Biden Expected to Win 54% of Popular Vote in 2020 US Presidential Election*

4% Confidence Interval Based on a Survey from June 2020

James Bao, Zakir Chaudry, Alan Chen, Xinyi Zhang

02 November 2020

Abstract

After a shocking upset in the 2016 US Presidential election, everyone has their eyes on the 2020 election to determine the leader of the free world for the next four years. In this Paper we trained a model using survey level survey results . Our model predicts that Joe Biden will win the 2020 Presidential election with a 53.5% of the popular vote and a confidence level of 95%. Our prediction as well as our breakdown of votes by demographic group could potentially provide the candidates of the election with information on how to target voters.

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

1 Introduction

2 Data

To train our model to predict the outcome of the 2020 US presidential election, we used Wave 49 of the Nationscape Dataset (results from the week of June 18-24, 2020). We will discuss how this data was collected, its key features, and what the data looks like in the section titled "Individual-level survey dataset."

To make predictions on the outcome of the 2020 US presidential election, we used the results of the 2018 American Community Survey. We will discuss how this data was collected, its key features, and what the data looks like in the section titled "Post-stratification dataset." The explanation of multilevel modelling with post-stratification can be found in the "Model" section.

2.1 Individual-level survey dataset

2.1.1 Data collection

The Nationscape Project is 16-month-long voter study (from July 2019 to January 2021) that conducts weekly surveys regarding the 2020 US presidential election. We will mainly discuss Wave 49 of the Nationscape Dataset for the reminder of this paper.

From June 18, 2020 to June 24, 2020, Nationscape collected data on public opinion about the 2020 presidential campaign and election by conducting 15-minute online interviews. Their target is the American "population." Unfortunately, the published information on their methodology is not more specific as to what constitutes a

*Code and data supporting this analysis are available at: https://github.com/JamesBond0014/sta304_ps4.

member of the American population. Presumably (based on analyzing the data), their target population is all adult individuals presently residing in the United States.

Nationscape used the audience of market research platform Lucid as its sampling frame. Sampling frames are lists of the units (individuals in our case) that will be selected for the survey sample, meaning that the survey respondents on Lucid form a list of a subset of the target population (from which a sample will be taken). Finally, a sample matching the demographics of the American population is selected from the frame using a purposive sampling method. This is a non-probability sampling method where the researcher decides which samples are most representative of the target population. More specific information about their sampling method was not provided (besides a statement that the sampling was not random). After being contacted by Lucid to take the survey, respondents are immediately redirected to Nationscape survey software where the questionnaire starts.

Nationscape reported that the nonresponse rate was about 17%. Another 8% of responses were removed for speeding (spending less than 6 minutes completing the survey) or for "straight-lining" answers (selecting the same response for all policy questions) resulting in a final sample size of 6,532 respondents. To reduce the effects of non-response bias and to ensure results were representative of the US population, survey responses were weighted using data from the 2018 American Community Survey (for demographic variables) and from the United States Elections Project and MIT Election Lab (for 2016 vote). This ensures that the discrepancy between the target population and survey responses is minimized. Lastly, Nationscape assessed the representativeness of the survey sample by including questions from the 2018 Pew evaluation of non-probability samples and comparing their results to Pew findings and government benchmarks. Overall, the difference between Nationscape results and government benchmarks was comparable to the difference between Pew findings and government benchmarks. Consequently, Nationscape concluded that estimates from their dataset should be considered sufficiently valid (at least in comparison to other political polling non-probability samples analyzed by Pew).

The strengths of Nationscape's survey methodology include pilot testing their questionnaire for several weeks, which allowed staff to finetune survey questions and respondent selection criteria. Along the same line, the survey strikes a good balance between being detailed enough to capture useful data while being short enough to hold respondent attention. Furthermore, the high frequency of the data collection process provides the dataset with a week by week breakdown of voter sentiment, potentially capturing changes in public political opinion as news or controversies break. Lastly, the response rate is extremely good for an online survey, indicating that the vast majority of the selected sample responded. In fact, a response rate of over 80% is very high and likely due to the distribution of the survey through the Lucid platform (and certain characteristics of or certain incentives for survey respondents on the platform).

On the other hand, a major weakness of the survey is that sampling was not conducted at random but rather demographic criterias were designed by the Nationscape staff. Another weakness is that the sampling frame is not necessarily representative of the American population (those who aren't members of survey panels or aren't comfortable sharing political opinions are likely not represented). Lastly, the results are likely subjected to response bias because of the subjective nature of the research topic. However, as previously mentioned, Nationscape addressed these weaknesses by comparing their results to the results from 2018 Pew evaluations on non-probability sampling (and found the accuracy and representativeness of their dataset to be comparable).

2.1.2 Data features and visualization

The full dataset for Wave 49 consists of 6,532 responses for over 260 variables. They cover topics ranging from the presidential candidates to government policies, current events, political views and respondent demographics. In the interest of brevity, we will focus our discussion on the explanatory and response variables relevant to our model. We aim to predict the winner of the popular vote in the 2020 US presidential election so our response variable of choice is vote_2020. We chose age, gender, race_ethnicity, state, and education as explanatory variables based on the demographic characteristics that are most important in determining user vote and our ability to match these variables with the post stratification dataset. In greater detail, here are the chosen variables:

- vote_2020: the vote of the respondent given that the Democratic nominee is Joe Biden and the Republican nominee is Donald Trump

Table 1: Respondent 2020 US presidential election vote distribution

vote_2020	Frequency
Donald Trump	2481
Joe Biden	2719
Someone else	250
I would not vote	374
I am not sure/don't know	651

- age: the age of the respondent in years at the time of the survey

Table 2: Respondent age statistics

Statistics	Values
Min.	18.00000
1st Qu.	31.00000
Median	43.00000
Mean	45.16546
3rd Qu.	59.50000
Max.	93.00000

- gender: the sex of the respondent (the options being "Male" or "Female")

Table 3: Respondent gender distribution

gender	Frequency
Female	3309
Male	3170

- race_ethnicity: the race of the respondent

Table 4: Respondent race distribution

race_ethnicity	Frequency
White	4816
Black, or African American	774
American Indian or Alaska Native	90
Asian (Asian Indian)	102
Asian (Chinese)	84
Asian (Filipino)	46
Asian (Japanese)	21
Asian (Korean)	14
Asian (Vietnamese)	13
Asian (Other)	37
Pacific Islander (Native Hawaiian)	10
Pacific Islander (Guamanian)	1
Pacific Islander (Samoan)	3
Pacific Islander (Other)	8
Some other race	460

- state: the state the respondent resides in (table omitted in the interest of space)

- education: the highest level of education completed by the respondent

Table 5: Respondent education distribution

education	Frequency
3rd Grade or less	11
Middle School - Grades 4 - 8	26
Completed some high school	638
High school graduate	1079
Other post high school vocational training	324
Completed some college, but no degree	1327
Associate Degree	570
College Degree (such as B.A., B.S.)	1477
Completed some graduate, but no degree	238
Masters degree	643
Doctorate degree	146

For the variables age, gender, race_ethnicity, state, and education, we did not find similar equivalents in the dataset. We did find that the variable trump_biden (the candidate that the respondent would support if the election was a contest between Donald Trump and Joe Biden) was similar to our selected variable vote_2020. However, as vote_2020 is more representative of the nature of the popular vote, we did not end up choosing trump_biden.

When cleaning the data, we merged some of the factors of the variables to match the granularity of the data in the post-stratification dataset. This included splitting age responses into bins of size 10, reducing education to two bins ("High School or Less" and "Post Secondary or More"), and combining Asian Indian, Korean, Filipino, Vietnamese, and Pacific Islander ethnicities into "Asian (Other)" (Chinese and Japanese remained their own factors because this is the level of specificity available in the post-stratification dataset). Lastly, we took a subset of the dataset where respondents had decided to vote for either Trump or Biden in the 2020 US presidential election (for the purposes of being able to predict the popular vote using a binary model).

The distributions of each of our cleaned variables (with the exception of gender) are shown in the following few pages.

Out of the respondents, approximately 50% were female and 50% were male (hence gender was not graphed). The distribution of respondent age approaches a bell shaped curve with much fewer "18 to 28" year-olds and individuals aged "79 and above" than any other age group (Figure 1). This is not too surprising for the older age division because their technology use is likely very limited and thus few older respondents are reached. However, it's surprising so few young adults participated in the survey. The majority of survey respondents were White (76%), followed by African American (11%), Asian (4%), Native American (1%), and some other race, meaning bi- or multi-racial (6%) (Figure 2). This mostly aligns racial demographic estimates from the US Census Bureau although they reported that only 2.8% of the American population is of two or more races while 13% are African American and 6% are Asian. This redistribution of minority frequencies is unfortunate but shouldn't end up significantly impacting our prediction (because our calculation of the popular vote is conducting using post-stratification data while we are only using survey data to train our model on the relationship between our explanatory and response variables).

71% of responses had some level of post secondary education (Figure 3), which doesn't line up with US Census estimates of only 32% of Americans having obtained a bachelor's degree or higher (again, emphasizing the need for prediction on post-stratification data). The state distribution has a lot of factors but basic sanity checks suggest that the distribution is proportional to the American population with the majority of respondents residing in California, Florida, New York, and Texas (Figure 4), which are the four most populous states. Lastly, 52% of respondents planned on voting for Joe Biden with Donald Trump behind by 4 points (Figure 5). We look forward to predicting the election results by training our model on this dataset and seeing how the voting distribution changes when generating predictions given post-stratification data.

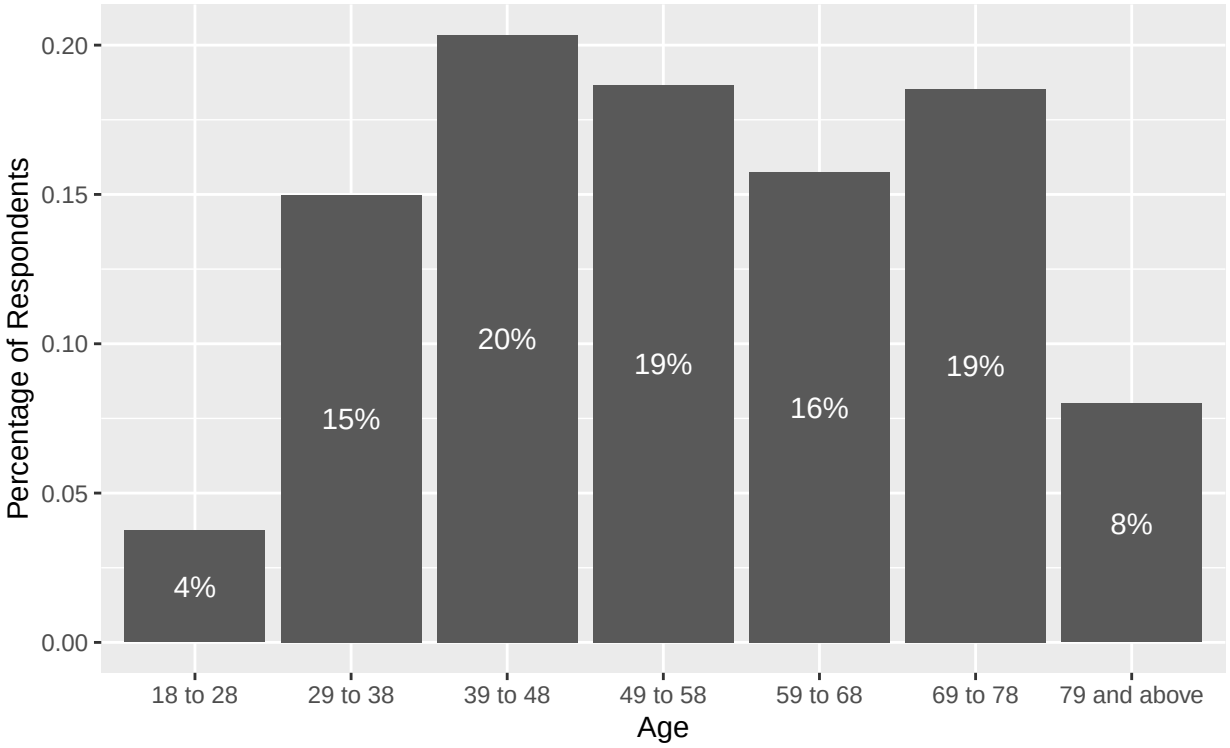


Figure 1: Distribution of the age of survey respondents in percentages.

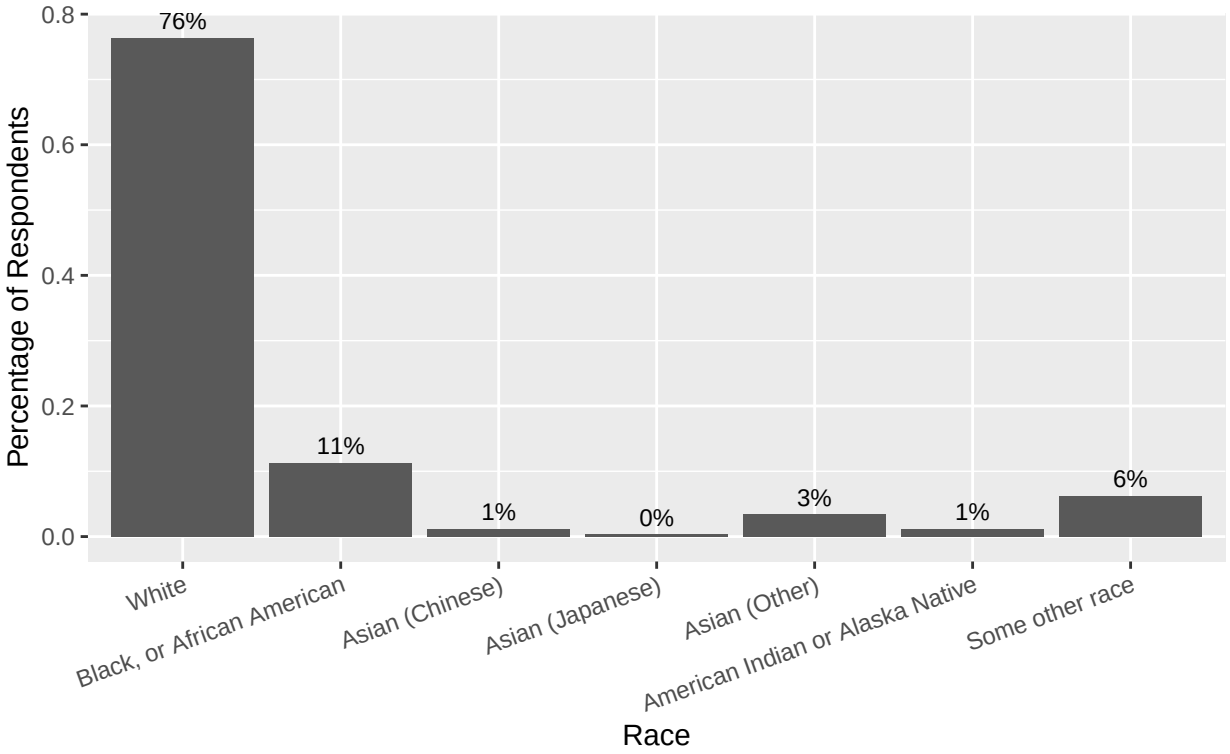


Figure 2: Distribution of the race of survey respondents in percentages.

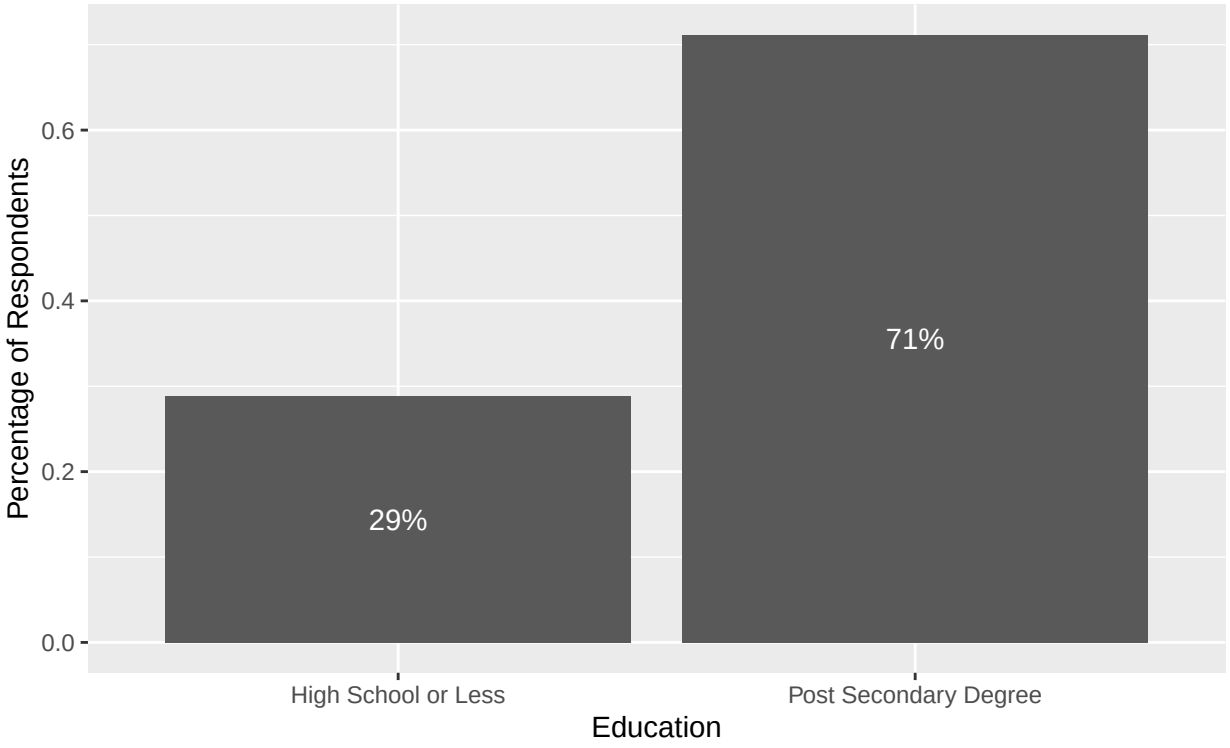


Figure 3: Distribution of the education of survey respondents in percentages.

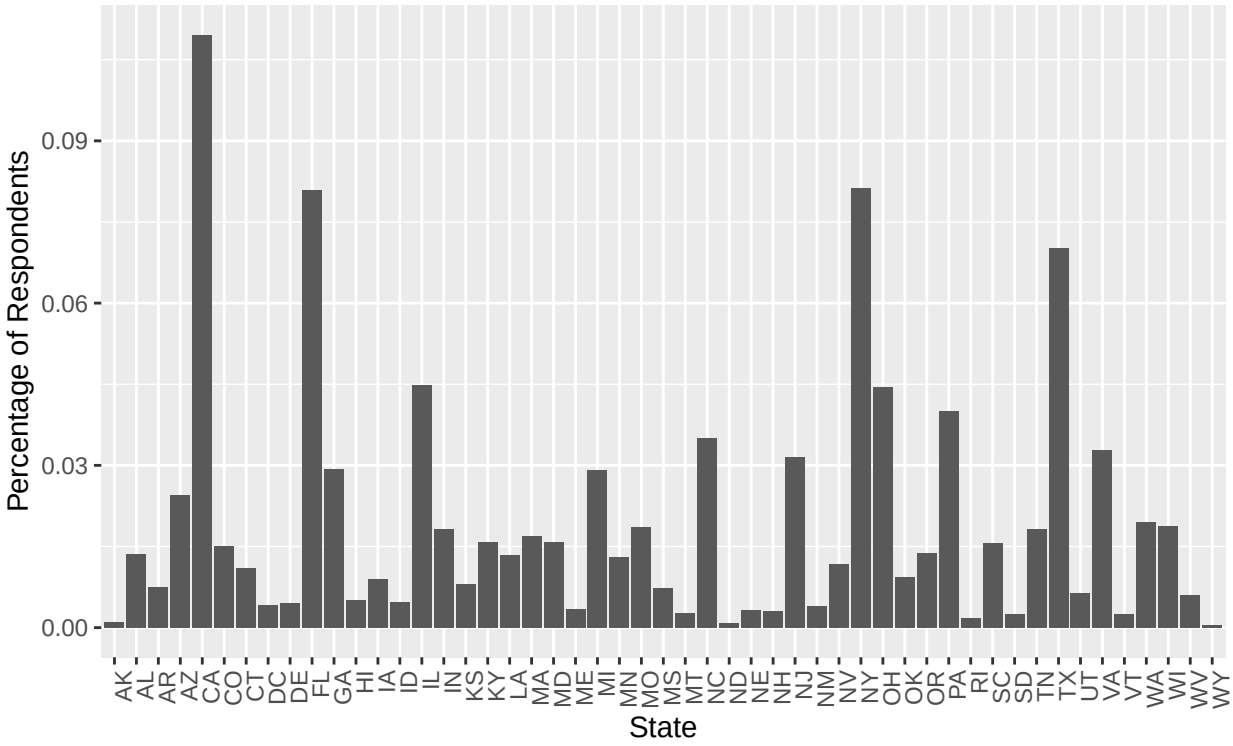


Figure 4: Distribution of the state of survey respondents in percentages.

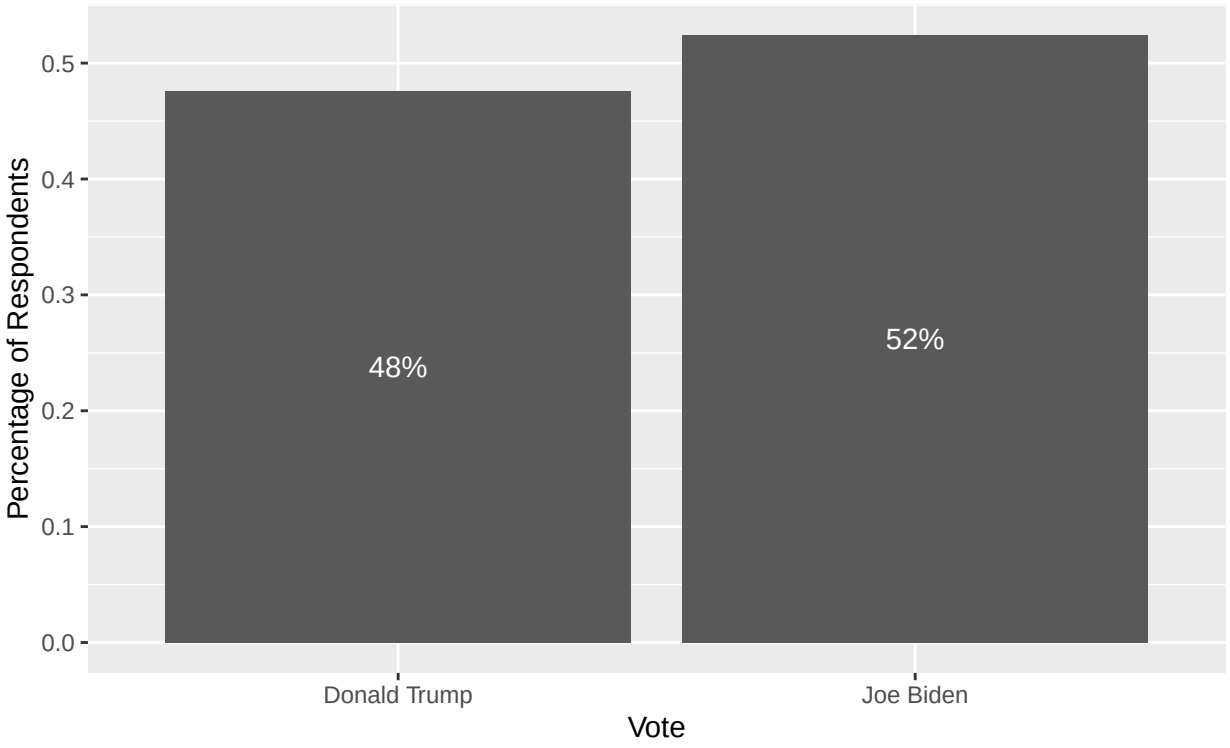


Figure 5: Distribution of the vote of survey respondents in percentages.

2.2 Post-stratification dataset

2.2.1 Data collection

The post-stratification dataset was gathered from the American Community Survey (ACS), a project aiming to mitigate issues stemming from the Census' 10-year intervals by providing an annualized version of data like that produced by the decennial census long form. The dataset used in this study is specifically that of the 2018 ACS data. The ACS data can be accessed at the IPUMS website. More details on attaining and cleaning data are found in 01-data_cleaning-post-strat.R in the scripts folder of the git repository. TODO

The target population, much like the Census, is essentially anyone who resides in a dwelling in the US. Following this, the sampling frame of the ACS is the Master Address File that is maintained by the US Census Bureau. Created for the 2000 Census, it was originally based on the 1990 Address Control File and the United States Postal Service's Delivery Sequence File. The maintaining and updating of this file is crucial to the efforts of the ACS and any other body that makes use of it. In addition, the ACS samples 2.5 percent of the population living in "Group Quarters", which are non-housing units (eg. nursing homes, prisons, college dorms, etc.). In total, the 2018 ACS data contains about 3.2 million observations, sampled from across the country.

Every month, a systemic sample is created for each US county or equivalent, where they are mailed the ACS survey at the start of the month. As of February 2002, the sampling rate of all counties has been 2.5%, except for Houston, Texas, which is sampled at 1% (to reduce cost due to the size of the population). For every site, the sampling is broken into two steps. The first step is sampling 17.5% of the population, which is then subsampled from to achieve the desired 2.5 or 1% of the county. All non-respondents are subsequently contacted by phone for a computer assisted telephone interview one month later. One third of non-respondents that have reached this point are then sampled from to be contacted for a computer assisted personal interview following the previous telephone interview attempt. Beyond this sample (referred to as the

National Sample or Supplemental Sample), data was also collected at 31 selected test sites to represent areas with various county population sizes or areas that were difficult to enumerate. The ACS data is weighted in order to ensure reliable and usable estimated regarding the population.

2.2.2 Data features and visualization

The ACS Questionnaire asks questions regarding every inhabitant in the residence it is sent to. However, the most information is required of "Person 1", the person whose name the residence is owned in, being bought in, or rented in (or any adult, if none of those labels apply). The ACS questionnaire is extremely like that of the US Census Questionnaire, given that it's meant to be a substitution for it. It was developed after the Census Bureau was provided with various subjects that other federal agencies justified as important, categorizing each subject as "mandatory", "required", or "programmatic", from which the ACS collected data for both the "mandatory" and "required" subjects. This approach illustrates itself in the questionnaire, as every question is answered by an objective fact, whether it's a checkbox, a number, or a short answer (such as the person's major). This allows little ambiguity in how the question must be answered, and makes each response fairly easy, which is a major strength of the questionnaire. However, depending on the number of people in the residence, the survey can be on the longer side (up to 11 pages). Combined with the fact that it is legally mandated that the response is filled out, with a potential fine of up to $5000, respondents may feel the need to rush through or give a fake response to questions they may not know immediately (such as when the residence was built). In addition, in the case that there is not one single person responsible for the residence, like in the case where multiple people are renting a room from one landlord who doesn't live at the residence, then the data of those who are not "Person 1" is not collected, despite the fact that they are on the same standing in the eyes of the survey. However, no one has been prosecuted for not filling out the survey since 1970 and the case outlined in the second point is relatively rare in the total population, so these weaknesses are fairly minor in the grand scheme of things.

As mentioned above, the variables we used were age (the age of the respondent), gender (what gender they identified as), race/ethnicity of the respondent, state (the state they lived in), and education (what their highest level of education was). We did not find any variables that were similar enough to the ones chosen. However, as the post-stratification data was a different dataset to that of the survey data, we had to transform the post-stratification data to match that, including general cleaning in the sense of converting non-numeric values to numeric, matching spelling/capitalization of certain responses, and constructing an age_groups variable from the ages given. For more information, please see 01-data_cleaning-post-strat.R in the scripts folder of the github repository. TODO

Due to the nature of our methodology, we are able to directly compare and contrast demographic distributions for the individual survey data and the post-stratification data. The gender distribution of the post-stratification data (Figure 7) is fairly similar to that of the survey distribution (50-50 split. The ACS distribution of age is closer to what we would expect from "18 to 28" year-olds (Figure 6), indicating that the Nationscape survey results are skewed towards an older demographic (Figure 1). Again, there is certain discrepancy between expected race distributions (Figure 8) and the estimates reported by the Census, which we may take into account when making predictions on the post-stratification dataset as a proxy for the popular vote.

There is a major difference in education distribution across surveys as shown in Figure 9 where only 53% of respondents had some level of post secondary education compared to 71% in Figure 3. This better matches the data reported in the US, indicating that the true level of post secondary education is much lower than represented in the Nationscape survey (Figure 3). Lastly, the state distribution looks proportional to the American population (Figure 10) after conducting the same sanity check as before. The discrepancy between the Nationscape survey data and the post-stratification data with regards to age as well as education indicates how important it is that we are using multi-level regression model with post-stratification to predict the popular vote.

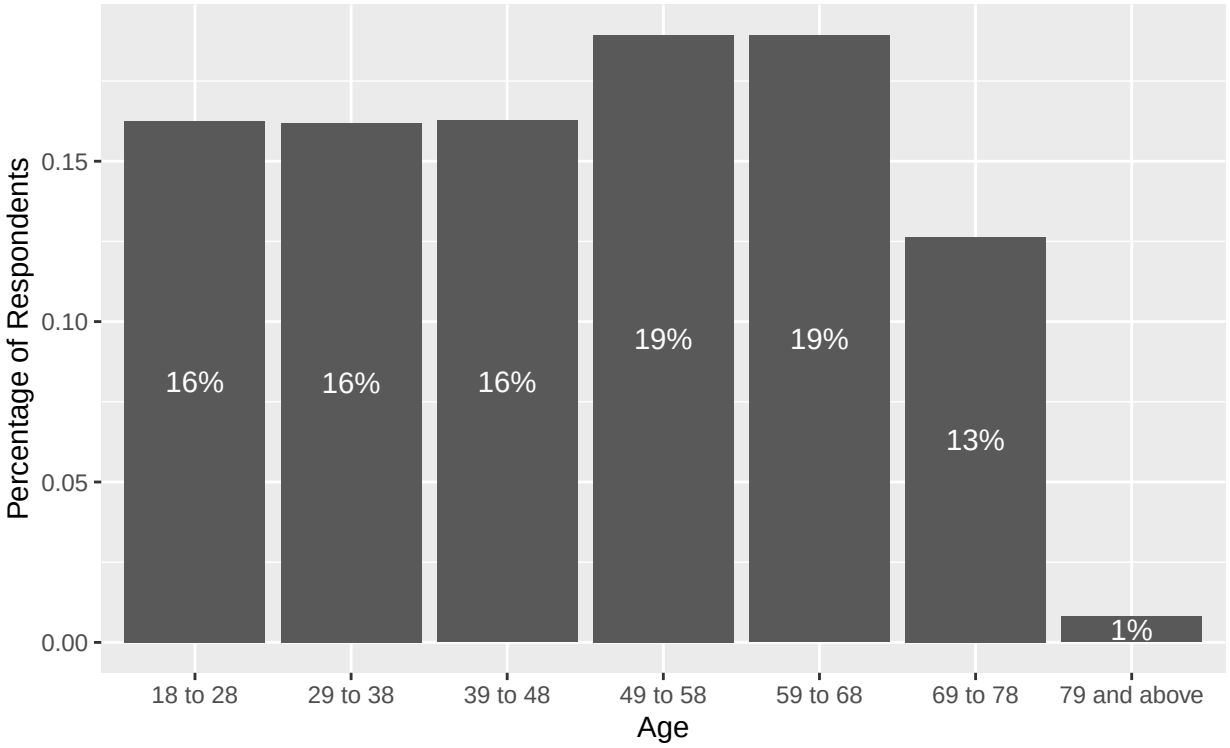


Figure 6: Distribution of the age of ACS respondents in percentages

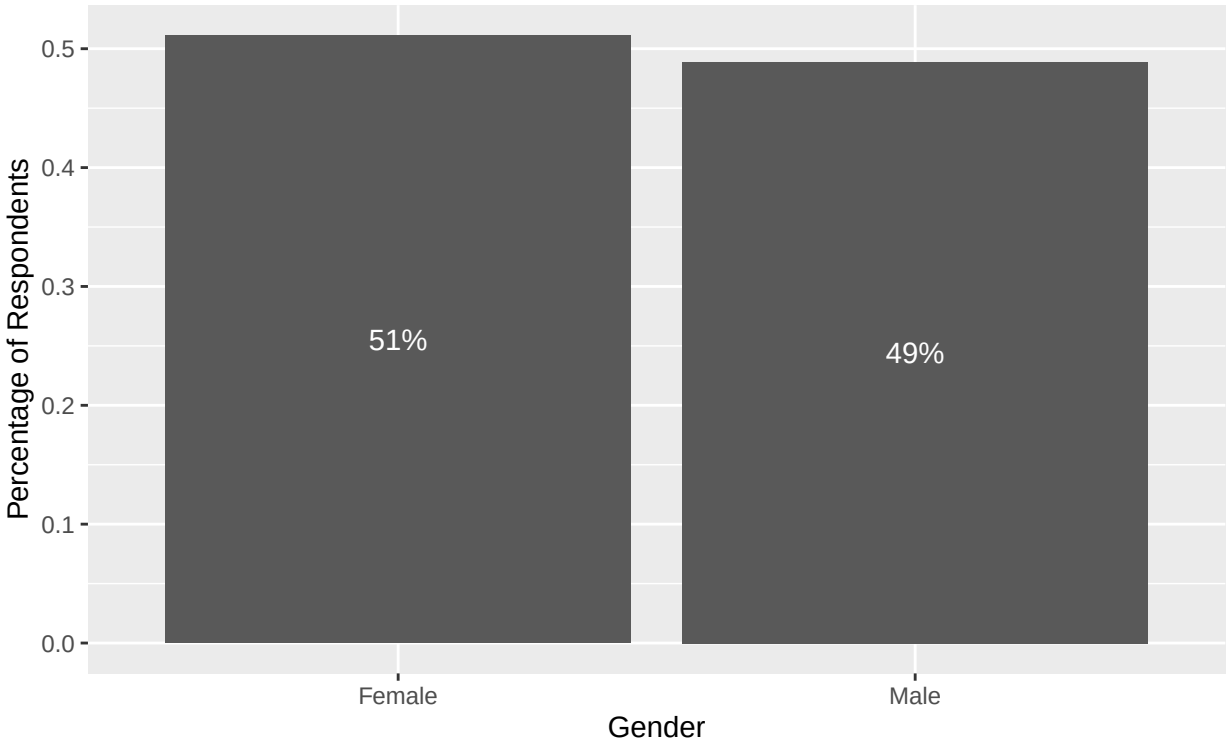


Figure 7: Distribution of the gender of ACS respondents in percentages

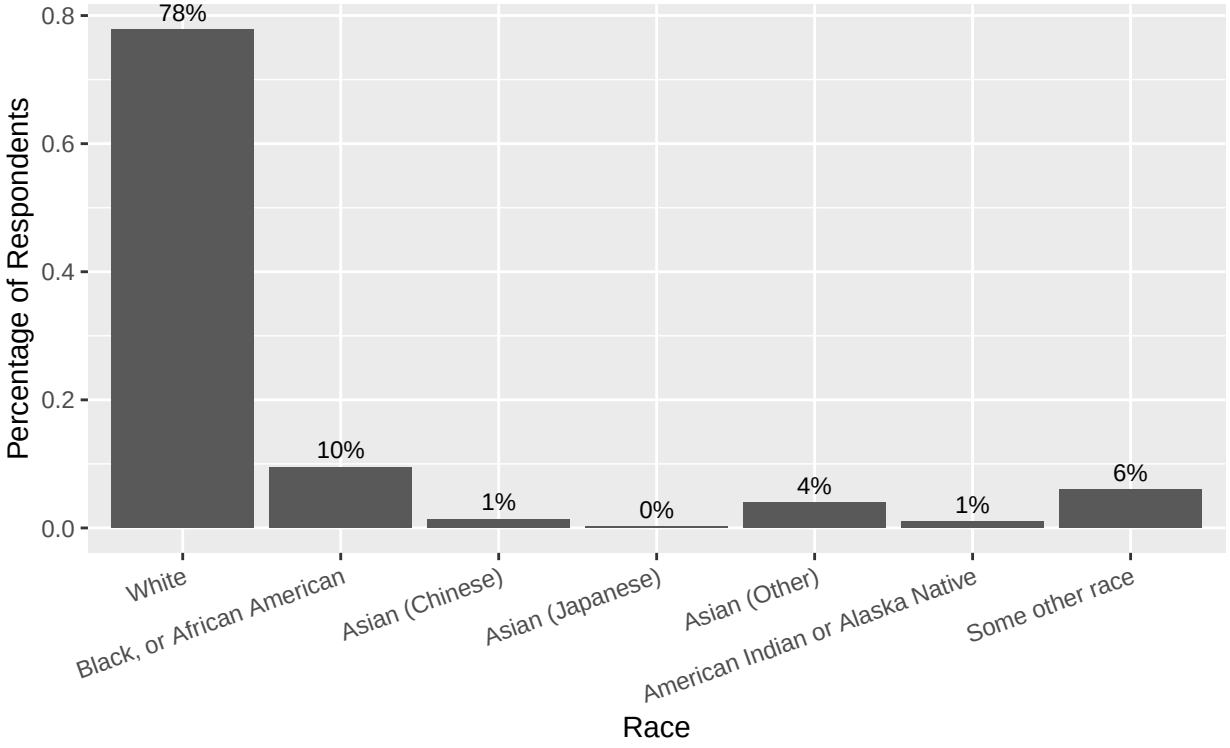


Figure 8: Distribution of the race of ACS respondents in percentages

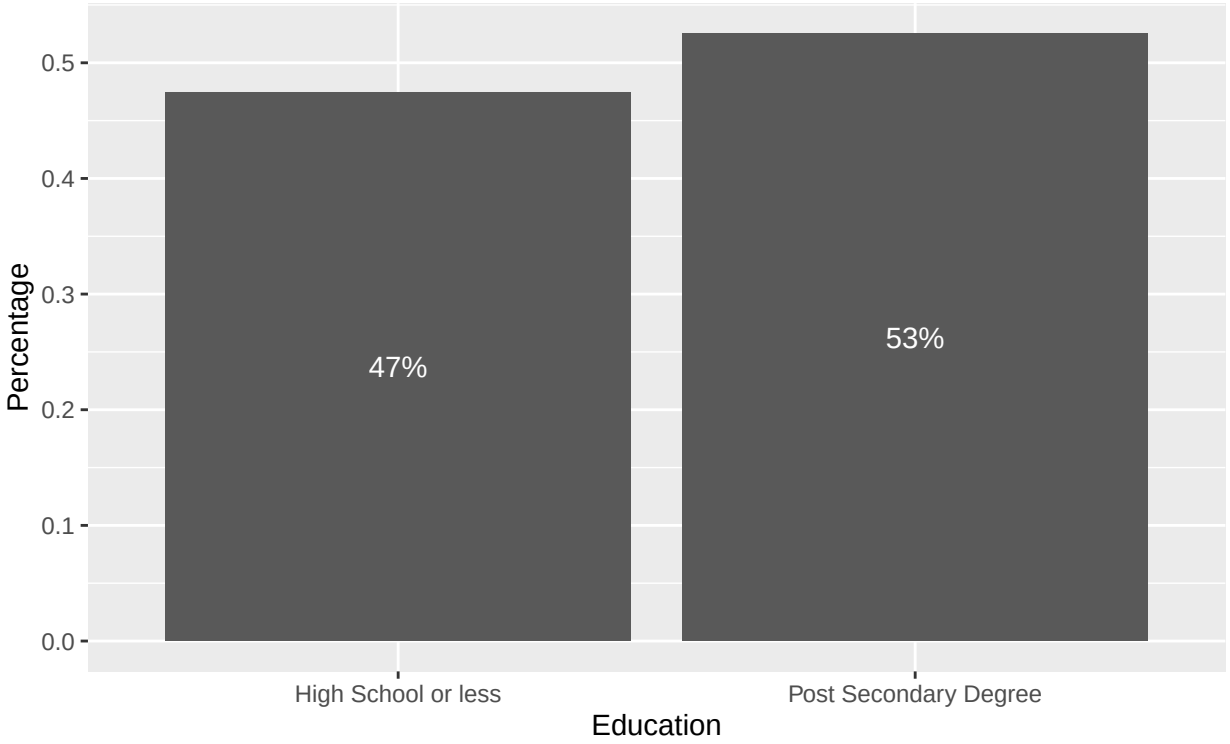


Figure 9: Distribution of the education of ACS respondents in percentages

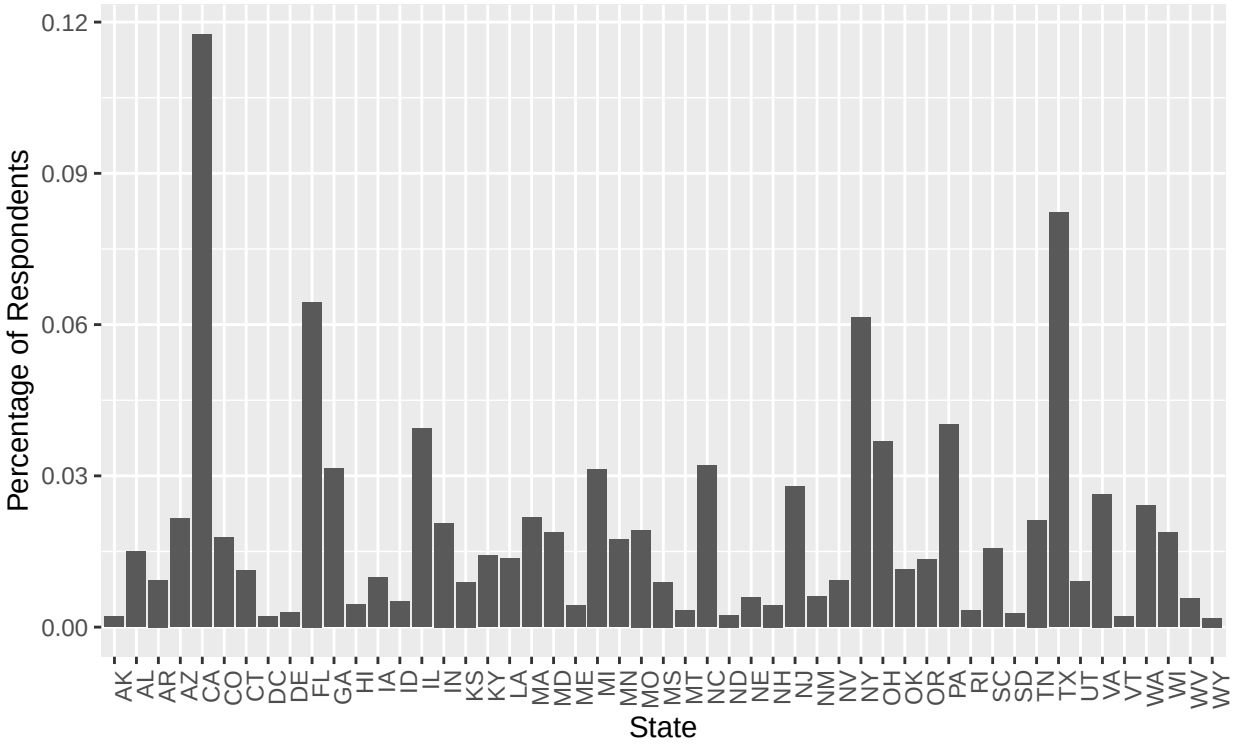


Figure 10: Distribution of the state of ACS respondents in percentages

3 Model

The purpose of our model is to predict a person's vote in the 2020 US election given their demographic characteristics. The dependent variable is a binary categorical variable, where 0 represents voting for Trump and 1 represents voting for Biden. The independent variables are the characteristics of the person (age, gender, race, state, and education). To avoid multicollinearity (when there is a relationship between the independent variables) which causes unreliable regression estimates, a few carefully selected independent variables were used for our model. One criteria for the selection was that the variables had to be both present in the Nationscapes dataset and the ACS dataset, so that they be included in post-stratification. Another criteria is that the characteristic should group people that share a similar perspective which may impact on their voting decision. Through literature research (Uppal and LaRochelle-Cote, 2015), we found that some helpful indicators were age, gender, race, state, education, and self-reported income. To finalize the selection, characteristics should not be highly correlated and the data provided should make logical sense. Since the data quality for income was generally poor and it worsened the model performance, income was ultimately excluded from the model. Therefore, the predictors used for the model are age, gender, race, state, and education.

We found when exploring and graphing the data in the section above that the survey data has underrepresented or overrepresented certain characteristics compared to the actual distributions in the population (namely younger individuals were underrepresented and individuals with higher education were overrepresented in the Nationscape survey data). Therefore, if we were to predict the outcome of the popular vote using the demographic proportions found in the Nationscape survey, we would find that the prediction is biased because the survey sample is not representative of the American population. Hence, we use multi-level regression with post-stratification (MRP) to adjust the influence of each subgroup of the respondents to get a better match on the actual population distribution. To do this, a multi-level model is required as well as another resource (ACS dataset in our case) for post-stratification.

3.1 Multi-level logistic regression model

Since we have represented predicting the popular vote as a binary classification problem using multiple explanatory variables, binary logistic regression is a suitable choice for the model. We use logistic regression to predict the probability of a person voting for Biden and determine their vote by rounding to the nearest represented binary categorical variable. Logistic regression takes independent variables (in our case age, gender, race, state, and education, state, and race) as inputs. Based on the assigned weights and a logit function, the output will be a probability [0,1] of voting for Biden/Trump. Equation (1) is the equation of the logistic regression model and this defines the multi-level regression part of MRP:

Equation 1: Logistic regression model

$$Pr(Y_i \in \{Trump, Biden\}) = logit^{-1}(\alpha^{\text{age}}_{a[i]} + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{edu}}_{e[i]} + \alpha^{\text{state}}_{s[i]} + \alpha^{\text{race}}_{r[i]}) \quad (1)$$

where Y_i represents the probability a respondent is likely to vote for Trump or Biden given various demographic information and the α's are age, gender, education, state, and race and the notation $\alpha_{a[i]}$ refers to the age group the i-th individual belongs to, $\alpha_{g[i]}$ refers to the gender group the i-th individual belongs to, and so forth.

The Nationscape dataset contains voter characteristics and their expected vote for the 2020 US Presidential election, so the model was trained using cross-validation on that dataset. 95% of the dataset was used as the training set to determine the best weights for the model and the remaining 5% of the dataset as the test set to verify the accuracy of the model. The vast majority of the data is used for the training set because we wanted to ensure that the data does not overfit due to a small sample size.

Furthermore, there were many adjustments that had to be done with the inputs to improve model accuracy. Apart from self-reported income, education was another independent variable that we debated whether it should be included in the model. Education severely impacted our cross-validation accuracy due to a wide range of values which were too specific for our purpose. These issues were resolved by summarizing all education levels into two levels: "High School or Less" and "Post Secondary or More". We were able to confirm that there was a significant difference in the voting preference between the two groups, consequently helping our model perform better.

Another variable that initially caused inaccurate results was age. There was no clear trend in having age as a numeric continuous value for predicting votes. As a result, the numerical age values were grouped into bins of size 10 (eg. 18-28, 29-38) in order to better represent different age groups. Another grouping we tried was splitting the data into youth, middle age, and seniors (18-35, 36-55, 55+); however, that split was not able to capture any significant voting pattern in the age groups because we found that the model accuracy actually worsened. So grouping age by bins of size 10 gave us the best results and is the configuration that was included in the model.

After finalizing the inputs that went into our model, we implemented the logistic regression model using the BRMS library written in the programming language R. We ran our script using the software RStudio. We did not run into any diagnostic issues when running our model. As briefly mentioned, we conducted cross-validation to check our model. The results are fittingly covered in the Results section.

3.2 Post-stratification

As we have previously established, we need to use post-stratification to correct our model estimate of the popular vote because of the discrepancy between the demographic distribution of and the demographics of the American population. We conduct post-stratification by calculating a weighted average of estimates for all combinations of explanatory variables. We found the weights to do so from the ACS dataset. However, we cannot immediately use the ACS distributions as the dataset may not represent American demographics accurately. For example, as we found in our data exploration, there is a discrepancy between the minority distribution found in the ACS dataset and the data reported by the US Census. If minorities tend to

democratically and for Biden, we would have a bias against Biden because they are underrepresented in the ACS dataset.

When we perform post-stratification on the ACS data, we find all combinations of our variables and find the weight of each combination representation within the US. Using the PERWT variable provided by the ACS, we can calculate how much of the population each combination represents within the United States. As per the IMPUMS webpage, "PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample", meaning we can find all the individuals with the same values for the combination of variables we are measuring and add their PERWT values together to estimate how much these specific values would weigh.

Using MRP provides many benefits. As mentioned, we can more accurately estimate the weight of our sample predictions in relation to the population without being heavily affected by any sampling bias. It also allows us to use a small sample for training, and apply the model to a much larger sample that better represents the population. In our example, we can use a small scale election survey from Nationscape and train a model which allows us to predict election results for the respondents in the American Community Survey consisting of over 3 million data. This is crucial, as collecting surveys with a useful sample size with regards to an upcoming election can be extremely expensive and time consuming, as these poll results may be time sensitive. The option to collect from a small sample size allows statisticians to save money and time while providing significant results about the larger population by applying their surveys to general census information. Another noted benefit is computational efficiency. Instead of individually predicting over 3 million people, we summarize all the comparisons and only predict on over 8000 individuals and multiply their proportions, which allows us to compute our predictions significantly faster.

However, there are also cons to using MRP. For instance, we are limited to only use variables that will be found in general census data. If we are interested in looking at religion but the post-stratification dataset does not have that field, we are unable to include religion in our model. This can be challenging as information such as an individual's vote in 2016 may be extremely important as a variable, but since the ACS dataset does not contain such information, we are use it to our advantage. Furthermore, if a variable is broken down to a different granularity, we must group by the common group. For example, if dataset 1 has age grouped in bins of size 3 and dataset 2 has age grouped in bins of 5 (eg. 5-8,9-12 vs 5-10,15-20) then we must map them into bins that are multiples of 15. We cannot use groups of 3, because we cannot accurately break down groups of 5 into groups of 3. Therefore we lose granularity in situations where it may be desired.

In this case, using MRP makes sense as we have voting information provided by Nationscape, however this dataset is not distributionally representative of the target population. We also have access to the ACS which contains a lot of distributional data about the American population but it does not contain any information on who an individual will vote for in the upcoming election. Using MRP allows us to use both datasets in order to predict the outcome of the popular vote of the 2020 US presidential election.

4 Results

Figure 1: Distribution of votes by age

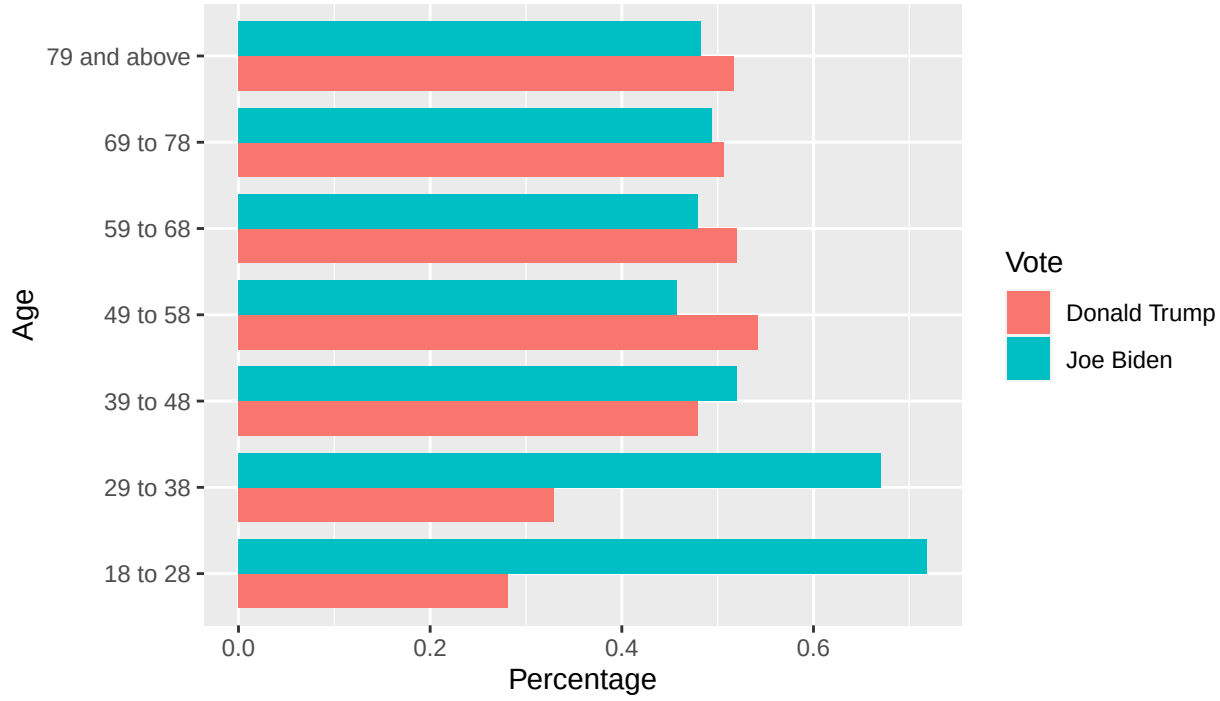


Figure 2: Distribution of votes by gender

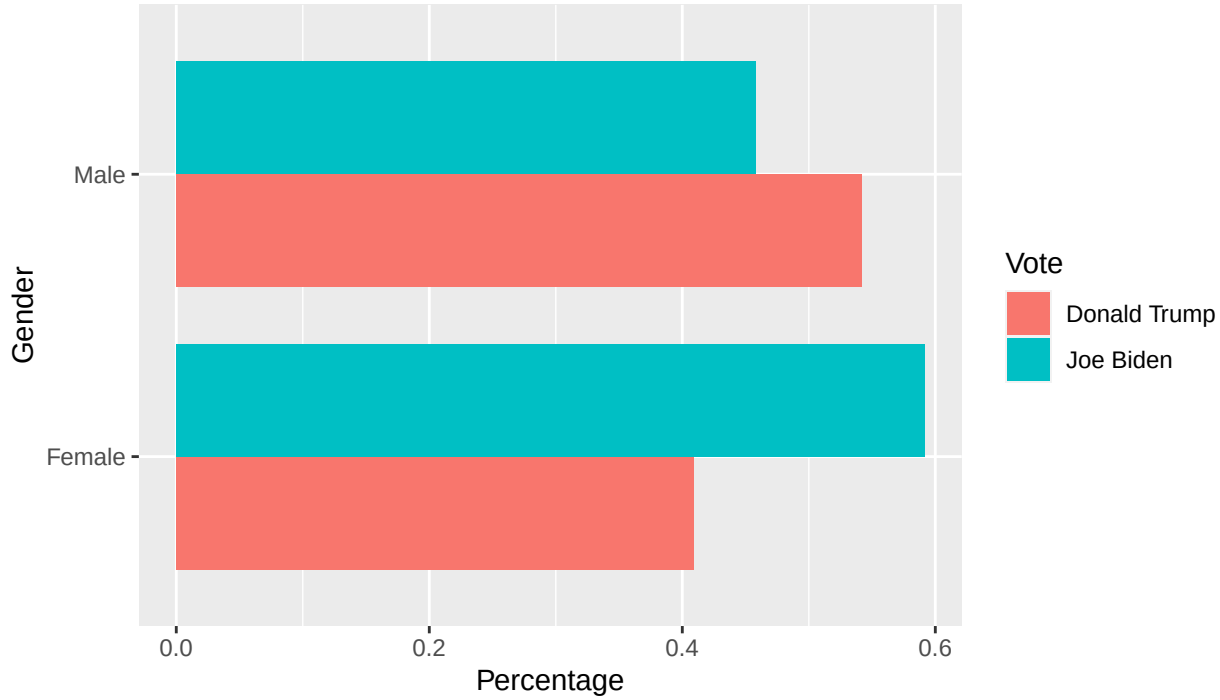


Figure 3: Distribution of votes by education level

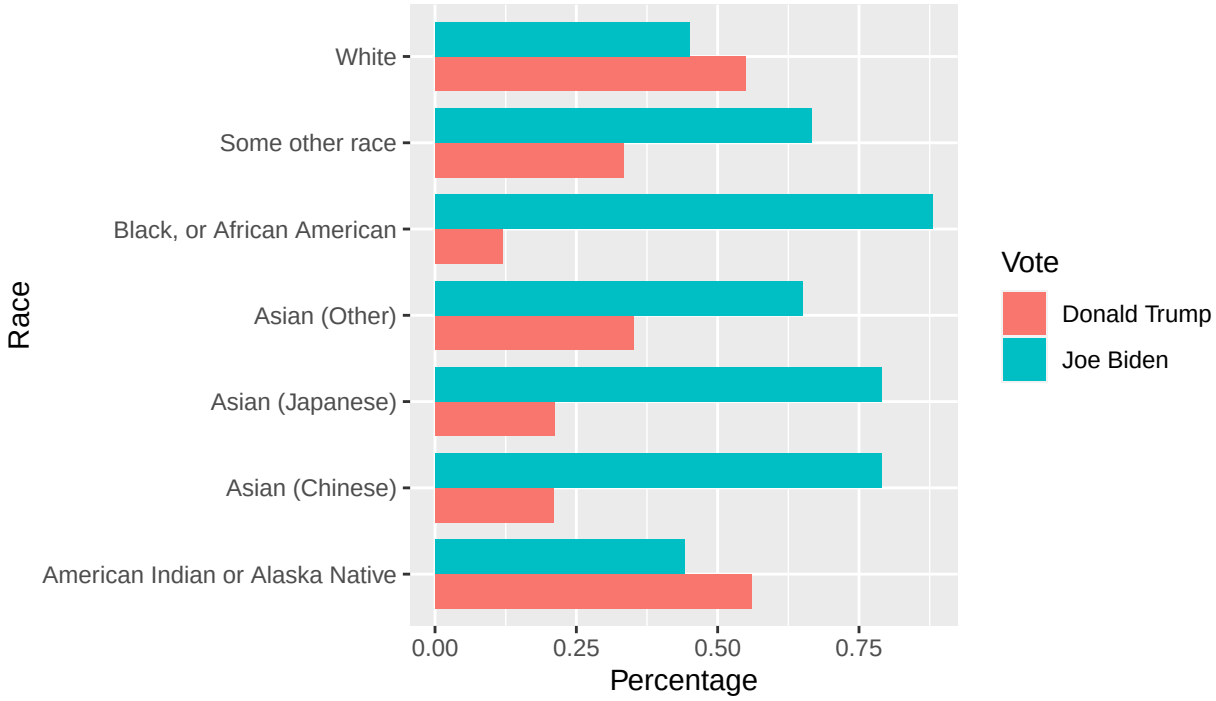
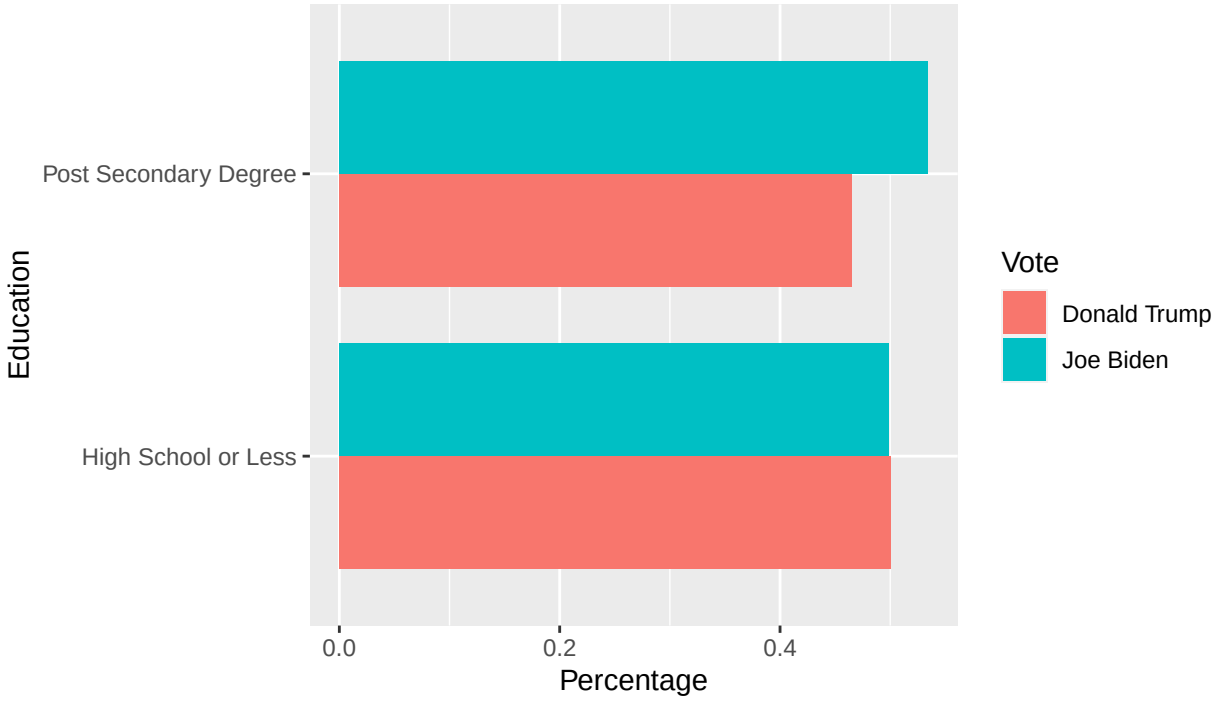


Figure 4: Distribution of votes by education level



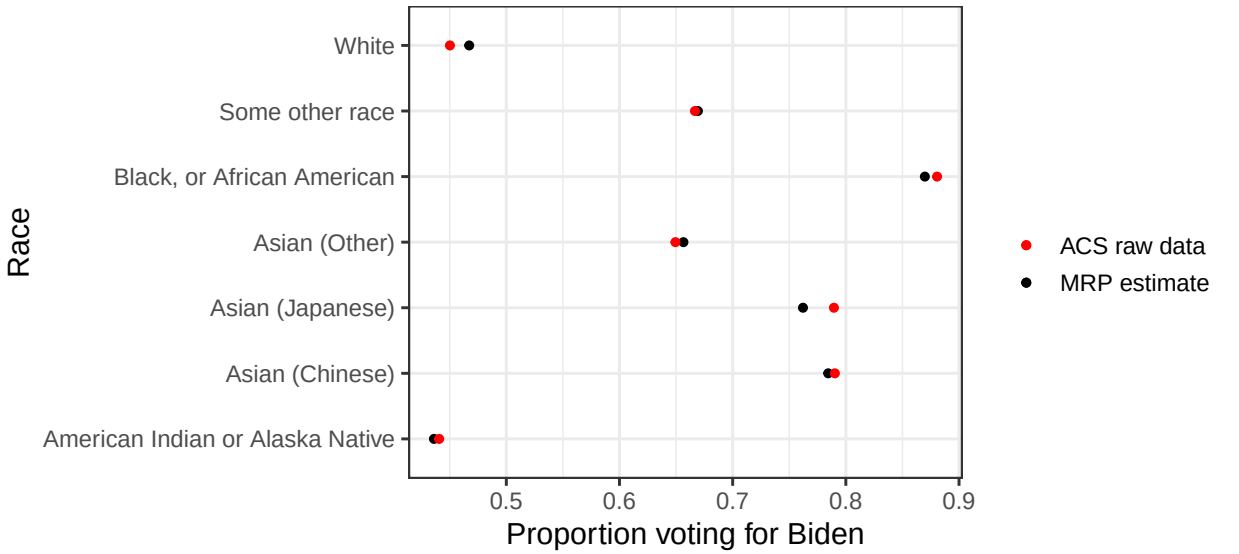
4.1 Post-stratification

```
cell_counts <- readRDS("../../inputs/cell_counts.RDS")
training_data <- readRDS("../../inputs/training_data.RDA")
model <- readRDS("../../model/4chains_3000iter_.rds")

race_res <- readRDS("../../processed_data/race_res.RDS")
gender_res <- readRDS("../../processed_data/gender_res.RDS")
state_res <- readRDS("../../processed_data/state_res.RDS")
age_res <- readRDS("../../processed_data/age_res.RDS")
education_res <- readRDS("../../processed_data/education_res.RDS")
```

The post-stratification data of the 2018 ACS data showed that most of the variables had very little biases since the MRP estimates were very close to the original 2018 ACS data. In Figure N shown below, we see that our raw data proportions was very close to the estimate proportions. Furthermore, this figure forecasts that almost 90% Black, or African American voters will vote for Joe Biden, 75% to 80% of Chinese and Japanese voters are predicted to vote for Biden, while approximately only 45% of white and American Indian or Alaska Native voters are projected to Vote for Joe Biden instead of Donald Trump. Overall, ethnic minorities are projected to vote for Biden over Trump.

Figure N: Proportion of Forecasted Votes for Biden



In Figure N show below, we show the projected voting proportions by each state. We see that most of the states are near the 50% mark, with states such as Arizona and Nevada having only around 30% votes for Biden where as Vermont is predicted to have almost 90% of its voters vote for Biden.

Similarly for age, there showed a general trend that the older population are less likely to vote for Trump compared to Biden, with 18-28 years old population projected to vote for Biden almost 70% of the time and just over 45% of the 19 to 58 age group predicted to vote for Biden

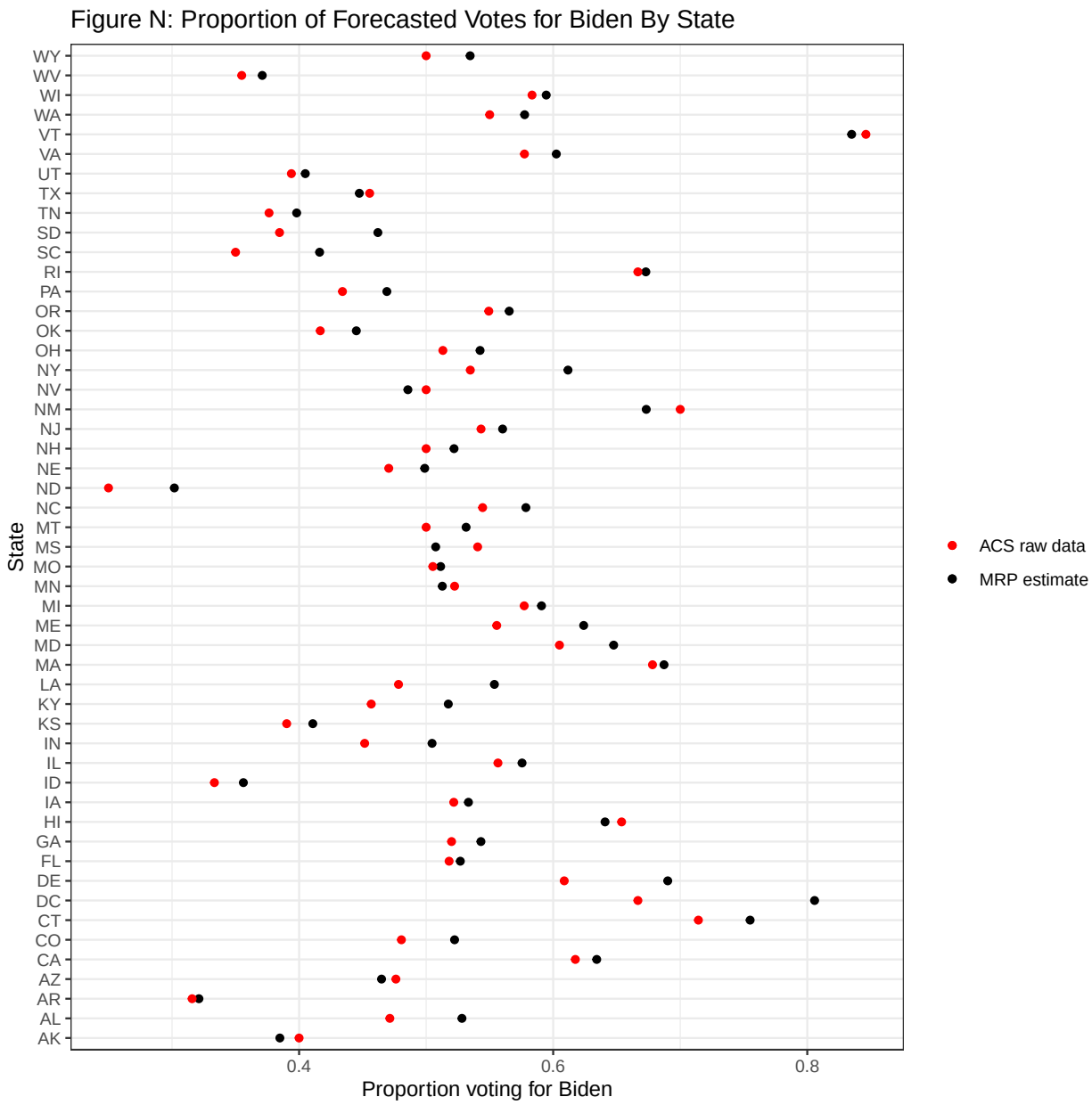
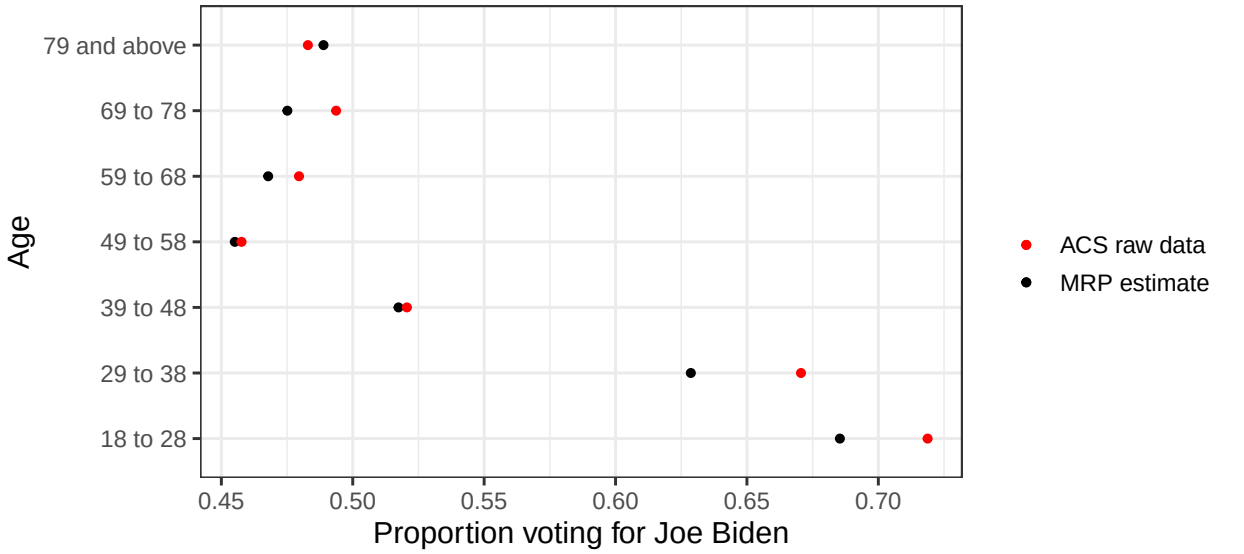


Figure 11: Proportion of Biden Votes by state

Figure N: Proportion of Forecasted Votes for Biden By Age grou

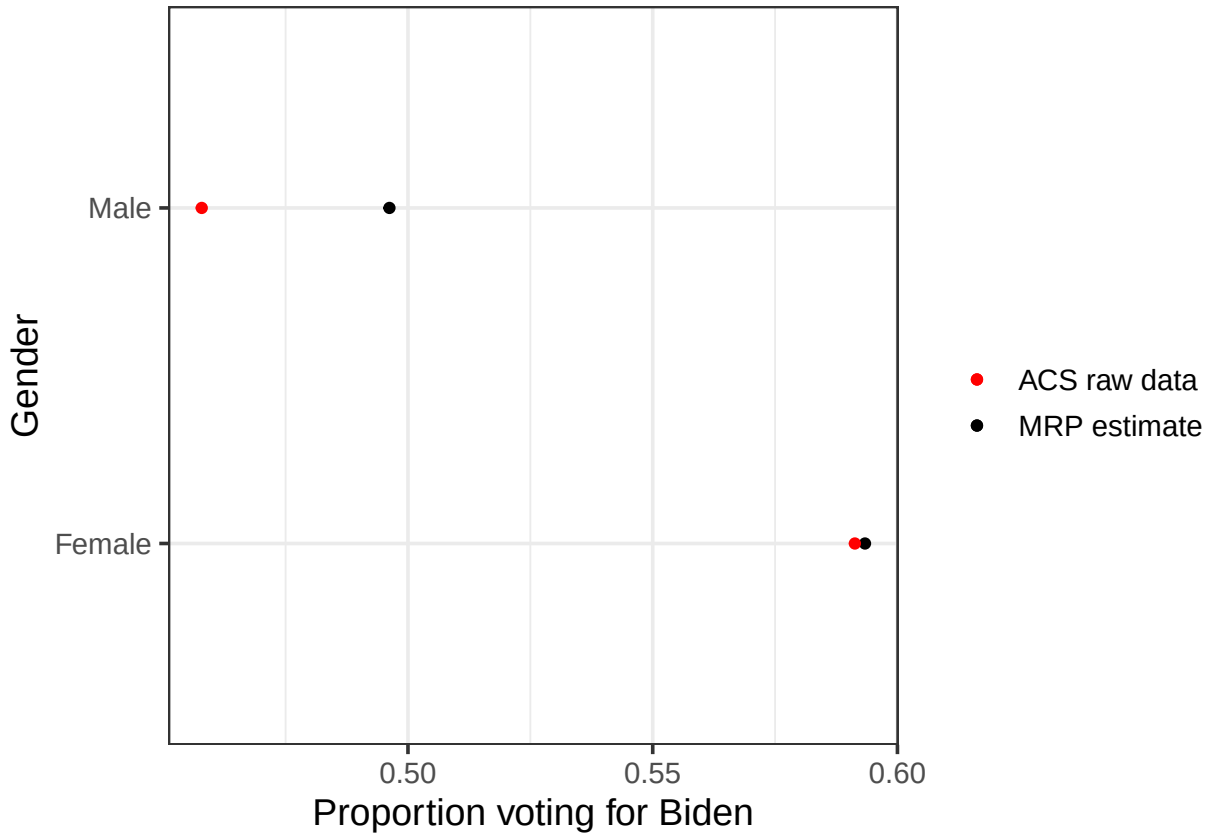


Finally in Figures N and N below, we show that Females are more likely to vote for Biden compared to Males, and a larger portion of individuals with at least some college education are projected to vote for Biden in comparison to High School educated and below.

```
gender_res %>%
  ggplot(aes(y = mean, x = forcats::fct_inorder(gender), color = "MRP estimate")) +
  geom_point() +

  ylab("Proportion voting for Biden") +
  xlab("Gender") +
  geom_point(data = training_data %>%
               group_by(gender, vote_biden) %>%
               dplyr::summarize(n = n()) %>%
               group_by(gender) %>%
               mutate(prop = n/sum(n)) %>%
               filter(vote_biden==1),
             aes(gender, prop, color = "ACS raw data")) +
  scale_color_manual(name = "", values = c("MRP estimate" = "black", "ACS raw data" = "red")) +
  theme_bw(base_size = 14) +
  coord_flip()
```

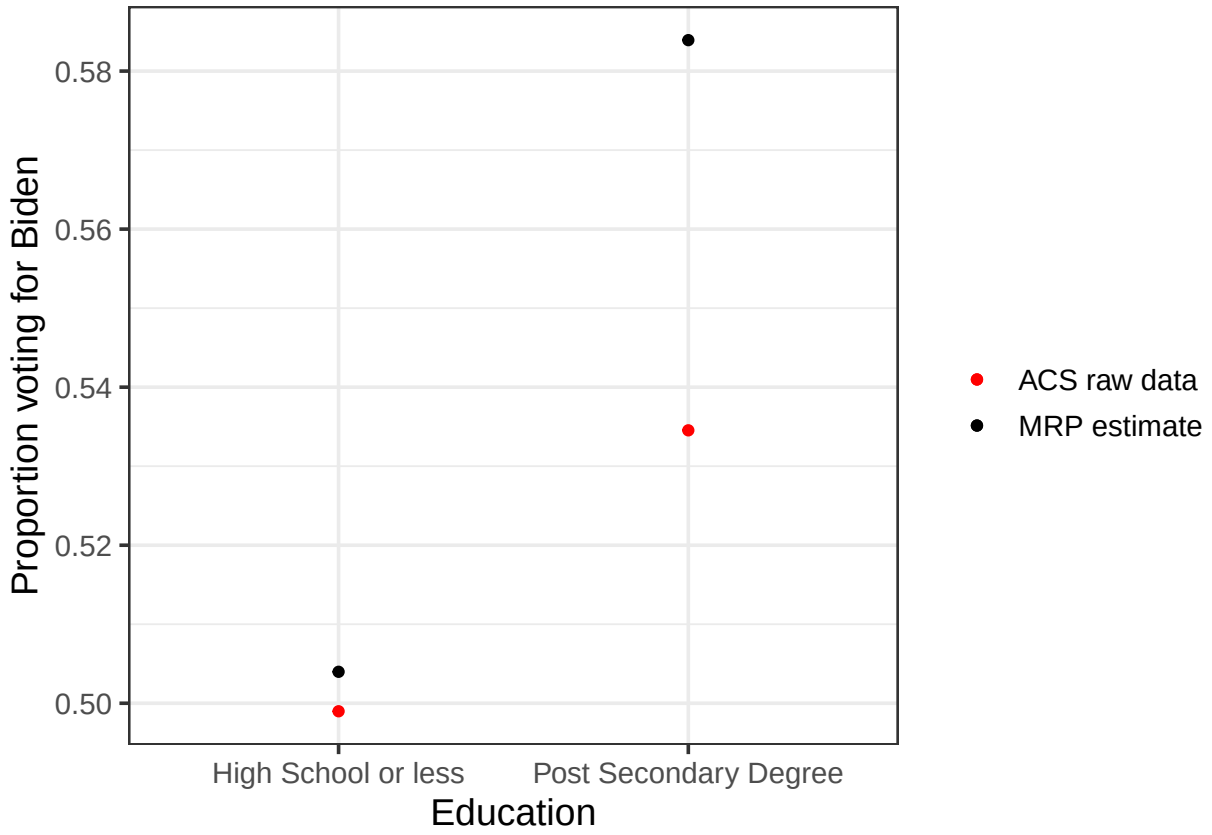
```
## `summarise()` regrouping output by 'gender' (override with `.groups` argument)
```




```
# education_res$age_group <- as.factor(education_res$age_group)

education_res %>%
  ggplot(aes(y = mean, x = forcats::fct_inorder(education), color = "MRP estimate")) +
  geom_point() +
  ylab("Proportion voting for Biden") +
  xlab("Education") +
  geom_point(data = training_data %>%
               group_by(education, vote_biden) %>%
               dplyr::summarize(n = n()) %>%
               group_by(education) %>%
               mutate(prop = n/sum(n)) %>%
               filter(vote_biden==1),
             aes(education, prop, color = "ACS raw data")) +
  scale_color_manual(name = "", values = c("MRP estimate" = "black", "ACS raw data" = "red")) +
  theme_bw(base_size = 14)
```

```
## `summarise()` regrouping output by 'education' (override with `.groups` argument)
```



For the US 2020 election between Biden and Trump, we forecast Biden to win the popular vote by 53.5% to 46.5% of the votes.

Overall, our model performs non-trivially with a cross-validation accuracy of approximately 63% (Figure 12). This is significantly better than a random guess (which would have an accuracy of 50%). Furthermore, by investigating the source of our inaccuracies (Figure 13), we see that of the approximately 21% of our predictions were false negatives (Predicted Trump vote when it should predict Biden) and 15% of our predictions were false positives (predicted Biden when it should have predicted Trump vote). This was significantly better than some of the other models we tried as discussed in the previously, as we had a large number of false negatives in many models when education was split into multiple groups rather than just College or no College.

5 Discussion

Figure 1 features the trend of the younger voters (between the ages of 18 and 38) to vote left-wing. There are many different reasons this can be the case. Younger people generally tend to be left wing (pew 1). In addition, given the current state of the economy, combined with the fact that in general millennials and Gen-Xers are poorer than their parents, many younger voters are in disillusion with the current government, leading to vote against the Republican Party (mckinsey).

Regardless of the reason most young voters support Biden, it's imperative to realize the repercussions of this. Relative to the 2016 election, young voters are coming out (metaphorically) in droves to vote, be it due to their own stronger opinions, voting drives and increased voting awareness online, or any other facet of their lives. In Texas alone, 1.3 million voters under 30 have already cast their ballot, surpassing the TOTAL number of young votes from 2016 (Texas). This can have huge effects nationally, creating a greater likelihood of a Biden victory.

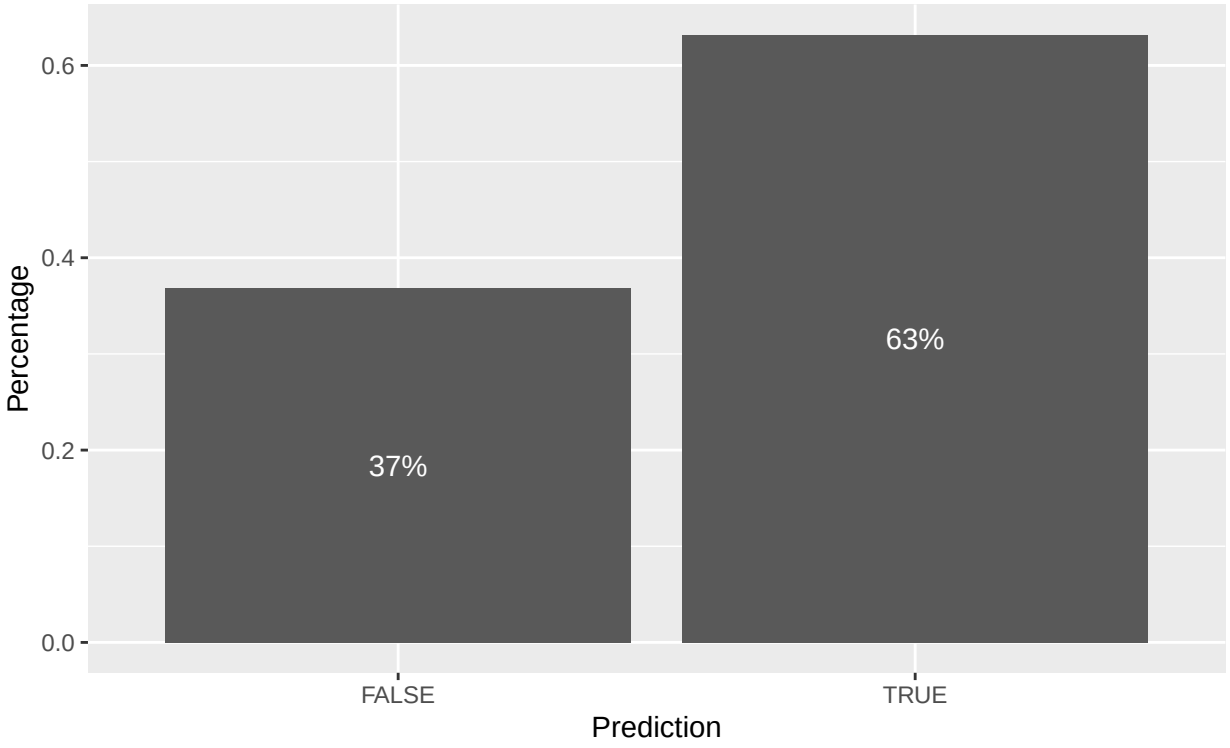


Figure 12: Cross-validation results

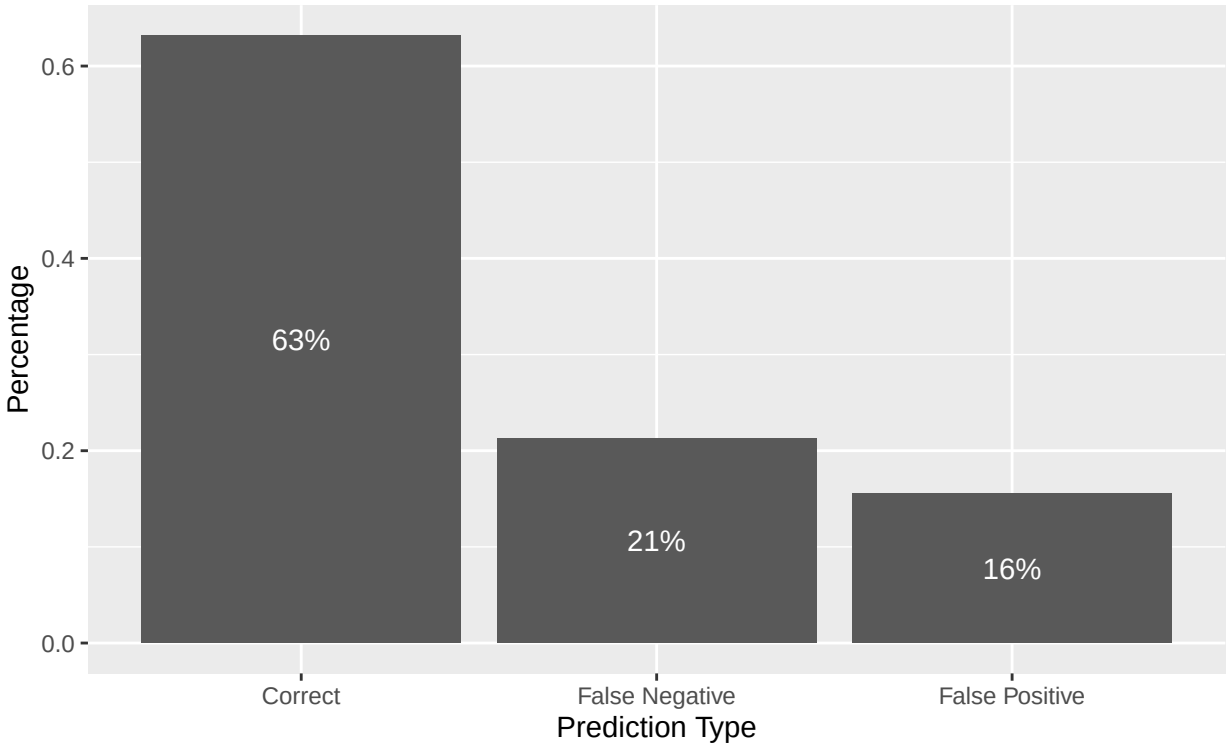


Figure 13: Prediction distribution after cross validation

Contrasting this are the older populations. The majority of voters ages 49 and above support Trump. However, a key point of note is that the difference between Biden support and Trump support is much lower than with younger voters. Given this, there's not as much change in the way of Trump. That said, in typical elections this does indeed make a huge difference regardless, as it's the older voters that vote the most (old). That's why it's so important that many young voters are actually voting compared to before, as this drive to vote acts a form of mitigation against the older voters' support of Trump.

• https://www.pewresearch.org/fact-tank/2014/07/09/the-politics-of-american-generations-how-age-affects-attitudes-and-voting-behavior/ • https://www.texasobserver.org/young-voters-texas-2020/ • https://www.mckinsey.com/featured-insights/employment-and-growth/poorer-than-their-parents-a-new-perspective-on-income-inequality# • https://pubmed.ncbi.nlm.nih.gov/10750310/

Figure 2's comparison of the votes by gender reveals some interesting findings. While the majority of male voters support Trump, the real highlight of this figure is the fact that nearly 60% of female voters support Biden, essentially twice the difference between Biden and Trump for male voters. Again, there's many reasons why this could be the case, but there are two major reasons that present themselves.

The first is the way that Trump has treated women in the past, a notable example being his quote "Grab 'em by the pussy" (tape). The other reason that stands out is the Republican Party's stance against abortion, an important women's right issue. This has only been further exemplified with the recent nomination and appointment of Amy Coney to the supreme court, who maintains a personal belief that abortion is immoral and clerked for Justice Scalia, who was a major critic of Roe v. Wade, the case that set the legal precedent for abortion (bbc).

Overall, while there in majority men vote support Trump, a greater percentage of women support Biden. In conjunction with the metric that women vote more than man, this finding supports a Biden victory. (gender diff)

• https://www.nytimes.com/2016/10/08/us/donald-trump-tape-transcript.html • https://www.bbc.com/news/election-us-2020-54512678 • https://cawp.rutgers.edu/sites/default/files/resources/genderdiff.pdf

5.1 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

6 References

Here's a dumb example of how to use some references: In paper we run our analysis in `R` (R Core Team 2020). We also use the `tidyverse` which was written by Wickham et al. (2019) If we were interested in baseball data then Friendly et al. (2020) could be useful. Tausanovitch and Vavreck (2019) R Survey and ACS dataset https://www.voterstudygroup.org/uploads/reports/Data/Nationscape-User-Guide_2020sep10.pdf https://www.voterstudygroup.org/uploads/reports/Data/NS-Methodology-Representativeness-Assessment.pdf Wu https://www.pewresearch.org/fact-tank/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/

Model Stuff: https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm

ACS stuff: https://www2.census.gov/geo/pdfs/education/Uhl_CAS_2011.pdf https://www.census.gov/programs-surveys/acs/technical-documentation/errata/120.html https://usa.ipums.org/usa/resources/voliii/formACS2018.pdf https://usa.ipums.org/usa/resources/codebooks/ACS_codebook.pdf https://www.politifact.com/factchecks/2014/jan/09/us-census-bureau/americans-must-answer-us-census-bureau-survey-law-/

https://money.usnews.com/money/retirement/aging/articles/why-older-citizens-are-more-likely-to-vote

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, June 10-17, 2020 (version 20200814). Retrieved from [URL].

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2020. *Lahman: Sean "Lahman" Baseball Database*. https://CRAN.R-project.org/package=Lahman.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tausanovitch, Chris, and Lynn Vavreck. 2019. "Democracy Fund + Ucla Nationscape." https://www.voterstudygroup.org/publication/nationscape-data-set.

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.