Understanding the Socio-Economic Determinants of Voter Behavior: An Analysis of the 2020 U.S. Presidential Election*

Hannah Yu Mary Cheng Yimiao Yuan

March 12, 2024

This paper delves into the dynamics of voter turnout of the 2020 United States presidential election. Utilizing dataset from the Cooperate Election Study (CES2020), we employ logistic regression analysis to explore the influence of the voter's race, region, and employment on their vote for presidential candidate. Our findings reveal distinct voting patterns among different demographic groups, with minorities showing strong support for Biden, the South predominantly favored Trump, and individuals reliant on unemployment benefits preferred Biden. These insights stress the need for customized outreach and policies to enhance voter engagement across diverse U.S. communities.

Table of contents

*Code and data are available at: https://github.com/hannahyu07/US-Election

1 Introduction

Four years ago, the United States once again was at the forefront of an important electoral contest between the Democratic and Republican parties in the 2020 presidential election. The Democratic and Republican parties are represented by Donald Trump and Joe Biden respectively. Donald Trump achieved a remarkable victory in the 2016 election against Hillary Clinton. On the other hand, Joe Biden, the former vice-president under Barack Obama's administration from 2008 to 2016, led the Democratic campaign. He aimed to regain control from the Republican party. In the end, Biden won the election by 51.3% majority votes (CNN (2020)), which Trump questioned its credibility. Due to the COVID-19 pandemic in 2020, the vote system changed to in-mail voting instead of in-person voting, which was why Trump believed that some votes were stolen (Longwell (2022)). Therefore, it is important for us to analyze why people voted for Biden and Trump in 2020 election and how various demographic, economic, and social factors may influence individuals' voting decisions. Perhaps then we could see why Biden eventually won. We analyze the 2020 election data set from Cooperative Election Study (CES) (Schaffner, Ansolabehere, and Luks (2021)) and build a prediction model for individual voting preference. This allows us to discover trends and address potential political, economic, and societal situations that may cause the eventual winning for Biden.

In this paper, a logistic regression model is used to forecast the outcome of the 2020 election, with data from the Cooperative Election Study. Logistic regression is a great choice since it is used to predict binary outcomes, such as election results (Trump or Biden). Our analysis focuses on estimating the likelihood of victory for either Trump or Biden, based on various demographic, geographic, and socioeconomic factors captured in the CES data set.We selected data features:race, region that they live, employment status, and who they voted for. The estimand in this paper is the number of people who support Trump and Biden in reality. However, it is difficult to measure the exact number of people who support Trump and Biden since there are millions of people in the United States and not all of them with the rights to vote would actually vote. Therefore, in this paper, we attempt to estimate the estimand using a logistic regression model which is trained using sample dataset from the 2020 election dataset from CES.

The logistic regression model shows that in the 2020 presidential election, Biden has strong support from racial groups including black, Hispanic, and Asian populations. Geographically, Biden gains greater support in the Northeast and West regions, while Trump secured more votes in the South. Additionally, both full-time and part-time workers, as well as students and those facing unemployment or layoffs, lean towards Biden.

The remainder of this paper is structured into different sections. Section 2 demonstrates the data used for our report and includes some tables and graphs to illustrate the different groups of people in our data. Section 3 builds the model and discusses its justification and explanation. Section 4 highlights the results of the predictions using tables and graphs. Section 5 contains discussions that conducted based on the findings, which addresses the voting prediction results based on race, region, employment status, and the influence of COVID-19 and in-mail voting systems.

2 Data

2.1 Data Source

In this report, we use the 2020 Cooperative Election Study (CES) (Schaffner, Ansolabehere, and Luks (2021)) as our primary dataset. This is a long-established annual political survey of US. CES contains information about how Americans view Congress and hold their representatives accountable in elections, how they vote and their electoral experiences, and how their behaviors and experiences vary with political geography and social context. In 2020, there were 61,000 American adults completed the survey.

2.2 Features

The original CES 2020 dataset, which shows in **?@tbl-raw**, contains 61000 observations and many variables. We chose to feature these 5 variables: “votereg”, “CC20_410”, “race”, “re-

gion", "employ", in our analysis.

1. votereg: whether the respondent is registered to vote.
2. CC20_410: the presidential candidate the respondent voted for.
3. race: the census region where the respondent lives.
4. region: the racial or ethnic group of the respondent.
5. employ: the current employment status of the respondent.

2.3 Data Measurement

Our dataset's quantitative representation of real-world phenomena is influenced by several factors, which inherently limit the accuracy of our measurements for the five variables.

Firstly, for the variable "votereg" indicating whether a respondent is registered to vote, the measurement process involves relying on self-reported information from survey respondents. The US voter registration system is known to be inaccurate and inefficient. While efforts are made to ensure accuracy, there may be instances of misreporting or misunderstanding of voter registration status, leading to potential inaccuracies in the data. According to (*Inaccurate, Costly, and Inefficient: Evidence That America's Voter Registration System Needs an Upgrade*, n.d.), there are millions of voter registrations that are no longer valid or inaccurate. There may also be cases when individuals falsely believed they are registered to vote because they reached the voting age but did not actually register at their local office.

Regarding "CC20_410," representing the presidential candidate the respondent voted for, data collection relies on self-reporting, which is subject to recall bias and social desirability bias. While the data collection occurred before the notorious January 6, 2021 United States Capitol Attack that resulted in international criticism and increased the unpopularity of Trump in mainstream media, Trump had made many speeches and taken actions throughout his presidency that contributed to his polarizing image. Among many voters, voting for Trump became associated with characteristics such as being uneducated, backward-thinking, and anti-science. As a result, individuals who voted for him may have felt social pressure to conceal or lie about their support, potentially leading to under-reporting of votes for Trump in the survey data.

For the variable "race," which denotes the racial or ethnic group of the respondent, categorization relies on self-identification. However, racial identity is complex and can be influenced by cultural, social, and historical factors. Mixed-race individuals, for instance, may choose to identify with one racial group over another based on social preferences or personal experiences. Moreover, the race variable includes a category for "two or more races" to classify individuals identifying with multiple racial backgrounds. While this acknowledges the diversity within

the population, it may not capture the nuances of each individual's identity. Additionally, the absence of a distinct category for "Indians or South Asians" presents a limitation. While these individuals may technically fall under the category "Asian," many Indian and South Asian Americans may prefer to identify separately due to differences in appearance and cultural background. The term "Asians" are more generally associated with East Asians such as Chinese, Japanese, and Koreans. As a result, there's a potential for misclassification, with some individuals opting to report themselves as belonging to the "other" race category instead.

In the case of "region," which represents the census region where the respondent lives, data collection relies on categorization based on geographical location. However, it's important to note that regional boundaries can sometimes be arbitrary and may not fully capture the diverse cultural, economic, and social affiliations of individuals. Additionally, people living in bordering states or areas near regional boundaries may have affiliations with multiple regions or may identify more strongly with a neighboring region. For example, states like Missouri and Kentucky are often considered part of the Midwest, but they also share cultural and economic ties with the South. This complexity can lead to challenges in accurately categorizing individuals based on their region of residence, potentially resulting in oversimplification or misrepresentation of regional identities and characteristics.

The "region" variable categorizes respondents based on their geographical location into four census regions: Northeast, Midwest, South, and West. However, regional boundaries can be arbitrary, and individuals living in bordering areas may have affiliations with multiple regions or identify more strongly with a neighboring region. For instance, states like Missouri and Kentucky are often considered part of the Midwest but also share cultural and economic ties with the South. This complexity can make individuals accurately reporting their region challenging.

Finally, for "employ", indicating the respondent's self-reporting employment status, can be influenced by various factors such as job changes, seasonal work, and personal circumstances. In addition, individuals may intentionally misreport their employment status due to societal pressures or to access benefits. For instance, some may falsely claim employment to avoid stigma, while others may report unemployment to qualify for unemployment benefits. Additionally, students working part-time jobs may selectively report their status based on their perceived benefit. These discrepancies can introduce bias and inaccuracies into the data, impacting the reliability of our analyses.

2.4 Methodology

Since it is difficult to observe such as large dataset, this report will only explore and analyze through specific aspects. The original dataset contains respondent's voter registration status and the presidential candidate the respondent voted for. This paper will only analyze respondents who are registered to vote and focus on the outcome of two candidates, Joe Biden representing the Democratic party and Donald Trump representing the Republican party.

The dataset is cleaned by renaming the column names, specifying the class of the columns, and changing the numbers in the table to the corresponding description in the codebook to improve the readability. In addition, NA employment status is removed in the cleaned data file since there are only 34 entries, which is a small amount of missing value compared with 61000 entries. After cleaning, 43534 rows of data with 4 data features remain.

The cleaned data will be analyzed and performed using R (R Core Team (2023)) with `tidyverse` (Wickham et al. (2019)), `here` (Müller (2020)), `rstanarm` (Brilleman et al. (2018)), `modelsummary` (Arel-Bundock (2022)), `ggplot2` (Wickham (2016)), `knitr` (Xie (2014)), `marginaleffects` (Arel-Bundock (2024)), `plotly` (Sievert (2020)), `tibble` (Müller and Wickham (2023)), `margins` (Leeper (2021)), `testthat` (Wickham (2011)) and `kableExtra` (Zhu (2021)). Table 1 shows a preview of the cleaned dataset.

Table 1: Preview of the cleaned 2020 CES dataset

voted_for	race	region	employ
Trump	White	Northeast	Permanently disabled
Biden	White	Midwest	Retired
Biden	White	Northeast	Permanently disabled
Trump	White	Midwest	Full-time
Trump	White	Midwest	Retired

Table 2 is a summary of the cleaned data, showing detailed statistics about the dataset. As we can see from the table, there are more people who support Biden. Respondents covered a wide range of races, with white people being the most heavily represented. Also, CES data came from a variety of locations, but the majority came from the southern region of the United States. The employment status of those who completed the survey varied, with most of them being full-time or retired.

Table 2: Statistics summary of the cleaned 2020 CES dataset

voted_for	race	region	employ
Trump:17548	White :33223	Northeast: 8374	Full-time :16671
Biden:25986	Black : 4123	Midwest : 9998	Retired :11726
NA	Hispanic : 3051	South :16184	Part-time : 4145
NA	Asian : 1150	West : 8978	Permanently disabled: 2933
NA	Middle Eastern : 858	NA	Homemaker : 2702
NA	Two or more races: 755	NA	Unemployed : 2655
NA	(Other) : 374	NA	(Other) : 2702

3 Model

3.1 Model Set-up

In our analysis, we utilized a Bayesian logistic regression model to examine the relationship between voter preferences and various demographic and socioeconomic factors. The model is formulated as follows:

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) && (1)\\
\text{logit}(\pi_i) &= \alpha + \beta_1 \times \text{race}_i + \beta_2 \times \text{region}_i + \beta_3 \times \text{employ}_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)
\end{aligned}
$$

In this model, y_i represents the binary outcome variable indicating whether an individual voted Biden (as opposed to Trump). The probability of voting for the Biden (π_i) is modeled using a logistic link function ($\text{logit}(\pi_i)$), which is a linear combination of the intercept (α) and the coefficients (β_1, β_2, β_3)) corresponding to the predictor variables race, region, and employment status, respectively. These predictor variables are denoted as race_i, region_i, and employ_i, where i indexes the individuals in the dataset.

The intercept (α) and coefficients (β_1, β_2, β_3) are assigned informative prior distributions to regularize the model. Specifically, we assume a normal distribution with a mean of 0 and a standard deviation of 2.5 for each parameter.

We chose this modeling approach for several reasons. Firstly, logistic regression is well-suited for binary outcome variables, making it appropriate for analyzing voting behavior. Additionally, Bayesian methods allow us to incorporate prior knowledge and uncertainty into our analysis, providing more robust estimates of the model parameters.

Alternative modeling approaches, such as linear regression models, were also considered. However, we chose Bayesian logistic regression because our result is a binary variable of voter's decision.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Brilleman et al. (2018). We use the default priors from `rstanarm`. Rstanarm employs Markov chain Monte Carlo (MCMC) techniques to estimate the posterior distribution of the parameters. To avoid exessive runtime, we randomly sampled 1000 observations to fit the model. Model diagnostics, including convergence checks and posterior summaries, are available in the supplementary materials (see -Section B).

3.1.1 Model Justification

We expect a positive relationship between individuals of Black, Asian, and Hispanic ethnicities and support for Biden. This expectation arises from Trump's history of spreading polarizing language and anti-immigrant sentiments, as well as his controversial plans such as building a border wall. These groups are more likely to align with Biden's policies, which prioritize inclusivity and diversity. White individuals with traditional family values and conservative leanings tend to support Trump. They are drawn to his emphasis on preserving traditional values and promises to uphold conservative principles, especially regarding immigration, law and order, and gun rights.

Conversely, we anticipate a negative relationship between voters in the South and Midwest regions and support for Biden. These regions have a stronger conservative presence and a history of supporting Republican candidates like Trump. States such as Texas and Florida, which are known Republican strongholds, are located in the South. Therefore, individuals in these regions may be less inclined to support Biden's progressive agenda.

Regarding employment status, we expect students, unemployed individuals, and those temporarily laid off to be more inclined to support Biden. Students are often exposed to diverse perspectives and progressive ideas in educational settings, making them more likely to endorse Biden's platform. Unemployed and laid-off individuals may favor Biden due to the Democratic Party's advocacy for social welfare programs and support for workers' rights.

The voting behavior of employed individuals is harder to distinguish. Some working individuals support Trump due to their opposition to higher taxes and prefer his promises of tax cuts and economic deregulation. Conversely, others lean towards Biden because they believe tax increases should primarily target the wealthy and not burden the middle class. Additionally, educated and liberal-leaning working professionals may prioritize issues such as healthcare, climate change, and social justice, aligning them with Biden's platform.

4 Results

4.1 Data Result

Figure 1 shows the relationship between race and voting preference. Overall, Biden gets more support from most racial groups, except for Native Americans and people of "two or more races," who tend to support Trump slightly more.

Figure 2 shows the more detailed relationship between race, region, and voting preference. In the Northeast, Midwest, and West regions of the United States, there is a consistent pattern where individuals across almost all racial groups tend to show greater support for Biden over Trump. Conversely, in the South, there is a distinct divide, with White individuals predominantly favoring Trump, while almost all other racial groups support Biden more than Trump.

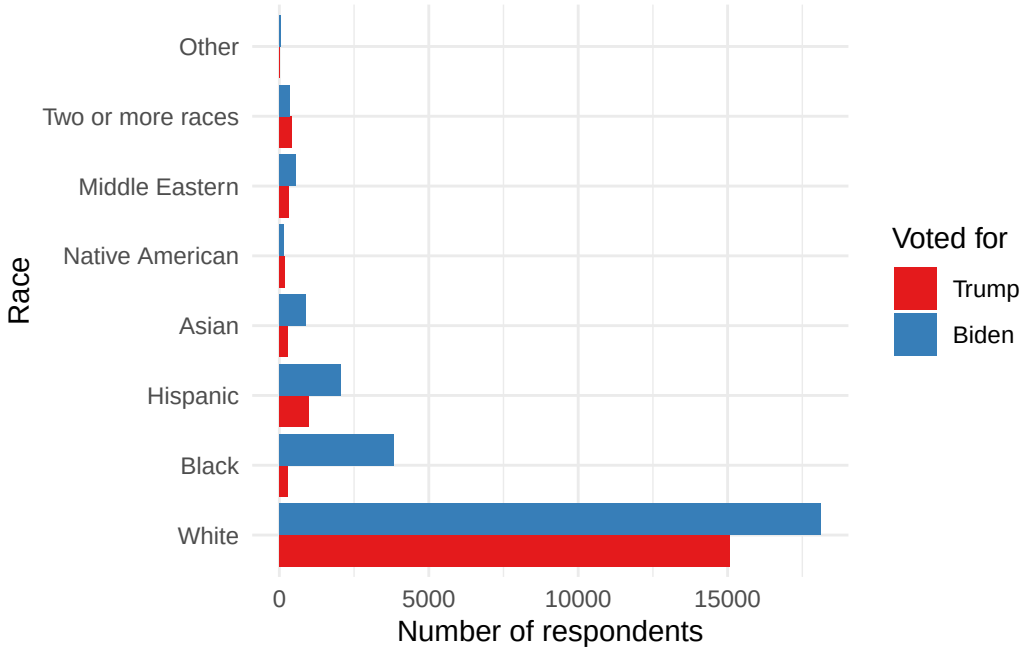


Figure 1: The distribution of presidential preferences by race

Overall, the four subplots all illustrate that among Black, Asian, and Hispanic communities, the proportion of support for Biden exceeds that for Trump. However, it is evident that White individuals constitute a significantly larger portion of the population compared to other racial groups, thereby exerting a greater influence on voting outcomes.

Figure 3 shows the relationship between employment and voting preferences. Across various employment statuses, there is a consistent trend of higher support for Biden compared to Trump. It can be observed that individuals holding full-time jobs show the most disparity in support between the two candidates, with Biden gaining substantially more support than Trump. The support for Biden and Trump among those with other employment statuses is relatively similar, with the support for Biden slightly higher than for Trump.

Figure 4 illustrates a more detailed relationship between region, employment status, and voting preferences for Trump and Biden. Across all regions, people in full-time and part-time employment show a greater percentage of support for Biden than to Trump. Among retired people, support for Trump and Biden is roughly comparable: with a slight tilt towards Biden in the Northeast, Midwest, and West, and towards Trump in the South. In addition, people with other employment statuses generally favor Biden over Trump, except for certain group (Homemaker) in the South.

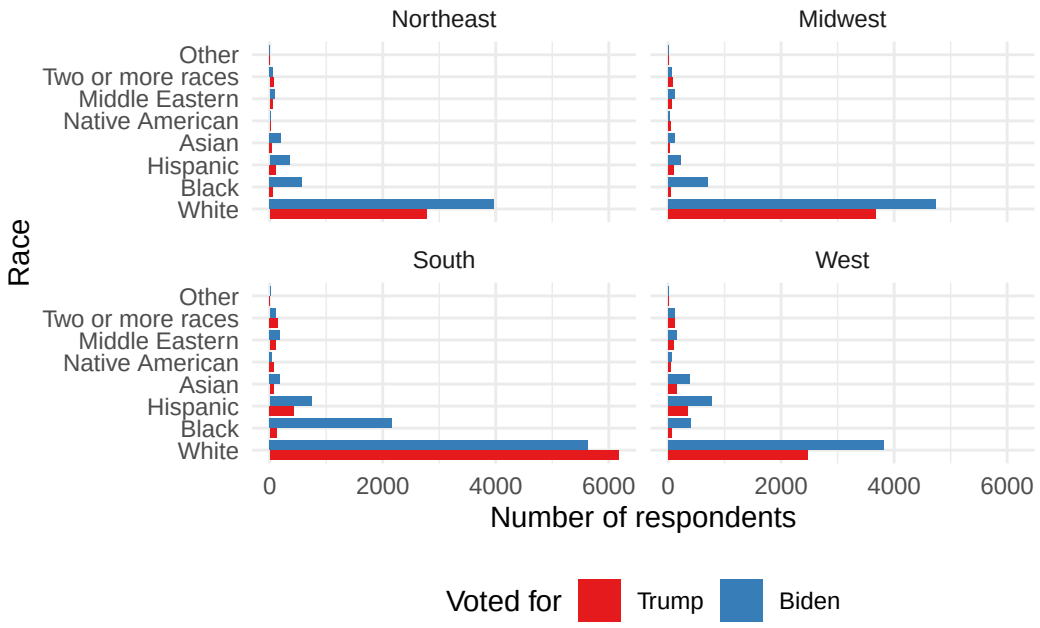


Figure 2: The distribution of presidential preferences, by region and race

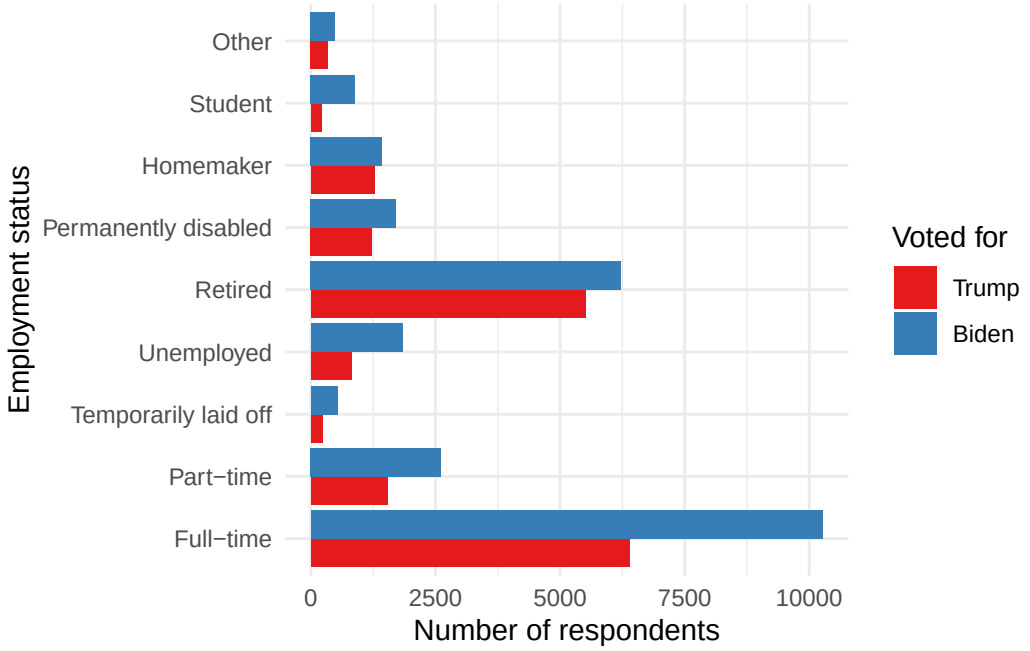


Figure 3: The distribution of presidential preferences by race

Table 3: Explanatory models Political Preferences (n = 1000)

	Support Biden
(Intercept)	0.806
	(0.438)
raceBlack	2.549
	(0.683)
raceHispanic	0.039
	(0.503)
raceMiddle Eastern	−0.162
	(0.649)
raceNative American	−2.791
	(1.288)
raceOther	37.999
	(33.885)
raceTwo or more races	−1.900
	(0.769)
raceWhite	−0.650
	(0.436)
regionNortheast	0.436
	(0.217)
regionSouth	−0.161
	(0.186)
regionWest	0.137
	(0.206)
employHomemaker	−0.187
	(0.297)
employOther	0.298
	(0.557)
employPart-time	0.433
	(0.270)
employPermanently disabled	−0.273
	(0.295)
employRetired	−0.200
	(0.168)
employStudent	0.831
	(0.501)
employTemporarily laid off	0.699
	(0.553)
employUnemployed	0.416
	(0.331)
Num.Obs.	1000
R2	0.118
Log.Lik.	−606.898
ELPD	−627.2
ELPD s.e.	11.8
LOOIC	1254.4
LOOIC s.e.	23.6
WAIC	1253.7
RMSE	0.46

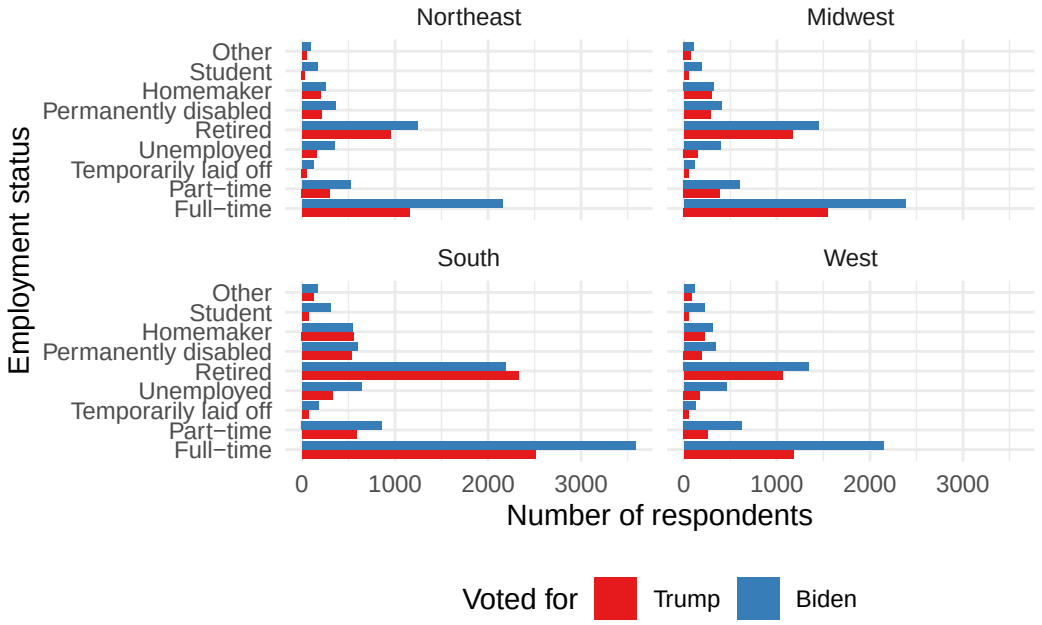


Figure 4: The distribution of presidential preferences, by region and employment status

4.2 Model Results

Our results are summarized in Table 3. Our results generally matches our expectation. To avoid multicollinearity, the model excluded three variables from each category: race Asian, region Midwest, and employed full-time.The intercept represents the estimated log-odds of supporting Biden when all other predictors are held constant at their reference levels. In this case, the estimated log-odds of supporting Biden for individuals who are Asian, live in the Midwest, and are employed full-time is 0.806.

Black individual's support for Biden is large. The estimated coefficient of 2.549 suggests that, holding all other variables constant, Black individuals are estimated to have a 2.549 unit increase in the log-odds of supporting Biden compared to the reference group. Hispanic races on average are also more likely to vote for Biden. We observe an outlier, individuals with race "other" has a 37.999 higher log-odds of supporting Biden compared to the reference group.

Voters preference for candidates also vary by regions and their employment status. Biden has more supports in the Northeast and West region while Trump gains a majority of his votes from the South. As anticipated, students exhibit strong support for Biden. Additionally, individuals who are laid off, unemployed, or employed part-time, potentially reliant on social benefits, also prefer for Biden.

?@fig-modelresults1 shows range of coefficient estimates of our model within the 90% prob-

ability. Due to the fact that the credibility interval for race "Other" is particularly large, we cannot observe the trend of the 90% credibility intervals of other variables. Therefore we created figure # with the x axis limited from -10 to 10.

Combing **?@fig-modelresults1** and **?@fig-modelresults2**, we observe statistical significance for the coefficient estimates for students, the Northeast region, individuals identifying with two or more races, individuals identifying with "Other" racial category, Native Americans, Black individuals, and the intercept, Asians in the Midwest region who are employed full-time. The estimates are significant if the intervals do not cross 0. The value for the estimates are in log-odds, indicating that if the coefficient is positive, the individual supports Biden, if negative, the individual supports Trump.

5 Discussion

5.1 Racial Variations

In our analysis, respondents of different races show varying levels of support, with a majority of blacks, Hispanics, and Asians supporting Biden, and a majority of Native Americans and multi-racials supporting Trump. The results for whites are more evenly distributed, with about the same percentage supporting Biden and Trump.

The reason for the majority of Hispanics supporting Biden may be due to the US-Mexico border wall that was built during Trump's reign. The wall was built over 452 miles, which replaced the old fence (Rodriguez (2021)). With this measure in place, Trump lost most of the Hispanic vote. Trump claims the wall reformed the immigration system and achieved the most secure southern border in US history. However, that wasn't true, and with it has come a surge in illegal border crossings. For Hispanics, Trump's long history of anti-immigrant rhetoric and this border wall have denied them the opportunity to seek asylum. While crossing the border illegally is not something to be promoted, it is the last hope for those living in dire straits.

2020 is the age of pandemics, as Trump's phrase "Chinese virus" led to a large number of Asians starting to support Biden. This is an extremely anti-Asian statement: although COVID-19 was first found in China, but it does not mean that virus was from China. Trump's comments quickly festered in public opinion and led to a dramatic increase in hate incidents in the Asian American community (Reja (2021)). The repeated use of the "Chinese Virus" by a president of the country has contributed to an environment of racial discrimination and hatred against Asians, which has directly led to more Asians supporting Biden, whose policies are more open and democratic, and who values racial equality more.

Most white people vote relatively evenly, due to the fact that they are not affected by immigration or other policies. Whites hold a major voice and position in the US, most of the ideas and policies Trump proposes are favorable to whites. Therefore, he can gain support from most whites. However, there are some whites who want a democratic and equal environment, and at the same time do not want to see racial discrimination and other phenomena, so they will choose to support Biden, who values equality.

5.2 Regional Variations

Our analysis highlights regional voting variations, with the Northeast, West, and Midwest favoring Biden, while the South supports Trump. These differences stem from diverse cultural, family, religious, and social perspectives.

Biden's support base is concentrated in the Mid-Atlantic states, New England, the West Coast, Southwestern states, and certain regions around the Great Lakes. This support is particularly evident in cities within these regions, driven by various factors including liberal ideals and

progressive policies. For instance, states like New York and Massachusetts support Biden due to their liberal-leaning populations, diverse demographics, and emphasis on social justices. Similarly, on the West Coast, states such as California and Washington are known for their progressive values and advocacy for issues like environmental conservation and LGBTQ+ rights, contributing to Biden's popularity in these areas. In the Southwest, states like Arizona and Nevada, with growing minority populations and concerns over immigration policies, also show significant support for Biden. Additionally, states around the Great Lakes, including Illinois and Michigan, align with Biden's platform due to their industrial heritage and strong labor unions.

Voters favoring Biden are predominantly concentrated in big cities and large suburban regions, whereas Trump supporters are dispersed across rural areas with lower population density. Consequently, Trump secured victories in 2,588 counties, whereas Biden won over just 551 (Frey 2021). Even in traditionally Republican states, major urban areas exhibit significant support for Biden. Notable examples include Charlotte, North Carolina; Salt Lake City, Utah; and Nashville and Memphis, Tennessee, among others (CNN 2020).

This trend reflects the impact of diverse demographics, progressive ideologies, and evolving societal norms in urban areas. Here, a mix of cultural diversity, acceptance of progressive values, and socio-economic factors favors Biden's platform. Additionally, urban centers, with their concentration of educational and cultural institutions, often align with Democratic policies on healthcare, climate change, and social justice.

The 2020 election witnessed intriguing shifts in traditionally conservative states, turning them into battlegrounds. For instance, Texas, a Republican stronghold, saw unprecedented Democratic gains, though ultimately remaining red. Notably, the five largest cities in Texas—Houston, San Antonio, Dallas, Austin, and Fort Worth—all backed the Democratic party ("Texas 2020 Election Results" 2020). Similarly, Georgia, a Republican stronghold since 1992, voted for the Democratic party, reflecting evolving demographic and ideological shifts ("Georgia 2020 Election Results" 2020).

Southern states, such as Alabama, Mississippi, and South Carolina with a large amount of Christian population, strongly adhere to conservative principles. For example, these states have implemented strict anti-abortion law and advocated for tight immigration policies. Furthermore, these states have historically been less supportive of LGBTQ+ rights. As a result, voters in these Southern states typically align with the Republican party represented by Trump, which prioritizes pro-life initiatives and traditional Christian values.

5.3 Employment Variations

maybe talk about employment? tax

5.4 COVID and Mail-in Ballots

In addition to traditional socio-economic factors like race, region, and employment, the unique circumstances surrounding the COVID-19 pandemic and the widespread use of mail-in ballots significantly influenced voter turnout and led to Biden's election victory. Trump's handling of the pandemic, including his skepticism towards mask-wearing and vaccination and disagreements with public health guidance, drew many criticism (Pollitz, Long, and Freed 2020). Many dissatisfied Republican voters voted for Biden instead.

Mail-in ballots also played a key role in shaping the election landscape. It increased voter participation, particularly among individuals who were reluctant to vote due to logistical constraints or health concerns amid the pandemic (Amlani and Collitt 2022). Mail-in ballots raised many election controversies. Initially, Trump led in some states, but as mail-in ballots were counted, the tide often turned in favor of Biden. Trump and his supporters raises concerns about the legitimacy of the ballots and claimed the election is "rigged" (Axelrod 2022). However, there is no solid proof that the election is "rigged".

5.5 Weaknesses and Next Steps

One weakness of using a logistic regression model is its inability to predict voting turnout for candidates beyond the Biden or Trump. Expanding the analysis to include other candidates would enhance its comprehensiveness. Furthermore, the reliability of our findings is based on the accuracy of the survey data. Respondents may provide inaccurate or misleading information, intentionally or unintentionally, leading to biased estimates and unreliable predictions. Due to many controversial opinions and actions of Trump and his supporters, many Republican who voted for him might be reluctant to admit their voting in the survey. Additionally, missing data (N/As) can introduce further challenges, as logistic regression requires complete data for accurate modeling.

Furthermore, it's important to recognize that our findings establish correlation, not causation. Because we only controlled few variables for election turnout, our results may lack essential variables and thus suffer from omitted variable bias. Critical factors such as the impact of mail-in ballots, the role of social media, and broader socio-political dynamics that may have impacted the election are not incorporated in the model.

For further studies, we could start by incorporating more socio-economical indicators such as income levels, education, and urbanization rate that attribute to the election turnout. Analyzing sentiment from social media platforms, news articles, and public forums could offer a real-time perspective on voter attitudes leading up to the election. We could also add in state dummies to observe for election turnouts by state. To infer causality, future research should employ advanced methodologies capable of addressing issues like selection bias and heterogeneity treatment bias.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 5, we implement a posterior distribution. This compares who people voted for in reality with the prediction results from the posterior distribution from our logistic regression model. It can be observed that the posterior distribution fits perfectly with the actual data, it suggests that the model accurately captures the observed data patterns. So the posterior is able to generate simulated data that closely resembles the actual data (Gelman and Modrák (2020)). This is a positive sign because it indicates that our logistic regression model is a good representation of the actual voting preferences in the 2020 election data.

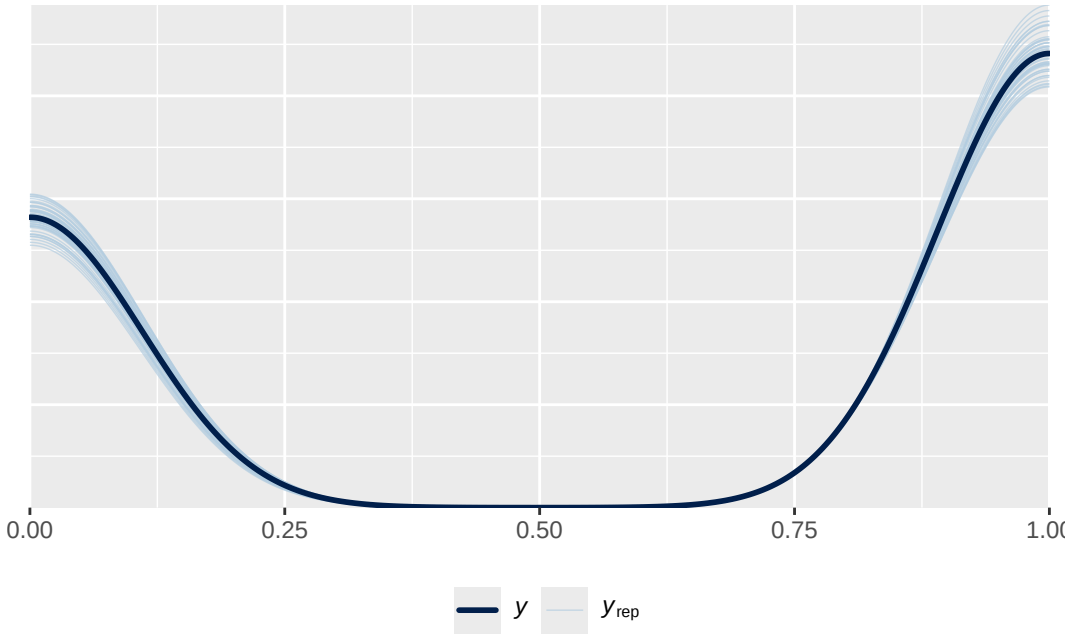


Figure 5: Posterior distribution for logistic regression model

Figure 6 compares the posterior with the prior. This compares the prior distribution of parameters with the posterior distribution of parameters in our logistic regression model. We can see that majority of the model parameters do not change even after data are taken into account. This shows that the observed data matches with our initial belief and expectation

about the voting preferences for 2020 presidential election. However, for "raceOther", the posterior distribution shifts from its prior after we input observed data. This is likely suggesting that the observed data for "raceOther" strongly contradicts our initial belief. Yet, this is not a big problem since the percentage of the racial group "raceOther" is 0.86% (473/43534 from Table 2) in our dataset.

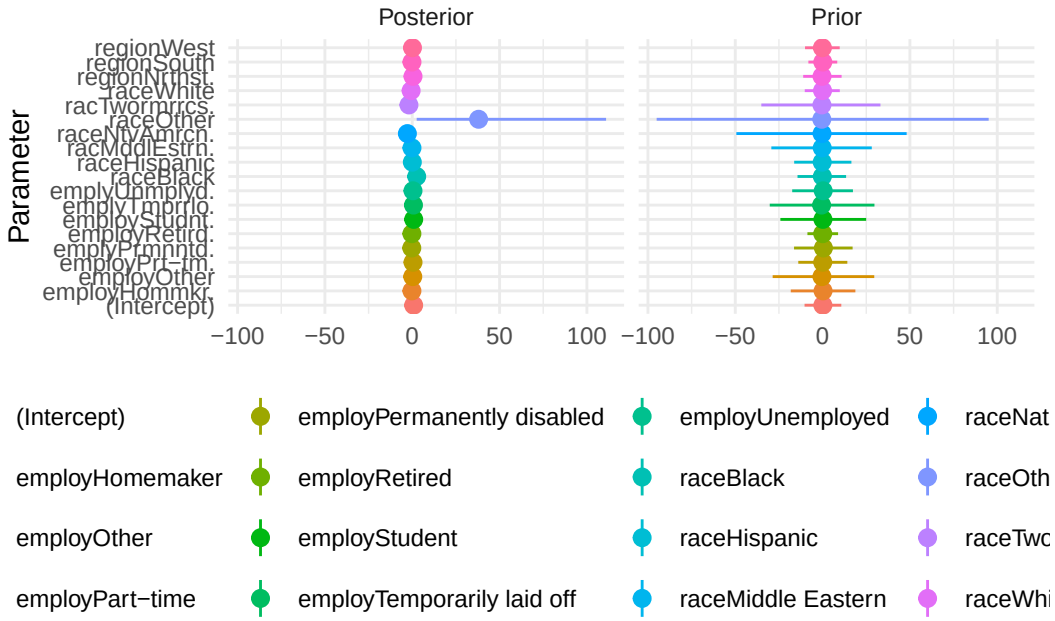


Figure 6: Comparing the posterior with the prior

B.2 Credibility interval

B.3 Credibility Interval

```
# #| echo: false
# #| eval: true
# #| warning: false
# #| message: false
# #| label: fig-modelresults1
# #| fig-cap: "Credible intervals for predictors of support for Biden 1"
#
# modelplot(political_preferences1, conf_level = 0.95, size = 0.1) +
#   labs(x = "90% credibility interval")
```

Figure 7: Comparing the posterior with the prior

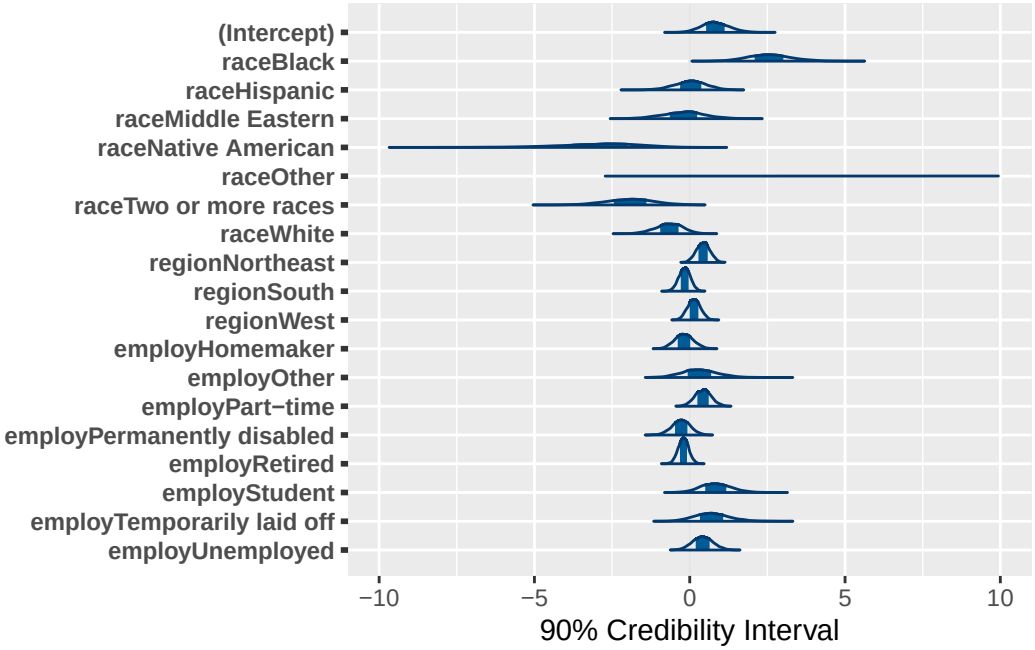


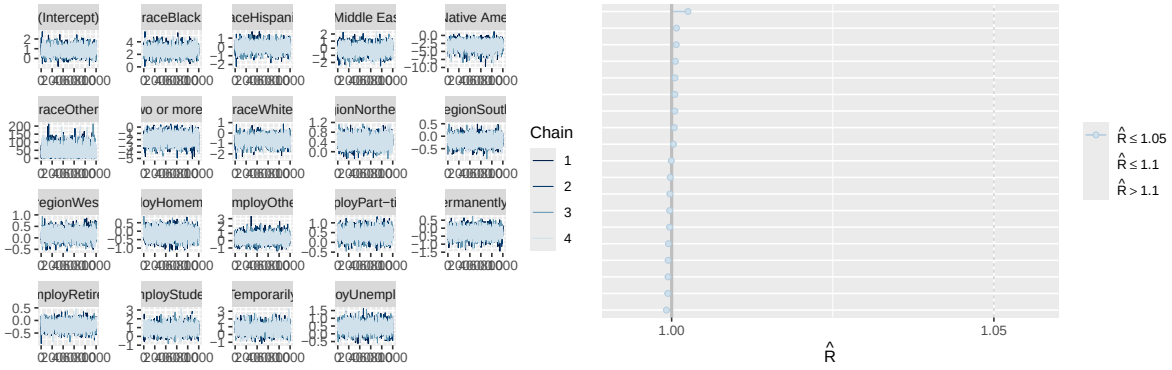
Figure 8: Comparing the posterior with the prior

```
# #| echo: false
# #| eval: true
# #| warning: false
# #| message: false
# #| label: fig-modelresults2
# #| fig-cap: "Credible intervals for predictors of support for Biden 2"
# # Create the model plot
# model_plot <- modelplot(political_preferences1, conf_level = 0.9)
#
# # Modify the x-axis limits
# model_plot + xlim(-5, 5) +  # Adjust the limits as needed
#   labs(x = "90% Credibility Interval")
```

B.4 Diagnostics

Figure 9a is a trace plot. It shows... This suggests...

Figure 9b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 9: Checking the convergence of the MCMC algorithm

References

Amlani, Sharif, and Samuel Collitt. 2022. "The Impact of Vote-by-Mail Policy on Turnout and Vote Share in the 2020 Election." *Election Law Journal: Rules, Politics, and Policy*, February. https://doi.org/https://doi.org/10.1089/elj.2021.0015.

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://CRAN.R-project.org/package=marginaleffects.

Axelrod, Tal. 2022. "A Timeline of Donald Trump's Election Denial Claims, Which Republican Politicians Increasingly Embrace." *ABC News*. https://abcnews.go.com/Politics/timeline-donald-trumps-election-denial-claims-republican-politicians/story?id=89168408.

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/.

CNN. 2020. "2020 Presidential Election Results." *CNN*. https://www.cnn.com/election/2020/results/president.

Frey, William. 2021. "Biden-Won Counties Are Home to 67 Million More Americans Than Trump-Won Counties." *Brookings*. https://www.brookings.edu/articles/a-demographic-contrast-biden-won-551-counties-home-to-67-million-more-americans-than-trumps-2588-counties/.

Gelman, Aki Vehtari, Andrew, and Martin Modrák. 2020. "Bayesian Workflow." https://doi.org/10.48550/arXiv.2011.01808.

"Georgia 2020 Election Results." 2020. *CNN*. https://www.cnn.com/election/2020/results/state/georgia.

Inaccurate, Costly, and Inefficient: Evidence That America's Voter Registration System Needs an Upgrade. n.d. https://www.pewtrusts.org/~/media/legacy/uploadedfiles/pcs_assets/2012/pewupgradingvoterregistrationpdf.pdf.

Leeper, Thomas J. 2021. *Margins: Marginal Effects for Model Objects*.

Longwell, S. 2022. "Trump Supporters Explain Why They Believe the Big Lie. The Atlantic." https://www.theatlantic.com/ideas/archive/2022/04/trump-voters-big-lie-stolen-election/629572/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Pollitz, Karen, Michelle Long, and Meredith Freed. 2020. "Comparing Trump and Biden on COVID-19." *KFF*. https://www.kff.org/coronavirus-covid-19/issue-brief/comparing-trump-and-biden-on-covid-19/.

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

Reja, Mishal. 2021. "Trump's 'Chinese Virus' Tweet Helped Lead to Rise in Racist Anti-

Asian Twitter Content: Study." *ABC News*. ABC News Network. https://abcnews.go.com/Health/trumps-chinese-virus-tweet-helped-lead-rise-racist/story?id=76530148.

Rodriguez, Sabrina. 2021. "Trump's Partially Built 'Big, Beautiful Wall." *POLITICO*. https://www.politico.com/news/2021/01/12/trump-border-wall-partially-built-458255.

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

"Texas 2020 Election Results." 2020. *CNN*. https://www.cnn.com/election/2020/results/state/texas.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 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.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.