Toronto Neighbourhood Crime

Examining the relationship between Neighbourhood Crime Rate and Demographic Data in Toronto's 140 historical neighbourhoods in 2016

24 April 2022

Contents

1. Introduction

This research study aims to explore the relationship between crime rate and the characteristics of neighbourhoods in Toronto in 2016. Past studies showed that neighbourhood characteristics were significant factors of the crime rate of an area (Foster et.al 2010[1]). These characteristics include: social-economic factors such as number of businesses, number of community spaces such as parks, green spaces and community centers, ease of transit access, and human factors such as community belonging and quality of local communities (Statistics Canada 2021[2]). In this study, we aim to explore the relationship between the demographics, neighbourhood amenities and crime rate of neighbourhoods. Understanding this relationship will help the government and local communities respond more effectively against emerging crime rates, by eliminating risk factors and improving the well-being of a neighbourhood.

From 1996 to 2021, the City of Toronto was divided into 140 (historical) neighbourhoods. On 12 April 2022, Toronto's social planning neighbourhoods changed which resulted in an increase of the number of neighbourhoods to 158. This research used the historical neighbourhoods planning, since all data were collected before 2022. The sources of data used come from the following sources: the City of Toronto's Open Data Portal, Statistics Canada, and several custom datasets from Kaggle. Crime rates and police-reported crimes were collected from the first source, while Statistics Canada provided census data via Census of Population, containing information about people and housing units in Canada by their demographics, social and economic characteristics[3].

We first identified main demographics factors which were thought to be highly correlated with neighbourhood crime rates, by examining past studies of similar research: crime rates were found to be lower in neighbourhoods with higher proportions of senior citizens and immigrants, and in neighbourhoods which were further away from the urban city centers[4]; the proportion of visible minorities and racial heterogeneity were also found to be correlated to neighbourhood crime rates[5].

In terms of social-economic factors such as community amenities and businesses of the neighbourhood, commercial activities were found to be correlated with higher crime rate, but this relationship was mostly due to the dining businesses which opened after midnight[6]. On the other hand, researchers discovered that short-term housing, such as Airbnb rentals, contributed to higher crime rate in the surrounding area, although the increase in crime often happened after a year or more following an increase in rental listings[7].

These studies provided preliminary insights for us to select the important variables and observations from thousands of features in the census data, as well as retrieving representative data from other sources.

This report outlines the process of data collection, exploratory data analysis and preliminary testing, modeling, interpreting results and findings.

[1]Foster, S., Giles-Corti, B., & Knuiman, M. (2010). Neighbourhood design and fear of crime: A social-ecological examination of the correlates of residents' fear in new suburban housing developments. In Health & Place (Vol. 16, Issue 6, pp. 1156–1165). Elsevier BV. https://doi.org/10.1016/j.healthplace.2010.07.007

[2]Statistics Canada. (2021). Neighbourhood characteristics and life satisfaction of individuals in lower-, middle-, and higher-income families in Canadian metropolitan areas. Government of Canada. https://doi.org/10.25318/36280001202100500006-ENG

[3]Government of Canada, S. C. (2020, July 17). Census of population. Surveys and statistical programs. Retrieved March 13, 2023, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3901

[4]Statistics Canada. (n.d.). Main article. Neighbourhood Characteristics and the Distribution of Police-reported Crime in the City of Toronto. Retrieved March 13, 2023, from https://www150.statcan.gc.ca/n1/pub/85-561-m/2009018/part-partie1-eng.htm

[5]Sun, Ivan Y.; Triplett, Ruth A.; and Gainey, Randy R., "Neighborhood Characteristics and Crime: A Test of Sampson and Groves' Model of Social Disorganization" (2004). Sociology & Criminal Justice Faculty Publications. 3. https://digitalcommons.odu.edu/sociology_criminaljustice_fac_pubs/3

[6]Twinam, T. (2017, May 25). Danger zone: Land use and the geography of neighborhood crime. Retrieved April 25, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S009411901730044X?dgcid=raven_sd_aip_email

[7]Ke, L., O'Brien, D. T., & Heydari, B. (2021, July 14). Airbnb and neighborhood crime: The incursion of tourists or the erosion of local social dynamics? PLOS ONE. Retrieved April 25, 2023, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0253315

2. Methods

2.1 Data Collection

To answer our research question, data were collected from The City of Toronto's Open Data Portal (the Portal), the official source for Toronto open data from city divisions and agencies[8]. The datasets of interest were "Major Crime Indicators" and "Neighbourhood Profiles". Both datasets were readily available for download on Open Toronto as a CSV, JSON or XML format, or by using the Open Data Portal API. We retrieved the data by downloading the CSV file[9]. The "Neighbourhoods" dataset was also retrieved as a geojson file from the Portal for creating a map visualization of crime rate: it contained the coordinates of the boundaries of the 140 historical neighbourhoods in Toronto. Lastly, two datasets were used from Kaggle, Airbnb Toronto Data and Toronto Neighborhoods Information. As their names suggest, the former provide listings of Airbnb rentals in Toronto and the latter used the Foursquare API (among others) to provide additional information about Toronto neighbourhoods.

2.1.1 Dataset - Major Crime Indicators

This dataset contains all Major Crime Indicators (MCIs) occurrences in 140 historical neighbourhoods in Toronto with reported date in between 1 January 2014 and 30 June 2022. The MCIs are divided into five categories: Assault, Break and Enter, Auto Theft, Robbery and Theft Over. The Indicators excludes the Sexual Violation offences.

The variables of interest include the MCI category (mci_category), the offence (Offence), the neighbourhood of occurrence (Neighbourhood), and the occurrence and reported date and time. According to the dataset description, the data is provided at the offence and/or victim level, thus one occurrence of a crime major appear several times, each associated with different MCI categories; that is, an offence could be categorized into several categories and reported multiple times in the data[10].

Table 1 in the Appendix shows the top four rows of the raw data. Descriptions of the variables can be found in the Public Safety Data Portal: Open Data Documentation, or in the "Data Cleaning" section below.

2.1.2 Dataset - Neighbourhood Profiles

This dataset includes census data of the 140 neighbourhoods in the City of Toronto in 2016. It contains over 2300 features of the neighbourhoods, spanning areas such as population, language, income, housing, education, and labour. There are over 2300 rows in the dataset, each representing a characteristic (feature) among the 50 topics, including marital status, income of individuals etc. Each characteristic has several sub-characteristics, for example, for the income response, there were different levels of income ($10000-$14999, $15000-$19999 etc.). Based on our findings on prior research on crime rate, these were the demographic characteristics of interest: age, level of education, employment status, immigration and citizenship status, average income tax, number of visible minorities, and population count.

The data were originally collected through the Census of Population conducted nationwide by Statistical Canada, which did not release the complete data at the level of individual neighbourhoods. However, it was aggregated by Toronto Open Data into the levels of neighbourhood and the result was summarized in this dataset. Nonetheless, a small amount of data released in the Census could not be aggregated to the levels of neighbourhoods, such as summaries of income distribution of residents in the City of Toronto as a whole. This did not deter our study as the amount of data not released at the level of neighbourhoods were significantly less than not, and the variables we chose were not any of affected variables.

[8] https://open.toronto.ca/

[9] We originally retrieved the data through the portal's API, by installing the R package opendatatoronto and following the instructions in the "For Developers" section (See Neighbourhood Profiles and Major Crime Indicators. However, the Portal updated their API and the format of the data (i.e. the column names), which rendered our old code unusable.

[10] https://open.toronto.ca/dataset/major-crime-indicators/

Note that Statistics Canada carry out random rounding reporting practices, which may have an effect on aggregated statistics such as median, percentages and mean of groups with low number of observations[11].

Table 2 in the Appendix below shows the top four rows of the raw data. Descriptions of the variables can be found in the Public Safety Data Portal: Open Data Documentation, or in the "Data Cleaning" section below.

2.1.3 Dataset - Airbnb Toronto Data

This dataset from Kaggle contained over 16000 observations of Airbnb rental listings in Toronto (uploaded on 4th November 2022). The author collected the data through the Inside Airbnb website, which provided quarterly data of Airbnb listings in cities around the world. The dataset was then cleaned by the author such that each listing was located in one of the 140 historical neighbourhoods of Toronto. The variables include the listing's owner, neighbourhood, the price, property and room type etcetera.

Although the dataset was collected in 2022, it still provided invaluable insights of the distribution of short term rentals in the city of Toronto, assuming that each region has roughly the same changes in the number of listings over the years. Note that data from previous years were available but required payments.

2.1.4 Dataset - Toronto Neighbourhood Information

Similar to our other data sources, the Kaggle author created this dataset in March 2021 by combining data from "Neighbourhood Profiles", "Neighbourhood Crime Rates" from Statistical Canada, and venue information from Foursquare API . The data extracted from the first two sources were heavily cleaned and preprocessed, which we did not use because we already obtained the original datasets from the official sources (as stated above). The only data we were interested in was the gym and venue information data from Foursquare API, which contained the number of gyms and points of interest in each neighbourhood.

Although this dataset was collected after 2016, it was still important to our research as the number of venues and points of interests should roughly remain the same throughout these years.

2.2 Data Cleaning and Wrangling

For all the datasets, we first inspected the datasets using R functions such as summary, str, dim, head, is.na and functions from the tidyverse package to group and filter the data to better understand the characteristics of the data by different groups.

2.2.1 Cleaning of "Major Crime Indicators"

Unwanted variables were first removed. They included: X_id (unique identifier), event_unique_id (event identifier), Division (Police Division where offence occurred), ucr_code (UCR code for offence), ucr_ext (UCR extension for offence), and Hood_ID (identifier of neighbourhood).

The remaining wanted variables contained information about the occurrence time and date (in variables occurrencedate, occurrenceyear, occurrencemonth, occurrenceday, occurrencedayofyear, occurrencedayofweek, occurrencehour), the reported time and date (in variables reporteddate, reportedyear, reportedmonth, reportedday, reporteddayofyear, reporteddayofweek, reportedhour), the location (in variables location_type, premises_type, Neighbourhood), and the crime offence (in variables Offence and MCI).

Observations not from 2016 were also removed, since we were only interested in reported offences which occurred in 2016. We also removed observations where the neighbourhood was missing, which was represented by "NSA" ("Not Specified Area") in the data.

[11] https://open.toronto.ca/dataset/neighbourhood-profiles/

2.2.2 Cleaning of "Neighbourhood Profiles"

The "Neighbourhood Profiles" dataset contains characteristics (features) which have sub-characteristics (sub-features). However, these sub-characteristics are all listed under the same column (same variable) as the parent characteristic, except being indented by varying degrees based on the sub-characteristic relationships (i.e. a sub-sub-characteristic is indented more than a sub-characteristics, and so on). Moreover, some of the rows in the data are in complete wrong order, which made data cleaning inefficient and prone to error. For example, two sub-characteristics with very similar description are placed in the same indentation under the two characteristics of similar type (in the wrong order), which makes it impossible to distinguish which sub-characteristics belongs to which parent. To tackle this problem, we first identified the main characteristics wanted before cleaning the sub-characteristics.

We first removed the "Data Source" column (which defines the data source, unrelated to our research) and divided the dataset into 50 tibbles, each representing a different topic with varying number of (sub-)characteristics. Then, we cleaned only the features (characteristics) wanted: age, level of education, employment status, immigration and citizenship status, average income tax, number of visible minorities, and population count. These features were kept in separate tables for the purpose of preliminary testing individually. Note that some of these subtables still suffered from faulty arrangements and we could only use a subset of these sub-tables in our study.

2.2.3 Cleaning of "Airbnb Toronto Data"

The dataset was already cleaned by the author. However, although the variable is named neighbourhood_cleansed, some neighbourhood names were not standardized. Thus, we manually cleaned the neighbourhood names so that they match with the column in the datasets from the Toronto Data Portal. For example, Mimico (includes Humber Bay Shores) was changed to Mimico.

2.2.4 Cleaning of "Toronto Neighbourhood Information"

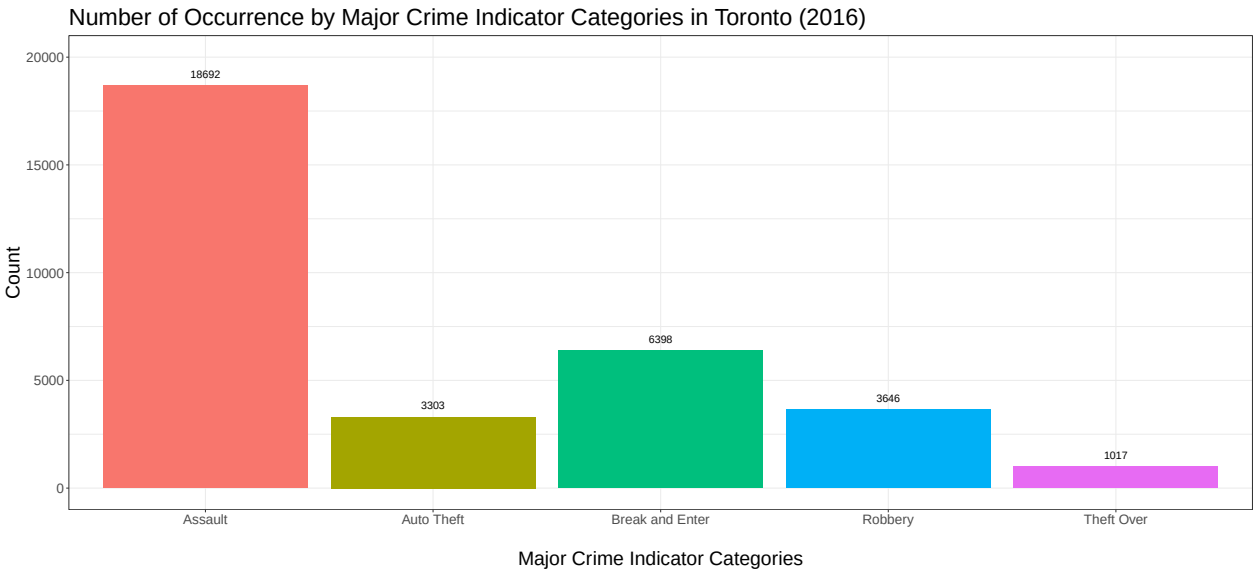
Similar to the Airbnb dataset, we standardized the neighbourhood names with the existing datasets. We also extracted the number_gyms and number_venues variable columns.

2.3 Data Exploration

After data cleaning, we performed data exploration on all the datasets, using the ggplot2 and leaflet libraries.

2.3.1 Distribution of Crime

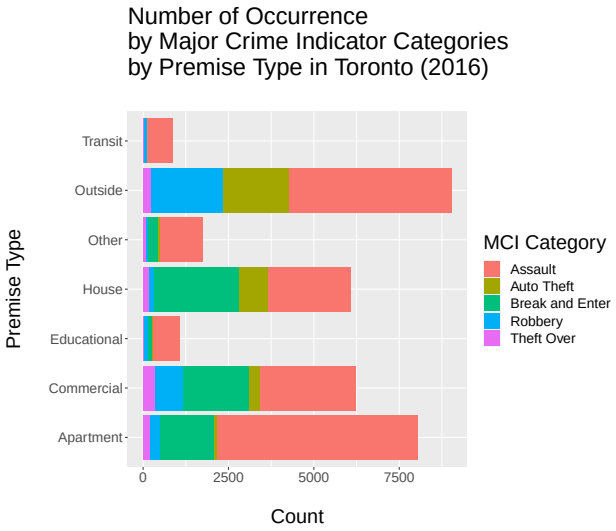
The plot below shows the number of occurrences of crimes by the five Major Crime Indicator categories in 2016. From the bar plot, assault was the most common crime in 2016, having 18692 occurrences among 33056 cases of offences (56.5%). The next category with highest occurrence was "break and enter", with a count of 6398 times (about one-third of "Assault" cases).



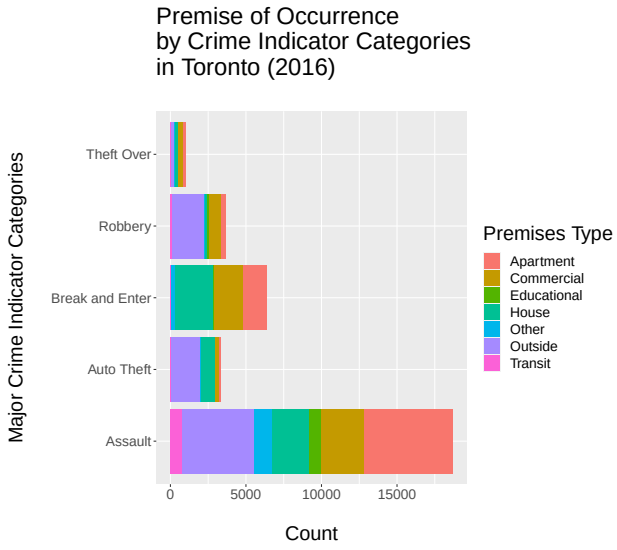
The bar plot shows the distribution of the five Major Crime Indicator categories in Toronto in 2016. Assault was the most frequent MCI offence.

The following stacked bar plots show the distribution of crime occurrences and where they occurred. From these plots, there were some interesting observations:

- "Assault" was the predominant crime in the seven premises type shown below, especially in Transit where it was about seven times more likely to occur than all the other four crimes. This suggested that neighbourhoods with large transit hubs were more likely to suffer from assault offences, or neighbourhoods with a higher population density (hence the need of more transit);
- Neglecting "Outside", Commercial areas and Apartment buildings have the most MCI offences occurred. This might mean that areas such as downtown area with a high density of commercial properties and apartment buildings have a higher crime count than in relatively suburb areas with more houses.
- Nearly half of the "auto theft" offences occurred outside. This suggested that neighbourhoods with more indoor or underground parking spaces might be less susceptible to auto theft
- The number of crimes occurred in houses were about 80% of that in apartments, but assuming that apartments usually house many more people than a house could, this suggested that the number of offences per person might be higher in houses (which appear in less busy areas) than apartments (which are often in city centers).

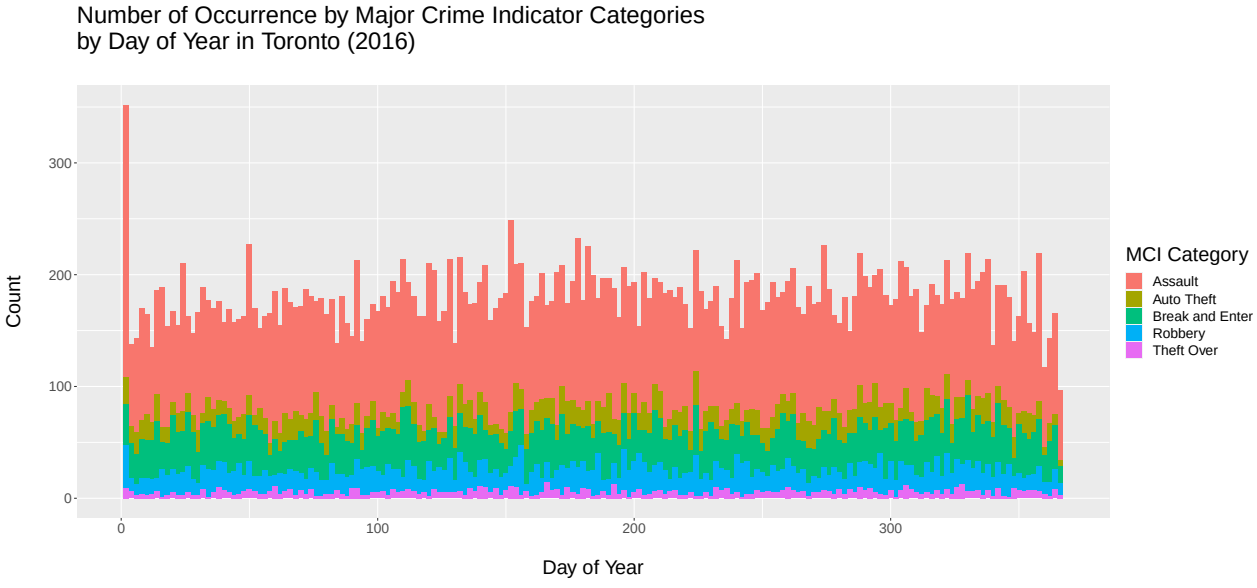


The stacked bar plot shows the distribution of the five Major Crime Indicator categories by premise type in Toronto in 2016. Commerical areas and Apartment buildings are all targets of offences.



The stacked bar plot shows the distribution of premise of occurence by the five Major Crime Indicator categories in Toronto in 2016. Robbery and Autotheft mainly occurred outside, while Break and Enter mostly occurred in Houses and Commercial areas.

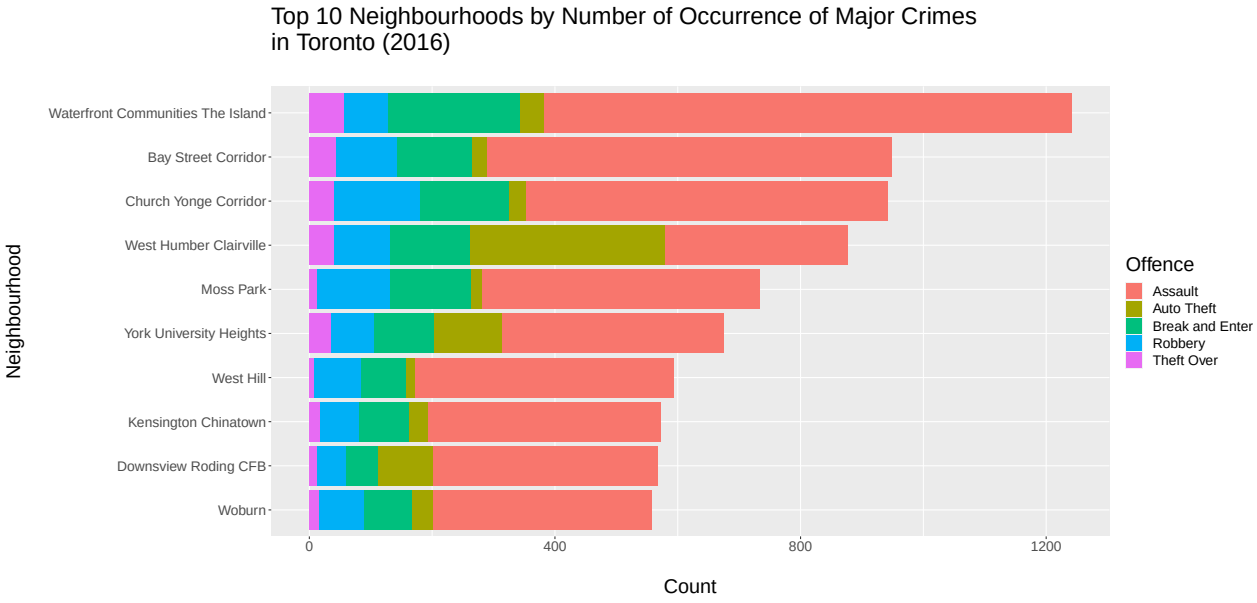
We also inspected the trend of offence occurrences with respect to time in 2016 with the histogram below. The trend shown in the plot suggests that there were no significant relationship between the day of year and the number of major crime occurred, as the plot appears to be close to a uniform distribution with only minor fluctuations periodically, with the exception of Day 1: the spike of assault cases on the first day of the year was interesting yet concerning. This spike does not seem to be from a data collection error, as the dataset had valid information on all the those cases on that day. This number might be related to new years celebration where people usually got drunk and high during celebration events and festivals, hence the increase in assault offences.



The stacked bar plot shows the distribution of the five Major Crime Indicator categories by day of year in Toronto in 2016. A large amount of assault offences were found on New Years Day.

The following barplot shows the Top 10 Neighbourhoods sorted by the number of major crime occurrences in 2016 in descending order. The results did not come as surprising, as the neighbourhood with the most population (Waterfront Communities Population: 65913) had the highest crime occurrences. Busy city centres were also in the plot (Bay Street, Church-Yonge). Interestingly, West Humber Clairville had significantly more auto theft offences compared to the other neighbourhoods, which was 80% of all the other 9

neighbourhoods combined (318 vs 385).



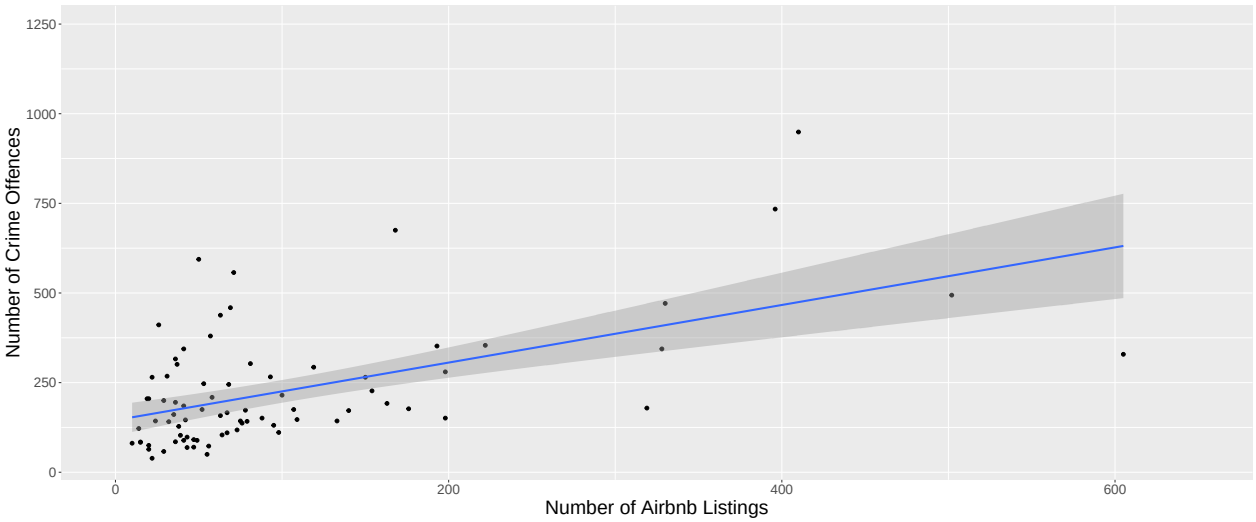
For more exploration on how the MCI offences were distributed in Toronto's neighbourhoods, see the Interactive Plots.

2.3.2 Airbnb Listings and MCI Offences

To investigate whether there was a correlation between the number of Airbnb listings and MCI offences, a scatterplot with a best-fitted line was created. The plot suggested that there was a moderate positive relationship between these two numbers. However, the variance of number of offences was much higher when the number of listings were high, which suggested that Airbnb listing was only one of the factors which affected the number of MCI offences.

Note that Waterfront Communities-The Island with 2754 listings and 1241 offences was omitted in the plot for aesthetic purposes.

Number of Airbnb Listings vs Number of Crime Offences by Neighbourhood in Toronto (2016)

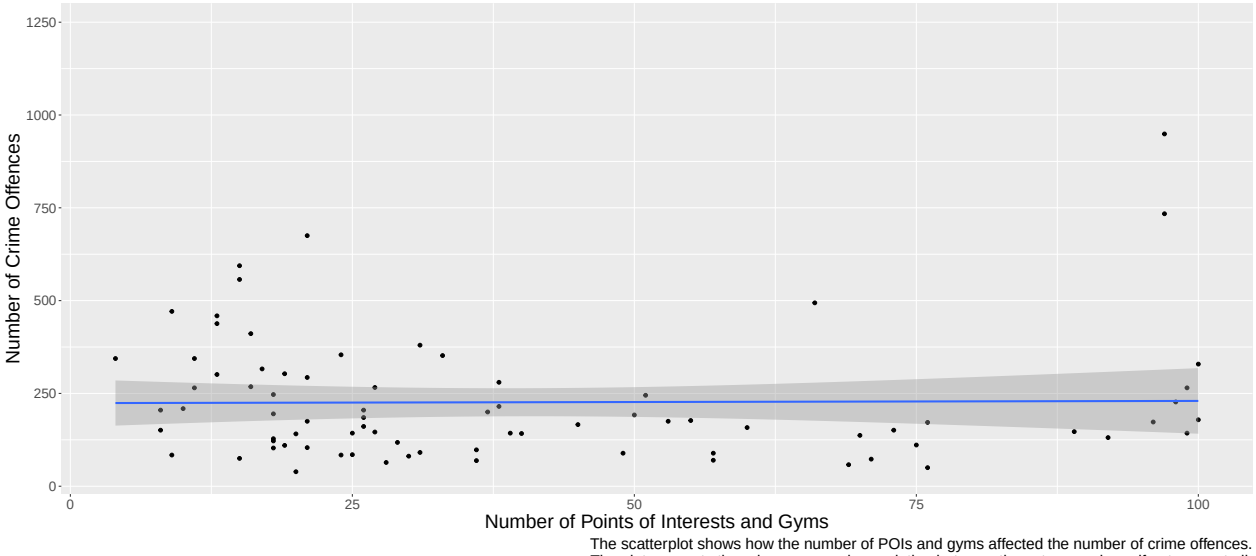


The scatterplot shows how the number of airbnb listings affected the number of crime offences.
The plot suggests there was a positive correlation between these two numbers.
Note that Waterfront Communities−The Island with 2754 listings and 1241 offences was omitted for aesthetic purposes.

2.3.3 Venues (POIs), Gyms, and MCI Offences

Next, we investigated if the number of MCI offences were related to the number of venues (POIs) and gyms by plotting a scatterplot with a best-fitted line. There was only a very weak positive relationship between these two factors, if not none at all. However, the plot shows that some outliers with high number of MCI offences occurred in neighbourhoods with the highest venues and gyms.

Number of POIs and Gyms vs Number of Crime Offences by Neighbourhood in Toronto (2016)



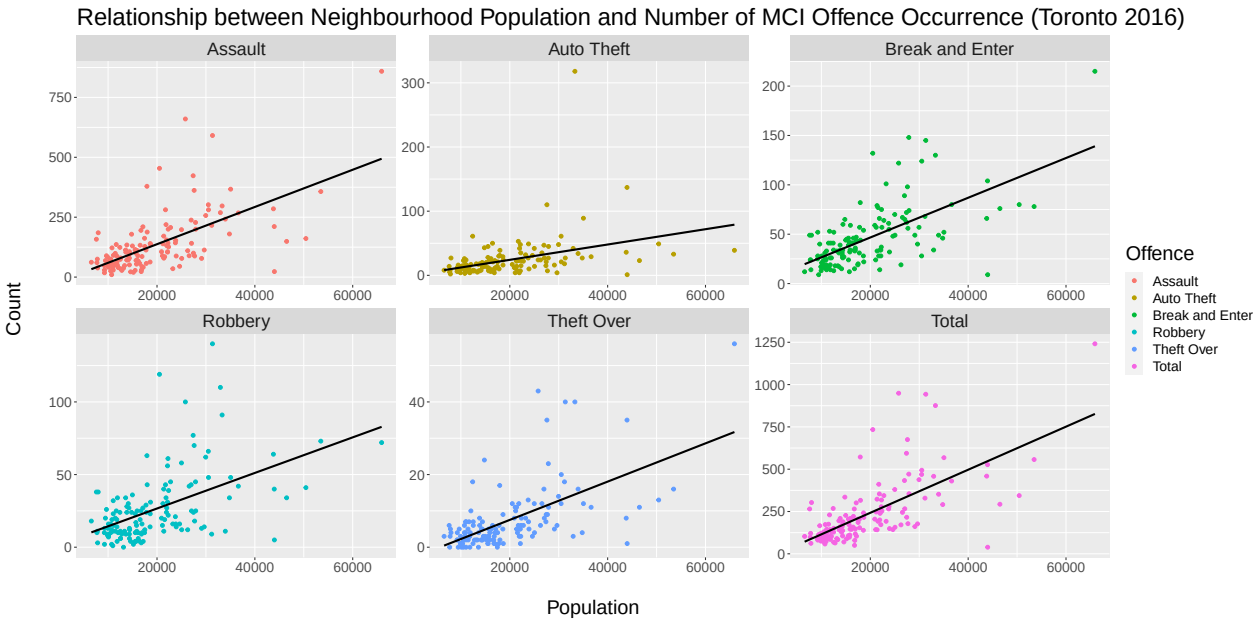
The scatterplot shows how the number of POIs and gyms affected the number of crime offences.
The plot suggests there is a very weak correlation between these two numbers, if not none at all.

2.3.4 Neighbourhood Demographics and Crime Rate

We inspected the relationship between the crime rate and the variables of interests, i.e. population size, population age, education, employment/labour, immigration and citizenship status, income taxes, and the level of visible minority. For each variable, we used the extracted sub-tables with only relevant fields created

in the "Data Cleaning" section. For each sub-table, we merged it with two other tables, one containing (Neighbourhood, Population) tuples, and one containing the number of reported crime per offence type.

2.3.4.1 Relationship between Population Size and Crime Rate It is not surprising that the best fitted lines shows that the number of offences increased with the number of population in a neighbourhood. However, in all the subplots, the majority of the neighbourhoods had a population of less than 30 thousand, and there are outliers and influential points in the plots. A better indicator of crime rate versus population might be plotting the crime rate against the population density instead, which is more representative of the scale of crime rate in terms of neighbourhood size: a neighbourhood with higher population density is more likely to have higher crime rate.



2.3.4.2 Relationship between Population Age and Crime Rate We fitted a basic linear model where the response variable was the number of offences and the variables were the proportion of different age groups in the neighbourhood. Note that in the original dataset, there were two additional age groups (Seniors (65+ years) and Older Seniors (85+ years)). However, we did not include them in this preliminary model as it would induce multicollinearity, since these proportions of age groups add up to 1.

From the p-values below, we can see that the proportion of age groups are all statistically significant. Based on the signs of the estimated coefficients, they confirm from our previous findings that a neighbourhood with larger proportion of senior residents are likely to have fewer crime rate (negative coefficient), while increasing number of youths (positive coefficient) is more likely to the crime rate[12].

However, the estimated coefficient for the intercept is much larger than the estimated coefficient for the age groups, thus any change in the percentages of age groups would only change the estimated crime rate slightly. This strongly suggested that there were other factors which affect crime rate, and the p-values show that age group is one of the significant factors.

[12]Ulmer, J. T., & Steffensmeier, D. (2014). The age and crime relationship: Social variation, social explanations. In The Nurture Versus Biosocial Debate in Criminology: On the Origins of Criminal Behavior and Criminality (pp. 377-396). SAGE Publications Inc.. https://doi.org/10.4135/9781483349114.n23

Table 1: Coefficient Estimates of a Linear Regression Model for Estimating MCI Offence Count (Age Groups)

Terms	Estimate	Std. Error	t-value	p-value
(Intercept)	57.993	21.730	2.669	0.009
`Children (0-14 years)`	-0.032	0.012	-2.660	0.009
`Youth (15-24 years)`	0.125	0.015	8.320	0.000
`Working Age (25-54 years)`	0.014	0.003	4.286	0.000
`Pre-retirement (55-64 years)`	-0.065	0.019	-3.351	0.001

2.3.4.3 Relationship between Education and Crime Rate We also fitted a basic linear model where the response variable is the number of MCI offences and the predictors are the proprtion of people with different levels of education. The baseline of education is "No certificate, diploma or degree". From the summary, we can see that an increase proportion of people with secondary and postsecondary education increases the number of crime rates, but the increase is around 3 times as big for an increase in proportion of people with only secondary education than those with postsecondary education.

Similar to the estimated coefficients for age groups, the estimated coefficient for the intercept (-581) is negative and the estimated coefficients for the factors were positive and much larger (708 and 2299). This strongly suggested that we should also look at other factors at the same time. The p-values suggest that education is a significant factor on crime rate alone, so it might be useful to combine education with other variables and check if the combinations affect crime rate altogether.

Table 2: Coefficient Estimates of a Linear Regression Model for Estimating MCI Offence Count (Education)

Terms	Estimate	Std. Error	t-value	p-value
(Intercept)	-581.034	256.482	-2.265	0.025
`Postsecondary certificate, diploma or degree`	708.739	250.138	2.833	0.005
`Secondary (high) school diploma or equivalency certificate`	2299.184	711.735	3.230	0.002

2.3.4.4 Relationship between Employment and Crime Rate The fitted linear model below shows that the effect of the number of people who did not work in a neighbourhood on crime rate was statistically significant. This result aligned with the common consensus that unemployment rate was positively correlated with crime rate (Farrington et al. 1986[13], John Howard Society of Ontario 2009[14]).

Table 3: Coefficient Estimates of a Linear Regression Model for Estimating MCI Offence Count (Employment Status)

Terms	Estimate	Std. Error	t-value	p-value
(Intercept)	57.258	28.187	2.031	0.044
`Did not work`	0.032	0.004	7.282	0.000

2.3.4.5 Relationship between Immigration and Citizenship Status and Crime Rate We plotted the number of MCI offences against the proprtion of immigrants in the 140 neighbourhoods, separated by offence type and also included a total count. From these plots, we can see that as the proportion of

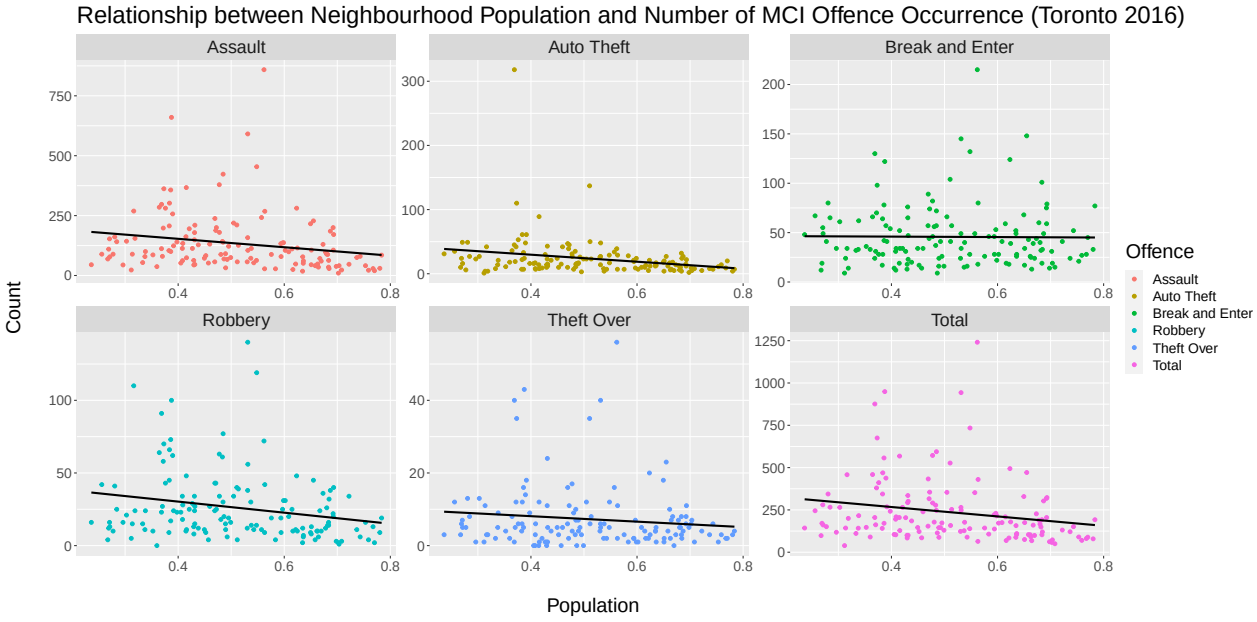
[13]Farrington, D. P., Gallagher, B., Morley, L., St. Ledger, R. J., & West, D. J. (1986). UNEMPLOYMENT, SCHOOL LEAVING, AND CRIME. The British Journal of Criminology, 26(4), 335–356. http://www.jstor.org/stable/23637076

[14]https://johnhoward.on.ca/wp-content/uploads/2014/09/facts-24-crime-and-unemployment-whats-the-link-march-2009.pdf

immigrants increased in a neighbourhood, the number of crimes committed decreases, with the exception of "Break and Enter" which stayed roughly the same. This aligned with our prior findings which concluded that the number of immigrants (statistically-)significantly reduced the number of crimes in the area[15].

Table 4: Coefficient Estimates of a Linear Regression Model for Estimating MCI Offence Count (Employment Status)

Terms	Estimate	Std. Error	t-value	p-value
(Intercept)	378.555	59.861	6.324	0.000
Immigration_Rate	-278.403	112.786	-2.468	0.015

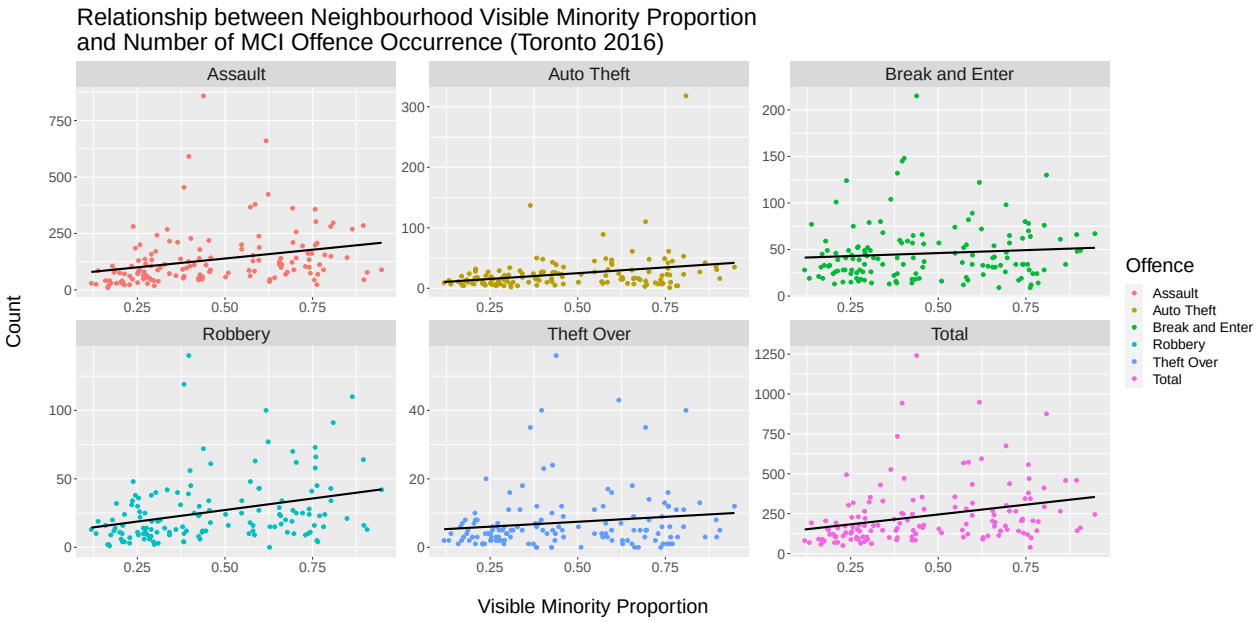


2.3.4.6 Relationship between Visible Minority and Crime Rate A racially-diversed neighbourhood were found to be linked to lower crime rate, especially in rural areas[16]. Hence, we chose to explore the relationship of neighbourhood crime rate and the proportion of visible minority because we believed that this feature was also related to immigration and citizenship status, which was found to be correlated with crime rate[17]. However, unlike the number of immigrants, we could see that crime rates were likely to be weakly positively correlated with the proportion of visible minority in the neighbourhood from the plots. The number of outliers in the plots might suggest why we obtained a different result than expected.

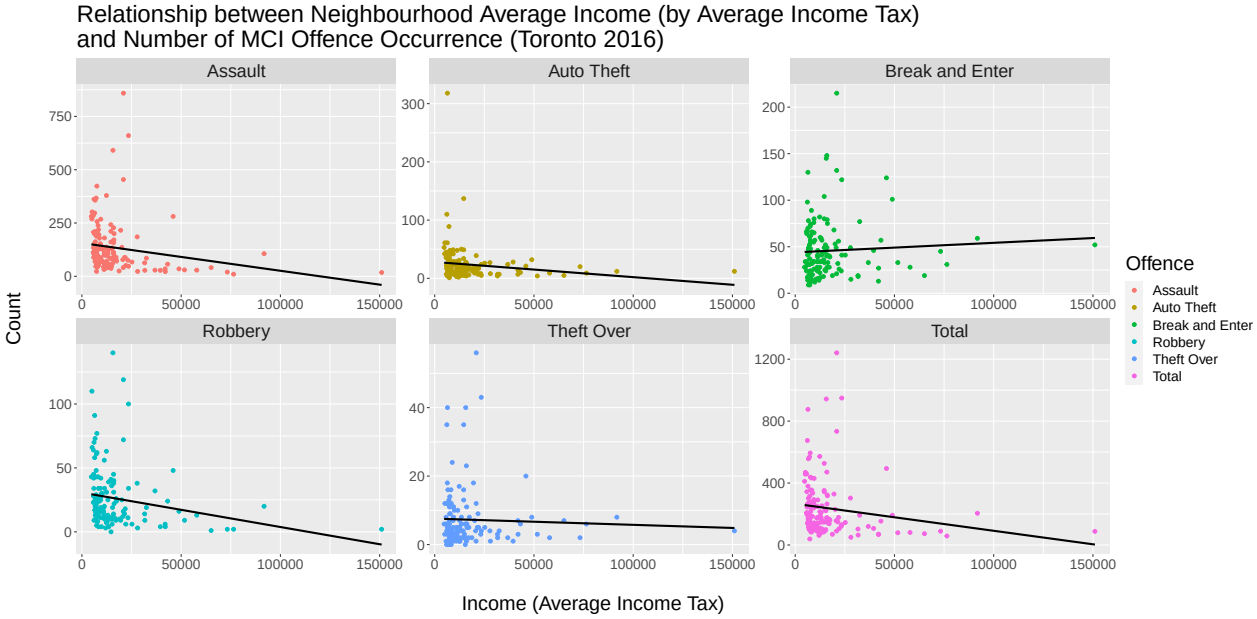
[15]Statistics Canada. (n.d.). Main article. Neighbourhood Characteristics and the Distribution of Police-reported Crime in the City of Toronto. Retrieved March 13, 2023, from https://www150.statcan.gc.ca/n1/pub/85-561-m/2009018/part-partie1-eng.htm

[16]Kim, Y.-A., & Wo, J. C. (2022, April 1). Racially diverse neighborhoods in diverse areas are linked to lower crime rates. Racially diverse neighborhoods in diverse areas are linked to lower crime rates Comments. Retrieved March 13, 2023, from https://blogs.lse.ac.uk/usappblog/2022/04/01/racially-diverse-neighborhoods-in-diverse-areas-are-linked-to-lower-crime-rates/

[17]Statistics Canada. (n.d.). Main article. Neighbourhood Characteristics and the Distribution of Police-reported Crime in the City of Toronto. Retrieved March 13, 2023, from https://www150.statcan.gc.ca/n1/pub/85-561-m/2009018/part-partie1-eng.htm



2.3.4.7 Relationship between Income and Crime Rate To estimate the effect of average individual income on a neighbourhood's crime rate, we plotted the number of offences against average income tax per offence type. The data for income groups were discovered to be corrupted during data cleaning (see Section 2.2.2). Thus, we used the average income tax as an indicator of household income level for each neighbourhood. In the subplots, the numbers of all offence types decrease as the average income taxes of the neighbourhood increase, with the exception of "Break and Enter". This might be explained by the fact that wealthy neighbourhoods were more likely to have better security, and relative risk associated with theft in wealthy neighbourhoods outweighted the possibility of stealing more expensive goods, thus deterring thefts[18].



[18]Chamberlain, A. W., & Boggess, L. N. (2016, September 26). Why disadvantaged neighborhoods are more attractive targets for burgling than wealthy ones. Why disadvantaged neighborhoods are more attractive targets for burgling than wealthy ones Comments. Retrieved March 13, 2023, from https://blogs.lse.ac.uk/usappblog/2016/09/26/why-disadvantaged-neighborhoods-are-more-attractive-targets-for-burgling-than-wealthy-ones/

2.4 Modelling

To investigate how much the factors had effects on crime rate, we trained several models to predict the number of MCI offences (total count) using age, level of education, employment status, immigration and citizenship status, average income tax, number of visible minorities, and population count, and number of airbnbs, gyms and venues.

The baseline models were Poisson regression model and negative binomial regression model. The machine learning models trained were regression tree and random forest.

With only 140 observations corresponding to the 140 neighbourhoods, a 80-20 split was used as this ratio gave us enough data for both training and testing.

A data frame with all variables of interest (response and predictors) was created using R's merge function, and new variables were created for regression: for each age group, the ratio between the size of that group and the size of working age group was created (working age group as reference); for each education level, the ratio between the number of people of that level and that of no education was created (non-working as reference); for each work status, similar ratio was created (ratio between the number of people working full-time/part-time and that who did not work); and the proportions of immigrants and visible minorities in each neighbourhood were created.

For Poisson regression and negative binomial regression, each had a model with all the predictors fitted, and the step function was used to perform backward stepwise regression with Akaike's Information Criterion (AIC) as the penalty. Afterwards, the variance inflation factors (VIF) of each remaining predictors were checked using R's vif function, and variables with VIF greater than 10 were removed and the model was refitted and rechecked. The final Poisson and negative binomial models were used as baselines for performance comparison against the machine learning models.

A regression tree was fitted using the rpart R library. The complexity parameter was chosen to be 0 (which resulted in the least amount of tree pruning). A variable importance plot was then created to investigate which factors affected the MCI offence counts to the greatest extents. Note that all predictors were used in the regression model, whereas in the Poisson and negative binomial models, some predictors were removed due to multicollinearity and/or insignificance in the stepwise regression.

Finally, a random forest was fitted using the randomForest package, which implemented Breiman's random forest algorithm[19]. The number of trees to grow was set to 5000. Similar to the regression tree model, a variable importance plot was created to interpret which predictors were the most significant on crime rate.

Although our focus was to investigate which factors affected crime rate the most, the model performances were evaluated using the mean-squared-error statistic: each model was used for prediction on both the training and testing data and the subsequent mean-squared-error for each data set was calculated. For the basic regression models, the AIC and the log-likelihood values were also calculated for reference.

3. Results

3.1 Poisson and Negative Binomial Regression

The final Poisson regression model fitted was created with all predictors with the exception of part_time_ratio (ratio between number of people working part time vs not working) and secondary_ratio (ratio between number of people with secondary level certificate vs no education/certificate).

[19] https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1/topics/randomForest

Table 5: Coefficient Estimates of Poisson Regression Model for Estimating MCI Offence Count

Terms	Estimate	Std. Error	t-value	p-value
(Intercept)	5.659	0.080	70.971	0.000
airbnb	0.000	0.000	7.897	0.000
gyms	-0.080	0.007	-11.191	0.000
venues	0.001	0.000	2.186	0.029
population	0.000	0.000	51.775	0.000
children_ratio	-1.700	0.089	-19.186	0.000
youth_ratio	4.113	0.118	34.733	0.000
pre_retirement_ratio	-2.961	0.208	-14.233	0.000
seniors_ratio	-0.400	0.097	-4.116	0.000
postsecondary_ratio	-0.031	0.002	-16.487	0.000
full_time_ratio	-0.100	0.023	-4.276	0.000
immigrants_ratio	-0.358	0.020	-18.242	0.000
visible_minority_ratio	0.021	0.003	7.526	0.000

Table 6: Coefficient Estimates of Negative Binomial Regression Model for Estimating MCI Offence Count

Terms	Estimate	Std. Error	t-value	p-value
(Intercept)	5.057	0.300	16.841	0.000
gyms	-0.072	0.033	-2.198	0.028
venues	0.003	0.002	1.768	0.077
population	0.000	0.000	12.022	0.000
children_ratio	-1.591	0.460	-3.460	0.001
youth_ratio	3.945	0.657	6.004	0.000
pre_retirement_ratio	-2.863	0.691	-4.143	0.000
postsecondary_ratio	-0.042	0.010	-4.057	0.000
immigrants_ratio	-0.253	0.077	-3.265	0.001

The tables above show the coefficient estimates of the predictors selected in the fitted Poisson and negative binomial models. Predictors which appeared in both model had highly similar coefficient estimates. All predictors in both models, with the exception of the number of venues in the negative binomial model, were all statistically significant at the 5% level.

In summary, age group had the highest impact on the crime rate, as reflected by the large magnitude of estimated coefficients. The ratio of visible to non-visible minorities, the number of venues all had positive impacts on the crime rate. The Poisson model suggested that the number of Airbnbs in the neighbourhood was not a significant factor of MCI offence counts. On the other hand, an increase in the number of gyms, the ratio of people with post-secondary education to one who did not have any certificates/education, and the ratio of immigrants to non-immigrants, were all linked with a decrease in crime count, with immigrants having the most impactful role.

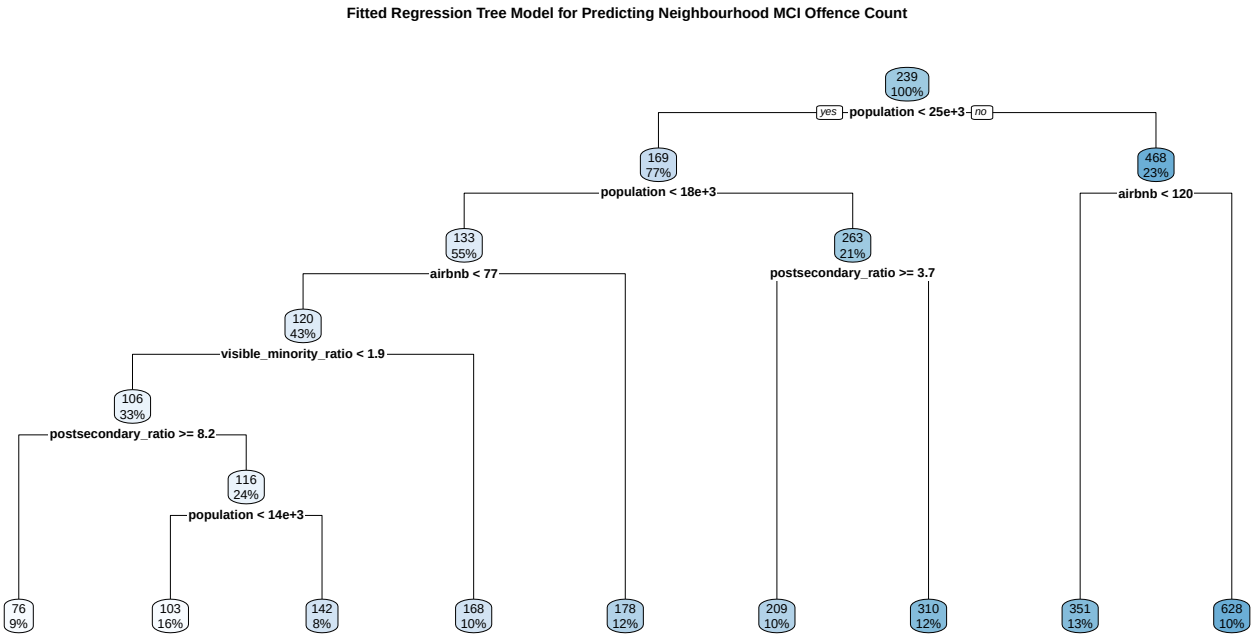
The estimated coefficients of the age groups were interesting: a one-unit increase in the ratio of the number of youths (15 - 24 years) to the number of working age people (25 - 54 years) had a multiplicative impact of 54 on the mean crime rate (a 53-times increase), while a one-unit increase in ratio of children (0 - 15 years) or pre-retirement ages (55 - 64 years) was predicted to decrease the mean crime rate by 80% and 95% respectively. The p-values of estimated coefficients of the age group predictors were all very small, which confirmed that the proportions of different age groups had a statistically significant impact on MCI offence counts in neighbourhoods.

Poisson and Negative Binomial Regression Model Statistics

Model	Train.MSE	Test.MSE	AIC	Deviance
Poisson	9247.67	15630.57	4167.50	3349.88
Negative Binomial	12793.42	17805.41	1286.32	113.71

The performance of the negative binomial regression model was much better than the Poission regression model. Although the former model had higher training and testing MSEs (12793.42 vs Poisson's 9247.67, 17805.41 vs Poisson's 15630.57), it had much lower AIC (1286.32 vs 4167.50) and deviance (113.71 vs 3349.88). Despite their differences, both models could be adapted as baseline models to evaluate the performances of the advanced machine learning models.

3.2 Regression Tree and Random Forest

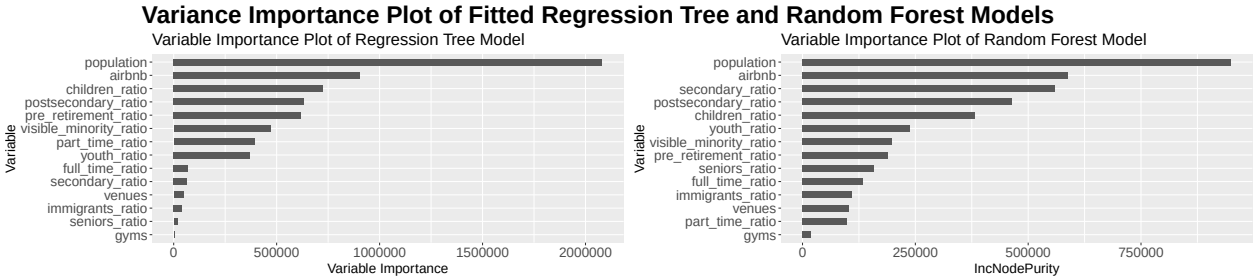


Regression Tree and Random Forest Model Statistics

Model	Train.MSE	Test.MSE
Regression Tree	15215.15	21061.30
Random Forest	3113.12	10966.88

The fitted regression tree model with all predictors were simple. The decision nodes were created by only a few predictors of interest: population, the number of Airbnbs, the ratio of people with postsecondary education vs no certificates/education, and the ratio between visible and non-visible minorities. The resulting tree had only 9 leaves, but the training and test mean-squared-error were both similar to the baseline regression models. However, the performance of the random forest was much better than all three model others: it had a training and test MSE of 3113.12 and 10966.88 respectively.

Despite the simplicity of the regression tree, it may be useful to use the model as a guideline to predict the MCI offence counts of a neighbourhood, or as a validation model to validate predictions from other models. However, it is not advisable to use the regression tree as a prediction model: the low number of leaves means the predicted values could be far from the true values.



In general, both the regression tree and random forest suggested that population and the number of airbnbs were the most important factors of crime count, followed by education, age and work status (with varying importance), then immigration and citizenship status, and the number of venues and gyms had the lowest significance on crime count.

From the variance importance plots, the order of predictors in terms of variable importance were very similar: population was the most important factor, which was not surprising as the crime count should be higher for a larger population. The second most important variable was the number of Airbnbs, which had the about half of the importance of population. What followed were education level (ratio between post-econdary/secondary level education and no certificates), the proportion (ratio) of children and youths in the neighbourhood, and the ratio of visible_minorities vs non-visible minorities. In both models, the number of venues and gyms were some of the variables with the lowest importance.

All four models suggested that age groups indeed had the most significant impact on crime rate. However, the number of Airbnbs had much higher significance (in terms of impact, not in the sense of p-value) in the fitted regression tree and random forest model, but not the regression baseline models. Nonetheless, all four models were similar in terms of variable (predictor) importances.

4. Conclusions and Summary

In this research study, we aimed to explore the relationship between Toronto's neighbourhood crime rate and the neighbourhood's demographics and the community amenities, and investigated which factors affected crime rate to the greatest extents. Poisson regression, negative binomial regression, regression tree and random forest models were deployed to understand this relationship. The fitted random forest model had the best performance in terms of training and testing mean-squared-error, followed by Poisson regression, negative binomial regression, and regression tree.

Our research found out that the population played the largest role on a neighbourhood's crime rate, followed by different age groups: the proportion of youths (15 - 24 years) were correlated with the highest increase in crime, while the proportions of children (0 - 15 years) and pre-retirement-aged people (55 - 64 years) were correlated with the largest drop in crime. The proportions of immigrants and visible minorities were also notable factors of crime, with neigbourhoods having higher proportion of immigrants and lower proportion of visible minorities having significantly fewer crime occurred. Moreover, people with post-secondary or secondary education are less likely to commit crime, while an increase in proportion of people with no certificates or education was connected to a higher crime rate.

Consistent with previous studies, the number of Airbnb's in the area was found to be highly correlated with an increase in crime rate. However, the number of venues (points-of-interests) and gyms did not seem to affect MCI offences in the area, contrary to prior studies which concluded that areas with larger business zonings and more tourists attractions were prone to higher crime rate.

5. Limitations and Next Steps

There were several limitations in our research study.

First, the small amount of corrupted/mis-placed observations in the "Neighbourhood Profiles" dataset prevented us from using the income groups directly to estimate the effect of average household and individual income on neighbourhood crime rate. Instead, we utilised the average income tax to represent the wealthiness of neighbourhoods. Although income tax was positively correlated with individual and household income, the data come in aggregated form (one number, "average tax income", for each neighbourhood), compared to the exact numbers of individuals/household in each household. This reduced the amount of information we could use and this could affect the final result in our preliminary testing.

Second, for the regression models, despite the low variance inflation factors of the predictors, there might be high correlations between predictors. For example, the proportion of visible minorities in a neighbourhood might be positively correlated with the number of immigrants in the area. Although we removed predictors with high VIFs in the baseline Poisson and negative binomial models, the underlying relationship between predictors still existed. This limitation was hard to remove entirely as statistics in demographic data were usually correlated with one another. In future studies, it might be beneficial to include a wider variety of data (even if demographic in nature) to reduce multicollinearity when fitting a regression model.

6. References

1. Foster, S., Giles-Corti, B., & Knuiman, M. (2010). Neighbourhood design and fear of crime: A social-ecological examination of the correlates of residents' fear in new suburban housing developments. In Health & Place (Vol. 16, Issue 6, pp. 1156–1165). Elsevier BV. https://doi.org/10.1016/j.healthplace.2010.07.007

2. Statistics Canada. (2021). Neighbourhood characteristics and life satisfaction of individuals in lower-, middle-, and higher-income families in Canadian metropolitan areas. Government of Canada. https://doi.org/10.25318/36280001202100500006-ENG

3. Government of Canada, S. C. (2020, July 17). Census of population. Surveys and statistical programs. Retrieved March 13, 2023, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3901

4. Statistics Canada. (n.d.). Main article. Neighbourhood Characteristics and the Distribution of Police-reported Crime in the City of Toronto. Retrieved March 13, 2023, from https://www150.statcan.gc.ca/n1/pub/85-561-m/2009018/part-partie1-eng.htm

5. Sun, Ivan Y.; Triplett, Ruth A.; and Gainey, Randy R., "Neighborhood Characteristics and Crime: A Test of Sampson and Groves' Model of Social Disorganization" (2004). Sociology & Criminal Justice Faculty Publications. 3. https://digitalcommons.odu.edu/sociology_criminaljustice_fac_pubs/3

6. Ulmer, J. T., & Steffensmeier, D. (2014). The age and crime relationship: Social variation, social explanations. In The Nurture Versus Biosocial Debate in Criminology: On the Origins of Criminal Behavior and Criminality (pp. 377-396). SAGE Publications Inc.. https://doi.org/10.4135/9781483349114.n23

7. Farrington, D. P., Gallagher, B., Morley, L., St. Ledger, R. J., & West, D. J. (1986). UNEMPLOYMENT, SCHOOL LEAVING, AND CRIME. The British Journal of Criminology, 26(4), 335–356. http://www.jstor.org/stable/23637076

8. Kim, Y.-A., & Wo, J. C. (2022, April 1). Racially diverse neighborhoods in diverse areas are linked to lower crime rates. Racially diverse neighborhoods in diverse areas are linked to lower crime rates Comments. Retrieved March 13, 2023, from https://blogs.lse.ac.uk/usappblog/2022/04/01/racially-diverse-neighborhoods-in-diverse-areas-are-linked-to-lower-crime-rates/

9. Chamberlain, A. W., & Boggess, L. N. (2016, September 26). Why disadvantaged neighborhoods are more attractive targets for burgling than wealthy ones. Why disadvantaged neighborhoods are more attractive targets for burgling than wealthy ones Comments. Retrieved March 13, 2023, from https://blogs.lse.ac.uk/usappblog/2016/09/26/why-disadvantaged-neighborhoods-are-more-attractive-targets-for-burgling-than-wealthy-ones/

10. Twinam, T. (2017, May 25). Danger zone: Land use and the geography of neighborhood crime. Retrieved April 25, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S009411901730044X?dgcid=raven_sd_aip_email

11. Ke, L., O'Brien, D. T., & Heydari, B. (2021, July 14). Airbnb and neighborhood crime: The incursion of tourists or the erosion of local social dynamics? PLOS ONE. Retrieved April 25, 2023, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0253315

7. Appendix

Table 1. Sample of Dataset - Major Crime Indicators

X_id	event_unique_id	Division	occurrencedate	reporteddate	location_type
1	GO-20141273318	D31	2014-01-03	2014-01-03	Apartment (Rooming House, Condo)
2	GO-20141274349	D42	2014-01-03	2014-01-03	Single Home, House (Attach Garage, Cottage, Mo
3	GO-20141274052	D22	2014-01-03	2014-01-03	Open Areas (Lakes, Parks, Rivers)
4	GO-20141276966	D53	2014-01-03	2014-01-03	Other Commercial / Corporate Places (For Profit,

Table 2. Sample of Dataset - Neighbourhood Profiles (Census)

X_id	Category	Topic	Data.Source	Characteristic
1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Nu
2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designa
3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016
4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011

Table 3. Sample of Dataset - Airbnb Toronto Data

id	listing_url	host_name	neighbourhood	neighbourhood_cleansed
27640141	https://www.airbnb.com/rooms/27640141	Liora	Toronto	Dovercourt-Wallace Emerson-Junc
27826009	https://www.airbnb.com/rooms/27826009	Brianne	Toronto	Waterfront Communities-The Islar
27647117	https://www.airbnb.com/rooms/27647117	Alexandra	Toronto	Playter Estates-Danforth
27647509	https://www.airbnb.com/rooms/27647509	Rosana	Toronto	High Park North

Table 4. Sample of Dataset - Toronto Neighbourhoods Information

Toronto Neighbourhoods Information

Neighborhood	Total.population	number.of.educated.people	number.of.15.45	number.of.employers
Agincourt North	30280	19805	11850	13230
Agincourt South-Malvern West	21990	14535	8840	9860
Alderwood	11900	7915	4520	6240
Annex	29180	23495	15095	16770

Table 5. Sample of Cleaned Dataset - Major Crime Indicators

occurrencedate	reporteddate	location_type	premises_type	offence
2016-01-01	2016-01-01	Apartment (Rooming House, Condo)	Apartment	Assault With
2016-01-06	2016-01-06	Single Home, House (Attach Garage, Cottage, Mobile)	House	B&E
2016-01-01	2016-01-01	Single Home, House (Attach Garage, Cottage, Mobile)	House	Assault With
2016-01-06	2016-01-06	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside	Robbery - Mu

Table 6. Sample of Cleaned Dataset - Airbnb Toronto Data

id	listing_url	host_name	neighbourhood	neighbourhood_cleansed
27640141	https://www.airbnb.com/rooms/27640141	Liora	Toronto	Dovercourt Wallace Emerson Junc
27826009	https://www.airbnb.com/rooms/27826009	Brianne	Toronto	Waterfront Communities The Islar
27647117	https://www.airbnb.com/rooms/27647117	Alexandra	Toronto	Playter Estates Danforth
27647509	https://www.airbnb.com/rooms/27647509	Rosana	Toronto	High Park North

Table 7. Sample of Cleaned Dataset - Toronto Neighbourhoods Information

Airbnb Toronto Data

Neighbourhood	number_gyms	number_venues
Agincourt North	0	26
Agincourt South Malvern West	0	34
Alderwood	1	17
Annex	3	63
Banbury Don Mills	2	14
Bathurst Manor	1	26
Bay Street Corridor	1	96
Bayview Village	3	37