Predicting Joe Biden Will Win 10.2% of the Popular Vote in 2020 US Election*

Analysis is Conducted Using Multilevel Regression with Post-Stratification

Jasmine Carlos, Matthew Caringi, Haeun Choi, Mahmoud Elsheikh

2 November 2020

Abstract

This study discusses the prediction of the overall popular vote of the 2020 US election. In this particular paper, we observed the UCLA data and the American Community Surveys (ACS) and investigated four variables: respondents favourable candidate, racial group, education, and age. From our model, we discovered that the white individuals with below highschool education are more likely to vote for Donald Trump and the colored population with higher education are supporting Joe Biden. We predicted that the percentage that would vote for Joe Biden in the upcoming 2020 US Election is 10.2%.

Key words: forecasting; US 2020 election; Trump; Biden; multilevel regression with post-stratification

Note that all code and data used in this analysis can be found at the following link: https://github.com/jasminekcarlos/US-2020-election-prediction

The following report is written using R (R Core Team [2020]). The file was compiled using rmarkdown (Allaire et al. [2020]). The following packages were used: tidyverse (Wickham et al. [2019]), readr (Wickham et al. [2018]), dplyr (Wickham et al. [2020]), knitr (Xie [2020b]), ggplot2 (Wickham [2016]), kableExtra (Zhu [2020]), broom (Robinson et al. [2020]), here (Müller [2017]), bookdown (Xie [2020a]) and haven (Wickham and Miller [2020]). The following data sets were used: UCLA Phase 2 June 25, 2020 survey (Tausanovitch and Vavreck [2019]) and 2018 1-year ACS (Jose Pacas and Matthew Sobek [2020]). The following are citations for websites that were referenced: CNN [b], noa, CNN [a], cit [a], cit [b], cit [c], Tausanovitch et al., admin, Bureau and NW et al. [2020], full links being in the references section.

1 Introduction

The United States of America holds an election for the governing party and presidency position every four years. This election consists of the population of the USA to vote for a political party that they would like to govern the country. Two of the most favourable political parties include the Democratic and Republican Party. The president of the United States is currently Donald Trump who is a politician that represents the Republican party. President Trump was elected back in November 2016 when he faced off against his opposition, Hilary Clinton, who represented the democratic party. Currently, the democratic party leader and opposition candidate is Joe Biden. These parties differ on several political issues which will not be covered in this paper, but each political party's website can be found at cit [a] and cit [b].

Using the June 25, 2020 Phase 2 weekly survey that was conducted by UCLA Nationscape, a prediction of the overall popular vote for the 2020 US election will be examined. This investigation was done using predictor variables of age, race, and education level. The distribution of these variables for the survey are shown below as well as their relation to who they will vote for in the 2020 election, which is the variable of interest. Poststratification based on the American Community Survey of these predictor variables were used

*Code and data are available at: https://github.com/jasminekcarlos/US-2020-election-prediction.

to more accurately represent the population, then these reweighed variables were the basis for the multilevel logistic regression model. Our model with the post-stratification data revealed the influence of the predictor variables in reference to how likely they will vote for Joe Biden. The estimate of our model displayed that Biden will gain 10.2% of the total votes in the 2020 election but does not accurately represent whether this will lead to the presidency position due to the United States Electoral College System (**?**). Future work will need to examine the popular vote in each electoral college in order to gain a more insightful prediction of who will win the 2020 election.

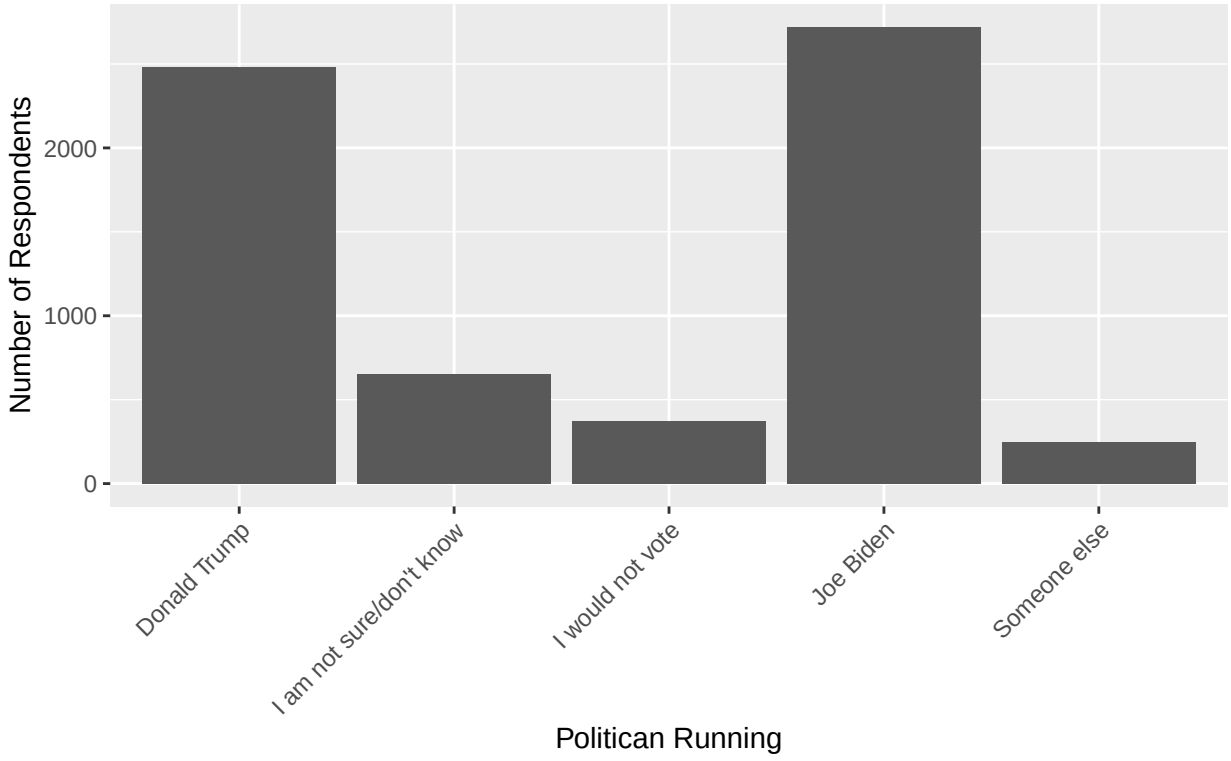
2 Data

The source of the data comes from UCLA Nationscape Phase 2 June 25, 2020 survey (Tausanovitch and Vavreck [2019]). To examine the validity of the survey, both the methodology and user guide are examined CITE (Tausanovitch et al.). The target population includes citizens who are 18 years old or older and reside in the United States of America. The sample population includes 6479 participants which exceeded UCLA Nationscape average weekly sample population for the survey which is roughly 6250. Collection of this data includes online interviews that were designed for a short time frame of 15 minutes. The participation rate for this survey was not mentioned but UCLA Nationscape did mention how their weekly surveys have roughly a 75% participation rate. Participants are found through Lucid, a market research platform, which assists many researchers to find participants by using suppliers, such as panels and communities, that represent a large variety of respondents (admin). The criteria of the respondents were created based on certain demographic characteristics that include race, age, education, and more. Collection of this data is described as a purposive sampling method which selects respondents based upon their characteristics in order to accurately represent the population of interest. The data was then processed and reweighed by a simple ranking technique to be more representative of the United States population. In the post stratification data, re-weighting of variables to illustrate the current population were based on previous official election results as well as the 2018 American Community Survey. One strength of this survey methodology includes the use of having suppliers to recruit large amounts of participants. This strength allows the reduction of nonrespondents which is a known major issue within interview studies. Additionally, the survey does not represent any leading or bias questions that may impact the survey data and focus on the main theme revolving aspects of the 2020 election. In opposition, one weakness about the sampling method includes that it does not contain any randomization on selecting respondents which introduces the influence of selection bias. Furthermore, the sample size of the survey is quite small for the target population and this is due to UCLA Nationscape conducting weekly surveys on the same topic of the 2020 election. Although combining these weekly surveys together can provide an exceptional sample size, unfortunately, only one survey is focused on.

Within the survey data (Tausanovitch and Vavreck [2019]), the variable of interest includes who the respondent will vote for in the 2020 election. The predictors variables include age, race/ethnicity, and education. In terms of the age and education variables, these categories were segmented into 5 groups. Additionally, the ethnicity variable was segmented into 6 groups. Within the "Asian" ethnicity group, this includes Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, and others that identify with the Asian ethnicity. Within the "Pacific Islander" ethnicity group, this includes Native Hawaiian, Guamanian, Samoan, and others that identify with the Pacific Islander ethnicity.

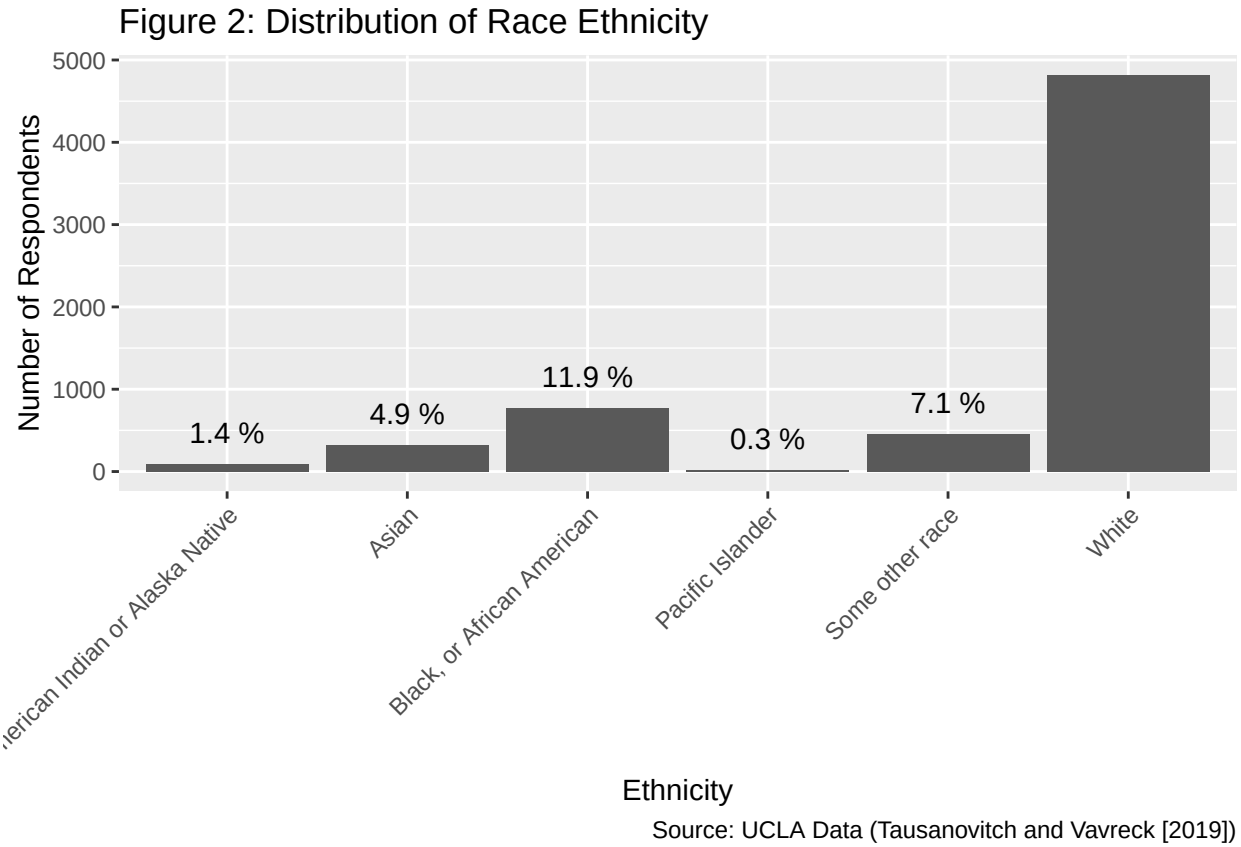
Using R Studio (R Core Team [2020]), the graphs listed below illustrate the distribution of the individual variables as well as their relation to the variable of interest.

Figure 1: Distribution of Favourable Candidates in 2020 Election

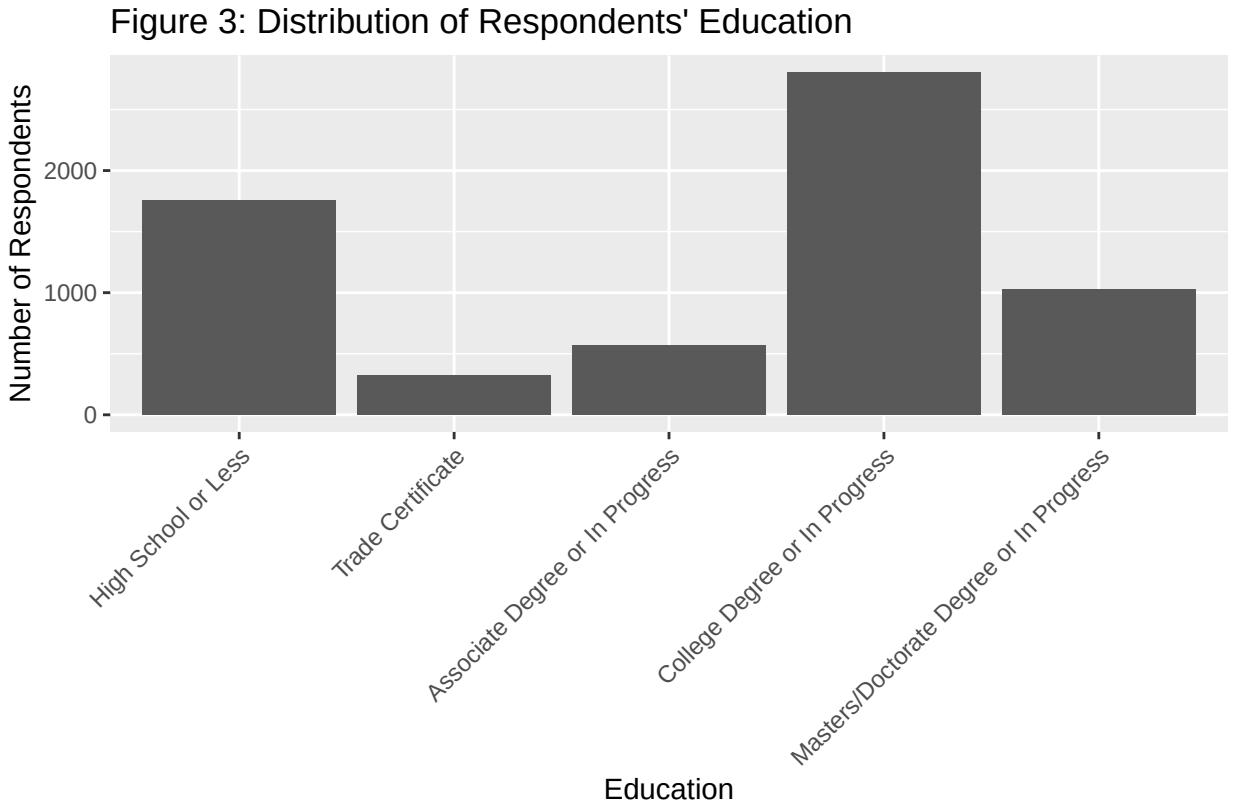


Source: UCLA Data (Tausanovitch and Vavreck [2019])

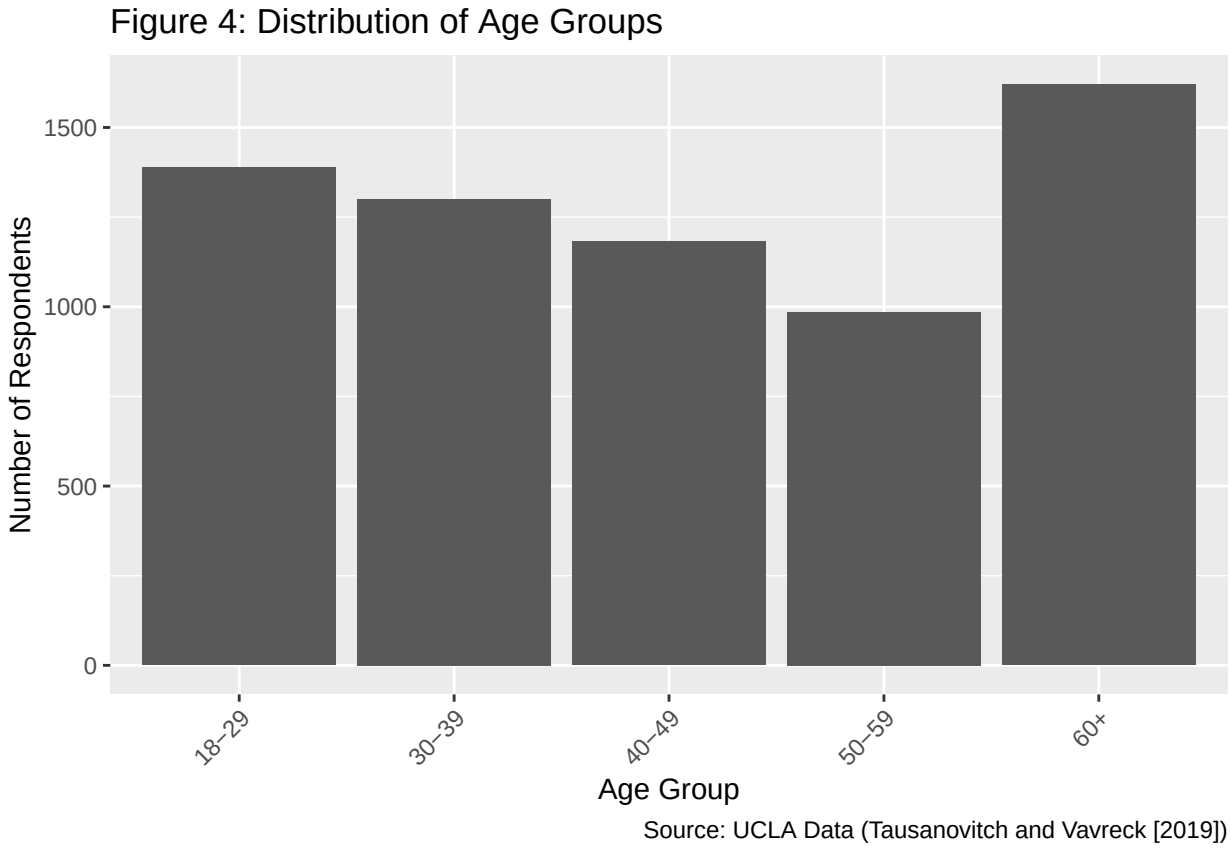
In Figure 1, the distribution of favourable candidates in the 2020 election are seen. The options on the x-axis include the politicians running as well as an option for non-voters and uncertain voters. Figure 1 shows Joe Biden leading Donald Trump by approximately 10%. The graph also shows a significant proportion of voters are not sure meaning the election could swing either way.



In Figure 2, the distribution of the respondents' ethnicity is displayed. This plot includes the proportions of each ethnicity in the sample population. The overwhelming majority of those surveyed are white at approximately 75% followed by African American's at 12%.



In Figure 3, the distribution of the participants' education is exhibited. The options on the x-axis include whether the respondents have completed the education level or are in process. This allows a more accurate distribution of education within the population, especially when the population includes the younger demographic. The largest proportion of those surveyed were college educated or in progress representing 44% of the surveyed subjects followed by high school or less representing 28% of respondents.



In Figure 4, the distribution of respondents' age is shown. It is acknowledged that the segmented groups are non-proportional to each other in terms of age bracket. This is due to the focus being more on the concept of comparing generations (young adults to seniors) rather than the focus being on the specific ages. In our bar graph representing the distribution of age groups, we have the age group of 60+ representing the largest proportion as the other groups are in 10-year intervals. The next largest age group in the 18-29 which is the demographic that is more inclined to vote for Joe Biden.

The poststratification data derives from the 2018 American Community Survey conducted by the United States Census Bureau (Jose Pacas and Matthew Sobek [2020]). The newest updated design and methodology report from 2014 will be examined (Bureau). The target population includes residents residing in the US and Puerto Rico. The frame of the sample includes households and group quarters which resulted in the sample population consisting of around 3.7 million participants. Participants were selected from the master address file (MAF) which consist of mailing addresses. Stratification sampling was used to segment blocks into multiple strata, where random samples were taken from each stratum. This sampling method allows under representative populations to be accounted for unlike in random sampling where these populations are overlooked due to chance. Collection of data was conducted by four different modes which include telephone, mailing, internet and personal interviews. Surveys are initially sent through mail and internet, when no response occurs then two possible options can happen. For addresses with available telephone numbers, then Computer Assisted Telephone Interviews (CATI) are conducted. If no telephone number is available or CATI was not completed, then Computer Assisted Personal Interviews (CAPI) are conducted. This process displays a strength in the survey methodology due to the multiple attempts to reach respondents which reduces the rate of nonresponses. This strength is shown with the 92.0% response rate for the 2018 ACS. Another strength includes the very large sample size which is appropriate for the target population of this survey. One weakness of the survey is the incorporation of data from Puerto Rico which can influence the aim of the target population and reduce its validity to accurately represent the US due to cultural differences.

Similar to the survey dataset, the poststratification data underwent grouping of the same variables. A matrix was created using four-loops to iterate all possible variable combinations in order to create proportions and

cell-level predictions. These predictions were displayed in a table to present the corresponding variables with their counts and portions.

3 Model

In order to carry out our analysis we began by constructing a model using the UCLA survey data. Note that the goal of this analysis is to be able to predict the outcome of the 2020 US election. In order to do this we fit a logistic regression, which is achieved by using the glm function in R. This function allows us to make use of a generalized linear model which is a flexible generalization of a regular linear regression. This model also allows us to clearly understand how predictors relate to the outcome as we are easily able to see coefficient and intercept estimates. Our model is of the binomial family as there are two outcomes, either Donald Trump or Joe Biden winning the election. While fitting the model we also made use of the as.factor command so that the categorical variables can easily be processed by the model. The model constructed is described by the following equation.

$$P(y = Biden) = \text{logit}^{-1}\left(\alpha^{race/ethn} + \alpha^{educ} + \alpha^{age}\right)$$

We are trying to calculate the probability of a given candidate to win so this was done by using the vote 2020 variable which indicates whether a respondent will vote for either Donald Trump or Joe Biden. We took these categorical variables and converted them into binary numerical variables in order to model them. Joe Biden is represented by the value of 1 while Donald Trump is represented by the value of 0. In order to predict which candidate would win we then used 3 predictor variables. The

$$\alpha^{race/ethn}$$

symbol in the model represents the variable of race/ethnicity, the

$$\alpha^{educ}$$

symbol in the model represents the variables of education and the

$$\alpha^{age}$$

symbol represented the variable of age.

Note that there were a few modification that were made for the sake of simplicity. In our model we have only one variables for race and ethnicity even though these can often be different. Also note that we chose to organize the data into just 6 racial groups while the original data had more groups. Something similar was done for age, filtering the age variables into age groups that span 10 years. This was once again done for the sake of simplification and being able to work with cleaner data. There are multiple variables we could have chosen but these are the ones we decided to input into the model as we believed they had a strong ability to predict the outcome of the election. This was proven to be true as all variables had a p-value less than 0.05 which means they were all significant to the model. There were other variables that may also be significant such as ideology but this was not included as there was not a way to reweigh this variable in the post-stratification step.

For our analysis we are not only fitting a logistical model, we are also using the method of multilevel regression with post-stratification (MRP). MRP is a method that is commonly used in order to adjust samples which are not representative in order to make more accurate predictions and conclusions. The first step in this method would be to build a regression model from survey data, which was what we did in our analysis. After this is done we use another set of data to relate individual survey responses to various different characteristics, training the model based on the new set of data. The largest advantage is that we are able to reweigh the data so that it is more representative of the population, this is particularly helpful when our initial model data is highly skewed. MRP is a very helpful method but there are also disadvantages. A notable one that

we encountered during our analysis is the fact that the data can easily become quite large. The reason for this is that we need proportions for every iteration of each variable. This makes it difficult to have a large amount of predictors in our model. We also ran into the problem of needing a reliable data set to use in the post stratification step. Something like a census is reliable for reweighing variables like age, race and gender but it can be difficult to effectively reweigh a variable such as a person's political ideology in order to better represent the population.

In our post stratification step we used the 2018 American Community Surveys. As previously mentioned in the data section we have our post-stratification data which includes the various cells, with corresponding counts and proportions for every iteration of the variables. We applied the post stratification data to our original model, creating estimates and using proportions in order to reweigh our variables to better represent the population. We were then able to calculate what proportion of the population is going to vote for Joe Biden in the upcoming 2020 US election.

4 Results

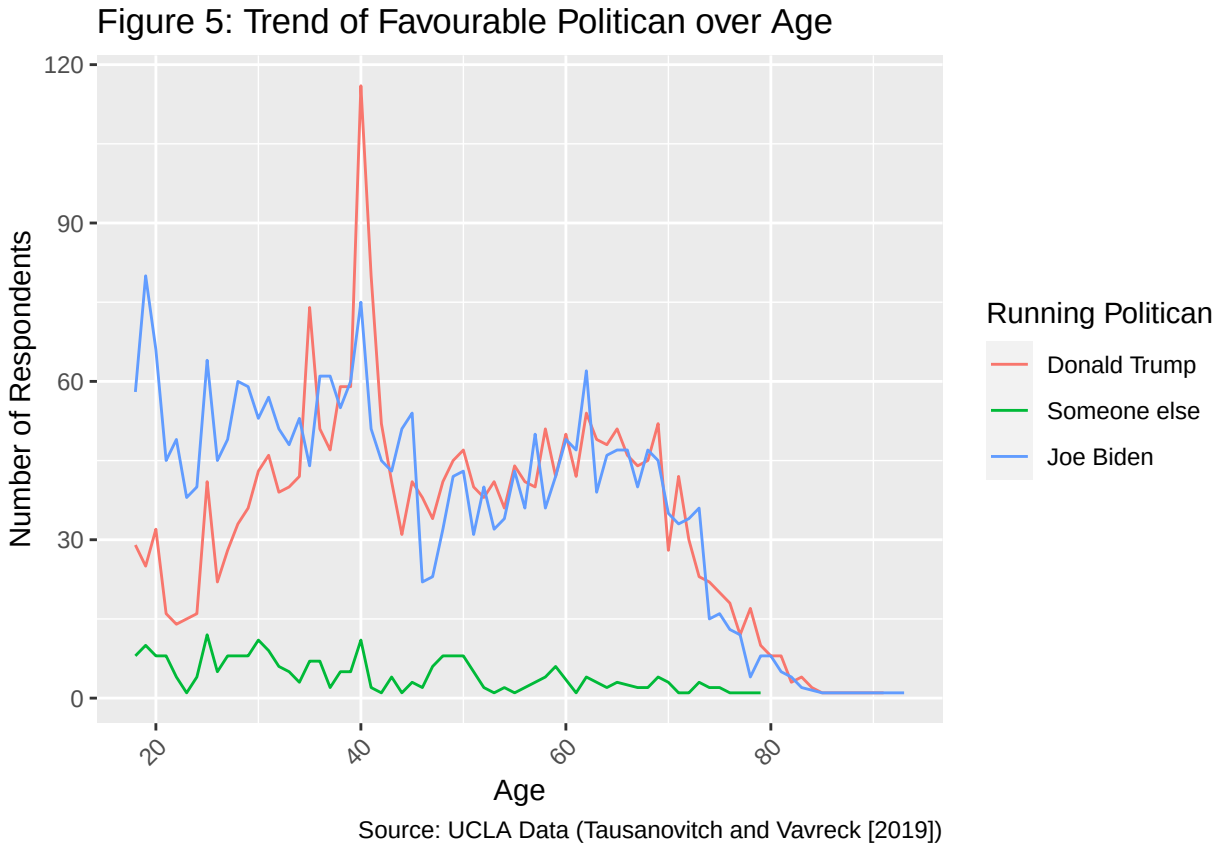
Table 1: Survey Summary Statistics

	vote_2020	race_ethnicity	education	age	age_grouped
	Length:6475	Length:6475	Length:6475	Min. :18.00	Length:6475
	Class :character	Class :character	Class :character	1st Qu.:31.00	Class :character
	Mode :character	Mode :character	Mode :character	Median :43.00	Mode :character
	NA	NA	NA	Mean :45.18	NA
	NA	NA	NA	3rd Qu.:60.00	NA
	NA	NA	NA	Max. :93.00	NA

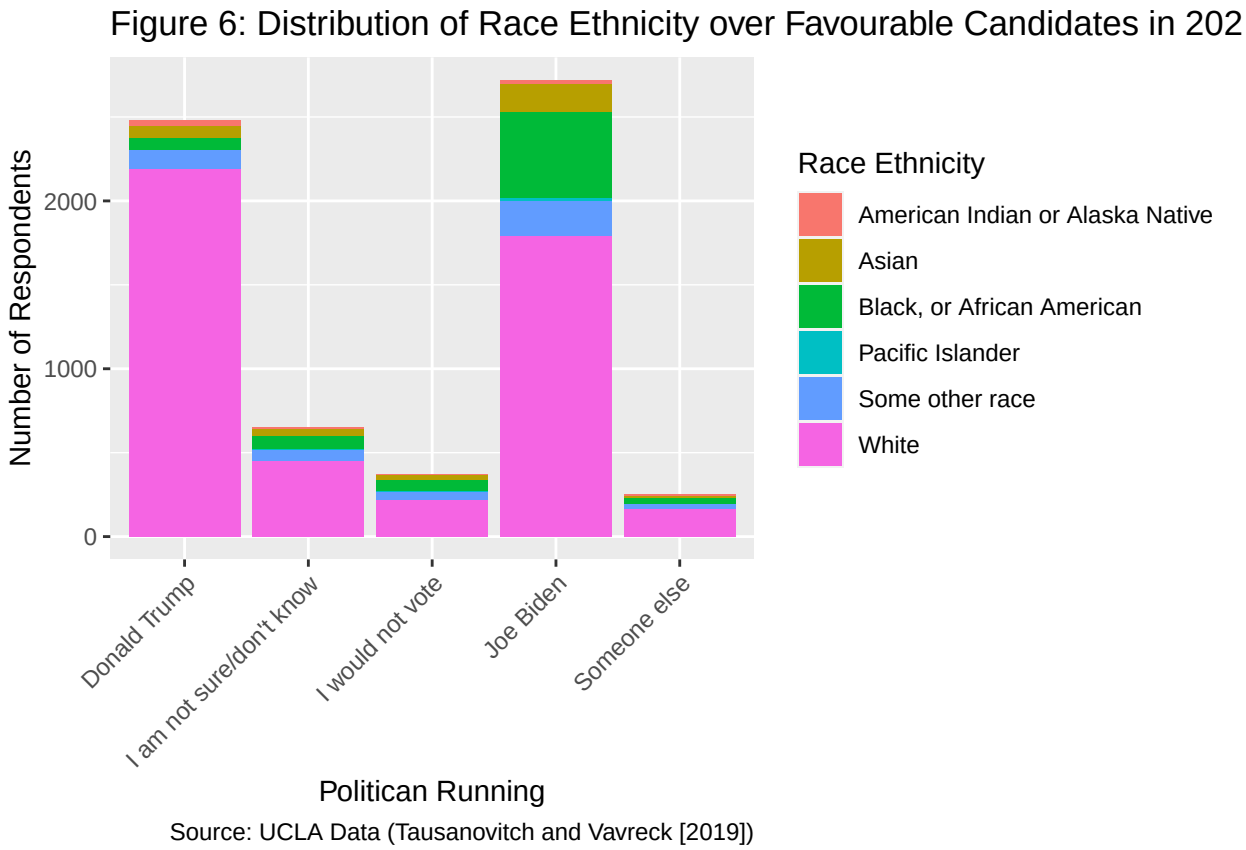
Table 2: Count For Preferred Candidate

vote_2020	count
Donald Trump	2481
I am not sure/don't know	651
I would not vote	374
Joe Biden	2719
Someone else	250

The summary statistics of the survey results show that Joe Biden had the popular vote by a margin of 53% to 47%. This survey was carried out with a total of 6475 entries whereas the census data used contained 460,000 entries.

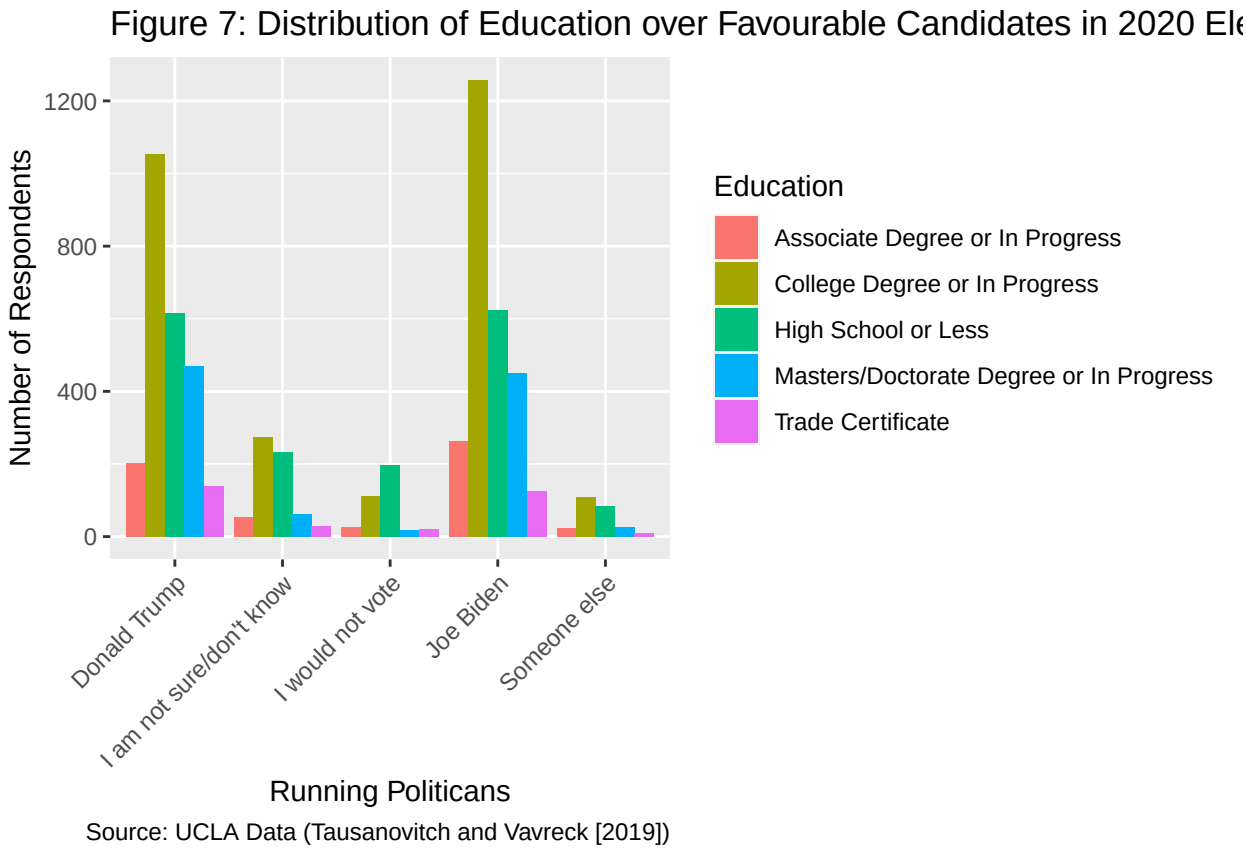


In Figure 5, the trend of favourable politicians are shown over age of the participants. It is to note that age is not grouped in this plot in order to visualize the trend more clearly. Joe Biden is also a favourite with younger voters in the survey as shown in the figure below. The line graph depicts that voters from the ages of 18-36 are more likely to vote for Joe Biden rather than Donald Trump.



Source: UCLA Data (Tausanovitch and Vavreck [2019])

In Figure 6, race ethnicity is shown in relation to favorable politicians. A stacked bar graph is used to visually support the emphasis of data from minority groups' instead of being overshadowed in a regular bar graph. One of the most significant demographic differences is the African American vote which disproportionately prefers Joe Biden clearly visualised in figure 6. Donald Trump has the majority of the white vote with a 55% to 45% advantage to Trump.



In Figure 7, education level is shown in relation to favourable politicians. Both candidates have approximately the same number of voters with a high school degree or less however this represents a larger proportion of Donald Trump's voter base than it does for Joe Biden.

Table 3: Summary of Model Statistics

term	estimate	std.error	statistic	p.value
(Intercept)	0.5431649	0.2858873	1.8999265	0.0574428
as.factor(race_ethnicity)Asian	0.9442532	0.3010353	3.1366855	0.0017087
as.factor(race_ethnicity)Black, or African American	2.2536875	0.2952376	7.6334707	0.0000000
as.factor(race_ethnicity)Pacific Islander	1.5825872	0.6309537	2.5082461	0.0121332
as.factor(race_ethnicity)Some other race	0.8949622	0.2913427	3.0718538	0.0021273
as.factor(race_ethnicity)White	0.0756269	0.2679453	0.2822477	0.7777536
as.factor(education)College Degree or In Progress	-0.1709640	0.1070026	-1.5977552	0.1100975
as.factor(education)High School or Less	-0.5371666	0.1164495	-4.6128737	0.0000040
as.factor(education)Masters/Doctorate Degree or In Progress	-0.2126279	0.1196515	-1.7770602	0.0755583
as.factor(education)Trade Certificate	-0.5099123	0.1637225	-3.1144919	0.0018426
as.factor(age_grouped)30-39	-0.5728486	0.0989846	-5.7872524	0.0000000
as.factor(age_grouped)40-49	-0.8039400	0.1023164	-7.8573893	0.0000000
as.factor(age_grouped)50-59	-0.7177855	0.1053844	-6.8111163	0.0000000
as.factor(age_grouped)60+	-0.5890993	0.0940859	-6.2612951	0.0000000

The model clearly shows the estimates of our variables age group, education and ethnicity to show how each demographic is predicted to vote. All our values are significant as they all have p-values under our threshold of 0.05. African American's showed a coefficient estimate of 2.25 showing a strong correlation between African Americans and their preference to vote for Joe Biden. White voters, on the other hand, show

a low correlation estimate to vote for Joe Biden at 0.0756 making them more inclined to vote for Donald Trump. The model also shows that more educated voters are more inclined to vote for Joe Biden with the strongest correlations being College degree at -0.171 and Masters/Doctorate at -0.21

Table 4: Percentage of Votes for Biden

win_predict
0.1023812

5 Discussion

The Presidential Election is often a significant event for many individuals. The 2020 US Election is particularly more in the spotlight since the world is going through a global pandemic and an economic slump so the US citizens are wanting a president who would take good care of the country. Since millions of Americans lost their jobs and the country is still suffering from the damages of the COVID-19, electing a candidate who would ensure the citizens' healthcare and return to their jobs would be supported. Since the world including America is in fighting through a global pandemic, this 2020 US Election is more sensitive and emotional than the previous elections.

In Figure 1, we can clearly see that Joe Biden is a more favourable candidate than Donald Trump which means that he is more popular among the American citizens. However, being the most popular candidate before the election does not mean that the candidate would win the electoral college.

If we take a look at Figure 5,6 and 7, the graphs display how Donald Trump is more popular among the white older population with high school diploma or less and Joe Biden is favoured by the younger, colored population that has higher education. In the 2016 American Election, older white voters played a great role in electing President Trump which means that in the 2020 American election, the voters who voted for Trump previously are highly likely to vote for him again.

In order to calculate the what percentage of the population would vote for Joe Biden in the upcoming 2020 US election, we applied the post stratification data to our original model. The estimate that was produced by our model predicts that 10.2% of the American voting population would vote for Biden. However, this is a rough estimate with limited variables, so in reality, Joe Biden would have a higher supporting percentage of the voters. Also, there were participants who preferred not to respond and also there were some N/A values which means that this estimate of the probability of Biden winning is not realistically true in the real world.

Joe Biden is popular among the female and minority groups since President Trump has been criticized for his racist comments throughout his presidency. Biden has a lot of colored supporters that believe the white people are benefiting more than the colored citizens in America. The group of young, and colored American citizens, are willing to support Biden because they have negative emotions towards Trump for his disrespect and discrimination towards other racial group than the white people. On the other hand, Trump's major voters are likely to be the supporters of Trump in the 2016 election. Since the older population of the America can align their beliefs and thoughts with President Trump, they have significantly shifted towards Donald Trump.

There were few weaknesses in this paper because of the modifications that were made in order to simplify the dataset. For example, we simplified the data of the racial groups into 6 specific groups but the original data had more groups. Similarly, in other variables, the data was filtered to create a cleaner data. However, the low p-values of the variables were all under 0.05 which means that the variables we chose are significant to this study.

In the future study, we might be able to add more variables to predict the 2020 American election. We did not include other variables such as voting intention and ideology because we couldn't find the variables in the census that can be used to reweight the two variables of the ideology and voting intention.

The limitations that we confronted in this paper is that popular vote does not necessarily tell us who is going to win the voting. For example, in the 2016 American Election, Hilary Clinton outpaced Trump by almost 2.9 million votes. Similarly, Hillary Clinton got more popular votes in 2016, compared to what Barack Obama got in 2012; however, Obama won on the actual Election day whereas Clinton did not. This indicates that predicting the 2020 US Election by investigating the popular votes might not be the most ideal way of predicting who is going to win more electoral votes.

References

We are the Democratic Party, a. URL https://democrats.org/.

Republican National Committee, b. URL https://gop.com/.

Presidential Election Process | USAGov, c. URL https://www.usa.gov/election.

Older White Voters Who Helped Trump Win May Be In Play. URL https://www.npr.org/2020/08/23/904755484/older-white-voters-who-helped-trump-win-may-be-in-play.

admin. MARKETPLACE. URL https://luc.id/marketplace/.

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2020. URL https://github.com/rstudio/rmarkdown. R package version 2.5.

US Census Bureau. Methodology. URL https://www.census.gov/programs-surveys/acs/methodology.html. Section: Government.

Analysis by Stephen Collinson CNN. Analysis: A nation in crisis faces a critical moment in history this election, a. URL https://www.cnn.com/2020/11/01/politics/election-2020-donald-trump-joe-biden-history/index.html.

Gregory Krieg CNN. It's official: Clinton swamps Trump in popular vote, b. URL https://www.cnn.com/2016/12/21/politics/donald-trump-hillary-clinton-popular-vote-final-count/index.html.

Steven Ruggles Sarah Flood Ronald Goeken Josiah Grover Erin Meyer Jose Pacas and Matthew Sobek. Ipums usa: Version 10.0 [2018 1-year acs], 2020. URL https://doi.org/10.18128/D010.V10.0.

Kirill Müller. *here: A Simpler Way to Find Your Files*, 2017. URL https://CRAN.R-project.org/package=here. R package version 0.1.

1615 L. St NW, Suite 800Washington, and DC 20036USA202-419-4300 | Main202-857-8562 | Fax202-419-4372 | Media Inquiries. Voters' Attitudes About Race and Gender Are Even More Divided Than in 2016, September 2020. URL https://www.pewresearch.org/politics/2020/09/10/voters-attitudes-about-race-and-gender-are-even-more-divided-than-in-2016/.

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

David Robinson, Alex Hayes, and Simon Couch. *broom: Convert Statistical Objects into Tidy Tibbles*, 2020. URL https://CRAN.R-project.org/package=broom. R package version 0.7.2.

Chris Tausanovitch and Lynn Vavreck. Democracy fund + ucla nationscape, oct 2019. URL https://www.voterstudygroup.org/downloads?key=3f8502d8-aa0d-4c59-a7f7-48c7f314d9b7.

Chris Tausanovitch, Lynn Vavreck, Tyler Reny, Alex Rossell Hayes, and Aaron Rudkin. Democracy Fund + UCLA Nationscape Methodology and Representativeness Assessment. page 19.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham and Evan Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, 2020. http://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat.

Hadley Wickham, Jim Hester, and Romain Francois. *readr: Read Rectangular Text Data*, 2018. URL https://CRAN.R-project.org/package=readr. R package version 1.3.1.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel,

Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2020. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.2.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2020a. URL https://github.com/rstudio/bookdown. R package version 0.21.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020b. URL https://yihui.org/knitr/. R package version 1.30.

Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2020. URL https://CRAN.R-project.org/package=kableExtra. R package version 1.3.1.