# Empirical evidinces of gender gap and education gap in choice of candidate in 2020 US presidential elections*

Thu Dong

March 17, 2024

Voters' preference of candidates running for president and their backgrounds have a significant impact on the political landscape. We are particularly interested in the potential relationship between voters' demographics, specifically gender and education, on their choices of candidate. This relation has been analyzed in numerous studies throughout the years and different results have been found. Given the scarce empirical evidence that exists regarding this topic, our paper looks at the relation between gender, education, and choices of candidates using the 2020 US presidential election data. We observed a statistically and economically significant that females are more likely to support Democrat representative candidates like Biden. Additionally, we witnessed how an increase in educational level affects the choice of candidates while the effect differs between two genders. Finally, we acknowledged the potential limitations and threats to the internal and external validity of our results.

## Table of contents

*Code and data are available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E9N6PH.

# 1 Introduction

The political scene is shaped by various factors and one of the most important factors that affect the outcome of elections is the demographics of voters. Understanding how the demographics of the voting population influence their candidate choices allows candidates to adjust political campaigns and policies. Therefore, this paper will discuss the effect of voters' demographics on the candidate they choose with empirical evidence and the implication of this effect on the outcome of the elections.

According to a 2023 Center for American Women and Politics study, there is a gender gap in voting which refers to a difference between the percentage of women and the percentage of men voting for a given candidate, generally the winning candidate (Center for American Women and Politics (CAWP) 2024). The study, conducted in the US, shows that one of the reasons the gender gap in voting exists is because the majority of women have preferred Democratic candidates. The evidence was found in the 1996 election, 2000 election, and 2008 election. Additionally, the same study shows that in the 2016 elections that shows that voters with a college degree or higher education are more likely to vote for the Democrat party representatives than those who do not have a college degree.(Pew Research Center 2016)

This paper will examine the influence of voters' demographics, specifically gender and education level, on voters' preference of candidates as they are shown to have a growing significance in election outcomes. Since no previous study on real data regarding this topic, this research aims to provide additional empirical evidence to this topic and explore how different demographics respond to candidates' campaigns, policies, and political views. Additionally, this paper target to contribute a deeper understanding of the population's view of the political landscape and the impact of population preference on presidential outcomes. Using the 2020 presidential election data provided by CCES in Harvard Dataverse, this paper found that there is empirical evidence that shows that the gender gap and education gap in voting exist and they exist mainly because of politician's proposals of policies during campaigns and party affiliation.

The estimand of the paper is the true relationship between voters' demographics, specifically gender, and education, on their choices of candidate. The remainder of this paper is structured as follows. Section 2 discusses the raw data, cleaning process, summary statistics, and visualization of the predictors. Section 3 shows the model used for empirical analysis and the rationale behind the model. Section 4 analyze the results of the models. Finally,Section 5 examines the implication of the findings on political scenes, the limitations of the research, and further steps to improve the reliability of the research.

# 2 Data

## 2.1 Raw data.

The data used in this paper is derived from CCES in Havard DataVerse. This was the final version of the 2020 Cooperative Election Study Common Content dataset. All the data analysis was done through R (R Core Team 2023) with the aid of the following packages:tideyverse (Wickham et al. 2019),boot (Davison and Hinkley 1997),broom.mixed(Bolker and Robinson 2022),collapse(Krantz 2024),knitr(Xie 2014), dataverse(Kuriwaki, Beasley, and Leeper 2023), gutenbergr(Johnston and Robinson 2023),janitor(Firke 2023), marginaleffects(Arel-Bundock 2024), modelsummary(Arel-Bundock 2022), rstanarm(Brilleman et al. 2018),and tidybayes(Kay 2023).

The raw data is published by Cooperative Election Study, a national stratified sample survey administered by YouGov. The data was gathered through two surveys: Pre-election ( September 29 to November 2, 2020) and Post-election ( November 8 to December 14, 2020) which were specifically gathered to study voters' behavior in the 2020 election. There was a total of 650 questions asked across both surveys. The surveys were completed by 60,000 American citizens. Each variable in the dataset corresponds to one of the 650 questions. The questions involved topics about their gender, race, socioeconomic status, political beliefs, etc. These questions aim to provide information on how Americans view Congress, how they voted, and how their behavior and experiences vary. The data included a large sample of 60,000+ representative Americans. The unit of observation is ' respondent'.

## 2.2 Cleaned data

Since the data constructed contained 650 variables, only the demographics variables, which are gender and education level were selected to analyze the effect of demographics on voters' preferences. To identify who the voters prefer in the 2020 election, it is important to select only the survey participants who registered to vote. Moreover, since the 2 most popular candidates in the 2020 presidential election are Trump, the representative candidate of the Republican Party, and Biden, the representative of the Democratic Party, only the participants who voted for either candidate will be selected for simplicity. Furthermore, some data points had missing attributes whereby an "NA" was put in place of the true value. Such entries were removed entirely in the data cleaning process as the number of observations was large and removing those entries will not have a significant impact on the outcome. The data now has 43,554 observations and 3 variables: candidates they voted for, gender, and education level. All 3 variables are factors variables and will further be explored below. Table 1 shows the data after the cleaning proccess.

Table 1: Voters demographics and choice of candidate in 2020 presidential election

| X | voted_for | gender | education |
|---|---|---|---|
| 1 | Trump | Male | 2-year |
| 2 | Biden | Female | 4-year |
| 3 | Biden | Female | 4-year |
| 4 | Trump | Male | Some college |
| 5 | Trump | Female | Some college |
| 6 | Trump | Female | High school graduate |

The variable in the data set is measured by using the results of the survey conducted pre and post-election. All 3 variables used are categorical. Preference for candidates is measured by whether the person is reported to vote for Trump or Biden. Other candidates running in the 2020 election were not considered. Additionally, gender is measured by the categorical variable Male or Female. No other gender was included in the data set. Finally, education is measured by 6 different categories that will be discussed below.

Table 2: Summary statistics of voters demographics and choice of candidate in 2020 presidential election

| voted_for | gender | education |
|---|---|---|
| Trump:17558 | Male :19251 | No HS : 689 |
| Biden:25996 | Female:24303 | High school graduate: 9814 |
| NA | NA | Some college : 9290 |
| NA | NA | 2-year : 4971 |
| NA | NA | 4-year :11518 |
| NA | NA | Post-grad : 7272 |

### 2.2.1 Preference of candidates:

The variable "vote_ for" represents the candidate the survey participants voted for during the 2020 presidential elections. The participants either voted for Trump or Biden. Table 2 shows summary statistics of the clean data. It shows that 17,558 survey participants voted for Trump which was approximately 40.3% of the voters while 25,996 people voted for Biden which was around 59.7% of voters.

### 2.2.2 Voters Demographics:

#### 2.2.2.1 Gender.

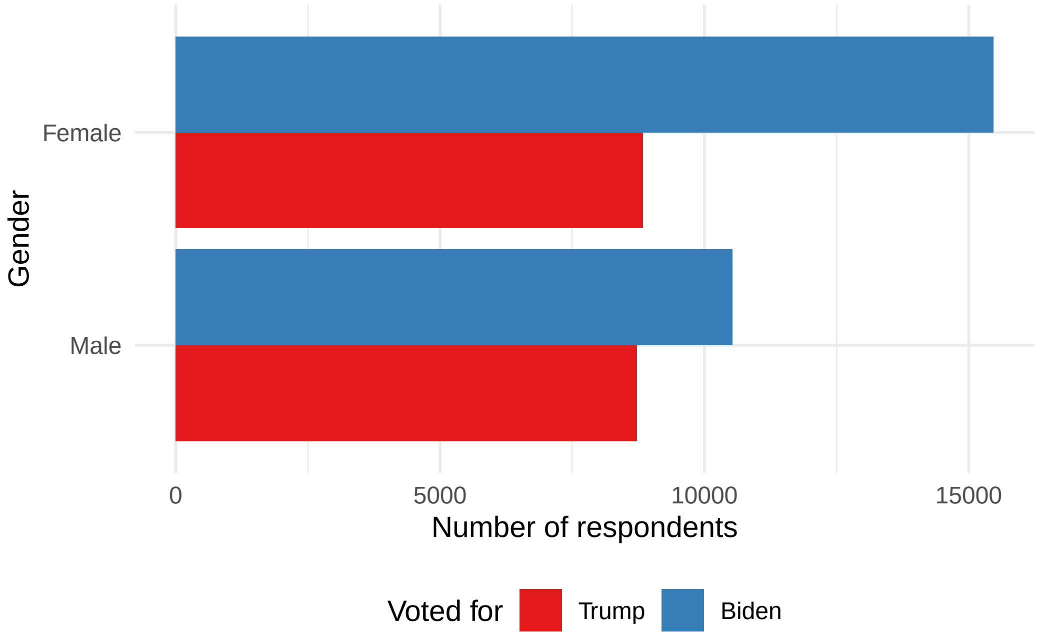

Figure 1: The distribution of presidential preferences, by gender

The gender of survey participants will take values of either “Male” or “ Female”. Of the 43,554 survey participants, 19,251 were male and 24,303 were female. Figure 1 shows a bar graph of who respondents vote for, grouped by gender. It is shown that around 36.4% of female participants voted for Trump while 63.6% voted for Biden. For their male counterparts, while 45.3% of male participants voted for Trump, 54.7% of male respondents voted for Biden. These number indicates that female voters are more likely to vote for Biden than male voters. However, this evidence is inconclusive and needs further testing to conclude the relationship between voters’ choice of candidate and their gender.

#### 2.2.2.2 Education level.

Participants’ education level is divided into 6 groups: No High School, High school graduates, some colleges, 2 years of college, 4 years of college, and post-grad as shown in Table 2. There are 689 out of 43,554 respondents who have not finished high school, and 9814 participants are high school graduates. Additionally, 9290 respondents have some college education, and 4971 have finished 2 years of college. Out of 43,554 participants, the largest group consists of individuals with a 4-year college degree, totaling 11,518 respondents. Finally, the group with the highest level of education is post-grads with 7272 respondents.

Figure 2 is a bar graph illustrating the candidates respondents vote for grouped by gender and education level. Overall, it is found that for the most part, females with higher education levels are more likely to vote for Biden while there is no clear trend in male counterparts. It shows

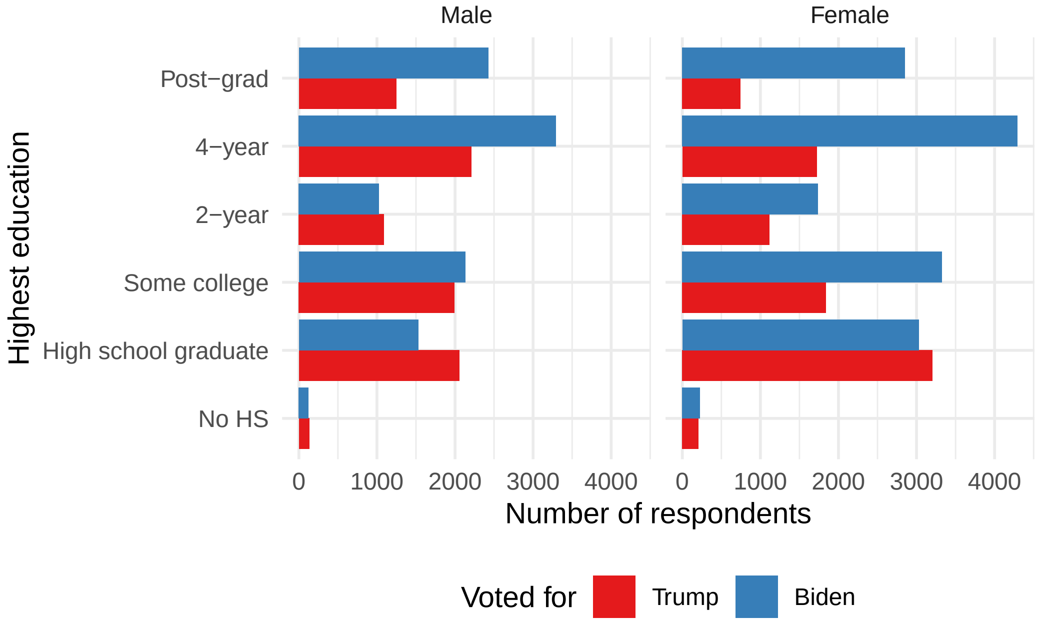

Figure 2: The distribution of presidential preferences, by gender and educational level

that 51.8% of male participants with no high school experience voted for Trump while only 48.2% voted for Biden. However, 47.5% of female respondents with no high school experience voted for Trump while 52.5% voted for Biden. The graph also demonstrates that 51.4% of female high school graduate participants voted for Trump while 57.3% of male high school graduate respondents voted for Trump. Interestingly, it is found that high school graduates, regardless of gender, are more likely to vote for Trump than Biden. Furthermore, Figure 2 shows that 51.5% of male participants with 2 years of college voted for Trump and 48.9% voted for Biden. However, 39.1% of female participants from the same educational background voted for Trump while 60.9% voted for Biden. Additionally, 35.6% of female respondents with some college experience voted for Trump while 48.3% of their male counterparts voted for Trump. The bar graphs also show that 28.6% of female respondents with 4-year college graduates voted for Trump while 40.1% of the male participants with the same educational level voted for Trump. Lastly, 34% of post-grads male participants voted for Trump while only 20.7% of the female counterparts voted for Trump. These statistics not only show that females are less likely to vote for Trump than males but also show that in general, but it also demonstrates that people with higher education levels are more likely to vote for Biden.

## 3 Model

The model used to analyze the relationship between the candidate respondents vote for and voters' demographics is logistics regression. Logistics regression provides a framework to analyze categorical outcome variables. Furthermore, logistics regression shows the probability of the occurrence of an event which is suitable for analyzing binary variables such as ones in our research.

## 3.1 Simple regression:

Simple regression was used as a guideline to examine the relationship between gender and choice of candidates.

$$\begin{aligned} y_i|\pi_i &\sim \text{Bern}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 * gender \\ \beta_0 &\sim \text{Normal}(0, 2.5) \\ \beta_1 &\sim \text{Normal}(0, 2.5) \end{aligned}$$

## 3.2 Multiple regression:

Multiple regression was used to deepen analysis and include the effect of educational level on voting choices.

$$\begin{aligned} y_i|\pi_i &\sim \text{Bern}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 * gender + \beta_2 * education + \beta_3 * gender * education \\ \beta_0 &\sim \text{Normal}(0, 2.5) \\ \beta_1 &\sim \text{Normal}(0, 2.5) \\ \beta_2 &\sim \text{Normal}(0, 2.5) \\ \beta_3 &\sim \text{Normal}(0, 2.5) \end{aligned}$$

The dependent variable $y_i$ represents the political preference of the respondent and is equal to 1 if Biden and 0 if Trump, $gender_i$ is the gender of the respondent with 1 equal to female and 0 if male and $education_i$ is the education of the respondent. Additionally, $gender_i$*$education_i$, which is the interaction term of gender and education, was included to enhance the accuracy of the model and account for the conditional relationship between the two independent variables.

$\pi_i$ is the probability that the ith respondent voted for Biden. Moreover, it is assumed that the distribution of coefficients $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ are normal distributions with a mean of 0 and standard deviations of 2.5. This assumption is made as the model follows a Bayesian framework which allows us to incorporate prior information into the model. Assuming that the distribution of coefficients $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ follows a normal distribution allows for a weakly informative prior which implies a neutrality on the possible value of the coefficients. Mean centering around 0 implies that there is no bias in the direction of all coefficients and

Table 3: Regression of gender on choice of candidates.

| | Support Biden |
|---|---|
| (Intercept) | 0.528 |
| genderMale | −0.340 |
| Num.Obs. | 1000 |
| R2 | 0.007 |
| Log.Lik. | −672.724 |
| ELPD | −674.8 |
| ELPD s.e. | 6.3 |
| LOOIC | 1349.5 |
| LOOIC s.e. | 12.7 |
| WAIC | 1349.5 |
| RMSE | 0.49 |

a standard deviation of 2.5 allows for a moderate level of variation in the predictor variable. Furthermore, this assumption also prevents overfitting of the model by constraining the coefficients to reasonable values.

## 4 Result

Table 3 above shows the regression of gender on the voting choice of survey participants. The intercept is 0.528 which represents the log odds of female participants supporting Biden. This number indicates that the probability that female participants supporting Biden is, on average, 1.7 times more than they supporting Trump. The log odds decrease for males as the coefficient is −0.340, which shows that the odds that male respondents support Biden are less than Female participants. It is found that the probability that the male participants support Biden is 1.26 times they support Trump. Table 3 shows empirical evidence that female participants overall are more likely to support Biden than male participants.

Table 4 demonstrates the effect of gender and education on voting choices. The intercepts represent the log odds of supporting Biden for female participants with 2 years of college experience which is 0.151, illustrating that the female participants with 2 years of college experience supporting Biden is approximately 16% more than those supporting Trump. However, with a B_1 of -0.312, on average, the odds of a male with 2 years of college experience supporting Biden is 0.312 less than a female with the same education level.

It was also found that females with an education level below an associate degree( except in some colleges) tend to vote less for Biden. The odds of females with a high school degree voting for Biden decreased by 0.477 compared to females with associate degrees. This number decreases even further when considering females with below-average education levels. Compared to

Table 4: Regression of gender and education on choice of candidates.

| | Support Biden |
|---|---|
| (Intercept) | 0.151 |
| | (0.238) |
| genderMale | −0.312 |
| | (0.387) |
| education4-year | 1.083 |
| | (0.312) |
| educationHigh school graduate | −0.477 |
| | (0.301) |
| educationNo HS | −0.668 |
| | (0.579) |
| educationPost-grad | 0.770 |
| | (0.339) |
| educationSome college | 0.677 |
| | (0.306) |
| genderMale × education4-year | −0.470 |
| | (0.474) |
| genderMale × educationHigh school graduate | 0.368 |
| | (0.489) |
| genderMale × educationNo HS | −32.525 |
| | (27.330) |
| genderMale × educationPost-grad | −0.195 |
| | (0.517) |
| genderMale × educationSome college | −0.314 |
| | (0.481) |
| Num.Obs. | 1000 |
| R2 | 0.072 |
| Log.Lik. | −642.907 |
| ELPD | −654.2 |
| ELPD s.e. | 9.7 |
| LOOIC | 1308.3 |
| LOOIC s.e. | 19.3 |
| WAIC | 1308.3 |
| RMSE | 0.48 |

females with an associate degree, the odds of females with no high school degree supporting Biden is on average 0.668 lower. Remarkably, the odds of a male with a high school degree voting for Biden is 0.056 higher than the odds of a female with the same education level voting for Biden while the difference is minimal. Notably, the odds of males with no high school degree voting for Biden are 32.837 lower compared to females with no high school degree which is a significant decrease.

Evidence also shows that female voters with higher education levels tend to support Biden more. Females with some college education have higher odds of supporting Biden than females with associate degrees, specifically 0.677 higher. Furthermore, the odds of females with 4 years of college education supporting Biden is 1.083 higher than females with 2 years of college education which indicates that the higher the education the women have, they are more likely to support Liberal candidates as discussed in the literature review. Notably, compared to females with associate degrees, the odds of females with post-graduation educations voting for Biden are 0.770 higher. On the other hand, the higher education level has a slightly different effect on their voting preference. It was found that the effect of education on males' voting behavior is much smaller. Compared to male voters with associate degrees, the odds that males with some college degrees support Biden is 0.363 higher. Additionally, the odds of men with 4 years of college education vote for Biden is 0.613 higher than the odds of supporting Biden for males with 2 years of college education. The odds of males voting for Biden with post Grad education is 0,575 higher than the odds of supporting Biden for males with 2 years of college education

# 5 Discussion

This paper has not only shown the existence of the gender gap and education gap in voting using empirical evidence, but it was also found that gender and education have a significant effect on the outcome of the election which illustrates that appealing to certain demographics can significantly change the outcome of the election.

## 5.1 The gender gap in election 2020:

As it has been shown Table 3 and Figure 1, even though, male voters are more likely to vote for Biden, female voters support Biden with a higher margin than men. One of the potential reasons why Biden is favored in the 2020 election is because of gender-related policies that he proposed during his presidential campaign. For example, Biden issued policies for reproductive rights, paternity leave, and pay equity which support women's rights(The White House 2022). These proposals in the campaign increase the possibility that women will vote for Biden. Moreover, another reason why women are more likely to support Biden is because of his stance on social issues. It is believed that Biden has an empathetic attitude towards gender equality, LGBTQ+ rights, and racial justice which allows both genders to support him more. Lastly, Trump's policies in his previous presidency could be a reason why he gained less support from the public. Specifically, his rhetoric and behavior during his precedency could have driven female voters to seek an alternative candidacy in Biden.

### 5.2 The education gap in voting preferences:

Figure 2 shows clearly the education gap in voting and how it affects each gender differently. Additionally, Table 4 provides how significant the effect of each level of education on voters' behavior. It is found that on average, when females are more educated, they are more likely to support Biden as they are more aware of the proposal of policies and social views of Biden and how it affects them as individuals. Moreover, in the 2020 elections, while males with higher education levels are more likely to vote for Biden, the effect of education and gender is much smaller compared to females. This could also be linked to Biden's policies that have a greater effect on female voters. This is not the first time the education gap in voting has existed. There was evidence of an education gap in presidential preference ever since 1992. According to the Pew Research Center, college graduates or more and those without college degrees have little difference in their voting choices. However, it has shown that this gap widens over time, and it is the most noticeable in 2016, especially in white voters. It was shown that white college graduates supported Clinton over Trump ( 47% Clinton vs 33% Trump) while those without a college degree supported Trump over Clinton ( 51% Trump vs 26% Clinton). ("Educational Divide in Vote Preferences on Track to Be Wider than in Recent Elections" 2016). One of the reasons why this phenomenon exists could be because of the different party affiliations in both groups. The voters who have higher educational levels are shown to be more likely to hold a Liberal view than people who have lower educational levels which indicates that they are more likely to support the Liberal Candidate.

### 5.3 Limitation

There are certain limitations to this model that can affect the internal and external validity of the research. Firstly, one of the external validity concerns is that since the data is recorded for the 2020 presidential election, the result may not be true for elections in different years. Additionally, there could be bias in sampling as there is a specific population that did not vote or is not willing to provide information which could lead to an inconclusive result for the population. Moreover, one of the internal validity concerns is confounding variables that are correlated to the outcome variables or predictors which can cause bias in the parameters of the model

### 5.4 Next step:

To address the limitations of the research, we will examine the population to ensure that the data that is used to train the regression model is representative of the whole population and perform specific methods to address missing data. Additionally, we'll perform the same analysis on data in different presidential elections to test if the result is conclusive for data in different periods. Lastly, we'll gather other demographic factors that can affect predictors in order to address any bias in parameters.

## Reference

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

———. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://marginaleffects.com/.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://github.com/bbolker/broom.mixed.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Center for American Women and Politics (CAWP). 2024. "Gender Gap: Voting Choices in Presidential Elections." Rutgers University. 2024. https://cawp.rutgers.edu/gender-gap-voting-choices-presidential-elections.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press. http://statwww.epfl.ch/davison/BMA/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://docs.ropensci.org/gutenbergr/.

Kay, Matthew. 2023. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Krantz, Sebastian. 2024. *Collapse: Advanced and Fast Data Transformation in r*. https://doi.org/10.5281/zenodo.8433090.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories*.

Pew Research Center. 2016. "Educational Divide in Vote Preferences on Track to Be Wider Than in Recent Elections." Pew Research Center. 2016. https://www.pewresearch.org/short-reads/2016/09/15/educational-divide-in-vote-preferences-on-track-to-be-wider-than-in-recent-elections/.

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