Exploratory Data Analysis of the Major Crime Indicators for the City of Toronto

Ishan Soni

Table of Contents

Executive Summary

According to the 2023 Global Peace Index, Canada is ranked as the 11th most peaceful country in the world. However, this statistic is hard to believe when we hear about the various crimes taking place on a daily basis. Open Data provided by the government serves as a tool to provide us with information regarding public safety and awareness.

An exploratory data analysis is performed on the Major Crime Indicators Dataset of the Toronto Police Service from 2014 to 2024. The dataset is downloaded as a CSV (comma separated values) file. Several recurring terms are defined for contextual understanding of the reader.

The dataset is uploaded in a Python notebook environment and converted to a data frame. The analysis is started by understanding the general characteristics of the data frame. This is followed by formulating questions to define the goal for the analysis. This analysis considers geospatial factors to gain insights on the crime happenings in the city. Going forward, the data frame is cleaned and prepared for analysis. Finally, an exploratory data analysis is performed to answer the questions. Visualizations such as heatmap, bar chart, and line graph are generated to summarize the different aspects of the data.

At the end, several conclusions are drawn pertaining to the observations and visualizations of the crime patterns based on the geospatial factors of the city which serve as insights for the Toronto Police Service to take the necessary steps and make the city a safe place for its residents.

1. Introduction

1.1 Background

Toronto is the financial and business capital of Canada and the most populous city in the country. These factors make it a potential hotspot for crime in the country. This report is an exploratory data analysis of the Major Crime Indicators Open Data provided by the Public Safety Data Portal of the Toronto Police Service. The Major Crime Indicators categories include Assault, Break and Enter, Auto Theft, Robbery and Theft Over. The dataset excludes sexual violations and occurrences that have been deemed unfounded. The term unfounded is defined as: “It has been determined through police investigation that the offence reported did not occur, nor was it attempted” (Statistics Canada, 2020).

The analysis will help us gain various insights into the crime patterns of Toronto and possibly draw conclusions and recommendations regarding public safety over the last ten years. The dataset is in the form of a comma separated values (CSV) file, and it is analyzed in Python using multiple data manipulation and visualization libraries such as pandas, matplotlib and seaborn.

1.2 Special Terms

When performing exploratory data analysis on the dataset, certain terms will be commonplace. Below is an overview of the more common words and abbreviations which will be used in the report later to explain the analysis:

Exploratory Data Analysis is a process to identify patterns/trends within a dataset

Data frame is a two-dimensional array of rows and columns, just like a spreadsheet, which stores data.

MCI is the abbreviation for Major Crime Indicators.

Fig. is the abbreviation for figure.

App. is the abbreviation for appendix.

Other less common terms will be defined as they are used in the analysis.

2. Analysis

2.1 Understand the Dataset

The dataset is available in various formats in the data portal – CSV (comma separated values), KML (Keyhole Markup Language), ShapeFile, and GeoJSON (Geospatial JavaScript Object Notation). This analysis uses the CSV format. The CSV file is transformed into a pandas data frame in Python.

This data frame is yet to be cleaned. However, some preliminary analysis can provide a brief understanding of the data frame. The dataset is regularly updated by the Toronto Police Department to include the latest information. As of April 5th, 2024, the data frame has 31 columns and 384687 rows. It includes fields to indicate the location coordinates, the occurrence and reporting times for the crime, approximate location data for where the crime occurred, and information about the nature of the crime.

2.2 Define the Goal for Analysis

Formulating the right questions assist in categorizing the crimes based on different factors, discover connections between the crimes and find patterns in their occurrences. The main question or the goal for this analysis is: How are the crime occurrences and crime types related to the geospatial aspects of the city such as location and neighbourhood? Numerous insights will be generated under the umbrella of this question.

2.3 Clean the Dataset

The data frame needs to be cleaned before analysis. Cleaning the data frame will optimize it for analysis. The data frame is checked for empty and duplicate records. Having empty and duplicate records can result in unforeseen errors in the future and possibly skew the results of the analysis. It is found that there are 0.031454% of empty records in the occurrence year, occurrence month, occurrence day, occurrence day of year and occurrence day of week columns. It is observed that there are no duplicate records present in the data frame. Finally, the empty records are dropped.

Moving further, the unclean data frame has a total of 31 columns out of which 17 columns are dropped. Some of these columns included the approximate location coordinates for the crimes, and a detailed breakdown of the reported date and occurrence date columns into various columns in terms of the day, month, year, day of year, day of the week, and hour. These columns are deemed to be redundant and useless because they do not aid in drawing conclusions which are aligned with the goals of the analysis. They are

useful when performing a more in-depth analysis defined by a different set of goals and questions. The data frame has 14 columns remaining after this process.

Additionally, the data types of the remaining fields are checked. Most of the columns have the “object” data type while a few are defined as “int64” integer data type. The report date and occurrence date columns are converted to the “datetime64[ns, UTC]” date data type. This ensures uniformity and correctness among all the records of the two columns. Finally, the pre-processed data frame has 14 columns and 384566 rows. A new CSV file is created, and the clean data frame is copied to this file for analysis.

2.4 Analyze the Dataset

There are a total of 54 unique location types identified in the dataset. A bar chart is plotted (Fig. 1 of the App.) to identify the top ten crime prone location types. Closed spaces such as apartments, single homes and houses have recorded the highest number of incidents. Commercial and profit places rank fourth on the chart. A striking feature of this visualization is that “Schools during Supervised Activity” is placed eighth on the graph.

A heatmap is plotted (Fig. 2 of the App.) to identify the distribution of the five MCI categories in these ten locations. It is observed that Assault cases has been reported the most in all these locations with the highest occurrences in Apartments, Streets and Houses. It can also be seen in the heatmap that apartments and houses have accounted for the highest number of break and enter cases. This examination adds weight towards the possibility that many of the break and enter cases might also result in the incidents of

assault. Further examination states that there are 25351 events where Assault and Break and Enter occurred simultaneously.

Following from the observation in the previous paragraphs, “Schools During Supervised Activities” have reported a total 5316 Assault cases. A line graph is plotted (Fig. 3 of the App.) to identify the trends in these cases over the period of 10 years. It is safe to say from the graph that schools reported around 600 cases each year for the initial half of the decade period. The number stooped below 200 during the year 2020 due to the Covid-19 nation-wide lockdown and rapidly rose to a peak of over 700 cases in 2023 in span of just 3 years. An interesting feature of the graph is that less than 100 cases were reported in the first quarter of 2024.

Moving forward, the Toronto Police Service have split the city into 158 neighbourhoods. Since the number of neighbourhoods is high, a bar chart is plotted (Fig. 4 of the App.) to identify the top fifteen crime prone neighbourhoods in the city. West Humber-Clairville recorded the highest number of cases followed by Moss Park, Downtown Yonge East, York University Heights, Yonge-Bay Corridor, and others. A line graph is plotted (Fig. 5 of the App.) to identify the trend in the incidents in West Humber-Clairville. Similar to the previous line graph, this graph also shows a spike in the crimes post Covid-19 lockdown.

A heatmap is generated (Fig. 6 of the App.) to showcase the relationship between the top 15 crime prone neighbourhoods and the five MCI categories. Analogous to the other heatmaps, this heatmap also depicted a majority of Assault cases in all the

neighbourhoods. It is also interesting to note that West Humber-Clairville reported 4897 cases of Auto Theft: almost 2000 more than Assault. Furthermore, it is observed that in most of these neighbourhoods, incidents of Break and Enter are higher than those of Auto Theft. Insights from these observations will be stated in the conclusion.

Conclusions

The observations from this exploratory data analysis can assist the authorities in identifying the crime patterns across different locations and neighbourhoods of the city and take the necessary steps. Several conclusions can be drawn on the basis of the above analysis.

Schools being ranked 8th out of 54 defined location categories sparks concern regarding the safety of the students in the city. High occurrence of crimes in residential locations poses a hazard to the safety of the residents and their family members in their homes. Moreover, there is a high probability that criminals will cause harm to the people when they break and enter into these spaces. It is also concluded that crime rates have rose to higher peaks in the post pandemic years. There can be several factors causing this spike including inflation, unemployment, and high interest rates. However, these aspects will need a deeper analysis and are out of the scope of this report.

The conclusions derived from this analysis serve as primary insights which should be taken into account by the Toronto Police Service department along with other concerning authorities to tackle crime and make the city a safe place for its citizens.

Appendix

Figure 1 - Top 10 Crime-Prone Location Types

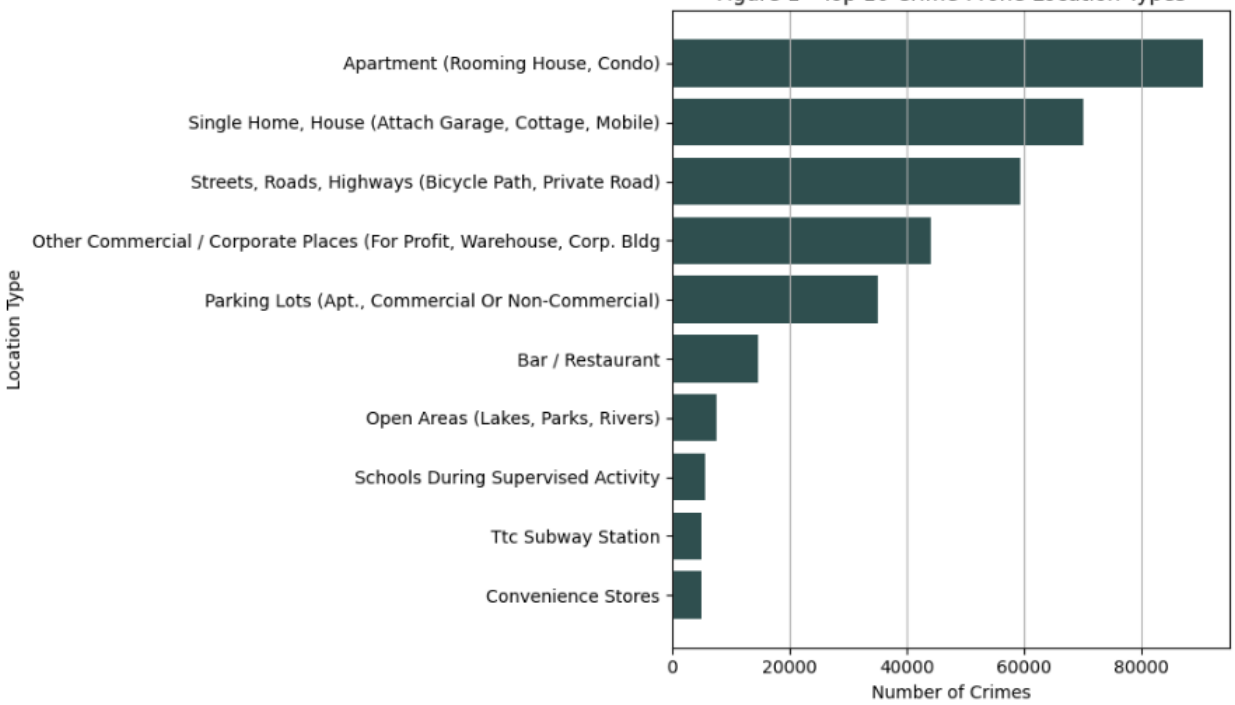

Figure 2 - Heatmap of MCI Categories vs. Location Types

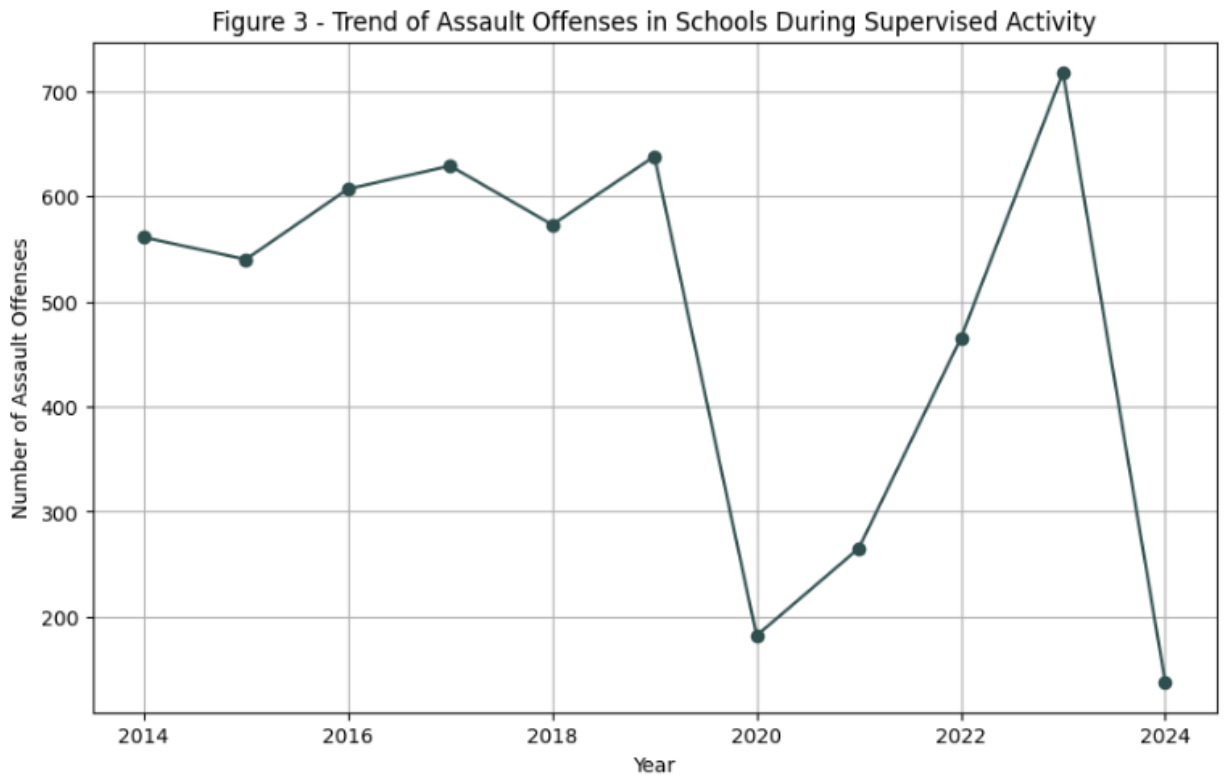
Figure 3 - Trend of Assault Offenses in Schools During Supervised Activity
Number of Assault Offenses
700
600
500
400
300
200
2014
2016
2018
2020
2022
2024
Year

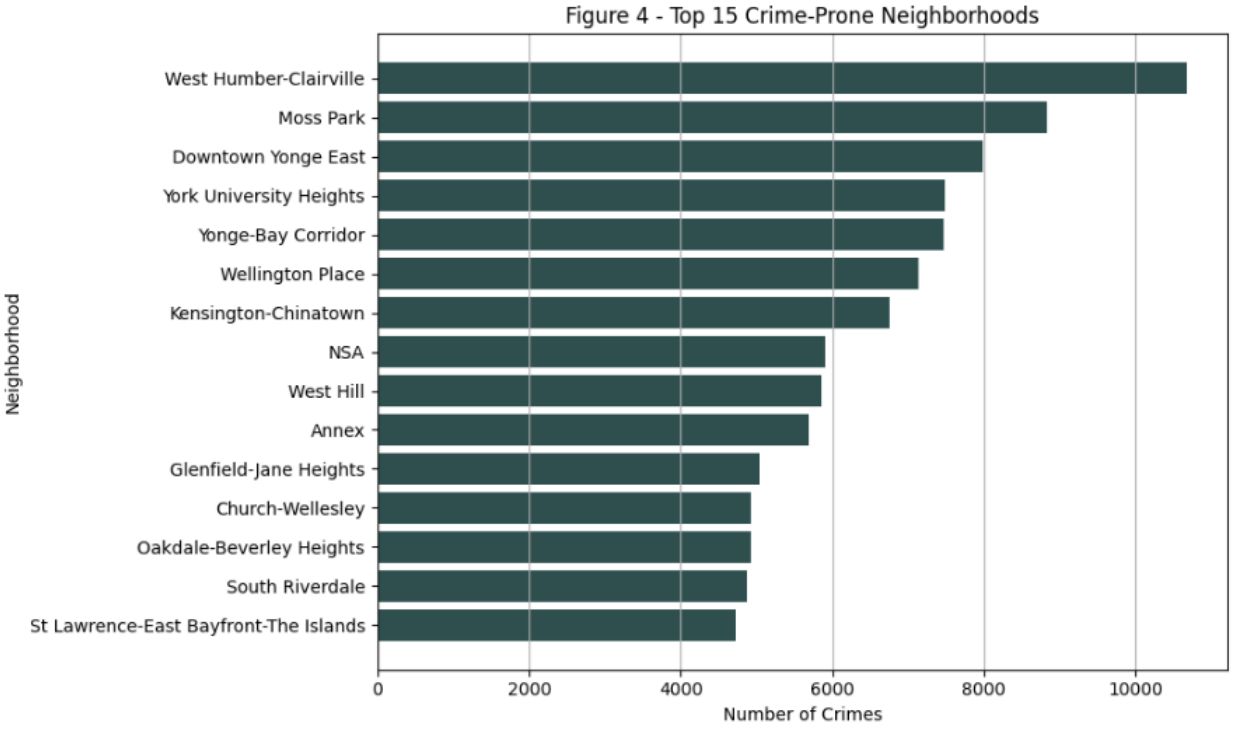
Figure 4 - Top 15 Crime-Prone Neighborhoods
Neighborhood
West Humber-Clairville
Moss Park
Downtown Yonge East
York University Heights
Yonge-Bay Corridor
Wellington Place
Kensington-Chinatown
NSA
West Hill
Annex
Glenfield-Jane Heights
Church-Wellesley
Oakdale-Beverley Heights
South Riverdale
St Lawrence-East Bayfront-The Islands
0
2000
4000
6000
8000
10000
Number of Crimes

Figure 5 - Trend of Incidents in West Humber-Clairville

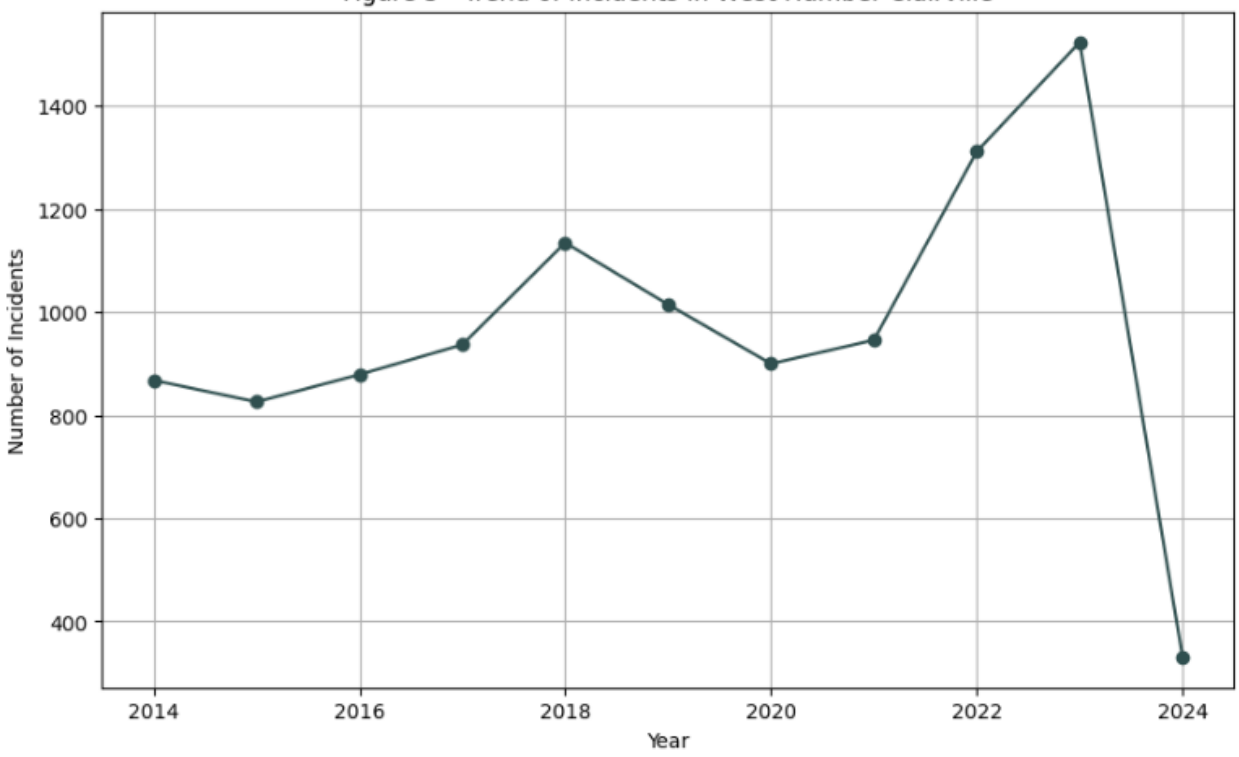

Figure 6 - Heatmap of MCI Categories vs. Neighbourhoods

Bibliography

Toronto Police Service (2023) *Major Crime Indicators Open Data*, *Toronto Police Service Public Safety Data Portal*. Available at: https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about (Accessed: 11 April 2024).

Institute for Economics & Peace (2023) *Global Peace Index 2023*, *Vision of Humanity*. Available at: https://www.visionofhumanity.org/wp-content/uploads/2023/06/GPI-2023-Web.pdf (Accessed: 11 April 2024).