Forecasting the Future: A Regression Analysis of Voter Behavior in the Trump vs. Biden Electoral Showdown*

Ben Li

April 17, 2024

This study employs logistic regression to analyze the factors influencing American voters in the 2020 presidential election, focusing on key demographics like education and income. We discover that higher education levels strongly correlate with voting for Biden, while income shows varied influences across different brackets. These insights underscore the complexity of political choices, influenced not only by socioeconomic status but also by deeper, less tangible societal currents. Ultimately, this paper sheds light on the intricate tapestry of American electoral behavior, signaling the need for political strategies that resonate with an electorate's diverse and multifaceted nature.

Table of contents

*Code and data supporting this analysis is available at: https://github.com/UofT-DailinLi/A-Regression-Analysis-of-Voter-Behavior-in-the-Trump-vs.-Biden-Electoral-Showdown.git

1 Introduction

The U.S. presidential election is one of the world's most closely watched political events, and its outcome impacts not only the U.S. political environment but also international relations. The context of the 2020 election is even more complex and challenging, including Covid-19 and racial issues(Baccini, Brodeur, and Weymouth 2021). Understanding voter preferences and the factors that can shape choices in the highly polarized context of contemporary U.S. politics is critical in both academic and practical applications. This paper discusses the application of logistic regression analysis to the upcoming U.S. presidential election in order to predict voters' choice between Joe Biden and Donald Trump. The aim is to systematically explore and quantify how real-world factors affect the likelihood of voting for a candidate, thus providing valid insights for predicting the election outcome in advance.

The reason why logistic regression was chosen to be used is because the dependent variable, the choice of vote, is inherently binary, voting for either Donald Trump or Joe Biden. Logistic regression is suitable for modelling situations where the outcome is binary, therefore effectively predicting the vote for the two primary candidates. The impact of multiple factors on the vote is captured by considering a range of independent variables, such as demographic characteristics (age, gender, education level, etc.) and social and economic background (income, health care, etc.). This analysis is not only of academic interest but also invaluable to political campaigners in understanding the impact of multiple factors on voter decision-making, campaign strategies, and targeted campaign messages. Policymakers can also learn from voter dynamics, such as those in swing states where small changes in voter preferences can change the outcome of an election.

This report is structured first to present the Section 2, which is used to analyze, process, and discover the data, and includes a data visualization of the variables of interest. The Section 3, which presents simple logistic regression models used to demonstrate the relationship between demographic characteristics and voter outcomes. The Section 4 is a parse of the model and an analysis of the model summary data. The Section 5 discusses the study's findings, this paper's shortcomings, and future directions and potential possibilities.

2 Data

2.1 Data Source

The dataset used in this study is from the American National Election Study (ANES) 2020 time series, which was obtained through the Intercollegiate Consortium for Political and Social Research (ICPSR)(Inter-university Consortium for Political and Social Research 2020). This site is open source and can be downloaded by visiting their website. The dataset was obtained through a questionnaire that included web, phone, and video contact with respondents. The ANES is designed to collect and analyze data on Americans' demographic characteristics, social attitudes, and political behaviour. The purpose is to understand voters' social backgrounds, values, and political opinions better. In the tradition of time-series studies, ANES conducts interviews during presidential election years. The present dataset is from 2020 and consists of two phases of interviews: pre-election interviews conducted two months before the November election and post-election interviews conducted two months after the election. This design approach aims to capture better the discrepancy between voters' expectations and the final outcome, providing a dynamic analytical perspective on the electoral process and its impact. The original dataset consisted of 257 variables and 7453 observations.

This paper was produced using the R statistical programming language (R Core Team 2022) and using here (Müller 2020) to reference file locations . The data was examined and cleaned using the packages janitor (Firke 2021), dplyr (Wickham et al. 2023), and tidyverse (Wickham et al. 2019). Tables were made knitr (Xie 2023) , and formatted with kableExtra (**kableExtra?**). ggplot2 (Wickham 2016) was used to plot and scale the graphs.

2.2 Variables of Interest

Before modelling, we need to select the variables of interest to be analyzed; this way, reducing the number of variables can help simplify the model and make it easier to interpret while maintaining stability. Similarly, we need to remove observations with missing data to reduce the impact of incomplete data on the model. By screening, variables that are irrelevant to the research question are selected to ensure that the factors that influence the voting are selected. Highly correlated independent variables are avoided to avoid distortions to the model and the influence of independent variables on each other. Ultimately, we found seven variables and 5832 observations after screening and cleaning the data.

From table 1 we can see the cleaned dataset, and the seven variables are

1. **Voted (A01)**: Indicates whether the respondent voted in the 2020 general election.
2. **Presidential_Vote (A02)**: Records the candidate for whom the respondent voted in the 2020 Presidential election. This variable includes categories for major candidates such as Joe Biden and Donald Trump.

Table 1: First Ten Row of Cleaned Election Data

Voted	Presidential_Vote	Marital_Status	Household_Income	Education_Level	Age	Sex
Voted	Jo Jorgensen	Married: spouse present	$50,000 to $74,999	Some post-high school, no bachelor's degree	35 to 44	Female
Voted	Joe Biden	Married: spouse present	$100,000 to $124,999	High school credential	35 to 44	Female
Voted	Joe Biden	Married: spouse present	$35,000 to $49,999	Some post-high school, no bachelor's degree	35 to 44	Male
Voted	Donald Trump	Widowed	$125,000 or more	Graduate degree	65 to 74	Male
Voted	Joe Biden	Divorced	Less than $20,000	Some post-high school, no bachelor's degree	65 to 74	Female
Voted	Donald Trump	Married: spouse present	$20,000 to $34,999	Some post-high school, no bachelor's degree	35 to 44	Female
Voted	Joe Biden	Married: spouse present	$50,000 to $74,999	High school credential	65 to 74	Female
Voted	Joe Biden	Separated	$50,000 to $74,999	Some post-high school, no bachelor's degree	35 to 44	Male
Voted	Joe Biden	Married: spouse present	$35,000 to $49,999	High school credential	35 to 44	Female
Voted	Joe Biden	Married: spouse present	$100,000 to $124,999	High school credential	55 to 64	Male

3. **Marital_Status (R13)**: Describes the marital status of the respondent, including categories such as Married, Widowed, Divorced, Separated, and Never married, each represented by a numeric code.
4. **Household_Income (R12)**: Represents the total household income bracket of the respondent, with categories ranging from "Less than $20,000" to "$125,000 or more", coded numerically.
5. **Education_Level (R11)**: Reflects the highest level of education achieved by the respondent, from "Less than high school credential" to "Graduate degree".
6. **Age (R08)**: Specifies the age range of the respondent, such as "18 to 24", "25 to 34", and so on, up to "75 or older".
7. **Sex (R09)**: Indicates the gender of the respondent.

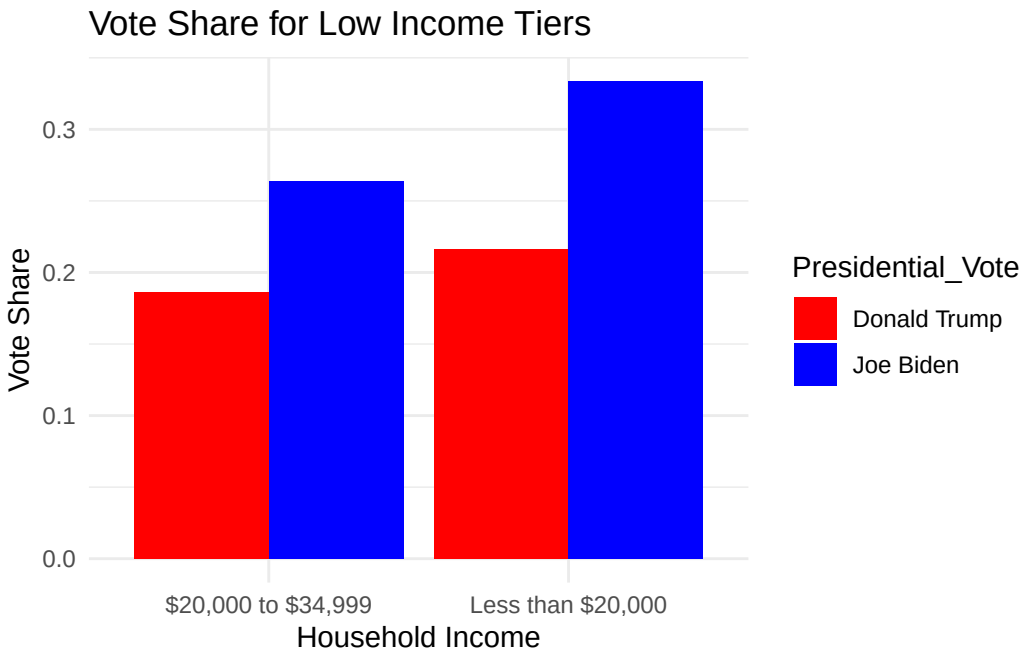

Figure 1: Vote Share for Low Income Tiers"

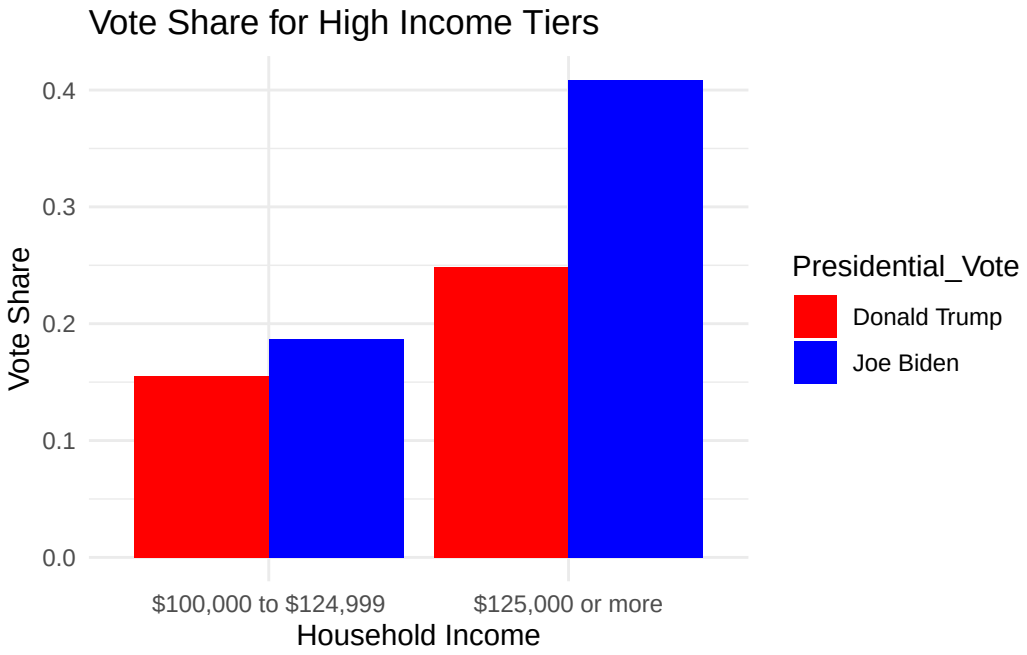

Figure 2: Vote Share for High Income Tiers

According to the two bar charts in Figure 1 and Figure 2, which show the vote share of

Trump and Biden's voting elections in the high-income and low-income tiers, respectively, We categorize "$20,000 to 34,999" and "less than $20,000" as the low-income tier, and "$100,000 to 124,999" and "$125,000 or more" as the high-income tier. Figure 1 represents the vote share of low-income individuals. From the income bracket of "$20,000 to 34,999," we can observe that Trump and Biden will be relatively close, with a more even vote share. Among voters with incomes "less than $20,000", we can see that Biden's vote share is significantly higher than Trump's, which means that people with the lowest incomes represented by the data would be more inclined to choose Biden as president. While Figure 2 represents the high-income people, we can find through the chart that in the "$100,000 to 124,999" income bracket, Trump and Biden's vote share is relatively close to Trump's (in red) and slightly lower than Biden's. In the higher income bracket of "$125,000 or more," Biden's vote share dramatically exceeds Trump's.

The data visualization shows that Joe Biden would have a higher percentage of votes in both the highest and lowest income brackets. In contrast, the percentage of votes in the middle and upper-income brackets would be relatively balanced. The visualization suggests that the economic factors behind household income may influence voting preferences in elections, with the two extremes of low and high incomes favouring Biden more. At the same time, the middle and upper-income tiers are more balanced. This result is also common sense.

3 Model

In our study, we choose logistic regression modelling because the outcome is a binary outcome. The probability of an individual voting for a specific presidential candidate such as Biden and Trump, "p," is modelled as a function of several independent variables. These variables include marital status, household income, education level, gender, and age. The model is as follows.

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) =& \beta_0 + \beta_1 \times \text{MaritalStatus} + \beta_2 \times \text{HouseholdIncome} + \\ & \beta_3 \times \text{EducationLevel} + \beta_4 \times \text{Age} + \beta_5 \times \text{Sex}\end{aligned}$$

Where:

- p represents the probability of voting for the specified candidate.
- β_0 represents the intercept of the logistic regression model, which is the log odds of voting for the candidate when all predictors equal zero.
- β_i (for i = 1, 2, ..., 5) represents the coefficients of the respective predictors in the model.
- Each of the independent variables (MaritalStatus, HouseholdIncome, EducationLevel, Age, Sex) are represented by their respective β_i coefficients.

In the field of predictive analytics, the logistic regression model is a stable and widely used model that can explain the relationship between a binary outcome variable and multiple independent variables well. Moreover, logistic regression is particularly suitable for binary situations, such as elections, where the direction of the study is to choose between two presidential candidates. In contrast, linear regression predictions may exceed the possible range of 0-1. Logistic regression does not and is more stable than linear regression when predicting binary outcomes. The principle of logistic regression is to provide a coefficient for each predictor. By quantifying its log probability impact on the outcome (in this study, the outcome of the vote for a particular candidate) and understanding how each factor has an impact on the voter's voting preference, and by converting log odds to probabilities, we can predict the voter's voting preference through a logistic regression model.

4 Results

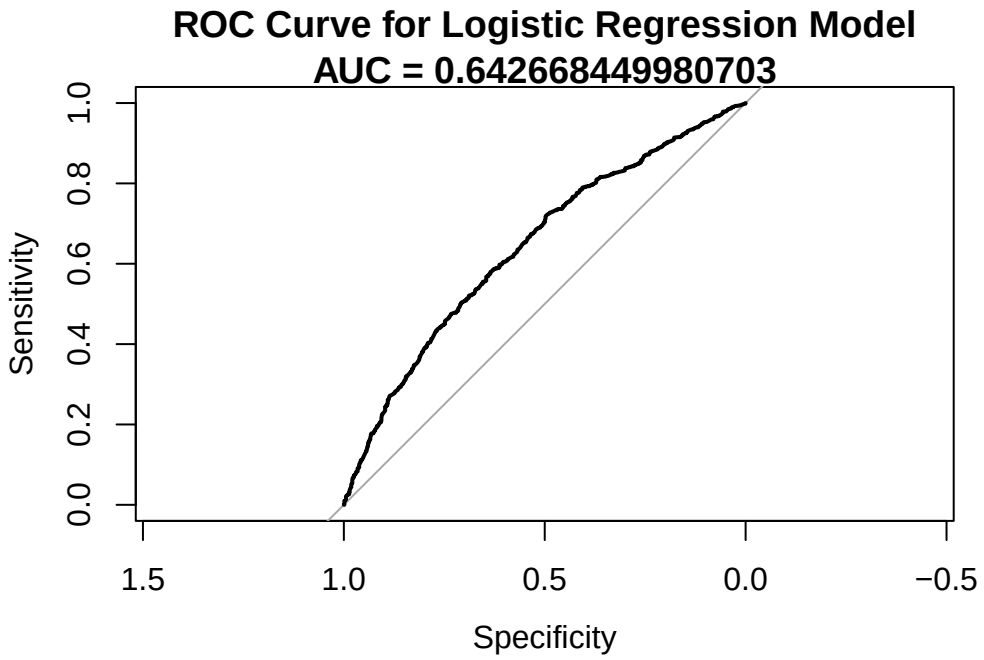

Figure 3: ROC Curve for Logistic Regression Model

Figure 3 shows the Receiver Operating Characteristic Curve (ROC) of a logistic regression model used to predict voter behaviour based on household income, gender, age, marital status, and education level, i.e., whether or not an individual voted for Biden. The dataset for this study utilized the typical 80-20 data split method, where 80% of the data was used to train the model, and the remaining 20% was used to test the model's accuracy. The roc curve above is used to illustrate how the diagnostic ability of a binary classifier system varies with its discrimination threshold. It includes two parameters:

- **True Positive Rate (Sensitivity)**: The proportion of actual positives (votes for Biden) correctly identified by the model, plotted on the y-axis.
- **False Positive Rate (1 - Specificity)**: The proportion of actual negatives (votes not for Biden) incorrectly classified as positive, plotted on the x-axis.

The "AUC" (Area Under the Curve) of this graph quantifies the overall ability of the test data to distinguish between voting for Biden (positive) and not voting for Biden, which is also known as supporting Trump (negative). The value of the AUC ranges from 0 to 1. When the value of the AUC is equal to 0.5, the model's accuracy tends to be a random guess, which means it cannot distinguish. When the value of AUC is equal to 1, it means that the model can distinguish the data completely. As shown in the Figure 3, the value of AUC is 0.6427, which indicates that the model has a moderate discriminative ability, stronger than random guessing but not excellent. According to Figure 3, the model has some predictive ability in predicting the voter's voting preference (Biden or Trump), but the accuracy is not very good. There is still room for improvement at this AUC level.

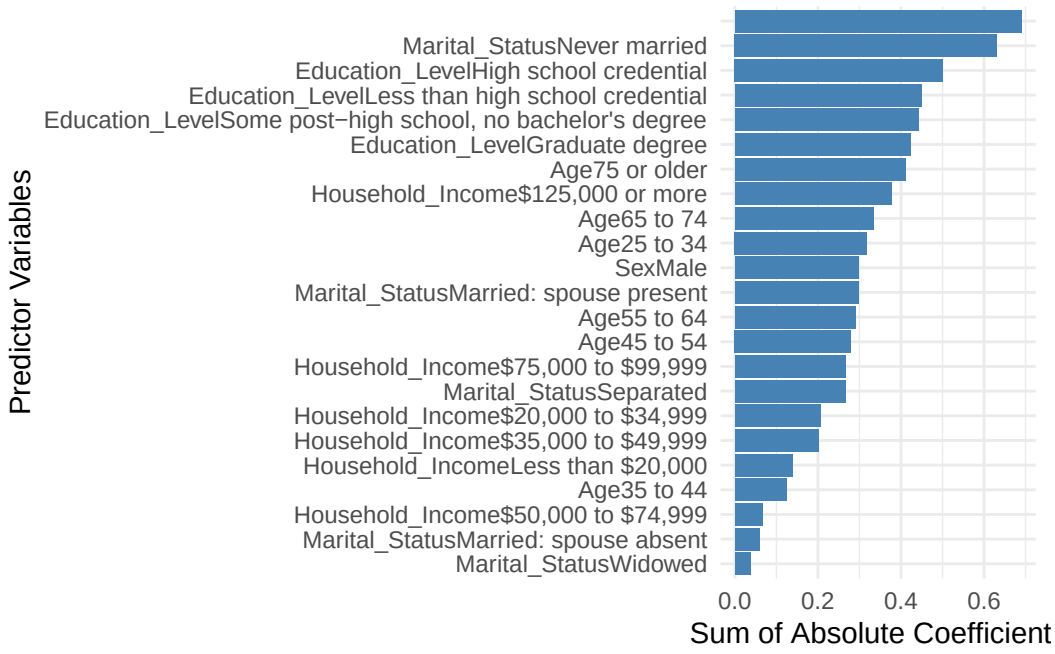

Figure 4: Variable Importance Plot

Figure 4 above is a plot of variable importance showing the sum of the absolute values of the coefficients in the logistic regression model. From the above Figure 4, we can see that the length of the bar graph indicates the relative impact of each predictor variable (or level in the case of categorical variables) on the results, with the more extended bar graph indicating a more significant impact on the model predictions. The graph is organized in order of importance from highest to lowest. Since the graph uses absolute values, the bar graph does not indicate direction (positive or negative). Looking at it, the level of education seems to be the most influential variable. This is because, from the chart above, the second, third, fourth, and fifth positions are all education levels. This Figure 4 allows for quick visual observation of which variable factors are most influential in the logistic regression model and which factors do not play a role, for example, age, all of which are at the bottom of the significance graph, which means that there is not much of a role. However, further analysis is needed to explain these effects fully.

5 Discussion

5.1 Summary of Findings

This study aims to use logistic regression to analyze the decision-making preferences of voters in the 2020 U.S. presidential election. By analyzing a dataset containing demographic and

socio-economic indicators of several voters, we illustrate the relationship between these factors and the likelihood of voting for Joe Biden or Donald Trump. Our modelling confirms that demographic characteristics, such as household income, education, etc., impact voters' decisions. According to Lewis-Beck and Rice's model, the prediction of economic status and candidate popularity on future votes was measured(Abramowitz 1988). Our conclusions coincide with Lewis-Beck and Rice's findings.

5.2 Insights into Voter Behavior

The analysis shows that education level is a significant predictor of voting preferences, suggesting that the educational background of voters heavily influences the 2020 presidential election. This is in line with contemporary discussions about the impact of education on political ideology and voting, while the model also points out that household income is another important influence and that different income brackets are likely to treat presidential candidates differently, as political views cannot satisfy all classes of people at the same time. These insights can be helpful to politicians.

5.3 Additional Learnings

The knowledge gained in this study is not only limited to statistics but also includes a better understanding of the factors influencing voters. Roc's curve analysis yielded medium AUC scores, which suggests that a large portion of influences are not captured by our regression model, which means that there is a large portion of unexplained factors. It also shows that voter behaviour is multifaceted and receives influence from several columns of factors, and many factors need to be quantifiable. In contemporary political analysis, it has been realized that while demographic factors are important, they are also influenced by more profound emotional responses and psychological influences(Gimpel 2003). Voters don't vote based on data but are influenced by life circumstances, energy, and political discourse. The role of the media is also very important, as the primary way we get our information nowadays is through various media, especially with the rise of digital platforms. The influence of the media in shaping public opinion and voting behaviour is also significant. In conclusion, while the regression model provides valuable insights into the factors that influence voting, the value of AUC pushes us to understand the electorate's broader, complex and unquantifiable dimensions.

5.4 Weaknesses & Next Steps

5.4.1 Weaknesses

Despite our insights, this analysis has limitations, and the model's discriminatory power is average in terms of AUC values. This suggests that the model has a lot of room for improve-

ment. In addition, the selection of variables relies mainly on perceived screening. It does not allow for a systematic and accurate analysis of which factors should be covered in the model using advanced machine learning models. So, by manual screening, we have limitations in selecting six demographic characteristics from 257 variables. At the same time, the nature of the data was obtained by self-reporting, a method that is likely to bring inaccuracies in the data and thus have an impact on the model and while the logistic regression coefficients may overemphasize certain levels in the categorical variables, thus skewing the predictions of the model.

5.4.2 Next Steps

Future research directions can be adapted to more complex models, e.g. through machine learning, and to consider the impact of public opinion and media influence on voting. For example, one could analyze the potential impact of twitter comments on voting during an election, and its possible to analyze deeper levels of people's mental activity as well as their emotions. Our research is going to try to incorporate more unquantifiable dimensions, trying to incorporate these multifaceted factors - psychological, media-driven, campaign-related, and faith-based factors. As our understanding of these dynamics deepens, so will our ability to more accurately predict election outcomes.

References

Abramowitz, Alan I. 1988. "An Improved Model for Predicting Presidential Election Outcomes." *PS: Political Science & Politics* 21 (4): 843–47.

Baccini, Leonardo, Abel Brodeur, and Stephen Weymouth. 2021. "The COVID-19 Pandemic and the 2020 US Presidential Election." *Journal of Population Economics* 34: 739–67.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gimpel, Jim. 2003. *Perspectives on Politics* 1 (3): 606–7.

Inter-university Consortium for Political and Social Research. 2020. "American National Election Study: 2020 Time Series Study." ICPSR - Inter-university Consortium for Political and Social Research. https://www.icpsr.umich.edu/web/ICPSR/studies/38313/summary.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*.

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. 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.

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

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.