Forecasting US Election: Trump or Biden[*]

TBD

Yi Su

30 October 2020

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

President election has been one of the most important political events in the United States of America which happens every 4 years. Most American citizens will be involved in this event and make their decision on which direction the county will go at least in the next 4 years. Meanwhile, the rest of the world will also keep their eye on the election because of the global political position of the US. The election is an invisible war between the parties of the US, among those parties, the Republican and the Democratic are the two oldest and dominant parties.

The Republican party and the Democratic party are the two dominant parties holding a large number of positions in congress. Throughout the history of the US, the competition between the president nominates of these two parties has never stopped. In 2016, Donald J. Trump won the election as the nominate of the Republican party and defeating his opponent from the Democratic party, Hilary Clinton. In 2020, Trump is the Republican nominate again and this time, his main opponent from the Democratic party is Joe Biden, a former vice president of the U.S. during 2009-2017. In general, these two nominates are more likely to win the election than nominates from other parties like the Green party.

In this report, we are interested in forecasting the 2020 U.S. election. Specifically, who is more likely to win among Trump and Biden? To do this, the support of a voter intent survey is essential, and we used the UCLA Nationscape survey dataset (Tausanovitch and. 2020) requested from the URL in reference. The discussion of this survey dataset will be included in *Section 2.1.* Among the many statistical techniques of forecasting election, we used multilevel regression and poststratification (MRP) to produce estimates of votes. The MRP involves partitioning the data into small cells based on demographic characteristics of our choice, then estimate voter intent (Trump or Biden) in the cell level using a multilevel regression model. To make the forecasts, we used a census-like dataset, the American Community Survey Dataset (Steven Ruggles and Sobek., n.d.). This is the dataset where partitioning into cells happens and we make predictions from here. The discussion on this dataset is included in *Section 2.2.*

The discussion on the specific procedure of this report and the multilevel regression model is included in *Section 3*, which we regressed the voting intent based on age, household income, race and states in the US. In *Section 4*, we present and discuss the resultant model and present estimates of the voting intent between Trump and Biden. Meanwhile, *Section 5* includes discussions of our forecast results as well as some weaknesses and future improvements on our procedure. This report was produced using `R` (R Core Team 2020), as well as some packages which will be mentioned in each section for usage.

The forecast results are divided into two main part, the first part is an overall support rate among the ACS dataset in favor of Trump. The second part is to calculate the mean support rate of Trump in each state and then calculate the number of electoral votes from each state that we forecast Trump to get.

[*] Code and data are available at: https://github.com/YiSu2000/US_Election_Yi_Su

2 Data

2.1 Democracy Fund + UCLA Nationscape Data

The Nationscape (Tausanovitch and. 2020) is a survey conducted from July 2019 to December 2020, collecting demographics of the respondent as well as their voting intent during the 2020 election. The survey samples are provided by Lucid, an online exchange platform focusing on market research. Specifically, the samples were drawn from the online platform based on a set of demographic quotas like age, region, and gender.

The Nationscape aimed at conducting 500,000 interviews in total and roughly 6,250 interviews per week. The survey took the form of an online survey using a survey software controlled by the Nationscape team, however, the respondents were sent to the software directly by the Lucid platform. Since only the respondent will only be directed to the survey software if they match on the Lucid platform, the non-response rate should be reasonably low. Although the quality of the responses is expected to be high from the Lucid platform, the representativeness of the population of interest still needs to be assessed. This is solved by comparing the Nationscape's results to the results of the Pew Research Center's evaluations of online non-probability samples in 2018. The Pew Research Center's 2018 report assessed how various choices impact the quality of the online survey.

After requesting the data on the June-25-2020 phase of the Nationscape survey dataset. We need to modify the original observation levels based on our needs. In this report, we are interested in modeling the voting intent between Trump and Biden by age, household income, race, and states of living. Only 4 variables were chosen because of some hardware limitations which will be discussed in *Section 3* in detail. The original survey data has race and income levels too narrow and thus might cause trouble in our regression.

First, we created a binary variable with 1 meaning vote for Trump and 0 meaning vote for Biden. This will be the response variable of our regression model. We deleted the observations that will vote for nominates other than Trump and Biden, and this caused a reduction in sample size for our model. Further discussion of the potential hazard of this reduction is included in *Section 5.4*.

Second, to make cells wider for a stable sample mean in the MRP process, we redefined some variables related to age, race, and household income. We reduced the household income levels from 24 to 6 wider income groups. For the race, the categories for races are reduced from 15 to 5 wider race groups. Similarly, age was redefined from a discrete variable to a categorical variable with 8 levels, each level representing a decade's group of age. For specific of the group names of each variables, refer to Table 2 which is frequency table that includes all variable group names except for the states. Since the state is an important factor in the election and we need individual state forecast, there will be no modification on it.

The decision on how many level or group there is for each explanatory variable is based on both the Nationscape dataset and the ACS dataset. For example, since household income is a categorical variable in the Nationscape dataset, we need to categorize the household income numeric variables in the ACS dataset so the two dataset has the exactly same variable categories for forecasting. Another good example is with race, since the ACS dataset groups Chinese and Japanese into one category, we need to group the Chinese and Japanese samples in the Nationscape dataset into one category as well.

Some example observations in the dataset are shown in Table 1). These modifications of the dataset were done with `R` (R Core Team 2020), through R packages `tidyverse` (Wickham et al. 2019), `naniar` (Tierney et al. 2020), `haven` (Wickham and Miller 2020) and `broom` (Robinson, Hayes, and Couch 2020). And Table 1) was created with `knitr` package (Xie 2020) and `kableExtra` package (Dowle and Srinivasan 2019).

Table 1: the first 6 rows of the dataset

Expected Vote in 2020		Income level	Race	Age group	State
Donald Trump	1	$70,000 to $99,999	White	Between 40 to 50	Wisconsin
Donald Trump	1	$175,000 to $249,999	White	Between 40 to 50	Virginia
Donald Trump	1	$35,000 to $69,999	White	Between 70 to 80	Texas
Donald Trump	1	Less than $35,000	White	Between 50 to 60	Washington
Joe Biden	0	$70,000 to $99,999	White	Between 20 to 30	Massachusetts
Joe Biden	0	Less than $35,000	Black/African American/Negro	Between 30 to 40	Texas

Next, we check the frequency of each group in our modified dataset to ensure variation in explanatory variables for the accuracy of the model. Table 2 displays the frequency of each group within each variable instead of plotting the distribution of each categorical variables. These frequencies tells us the distribution of each categorical variables and provides a more compact view to the problems within each variable. It is also a good display of all variable sub-group names. However, since there are 50 states, we only display the first 6 states in alphabetic order.

Although only 6 states are shown, the problem is clear. Some states like Alaska and Arkansas have a significantly fewer number of respondents than large states like California. The same situation happened in all other variables as well, White people have way more number of observations than Chinese or Japanese and Alaskan natives, the number of respondents above 80 is way less than others, and the number of respondents with higher household income decreases as income level increase.

Unfortunately, we can only redefine the age groups since the other variables groups need to line up exactly the same between the two datasets. And some variables are already at the base-line level and could not be further modified. We can choose to specify states into different regions, but that would obey one of our goal of forecasting election base on electoral votes which is highly dependent on winning in each state.

After we joined the age group of above 80 into the age group between 70 to 80. The remaining differences in other variables would be small enough to be compensated by our Bayesian multilevel modeling approach. This approach pools the effect of the minor groups with other more major group cells, which is very beneficial for the states with less than a hundred respondent observations since we cannot redefine state groups.

Table 2: Frequency of each group

Age Group	Count	Income level	Count	Race	Count	State	Count
Under 20	155	Less than $35,000	1648	American Indian or Alaska Native	55	Alabama	71
Between 20 to 30	687	$35,000 to $69,999	1303	Black/African American/Negro	529	Alaska	6
Between 30 to 40	966	$70,000 to $99,999	665	Chinese or Japanese	66	Arizona	122
Between 40 to 50	873	$175,000 to $249,999	228	Other Asian or Pacific Islander	158	Arkansas	31
Between 50 to 60	749	$100,000 to $174,999	797	Some other race	291	California	520
Between 60 to 70	884	More than $250,000	112	White	3654	Colorado	67
Between 70 to 80	387					Connecticut	54
Above 80	52					Delaware	20

Further concerns with this pick of variables are any underlying strong multicollinearity between age and income. We roughly check this by Figure 1, which was created with the `ggplot2` package (Wickham 2016) and `data.table` package (Zhu 2020). We do see an increasing income level trend with aging in general. However, the degree of multicollinearity is not sufficient to violate the assumption of no perfect multicollinearity between explanatory variables since the general proportion of incomes remains steady except for the age group between 20 to 30.

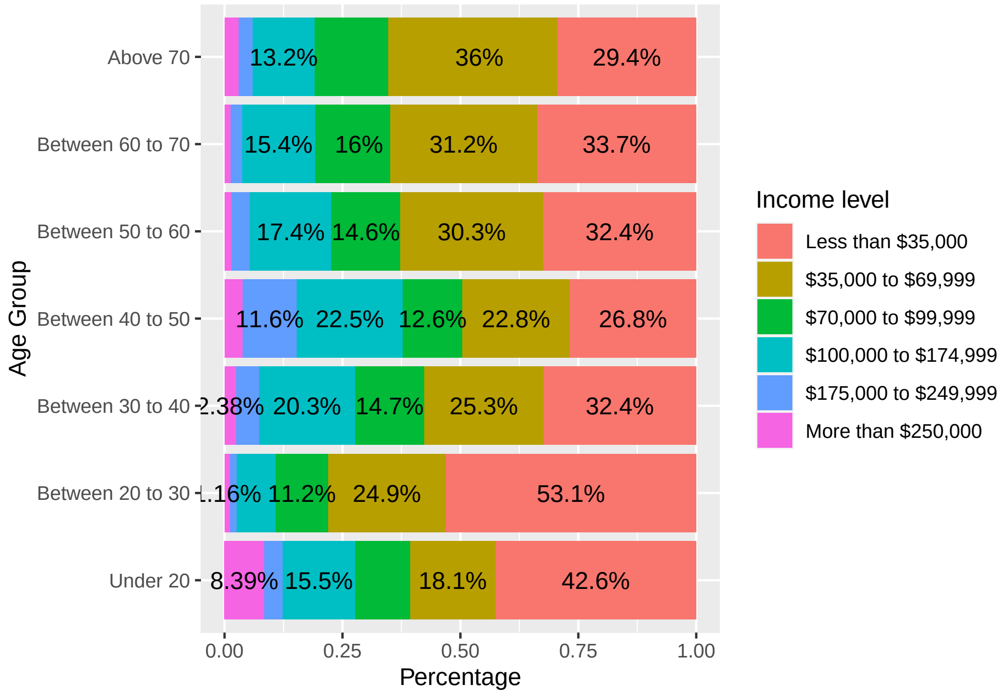

Figure 1: Percentage of income levels for age groups

2.2 American Community Surveys Data

The American Community Survey (ACS) (Steven Ruggles and Sobek., n.d.) are monthly surveys on rolling households that was designed as a substitute of a census. The target population of the ACS is all American households since it is designed to replace a census, and at full implementation, the sample would include about 3 million households across the US. The samples were extracted from the Census Bureau's larger internal data files, and thus shares the same sampling errors. For each sampled household, the ACS records the response from all member of the family. Privacy was protected by restricting geographic variables to state level and some of the individual variables are Top coded.

The ACS sample design took form of systematically sampling one household to represent each U.S. county each month. And that monthly sample will receive the ACS survey though mail at the beginning of the month. Non-respondents of the mail were contacted via telephone for a computer assisted telephone interview one month later. If the household is still not responding, one third of the non-respondents to the mail or telephone survey are contacted in person for a computer assisted personal interview one month following the last attempt. Since the ACS is sampling on a county level monthly, this makes it close to a census dataset. However, the initial attempt of survey through mailing may increase the initial non-response rate and decrease quality of response. Mailing back takes more effort as the respondent and mailing also have the uncertainty of lost mails during delivery. In the most extreme cases, the survey mail may be miss-classified as a junk mail and thus discarded. However, the latter attempts would likely to solve this issue.

The ACS dataset is as good as a census dataset and thus appropriate to our goal of forecasting for the whole nation's vote. For our purpose, we need to filter out all sample observations with age under 18 so the remaining all reaches the age to vote, and then assume all observation in the remaining dataset will vote between Trump and Biden. The next step is to create variable groups that are identical to ones we created in the Nationscape survey dataset. Ensuring the variables in both dataset lines up is how we could implement our model trained from the Nationscape dataset to the ACS dataset and create the estimated forecasts. After doing this, we have a modified ACS dataset that have the same variables as the modified Nationscape dataset but with a lot more observation. Thus, Table 1 can also be used as example observations of the modified ACS dataset.

Table 3 shows the frequency of each group within the explanatory variables in a similar fashion of Table 2 in the previous section. The reason why use a frequency table instead of plotting each variable is the same as for Table 2.

Table 3: Frequency of each group

Age Group	Count	Income level	Count	Race	Count	State	Count
Under 20	60441	Less than $35,000	498152	American Indian or Alaska Native	24013	Alabama	36288
Between 20 to 30	341839	$35,000 to $69,999	620021	Black/African American/Negro	213941	Alaska	4511
Between 30 to 40	372096	$70,000 to $99,999	419774	Chinese or Japanese	41604	Arizona	52176
Between 40 to 50	368195	$175,000 to $249,999	172624	Other Asian or Pacific Islander	95835	Arkansas	22373
Between 50 to 60	437794	$100,000 to $174,999	551253	Some other race	134711	California	284580
Between 60 to 70	432402	More than $250,000	149052	White	1900772	Colorado	42110
Above 70	398109					Connecticut	27132
						Delaware	7178

The distribution of the groups are similar to the Nationscape dataset, and the unequal distribution of frequency in each group still exist as expected. There still exist states that have significantly lower frequency comparing to larger states, the higher house income level groups still have lower frequency among the dataset. Chinese or Japanese and Alaska Native or American Indians still have a significantly lower share of the distribution comparing to the number of White people.

However, this similar pattern is a good sign which indicate that our regression model trained on the Nationscape dataset would be appropriate to use on the ACS dataset. This similarity also indicates the reliability of both dataset in term of response quality. Since the distribution of variables though the two survey match, we have more reason to believe that the two survey captured the true population distributions in general.

The next step is the poststratification part of the MRP approach. The first step is to sum up all the observations in the dataset sharing the same age group, income group, race and state. Then we can create new datasets with calculated proportion of such combination of variables with respect to the specific variable we are interested in. Table 4 shows one of such sub-dataset for proportion with respect to age groups. For example, for the first row, the proportion variable means that the proportion of American Indian or Alaska Native living in Alabama with income between $100,000 to $174,999 at age above 70 is only 0.00025% of all people with age above 70. This could also be done by calculating the proportion of that cell with respect to the whole ACS dataset samples, that will make the proportion be $n/sum(n)$. The former group wise proportion is useful when making predictions within each group, while the latter is useful when making a final prediction of proportion of Trump vote among all the samples in the ACS dataset.

Table 4: Proportion with respect to Age

Age Group	Race	State	Income Group	Count	Proportion
Above 70	American Indian or Alaska Native	Alabama	$100,000 to $174,999	1	2.50e-06
Above 70	American Indian or Alaska Native	Alabama	$35,000 to $69,999	4	1.00e-05
Above 70	American Indian or Alaska Native	Alabama	Less than $35,000	7	1.76e-05
Above 70	American Indian or Alaska Native	Alabama	More than $250,000	1	2.50e-06
Above 70	American Indian or Alaska Native	Alaska	$100,000 to $174,999	9	2.26e-05
Above 70	American Indian or Alaska Native	Alaska	$175,000 to $249,999	8	2.01e-05

3 Model

$$Pr(y_i = 1) = logit^{-1}(\alpha_{a[i]}^{age\,group} + \alpha_{a[i]}^{income\,group} + \alpha_{a[i]}^{race} + \alpha_{a[i]}^{state}) \tag{1}$$

Equation (1)

4 Results

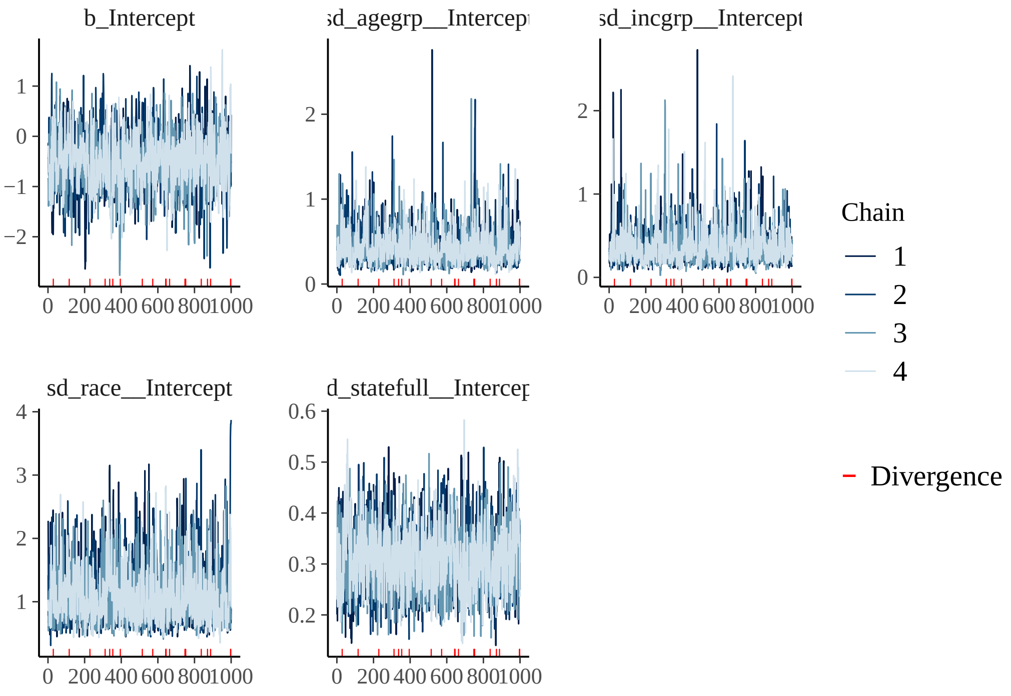

Figure 2: Convergence of Model Parameters

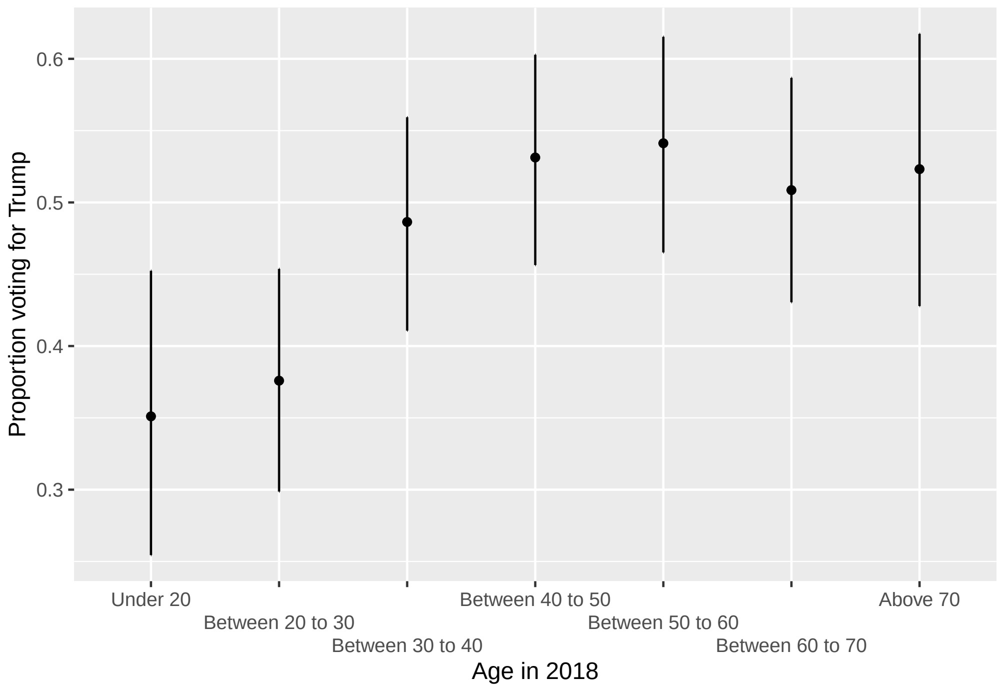

Figure 3: Forcasting Votes for Trump by Age Groups

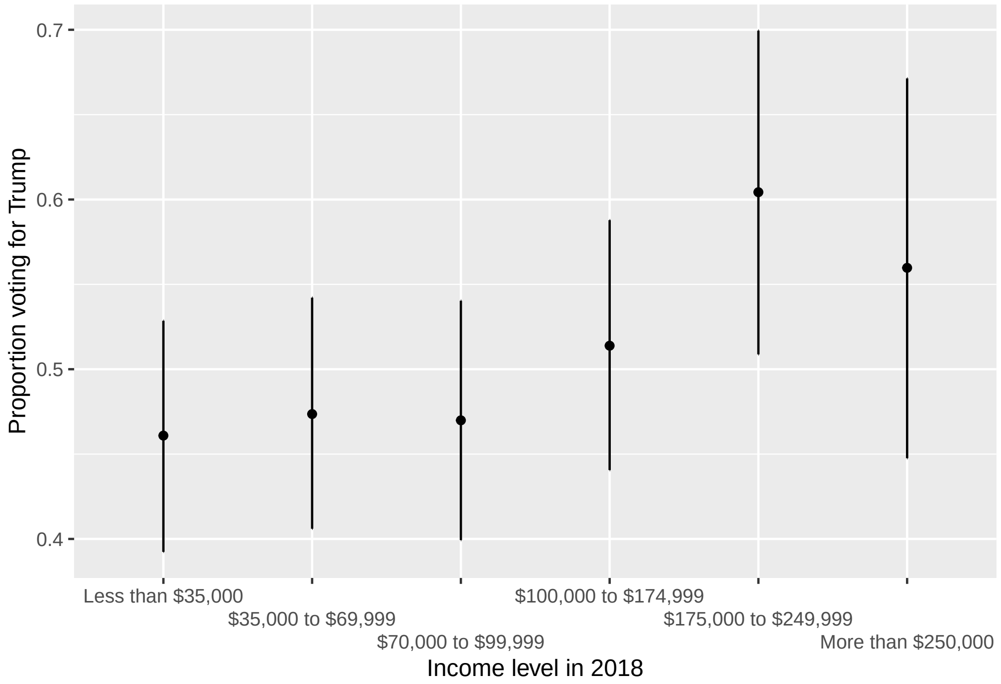

Figure 4: Forcasting Votes for Trump by Household Income Groups

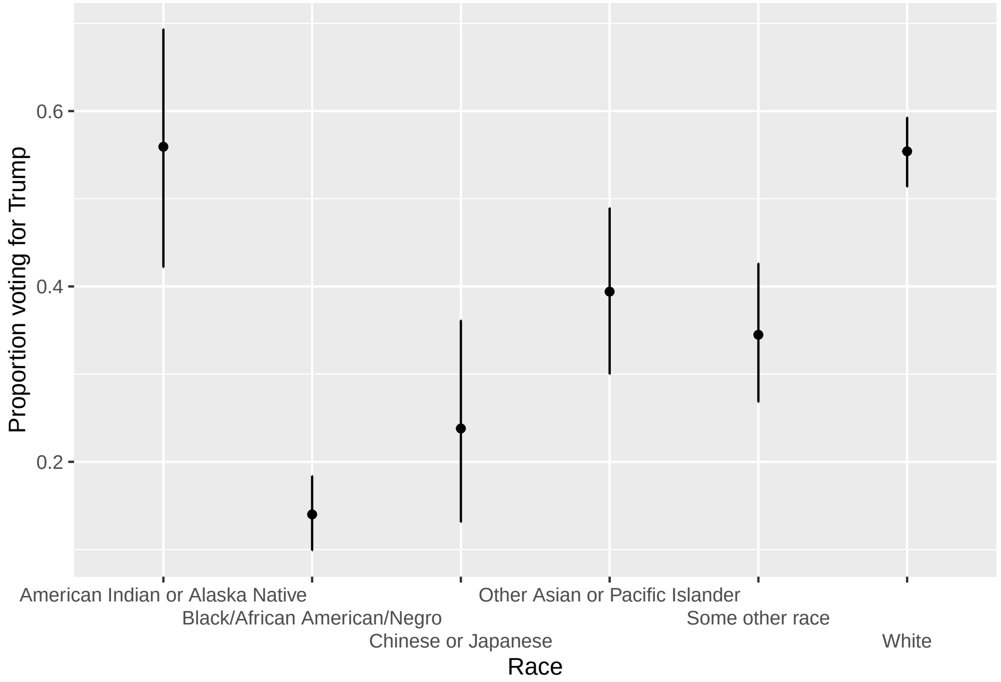

Figure 5: Forcasting Votes for Trump by Race

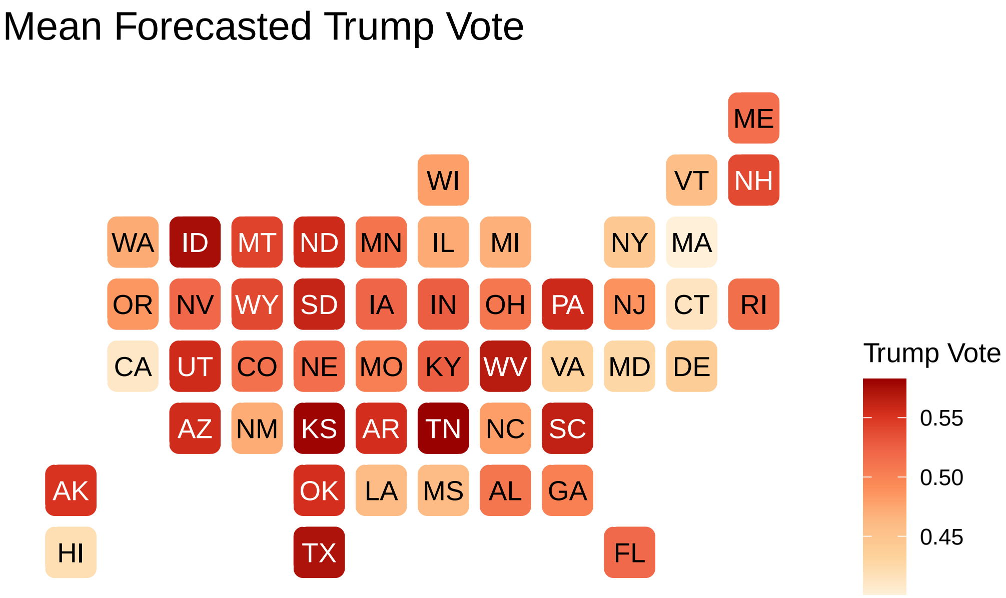

Figure 6: Forcasting Votes for Trump by States

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

References

Dowle, Matt, and Arun Srinivasan. 2019. *Data.table: Extension of 'Data.frame'.* https://CRAN.R-project.org/package=data.table.

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

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

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. n.d. "IPUMS Usa: Version 10.0 [Dataset]." Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and.Lynn Vavreck. 2020. "Emocracy Fund + Ucla Nationscape, October 10- 17, 2019 (Version 20200814)." https://www.voterstudygroup.org/publication/nationscape-data-set.

Tierney, Nicholas, Di Cook, Miles McBain, and Colin Fay. 2020. *Naniar: Data Structures, Summaries, and Visualisations for Missing Data.* https://CRAN.R-project.org/package=naniar.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

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.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files.* https://CRAN.R-project.org/package=haven.

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

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