Forecasting the 2024 U.S. Presidential Election*

President Joe Biden Projected to win the Popular Vote Based on MRP Analysis

Talia Fabregas

April 18, 2024

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

1 Introduction

Every four years, on the first Tuesday of November, Americans head to the polls to elect their president. The 2024 United States presidential election will take place on Tuesday November 5, 2024. In an era of unprecedented political polarization and distrust in democratic institutions, America will see a rematch of the 2020 election. President Joe Biden will seek a second term and former president Donald Trump will try to become the second president to serve two non-consecutive terms. The only U.S. president to return to office after losing his re-election bid is Grover Cleveland in 1893 (Waxman 2022).

This report builds on the lessons that I learned when I completed my first US Election Forecast last month. The survey data set that I used was provided by the Cooperative Election Study (CES) Dataverse. The CES is a nationally representative survey of 60,000 American adults, conducted before and after U.S. presidential and midterm elections (Schaffner, Ansolabehere, and Shih 2023). It aims to study the voting behavior of American adults and how it is influenced by geographic and demographic factors (Schaffner, Ansolabehere, and Shih 2023). The post-stratification data that I used was provided by the Integrated Public Use Microdata Series (IPUMS) USA online database. IPUMS provides American survey and census data, dating back to 1850, with the help of 105 statistical organizations (Ruggles et al. 2024). I selected a subset of the 2022 American Communities Survey (ACS) to use as my post-stratification data. I performed MRP analysis to estimate the 2024 U.S. presidential election results. This involves using a smaller survey dataset (~10,000 respondents) to fit a model to predict vote preference based on geographic and demographic characteristics and then applying it to a larger post-stratification dataset (~500,000 respondents). The model will learn how to classify respondents as Trump or Biden voters using the survey dataset. It will then use what

*Code and data are available at: https://github.com/taliafabs/us-election-analysis.git

it has learned to classify ACS respondents as Trump or Biden voters when applied to the post-stratification dataset. I will use these results to estimate the popular vote and electoral college results of the 2024 U.S. presidential election.

This report features four sections. In Section 2, I discuss the context of the survey and post-stratification dataset, present the results of exploratory data analysis, and use tables and visualizations to show what the variables look like and explain how they interact. In Section 3 I explain, outline, and justify the Bayesian logistic regression model that I used to predict vote preference. In Section 4, I use tables and graphs to present the results of my MRP analysis, which include a popular vote and an electoral college prediction for the 2024 U.S. election. In Section 5, I discuss how my analysis was conducted in more detail, what we can learn from the popular vote and electoral college predictions, the limitations of my datasets and model, why logistic regression was used and the case for (and against) SoftMax regression in U.S. election analysis, and how I hope to extend and improve this report in the future.

I used R programming language (R Core Team 2023) and the `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot` (Wickham 2016), `knitr` (Xie 2014), `readr` (Wickham, Hester, and Bryan 2023), `arrow` (Richardson et al. 2023), and `rstanarm` (Goodrich et al. 2022) packages to clean my survey and post-stratification datasets, create my data visualizations, fit my logistic regression model, and apply my logistic regression model.

2 Data

My survey data comes from the 2022 Comprehensive Election Study (CES), provided by Harvard Dataverse. Sixty research teams participated in the 2022 CES. It is part of the ongoing Cooperative Election Study (CES), which has been conducted every year since 2006 to study elections in the United States using large-scale survey datasets (Schaffner, Ansolabehere, and Shih 2023). Schaffner, Ansolabehere, and Shih (2023) explain that the CES aims to build off of the work of the 2005 Massachusetts Institute of Technology Public Opinion Research and Training Lab (PORTL) study. The CES was known as the Cooperative Congressional Election Study (CCES) before 2020 (Schaffner, Ansolabehere, and Shih 2023).

2.1 Survey Data

I selected a random subset of the final release of the 2022 CES Common Content Dataset as my survey data. The 2022 CES Common Content Dataset is a nationally representative sample of 60,000 American adults and it includes sample identifiers such as state and congressional district, demographic profile questions, pre-election questions, post-election questions, and questions about candidate and party preferences (Schaffner, Ansolabehere, and Shih 2023). Most survey respondents are registered with the YouGov panel to receive notifications about surveys and are rewarded with points that can be exchanged for gift cards, however some come from other survey platforms and online ads (Schaffner, Ansolabehere, and Shih 2023).

As outlined by Schaffner, Ansolabehere, and Shih (2023), not all responses are included in the 2022 CES Common Content Dataset to ensure that it is as representative as possible. I use a random subset of 10,000 of the 60,000 respondents. I explain my sub-setting process and why my random subset is representative of the 2022 CES Common Content Dataset in Section A.

I do not use the entire questionnaire in my study; I focus on survey questions about demographics, party identification, and vote choice. The variables from the 2022 CES Common Content Dataset that I selected include: `pid3`: 3-point party identification, `presvote16post`: who the respondent voted for in the 2016 U.S. presidential election, `presvote20post`: who the respondent voted for in the 2020 U.S. presidential election, `state`: the state that the respondent lives in, `gender4`: gender identity, `birthyr`: year of birth, `race`: race or ethnicity, `hispanic`: whether the respondent identifies as Hispanic, `educ`: highest level of education completed, and `urbancity`: the type of area that the respondent lives in. `gender4` initially included male, female, non-binary, and other. I re-named it to `sex` and only used male and female because the `sex` variable in my post-stratification dataset provided by IPUMS only includes male and female. This is one example of how data science can ignore LGBTQ+ identities and reduce humanity to something that can easily be quantified (Keyes 2019).

The 2022 CES Common Content Dataset focuses on the periods leading up to (September 29 to November 8) and following (November 10 to December 15) the 2022 U.S. midterm elections (Schaffner, Ansolabehere, and Shih 2023). While the survey contains specific questions about party identification, strength of party identification, and votes congressional, senate, and gubernatorial elections, it does not ask about preferred 2024 presidential candidate. The survey was conducted in late 2022, when the Republican 2024 presidential nominee was not yet known. Party identification and recent vote choices, especially in the 2020 presidential election where the nominees were also Joe Biden and Donald Trump, are strong indicators of 2024 vote choice. I used `pid3`: party identification, `presvote16post`: 2016 presidential vote, and `presvote20post` to determine whether each respondent would vote for Joe Biden or Donald Trump in the 2024 presidential election and construct the `vote_biden` binary indicator variable. `vote_biden` is equal to 1 if a respondent's preferred 2024 presidential candidate is Joe Biden, and 0 if it is Donald Trump. I discuss the creation of the `vote_biden` variable in more detail in Section A.

Table 1: Presidential preferences of survey subset respondents living in urban, rural, and suburban areas

	Biden %	Trump %
Urban	74.58	25.42
Suburban	59.19	40.81
Rural	44.31	55.69

Figure 5 illustrates the proportion of subset survey respondents in each state who plan to support President Biden in the 2024 election. My subset survey dataset appears to show

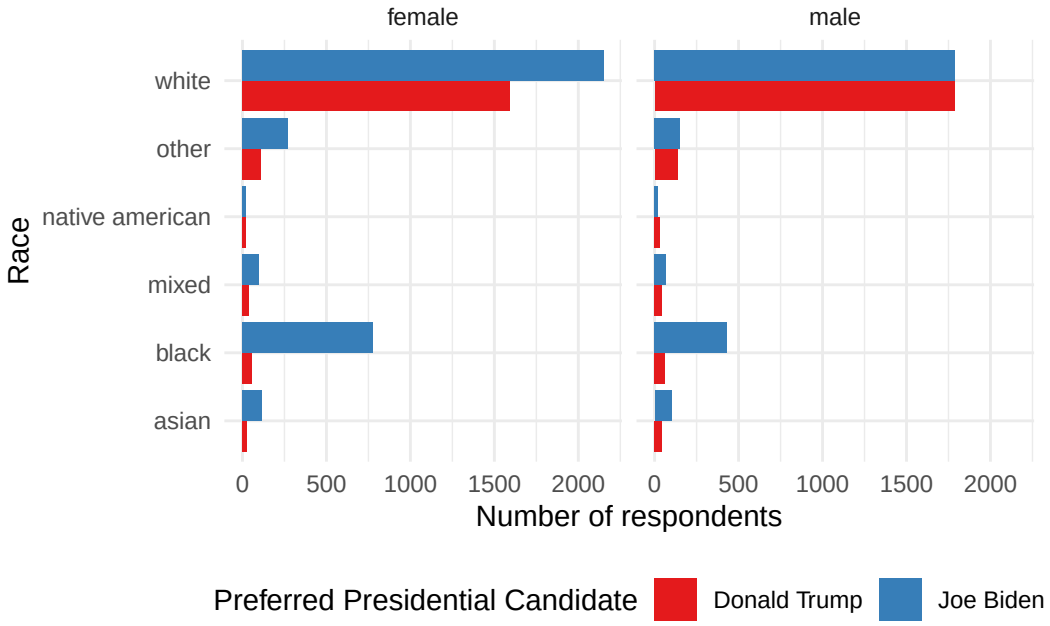


Figure 1: Preferred presidential candidates of survey subset respondents, by gender and race

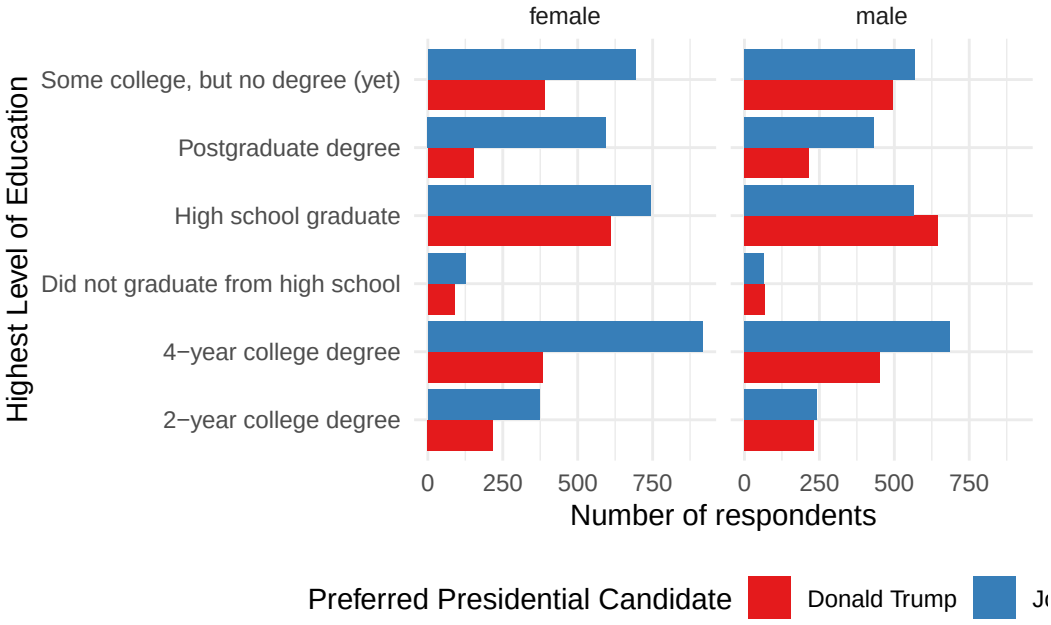


Figure 2: Preferred presidential candidates of survey subset respondents, by highest level of education

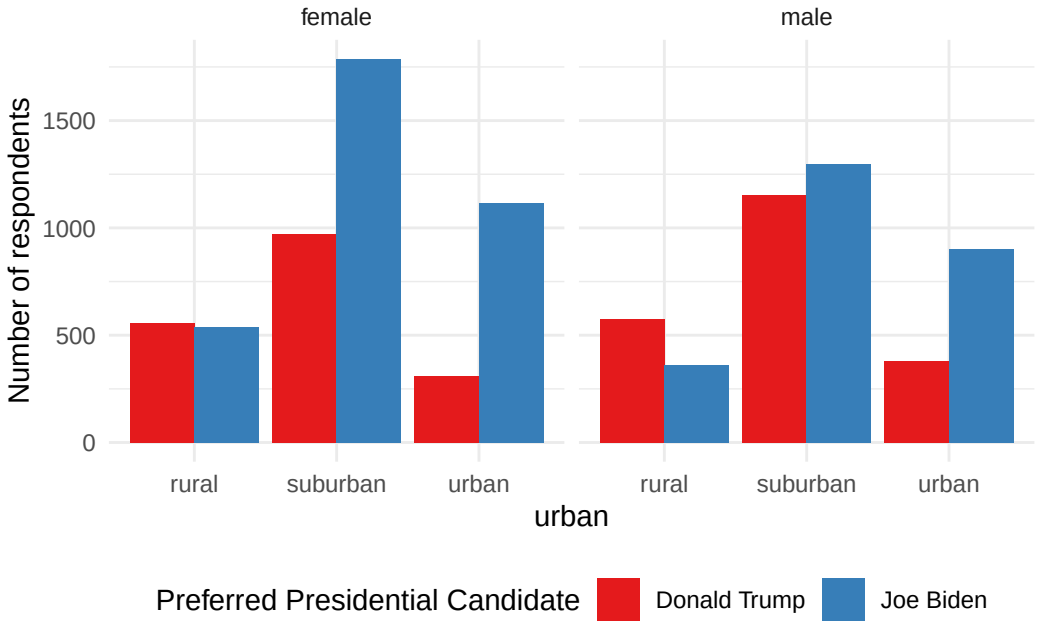


Figure 3: Preferred presidential candidate of subset respondents living in urban vs rural areas

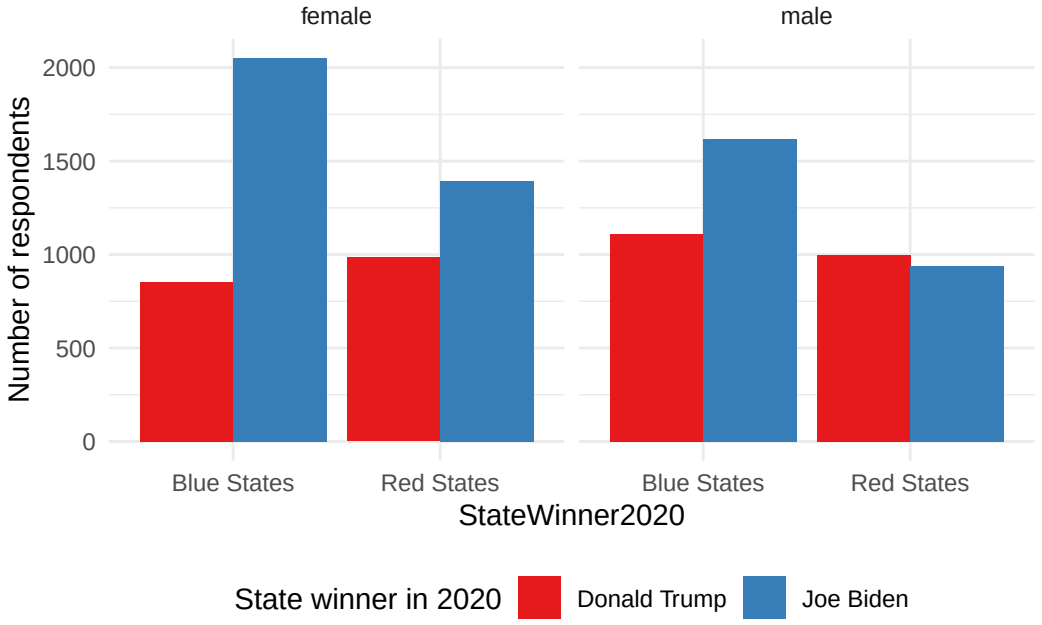


Figure 4: Preferred presidential candidate of subset respondents in states won by Trump vs Biden in 2020

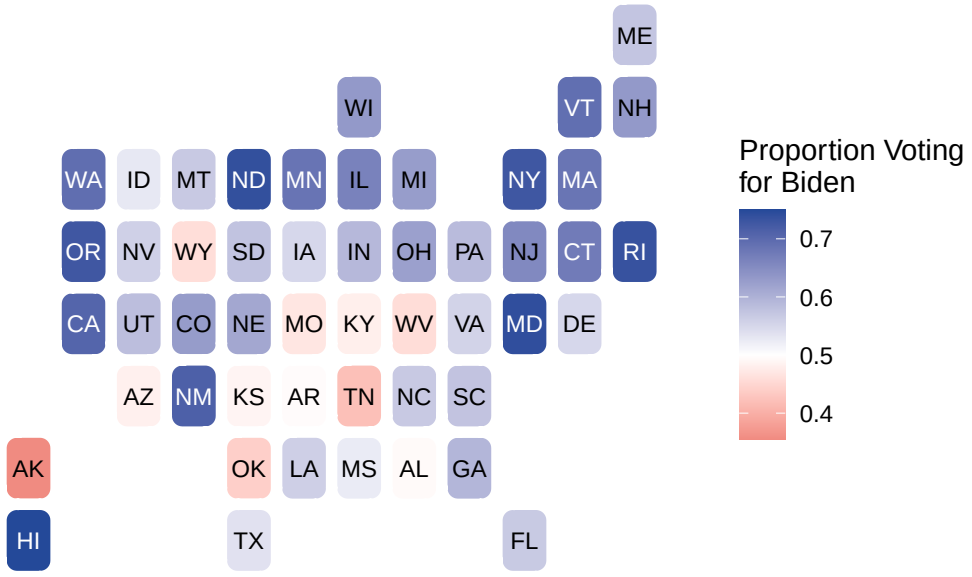


Figure 5: Electoral college map based on the subset survey data

Table 2: Popular vote and electoral college based on subset survey data

Survey Estimate:	Biden	Trump
Num Votes	6003.00	3946.00
% Votes	60.34	39.66
Electoral College	460.00	78.00

stronger support for President Joe Biden than the general U.S. electorate both overall and at the state level. As shown in Section B, this is not unique to the subset data; the complete 2022 CES Common Content Dataset shows strong support for President Biden and the Democratic Party based on 2016 vote choice (`presvote16post`), 2020 vote choice (`presvote20post`), and party identification (`pid3`). However, as seen in subset-state2020_subset, respondents who live in states won by Biden in 2020 (blue states) were more likely to support him over Trump in 2024 than those who live in states won by former president Trump in 2020 (red states). To account for this difference, and the fact that the survey dataset I used has strong support for President Biden, I created the `biden_won` variable, which is equal to 1 if a state was carried by President Biden in 2020 and 0 if it was carried by former President Trump in 2020. The data cleaning steps that I used to create the `biden_won` variable are outlined in Section A.

2.2 Poststratification Data

The post-stratification data that I used is a subset of the 2022 American Community Survey (ACS) from IPUMS (Ruggles et al. 2024). It includes 500,000 census respondents and was

downloaded from the IPUMS online database. I selected variables that match the ones in my survey dataset so that my model could easily be applied to my post-stratification data. The variables that I selected include:

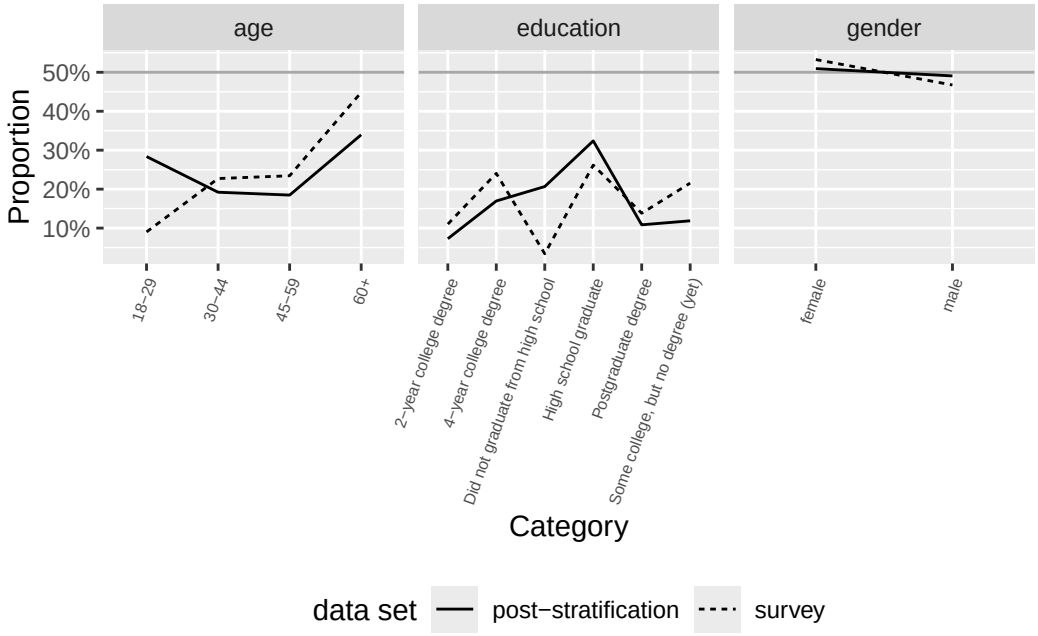


Figure 6: Survey vs post-stratification voter demographics

3 Model

I performed multi-level regression with post-stratification (MRP) to predict support for president Joe Biden and former president Donald Trump in the 2024 U.S. presidential election. To perform MRP analysis, I fit the model on the survey data and applied it to the post-stratification data. Fitting the logistic regression model on the survey data teaches it to classify each respondent as a Biden or Trump voter based on the state that they live in, whether Biden or Trump won that state in 2020, their sex, age bracket, race, highest level of education, and whether they live in an urban area. I then apply the model to my post-stratification dataset (Ruggles et al. 2024) to forecast the popular vote and electoral college results for the 2024 U.S. presidential election. When applied to my post-stratification dataset, the logistic regression model uses the same variables (state, whether Biden or Trump won that state in 2020, sex, age bracket, race, highest level of education, and urban) and what it learned from being fit on the survey dataset to classify each census respondent as a Biden or Trump voter.

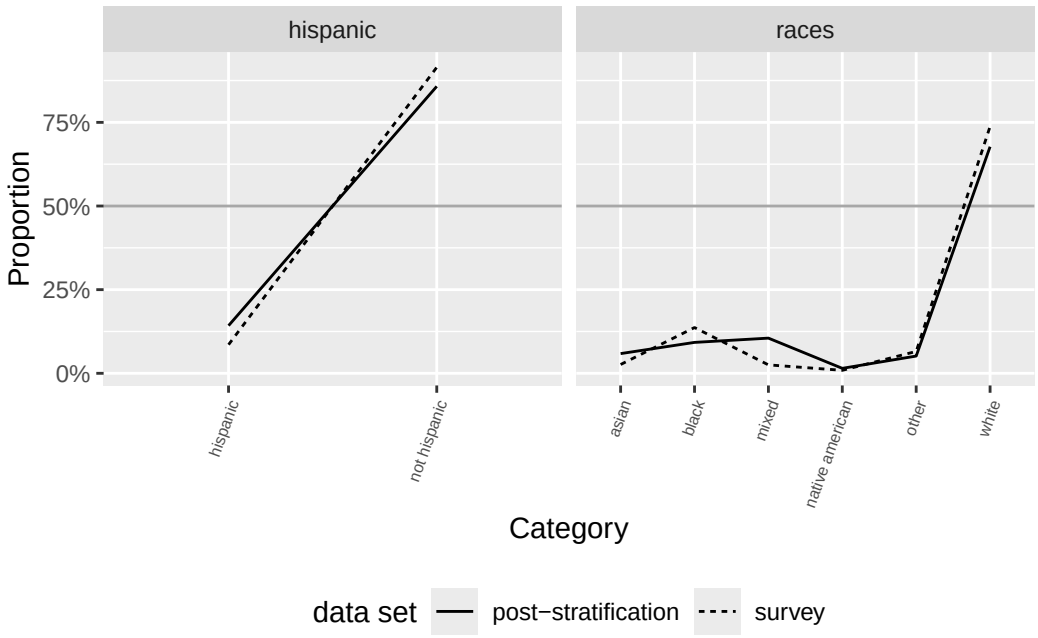


Figure 7: Survey vs post-stratification voter race demographics

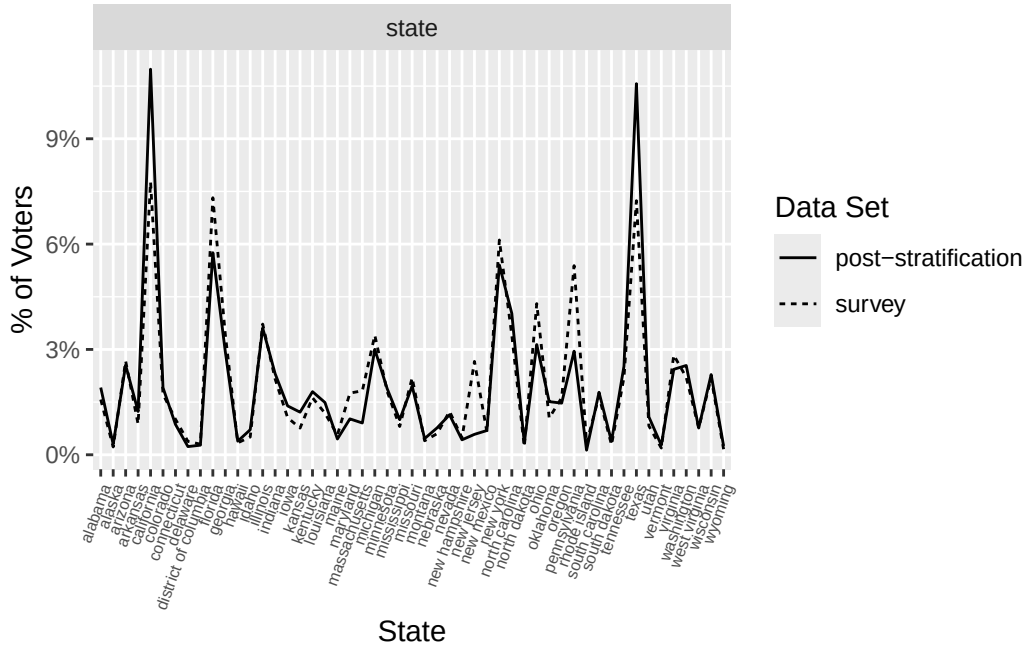


Figure 8: Survey and Post-Stratification Data Proportion of Voters by State

3.1 Model set-up

I built my Bayesian Logistic Regression model in R (R Core Team 2023) using the `stan_glm` function and the default priors of the `rstanarm` package (Goodrich et al. 2022). My model is as follows:

$$
\begin{aligned}
vote\_biden_i|\pi_i &\sim \text{Bern}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \text{state}_i + \beta_2 \text{biden\_won}_i + \beta_3 \text{sex}_i + \beta_4 \text{age\_bracket}_i \\
&\quad + \beta_5 \text{race}_i + \beta_6 \text{hispanic}_i + \beta_7 \text{educ}_i + \beta_8 \text{urban}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5) \\
\beta_4 &\sim \text{Normal}(0, 2.5) \\
\beta_5 &\sim \text{Normal}(0, 2.5) \\
\beta_6 &\sim \text{Normal}(0, 2.5) \\
\beta_7 &\sim \text{Normal}(0, 2.5) \\
\beta_8 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

where the binary indicator variable vote_biden_{i} is equal to 1 if the respondent's preferred 2024 presidential candidate is President Joe Biden (D) , or 0 if their preferred candidate is former President Donald Trump (R). My model uses logistic regression, it is not without tradeoffs. Firstly, logistic regression can only be used to predict a binary outcome variable. As far as my model is concerned, the only two vote choices for U.S. adults in the upcoming presidential election are Joe Biden and Donald Trump. The possibilities of voting third-party or not voting at all are not considered. I discuss the tradeoffs, benefits, weaknesses, and limitations associated with the use of a logistic regression model to forecast the U.S. election in more depth in Section 5.4.

3.2 Model justification

4 Results

4.1 Popular Vote Prediction

I got my popular vote prediction by applying the model outlined in Section 3.1 to my post-stratification data set to predict the 2024 preferred presidential candidate of each ACS 2022

respondent (Ruggles et al. 2024). I estimated the upper, lower, and mean quantile of the proportion of voters in each state supporting President Biden. The lower quantile, mean, and upper quantile@tbl-poststrat_results1

Table 3: 2024 U.S. election popular vote estimates based on post-stratification analysis

Estimate:	Biden %	Trump %
Lower Estimate	48.55	51.45
Mean Estimate	55.47	44.53
Upper Estimate	62.56	37.44

4.2 Electoral College Prediction

Table 4: 2024 U.S. election electoral college estimates based on multilevel regression with post-stratification (MRP) analysis

Electoral College Estimate:	Biden	Trump
Lower Estimate	220	318
Mean Estimate	413	125
Upper Estimate	517	21

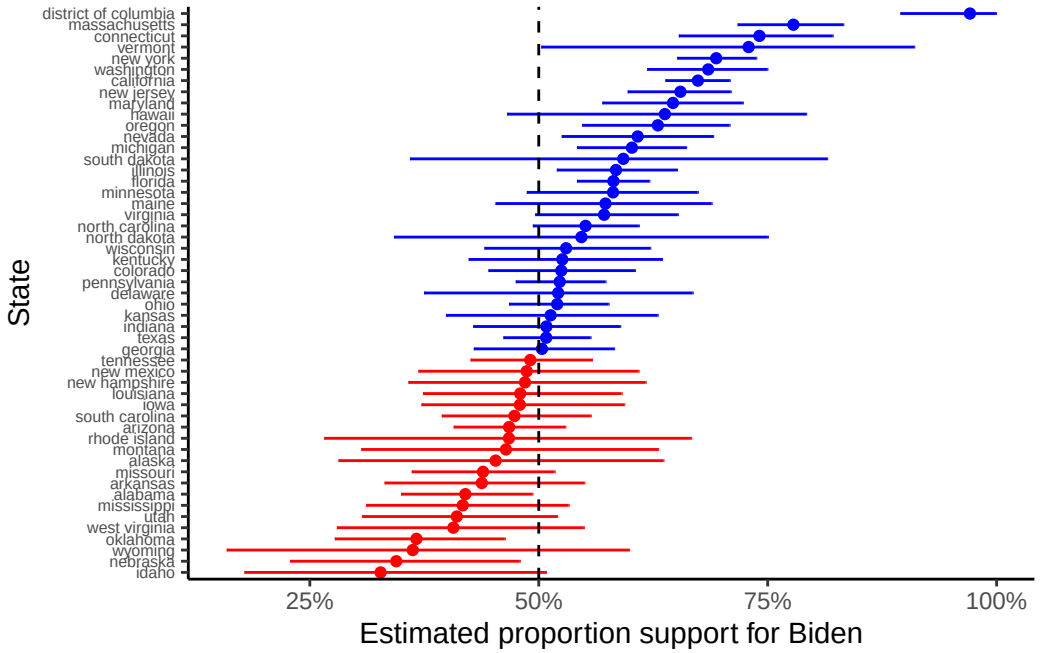


Figure 9: Estimated proportion of each state voting for Biden in 2024 based on MRP analysis

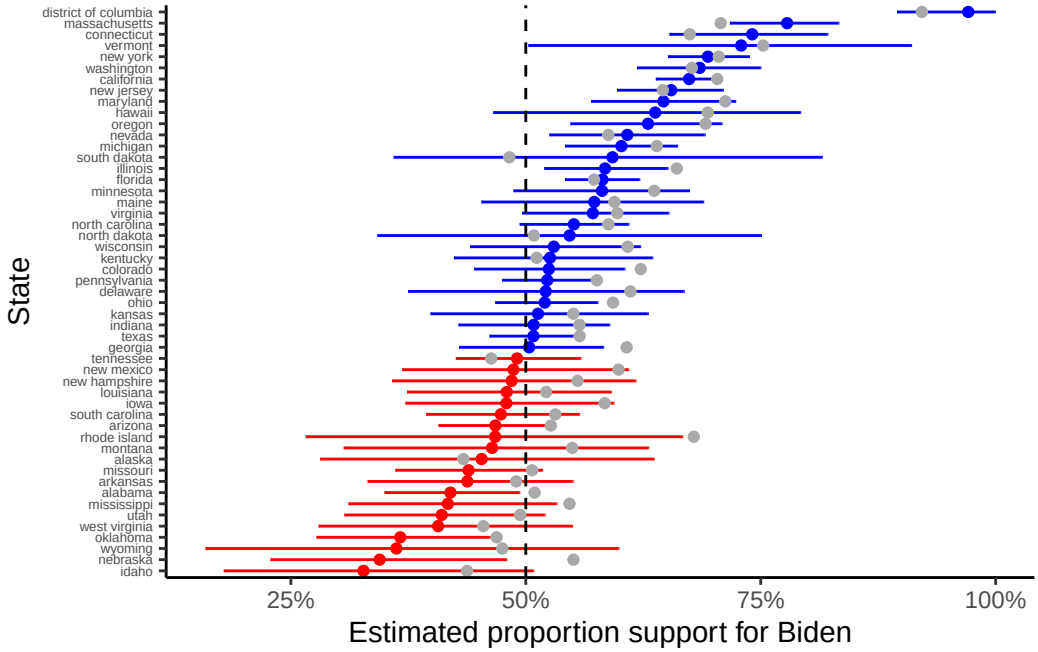


Figure 10: Estimated proportion of each state voting for Biden in 2024 Post-Stratification vs subset Survey Data

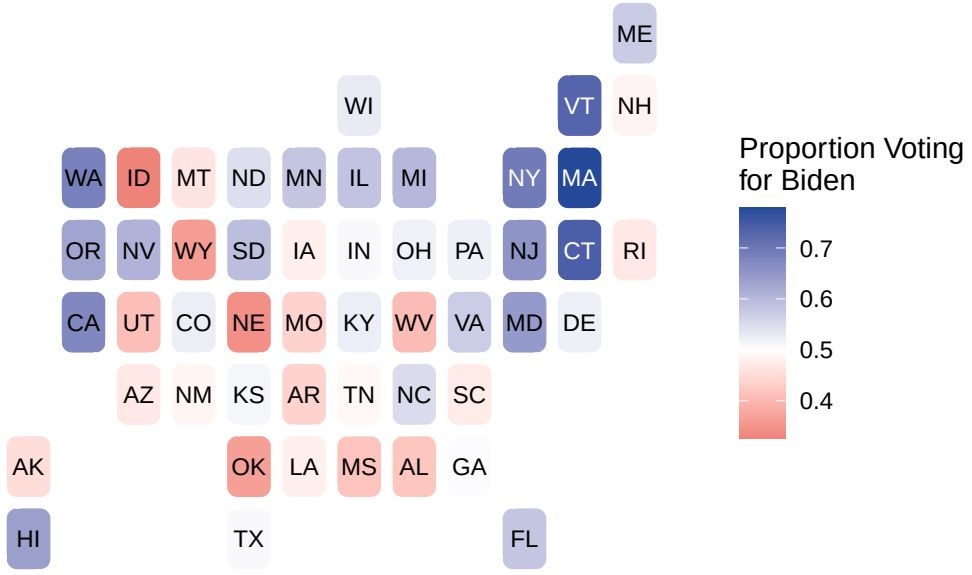


Figure 11: Electoral map based on MRP analysis

5 Discussion

5.1 Popular Vote Prediction

5.2 Swing States had Close Margins and Large Error Ranges

Swing states and error ranges in the MRP analysis electoral college prediction

5.3 Weaknesses and Limitations of the Survey Dataset

5.4 The Limitations of Logistic Regression and the Case for (and against) SoftMax Regression

Logistic regression was completely appropriate in this case. SoftMax has more parameters and is more complex; this means that a SoftMax regression model is more likely to overfit.

5.5 Next Steps

Python, more recent data set, softmax regression, gradient descent Split the survey data into training, validation, and test Use gradient descent to find the optimal weights to maximize validation accuracy Apply the model to the post-stratification data Softmax regression does risk overfitting

Appendix

A Additional data cleaning details

I created a binary variable, `vote_biden` that is equal to 1 if a respondent's preferred 2024 presidential candidate is Joe Biden, or 0 if it is Donald Trump. My Cooperative Election Study Common Content survey dataset was put together in 2022, so respondents were not asked about their preferred 2024 presidential candidate. However, they were asked about their party identification (`pid3`), who they voted for in the 2016 presidential election (`presvote16post`), and who they voted for in the 2020 presidential election (`presvote20post`). I started by filtering out respondents who have no party affiliation and did not vote for either major party nominee in the 2016 and 2020 presidential elections. This was done because I am using a logistic regression model, which is only capable of performing binary classification. I used this information to create the `vote_biden` indicator variable; if a respondent's `pid3` is Democratic, or they voted for the Democratic nominee in 2016 or 2020, I label them as a Biden voter (`vote_biden = 1`). Otherwise, I label them as a Trump voter (`vote_biden = 0`). As previously stated, respondents who both voted third-party and have no party affiliation were not considered because my model is not capable of performing multi-class classification and there is no indication of whether they prefer Joe Biden or Donald Trump.

I downloaded the 2020 National Popular Vote Tracker from Wasserman et al. (2020), cleaned it to change the state column to match my survey and post-stratification data sets, and selected the `state`, `biden_won`, `dem_votes`, `rep_votes`, `other_votes`, and `dem_percent` columns. I then left-joined the 2020 National Popular Vote Tracker with both my survey and post-stratification datasets so that I could include Biden's performance in each state in the 2020 election to my model as a predictor for `vote_biden`. I added the binary variable, `biden_won`, to both my survey and post-stratification analysis datasets. `biden_won` is equal to 1 if Joe Biden won the electoral college votes of the state that the respondent lives in in the 2020 presidential election, and 0 if Donald Trump won the state in the 2020 presidential election. Maine and Nebraska have split electoral college votes; this means that the presidential nominee who wins each congressional district receives an electoral college vote, and an additional 2 electoral college votes are awarded for winning the statewide popular vote. In the 2020 presidential election, Joe Biden won the statewide popular vote in Maine, while Donald Trump won the statewide popular vote in Nebraska, although Biden won one congressional district in Nebraska, and Trump won one congressional district in Maine. `biden_won` corresponds to the presidential nominee who won the statewide popular vote in the 2020 election, so `biden_won = 1` for respondents who live in Maine, and `biden_won = 0` for respondents who live in Nebraska. I use information from the 2020 National Popular Vote tracker in various places throughout my analysis and discussion.

B Additional survey data details

The original survey dataset contained over 50,000 responses, but it was subset to 10,000 so that R and `rstanarm` could handle it (R Core Team 2023). The `glm` function of the `rstanarm` package was used to fit the logistic regression model to predict 2024 presidential vote choice based on state, whether Biden or Trump won that state in 2020, sex, age bracket, race, and highest level of education completed. Although Schaffner, Ansolabehere, and Shih (2023) advised against sub-setting the 2022 CES Common Content Dataset, this was a necessary step for me. When I tried to fit the model using the original survey data set, it took hours to run and I was not able to post-stratify due to the following error: `Error: vector memory exhausted (limit reached?)`. 10,000 of the more than 50,000 responses were randomly selected using the `sample` function of base R. The visualizations below show the results of the exploratory data analysis conducted on the original survey data set. As seen in Figure 12, Figure 13, Figure 14, and Figure 16, the results are similar to the ones shown in Section 2.1. For the reasons outlined above, I am confident that my random subset is representative of the original 2022 CES Common Content Dataset (Schaffner, Ansolabehere, and Shih 2023).

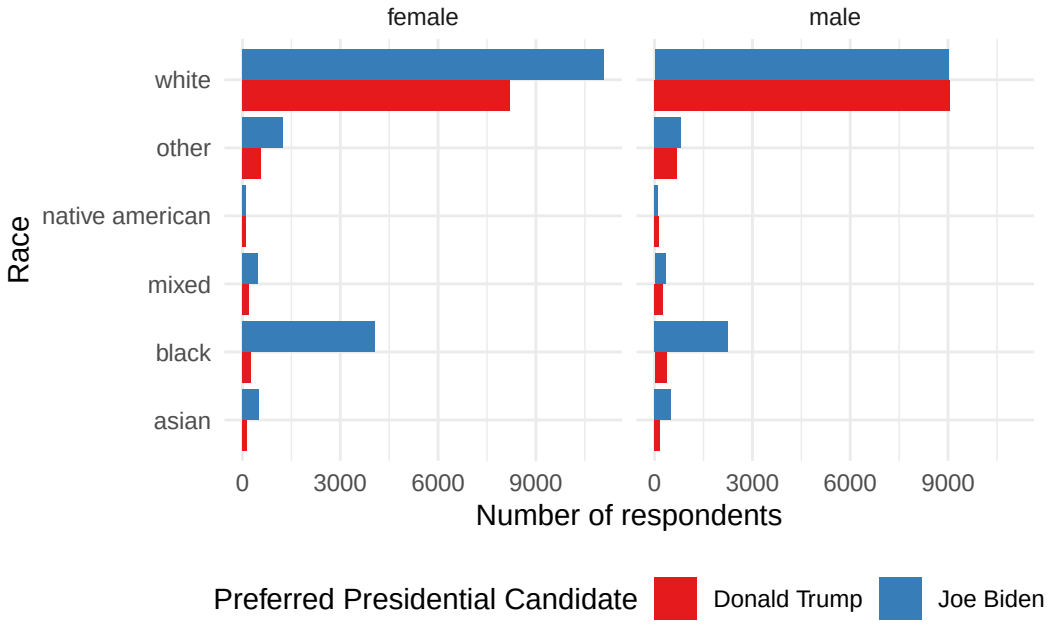


Figure 12: Preferred presidential candidates of survey respondents, by gender and race

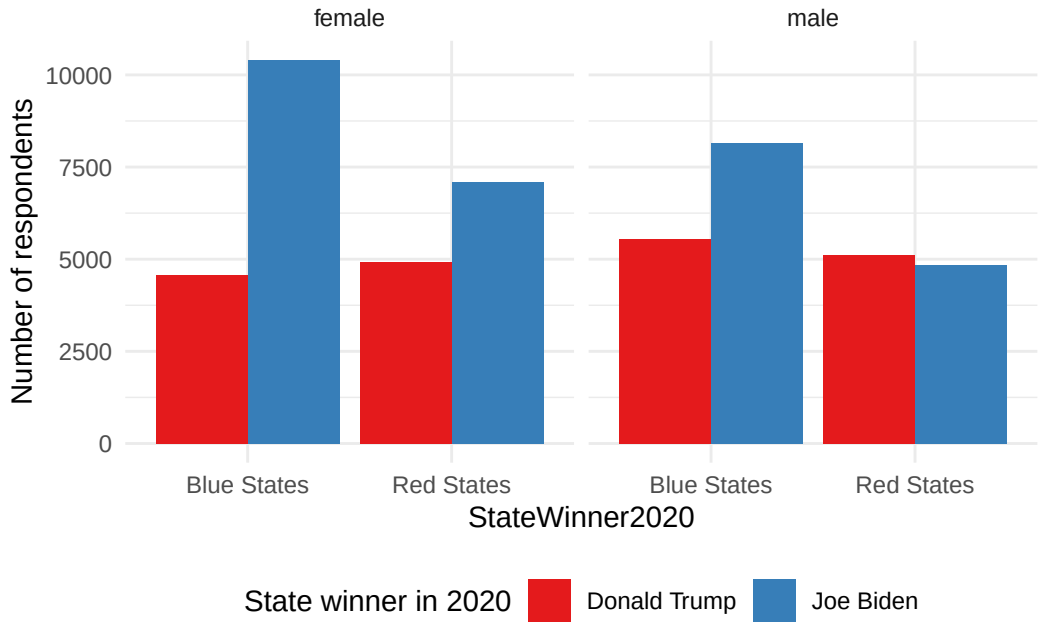


Figure 13: Preferred presidential candidate of survey respondents in states carried by Trump vs Biden in 2020

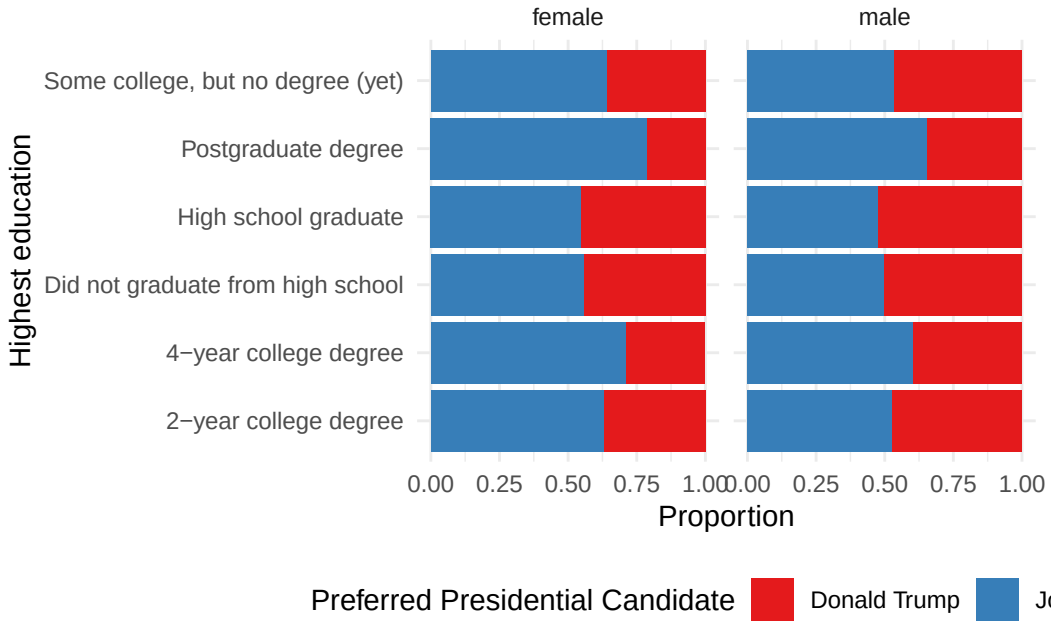


Figure 14: Preferred presidential candidates of survey respondents, by highest level of education

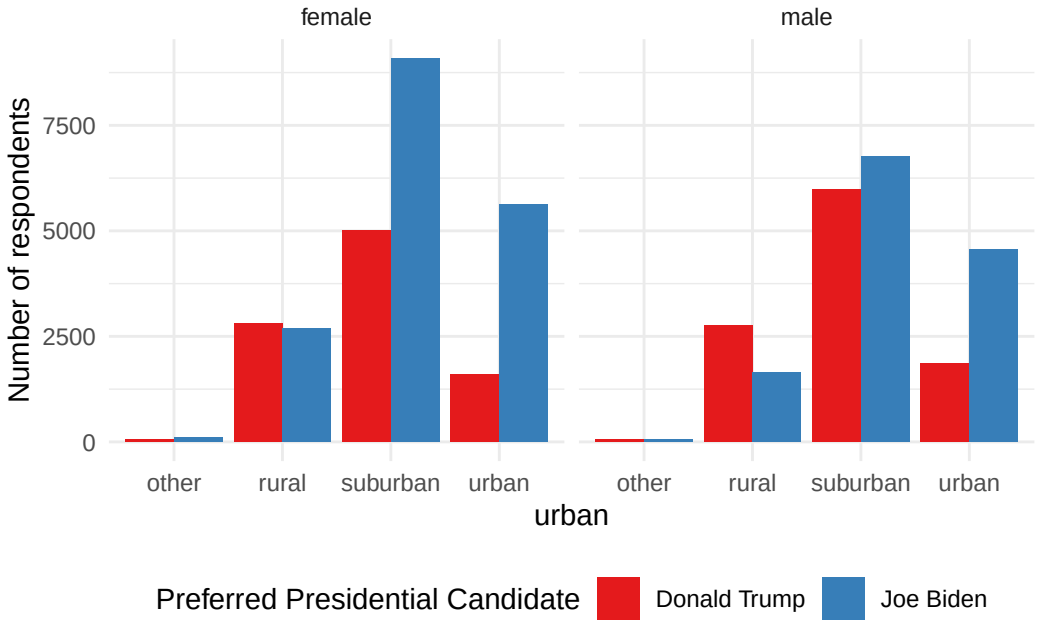


Figure 15: Preferred presidential candidate of survey respondents living in urban vs rural areas

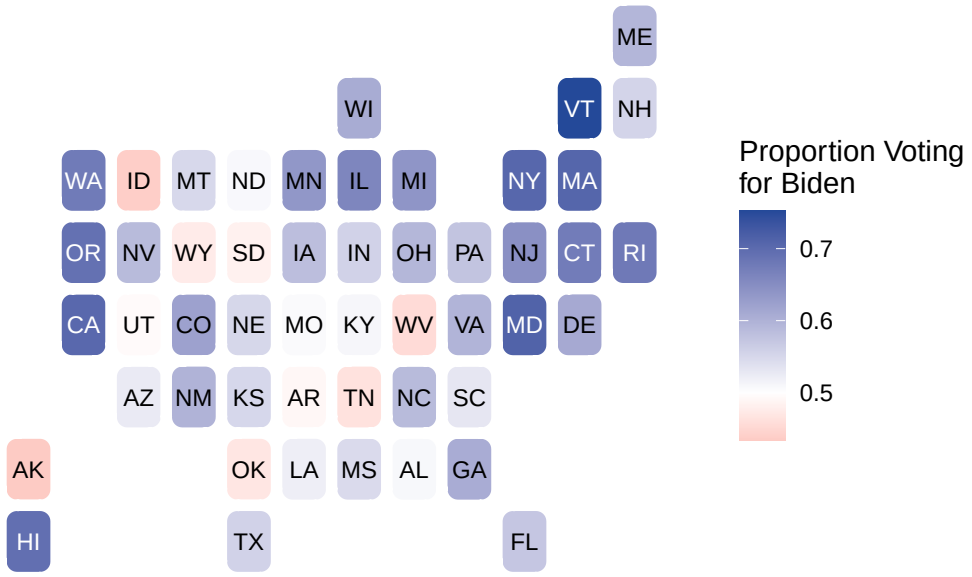


Figure 16: Electoral college map based on the survey dataset

Table 5: Presidential preferences of survey respondents living in urban, rural, and suburban areas

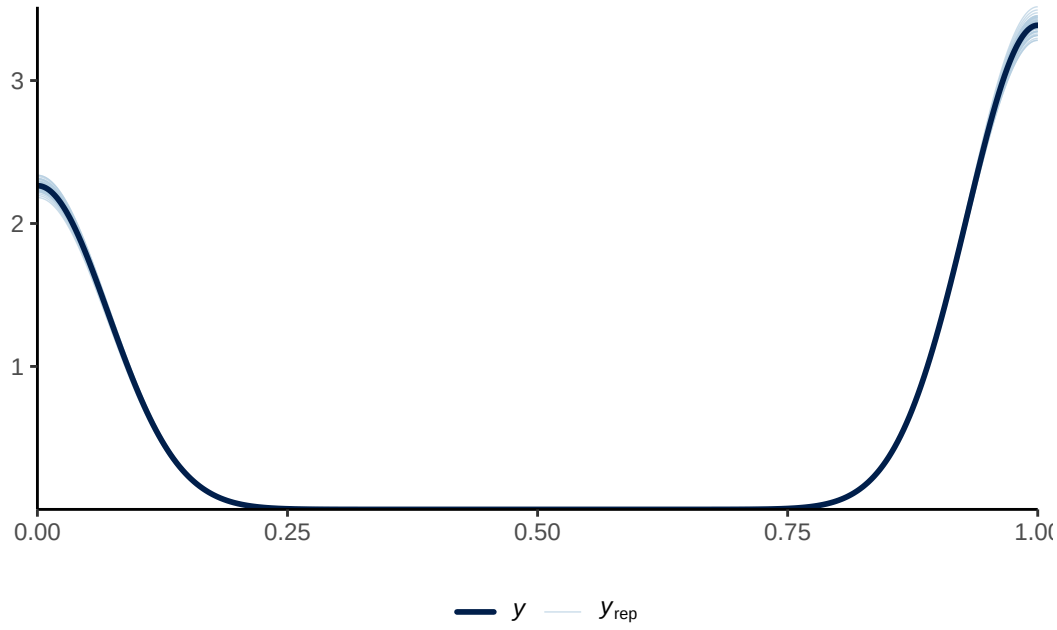
	Biden %	Trump %
Urban	74.67	25.33
Suburban	59.02	40.98
Rural	43.81	56.19

C Model details

C.1 Posterior predictive check

C.2 Markov Chain Monte Carlo

C.3 Credibility intervals

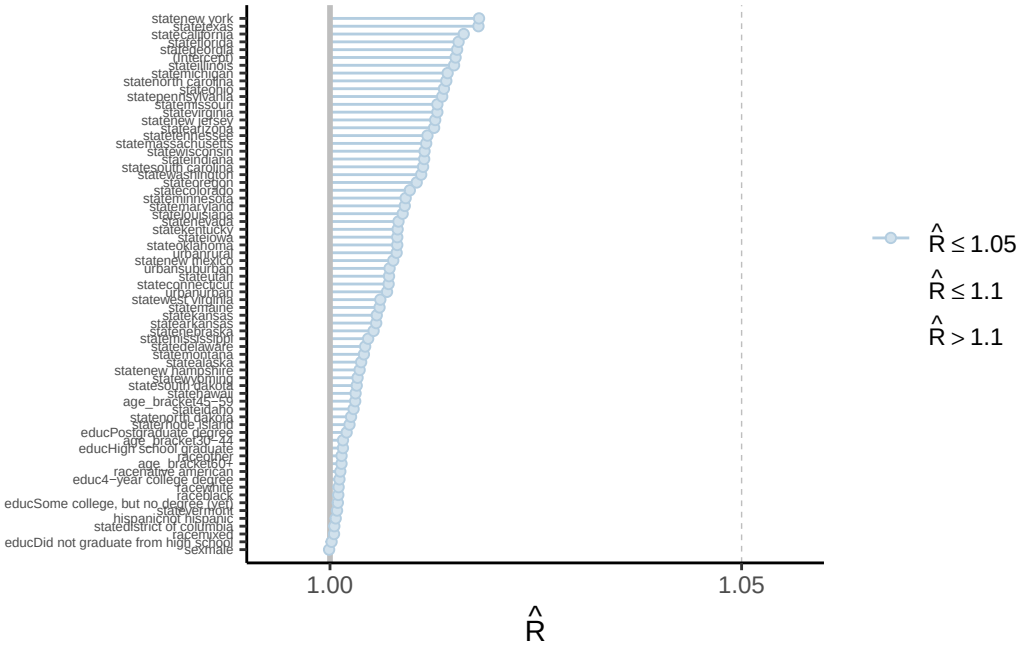


(a) Posterior prediction check

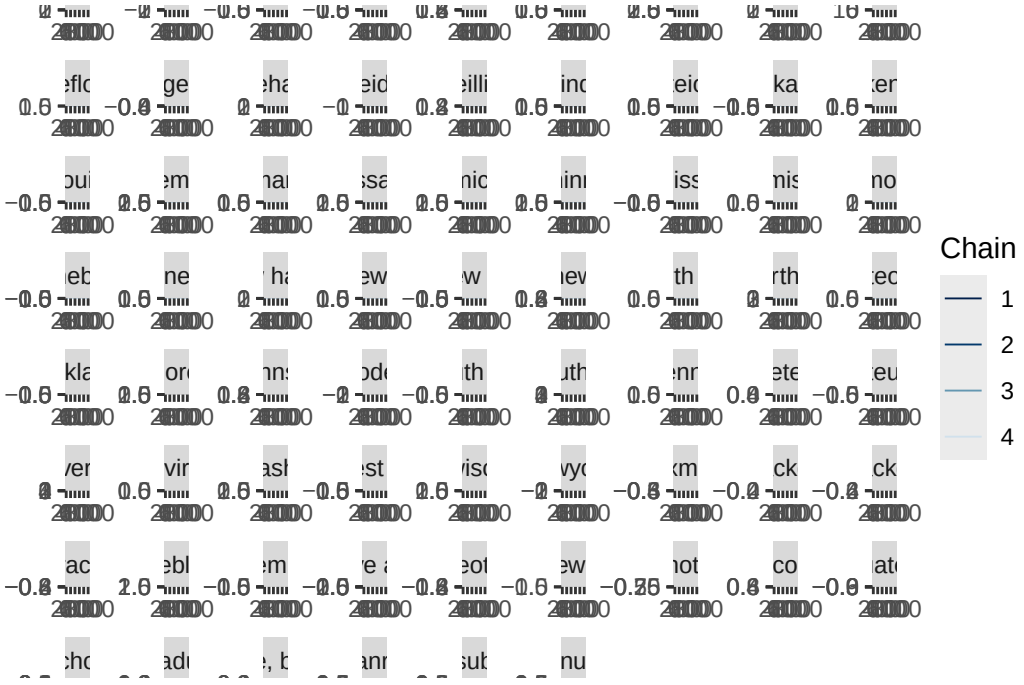
bracket30−44	sexmale	stateindiana	s
bracket45−59	statealaska	stateiowa	s
bracket60+	statearizona	statekansas	s
4−year college degree	statearkansas	statekentucky	s
Did not graduate from high school	statecalifornia	statelouisiana	s
High school graduate	statecolorado	statemaine	s
Postgraduate degree	stateconnecticut	statemaryland	s
Some college, but no degree (yet)	statedelaware	statemassachusetts	s
unicnot hispanic	statedistrict of columbia	statemichigan	s
olack	stateflorida	stateminnesota	s
nixed	stategeorgia	statemississippi	s

(b) Comparing the posterior with the prior

Figure 17: Examining how the model fits, and is affected by, the data



(a) Rhat plot



(b) Trace plot

Figure 18: Checking the convergence of the Markov Chain Monte Carlo (MCMC) algorithm

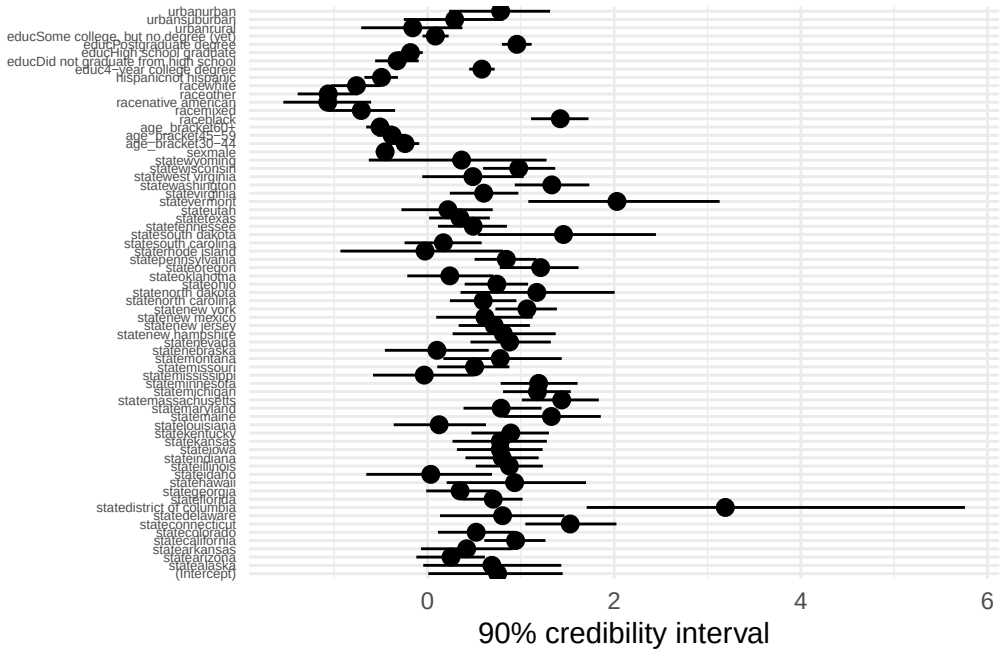


Figure 19: 90% Credibility intervals for the predictors of vote_biden

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Keyes, Os. 2019. *Counting the Countless*. https://reallifemag.com/counting-the-countless/.

R Core Team. 2023. *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. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Renae Rodgers Grace Cooper Stephanie Richards, and Megan Schouweiler. 2024. *IPUMS USA: Version 15.0 [ACS 2022]*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. "Cooperative Election Study Common Content, 2022." Harvard Dataverse. https://doi.org/10.7910/DVN/PR4L8P.

Wasserman, David, Sophie Andrews, Leo Saenger, Lev Cohen, Ally Flinn, and and Griff Tatarsky. 2020. *2020 National Popular Vote Tracker*. https://www.cookpolitical.com/2020-national-popular-vote-tracker.

Waxman, Olivia. 2022. *Donald Trump, Grover Cleveland, and the History of Trying to*

Win Back the White House. TIME. https://time.com/6234562/nonconsecutive-terms-president-grover-cleveland-donald-trump/.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.