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

Ben Li

April 17, 2024

This paper investigates the development of American Sign Language in children by observing the relationship between the range of ages during developmental years and the various types of vocabulary terms learned. Using the data from the open database, Wordbank, we found that children learning to speak a visual language share a similar trend in vocabulary development to those learning spoken languages. This highlights the importance of learning American Sign Language at an early age as it provides the ability for those with hearing impairments or disabilities to participate in society like any other able-bodied individual.

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.

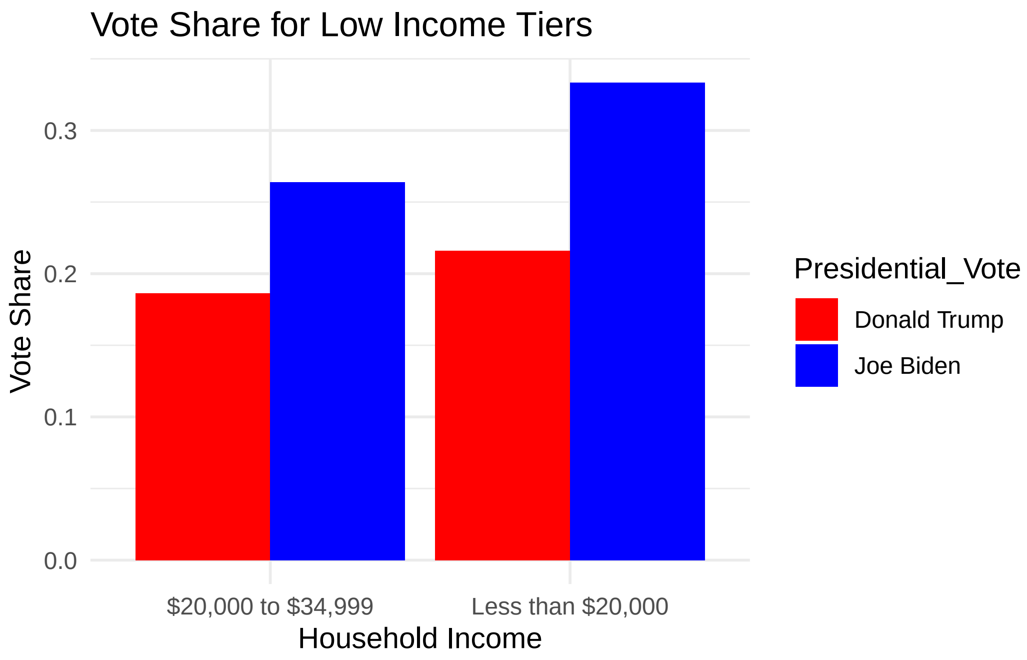

Figure 1: Vote Share for Low Income Tiers”

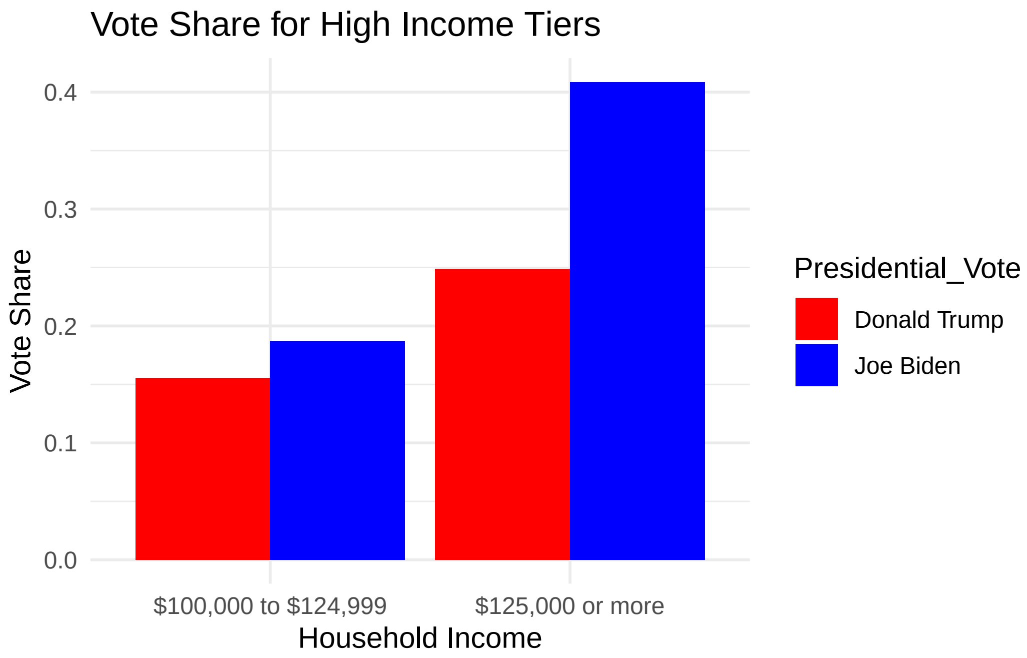

Figure 2: Vote Share for High Income Tiers

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

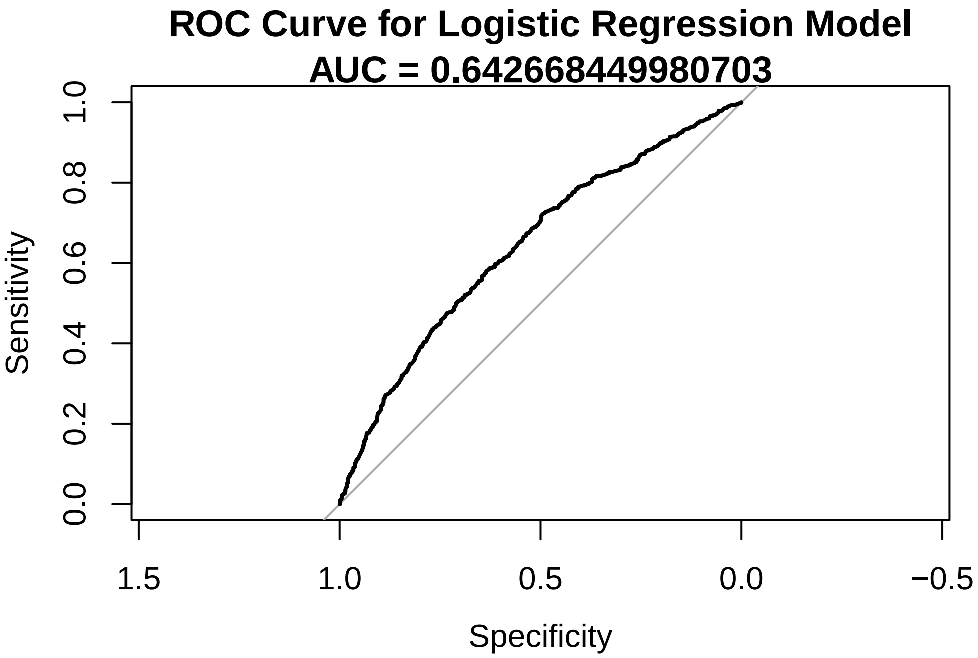

Figure 3: ROC Curve for Logistic Regression Model

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.

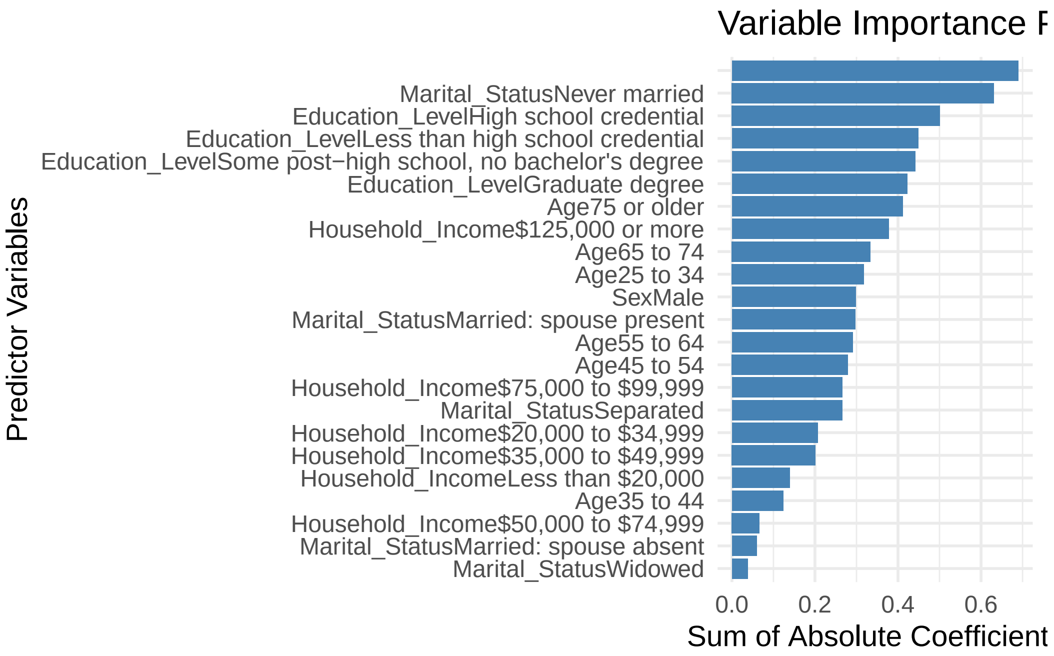

Figure 4: ROC Curve for Logistic Regression Model

5 Discussion

5.1 Learning sign language, especially at an early age, creates a more accessible world for the deaf and hard-of-hearing community

Learning sign language at an early age expands children's understanding of diverse individuals and communities, fostering their ability to be more empathetic and accommodating towards accessibility. Being able to communicate through sign languages, such as American Sign Language, provides more accessible and effective opportunities for individuals who are deaf or hard of hearing (**cite1whatisASL?**). Sign language is a highly accessible form of communication that enables individuals with hearing impairments to engage in conversations on equal grounds with others, ultimately accommodating diversity and universal inclusion.

5.2 At early ages, children have strong capabilities of learning multiple languages, including sign language

The timeline of language development for spoken languages during the emerging years of children has a similar growth trend to the data analysis in this paper, with a more significant upward trend. While children learning to speak a spoken language are able to produce their first words between the ages of 6 to 11 months old (**cite2speechmilestones?**), those learning to communicate through sign language produce one-word vocabulary by the age of 12 months (**cite3signmilestones?**). Afterwards, they both follow the same trend of producing 2-word phrases around the ages of 18 to 24 months old and by 2-3 years of age, they are able to use 3-word phrases. These highlighted milestones indicate that the vocabulary development of spoken languages shows a similar linear growth trend to the vocabulary development of sign languages.

Early learning of sign language is beneficial as it helps children have a better comprehension of the language while simultaneously improving their cognitive and social development. Research suggests that the early years of a child's life are the most crucial in terms of the development of their language skills (**cite1whatisASL?**). Children at early ages have the capability to learn multiple languages at a time without confusion or any sort of delay in their development, as research has shown that their brains are able to handle dual language development (**cite4signengbilingual?**). They are able to expand their communication and linguistic skills as well as provide access to a language that meets the needs of deaf individuals.

5.3 Learning sign language enhances visual skills and improves cognitive abilities

Expanding on the cognitive and social benefits that coincide with the development of sign languages in children, it also enables children to enhance their visual skills and cognitive abilities. Sign languages allow children to communicate and express themselves effectively. For instance, when a child is unsure of a sign in the language, they are able to work around this barrier by using the letters of the alphabet to express what they are trying to say (**cite5signbenefits?**). Additionally, because sign language involves the physical movement of the body, the use of muscle memory helps increase a child's vocabulary (**cite5signbenefits?**). It helps perform words automatically without conscious effort or attention, demonstrating a retained memory of vocabulary. Alongside movement, sign languages are also visual languages. This helps in enhancing the visual learning experience of children during their developmental stages (**cite5signbenefits?**). With this, they are able to see words in motion and retain both the motion of the word or phrase and the vocabulary word itself. Teaching children sign language, whether they are experiencing hearing loss or are able to communicate via speech, can have a positive impact on their language development as well as their social interactions and relationships.

5.4 Weaknesses & Next Steps

5.4.1 Weaknesses

The limitations of using data from Wordbank are that its primary focus is on typical development and its datasets focus on monolingual acquisition. In addition, the sample of languages is restricted by data accessibility. Additionally, there are a lot of NA responses within the dataset. This invites the problem of missing data, which can reduce the sample size and have an overall reduced statistical power. Missing data could also introduce bias into the analysis, especially with the main data collection being the survey reports from parents or primary caregivers. This can affect the representativeness of the sample. Moreover, due to the limited number of unique children for the data on American Sign Language development, a large number of observations had to be simulated for the model and its analysis.

5.4.2 Next Steps

A possible next step could be to explore how different categories of vocabulary words are developed over the age range and their varying levels of comprehension and production of those words. With that being said, we could also extend our linear regression model to be a multiple linear regression model, to accommodate the different categories. Regarding the NA responses of the dataset, what could be done to improve the overall analysis would be to fill in missing values with estimated values based on other the responses. This would be beneficial for both analyzing trends of the data and generating linear regression models. Additionally, since we based our timeline of milestones for learning spoken languages on secondary research, providing data visualizations and a model to compare and contrast the development of visual languages with spoken languages would enhance the analysis of this paper.

References

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.

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/.