A Detailed Demographic Analysis of 2020 U.S. Presidential Election Through Multilevel Regression and Post-Stratification*

The Impact of Gender, Educatoin, Geographic location on Voter Preferences

Jeongwoo Kim Jiwon Choi

March 16, 2024

In this study, we analyzed the 2020 U.S. Presidential Election outcomes using a sophisticated statistical technique called multilevel regression with post-stratification (MRP), applied to a large-scale national survey dataset and census data. Our findings reveal how demographic factors such as gender, education, and geographic location influence voting behavior, with particular emphasis on voter preferences for Donald Trump. The analysis highlights significant variations in voter support across gender, states and educational backgrounds and how they shape electoral outcomes. This research contributes to our understanding of the American electoral landscape, demonstrating the critical role of demographic diversity in determining the direction of political preferences and election results.

1 Introduction

The U.S. Presidential Election is a significant political issue given its impacts on social, economic and cultural dimensions across the world. The 2020 Election was unique in that it was held during COVID-19, characterized by unprecedented voter turnout and a divided political gap between the Republicans and the Democrats. Joe Biden was elected as the President of the United States after 4 years of the former President Donald Trump era. Our research uses statistical methodologies to analyze how demographic factors such as education, state, and gender influence voter behavior and result in the victory of the Democratic party. By utilizing multilevel regression with post-stratification (MRP) to large-scale national survey data and census information, we explain the roles of gender, education, and geographic location in shaping voter preferences, focusing on support for Donald Trump.

*Code and data are available at: https://github.com/Kjeongwoo99/2020_US_Election_Result_Analysis

Understanding the electoral outcomes is crucial, not only for politicians but for the broader public and policymakers aiming to grasp the evolving landscape of American politics. Traditional analyses often fall short of capturing the relationships between voter demographics and electoral preferences. There are growing gaps between the groups in each education, gender, and state level and this study responds to this need by employing MRP. This technique allows accurate inferences about the voting behavior of diverse population segments. The estimand of our research is the true effect of demographic factors— gender, education, and geographic location—on voter preferences for Donald Trump in the 2020 U.S. Presidential Election. This effect represents the real-world influence that the demographic variables affect the likelihood of an individual voting for Donald Trump with all other variables held constant. It shows the estimated shift in voter preference towards Donald Trump to one unit change in each demographic factor. We aim to estimate this true effect as accurately as possible using available data.

Our findings reveal significant variations in voter support across different states and educational backgrounds and gender. The analysis points to a divergent pattern in voting behavior among different educational groups, challenging stereotypical narratives about educational attainment and party allegiance. We find that there is an increase in support for the Republicans from the voters with high education levels in this election. We also find the gender gap between men and women where men are more supportive of Trump while women tend to favor Biden. Lastly, we discover the the strong preference for Biden from big cities in the West and the East Coast while Trump's supporters are concentrated in the mid-west and suburban cities.

The importance of our study extends beyond academic, offering valuable insights for political strategists, journalists, and citizens. By providing a clearer picture of the demographic factors that influenced the 2020 election, our research aids in the development of more informed political strategies and fosters a deeper understanding of American democracy.

The paper begins by introducing the broader context and motivation behind our study. It then discusses the specifics of the data sources and the variables that we considered important. This is followed by the presentation of the data and a discussion of our findings regarding education, gender, and state in voter preference.

2 Data

We have used two datasets for this study. One is the U.S. election survey data of Democracy Fund + UCLA Nationscape dataset from the Voter Study Group (Tausanovitch and Vavreck 2020), conducted on October 3, 2019. Second is the census data from IPUMS America Census Service (Ruggles et al. 2019), which is used as the post-stratification data for the survey data to adjust the weight. Data was collected and analyzed using R statistical programming software (R Core Team 2023), with additional packages like tidyverse (Wickham et al. 2019), rstanarm (Goodrich et al. 2022) knitr (Xie 2020), here (Müller 2020), and many others for support.

2.1 Survey Data

This survey data is an 18-month election study conducted by UCLA researchers with roughly 6250 online interviews each from July 2019 to February 2021 (Tausanovitch and Vavreck 2020). The sample is weighted to represent the U.S. adult population (Tausanovitch and Vavreck 2020). Nationscape groups weight on the following important factors: gender, the four major census regions, race, Hispanic ethnicity, household income, education, age, language spoken at home, nativity, 2016 presidential vote, and the urban-rural mix of the respondent's ZIP code (Tausanovitch and Vavreck 2020). According to the data, males make up 48.3% while females make up 51.3% (Tausanovitch and Vavreck 2020). 74.2% of the respondents are White, 6.8% are Asian/Pacific, and 12% are Black (Tausanovitch and Vavreck 2020). 20.4% are those between 18-29, 33.4% are 30-49, 32.4% are 50-69, and 3.3% are 70+ (Tausanovitch and Vavreck 2020). On average, 5.1% declined immediately among those who were selected for the survey. 16.7% of the respondents did not complete the survey. Another 5.9% were categorized as speeding or straight-line which means they completed the survey in less than 6 minutes or selected the same response for every question in the three policy question batteries (Tausanovitch and Vavreck 2020). Leaving these out leave 72.4% of the original sample for the analysis.

The Nationscape survey's strength lies in its methodological rigor - the effectiveness of collecting large samples from U.S. citizens and its weighting strategy designed to mirror the U.S. adult population by including weight factors such as age, gender, race, and more. As they filter out inaccurate or missing data, it makes sure that the data collected are accurate and ensures data integrity. While other datasets such as the General Social Survey (GSS) and the American National Election Studies (ANES) are available, the Nationscape dataset's frequency (surveys collected every week) gives it an advantage in analyzing electoral trends and shifts in real-time. Its' extensive sample size also justifies the choice of this dataset.

For our analysis, we decided to choose five demographics: age, gender, education, race, and state and focus on gender, education and state. Age is important because in general, voters tend to become more conservative as they get older. To account for the age difference, we divided the age group into four categories: 18-29, 30-49, 50-69 and 70+.

Gender is also an important category because in general, men tend to be more conservative and women tend to be more liberal. Recently, gender issues are growing social issues and this may affect the election, hence we wanted to explore how this affects our model.

Education is also an interesting factor. In the past, non-college white voters used to support Democrats while college-educated white voters supported Republicans (Harris 2018). However, there has been a switch in this trend as 61% of non-college white voters showed their support whereas just 45% of college-educated white voters did in the exit polls (Harris 2018).

Race also needs some attention because normally non-white groups are highly in favour of Democrats regardless of candidates and white swing by depending on candidates. According to the statistics collected in 2016, 93% of black, 71% of Latino, 68% of Asian support democrats

while only 41% of white support democrats (Prokop 2021). As white voters make up 74% of the voting population, it is really important for both parties to attain this demographic group.

Lastly, states are very important as some states historically favor conservatives while some states vote for democrats. In general, the west and the east coasts are democrat supporters whereas south are conservative supporters.

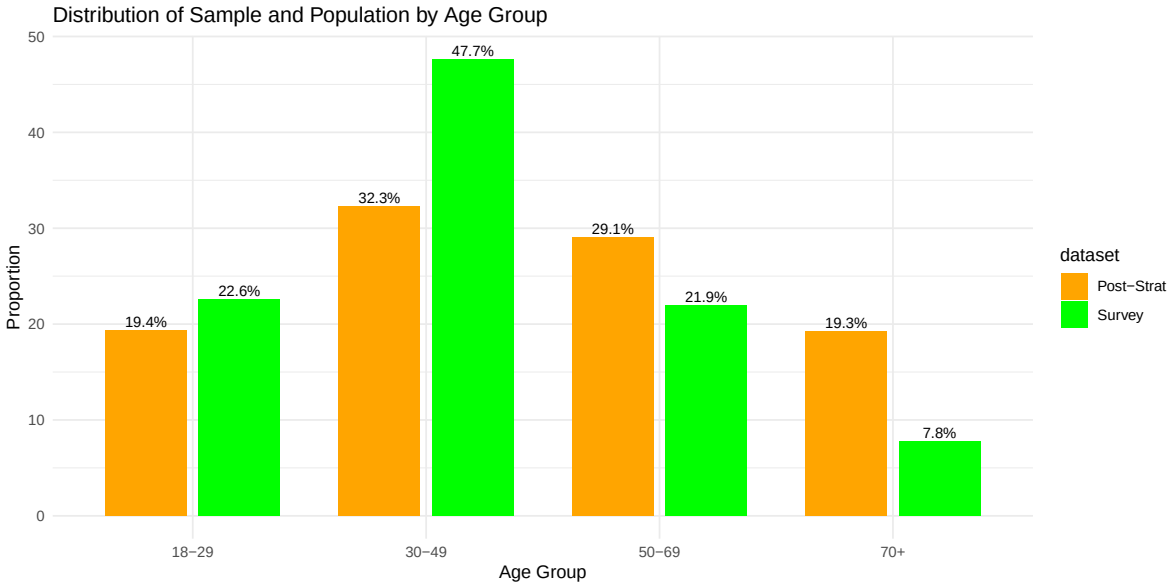

Figure 1: Distribution of Sample and Population by Age Group

2.2 Post-stratification Data

IPUMS ("Integrated Public Use Microdata Series") is a website that offers a database of samples of the American population from the American Community Surveys of 2000-present. These samples provide rich qualitative information on the long-term changes in the population. We selected the '2019 ACS' data (Ruggles et al. 2019) as the post-stratification dataset for our research to avoid any potential effects of COVID-19. The ACS is an ongoing survey that collects data monthly, which is then combined into 1-year, 3-year, and 5-year aggregates. It then uses stratified sampling where the U.S population is broken down into sub-groups and initial weights are assigned to each respondent.

One strength of the IPUMS survey is the fact that it provides a data with detailed demographic of the U.S. population with social, economic and housing characteristics, which is very useful in our analysis of the 2020 U.S presidential election forecast. The longitudinal data of this survey also allows researchers to analyze trends over time. The U.S. Census Bureau offers credibility of the data with high quality checks. The post-stratification process ensures correcting for sampling biases and non-response. On the other hand, since the survey relies on self-report,

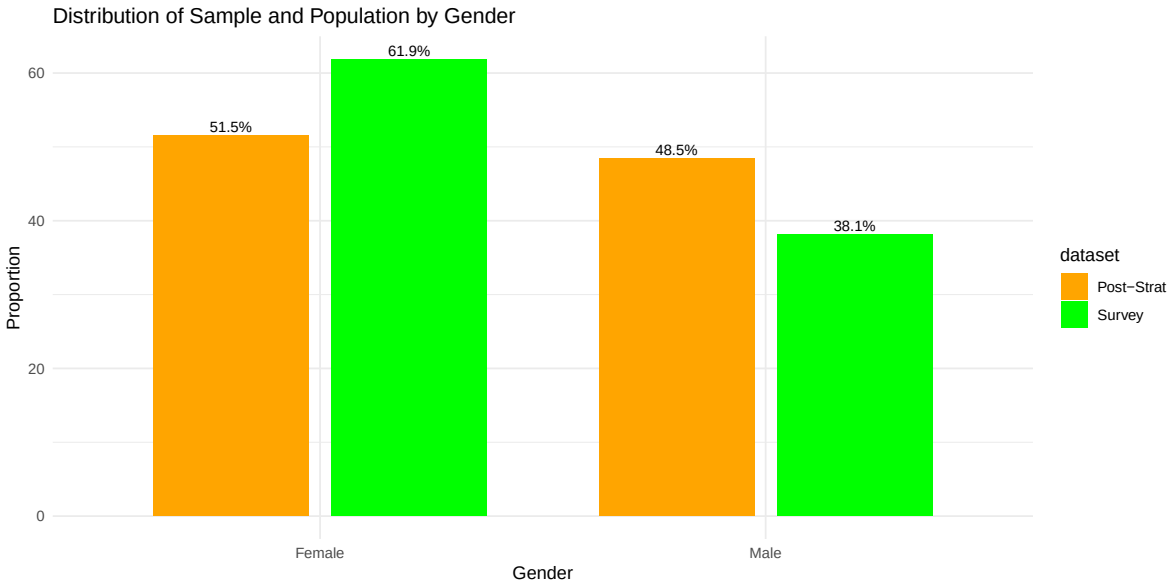

Figure 2: Distribution of Sample and Population by Gender

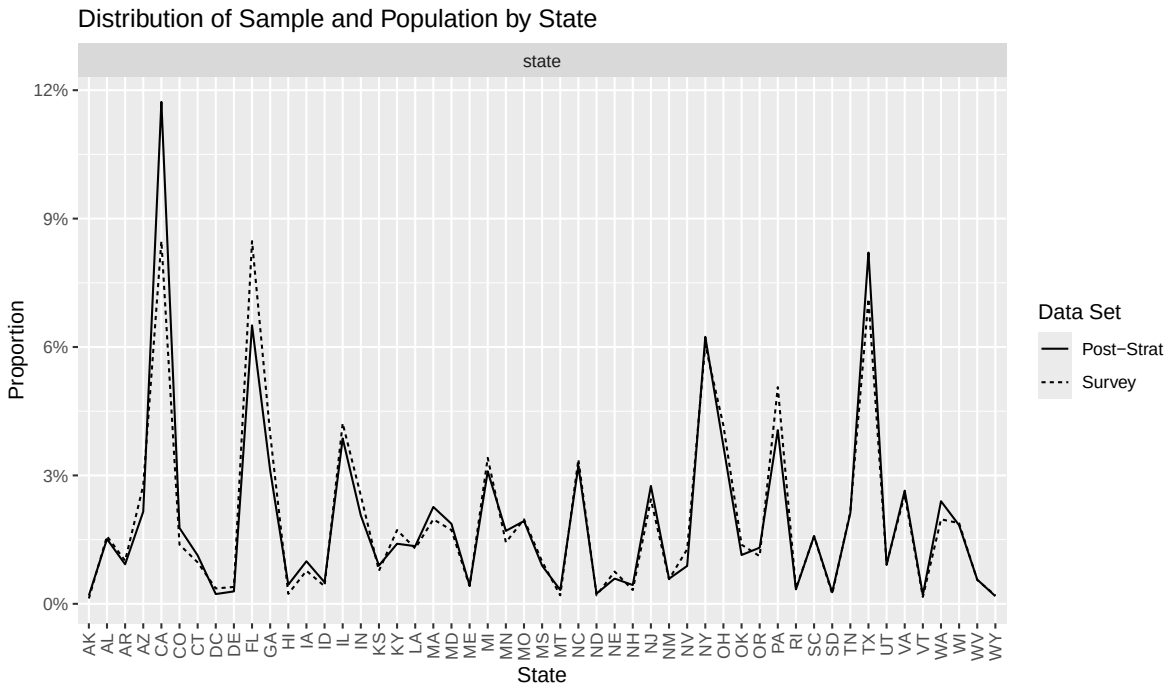

Figure 3: Distribution of Sample and Population by State

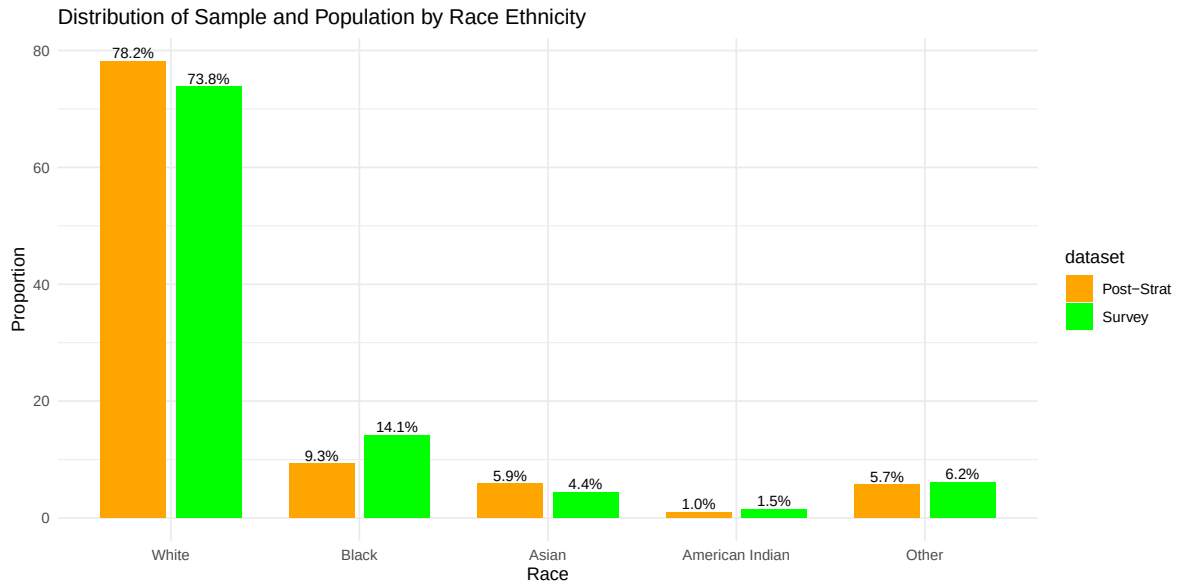

Figure 4: Distribution of Sample and Population by Race Ethnicity

there lies a risk of response bias inherently. While it is an ongoing survey, there is still a time lag between the data collection and data availability. However, the large sample size, consistency and reliability of the data collection, the integrated data over time with post-stratification can justify the decision to utilize IPUMS data over other sources.

In processing the raw post-stratification dataset, which initially contained approximately 3.2 million records, we refined it down to about 2.3 million records. This was achieved through a meticulous selection process, ensuring the data's integrity and relevance for our analysis. In our analysis, we've selected the variables 'sex', 'race', 'stateicp', 'age', and 'educd' from the dataset. To simplify our analysis, respondents who indicated 'other' or provided no data for their sex have been excluded. Consequently, 'sex' has been categorized strictly as 'Male' and 'Female'. We've refined 'race' into five categories: 'White', 'Black', 'Asian', 'American Indian', and 'Other', based on the composition of the U.S. population, with White, Black, and Asian categories accounting for approximately 93 percent of the total.

The 'stateicp' variable encompasses all U.S. states, using their standard abbreviations (e.g., 'CT' for Connecticut), and extends to 55 values to include 'Puerto Rico', 'State groupings (1980 Urban/rural sample)', 'Military/Military Reservations', 'District of Columbia', and an 'State not identified' category. Age has been grouped into four categories: '18-29', '30-49', '50-64', and '70+'. For educational attainment ('educd'), we've created four categories: 'High school or less', 'Some college', 'College degree', and 'Postgrad'.

We excluded any unknown responses to ensure clarity and accuracy in categorization and to enhance clarity and align with survey data, we've renamed 'sex', 'stateicp', 'age', and 'educd'

to 'gender', 'state', 'age_group', and 'education', respectively. This restructuring aims to streamline our analysis by ensuring each respondent is accurately categorized.

All figures showed in the data section illustrate comparisons between survey data and post-stratification data across different variables. First, the data comparison by state shows that the correspondence between the survey and post-stratification data for each state is quite accurate, further emphasizing the overall reliability of the survey methodology in capturing diverse demographic characteristics. (see Figure 3) In other four figures, orange bars represent post-stratification data, while green bars signify survey data. The percentages displayed on each figure are rounded to the nearest tenth, introducing a potential margin of error of $\pm 0.1\%$ in the total values. Generally, the survey data aligns closely with the post-stratification data, maintaining a discrepancy of about 10% across most categories. Notable exceptions are observed in the '30-49' age group and the 'Some college' education level, where the differences exceed this margin. (see Figure 1, Figure 2, Figure 4, and Figure 5)

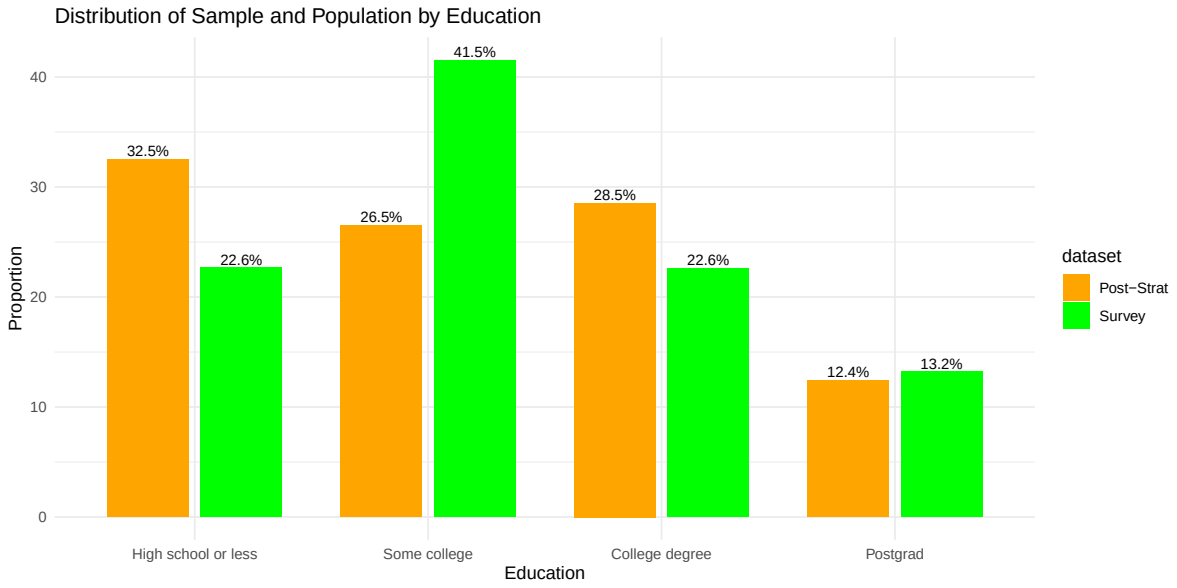

Figure 5: Distribution of Sample and Population by Education

Table 1 shows the proportion of voters who intend to vote for Donald Trump or not. More than half of the respondents, which is 55.30% choose not to vote for Donald Trump. Also, Table 2 shows the proportion of voters' supporting each party. The data used to create these table is from the Democracy Fund + UCLA Nationscape (Tausanovitch and Vavreck 2020). We see that Donald Trump and his party Republican are not expected to win the popular vote before we implement the model.

Table 1: Voters Intention to Support Trump

Response	Number of Respondents	Proportion (%)
Yes	1908	34.25
No	3080	55.30
Other	582	10.45

Table 2: Voters Intention of Their Primary Party

Party Preference	Number of Respondents	Proportion (%)
Democratic	2180	39.14
Republican	1533	27.52
Other	1857	33.34

3 Model

For our study, we employ a technique called multilevel regression with post-stratification (MRP). This approach involves creating a model based on a smaller data set, such as our survey data, and then extending the model's findings to a larger population.

The key steps in MRP involve initially selecting a dataset for model development. In this case, we utilized survey data from the Voter Study Group (Tausanovitch and Vavreck 2020). The next step is to construct a model with this smaller dataset; here, we employed a logistic regression based on the survey data, formulated as seen in equation 1. It is then followed by appling post-stratification to a broader dataset to estimate population characteristics. For our analysis, Census data from IPUMS (Ruggles et al. 2019) served as this larger dataset. Given that logistic regression is suited for binary outcomes, we've introduced a variable, 'consider_trump', which assigns a 1 if the respondent indicates a willingness to vote for Donald Trump, and a 0 for intentions to vote for other candidates, with 0 encompassing both "No" and "Other" responses.

The logistic regression model takes the form of:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{gender} + \beta_2 x_{agegroup} + \beta_3 x_{race} + \beta_4 x_{state} + \beta_5 x_{education} \quad (1)$$

In equation 1, each β represents a coefficient determined through regression analysis. The variables chosen for this project are gender, age, race, education, and state. These were selected because gender, age, and race are proven to be reliable indicators of voting preferences. This decision is based on patterns such as certain states consistently favoring the Republican party, while others alternate between Democratic and Republican. Education was chosen over income because it provides a clearer picture of an individual's background than income does.

Once the logistic regression model is developed, we'll use the `predict()` function in R (R Core Team 2023) to apply our model to Census data (Ruggles et al. 2019), breaking down the dataset into categories based on gender, race, age group, education level, and state. This will give us the likelihood of individuals within each category voting for Donald Trump. These predictions allow us to analyze potential outcomes like the popular vote winner or the electoral college vote distribution.

We use the `stan_glm()` function in R (R Core Team 2023) for our regression analysis, specifically because we are dealing with a binary outcome: whether a voter supports Donald Trump or not. The nature of our data suggests an S-shaped distribution rather than a linear one, making logistic regression a better fit than linear regression. This approach is advantageous, especially when paired with post-stratification, as it allows us to better represent under-represented groups in our analysis. For instance, despite having only 7 responses from Alaska in our survey data (Tausanovitch and Vavreck 2020), through multilevel regression and post-stratification, we can adjust this to effectively represent over 4500 individuals.

However, there are limitations to our model. The binary outcome does not allow for consideration of third-party candidates or non-voters, although this limitation is mitigated by our focus on the main candidates. More critically, our model's accuracy is heavily dependent on the quality of our survey data. Any inaccuracies or the need for adjustments in the survey can significantly impact our findings.

4 Results

Table 3 shows the estimates for the coefficients that will fit into our logistic regression equation. These coefficients will fit into Equation 1, and were calculated using data from the Voter Study Group (Tausanovitch and Vavreck 2020). The table is made using `kable` from `knitr` (Xie 2020) and is formatted using `kableExtra` (Zhu 2020).

Figure 6 presents the coefficients derived from logistic regression on the survey data (Tausanovitch and Vavreck 2020). It also includes error bars, indicating the confidence interval for each coefficient estimate. In interpreting these coefficients, it is essential to understand that positive values suggest a greater likelihood of voting for Donald Trump, whereas negative values indicate a tendency to vote for other candidates, such as Joe Biden.

By utilizing the results from our logistic regression model, we can formulate an equation that adheres to the structure outlined in equation 1, incorporating specific β coefficients for each variable. Given the number of variables, detailing the equation fully is challenging. Essentially, the equation integrates the β value of a variable if an individual's characteristic matches that variable. Table 4 offers examples of how the probability varies based on different variables.

Figure 7 shows us the estimated of proportion of support for Trump or not by state using MRP with the inclusion of error terms. Each dot represents the point estimate of the proportion of support for Trump (red) or support for any other candidates including Joe Biden (blue) in

Table 3: Coefficients from the Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	-1.1269207	0.8231922	-2.5930365	0.1878863
genderMale	0.6454282	0.0605598	0.5460791	0.7446367
educationHigh school or less	-0.0215196	0.0938906	-0.1785914	0.1291637
educationPostgrad	-0.0392885	0.1022382	-0.2097959	0.1332047
educationSome college	0.1320413	0.0781562	-0.0004280	0.2590242
age_group30-49	0.5119398	0.0850042	0.3699014	0.6477017
age_group50-69	0.6769174	0.0941263	0.5119061	0.8319957
age_group70+	0.8993021	0.1253618	0.6878129	1.1100246
raceAsian	-0.7916706	0.3008511	-1.2762432	-0.2728760
raceBlack	-1.7052759	0.2752683	-2.1485129	-1.2387060
raceOther	-0.9723473	0.2865172	-1.4448648	-0.4727599
raceWhite	0.1901461	0.2483476	-0.2079037	0.6111706
stateAL	0.1759303	0.8262695	-1.1488110	1.6604503
stateAR	-0.0657478	0.8344496	-1.4419087	1.4019364
stateAZ	-0.4574335	0.8034301	-1.7451418	0.9887553
stateCA	-0.5812949	0.7941797	-1.8467363	0.8398753
stateCO	-0.4103942	0.8165795	-1.7516658	1.0454097
stateCT	-0.8980631	0.8694824	-2.2666010	0.6056393
stateDC	0.1333770	0.9378865	-1.4644468	1.7361204
stateDE	-0.0329808	0.9211048	-1.5537404	1.5722435
stateFL	0.0218098	0.7992711	-1.2561402	1.4587686
stateGA	0.2273150	0.8093676	-1.0584034	1.6656086
stateHI	-0.4022495	1.0532898	-2.2097550	1.3718501
stateIA	-0.2792869	0.8523817	-1.6812761	1.2203228
stateID	-0.5715375	0.9117792	-2.1224329	1.0171239
stateIL	-0.4539897	0.7908061	-1.7369158	0.9760011
stateIN	-0.1124579	0.8034814	-1.3759652	1.3237964
stateKS	-0.9154592	0.8712311	-2.3169099	0.6127700
stateKY	0.1891792	0.8196904	-1.1163833	1.6735899
stateLA	0.5085951	0.8454838	-0.8286095	2.0124852
stateMA	-1.0144199	0.8087735	-2.3278399	0.4318518
stateMD	0.2540933	0.8153424	-1.0472043	1.7133075
stateME	0.3319872	0.8983963	-1.1254386	1.8934577
stateMI	-0.6201596	0.8147269	-1.9179094	0.8278040
stateMN	-0.1735514	0.8315900	-1.5049968	1.2975357
stateMO	0.2568405	0.8204754	-1.0284865	1.7210137
stateMS	1.0045240	0.8628145	-0.3393123	2.4995599
stateMT	0.9169121	1.0397878	-0.7536536	2.7395949
stateNC	-0.3230403	0.8028628	-1.6358424	1.1271320
stateND	-1.3821718	1.1464287	-3.4190661	0.5101015
stateNE	-0.7427467	0.8529747	-2.1674568	0.7907564
stateNH	-1.1627679	1.0026527	-2.8324897	0.5333387
stateNJ	-0.2755491	0.8195828	-1.5727529	1.1896796
stateNM	0.1284793	0.8789283	-1.2857051	1.6757967
stateNV	-0.4946525	0.8344399	-1.8263553	0.9859628
stateNY	-0.4028180	0.7889157	-1.6920330	1.0308732
stateOH	-0.2812188	0.8031339	-1.5499012	1.1728527
stateOK	-0.5510119	0.8230695	-1.8876618	0.9132757
stateOR	-0.6065868	0.8410369	-1.9839289	0.8792515
statePA	-0.3623052	0.8042335	-1.6496183	1.0852955
stateRI	-0.8718240	0.9623059	-2.4641792	0.7952202
stateSC	0.0476201	0.8259632	-1.2557306	1.5040781
stateSD	-0.1283511	0.9993491	-1.7289774	1.5773496
stateTN	0.0825080	0.8065814	-1.2181768	1.5502102
stateTX	-0.1073905	0.7918389	-1.3699153	1.3353886
stateUT	0.2037785	0.8368944	-1.1454047	1.6788839
stateVA	-0.0515024	0.7999330	-1.3606166	1.3835400
stateVT	-0.4688927	1.1001454	-2.2945844	1.3947062
stateWA	-0.7231677	0.8189467	-2.0323499	0.7530020
stateWI	-0.2468407	0.8044324	-1.5596314	1.1970349
stateWV	-0.1556233	0.8788915	-1.5628548	1.3558063
stateWY	0.8229283	1.0259232	-0.8515683	2.5749520

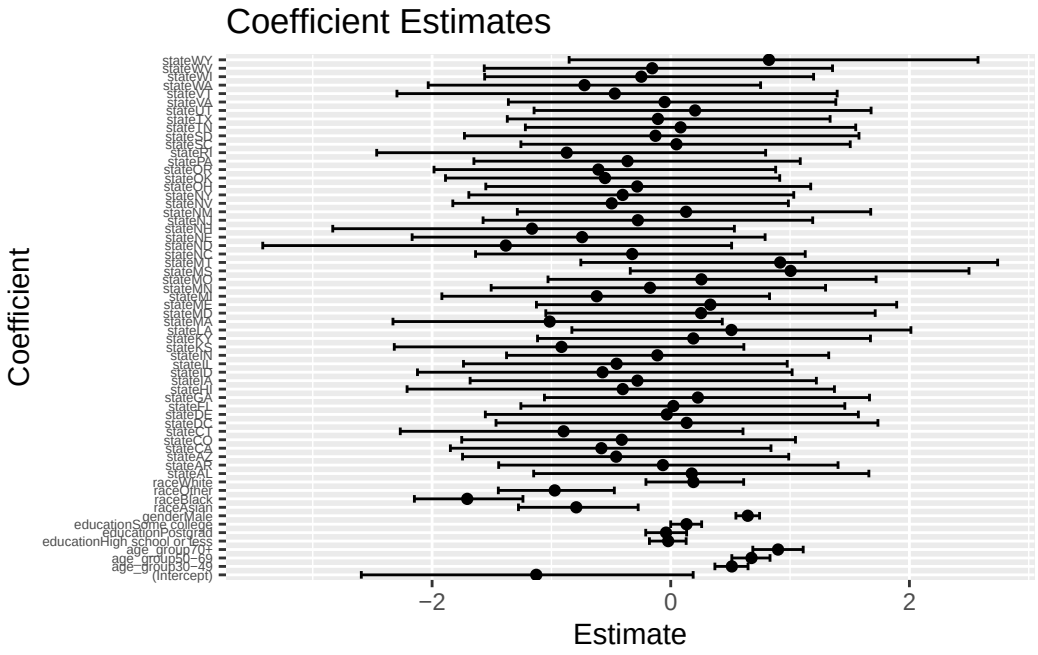

Figure 6: Coefficient Estimates

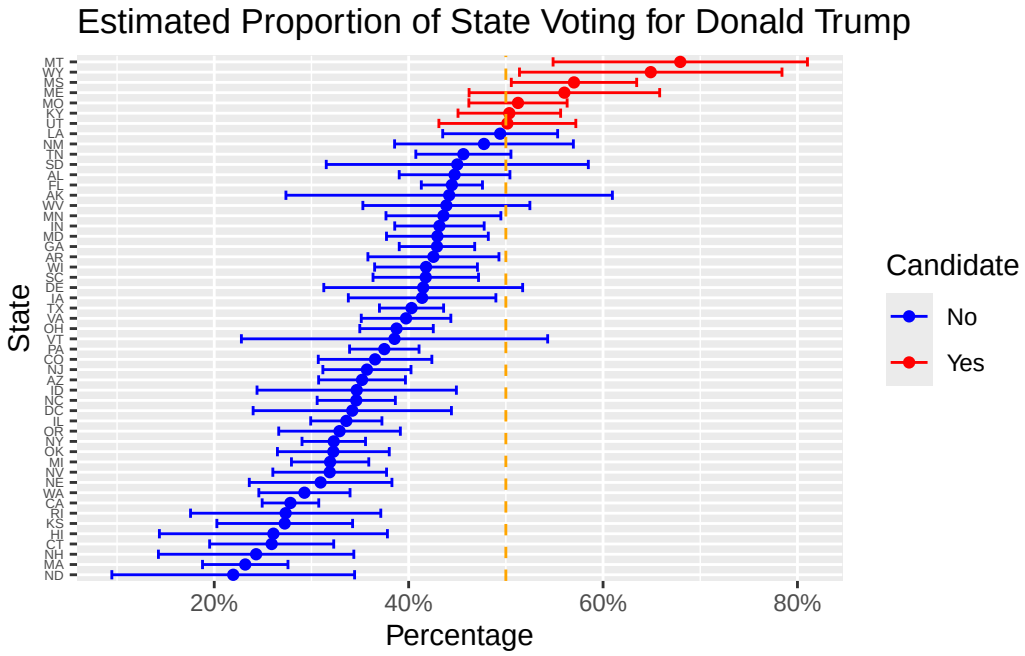

Figure 7: Distribution of Sample and Population by State

Table 4: Example of Prediction model

gender	race	state	age_group	education	predicted_consider_trump_probability
Male	Black	MD	18-29	Some college	0.1463268
Female	White	TX	50-69	Some college	0.4450551
Male	Black	DC	70+	High school or less	0.2471943
Female	White	NY	30-49	Some college	0.3363712
Female	White	AL	30-49	College degree	0.4419403
Male	White	MI	50-69	High school or less	0.4404717
Female	White	MN	30-49	College degree	0.3594809
Male	White	CA	70+	High school or less	0.5060479
Female	Black	TN	70+	High school or less	0.1369608
Female	White	FL	18-29	College degree	0.2902058

each state. Horizontal lines extending from the dots represent error bars for these estimates. The length of each line indicates the uncertainty associated with each estimate. For instance, we can see that this uncertainty lies between 55% to slightly higher than 80% for Trump in MT (Massachusetts). The dashed orange line in the middle at the 50% mark represents the threshold for majority support. On the y-axis, each state is listed with its abbreviations and is ordered based on the proportion of support for Trump from the highest to the lowest.

From Figure 7, it seems that the majority of the states do not support Trump. Only 7 states out of 51 have a point estimate greater than 50% for Trump. The horizontal lines of confidence intervals of some states overlapping the green mark give some hope for the Republicans. Excluding these contesting states, however, our model suggests that only 3 states are definitely in favor of Trump whereas 35 states are not considering supporting for Trump.

Figure 8 presents the estimated proportion of voters for Trump by education level, divided into four categories: 'High school or less', 'Some college', 'College degree', and 'Postgrad'. Each black dot represents the point estimate of the proportion of voters within the corresponding education category who are predicted to vote for Trump. The horizontal lines extending to the left and right of each dot represent the error terms around the estimate, which reflect the uncertainty.

It shows that regardless of education level, the level of support for Trump lies below 40% The Republican party does not have the majority, including the error bars across all education levels. Voters with a "High school or less" education level appear to have the lowest estimated support for Trump, which does not align with various exit polls and analyses from the 2020 election suggesting that Trump had substantial support among voters without a college degree. Conversely, Voters with 'Some college' and 'Postgrad' education are the two groups that are more in support of Trump, which is exactly the opposite of what we expected.

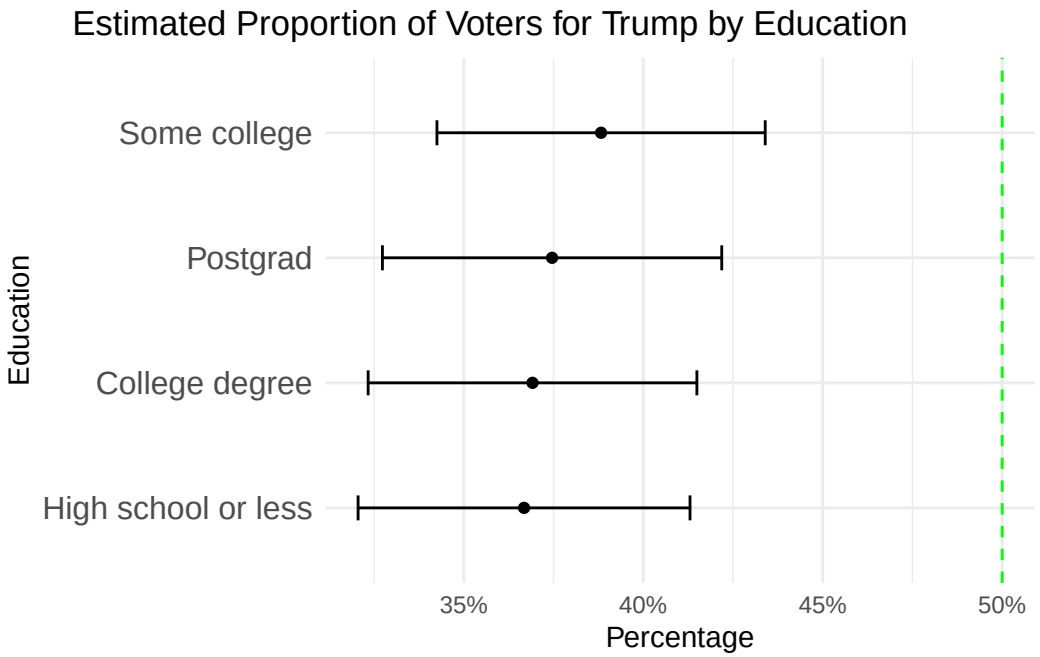

Figure 8: How different education levels of voters affect voting for Trump

5 Discussion

5.1 Analyzing Voter Preferences Across Education

The analysis reveals critical insights into how education influence voting behavior. Contrary to conventional wisdom and past electoral analyses, our findings indicate a divergent pattern in voting behavior among different educational groups. The shift in party allegiance among non-college voters from Republicans to Democrats and the dominance of popularity of the Republicans with 'Some college' and 'Postgrad' education levels challenge the stereotypical narratives that educational attainment leads to an increase in support for the Democrats. This group, especially the highly educated 'Postgrad' group being the second highest supporters suggests a reevaluation of Democrats' strategies to appeal to educated voters concerned with economic policies and national security. However, according to our analysis, the Republicans failed to attract voters as the highest proportion of support among the four educational groups remains under 40%. This almost certainly predicts that Joe Biden will become the president of the United States. This coincides with the election results where Biden won 55% of the votes from the college graduates while Trump gained 43% (Weigel 2020). Our analysis of 37% of voters with college degrees and 37.5% of postgrad suggests this is in line with the results. However, this model has failed to estimate the winning of Trump from the less educated group since we estimated 39% from voters with 'some college degree' and 36.5% from voters with

'High school or less' whereas the post-election results show 50% for Trump from 'Some college or less' (Weigel 2020). This is odd because the higher the education completed, the more likely an individual to vote for democratic party in general. One possible explanation for this is that there are some issues with the data collection where voting for Trump is underestimated as the the proportion of voters for him remain under 40% across all educational groups. Another possibility is that the gap between educational groups in party preference has actually been reduced from the past. In fact, support from white men without a college degree reduced somewhat although he still won a majority of this demographic (Ruth Igielnik 2021).

5.2 Voter Preferences Across Gender

Table 3 shows how males were more likely than females to have a favorable view of Trump, which correctly captured the tendency of voters by sex since 53% of men favored Trump whereas 42% of women favored Trump (Weigel 2020). Gender does not act as a uniform lens through which policy positions are viewed, however, it is still an important issue especially regarding Trump in this election due to his words and actions that may viewed as sexist. Historically, women have been leaning more liberal than men. This gender gap has varied across different elections but has remained a consistent feature of the American political landscape since the 1980s (Lizotte 2017). While men's preferences for a smaller government remained stable, women's preference for a bigger government increased (Hartig 2019). This shift contributed to a widening gap in presidential job approval with men remaining more supportive attitude of Trump. Solanas suggests that Trump's administration increased investment in health, education, and security for girls, teenagers, and migrant women, and support measures for pregnant women, as well as the promotion of women in STEM fields and efforts against gender violence (Solanas 2018). However, this is no secret that Trump is very conservative in treating gender equality. His openly sexist comments and behaviours presaged potential limitations on women's rights and freedoms, which may have contributed to the gender gap further (Solanas 2018).

5.3 Geographical Divides and Electoral Preferences

From Figure 7, we can observe that Trump leads in the mid-west of America, Montana (MT), Wyoming (WY), Mississippi (MS), Maine (ME), Missouri (MO), Kentucky (KY), and Utah (UT). Even the competing states like Louisiana (LA), New Mexico (NM), and Tennessee (TN) are also mid-west part of the country, which are rural areas with low population density. This is correctly estimated since in the real world where Trump won in these states: MT, WY, MS, MA, MI, KY, and UT. He also won in LA and TN, which we considered were competing states. We can go more in-depth into this study with the benefit of hindsight. We estimated that the biggest victory for Trump would be Montana (MT) with just below 70%. In reality, he won by 59.6%, which is lower than we thought but still sits within our error values. The next state with the highest proportion of support for Trump was Wyoming (WY) from our

model with 65%, which is close to the election poll result of 69%. We were very successful with Mississippi (MS) with an estimation of 57% and the poll result of 57.5%. Maine (ME) was another very close estimate with 52% and the poll result of 53.1%. 56.8% voted for Trump in Missouri (MO), which was slightly higher than our estimation of 51%. We estimated that only 26% of voters in Hawaii (HI) would vote for Trump with error margins of ±10%, which matched 34.3% for Trump in the election. 41.9% for Trump in Colorado (CO) lie within our estimate of 37% with ±5% margins.

There were some cases where our estimation was incorrect in terms of who won the state but the results lie within our error margins. For instance, we estimated Utah (UT) in favor of the Democrats with a slight margin of 51% for Biden but Biden only received 37.6% in the election. This still lies within our error values which we estimated to be from 43% to 58%. However, there were some errors in our estimation where our model failed to predict the Trump winning in the number of states in general. Our model predicted Louisiana (LA) for Biden with 51% but he only received 39.9% of the votes. This value is lower than our error margins. 62% of the residents in Alabama, a Republican stronghold in the last eleven presidential elections had voted for Trump while our model estimated 45% of votes for Trump. The Republicans won in Florida (FL) with a slight margin of 51.2% but we predicted an estimation of 45% with ±2% margins. What we estimated to be the strongest Democrat state was North Dakota (ND) with an estimation of only 23% with ±12% margins but in reality, 65.1% of the residents voted for Trump. In general, big cities and large suburban areas along the west coast and the east coast favored Biden whereas Trump supporters are mostly in the rural areas in the mid-south and mid-north states from the election results.

Our model was able to capture some challenges of long-held assumptions about electoral tendency by education. Biden improved his performance among suburban voters and White non-college voters, demonstrating a broadening appeal across different demographic groups, which is what our model suggested (Ruth Igielnik 2021). Trump maintained strong support among rural voters and White evangelicals, which emphasizes the importance of educational divides, with significant differences in candidate support based on education levels (Ruth Igielnik 2021). In addition, while Trump's stronghold among White men without a college degree loosened somewhat, he still won a significant majority of this demographic, which is not captured in our model (Ruth Igielnik 2021).

We were able to estimate the winning states for each candidate with some success. From the data and analysis outlined above, we believe that our model suggests that Biden would have won both the popular vote and the electoral votes and become the president of the United States. This was partially right in that he did win both but with a smaller margin than we had expected. In the end, he won 306 electoral votes and won 51.3% of the votes to become the 46th US president (CNN 2020). However, our model underestimates the proportion of Trump voters around 8 - 10%, compared to the actual 48% of votes he got in the election. In the cleaning process, we had to reduce both the number of survey data and post-stratification data, that could have led to this difference.

5.4 Weaknesses and Implications for Future Research

One major limitation of this research is the reliance on the quality of the survey data collected in October 2019. Any inaccuracies of COVID-19 affected the collection of the survey data, so we had to use the 2019 data. The election was held in November 2020, so there exists a time gap of one year between the data collection and the election, which may have affected the voting as one year gives enough time for voters to change their minds. In addition, any inaccuracies or biases within the survey could significantly impact the model's forecasts, especially, through the data cleaning process, which involved dropping observations with missing data. We reduced the survey dataset from 6,146 to 5,570 observations and the post-stratification dataset from 3,239,553 to 2,334,234 observations. This resulted in a loss of approximately 10% of the survey data and 30% of the post-stratification data. Such reductions could potentially impact the accuracy of our forecast results.

Our analysis also involves some weaknesses of using predictive modeling to fully grasp the complexity of how elections work. Although our model picks up on some trends and changes in voter demographics, it also shows that predicting election outcomes accurately, particularly in states where the results could go either way, is difficult. The differences between what our model predicted and the actual election results underline the importance of continuously improving our models. This means adding more detailed data and possibly trying different methods to better understand how people vote. This experience has made it clear that to make our election predictions more accurate and trustworthy in the future, we need to take into account various factors, such as education levels, differences between urban and rural areas, and the evolving political scene. Therefore, future research should explore models that can accommodate multiple electoral outcomes, including third-party voting and more variables if possible. Using more accurate or applicable real-time data sources could complement traditional survey methods and offer fresh insights into voter preferences. Societal norms and political landscapes change over time. Therefore, future research should focus on emerging social issues that may influence voter behavior. In the 2024 Presidential election, climate change, technology policy, and social justice could play more significant roles in shaping electoral outcomes.

References

CNN. 2020. "PRESIDENTIAL RESULTS." https://www.cnn.com/election/2020/results/president.

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

Harris, Adam. 2018. "AMERICA IS DIVIDED BY EDUCATION." *The Atlantic*. https://www.theatlantic.com/education/archive/2018/11/education-gap-explains-american-politics/575113/.

Hartig, Hannah. 2019. "Gender Gap Widens in Views of Government's Role – and of Trump." https://www.pewresearch.org/short-reads/2019/04/11/gender-gap-widens-in-views-of-governments-role-and-of-trump/.

Lizotte, Mary-Kate. 2017. "GENDER DIFFERENCES IN AMERICAN POLITICAL BEHAVIOR." *Scholars Strategy Network*. https://scholars.org/contribution/gender-differences-american-political-behavior.

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

Prokop, Andrew. 2021. "A New Report Complicates Simplistic Narratives about Race and the 2020 Election." *Vox*, May. https://www.vox.com/2021/5/10/22425178/catalist-report-2020-election-biden-trump-demographics.

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

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. *IPUMS USA: VERSION 9.0*. Minneapolis: University of Minnesota. https://www.ipums.org/projects/ipums-usa/d010.v9.0.

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

Solanas, María. 2018. "Gender Equality in Trump's America." https://www.realinstitutoelcano.org/en/analyses/gender-equality-in-trumps-america/.

Tausanovitch, Chris, and Lynn Vavreck. 2020. *Democracy Fund + UCLA Nationscape*. October 10-17, 2019 (version 20200814). https://www.voterstudygroup.org/publication/nationscape-data-set.

Weigel, David. 2020. "Exit Poll Results and Analysis for the 2020 Presidential Election." https://www.washingtonpost.com/elections/interactive/2020/exit-polls/presidential-election-exit-polls/.

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. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.