My title*

My subtitle if needed

Justin Teng, Mohammad Sardar Sheikh, Danur Mahendra

10 ¥|¤ë 2022

Abstract

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

1 Introduction

This report predicts the popular vote result of the 2020 United States presidential election using multilevel regression and post-stratification in R(R Core Team (2020)). As one of the most influential countries in the world, the US election does not only affect American citizens, it also impact those nations who depend on the United States for assistance, security, or trade, etc. This presidential election is between the incumbent Republican president Donald Trump and Democratic candidate Joe Biden, who was the former vice-president serving from 2008-2016. Trump's presidency began after the 2016 presidential elections against Democratic candidate Hillary Clinton. The 2016 presidential elections surprised the entire world, for Clinton had been heavily favored to win.(Shane Goldmacher (2016)) she led national polls and in most battleground states heading into the election. However, Trump's victory shocked major news outlets who considered him a significant underdog against Clinton.(cite) Trump will now attempt to win a second term of presidency against Biden, who many analysts again see Trump as the underdog following his controversial time in office (). His tenure was met with criticism following his controversial stance or remarks on racial inequality, diplomatic relations, and inefficient spending (cite). Trump's turbulent presidency can be attributed to his lack of political experience and thus, many believe that a second term will not see much improvement. Ultimately, poll analysis and forecast outlets will once again favour the Democrat over the Republican in this presidential election. This paper will attempt to use R to forecast our own prediction and analyse the main driving factors behind each vote.

For performing our prediction, we will be utilizing the data from U.S.presidential election 2020 survey data from Nationscape, which was conducted on June 25, 2020.

2 Data

This task requires multiple datasets to accurately predict the results of the 2020 United States popular vote. We first used Democracy Fund + UCLA Nationscape's December 2020 (Wave 76) data set to construct our multilevel regression method and followed it up by applying a post-stratification method using the IPUMS American Community Survey 2020 dataset. The ACS dataset would allow us to use our findings from the smaller Nationscape dataset to more accurately represent a much larger population.

Nationscape is a weekly online survey conducted by LUCID for Democracy Fund and UCLA researchers. Data for this wave's dataset was collected between December 24 - 30, 2020 and received 6,692 samples. Each

*Code and data are available at: LINK.

wave must collect a set of demographic quotas based on the respondents' age, gender, race/ethnicity, region, income, and education. The quotas are based on the U.S. adult population in 2017 provided by the U.S. Census Bureau. Respondents submit their responses through online survey software provided by LUCID.

The ACS survey is a monthly rolling survey used to update census estimation for the Census Bureau. The ACS uses two sampling frames both provided by the Census Bureau, housing unit (HU) addresses and residents of group quarter (GQ) facilities. Samples were collected by a method of stratified sampling. Respondents were then contacted to complete the survey via either Computer-Assisted Personal Interview (CAPI) or Computer-Assisted Telephone Interview (CATI). The ACS samples include roughly 3 million households with each sampling unit representing a household and all persons residing in the household.

The original datasets that we received needed to be cleaned so that they could be used effectively in our analysis. We used separate scripts to clean the datasets. First, we read in the datasets and then we choose the variables that we require. Since we are planning to do a regression with the data that we have, we cannot have any NA (missing) values in the cleaned datasets, so we remove all the rows that contain NA values. For the UCLA dataset, we create a new variable that equals 1 if a respondent chooses to vote for Donald Trump, and 0 if the respondent plans to vote for Joe Biden. We filter out all the observations that do not refer to either of these two candidates, for the sake of more accurate results. We then need to make sure that the common variables in both the datasets have equivalent values. We factorise race as a variable and we categorise it according to research and looking at the codebook for both the datasets. We categorise race into 5 parts, white, african american or black, asian and pacific islanders, native american, and others.

Regarding education, we split it into 3 categories. The first one is pre-high school (the respondent has not received a high school diploma), the second is high school diploma or equal, and the third is college diploma or higher. Since we are interested in how the Hispanic community votes, we convert being hispanic into a binary variable, much like voting for Trump, where a value of 1 signals that the respondent is Hispanic, and 0 says otherwise. We do a very similar thing with gender, choosing to classify females as 1 and males as 0.

Finally, we want to make sure that the value of States are the same for both the datasets. For this, all the unique strings need to match up accordingly, so "New York" has to be the same in both the datasets. It wont work if its "New York" for one and "new york" for the other.

3 Model

We are interested in forecasting the popular vote result of the 2020 United States presidential election. In Particular, we want to predict the proportion of voters that will vote for the Republican candidate, Donald Trump. To achieve this, we used the December 2021 Nationscape dataset to find a relationship between population characteristics and their vote intention. We then used the 2020 ACS dataset to apply a post-stratification method. This technique allows us to fit a smaller data set to match one that would more accurately represent a much larger population. In this case, the population we are trying to represent is the American voting population.

We used a logistic regression model to estimate the probability of a voter voting for Trump given certain characteristics (represented by predictor variables). An individual's voting intention is represented by a binary response variable in our data set thus, we believe that a logistic regression model best suits this task. The reason we don't choose a linear regression model is because we believe that a straight line will not accurately represent the data. A logistic regression line has a curved 'S' shape, and this is more suited to predict and calculate binary values that can only either take 0 or 1 as responses. This is crucial as the post-stratification process is contingent on choosing the most appropriate model. Choosing a model that breaks assumptions may render our post-stratification and prediction inaccurate. The response variable in our model denotes whether the voter intends to vote for Republican candidate Trump. We get a probability (p) as the response so we assume that (1-p) is the probability that a respondent intends to vote for Joe Biden.

The predictor variables used are the voter's age, state of residence, gender, race, income, whether the voter is Hispanic or not, and highest level of education attained as we believe they are key factors that form one's

political views. We believe that age is an important variable as most people's political views change with age, with some particular age groups preferring one candidate over the other. States are a big part of the model, there are some states that are very pro Trump, and some states that are very anti-Trump. We feel like incorporating these differences in views is a crucial part for the model. Another variable that we feel is important to include is gender. A respondents gender can influence the way they perceive society and can affect their political views. There is a slight difference in the gender make-up of the country so we feel like we should include this as a variable. Education is another important variable that we want to include in the model. Educated people are more aware of the way society works and what is going on around them. They are less likely to make decisions based on passion and feelings and are more likely to think before they do something important, such as decide which candidate to support. The difference in viewpoints that they bring with them seems like an important thing to not include in our model. We decide to include whether a respondent is Hispanic or not in the model as we want to see whether these specific people will choose to vote for Trump, even after all the statements that he has made against them such as "I would build a great wall, and nobody builds walls better than me, believe me, and I'll build them very inexpensively. I will build a great great wall on our southern border and I'll have Mexico pay for that wall." (Donald Trump). We want to see whether statements like these have had any effect on his popularity. Finally, we incorporate race into the model. People from different racial backgrounds have their own personal agendas, choosing to side with a specific candidate over the other.

Incorporating all of these variables, our logistic regression model can be represented by the following formula:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{gender} + \beta_2 x_{age} + \beta_3 x_{race} + \beta_4 x_{stateicp} + \beta_5 x_{education_category} + \beta_6 x_{hispanic} \qquad (1)$$

Equation (1) is not the full representation of the model as we have not shown the full model to account for space. $\beta_4 x_{stateicp}$ actually refers to all the different states that we have in the data set (50 of them), and each state has its own unique β value. Similar for education and race.

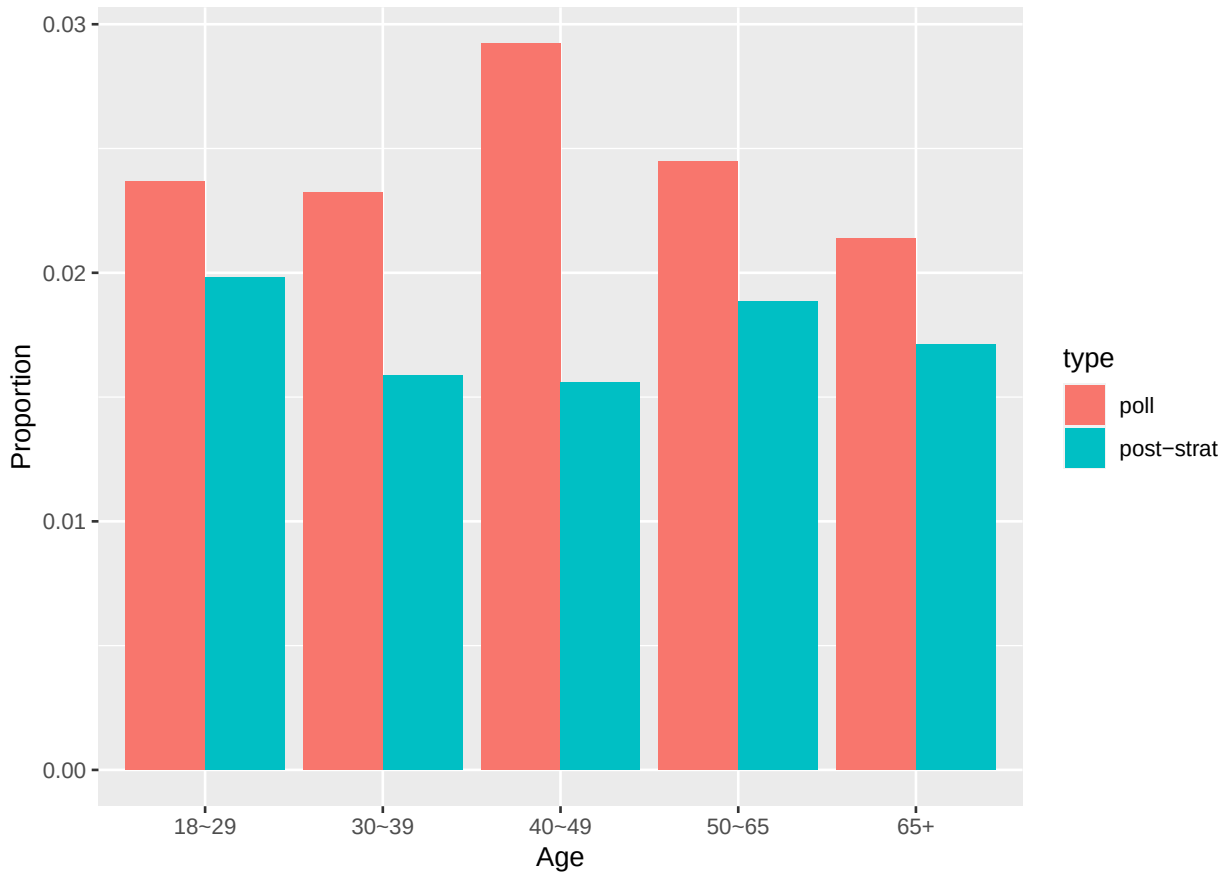

Figure 1: Voter's Demographic: Age (poll)

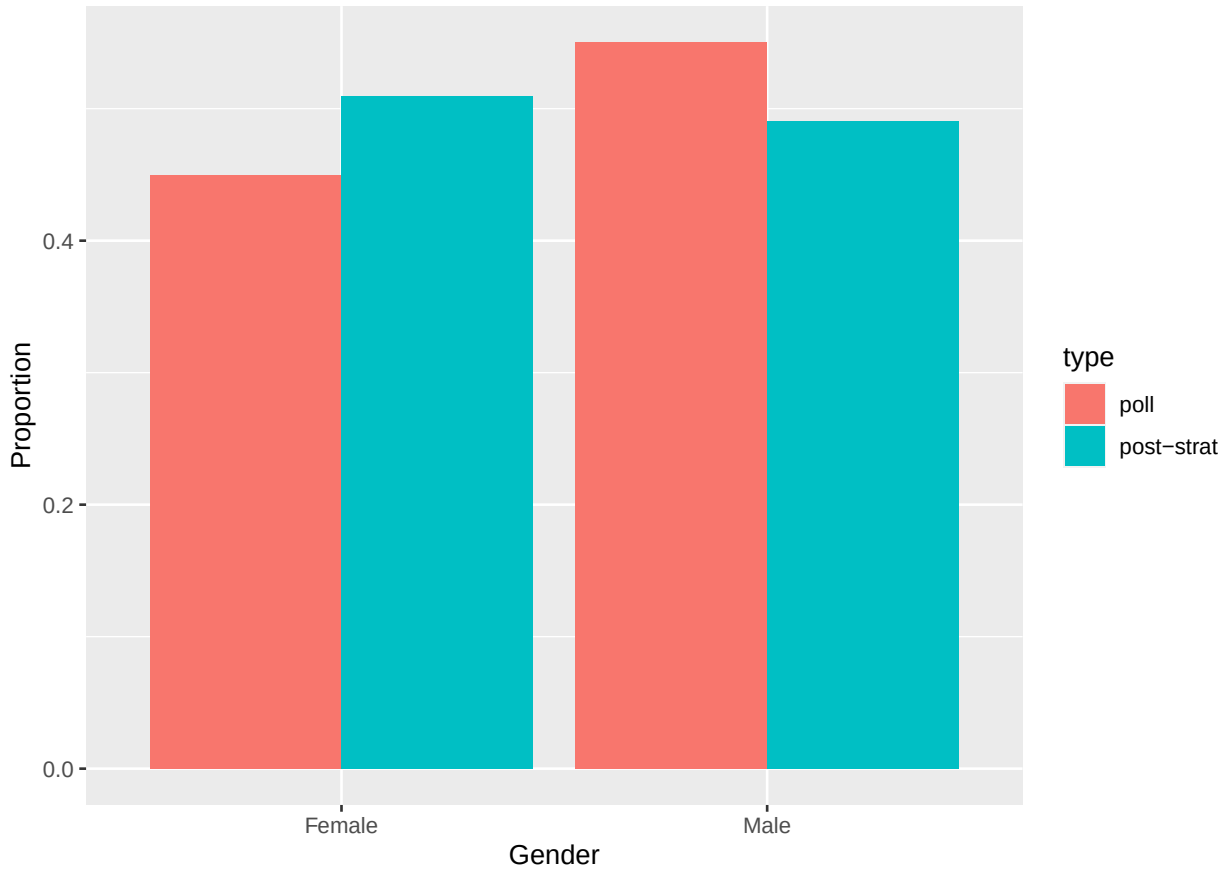

Figure 2: Voter's Demographic: Gender

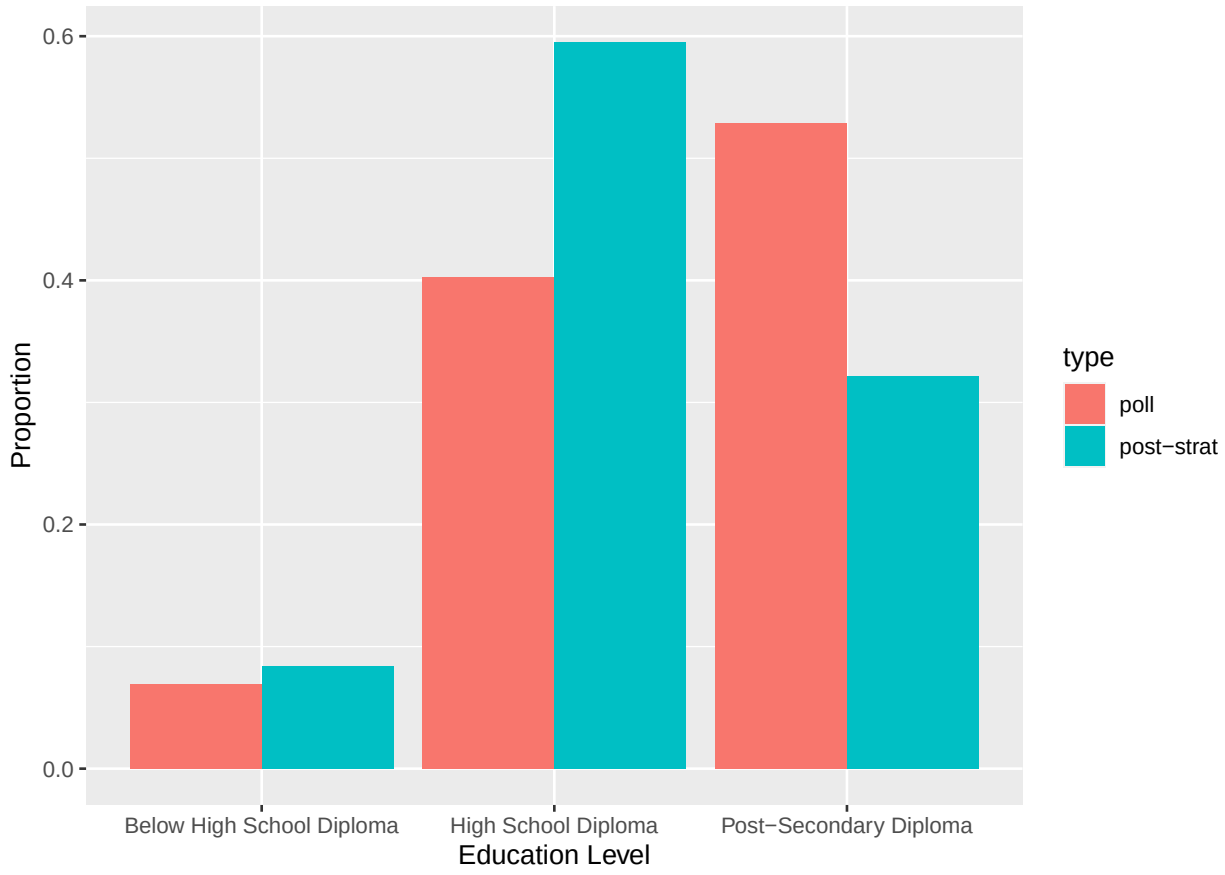

Figure 3: Voter's Demographic: Education

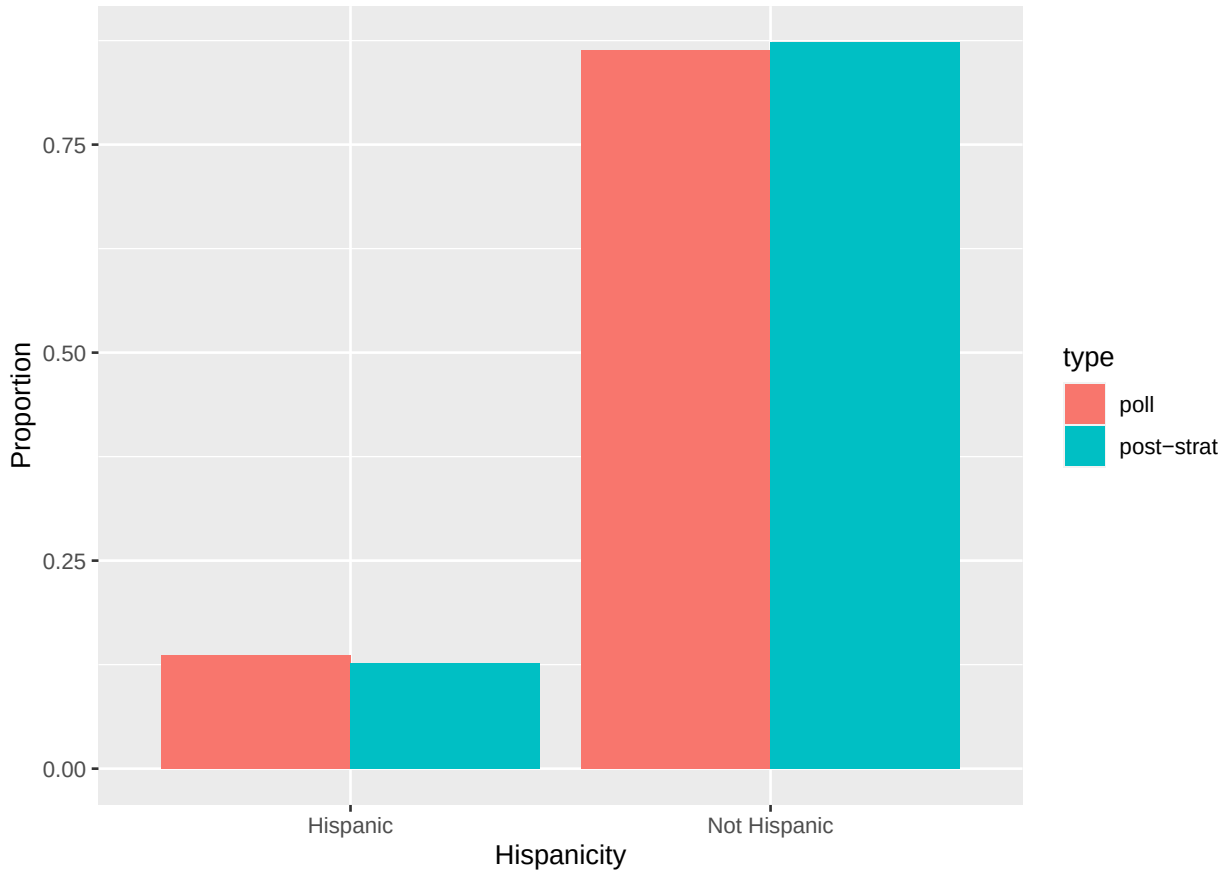

Figure 4: Voter's Demographic: Gender

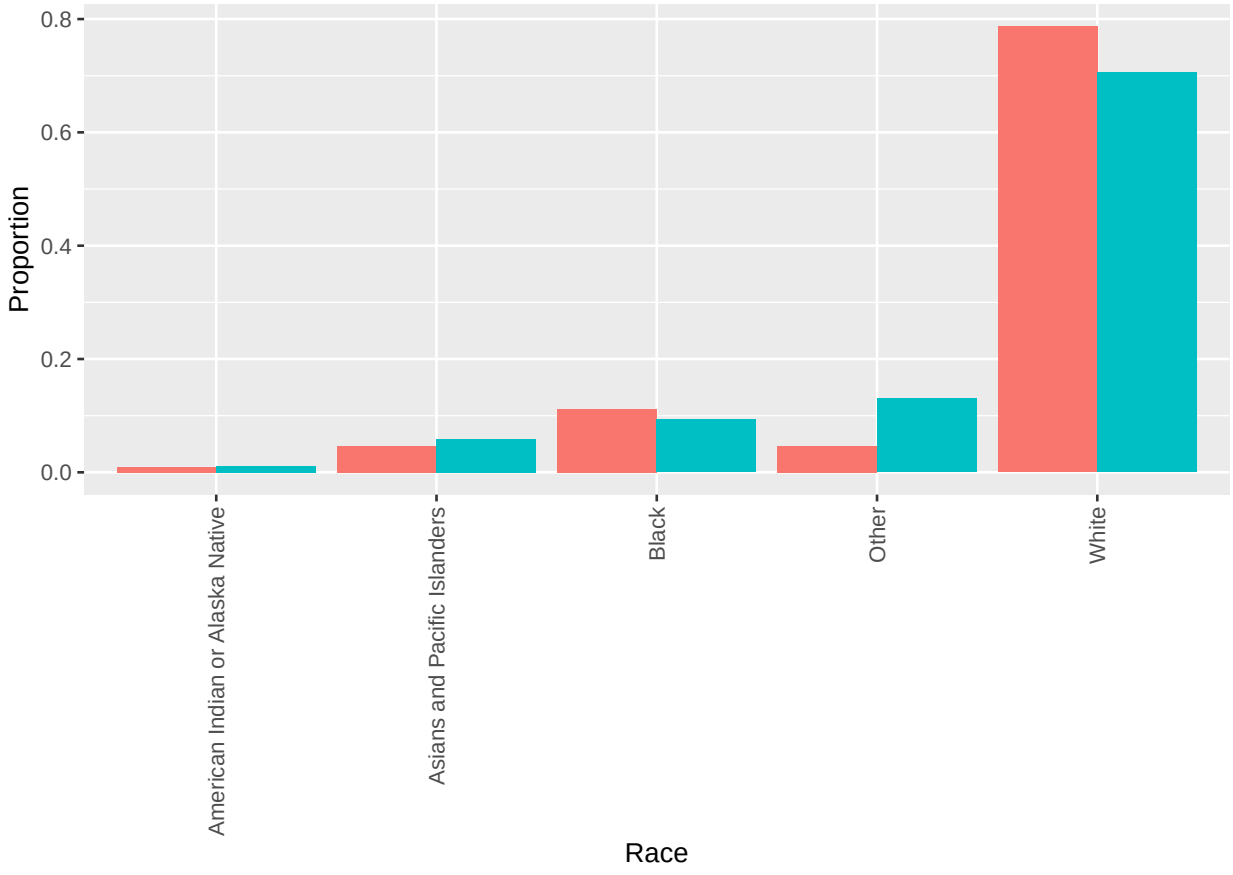

Figure 5: Voter's Demographic: Race

Table 1: Coefficients from the Model

predictor	estimate	standard error	statistic	p-value
(Intercept)	1.165	0.340	3.429	0.001
age	0.014	0.002	6.778	0.000
as.factor(race)2	-2.007	0.146	-13.764	0.000
as.factor(race)3	-0.304	0.322	-0.946	0.344
as.factor(race)4	-1.078	0.184	-5.844	0.000
as.factor(race)5	-0.885	0.165	-5.354	0.000
as.factor(education_category)2	-0.291	0.131	-2.212	0.027
as.factor(education_category)3	-0.769	0.130	-5.890	0.000
sex	-0.347	0.063	-5.497	0.000
stateicpAlaska	-1.151	0.690	-1.669	0.095
stateicpArizona	-1.297	0.365	-3.555	0.000
stateicpArkansas	-0.425	0.484	-0.879	0.379
stateicpCalifornia	-1.406	0.323	-4.350	0.000
stateicpColorado	-0.959	0.364	-2.633	0.008
stateicpConnecticut	-1.866	0.435	-4.292	0.000
stateicpDelaware	-1.740	0.636	-2.737	0.006
stateicpFlorida	-1.354	0.326	-4.151	0.000
stateicpGeorgia	-0.901	0.350	-2.573	0.010
stateicpHawaii	-0.423	0.555	-0.762	0.446
stateicpIdaho	-0.600	0.520	-1.155	0.248
stateicpIllinois	-1.316	0.338	-3.889	0.000
stateicpIndiana	-1.294	0.366	-3.533	0.000
stateicpIowa	-1.314	0.420	-3.128	0.002
stateicpKansas	-0.875	0.427	-2.051	0.040
stateicpKentucky	-0.678	0.399	-1.700	0.089
stateicpLouisiana	-1.075	0.451	-2.382	0.017
stateicpMaine	-1.209	0.563	-2.147	0.032
stateicpMaryland	-1.624	0.399	-4.069	0.000
stateicpMassachusetts	-2.126	0.401	-5.296	0.000
stateicpMichigan	-1.385	0.342	-4.050	0.000
stateicpMinnesota	-1.266	0.379	-3.344	0.001
stateicpMississippi	-0.653	0.511	-1.277	0.202
stateicpMissouri	-0.924	0.366	-2.524	0.012
stateicpMontana	-0.721	0.730	-0.987	0.324
stateicpNebraska	-0.981	0.490	-2.001	0.045
stateicpNevada	-1.710	0.426	-4.016	0.000
stateicpNew Hampshire	-1.762	0.711	-2.476	0.013
stateicpNew Jersey	-1.656	0.361	-4.585	0.000
stateicpNew Mexico	-1.363	0.530	-2.572	0.010
stateicpNew York	-1.520	0.331	-4.598	0.000
stateicpNorth Carolina	-1.244	0.353	-3.522	0.000
stateicpNorth Dakota	-0.241	0.760	-0.317	0.751
stateicpOhio	-1.168	0.338	-3.452	0.001
stateicpOklahoma	-0.479	0.437	-1.095	0.274
stateicpOregon	-1.029	0.393	-2.620	0.009
stateicpPennsylvania	-1.215	0.337	-3.608	0.000
stateicpRhode Island	-2.402	0.655	-3.669	0.000
stateicpSouth Carolina	-0.614	0.379	-1.619	0.105
stateicpSouth Dakota	-1.299	0.643	-2.020	0.043
stateicpTennessee	-0.678	0.374	-1.810	0.070
stateicpTexas	-1.179	0.328	-3.596	0.000
stateicpUtah	-0.258	0.467	-0.551	0.581
stateicpVermont	-14.314	216.129	-0.066	0.947
stateicpVirginia	-1.207	0.362	-3.338	0.001
stateicpWashington	-1.823	0.358	-5.087	0.000

4 Results

4.1 Voter's Demographics

4.1.1 Age

4.1.2 Gender

4.1.3 Education

4.1.4 Hispanic

4.1.5 Race

4.2 States

4.3 Regression Results

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Appendix

A Additional details

References

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

Shane Goldmacher, Ben Schrekinger. 2016. "Trump Pulls Off Biggest Upset in u.s. History." *Politico*. Politico. https://www.politico.com/story/2016/11/election-results-2016-clinton-trump-231070.