An Analysis on the Crime Counts in York University Heights from 2018 to 2020*

Using Data Extracted from Open Data Toronto

Yujun Jiao

4/27/2022

Abstract

Has the number of crime cases increased in the past few years as a result of many influential global events? The dataset used in this report is extracted from Open Data Toronto, and it contains the number of different types of crimes that have taken place in Toronto from 2014 to 2020. Even though the datasets contain information for 140 neighborhoods, this report focuses only on the York University Heights area, which is famous for its large proportion of the young population(n.d.). Also, the analysis will study only the years 2018, 2019 and 2020. Using R studio (R Core Team 2020), this report will explore the trends in the data by building models and graphs.

Contents

1 Introduction **1**

2 Data **2**
- 2.1 Variables and Data Methodology 2
- 2.2 Combining Variables 2

3 Variables: Types of Crimes **2**

4 Results **3**

5 Discussion **9**
- 5.1 What is done in this report? 9
- 5.2 What is something we learn about the world? 9
- 5.3 Weakness 9
- 5.4 Future directions 9

6 Appendix **10**

References **10**

1 Introduction

This dataset is extracted from Open Data Toronto, and it contains data for all types of crimes such as assault, break-enter, robbery, homicide and shooting. Due to the raging of COVID-19 and other unexpected global events, many things have changed in the past couple of years. The lives of ordinary people were altered, and some of them might have chosen the wrong path. Thus, this report aims to discuss the counts of these crimes

*https://github.com/bellajiao1999/Toronto_Crimes in the past few years in York University Heights. In hopes of discovering the obvious trends in the count of crime

The factors, such as the pandemic, economy, parades, environment and policy changes, will also be considered and discussed. This report is extremely important as it outlines the factors that influence the crime rates. Secondly, the graphs, models and results of this analysis can also be used to predict future trends. Thus, the government and the police can be better prepared for future plans to maintain a peaceful, stable and prosperous community.

This report is divided into three sections: data, results and discussion. The data section will explore all the variables used in this analysis. The result section will analyze the models that examine the correlations between the variables using R (R Core Team 2020). The discussion will discuss what has been done in this paper, the weakness of the report, and future directions. The code and data that support this analysis can be found in the Github repository: Toronto_crimes.

2 Data

2.1 Variables and Data Methodology

The data used in this report is extracted from Open Data Toronto Portal("Open Data Dataset," n.d.). It contains both the count and the rate for 8 types of crimes that happened in 140 neighborhoods in Toronto. As we only want to study the count of six crimes, we're only analyzing the number of assault, break and enter, auto theft, robbery, shooting, homicide. Besides, the data set also outlines the years in which each crime has taken place. Since we're only focusing on the trends in the crimes in recent years, we're only studying the data from 2018, 2019 and 2020. Lastly, as the data set contains data for 140 neighborhoods, this analysis will only focus on the crime that happens in the York University Heights region.

This data was directly provided and published by Toronto police Services, and it was licensed by the Open Government. Hence, it was collected by the police services and it's based on real-life cases. The rate of the crime was calculated by dividing the total number of crimes by 100,000, which is the population estimate that is provided by Environics Analytics.("Open Data Dataset," n.d.)

2.2 Combining Variables

In the original data set, each type of crime of a particular year occupies a whole column, and is recorded as a variable by itself. For example, the counts of robbery in 2018 serves as a variable, and it's named as "Robbery_2018" in the original data set. During the research stage of the project, our team has realized that this way of organizing variables is not very proficient in the later analyzing and visualizing processes. Hence, in order to facilitate the data analysis process, we have decided to combine all the years of one particular type of crime into one variable. In addition, a new variable called "year" was also created. By doing so, the variables are more organized and easier to graph.

3 Variables: Types of Crimes

In this report, we're interested to see if the criminal cases have changed in the last couple of years. Thus, we have cleaned the original data set by extracting the variables that are relevant to our research question. To help the audience to better comprehend the data, bar charts were made to show the trends that exist within the data.

Figure 1 shows the number of assault cases from 2018 to 2020. The year in which the crime was committed was labeled on the x-axis and the count of crime cases was labeled on the y-axis. In Figure 1, there's no obvious trend in the number of cases, as the count for assault cases was highest in 2019. As we know, many social changes such as the pandemic and elections have occurred in 2020, and 2019 was regarded as a peaceful year in comparison to the later years. Thus, the graph for assault is not coherent to our assumption.

Secondly, in Figure 2, the number has actually decreased from 2018 to 2019. The number of people who broke and entered other people's houses was lowest in 2020. This is also incoherent to our expectation as we believe that this number would increase in 2020 as people are afraid to stay outside due to the raging pandemic. Next, Figure 3 shows a similar pattern to the Figure 2. The number of robbery cases has decreased since 2018, and was lowest in 2020.

In contrast, the auto theft cases shown in Figure 4 have demonstrated an increasing pattern in the count of cases. The number of reported cases of auto theft has shown a proportional relationship with time, and the number has changed significantly from 91 to 184. Similarly, the high number of crime cases in 2020 in Figure 5. The number of shooting cases has increased from 6 to 12 from 2019 to 2020. Obviously, the shooting cases have increased immensely in the last couple of years. Finally, the number of homicide cases that happened within the York University Heights has not shown useful results for our research question. As homicide is a very serious crime, it happens rarely, and there was only 1 case in 2018, 0 case in 2019 and 1 case in 2020.

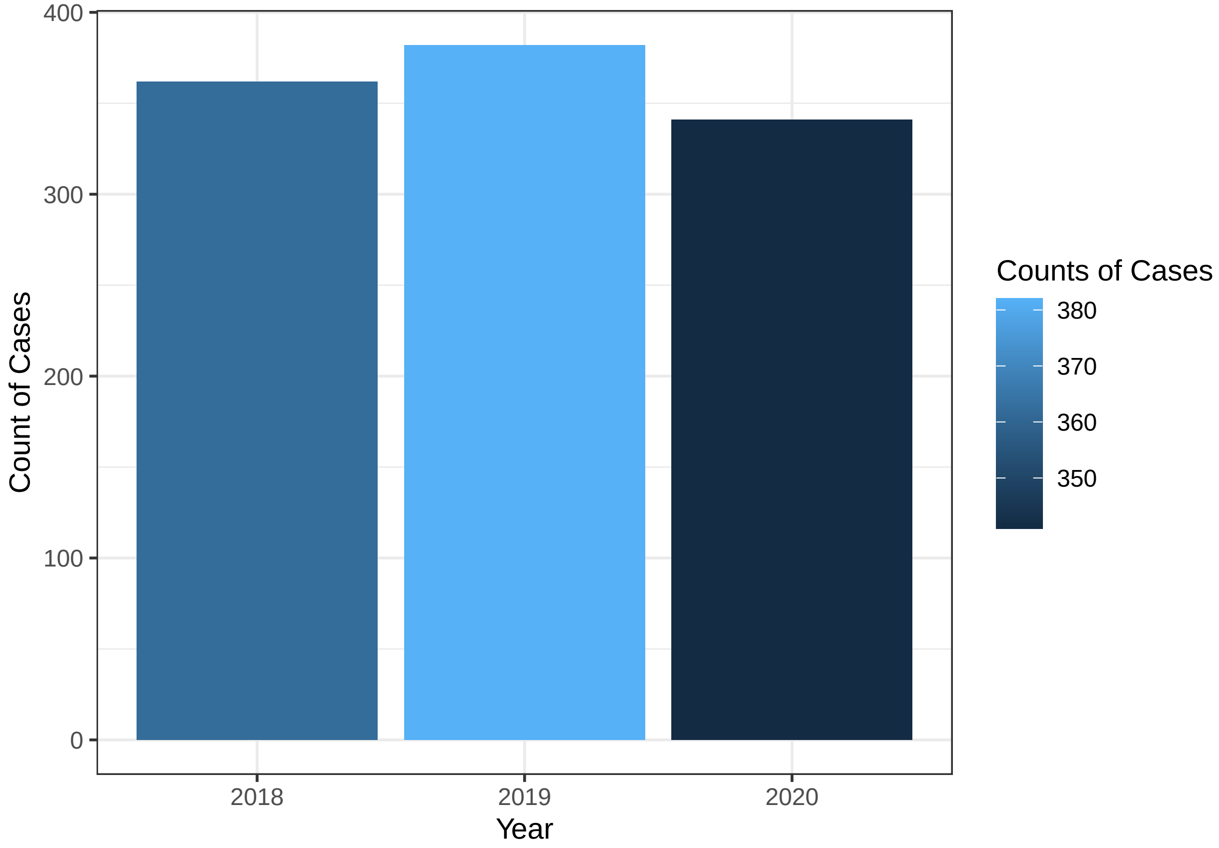

Figure 1: Count of Assaults from 2018 to 2020

4 Results

In the previous section, plots were made to show the correlation between the number of criminal cases and year. In conclusion, assault and homicide didn't show any trend in the numbers; the break entry and robbery cases have decreased as time passed by, and auto theft and shooting cases have increased in the most recent years. Unfortunately, with the current data, it's difficult to make solid assumptions about the relationship between time and the crime counts.

We have decided to build linear regression models as a way to monitor the relationship between these variables. Hence, we have chosen the count of each type of crime as our response variable, and have decided to use

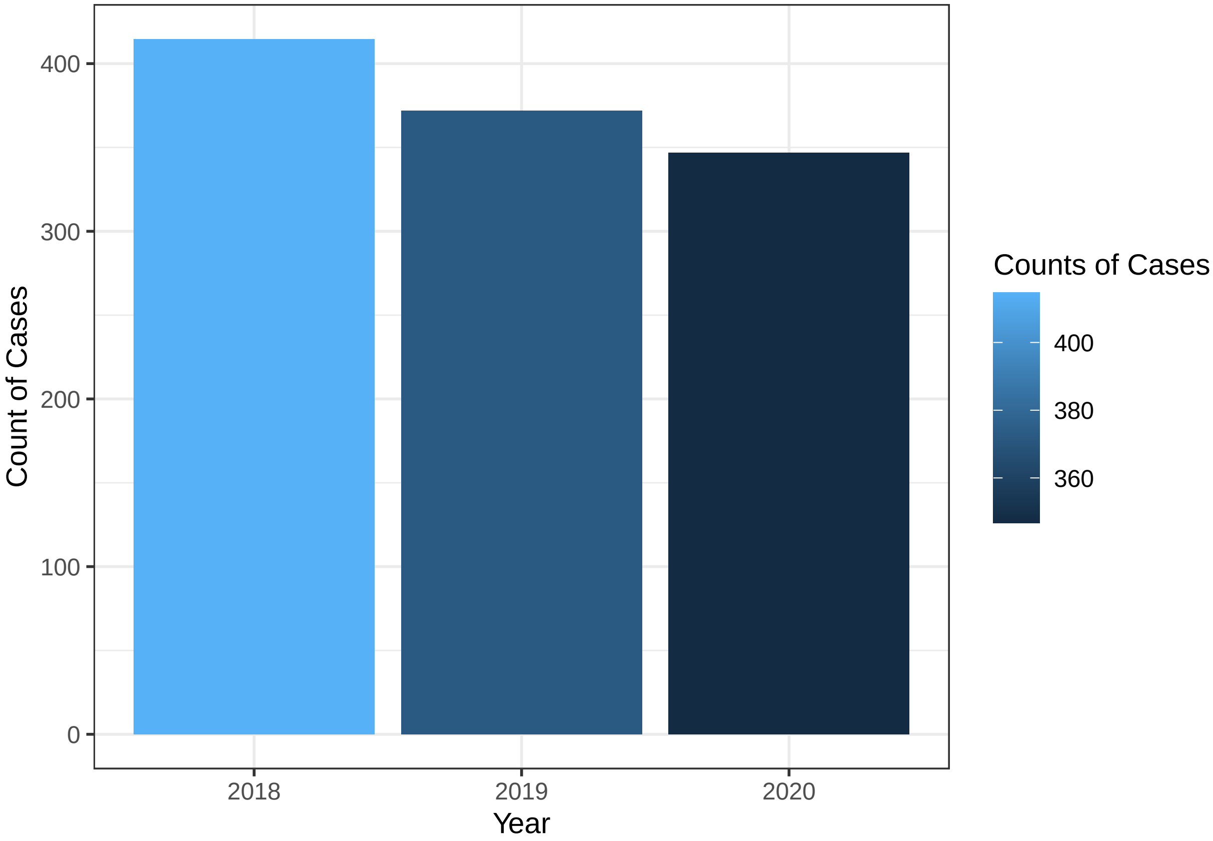

Figure 2: Count of Break & Enter cases from 2018 to 2020

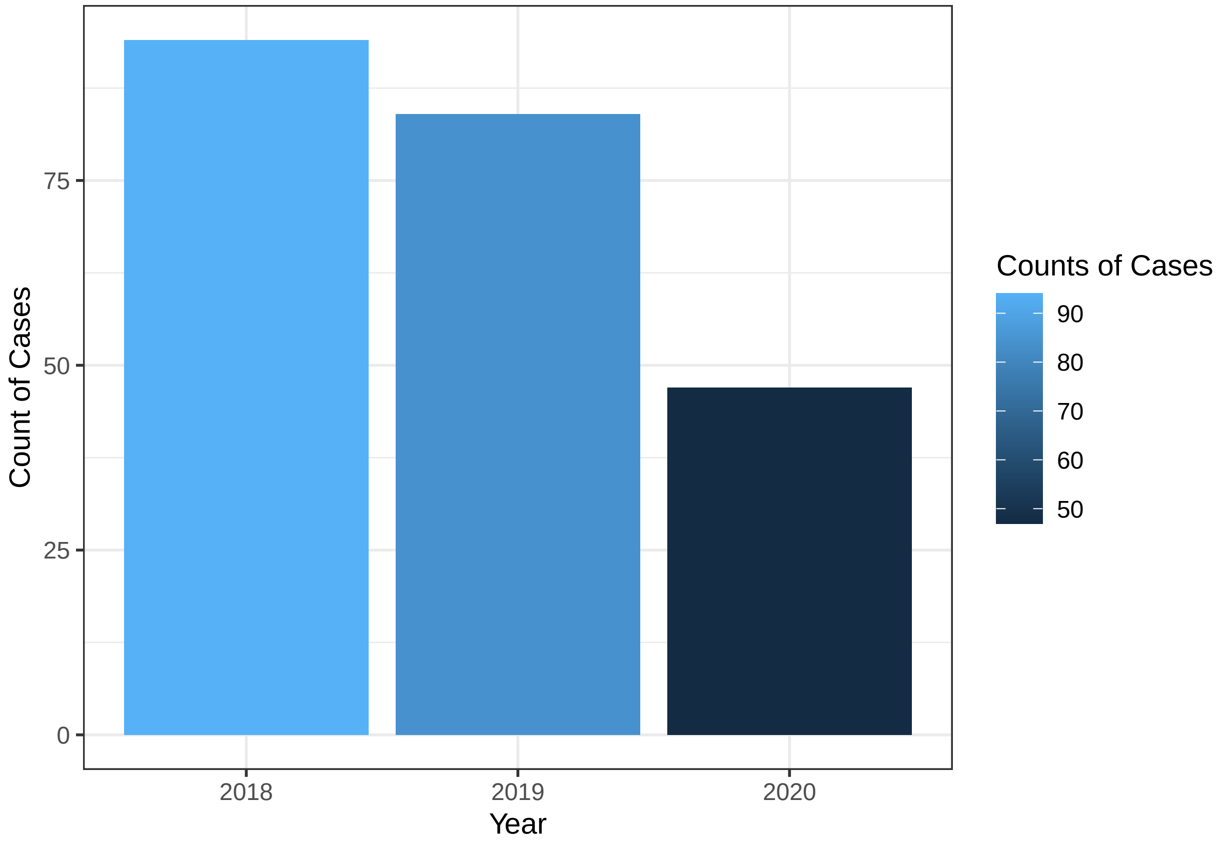

Figure 3: Count of Break & Enter cases from 2018 to 2020

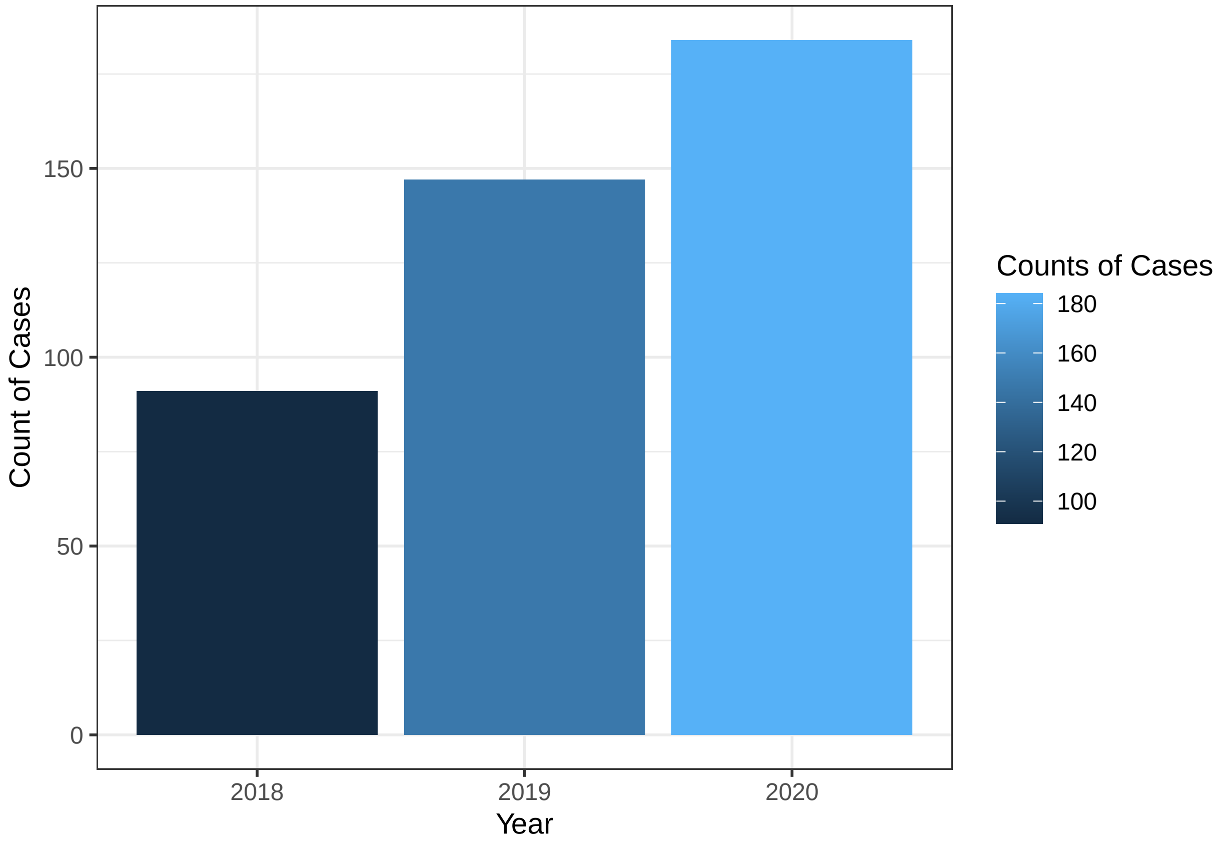

Figure 4: Count of Auto Thefts cases from 2018 to 2020

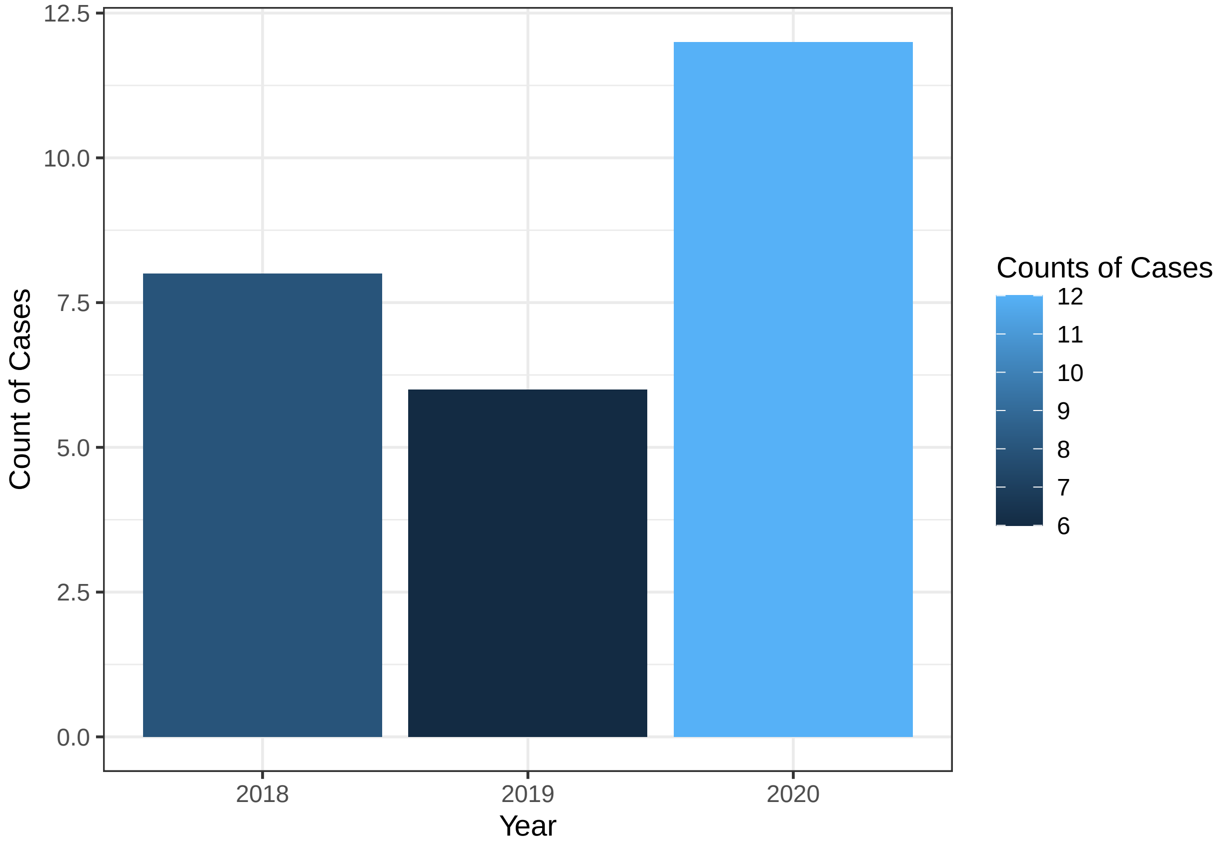

Figure 5: Count of shooting cases from 2018 to 2020

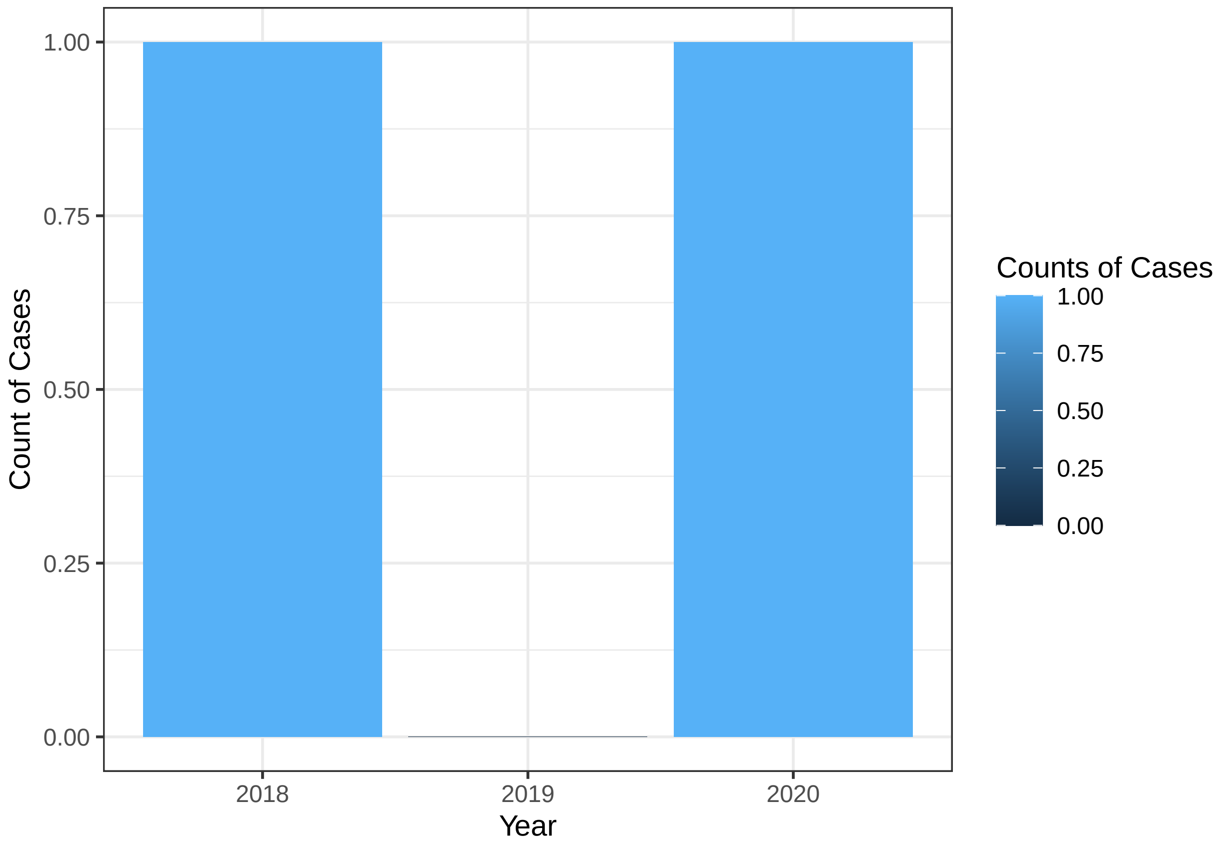

Figure 6: Count of homicide cases from 2018 to 2020

“year” as our predictor variables. For example, for the robbery model, the formula of the model is: Robbery = a + b(year). Robbery is the response variable, year is the predictor variable, and a and b are intercept and slope. Thus, we have come up with six models for six dependent variables. However, it’s very difficult to make models with only one independent variable each time, so the p-value for each model was very high. For assault, the p-value was 0.658 and for shooting, it was 0.546. The p-value was smallest for break entry, which was 0.0949. We know that if the p-value is higher than 0.05, then it would be unwise to state that the factors have a significant impact on the response variable. Thus, for this report, we claim that the variables used in the model are not significant to the year. In order words, there’s no evidence that the number of crimes has increased with time.

5 Discussion

5.1 What is done in this report?

The original dataset published on Open Data Toronto Portal was very lengthy and contained a lot of information. This paper has cleaned the old data set and created a new data set that is subject to the research question. In order to obtain an answer to this question, research has been made to acknowledge the background of York University Heights. After a series of analyses, the trends in the number of six types of crimes such as assault, break entry, auto theft, shooting and homicide were plotted and studied. In this report, we came to the conclusion that some

5.2 What is something we learn about the world?

We have learned a lot about the world by conducting this research project. Most importantly, we now realize that society is considerably sensitive to influential events, policy changes, and other unexpected factors. Economy, politics, public health and other factors may easily affect the stability of the society and result in severe consequences. In fact, the crime rate is not the only consequence of an unstable society. If the impact of a certain event is not represented in the number of crime cases, it may be shown in other factors such as mortality rate, health screening surveys, financial conditions, etc. Thus, it’s interesting to see that for some types of crimes, the numbers have or have not changed significantly in the past few years.

5.3 Weakness

There’re many weaknesses and flaws in this report. First of all, this report was limited to only one neighborhood in the whole Toronto region. Even though North York Heights is highly populated and contains a lot of criminal cases, it is still not sufficient for us to make confident answers to the question. Also, sampling and non-sampling bias may also exist in the data set. As the police services can only record the cases that have been reported to the police station, those crimes that were kept secret and not reported are not included in this data. It’s very important for the readers of this report to acknowledge the flaws and weaknesses of the data set and the report, as they may potentially affect the clarity, accuracy and correctness of the statements that are being made.

5.4 Future directions

In order to better comprehend the impacts and consequences of large-scale global events such as political changes, pandemics and other social events, more reports should be made to support further analysis and conclusions. More precisely, our team aims to do more analysis on other neighborhoods within or outside the Toronto areas. As different neighborhoods symbolize different population density, percentages of young population and income levels, it is essential to study all the neighborhoods as a whole to better explore the consequences of the global impacts. Also, it’s also interesting to dig deeper into the factors that have affected the crime rates. Thus, in future work, our team will look at the actual factors that have played important roles in the criminals’ lives that have caused them to choose the wrong path.

6 Appendix

References

Iannone, Richard, Joe Cheng, and Barret Schloerke. 2022. *Gt: Easily Create Presentation-Ready Display Tables*.

"Open Data Dataset." n.d. *City of Toronto Open Data Portal*. https://open.toronto.ca/dataset/neighbourhood-crime-rates/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

n.d. https://www.toronto.ca/ext/sdfa/Neighbourhood%20Profiles/pdf/2016/pdf1/cpa26.pdf.