2020 US Presidential Election Bayesian Model*

Result based on popular vote calculated through multi-level regression with post-stratification.

Guannan Shen

01 November 2020

Abstract

In the US 2020 election, polling the US electorate is a very important method in which experts go about forecasting election results, especially in the age of COVID-19 where traditional practices of large events are no longer common practice for certain candidates. In this paper, we did a multi-level regression with post-stratification in order to determine the probability of Biden winning the popular vote over Trump, found through the use of BRMS that Biden had a 51% probability. This of course is an exceedingly narrow result and rife with weaknesses-for instance, the presidency is not decided by popular vote but rather who wins the electoral college, and the data is based o˙ a survey by the Democracy Fund Voter Study Group in the week of June 25 to July 1st, suggesting perhaps the data is outdated for the election on November 3, 2020. Next steps include accounting for the electoral college, and perhaps using more predictors such as income or timely issues such as views on the pandemic to determine voter intention.

1 Introduction

In this paper, we use statistical software R [R Core Team, 2020] to analyse the US 2020 Presidential Election.

Polling has an exceedingly important role to play in the 2020 US election between the Republican candidate Donald Trump and the Democratic candidate Joe Biden. Given the unique circumstance in which 2020 finds itself—in the midst of a global pandemic in which the US is an epicenter, with nearly 15,000 cases reported daily as of the week of October 18 to 23 [Hipes, 2020]. This means that certain other metrics of voter enthusiasm, such as door-to-door campaigning and large-scale campaign events such as rallies, are no longer as indicative of a candidate's popularity given the Biden campaign is practicing the idea of "remote campaigning" virtually [Forgey, 2020]. However, there is a lot of speculation regarding the accuracy of polls, especially in the wake of the 2016 election in which Democratic candidate Hillary Clinton was projected to win the presidency over then-candidate Donald Trump with the percentage of her win considered from "70 to as high as 99%" [Mercer et al., 2016]. Clearly the level of support for Donald Trump had been underestimated by pollsters and pundits alike, suggesting more care has to be put into the integrity of polling to ensure there is no pre-assumed conclusion that will drive the statistical analysis. Indeed, our paper strived to do just that—it investigates Biden and Trump's chances of winning the popular vote (an important caveat as the US election is determined by the electoral college) through the lens of the electoral population's gender, race, education, age and state distributions.

This paper specifically makes use of data collected by the election polling done by the Democracy Fund Voter Study Group, for the week of June 25th to June 1st in order to create a logistical model which

determines the probability that one would vote for Trump or Biden depending on their age, race, gender, state and education as explanatory variables. This model, created from a survey which does not perfectly align with the American electorate's distribution of age, race, gender, state and education (and thus is a "non-probability sample") is then re-weighted to data from the American Community Surveys published by IPSUM. Through multilevel logistic regression with post-stratification, the response variable being who one would vote for, this paper finds Joe Biden's probability of winning the popular vote to be 51%, indicating that he would narrowly win. Regarding predictors, it was shown that men, people who hadn't graduated high school, and the older population were more likely overall to vote for Trump over Biden, whereas each state had individual localized effects. For instance, Californian voters were clearly much more supportive of Biden than Arizona. In terms of race, Biden had a much more palpable lead in terms of winning African American/Black votes than his opponent. All these various effects from the logistical model culminated in the narrow victory mentioned earlier.

The outline of the paper is as follows. In the Data section of this document, the paper discusses the nature of the survey and stratification data that we collected from the Democracy Fund Voter Study Group and IPSUM respectively. Such details include the strengths, weaknesses and natures of the datasets, the variables the paper makes use of, and the methodology of the data collection. Moreover, it briefly touches on the cleaning of the data that had to be done such that the two datasets could be referenced with relation to each other, such as the mapping of 'sex' in the survey data to 'gender' in the stratification data, but those intricacies are discussed further in the Discussions section. An analysis of the raw data (in terms of the variables age, gender, state, education and race) will be found here as well. In the Model section, the paper defines and explains the statistical methodology regarding how the logistic regression was structured, as well a justification for its appropriateness in the current context. In the Results section there will be a summary and graphical representations of our logistic regression, and these results will be further discussed in the Discussions section regarding what the conclusion of Biden's narrow victory actually entails. This study of course has its fair share of weaknesses and thus next steps—it predicts the winner of the popular vote, not necessarily the electoral college. As it is completely possible for one to win the first but not the second, a next step would be to consider applying this stratification to a theoretical electoral college vote. Moreover, this survey data was conducted in June and the context of the election in particular has been the rapidly worse situation that the US has been in with the pandemic in recent weeks [Hipes, 2020]. Thus, perhaps a stronger analysis would consider the effect COVID-19 has on Trump's election efforts given it appears to be a prominent election issue.

2 Data

2.1 Survey Data

The source of the survey data is the Democracy Fund Voter Study Group, which conducts weekly interviews of American voters from July of last year to December of this year [Dem, 2020]. The most recent data was from the week of July 25. This data originally had 6,479 observations from voters, but in order to construct a binary response for our response variable of the 2020 presidential vote, this paper omitted those who were not voting, were voting third party, or were unsure. As a result, there remained 5,200 observations in which we could determine whether our explanatory variables could determine one's inclination to vote for one of the two major candidates.

The population of this dataset is the electorate of the United States, but the frame in which this study is done are the online respondents from a market research platform. The sample consists of the 6,479 observations. These online respondents complete an "attention check" before doing the survey, which decreases the proclivity towards non-response as the respondents must prove they are paying attention, and then the survey data is weighted by the 2017 American Community Survey that is done by the United States Census Bureau [Dem, 2020]. The respondents are picked specifically based on their characteristics, so the sampling isn't random but rather purposive, in order to meet a target that is vaguely representative of the United States population. This does mean that the error is arguably, according to Dem [2020], greater than if

the respondents were truly randomly picked. Regarding weighting, these "weights" artificially create more data through sampling within underrepresented subgroups, as well as through interaction between different variables. For the purpose of this analysis, the interactions of note are: gender by race, education by gender and race by education.

As this dataset was processed through the software provided by R Core Team [2020], all instances of non-response for any of the explanatory variables or response variables were omitted. A weakness of this survey for the sake of our study and in general is that it is done so far in advance of the election that given the dynamic nature of US politics, one must wonder if these data would be truly helpful in being predictive of voting intention in 2020. Another is that in terms of "gender", it appears there are only two options available, and for those outside the gender binary it means they are not properly represented within the survey. The weighting, if there is a non-representative pattern inside a small sub-group, may enhance the prominence of that pattern in an inappropriate way. That being said, a strength is that the survey itself removes 8% of responses because of too speedy responses or just picking the same option (ex. the first listed option) for all the questions, meaning the data is more clean and useful. Moreover, the size of the survey is large enough that there is specific data on specific subgroups of the population.

Regarding variables, we picked state, age, education, race and gender as predictors for one's vote in 2020. Gender, race and age according to the figures below play a notable role in voting patterns, which falls in line with the pre-existing literature on the topic.

As you can see from Figure 1, from the raw data it appears that women want to vote for Joe Biden in greater numbers in a larger margin than Trump's increase over votes from men. In Figure 2, according to percentages, 59% of women wish to vote for Joe Biden out of decided voters in our population, giving Biden an 18% increase amongst women versus Trump's 10% increase in support amongst men. This validates the premise that one's gender does correlate with their vote and deserves to be considered in our model. We must note, however, that this data does not include information regarding non-binary voters, which is mitigated somewhat by the low percentage of non-binary individuals in the American electorate.

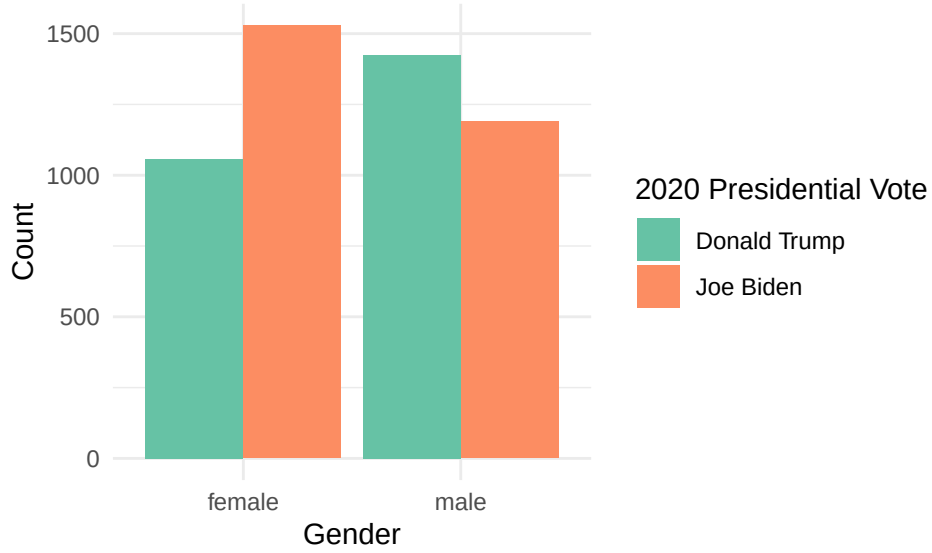


Figure 1: 2020 Presidential Votes by Gender

Regarding race, according to Figure 3 we can see that a large percentage of the respondents in the sample are white and appear to support Trump by a wide margin to Biden. However, in contrast, the Black/African American vote goes overwhelmingly to Biden (shown in Figure 4 with an overwhelming 76% lead), whereas other racial categories are much more contested. It must be noted that in order to format this data for post-stratification with the ACS and ensure the subgroups that our eventual model would be enacted upon were big enough, several groups, such as Hawaiian, had to be conflated to the "other asian or pacific islander"

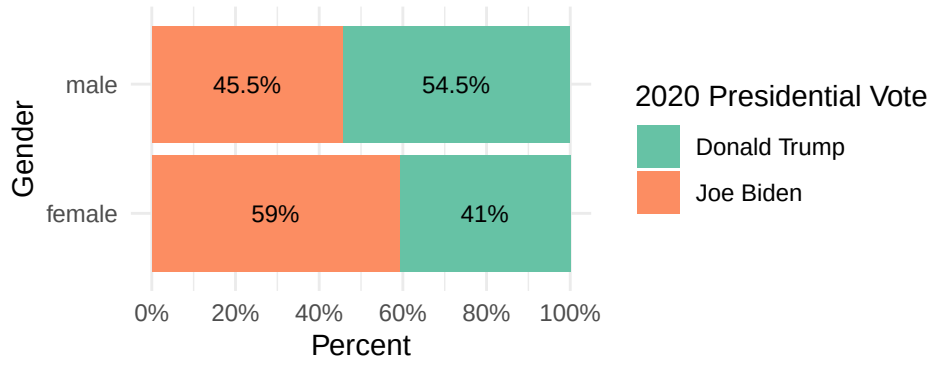


Figure 2: 2020 Presidential Vote Distribution by Gender

category, justified as they were a small percentage of the population and their voting patterns tended to correlate. We nearly used whether a respondent was hispanic within the model, but in the end decided not to because the subcells were getting too small and we believed race in turn would suffice as a predictor regarding background.

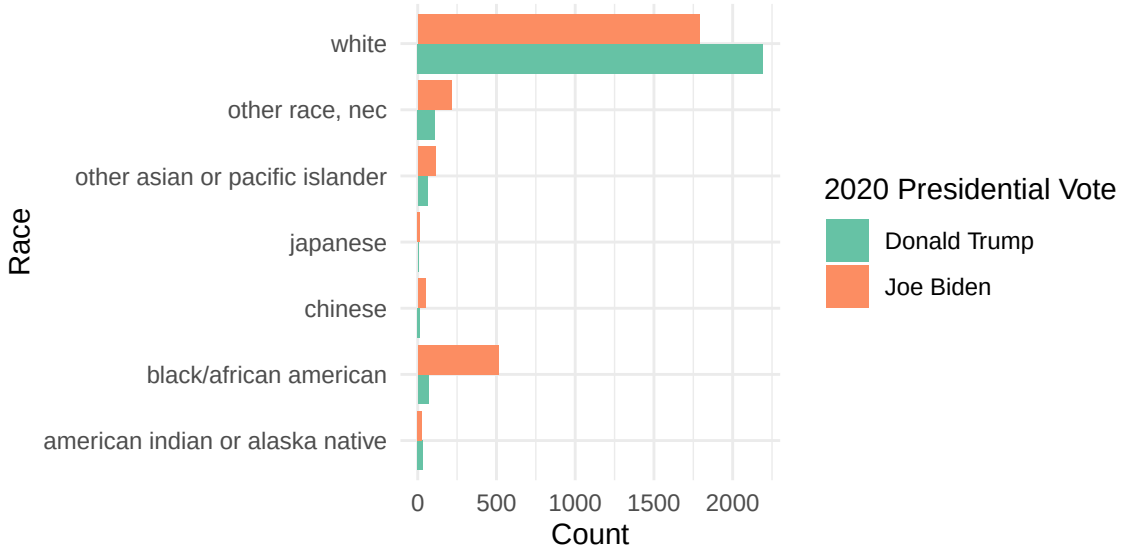


Figure 3: 2020 Presidential Votes by Race

For reasons due to ensuring that our subgroups remained large enough that there were a significant pool of samples for each, the education predictor was conflated to the categories: didn't graduate high school (from "3rd Grade or less", "Middle School - Grades 4 - 8" and "Completed some high school"), graduated high school, didn't graduate college, and graduated college (from "College Degree (such as B.A. or B.S.)", "Master's Degree", "Doctorate Degree", "Associate Degree", "Completed some graduate, but no degree"). This is because the combined subgroups tended to follow similar voting patterns and because given the small sizes of some of the variables in terms of observations, they did not affect the overall data significantly.

Figure 5 seems to demonstrate an overall trend that those who have graduated college tend to vote more for Biden than Trump, Figure 6 showing a 6.2% difference. That being said, the other categories appear highly competitive, so perhaps education is not as strong a predictor as certain other variables mentioned earlier.

According to Figure 7, it appears that Biden has a huge lead amongst young voters, and there is a direct correlation between the age group one belongs to and the likelihood that they'll vote for a candidate. For

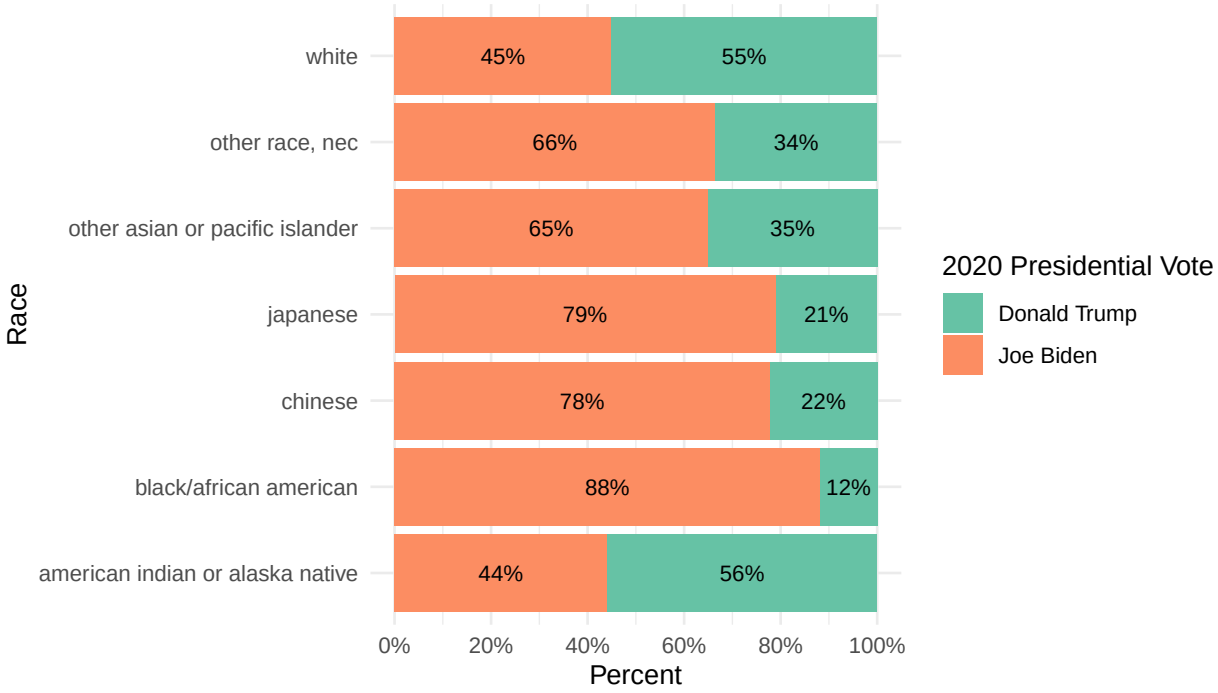


Figure 4: 2020 Presidential Vote Distribution by Race

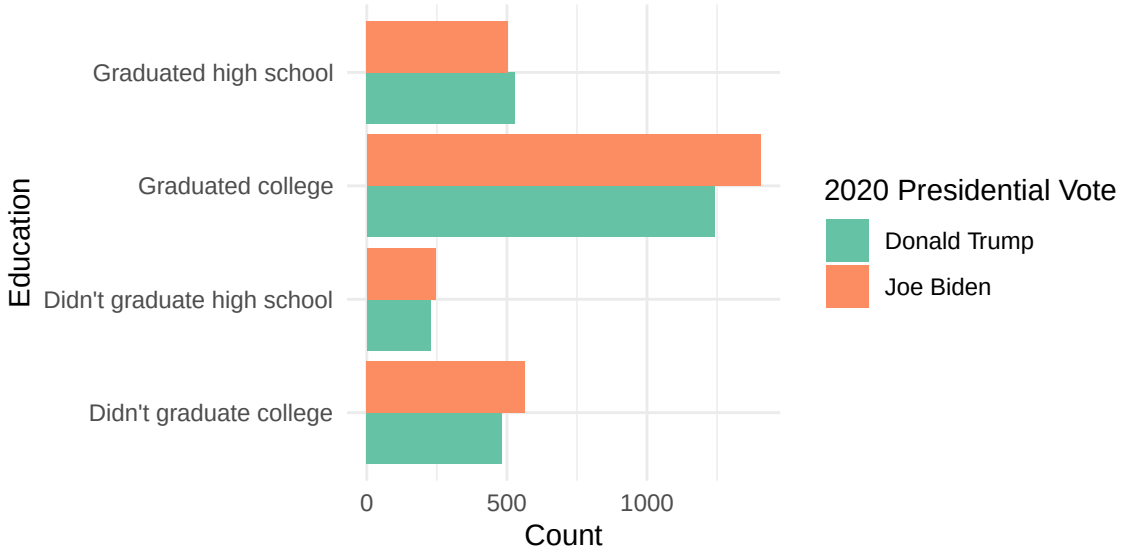


Figure 5: 2020 Presidential Votes by Education

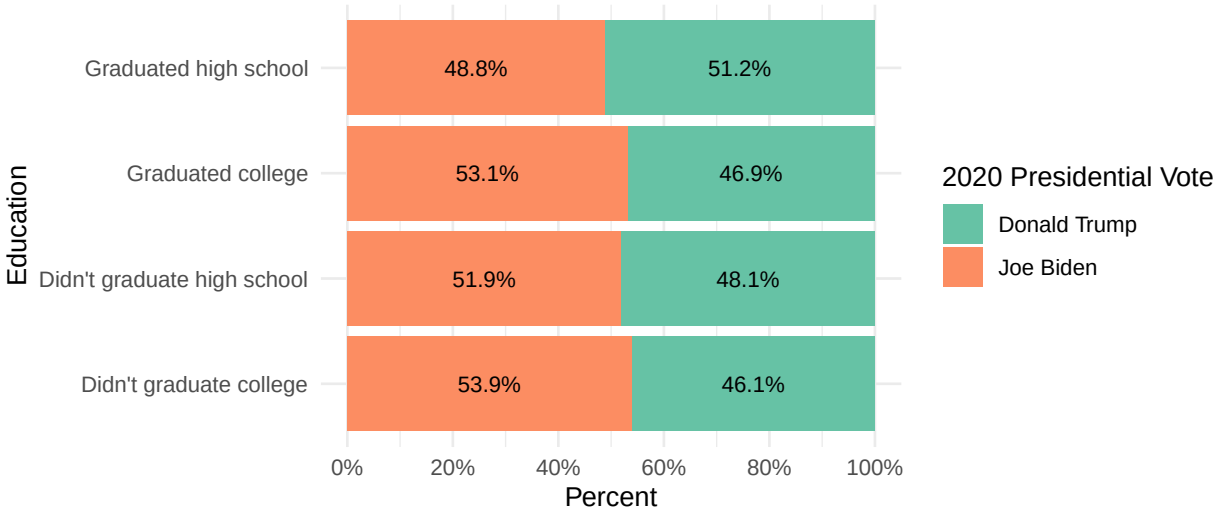


Figure 6: 2020 Presidential Vote Distribution by Education

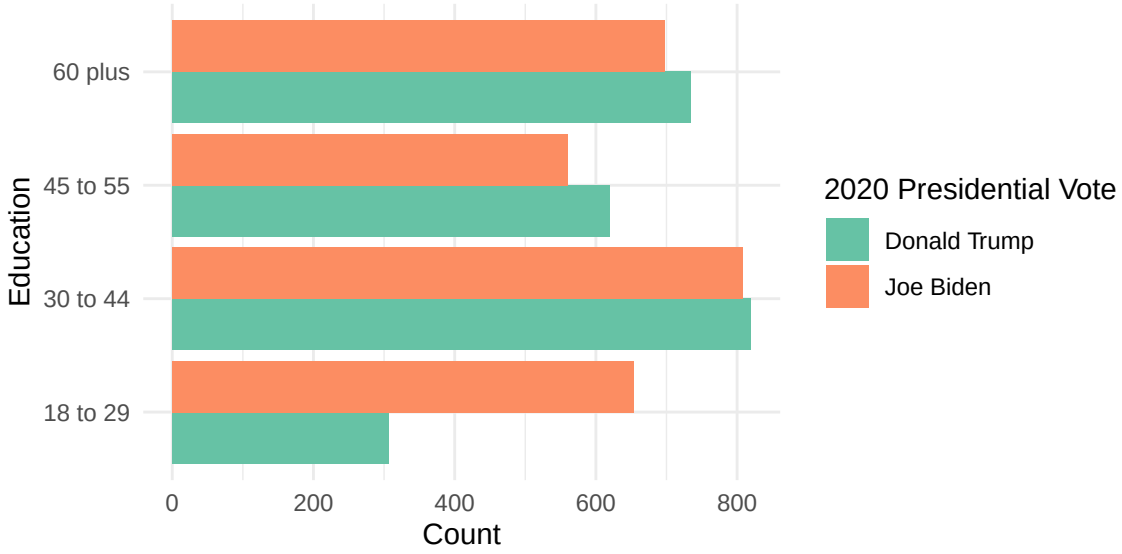


Figure 7: 2020 Presidential Votes by Age

instance, Biden's lead is 36% amongst voters 18 to 29 in Figure 8, very tightly tied with Trump for the age group 30 to 44, and appears to be somewhat behind Trump with older voters. This suggests age is an appropriate variable to gauge whether someone would vote for a particular candidate or not.

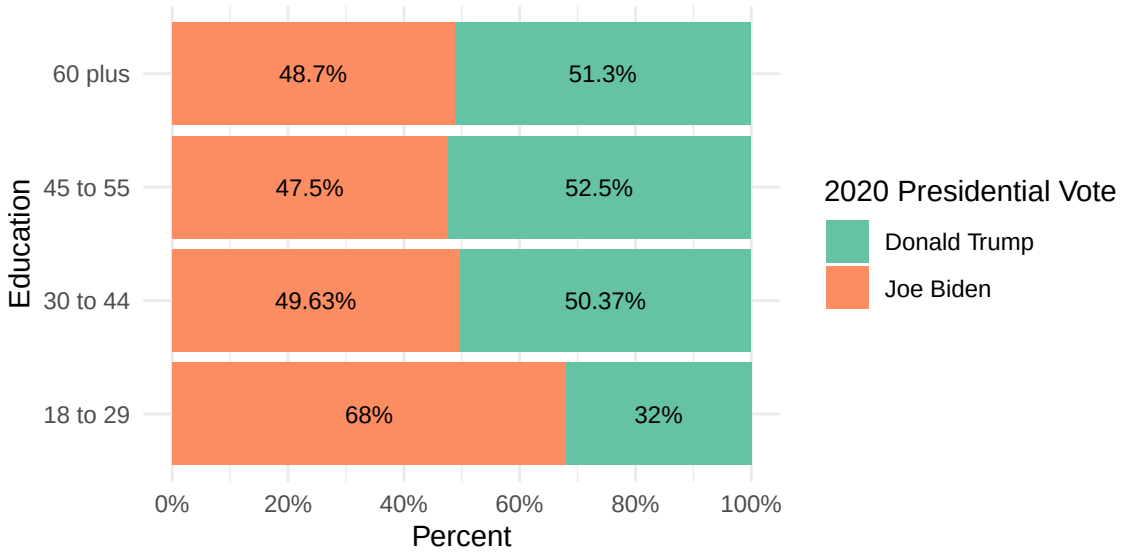


Figure 8: 2020 Presidential Vote Distribution by Age

America is divided into states, and the regional politics and situations within those states appears to play an important role, according to Figures 9 and 10, of which candidate that one would vote for. For instance, some states like California are 62% skewed towards Biden out of decided voters, while others like Arizona are 68% in favour of voting for Trump. Others like Wyoming are 50% split exactly, so this shows that one's state has a hugely important influence/pressure on what candidate one is likely to vote for.

2.2 Post-Stratification Data

This data is offered by IPUMS, based on the American Community Surveys from the U.S. Census Bureau. In this project, we use the data collected from the American Community Survey 2018 [Ruggles et al., 2020].

The target population in this case is every resident in the United States (not necessarily every voter in the United States, something we kept in mind when we engaged in the process of data-cleaning through taking out underage respondents who cannot vote, for instance). The sampling population is all the living quarters in the United States (in which both housing units and group quarters, such as college residence halls, military barracks, faculties for the homeless, correctional facilities, group homes, nursing facilities and more, are considered "living quarters"), and thus the sampling frame is the list of living quarters that were picked out from the sampling population.

There are two phases of determining the samples for housing units: in the first, new samples are selected (in September/October of the previous year), and in the second, non-responding addresses are selected for personal interviewing (in January of the current year). This removes the percentage of non-respondents. These samples are selected from the sampling population, found via the US Census Bureau's "Master Address File", which has a list of all living quarters and certain non-residential buildings [cit, 2014].

For group quarters, certain exclusions are applied for privacy and feasibility reasons (ex. domestic violence shelters, dangerous encampments and more). Field representatives do interviews of qualifying group quarters, in which people are organized into groups of 10 [cit, 2014].

For housing units, data collection is done through Internet, mail, telephone and personal visit. This includes emailing respondents or offering paper questionnaires for ease of accessibility. In the case of non-response,

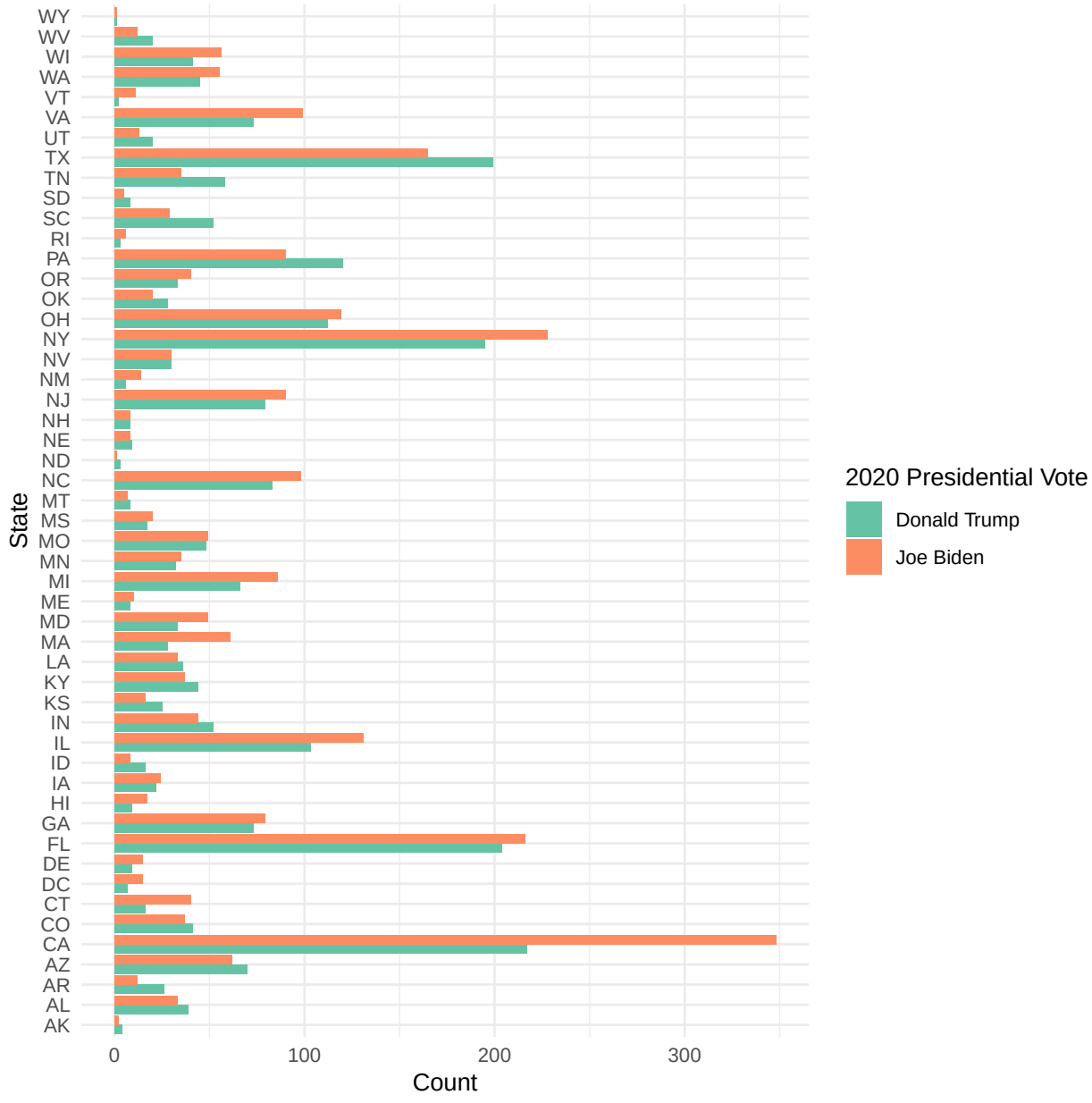


Figure 9: 2020 Presidential Votes by State

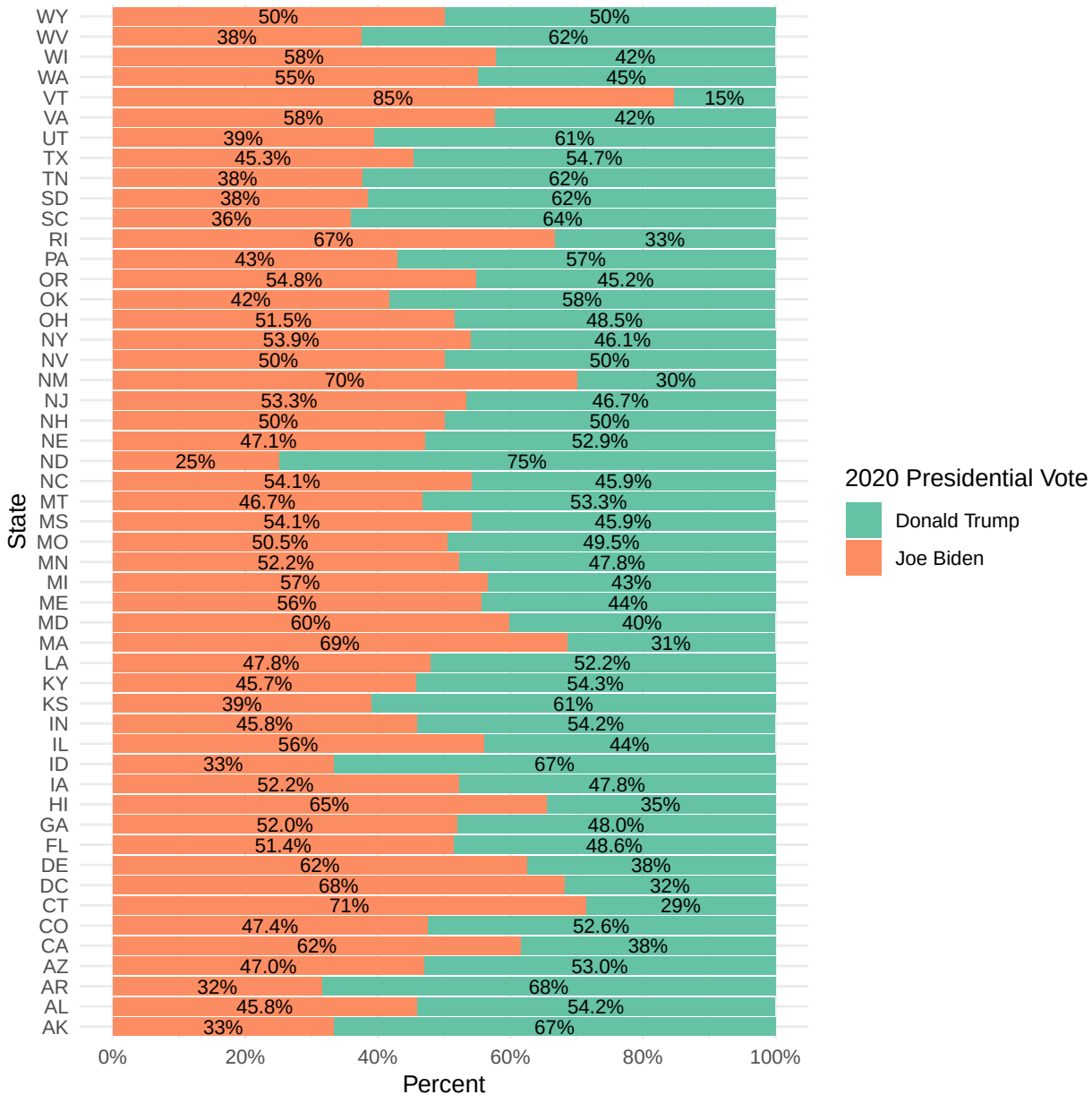


Figure 10: 2020 Presidential Vote Distribution by State

there is a follow-up with a computer-assisted telephone interview. This data collection is made further accessible with language assistance, such as translated resources and interviewers that can speak several languages through telephone contact [cit, 2014].

This methodology has several pros and cons—in terms of pros, a lot of care has been taken to lower the level of non-response. Not only through computer-assisted interviewing in order to follow up with those who do not respond, but this survey also takes into account group quarters for those who may not have access to a residential home. This ensures that the underprivileged are not overlooked in the survey. Because of the care taken to make sure those not proficient in English can access the survey, there will be greater representation of the less dominant culture given America is primarily an English-speaking country. The second stage, which involves contacting people once again who didn't respond the first time, offers even more accuracy. However, this does not guarantee complete response in the least, and this data is accurate to 2018. The population distribution may have changed since then and the data could be outdated to the current demographics. Moreover, this dataset asks for the respondents' sex rather than their gender, providing then ignoring the demographic of non-cisgender and gender non-conforming people in their methodology [cit, 2014].

As mentioned in the previous section, the variables at play are the state, the gender (taken from the sex and assumed to the same), age, race and education level of the respondant. Variables that were created for the sake of post-stratification are the number of cases under each combination of the five variables (so the size of the cell), and the proportion of each cell with regard to the total population [cit, 2014].

Like with the previous dataset for the survey, this one had to combine race into particular categories. In this one, observations such as "Two major races" had to be combined into "Other race, nec" as it was the most accurate category. Education was split into the same categories mentioned earlier, associate degrees and such all folded under college degree for the sake of simplicity.

Below are some distributions of the variables of interest.

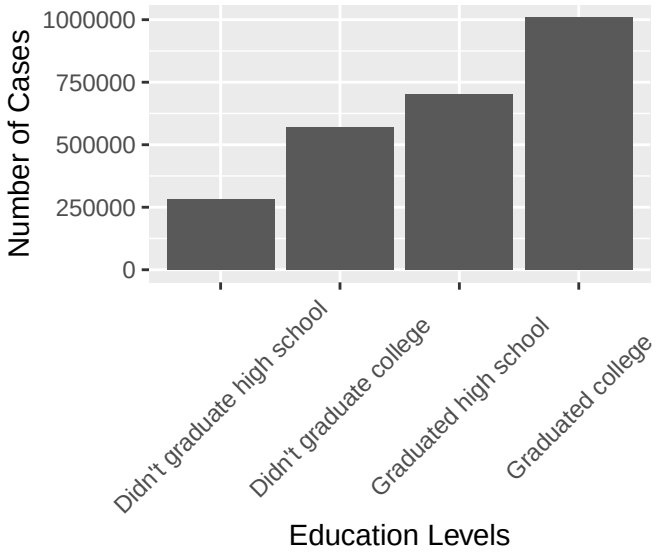


Figure 11: Distribution of Education

As we can identify in Figure 11, there are more people who have graduated from college than simply high school. Moreover, there are more people who didn't graduate from college than didn't graduate high school. Thus, a sizable proportion of the US electorate is relatively educated. In Figure 12, people who are 60+ are by far the largest age group, showing an aging population. Given that this population tends to support Trump, this is an important statistic to keep an eye on.

You can see from Table 1 that the total number of people who completed high school is 65.83% of the overall

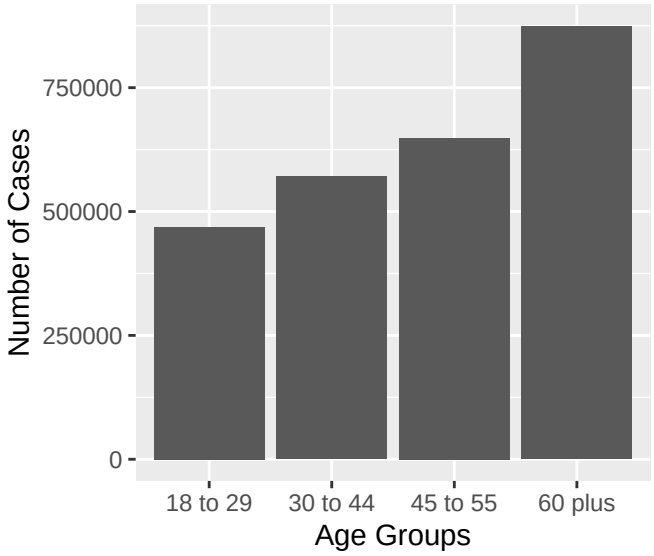


Figure 12: Distribution of Age

population, whereas the proportion of people who completed college is 38.75%. Based on Table 2, the people whose age are 60 and above has 35.4% of overall population, which is the largest portion in US.

Figure 13 shows the distribution of population amongst US states. The largest by far is California (shown earlier to skew towards Biden), the second largest Texas (which tends to skew towards Trump). Figure 14 shows that the population is highest amongst coastal and southern states, and this may skew the popular vote from the electoral vote.

Figure 15 has the distribution of gender, shown to be 51.56% for women and 48.44% for men. Figure 16 shows that much like the survey data from Dem [2020], the largest race group in the US is white, and the second largest race group is Black/African American. Based on Table 4, the white population has 2,031,750 people, and 78.05% of the overall population. The Black/African American population has 249,642 people, and 9.59% of the overall population. This indicates they are important sectors of the population for the candidates to campaign for.

3 Modeling with MRP

Different levels of the variables of interest in the post-stratification process show different responses for presidential votes (support for Joe Biden). The survey's demographics differ from the actual population, therefore certain groups of individuals were under/overrepresented in the survey, leading to sampling bias. Therefore, we use the survey sample data to infer presidential votes on the census population in the US using post-stratification variables. In order to address the over/underrepresentation of certain demographics within the survey results, we perform multilevel regression and poststratification (MRP) on the survey data from the Democracy Fund Voter Study Group with census data from the ACS provided by IPUMS. In doing so, we calculate the proportions of each group of individuals from the actual population distribution, the census data, and apply the results of the survey onto these proportions. In doing so, we basically reweigh the results from the survey to post-stratify our results to produce an estimate of the 2020 US election that more accurately represents the actual population distribution. This method is effective for using non-representative samples to make predictions, however can be quite troublesome when cells created of each group are too small, creating unstable results for each variable [Alexander, 2019]. Throughout our cleaning process, we had to alter the way each group was levelled off in order to create cells that were large enough to produce stable results.

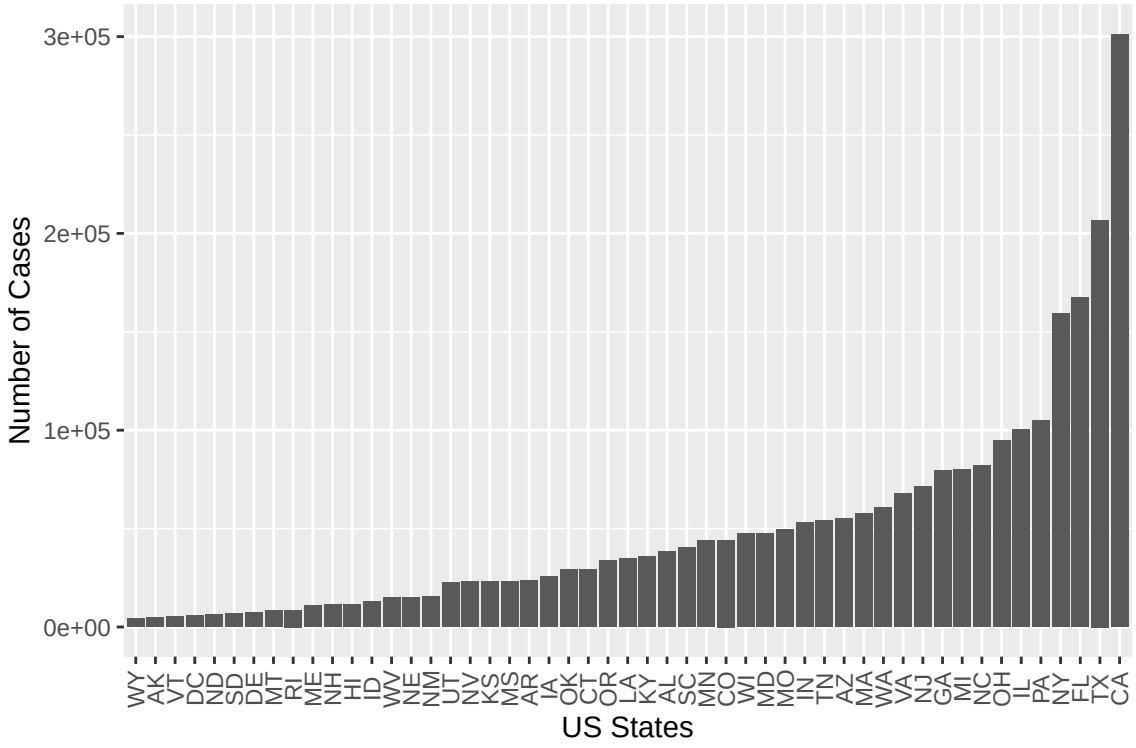


Figure 13: Population for Each State

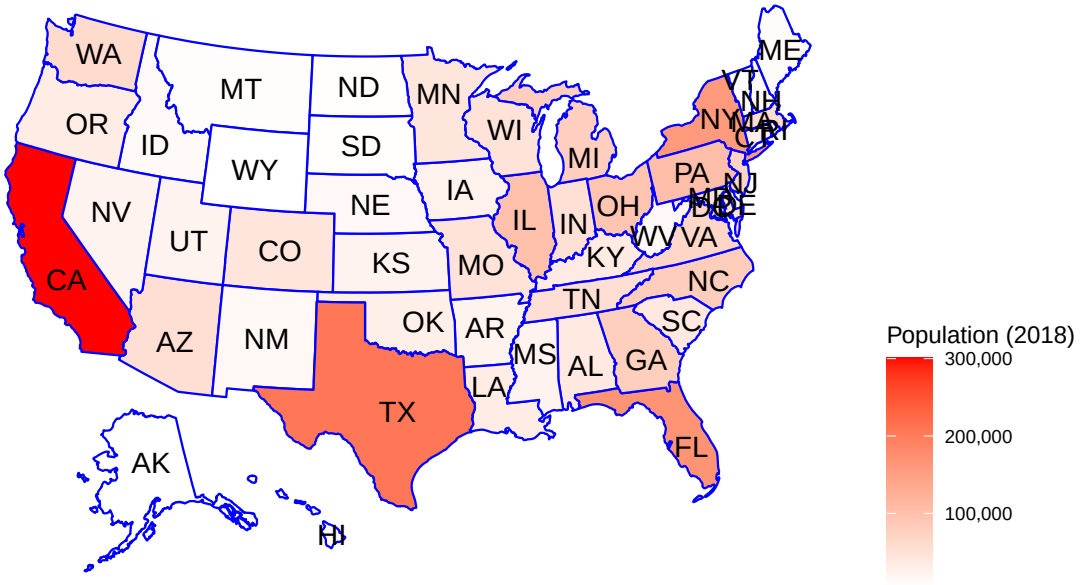


Figure 14: States Population Map

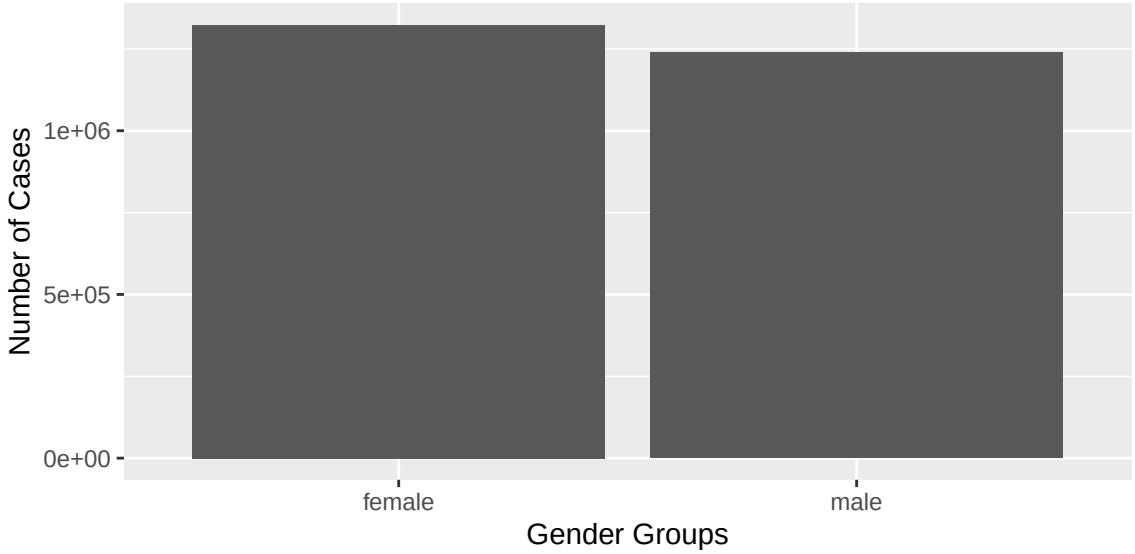


Figure 15: Distribution of Gender

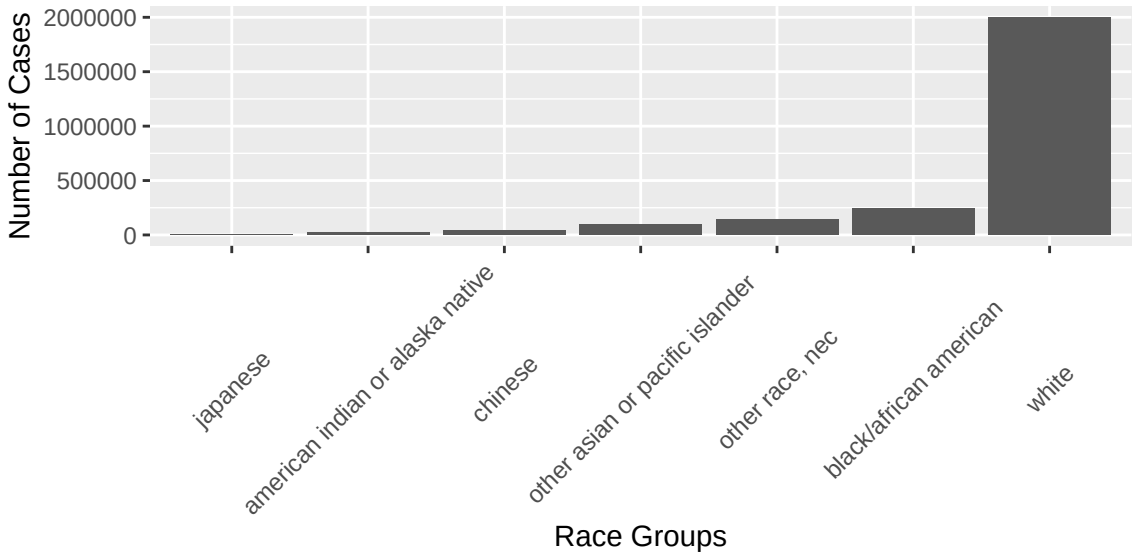


Figure 16: Population for Races

Steps of MRP

1. Gather sample data(survey data) with certain key variables in mind (state, age, gender, education, race).
2. Gather population data(census data) with the same key variables.
3. Match the variables of each dataset so each of the cells will correspond with each other, if necessary.
4. Decide the quantity you would like to measure in the sample (support for Joe Biden = 1, support for Donald Trump = 0).
5. Estimate quantity of interest in the population using MRP to predict using the key demographics from the sample.
6. Apply this model to the new data (census data) and make estimates of predictions for the population.

Our first approach in modelling predictions was to look at linear regression on the *dummy_vote* variable and the five explanatory variables selected using the function *lm()*. We chose to use a frequentist approach, where parameters of interest are fixed, first to explore both datasets. There are assumptions made about the data when choosing to perform ordinary linear regression on the survey data . 1. The relationship between all explanatory variables and the response variable follows a linear pattern. 2. The variance of the residuals is constant for all values/levels of the explanatory variables. 3. Observations are independent of each other. 4. All variables are normally distributed. The estimates for the linear regression model and the generalized logistic regression model provide the same estimates for the explanatory variables, in our case, the proportion of votes for Mr. Joe Biden, however, the variances for each explanatory variable will differ.

Linear regression general formula:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Since the outcome of the response variable is either 1 or 0 (1 indicating a vote for Joe Biden), a logistic regression can be used to address the binary response variable [Andrew, 2020]. The function *glm* from the allows defining of the response variable to a binomial distribution to the logistic regression. For p as the probability of support for Joe Biden, the logistic regression can be modelled in this formula:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_2 x_2 + ... + \beta_k x_k$$

The coefficients in this logistic regression model represent the change in log odds of support for Joe Biden. Since gender was only recorded as either 'male' or 'female, this explanatory variable is binary. Race, education, state and age group were all categorical variables By conducting logistic regression on the multiple explanatory variables with multiple levels, a few things jumped out. Since many variables and cells are being represented in the model, we wanted to find out which explanatory variables were statistically important to the raw data obtained from the sample in order to find the best fitting models for predicting on other populations. Certain variables could be identified as statistically important by evaluating the p-values for each, and the estimate for each coefficient. Our second approach in modelling predictions was to fit a Bayesian model to address the fact that we are predicting a variable in the census data based on the survey data. In using a Bayesian model, we intended to compare our prediction results with a frequentist model to compare how much the prediction was influenced by the census proportions. We used the *brm* function from the **brms** package [Bürkner, 2017, Stan Development Team [2020]] and organized the data and results with the **tidybayes** [Kay, 2020] ,**tidyverse** [Allaire et al., 2019], **magrittr** [Bache and Wickham, 2014], **gridExtra** [Auguie, 2017], usmap [Di Lorenzo, 2020], **scales** [Wickham and Seidel, 2020], **haven** [Wickham and Miller, 2020], **broom** [Robinson et al., 2020] and **here** [Müller, 2017] packages. Bayesian inference assumes certain priors about the parameters of interest, meaning that the parameters are not fixed, rather they follow some type of distribution [Christian-Burkner, 2017]. Since our parameter of interest is being formed in the post-stratification data, our new parameter of interest is the proportion of Biden support in the census population, which also has a random binary outcome. In the Bayesian model, we assume a bernoulli distribution for the support of Joe Biden in the US 2020 election.

Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)}$$

A is the prior distribution that is assumed about the parameter of interest, in our case the distribution of the proportion of Biden support in the census data whereas B is the census data [Simpson, 2019].

Bernoulli Distribution:

$$f(k;p) = \begin{cases} p & \text{if k= 1,} \\ 1-p & \text{if k = 0.} \end{cases}$$

There are two main things to look out in this comparison of estimates based on the raw data and the MRP: how the post-stratification changes the estimate obtained from the raw data, and how much variance each estimate has based on each level of the variable. In addition, analyzing the proportion of the census data and the survey data will provide additional insight on how the census proportions of groups affect the MRP estimate. As shown in Figure 19, where the estimates based on each State are being compared, the MRP estimate does not differ significantly from the raw data. This indicates that the proportion of individuals surveyed residing in each state was similar to what was observed within the census data. The red bars indicate the 97.5% confidence interval for each estimate.

4 Results

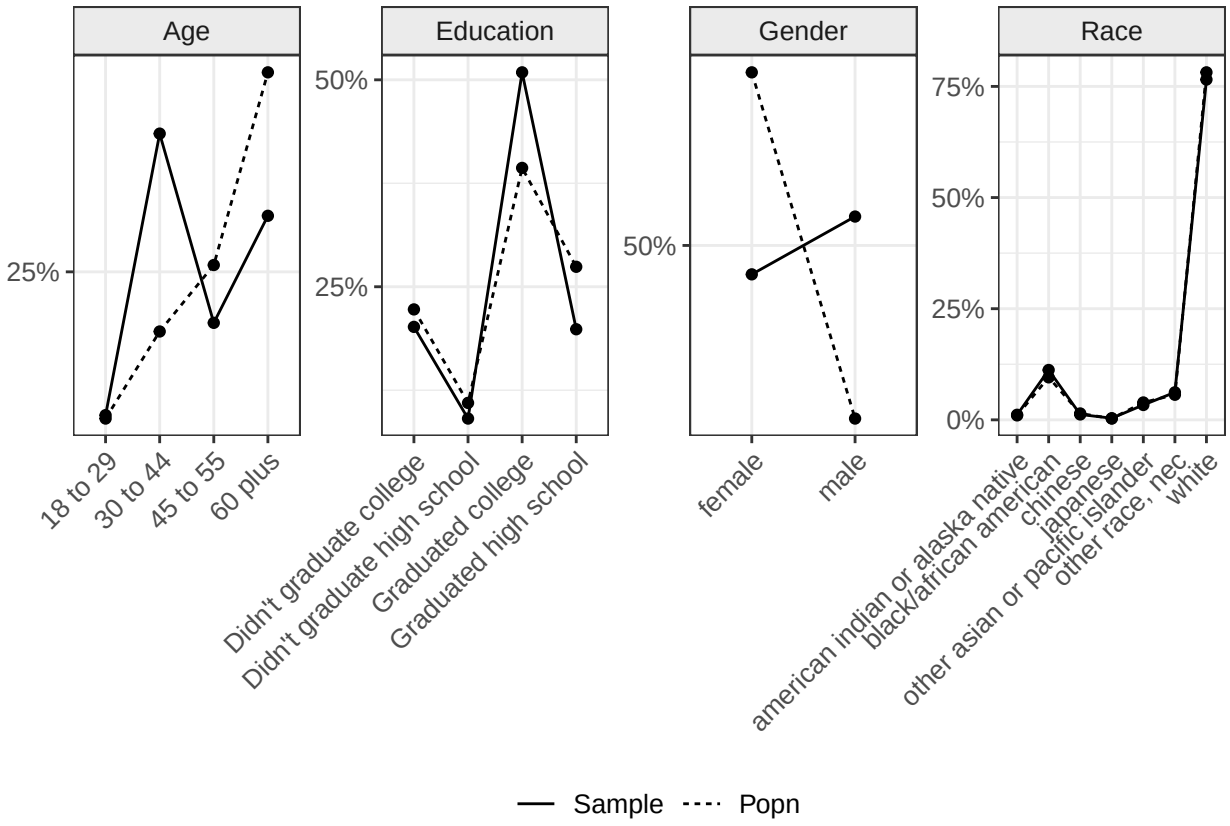


Figure 17: Demographic Proportions

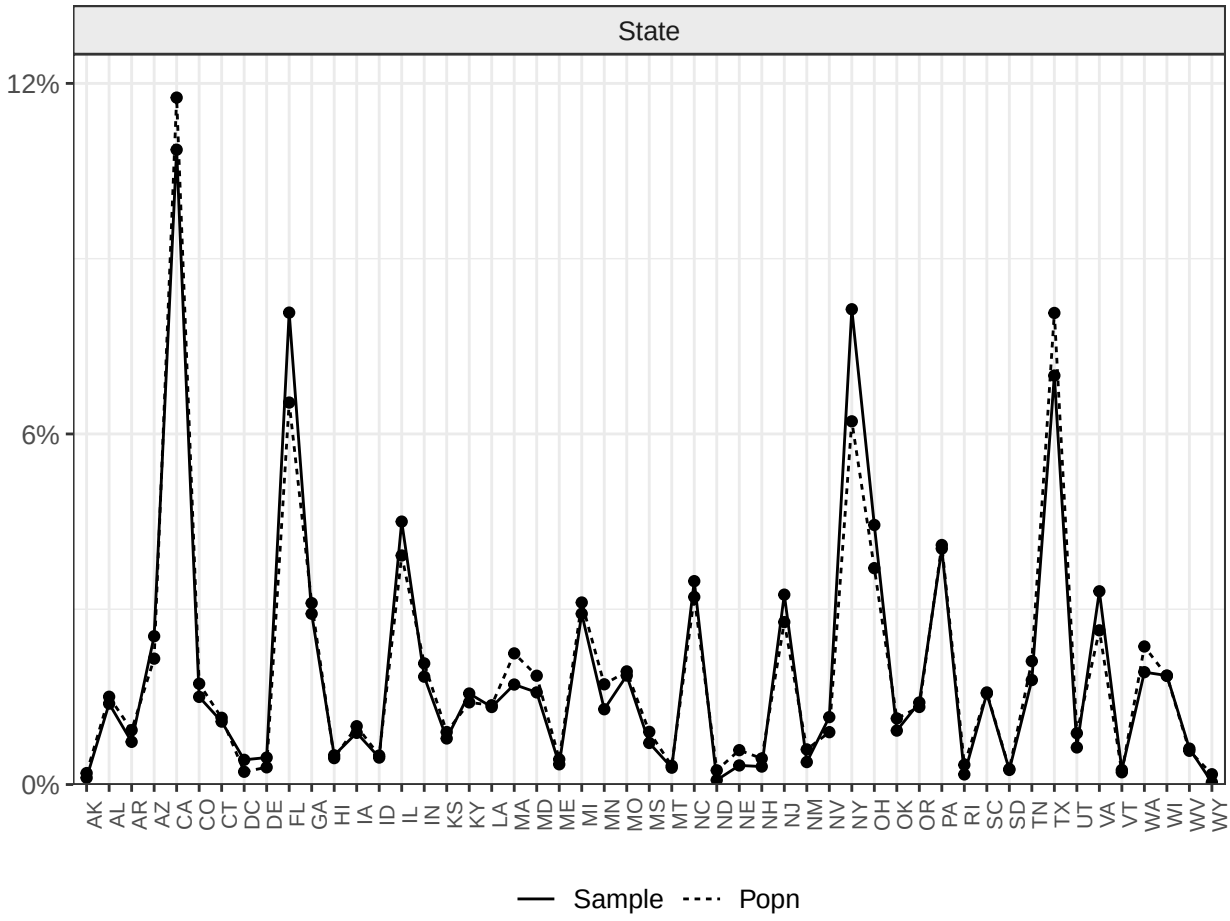


Figure 18: Proportions by State

4.1 Bayesian Approach to MRP

Figures 19-23 are graph the distibution of estimates using Bayesian inference in the **brms** package [Bürkner, 2017].

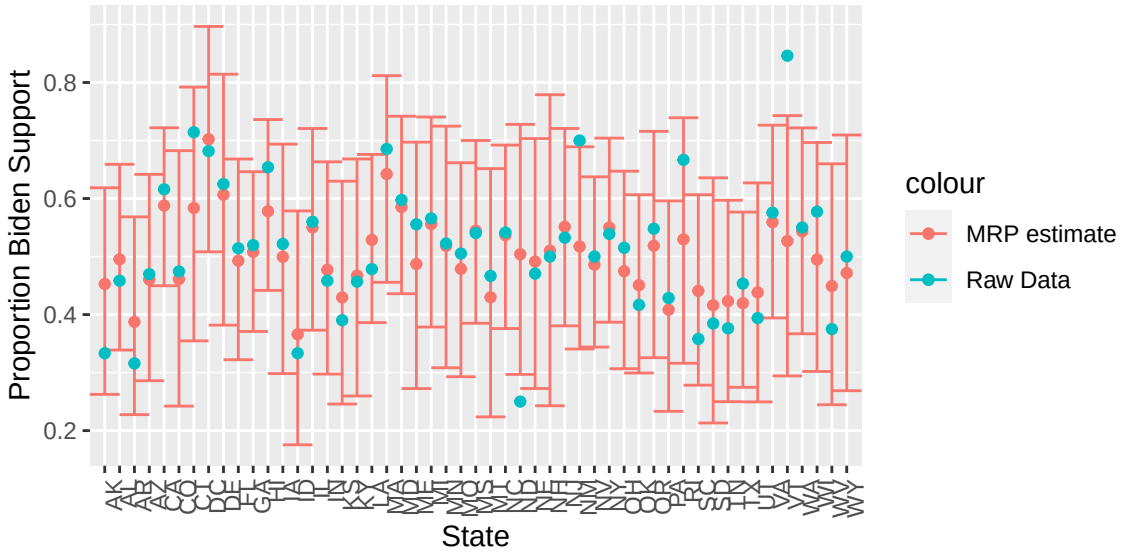


Figure 19: Comparing Survey Estimates with MRP Estimates by State

The prediction of proportion of support for Biden in the post-stratification across each state shown in Figure 19 does not seem to differ much from the original proportions found in the sample distribution, considering that each state has relatively the same amount of proportions across both datasets, this is not surprising. However, there were a few states that had quite different estimates. Vermount, Rhode Island and New Mexico were estimated to be less supportive of Joe Biden in the MRP modelling whereas Alaska and North Dakota were more supportive of Joe Biden than in the original raw sampling data.

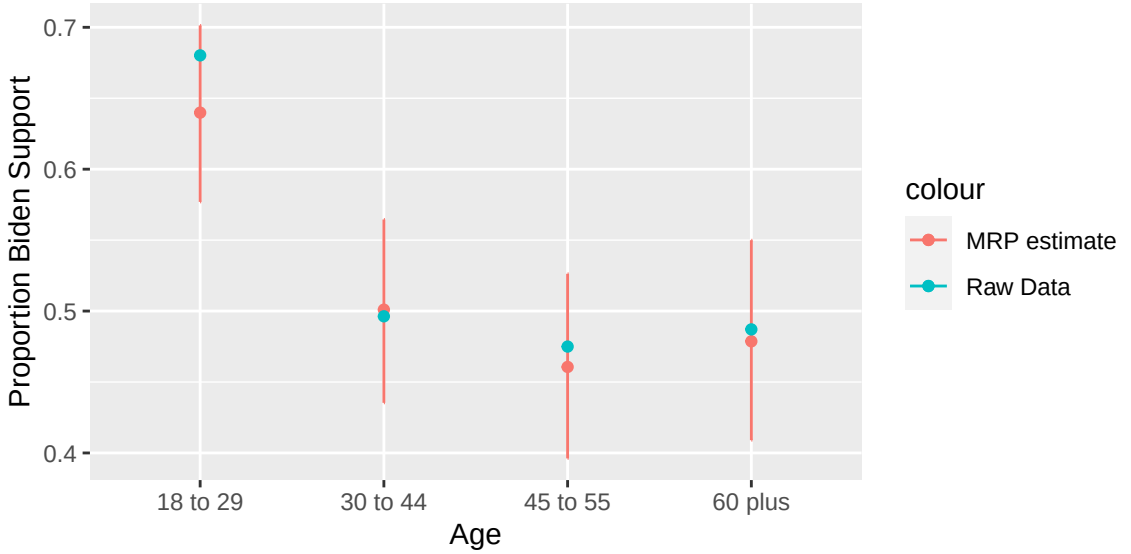


Figure 20: Comparing Survey Estimates with MRP Estimates by Age

The estimate for the Biden support for individuals in age group 18-19 were slightly higher in the raw data shown in Figure 20. However, this age group was still particularly in high favor of Joe Biden as both estimates were over 60%. Individuals in the 30-45 age category seemed to respond more to the polling survey, thus creating slightly biased results. However, the estimates for individuals in the 30-45 age category were consistent in the post-stratification data and the survey data. It appears that among both groups, older age groups appear to be less supportive of Joe Biden, but only marginally.

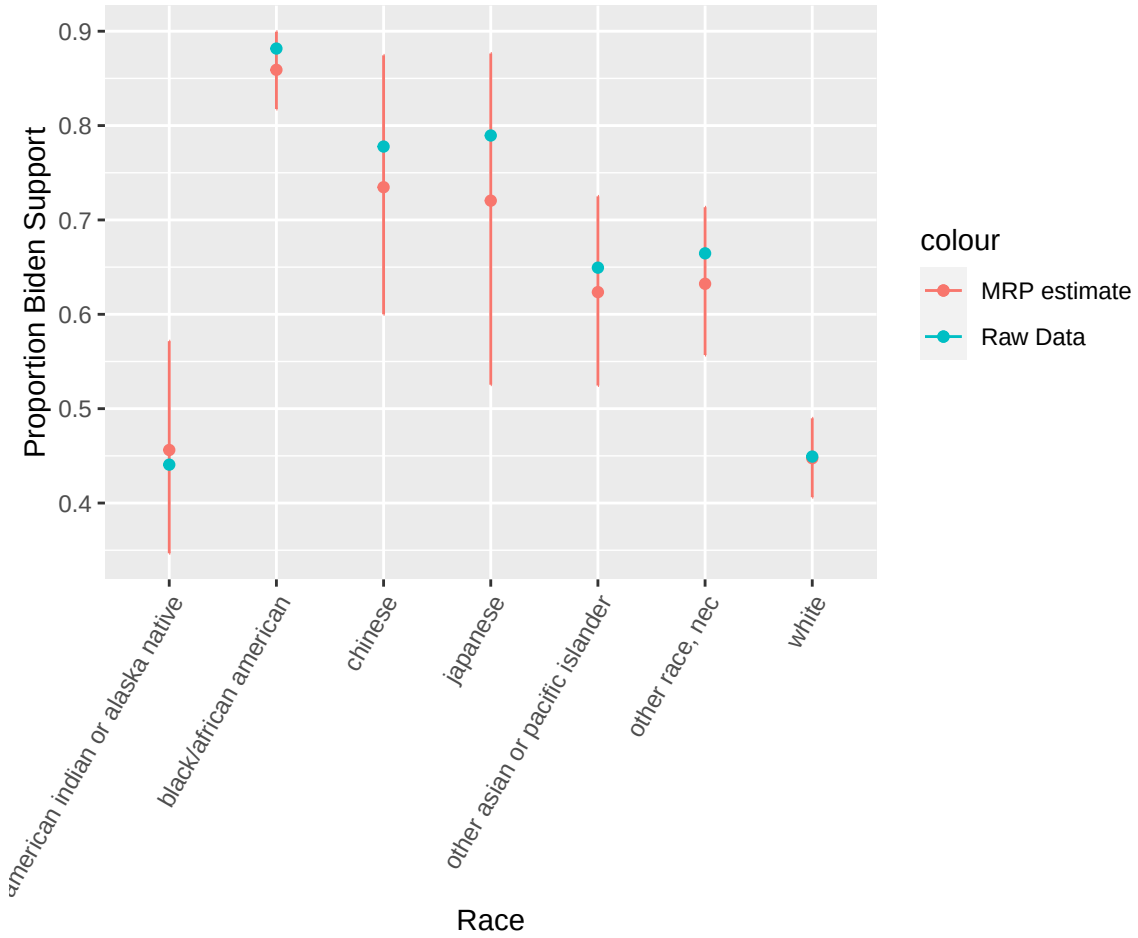


Figure 21: Comparing Survey Estimates with MRP Estimates by Race

Estimates between survey data and post-stratification data remain relatively the same for race shwon in Figure 21. Among black/African Americans, Chinese and Japanese age groups, Biden appears to be well supported with estimates exceeding 70% for each age group. American Indian and Alaska Natives as well as white individuals were slightly less supportive. Based on this graph alone it would appear that Joe Biden would have a significant lead, however, the proportion of white individuals in the census population and the survey population exceeds 75%, comprising a significant portion of the population.

On average, females tend to be more supportive of Joe Biden while males tend to support Donald Trump as shwon in Figure 22. Although females represent more of the census population than males, this is not reflected in the survey where males had a slightly higher response rate. This is reflected in the difference between estimates of females being decreased in the MRP estimate by almost 5%.

It appears that there is an overrepresentation of individuals who have graduated college who responded to the surveys reflected in Figure 23. In addition, there is an underrepresentation of those who had just graduated high school. The MRP estimate was lower than the raw data for both individuals who graduated

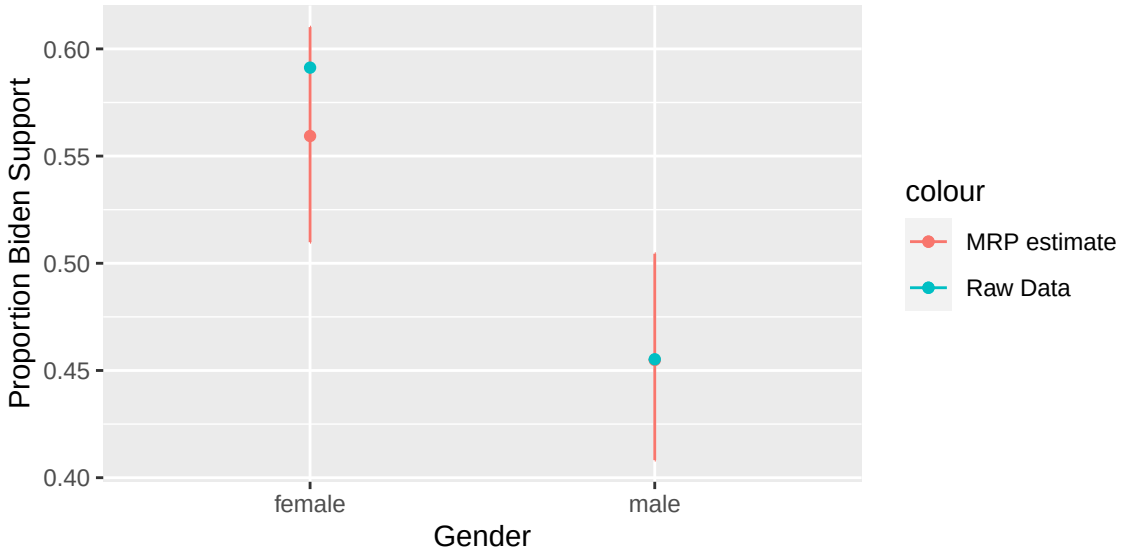


Figure 22: Comparing Survey Estimates with MRP Estimates by Gender

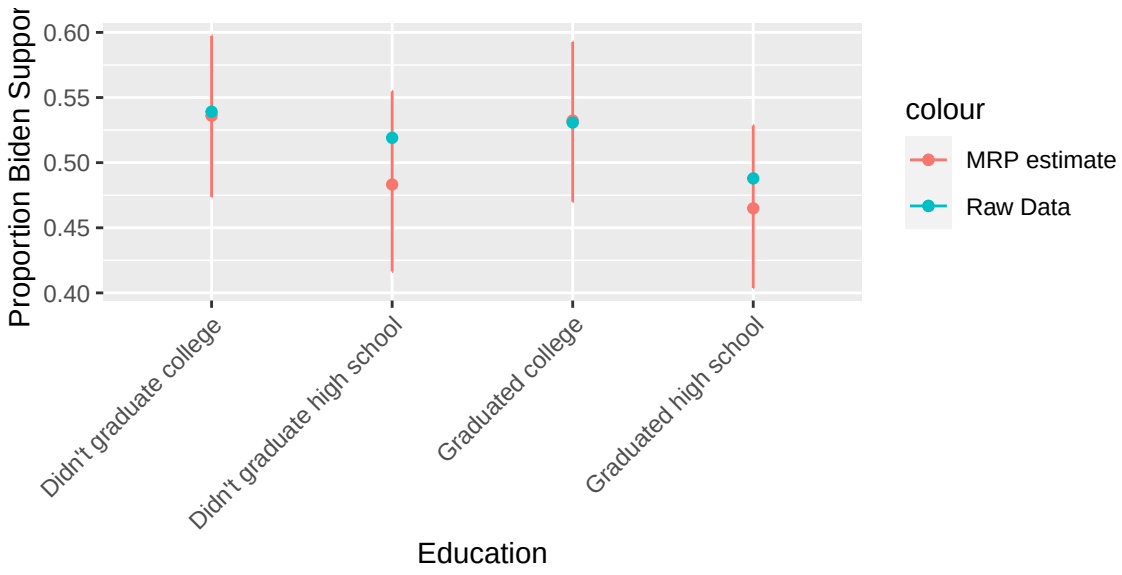


Figure 23: Comparing Survey Estimates with MRP Estimates by Education

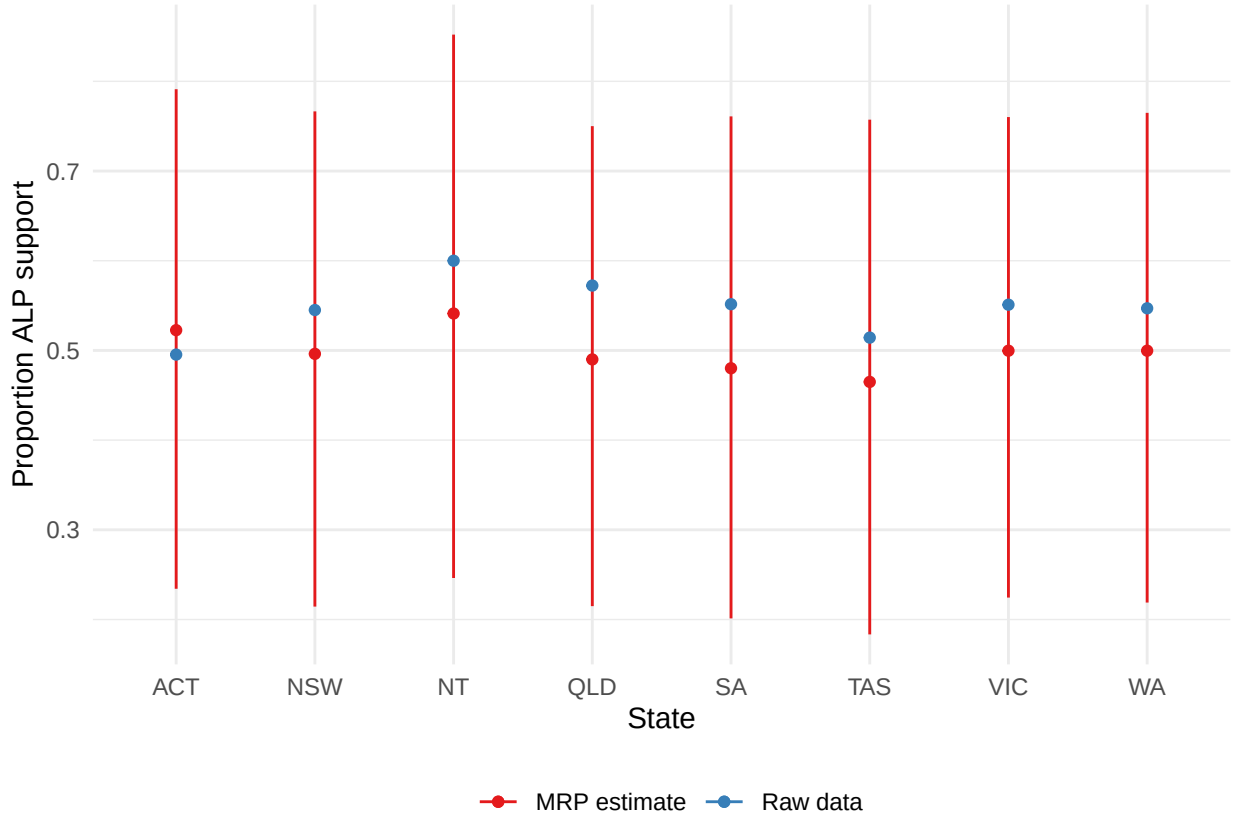
8 Communication

There are many interesting aspects that we may like to communicate to others. For instance, we may like to show how the model is affecting the results. We can make a graph that compares the raw estimate with the model estimate.

```
post_stratified_estimates %>%
ggplot(aes(y = mean, x = forcats::fct_inorder(state), color = "MRP estimate")) +
geom_point() +
geom_errorbar(aes(ymin = lower, ymax = upper), width = 0) +
ylab("Proportion ALP support") +
xlab("State") +
geom_point(data = example_poll %>%
               group_by(state, supports_ALP) %>%
               summarise(n = n()) %>%
               group_by(state) %>%
               mutate(prop = n/sum(n)) %>%
               filter(supports_ALP==1),

             aes(state, prop, color = "Raw data")) +
  theme_minimal() +
  scale_color_brewer(palette = "Set1") +
  theme(legend.position = "bottom") +
  theme(legend.title = element_blank())
```

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



Similarly, we may like to plot the distribution of the coefficients.[2]

```
model %>%
  gather_draws(`b_.*`, regex=TRUE) %>%
  ungroup() %>%
  mutate(coefficient = stringr::str_replace_all(.variable, c("b_" = ""))) %>%
  mutate(coefficient = forcats::fct_recode(coefficient,
                                           Intercept = "Intercept",
                                           `Is male` = "genderMale",
                                           `Age 30-44` = "age_groupages30to44",
                                           `Age 45-59` = "age_groupages45to59",
                                           `Age 60+` = "age_groupages60plus"
                                           )) %>%

# both %>%
  ggplot(aes(y=fct_rev(coefficient), x = .value)) +
  ggridges::geom_density_ridges2(aes(height = ..density..),
                                 rel_min_height = 0.01,
                                 stat = "density",
                                 scale=1.5) +
  xlab("Distribution of estimate") +
  ylab("Coefficient") +
  scale_fill_brewer(name = "Dataset: ", palette = "Set1") +
  theme_minimal() +
```

[2]You can work out which coefficients to be pass to gather_draws by using tidybayes::get_variables(model). (In this example I passed 'b_', but the ones of interest to you may be different.)

```
theme(panel.grid.major = element_blank(),
      panel.grid.minor = element_blank()) +
theme(legend.position = "bottom")
```

high school and who did not graduate high school.

5 Discussion

Information of interest collected through surveys is often not representative of the population that we are looking into. Due to the nature of traditional surveys, accurately representing the actual population of interest is difficult, expensive, timely and perhaps impossible when dealing with very large-scaled populations. Luckily, there are some ways to address this issue by using MRP to correct misproportions in samplings. Survey data collected from the Democracy Fund Voter Study Group employed stratified sampling. Stratified sampling divides the population into subpopulations based on certain characteristics or attributes. These subpopulations are called strata. Stratified sampling helps to somewhat control the proportions of certains groups being included in the sample instead of having all groups of individuals having an equal chance to be represented in the survey. Thus the sampling data is very extensive and somewhat pre-weighted due to stratified sampling, however, looking at the differences in proportions between each group there still exists some sampling biases for certain groups, such as age. Indicated in the logistic regression, certain factors such as age, gender and race had lower p-values, indicating they were more statistically important to the model than variables such as state. Only certain states reported low p-values, such as Connecticut, Massachusetts and Vermont, indicating statistical importance on the logistical model. In the logistic regression and Bayesian model, we found that Biden had a predicted vote proportion of 50.1% and 51.03%, respectively . Considering that the proportions of all the levels of variables did not differ greatly, except across age groups, the predictions remained relatively similar to each other. The survey data was already pre-weighted to a certain extent using stratified sampling, therefore proportions across race for each of the datasets were very similar. The main difference between proportions in the datasets was within the age group, where younger individuals tended not to respond to surveys as much as middle-aged individuals. The MRP estimates that were quite different from the raw data for each state were moste likely affected by the different age groups sampled and possibly even gender. We must also not forget about particular cells to identify any cross sectional relevance of the variables and if there is any type of relationship for certain groups. Conducting MRP over 5 variables of interest helped to identify and narrow down key variables and levels that could further influence voting choice in the US 2020 elections.

It is worthwhile to mention that the survey used was collected back in June. Since then, many events surrounding the election have changed public opinion of both candidates, and therefore will ultimately affect the final outcome of the US 2020 Presidential Election taking place on November 3rd, 2020. Further work might look into conducting bayesian inference models across time to model changes over periods of time leading up to the election to provide further insight. Time series analysis can be helpful in understanding trends that are relative to the election date.

MRP can be used as a tool when a population of interest is identified with key variables that can affect the variable of interest, which can be matched to a sample population to formulate inferences. Comparing the sample estimate of Joe Biden support with the forecasted estimates on the census population provides further tools for research to explore nuances in the data and additional differences. MRP ultimately can identify certain variables of interest as predictions to perform smaller scaled polling to include more levels of variables of the data that were hidden in the way we categorized individuals within each of the groups, such as race and education in our case. It is relatively quick, easy and cost effective to perform on data given a survey with matching key variables with a more representative population.

MRP has lead many studies to significant and representative models on actual populations from samples, however it does not come without flaws. Insufficient estimates can be caused by lack of some demographic predictors, insufficient data from surveying as well as lack of regularization. In some cases, selecting certain variables as predictors lead to extremely small cell counts for certain groups, therefore producing unstable estimations. Through our cleaning process of both datasets, some tweaking of the variables and levels had to be made in order to match all variables to perform MRP. The sample data was extremely extensive and all-inclusive of any possible variable, however this is not always the case. The way that information

demographic variables and in the sample and census data are sometimes collected and recorded in different forms. [Kennedy, 2020] [Gelman, 2020].

6 Appendix:

Table 1: Level of Education Statistics

	Number of Cases	Proportion
Didn't graduate college	570553	0.2226
Didn't graduate high school	281009	0.1097
Graduated college	1008593	0.3936
Graduated high school	702436	0.2741

Table 2: Age Statistics

	Number of Cases	Proportion
18 to 29	469102	0.1831
30 to 44	570930	0.2228
45 to 55	648619	0.2531
60 plus	873940	0.341

Table 3: Gender Statistics

	Number of Cases	Proportion
female	1322620	0.5161
male	1239971	0.4839

Table 4: Race Statistics

	Number of Cases	Proportion
american indian or alaska native	25700	0.01003
black/african american	244965	0.09559
chinese	36838	0.01438
japanese	7442	0.002904
other asian or pacific islander	100149	0.03908
other race, nec	144585	0.05642
white	2002912	0.7816

References

Design and Methodology Report, 2014. URL https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html. Version 2.0.

Monica Alexander. *Analyzing name changes after marriage using a non-representative survey*, 2019. URL https://www.monicaalexander.com/posts/2019-08-07-mrp/.

JJ Allaire, Jeffrey Horner, Yihui Xie, Vicent Marti, and Natacha Porte. *markdown: Render Markdown with the C Library 'Sundown'*, 2019. URL https://CRAN.R-project.org/package=markdown. R package version 1.1.

Andrew. *Linear or Logistic Regression with Binary Outcomes*, 2020. URL https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/.

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2014. URL https://CRAN.R-project.org/package=magrittr. R package version 1.5.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

Paul Christian-Burkner. *Advanced Bayesian MultiLevel Modeling with the R package brms*, 2017. URL https://arxiv.org/pdf/1705.11123.pdf.

Nationscape Data Set. Democracy Fund Voter Study Group, Washington, USA, 2020. URL https://www.voterstudygroup.org/publication/nationscape-data-set.

Paolo Di Lorenzo. *usmap: US Maps Including Alaska and Hawaii*, 2020. URL https://CRAN.R-project.org/package=usmap. R package version 0.5.1.

Quint Forgey. *Biden: 'I reject the premise that' remote campaigning is hurting my White House bid*, 2020. URL https://www.politico.com/news/2020/05/12/joe-biden-remote-campaign-not-hurting-white-house-bid-251142.

Andrew Gelman. *Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample*, 2020. URL https://arxiv.org/pdf/1906.11323.pdf.

Patrick Hipes. *U.S. Coronavirus Update: Infections in America Hit High Not Seen Since July Peak As Hospitals In Some Smaller States Are Overwhelmed*, 2020. URL https://deadline.com/feature/coronavirus-deaths-united-states-1202874446/.

Matthew Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2020. URL http://mjskay.github.io/tidybayes/. R package version 2.1.1.

Lauren Kennedy. *Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample*, 2020. URL https://arxiv.org/pdf/1906.11323.pdf.

Andrew Mercer, Claudia Deane, and Kyley McGeeney. *Biden: 'I reject the premise that' remote campaigning is hurting my White House bid*, 2016. URL https://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/.

Kirill Müller. *here: A Simpler Way to Find Your Files*, 2017. URL https://CRAN.R-project.org/package=here. R package version 0.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. 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.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. *IPUMS USA: Version 10.0 us00001.dta*. IPUMS, Minneapolis, US, 2020. URL https://doi.org/10.18128/D010.V10.0.

Dan Simpson. *Multilevel (structured) regression and post-stratification*, 2019. URL https://statmodeling.stat.columbia.edu/2019/08/22/multilevel-structured-regression-and-post-stratification/.

Stan Development Team. RStan: the R interface to Stan, 2020. URL http://mc-stan.org/. R package version 2.21.2.

Hadley Wickham and Evan Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, 2020. URL https://CRAN.R-project.org/package=haven. R package version 2.3.1.

Hadley Wickham and Dana Seidel. *scales: Scale Functions for Visualization*, 2020. URL https://CRAN.R-project.org/package=scales. R package version 1.1.1.