Is Our Community Getting Safer in Recent Years: A Data-Driven Study of Local Crime Incidents in Toronto*

Huayan Yu

September 24, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Data Source and Computational Environment

The dataset used in this study, "Police Annual Statistical Report: Victims of Crime," is provided by the City of Toronto through its open data initiative. [1] It reflects crime incident reports, focusing on the experiences of victims from 2019 to 2023. The data is publicly accessible, ensuring transparency in city governance and fostering community engagement in crime monitoring. Ethically, the data adheres to the Open Government License, which mandates privacy protections for individuals involved, promoting responsible use in academic and policy analysis. [2] Statistically, the dataset offers insights into crime trends, aiding in evidence-based policymaking.

All data analysis and visualization in this paper were performed using the R programming language (R Core Team 2023) along with the following packages: `tidyverse` (Wickham et

*Code and data are available at: https://github.com/huayan1998/toronto_crime_analysis-main

al. 2019) which includes the `dplyr` component (Hadley Wickham 2023) for data frame manipulation and `ggplot2` (Wickham 2016) for visualizations, as well as `knitr` (Xie 2023) for table formatting.

2.2 Data Description

The dataset consists of 1244 records from 2014 to 2023, with each column representing critical information as follows:

- `REPORT_YEAR`: The year in which the crime was reported.
- `CATEGORY`: The general classification of the crime, such as "Crimes Against the Person."
- `SUBTYPE`: A more detailed breakdown of the crime, including types like "Assault" or "Robbery."
- `SEX`: Gender of the victim, represented as Male (M), Female (F), or Unknown (U).
- `AGE_GROUP`: The age range of the victim.
- `AGE_COHORT`: Specific age groupings (e.g., 25 to 34) for more detailed demographic analysis.
- `COUNT_`: The number of occurrences for each record.

In similar studies, other datasets like overall crime reports which usually focus more on offenders could have been used. However, these datasets were not appropriate for our study because this dataset uniquely focuses on victims, providing insights specific to victim demographics and experiences that are important for understanding crime impacts within Toronto local communities.

```
# Load the data
data <- read.csv("../data/raw_data/toronto_crime_victims_raw.csv")

# Convert necessary columns to factors
data$CATEGORY <- as.factor(data$CATEGORY)
data$SEX <- as.factor(data$SEX)
data$AGE_COHORT <- as.factor(data$AGE_COHORT)
data$AGE_GROUP <- as.factor(data$AGE_GROUP)
data$SUBTYPE <- as.factor(data$SUBTYPE)
data$ASSAULT_SUBTYPE <- as.factor(data$ASSAULT_SUBTYPE)

# Table 0: data summarization table
# Replace NA values with empty strings for the selected columns
summary_data <- summary(data[,c(3, 4, 6, 8)])
summary_df <- as.data.frame.matrix(summary_data)
summary_df[is.na(summary_df)] <- ""
kable(summary_df, caption = "Number of Records in Each Categorical Variable", row.names =
```

Table 1: Number of Records in Each Categorical Variable

CATEGORY	SUBTYPE	SEX	AGE_COHORT
Crimes Against the Person:1244	Assault :617	F:518	Unknown :193
	Other :222	M:570	25 to 34:165
	Robbery :201	U:156	35 to 44:154
	Sexual Violation:204		45 to 54:148
			18 to 24:143
			55 to 64:128
			(Other) :313

Table 1 above gives an overview of the distribution of key variables, showing that all crimes recorded in the dataset fall under the "Crimes Against the Person" categories. "Assault" is the most common subtype (617), with a fairly even gender distribution (518 females, 570 males). Age cohorts show a diverse range, with "Unknown" being frequent, and many victims in the 25 to 34 age range (165).

The dataset has no missing values in each of its column,and we will not be analyzing the column `ASSAULT_SUBTYPE` because limited information is found in the Open Data Toronto data description page. [1] Hence no further data cleaning is needed other than removing the `ASSAULT_SUBTYPE` column. This comprehensive dataset facilitates focused analysis on victim profiles and crime trends, making it valuable for understanding community safety.

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

3 References

[1]. About Police Annual Statistical Report - Victims of Crimes (https://open.toronto.ca/dataset/police-annual-statistical-report-victims-of-crime/) [2]. Open Government Licence – Ontario (https://www.ontario.ca/page/open-government-licence-ontario) [3]. R [4]. tidyverse [5]. dplyr [6]. ggplot2 [7]. knitr

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.