Young age groups, women, and lower levels of education are more likely to vote Democractic*

Victor Ma

March 16, 2024

In this study, I investigate demographic patterns which influence political support for the US 2020 election, specifically focusing on whether demographics of people tend to vote for Joe Biden or Donald Trump. I use a logistic regression model with the variables age, gender, and education to predict which categories in each demographic have the highest chance of voting for Joe Biden. I find that the model accurately predicts the trend of younger people, women, and those with the lowest level of education being more likely to support Joe Biden.

1 Introduction

The 2020 United States presidential election was not only a pivotal political event but also a reflection of a nation grappling with unprecedented challenges. The backdrop of the COVID-19 pandemic, which affected millions and significantly altered everyday life, coupled with nationwide protests following the death of George Floyd, underscored the deep social and political divisions within the country (Foundation 2020; Cohn 2020). These events created a heightened sense of importance around the election that had potential to shape the direction of national policies. 2020 was a year of heightened misinformation and polarization, with many people challenging the electorate's ability to navigate complex issues and make informed decisions (Lazer et al. 2018; Allcott and Gentzkow 2017).

Understanding how demographics are correlated with political alignment in such a pivotal year reveals a lot about societal trends. For example, younger voters might prioritize progressive issues like climate change and social justice, signal a shift towards liberalism. Older voters, emphasizing economic and healthcare concerns, show more conservative preferences. Gender disparities in voting reveal women's tilt towards equality-focused candidates, while varying

*Code and data are available at: https://github.com/bestmustard/Determinants-of-Political-Support-in-the-United-States.

education levels demonstrate that higher education correlates with liberal views, highlighting the role of informed decision-making in political orientation. This demographic analysis not only maps the political landscape of 2020 but also teaches us about evolving societal values.

In this study I use a logistic regression model in order to predict which candidate will win more votes based on demographics such as age, gender, and level of education. The results of this model gives us key insights into how much demographic variables can influence voter preferences in a polarized climate. In particular, logistic regression models are used to model binary outcomes, so our outcome will be whether Joe Biden or Donald Trump is forecasted to have more votes.

The data we are using is the 2020 Cooperative Election Study (CCES), an American stratified sample survey administered by YouGov (Schaffner, Ansolabehere, and Luks 2021). The data is directly sourced from the Harvard dataverse, an online data repository running on the open-source web application dataverse. For my purposes I will be focusing on the variables of age, gender, and level of education as explanatory variables in the logistic regression model. I will also be referencing data from similar polls collected for the 2020 election, such as from American National Election Studies, Gallup Polls, and Pew Research Surveys (American National Election Studies 2021; Igielnik, Keeter, and Hartig 2021; "How Does Gallup Polling Work?" 2024)

My report is structured into four main sections following the introduction. In the first section, I describe the data utilized for my analysis, highlighting the CCES 2020 dataset and presenting graphs that show the distribution of key demographic variables. The second section details the logistic regression model, including the rationale for its use and an interpretation of preliminary findings. Next I will analyse the demographic factors' impact on election outcomes through the use of graphs and specific numerics from my results. Finally, I discuss the implications of my findings, address potential weaknesses in my study, and suggest directions for future research.

This analysis is conducted using R, using several R packages to facilitate my analysis and presentation. This includes tidyverse for data manipulation and visualization, dataverse for accessing the CCES data, knitr for report generation, modelsummary for model interpretation, and rstanarm for Bayesian regression modeling (R Core Team 2021; Leeper 2021; Xie 2021; Arel-Bundock 2021; Goodrich et al. 2022; Kay 2021; Wickham et al. 2021).

2 Data

The dataset I used was the 2020 Cooperative Election Study (CCES) (Schaffner, Ansolabehere, and Luks 2021). This data is provided by the Harvard dataverse and is conducted yearly as a survey of US political opinions, with the 2020 iteration including 61,000 respondents Schaffner, Ansolabehere, and Luks (2021). The release also includes a full guide to the data, and the questionnaires used. Vote validation was conducted by Catalist, a largescale organization with information on over 240 million individuals in the United States (Catalist 2017). ## Strengths

2.0.1 Sample Size

Sources for Figure 1 below: (American National Election Studies 2021),(Igielnik, Keeter, and Hartig 2021),("How Does Gallup Polling Work?" 2024), (Schaffner, Ansolabehere, and Luks 2021).

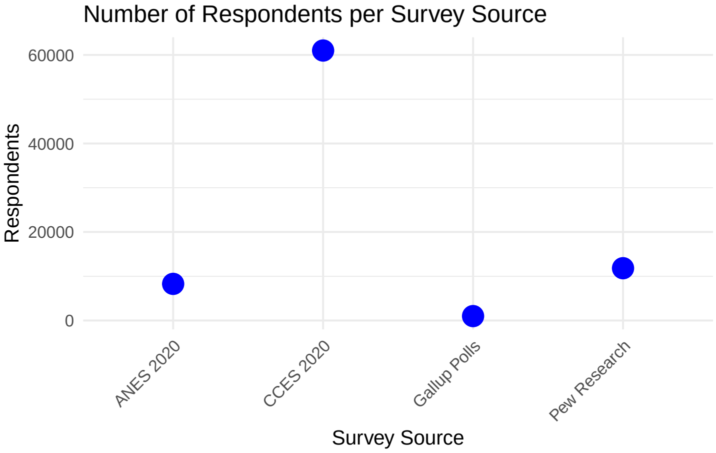

Figure 1: Number of respondents for various 2020 election survey sources

From Figure 1, we can see that the CCES poll has the largest respondent pool compared to similar surveys. This ensures that analyses can be sufficiently granular, allowing us to analyze smaller subgroups that a smaller dataset may not represent accurately (Ansolabehere and Schaffner 2020).

2.0.2 Data Collection Phases

The CCES collected data in two waves, immediately before and after the 2020 election. This helps capture the shifts in voter attitudes and preferences during the critical final stages of the election campaign (Schaffner, Ansolabehere, and Luks 2021). Pew Research and Gallup each conduct multiple polls of much smaller scale leading up to the election, which will give a snapshot in time of the voter sentiment but does not capture as broad of a perspective as the CES survey (Igielnik, Keeter, and Hartig 2021), ("How Does Gallup Polling Work?" 2024).

2.0.3 Sampling Methodology

The sampling methodology and vote validation protocols are well outlined in the guide. This dataset employs sample matching, a two-step process designed to create a survey sample that represents a larger population as closely as possible. First, a sample that reflects the broader population (called the target sample) is selected based on certain characteristics like age, race, and gender. Then, for each person in this target sample, someone with the closest matching profile from a pool of survey volunteers is chosen. The match is based on many characteristics available in databases, ensuring the survey sample mirrors the target population.

After the matched sample is created, it's given weights to adjust for any minor differences between it and the target population, ensuring the final results are representative. The weights are calculated in two stages using data from the American Community Survey and validated voter registration records, accounting for factors such as demographics and voter behaviors.

The CES then uses state-level samples and compares them with actual election results to validate the sample's accuracy and the weighting process. If survey estimates closely align with the actual votes, it suggests that the CES sample is accurate and representative (Schaffner, Ansolabehere, and Luks 2021).

This methodology is well fitted for the online opt-in nature of the survey, and helps combat the issues that arise with random or quota sampling which lead to results that are uncharacteristic of the target sample (Rivers 2007).

2.1 Limitations

While the data is largely reliable, there exist some smaller errors and limitations. The guide details an error affecting 925 respondents from North Carolina, who were shown candidate names for incorrect districts. This could influence the accuracy of analyses related to House races, necessitating verification methods and adjustments in studies which focus on these aspects.

As with any survey-based research, the CCES 2020 relies on self-reported data, which can introduce biases such as social desirability bias or recall bias. Additionally, pre-election surveys may not capture last-minute shifts in voter sentiment, which is particularly relevant in a rapidly evolving and polarized political climate like that of 2020.

The survey was conducted using an online survey platform hosted by YouGov (Schaffner, Ansolabehere, and Luks 2021). Online surveys overcome geographical and physical barriers that traditional survey methods might face, allowing for more diverse participant engagement. However, certain demographic groups might be underrepresented in online panels due to varying levels of internet access and technological literacy, potentially introducing biases into the dataset. Additionally, the impersonal nature of online surveys could affect the quality of responses, since the lack of a direct interviewer may lead to less thoughtful answers from participants.

2.2 Variables of Interest

The selection of age, gender, and education as predictor variables for analyzing political preference in the 2020 CCES data is informed by specific findings from prior research that shows their impact on voting behavior.

2.2.1 Age

The impact of age on political preferences is significant, as demonstrated in the 2016 election context. An article by Pew Research (2018) shows that 58% of validated voters aged 18-29 voted for the Democratic Party (Hillary Clinton), compared to 28% voting Republican (Trump). These numbers slowly begin to favor Trump with older age groups, where 53% of validated voters aged 65+ voted for Trump and only 44% voted for Hillary Clinton (Pew Research Center 2018). Research by Iyengar and Krupenkin (2018) illustrates that younger voters, influenced by contemporary sociopolitical developments, exhibit a marked preference for progressive policies and candidates (Iyengar and Krupenkin 2018). This tendency is often contrasted with older voters who may prioritize different issues based on their lived experiences.

2.2.2 Gender

The gender gap in political preferences is a well-documented phenomenon, with women more likely to support democratic candidates and policies, in particular social welfare and healthcare (al. 2004; Dolan 2014). Pew Research found 11% more men voted for Trump compared to Clinton, while 16% more women chose the Democratic candidate Clinton (Pew Research Center 2018).

2.2.3 Education

Level of education is a strong predictor of political knowledge and ideological orientation. Individuals with higher education levels are generally more politically engaged, exhibit greater political efficacy, and tend to support liberal ideologies (Highton 2009). We saw this play out in the 2016 election as generally college-grads were more favoured toward the Democratic

Party, while remarkably more (64% vs. 28%) non-grads favoured Trump (Pew Research Center 2018).

2.3 Data Preparation and Cleaning

The 2020 CCES data was first obtained via the dataverse package directly from Harvard Dataverse, and then saved in parquet format using the arrow package for efficient storage and access (Richardson et al. 2024). The cleaning process involved filtering for registered voters who chose either Biden or Trump, in order to treat presidential votes as a binary outcome.

The respondent variables (age, gender, and education) were transformed for clarity. Age was categorized into groups (“65+”, “45-64”, “30-44”, “18-29”), gender was labeled (“Male”, “Female”), and education levels were delineated (“No HS”, “High school graduate”, “Some college”, “2-year”, “4-year”, “Post-grad”). These steps utilized the dplyr package for data manipulation.

The cleaned dataset was then saved in both CSV and parquet formats.

The distributions for each explanatory variable are illustrated in Figure 2 and Figure 3 below:

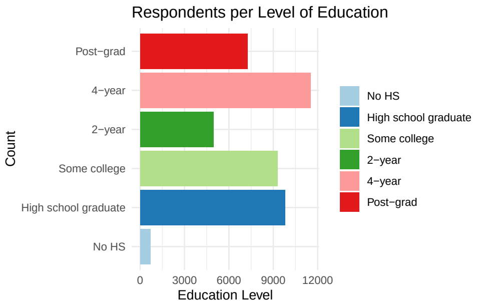

Figure 2: Number of respondents for various 2020 election survey sources

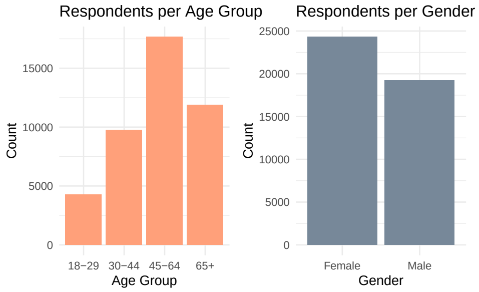

Figure 3: Number of respondents for age group and gender

Figure 2 tells us that 4-year degree graduates are the most well represented in the polls, followed by high school graduates, people with some college, and then 2-year degree graduates. People who have not graduated high school are hardly represented at all.

In Figure 3 we see 45-64 being the most dominant age group, with much less representation in the 18-29 range. Also, about 20% more females participated in the study than males.

Below are some figures representing who people voted for based on their age and gender, followed by education level and gender.

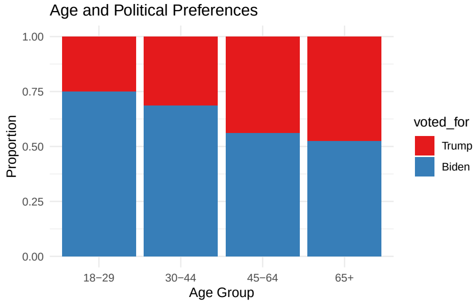

Figure 4: Proportion of votes at each age group

Figure 4 shows a negative correlation between votes for Joe Biden and older age groups.

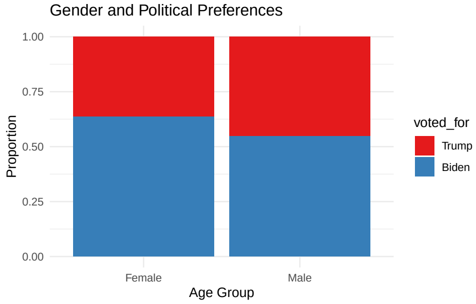

Figure 5: Proportion of votes by gender

We can see that in Figure 5, more women voted for Joe Biden than men.

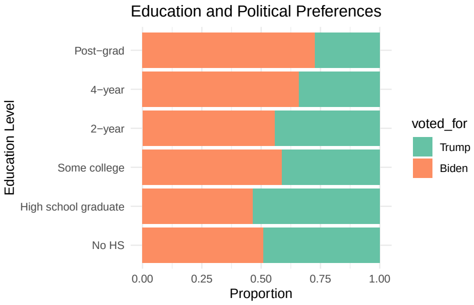

Figure 6: Proportion of votes at each level of education

Figure 6 is more interesting in the sense that the number of voters for Biden generally trends positively with higher levels of education, but the "No HS" and "Some College" groups had more votes for Biden respectively than the next level of education. It may be worth to note that "Some college" does not specify how much time they spent in college.

3 Model

Logistic regression is a model used when the outcome or dependent variable is binary, which fits this scenario perfectly as we are modelling votes for Joe Biden vs. Donald Trump.

The regression model will calculate the log odds of the probability that a respondent votes for Joe Biden, and then map it to a probability between 0 and 1 through the logistic function.

The standard logistic function $\sigma(t)$ for a real-valued input t is defined as:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

The graph of the logistic function is an S-shaped curve known as a sigmoid curve. It approaches 1 as t goes to positive infinity and approaches 0 as t goes to negative infinity.

In logistic regression, the input t is the linear combination of predictors including the intercept, which can be represented as $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. The logistic function then translates this into a probability that the dependent variable is 1 (voted for Joe Biden).

In this situation, I will be using the predictors age, education, and gender and then applying the logistic function to get the probability $P(Y_i = 1)$ that a respondent i supports Joe Biden.

This model is particularly strong at handling categorical dependent variables, which each of my explanatory variables fall under (Jr., Lemeshow, and Sturdivant 2013).

3.1 Model Specification

The logistic regression model is defined as:

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{\text{age},i} + \beta_2 X_{\text{education},i} + \beta_3 X_{\text{gender},i}$$

3.2 Model set-up

- Y_i is the binary indicator of support for the Democratic candidate Joe Biden (1) versus the Republican candidate Donald Trump (0) for respondent i.
- $X_{\text{age},i}$, $X_{\text{education},i}$, and $X_{\text{gender},i}$ are the age, education level, and gender of respondent i, respectively.
- β_0 represents the model intercept, while β_1, β_2, and β_3 are coefficients quantifying the effects of age, education, and gender on the likelihood of Democratic support.

I fit my logistic regression model to the data using 'stan_glm()' function from the 'rstanarm' package in R Goodrich et al. (2022). This function will automatically determine each of the β coefficients in the model, using a smaller slice sample of 3000 from the 2020 CCES data we processed. This function also uses Bayesian logistic regression with the default priors from 'rstanarm'.

3.2.1 Model Justification

In order to better interpret the results of the model, I can create a coefficient plot to visually see the effect sizes of the predictor variables on the likelihood of an individual supporting Biden.

Figure 7 maps each predictor variable on the y-axis to an effect size and confidence interval on the x-axis. The effect size is the change in log-odds of supporting Biden for a one-unit increase in the predictor variable, which is essentially how much impact each demographic characterstic has an effect on voting for Biden.

The confidence intervals tell us the range within we can be confident that the true effect lies. Smaller confidence intervals means there is a higher level of precision in the estimate of the effect size. I am not as interested in the intervals that cross zero because that means that there is data to support each side (Biden or Trump) and so they are less statistically significant.

```
[1] "term"      "estimate"  "std.error"
```

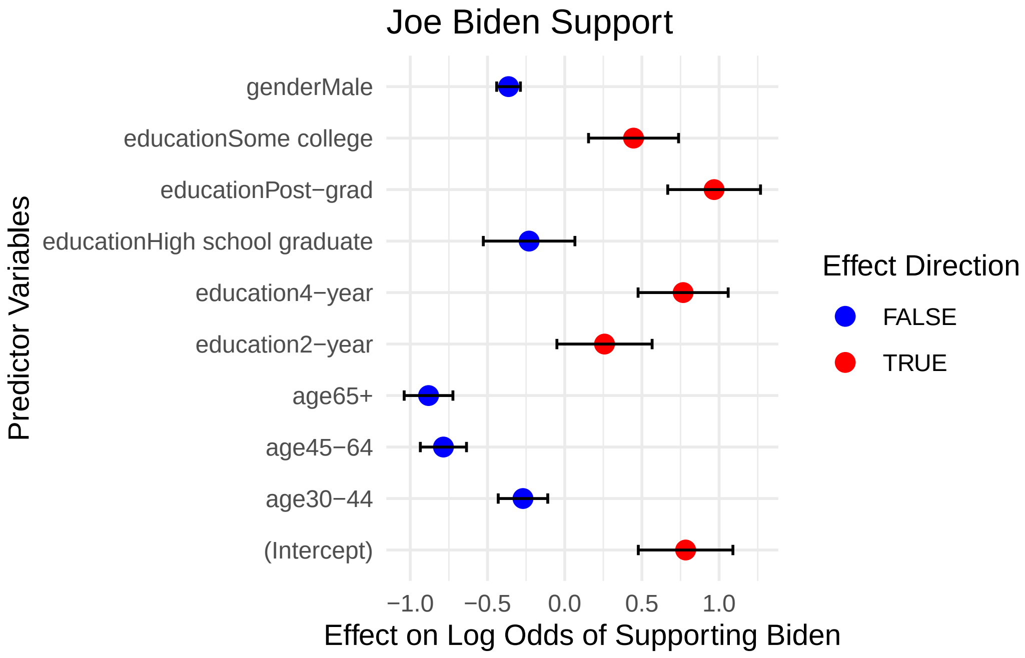

Figure 7: Coefficient plot of demographics

The conclusions I can draw from Figure 7 align with what I expected given the dataset. It's clear that higher education levels, specifically 4-year degrees and post-grad degrees, are positively associated with support for Biden. Higher age groups, specifically 45-64 and 65+, are strongly tied to voting for Trump. Men are less likely to vote for Biden than Women.

4 Results

Figures Figure 8, Figure 9, and Figure 10 are recreations of the earlier graphs we saw showcasing the proportion of voters for Joe Biden by groups within each demographic, where the bar is the data from CCES 2020 while the point is the prediction generated by the model.

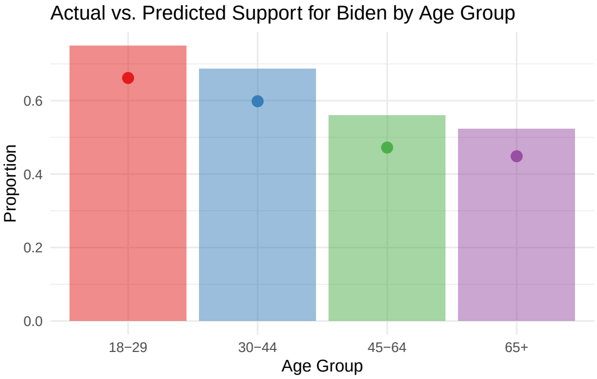

Figure 8: Model prediction for votes by age vs. CCES 2020 data

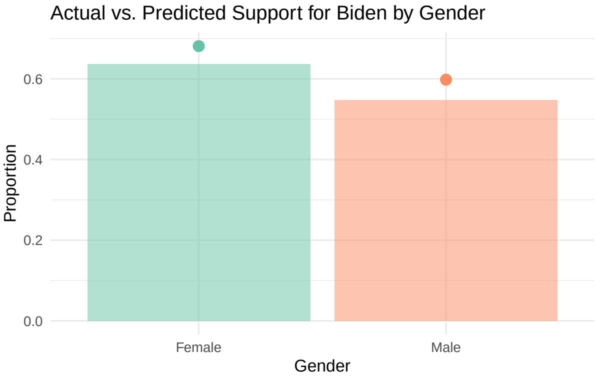

Figure 9: Model prediction for votes by gender vs. CCES2020 data

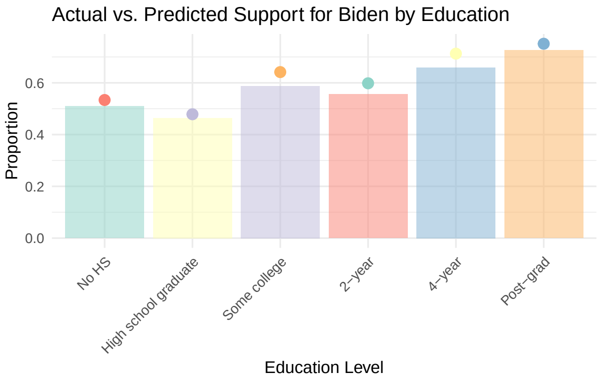

Figure 10: Model prediction for votes by education level vs. CCES2020 data

By observation, it is evident that the model's prediction was fairly close to what the CCES data showed. The trend remains the same for each result. However, gender and education were predicted to have a higher sentiment of voting for Joe Biden while age had a lower sentiment.

It is worth nothing though that in Figure 8, all age categories showed polling data of above 50% to vote for Biden. Now with the lower average vote, the older age categories of "45-64" and "65+" are leaning towards Trump. On Figure 10 for the lower education groups, the "No-HS" group which was at just about 50% is now above it, and the "High School Graduate" is also closer to voting for Biden. While this may seem marginal, close elections where voter sentiment is close to 50% would be largely impacted by changes like this. Although ultimately this is not an accurate predictor for who would win the election, only which groups would vote for who. # Discussion

This analysis demonstrates the demographics play in political preferences, with results that align with the existing literature. Our data suggests that age, education, and gender significantly influence electoral outcomes. In particular, the model's predictions closely mirrored CCES 2020 data trends, reinforcing the importance of these variables in understanding voter behavior. High education and younger ages showed high correlation with voting for joe biden.

It is important to note that based on the sample size of our data (as seen in Figure 2 and Figure 3), the majority age groups which participated in the survey were "45-64" and "65+", while the variables of education and gender were much more closer together. If this were representative of the actual population then it would suggest that the age groups are skewed more towards those that support Trump, so despite most of the proportions being towards Biden, it can be a lot more nuanced of there are more groups in demographics that support Trump.

5 Weaknesses and Next Steps

Although valuable, this study provides an intriguing analysis of how demographic influences affect political preferences; however, there are additional limitations to consider. Firstly, the study's analysis model, which used logistic regression, may have simplified the understanding of complex dynamics and could have overlooked some interactions between demographic variables that would provide a better understanding of voter behavior. Future research could utilize more advanced tools, such as multilevel or mixed-effects models, to more accurately capture the interplay of these demographic variables.

Secondly, while detailed, the study's dataset may not fully account for the subtleties in voter behavior across all demographics. For example, the impact of regional urban-rural political contexts on voting patterns could have nuances beyond the current analysis's reach. Incorporating geographic and socioeconomic variables could further improve and strengthen the model, offering deeper insight into the electoral impact of demographic factors.

Lastly, the study's focus on the 2020 election, though timely, limits its ability to capture longitudinal trends in voter behavior. A comparative analysis developed over multiple election cycles might reveal changes in demographic influences on political preferences over time, providing a richer context for interpreting the 2020 election results. In summary, this paper not only sheds light on the demographic determinants of voter preferences in a pivotal election year but also encourages reflection on the broader societal shifts these patterns indicate. As America faces deep social and political challenges, understanding the demographic underpinnings of political allegiance becomes increasingly important. Further research, building on the insights provided in this paper and addressing the weaknesses outlined, has the potential to offer even better explanations of the complex tapestry of American electoral politics.

References

al., Janet M. Box-Steffensmeier et. 2004. "The Emergence and Evolution of the Gender Gap in Political Ambition." *Political Research Quarterly* 57.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2). https://doi.org/10.1257/jep.31.2.211.

American National Election Studies. 2021. "ANES 2020 Time Series Study Full Release." https://www.electionstudies.org.

Ansolabehere, Stephen, and Brian Schaffner. 2020. "The Cooperative Congressional Election Study." *Harvard Dataverse*. https://doi.org/10.7910/DVN/ZSBZ7K.

Arel-Bundock, Vincent. 2021. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Catalist. 2017. "About Catalist." http://web.archive.org/web/20171028000000*/https://catalist.us/about/.

Cohn, Nate. 2020. "How Public Opinion Has Moved on Black Lives Matter." *The New York Times*. https://www.nytimes.com/interactive/2020/06/10/upshot/black-lives-matter-attitudes.html.

Dolan, Kathleen. 2014. "When Does Gender Matter? Women Candidates and Gender Stereotypes in American Elections."

Foundation, Kaiser Family. 2020. "KFF Health Tracking Poll - October 2020: The Role of Health Care in the 2020 Election and Attitudes Toward COVID-19." *KFF*. https://www.kff.org/health-reform/poll-finding/kff-health-tracking-poll-october-2020/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Highton, Benjamin. 2009. "Revisiting the Relationship Between Educational Attainment and Political Sophistication." *Journal of Politics* 71.

"How Does Gallup Polling Work?" 2024. 2024. https://news.gallup.com/poll/101872/how-does-gallup-polling-work.aspx.

Igielnik, Ruth, Scott Keeter, and Hannah Hartig. 2021. "Behind Biden's 2020 Victory." 2021. https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory/.

Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39.

Jr., David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. New York: John Wiley & Sons.

Kay, Matthew. 2021. *Tidybayes: Tidy Data and Geoms for Bayesian Models*. https://CRAN.R-project.org/package=tidybayes.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. 2018. "The Science of Fake News." *Science* 359. https://doi.org/10.1126/science.aao2998.

Leeper, Thomas. 2021. *Dataverse: Client for Dataverse 4 Repositories*. https://CRAN.R-project.org/package=dataverse.

Pew Research Center. 2018. "An Examination of the 2016 Electorate, Based on Validated

Voters." https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Rivers, Douglas. 2007. "Sample Matching." https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

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

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.