Biden Predicted to Secure a Minimum 10% Lead in Popular Votes Over Trump in the 2024 Election*

Boxuan Yi

11 March 2024

Abstract

The dominance of only two political parties in the current US political system adds an extra layer of competitiveness and interest to the presidential election. This paper uses multilevel regression with post-stratification to predict the outcome between the two candidates, Biden and Trump. The model, along with the post-stratification, indicates that Biden is poised to secure a substantial lead of at least 10% in popular votes, suggesting a strong likelihood that the US will continue on a democratic pathway.

Table of contents

*Code and data in this analysis is available at: https://github.com/Elaineyi1/2024prediction_USPresidentialElection

1 Introduction

The forthcoming 2024 United States presidential election, scheduled for Tuesday, November 5, 2024 (270ToWin), will see US citizens electing to determine the nation's president and vice president, shaping the pathways of the country for the next four years. The current president Joe Biden, a member of the Democratic Party, is running for re-election, and his predecessor, Donald Trump from the Republican Party, is running for re-election for a nonconsecutive term. Throughout campaigns, political parties articulate their stances and future policies on broader issues such as abortion, immigration, safety, and LGBTQ+ rights. The Democratic and Republican parties typically hold distinct yet steadfast viewpoints, attracting a significant number of supporters among US citizens. However, the post-COVID era has brought about significant changes economically, socially, and politically, making it particularly intriguing to see how these shifts will influence the upcoming 60th quadrennial presidential election.

I will use multilevel regression with post-stratification (MRP) to model and then predict whether Biden or Trump will win the popular votes, which is the total number of votes cast by individual citizens. MRP in this paper involves collecting individual-level survey data from the Polarization Research Lab, analyzing the proportion of support among different demographical groups, and applying the proportion to a larger census dataset from the IPUMS USA, 2022 American Community Survey (Team). One thing that needs to be pointed out is that rather than just depending on the popular votes, the United States elects its president though electoral college. The 51 states are allocated 538 electoral-college votes. The numbers of votes granted to each state, which are different, depend on its representation in the Senate and House of Representatives, and usually every state gives all of its electoral-college votes to whomever wins the popular vote in that state (TheEconomist 2016). In other words, it is possible that one candidate wins the popular vote, representing the collective choice of voters, but loses the presidential election because of the distribution of electoral college votes, like the election in 2016 where Hillary Clinton won the popular votes (48%) while losing the election due to the states that supported her have fewer electoral college votes than the ones supporting Trump (TheNewYorkerTimes 2017).

In this paper, I will only predict the popular votes for two candidates, Joe Biden and Donald Trump, utilizing predictors as follows: race, gender, age group, highest level of education, the state of residence. Two datasets will be generated from the survey data —— one being simply cleaned, and the other cleaned and balanced. The latter involves adjusting the survey respondents' votes in 2020 to align with the actual 2020 results while keeping other characteristics the same. Since the balanced dataset has fewer observations than the original one, this paper focuses on the unbalanced dataset to avoid potential sampling biases. Using the unbalanced dataset, my prediction suggests that Joe Biden is poised to win 62.96% of the popular votes, and 57.61% using the balanced one. Both percentages indicate that Biden is going to secure a minimum 10% lead in the popular votes, which is a pretty large lead in election history.

In this paper, I will delve into the data utilized for analysis, encompassing both survey and census data, in the upcoming section. Following that, I will introduce the model designed to generalize patterns and forecast election outcomes. I will also talk about the predicted results derived from the model. Lastly, I will discuss the results in a broader context and address weaknesses of this paper as well as the future explorations. This paper use the programming language R (R Core Team 2022). The analysis, the model and the visualizations use the following packages: ipumsr (Greg Freedman Ellis, Derek Burk, and Finn Roberts 2024), dplyr (Wickham et al. 2023), readr (Wickham, Hester, and Bryan 2024), tidyr (Wickham, Vaughan, and Girlich 2024), arrow (Richardson et al. 2024), ggplot2 (Wickham 2016), knitr (Xie 2014), statebins (Rudis 2020), modelsummary (Arel-Bundock 2022), broom (Robinson, Hayes, and Couch 2023), here (Müller 2020), and kableExtra (R Core Team 2022 kableExtra).

2 Data

2.1 Survey Data and Variables Selection

The survey data I use is from the America's Political Pulse Survey by the Polarization Research Lab (Shanto Iyengar 2023). This open-access online survey contains a series of core questions on politics and special

questions designed by political science researchers, posing inquiries to a thousand Americans every week. All the respondents, the population, are paid survey takers from the YouGov survey platform. It collects demographic information, party each respondent support, and core questions like Democrat/Republican thermometer or pride in being American.

The datasets are available weekly. For this paper, I choose 20 weeks of data, spanning from October 6th, 2023 to February 23rd, 2024, forming a dataset with 20000 observations and 69 variables. The variable 'pid3' shows the answers to 'Generally speaking, do you think of yourself as a...?', with responses being Democratic, Republican, or Independent. For this paper, I will use this variable as supporting the Democratic Party/Biden or Republican Party/Trump for analysis.

In order to predict the 2024 presidential election result using several predictors, I first need to choose variables that can effectively help distinguish between Democrats and Republicans. The survey dataset results show that females tend to support Democracy more than males, with 63.9% female and 58.2% male respondents supporting Democracy. People with higher levels of education tend to support the Democratic Party, especially those who completed post-graduate studies. 55.1% of high school graduates and 60.1% of people with no high school support the Democratic Party, while 72.8% of post-grads support it. More widowed and married people consider themselves as Republicans, while more people who are divorced, never married and separated consider themselves as Democrats. Different ethnicities also show different preferences. Around 74.8% of Asians and 79.7% of African Americans support Biden, compared to 55.7% of white people. Additionally, since the presidential election is voted by state, I would choose race, gender, marital status, highest level of education and state of residence as the variables to explain the binary support for Biden.

The voter turnout rates for different age groups are different. Most of the time, the turnout rate among 60+ years old is the highest, followed by people between 45 to 59, but this is still about 10% higher than the rate among 30-44 years old, and the people from 18 to 29 have the lowest rate of voting for the past 30 years (OurWorldinData). Hence, I will also use age groups as another factor that affects the popular vote outcome, and the grouping is 18-29, 30-44, 45-59, 60+.

About 22% of individuals identifying with both the Democratic and Republican parties consider voting important, and a significant 70% from each party consider it very important. In this survey, both Democrats and Republicans exhibit similar perceptions of voting importance, indicating that the the survey doesn't inaccurately estimate voting due to differing levels of importance attached to it by the parties. For all the Republican respondents, 55.7% are extremely proud to be Americans, 27.7% are very proud, while only 1.4% are not proud. For all the Democrat respondents, 33.3% are extremely proud, 25.5 are very proud, and 7.3% are not at all proud. Republicans seem to be more polarized as 52.5% of them consider themselves as MAGA Republicans. 22.9% of Democrats think they are very liberal, 36.1% are liberal, while 27.65% of Republicans consider themselves as very conservative, and 44.0% are conservative.

There are also significant differences in religious beliefs. 80% of Republicans think religion is important in their lives (54% very important, 26% somewhat important), while 54.8% of Democrats see it as important (31.1% very important, 23.7% somewhat). Half of the Republican respondents think they are "born-again," or evangelical Christian, while only 26.5% of Democrats think so. 38.3% of Democrats are 'Agnostic,' 'Atheist,' or 'Nothing in particular,' and 47.6% are 'Protestant' or 'Roman Catholic.' However, only 17.2% of Republicans are 'Agnostic,' 'Atheist,' or 'Nothing in particular,' and 71.2% are 'Protestant' or 'Roman Catholic.' Since religion is not collected in the census data, I will not use this factor, but this could be a potential predictor as well.

2.2 Post-Stratification Data

IPUMS USA is a website that contains decennial census from 1790 to 2010, as well as survey data from the American Community Surveys (ACS) and the Puerto Rican Community Surveys from 2000 to the present (Team). The annual ACS public-use samples typically include housing data, demographic data, economic data, and some other individual characteristics. For this analysis, I will use the most recent American Community Survey, the 2022 Sample, as the post-stratification dataset to represent the voter population and adjust the weight of the survey data. This weighted survey used cluster sampling to obtain 1-in-100

national random sample of the population. The population for this sample was the entire US population, categorized into households as subgroups, and the responses of the sampled subgroups were collected.

ACS 2022 dataset consists of 3,373,378 observations, making it a large and relatively representative dataset. The variables I select are gender, age, marital status, race, education and stateicp (ICPSR codes of states) to match with the survey data variables I choose. I then clean and rename the variables to maintain unity with the survey data. Figure 1 and Figure 2 display the voters' demographics in both survey data (pre-strat) and census data (post-strat), with solid lines representing the post-stratification data and dotted lines representing the survey data. The differences in the two lines are due to weight adjustment. Some noticeable adjustment in Figure 1 include more people aged 30 to 44, more females, fewer married people, fewer asians, and more African American completing the survey. Figure 2 classifies people based on education and state of residence. There are more people from Florida, New York and Texas in the survey, while states like Wyoming have very few participants, potentially bias.

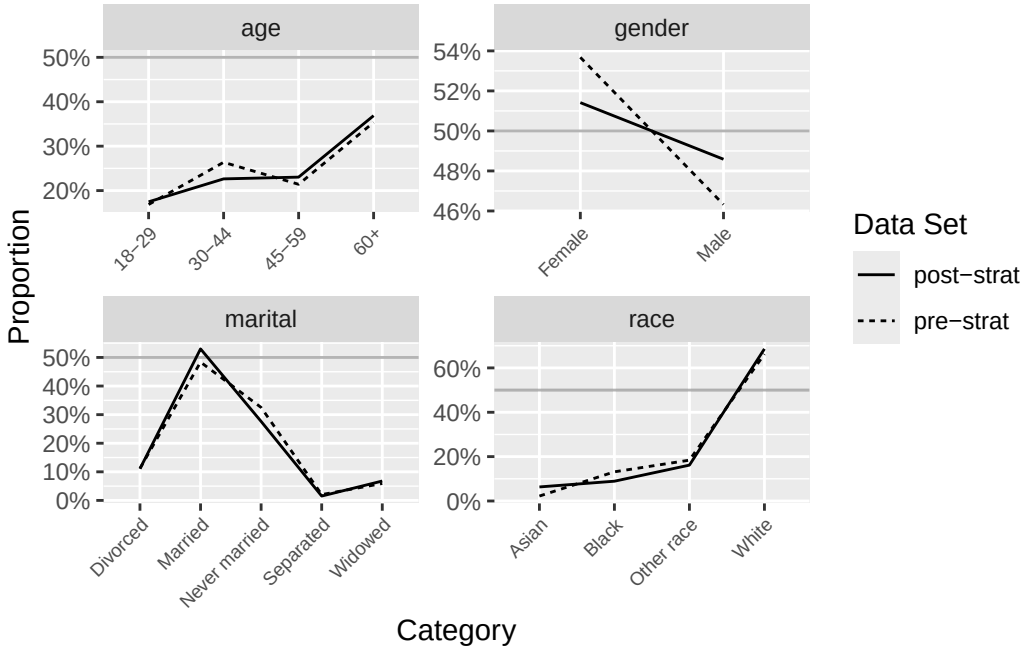


Figure 1: Voter Demographics (Age, Gender, Marital Status, Race)

Figure 3 depicts the proportion of support for both parties among the survey respondents. 61% support the Democratic Party and 39% support the Republican Party. This ratio is not accurate compared to the actual ratio. To further visualize it, Figure 4 uses a gradient colour scale to present the US map of support by state, with blue representing support for Biden, and red representing support for Trump. States like Wyoming (WY), Kansas (KS), Nebraska (NE) and Alabama (AL) are blue in Figure 4 while they are reliably red states in recent decades (PewResearchCenter). Therefore, to adjust this sample bias, I also create a balanced survey dataset.

2.3 US Election 2020 Dataset and the Balanced Survey Dataset

Among the respondents in the survey, those who voted for Biden in 2020 (44.7%) are approximately 1.5 times as many as those who voted for Trump (30.7%) as shown in Table 1. This difference is substantial when compared to the actual popular vote ratio in 2020, which is 51.3%:46.9% (Walter). I adjust the survey ratio of 2020 votes to match the actual 51.3%:46.9% ratio to create a balanced dataset. To be more specific, I retain all the observations that supported Trump in 2020 but randomly select a certain number of observations that supported Biden in 2020. The number of rows selected is determined by the floor value of 51.3/46.9 times the number of rows that supported Trump.

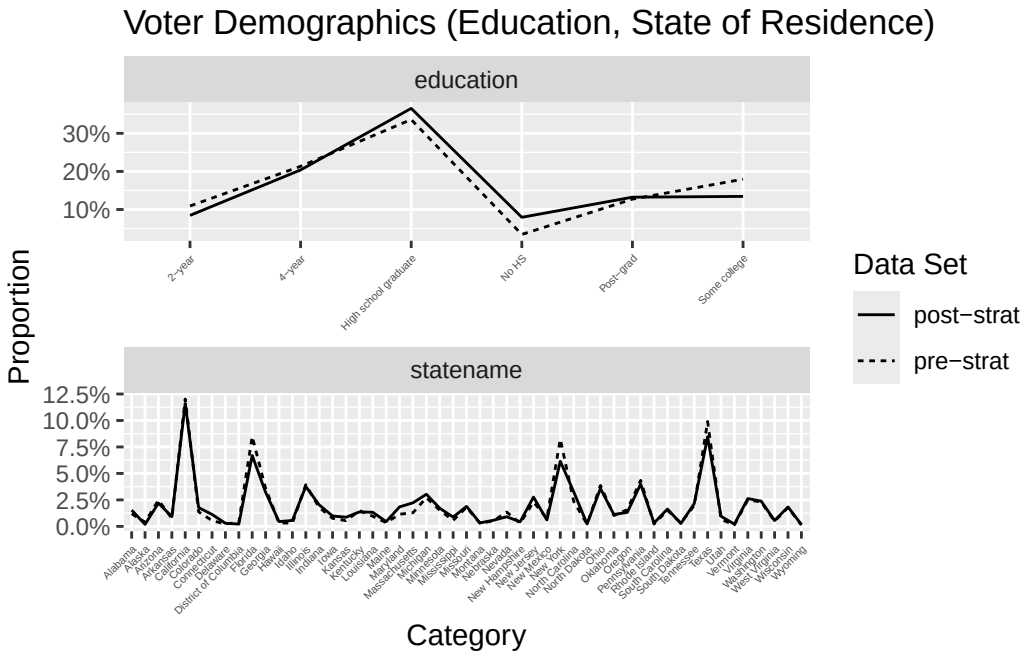


Figure 2: Voter Demographics (Education, State of Residence)

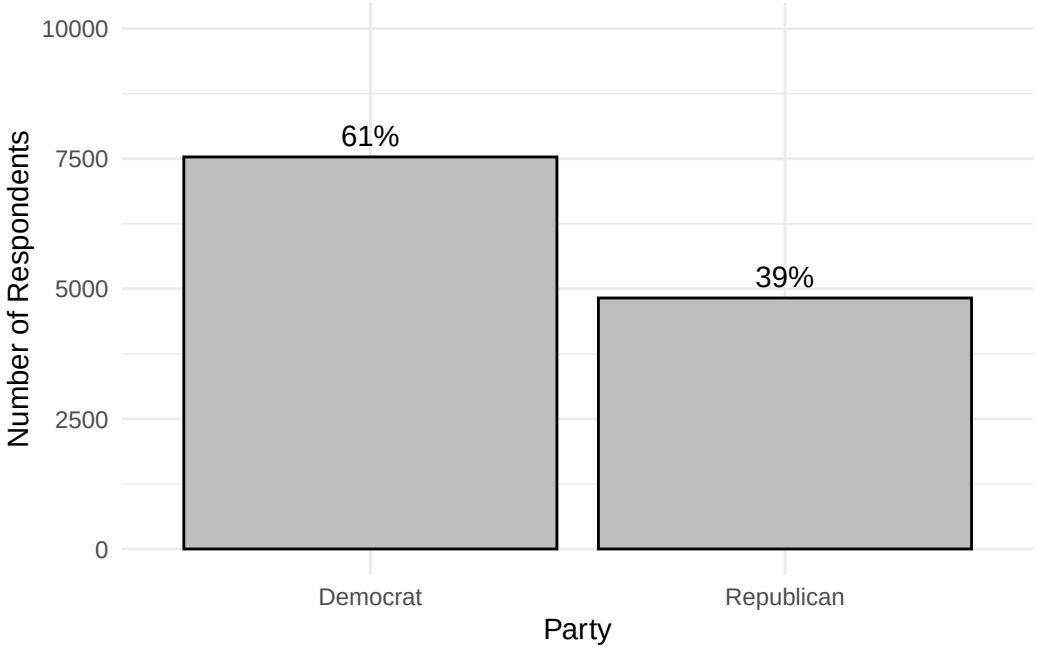


Figure 3: Proportion of the Two Parties in the Survey

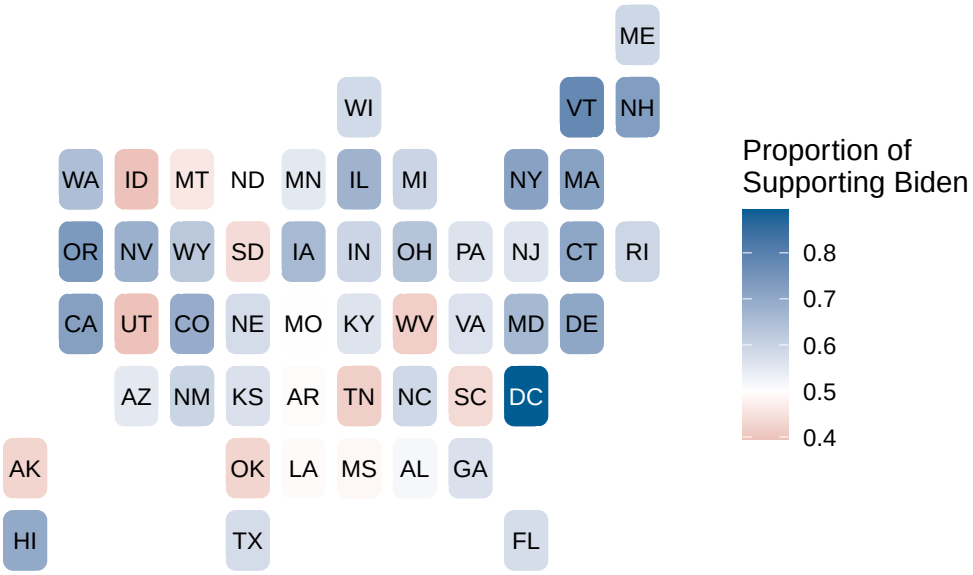
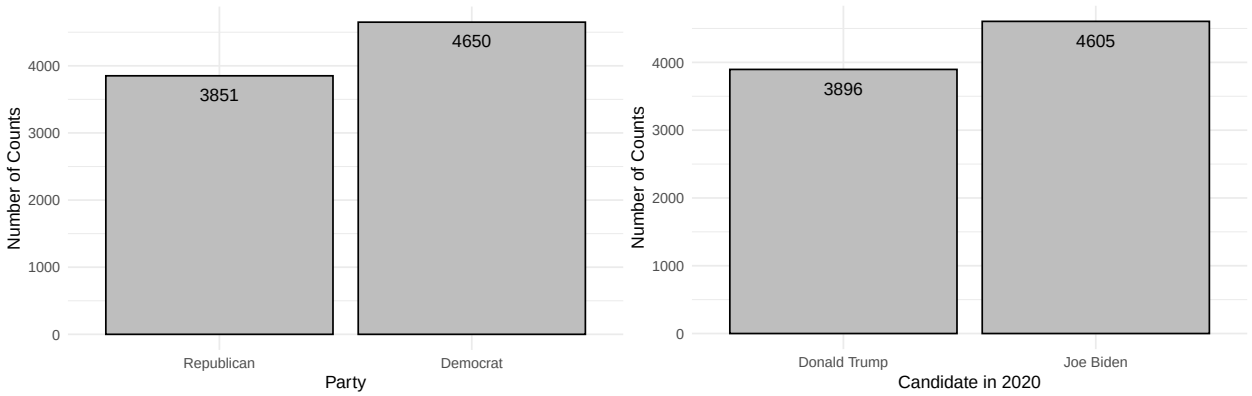


Figure 4: Support for Joe Biden and Donald Trump by State in the Sample

Table 1: Votes in 2020 from the Survey Respondents

Candidates	Number of Votes	Proportion (%)
Did not vote for President	4471	22.355
Donald Trump	6020	30.100
Howie Hawkins	92	0.460
Jo Jorgensen	259	1.295
Joe Biden	8942	44.710
Other	216	1.080

The limitation of this method is that the balanced sample is smaller than the unbalanced sample, causing potential biases and underrepresentation for some groups. The strength is that the balanced dataset contains a more reasonable, rather than a disparate, proportion of supporters. Figure 5 (a) shows 3851 respondents consider themselves as Republicans, and 4650 respondents consider themselves as Democrats. Figure 5 (b) indicates 3896 people voted for Trump and 4605 voted for Biden in the last election. These two groups of numbers are more balanced.

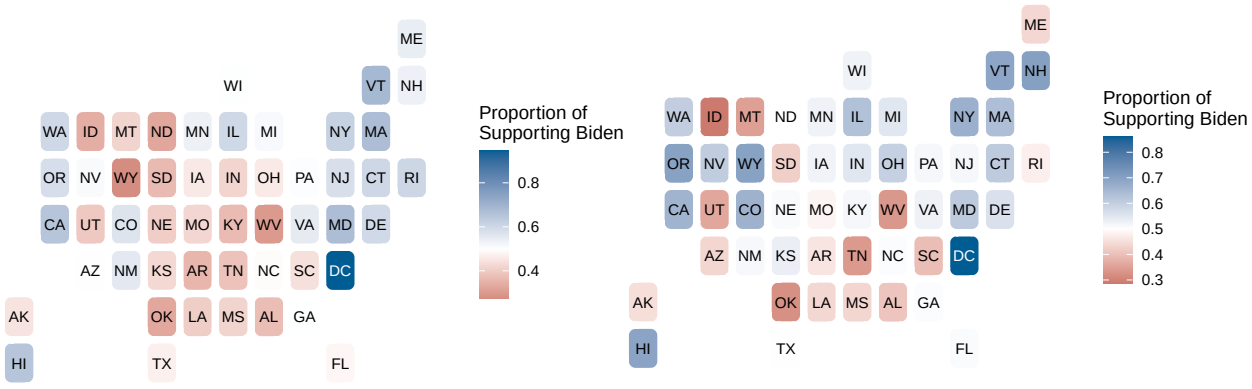


(a) Parties Respondents Support Now

(b) Respondents' votes in 2020

Figure 5: Distribution of Supporting Parties Now and Votes in 2020 from the Survey Respondents

The data creating Figure 6 (a) is a US Election 2020 dataset from Kaggle (Fontes 2021), which contains county-level data from the 2020 US Election, including votes, candidates and whether each candidate wins the election. Figure 6 shows two maps, one presents the 2020 actual voting results, and one depicts the 2020 voting results from the balanced survey. The distribution of votes from the balanced dataset is more accurate. Alabama (AL) is red in Figure 6, Kansas (KS) and Nebraska (NE) are not as blue as in Figure 4. I will use the ACS dataset to adjust balanced survey dataset as well. A prediction map using the balanced dataset will be included at the end of the Results section.



(a) 2020 Actual Voting Results by States

(b) 2020 Voting Results From Balanced Survey Respondents by States

Figure 6: 2020 Election Votes

3 Model

To predict the 2024 election result, I will use Multilevel Regression with Post-Stratification (MRP), a statistical technique that combines the strengths of multilevel modelling and post-stratification. MRP takes a sample, usually a large-scale poll, and uses the poll results to generate a model. This model is then applied to a post-stratification dataset, typically a census or another larger sample (Alexander 2023). In this paper, the sample is the America's Political Pulse dataset (Shanto Iyengar 2023), and the census is the IPUMS ACS dataset (Team). Logistic regression will be used to create the model.

It is important to note that, if we had perfect data, we would not need a model (Alexander 2023). The purpose of a model is to generalize patterns from the data and make predictions, so MRP is not guaranteed to provide perfect accuracy. I assume there exists a relationship between predictors (race, gender, marital status, age group and highest level of education) and the voting outcome, and this relationship is consistent between the sample and the post-stratification dataset. Therefore, the model is not perfectly flawless, and given that the accuracy of MRP depends on the accuracy of the model, uncertainties and imprecision are inevitable.

Logistic regression is highly useful when dealing with a binary outcome variable. Instead of using 'pid3' variable in the survey, which gives either 'Democrat' or 'Republican', I create a new variable named 'vote_biden'. In this new setup, 'vote_biden' will be assigned a value of 1 if 'pid3' is Democrat and 0 if 'pid3' is Republican. For instance, if the average "vote_biden" equals 0.5, then half of the population supports Biden.

The logistic regression equation for predicting the binary outcome of support for Biden (where 1 indicates support) can be expressed as follows:

$$\text{logit}(\hat{p}) = \beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{race} + \beta_3 \times \text{state} + \beta_4 \times \text{age} + \beta_5 \times \text{marital} + \beta_6 \times \text{education} \tag{1}$$

where:

- $\text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right)$
- $\hat{p}$ is the predicted probability of supporting Biden occurring
- β_0 is the intercept term, representing the value if all the predictors are zero
- β_1, β_2, β_3, β_4, β_5, β_6 are the coefficients corresponding to the predictor variables (gender, race, state, age, marital, education).

The predicted probability, $\hat{p}$, of supporting Biden can be obtained by applying the logistic function to the log-odds, shown in Equation 2:

$$\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{race} + \beta_3 \times \text{state} + \beta_4 \times \text{age} + \beta_5 \times \text{marital} + \beta_6 \times \text{education}}} \tag{2}$$

Table 2: Coefficients of the Voting Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.56	0.25	2.27	0.02	0.08	1.04
raceBlack	0.71	0.16	4.45	0.00	0.39	1.01
raceOther race	-0.05	0.15	-0.34	0.73	-0.36	0.24
raceWhite	-0.44	0.15	-2.97	0.00	-0.73	-0.15
genderMale	-0.30	0.04	-7.57	0.00	-0.38	-0.22
education4-year	0.42	0.07	5.78	0.00	0.28	0.56
educationHigh school graduate	-0.10	0.07	-1.52	0.13	-0.23	0.03
educationNo HS	-0.07	0.12	-0.57	0.57	-0.30	0.17
educationPost-grad	0.84	0.08	10.04	0.00	0.67	1.00
educationSome college	0.17	0.07	2.24	0.02	0.02	0.31
statenameAlaska	-0.63	0.35	-1.81	0.07	-1.32	0.05
statenameArizona	0.30	0.21	1.42	0.16	-0.11	0.71
statenameArkansas	0.09	0.28	0.32	0.75	-0.46	0.64
statenameCalifornia	0.89	0.18	4.82	0.00	0.53	1.25
statenameColorado	0.86	0.24	3.53	0.00	0.39	1.34
statenameConnecticut	0.89	0.32	2.74	0.01	0.26	1.54
statenameDelaware	1.05	0.46	2.27	0.02	0.17	2.00
statenameDistrict of Columbia	1.72	0.51	3.38	0.00	0.81	2.84
statenameFlorida	0.38	0.19	2.07	0.04	0.02	0.75
statenameGeorgia	0.10	0.20	0.48	0.63	-0.30	0.49
statenameHawaii	0.66	0.40	1.66	0.10	-0.10	1.46
statenameIdaho	-0.29	0.38	-0.75	0.45	-1.05	0.46
statenameIllinois	0.73	0.20	3.64	0.00	0.34	1.12
statenameIndiana	0.57	0.22	2.56	0.01	0.14	1.01
statenameIowa	0.85	0.28	3.03	0.00	0.31	1.41
statenameKansas	0.40	0.31	1.29	0.20	-0.20	1.01
statenameKentucky	0.44	0.23	1.89	0.06	-0.02	0.90
statenameLouisiana	-0.09	0.26	-0.35	0.73	-0.60	0.42
statenameMaine	0.67	0.36	1.87	0.06	-0.03	1.39
statenameMaryland	0.66	0.25	2.61	0.01	0.17	1.17
statenameMassachusetts	1.00	0.25	3.97	0.00	0.51	1.49
statenameMichigan	0.42	0.21	2.03	0.04	0.01	0.83
statenameMinnesota	0.33	0.23	1.45	0.15	-0.12	0.79
statenameMississippi	-0.13	0.31	-0.42	0.67	-0.73	0.47
statenameMissouri	0.11	0.22	0.50	0.62	-0.33	0.55
statenameMontana	-0.14	0.39	-0.37	0.71	-0.91	0.62
statenameNebraska	0.40	0.31	1.27	0.20	-0.21	1.02
statenameNevada	0.83	0.24	3.40	0.00	0.35	1.31
statenameNew Hampshire	1.24	0.41	2.98	0.00	0.45	2.09
statenameNew Jersey	0.24	0.21	1.10	0.27	-0.18	0.65
statenameNew Mexico	0.47	0.28	1.69	0.09	-0.07	1.01
statenameNew York	0.89	0.19	4.73	0.00	0.52	1.26
statenameNorth Carolina	0.30	0.21	1.41	0.16	-0.12	0.72
statenameNorth Dakota	0.06	0.44	0.14	0.89	-0.81	0.94
statenameOhio	0.70	0.20	3.51	0.00	0.31	1.09
statenameOklahoma	-0.12	0.25	-0.47	0.64	-0.62	0.38
statenameOregon	1.17	0.24	4.87	0.00	0.70	1.65
statenamePennsylvania	0.43	0.20	2.18	0.03	0.04	0.81
statenameRhode Island	0.46	0.44	1.04	0.30	-0.40	1.34
statenameSouth Carolina	-0.30	0.23	-1.29	0.20	-0.75	0.15

Table 2: Coefficients of the Voting Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
statenameSouth Dakota	-0.08	0.37	-0.22	0.83	-0.81	0.64
statenameTennessee	-0.17	0.22	-0.79	0.43	-0.61	0.26
statenameTexas	0.26	0.18	1.43	0.15	-0.10	0.62
statenameUtah	-0.39	0.29	-1.33	0.18	-0.96	0.18
statenameVermont	1.38	0.61	2.27	0.02	0.27	2.71
statenameVirginia	0.23	0.21	1.11	0.27	-0.18	0.65
statenameWashington	0.71	0.22	3.29	0.00	0.29	1.13
statenameWest Virginia	-0.02	0.32	-0.07	0.94	-0.65	0.60
statenameWisconsin	0.55	0.22	2.48	0.01	0.12	0.99
statenameWyoming	0.51	0.55	0.92	0.36	-0.56	1.65
maritalMarried	-0.43	0.07	-6.58	0.00	-0.56	-0.30
maritalNever married	0.06	0.08	0.84	0.40	-0.08	0.21
maritalSeparated	0.01	0.15	0.08	0.93	-0.28	0.31
maritalWidowed	-0.26	0.10	-2.72	0.01	-0.46	-0.07
age30-44	-0.08	0.07	-1.25	0.21	-0.21	0.05
age45-59	-0.31	0.07	-4.40	0.00	-0.45	-0.17
age60+	-0.27	0.07	-3.86	0.00	-0.40	-0.13

I use the binomial distribution to model the probability of the binary outcome in the survey. Table 2 shows the estimates for the coefficients that will fit into the logistic regression equation, as well as the standard error, statistic, p-value, and the 95% confidence level for each coefficient. The statistic is the t-statistic for testing whether the corresponding coefficient is significantly different from zero. A 95% confidence level means that if I were to repeat the sampling and estimation many times, I would expect the true value of the coefficient to fall within this interval in approximately 95% of repetitions. The 'state' variable has relatively larger p-values and standard errors when compared to other predictors in Table 2. Terms like 'statenameNorth Dakota', 'statenameSouth Dakota', 'statenameWest Virginia' and 'maritalSeparated' have p-values greater than 0.8, implying their evidence against the null hypothesis is weak. In simpler terms, this suggests that these characteristics may not strongly influence the likelihood of individuals supporting Biden, as the statistical evidence supporting such an influence is not particularly strong. The greater the estimate, the higher the likelihood that people with these characteristics will vote for Biden.

Figure 7 visualizes the estimates of coefficients as black dots and the confidence intervals as blue lines from a logistic regression model. The length of each blue line (error bar) reflects the confidence interval for the corresponding coefficient. The zero on the x-axis denotes neutrality. For example, an estimate of -0.1 implies a slight inclination toward voting for Trump. Most data points in Figure 7 are situated to the right of the zero on the x-axis, indicating a higher overall inclination toward supporting Biden. The further to the right both the points and blue lines extend, the greater the likelihood that individuals with these characteristics will vote for Biden. Conversely, points and lines extending to the left suggest a higher probability of supporting Trump. From Figure 7, a person from the District of Columbia is most likely to support Biden, a person from Alaska is most likely to support Trump, and someone who is separated is swing and unpredictable.

After constructing a model, I categorize individuals in both survey and census datasets according to their race, age, gender, marital status, education level and state of residence. I compute both the total count and the proportion for each group. Next, utilizing the survey data, I calculate the proportion of voters that tend to support Biden within each demographic group including the confidence intervals, and apply this proportion to the census dataset for weight adjustment, ensuring a more representative population composition.

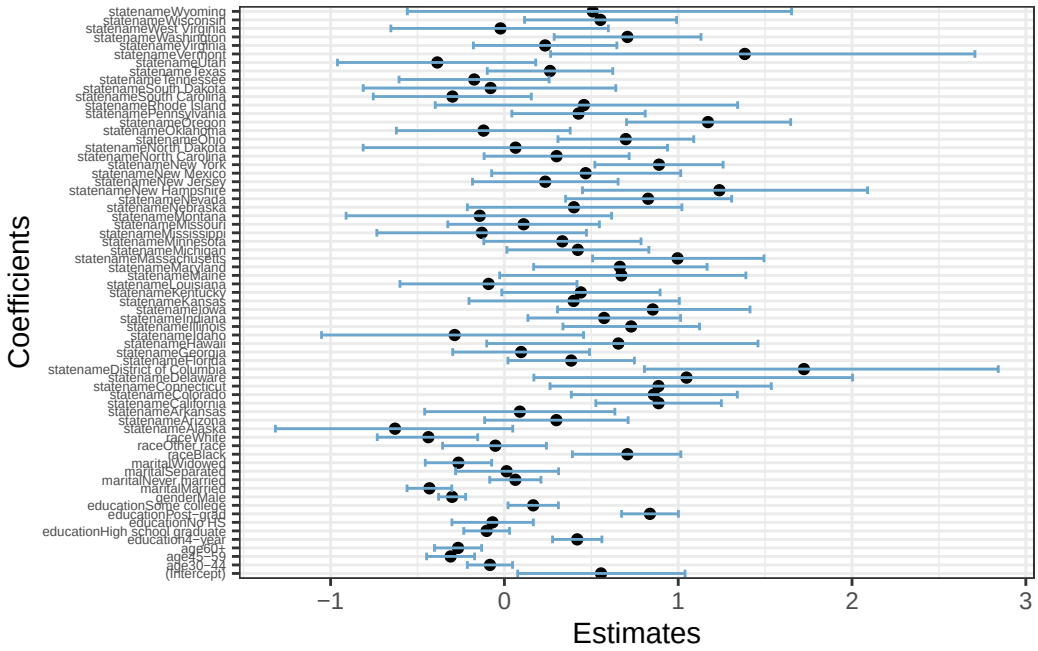


Figure 7: Estimates of Coefficients in the Model

4 Results

4.1 Results Using the Original Dataset

Figure 8 illustrates the proportion of voters in each state supporting Biden after post-stratify. The red points represent the percentage of voting for Biden from the survey data. The black points represent the predicted voting percentage based on the model generated from the survey, in other words, the black points are post-stratified. To enhance clarity and visualize the fluctuations, solid black lines and dashed red lines are constructed to connect the points from the corresponding data source. A purple ribbon, denoting the range of potential outcomes or variations around the predicted values, is included. This visual representation of possible deviations employs the 95% confidence interval. Moreover, a purple line at 50% serves as a reference. States with points above this line, such as Wisconsin and Massachusetts, indicate more people supporting Biden, while states below the line, including Utah and Montana, have more residents voting for Trump. States like Minnesota, where the voting percentage hovers around 50% and the confidence interval spans both above and below the purple line, are commonly referred to as swing states. Overall, the predicted outcomes closely align with the survey results. With the unbalanced dataset, it appears that a greater number of states favour Biden over Trump. Based on this observation, I predict that Biden will establish a comfortable lead in the popular votes in the upcoming Presidential Election.

Figure 9 is a state bin plot of the predicted support for Joe Biden and Donald Trump by states in a gradient colour scale. The colour intensity indicates the level of support. The redder a state appears, the higher the predicted proportion of support for Trump. The bluer a state is, the higher the predicted proportion of support for Biden. Once again, it appears that a greater number of states prefer Biden over Trump, with an overall 62.96% supporting Biden.

4.2 A Quick Overview of Results Using the Balanced Dataset

Similar to the summary of Figure 7, I use the balanced dataset, along with the logistic regression and Equation 1, to create another model, and Figure 10 displays the estimates of coefficients. Black points represent estimates, the orange lines depict confidence intervals, and the zero on the x-axis denotes neutrality. In Figure 10, New Hampshire and the District of Columbia have the most citizens supporting Biden, while

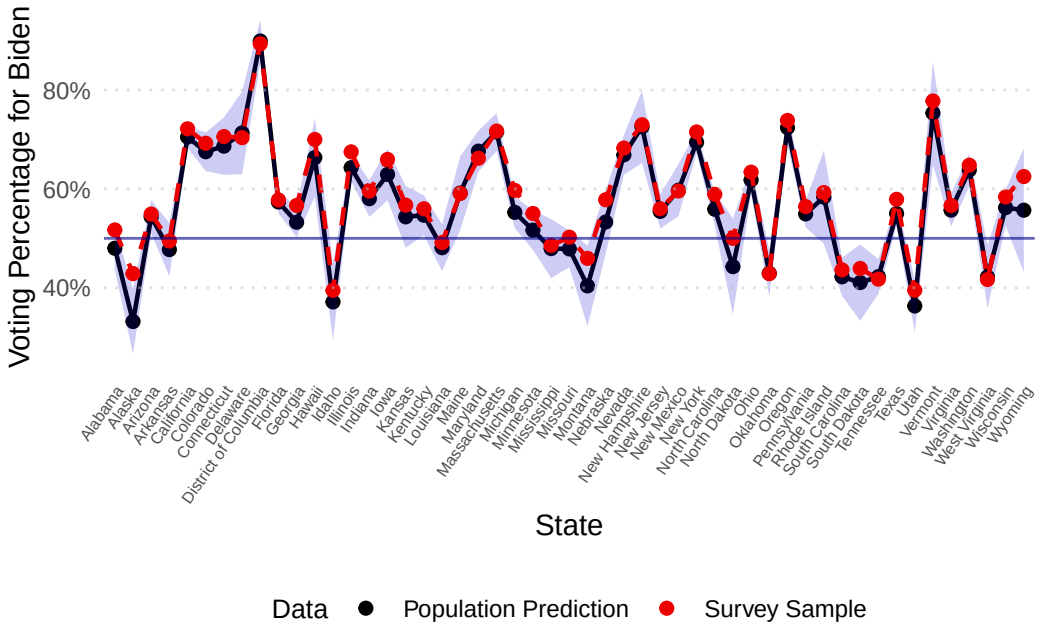


Figure 8: Proportion of Voters in Each State Voting for Biden

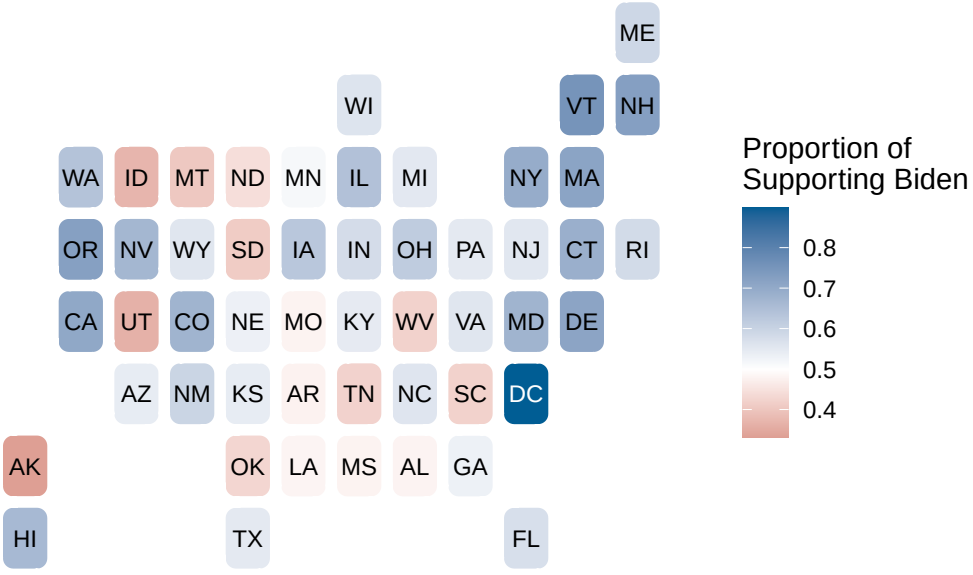


Figure 9: Predicted Support for Joe Biden and Donald Trump by State

white people and married people are most likely to support Trump.

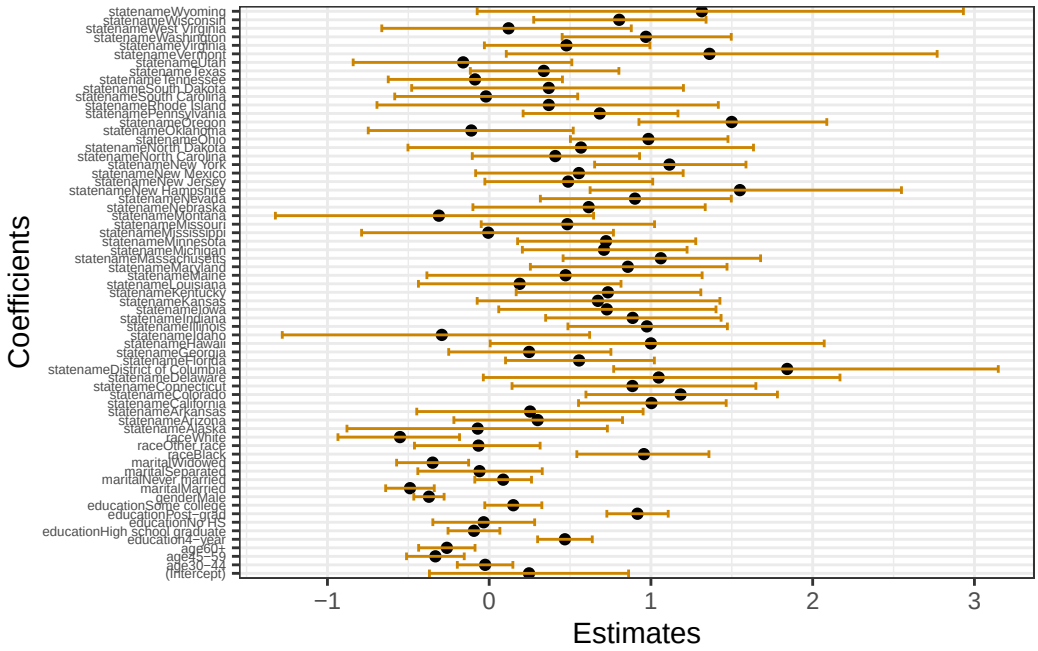


Figure 10: Estimates of Coefficients in the Model (Balanced)

Figure 11 shows the result using MRP, indicating the proportion of voters in each state voting for Biden. Black represents population prediction, and red represents the survey. The orange ribbon shows the range of potential variations around the predicted values using a 95% confidence interval. The predicted outcomes are close to the survey. Compared to Figure 8, Figure 11 is more balanced and reasonable, as it presents comparable numbers of states above and below the orange line which represents 50% support for Biden. Traditionally Republican-leaning states like Georgia, Texas, and Nebraska turn red in Figure 11.

Figure 12 is a map of the US, using colours to show the predicted support. The redder a state appears, the higher the predicted proportion of support for Trump. The bluer a state is, the higher the predicted proportion of support for Biden. There are 23 red states and 28 blue states, but still, 57.61% of the total population will vote for Biden, which is more than half.

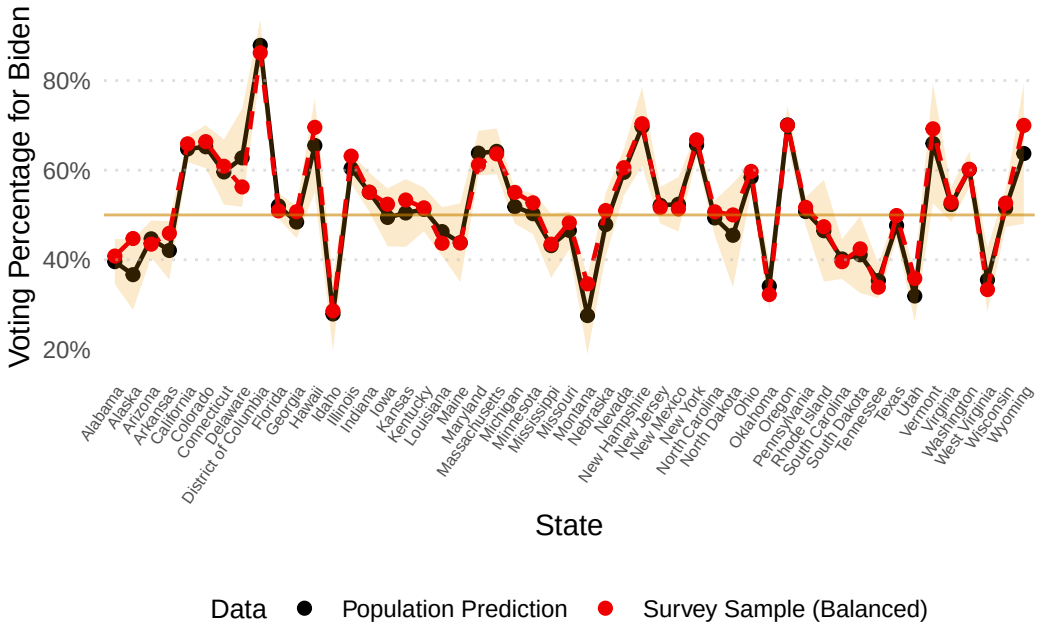


Figure 11: Proportion of Voters in Each State Voting for Biden (Balanced)

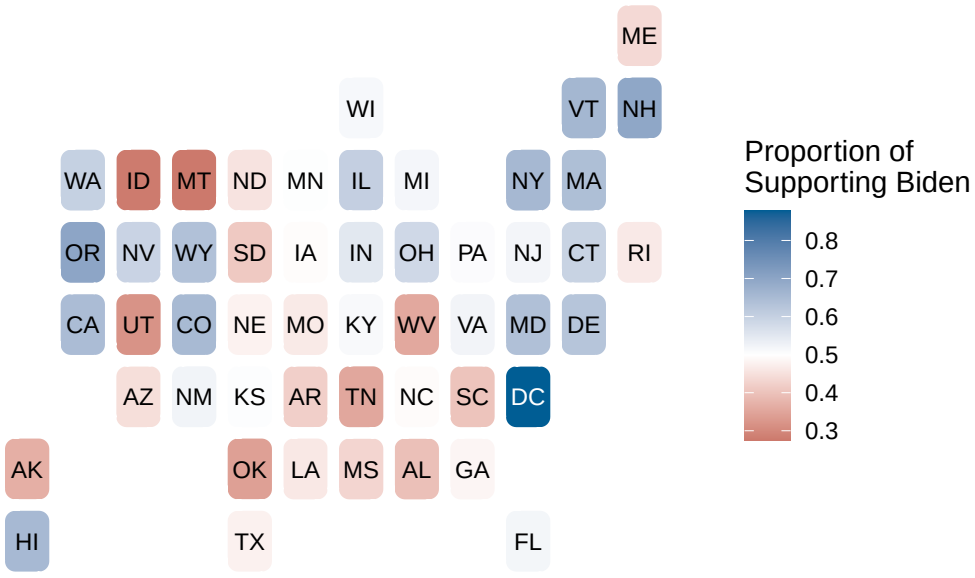


Figure 12: Predicted support for Joe Biden and Donald Trump by State (Balanced Survey Results)

5 Discussions

As noted in the introduction, the outcome of the presidential election depends on electoral college votes, and unexpected outcome, like the one in 2016, may occur. Whether using the balanced or unbalanced datasets, both suggest a substantial lead for Biden in popular votes, surpassing 10%. This significant advantage makes it highly likely for him to secure more electoral college votes and thus, ensure his re-election as the President of the United States.

5.1 Geography, Ethnicity and Gender

Based on both Figure 9 and Figure 12, states that support Trump are more concentrated in the middle and Midwest parts of the US, whereas coastal states and the Northeast tend to favour Biden. Although Republicans and Democrats currently appear competitive, this wasn't the case thirty years ago. In the 1980, 1984, and 1988 presidential elections, the Republican Party was strongly favoured, with a maximum of 11 out of 51 states supporting the Democratic Party (270ToWin). However, a huge change occurred in 1992 when 33 states gave their electoral college votes to the Democratic Party, including influential states like California with 54 votes and New York with 28 votes. Since then, the two parties have been competitive, solidifying the pattern of the Midwest preferring the Republicans and coastal states leaning towards the Democrats (270ToWin).

The shift in voting patterns during the 1990s, in addition to Bill Clinton's successful presentations, can be attributed to the rising number of immigrants during that period. There was a rapid increase in the number of immigrants in the 1990s because of immigration act of 1990 (101stCongress 1990), while Table 2 has demonstrated a preference for the Democratic Party among Asians and other racial groups over white individuals. Furthermore, nowadays, blue states have more Asians and people in other race (WorldPopulationReview). For example, 14.83% of population in the blue state California are Asian, 22.23% are other races, while only 2.37% of the residences in Indiana are Asians and 5.67% are other races (WorldPopulationReview).

The Democrats are the liberal political party, with an emphasis on civil rights, safety issues and student loan forgiveness (StudentAid). The Republicans are the conservative party, which is also known as the Grand Old Party. It has stood for lower taxes, border issues and gun rights.

The geographic distribution of red states along the southern border, as illustrated in Figure 9 and Figure 12, is noteworthy. This alignment could be linked to Republicans addressing broader issues. Furthermore, Trump's emphasis on improving the economy in the Midwest, where states traditionally lean Republican, provides another explanation for their preference for Trump compared to wealthier coastal states. This may also be correlated to the rural and suburban demographics.

In terms of abortion rights, women's rights, and LGBTQ+ rights, the Democratic Party typically embraces more progressive positions, advocating for gender and LGBTQ+ equality to a greater extent. Conversely, the Republican Party tends to have more conservative stances on social issues, advocating for restrictions on abortion and often holding a more conservative stance on LGBTQ+ rights. As a result of these differing policy positions, there is a tendency for females to show greater support for Democratic candidates, such as Biden, compared to their male counterparts as shown in Table 2.

It is essential to note that these are generalizations. Individual political preferences can vary significantly within each demographic group.

5.2 Age Group, Education and Marital Status

Using Table 2, we observe a trend where a higher proportion of younger individuals express support for Biden, while the older generation tends to favour Trump. This difference could be attributed to the younger generation's emphasis on social justice and LGBTQ+ rights. Given there are definitely more sexual minorities among the youth, there is a tendency towards their alignment with the Democratic Party and Biden. On the other hand, the older generation may be influenced by considerations related to national security and conservative values, which are often associated with the Republican Party. Widowed and married people, compared to single, divorced and separated individuals, prefer Trump is also observed in Table 2. This

can be influenced by several factors, including the higher average age of widowed and married individuals compared to the single ones. Additionally, some single individuals that consciously choose not to marry are more likely to be untraditional.

Table 2 also shows people with a higher level of education favour Biden more often than people with lower level of education. Considering Biden's proposed student debt relief plan, which aims to offer targeted debt relief to low- and middle-income families, voters with a strong concern for education may find themselves inclined to support Biden. The Northeast states have a lot of top universities including Ivy League. As seen in Figure 9 and Figure 12, the strong support for Biden in the Northeast could be correlated to the average education level of the people in this region.

Again, it is crucial to acknowledge that these are generalizations, and individual political preferences can vary significantly.

5.3 Biases and Weaknesses

The sample survey was collected using participants from the YouGov panel, which may introduce potential biases, as online survey participation is not evenly distributed across the entire population. This could result in the underrepresentation of certain demographic groups, especially older individuals and those with lower socioeconomic status. Additionally, the self-selection bias of individuals choosing to participate in paid online surveys may lead to distinct characteristics among respondents, further skewing the representation. The active political engagement of YouGov participants could overemphasize the views of individuals that are interested in politics, potentially misrepresenting the total population.

An evident weakness is the imbalance between the proportions of Democratic and Republican supporters in the survey, as highlighted by Figure 3. Among the respondents in the survey, 44.7% voted for Biden in 2020 and 30.7% voted for Trump as shown in Table 1, while the actual popular vote ratio in 2020 is 51.3%:46.9% (Walter). This imbalance permeates figures generated from the unbalanced dataset, including Figure 7, Table 2, Figure 8 and Figure 9, potentially skewing the results toward Democrats. In order to adjust this, I created a balanced dataset, but this balanced sample has fewer observations than the unbalanced sample.

Illustrating this weakness with a specific example, consider the small representation from Wyoming in the original, uncleaned survey dataset. Out of 20,000 observations, only 30 are from Wyoming, a traditionally red state with 69.9% population supporting Trump and 26.6% supporting Biden in 2020 (Walter). However, for those 30 participants from Wyoming, only 6 of them consider themselves as Republicans, and 10 are Democrats, 14 are Independent or other. Another example is that, 140 out of 383 respondents living in Indiana are Democrats, 97 of them are Republicans, while Trump had 16% more Indiana popular votes than Biden in the last election (Walter). This type of disparity, where the sample composition greatly deviates from the state's political leanings, was even enlarged after excluding those who are not Democrats or Republicans. While MRP and post-stratification attempt to adjust this imbalance, they do not fully eliminate the underrepresentation, as both Wyoming and Indiana are pretty blue in both Figure 9 and Figure 12.

Another weakness arises from variations in categorization. For example, the 'race' variable has different options in the survey and census datasets, and the options cannot be perfectly combined. Merging these variables introduces complexities and reduces the precision of the model. Moreover, the marital status variable in the survey has another option as Domestic/civil partnership, but there is no same or similar generalization in the census. The exclusion of this marital status option from the survey dataset further contributes to this challenge. Despite efforts to accurately merge based on questionnaire descriptions, these differences bring limitations to the analysis's accuracy. Detailed information on these discrepancies is provided in the Appendix.

5.4 Next Steps

The predictive model generated from this analysis, with its strengths and limitations, serves as a foundation for ongoing discussions on campaign strategies, policy implications, and the broader implications for

democracy. While this analysis provides some understandings of the factors influencing voter preferences, it is crucial to acknowledge the inherent uncertainties in forecasting political events.

For future exploration, I would suggest that survey companies use consistent demographic generalizations as those used in census data to minimize additional biases. Special attention should be given to groups that have been underrepresented to enhance the precision of predictive models. It is recommended that future research delves into the dynamics of electoral behaviour, particularly by examining the outcome of 2024 election and the repercussions of the COVID-19 pandemic. This will enable a comprehensive understanding of unforeseen events that might alter political preferences among citizens.

A Appendix

A.1 Data Cleaning

In the America's Political Pulse survey data, I only select variables 'inputstate', 'gender', 'race', 'educ', 'pid3', 'marstat', 'birthyr', 'presvote20post'. In the IPUMS census data, I select variables 'SEX', 'MARST', 'STATEICP', 'RACE', 'EDUC', 'AGE'. The 'inputstate' variable from the survey contains 51 states, and the 'STATEICP' from the census contains 51 codes of states using the coding scheme developed by the Inter-University Consortium for Political and Social Research (ICPSR). To ensure consistency between these datasets, I have created three additional variables in both cleaned datasets: 'statename' (state name), 'state_abb' (state abbreviation), and 'stateicp' (state codes).

In the 'race' variable, the survey encompasses eight categories: White, Black, Hispanic, Asian, Native American, Two or more races, Other and Middle Eastern. The 'RACE' from the census has a different set: White, Black/African American, American Indian or Alaska Native, Chinese, Japanese, Other AsiaFn or Pacific Islander, Other race/nec, Two major races, Three or more major races. Since both Asian and Middle Eastern, both White and Hispanic are listed in the survey, to harmonize them, I have grouped 'Hispanic', 'Native American', 'Two or more races', 'Other', and 'Middle Eastern' from the survey as 'Other race'. In the census, 'Chinese', 'Japanese', 'Other Asian or Pacific Islander' are combined into 'Asian', and 'American Indian or Alaska Native', 'Other race/nec', 'Two major races', 'Three or more major races' are combined into 'Other race' using the explanation on IPUMS website (Team). Therefore, the renamed column 'race' from both datasets only contains 'White', 'Black', 'Asian', 'Other race', which is the only way merge these two variables.

Regarding the 'marstat' variable from APP, I have merged 'Married (spouse present)' and 'Married (spouse absent)' into 'Married' and excluded instances with 'Domestic/civil partnership.' Thus, the renamed 'marital' column contains 'Married', 'Separated', 'Divorced', 'Widowed', 'Never married' in both datasets.

I only keep individuals older or equal to 18, and individuals who support either the Democratic Party or the Republican Party. To align 'AGE' in IPUMS with 'birthyr' in APP, a new variable 'Age' is created by subtracting 'birthyr' from 2024. Age groups are then formed as 18-29, 30-44, 45-59, and 60+. The gender variables are renamed to 'gender' in both cleaned datasets. A new variable called 'vote_biden' is created to give 1 for individuals consider themselves as Democrat in pid3 column, and give 0 for individuals consider themselves as Republican in pid3 column.

For 'EDUC', I have standardized values in both datasets by creating a new 'education' column, representing 'No HS', 'High school graduate', 'Some college', '2-year', '4-year', and 'Post grad' based on IPUMS and survey questionnaire details, using the questionnaire text from IPUMS (Team) and the survey questions details (Shanto Iyengar 2023).

To address the imbalance in the survey's 2020 vote ('presvote20post') ratios, I have adjusted them to match the actual 51.3%:46.9% ratio, creating a balanced dataset while keeping other variables consistent.

Lastly, for the US Election 2020 dataset, I combine observations within the same states, creating two columns 'Joe_Biden' and 'Donald_Trump', with corresponding votes in each state.

A.2 Coefficients Estimates Table Using Balanced Dataset

Table 3: Coefficients of the Voting Model (Balanced)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.25	0.31	0.79	0.43	-0.37	0.86
raceBlack	0.96	0.21	4.60	0.00	0.54	1.36
raceOther race	-0.07	0.20	-0.33	0.74	-0.46	0.31
raceWhite	-0.55	0.19	-2.88	0.00	-0.93	-0.18
genderMale	-0.37	0.05	-7.79	0.00	-0.47	-0.28
education4-year	0.47	0.09	5.43	0.00	0.30	0.64

Table 3: Coefficients of the Voting Model (Balanced)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
educationHigh school graduate	-0.09	0.08	-1.15	0.25	-0.25	0.07
educationNo HS	-0.03	0.16	-0.21	0.83	-0.35	0.28
educationPost-grad	0.92	0.10	9.48	0.00	0.73	1.11
educationSome college	0.15	0.09	1.66	0.10	-0.03	0.33
statenameAlaska	-0.07	0.41	-0.17	0.86	-0.88	0.73
statenameArizona	0.30	0.27	1.13	0.26	-0.22	0.82
statenameArkansas	0.25	0.36	0.71	0.48	-0.45	0.95
statenameCalifornia	1.00	0.23	4.32	0.00	0.55	1.47
statenameColorado	1.18	0.30	3.92	0.00	0.60	1.78
statenameConnecticut	0.89	0.38	2.31	0.02	0.14	1.65
statenameDelaware	1.05	0.55	1.89	0.06	-0.04	2.17
statenameDistrict of Columbia	1.84	0.59	3.11	0.00	0.77	3.15
statenameFlorida	0.56	0.23	2.37	0.02	0.10	1.02
statenameGeorgia	0.25	0.26	0.97	0.33	-0.25	0.75
statenameHawaii	1.00	0.52	1.92	0.05	0.01	2.07
statenameIdaho	-0.29	0.48	-0.61	0.54	-1.28	0.62
statenameIllinois	0.97	0.25	3.88	0.00	0.49	1.47
statenameIndiana	0.89	0.28	3.21	0.00	0.35	1.43
statenameIowa	0.73	0.34	2.13	0.03	0.06	1.40
statenameKansas	0.67	0.38	1.76	0.08	-0.07	1.43
statenameKentucky	0.73	0.29	2.52	0.01	0.17	1.31
statenameLouisiana	0.19	0.32	0.59	0.55	-0.44	0.81
statenameMaine	0.47	0.43	1.09	0.27	-0.39	1.32
statenameMaryland	0.86	0.31	2.77	0.01	0.25	1.47
statenameMassachusetts	1.06	0.31	3.41	0.00	0.46	1.68
statenameMichigan	0.71	0.26	2.74	0.01	0.21	1.22
statenameMinnesota	0.72	0.28	2.57	0.01	0.18	1.28
statenameMississippi	-0.01	0.40	-0.01	0.99	-0.79	0.77
statenameMissouri	0.48	0.27	1.77	0.08	-0.05	1.02
statenameMontana	-0.31	0.50	-0.62	0.53	-1.32	0.65
statenameNebraska	0.62	0.37	1.68	0.09	-0.10	1.34
statenameNevada	0.90	0.30	3.00	0.00	0.32	1.50
statenameNew Hampshire	1.55	0.49	3.19	0.00	0.62	2.55
statenameNew Jersey	0.49	0.26	1.85	0.06	-0.03	1.01
statenameNew Mexico	0.55	0.33	1.70	0.09	-0.08	1.20
statenameNew York	1.11	0.24	4.68	0.00	0.65	1.59
statenameNorth Carolina	0.41	0.26	1.55	0.12	-0.10	0.93
statenameNorth Dakota	0.57	0.54	1.05	0.29	-0.50	1.63
statenameOhio	0.98	0.25	3.97	0.00	0.50	1.48
statenameOklahoma	-0.11	0.32	-0.34	0.73	-0.75	0.52
statenameOregon	1.50	0.30	5.07	0.00	0.93	2.09
statenamePennsylvania	0.68	0.24	2.80	0.01	0.21	1.17
statenameRhode Island	0.37	0.53	0.69	0.49	-0.69	1.42
statenameSouth Carolina	-0.02	0.29	-0.07	0.94	-0.58	0.55
statenameSouth Dakota	0.37	0.43	0.86	0.39	-0.48	1.20
statenameTennessee	-0.09	0.27	-0.32	0.75	-0.62	0.45
statenameTexas	0.34	0.23	1.44	0.15	-0.12	0.80
statenameUtah	-0.16	0.34	-0.47	0.64	-0.84	0.51
statenameVermont	1.36	0.66	2.05	0.04	0.11	2.77
statenameVirginia	0.48	0.26	1.83	0.07	-0.03	0.99

Table 3: Coefficients of the Voting Model (Balanced)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
statenameWashington	0.97	0.27	3.64	0.00	0.45	1.50
statenameWest Virginia	0.12	0.39	0.31	0.76	-0.66	0.88
statenameWisconsin	0.80	0.27	2.96	0.00	0.28	1.34
statenameWyoming	1.31	0.74	1.77	0.08	-0.07	2.93
maritalMarried	-0.49	0.08	-6.39	0.00	-0.64	-0.34
maritalNever married	0.09	0.09	0.97	0.33	-0.09	0.26
maritalSeparated	-0.06	0.20	-0.30	0.76	-0.44	0.33
maritalWidowed	-0.35	0.11	-3.08	0.00	-0.57	-0.13
age30-44	-0.02	0.09	-0.28	0.78	-0.20	0.15
age45-59	-0.33	0.09	-3.65	0.00	-0.51	-0.15
age60+	-0.26	0.09	-2.94	0.00	-0.44	-0.09

References

101stCongress. 1990. "Justice.gov." https://www.justice.gov/sites/default/files/eoir/legacy/2009/03/04/IMMACT1990.pdf.

270ToWin. "270toWin - 2024 Presidential Election Interactive Map — 270towin.com." https://www.270towin.com.

Alexander, Rohan. 2023. "Telling Stories with Data - 16 Multilevel Regression with Post-Stratification — Tellingstorieswithdata.com." https://tellingstorieswithdata.com/15-mrp.html.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Fontes, Raphael. 2021. "US Election 2020 — Kaggle.com." https://www.kaggle.com/datasets/unanimad/us-election-2020?resource=download.

Greg Freedman Ellis, Derek Burk, and Finn Roberts. 2024. *Ipumsr: An r Interface for Downloading, Reading, and Handling IPUMS Data*. https://tech.popdata.org/ipumsr/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

OurWorldinData. "Election Voter Turnout Rate by Age in the United States — Ourworldindata.org." https://ourworldindata.org/grapher/voter-turnout-rate-by-age-usa.

PewResearchCenter. "Religious Landscape Study — Pewresearch.org." https://www.pewresearch.org/religion/religious-landscape-study/compare/party-affiliation/by/state/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

Rudis, Bob. 2020. *Statebins: Create United States Uniform Cartogram Heatmaps*. https://gitlab.com/hrbrmstr/statebins.

Shanto Iyengar, Sean Westwood, Yphtach Lelkes. 2023. "America's Political Pulse." https://polarizationresearchlab.org/americas-political-pulse/.

StudentAid. "The Biden-Harris Administration's Student Debt Relief Plan Explained." https://studentaid.gov/debt-relief-announcement.

Team, MPC UX/UI. "IPUMS USA — Usa.ipums.org." https://usa.ipums.org/usa/.

TheEconomist. 2016. "How Does America's Electoral College Work? — Economist.com." https://www.economist.com/the-economist-explains/2016/11/07/how-does-americas-electoral-college-work?utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=18798097116&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gad_source=1&gclid=CjwKCAiA0bWvBhBjEiwAtEsoW7R8GvcqOKRKnLX29Pl8tLldvf_s8RKroJg0z4vCxdJahlTBT9f0pRoCmUAQAvD_BwE&gclsrc=aw.ds%20.

TheNewYorkerTimes. 2017. "2016 Presidential Election Results — Nytimes.com." https://www.nytimes.com/elections/2016/results/president.

Walter, Amy. "2020 Popular Vote Tracker | The Cook Political Report — Cookpolitical.com." https://www.cookpolitical.com/2020-national-popular-vote-tracker.

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

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org.

WorldPopulationReview. "US States by Race 2024 — Worldpopulationreview.com." https://worldpopulationreview.com/states/states-by-race.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Re-*

producible Computational Research, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.