51% of the Popular Vote Goes to Trump According to Socioeconomic Factors

A multilevel linear regression approach towards the election outcome

Alex (Zikun) Xu and Yitian Zhao

03 November 2020

Abstract

With the upcoming 2020 US Election this Tuesday, many people around the world are looking forwards to the presidential race between Trump and Biden. We are analyzing this election and making a forecasting to predict who is most likely to win the election based on socioeconomic factors such as the age, education, income, region, and race of voters. Our data is retrieved from UCLA Nationscape Survey and the American Community Survey to create a multilevel regression with post stratification. We will have a chance to see the attitudes of American voters within the past 2 years to determine how their values have changed to decide on this year's election.

Introduction

In this paper, we used various demographics of the American population to help model and predict the outcome of the 2020 U.S. presidential election. The question lies as to which candidate will fall in position as the head of the U.S. government, Donald Trump vs. Joe Biden. This is the sort of election that will impact the future of the United States both in the short-term, especially with the COVID-19 pandemic situation, as well as the long term in determining the reputation and impact the United States will have on a global scale. The world's eyes will be tuning in on November 3, 2020 to watch the outcome of the election as it is going incredibly close between the two controversial candidates.

As students trying to get a better sense of the election outcome, it is incredibly difficult to have any grounded confidence in either one of the candidates due to the overabundance of media sources with widely different opinions on the election. There is added complexity in gaining a clear result forecasting due to the complicated layers involving states and ridings which makes it difficult to get an outcome for the whole election by only observing individual voter preferences.

With that in mind, we want to make a forecasting for the 2020 election based on socioeconomic variables including age, education, race, income, and region to determine the popular vote. The data that we are using to determine a forecasting include the UCLA Nationscape data set from June 2020 and the 2018 American Community Survey from IPUMS. We are using the candidate preference along with the demographic information from the Nationscape data in order to create multilevel linear regression for the odds of Trump winning and then applying this model on the ACS to obtain a forecasting. Our results show that there the popular vote is slightly in Trump's favor with the most significant variables being voters in older age groups, voters with high income, voters in the Southern census region, and voters who belong to a non-white race group. Most of these factors make a positive contribution to our model with the exception of the preference of non-white voters.

Since there are so many different angles to approach studying the US election, we have to bear in mind that our forecasting is by no means a definitive conclusion for who is going to win. Especially since our results show that there is only a slight favor for Trump's chances, we have to understand that there remains many more variables that we did not take into account that might sway our conclusion towards either candidate. Nonetheless, we are taking a socioeconomic approach towards creating our statistical model. We will first discuss the data, where it came from, and what a preliminary look at the variables might suggest. We then create our model and discuss its significance. Finally, we will discuss the results from our model and why it may turn out the way it did. For future reference, there remain many approaches towards studying something as complex as the federal election in any nation. Our work in this paper helps us gain a better understanding of an outcome with respect to the variables we choose to approach it from.

Data

For our data, we are using the American Community Survey (ACS) and the UCLA Nationscape Dataset. We will discuss the key features and methodology for each of these surveys.

Survey Dataset

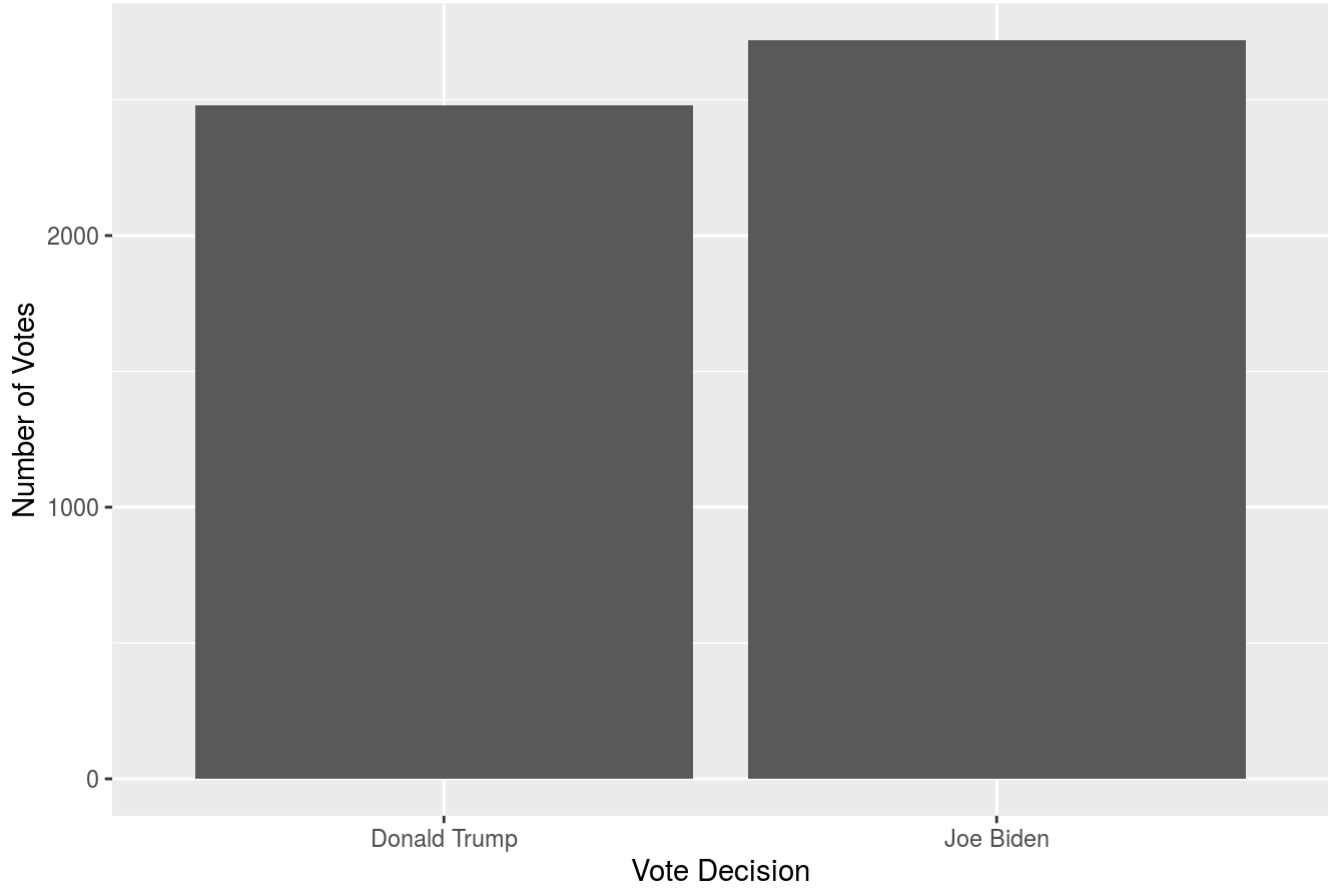
The individual-level survey dataset was obtained from the collaboration between the Democracy fund Voter Study Group and the UCLA Nationscape's public opinion survey. Specifically, UCLA Political Scientists Chris Tausanovitch and Lynn Vavreck. The Nationscape samples are obtained from Lucid, a market research platform which is an online platform for survey respondents. The results of the questionnaire by Lucid are then directly sent to a survey software operated by the UCLA Nationscape team.

The Nationscape survey has a target audience of American citizens (reasonably assuming) and conducted over 500,000 online interviews from the periods of July to December in 2020. This survey commenced in July 2019 and has approximately 25,000 respondents per month. The demographics of the survey results are age, gender, ethnicity, region, income, and education, which are used as explanatory variables to help model our outcome. The Nationscape survey data are weighted and is generated for each week's survey. The targets of the weights, conducted by Nationscape, is derived from the 2017 American Community Survey (ACS) of Census Bureau. A key feature that they included was the results of the 2016 presidential vote results to explain how past election results were affected by various demographics. This helps us analyze based on past results, the possibility of the respondent voting for the same candidate – Donald Trump. Weaknesses of this survey include divergence between Nationscape datasets and government provided datasets may be dependent on the type of questions asked. Respondents have the choice of not answering certain questions, such as income, thus non-respondents are ultimately not weighted in the targeted weights for this survey.

In order to present the data in a detailed way, we provided seven plots of the seven different variables used in the data set – five of those variables are used for modelling, specifically, age group, education level, household income, ethnicity, and region of respondents. In figure 6 and table 4, we showed the proportion of registered respondents that have the ability to vote in the upcoming election and we found that over 10% of survey respondents are either not registered to vote or do not know the eligibility of their registration status yet. That means that these respondent's answers for the other variables such as their education level, ethnicity, household income etc… are not significant towards the prediction of the 2020 presidential election result. Therefore, these respondents were omitted from the dataset in order to provide a model with lower bias to make a better prediction of the upcoming election outcome. From figures 1 to 5 and tables 1 to 3, we plotted the variables that will be used for modelling of the cleaned Nationscape dataset to have a better understanding of the variable compositions. Surprisingly, we find that in Figure 1, slightly more respondents chose to vote for Joe Biden than

Donald Trump (50.3% vs 49.7%) which contradicts the result that our model predicted. Other demographics such as age group, gender, ethnicity, and census region found from figure 2 to figure 6 provides a visual representation of the actual data which are used for modeling later on.

Fig 1. Who Respondents are Voting For - from the Nationscape survey




```
## [1] "Table 1. Proportion of votes for each decision"
```

```
##
##             Donald Trump                Joe Biden             Someone else
##                0.4771154                0.5228846                0.0000000
##         I would not vote I am not sure/don't know
##                0.0000000                0.0000000
```

Fig 2. Age Group Regions of Voters - from the Nationscape survey

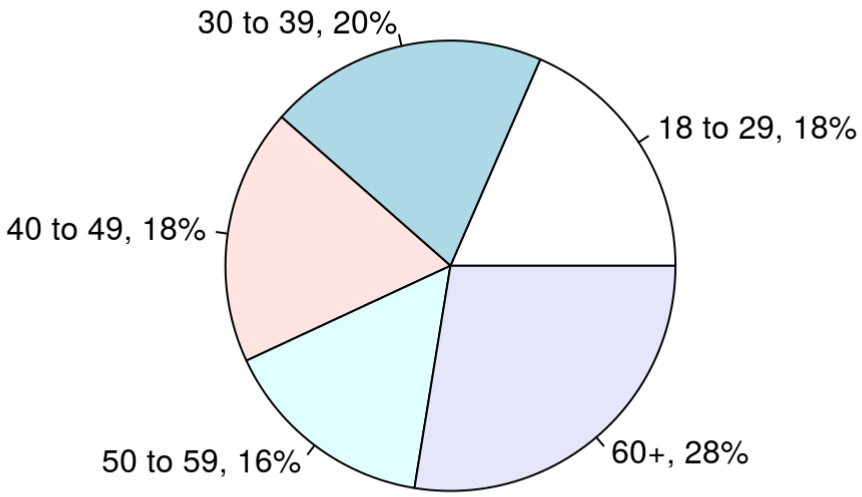
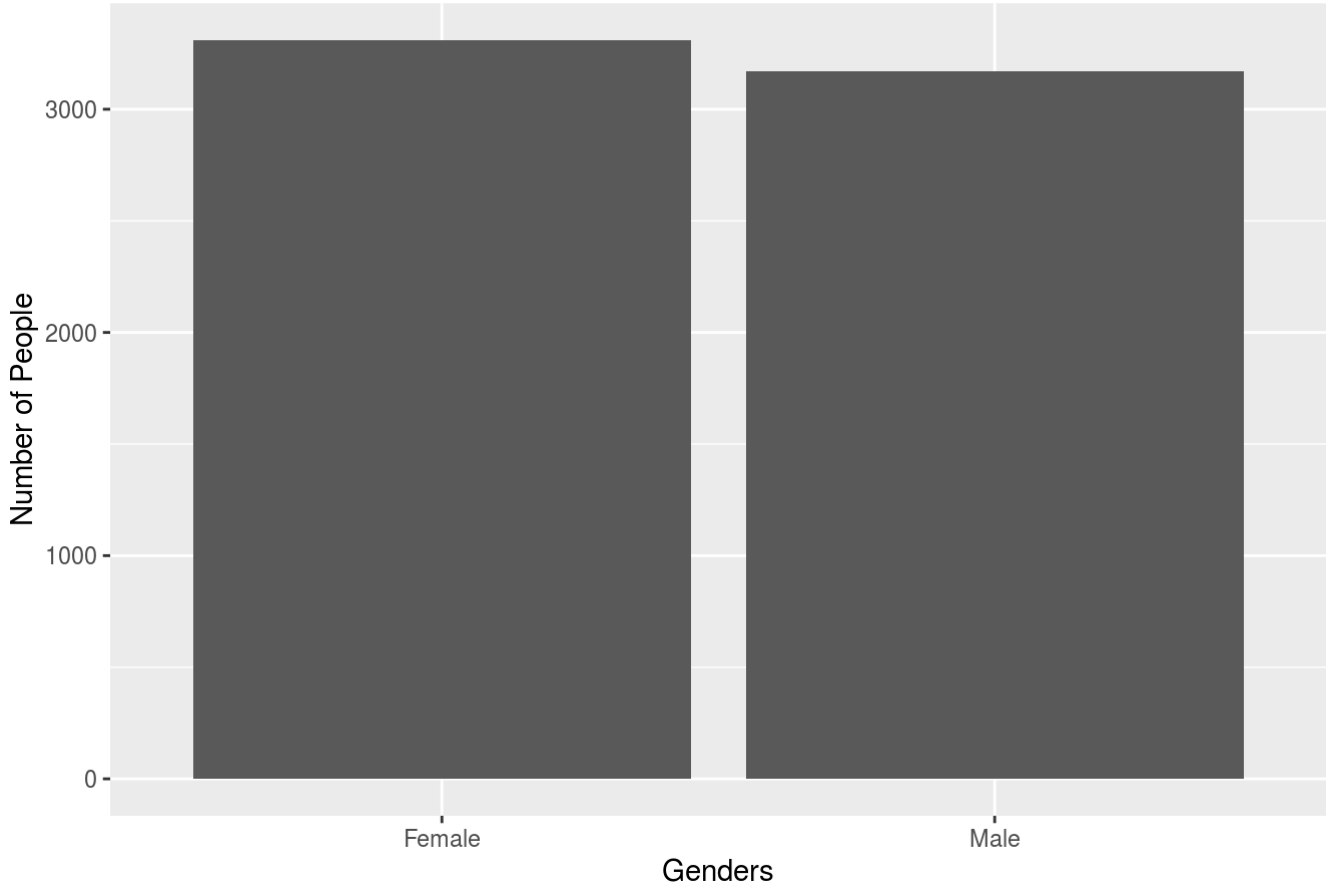
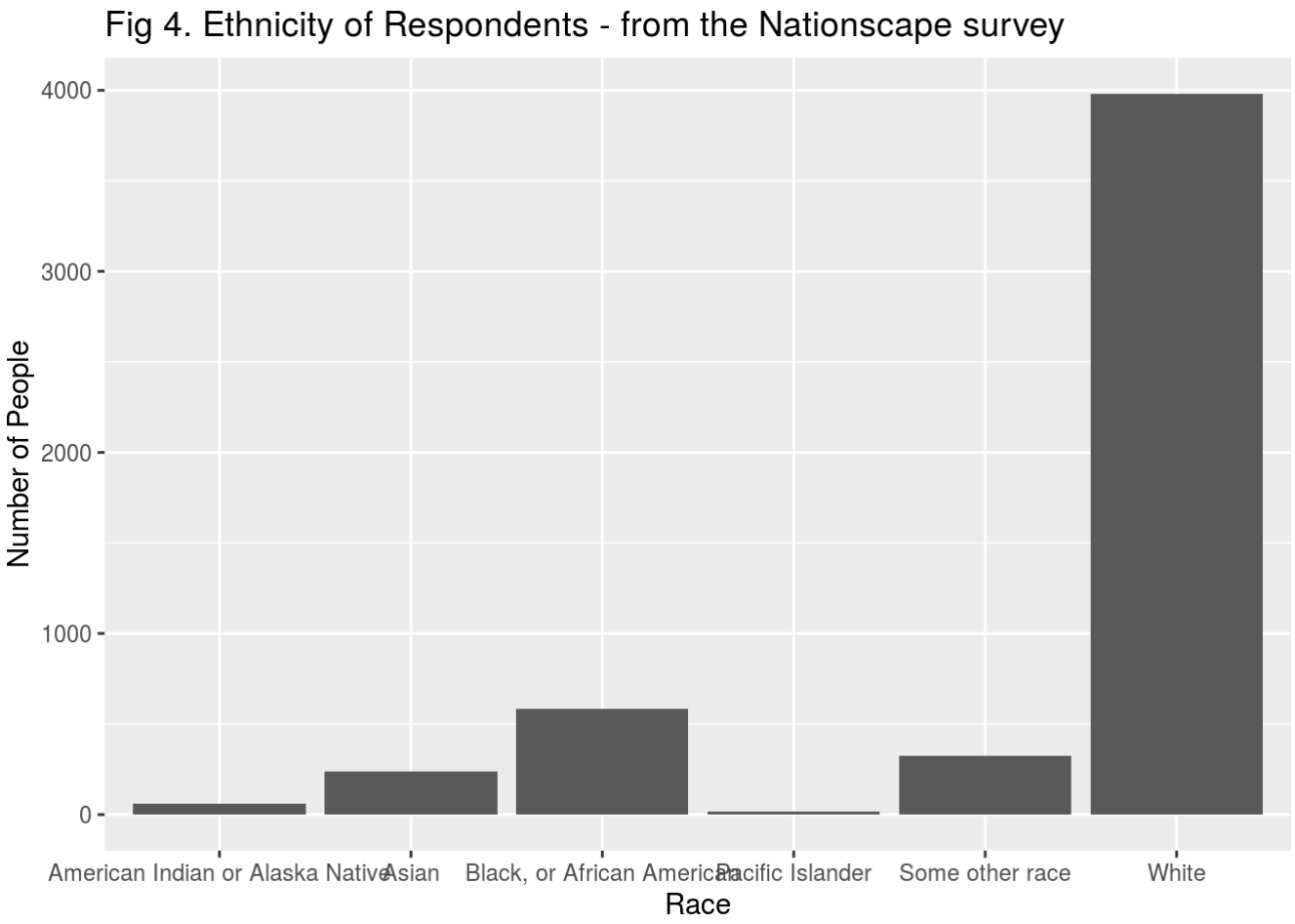


Fig 3. Gender of Respondents - from the Nationscape survey



```
## [1] "Table 2. Proportion of Gender"
```

```
##
##   Female     Male
## 0.510727 0.489273
```




```
## [1] "Table 3. Proportion of Race Ethnicity"
```

```
##
## American Indian or Alaska Native                    Asian
##                      0.011346154                0.045769231
##       Black, or African American           Pacific Islander
##                      0.112115385                0.003461538
##                  Some other race                    White
##                      0.061923077                0.765384615
```

Fig 5. Census Regions of Voters - from the Nationscape survey

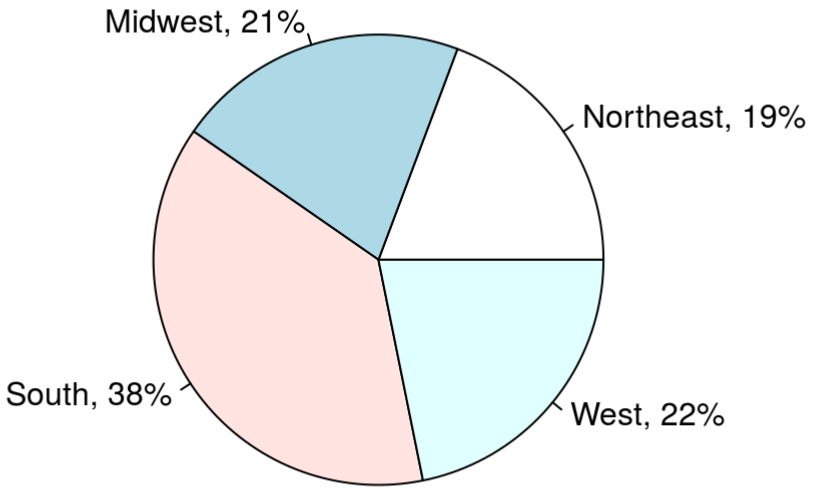
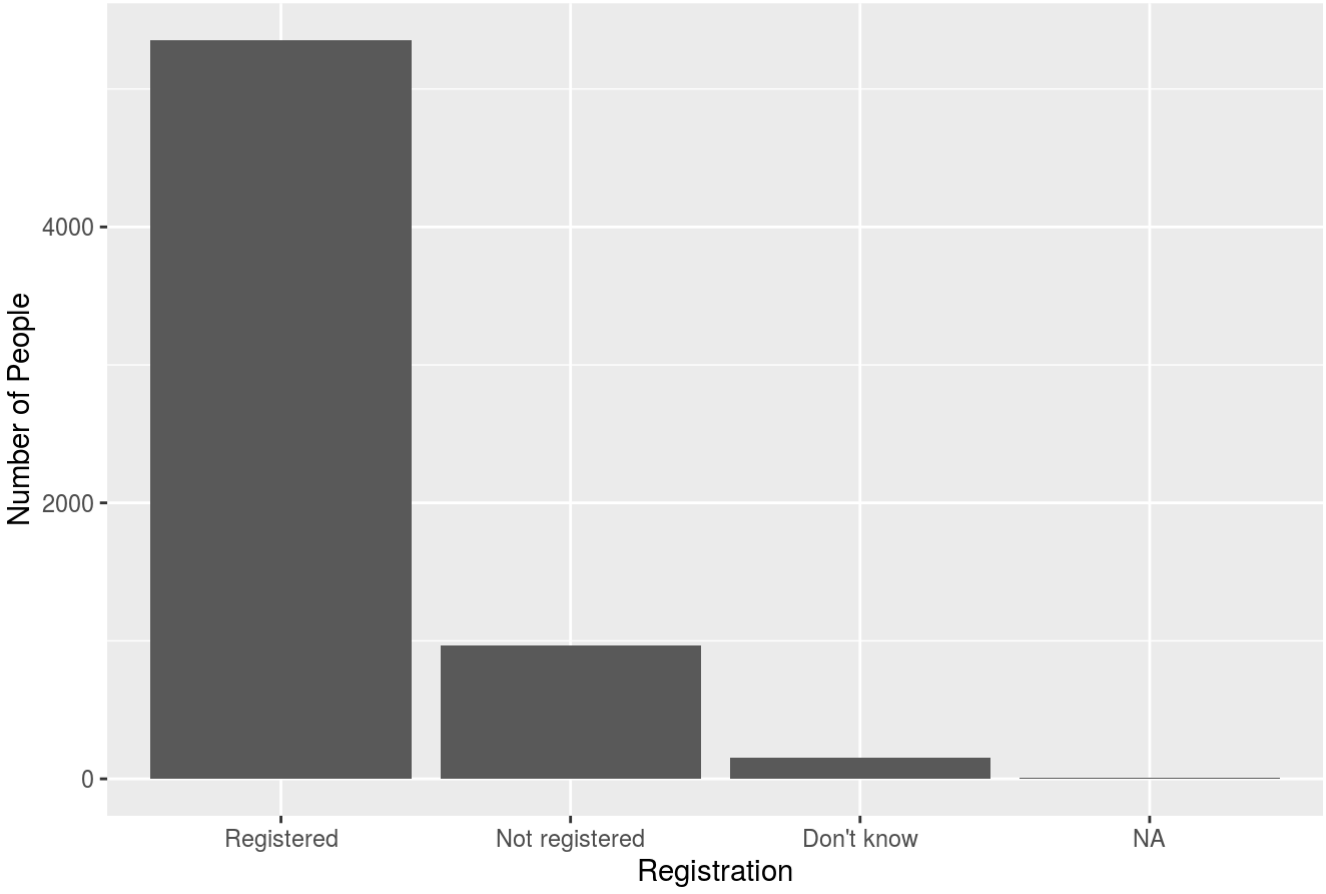


Fig 6. Registration of Respondents - from the Nationscape survey



```
## [1] "Table 4. Proportion of Registered voters"
```

```
##
##     Registered Not registered     Don't know
##     0.82722918     0.14943594     0.02333488
```

Post-Stratification Dataset

The Post Stratification dataset was gathered from the American Community Survey who has a target audience of the American resident population, living in housing units and group quarters (nursing, group homes, workers' dormitories, etc.) facilities. The ACS survey data for this paper was provided by IPUMS USA on their website, and we specifically used the 2018 ACS sample surveys drawn from federal censuses. The 2018 ACS Survey uses a series of monthly weighted samples to result in annual estimates with a 1-in-100 ratio national random sample of the population.

The population of this dataset is basically any person who is living in the United States. The sample frame, in other words the source of the material of which the ACS sample data was drawn from, is from the Master Address File (MAF) of Census Bureau. The MAF, that was initially created for Census 2000, uses multiple sources including the 1990 Address Control File, the United States Postal Service Delivery Sequence File and various computer assisted clerical field operations. The intent of the MAF was to source the American Community Survey, other Census Bureau demographic surveys, and the decennial census. The ACS data set samples approximately 2.5% of the expected number of residents in "Group Quarter" facilities. They aim to collect a sample of 3.54 million residents in the United States and 36,000 in Puerto Rico. There are sixteen designated sampling strata which consists of sampling within geographic entities such as counties, school districts, and other areas with functioning local governments. Blocks within each sampling entity fall within one of the sampling strata organized by size. Post-stratification is used to make sure that blocks assigned within smaller stratum are relatively represented and justified in comparison to the larger stratum.

The ACS consists of 12 monthly independent samples, sampling by address with personal visits in its first main stage of sample selection, then by phone, and the final phase by mail or the internet. If additional samples are required, the process is done for new addresses to be categorized within required stratums and sampling entities. Ever since 2006, the ACS has interviewed the resident population who are living in both housing units (HUs) or group quarters (GQ) – which are classified as "living quarters" by the Census Bureau. Data collection for housing units, done by Census Bureau are carried out through internet survey, mail, telephone, or personal visit. Non-responses or unreachable addresses are then reattempted for phone in contact in a second stage of sample collection later on, either followed up with computer assisted telephone interviewing or by computer-assisted personal interviewing.

Some key features of this dataset are that many demographic characteristics are covered through the survey and since the sample selection is updated monthly, most recent data are updated with the exception of the current month. The Census Bureau provides the frame of the ACS survey and produces/publishes up to three sets of estimates based on its total geographical population. In particular, the ACS uses ratio estimation to take advantage of independent population estimates by sex, age, race, and Hispanic origin, and estimates of total housing units produced by statistical estimates performed by the Census Bureau.

Since the post-stratification data set differs from the survey dataset from the previous section, our team had to filter out the variables and clean them in order to match the structure of both data sets. We mainly chose five variables to use in our modelling which individually are: age group each respondent is categorized in, highest

education of the respondent, race ethnicity of the respondent, and household income of the respondent, and which region of America the respondent was from.

Fig 7. Age Group Regions of Voters

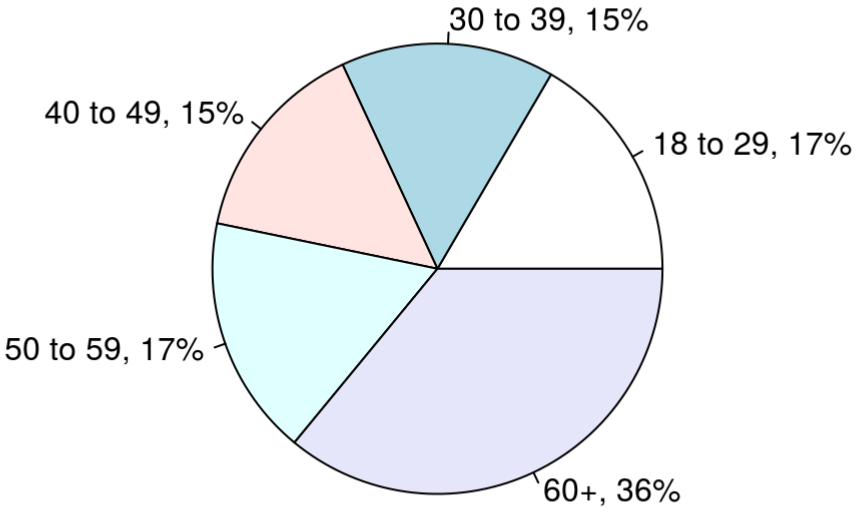


Fig 8. Race/Ethnicity of Voters

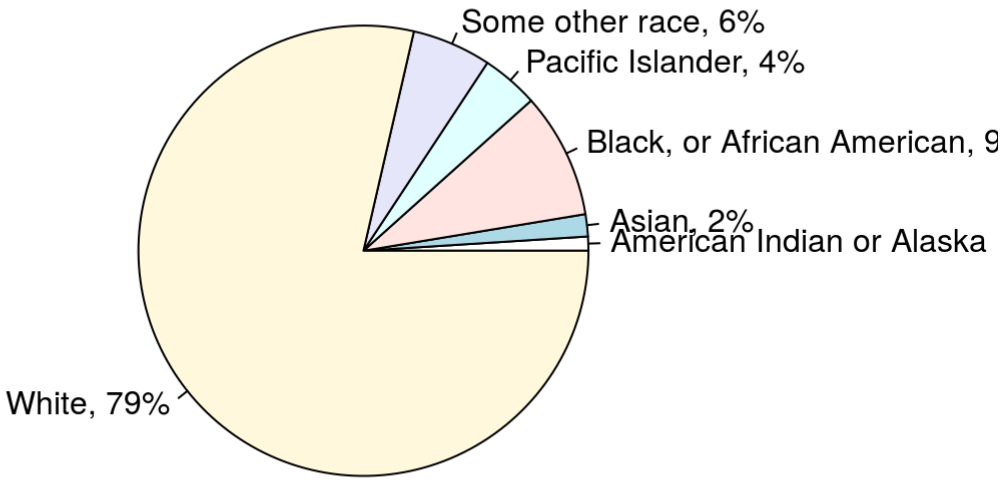
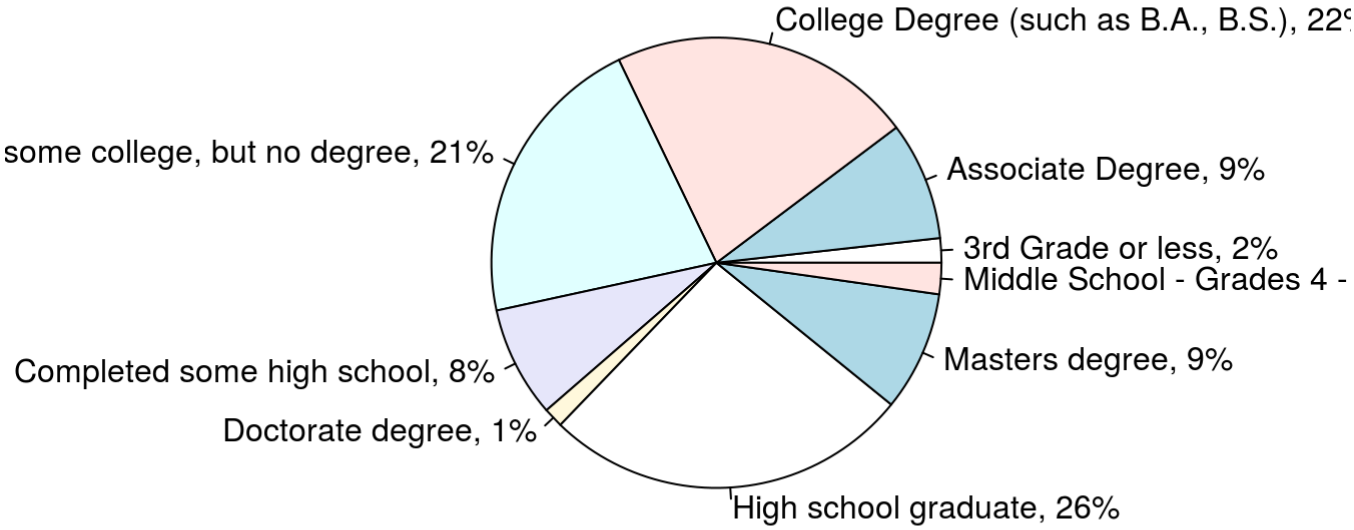


Fig 9. Education Level of Voters



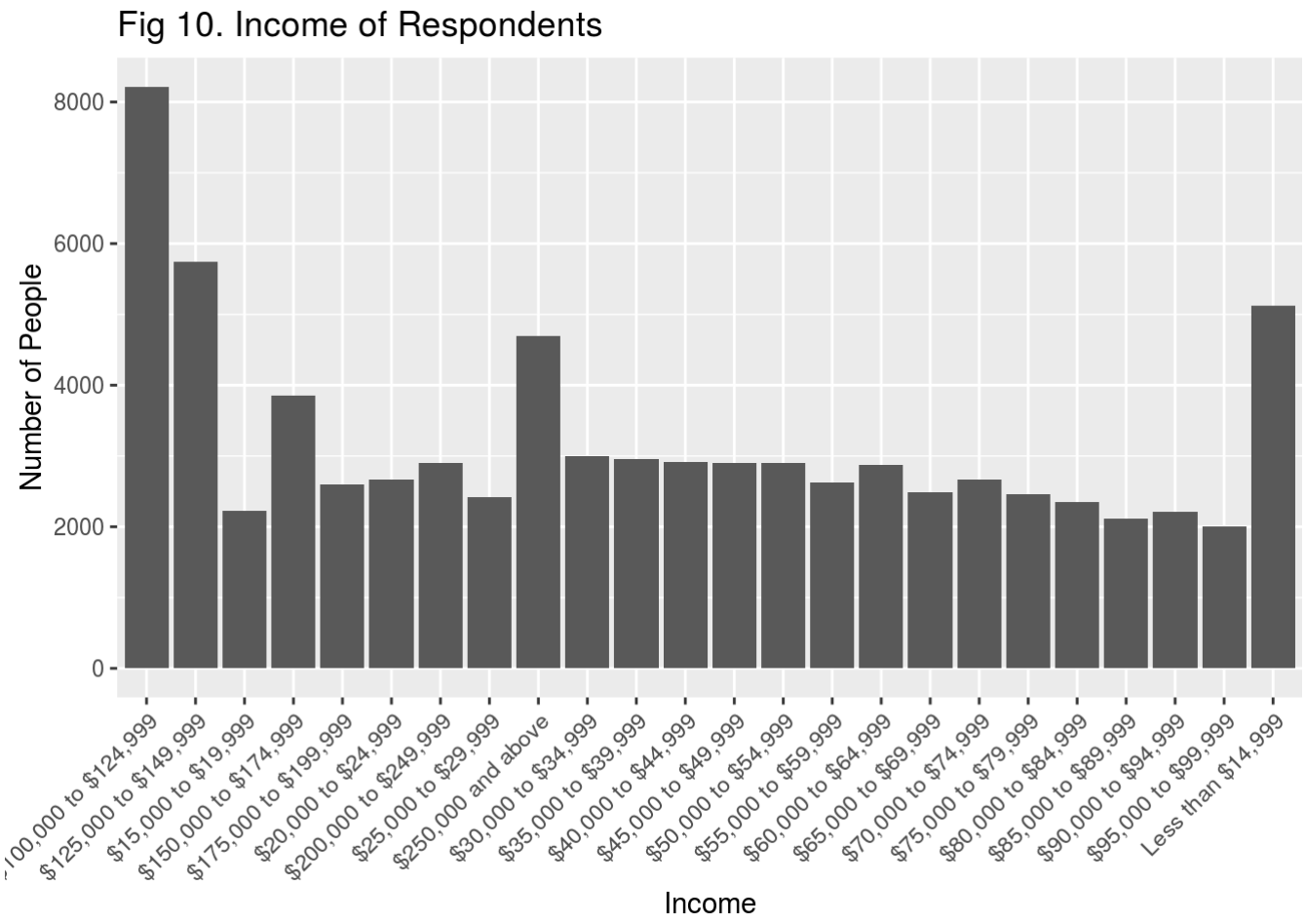
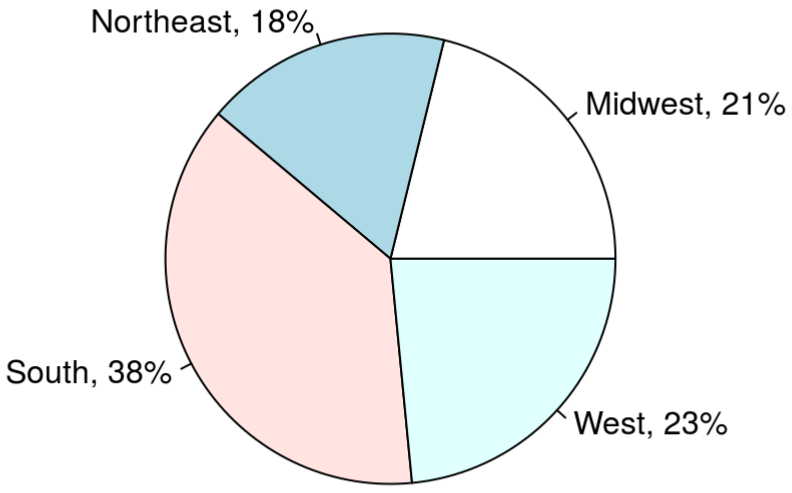


Fig 11. Census Regions of Voters



Model

```
##
## Call:
## glm(formula = as.numeric(vote_2020 == "Donald Trump") ~ age +
##     education + household_income + race_ethnicity + census_region,
##     family = "binomial", data = reduced_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.880  -1.137  -0.409   1.071   2.454
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                     -0.47116    0.71861  -0.656
## age30 to 39                                      0.56253    0.10398   5.410
## age40 to 49                                      0.74557    0.10852   6.870
## age50 to 59                                      0.70937    0.11098   6.392
## age60+                                           0.61768    0.10056   6.142
## educationMiddle School - Grades 4 - 8            0.44351    0.88532   0.501
## educationCompleted some high school              0.20352    0.67471   0.302
## educationHigh school graduate                    0.14769    0.67150   0.220
## educationOther post high school vocational training  0.08970    0.68003   0.132
## educationCompleted some college, but no degree  -0.19339    0.66988  -0.289
## educationAssociate Degree                       -0.41108    0.67444  -0.610
## educationCollege Degree (such as B.A., B.S.)    -0.39626    0.66981  -0.592
## educationCompleted some graduate, but no degree -0.38960    0.68548  -0.568
## educationMasters degree                         -0.48728    0.67399  -0.723
## educationDoctorate degree                       -0.15979    0.69874  -0.229
## household_income$15,000 to $19,999              -0.09976    0.16494  -0.605
## household_income$20,000 to $24,999               0.20510    0.15863   1.293
## household_income$25,000 to $29,999               0.10955    0.16020   0.684
## household_income$30,000 to $34,999               0.10320    0.15960   0.647
## household_income$35,000 to $39,999               0.03811    0.16526   0.231
## household_income$40,000 to $44,999              -0.01347    0.17387  -0.077
## household_income$45,000 to $49,999               0.24082    0.16799   1.434
## household_income$50,000 to $54,999               0.25386    0.15758   1.611
## household_income$55,000 to $59,999               0.31233    0.20168   1.549
## household_income$60,000 to $64,999               0.06431    0.20197   0.318
## household_income$65,000 to $69,999               0.15813    0.22054   0.717
## household_income$70,000 to $74,999               0.15986    0.18961   0.843
## household_income$75,000 to $79,999               0.33836    0.19389   1.745
## household_income$80,000 to $84,999              -0.05239    0.23518  -0.223
## household_income$85,000 to $89,999               0.05427    0.25026   0.217
## household_income$90,000 to $94,999               0.23752    0.27185   0.874
## household_income$95,000 to $99,999               0.01981    0.20712   0.096
## household_income$100,000 to $124,999             0.53423    0.14522   3.679
## household_income$125,000 to $149,999             0.42064    0.15948   2.637
## household_income$150,000 to $174,999             0.33180    0.19601   1.693
## household_income$175,000 to $199,999             0.77017    0.23524   3.274
## household_income$200,000 to $249,999             1.13235    0.22480   5.037
## household_income$250,000 and above               0.69216    0.22087   3.134
## race_ethnicityAsian                             -0.82407    0.30948  -2.663
## race_ethnicityBlack, or African American        -2.29571    0.30165  -7.611
## race_ethnicityPacific Islander                  -1.46827    0.64355  -2.282
```

```
## race_ethnicitySome other race                          -0.87027    0.29796  -2.921
## race_ethnicityWhite                                    -0.07857    0.27337  -0.287
## census_regionMidwest                                    0.10023    0.09620   1.042
## census_regionSouth                                      0.41143    0.08629   4.768
## census_regionWest                                       0.03536    0.09484   0.373
##                                                        Pr(>|z|)
## (Intercept)                                            0.512044
## age30 to 39                                            6.31e-08 ***
## age40 to 49                                            6.41e-12 ***
## age50 to 59                                            1.64e-10 ***
## age60+                                                 8.12e-10 ***
## educationMiddle School - Grades 4 - 8                  0.616398
## educationCompleted some high school                    0.762932
## educationHigh school graduate                          0.825915
## educationOther post high school vocational training    0.895058
## educationCompleted some college, but no degree         0.772817
## educationAssociate Degree                              0.542185
## educationCollege Degree (such as B.A., B.S.)           0.554122
## educationCompleted some graduate, but no degree        0.569794
## educationMasters degree                                0.469695
## educationDoctorate degree                              0.819120
## household_income$15,000 to $19,999                     0.545275
## household_income$20,000 to $24,999                     0.196048
## household_income$25,000 to $29,999                     0.494062
## household_income$30,000 to $34,999                     0.517867
## household_income$35,000 to $39,999                     0.817636
## household_income$40,000 to $44,999                     0.938244
## household_income$45,000 to $49,999                     0.151695
## household_income$50,000 to $54,999                     0.107182
## household_income$55,000 to $59,999                     0.121468
## household_income$60,000 to $64,999                     0.750168
## household_income$65,000 to $69,999                     0.473370
## household_income$70,000 to $74,999                     0.399173
## household_income$75,000 to $79,999                     0.080957 .
## household_income$80,000 to $84,999                     0.823720
## household_income$85,000 to $89,999                     0.828316
## household_income$90,000 to $94,999                     0.382278
## household_income$95,000 to $99,999                     0.923821
## household_income$100,000 to $124,999                   0.000234 ***
## household_income$125,000 to $149,999                   0.008352 **
## household_income$150,000 to $174,999                   0.090506 .
## household_income$175,000 to $199,999                   0.001060 **
## household_income$200,000 to $249,999                   4.73e-07 ***
## household_income$250,000 and above                     0.001726 **
## race_ethnicityAsian                                    0.007750 **
## race_ethnicityBlack, or African American               2.73e-14 ***
## race_ethnicityPacific Islander                         0.022519 *
## race_ethnicitySome other race                          0.003492 **
## race_ethnicityWhite                                    0.773812
## census_regionMidwest                                   0.297489
## census_regionSouth                                     1.86e-06 ***
## census_regionWest                                      0.709275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6842.9  on 4942  degrees of freedom
## Residual deviance: 6195.4  on 4897  degrees of freedom
##   (257 observations deleted due to missingness)
## AIC: 6287.4
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 0.5110307
```

```
##                                                          2.5 %     97.5 %
## (Intercept)                                          -1.90133262  0.9758500
## age30 to 39                                           0.35922432  0.7669231
## age40 to 49                                           0.53343670  0.9589372
## age50 to 59                                           0.49242076  0.9275547
## age60+                                                0.42118448  0.8154624
## educationMiddle School - Grades 4 - 8                -1.31446751  2.1970060
## educationCompleted some high school                  -1.16150205  1.5540088
## educationHigh school graduate                        -1.21170385  1.4921961
## educationOther post high school vocational training -1.28513232  1.4497747
## educationCompleted some college, but no degree       -1.55003489  1.1480538
## educationAssociate Degree                            -1.77611170  0.9385252
## educationCollege Degree (such as B.A., B.S.)         -1.75299310  0.9448615
## educationCompleted some graduate, but no degree      -1.77454803  0.9802566
## educationMasters degree                              -1.85167497  0.8613395
## educationDoctorate degree                            -1.56846031  1.2347524
## household_income$15,000 to $19,999                   -0.42398847  0.2229888
## household_income$20,000 to $24,999                   -0.10580859  0.5163769
## household_income$25,000 to $29,999                   -0.20466540  0.4236692
## household_income$30,000 to $34,999                   -0.20980637  0.4161837
## household_income$35,000 to $39,999                   -0.28637539  0.3618666
## household_income$40,000 to $44,999                   -0.35501711  0.3271227
## household_income$45,000 to $49,999                   -0.08830172  0.5706424
## household_income$50,000 to $54,999                   -0.05500415  0.5630104
## household_income$55,000 to $59,999                   -0.08320834  0.7083497
## household_income$60,000 to $64,999                   -0.33255138  0.4601919
## household_income$65,000 to $69,999                   -0.27453663  0.5915213
## household_income$70,000 to $74,999                   -0.21236160  0.5317026
## household_income$75,000 to $79,999                   -0.04109275  0.7197598
## household_income$80,000 to $84,999                   -0.51678109  0.4071717
## household_income$85,000 to $89,999                   -0.44014447  0.5434188
## household_income$90,000 to $94,999                   -0.29497351  0.7746129
## household_income$95,000 to $99,999                   -0.38826783  0.4247660
## household_income$100,000 to $124,999                  0.25015854  0.8196083
## household_income$125,000 to $149,999                  0.10842780  0.7338789
## household_income$150,000 to $174,999                 -0.05219942  0.7168911
## household_income$175,000 to $199,999                  0.31393928  1.2380250
## household_income$200,000 to $249,999                  0.69735707  1.5799548
## household_income$250,000 and above                    0.26163968  1.1287783
## race_ethnicityAsian                                  -1.43677462 -0.2198435
## race_ethnicityBlack, or African American             -2.89411402 -1.7076191
## race_ethnicityPacific Islander                       -2.84899979 -0.2738039
## race_ethnicitySome other race                        -1.46045721 -0.2882916
## race_ethnicityWhite                                  -0.62131729  0.4556182
## census_regionMidwest                                 -0.08827377  0.2889028
## census_regionSouth                                    0.24251589  0.5808421
## census_regionWest                                    -0.15047342  0.2213436
```

We ran a generalized linear regression model to form a relationship between a voter's preferred candidate against their socioeconomic factors including age, income, education, region, and race. This would regress one's choice of presidential candidate as the dependent variable against our 5 independent variables mentioned as such:

$$Z_{\hat{Y}} = \beta_0 + \beta_1 Z_{X_1} + \beta_2 Z_{X_2} + \beta_3 Z_{X_3} + \beta_4 Z_{X_4} + \beta_5 Z_{X_5}$$

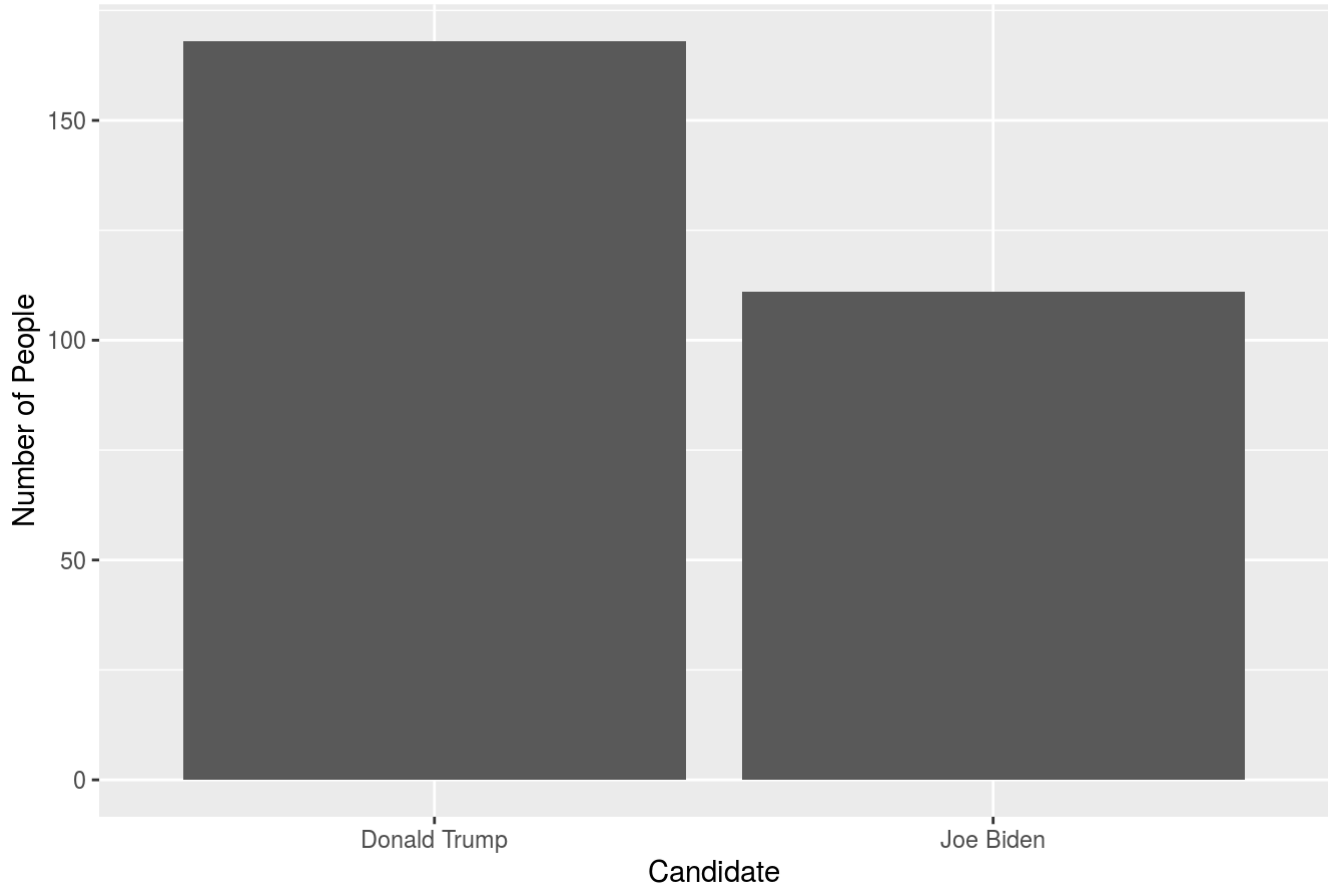
Our Z_y value represents the candidate choice while Z_x1, Z_x2, Z_x3, Z_x4, and Z_x5 correspond to age, income, region, race, and education respectively. Our independent variables are regressed as factors since education, region, and race don't make sense to be treated as quantitative variables and income was surveyed in brackets which leaves us without exact values for an individual. Age is sorted and organized into brackets and regressed as a factor in order to remain consistent with the other variables and allow for our trained model to be used for forecasting with the post stratification data.

When considering other models that could potentially be used, we decided to choose using a generalized linear since the nature of this study provides us with independent responses for our variables. While we are limited with analyzing the interactions between our independent variables, we are not restricted to using variables with normally distributed errors which works out in our case of regressing by factors. We are also only interested in looking at the result of two total outcomes for our candidate preference so we are using a binomial model. When considering an alternative such as Bayesian regression, it is harder to make a strong conclusion from our model only using non-informative priors as we do not have priors for our data.

When we fit our model trained from the UCLA data onto the ACS data, we get a forecasting that Donald Trump will get 51% of the popular vote over Joe Biden. Looking at the variables with significantly small p-values of in our model, we generally see that older voters (30+), high income earners ($100k+), race groups (except white), and the Southern census region are the factors making the most impact. Taking a look at the confidence intervals for these significant factors, we also see that a majority of them lie within positive values suggesting that there is a strong case to be made for the voting popularity being in Trump's favor.

Results

Fig 12. Preference Among Voters With Key Factors



As discussed in the model section, we see there are quite a few variables that have significant positive impacts on our model due to their small p-values. This voter preference (Fig 12) comparison only includes voters who are older (30+), have an income of over $100k per year, and live in the southern census region. We see that there is a significantly greater proportion of voters who favor Trump over Biden. Despite more people favoring Biden in the entire sample, the factors that greatly affect the model clearly shows a significant number of people prefer Trump. Other factors that should be significant include the non-white race groups, which seem to mainly negatively impact the model, make a very small amount of the voter base as seen in the race group distribution (Fig 4) in the data section. Therefore, a slight favoring towards Trump's re-election makes sense when so many voters fitting the significant demographics prefer him over Biden.

Discussion

Our model consists of multilevel logistic linear regression with five main independent categorical variables that help us predict a binary outcome, which is the percentage of votes that Donald Trump will receive in the 2020 U.S .Presidential Election. With the intent of predicting the outcome of an election, we chose these five variables – age, education, income, ethnicity, and region – to be the most influential factors that determine the final decision of a registered voter. Given the ACS dataset, it did not provide the independent variable that we are looking for, which is the final vote decision of the respondent. In order to predict final outcomes given the demographic characteristics of respondents in the ACS dataset, we chose to create a generalized linear regression model using the UCLA dataset and apply it to same variables within the ACS dataset. This results in a more accurate representation of the voters across America, since the post-stratification dataset minimizes bias and variance, and also makes a more accurate prediction of the election.

When we look at the different demographics of registered voters, we found that voters that are between 30 and 59 years of age have the highest likelihood to vote for Trump due to the highest estimate of coefficient in the model and the extremely low p-values justifying that these coefficients of the model are highly significant. This is due to Trump's promises to protect current Medicare, Social Security plans, and other government benefits of older voters. This has led to large support from middle-aged working adults; but on the other hand, as of June 2020, voters that are 60+ of age have not been satisfied with how Trump is handling the on-going pandemic, especially with the increase of mortality rates among elders and the inability to deliver effective medicine/treatment for the COVID-19 virus. This has led to less people in favor of Trump's reelection within elders, which is reflected in our model with second lowest estimate of the age-group coefficient. For young voters between 18 and 29, have a different opinion of Trump's reelection. This could be due to the fact that the economy and finance is generally the young voters main concern, while health and social insurance is the main concern for older voters. On the other hand, Biden has been proposing various plans for student debt and education policies, such as making college/university tuition free for all families with income less than $125,000 and targeting additional financial support for low to middle class individuals. These propositions justify why voters for Biden are more than Trump for the younger generation as financial stress may potentially be relaxed in terms of student debt.

Given the Biden's propositions of loosening financial stress for younger voters, we transition into the next topic, education. We initially thought that the education level of respondents plays a significant role when voters make the final decision, which we justified by looking at the demographics of the 2016 U.S. election, where 48% of college educated white voters supported Trump and comparatively 66% of non-college educated white voters did not. As we explored the regression model of the given dataset, we concluded that all the education explanatory variables were insignificant with high p-values. Meaning that no matter if the voter did or did not complete high school or university has little affect as to which presidential candidate, he/she would vote for. While we can say

that college students that are constraint by student debt would prefer Biden's campaign title "The Biden Plan for Education Beyond High-School", we found that other characteristics of the voter are more significant than their education level when making a voting decision.

As income-gap distinction increases over the years, policies in terms of taxes are different between the two candidates. Former vice-president Joe Biden proposed to not raise taxes on Americans with an annual income of $400,000, yet raising income tax rates to almost 40% for those who make over that threshold income. Middle income Americans could also take an indirect hit from Biden's proposed tax policies since large corporations would be taxed more heavily, corporations may balance this cost by potentially increasing prices to consumers or lower wages of workers. Trump promised to cut taxes by an additional 300 billion by the end of his second term, primarily for the wealthy and corporate Americans. These two very different proposed tax policies impact the decision of voters, since the amount of annual income voters make directly affect the amount of tax, they are willing to pay. Simply concluding that people with a higher income are more likely to vote for Trump and people with a lower income are more likely to vote for Biden. From our regression, we can justify this result by looking at the house-hold income estimates and their corresponding p-values to see if it is significant enough to include in our model. Since our model targeted the likeliness to vote for Trump, we see that from the coefficients, voters with an income of less than $100,000 have very high p-values meaning that it is inconclusive of their voting result. On the other hand, people with household incomes of over $100,000 have significantly small p-values with a positive coefficient meaning that it is very likely that these individuals will vote for Trump.

As Trump continue to revoke policies implemented by the previous President, Barack Obama, including policies that were aimed to ensure the safety and minimize racial disparity of minorities amongst school, through our model we can see that voters that have the ethnicity of either Asian, African American, Pacific Islander, or an other race, will not vote for Trump. This is justified by the negative coefficients of these variables and the significantly low p-values, indicating that voter's that are of the race mentioned above will likely not vote for Trump. While majority of Trump supporters are White, our model suggests that the coefficient of the voter's race being white, has no significance due to the high p-value of 77.3%. This means that voters that are of the white descent will not have a clear distinction in terms of whom they will be voting for.

In terms of demographic region of the U.S., the southern region is mainly composed of red states whose voters predominantly choose the Republican party such as Alabama, Kentucky, Texas, Mississippi etc. Based on the results of past elections, the southern region is mostly comprised of states which are carried by the Republicans in the past four elections. This outcome can be seen by our model where the census region of the south being the only significant coefficient with a positive coefficient and a significantly small p-value. This means that voters of the southern region are more likely going to vote for Trump than Biden based on their demographic region and what the state mostly supports. On the other hand, the other census regions have a high p-value meaning that voters of the other regions are not significant enough to determine their voting outcome.

Weaknesses and next steps

A prominent weakness in our study consists of the fact that a conclusion on the overall popular vote doesn't necessarily coincide with the electoral votes. While the popular vote might be for Trump as concluded from our model, we are only conducting our survey on individual voter responses which makes it difficult to structure accordingly with electoral votes. This sort of drawback definitely hinders the confidence in our conclusion which is why a potential next step would be to organize the data in according to electoral districts and ridings for next time.

In addition, we also used a general linear model to make our forecasting. While this displays a relationship between all our independent variables with voter preference, it is hard to define clear relationships among our variables which could provide further insight into our conclusion. This especially coincides with our first weakness

since a potential angle of approach would have been to analyze the other socioeconomic factors with our region variable to get a sense of confidence for the candidates among different areas of the US. Manipulating the data this way would've strengthened our conclusion for a candidate as we can pinpoint where their votes are coming from and the type of supporters in different locations. Ultimately, the United States is a very large and diverse country with a variety different sub-cultures and issues too specific to generalize the country as a whole.

References

- R Core Team (2020). R: A language and environment for statistical computing. R, Foundation for Statistical Computing, Vienna, Austria. URL, https://www.R-project.org/ (https://www.R-project.org/).
- Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set (https://www.voterstudygroup.org/publication/nationscape-data-set).
- 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 (https://doi.org/10.18128/D010.V10.0)
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)
- Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven (https://CRAN.R-project.org/package=haven)
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr (https://CRAN.R-project.org/package=dplyr)
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Alexander, C., Dahl, S., & Weidman, L. (1997). Making Estimates from the American Community Survey. JSM Proceedings, Social Statistics Section (pp. 88-97). Alexandria, VA: American Statistical Association.
- Ball, M. (2016, October 25). Trump's Graying Army. Retrieved November 02, 2020, from https://www.theatlantic.com/politics/archive/2016/10/trumps-graying-army/505274/ (https://www.theatlantic.com/politics/archive/2016/10/trumps-graying-army/505274/)
- Daugherty, G. (2020, October 26). Biden's Plan for Student Debt and Education Policy. Retrieved November 03, 2020, from https://www.investopedia.com/biden-s-plan-for-student-debt-and-education-policy-5084094 (https://www.investopedia.com/biden-s-plan-for-student-debt-and-education-policy-5084094)
- Gandel, S. (2020, October 30). Comparing the Biden and Trump tax plans: Will you pay more? Retrieved November 03, 2020, from https://www.cbsnews.com/news/biden-tax-plan-comparison-trump/ (https://www.cbsnews.com/news/biden-tax-plan-comparison-trump/)
- Green, E., & Benner, K. (2018, December 17). Trump Officials Plan to Rescind Obama-Era School Discipline Policies. Retrieved November 03, 2020, from https://www.nytimes.com/2018/12/17/us/politics/trump-school-discipline.html (https://www.nytimes.com/2018/12/17/us/politics/trump-school-discipline.html)
- Gregorian, D. (2019, July 20). NBC News poll of the South: Voters' support for Trump grows, residents see race relations improving. Retrieved November 03, 2020, from https://www.nbcnews.com/politics/politics-news/nbc-news-poll-south-voters-support-trump-grows-residents-see-n1031851

(https://www.nbcnews.com/politics/politics-news/nbc-news-poll-south-voters-support-trump-grows-residents-see-n1031851)

- Harris, S. (2018, November 12). America Is Divided by Education. Retrieved November 03, 2020, from https://www.theatlantic.com/education/archive/2018/11/education-gap-explains-american-politics/575113/ (https://www.theatlantic.com/education/archive/2018/11/education-gap-explains-american-politics/575113/)
- https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html (https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html)