**Exploratory Data Analysis of Major Crimes in Toronto using R**

Raviteja Bodla

Metro College of Technology

August-2018

1. **Introduction**

Crime in Toronto has been relatively low in comparison to other major cities. A 2017 ranking of 60 cities by The Economist ranked Toronto as the fourth safest major city in the world, and the safest major city in North America. However, there has been a recent increase in crime rates in Toronto. This project is conducted to explore the major crime incidents in Toronto other than homicide. The objectives of the project are to:

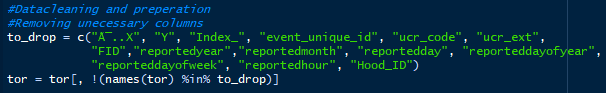
* Find the data regarding the major crimes in Toronto.
* Clean the data.
* Determine the major crime type.
* Determine the months with high crime rate.
* Determine the hour of the day with high crime rate.
* Determine the neighbourhood with high crime rate.

1. **Data Acquisition and Cleaning**

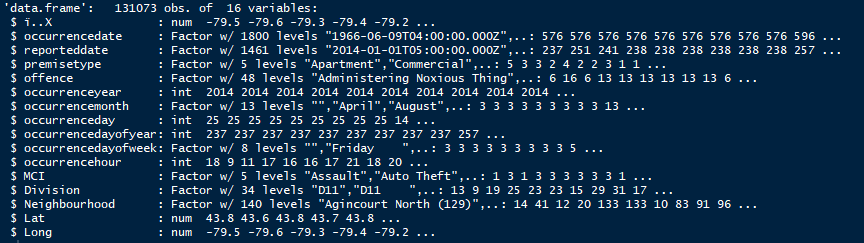
One dataset was used for the analysis. The dataset contains major crime incidents (MCI) for the city of Toronto between the years 2014-2017.

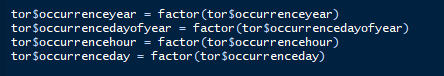
The dataset was acquired from Toronto Police Service public safety data portal (<http://data.torontopolice.on.ca/pages/open-data>). The website contains crime data reported between 2014-2017. The MCI dataset originally had 131073 observations and 29 variables. The main aspects to be cleaned in the data included non-factor to factor conversion, factor to non-factor conversion, dealing with missing data, redundant variables. Upon search for duplicate cases, none were found.

* Out of the 29 variables originally uploaded into R, latitude was present as ‘i…X’ and ‘Lat’, and longitude was present as “Y” and “Long”, so ‘i…X’ and “Y” were deleted. There were a total of 6 ID columns ("Index\_", "event\_unique\_id", "ucr\_code", "ucr\_ext", "FID", “Hood ID”), which were not required for data analysis, hence were deleted. Variables "reportedyear","reportedmonth", "reportedday", "reporteddayofyear", "reporteddayofweek" were removed as the reported date was already present and all the data analysis will be done on occurrence dates. A total of 14 columns were deleted leaving the total number of variables to 15.

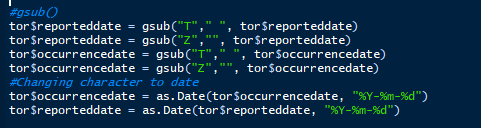


* 'occurrenceyear', 'occurrencedayofyear', 'occurrencehour', 'occurrenceday' were imported as integers, however should be factors. The above-mentioned non-factors were converted to factors using factor(function).

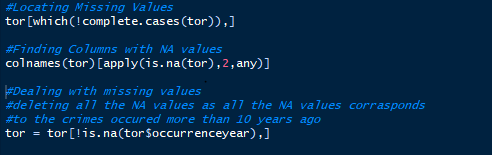




* ‘occurrencedate’ and ‘reporteddate’ were uploaded as strings. There were unwanted alphabets “T” and “Z” present in the date, which were removed using gsub() and the string was converted to date using as.Date() function.



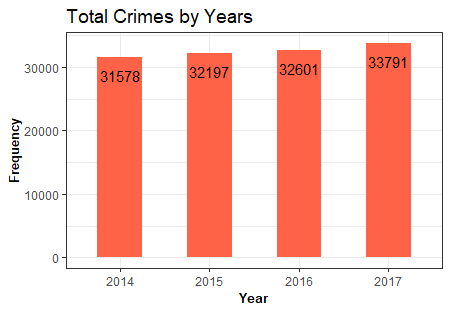
* Upon use of “!completed.cases()” function on the dataset, it was found that the dataset contained columns with missing data. Using apply() function, it was found that the variables "occurrenceyear", "occurrenceday", "occurrencedayofyear" had missing values. All the columns with missing values were deleted.



1. **Data Exploration**

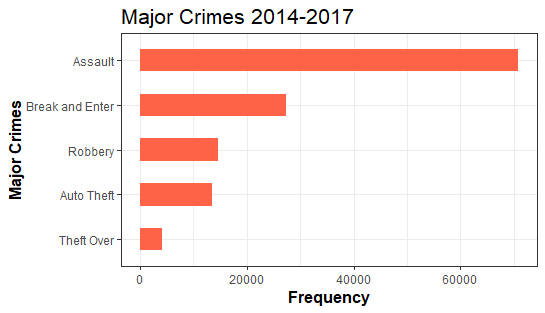
There were 131073 observations and 15 variables in the cleaned dataset for the crimes that were reported in 2014, 2015, 2016, 2017 respectively. The crimes were divided into five categories (Assault, Break and Enter, Robbery, Theft Over and Auto Theft) and 140 neighbourhoods in Toronto.

**Crime rates by year:**



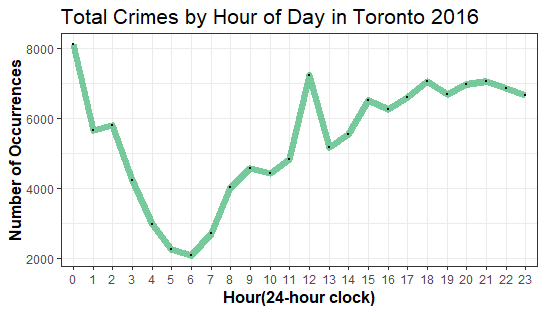
Crime rates constantly increased from 2014-2017, but the rate of increase was not so high.

**Major Crimes:**



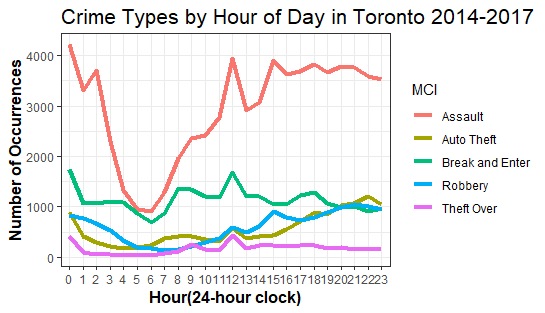
Assault is the most prevalent form of violent crime in Toronto. What is assault? In criminal and civil law, assault is an attempt to initiate harmful or offensive contact with a person, or a threat to do so.

* **Crime by time of day**



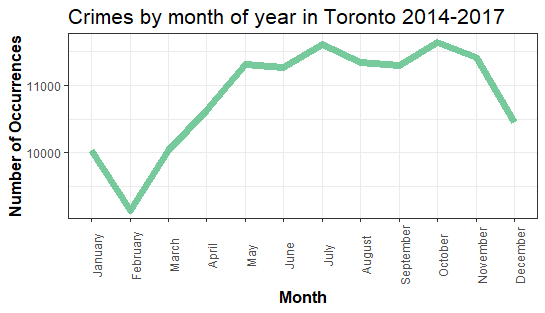
The worst hour is around the midnight, another peak time is around the noon, then again at around 8pm.

* **Types of crimes by hours:**



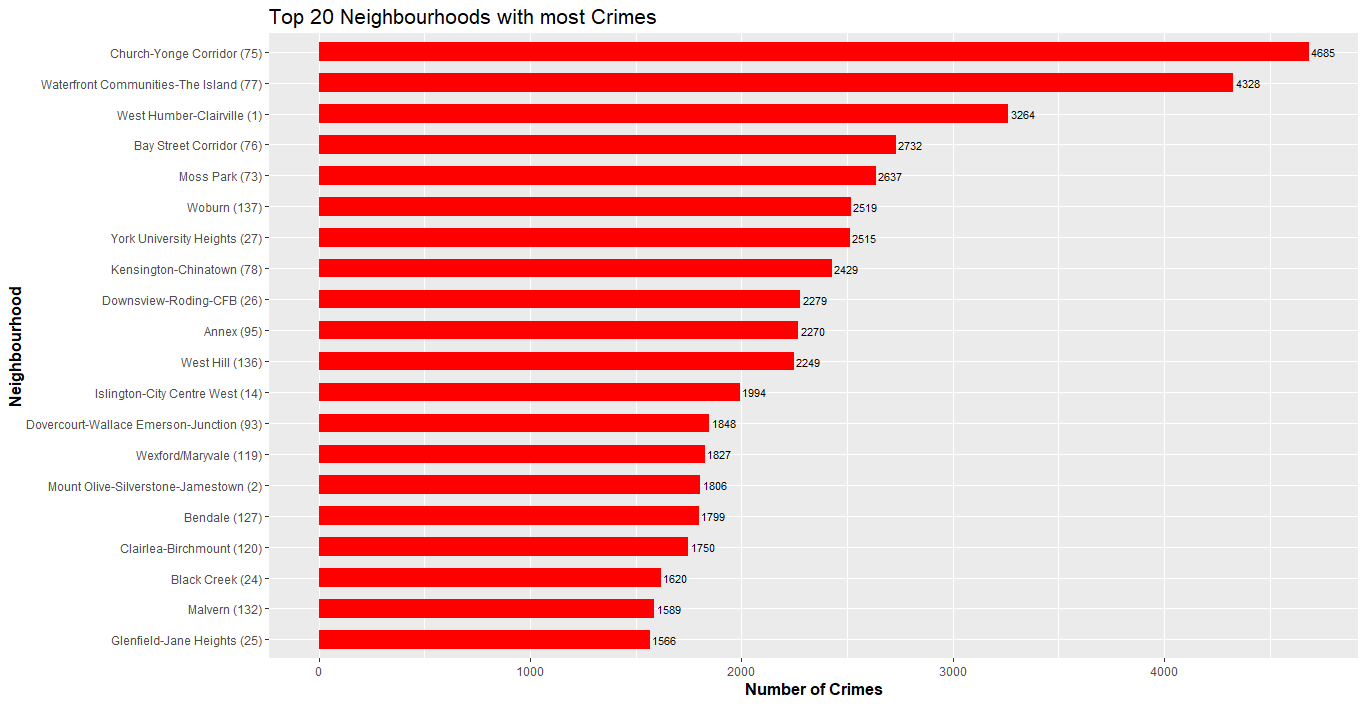
Assaults are the top crimes almost all the time, they happened more often from the late mornings till nights than the early mornings. On the other hand, break and enter happened more often in the mornings and at around the afternoon (when no one at home or office). Robberies and auto thefts are more likely to happen in the late evenings till the nights. They all make sense.

* **Crime rate by month:**



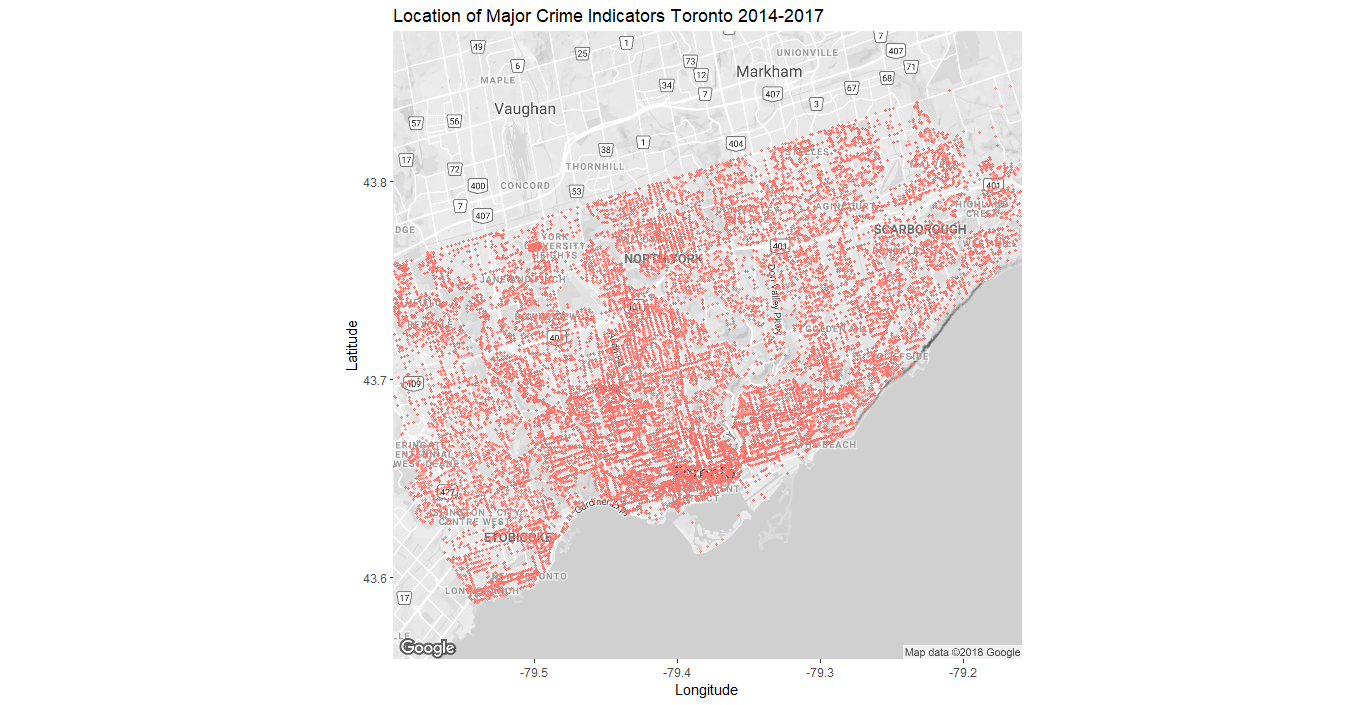
The crime rates decrease from November to February, gradually increase from February to May and is high from May to October.

* **Crime rates by location:**



The most dangerous neighbourhood Church-yonge Corridor, followed by Waterfront.

* **Map of Toronto Crimes:**



It’s clear to see where the major crimes in the city occur. A large concentration in the Waterfront area, Etobicoke is more peaceful than any other areas.

1. **Conclusions:**

* Crime rate increased from 2014-2017, but the rate of increase was not significant.
* The major type of crime is assault with 70817 cases followed by auto-theft with 13435 cases.
* Majority of crimes takes place during mid nights and early mornings. Break in and enter seems to also occur during afternoons (when people are at office).
* May-November are the months with high crime rate and November-February are the months with low crime rates.
* Church-yonge Corridor and Waterfront are the neighbourhoods with high crime rates.
* Further study has to be done based on population density and crime rates to accurately determine the neighbourhoods with high crime rates.