How Has The Number Of Shootings in Toronto Changed?*

An Analysis on the Implications of Dataset Bias

Fatimah Yunusa

January 23, 2024

Recently, Canada has been named the safest country in the world. A major component of Canada's safety comes from the safety of Toronto. Data detailing the number of shootings that have occured between the years of 2014 to 2019 have been obtained an will be thouroughly analysed within this paper through the use of various figures. Our findings show that in general, the number of shootings have increased in Toronto. From this paper, we learn that although Canada is considered a safer county overall, the number of shootings in Toronto have increased and we also learnt that datasets do have certain biases attached to them. This paper will also emphasise the discrepancies and biacies that exist within the crime reporting sphere.

1 Introduction

Over the years, having having data that describes and shows explicitly the number of shootings that have occurred has proven to be extremely valuable. It it provides information about how well a community is doing, law enforcement efficiency and the general well being of the area. In terms of Toronto, information about shooting occurrences will help shape policy and inform law makers and community shareholders where to focus their attention on. This paper focuses on shooting occurrences between 2014-2019 in Toronto and aims to analyse shooting occurrences and their geographical locations between this time period.

It is important to note that current literature on this issue does not go beyond the surface of the issue, but rather, it just focuses on the number and the general trend. This study aims to bridge the gap and provide a nuanced and detailed explanation as to why these trends occur. This paper is structured using the following headings:Data,Discussion & Conclusion. The Data

*Code and data are available at: https://github.com/fatimahsy/Shootings.git section focuses on providing a brief explanation of the data and how they were obtained from opendatatoronto (Gelfand 2022) including a brief discussion of the data cleaning process this section will focus on the trends found after performing statistical analysis. The discussion section will provide further insight and detail onto the results of the analysis and dicuss the implications of the data that is missing wthin this dataset. Lastly, the conclusion section will sum everything up and highlight the most important findings of this paper.

2 Data

Throughout this paper, we use data that has been obtained from the city of Toronto's opendatatoronto (Gelfand 2022). Mainly, we used the Police Annual Statistical Report - Shooting Occurrences, 2022 (Toronto 2022). To collect and analyse this data, we use the Statistical program R (R Core Team 2022) and additional packages such as: tidyverse (Wickham et al. 2019),ggplot2 (Wickham 2016), knitr (Xie 2014), janitor (Firke 2023),readr (Wickham, Hester, and Bryan 2024), stringr (Wickham 2022), here (Müller 2020), kableExtra (Zhu 2021),lubridate (Grolemund and Wickham 2011), dplyr (Wickham et al. 2023), and finally tibble (Müller and Wickham 2023).

The report Police Annual Statistical Report - Shooting Occurrences, 2022 (Toronto 2022) contains data on the number of shootings in Toronto from 2014-2019.The main data features and variables of the data set we will be focusing on are: OccurredYear, GeoDivision, & count. In this data set with 96 observations, they categorize a shooting occurrences any incident which a projectile is discharged from a firearm Police Annual Statistical Report - Shooting Occurrences, 2022 (Toronto 2022). This data is collected by police all around the city. They follow strict reporting rules and ensure their accounts and data collection is accurate.

In spite of this, the data consists of many issues which are deep rooted in the system. The data might be accurate in terms of collection but the issue with these reported shootings comes from many socioeconomic and institutional biases that have existed within the system.If this data was grouped according to race, it would tell a completely different story because colored people are more likely t be stopped by the the police. The implications of this act morose like selection bias where your results are more likely to be one thing as a result of respondents being of a particular group or having similar traits.This means that our data even though is accurate, stems from many biases.It is important to note that even with a biased data set, we are able to deduce many things from it.

This data includes six variables and 96 observations. the first two variables _id and Index are identifier variables. the third variable OccuredYear is the year the reported shooting took place and for our data it is between 2014 - 2019. The next variable GeoDivision is predetermined geographic zone of the police division that the shooting was reported in.The next variable Category represents what kind of crime occurred and for our analysis purposes, this variable is always a shooting occurrence.The last variable is count which counts the number of shootings in that year and division. A sample view of their data set we used is include below.

```
# A tibble: 6 x 6
  `_id` Index_ OccurredYear GeoDivision Category            Count_
  <int>  <int>        <int> <chr>       <chr>                <int>
1     1      1          2014 D11         Shooting Occurrence      2
2     2      2          2014 D12         Shooting Occurrence     20
3     3      3          2014 D13         Shooting Occurrence      2
4     4      4          2014 D14         Shooting Occurrence      7
5     5      5          2014 D22         Shooting Occurrence      8
6     6      6          2014 D23         Shooting Occurrence     18
```

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://cran.r-project.org/web/packages/opendatatoronto/index.html.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto, Open Data. 2022. *Police Annual Statistical Report - Shooting Occurrences*. https://open.toronto.ca/dataset/police-annual-statistical-report-shooting-occurrences/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

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://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.