Predicting Future Crimes in Toronto*

A Statistical Analysis of Crime Indicators

Yiyi Yao

September 23, 2024

This paper analyzes crime data from Toronto to predict future crime trends using a linear regression model. We found that certain crime categories have fluctuated significantly over time, with some increasing and others decreasing. By understanding these trends, we can make more accurate predictions about future crime rates and guide law enforcement efforts more effectively. This is important because it allows for better resource allocation and proactive crime prevention, ultimately helping to improve public safety in Toronto.

1 Introduction

Crime is a major concern for urban societies, with cities like Toronto constantly seeking ways to understand and reduce criminal activity. As crime patterns fluctuate over time, it becomes critical for law enforcement and policymakers to track these changes, predict future trends, and allocate resources efficiently. Understanding historical crime data is essential for spotting trends and identifying emerging challenges. The use of data analytics and predictive models offers a powerful tool to assess the factors contributing to crime rates and to provide insights that can guide effective strategies to improve public safety.

This paper focuses on analyzing the reported crime data from Toronto, using statistical methods to predict future crime trends. The dataset includes reported crimes across multiple categories over a range of years, making it a rich source for exploring how crime patterns evolve. By building a predictive model using linear regression, we aim to estimate future crime rates based on historical data. The key variables considered in this analysis include the year of reporting, crime categories, and geographic divisions within Toronto. The model provides a clearer picture of how crime has changed over time and which crime categories are likely to grow or shrink in the future.

[*] Code and data are available at: https://github.com/Yaoee111/Toronto-crimes. DOI: DOI 10.17604/OSF.IO/JS83Y.

A gap in the current approach to crime prediction is that many models overlook the combination of time-based trends with specific crime categories. Furthermore, existing studies often lack a comprehensive focus on crime trends specific to urban centers like Toronto, which face unique challenges. This paper addresses this gap by developing a model that takes into account both temporal changes and crime type, providing more accurate insights into future crime patterns.

Our analysis reveals important findings about how crime in Toronto has changed over time, particularly the rise and fall of certain crime types. The linear regression model shows clear trends and provides predictive power for future crime rates. This analysis is crucial because it gives law enforcement agencies and policymakers data-driven insights that can help shape crime prevention strategies and allocate resources more effectively. By understanding these trends, we contribute to a proactive approach in ensuring public safety and reducing criminal activity in Toronto.

The rest of the paper is organized as follows: The Section 2 describes the data, variables, and methods used in the study and the rationale for the selection of this dataset. The data is presented through graphs. The Section 3 describes how to fit a linear regression model. The Section 4 presents the results of the analysis. The Section 5 provides an in-depth discussion of our findings and reflections on the research process. Finally, the Section A adds details of the data.

2 Data

2.1 Overview

This dataset is from Open Data Toronto (Gelfand 2020). It includes all reported crime incidents, organized by the date they were reported and grouped by police division. It covers all crimes reported to the Toronto Police Service, even those later found to be unfounded, those happening outside Toronto, or those without a confirmed location.

In line with the Municipal Freedom of Information and Protection of Privacy Act, the Toronto Police Service has ensured that the privacy of individuals involved is protected. No personal details about anyone involved will be shared as part of the open data. The information is summarized by year, crime category, subtype, and division.

If an incident involves more than one type of offence, it will appear in multiple categories. The numbers shown do not represent the total number of unique incidents.

2.2 Data Management

I use the statistical programming language `R` (R Core Team 2023). In the data analysis and visualization process, I also made use of the following `R` packages: `readr` (Wickham et al. 2024), `dplyr` (Wickham et al. 2014), `knitr` (Xie 2018), `kableExtra` (Zhu et al. 2019), `ggplot2` (Wilkinson 2011), `tibble` (Müller and Wickham 2023).

2.3 Variable

This dataset contains detailed variables for evaluating each property. I will focus on some most vital variables. Details are in (Table 1)

Table 1: Key Variables in the Toronto Crimes Dataset

Variable	Description
_id	Unique row identifier for Open Data database
REPORT_YEAR	Year crime was reported
DIVISION	Geographic division where crime took place
CATEGORY	Crime category
SUBTYPE	Crime category subtype
COUNT_	Total number of crimes

2.4 Data cleaning and preparation

The data cleaning and preparation process involved several important steps to ensure the dataset was accurate and ready for analysis. First, missing data was addressed by either removing incomplete records or imputing values where appropriate. Unnecessary columns were removed to simplify the dataset and focus on key variables. The data was then aggregated to consolidate crime counts for similar incidents. These steps ensured that the dataset was clean, organized, and ready for further analysis.

More details about data cleaning are included in Appendix B.

2.5 Rationale for Selection

This dataset was chosen because it offers a complete record of reported crimes in Toronto, covering several years. It includes key variables like `REPORT_YEAR`, `CATEGORY`, and `SUBTYPE`, which make it suitable for statistical analysis and prediction. The dataset covers a wide range of crime types, allowing for a balanced view of both property and personal crimes. Its detailed structure, with data grouped by police division and crime type, provides useful insights for analyzing crime trends and forecasting future crime rates. This makes it a strong foundation for understanding and addressing crime patterns in Toronto.

2.6 Limitation

One limitation of this dataset is the presence of "No Specified Address" (NSA) entries, which include crimes reported outside Toronto or incidents without a verified location. Additionally, the dataset includes reported crimes that were later deemed unfounded, which may not accurately reflect actual crime occurrences. The data is filtered by the year the crime was reported, not necessarily the year it occurred, which could affect trend analysis. Finally, incidents categorized under "Crimes Against the Person" do not include cases where the victim's name is missing, which could leave out important data.

2.7 Plots

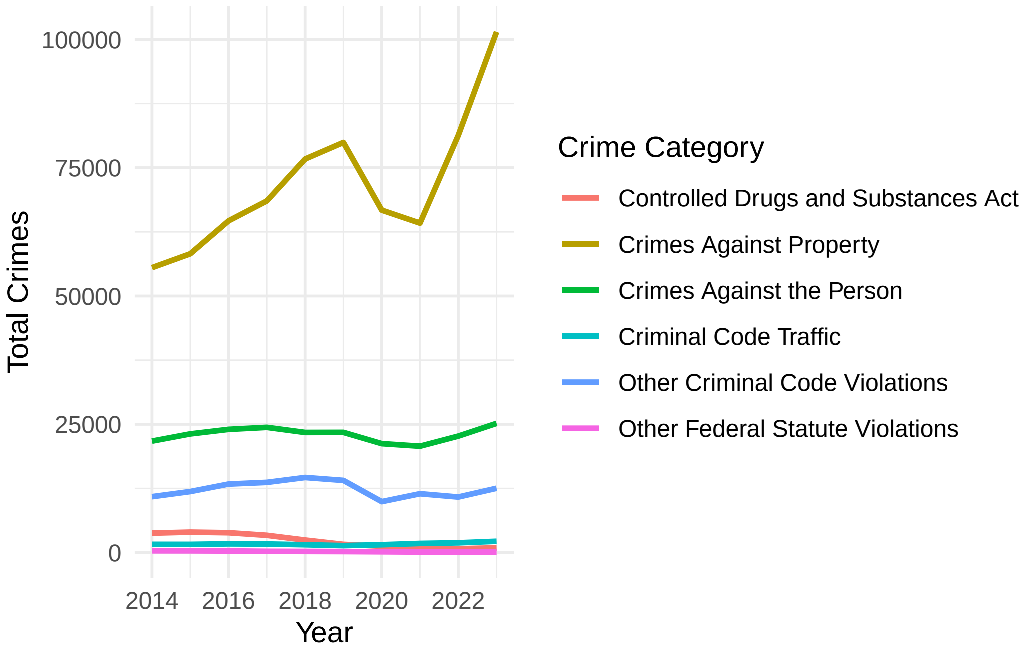

Figure 1: Crime Trends in Toronto by Year and Category

Figure 1 shows the crime trends in Toronto over time, broken down by different crime categories. The x-axis represents the `REPORT_YEAR`, indicating the year in which crimes were reported. The y-axis represents the total number of crimes in each year. Each line in the plot represents a different crime category, with colors used to differentiate between these categories.

The line plot allows us to easily compare how the volume of reported crimes has changed over time within each category. For instance, we can observe whether certain types of crimes, such as property-related crimes or personal offences, have increased or decreased in frequency. The visualization provides a clear overview of the year-to-year fluctuations in crime, making it easier to identify any trends, spikes, or declines in specific types of criminal activity across Toronto.

This information could be useful for understanding how crime patterns have evolved and for predicting future crime trends.

3 Model

The goal of my modelling is twofold. Firstly, I want to Explore the relevance of various variables to crime indicates. Secondly, I want to fit a linear regression model to predict the crimes in Toronto.

3.1 Model set-up

To predict crimes, I will fit a linear regression model. The general form of a linear regression model is:

$$\text{Total Crimes} = \beta_0 + \beta_1 \cdot \text{Year} + \beta_2 \cdot \text{Category}_1 + \beta_3 \cdot \text{Category}_2 + ... + \beta_n \cdot \text{Category}_n + \epsilon$$

where: - Total Crimes is the dependent variable (Y), which represents the total number of crimes for a given year and category. - β_0 is the intercept of the model, representing the expected value of total crimes when all independent variables are zero. - $\beta_1 \cdot \text{Year}$ represents the effect of the year on the total number of crimes. The coefficient β_1 indicates how much the total number of crimes changes for each additional year. - $\beta_2 \cdot \text{Category}_1 + \beta_3 \cdot \text{Category}_2 + ... + \beta_n \cdot \text{Category}_n$ represent the effect of different crime categories (such as property crime, violent crime, etc.) on the total number of crimes. Each β_n coefficient shows how the total number of crimes is expected to change when that specific category is present. - ϵ is the error term, representing the difference between the observed and predicted total number of crimes.

3.2 Model summary

The multiple linear regression model aims to predict the total number of crimes based on the year and crime categories. Each coefficient in the model represents the impact of a particular variable on the total crime count. For example, the coefficient for the year (β_1) shows how crime changes over time, with a positive value indicating an increase in crime and a negative value indicating a decrease. The coefficients for crime categories (β_2, β_3, ..., β_n) reflect how each type of crime contributes to the overall crime rate. The model also includes an error term (ϵ), which accounts for the difference between the actual crime count and the predicted value. We expect the model to show how well the year and crime types explain changes in total crime over time. A good fit will indicate that these factors significantly influence crime trends, allowing us to predict future crime rates based on historical data. This model can help in understanding key crime patterns and can be useful for public safety planning and resource allocation.

4 Results

4.1 visualize relationship between variables

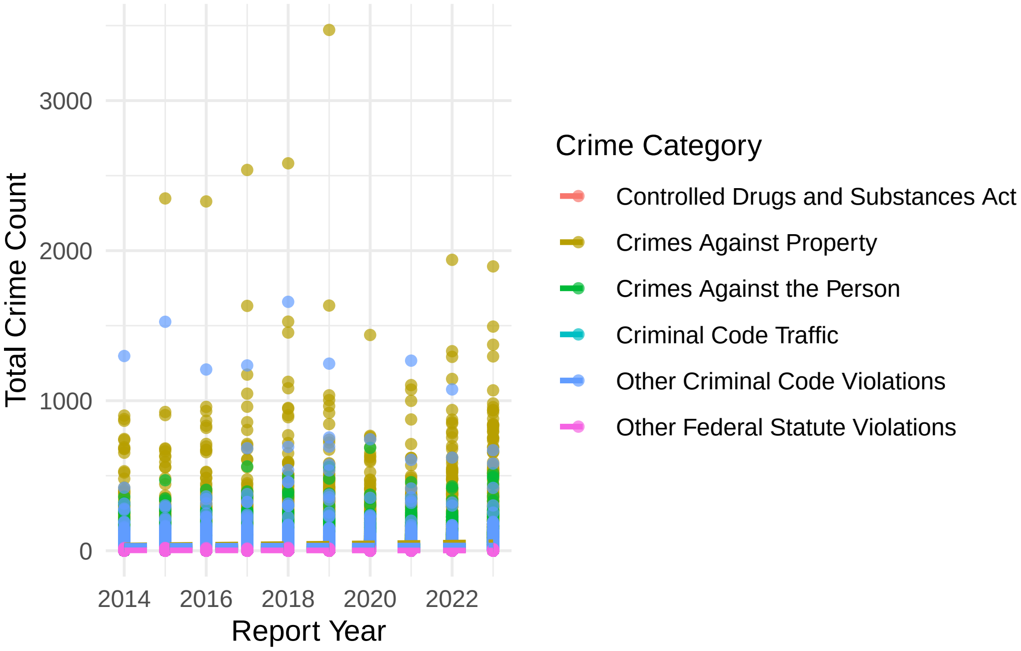

Figure 2: Relationship between Report Year and Crime Count by Category

The scatter plot (Figure 2) illustrates the relationship between the REPORT_YEAR and the total crime count (COUNT_), with different colors representing various crime categories. Each point on the plot represents the total number of crimes reported for a specific year and category. A linear trend line is added to each category to show the general direction of the relationship between the year and crime counts. This plot helps to identify whether crime rates for different categories are increasing or decreasing over time. For example, a positive slope on the trend line indicates that crimes in that category have been increasing, while a negative slope suggests a decrease. This visualization provides a clear overview of how crime patterns have evolved in Toronto across different categories over the years.

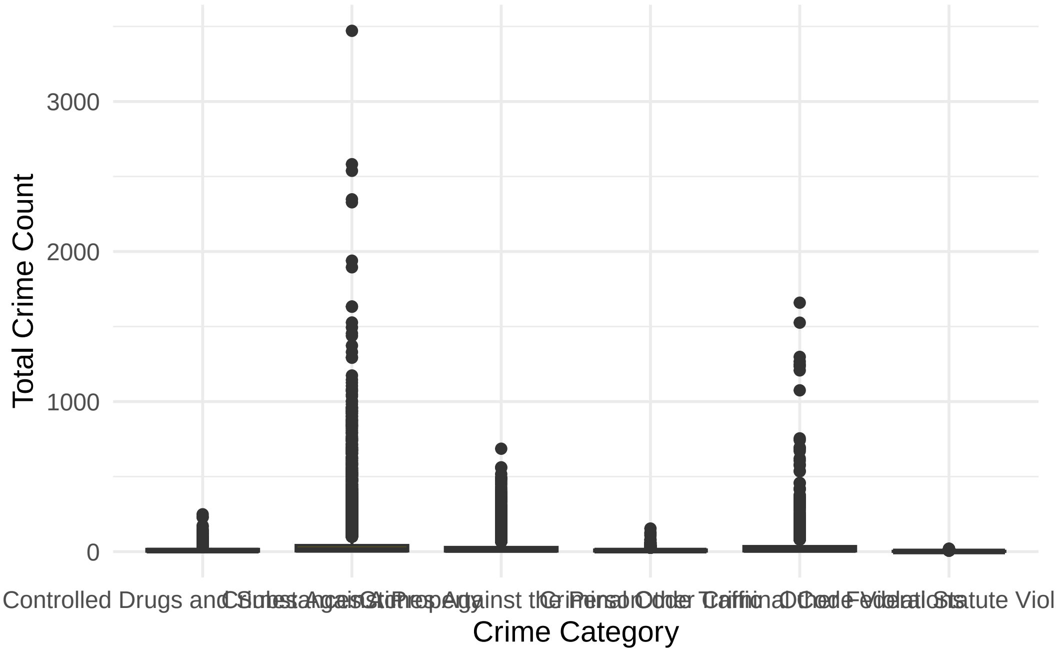

Figure 3: Crime Count Distribution by Category

The box plot (Figure 3) compares the distribution of crime counts across different categories. Each box represents the spread of total crime counts within a specific category, showing the median, interquartile range (IQR), and any potential outliers. The width of the box indicates the concentration of data points, and whiskers extend to the smallest and largest values within 1.5 times the IQR. This plot makes it easy to compare how different categories vary in terms of crime counts. Categories with taller boxes or wider IQRs show greater variability in crime counts, while more compact boxes suggest a more consistent pattern of crime. The plot also highlights any outliers, which could indicate unusual spikes or drops in crime within certain categories.

4.2 Visualize linear regression model

The plot (Figure 4) visualizes the linear regression model that predicts the total number of crimes in Toronto over time based on the `REPORT_YEAR`. The blue points represent the actual crime counts for each year, showing how the number of reported crimes has varied over time. The red line represents the fitted linear regression model, which captures the overall trend in crime rates as time progresses. The shaded area around the line indicates the confidence interval, which shows the uncertainty around the predicted values.

If the red line has a positive slope, it suggests that crime rates have increased over time. A negative slope would indicate a decrease in crime rates. The confidence interval helps highlight

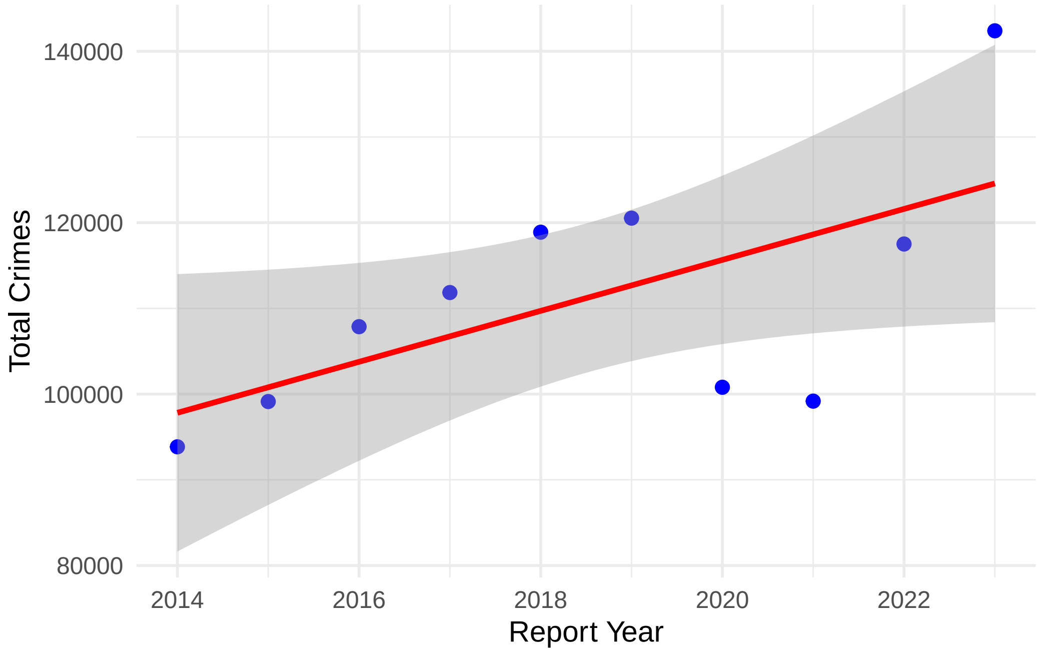

Figure 4: Linear Regression of Total Crimes Over Time

how confident we are in the predictions made by the model, with wider shaded areas indicating more uncertainty. This visualization provides a clear understanding of how crime has changed over the years and how well the linear model fits the observed data.

5 Discussion

5.1 Things we have done in this paper

In this paper, we analyzed the reported crime data in Toronto to predict future crime trends. We began by selecting and cleaning the dataset, removing unnecessary columns and handling missing or duplicate entries. We then aggregated the data by key variables, such as `REPORT_YEAR`, `CATEGORY`, and `COUNT_`, to prepare it for analysis.

Next, we explored the data visually, generating scatter and box plots to examine the relationship between crime counts and other variables. We fitted a linear regression model to predict total crime counts over time, using `REPORT_YEAR` as the primary predictor. The regression model was visualized to show the relationship between the actual crime counts and the predicted values.

Throughout the paper, we interpreted the results of our analysis, identifying key trends and discussing how crime patterns have evolved in Toronto. We also assessed the performance of our regression model and visualized the fit with a confidence interval to understand the prediction uncertainty.

These steps have provided a comprehensive understanding of the crime data and helped in building a model that can be used to predict future crime trends in Toronto.

5.2 Something we learn about the world

From this analysis, we gained insights into how crime trends in Toronto have evolved over time and how statistical models can be applied to predict future crime rates. One key takeaway is that crime patterns can show both gradual and sudden shifts, which may be influenced by various factors such as social, economic, or policy changes. By studying how different categories of crime have fluctuated, we can better understand which areas of law enforcement may need more attention or resources.

We also learned that linear regression, while a powerful tool, has limitations when it comes to capturing complex crime patterns. Crime rates are influenced by many factors beyond time, such as economic conditions, population changes, and local law enforcement policies. Although our model shows a general trend, it highlights the need for more sophisticated approaches to fully capture and predict crime dynamics.

This analysis underscores the importance of data-driven approaches to public safety, showing how predictive models can help inform decisions about where to focus resources and efforts in the fight against crime. By leveraging past data, we can make more informed predictions about future trends and improve community safety planning.

5.3 Weaknesses

While this analysis provided valuable insights, there are several limitations that need to be addressed. First, the linear regression model we used is fairly simplistic and assumes a linear relationship between crime rates and time. In reality, crime trends are influenced by many complex factors, such as economic conditions, demographic changes, and law enforcement policies, which were not included in this model. As a result, the model may oversimplify the patterns and miss key drivers of crime.

Additionally, the dataset includes crimes that were later deemed unfounded, which may skew the analysis. Furthermore, the data is limited to reported crimes, meaning that unreported crimes, which may represent a significant portion of overall crime, are not accounted for. Another limitation is that the dataset only includes crimes within the jurisdiction of the Toronto Police Service, excluding potential influences from neighboring areas. (Westin 2021)

Lastly, the exclusion of some incidents from "Crimes Against the Person" where victim names are missing introduces potential bias. These weaknesses suggest that while the model provides useful trends, more complex models and data enrichment would be needed for more precise predictions and a fuller understanding of crime dynamics.

5.4 Next steps

To improve upon the analysis and gain deeper insights, several next steps should be considered. First, we could explore more complex predictive models, such as time series analysis (e.g., ARIMA) or machine learning techniques like random forests or neural networks. These models could capture non-linear relationships and consider more factors affecting crime rates, improving prediction accuracy.

Additionally, incorporating external data sources, such as demographic information, economic indicators, or unemployment rates, would allow us to understand the broader social and economic factors driving crime. This would also help in developing a more holistic model.

Another important step would be to focus on geographical analysis by examining crime rates at a more granular level, such as neighborhoods or districts, instead of relying solely on divisions. This would enable us to identify high-crime areas and target interventions more effectively.

Lastly, addressing some of the data limitations, such as filtering out unfounded crimes and accounting for unreported crimes, would enhance the quality and reliability of the analysis. By taking these next steps, we can develop a more sophisticated and accurate model to predict future crime trends and support better decision-making in public safety and law enforcement.

A Appendix

B Data-Cleaning

The data cleaning process involved several detailed steps to ensure the dataset was ready for analysis:

B.1 Handling Missing Data

The first step was to check for any missing values across the dataset. Missing values were identified using the `is.na()` function, and the extent of missing data was evaluated. For columns with minimal missing data, the rows were removed to maintain dataset integrity. If a column had substantial missing data but was deemed essential, imputation techniques (like replacing with the mean, median, or mode) were considered. However, if a column was found to be irrelevant for the analysis, it was removed entirely.

B.2 Removing Unnecessary Columns

Columns that were not useful for the analysis were dropped. Specifically, `Division` and `Count_Cleared` were removed because they were not contributing to the predictive modeling objectives. This step helped to reduce dataset complexity and focus on the key variables necessary for the analysis.

B.3 Aggregating Data

The data was aggregated to simplify the analysis and avoid overloading the model with granular details. Crime occurrences were grouped by key variables such as `REPORT_YEAR`, `CATEGORY`, and `SUBTYPE`, and the total counts for these groupings were summed. This allowed for an easier examination of crime trends over time and by category, without repetitive data entries.

B.4 Verifying Data Consistency

Finally, the cleaned dataset was checked for consistency. This involved reviewing the range of values in each column, ensuring the absence of outliers or unexpected data points, and confirming that the dataset accurately reflected the underlying data without any integrity issues.

Overall, these detailed steps ensured the dataset was accurate, consistent, and suitable for analysis, allowing the predictive model to perform optimally.

References

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/, https://github.com/sharlagelfand/opendatatoronto/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*.

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

Westin, Morgaine. 2021. "Reported Crime Statistics in Toronto Can Be Misleading."

Wickham, Hadley, R Francois, L Henry, and K Müller. 2014. "Dplyr." *A Grammar of Data Manipulation 2020 [Last Accessed on 2020 Aug 12] Available from*, Rproject.

Wickham, Hadley, Jim Hester, Romain Francois, Jennifer Bryan, Shelby Bearrows, J Jylänki, and M Jørgensen. 2024. "Package 'Readr'." *Read Rectangular Text Data. Available Online: Https://Cran. R-Project. Org/Web/Packages/Readr/Readr. Pdf (Accessed on 23 August 2023)*.

Wilkinson, Leland. 2011. "Ggplot2: Elegant Graphics for Data Analysis by WICKHAM, h." Oxford University Press.

Xie, Yihui. 2018. "Knitr: A Comprehensive Tool for Reproducible Research in r." In *Implementing Reproducible Research*, 3–31. Chapman; Hall/CRC.

Zhu, Hao, Thomas Travison, Timothy Tsai, Will Beasley, Yihui Xie, and GuangChuang Yu. 2019. "Package 'kableExtra'."