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

First sentence, second sentence, third sentence, fourth sentence

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

1 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.

2 Introduction

3 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."

TODO: To make predictions on the outcome of the 2020 US presidential election, we used... 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.

3.1 Individual-level survey dataset

3.1.1 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.

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

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 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).

3.1.2 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 distribution of each of our cleaned variables (with the exception of gender and state) are shown in the following two pages.

Figure **??** Out of the respondents, approximately 54% were female and 46% were male (Figure 2). Most of the respondents were in their 60's and 70's, while the next most common demographic were respondents in their 40's and 50's (Figure 1). This preliminary look at the dataset is fairly consistent with Canadian demographics according to the 2016 Census, with the female response being approximately 3% higher than expected and the average age being approximately 11 years older than expected (the average Canadian age is 41 while the average respondent age was 52). This older demographic makes sense as individuals less than 15 years of age were not eligible to respond to the survey and are therefore not represented here.

Of the 20,602 responses, 194 rows were dropped if the value for a variable was not available. For our purposes of attempting to model mental health, we consequently removed these individuals from the dataset we used in generating our model. Furthermore, according to Figure 5, the responses are heavily skewed towards positive responses, with 30% of respondents replying with 'Excellent' and 34% replying 'Very good'. 28% rated their mental health as 'Good' with the remaining 8% split 6 to 2 with regards to 'Fair' and 'Poor', respectively. These results overwhelmingly indicate that a large proportion of the sampled population feel that their mental is very strong. However, we proceed with modeling in the next section of this paper to better understand the contribution of the chosen demographic and family factors on self-rated mental. Is there a pattern of traits that separate "Excellent", "Very good", and "Good" ratings? What are the biggest distinctions between an individual with good mental health and poor mental health? These are some of the motivating questions we strive to answer with our model.

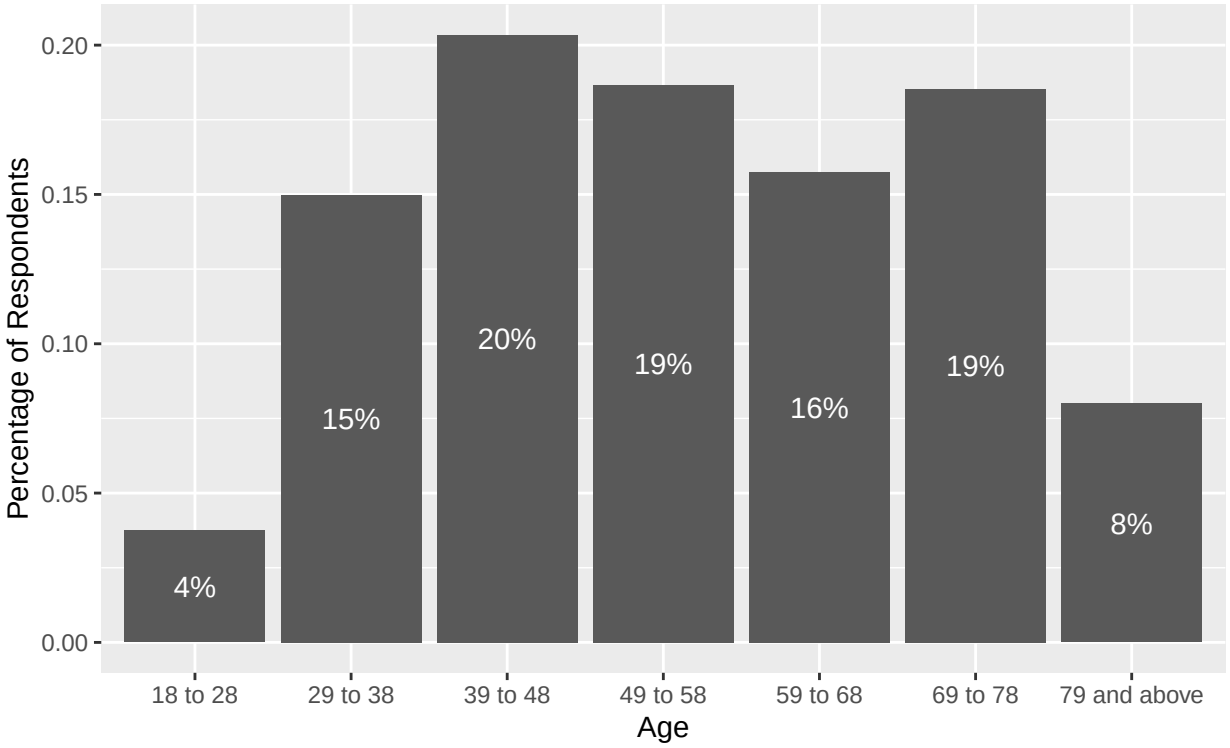


Figure 1: Distribution of the age of respondents in percentages.

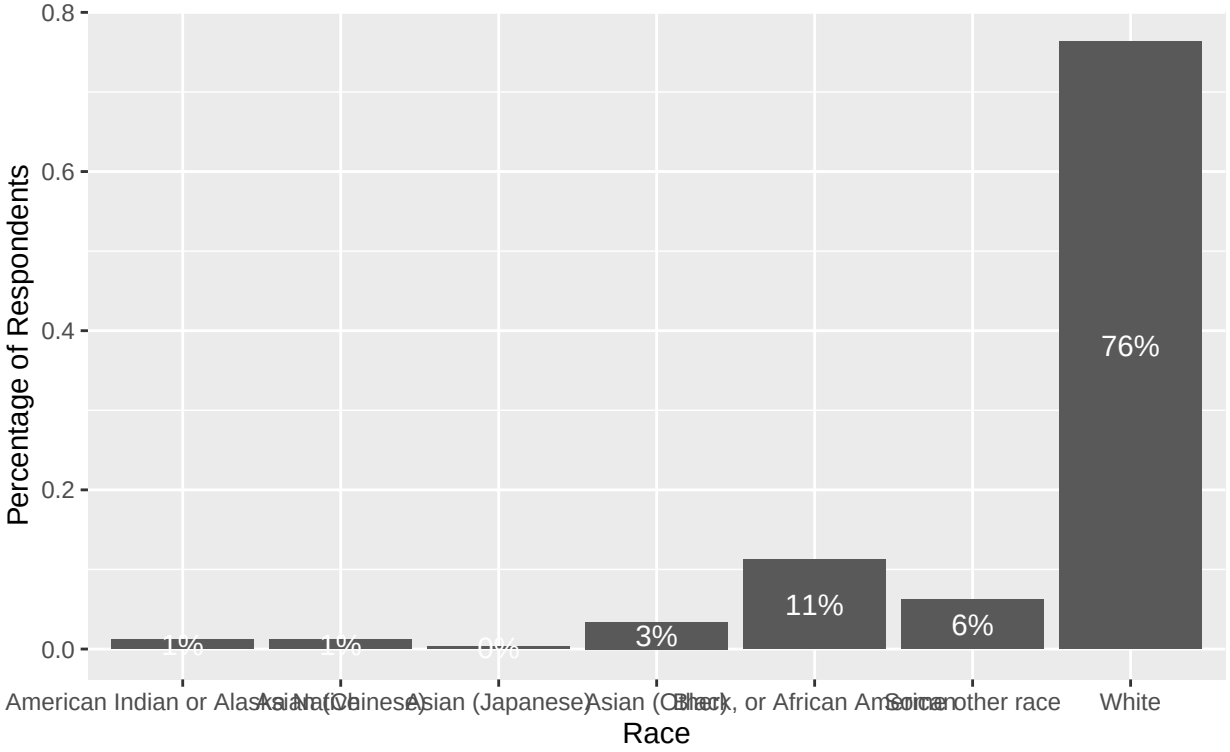


Figure 2: Distribution of the race of respondents in percentages.

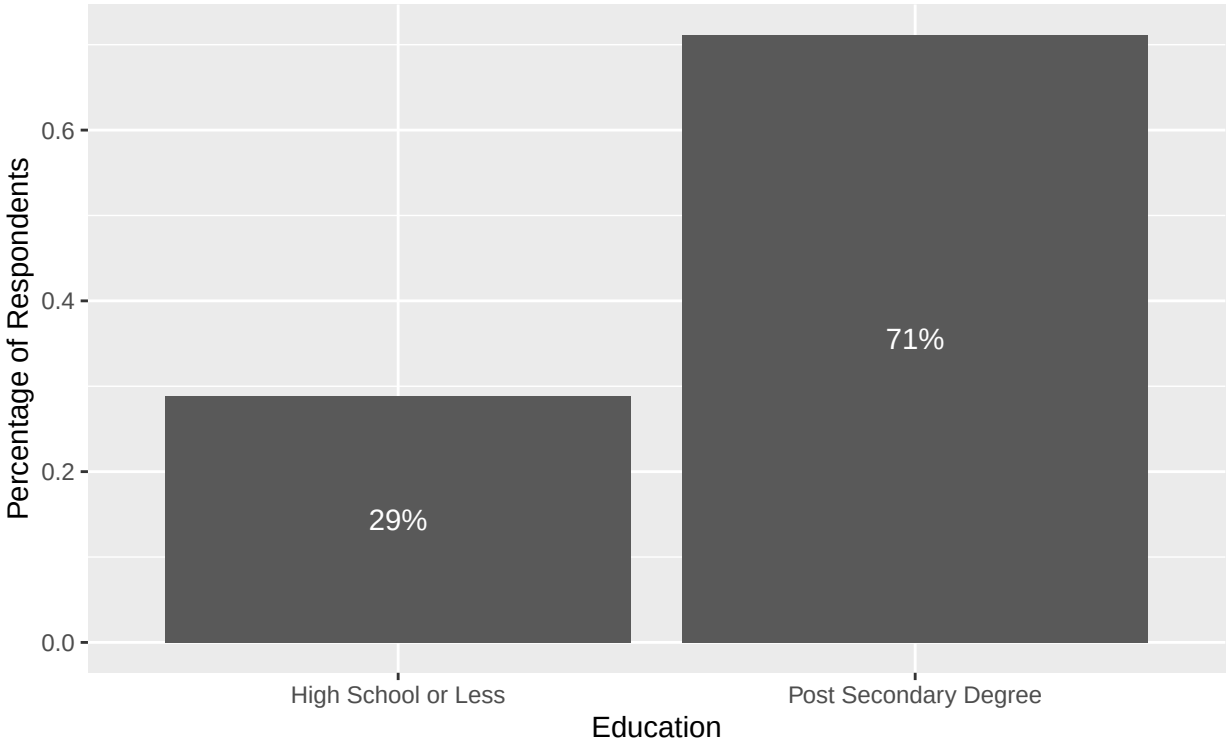


Figure 3: Distribution of the education of respondents in percentages.

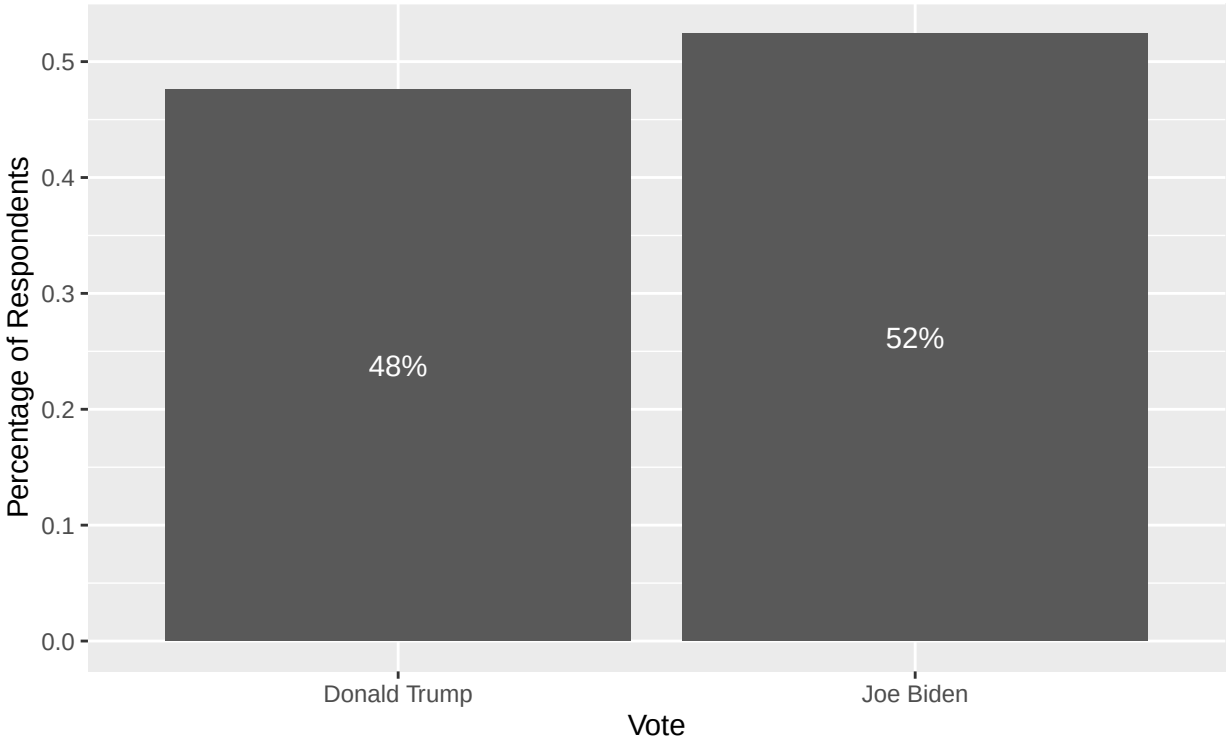


Figure 4: Distribution of the vote of respondents in percentages.

3.2 Post-stratification dataset

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.

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 "Group Quarters", 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% (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 that desired percentage. 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.

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), state (the state they lived in), gender (what gender they identified as), education (what their highest level of education was), and race/ethnicity of the respondent. 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 up 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. <– This section is to be updated

accordingly

```
# Plot Race
perc_race <- cleaned_data %>% count(race_ethnicity) %>% mutate(perc = n/nrow(cleaned_data))
race <- perc_race %>% ggplot(aes(x = race_ethnicity, y = perc)) + geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) + labs(title = "Race/Ethnicity
                                                                                x = "Race/Ethnicity", y =
race

# Plot Gender
perc_gender <- cleaned_data %>% count(gender) %>% mutate(perc = n/nrow(cleaned_data))
gender <- perc_gender %>% ggplot(aes(x = gender, y = perc)) + geom_bar(stat = "identity") +
  labs(title = "Gender of Respondents in 2018 ACS data", x = "Gender", y = "Percentage", subtitle = "Fi
gender

# Plot education
perc_education <- cleaned_data %>% count(education) %>% mutate(perc = n/nrow(cleaned_data))
perc_education$education <- perc_education$education %>% factor(levels = c("High School or less", "Post
education <- perc_education %>% ggplot(aes(x = education, y = perc)) + geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) + labs(title = "Education of Re
                                                                                x = "Education", y = "Pe
summary(cleaned_data$education)
education

# Explain 88 Olds are shit


# Plot state
perc_state <- cleaned_data %>% count(state) %>% mutate(perc = n/nrow(cleaned_data))
perc_state$state <- perc_state$state %>% factor(levels = sort(as.character.factor(perc_state$state)))
state <- perc_state %>% ggplot(aes(x = state, y = perc)) + geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) + labs(title = "State of Respon
                                                                                x = "State", y = "Percen
state

# Plot age_group
perc_age_group <- cleaned_data %>% count(age_group) %>% mutate(perc = n/nrow(cleaned_data))
#perc_age_group$age_group <- perc_age_group$age_group %>% factor(levels = sort(as.character.factor(perc
age_group <- perc_age_group %>% ggplot(aes(x = age_group, y = perc)) + geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) + labs(title = "Age of Responde
                                                                                x = "Age Group", y = "Pe
age_group
```

4 Model

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \tag{1}$$

Equation (1) seems useful, eh?

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)

5 Results

Figure 1: Distribution of votes by age

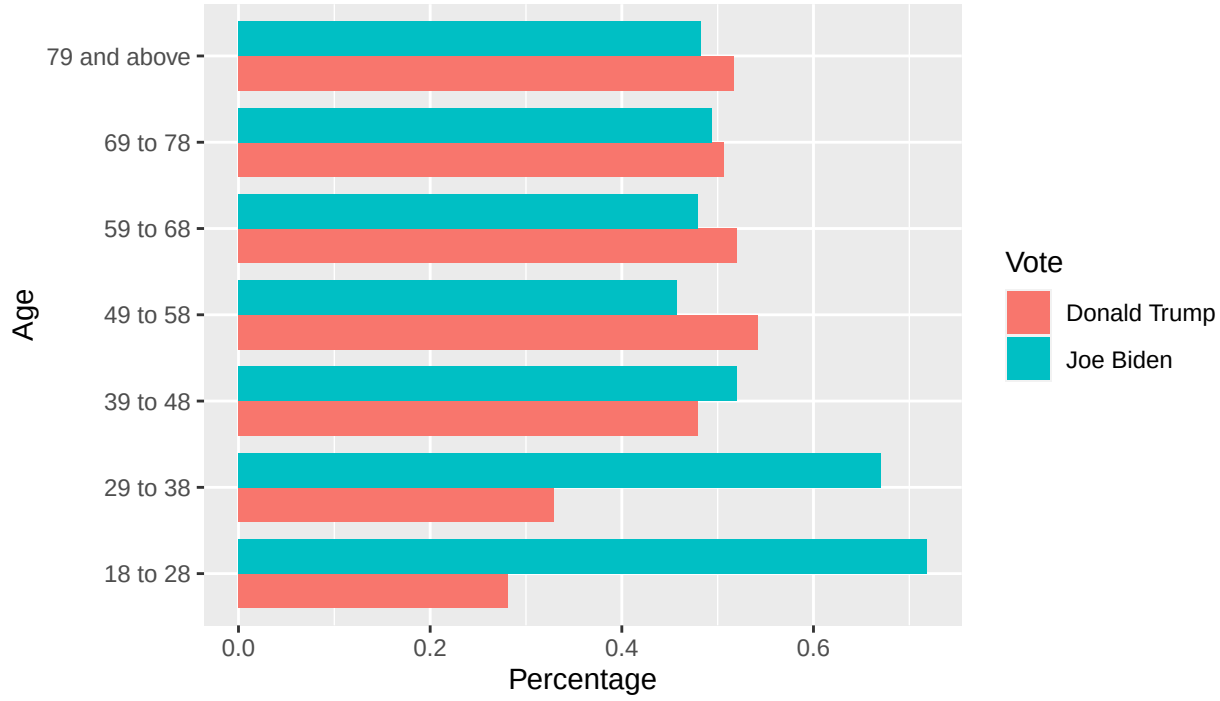


Figure 2: Distribution of votes by gender

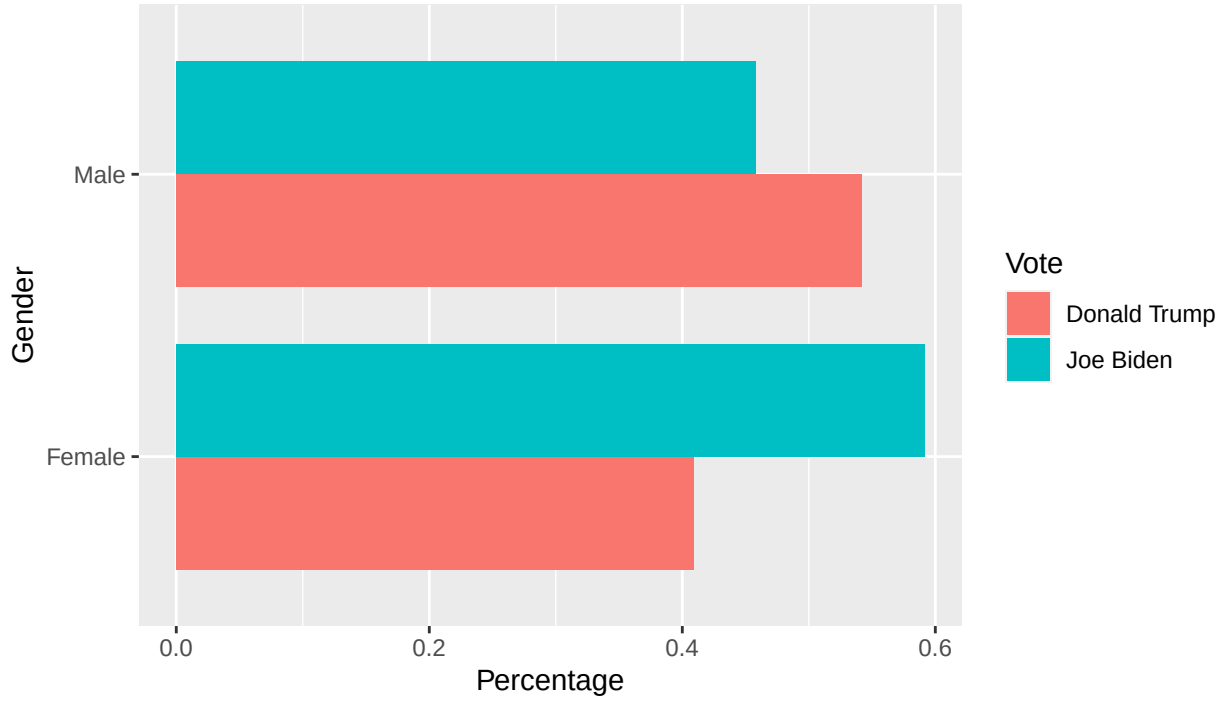


Figure 3: Distribution of votes by education level

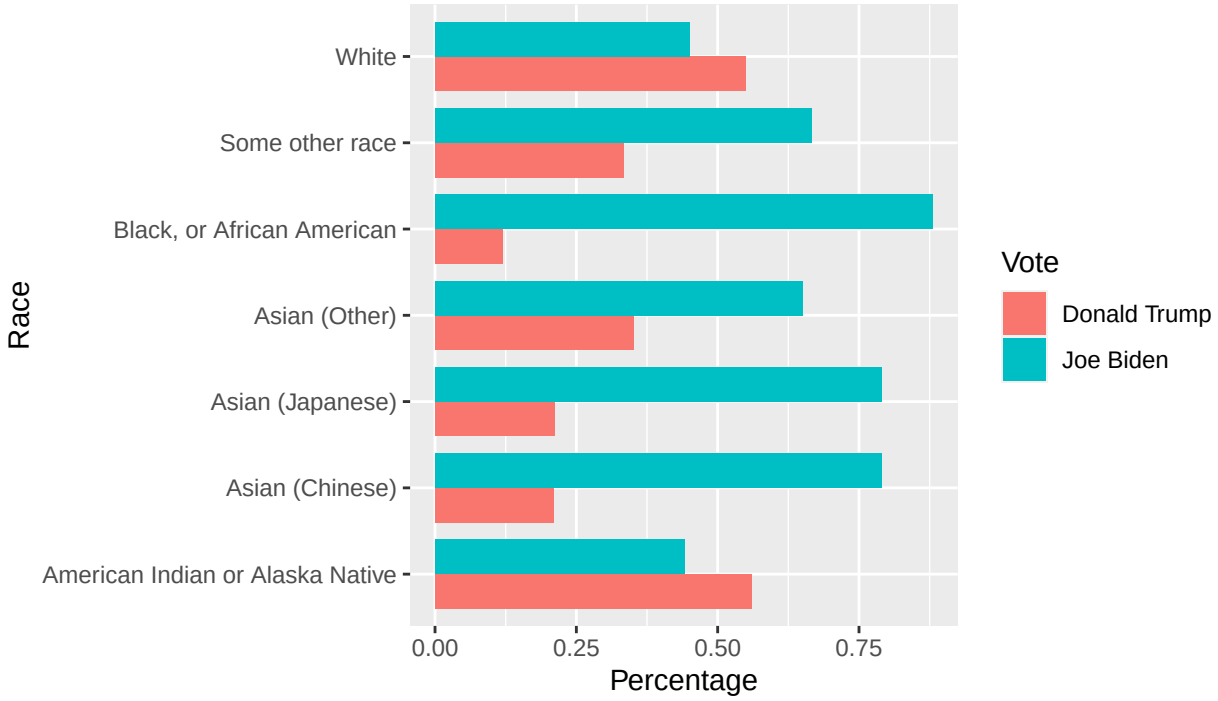
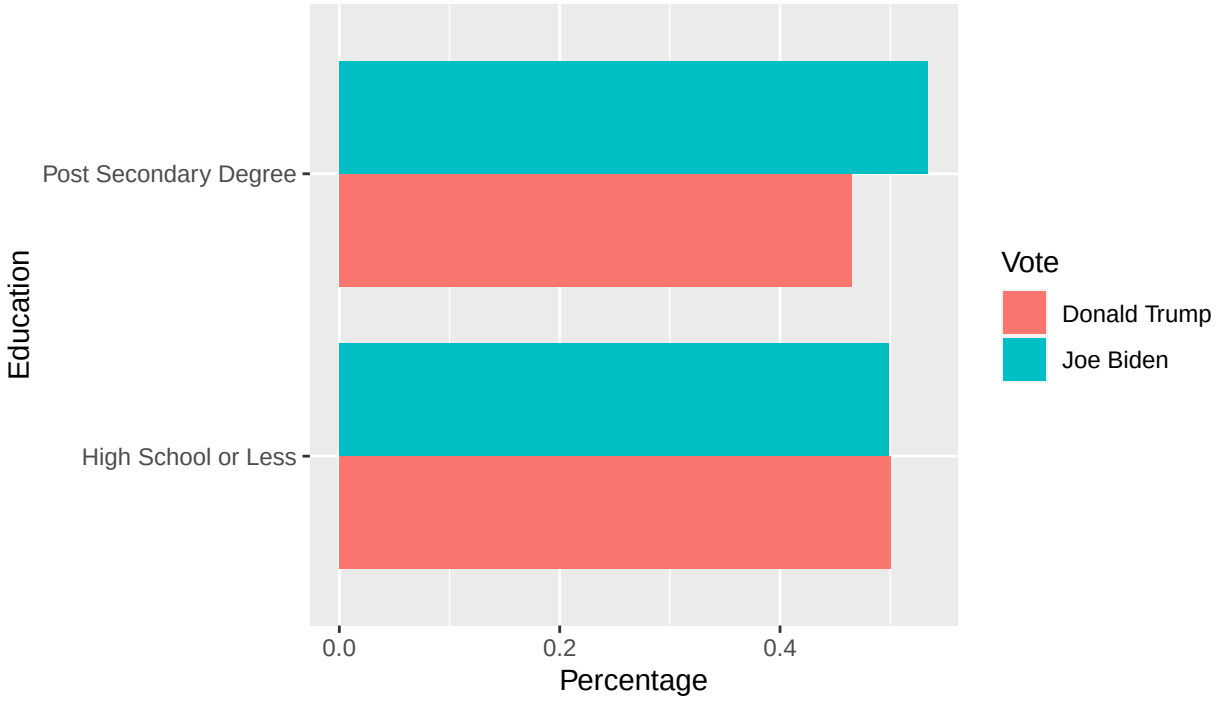


Figure 4: Distribution of votes by education level



Our data is of penguins (Figure 5).

Talk more about it.

Also bills and their average (Figure 6). (Notice how you can change the height and width so they don't take the whole page?)

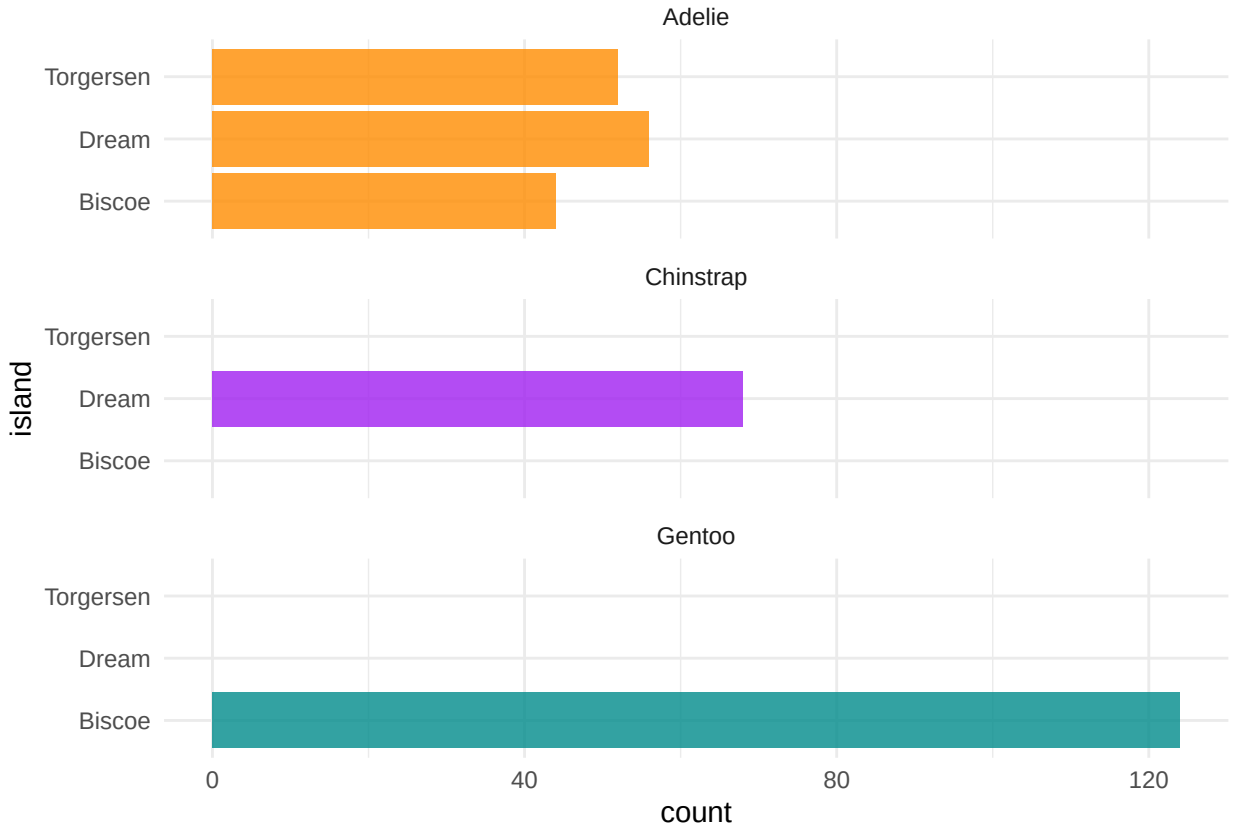


Figure 5: Bills of penguins

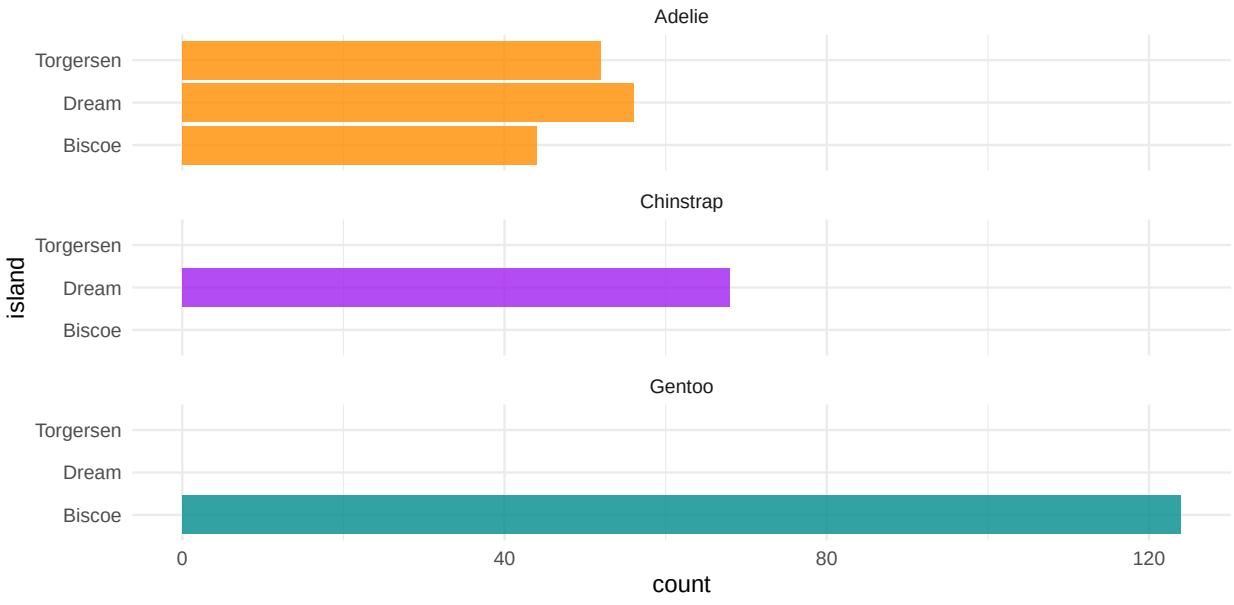


Figure 6: More bills of penguins

Talk way more about it. # Discussion

5.1 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

6 References

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/ Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (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.