crime counts per premises type.

The was a progressive decrease in the number of incidents per hour from midnight to 7 am, after which incident counts increased with a significant peak between the hours of noon and 1 pm (Figure 13). A day-time analysis of crime indicated that most incidents occurred between midnight and 1 am on Saturday and Sunday, and the lowest crime numbers were noted for every weekday between the hours of 3 am and 11 am (Figure 14).

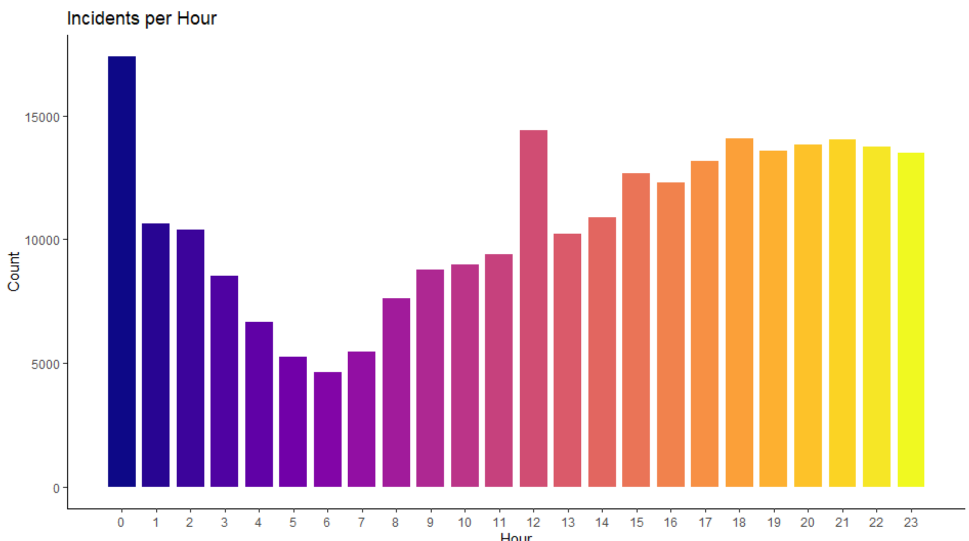


Figure 13. Trend in Toronto crime counts per hour.

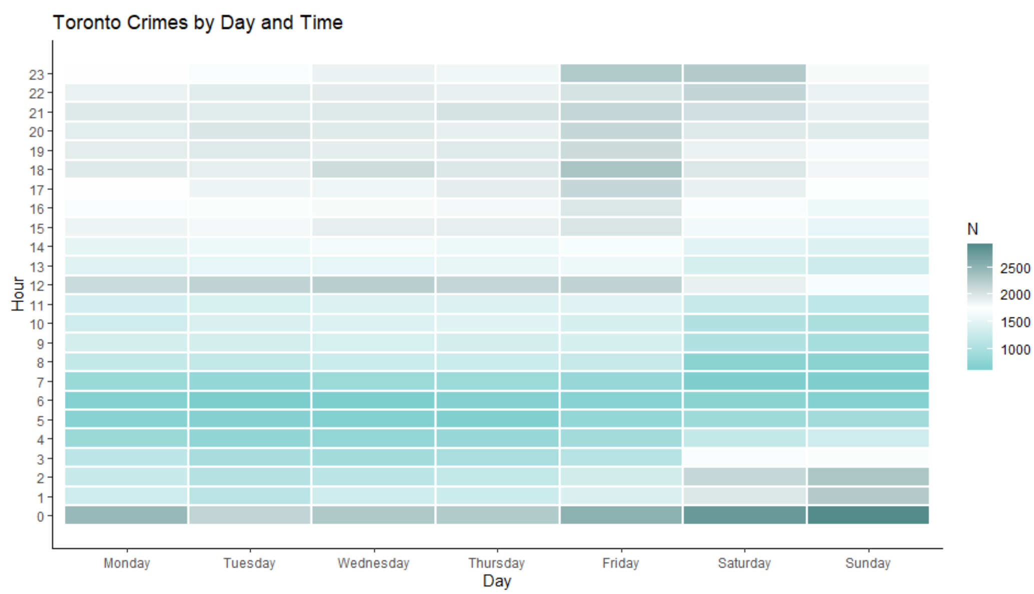


Figure 14. Day/Time trends in Toronto crime counts.

Incidents per neighbourhood ranged from 388 for Lambton Baby Point to 9622 for the Waterfront Communities, with a median value of 1371.5 (Figure 15). Outlier values were observed for the neighbourhoods of Waterfront Communities (9622), Church-Yonge Corridor (8606), Bay Street Corridor (7300), West Humber-Clairville (6874), Moss Par (6479), York University Heights (5056), Kensington-Chinatown (4960), Downsview-Roding-CFB (4931), Woburn (4547), and West Hill (4143). The outlier values were retained as neighbourhood-specific analysis related to crime counts were conducted; it is anticipated that rebalancing of the dataset will minimize the effect of the outliers. The top 10 neighbourhoods with the highest and lowest crime counts are shown on Figures 16 and 17.

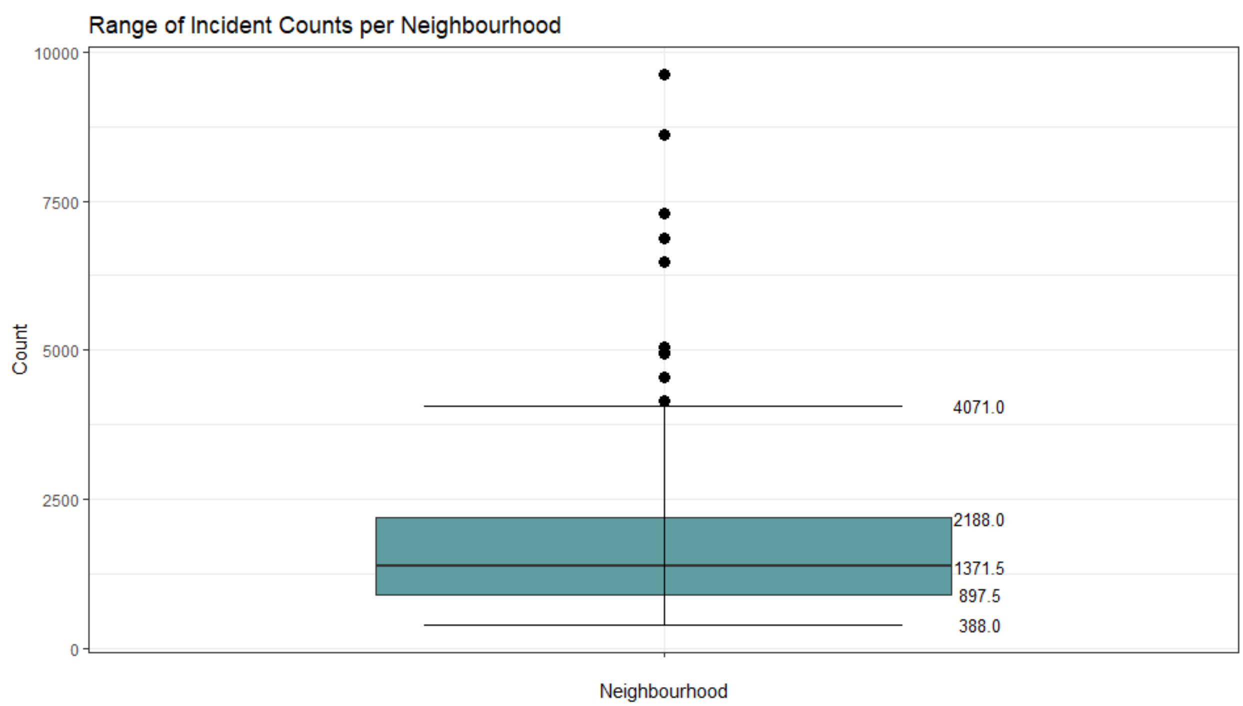


Figure 15: Box plot of incident counts per neighbourhood.

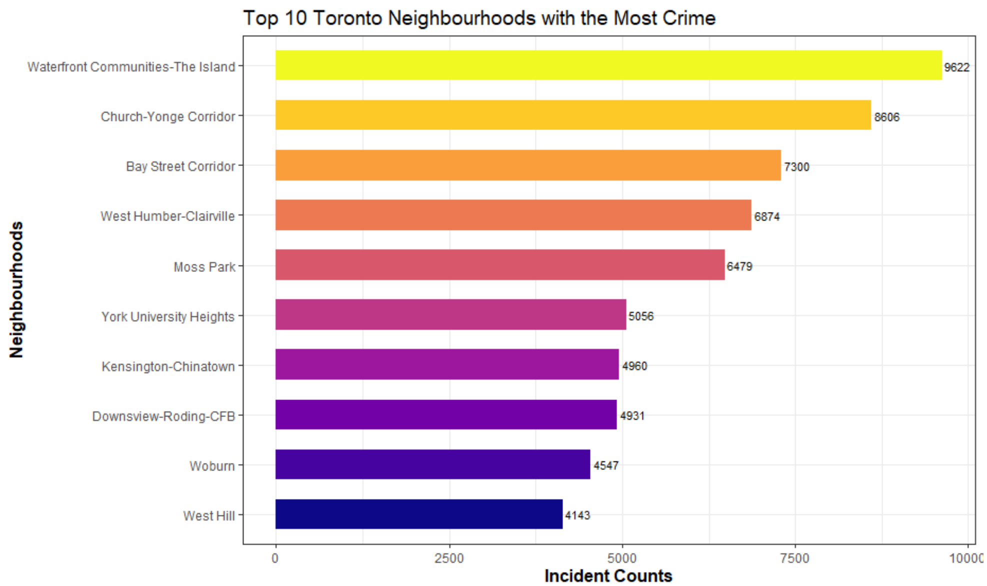


Figure 16: Toronto neighbourhoods with the highest crime counts – Top 10.

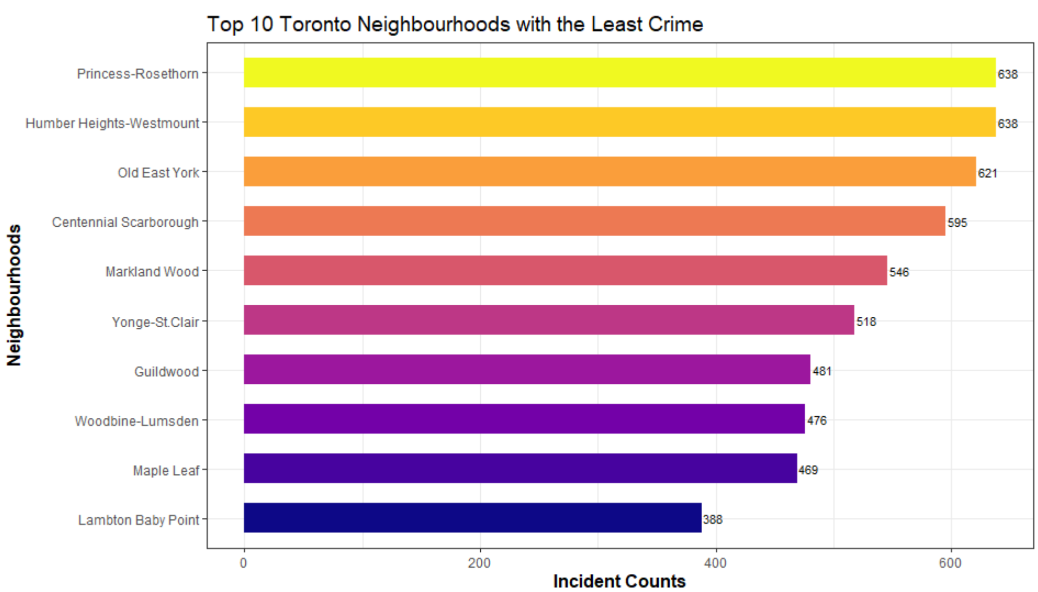


Figure 17: Toronto neighbourhoods with the lowest crime counts – Top 10.

Following the broad review of the crime dataset, the MCI data was examined in more detail to observe whether there were variations in the crime types in terms of time and location. There is a notable imbalance in the dataset with assaults clearly outnumbering the other MCI types (Figure 18) indicating that the data will be balanced prior to classification. A more detailed view of the offense types also shows assault as the primary offence type followed by break and enter, theft of motor vehicle and assault with a weapon (Figure 19). Assault was also the top MCI type per neighbourhood (Figure 20). The type of MCI shows variation with respect to premises type with assault occurring primarily at apartments and outside, robberies were outside, break and enters were committed at apartments, commercial properties, and houses, and theft over $5000 effectively the same across all types (Figure 21).

Assaults increased from 2014 to 2019 after which the number of incidents decreased. Break and enters exhibited minor variation between 2014 and 2018 with incidents increasing between 2019 and 2021. Similarly, auto theft increased between 2014 and 2017 and increased between 2018 and 2021. Theft over $5000 and robbery showed little annual variation (Figure 22). Apart from assaults, which occurred primarily between May and October, MCI categories showed little monthly variation (Figure 23). The was little variation in the MCI type per day; break and enters exhibited a subtle increase on Friday and assaults a slight increase on Saturday and Sunday (Figure 24). Assaults peaked between the hours of midnight and 1 am, and noon and 1 pm, with a distinct low between 3 am and 8 am. Break and enters showed a similar trend to assaults with maximums between midnight and 1 am and noon and 1 pm. Robbery and theft over $5000 exhibited little hourly variation while auto theft progressively increased from 7 pm to midnight (Figure 26).

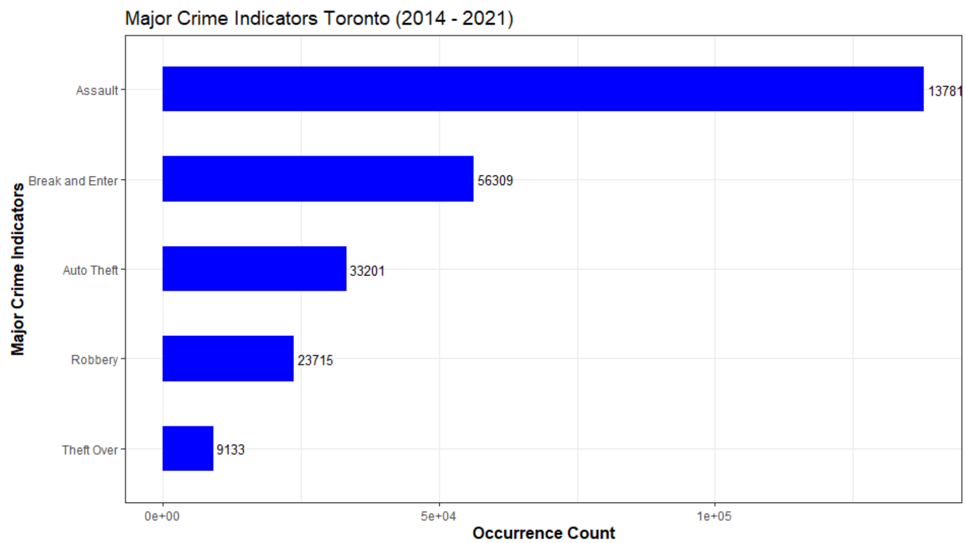


Figure 18: Trend in MCI categories.

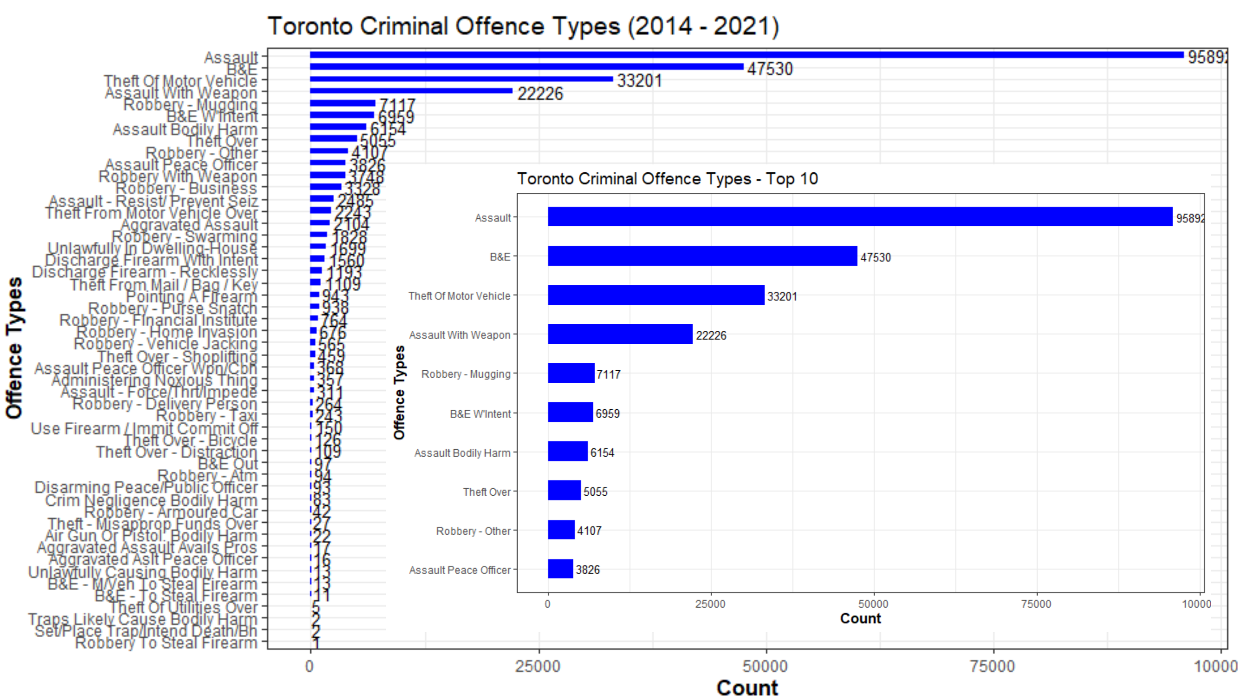


Figure 19: Toronto crime broken down by offence type with the top 10 shown in the inset plot.

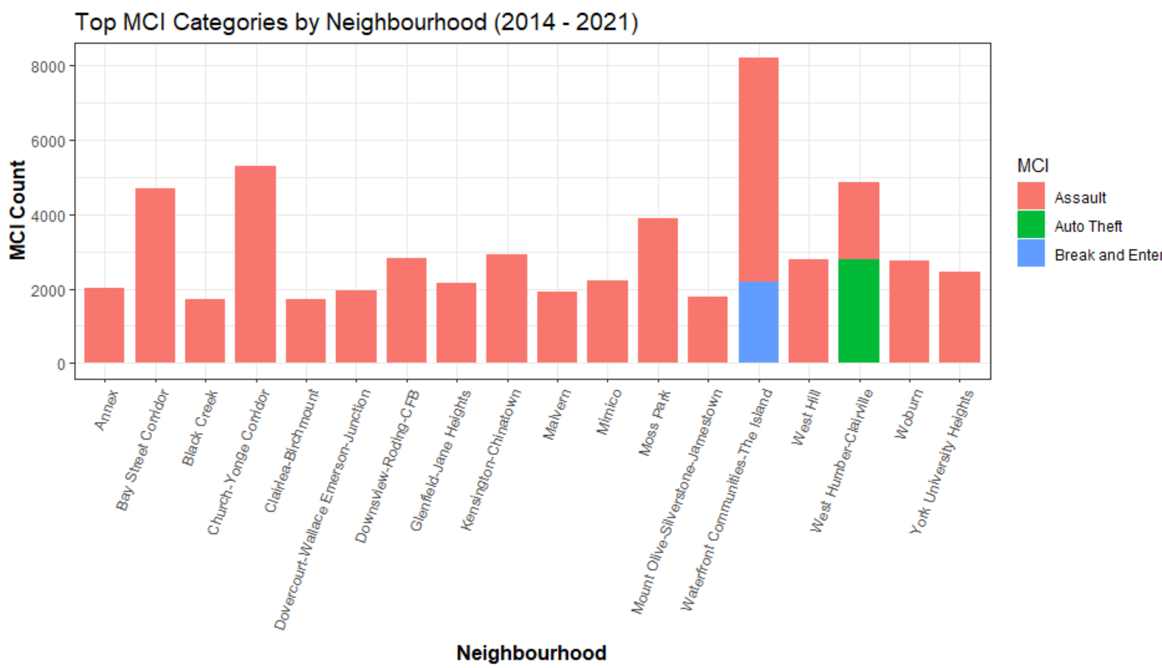


Figure 20: Primary MCI type per neighbourhood (first 20 neighbourhoods shown only).

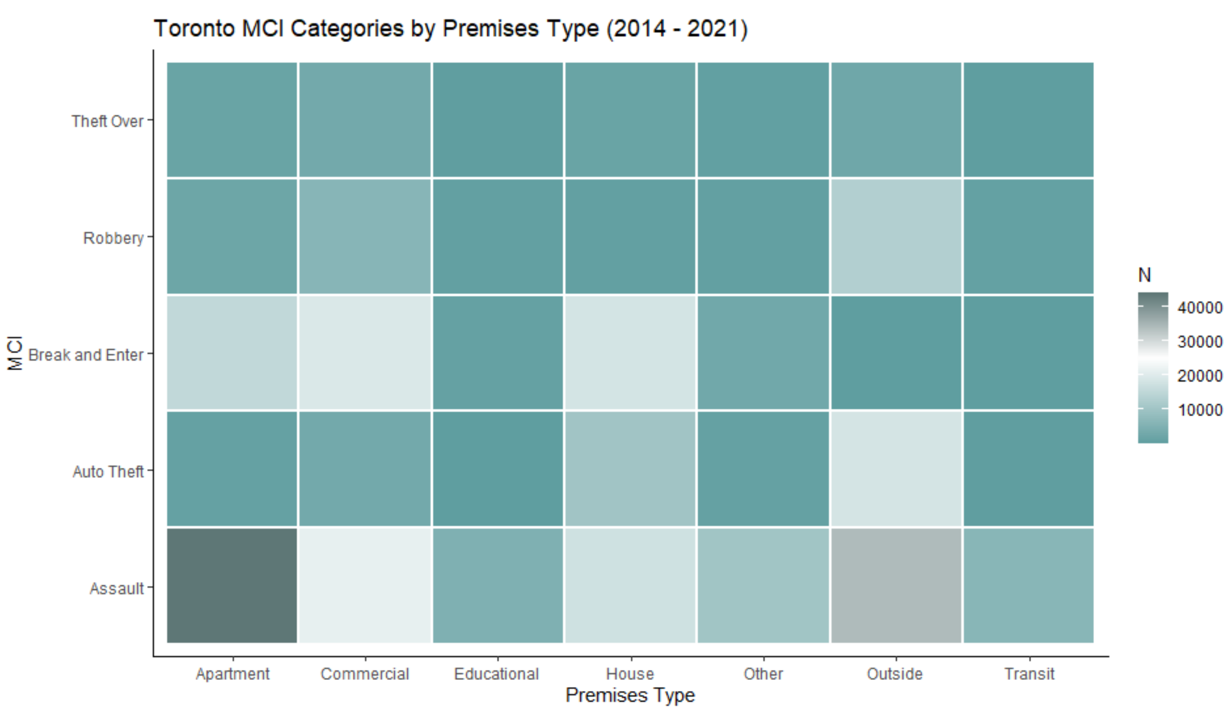


Figure 21: Variation in MCI category by premises type.

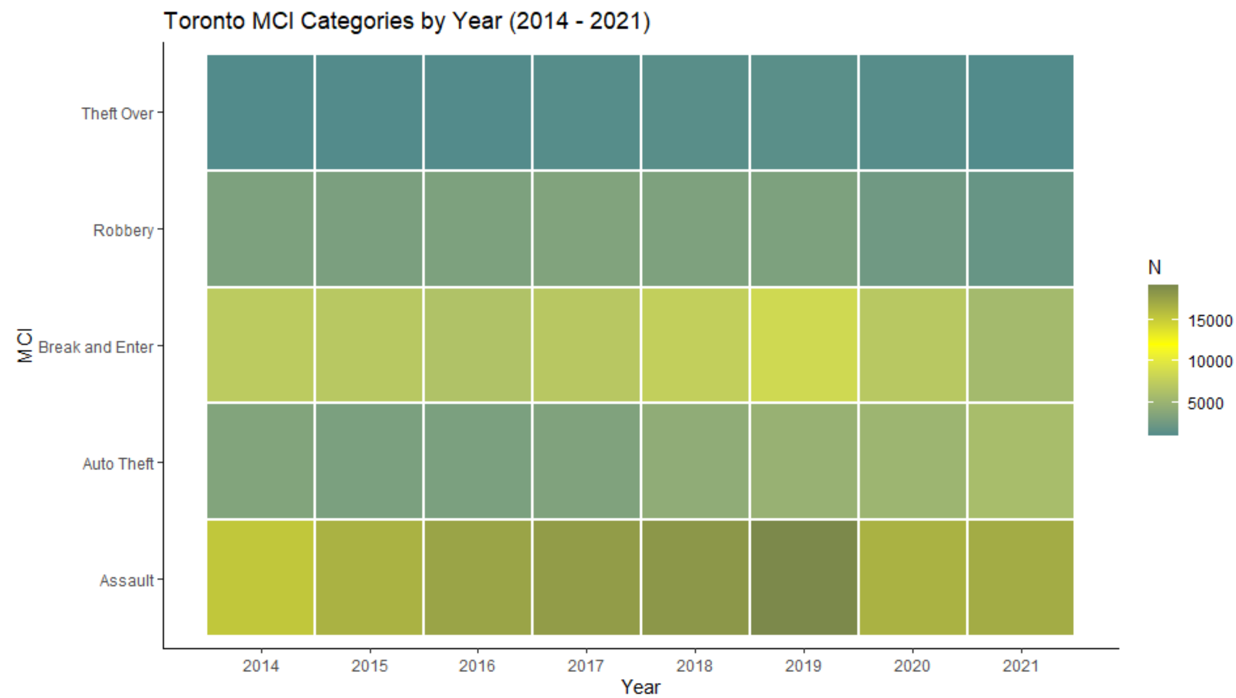


Figure 22: Variation in MCI categories per year.

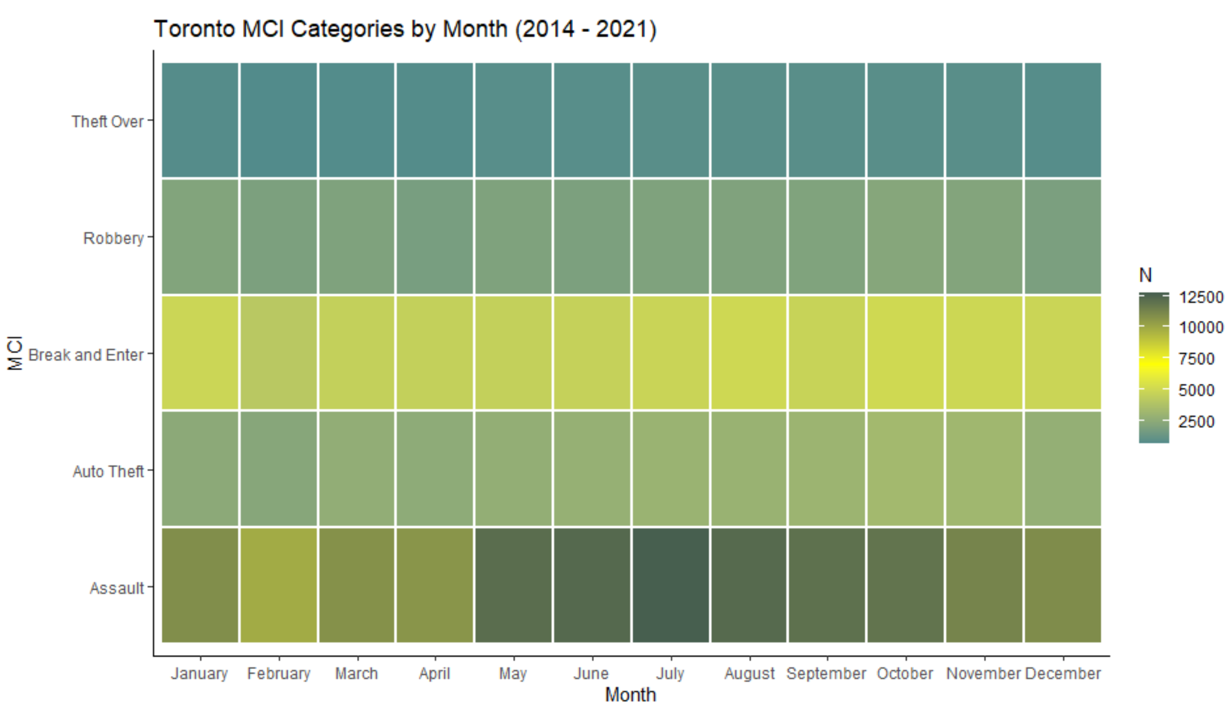


Figure 23: Variation in MCI categories per month.

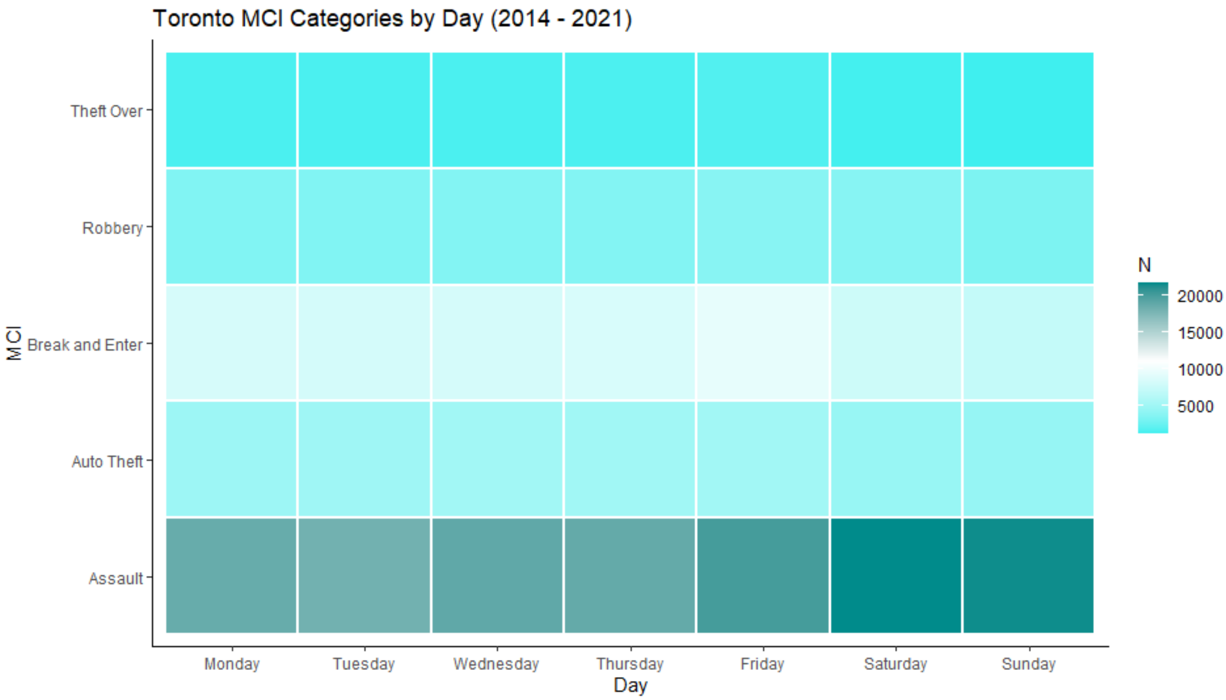


Figure 24: Variation in MCI categories per day.

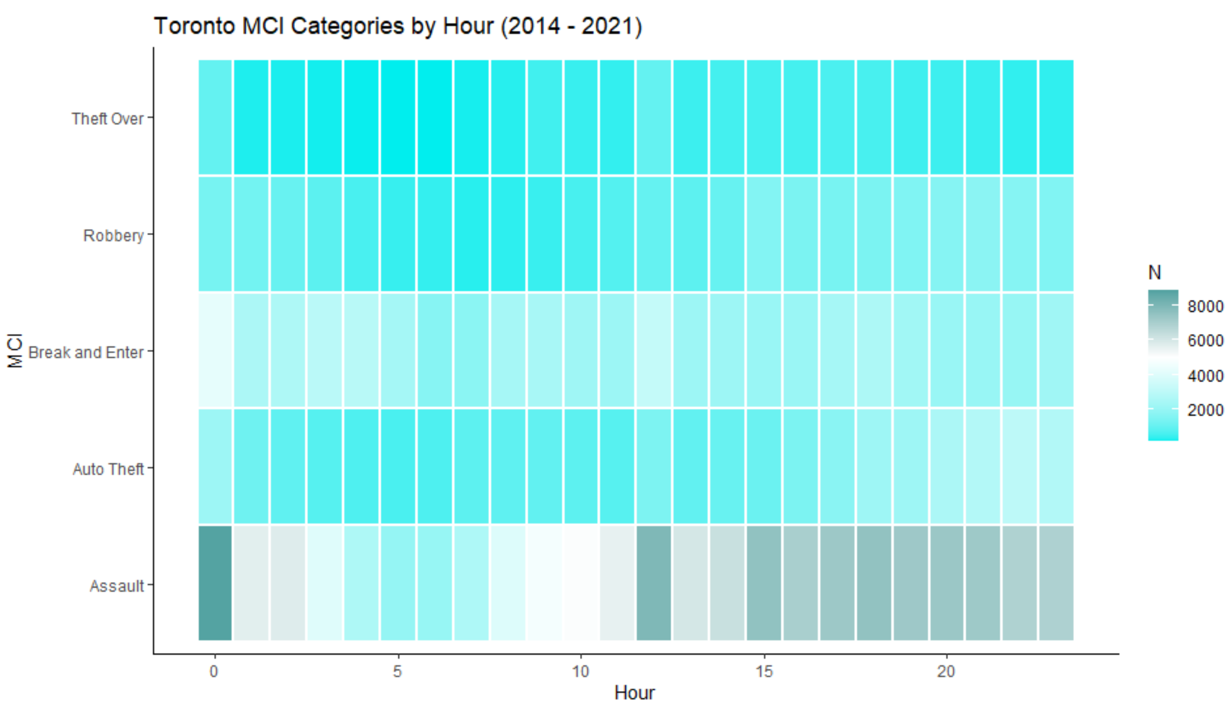


Figure 25: Variation in MCI categories per hour.

5. Approach

The project methodology is anticipated to follow the steps shown in Figure 26. Double arrows denote those project stages that could have overlap.

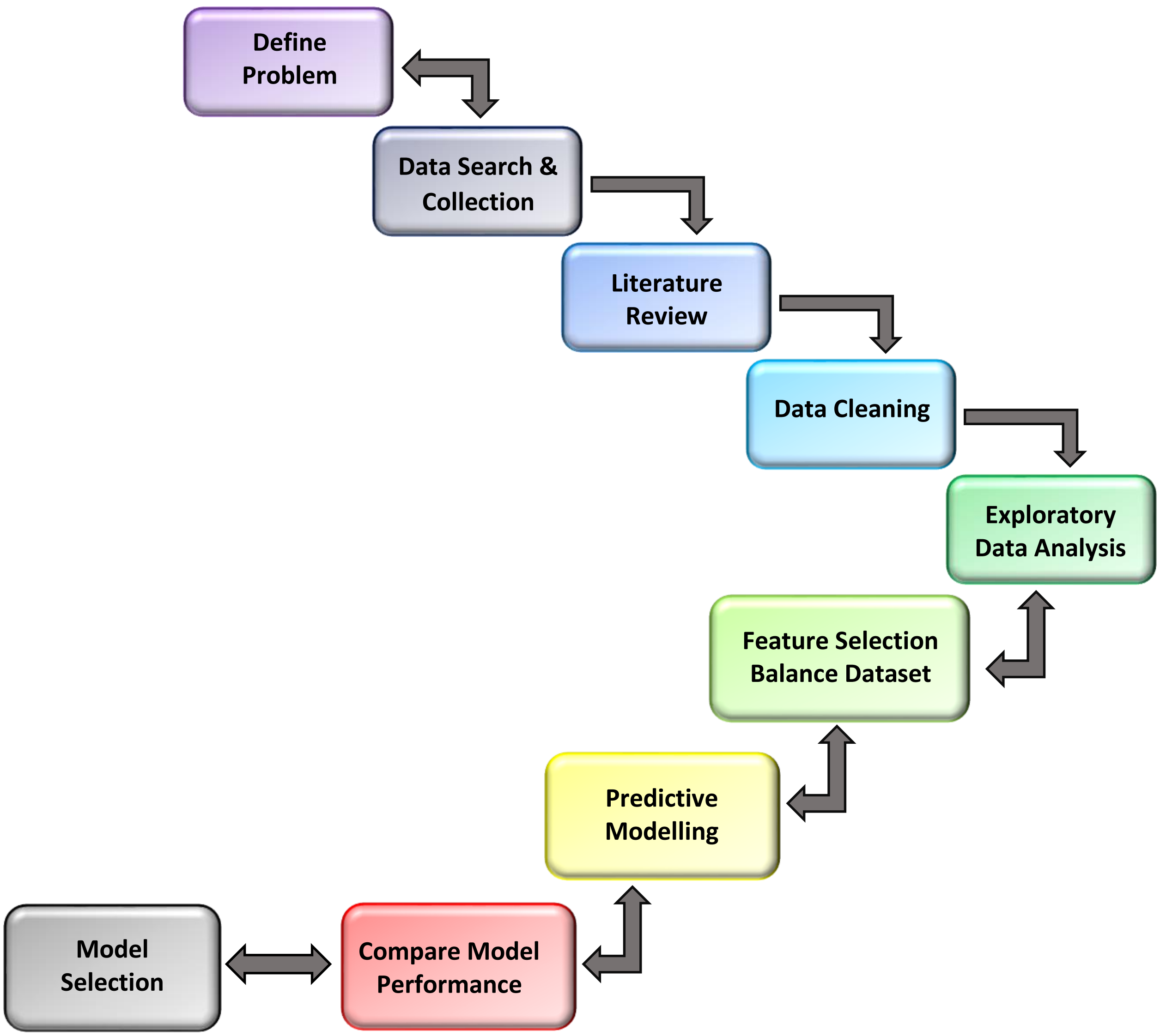


Figure 26: Tentative project methodology.

6. Initial Results & Code

Following preliminary assessment, exploratory data analysis and cleaning of the MCI dataset, the data was further reviewed, and manual feature reduction was performed. The additional variables removed were either redundant or clearly not predictive and their presence would only serve to increase processing time. These variables included:

- Event ID (not a predictive variable)
- Weight (weight column was used with UCR to compile Neighbourhoods dataset and not useful an MCI predictor)
- UCR (UCR and offense type used to categories MCI and detailed review of UCR not being conducted)
- Offense (offense type was used to categorize MCI and detailed review of offenses not being conducted)
- Hood ID (removed as there was a Neighbourhood variable)
- Occurrence date (removed as the dataset contains year, month, and day columns)

Following the removal of redundant or non-predictive variables, the final MCI dataset contained 255951 records and 11 variables (Figure 27). The full cleaned dataset can be viewed via the following link to the GitHub repository:

(https://github.com/kmault/CIND820_Capstone/blob/main/MCI_WorkingDataset.csv).

```
'data.frame':   255951 obs. of  11 variables:
 $ Division           : Factor w/ 16 levels "D11","D12","D13",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ premises_type      : Factor w/ 7 levels "Apartment","Commercial",..: 4 2 2 4 2 4 5 1 6 1 ...
 $ occurrenceyear     : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ occurrencemonth    : Factor w/ 12 levels "January","February",..: 3 1 1 1 1 1 1 1 1 1 ...
 $ occurrencedayofweek: Factor w/ 7 levels "Monday","Tuesday",..: 7 5 3 3 1 5 7 7 7 1 ...
 $ occurrencehour     : int  8 10 2 11 0 18 10 21 22 9 ...
 $ Neighbourhood      : Factor w/ 140 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Long               : num  -79.6 -79.6 -79.6 -79.6 -79.6 ...
 $ Lat                : num  43.7 43.7 43.7 43.7 43.7 ...
 $ season             : Factor w/ 4 levels "Winter","Spring",..: 2 1 1 1 1 1 1 1 1 1 ...
 $ MCI                : Factor w/ 5 levels "Assault","Auto Theft",..: 1 3 5 3 5 3 1 1 1 1 ...
```

Figure 27: MCI dataset after manual reduction of variables deemed redundant or non-predictive.

6.1 Feature Selection

Prior to moving forward to the feature selection step, the dataset was assessed for correlation among variables. Certain correlations among predictors were expected, such as season to month, division to neighbourhood and geographic coordinates (Figure 28, 29). The dataset remained unchanged, and methods of feature selection were considered to provide the most effective subset of predictive variables.

	Division	premises_type	occurrenceyear	occurrencemonth	occurrencedayofweek	occurrencehour	Neighbourhood
Division	1.000000000	-0.0098227121	0.006695748	-0.0043328302	-2.925497e-03	-0.008237769	0.244741559
premises_type	-0.009822712	1.0000000000	0.004255985	0.0281614464	-6.957376e-03	0.099525626	-0.017463598
occurrenceyear	0.006695748	0.0042559853	1.000000000	-0.0050735401	-6.236548e-03	0.009280711	-0.014817682
occurrencemonth	-0.004332830	0.0281614464	-0.005073540	1.0000000000	1.689371e-03	0.005496235	-0.008642553
occurrencedayofweek	-0.002925497	-0.0069573758	-0.006236548	0.0016893707	1.000000e+00	-0.044392635	0.005954492
occurrencehour	-0.008237769	0.0995256259	0.009280711	0.0054962352	-4.439264e-02	1.000000000	-0.020324976
Neighbourhood	0.244741559	-0.0174635976	-0.014817682	-0.0086425527	5.954492e-03	-0.020324976	1.000000000
Long	0.027713171	-0.0002520007	0.015260252	-0.0006506979	5.935352e-05	0.001207821	0.035934222
Lat	0.001999267	-0.0000724674	-0.017296412	0.0002155488	-1.025999e-03	0.000286215	0.003810745
season	-0.003711598	0.0459329289	-0.002730476	0.5832411176	1.333082e-03	0.008749852	-0.004482305

	Long	Lat	season
Division	2.771317e-02	0.0019992666	-0.0037115980
premises_type	-2.520007e-04	-0.0000724674	0.0459329289
occurrenceyear	1.526025e-02	-0.0172964118	-0.0027304762
occurrencemonth	-6.506979e-04	0.0002155488	0.5832411176
occurrencedayofweek	5.935352e-05	-0.0010259995	0.0013330815
occurrencehour	1.207821e-03	0.0002862150	0.0087498517
Neighbourhood	3.593422e-02	0.0038107446	-0.0044823049
Long	1.000000e+00	-0.9978234771	-0.0011451850
Lat	-9.978235e-01	1.0000000000	0.0006131715
season	-1.145185e-03	0.0006131715	1.0000000000

Figure 28: Correlation among the MCI dataset variables.

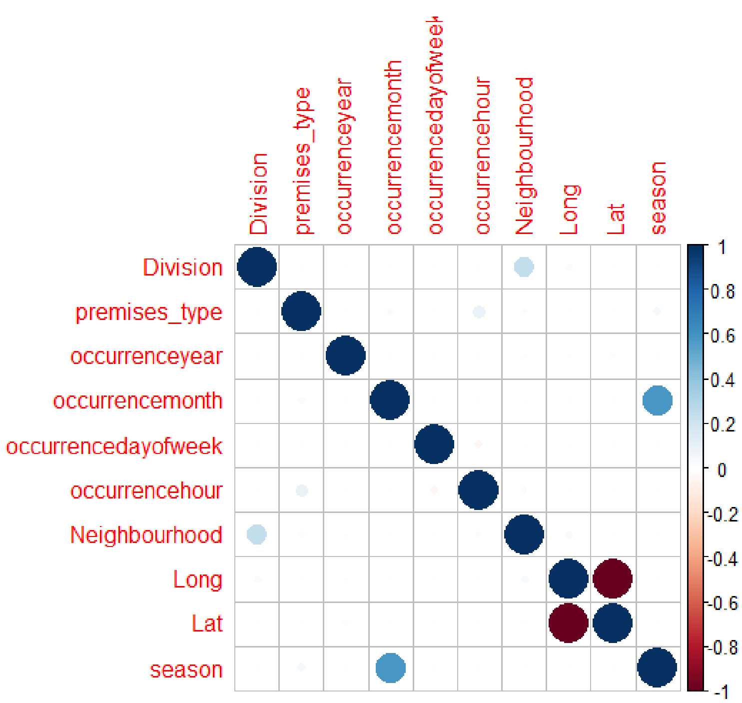


Figure 29: Correlation matrix for variables within the MCI dataset.

6.1.1 Boruta Method

Following review of machine learning feature selection methods, the decision was taken to apply the random forests based Boruta algorithm. Despite being resource-heavy, with a 3-hour run time, the output provided a succinct ranking of variables by importance. Although all variables were considered to have some significance, premises type, occurrence hour, latitude, and longitude demonstrated the highest impact as predictor variables (Figure 30). Boruta feature selection was performed using the Boruta, mlbench and randomforest libraries in R.

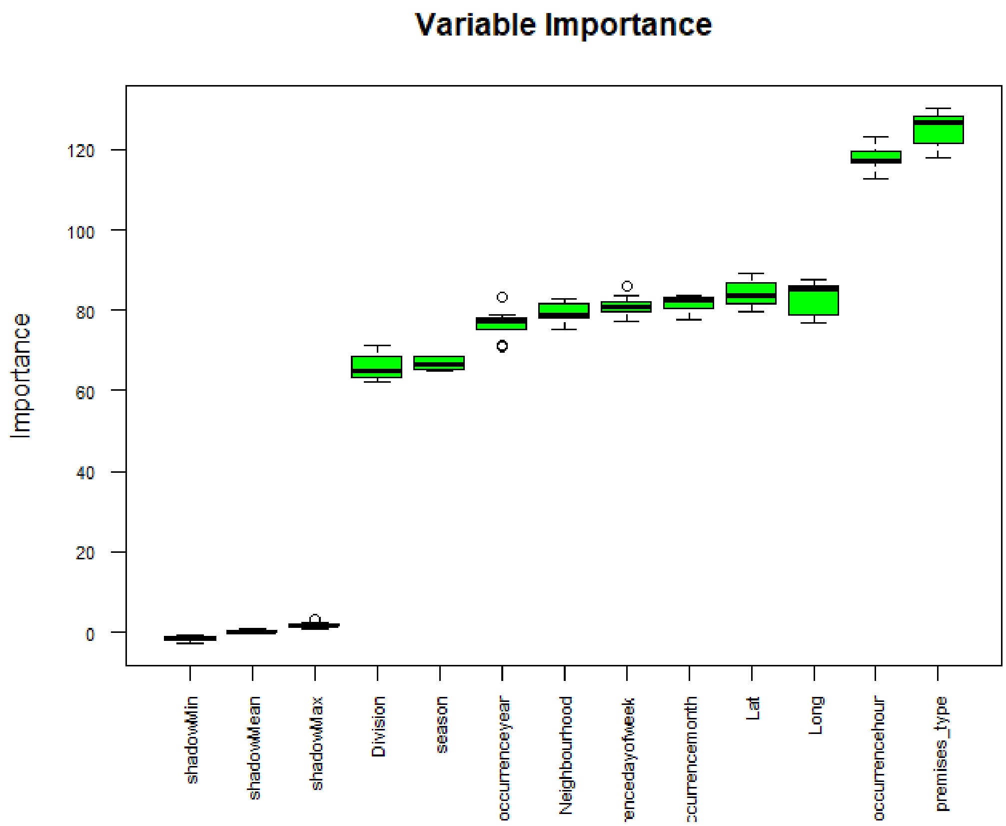


Figure 30: Results of Boruta feature ranking of the MCI dataset.

6.1.2 Stepwise and Backward Selection

The stepwise forward and backward elimination method for feature selection was implemented as a comparison to the Boruta method. This algorithm was not as costly as the Boruta algorithm and returned results within a few seconds. Stepwise and backward selection algorithms provided comparable results to the Boruta algorithm, with the stepwise selection indicating that all 10 predictive variables would produce the highest performing model (Figure 31) while backward selection indicated that 7 variables were sufficient (Figure 32). Stepwise and backward selection methods were conducted using the caret and randomforest libraries in R. Given that the MCI dataset contains significantly more samples than variables, the backward selection method was preferred.

```
Subset selection object
10 Variables  (and intercept)
                    Forced in Forced out
Division                FALSE      FALSE
premises_type           FALSE      FALSE
occurrenceyear          FALSE      FALSE
occurrencemonth         FALSE      FALSE
occurrencedayofweek     FALSE      FALSE
occurrencehour          FALSE      FALSE
Neighbourhood           FALSE      FALSE
Long                    FALSE      FALSE
Lat                     FALSE      FALSE
season                  FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: 'sequential replacement'
          Division premises_type occurrenceyear occurrencemonth occurrencedayofweek occurrencehour Neighbourhood Long Lat season
1  ( 1 )  " "      " "           " "            " "             " "                 "*"            " "           " "  " " " "
2  ( 1 )  " "      " "           " "            " "             "*"                 "*"            " "           " "  " " " "
3  ( 1 )  " "      " "           "*"            " "             "*"                 "*"            " "           " "  " " " "
4  ( 1 )  " "      "*"           "*"            " "             "*"                 "*"            " "           " "  " " " "
5  ( 1 )  "*"      "*"           "*"            "*"             "*"                 " "            " "           " "  " " " "
6  ( 1 )  "*"      "*"           "*"            " "             "*"                 "*"            " "           " "  " " "*"
7  ( 1 )  "*"      "*"           "*"            "*"             "*"                 "*"            "*"           " "  " " " "
8  ( 1 )  "*"      "*"           "*"            "*"             "*"                 "*"            "*"           "*"  " " " "
9  ( 1 )  " "      "*"           "*"            "*"             "*"                 "*"            "*"           "*"  "*" "*"
10  ( 1 ) "*"      "*"           "*"            "*"             "*"                 "*"            "*"           "*"  "*" "*"
```

Figure 31: Results of stepwise feature selection.

```
Subset selection object
10 Variables  (and intercept)
                   Forced in Forced out
Division               FALSE      FALSE
premises_type          FALSE      FALSE
occurrenceyear         FALSE      FALSE
occurrencemonth        FALSE      FALSE
occurrencedayofweek    FALSE      FALSE
occurrencehour         FALSE      FALSE
Neighbourhood          FALSE      FALSE
Long                   FALSE      FALSE
Lat                    FALSE      FALSE
season                 FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: backward
         Division premises_type occurrenceyear occurrencemonth occurrencedayofweek occurrencehour Neighbourhood Long Lat season
1  ( 1 ) " "      " "           " "            " "             " "                 "*"            " "           " "  " " " "
2  ( 1 ) " "      " "           " "            " "             "*"                 "*"            " "           " "  " " " "
3  ( 1 ) " "      " "           "*"            " "             "*"                 "*"            " "           " "  " " " "
4  ( 1 ) " "      "*"           "*"            " "             "*"                 "*"            " "           " "  " " " "
5  ( 1 ) " "      "*"           "*"            " "             "*"                 "*"            " "           "*"  " " " "
6  ( 1 ) " "      "*"           "*"            " "             "*"                 "*"            " "           "*"  "*" " "
7  ( 1 ) " "      "*"           "*"            " "             "*"                 "*"            " "           "*"  "*" "*"
```

Figure 32: Results of backward feature selection.

6.1.3 Final Selected Features

Both Boruta and backward feature selection algorithms indicate that premises type, occurrence hour, latitude and longitude are the most significant features associated with major crime indicators; hour, day of week, season, and year, while being weaker features, also show some importance. Based on apparent variable importance provided by the 2 methods, the attributes of premises type, occurrence hour, latitude, longitude, occurrence month, and occurrence day of week were selected to build the final dataset for classification (Figure 33).

```
'data.frame':   165333 obs. of  7 variables:
 $ premises_type      : int  2 4 2 4 2 4 1 2 2 4 ...
 $ occurrencemonth    : int  3 3 3 3 3 3 3 3 3 3 ...
 $ occurrencedayofweek: int  3 6 6 6 7 6 7 1 1 2 ...
 $ occurrencehour     : int  2 3 14 22 1 18 9 7 21 21 ...
 $ Long               : num  -79.6 -79.6 -79.6 -79.6 -79.6 ...
 $ Lat                : num  43.7 43.7 43.7 43.8 43.7 ...
 $ MCI                : int  5 1 4 1 1 3 1 3 3 3 ...
```

Figure 33: Structure of the final MCI dataset to be used for classification.

6.2 Class Imbalance & SMOTE Oversampling

A review of the MCI category proportions revealed that the dataset was significantly imbalanced with assault comprising 53% of samples and Theft Over only 4% (Table 2) and would require balancing prior to initial modeling otherwise classification would be biased toward the majority class of Assault. The SMOTE oversampling method was selected as a means of addressing class imbalance and was implemented using the smotefamily package.

MCI Category	Count	%
Assault	86909	53
Auto Theft	23279	14
Break & Enter	35384	21
Robbery	13844	8
Theft Under $5000	5917	4

Table 2: Class distribution of the MCI dataset.

Following the implementation of the SMOTE method, the dataset nearly doubled in size, increasing from 165333 records to 313258 (Figure 34) with a significant change in the distribution of the feature classes (Table 3). Based on the new distribution of classes, it can be inferred that Assault, Theft Under, and Break & Enter will be the most accurately classified among the MCI categories with the others likely returning lower classification accuracies.

```
data.frame':   313258 obs. of  7 variables:
$ premises_type      : num  4 6 1 2 1 1 1 2 2 1 ...
$ occurrencemonth    : num  10 12 8 10 6 6 2 7 1 2 ...
$ occurrencedayofweek: num  7 4 5 4 3 1 5 2 1 3 ...
$ occurrencehour     : num  12 17 4 7 6 0 1 19 15 12 ...
$ Long               : num  -79.4 -79.3 -79.4 -79.5 -79.5 ...
$ Lat                : num  43.7 43.7 43.8 43.6 43.7 ...
$ MCI                : num  5 5 5 5 5 5 5 5 5 5 ...
```

Figure 34: Structure of the balanced MCI dataset.

MCI Category	Count	%
Assault	86909	28
Auto Theft	23279	7
Break & Enter	35384	11
Robbery	13844	4
Theft Under $5000	15384	49

Table 3: Class distribution of the balanced MCI dataset

Once the dataset was balanced, it was split into training and test sets at a ratio of 80% to 20% using the catools library.

6.3 Model Selection and Initial Results

The algorithms chosen to classify the MCI dataset were decision tree (J48), multivariate logistic regression, Naïve Bayes, and Random Forest. Previous classification of the Toronto MCI dataset (Sundar, 2020) returned accuracies of ranging from a minimum of 40% (Naïve Bayes) to a maximum of 52% (Random Forest).

6.3.1 Decision Tree Classification

The J48 decision tree algorithm was executed using the RWeka package, had an 8-minute run time and was evaluated using 10-fold cross validation. The training model correctly classified 76.19% of instances and had a Kappa value of 0.635 indicating there was moderate agreement between actual and predicted classes (McHugh, 2012). Evaluating the predictive capability of the model returned similar accuracy and kappa values with an overall accuracy of 76.36% and kappa value of 0.637. While the model performed well overall, it could more accurately classify Assault and Theft Under $5000 compared to the other MCI classes (Table 4).

MCI Category	Sensitivity	Specificity	Balanced Accuracy
Assault	0.627	0.902	0.764
Auto Theft	0.454	0.950	0.702
Break & Enter	0.524	0.933	0.729
Robbery	0.285	0.959	0.621
Theft Under $5000	0.965	0.950	0.958

Table 4: Selected performance measures of the J48 Decision Tree algorithm on the test data.

6.3.2 Multivariate Logistic Classification

The multivariate logistic algorithm was executed using the nnet package, had a 3-minute run time and was evaluated using 10-fold cross validation. The training model had an accuracy of 49% of instances and a Kappa value of 0.0039 indicating there was little to no agreement between actual and predicted classes. Evaluating the predictive capability of the model returned similar accuracy and kappa values with an overall accuracy of 49.2% and kappa value of 0.0038; the model misclassified more values that were correctly identified. The model could most accurately classify Assault and Theft Under $5000, with accuracies of 56.4% and 61%, respectively, and could not or barely classify the other MCI classes (Table 5).

MCI Category	Sensitivity	Specificity	Balanced Accuracy
Assault	0.404	0.723	0.564
Auto Theft	0.00	0.925	0.462
Break & Enter	NA	0.887	NA
Robbery	NA	0.955	NA
Theft Under $5000	0.493	0.727	0.610

Table 5: Selected performance measures of the Multivariate Logistic Classification algorithm on the test data.

6.3.4 Naïve Bayes Classification

The Naïve Bayes algorithm was executed using the naivebayes package, had a 10 second run time and was evaluated using 10-fold cross validation. The training model had an accuracy of 61.6% of instances and a Kappa value of 0.348 indicating there was good agreement between actual and predicted classes. Assessing the predictive capability of the model returned similar accuracy and kappa values with an overall accuracy of 61.9% and kappa value of 0.356. Like the other models, the MCI categories most accurately classified were Assault and Theft Under $5000, with accuracies of 68.2% and 75.8%, respectively, with the other classes show poor to moderate classification accuracies (Table 6).

MCI Category	Sensitivity	Specificity	Balanced Accuracy
Assault	0.528	0.836	0.682
Auto Theft	0.00	0.925	0.462
Break & Enter	0.458	0.903	0.681
Robbery	0.00	0.955	0.477
Theft Under $5000	0.673	0.842	0.758

Table 6: Selected performance measures of the Naïve Bayes Classification algorithm on the test data.

6.4 Comparative Performance of Classification Algorithms – Preliminary Results

Initial classification of the Toronto MCI dataset indicated that the MCI dataset could be classified with reasonable accuracy, and that the Decision Tree algorithm notably outperformed predictive models generated using Multivariate Logistic Regression and Naïve Bayes algorithms (Table 7). The MCI categories of Assault and Theft Under $5000 were the most consistently correctly classified compared to Auto Theft, Break & Enter, and Robbery which were either poorly classified or not classified at all (Table 8). The higher classification accuracies for Assault and

Theft Under would have been influenced by their proportions within the oversampled dataset, 28% and 49%, respectively.

MCI Category	Accuracy	Kappa	Training Time
J48 Decision Tree	76.4%	0.637	8 min
Multivariate Logistic Classification	49.2%	0.0038	3 min
Naïve Bayes	61.9%	0.356	10 sec

Table 7: Comparative performance of the 3 classification algorithms.

MCI Category	DT	MLR	NB
Assault	0.764	0.564	0.682
Auto Theft	0.702	0.462	0.462
Break & Enter	0.729	NA	0.681
Robbery	0.621	NA	0.477
Theft Under $5000	0.958	0.610	0.758

Table 8: Comparative accuracy of the Decision Tree, Multivariate Logistic Regression and Naïve Bayes algorithms in classifying the MCI categories.

References

Akila, G., and Mohana, M. (2017). *Crime Analysis Using R.* IJECRT – International Journal of Engineering Computational Research and Technology, volume 2, Issue 1, December 2017. https://www.researchgate.net/publication/337414019_Crime_Analysis_Using_R.

Alves, L.G.A., Ribeiro, H.V., and Rodrigues, F.A. (2018). *Crime prediction through urban metrics and statistical learning.* Physica A: Statistical Mechanics and its Applications, volume 5.5, 1 September 2018. https://www.researchgate.net/publication/321758289_Crime_prediction_through_urban_metrics_and_statistical_learning.

Hvistendahl, M. (2016, Sept 28). *Can 'predictive policing' prevent crime before it happens?*. Science. https://www.science.org/content/article/can-predictive-policing-prevent-crime-it-happens.

Iqbal, R., Murad, M.A.A., Mustapha, A., Panahy, P.H.S., and Khanahmadliravi, N. (2013) *An Experimental Study of Classification Algorithms for Crime Prediction.* Indian J Sci Technol 6(3):4219–4225. https://doi.org/10.17485/ijst/2013/v6i3.6.

Kim, S., Joshi, P., Singh Kalsi, P., and Taheri, P. (2018). *Crime Analysis Through Machine Learning*. Conference: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). https://www.researchgate.net/publication/330475412_Crime_Analysis_Through_Machine_Learning.

Lau, T. (2020, April 1). *Predictive Policing Explained*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained.

Li, S. (2017, Oct 13). *Exploring, Clustering and Mapping Toronto's Crimes*. Towards Data Science. https://towardsdatascience.com/exploring-clustering-and-mapping-torontos-crimes-96336efe490f.

McHugh, M.L., (2012). *Interrater reliability: the kappa statistic.* Biochemia Medica 22(3):276-282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/.

Obuandike, G.N., Isah, A and Alhsan, J. (2015). *Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data.* IJARAI International Journal of Advanced Research in Artificial Intelligence, vol. 4., No.12, 2015. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.736.265&rep=rep1&type=pdf.

Oliveira, C. (2021). *Major Crime Indicators in Toronto in 2019.* Medium|Analytics Vidhya. https://medium.com/analytics-vidhya/major-crime-indicators-in-toronto-2019-dbbeaeb01d12. Source code also shared via GitHub: https://github.com/CROliveira/MCI_Toronto2019.

Palanivinayagam, A., Gopal, S.S., Bhattacharya, S., Anumbe, N., Ibeke, E., and Biamba, C. (2021). *An Optimized Machine Learning and Big Data Approach to Crime Detection.* Hindawi –

Wireless Communications and Mobile Computing, Volume 2021, Article ID 5291528. https://www.hindawi.com/journals/wcmc/2021/5291528/.

Social Development, Finance & Administration (2022). *Neighbourhoods* [Shapefile]. Open Data – City of Toronto. https://open.toronto.ca/dataset/neighbourhoods/.

Social Development, Finance & Administration. (2022). *Neighborhood Profiles* [Data set]. Open Data – City of Toronto. https://open.toronto.ca/dataset/neighbourhood-profiles/.

Statcan. (2021). *Crime Severity Index Weights*. [Data file]. Received via email request to Canadian Centre for Justice and Community Safety Statistics (statcan.ccjcss-ccsjsc.statcan@statcan.gc.ca).

Stodulka, J. (2021). *Toronto Crime and Folium.* https://www.jiristodulka.com/post/toronto-crime/.

Shojaee S, Mustapha A, Sidi F, Jabar MA (2013) *A Study on Classification Learning Algorithms to Predict Crime Status.* Int J Digital Content Technol Appl 7(9):361–369, May 2013. https://www.researchgate.net/publication/266971832.

Sundar, V. (2020). *Toronto Crime Analysis & Prediction - An effort to predict crime with supervised machine learning algorithms.* CKME136 Capstone Project, Ryerson University. https://github.com/tugga82/Toronto-Crime-Analysis_Machine_Learning.

Taneja, K., Gulati, P., Rajasekar, G and Nagi, A.S. (n.d.). *Toronto Crime Data Analysis Using Unsupervised Learning.* https://github.com/ganesh292/TorontoCrimeDataAnalysis.

Toronto Police Services. (2022). *Major Crime Indicators* [Data set]. Toronto Police Service Public Safety Data Portal. https://data.torontopolice.on.ca/search?q=crime.

Toronto Police Services. (2022). *Patrol Zones* [Shapefile]. Open Data – City of Toronto. https://open.toronto.ca/dataset/patrol-zones/.

Uwoghiren, J. (2020). Analysis of Toronto Neighbourhoods using Machine Learning: A New Immigrant's Guide to Settling in the City of Toronto. Towards Data Science. https://towardsdatascience.com/analysis-of-toronto-neighbourhoods-using-machine-learning-291b942578f2.

Vempala, N. (2016). Exploring Neighbourhood Crime in the City of Toronto using Open Data – An IPython Tutorial for Beginners. LinkedIn.com. https://www.linkedin.com/pulse/exploring-neighbourhood-crime-city-toronto-using-open-vempala-ph-d-.

Wibowo, A.H., and Oesman, T.I. (2020). *The comparative analysis on the accuracy of k-NN, Naïve Bayes, and Decision Tree algorithms in predicting crimes and criminal actions in Sleman Regency.* J. Physis.: Conf.Ser. 1450 012076. https://iopscience.iop.org/article/10.1088/1742-6596/1450/1/012076/pdf.

Yu, J., Ward, M.W., Morabito, M., and Ding, W. (2011) *Crime forecasting using data mining techniques.* In: Paper presented at the 2011 IEEE 11th international conference on data mining workshops. IEEE, Vancouver 11-11 December 2011. https://doi.org/10.1109/ICDMW.2011.56.

Zhang, X., Lui, L., Xiao, L., and Ji, J. (2020). *Comparison of Machine Learning Algorithms for Predicting Crime Hotspots*. IEEE Access. https://ieeexplore.ieee.org/document/9211482.