Joe Biden projected to win Popular Vote in 2020 US Election with 51% of Vote

Results accurate within +/- 4 percentage points

Alen Mitrovski, Xiaoyan Yang, Matthew Wankiewicz

November 2nd, 2020

Abstract

It seems like everyone has been waiting for the 2020 election, almost as soon as Donald Trump won in 2016. In our report, we use a logistic regression model along with multilevel regression with post-stratification to predict who will win the election. According to our model, we predict that Joe Biden will win the popular vote over Donald Trump, 51% to 49% (results accurate within +/- 4%). We also find that although Joe Biden wins the popular vote, he is projected to lost the electoral college 278 to 260, meaning Donald Trump will be in office for another four years.

Keywords: Forecasting, US 2020 Election, Trump, Biden, multilevel regression with post-stratification;

1 Introduction

The 2020 United States presidential election will be on November 3rd, 2020 and once again, it is a battle between the Democrats and the Republicans. The Republican party is represented by the current president, Donald Trump who won the 2016 election against Hillary Clinton with a surprising come-from-behind victory. His success on election day has brought skepticism to American polling data as he was projected to lose by a wide margin up until the morning of the election. It will be interesting to see if this trend continues in the 2020 presidential election. Likewise, former vice-president, Joe Biden is representing the Democratic party. Joe Bidan served as the vice-president under former president Barack Obama during the years 2008-2016. His party will be looking to regain their control over the country after a polarizing four-year reign of president Trump. This year's election is particularly interesting due to unforeseen circumstances of the COVID-19 pandemic and the civil unrest sparked by the killing of George Floyd. The US has the potential to go on to two very different pathways depending on who wins the election. Thus, it will be important to be aware of the possible scenarios that can occur on election day.

For our report, we are using multilevel regression with post-stratification (MRP) to predict who will win the 2020 election. In short, MRP uses two different data set, creates a model using one and applies the results of the model on the larger data set. In our results, we look at which candidate won the popular vote (ie. had a higher percentage of votes). Also, we have calculated the total votes within each of the 50 American states to see how many points each candidate can accumulate for the electoral college. The electoral college is a system that was implemented in the United States to ensure that states with higher populations do not overwhelm the vote and provide greater power to individuals in historically smaller states (West [2020]). The presidential candidate with the most votes in a particular state wins all the electoral college votes for that state (West [2020]). The only exceptions to this rule are Nebraska and Maine who allocate their votes between both candidates based on the congressional district method (West [2020]). The electoral college has historically been viewed as a critical part of American democracy. However, its modern application has come into question in 2016 when Republican candidate Donald Trump won the election by accumulating the most electoral college votes despite losing the popular vote.

The survey data for this study was obtained from the Democracy Fund Voter Study Group; an organization that took the collaborative efforts of many scholars and analysts to get a better understanding of the evolving behaviour of American voters (Tausanovitch and Vavreck [2020]). The post-stratification data for this study was obtained from the Integrated Public Use Microdata Series (IPUMS). IPUMS is an organization that provides survey and census data through the collaboration of 105 national statistical agencies (Ruggles et al. [2020]). For this report, we utilized the data from the 2018 American Community Survey. We selected state, race, age, sex, education level and whether an individual was Hispanic or not as key variables to use for the post-stratification of the dataset.

In our report, we have 4 sections, not including the introduction. In the first section, we look at the data used to carry out our report and have included some plots to help show the distribution of our demographics. In the next section, we begin to run our model and interpret what some of the values we obtain mean. Next, we display the results of our predictions, using tables, maps and plots. Lastly, we discuss our results and address some weaknesses that our report encountered. Our report is carried out using R (R Core Team [2019]), and the library `Tidyverse` (Wickham et al. [2019]) was used the most during the report. The report was then compiled using R markdown (Allaire et al. [2020])

2 Data

We utilized the U.S.presidential election 2020 survey data from Nationscape conducted on June 25, 2020, as the study data to predict the 2020 election winner between Trump and Biden. The census data from IPUMS America Census Service was used as the post-stratification data for the survey data weighting adjustment. In the following subsections, we will introduce the variable selection basis and the main data differences between the survey and the stratification data.

According to the research (Pew [2020]) released on August 13, 2020, there are stark differences in how registered voters who support Donald Trump and Joe Biden view the importance of top voting issues. Comparable shares of Biden supporters and Trump supporters view Foreign policy and Supreme Court appointments as very important issues. The top five voting issues from Trump supporters are economy (88%), violent crime (74%), immigration (61%), gun policy (57%), and foreign policy (57%). By contrast, the largest shares of Biden supporters view health care (84%), coronavirus outbreak (82%), and racial and ethnic inequality (76%) the top three voting issues. There are substantial differences between Trump and Biden supporters on the importance of most issues, the widest gaps are on climate change (57%) and racial and ethnic inequality (52%). Based on these facts, we selected age, gender, race, hispanic, education, and states as the explanatory variables in the model. A binary variable with values "Trump" or "Biden" was used as the respondent variable. The values were extracted from the survey variable `vote_2020`. The variable had 5 selections: "Donald trump", "Joe Biden", "I am not sure/don't know", "I would not vote", and "Someone else". We kept the data with "Trump" or "Biden" only and removed the remaining data with other selections.

For our analysis, we decided to focus on 6 main demographics of respondents; gender, age, education level, hispanic, race and state.

We decided to use age because seniors have been consistent voters and normally decide to vote Republicans. A recent survey from October reported that seniors were shifting to vote for the Democrats (Frey [2020]). Young people are also breaking tradition as it's been reported that 51% of voters aged 18-34 said that they are very enthusiastic to vote (Janfaza [2020]). After looking at these numbers, we divided the age group into 4 categories: <=35, 36-50, 51-65, and 65+.

Next, we wanted to analyze gender because in the past, men had consistently held more conservative positions than women on a range of issues and as the parties became more ideological the gender gap kept growing from 8% in 1980 to 12% in 2000, to 13% in 2016, for women being more likely to vote for the Democrats [Pew, 2018]. Since this trend has been occurring, we decided to see how that would hold up in our model.

Next, we decided to focus on education level because of the 2016 election. White people with a four-year college degree or more education made up 30% of all validated voters. Among these voters, far more (55%)

said they voted for Clinton than for Trump (38%). Also, the much larger group of white voters who had not completed college (44% of all voters), Trump won their vote by more than two-to-one(64% to 28%) [CAWP, 2020]. Given this information, we decided to use education and we grouped it into three categories: High school or lower, Post Secondary or Higher and some Post Secondary.

We also focused on whether the respondent was hispanic or not. This variable was important in our opinions because of how Donald Trump has treated the hispanic community and how he claimed he was going to build a wall on the US-Mexico border in 2016. We also note that Professor Raul Madrid from the University of Texas said that the majority of hispanic people tend to vote Democrat (de Leon [2020]) and we feel that could be crucial for our model. For this variable, we separated people into two groups: hispanic or not hispanic.

We also want to look at race to see if that had any influence. Race is another important category to analyze because usually, people of a certain race will be more connected to a candidate, such as this year, many people of African-American descent are planning to vote for Joe Biden because of the hateful actions Donald Trump has made. Our race variable has been grouped into 4 categories: White, Black, Asian or other.

Lastly, we wanted to look at states because there are some states which consistently vote Republican or Democratic and they are important to take into account for our model. Generally, the east and west coasts tend to vote Democrat and the mid-west and south tend to vote Republican. We wanted to see if this same trend is going to occur in 2020.

2.1 Survey Data

The survey data was provided by Nationscape. Nationscape is a featured project of the Democracy Fund Voter Study Group. Democracy Fund + UCLA Nationscape is one of the largest public opinion surveys ever conducted - interviewing people in nearly every county, congressional district, and mid-sized U.S. city in the leadup to the 2020 election [Tausanovitch and Vavreck, 2020]. Data collection is performed by LUCID, Inc., which collaborated with Nationscape personnel on aspects of the study's design, sampling plan, and feasibility. Nationscape is a 16-month election study conducted by researchers at UCLA.

Samples are collected using stratified sampling. Nationscape groups potential respondents by demographics like age, gender and race and try to make their results correlate with the census data by using weights on the responses (Tausanovitch and Vavreck [2020]).

The population of the survey is people that live in the United States. The frame of the survey were Americans placed into sub-groups based on certain demographics. The sample of the survey were the people who responded to the survey and completed it properly. The surveys started in July of 2019 and will conclude in January of 2021. It includes interviews with roughly 6,250 people with completed surveys per week [Tausanovitch and Vavreck, 2020]. Interviews are conducted online wherever the respondent has access to a networked computer or mobile device.

To make sure results are as accurate as possible, the weekly questionnaires are designed for a 15-minute administration time. On average across all surveys, roughly 12% decline immediately. Another 5% or so drop off elsewhere in the survey. Nationscape removes about 8% for speeding or straight-lining through the survey. Speeding is defined as completing the survey in fewer than 6 minutes and straight-lining as selecting the same response for every question in the three policy questions [Tausanovitch and Vavreck, 2020]. Nationscape samples are provided by Lucid, a market research platform that runs an online exchange for survey respondents. The samples were drawn from this exchange match a set of demographic quotas on age, gender, ethnicity, region, income, and education. The survey data are then weighted to be representative of the American population. One set of weights is generated for each week's survey. The targets to which Nationscape is weighted are derived from the adult population of the 2017 American Community Survey of the U.S. Census Bureau [1] [Tausanovitch and Vavreck, 2020]. Nationscape uses the most recent census source available, their source is usually a year old or more. It may not reflect the existing population.

[1] Nationscape puts weights on the following factors: gender, the four major census regions, race, Hispanic ethnicity, household income, education, age, language spoken at home, nativity (U.S. - or foreign-born), 2016 presidential vote, and the urban-rural mix of the respondent's zip code

The Nationscape survey is useful because of how results are collected. As seen above, they make sure that responses are given accurately and make sure people take their time to complete their responses. This is extremely useful. After all, it means that most of the responses will be as accurate as possible because people that rush to fill in incorrect responses won't be included which gives us a strong set of data.

The survey does encounter some weaknesses though. The major weakness it encounters is that the survey data can be prone to sampling bias. Since the counties and cities are randomly selected, the city selected could have had a more republican or democratic leaning population. Since they could be sampling high proportions of supporters for one party, the results for that city or even that state could be skewed.

The survey data from June 25, 2020, was selected as the study data. It had 6479 responses and 265 variables. The survey was designed to cover age, gender, race, education, religion, remarks on Trump's performance in his first term, economy, propensity to the major political parties, their leaders, or candidates, ethnicity inequality, gun policy/issues, environment issues, green energy, health care/insurance, income tax, college tuition, abortion, international trade, immigration, legalizing marijuana, and COVID-19 pandemic (Tausanovitch and Vavreck [2020]). Our goal is to select variables that can explain and predict the winner of the U.S. presidential election 2020 by applying a logistic multilevel regression model. It is crucial to understand the most important voting issues, especially the differences between Trump and Biden's supporters, and hence the selected variables must have the power to distinguish the two camps.

2.2 Post-Stratification Data

The U.S. IPUMS America Census Service (ACS) is a project of the U.S. Census Bureau that has replaced the decennial census as the key source of information about the American population and housing characteristics [Ruggles et al., 2020]. The IPUMS database contains the annual samples from the 2000-2018 ACS [Ruggles et al., 2020]. We selected the most recent ACS 2018 data as the post-stratification data set.

The ACS data is collected through cluster samples, in other words, using stratified sampling. Stratified sampling is a method of sampling where the population is broken down into smaller sub-groups and then those subgroups are sampled and their responses are collected. For IPUMS, the population for their sample was the entire US population. The sampling frame is Americans who have completed the census. While the sample was taken every 1 in 100 persons on a national level.

IPUMS breaks down samples into households or dwellings as opposed to individuals because some topics require information from multiple people (Ruggles et al. [2020]). When the data is collected, samples are stratified based on certain characteristics into groups, and then they are sampled from those groups. Some variables that respondents are stratified by are geography, household size, and race.

Some strengths of the survey include how many people they can reach, and how accurate their data is. The sample that we used for our results was about 2.5 million people but IPUMS has access to more than that amount. This is useful because it allows us to get a really strong picture of the population. After all, we can observe trends in certain groups. The accuracy is also a strength because we know that we apply the results to the whole population. IPUMS uses a number called "design factor" which analyzes the errors in observed data vs sampled data and most of their results are around 1 which means the results are almost identical (Ruggles et al. [2020]).

One major weakness that the IPUMS data encounters is that it is prone to people submitting fake results. When cleaning our data we encountered many observations where 2-year-olds were making millions of dollars. This is an issue that is annoying but simple to fix, as we can just filter out the fake responses and focus on the correct ones. The difficult issue is that sometimes we cannot be sure which data is real or fake, there are some obvious ones like children being millionaires but other incorrect responses just have to be kept in the data.

According to their website, IPUMS says, "Non respondents are contacted via telephone for a computer-assisted telephone interview (CATI) one month later. One-third of the non respondents to the mail or telephone survey are contacted in person for a computer-assisted personal interview (CAPI) one month following the CATI

attempt" (Ruggles et al. [2020]). This means that IPUMS ensures that they keep track of non-respondents and encourage them to complete their surveys.

The post-stratification has about 2.5 million records. To maintain unity with the survey data, we selected age, gender, race, hispanic, education as the target variables. We cleaned and regrouped these variables to align with the survey data. Figures 1 and 2 show the corresponding variable comparisons between the survey and the post-stratification data. The solid line and dotted line represents the survey and the post-stratification data, respectively.

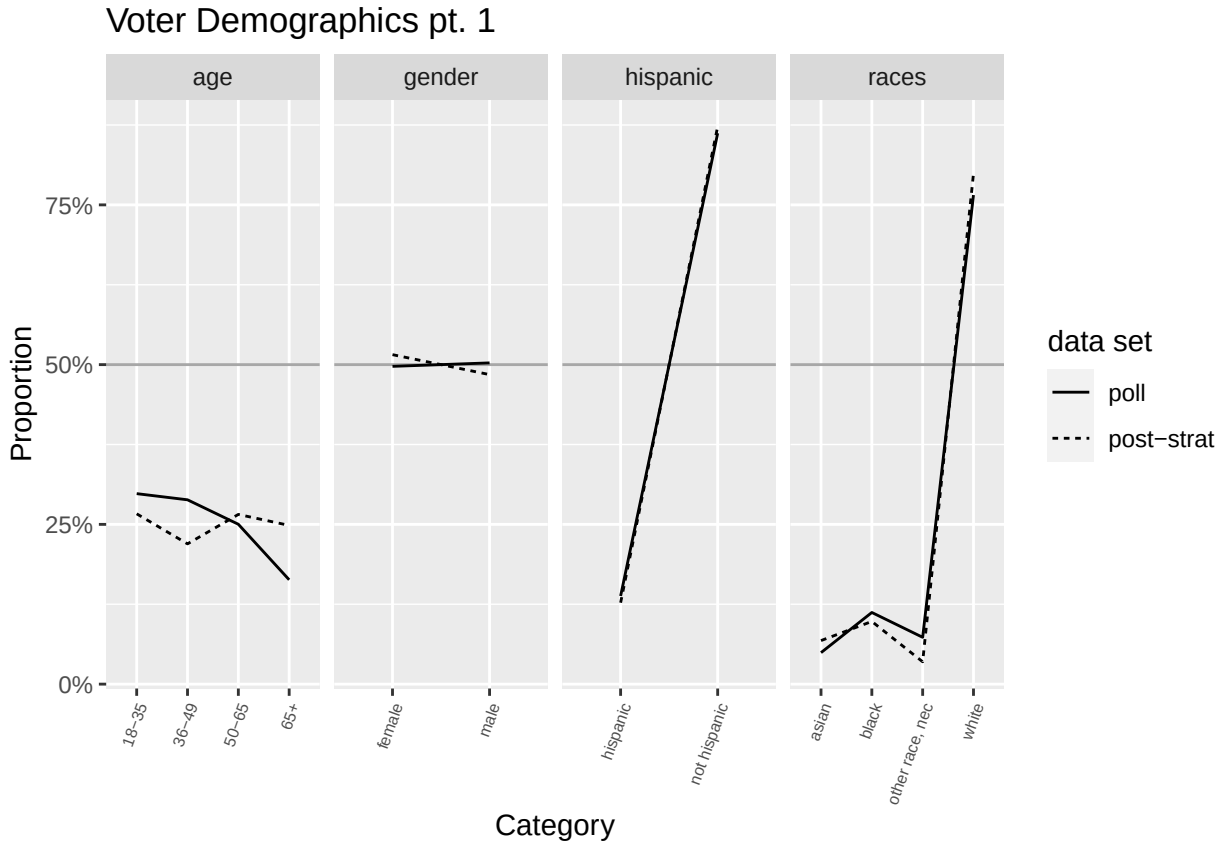

Figure 1: Demographics of Sample and Population

Figure 1 show us the voter demographics from the VSG data (Tausanovitch and Vavreck [2020]) vs the ACS data (Ruggles et al. [2020]). For the most part, we see from Figure 1 that the polling data matches up with our post-stratification data. The major differences present are that more respondents were Asian and less were other in the post-stratification data compared to the polling data. We also see that ages are slightly different as well, the polling data appeared to collect more responses from younger people while the post-stratification data had a roughly even distribution.

Table 1: Who decided voters plan to support (Polling Data)

Candidate	Number of Respondents	Proportion (%)
Donald Trump	2481	48
Joe Biden	2719	52

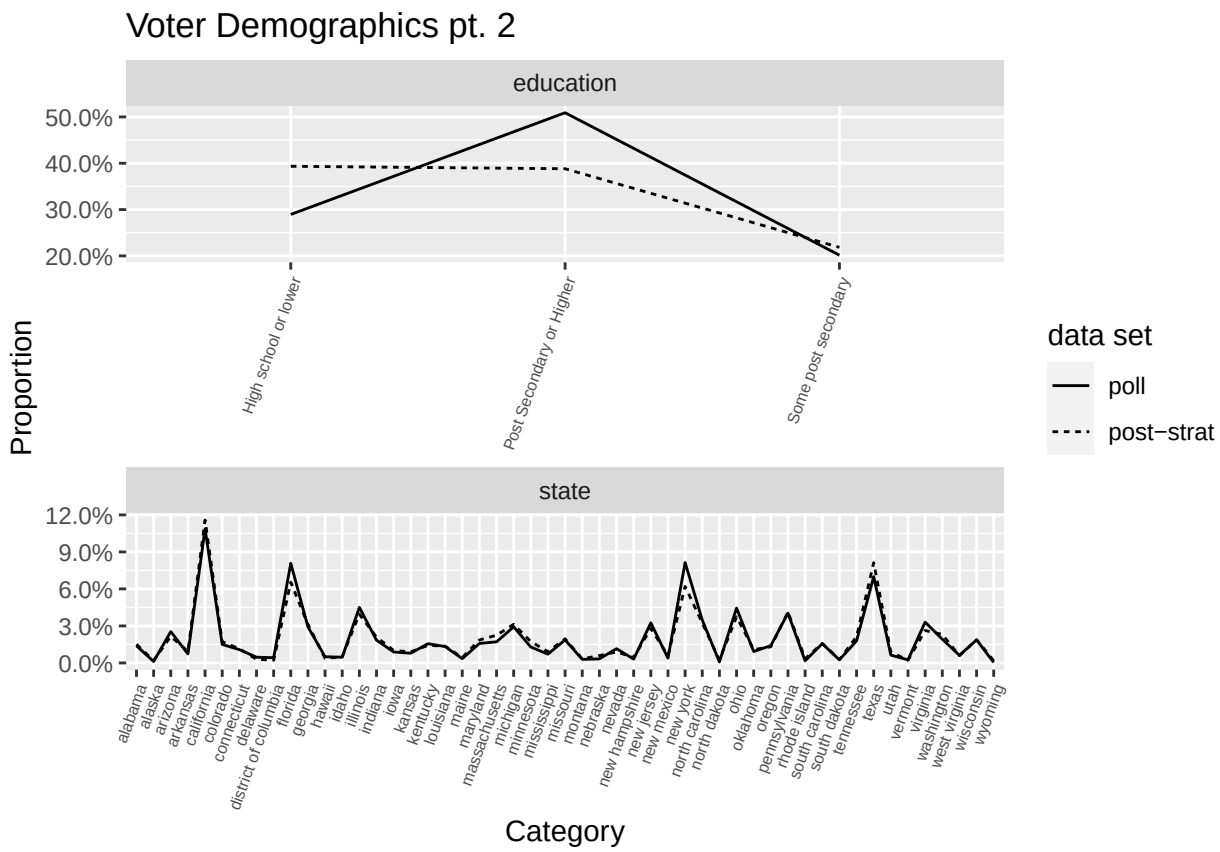

Figure 2: More Demographics of Sample and Population

Figure 2 shows us more of the voter demographics from the VSG data (Tausanovitch and Vavreck [2020]) vs the ACS data (Ruggles et al. [2020]). We can see that the polling data has more Post-Secondary graduates and less High School graduates compared to the post-stratification data. It also illustrates the population distributions by states. Overall, the curves present very similar patterns. In most states, the weights are very close. The noticeable gaps are in Florida, New York, and Texas. The survey data over-represents the proportions of respondents in Florida and New York by 1.5% and 2.8% respectively. In Texas, the survey data under-represents the population by 1.1% compared with the post-stratification data.

Table 1 shows the proportion of decided voters who plan to vote for Donald Trump or Joe Biden. The data used to create this table is from the Voter Study Group (Tausanovitch and Vavreck [2020]). We see that Joe Biden is expected to win the popular vote before we implement the model.

In general, by comparing with the survey data and the post-stratification data, the largest gap is in education variables. Age ranks the second one. The other variables do not show a large percentage of discrepancies between the two data sets. The post-stratification method is applied to make the survey results more representative for the whole population.

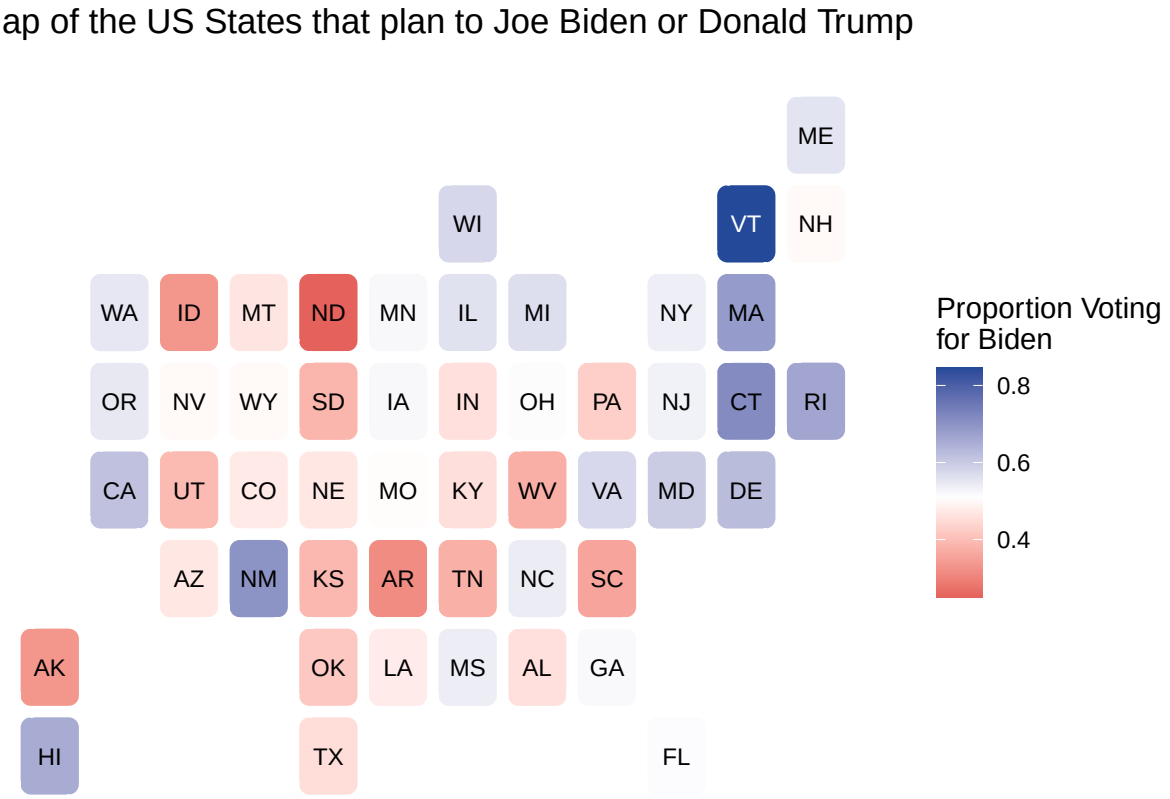

Figure 3: Proportion of voters choosing Biden from Polling Data

Figure 3 shows the proportion of voters from each state who plan to vote for Joe Biden, this is using the polling data. We see that the east coast favours Biden much more than the mid-west does which is normal. We also see that Florida and Georgia are both on the fence on who they plan to vote for and could potentially be favouring Biden. This is significant because the last time that Georgia voted for the Democratic party was in 1992 (Britannica [2020]), showing a major shift to the blue.

3 Model

For our analysis, we plan to use multilevel regression with post-stratification. Multilevel regression with post-stratification (MRP) is a type of analysis where we fit a model using a smaller data set, in this case, our polling data and then use the results of the model to apply it to a larger population.

The main steps for MRP are: Find the data set you want to use to create your model. For our scenario, we used the polling data from the Voter Study Group (Tausanovitch and Vavreck [2020]). Next, you must create a model using your smaller sample. We used logistic regression and the data used was the polling data. Our equation takes the form of equation (1), as seen below. Once you have your model, you must apply it to your larger data set to give an idea of the population. For our report, we used the Census data from IPUMS (Ruggles et al. [2020]).

MRP is extremely useful not only because of how simple it is but also because it allows us to estimate the preferences of a population using individual responses from surveys. Also, as Kennedy and Gelman discuss in their paper "Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample", MRP works best when you have a population and variables of interest and want to apply these variables using two data sets with different characteristics (Kennedy and Gelman). Lastly, concerning surveys, since you only use one survey to create

your model, you don't encounter issues with having to have multiple surveys for each region/state and lets under-sampled populations to be represented through post-stratification.

MRP does encounter some weaknesses as well. Once again we can look to Kennedy and Gelman's paper where they say that MRP is dependent on how accurate the model is (Kennedy and Gelman). If the generated model makes incorrect predictions or assumptions, your post-stratification will be incorrect and will hurt your results.

To predict whether or not a person plans to vote for Joe Biden or Donald Trump, we plan to build a logistic regression model using data from the Voter Study Group (Tausanovitch and Vavreck [2020]) and then post-stratify it using Census Data (Ruggles et al. [2020]). Since logistic regression only works for binary response variables, we created a variable called `vote_biden` which returns a 1 if the respondent plans to vote for Joe Biden and a 0 if they plan to vote for Donald Trump.

The logistic regression model takes the form of:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{agegroup} + \beta_3 x_{race} + \beta_4 x_{state} + \beta_5 x_{education} + \beta_6 x_{hispanic} \tag{1}$$

Once we have our regression model, we will use the `predict` function in R (R Core Team [2019]), to apply our model to the Census data (Ruggles et al. [2020]). We do this by grouping the Census data by the demographics we plan to analyze (sex, race, age, education level, hispanic or not, state), and then applying the model to each of those groups. After applying the model, we will receive probabilities that a person in that group will vote for Joe Biden. Once the predictions are complete, we can use them to find out who will win the popular vote or how many electoral colleges a candidate will win. We also can use a 95 percent confidence interval, which means that we are 95 percent certain that the true value for the population (in this case popular vote percentage), is within the range we obtain. From this, we can also conclude that our results will be accurate between +/- 4%.

In equation (1), each β represents a coefficient that the regression model will compute for us. As for our variables, we have chosen to use sex, age, race, education, state, and whether the respondent is hispanic. We decided to use the first 3 because they are generally strong predictors of which candidate a person would support, such as how some states tend to vote republican year after year while some states flip between democratic and republican almost every election. We also included education level, initially we were going to include income but decided that education level is more concrete on describing a person, as opposed to income. Lastly, we wanted to focus on whether the respondent was hispanic and if so, where they were from. This variable was important for our predictions because we know how poorly Donald Trump has spoken of hispanic people and we believe they could have a strong impact on the election.

The output of the logistic regression model will give us a probability of whether or not a voter plans to vote for Joe Biden or not. To find this probability, we take the sum of the right side of equation (1) and plug it into the equation below:

$$\frac{e^{sum}}{1+e^{sum}} \tag{2}$$

Equation (2) is just a manipulation of equation (1), where e is the exponential equation and sum is the sum of the right side of equation (1). We see that as the sum of the right side increases, the probability that a person will vote for Joe Biden increases as well. We are running our regression model using the `glm()` function in R (R Core Team [2019]). The decision to run this model over other models like linear regression was made by the fact that we were predicting a binary variable about a voter's decision. Since there are only two possible options our data will likely follow an S shape and a straight line equation will not be helpful to model this relationship. Another strength present for logistic regression is that when combined with post-stratification it allows us to take information from under-represented populations and it allows their views to be accounted for more greatly. For example, our polling data (Tausanovitch and Vavreck [2020]),

includes only 2 observations from Wyoming, but using multilevel regression with post-stratification, we can have that expanded to over 3000 people!

Our model does have some weaknesses, since the output must be binary, we cannot account for other candidates or a person deciding not to vote. This issue isn't too large because our main goal is to determine which of the two main candidates will be chosen by the people of America. Another weakness we do encounter with our model and multi-level regression with post-stratification is that it has a strong dependence on the survey data. This is a weakness because if the survey has any gaps or there are any tweaks we need to make, it can change the course of results.

Table 2 (below), 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])[2]. We can see that our p-values for variables like gender, age, hispanic and education return very significant results while the states have varying levels of significance. What our p-values tell us is whether or not our variable is statistically significant in impacting our response variable, in this case, whether the respondent will vote for Biden or not. When we look at our p-values, we want them to be as small as possible and can assume they're significant if it's under 0.05. As for the coefficients, the value we receive is in terms of log odds. When the log odds are positive, it means that the person will likely vote for Biden and if it's negative it means they'll likely vote for Trump.

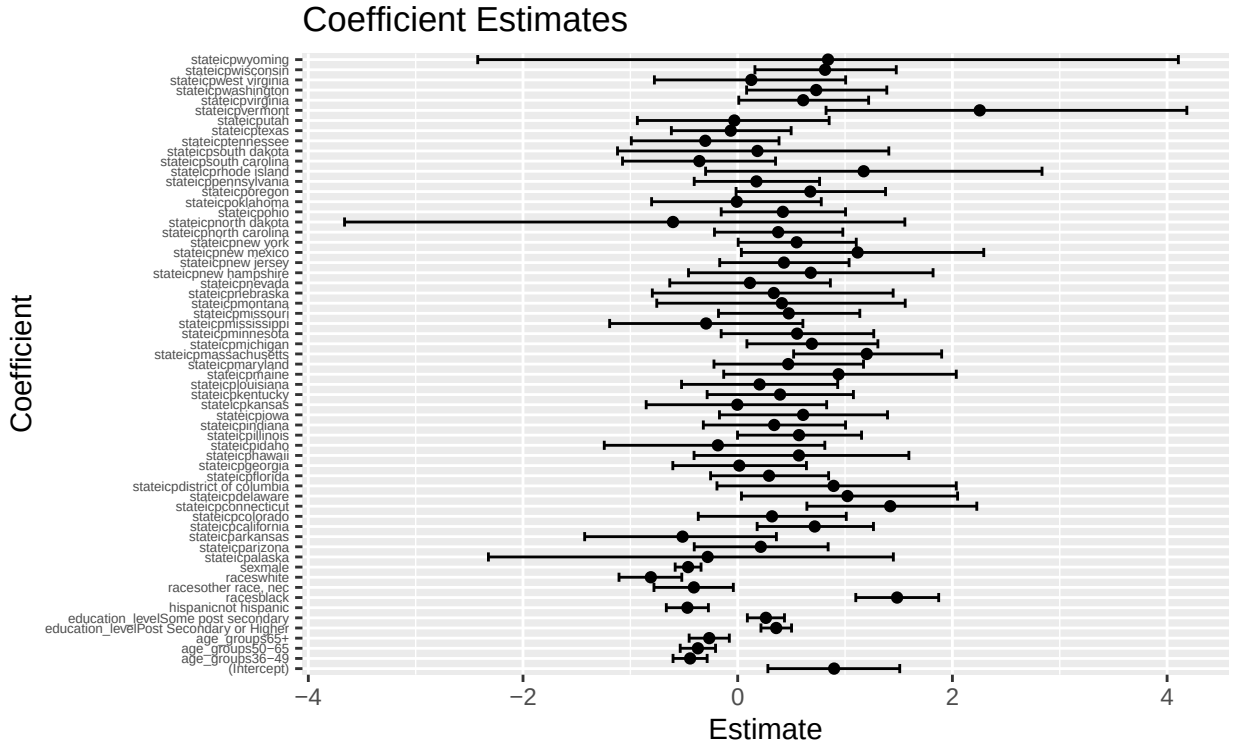

Figure 4: Coefficient Estimates

Figure 4 shows us the coefficients that would fit into equation (1) using the polling data (Tausanovitch and Vavreck [2020]). We also have error bars present, which show the upper and lower estimates for the coefficients. What we have to look out for in this scenario is that coefficients with negative values would mean that the person is more likely to vote for Donald Trump (with that characteristic) and positive values mean the person is more likely to vote for Joe Biden. Table 2 shows a numerical view of Figure 4, along with p-values.

[2]data was cleaned using the `tidy` function called `broom` (Robinson et al. [2020])

Table 2: Coefficients from the Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.90	0.31	2.87	0.00	0.28	1.51
sexmale	-0.46	0.06	-7.52	0.00	-0.58	-0.34
age_groups36-49	-0.44	0.08	-5.47	0.00	-0.60	-0.28
age_groups50-65	-0.37	0.08	-4.42	0.00	-0.54	-0.21
age_groups65+	-0.27	0.10	-2.78	0.01	-0.45	-0.08
racesblack	1.48	0.20	7.55	0.00	1.10	1.87
racesother race, nec	-0.41	0.19	-2.17	0.03	-0.78	-0.04
raceswhite	-0.81	0.15	-5.43	0.00	-1.11	-0.52
stateicpalaska	-0.28	0.91	-0.31	0.76	-2.32	1.45
stateicparizona	0.22	0.32	0.68	0.50	-0.41	0.84
stateicparkansas	-0.51	0.45	-1.13	0.26	-1.43	0.36
stateicpcalifornia	0.72	0.28	2.60	0.01	0.18	1.26
stateicpcolorado	0.32	0.35	0.91	0.36	-0.37	1.01
stateicpconnecticut	1.42	0.40	3.53	0.00	0.64	2.23
stateicpdelaware	1.02	0.51	2.01	0.04	0.03	2.05
stateicpdistrict of columbia	0.89	0.56	1.59	0.11	-0.19	2.04
stateicpflorida	0.29	0.28	1.04	0.30	-0.25	0.85
stateicpgeorgia	0.01	0.32	0.05	0.96	-0.60	0.64
stateicphawaii	0.57	0.51	1.12	0.26	-0.41	1.59
stateicpidaho	-0.18	0.52	-0.36	0.72	-1.24	0.81
stateicpillinois	0.57	0.29	1.94	0.05	0.00	1.15
stateicpindiana	0.34	0.34	1.01	0.31	-0.32	1.00
stateicpiowa	0.61	0.40	1.53	0.13	-0.17	1.40
stateicpkansas	0.00	0.43	-0.01	0.99	-0.85	0.83
stateicpkentucky	0.39	0.35	1.14	0.26	-0.28	1.08
stateicplouisiana	0.20	0.37	0.55	0.58	-0.52	0.93
stateicpmaine	0.94	0.55	1.72	0.09	-0.13	2.04
stateicpmaryland	0.47	0.35	1.33	0.18	-0.22	1.17
stateicpmassachusetts	1.20	0.35	3.42	0.00	0.52	1.90
stateicpmichigan	0.69	0.31	2.22	0.03	0.09	1.31
stateicpminnesota	0.55	0.36	1.53	0.13	-0.15	1.27
stateicpmississippi	-0.29	0.46	-0.64	0.52	-1.19	0.61
stateicpmissouri	0.48	0.34	1.42	0.16	-0.18	1.14
stateicpmontana	0.41	0.58	0.71	0.48	-0.75	1.56
stateicpnebraska	0.34	0.57	0.59	0.55	-0.80	1.45
stateicpnevada	0.11	0.38	0.30	0.77	-0.63	0.86
stateicpnew hampshire	0.68	0.57	1.19	0.24	-0.46	1.82
stateicpnew jersey	0.43	0.31	1.40	0.16	-0.17	1.04
stateicpnew mexico	1.12	0.57	1.96	0.05	0.03	2.29
stateicpnew york	0.55	0.28	1.96	0.05	0.00	1.10
stateicpnorth carolina	0.38	0.30	1.24	0.22	-0.22	0.98
stateicpnorth dakota	-0.60	1.20	-0.50	0.62	-3.66	1.56
stateicpohio	0.42	0.29	1.43	0.15	-0.15	1.00
stateicpoklahoma	-0.01	0.40	-0.02	0.98	-0.80	0.78
stateicporegon	0.68	0.35	1.91	0.06	-0.02	1.38
stateicppennsylvania	0.17	0.30	0.59	0.56	-0.41	0.76
stateicprhode island	1.17	0.78	1.51	0.13	-0.30	2.83
stateicpsouth carolina	-0.36	0.36	-0.98	0.32	-1.07	0.35
stateicpsouth dakota	0.18	0.63	0.29	0.77	-1.12	1.41
stateicptennessee	-0.30	0.35	-0.86	0.39	-0.99	0.39
stateicptexas	-0.06	0.28	-0.23	0.82	-0.62	0.50
stateicputah	-0.03	0.45	-0.07	0.95	-0.93	0.85
stateicpvermont	2.25	0.82	2.76	0.01	0.82	4.18
stateicpvirginia	0.61	0.31	1.98	0.05	0.01	1.22
stateicpwashington	0.73	0.33	2.20	0.03	0.08	1.39
stateicpwest virginia	0.13	0.45	0.28	0.78	-0.78	1.01
stateicpwisconsin	0.81	0.34	2.43	0.02	0.16	1.48

Using the outputs of the logistic regression model, we can get an equation that follows the form of (1), but with the β values filled out. This equation is difficult to write out because of the many variables, but in short, if the person's characteristic fits in with a certain variable, it is used in the equation. Then the equation is summed up and the probability is found using equation (2).

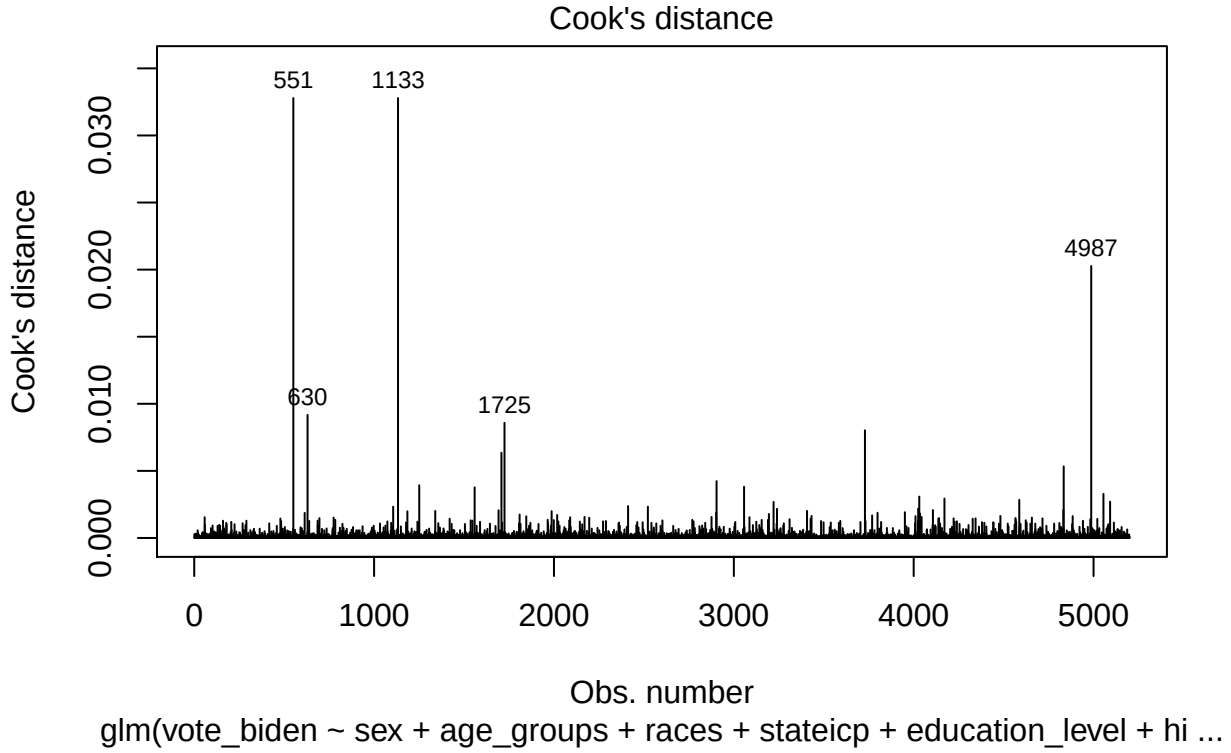

Figure 5: Cook's Distance Plot for Model

Figure 5 shows the cooks distance for observations in our polling data set. Cook's distance is useful for checking if a model is working correctly because it tells us how far a point is from the predicted value of it, telling us which points negatively impact our model. In our scenario, the distance between values would be the true value using our `glm` model vs what the `predict.glm` function returned. Figure 5 takes all observations of our polling data (Tausanovitch and Vavreck [2020]), and calculates the Cook's distance for it, showing which points can hurt the results of our model. We find that out of the 5200 observations, there aren't too many points that deviate from what we predict. This makes sense for our model because, in the real world, some people who will support Donald Trump or Joe Biden, even if they don't "fit" the usual voter demographic for the candidate. Since the number of observations with large Cook's Distances is low, we can conclude that our model is fairly strong.

We can also look at some common assumptions for logistic regression and check if our model follows them or not. One assumption is that the data set used is large. The data we used to create the model had about 5000 observations, which is high but maybe in the future, we could use a larger data set to increase the accuracy of predictions. Another assumption that logistic regression makes is that the observations are independent of each other. We know that the observations used for our model were independent, not only because the Voter Study Group ensures that there aren't duplicates but because they assign each respondent a unique ID number.

Lastly, using our coefficients (Table 2), we can calculate the upper and lower bounds for the probabilities predicted from our model. For our lower probability (white male, from North Dakota, aged 36-49, with some post-secondary education and not hispanic) we get a probability of 22% for supporting Biden. For our upper probability (black woman, from Vermont, aged 18-35, with post-secondary or higher education and considers

themselves hispanic) we get a probability of 99% for supporting Joe Biden.

4 Results

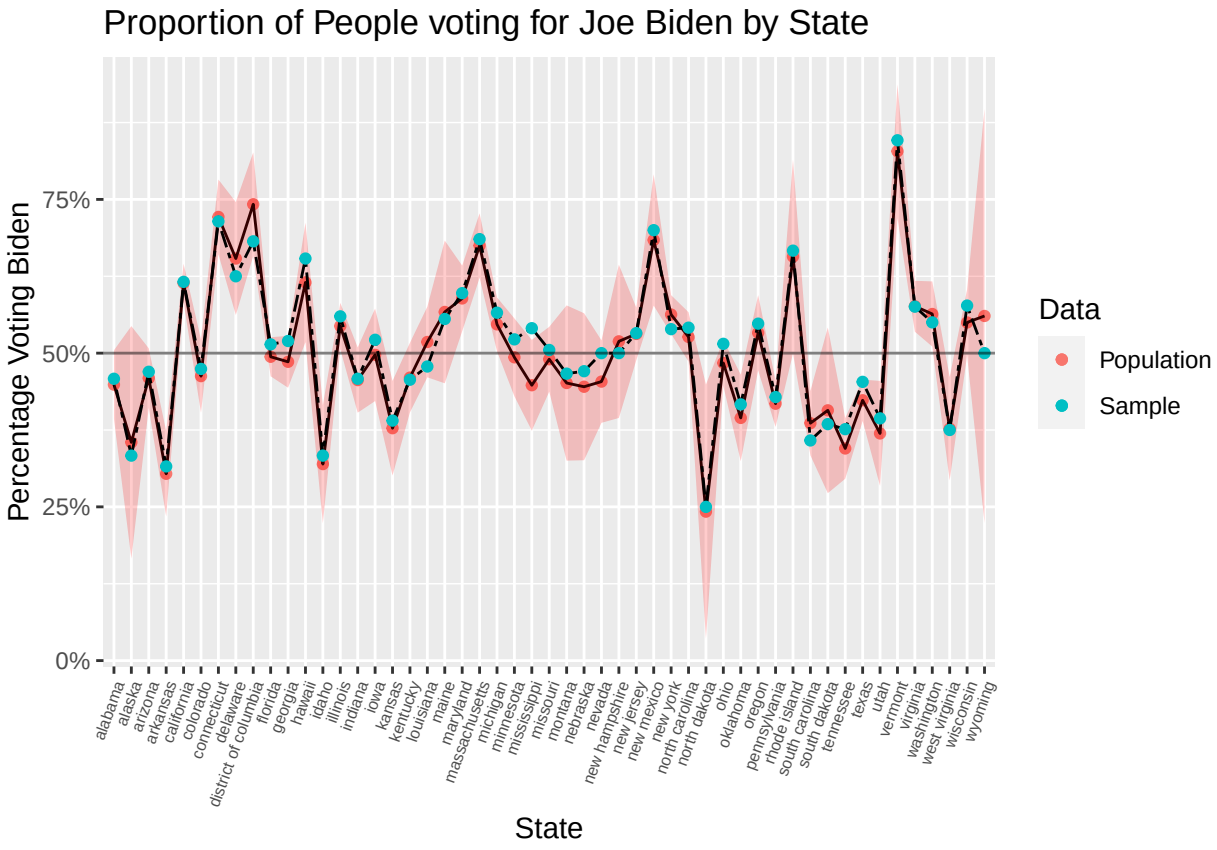

Figure 6: Proportion of each State Voting for Biden

Figure 6 shows us the proportion of respondents voting for Joe Biden broken down by state, along with the predicted proportions using MRP. We have also included the errors for each predicted value. We can see that the predicted proportions are fairly close to what the polling data showed. But, we also have to acknowledge the errors present, we can see that many states' errors cross over the 50% line which could be pivotal for each candidate because of the winner take all nature of the electoral college. We can see that some states like Florida and Georgia have small errors but some states who have very large ones like North Dakota and Wyoming.

Table 3: Joe Biden Voting Result Estimates

	Lower Estimate	Mean Estimate	Upper Estimate
Number of Colleges	191	260	423
Proportion of Vote (%)	46	51	56

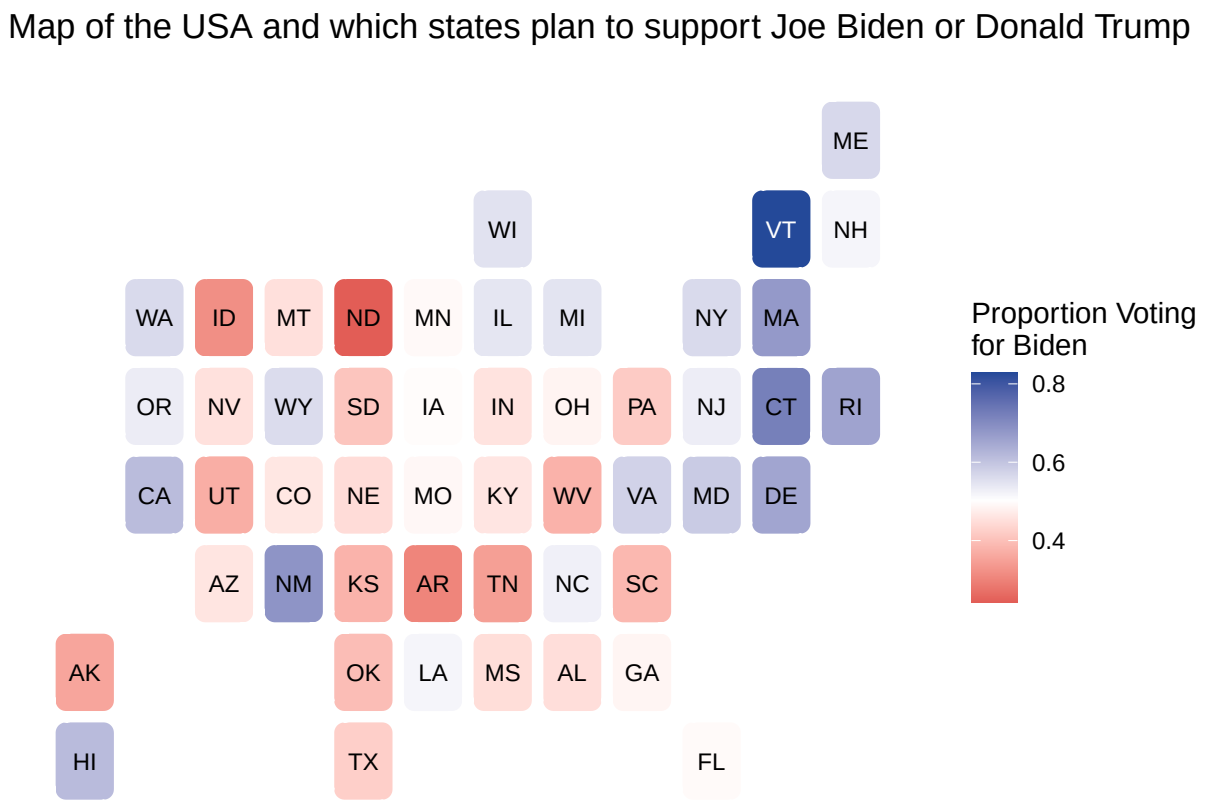

Figure 7: Proportion of Voters from Each State Voting Joe Biden

Figure 7 is an amazing view of all the states and which candidate they are leaning towards voting for using `statebins` (Rudis [2020]). We see that states like Vermont, Connecticut and California are some of the more "blue" states, meaning they plan to vote for Joe Biden, while North Dakota, Arkansas and Idaho are the "red" states, planning to vote for Donald Trump. The states that are more white can be the most important for the race when it comes to deciding an actual winner through the electoral college. We see that Florida, Louisiana and New Hampshire are some of the more undecided states, and a switch in these states can influence the election greatly.

We can also take note of Figure 3. We noted earlier that that Florida and Georgia were undecided from our polling data but after applying the model to the post-stratification data, we see that Florida and Georgia are actually leaning red. We do see some other states change their support as well, such as Louisiana, Missouri and Ohio, to name a few.

Table 3 shows the lower, mean and upper predictions for the results of the election for Joe Biden. We see that on the lower end, Biden can expect to get 46.1% of the popular vote while only getting 191 electoral colleges. We also see that our middle estimate says Biden will get 51% of the popular vote while still losing the election by getting only 260 colleges. Lastly, on the upper estimate for Biden's results, he can get 56% of the popular vote while getting 423 colleges! This truly shows how close this election is, as we can see in Figure

6, many states are hovering around the 50% mark, which shows the colleges can go either way. The number of electoral colleges by state were found on the Britannica Encylopedia's website (of Encyclopaedia Britannica).

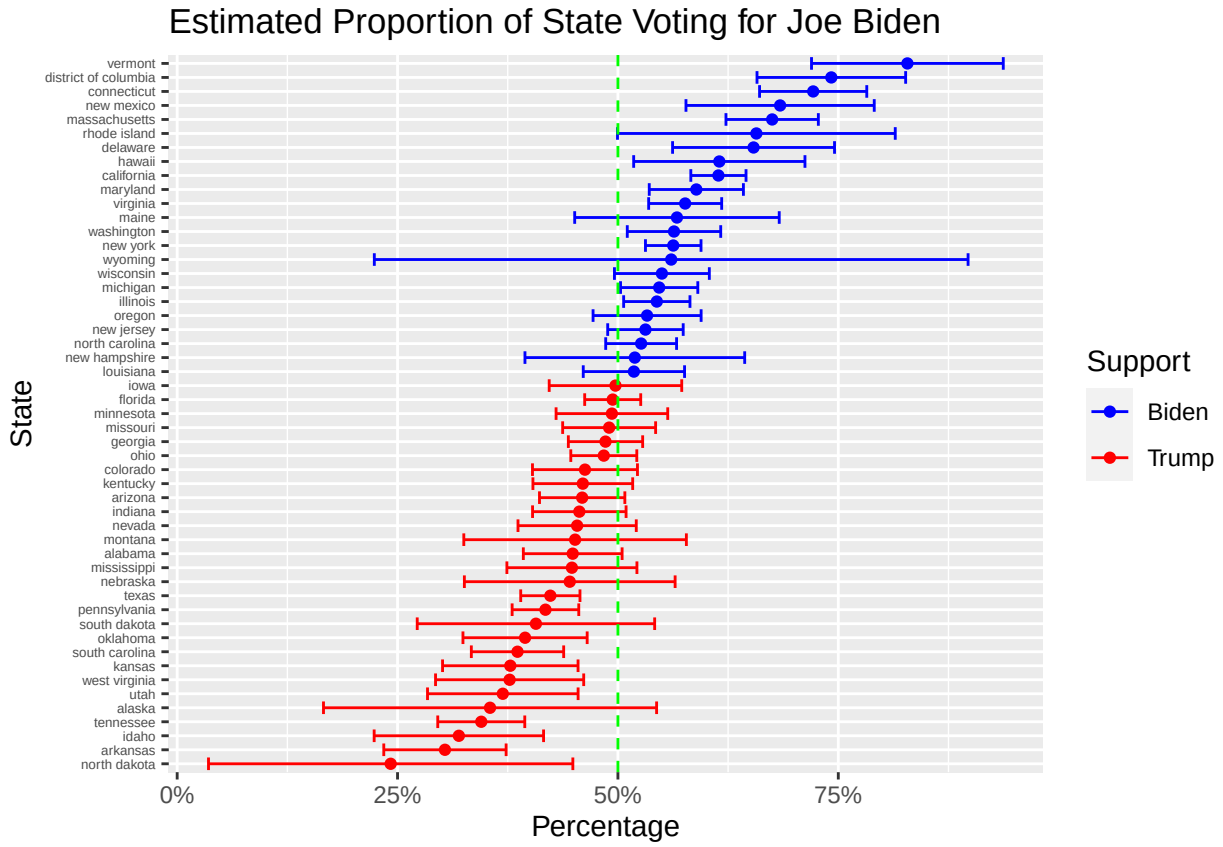

Figure 8: Proportion of Biden votes by state

Figure 8 is a different view of Figure 6, this time only focusing on the predictions for each state with the errors included. We see that many states have error bars overlapping the 50% line, showing that many states are a toss-up, given the nature of the electoral college.

5 Discussion

As seen in Table 3, the mean estimate is that Joe Biden will win the popular vote with 51% of the vote. However, the American election operates according to the electoral college and as such, it is possible that a candidate wins the popular vote but loses the electoral college. This has happened on two occasions: in 2000 when Al Gore lost to George W. Bush and in 2016 when Donald Trump won over Hillary Clinton (Blum [2020]). According to our results, Biden would win the popular vote but lose the electoral college vote with a mean estimate of 260 points (270 is the threshold for victory). The abolishment of the electoral college has been considered, however, professor Opal from McGill university notes that it would require a constitutional amendment to pass this unlikely change (Blum [2020]).

Through Figure 7, we can visualize numerous national trends of voting behaviour. First, it is observable that Joe Biden and the Democrats have a strong hold of the west coast states. However, this was not always the case. Up until 1988, the Republicans consistently had the support of California voters, who make up 55 electoral college points (Krishnakumar et al. [2016]). However, California had a massive spike in Hispanic and Asian immigrants during the 1990s (Krishnakumar et al. [2016]). This growing voter base was put off by

the anti-immigration stance of the Republican party; Particularly in 1994, when the Republicans attempted to pass a law that would deny public services to illegal immigrants (Krishnakumar et al. [2016]). As a result of information, it comes as no surprise that Biden has a strong lead in California in our data and collects the 55 electoral college points.

Secondly, based on Figure 7, Republican party leader Donald Trump performs very well in the American mid-west. This would follow-up Trump's 2016 victory in the historically democratic states of Iowa, Ohio, Michigan and Wisconsin (Lauck [2017]). Many of these states had not voted Republican since the '80s and were considered pillars of the Democratic voter base (Lauck [2017]). However, blue-collar workers in these states have flipped their votes over their dissatisfaction with the economy and manufacturing sector under Democratic leadership (Wilson [2017]). Considering the importance of an improving economy to mid-west voters, the economic growth under the Trump administration should lead to another set of victories in this part of America (News [2020]). This information would explain the results in our data for these mid-west states.

Thirdly, as seen in Figure 7, Joe Biden seems to possess a strong lead in north-eastern coastal states. This geographic area of America has historically, and consistently, been comprised of democratic-leaning states. This is the case because Democrats tend to gather a large percentage of the votes in big cities (Thompson [2019]). The north-eastern coast includes many of these populous cities such as New York, Boston, Washington D.C. and Newark. As these cities have become more ethnically diverse, less manufacturing-oriented and more educated over the years, the voter base has shifted towards the Democratic party (Thompson [2019]). Like California on the west coast, these big east coast cities are focal points for incoming immigrants who tend to vote for the democratic party (Thompson [2019]). As a result, the results from Figures 7 and 6 for north-eastern coastal states fall in line with prior voting behaviour and tendencies.

In Figure 4, we see that Hispanic voters are more likely to vote for, democratic nominee, Joe Biden because the log-odds are greater than 0. As previously mentioned, the Republican's willingness to pass Proposition 187 to restrict public services to illegal immigrants was poorly received by the Hispanic community, in particular (Krishnakumar et al. [2016]). However, modern circumstances surrounding President Trump have also played a role in the aversion of the Hispanic community towards the Republican party. In the 2019 paper by Gutierrez et al., Gutierrez cites that Mexican Americans and immigrants felt threatened and targeted by Trump's proposals and rhetoric (Gutierrez et al. [2019]). Despite comments specifically targeting illegal immigrants from Mexico, a focus group in Florida composed of Puerto Rican voters noted that they feel Trump is villainizing the entire Latino population (Gutierrez et al. [2019]). This could explain why Hispanic voters lean towards voting for Biden based on results seen in Figure 5. Gutierrez also follows-up by suggesting that the anger towards the 2016 election result, which saw Trump become president, could mobilize the Hispanic community to vote in greater numbers during the 2020 election (Gutierrez et al. [2019]).

Based on our results, white voters have a strong propensity to vote for Trump in the 2020 election (Table 2). This is important as white voters make up over 75% of the sample population and the post-stratification data (Figure 1). In the paper by Major et al., the researchers suggest that white voters are strongly supporting the Trump campaign as a result of the changing demographics in the United States (Major et al. [2016]). As forecasted by the U.S. Census Bureau, the proportion of non-white individuals in the United States will be considerably higher than white individuals by the year 2060 (Census [2012]). When reminded of this projection, Major et al. report that white voters with high ethnic identification became more concerned with their declining influence and have begun supporting Trump's anti-immigration policies, as a result (Major et al. [2016]). In contrast, our results show that black voters have a high propensity to vote for Biden (Table 2). As suggested by researchers, this support is likely due to their resentment towards Trump. Trump's opposition to the Black Lives Matter group and failure to condemn racism at key moments is viewed as threatening among African American voters (Towler and Parker [2018]). Black voters may represent a smaller percentage of the vote than white individuals, however, their support for Joe Biden has the potential to impact the election on a grand scale.

Based on the polling data in Figure 1, male and female voting is nearly equal at 50%. However, our results suggest that men are more likely to vote for Republican nominee Trump, while women are more likely to vote for Democratic nominee Joe Biden (Table 2). It is important to note that in the 2016 election, a significantly

greater number of women voted in the election than men (Pew [2018]). Furthermore, 54% of women voted for the Hillary Clinton of the Democratic party Pew [2018]). This disposition to vote for the Democratic party has been a staple of American women for the past 40 years. When president Raegan was nominated in 1980, the Republican party adopted right-leaning viewpoints on several issues relevant to women (Thompson [2020]). In particular, the Republican party began opposing abortion rights for women and ceased supporting the Equal Rights Amendment (Thompson [2020]). Even though Raegan won that election, the decline of female voting support of the Republican party was commencing. This trend could serve as an explanation for the female support of Joe Biden in our results.

Contrary to classic convention, our results indicate that older individuals are progressively more likely to vote for Democratic nominee Joe Biden (Figure 3). One possible explanation for this change in voter behaviour could be the COVID-19 outbreak that has plagued the entire world this year. As per the Centers for Disease Control and Protection, COVID-19 deaths are increasingly severe with age and disproportionately higher past the age of 55 (CDC [2020]). As a result, the COVID-19 outbreak is viewed as a critical issue in this year's election for the older percentage of voters (PEW [2020]). Trump's downplaying of the virus and refusal to wear a mask may have upset the older voter population; culminating in Biden's surprising lead over older individuals in our research (Happe et al. [2020]). Also, these voters may be impressed with Biden's plan to provide free testing to all Americans and financial support to those that lost their job due to the pandemic (Happe et al. [2020]).

Our results indicate that the higher education level completed, the more likely an individual will vote for Democratic nominee Joe Biden during the 2020 presidential election (Figure 4). A possible explanation for this general trend is that large, urban cities contain more universities and educational complexes compared to their rural counterparts. As a result, cities are more accessible for their inhabitants to complete higher levels of education. It is well understood that these large, urban cities tend to vote for the Democratic party while rural areas lean towards the Republican party (Thompson [2019]). As such, Republican-leaning states in the rural, mid-west will naturally be inhabited by less educated individuals than the urban individuals living in Democrat-voting coastal states.

We are very confident in our results for states that have p-values less than 0.05 such as California, Connecticut, Michigan and others (Figure 4). The results from these states are statistically significantly and provide stronger evidence that we should reject the null hypothesis that Joe Biden will not win the popular vote in the 2020 election. In contrast, we are less confident in states that have p-values greater than 0.5. For example, the states of Oklahoma, Utah and Georgia have p-values that are relatively high compared to other states (Table 2). These results distort our ability to assume that their results are significant in predicting the winner.

Considering the concurrences between our data results and the historical information we just outlined, we believe that Joe Biden will win the popular vote but ultimately lose the election through the electoral college. We believe Biden will gain the support of big-city residents to win the popular vote but fail to make up ground in the rural, blue-collar mid-west that Trump had won in 2016. However, it is possible that our prediction doesn't hold due to the incredibly close results in certain states, such as Florida, Georgia and others (Figure 7). Also, considering the weaknesses and limitations of this study (Outlined in the next section), the validity of the data used in this research is subject to debate.

Lastly, we can take a look at the state of Ohio. Ohio has been one of the best predictors for who will win the presidential election. Ballotpedia calls the state a "bellwether" because, since 1900, Ohio has voted for the winning candidate 93.33% of the time (Ballotpedia). This is a remarkable number and as seen in 7, Ohio appears to be redder than blue which may give us evidence that although Biden won the popular vote, Trump may scrape out a victory.

Overall, this election is going to be extremely close and according to our results including Figures 3, 7, 6 and Tables 3 and 2, November 3rd is going have a long night!

5.1 Weaknesses and Next Steps

As seen in the 2016 election, the polls leading up to the presidential election can be inaccurate. Typically, this has not been an issue for the American polling companies. However, since becoming a presidential candidate, Donald Trump has become a polarizing figure in the media and among voters. As a result, some Trump voters may be unwilling to show their support for the Republican candidate.

In addition, 2020 has been plagued by the COVID-19 pandemic and has changed the landscape of the presidential election. Not only has it become a key concern for voters when deciding their next president, but the virus has also changed the campaigning procedures and influencing voting mediums. For example, democratic representative Joe Biden has been holding parking lot rallies where supporters are social distancing within their cars while showing their support. Likewise, Trump rallies have seen a lower number of supporters than his previous campaign trails. However, most importantly, voters are changing their approach for voting with safety precautions regarding the pandemic. 91.6 million votes have already been cast through the mail to avoid coming in contact with others at the polling booths. As a result, this election is far from usual or predictable. Various aspects of the election are critically affected by the pandemic virus and in turn, the polling results from ACS should not be considered reliable or representative of the population as they may have been in previous years.

A limitation of the survey dataset we are using is that it is from June 2020. As a result, the data is missing four months where voters may have changed their opinions and political allegiances. Healthcare, in particular, is an issue of very high importance for Americans during this election (PEW [2020]). As of September 2020, four Covid-19 vaccines have entered phase 3 of clinical trials and a vaccine may arrive by early 2021 (NIH [2020]). This kind of impactful development, among others, may influence voter behavior and alter the dataset if collected now.

For future studies, we would suggest allocating more of the polling company budget towards examining the education level of individuals to get a better estimate of the results and gain a clearer picture of what the American population is leaning towards before the election. As seen in Figure 4, the estimate for this variable is not particularly strong in favor, nor against, voting for Joe Biden. In addition, there is little separating the estimates for the various categories of education level in our research.

6 Code

Code supporting this analysis can be found at: https://github.com/matthewwankiewicz/US_election_forecast

References

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2020. URL https://github.com/rstudio/rmarkdown. R package version 2.3.

Ballotpedia. Presidential voting trends in ohio. URL https://ballotpedia.org/Presidential_voting_trends_in_Ohio.

Benjamin Blum. Understanding the electoral college and how the u.s. president is really elected, Oct 2020. URL https://www.cbc.ca/news/world/electoral-college-explainer-1.5768507.

Britannica. United states presidential election of 1992, Oct 2020. URL https://www.britannica.com/event/United-States-presidential-election-of-1992.

CAWP. 2020 presidential gender gap poll tracker, Sep 2020. URL https://cawp.rutgers.edu/presidential-poll-tracking-2020.

CDC. Covid-19 provisional counts - weekly updates by select demographic and geographic characteristics, Oct 2020. URL https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm.

U.S Census. U.s. census bureau projections show a slower growing, older, more diverse nation a half century from now, Dec 2012. URL https://www-census-gov.myaccess.library.utoronto.ca/newsroom/releases/archives/population/cb12-243.html.

Luis de Leon. 2020 election: The impact of the latino vote in texas, Oct 2020. URL https://www.kvue.com/article/news/politics/vote-texas/election-2020-texas-latino-vote-impact/269-e27ced91-8f67-4bcf-bcc9-f0f56c5bc67f.

William H. Frey. Older voters may secure a biden victory in 2020's swing states, Oct 2020. URL https://www.brookings.edu/blog/the-avenue/2020/10/28/older-voters-may-secure-a-biden-victory-in-2020s-swing-states/.

Angela Gutierrez, Angela X. Ocampo, Matt A. Barreto, and Gary Segura. Somos más: How racial threat and anger mobilized latino voters in the trump era. *Political Research Quarterly*, 72(4):960–975, 2019. doi: 10.1177/1065912919844327.

Mackenzie Happe, Kate Sullivan, and Max Pepper. Where trump and biden stand on major policy issues, Oct 2020. URL https://www.cnn.com/2020/09/29/politics/trump-vs-biden-2020-election-policies/index.html.

Rachel Janfaza. Data suggests 2020 could see a surge in young u.s. voters, Oct 2020. URL https://www.ctvnews.ca/world/america-votes/data-suggests-2020-could-see-a-surge-in-young-u-s-voters-1.5159607.

Lauren Kennedy and Andrew Gelman. Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample. URL https://arxiv.org/pdf/1906.11323.pdf.

Priya Krishnakumar, Maloy Moore, and Armand Emamdjomeh. After decades of republican victories, here's how california became a blue state again, Dec 2016. URL https://www.latimes.com/projects/la-pol-ca-california-voting-history/.

Jon K Lauck. Trump and the midwest: The 2016 presidential election and the avenues of midwestern historiography. *Studies in Midwest History*, 03(01), Jan 2017.

Brenda Major, Alison Blodorn, and Gregory Major Blascovich. The threat of increasing diversity: Why many white americans support trump in the 2016 presidential election. *Group Processes amp; Intergroup Relations*, 21(6):931–940, 2016. doi: 10.1177/1368430216677304.

BBC News. Us 2020 election: The economy under trump in six charts, Sep 2020. URL https://www.bbc.com/news/world-45827430.

NIH. Fourth large-scale covid-19 vaccine trial begins in the united states, Sep 2020. URL https://www.nih.gov/news-events/news-releases/fourth-large-scale-covid-19-vaccine-trial-begins-united-states.

Editors of Encyclopaedia Britannica. United states electoral college votes by state. URL https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124.

Pew. An examination of the 2016 electorate, based on validated voters, Aug 2018. URL https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/.

Pew. Important issues in the 2020 election, Oct 2020. URL https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/.

PEW, Oct 2020. URL https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

David Robinson, Alex Hayes, and Simon Couch. *broom: Convert Statistical Objects into Tidy Tibbles*, 2020. URL https://CRAN.R-project.org/package=broom. R package version 0.7.1.

Bob Rudis. *statebins: Create United States Uniform Cartogram Heatmaps*, 2020. URL https://CRAN.R-project.org/package=statebins. R package version 1.4.0.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. *IPUMS USA: Version 10.0 [ACS 2018]*. Minneapolis, MN: IPUMS, 2020. URL https://doi.org/10.18128/D010.V10.0.

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

Derek Thompson. How democrats conquered the city, Sep 2019. URL https://www.theatlantic.com/ideas/archive/2019/09/brief-history-how-democrats-conquered-city/597955/.

Derek Thompson. Why men vote for republicans, and women vote for democrats, Feb 2020. URL https://www.theatlantic.com/ideas/archive/2020/02/how-women-became-democratic-partisans/606274/.

Christopher C. Towler and Christopher S. Parker. Between anger and engagement: Donald trump and black america. *The Journal of Race, Ethnicity, and Politics*, 3(1):219–253, 2018. doi: 10.1017/rep.2017.38.

Darrell M West. It's time to abolish the electoral college, Mar 2020. URL https://www.brookings.edu/policy2020/bigideas/its-time-to-abolish-the-electoral-college/.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Reid Wilson. How the midwest slipped away from dems, Aug 2017. URL https://thehill.com/homenews/state-watch/347414-how-the-midwest-slipped-away-from-dems.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020. URL https://yihui.org/knitr/. R package version 1.30.

Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2020. URL https://CRAN.R-project.org/package=kableExtra. R package version 1.3.1.