Prediction of the 2020 Presidential Election*

Spoiler Alert: Trump Wins

Anees Shaikh, Jaffa Romain, Lu Mu, and Cameron Fryer

02 November 2020

Abstract

The purpose of this analysis is to forecast the results of the upcoming United States 2020 presidential election. In this paper, we first consider the Democracy Fund + UCLA Nationscape Wave 50 dataset, which contains the results of a survey (conducted June 25-July 1, 2020) on American voter attitudes. Thereafter, the survey data is used to train our model relating voter intent to 4 explanatory variables. The model is then applied to the post-stratification dataset; namely, the results of the 2018 1-year American Community Survey (ACS), to get predictions representative of the target population. Since the ACS data pertains to individual persons and their characteristics, the associated use of our model allows us to conclude that Donald Trump will be relected as the President of the United States, winning the popular vote by _____ with a margin of error of _____

1 Introduction

. . . ..

2 Data

Individual Survey Data

As previously mentioned, our data is from two different datasets; the Democracy Fund + UCLA Nationscape Wave 50 dataset, and the 2018 1-year American Community Survey (ACS) dataset. They each contain both qualitative and quantitative data. To forecast the results of the presidential election, our model was fitted using the Nationscape survey data. The Democracy Fund + UCLA Nationscape is "one of the largest public opinion surveys ever conducted – interviewing people in nearly every county, congressional district, and mid-sized U.S. city in the lead up to the 2020 election." **?** Consequently, interviewing "roughly 6,250 per week" was part of the survey methodology employed by Nationscape. As such, the Nationscape Wave 50 dataset contains results for 6,479 interviews conducted during the week of June 25-July 1, 2020. It can be accessed by visiting https://www.voterstudygroup.org/publication/nationscape-data-set and submitting a request to access the data **?**. By providing your full name and email address, you will then be emailed a link to the page where it can be downloaded.

Nationscape samples are provided by Lucid, a market research platform that uses their software for conducting online surveys. The Nationscape surveys were completed by participants online in English **?** . The survey data is then weighted using a simple raking technique, where one set of weights is generated for each week's survey, in an attempt to have data that is representative of their target population. The survey data is weighted on many demographic factors such as gender, age, and nativity.

*Code and data are available at: https://github.com/aneesshake/elections_prediction

At first glance of the Wave 50 data, we noted that it contains 265 variables. This is noteworthy because it allows a more complete picture of voter attitude to be derived from the data compared to if there were less variables. Another strength of this dataset is the overall quality of responses from those surveyed. More specifically, the data does not contain any duplicate records, inappropriate entries, or poor entries (e.g. typos, misspellings); hence, making it easier to work with. This high quality of survey responses can be at least partially attributed to the fact that "Nationscape samples are provided by Lucid, a market research platform that runs an online exchange for survey respondents. **?**" What is more, the individuals belonging to the sample are required to "complete an attention check before completing the [online] survey **?**." With this method of sample selection, along with the precaution taken to ensure their attentiveness, it is unlikely for survey respondents to submit low-quality responses.

The main weakness of the Nationscape Wave 50 dataset is the likelihood that it includes sampling error and/or coverage error. While the goal of the Nationscape survey is to determine American voter attitudes, making inferences based only on the Wave 50 dataset would result in sample bias. This is because the target population is all American citizens age 18 years and older, whereas the sample frame is Lucid's online exchange for survey respondents, and the sample is a set of 6,479 individuals from said online exchange which match "a set of demographic criteria **?**." It is unlikely that a survey sample of 6,479 American citizens during a single week is representative of American voter attitudes as a whole. This may lead one to wonder why Nationscape chose to use the described sampling approach. Apart from the fact that the survey was conducted every week, the reason is trade-offs. For example, although there is potential coverage error in the data–a common result of online surveys–some trade-offs are a lesser rate of non-response, lower cost, convenience and accuracy. Whilst on the topic of non-response, it is worth mentioning that the variable with greatest amount of non-response in the Wave 50 data set is household income at roughly 5.5%. Nonetheless, this is a much smaller percentage of non-response in household income than is typically seen in political surveys, and Nationscape handles it by not weighting income for non-respondents. Similarly, a statement from Nationscape in its "Representativeness Assessment" implies an overall lack of concern for non-response in the data, as "previous evaluations of the samples Lucid provides have found them to be of high quality **?**."

The last aspect of the Nationscape Wave 50 dataset to be considered is the survey questionnaire itself. Strengths of the questionnaire include its structure, the design and wording of the questions being asked, and the variety of choices for closed-ended questions. As to the structure of the questionnaire, it contains an extensive amount of questions from which differing political views are revealed – the ideal circumstance for reflecting on American voter attitudes. Furthermore, the survey questions are asked in an order that does not influence the answer of respondents to subsequent questions. In terms of the design and wording of the questions, they are closed-ended, and avoid the use of tricky wording. This is a strength of the questionnaire because it makes comparing the responses of individuals and thus, statistical analysis, much simpler than it would be with open-ended questions and/or troublesome language. The diverse number of options available to select as answers to the questions is a strength of the questionnaire because they are both mutually exclusive and exhaustive; hence, ensuring survey respondents are able to answer every question. With regard to weaknesses of the questionnaire, it seems that there are none other than its sheer length.

The data was cleaned to get our target population of eligible voters that intend to vote in the 2020 elections. Voters who didn't plan to vote or not eligible were filtered from the data set, and the variable pertaining to the 2020 election vote was cleaned to consist only of voters who plan on voting for either Trump or Biden. Before the process of fitting our model, we selected variables of interest that we felt gave insight on how different demographics vote. The resulting data set consisted of 3541 observations, and later reduced to the variables state, age, sex, Hispanic, race, and intended 2020 vote.

Age: The age of a respondent. Age was originally a nominal value(numerical), but the data was later grouped into categories for our predictive model, and to match the post-stratification data. This allowed for us to look at how age groups might influence the vote. The 2020 election marks the first time that Millennials and Gen Z will have an equal share of the number of eligible voters as Baby Boomers and earlier generations **?** . Older voters in the past have had higher turn out rates, so looking at how how age categories tend to vote can give insight on how this will play a factor in the future election. We can see in (Figure **??**) that there is a much smaller proportion of younger voters, but this is the age group that seems to have the strongest

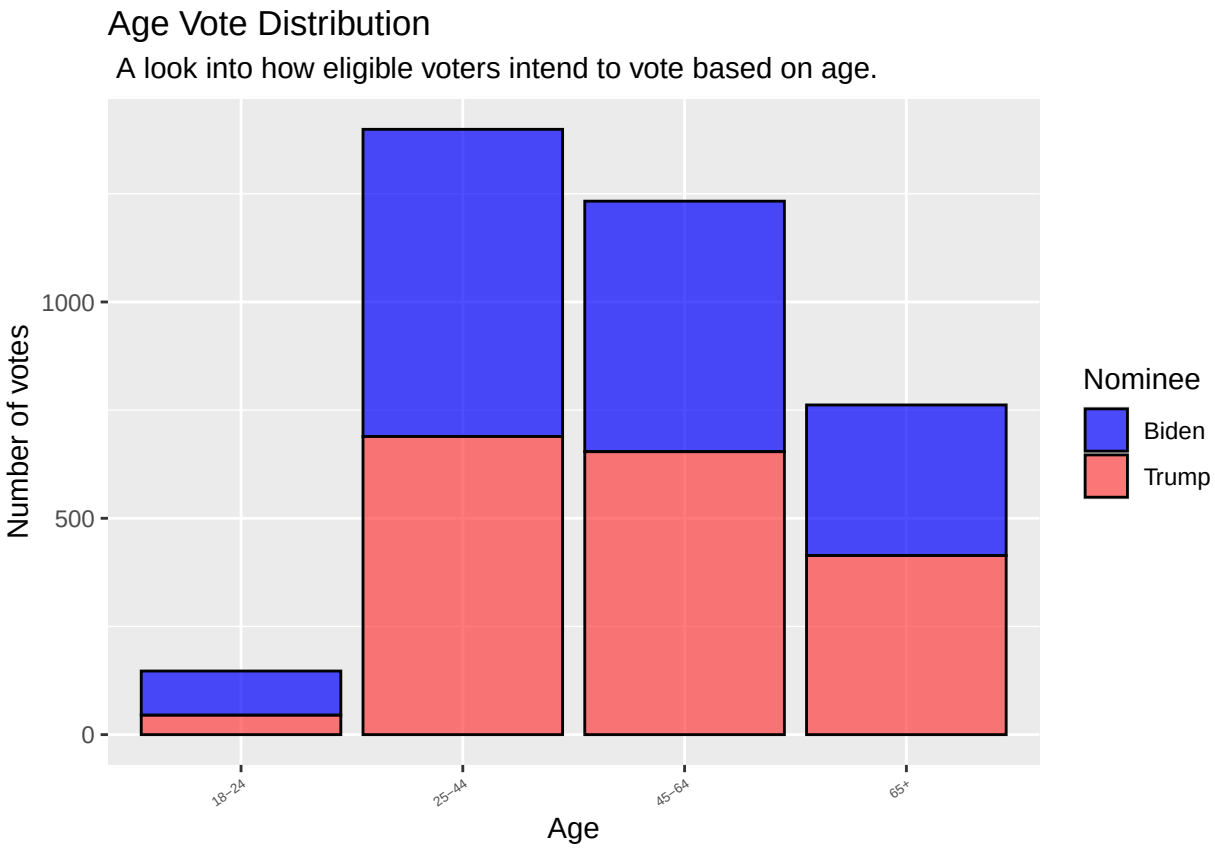


Figure 1: Distribution of Age in Nationscape data

Democratic support.

Sex: The sex of a respondent. Certain policies currently being such as those on anti-Abortion largely impact the female population in the United States. It seems that sex can be an important indicator of how voters respond to the platforms of the parties with regards to sex-specific policies.

Hispanic: This variable looks at whether a respondent is Hispanic or not. In the 2020 election, Latinos will be the largest racial/ethnic minority in the electoral. With Trump's previous comments on countries such as Mexico, it seems important to see how pivotal the vote of this group will be in the election **?**.

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

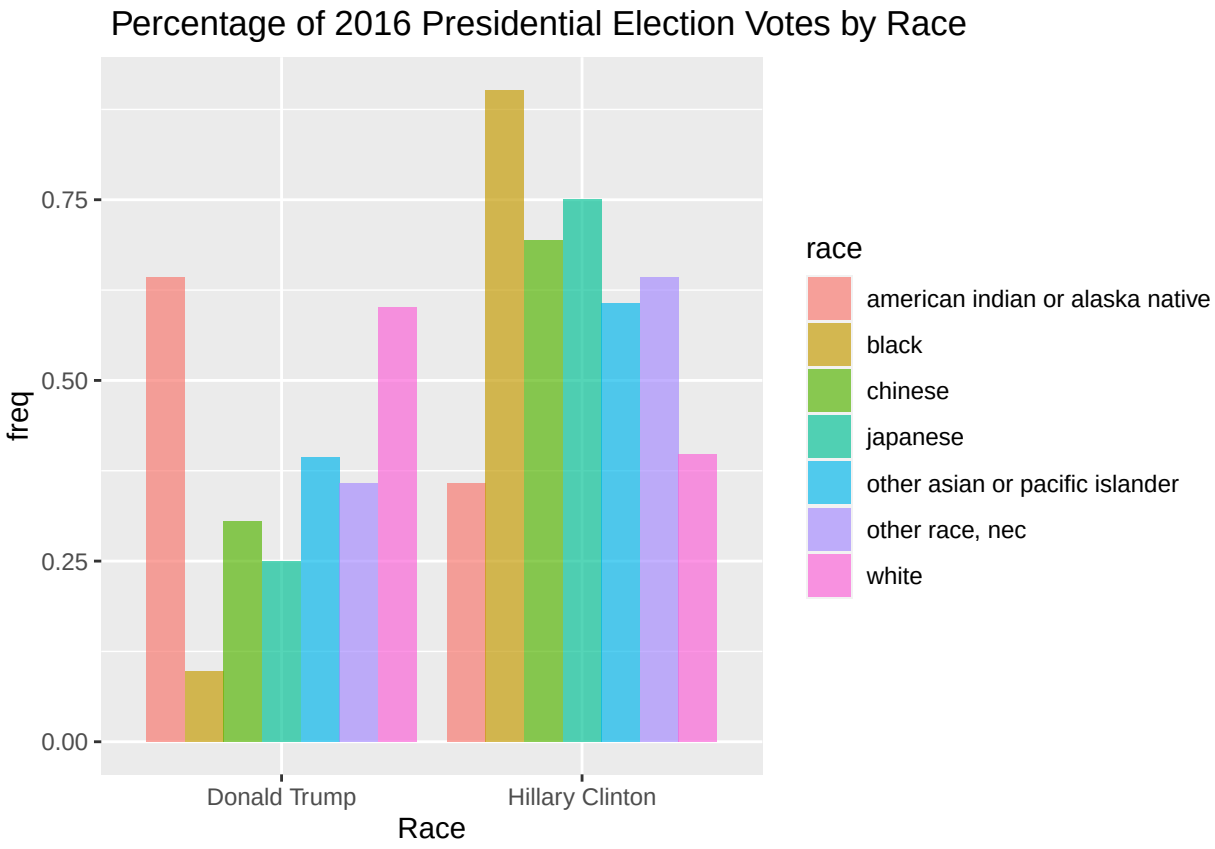


Figure 2: Votes by Race in Nationscape data

Race: (Figure **??**) shows the distribution of votes in the 2016 federal election; specifically the Democratic candidate, Hillary Clinton, and the Republican candidate, Donald Trump. We can see that certain groups had a significant different in their voting preference. for example, majority of Black voters voted for Clinton rather than Trump, as well as the population of Chinese voters.

State: When forecasting the election outcome, we will be predicting by State to see the majority vote in each. Certain states have been dominated by one particular party. For example, the Democratic party hasn't won the majority vote in the presidential election since 1992 \cite\{\{@cite270towin}. Since the election is based on electoral colleges as well, the number of Sates won by a candidate should give us a clear prospect for the dominating candidate.

Post-stratification Data

For our next step in our analysis, we use post-stratification to overcome any bias and have data representative of our target population. Conducted on a yearly basis, the American Community Survey "is a nationwide survey that collects and produces information on social, economic, housing, and demographic characteristics" of the United States population. In order to do so, the United States Census Bureau "mails questionnaires to approximately 295,000 addresses a month across the [country]." It should be noted that all the addresses receiving a letter to participate in the ACS are from a random sample chosen by the Census Bureau each month. Furthermore, every single address in the United States has approximately a 1 in 480 chance of being randomly selected to participate in the ACS during a given month. The exception to this comes from addresses that have already been selected to participate in the last 5 years; they are not allowed to be selected for another monthly sample. To access the 2018 1-year ACS dataset, the first step is to visit https://usa.ipums.org/usa/ and create an account. From there, IPUMS will send you an email to verify your email address. By clicking the link contained in the email, you will be redirected to the IPUMS website and your email will be verified. Once on the website, you will be able to access the page where ACS sample and variables can be selected. To choose the 2018 1-year ACS as your sample, press "Select Samples". For this analysis, the single 2018 sample was chosen. The website allows for the user to choose the sample, as well as any variables of interest. After selecting the sample, you will be taken to a page wherein variable categories are listed. In addition, there is also an option to search for variables of interest. The variables that we selected for the dataset used in this paper are: region, stateicp, hhincome, sex, age, marst, race, hispan, bpl, citizen, educ, labforce, and ftotinc. When finished selecting variables, click "View Cart" near the top right of the page, and you will arrive at the cart page. This is your chance to look over all of the chosen variables. If you are happy with them, press the blue button near the middle of the page that reads "Create Data Extract." On the new page that you arrive at, there is a blue button near the bottom of the page with "Customize Sample Sizes" displayed in it. If you wish to change the size of your sample, then click the button. Now that our dataset has been fully customized, press the blue "Submit Extract" button at the very bottom of the page. Finally, you have reached the page from which your dataset can be downloaded, and the final step is to wait for the sample to be ready.

Regarding the survey methodology employed by the Census Bureau for the ACS, they do not find people to survey, but instead find households. The target population of the ACS is the population of the United States, and the sampling frame consists of all Americans living at a specific address. The sample itself includes all Americans living at one of the 295,000 addresses that are selected to participate in the ACS during a given month. Furthermore, the Census Bureau designs the sample in a way that guarantees favourable geographic inclusion, while also utilizing random sampling to select the "addresses to be included in the ACS." Every resident of the addresses which the invitation to complete the ACS was mailed is required to complete either the online survey, or "to mail the completed paper questionnaire." If neither requirement is satisfied after about a month, the Census Bureau "will mail an additional paper survey questionnaire." If the second paper questionnaire goes uncompleted too, then roughly 1 month after it was sent, someone will show up at your address to conduct an in-person interview with you. Due to the fact that completing the ACS is the law in the United States, non-response does not appear to be a very big issue. In terms of the ACS sampling approach, there are definitely some trade-offs made. For instance, although random sampling can cause sampling bias in the data, this concern is outweighed by the benefits of the sampling approach used. A few obvious benefits of the sampling approach that come to mind are a lower cost of administering the ACS, as well as a widespread geographic representation in the data.

In addition to the strength of the ACS dataset with respect to its customizability, another is the timeliness of the data. To be precise, survey data was continuously collected throughout 2018, and the results were shared in 2019. This is very important as it means there is opportunity to gain a more thorough understanding of current characteristics in the United States than has ever before been possible. One of the only obstacles in achieving this comes from the main weakness of the ACS dataset–poor data quality. For example, there are numerous duplicate records and inappropriate entries contained in the data. Some potential explanations for the low standard of data are measurement error, processing error, and/or adjustment error. More explicitly, measurement error is a possible reason for the bad data because when American citizens are selected to complete the ACS survey, they are required to complete it by law. Although the survey can be completed online, or the completed questionnaire can be mailed back, for someone who never wanted to complete the survey to begin with, it would not be a big surprise for them to quickly record poor quality answers and

submit it. Processing error could be contained in the data as a result of either an error in coding with respect to the online survey, or an error in data entry by the person inputting completed paper questionnaires to the computer. Lastly, adjustment error serves as a possible explanation for the low quality data since any adjustments made to past survey results could have had an adverse affect.

3 Model

In order to forecast the popular vote of the U.S 2020 elections, multilevel logistic regression with post-stratification (MRP) was used. MRP is useful for when generalizing from a possibly non-representative poll **?**. The individual survey gives sample data on voters of the general U.S population, and post-stratification allows us to re-weight estimates, adjusting bias between our sample and the target population, so that we have a representative sample of likely voters **?**. To achieve this, cells are constructed using variables in the cleaned Nationscape survey such as age, household income, and race. The model is then trained on the survey using the proportions given by the ACS post-stratification data set. This approach gives us an advantage when attempted to forecast voting, as we can use a broad survey to speak to subsets in the population, and also tends to be less expensive to collect than non-probability samples **?**. However, using this approach places limitations on how much we can interpret from the estimates. Although we can estimate voting intention in different demographic groups, it doesn't tell us any qualitative information on voting patterns. For example, we might be able to see how different age categories vote, but we won't have a measure of the policy preferences, or their views on the candidates.

Using the Nationscape data, a multilevel logistic regression model can be fitted to the survey data set to model the proportion of voters who will vote for Trump. Logistic regression can be used to model a binary dependent variable, which is the choice between the Republic candidate, Donald Trump, and the Democratic candidate, Joe Biden in our case. This would be the main reason for choosing this model, over another model such as a linear regression model, where the output isn't necessarily binary. However, this imposes limitations. As we cannot take into consideration any other candidates running in the election due to the binary outcome. The variable 'vote_2020' in the data set was used as the dependent variable (the variable we are trying to predict), with the variable returning 1 for voters intending to vote for Trump, and 0 if they are voting for Biden.

Our model uses the variables pertaining to age, state, race, sex, household income, and whether a voter is Hispanic to predict whether a respondent will vote for Trump or Biden in the elections. The model takes the form:

$$Pr(y_i = 1) = logit^{-}1(X_i * \beta)$$

Equation from: **?**

where $y_i = 1$ represents a voter voting for Trump in the election. $Pr(y_i = 1)$ is the probability of a voter choosing to vote for Trump. The $X_i\beta$ are the linear predictors, where β represents the fitted coefficients for each independent variable X_i in the model. The coefficient values signify the mean variable changes in a one unit shift in the given variable β.

The logit function $logit(x) = log(x/1-x)$ maps the range (0, 1) to $(-\infty, \infty)$. Its inverse function $logit^{-}1(x) = e^x/(1 + e^x)$ maps back to the unit range. The model's output is bounded between 0 and 1, mapping the outputs into a binary outcome. The output tells us the probability that some person i, will vote for Trump depends on their age, sex, state, household income, and whether the voter is Hispanic.

We fit our regression model using the `glm()` function in [`R`]. We can see from table 1, which was made using `kable` from `knitr`, that the fitted model results in the following coefficients:

In logistic regression, collinearity implies that we have predictor variables that are highly correlated; that is, they have a linear relationship. This is possible when you have a larger number of variables in a model, but having collinearity in multiple variables can lead to unstable estimates and inaccurate variances, which can affect confidence intervals. This could lead to incorrect inferences about the relationship between our response variable and explanatory variables. Thus, it is imperative to check for any correlated variables in

Table 1: Table 1: Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	1.52	0.71	2.15	0.03
sexmale	0.40	0.07	5.45	0.00
age25-44	0.35	0.20	1.73	0.08
age45-64	0.46	0.21	2.22	0.03
age65+	0.38	0.21	1.81	0.07
raceblack	-3.13	0.45	-6.98	0.00
racechinese	-2.29	0.59	-3.87	0.00
racejapanese	-1.97	0.79	-2.50	0.01
raceother asian or pacific islander	-1.38	0.47	-2.95	0.00
raceother race, nec	-1.07	0.44	-2.42	0.02
racewhite	-0.55	0.41	-1.33	0.18
hispanmexican	-1.86	0.57	-3.26	0.00
hispannot hispanic	-1.22	0.55	-2.22	0.03
hispanother	-1.45	0.58	-2.51	0.01
hispanpuerto rican	-1.73	1.47	-1.18	0.24

Table 2: Checking Model Collinearity

	GVIF	Df	GVIF^(1/(2*Df))
sex	1.08	1	1.04
age	1.24	3	1.04
race	1.64	6	1.04
household_income	1.81	23	1.01
hispan	1.43	4	1.05
state	2.24	50	1.01

our model to avoid these issues. The value of the Variance Inflation Factor(VIF) in for the explanatory variables can give us a measure of collinearity our model. VIF values should be less than 5 to guarantee that collinearity is not an issue. We can see in table 2 below that all VIF values are less than 5, so collinearity will not be a concern for our model.

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  final_model$y, fitted(final_model)
## X-squared = 3.5112, df = 8, p-value = 0.8983
```

Before the model can be used for any predictions, we must also check that our model fits our data well, and meets any assumptions made by the model. The Hosmer-Lemeshow goodness of fit test assesses whether or not the observed events match expected events in subgroups of the model. For this test, the p-value can be used to assess goodness of fit. A smaller p-value indicates strong evidence against a good fit. In the case of our model, we can see in table 4 that we have a very large p-value of 0.9, indicating no evidence of poor fit.

Table 4: Hosmer and Lemeshow goodness of fit (GOF) test

X-squared	df	p-value
3.5	8	0.9

```
## # A tibble: 7,131 x 8
##    state_name  sex   age   race          hispan  num_records    prop state_count
##    <fct>       <fct> <chr> <chr>         <fct>         <int>   <dbl>       <int>
##  1 connecticut male  18-24 american ind~ not hi~           1 4.00e-5       24986
##  2 connecticut male  18-24 american ind~ puerto~           2 8.00e-5       24986
##  3 connecticut male  18-24 american ind~ other             1 4.00e-5       24986
##  4 connecticut male  18-24 black         not hi~          89 3.56e-3       24986
```

4.1 comparison on age

Looking at (Figure **??**), we observe a shift in age distributions between the two datasets. There is a slight peak in the 25-44 age group in the nationscape data. We see this flatten and then observe a significant increase in the 65+ age group, and a minor increase in the 18-24 age groups.

4.2 comparison on race

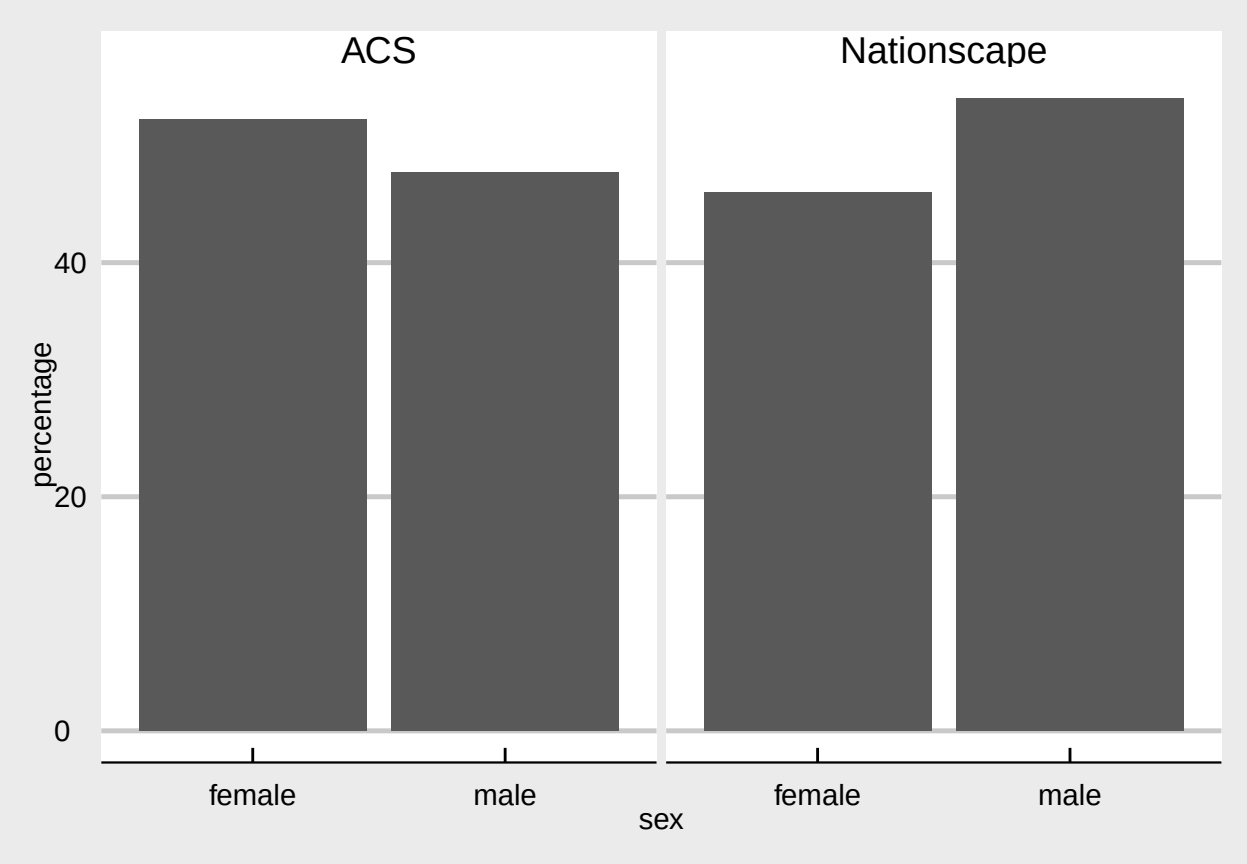
```
individual_sex_dist<- individual_survey %>%
  group_by(sex) %>%
  summarise(num_records = n()) %>%
  mutate(percentage = 100 * num_records/sum(num_records),
         total_num = sum(num_records),
         data_source = "Nationscape")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
post_strat_sex_dist <- post_strat %>%
  group_by(sex) %>%
  summarise(num_records = n()) %>%
  mutate(percentage = 100 * num_records/sum(num_records),
         total_num = sum(num_records),
         data_source = "ACS")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
individual_sex_dist %>%
  union(post_strat_sex_dist) %>%
  ggplot(aes(x = sex, y = percentage)) +
  geom_col() +
  facet_grid(~data_source) +
  theme_economist_white()
```




```
#figure to see 2020 voting by sex
vote_2020_sex <- individual_survey %>%
  mutate(vote_2020 = if_else(vote_2020 == 1,"Donald Trump","Joe Biden")) %>%
  group_by(sex,vote_2020) %>%
  tally() %>%
  ggplot(aes(x=sex, y = n)) +
  geom_col() +
  facet_grid(~vote_2020)


#figure to see 2016 voting by sex
vote_2016_sex <- individual_survey %>%
  group_by(sex,vote_2016) %>%
  tally() %>%
  ggplot(aes(x=sex, y = n)) +
  geom_col() +
  facet_grid(~vote_2016)



library(patchwork)
vote_2020_sex + vote_2016_sex
```

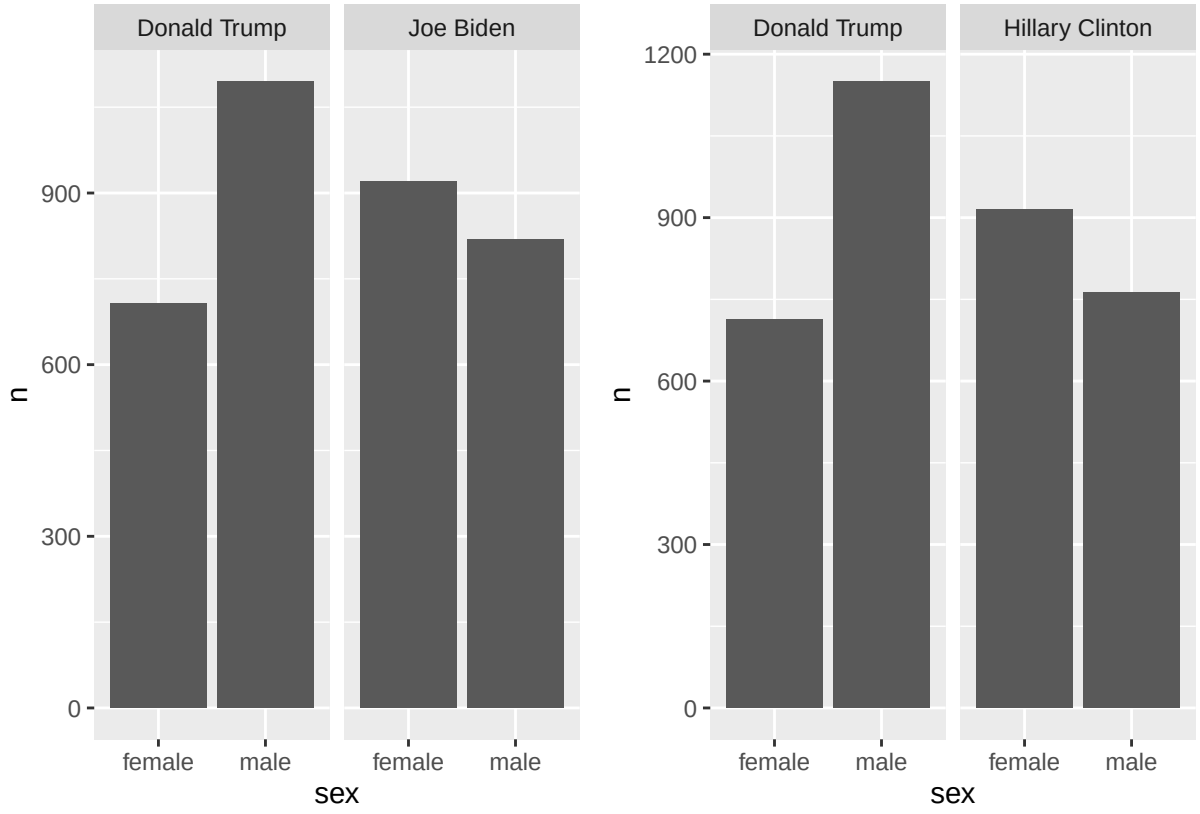


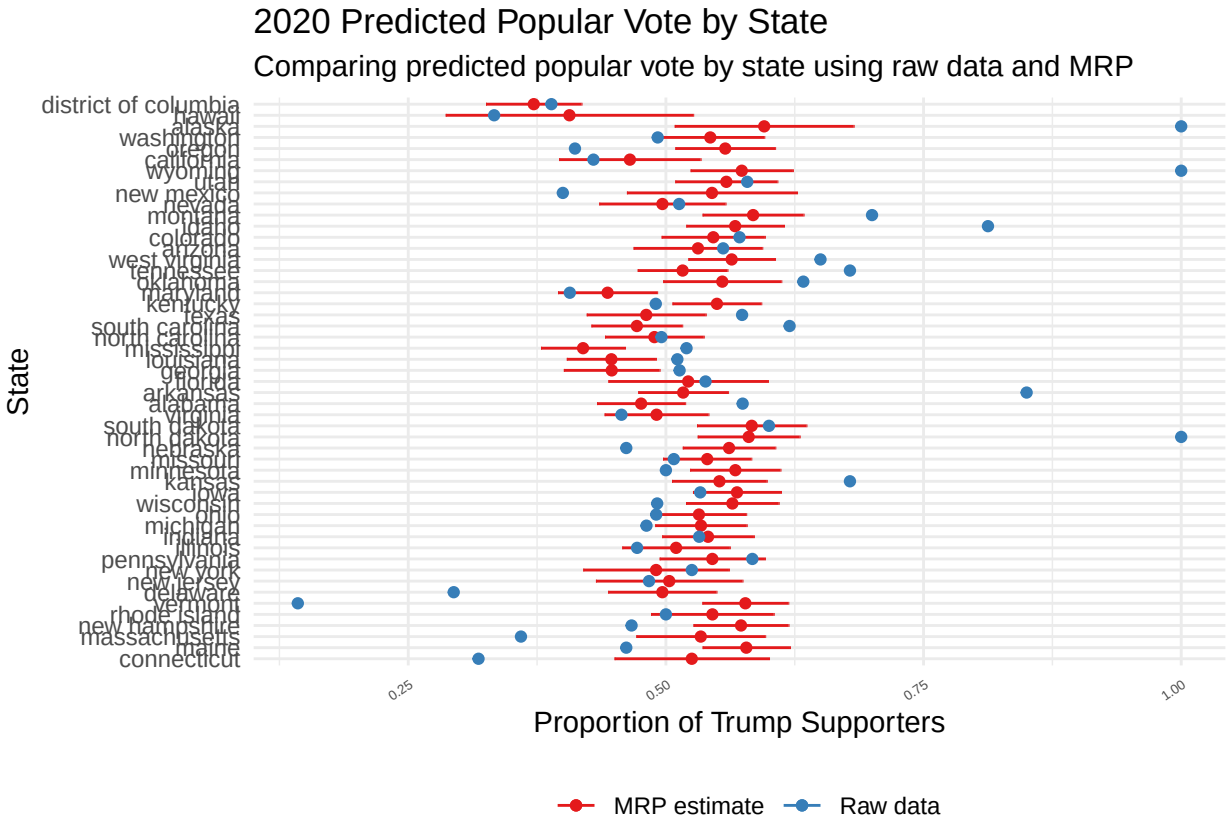
Figure (1) indicates that a man is much more likely to vote for Trump than a woman is. This tren

(1)

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

Table 4: Outcome of election

election_status	num_states
loss	24
toss-up	17
win	10




```
## Warning in validate_states(state_data, state_col, merge.x): Found invalid state
## values: District Of Columbia
```

Figure(**??**) shows the predictions of our model on the post-stratified dataset. This figure ignores close calls and has a binary Trump or Biden result.

```
post_strat_estimates %>% group_by(conservative_election_status) %>% summarise(num_states = n()) %>% rena
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

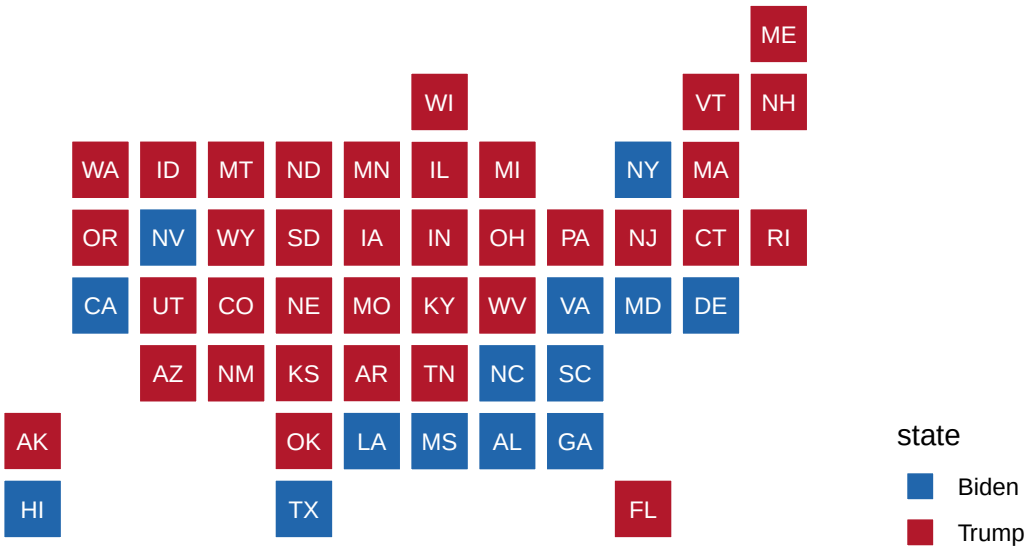


Figure 3: Heatmap showcasing model predictions at a state-level

4.3 comparison on age

4.4 comparison on race

4.5

5 Discussion

5.1 First discussion point

Talk about correlations between education and result

To begin this discussion, we begin with a quick explanation of the US Federal voting system. This context is important for us to understand some key points and drawbacks with interpreting our model.

Talk about previous research that has been done with similar models and the output of this model. Our model seems to indicate that Biden is going to win the popular vote. Throw some crap in here about 538 has certain polls that show a slight trump win. Show falling trump approvals and other things like that. This model has been thoroughly researched and has shown to be effective in multiple other contexts(needs citation)

One of the key things we have to explain here is how the electoral college works. Readers may remember the 2016 US election where Hillary Clinton won the popular vote, however, this didn't mean that she won the federal election. This was due to her not winning the electoral college. The electoral college is a body that is elected to determine who will be the next president of the US. This happened in 2000 when Bush didn't win the popular vote, but did end up winning the electoral college ## Second discussion point

5.2 Third discussion point

5.3 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

6 References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [URL].

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