

# Are some women more susceptible to crime than others in Toronto?\*

A visualisation of crime rates over the years, and the ages they affect most

Sehar Bajwa

January 25, 2024

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

## Table of contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Data</b>	<b>2</b>
2.1	Data Overview . . . . .	2
2.2	Data Cleanup and Processing . . . . .	3
2.2.1	Missing Data Points . . . . .	3
2.2.2	Categorisation of Crimes . . . . .	3
2.2.3	Group-wise summation: The sumcount Variable . . . . .	3
<b>3</b>	<b>Visualisation</b>	<b>4</b>
3.1	Changing crime trends in Toronto over the years . . . . .	4
3.1.2	Discussion . . . . .	5
3.1.3	Model justification . . . . .	5
<b>4</b>	<b>Results</b>	<b>5</b>
<b>5</b>	<b>Discussion</b>	<b>6</b>
5.1	First discussion point . . . . .	6
5.2	Second discussion point . . . . .	6
5.3	Third discussion point . . . . .	6
5.4	Weaknesses and next steps . . . . .	6

---

\*Code and data are available at: <https://github.com/SEHB2012/crime-victims-toronto>

<b>Appendix</b>	<b>7</b>
<b>A Additional data details</b>	<b>7</b>
<b>B Model details</b>	<b>7</b>
B.1 Posterior predictive check . . . . .	7
B.2 Diagnostics . . . . .	7
<b>References</b>	<b>8</b>

## 1 Introduction

You can and should cross-reference sections and sub-sections.

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

## 2 Data

This section delves into the characteristics of note and outlines the process undertaken to create a ready-to-analyse dataset, including addressing missing values and standardizing variable names.

### 2.1 Data Overview

Published by the Toronto Police Service on the Toronto Open Data Portal, the Toronto Police Annual Statistical Report on Crime Victims covers all crimes committed against the person, including those deemed unfounded post-investigation. The entries are filtered by the reported year, and each year is associated with the following features: type of crime, age cohort of victim, gender, and counts of the exact crime. The dataset was initiated in 2014 and is marked ‘updated annually’, but the latest recorded year available is 2022.

With a focus on demographic insights, the dataset stands at 1111 data points across 9 variables. Age cohort is a defining feature of this dataset, and refers to one of 8 age bins that are as follows: “>12”, “12 to 17”, “18 to 24”, “25 to 34”, “35 to 44”, “45 to 54”, “55 to 64”, “65+”. Understandably, efforts were made to attribute recognizable labels to these cohorts, but there was admittedly difficulty distinguishing between the middle aged adults that arguably spanned three of the 9 age cohorts. Therefore, the original identifiers were retained.

## 2.2 Data Cleanup and Processing

R (R Core Team 2021) was the language and environment used for the bulk of this analysis, alongside the dplyr (Wickham et al. 2021), tidyverse (Wickham et al. 2019), janitor (Firke et al. 2021), heatmap(cite). To enhance analytical readiness, missing values were addressed, variable names are standardized and new summation columns created.

### 2.2.1 Missing Data Points

The dataset featured incomplete data points, which are discounted at this time by filtering and subsequent removal. Though making analysis remarkably easier, this poses a compromise regarding validity and poses significant implications for subsequent analyses, a concern explored in detail later.

### 2.2.2 Categorisation of Crimes

The dataset features 4 major crime types (Sexual Violation, Robbery, Assault, and Other). There is an additional data point labelled Assault subtype, providing subcategories only for this of the four major crime types. These subtypes include Assault/Aggravating Peace Officer, Bodily harm, and resist Arrest. The subtype column has been discounted after deliberation since it skews the dataset in terms of specificity. While the first three labels are forthright and succinct, Other is vague, and trying to find it's intended definition proved a trickier task than expected. According to the glossary accompanying the Annual Police Statistical report, Other criminal code violations are defined as Non-traffic Criminal Code violations that are classified as neither violent nor property violations. The label Other is retained, but this definition proves useful to mention to viewers trying to gain an absolute understanding of the data.

### 2.2.3 Group-wise summation: The sumcount Variable

Removing the crime subtype column leaves multiple different entries for the same age group, crime and year. For example, An assault could have been recorded in 2015 against a 65 + female, but it could have been a resisted arrest, or bodily harm. Summing up the crime counts per category removes duplicate entries with matching data (with the only differentiator being crime counts). A new variable, labelled sum count was initiated for this purpose, and added as a new column.

Report Year	Crime Type	Age Cohort	Sum Count
2015	Assault	25 to 34	4
2015	Assault	35 to 44	7
2015	Assault	45 to 54	2
2015	Assault	65+	2
2015	Assault	25 to 34	1
2015	Assault	18 to 24	10

Cleaned data featuring the sum count variable

## 3 Visualisation

### 3.1 Changing crime trends in Toronto over the years

A dataset featuring years and counts inherently presents the initial statistical inquiry into the temporal dynamics and trends over the observed period. Naturally, this begs the question “How have the counts of different crimes varied over the years?”. To begin formulating an answer, one would envision a scatter plot with Years plotted against the number of crimes and 4 lines for each of the Crime types. Using the ‘group by’ function in the dplyr package, data was grouped by the report year, and based on the type of crime. The ‘summarise’ function then added all counts for different age groups for the crime type per year, and the resulting table is shown below.

Report Year	Crime Type	Sum Count
2014	Assault	7132
2014	Other	2618
2014	Robbery	904
2014	Sexual Violation	1810
2015	Assault	7346
2015	Other	3008

Group by crime

List of 1

```
$ legend.position: chr "right"
- attr(*, "class")= chr [1:2] "theme" "gg"
- attr(*, "complete")= logi FALSE
- attr(*, "validate")= logi TRUE
```

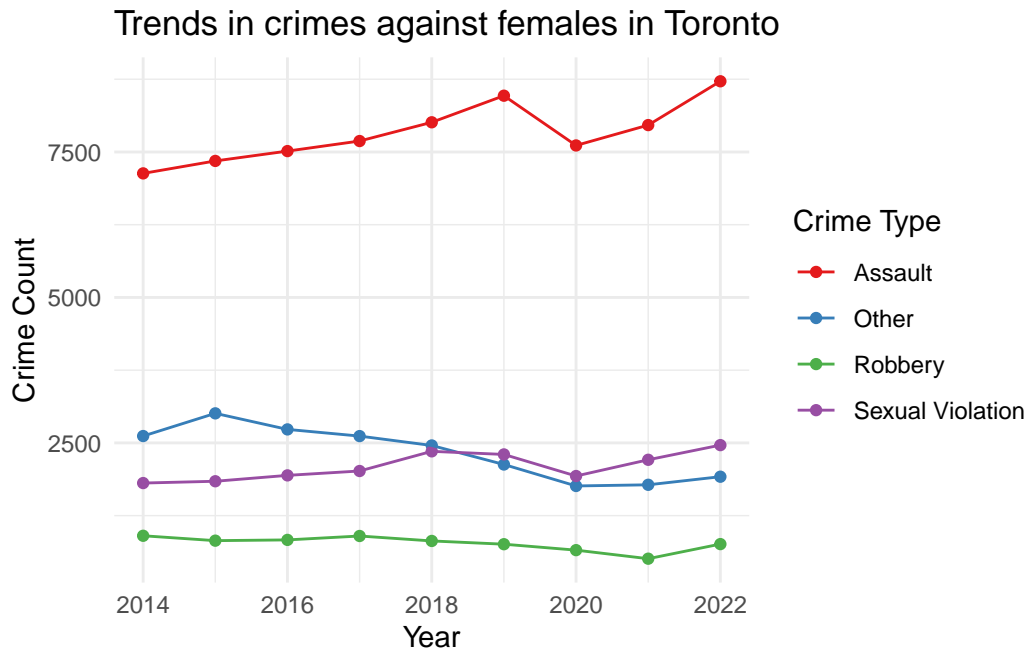


Figure 1: Changing crime trends over the years

### 3.1.1

### 3.1.2 Discussion

We run the model in R (R Core Team 2022) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

### 3.1.3 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance  $\theta$ .

## 4 Results

Our results are summarized in `?@tbl-modelresults`.

## **5 Discussion**

### **5.1 First discussion point**

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

### **5.2 Second discussion point**

### **5.3 Third discussion point**

### **5.4 Weaknesses and next steps**

Weaknesses and next steps should also be included.

## Appendix

### A Additional data details

### B Model details

#### B.1 Posterior predictive check

In `?@fig-ppcheckandposteriorvsprior-1` we implement a posterior predictive check. This shows...

In `?@fig-ppcheckandposteriorvsprior-2` we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected  
by, the data

Figure 2: `?(caption)`

#### B.2 Diagnostics

`?@fig-stanareyouokay-1` is a trace plot. It shows... This suggests...

`?@fig-stanareyouokay-2` is a Rhat plot. It shows... This suggests...

## emale age group is most susceptible to crimes in Toronto?

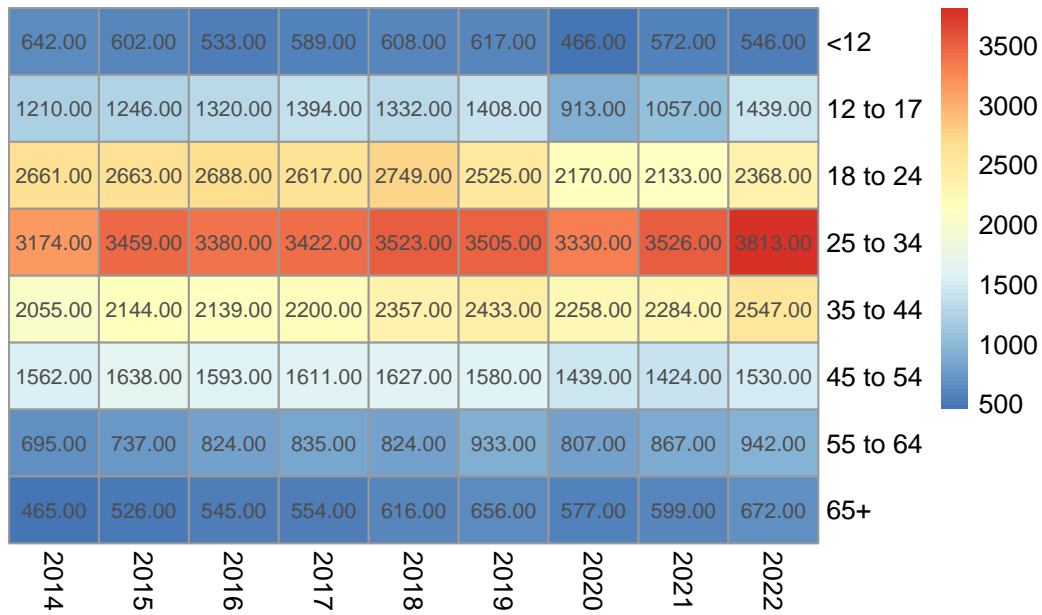


Figure 3: Relationship between wing length and width

## References

- Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. “Rstanarm: Bayesian Applied Regression Modeling via Stan.” <https://mc-stan.org/rstanarm/>.
- R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.