TBD*

TBD

Arjun Dhatt, Benjamin Draskovic, Yiqu Ding, Gantavya Gupta

02 November 2020

1 Abstract

In this report, we will be forecasting the winner of the upcoming US Presidential Elections 2020 between incumbent Donald Trump of the Republican Party and Joe Biden of the Democratic Party, by looking at different factors that affect a citizen's voting intentions leading to Election Day. We cleaned the Nationscape dataset and chose the variables age, education, race, family income, geographic region, and mask-wearing decision which will form the basis of our Elections prediction. Then we used an MRP model (a combination of multilevel logistic regression and post-stratification) to analyse the factors and it also helps us in achieving a much narrower confidence interval for our prediction. After our analysis, we concluded that Joe Biden will win the overall popular vote of the 2020 American presidential election.

2 Keywords

- Forecasting
- US 2020 Election
- Multi-level Regression with Post-Stratification
- Trump
- Biden

3 Introduction

The US 2020 Election is considered by many to be one of the most crucial elections in recent history. American voters will be deciding to elect Democratic Leader, Joe Biden or Republican leader, Donald Trump. In this report, we will be forecasting the results of the US 2020 election using multi-level regression with post stratification.

We are interested in finding out who will win the overall popular vote in the US 2020 election between Donald Trump and Joe Biden. To begin, we started by cleaning the dataset provided by Nationscape. The data we are using was collected from June 25th, 2020 to July 1st, 2020. Using the dataset, we plan to closely analyze the following variables to use them as predictors in predicting the results of the 2020 US election: - X - X - X - X After using multi-level regression with post stratification, we determined that Joe Biden will be the winner of the US 2020 Election. Predicting the results of the US 2020 election is important because it helps give those residing in the USA certainty about who their next president will be. Information about the winner of the election is highly sought after because it can be used to predict upcoming policy. This is

*Code and data are available at: https://github.com/STA304-PS4/Trump-vs-Biden.

useful for individuals and organizations when planning their lives and can have a large impact on the rights of some individuals. Additionally, collecting information related to the election allows us to analyze different demographic groups and determine how they are voting to further identify areas they find important; this information is important from a policy design perspective and also to analyze how society is interacting.

In this report, we will begin by discussing the data and then modeling our results. Lastly, we will discuss the results of our analysis and include an appendix.

4 Data

In this analysis, we collected data from the Nationscape dataset — a collaboration between the Democracy Fund Voter Study Group and UCLA Political Scientists. The Nationscape dataset that we will be analyzing was collected from June 25th, 2020 to July 1st, 2020. Using the data from Nationscape, we would like the predict the overall popular vote of the 2020 American presidential election using multilevel regression with post-stratification.

The samples are provided by Lucid, "a market research platform that provides access to authentic, targeted audiences". In order to ensure that an appropriate sample is selected, samples are selected based on a set of demographic quotas based on a multitude of factors such as age, gender, ethnicity, etc. Respondents of the survey are sent directly to an online survey (conducted in English) operated by the Nationscape team and forced to complete an attention check to ensure the answers they provide are meaningful. Lastly, the data is weighted to be representative of the American population using factors such as — gender, four major census regions, race, etc.

The population in this dataset is all voting residents in the USA that are eligible to vote. This does not include a large population of those that reside in the US such as non-citizens or those who are incarcerated, but this is representative of a real election. The population frame is all voting residents in USA that have access to a computer in order to complete the online survey. As mentioned before, the survey is taken online, therefore it excludes those without access to a device that can take the online survey. The sample in this case is all respondents who complete this survey online.

The data is collected using a convenience sample selected based on a set of demographic criteria. Because convenience sampling was used over random sampling, this can result in some biases in our dataset.

The respondents of the survey are found through Lucid, which "runs an online exchange for survey respondents". The respondents from the survey on Lucid are sent to the online Nationscape survey where they are prompted to complete the survey. Respondents who complete the survey not choose not to disclose their household income; this results in a large number of non-respondents because they may not be comfortable providing sensitive information. In an effort to account for the respondents who choose not to provide income, non-respondents are not weighed for income.

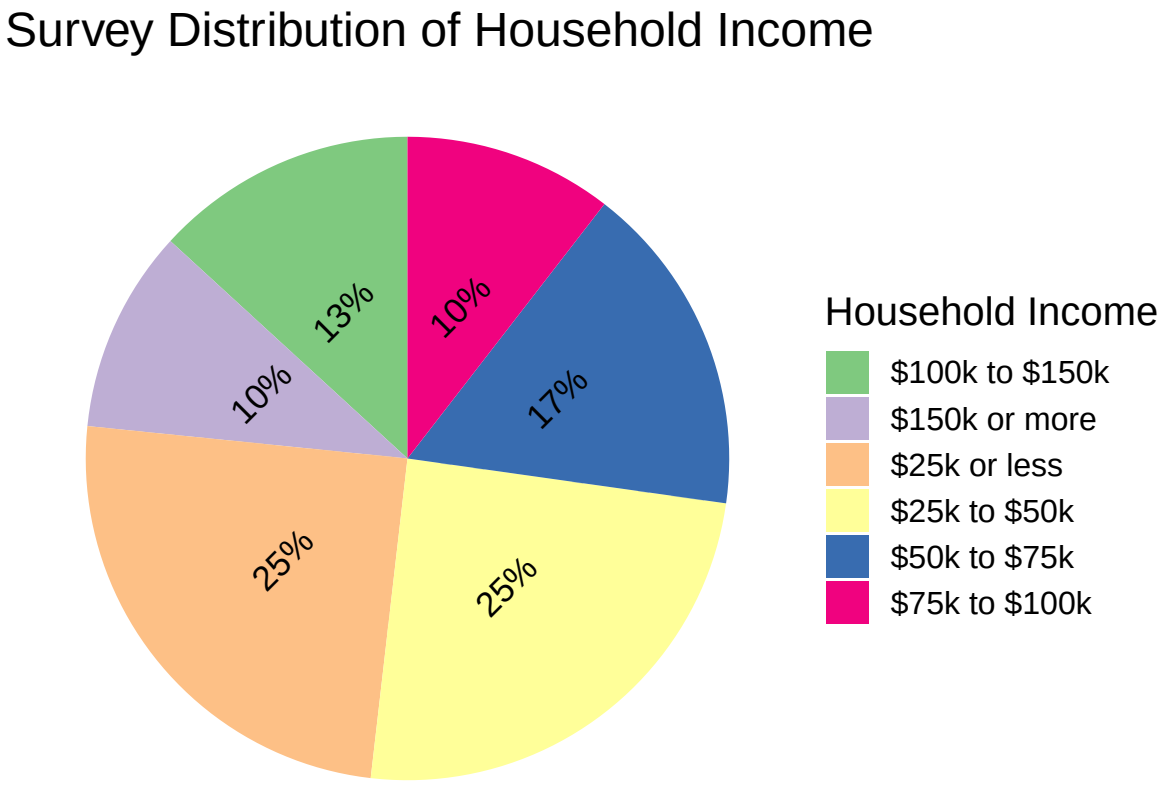
The variables that we used in our dataset are as follow: - age o "What is your age" o It provides an integer value greater than or equal to 18 - education o "What is the highest level of education you have completed?" o An extensive list of options are provided - Hispanic o "Are you of Hispanic, Latino, or Spanish origin?" o An extensive list of options are provided - household_income o "What is your current annual household income before taxes?" o Provides an extensive list of options beginning with $5,000 increments and then increases to $25,000 increments o This is the variable in our dataset that many people did not feel comfortable answering and was left as N/A. When cleaning the dataset, we removed all responses that left this answer as N/A
- race_ethnicity o "What is your race" o An extensive list of options are provided. - extra_covid_worn_mask o "Have you done any of the following in the past week? - Worn a mask when going out in public" o Yes or No options are available for the respondent to fill out - state o "Respondent's state, as a two-character postal abbreviation. Calculated based on entered ZIP code." o Manually filled out by respondent - trump_biden o "If the general election for president of the United States was a contest between Joe Biden and Donald Trump, who would you support? " o Options are either 'Joe Biden', 'Donald Trump', or 'Don't Know'

We created an additional variable called 'binary' which outputs the result of the variable 'trump_biden' as a 1 or a 0 making it binary, therefore easier to analyze. Since some variables had an extensive list of options (e.g. age, household income), we grouped the responses for those variables so that they are simpler and easier to analyze.

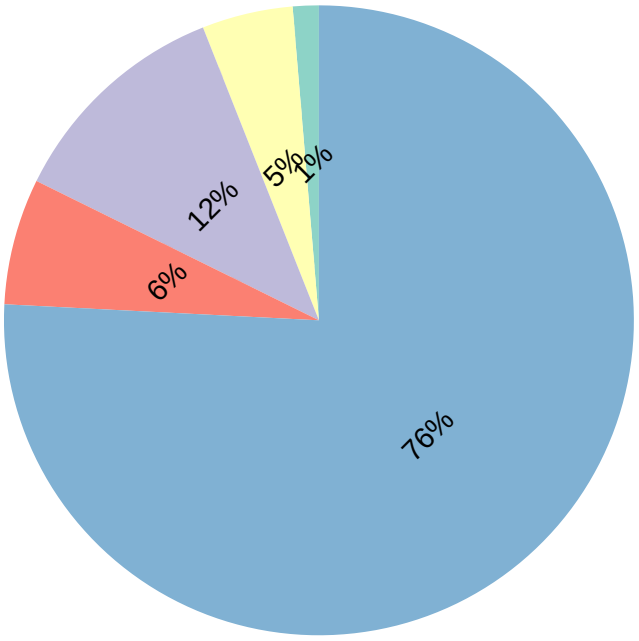
The first few rows of the cleaned dataset can be viewed below:

Age Group	Race	Household Income	Presidential choice
36-55	white	$75k to $100k	Donald Trump
36-55	white	$100k to $150k	Donald Trump
36-55	white	$150k or more	Donald Trump
56-75	white	$50k to $75k	Donald Trump
36-55	white	$25k or less	Donald Trump
36-55	white	$25k or less	Joe Biden

In Figure [] below, we can see that the distribution is largely skewed towards families with income less than

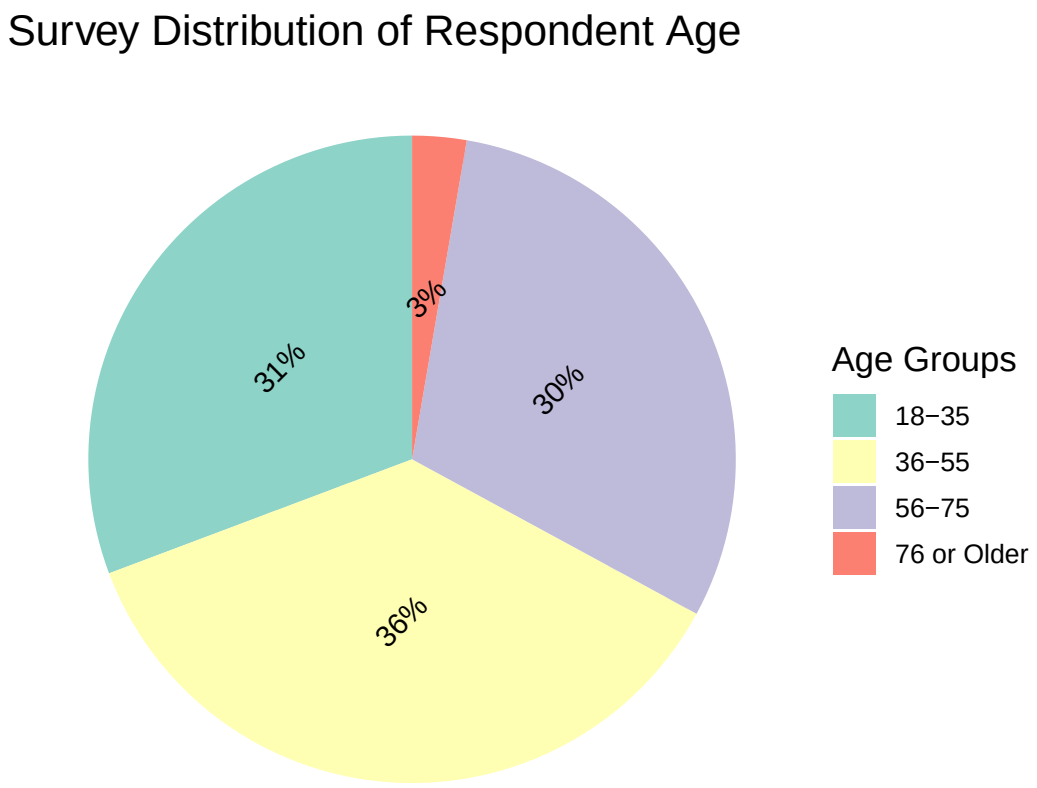


$50,000:
Because of the skewed distribution of income, we can justify using post stratification. Clearly, those with income less than $50,000 are overrepresented in the dataset as they take up 50% of the responses, therefore it is appropriate to use post stratification.



We can see a similar skewed distribution in Figure [] below:
Looking at the distribution of race, we can see that 75% of the respondants from the survey are white. Since the distribution does not represent minorities, we can justify using post stratification.

Lastly, in Figure [] below, we can see how the population aged older than 76 are barely represented:



The fact that those age 76+ make up only a 3% of the data could be a product of the fact that the survey was completed online. Clearly, the survey doesn't accurately represent those aged older than 76 posing the question — does our data represent the United States? The answer is no. To adjust for the fact that the population aged 76+, we will use post stratification.

The dataset that we used contains many key features. Many of the variables in the dataset contain an extensive list of options assuring us that the data represents the participant's view accurately. Despite the fact that we cleaned the dataset to make it simpler to view, we still believe that the detailed responses that the participants provided accurately reflect their views on the 2020 election, making our analysis meaningful. Another key feature is the fact that the data was collected over a short period of time— from June 25th, 2020 to July 1st, 2020. This significantly reduces the negative time based effects that may have affected the data ensuring that all respondents are responding on a similar ground and no one's opinion dramatically differed due to the survey taking place over a large period of time.

The dataset we used contains many strengths and weaknesses. One of the strengths is due to the fact that the survey is distributed online. In today's age where most things are online, this ensures that we can reach a wide part of the United States. At the same time, this can also be viewed as a weakness because certain age groups of the United States may not be tech-savvy enough to complete an online survey. The older population of the united states (e.g. 75+) may not be able to complete the survey because it is online which means that our data may not accurately reflect the older populations views regarding the 2020 election. A potential solution to this weakness is to use administer multiple forms of the survey to ensure that all people can access the survey.

Another weakness of the data comes from the fact that we use convenience-based sampling which inherently contains bias. Because the sampling method is based on convenience, we may be unlucky when sampling and get a sample group that inaccurately represents the population of the US making our results meaningless.

Lastly another weakness of our dataset is due to the fact that the survey data we are analyzing is collected from June 25th, 2020 to July 1st, 2020. Between July and November (the month of the election), many

voters may not have an informed vote and their opinions may still be swayed after watching crucial debates and hearing what the leaders have to say; our data doesn't take into consideration the people whose votes may change in the most crucial months of the election. A solution to this weakness can be to analyze data that may have occurred closer to the actual election so that the participants have an informed vote.

5 Model

The model we use to predict the result of the election is MRP, multilevel regression with post-stratification. The advantage of this model is that it has better performance when estimating behaviors of a particular subgroup of the population. One can interpret MRP as a combination of multilevel logistic regression and post-stratification. Logistic regression allows us the simplicity of analyzing the categorical response variable while incorporating both numerical and categorical explanatory variables; while combined with post-stratification, it yields a much narrower confidence interval for the estimation.

We are interested in how variables such as age, education, race, family income, geographic region, and mask-wearing decision affect citizens' voting intentions, particularly between the major party candidates, Donald Trump for the Republicans and Joe Biden for the Democratic Party. The response variable that we are interested in is the chances of a citizen voting for Trump in the coming presidential election. It is a categorical variable with two levels: 1 represents intending to vote for Donald Trump, and 2 means planning to vote for Joe Biden.

Ideally, we ought to run a pre-analysis to determine the variables of interest. However, we will leave that open for further analysis and future elections due to the limited time and budget. We chose the explanatory variables based on our understanding of a respondent's significant characteristics that could potentially affect his/her voting intention. Except for respondents' age, education, income level, and state(common in statistical studies), we are very interested in the respondent's habits towards mask-wearing. Firstly because the US is the country with the most COVID cases globally, the influence of COVID on the electoral votes is inevitable. Secondly, since Trump and Biden hold quite opposites opinions towards handling the pandemic, this parameter will likely represent part of the respondent's voting intention in the same way as the more familiar political indicators.

First, we train a multilevel logistic regression model using raw data from the voter study group towards the variable of interest using the seven explanatory variables. We then apply this model to our post-stratification data set to get the estimates.

Logistic regression estimates $\beta_0 ... \beta_k$ in (1)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1}$$

where p is the probability of event A that we are interested in, $\beta_0$is the intercept, $x_1 ... x_K$ are our variables of interest and $\beta_1 ... \beta_k$ are parameters for each of these variables. Based on the result, we are able to estimate p for a particular case given all the variables. We use `as.factors()` to incorporate dummy variables for all the categorical variable.

The main logic behind post-stratification is that we divide the sample data into strata based on a few attributes such as state, income, education level. Then we come up with a weight for each stratum to adjust its influence towards our prediction. Namely, if a stratum is over-represented, we want to reduce its impact on the prediction result; if a stratum is under-represented, we want to increase its influence. In this sense, the combination with logistic regression allows subgroups with a relatively small size to 'reference' from other subgroups with similar characteristics.

To mimic the population as closely as possible, we need a post-stratification dataset large enough to refer to. In this report, we use the ACS because the census happens less frequently than the ACS which is updated annually.

The post-stratification estimate is defined by $\hat{y}_{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. An illustration using the education variable is $\hat{y}_{edu}^{PS} = \frac{\sum_{j \in J_{edu}} N_j \hat{y}_j}{\sum_{j \in J_{edu}} N_j}$, we can get an estimation given any level of education of respondents, J_{edu} denotes all subsets of education levels, $\cup_{i \in J} J_i^{edu}$ = all possible education levels.

(1) produces a proportion for the post-stratification based on regression, instead of merely averaging the sample in each stratum. Specifically, all possible cells are determined from combining:

- 4 different age groups;
- 6 education levels
- 5 races;
- 6 income levels;

That is 720 cells in total. We fit the logistic regressions for estimating candidate support in each cell. We used `function` from `package` to fit (2).

$$P(Trump) = logit^{-1}(\beta_0 + \beta_a ge + \beta_{j[i]}^{edu} + \beta_{j[i]}^{race} + \beta_{j[i]}^{income} + \beta_{j[i]}^{state} + \beta_{j[i]}^{mask}) \tag{2}$$

(2) is the model that we fit using the survey data, with $\beta_{j[i]}^{var} \sim N(0, \sigma_{var}^2)$. β_0 is the intercept,each of $\beta_{j[i]}^{var}$ represents the parameter for the i-th respondent in j-th cell. For example, $\beta_{j[i]}^{edu}$ can take values from β from any of the six education levels. Then we predict the results by summing up estimates for each cell to the population. The results of the logidtic regression is summarised in table 1.

6 Results

6.0.1 Summary Table

Table 2: statistics summary for logistic regression

variable	$\hat{\beta}$	Std. Error	p-value
intercept	-0.00	0.26	0.99
$36 < age < 55$	0.42	0.07	$6.06e^{-9}$
$56 < age < 75$	0.37	0.07	$8.91e^{-7}$
$age > 76$	0.62	0.18	0.00
Asian or Pacific Islander	-0.75	0.27	0.01
black	-1.89	0.26	$1.25e^{-12}$
nec	-0.75	0.26	0.01
white	0.09	0.24	0.68
$income > 150k$	0.14	0.11	0.22
$income < 25k$	-0.38	0.10	0.01
$25k < income < 50k$	-0.27	0.10	0.01
$50k < income < 75k$	-0.19	0.10	0.06
$75k < income < 100k$	-0.31	0.11	0.01

7 Discussion

education_grouped	president_predict
Completed 1 Higher Education Degree	0.0954821
Completed 1 Higher Education Degree or more	0.0634028
Completed HS	0.1879795
Completed Some HS	0.0207523
Less than HS	0.0190661
Participated in Higher Education	0.1142789

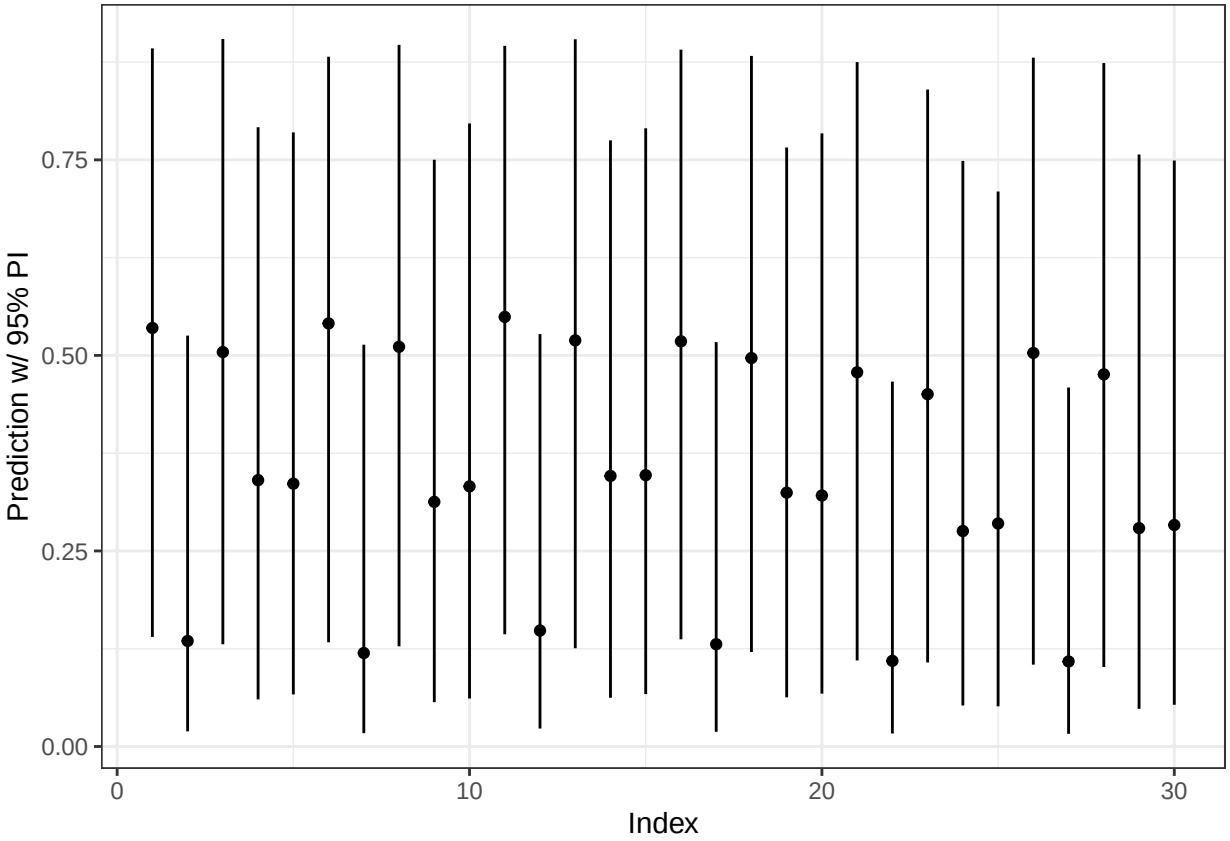


Figure 1: 95% Prediction Intervals for forcasting the Chances that Donald Trump wins the election

Weighing the model is important due to the survey collection weakness listed earlier in this paper. Survey weaknesses such as voluntary biases, which could see only those who can afford to spend time on a survey filling it out. Also, survey access, which would be affected by tech access, tech literacy, and literacy in general as the survey was only written and distributed via online ads.

This can be seen in how our demographic values compared to estimated national demographic values.

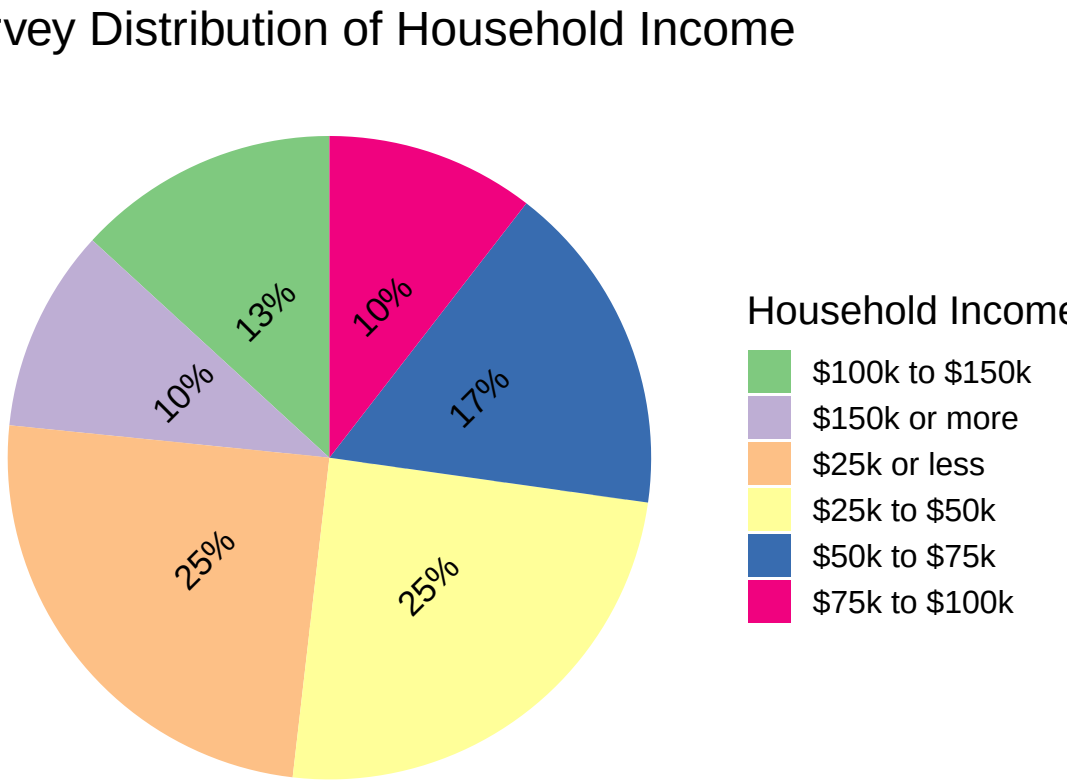
Before comparing those numbers how, was this analysis weighted? The method used for this was MRP or Multi-level Regression with Post-Stratification. This methodology was discussed in further detail above. But of note the data used for this weighting process was received through the IPUMS USA service. Specifically, their repository of American Community Surveys collected by the U.S. Census Bureau. The data was from the 2018 ACS. This survey is a 1-100 national random sampling survey of all U.S. addresses within communities of 65,000 people or more. It is run over the full 12-month period of the year and received 3.214 million responses for the year of 2018 (U.S. Census Bureau 2019). While there are some sampling errors

in the data they have been weighted across multiple years and the immense sample size helps to keep the risk of error low. Looking at a Confidence Interval calculation based on the sample size of 3.214 million and the U.S. population in 2018 of 327.2 million if looking at a confidence level of 99% the Confidence Interval is only 0.45 of a percent (CRS 2020). Meaning that there is a 99% certainty that the predicted percentage values of the ACS is with in 0.45% of the population value.

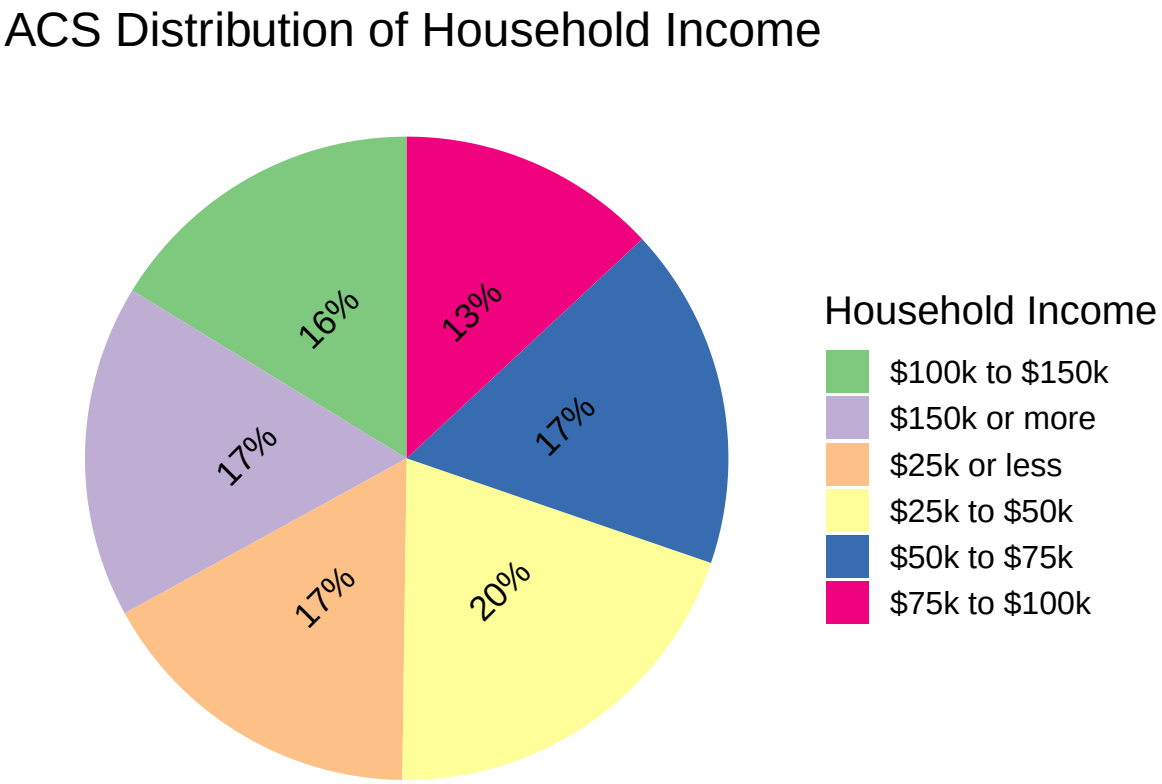
For more information on the ACS the user guide can be found in the input folder of the attached git repository.

However, to continue this discussion let's begin the comparison.

Looking at the income break down of our data it is clear when comparing Figure 1 and Figure 4 that the surveyed population tended towards a lower income on average. This is an interesting occurrence and may have to do with the tech literacy having a larger survey biasing effect than literacy and tech access. On average older individuals have higher incomes so it could be that the data is skewing younger and that is what is causing this. ## Figure 1

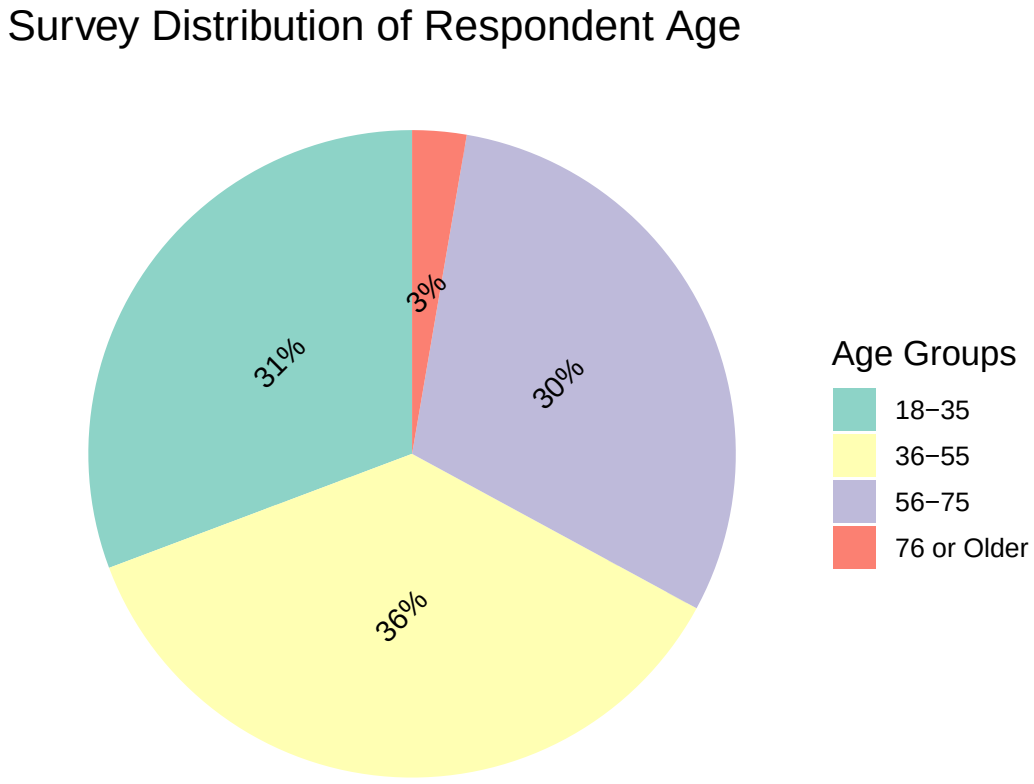


7.1 Figure 4

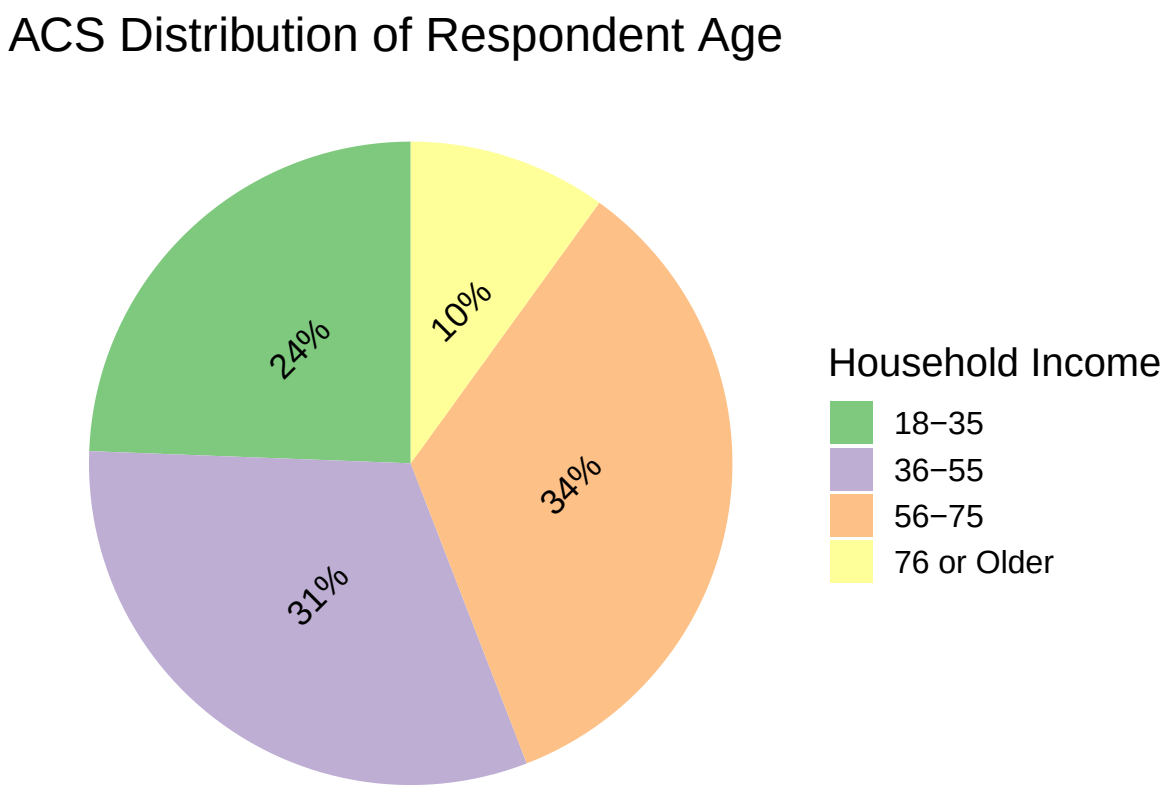


Examining the age group, we see that this could certainly be the case. As the Surveyed population is significantly younger than the general population. Most likely playing into the effects of tech literacy among older generations. This can be clearly seen in Figure 2 and 5, where 18-35-year-old respondents are greatly overestimated and 76 and older respondents greatly under-represented.

7.2 Figure 2

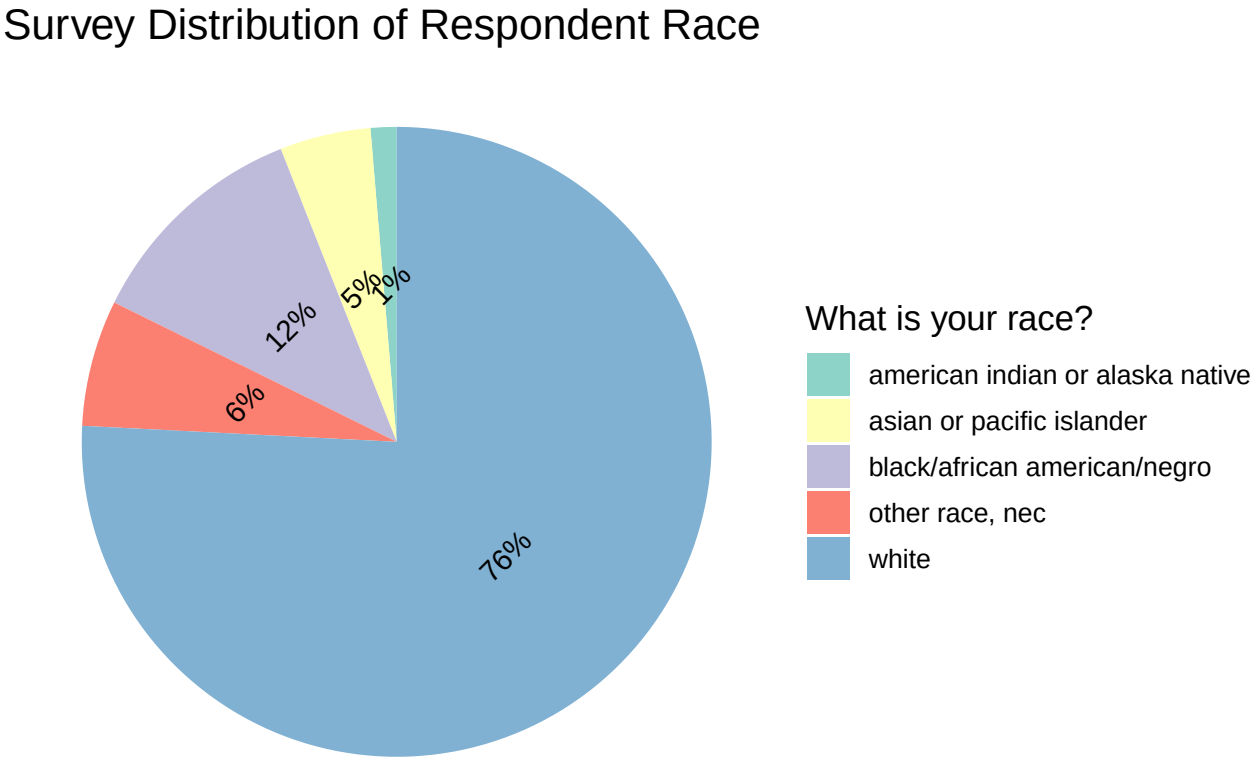

Survey Distribution of Respondent Age
31%
3%
30%
36%
Age Groups
18−35
36−55
56−75
76 or Older

7.3 Figure 5

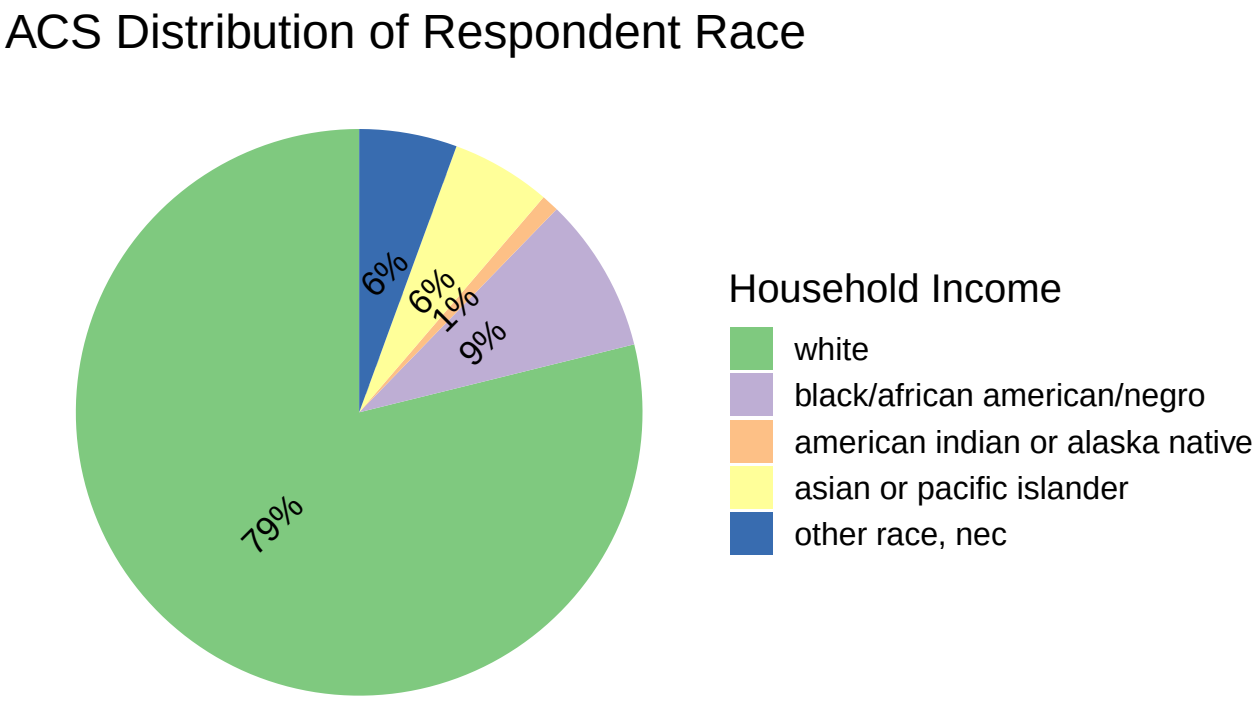


Finally, when looking in to the ethnic and racial breakdown of the survey data African American are underrepresented and White Americans overrepresented. This is displayed in Figure 3 and 6. This could be for a myriad of reasons, but it could be linked to tech access as a higher portion of African Americans experience poverty when compared to their White counterparts.

7.4 Figure 3


Survey Distribution of Respondent Race
1%
5%
12%
6%
76%
What is your race?
american indian or alaska native
asian or pacific islander
black/african american/negro
other race, nec
white

7.5 Figure 6



By separating all respondents of the ACS into subgroups based upon our independent variables we can then create proportions. This allows for the weighting of data through the model and adjusts our sample in order to account for the skewed demographics. An example of these proportions is displayed below.

7.6 Table 6

Age Group	Race	Household Income	Number of Repondents	Proportion of Responses
18-35	white	$100k to $150k	604	0.000250
18-35	black/african american/negro	$100k to $150k	64	0.000027
18-35	american indian or alaska native	$100k to $150k	9	0.000004
18-35	asian or pacific islander	$100k to $150k	72	0.000030
18-35	other race, nec	$100k to $150k	172	0.000071
18-35	white	$100k to $150k	2678	0.001109

Once the data is adjusted what are the findings?

Based on adjusted values Biden is projected to win the popular vote by a significant margin, ... This holds true and follows other surveys conducted in this time period (Financial times 2020).

Significantly though the data of this analysis shows a few key demographic group differences in voting patterns.

Firstly, race was a highly significant factor in voting behavior for non-white Americans. White Americans

were not significantly influenced by there race and tended to vote based on other factors. But individuals who described themselves as Black, Asian, a Pacific Islander, or another race all were significantly more likely to vote for Joe Biden than Donald Trump. The largest of these effects was among African Americans. Compared even to Asian Americans their coefficient was twice as large, meaning they were affected by racial effects twice as much.

Only 12.5% of Black respondents intended to vote for trump.

This aversion from non-white voters makes sense and may not necessarily be tied to a difference in policy opinion. Trump has time and time again been caught in racially charged situations where he has displayed disrespect to non-white communities. Either in calling African Nations "Shithole Countries" (Sorkin 2018), being resistant to denouncing white supremacist groups that endanger non-white Americans (Bump 2020), or in his reaction the Black Lives Matter movement (Liptak 2020).

So, it is no surprise that race has become a powerful factor on voting for people of color who may feel more strongly threatened than in other Presidential elections.

Secondly, age plays a significant effect. It is an oft said adage that the old are more conservative than the young. "If it ain't broke don't fix it." This adage appears to hold true for this presidential election as it has for many before it. The eldest in the population are far more likely than the young to vote for Trump than Biden. However, something of note is those aged 36-55 are slightly more likely to vote trump than those aged 56-75.

This is an interesting trend that is not often a part of election data and deserves further investigation. Some possible effects that could be hidden in this data is a worry for retirement as those approaching retirement age worry about whether retirement programs will still exist in the same capacity as they age into them. Or it could represent the effect the Covid pandemic is having on this election. Those who are older are at higher risk and will on average no more people who are at higher risk. So, it could be that people in the 56-75 age range are more worried about Trump's response to Covid which has been widely criticized. So, their age is playing less of a role than their opinion on health care.

Thirdly, income plays a significant effect on voting until an individual begins making over $150k. Then the effect begins to weaken. Looking at the data there is a clear trend that lower income individuals tend to prefer Biden to Trump. With a very linear relationship up until looking at individuals who live in households that make $75k to $100k. Where there is a sudden uptick in Biden support before it drops back down. And those making $100k and more tend to lean more towards Trump.

There are tow interesting trends in this data. First the uptick at $75k is interesting. Could this be an uptick from the confounding variable of Education as college educated individuals tend to be more liberal? Or is there another confounding variable that is affecting this uptick in Biden support like certain policies aimed at benefiting the middle class? Second the significance of the effect income has is strongest at lower incomes but weakens at higher income. Could this represent the general need for government support those with lower income experience? Meaning that those at lower income are more invested in policies that effect their incomes, than those with higher income. Or is this based on confounding variables that begin to exist at higher levels of income like education?

7.7 Weaknesses and next steps

7.7.1 1: Nested Bayesian Model

7.7.2 2: Pre-analysis

Appendix

8 References