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

A visualization 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

*Code and data are available at: https://github.com/SEHB2012/crime-victims-toronto. A special thanks to my mom for fervently arguing that 45 -54 was just middle-aged, not middle-middle aged, and hence helping make the decision to retain age cohort labels

1 Introduction

Toronto is a scary city.As a women living alone, it helps to know with certainty on which birthday I may be attacked.

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(City of Toronto, 2014) 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 2022) was the language and environment used for the bulk of this analysis, alongside the dplyr (Wickham et al. (2023)),tidyverse(Wickham et al. 2019), janitor(Firke 2023), pheatmap(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 female victimisation rates 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

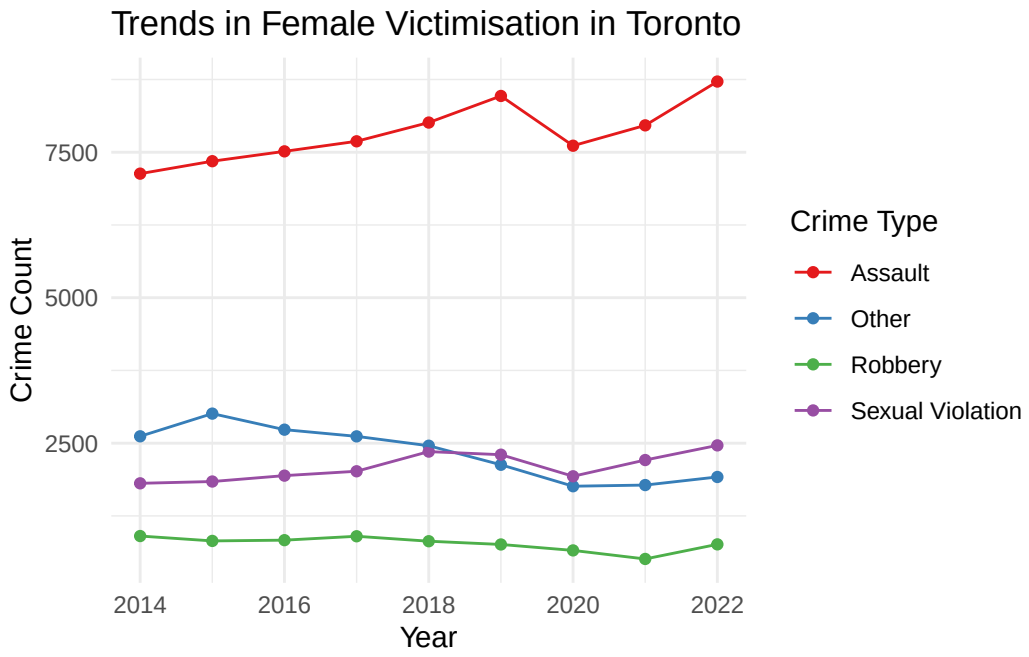

Figure 1: Changing crime trends over the years

Immediate observations indicate alarmingly high levels for assault compared to other types of crime, and this could be attributed to assault being a summed group including different assault subtypes. Assault has been growing steadily since 2014 and is at an all-time high in 2022, with a slight decline that can be attributed to Covid 19 restrictions imposed on the city in 2020.(cite) Sexual Violation and 'Other' crime numbers were dented in 2020 and saw a resurgence in the year after, but Robbery decreased in 2021, the second year of the pandemic. A reasonable assumption could be that the first year had stricter stay-at-home policy implementations, resulting in families at home, making it significantly challenging for break-ins to go unnoticed, thus dissuading robbers from attempting them. This dataset has a relatively short, 9 year range, and is perhaps too narrow to predict with accuracy any lasting trends for the decades to follow, but it presents fascinating, magnified data of how a worldwide pandemic impacts the number reported Crime victims.

3.2 Rate of female victimisation for various age groups in Toronto

The second line of questioning begins with trying to understand which female age groups are most at risk, and resolves the underlying personal motivation outlined before. Using group and summarise again, the data was grouped by report year and based on Age cohort, and summed as before, resulting in the following table.

Report Year	Age Cohort	Sum Count
2014	<12	642
2015	<12	602
2016	<12	533
2017	<12	589
2018	<12	608
2019	<12	617

Group by crime

Creating a heatmap was an inspired decision for this dataset, with the rationale being a visual that was inherently self-explanatory in how it represented trends and the relationships between the variables. The original dataset was transformed from its long format (1 observation per row, variables in different columns) to wide format (1 observation per row, all pertaining variables in that same row) by using the pivot_wider() function in the tidyr package. The pivoted data is presented below.

The dataset also needs to be converted to a numeric matrix before the heatmap can be created, and this step utilised the as.matrix() function to coerce all data types to numbers. The pheatmap package was installed and loaded(an appropriate abbreviation for what was indeed a 'pretty heatmap') to generate a visually appealing and customisable chart that allowed for clustering and custom labels.

Table 1: Explanatory models of flight time based on wing width and wing length

Age Cohort	2014	2015	2016	2017	2018	2019	2020	2021	2022
<12	642	602	533	589	608	617	466	572	546
12 to 17	1210	1246	1320	1394	1332	1408	913	1057	1439
18 to 24	2661	2663	2688	2617	2749	2525	2170	2133	2368
25 to 34	3174	3459	3380	3422	3523	3505	3330	3526	3813
35 to 44	2055	2144	2139	2200	2357	2433	2258	2284	2547
45 to 54	1562	1638	1593	1611	1627	1580	1439	1424	1530
55 to 64	695	737	824	835	824	933	807	867	942
65+	465	526	545	554	616	656	577	599	672

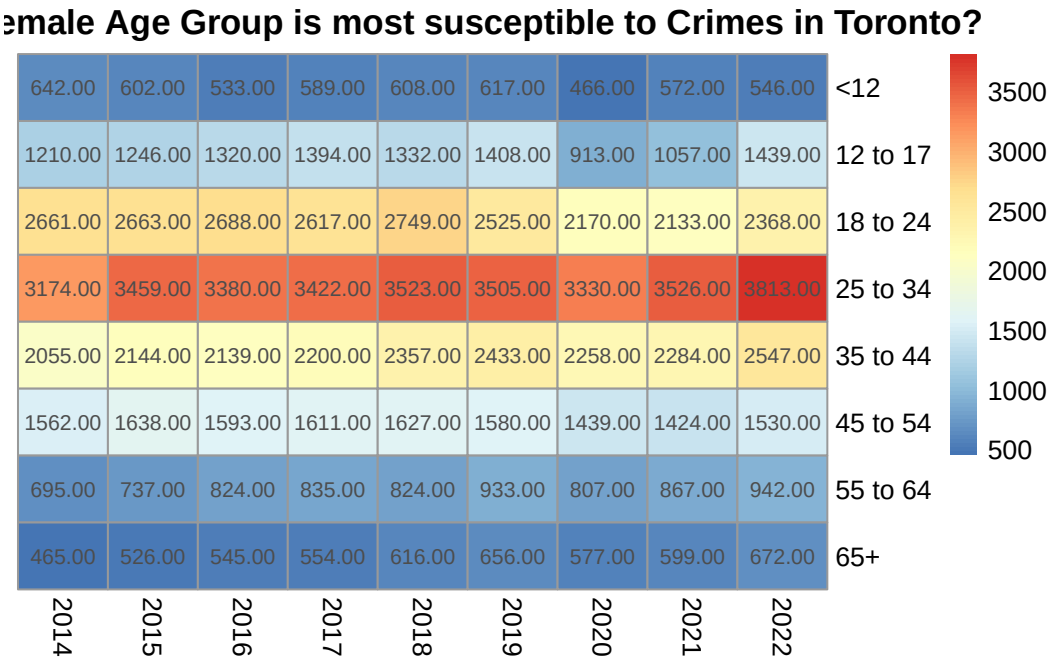

Figure 2: Relationship between wing length and width

Immediately noticeable is that women from the ages of 25 to 34 have been most susceptible to crimes regardless of year, with an all-time high in 2022. The other end of the spectrum is that children have never been safer than they were during COVID-19, staying at home. Moving vertically away from this high-risk group signals a reduced number of crime victims in both directions. This provides a possible inference that converging towards the transitional years that are 25 to 34 years of age, women are most in jeopardy, but is far from conclusive.

4 Discussion

4.1 Limitations

4.2 Future Work

References

City of Toronto,. 2014. "Open Data Dataset." *City of Toronto Open Data Portal*. Open data catalogue. https://open.toronto.ca/dataset/police-annual-statistical-report-victims-of-crime/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

R Core Team. 2022. *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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.