Biden Projected to Win the 2024 US Presidential Elections*

Analysis using data from American Community Survey 2022 and Ipsos Poll 2023 leveraging Bayesian Logistic Regression with Post-Stratification

Monica Sainani Kaavya Kalani

March 16, 2024

We analyze American voting behavior, focusing on demographic, socioeconomic, and geographic factors. A Bayesian Logistic Regression model with Post-Stratification is used to predict the election outcome between Biden and Trump. We find a strong relationship between age, family income, education, state of residency, sex and likely voting outcome. Our central forecast is a Biden re-election. This research provides insights into the factors affecting political engagement in the U.S., potentially guiding future electoral strategies and policy discussions.

Table of contents

*Code and data supporting this analysis is available at: https://github.com/sainanim/Predictions-for-the-US-2024-Elections

1 Introduction

The United States stands as one of the most influential political entities globally. This status implies that the leadership of the nation can cause indirect influence on global affairs (Barnes 2021). Since its inaugural election in 1788, the US Presidential elections have been a quadrennial tradition, building the course of the nation's history every four years. This year, the 60th presidential election will be held on Tuesday, November 5, 2024. As the nation of the United States approaches their 2024 presidential election, the political landscape is likely to undergo significant shifts and intensification.

The 2020 election brought upon extreme conflict, serious economic and social challenges which changed the societal dynamics for Americans (Barnes 2021). Although the decisions are made every 4 years, the United States conducts surveys every year to determine the opinions of citizens and predict the outcome of the elections. In our paper, we conduct a statistical analysis to forecast the major decision by focusing on the two major political parties, Democratic and Republican.

The rivalry between the Democratic and Republican parties, as embodied by their opposing candidates, Joe Biden and Donald Trump, has been an iconic feature of the modern American political scene. The Democratic Party's nominee is the veteran politician and current president Joe Biden. With an emphasis on middle-class families, Biden's program prioritizes measures that advance racial justice, healthcare reform, climate action, and economic regeneration. He is an advocate of a more compassionate and inclusive style of governing, hoping to bring the country back together and bring decorum back into political conversation.

On the other hand, the current Republican Party standard-bearer and former president Donald Trump represents a more populist and nationalist agenda. A major focus on economic nationalism, immigration restrictions, tax cuts, and deregulation have been highlights of Trump's leadership. His administration prioritized themes like law and order, trade protectionism, and border security as part of its conservative agenda. Trump's aggressive approach to politics is a

defining characteristic of his leadership style; it frequently divides public opinion and provokes strong emotions from both supporters and opponents.

Our estimand is the relationship between demographic, socio-economic, and geographic variables (such as age, income, education, state of residence, and sex) and political preferences. Using this extensive dataset, our goal is to identify trends and factors that influence voting behavior, clarifying the complex relationship between personal traits and local factors that influences election results. Eventually, we aim to predict will Joe Biden win the US 2024 Presidential Elections or will it be Donald Trump.

We use the survey data from the Ipsos Poll [Roper #31120637] along with census data from the American Community Survey (ACS) 2022 sample to understand voting preferences and predict the US 2024 Presidential Elections winner. This is done by leveraging a Bayesian Logistic Regression model with our survey data, which is then post-stratified with our census data. We evaluate voting preferences over some demographic, socioeconomic, and geographic aspects across the nation and then make a final prediction.

Our analysis led to us finding that overall, the younger generations, females, and people of the Midwest and southern states are highly supportive of the Democratic Party led by Biden. Our overall forecast is a Joe Biden re-election in the 2024 US Presidential Elections.

Our research adds to a greater understanding of the factors influencing voting decisions through comprehensive statistical analysis, yielding insightful data that may influence future election campaigns and policy discussions. Through highlighting the subtleties of voting patterns, our research provides a critical insight on the condition of American democracy and the factors motivating involvement in politics and engagement.

The paper is further organized into four sections: Data, Model, Results, and Discussion. In the Data section, we discuss how the datasets to be used for the analysis were obtained and pre-processed. We will explain the variables of interest in the datasets for the analysis. The Model section describes the model being used for the analysis. The Results section will then highlight and discuss the trends and associations found during the analysis. Lastly, the Discussion section will talk about some interesting trends found in Results in depth, link it to the real world and also highlight the weaknesses and future of our analysis.

2 Data

For this analysis, we have used two datasets. The datasets were cleaned and analysed using the statistical programming software R (R Core Team 2023) along with the help tidyverse (Wickham et al. 2019), knitr (Xie 2014), ggplot2 (Wickham 2016), here (Müller 2020), dplyr (Wickham et al. 2023), arrow (Richardson et al. 2024), rstanarm (Goodrich et al. 2024), broom.mixed (Bolker and Robinson 2022), modelsummary (Arel-Bundock 2022) and kableExtra (Zhu 2024).

2.1 Survey Data

Our survey data is from the Ipsos Poll [Roper #31120637]. It was collated from a web-based survey conducted on November 3 - 4, 2023 using KnowledgePanel. The survey is conducted by Ipsos and the sponsor is ABC News. This poll is based on a nationally representative probability sample of 949 adults aged 18 or older with oversamples among Black and Hispanic respondents. The poll is a collection of demographic information and political preferences.

A person becomes an entry in this dataset if they were 18 years or older on November 3-4, 2023, an American national and participated in this poll. They are recruited through a scientifically developed addressed-based sampling methodology using the latest Delivery Sequence File of the USPS – a database with full coverage of all delivery points in the US. Households invited to join the panel are randomly selected from all available households in the US. People in the sampled households are invited to join and participate in the panel. Those selected who do not already have internet access are provided a tablet and internet connection at no cost to the panel member. Those who join the panel and who are selected to participate in a survey are sent a unique password-protected log-in used to complete surveys online. Panelists receive a unique login to the survey and are only able to complete it one time. No reminder emails were sent for this study.

The results from this dataset are used for our model. Among the range of variables available, we selected age, income, education, state and sex as our independent variables.

- Age is the age group the panelist falls into. The poll had the ages of the panelists at the time of the poll which we then classified into one of the four age groups: 18-29, 30-44, 45-59, 60+
- Income is the annual household income reported by the panelist. They were given the option of seven categories: Less than $10,000, $10,000 to $24,999, $25,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $149,999 and $150,000 or more.
- Education is the highest level of education attained by the panelist. They were given the option between: Less than high school, High school, Some college and Bachelor's degree or higher. We then classified these into three categories: Less than high school, High school, More than high school.
- State is the state of residence of the panelist.
- Sex is the sex reported by the panelist and it is either male or female.

For the variable we are predicting, we created a binary variable which is 1 if the panelist favors Biden and 0 if the panelist favors Trump. There was no question in the poll which asked this direct question so we used "Overall, do you have a favorable or unfavorable impression of...Donald Trump?". We acknowledge that not favoring Donald Trump does not necessarily imply favoring Joe Biden, but since they are the two major potential candidates, we classify the panelists who chose "Unfavorable" as supporting Biden, i.e. not supporting Trump.

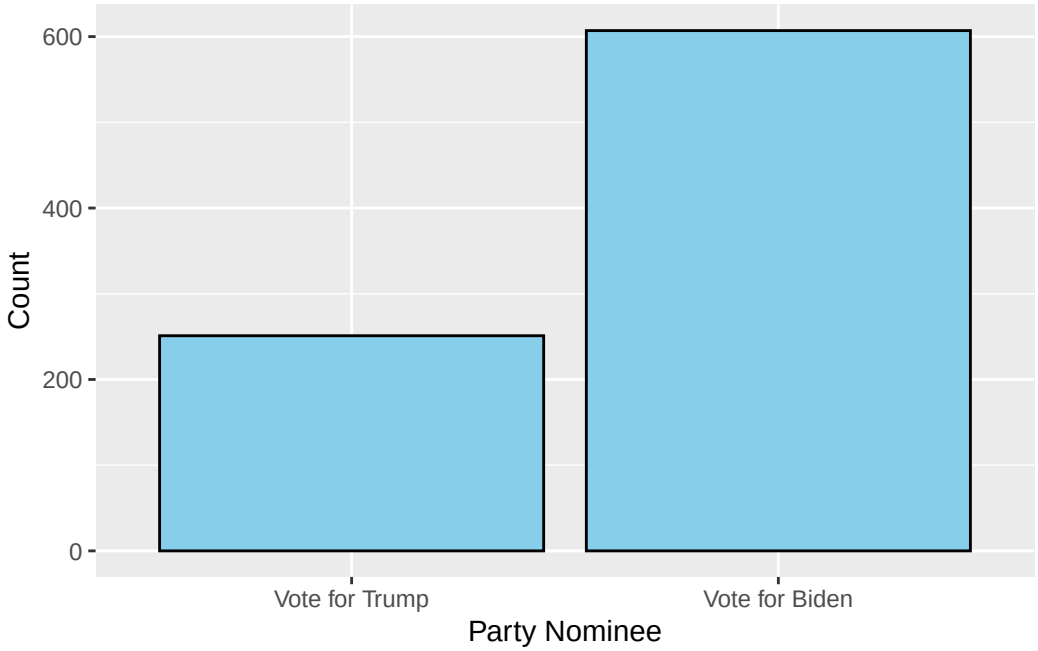
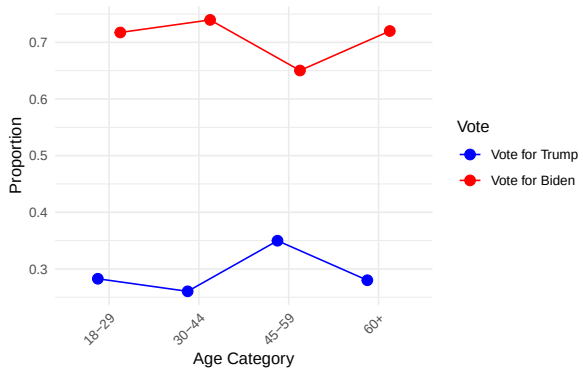


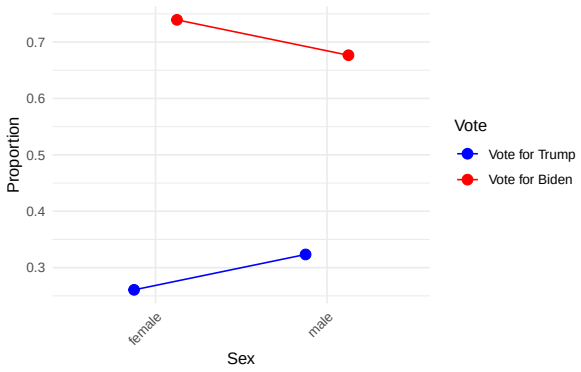
Figure 1: Vote count distribution in survey data

Figure 1 shows the distribution of vote count in our survey data. We see that the survey panelists are heavily in favor of Joe Biden with over twice the panelists not favoring Trump. The bias can be a potential weakness of the dataset and we aim to uncover the factors influencing these political preferences.

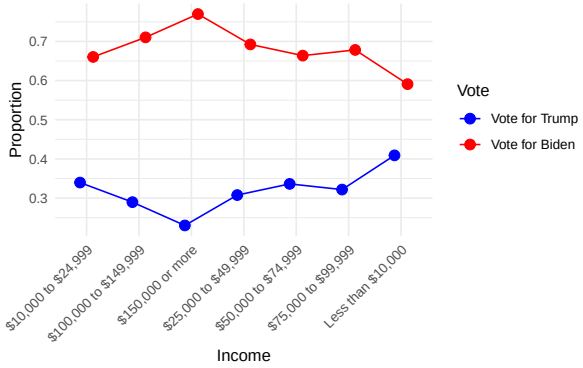
Figure 2 shows the vote proportion distribution for our chosen variables of interests. We notice a noticeable difference in the proportions that favor Trump or Biden. Some striking trends we notice in our survey data are more male panelists supporting Trump, majority of Trump supporters being in the 45-59 age group and Trump support declining with increase in education levels. While most states seem to support Biden, we see that some states support Trump over Biden. This is a fresh change to see over all other factors showing a clear preference for Biden in all categories.



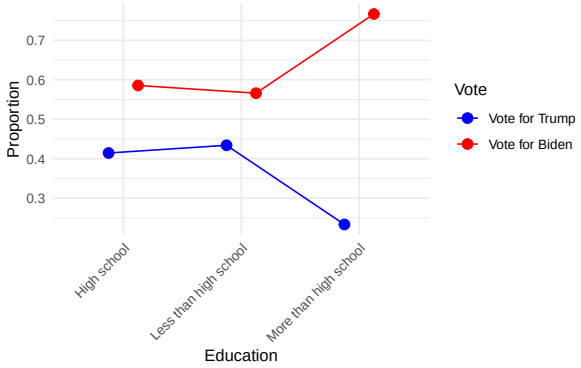
(a) Age



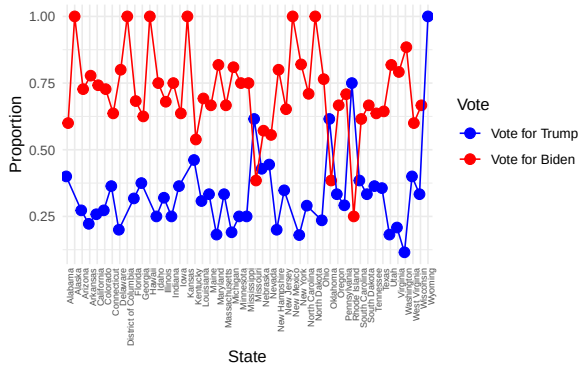
(b) Sex



(c) Income



(d) Education



(e) State

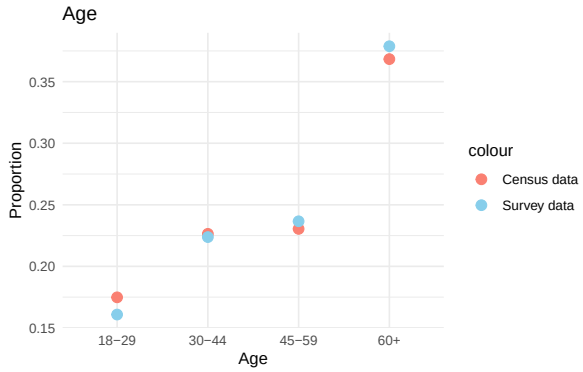
Figure 2: Variables of interest and the vote proportion distribution

2.2 Post-stratification Data

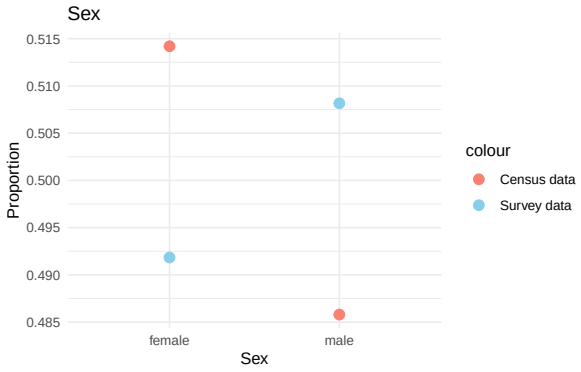
Our post-stratification data is from the most recent American Community Survey (ACS), the 2022 Sample. We use this to represent the voter population and adjust the weight of the survey data. The population for this sample was the entire US population, categorized into households as subgroups, and the responses of the sampled subgroups were collected. ACS 2022 dataset consists of 3,373,378 observations. Using a 1-in-100 national random sample, the ACS 2022 sample provides a thorough overview of the demographic, socioeconomic, and geographic features of the US population. Most importantly, this dataset includes people who live in both home and group quarters, offering a comprehensive view of the American population. A person becomes an entry in this dataset if they are a part of the sampled subgroups and have responded to the survey.

There are a large number of different features available and the variables we chose are age, income, education, state and sex to match with the variables from our survey. These variables were cleaned and the dataset was preprocessed so the values and categories match with the survey dataset.

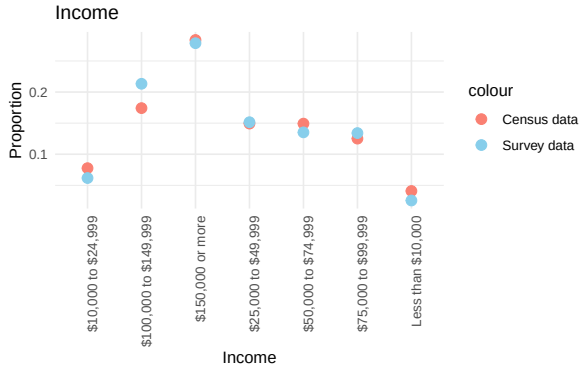
Figure 3 show a comparison of the survey and census data proportions. We notice that for all the variables of interest, the survey is not entirely representative of the population. For example, we see that females are underrepresented in the survey with the survey having about 49% females whereas the census shows 51% females. We also see how population with highest education attained being high school are under-represented whereas population with highest education attained being more than high school are over-represented. Difference in proportions like these can be seen in all the variables of interests and this is why post-stratifying is important to ensure precision and better representation of the population.



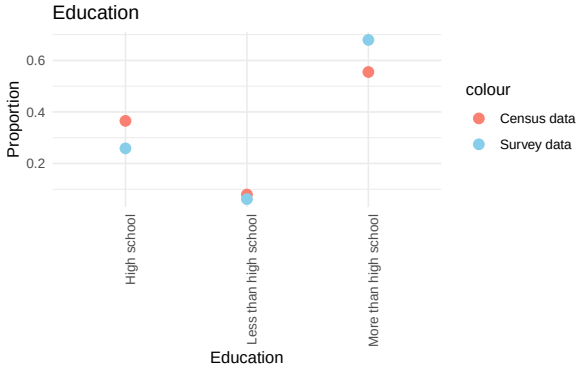
(a) Age



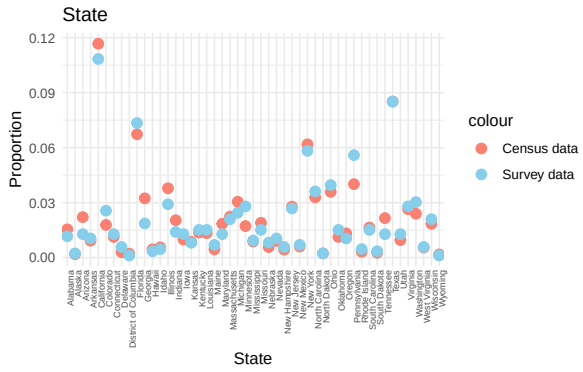
(b) Sex



(c) Income



(d) Education



(e) State

Figure 3: Comparison of Survey and Census Data Proportions

3 Model

We used a Bayesian Logistic Regression model with Post-Stratification to predict the 2024 US Presidential Elections. Logistic regression is a method used for binary classification to predict the probability of a categorical dependent variable.

For our analysis, a logistic regression model will be first used to model the proportion of the voters who will vote for Biden. The model will be based on five independent demographic variables: state, sex, age, education and income.

The logistic regression model we will be using is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 \times \text{state} + \beta_2 \times \text{sex} + \beta_3 \times \text{age} + \beta_4 \times \text{education} + \beta_5 \times \text{income} \quad (1)$$

$$\begin{aligned}
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5)
\end{aligned}$$

where,

- $\hat{p}$ represents the probability that someone will vote for Biden
- β_0 represents the intercept term of this logistical regression. It is the probability that someone will vote for Biden if the predictors' values are zero
- β_1 is the coefficient corresponding to state
- β_2 is the coefficient corresponding to sex
- β_3 is the coefficients corresponding to age
- β_4 is the coefficients corresponding to education
- β_5 is the coefficients corresponding to income

In our model, normal priors with a mean of 0 and a standard deviation of 2.5 are used for both the coefficients and the intercept. Setting the mean of the priors to 0 implies that there is no expectation of a particular direction or magnitude for the coefficients or intercept. We chose this as we have no expectation of the same. The standard deviation of 2.5 reflects the uncertainty or variability in the prior beliefs. We chose a moderately wide prior to allow for a reasonable amount of uncertainty.

The chosen priors allow the data to largely determine the posterior distribution as they are relatively non-informative. They don't heavily influence the results unless the data provide strong evidence to the contrary.

The use of moderately wide priors can also help regularize the model, preventing overfitting and providing more stable estimates, particularly when dealing with limited data.

Table 1: Coefficients of the model

term	estimate	std.error	conf.low	conf.high
(Intercept)	0.66	0.84	-0.65	2.20
sexmale	-0.39	0.17	-0.67	-0.11
age30-44	-0.14	0.27	-0.60	0.32
age45-59	-0.59	0.28	-1.04	-0.14
age60+	-0.31	0.26	-0.73	0.10
stateAlaska	34.46	29.95	3.36	99.13
stateArizona	0.18	1.02	-1.60	1.91
stateArkansas	0.68	1.12	-1.22	2.68
stateCalifornia	0.55	0.75	-0.88	1.73
stateColorado	0.33	0.89	-1.19	1.77
stateConnecticut	0.09	1.03	-1.64	1.77
stateDelaware	1.00	1.47	-1.27	3.74
stateDistrict of Columbia	47.85	42.67	3.33	144.28
stateFlorida	0.08	0.76	-1.38	1.30
stateGeorgia	-0.14	0.93	-1.73	1.33
stateHawaii	27.16	23.52	2.95	80.54
stateIdaho	0.73	1.53	-1.73	3.62
stateIllinois	0.17	0.85	-1.34	1.55
stateIndiana	0.40	0.99	-1.32	2.04
stateIowa	-0.04	0.99	-1.81	1.57
stateKansas	19.60	15.49	3.50	55.26
stateKentucky	-0.19	0.95	-1.87	1.30
stateLouisiana	0.22	0.95	-1.37	1.85
stateMaine	-0.06	1.18	-1.99	2.01
stateMaryland	0.83	1.11	-1.10	2.74
stateMassachusetts	-0.06	0.91	-1.65	1.39
stateMichigan	0.82	0.95	-0.86	2.36
stateMinnesota	0.52	0.88	-1.02	1.95
stateMississippi	0.95	1.16	-1.05	2.97
stateMissouri	-1.25	0.98	-2.98	0.28
stateNebraska	-0.08	1.13	-2.03	1.79
stateNevada	-0.58	1.05	-2.35	1.09
stateNew Hampshire	0.81	1.50	-1.51	3.79
stateNew Jersey	0.06	0.85	-1.48	1.39
stateNew Mexico	19.80	16.11	2.96	57.45
stateNew York	0.82	0.82	-0.64	2.15
stateNorth Carolina	0.26	0.84	-1.27	1.58
stateNorth Dakota	34.97	29.62	3.47	99.30
stateOhio	0.67	0.83	-0.86	2.02

term	estimate	std.error	conf.low	conf.high
stateOklahoma	-1.14	0.98	-2.80	0.40
stateOregon	-0.19	1.02	-1.99	1.55
statePennsylvania	0.35	0.77	-1.09	1.61
stateRhode Island	-1.96	1.61	-4.95	0.43
stateSouth Carolina	0.13	0.94	-1.52	1.71
stateSouth Dakota	0.54	1.63	-2.15	3.64
stateTennessee	-0.06	1.01	-1.77	1.67
stateTexas	-0.01	0.76	-1.45	1.18
stateUtah	0.93	1.12	-1.01	2.89
stateVirginia	0.70	0.90	-0.89	2.17
stateWashington	1.53	0.96	-0.12	3.13
stateWest Virginia	-0.37	1.25	-2.40	1.75
stateWisconsin	0.23	0.88	-1.33	1.73
stateWyoming	-51.51	45.23	-147.52	-5.80
educationLess than high school	0.04	0.33	-0.54	0.61
educationMore than high school	0.98	0.19	0.66	1.29
income$100,000 to $149,999	-0.19	0.38	-0.84	0.42
income$150,000 or more	0.03	0.38	-0.62	0.67
income$25,000 to $49,999	-0.05	0.39	-0.74	0.59
income$50,000 to $74,999	-0.32	0.40	-1.00	0.32
income$75,000 to $99,999	-0.31	0.40	-1.00	0.34
incomeLess than $10,000	-0.64	0.63	-1.66	0.38

Table 1 shows the coefficients for our model that can be used in the equation along with the standard error and the 95% confidence interval. The standard error (SE) is a measure of the precision with which a sample statistic estimates a population parameter. It quantifies the variability of sample statistics around the population parameter. A 95% confidence interval means that if you were to take many samples and construct confidence intervals in the same way, approximately 95% of those intervals would contain the true population mean.

After we created the model, we then post-stratify these proportions and to conclude who we predict to win the 2024 US Presidential Elections. Post-stratification is a statistical technique employed to refine survey estimates by aligning them with known population characteristics. It helps us ensure that our findings accurately represent the diversity of the entire population.

In this process, we first the organize the survey data based on the demographic factors we decided to investigate. This involved counting individuals in each group to understand the sample distribution and computing proportions to depict the relative contribution of each demographic group to the overall survey sample.

Predictive model is then employed to generate probabilities for each demographic group, capturing the nuanced political preferences within distinct strata. These probabilities were sum-

marized by aggregating them within each demographic group.

Applying this post-stratification process results in enhanced predictions that align with the known population distribution within each stratum. These adjusted values contribute to a more accurate representation of party votes within specific demographic categories, ensuring our findings are reflective of the broader population.

4 Results

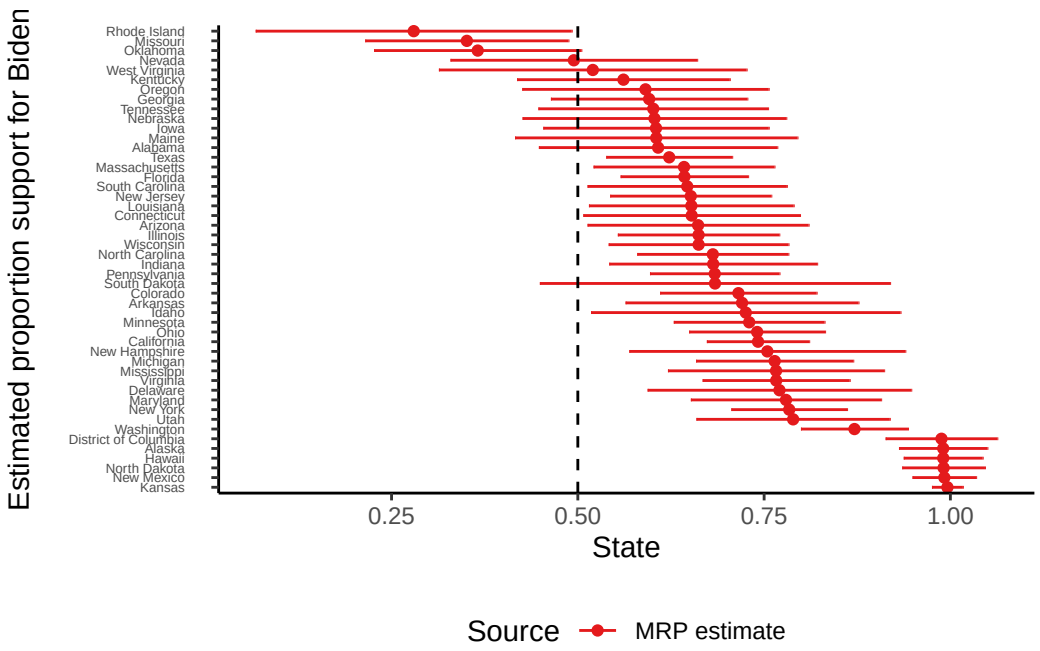


Figure 4: Estimated proportion support for Biden by State

In Figure 4 we see the percentage of the votes for Biden/Democratic party. This graph uses the data we collected after the post-stratification. Here, the red dots on the graph represent the voting percentage and the red lines are the error bars. The black vertical line with dashes indicates the line of proportion. Using that, we see that majority of the states of United States are in support for Biden and only 4 states are in support for the opposing candidate, Trump.

The states that have points to the left of the black vertical line are the states voting for Trump, and the states on the right are voting for Biden. We see that the majority of the votes for the democratic party are found from the Midwest and the southern states while only Rhode Island, Oklahoma, Missouri and Nevada are inclined to vote for Trump. Rhode Island is almost 75% more likely to vote for Trump, but in competition, most of Alaska, Hawaii, North Dakota, New Mexico and Kansas are inclined to vote for Biden.

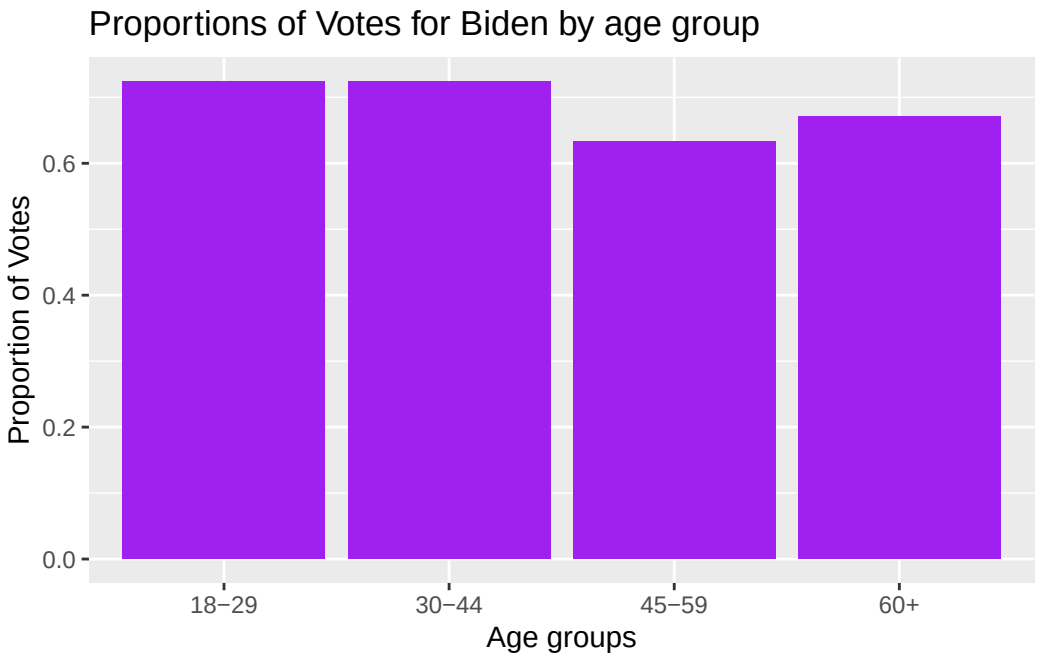


Figure 5: Estimated proportion support for Biden by Age Groups

In Figure 5, we see the proportion of the voters who would vote for the Democratic party categorized by age. We see that the younger generation are more likely to vote for Biden compared to the older. Americans under the age of 45 are 70% likely to vote for Trump. The support for Trump drops to around 60% in the 45-59 age group. While this is favorable towards Biden, it does indicate that more people between 45 to 59 years are likely to vote for Trump compared to other age groups.

Figure 6 shows the proportion of female voters that would vote for the Democratic Party as compared to the male voters. We see a 10% difference in the proportion where women are 10% more likely to vote for Biden.

Figure 7 looks at the division of voters for each party as an overall. We see that even after applying post-stratification, the preference has only changed by a negligible amount with a large majority of the population likely to vote for Biden. This graph clearly shows the conclusion we can deduce from our research, the population of the United States is more inclined towards the Democratic Party.

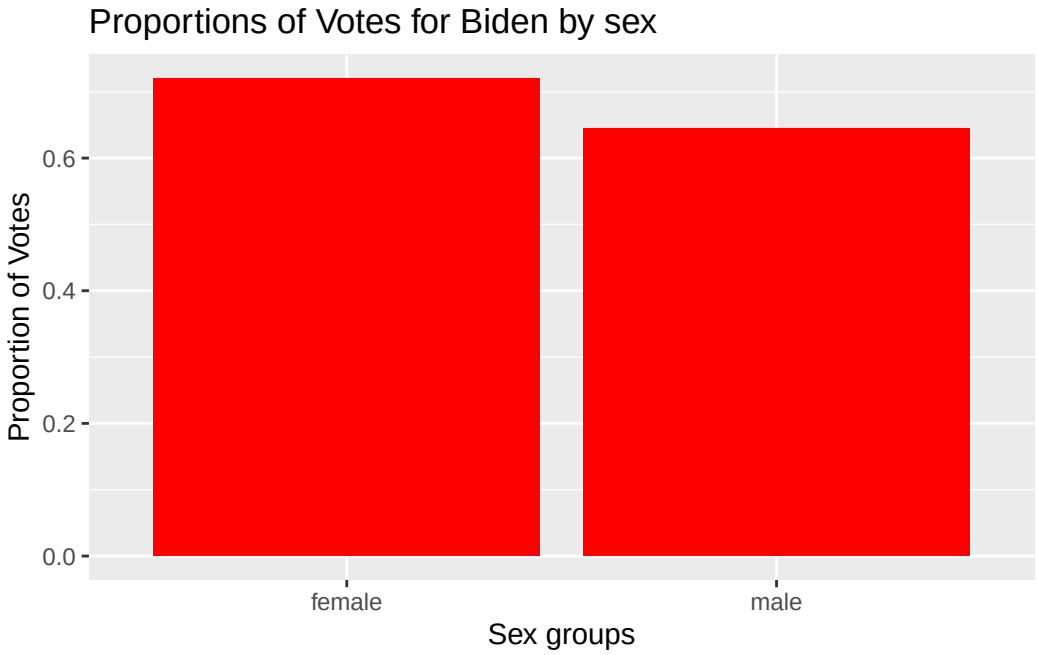


Figure 6: Estimated proportion support for Biden by Sex

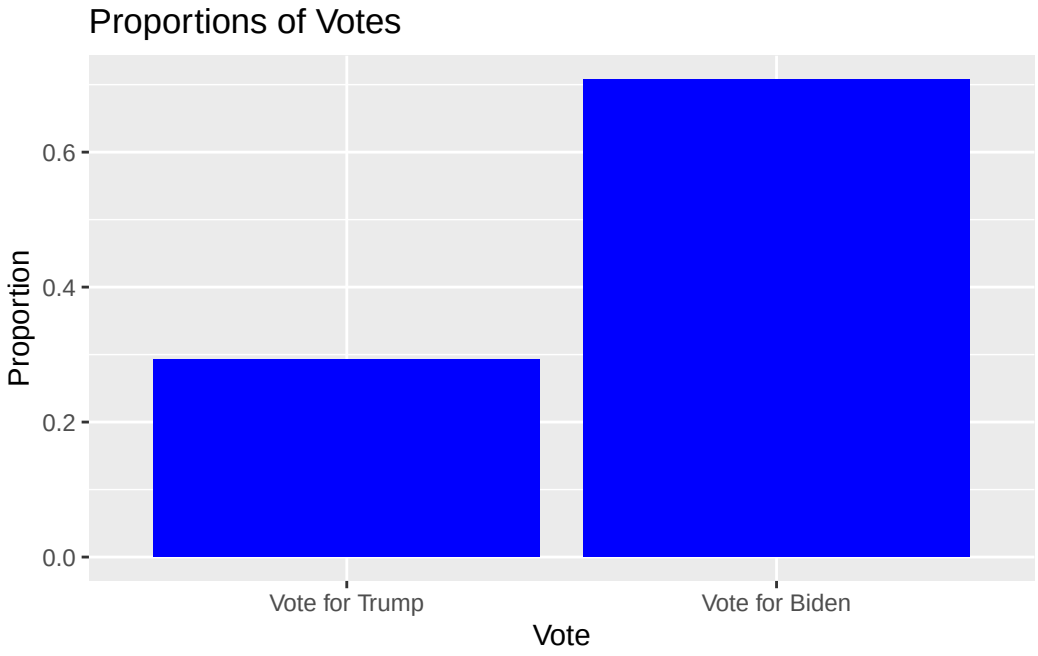


Figure 7: Estimated proportion support for Biden

5 Discussion

Our analysis modeled the variables of interest from the survey data, followed by post-stratification focusing on key variables. We plotted the trends for the post-stratified voting preferences by age, state, and sex. These visualizations yielded valuable insights into voter behavior. Notably, our findings indicated an inclination within the population towards voting for Biden from the Democratic Party.

5.1 Political preference trends in different states

In the graph, Figure 4, we notice that all states in the U.S. other than Rhode Island, Missouri, Oklahoma and Nevada are more inclined to vote for the Democratic Party. We see that as we go towards the Midwest and southern regions, the support for Biden increases as more than 50% of the state is likely to vote for them. This is similar to the findings in (James E. Campbell et al. 2016) that address the same prediction for 2024. Our model and the article predict that states that support Trump are from the middle and Midwest parts of the US, whereas coastal states and the Northeast tend to vote for Biden.

Historically, rural, conservative-leaning towns with strong cultural ties to traditional values and industries like manufacturing and agriculture have defined the middle and Midwest regions of the United States. These areas frequently support gun rights, minimal government intrusion, and religious liberty—issues that are strongly aligned with the agenda of the Republican Party, which Trump advocated. In contrast, the Northeast and coastline states, which are renowned for their progressive views on culture, urban facilities, and diverse populations, have historically tended toward more liberal policies and social ideals that are more in line with the Democratic Party platform, which Biden advocated.

5.2 Political preference trends in different age groups and sex categories

Figure 5 showcases different age groups and what their voting preferences are. We noticed that the youngest group of voters (age 18-29) and mid-age adults (30-44) are more than 70% inclined to vote for Biden. The emphasis on social justice and LGBTQ+ rights among the younger generation may be the cause of this difference. The democratic party led by Biden prioritizes these issues and speaks for the justice of the minorities (Brownstein 2023). Since there are more young people who identify as sexual minorities and they make up 10% of the U.S. population (Meyer 2024), there is a trend for them to support Biden and the Democratic Party.

The middle-aged generation i.e. individuals of age group from 30-44 years old were inclined to voting for Biden as well. Middle-aged voters may align more closely with Biden's policy proposals on issues such as healthcare, economic recovery, climate change, and social justice (Igielnik 2021).

Additionally, we notice almost a 10% decline in the votes for Biden from older age groups (45+ years old). The National security and conservative principles, which are frequently connected to the Republican Party, may have an impact on the elder generation.

Figure 6 shows a slight bias towards the democratic party from the females than the male gender. Predictions suggest that women are 10% more likely to vote for the Democratic Party than men. Studies have consistently demonstrated that there is a gender difference in political ideology, with women generally favoring progressive policies over those of males (Igielnik 2021). This discrepancy is sometimes linked to the fact that women prioritize some issues more than men do, with women giving social welfare, healthcare, education, and reproductive rights—all of which are more in line with the Democratic Party program (Igielnik 2021). With its focus on women's rights, healthcare reform, and inclusivity, Biden's campaign seems to have connected more deeply with female voters, resulting in higher levels of support.

5.3 General conclusion

Figure 7 shows that in general the population of the United States is a lot more inclined to vote for Biden. We see a drastic difference in opinion of almost 40% i.e. people are 40% more likely to vote for Biden than vote for Trump. This conclusion highlighted the final prediction from our analysis. Furthermore, it sheds light on the prevailing social values and the direction individuals envision for their society. Some of the key issues that Americans wanted to see change were the healthcare reforms, climate change actions and racial justice and social equity. They wanted better national response to the COVID-19 pandemic, more ambitious climate policies and investments in renewable energy and lastly, a demand for reforms addressing systemic racism and criminal justice (Brownstein 2023).

5.4 Contributions to Knowledge

By focusing our analysis on specific demographic variables, we contributed to a deeper understanding of the underlying determinants shaping electoral outcomes. Our study corroborates existing literature suggesting that demographic factors such as age, income, education, and geographic location play pivotal roles in influencing voter preferences (Gelman 2023).

The utilization of post-stratification techniques in our analysis allowed for a more granular examination of voting behavior, offering insights into the heterogeneity of voter sentiment across different demographic groups.

Lastly, it highlights the high support of the Democratic Party led by Biden from the younger generations, females, and people of the Midwest and southern states.

5.5 Weaknesses and Limitations

While our analysis provides valuable insights into voter behavior, it is important to acknowledge several limitations.

Firstly, our reliance on survey data from the Ipsos 2023 Poll introduces a sampling bias, as the sample does not fully represent the entire US population.

On analyzing the counts in the dataset, we see that some of the states have a count of 1 or 2 records in the dataset. For example, the District of Columbia and Wyoming, both have one record each in the dataset. While these states are not very populated so it is a given that the representation will be lesser, having only one representative leads to a very biased preference from that state. The political preference of the state will be considered as being what that one person's preference is.

If we consider Alaska which has 2 votes, both the representatives from there are supportive of Biden leading to the biased conclusion that Alaska is highly supportive of the Democratic Party. This is a contradiction to the fact that Alaska is considered a state which is highly of the Republican Party (Wikipedia 2024a).

Additionally, the survey data does not contain any participants from the state of Vermont and Montana. While these states are not very populated (Wikipedia 2024b), their opinions are still important to accurately analyze the preference splits throughout the country.

Consequently, these issues could lead to inaccuracies in our predictions and limit the generalizability of our findings to the broader population.

For our poststratification dataset from ACS, there was no dataset available from 2023. Our poststratification data relied on the survey that was conducted in 2022, and Americans' viewpoints on their choices for political parties may have changed in a year.

Additionally, the education variable from ACS had a vague description and only provided a certain level of education. The levels that were provided in the dataset were ranging from Grade 5 to 5+ years of college. This led to our education levels to be not very diverse but restricted to three major categories: Less than High School, High School and More than High School.

Lastly, while we focused on key demographic variables, other influential variables such as political ideology and campaign messaging were not fully accounted for in our model. Political parties use various data from the population such as media trends and ongoing cultural differences in the society. They also use micro-targeted political ads to voters on social media and online platforms (Funakoshi 2020). This means that our research has the inaccuracy of consideration of incomplete variable sets.

5.6 Future Directions

Moving forward, further research is needed to address the limitations identified in our study and data to advance our understanding of electoral dynamics. Future studies could explore the integration of additional data sources, such as social media sentiment analysis and polling data, to enhance the accuracy of election predictions.

Moreover, there are new strategies opted by political parties in every election that are facilitated by new technologies. Exploring the role of media exposure, including news sources, social media platforms, and political messaging, in shaping citizens' attitudes, expectations, and priorities is something that should be considered in shaping the opinions of Americans.

6 Appendix

6.1 Cleaning

For the survey dataset, the cleaning steps we took were:

1. Initial cleaning involved filtering the dataset to retain only responses relevant to the analysis. Responses were filtered based on the participant's stance on Q1_1 which is "Overall, do you have a favorable or unfavorable impression of…Donald Trump?", specifically selecting those labeled as "Favorable" or "Unfavorable".
2. Age groups were defined based on the participants' reported age (ppage). Participants were categorized into four age groups: "18-29", "30-44", "45-59", and "60+". The final was named age.
3. Respondents' income information (ppinc7) was retained in its original format to provide insights into income distribution. The final column was named income.
4. Education levels were standardized to facilitate analysis. Respondents' education information (ppeducat) was categorized into three groups: "High school", "More than high school" (including "Bachelor's degree or higher" and "Some college"), and "Less than high school". This was done to match the education level separation available in the poststratification dataset. The final column was named education.
5. The state of residence for each participant (ppstaten) was preserved for regional analysis. The final column was named state.
6. Gender information (ppgender) was just had a change of turning the value to lower case. The final column was named sex as just having these two criterias implies that and to maintain with the poststratification data reporting sex as well.

7. We created a binary variable for favorability of Biden. Responses from Q1_1 labeled as "Favorable" were coded as 0, while responses labeled as "Unfavorable" were coded as 1. The final column was named vote_biden.

8. The final survey dataset contained the above mentioned 6 columns: age, education, income, state, sex, and the coded variable for Biden vote.

For the poststratification dataset, the cleaning steps we took were:

1. Initially, the "AGE" variable was converted to numeric format. Subsequently, data points were filtered to retain only those individuals aged 18 and above, excluding individuals from Vermont and Montana and cases where total household income (FTOTINC) was recorded as less than zero. Individuals from Vermont and Montana were excluded from the dataset to maintain consistency with the survey data as the survey data does not involve any records from these two states.

2. Participants' ages (AGE) were categorized into four groups: "18-29", "30-44", "45-59", and "60+". The final column was named age.

3. Education levels were standardized into four categories: "High school", "More than high school", and "Less than high school". This categorization was based on the reported education level (EDUC), with corresponding numerical codes for each category. The final column was named education.

4. States were labeled according to their corresponding numerical codes (STATEICP), assigning each code to its respective state name. The final column was named education.

5. Sex information (SEX) was changed for the codes to their respective values: "male" and "female". The final column was named sex.

6. Household income (FTOTINC) was categorized into the same division of income brackets as the survey data: Less than $10,000, $10,000 to $24,999, $25,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $149,999 and $150,000 or more. The final column was named income.

7. The final poststratification dataset contained the above mentioned 5 columns: age, education, income, state and sex.

Now we are ensured that both the datasets have the same variable names and respective categories.

6.2 Survey dataset

Here is a glimpse of the survey data set used

age	education	income	state	sex	vote_biden
60+	High school	$25,000 to $49,999	Wisconsin	male	1
60+	More than high school	$150,000 or more	Florida	male	1
60+	More than high school	$50,000 to $74,999	Maine	male	1
60+	More than high school	$150,000 or more	Massachusetts	female	1
60+	High school	$10,000 to $24,999	Virginia	female	0
60+	More than high school	$100,000 to $149,999	Colorado	male	0

6.3 Poststratification dataset

Here is a glimpse of the poststratification data set used

age	education	state	income	sex
60+	More than high school	Alabama	$150,000 or more	female
45-59	High school	Alabama	$150,000 or more	male
30-44	Less than high school	Alabama	$150,000 or more	female
60+	Less than high school	Alabama	$150,000 or more	male
45-59	More than high school	Alabama	$150,000 or more	male
30-44	High school	Alabama	$150,000 or more	male

References

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.

Barnes, Melody. 2021. "Election 2020 and Its Aftermath." *Miller Center*, January. https://millercenter.org/election-2020-and-its-aftermath.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Brownstein, Ronald. 2023. "Why Older Voters Have Stuck with Biden More Than Younger Generations ..." *CNN News*, December. https://www.cnn.com/2023/12/19/politics/biden-older-voters-2024/index.html.

Funakoshi, Minami. 2020. "How Political Campaigns Use Your Data." *Reuters*, October. https://www.reuters.com/graphics/USA-ELECTION/DATA-VISUAL/yxmvjjgojvr/.

Gelman, Andrew. 2023. "How the Economist Presidential Forecast Works." *The Economist*. https://projects.economist.com/us-2020-forecast/president/how-this-works.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Igielnik, Ruth. 2021. "Behind Biden's 2020 Victory." *Pew Research Center - U.S. Politics & Policy*, June. https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory/.

James E. Campbell, Thomas E. Mann, David Rothschild Justin Wolfers, Thomas E. Mann, Nicol Turner Lee Norman Eisen, Steven Overly Darrell M. West, and William A. Galston. 2016. "What Do the Models Say about Who Will Win in November?" *Brookings*, September. https://www.brookings.edu/articles/what-do-the-models-say-about-who-will-win-in-november/.

Meyer, Ilan H. 2024. "How Do You Measure the LGBT Population in the u.s.?" *Gallup.com*, February. https://news.gallup.com/opinion/methodology/259457/measure-lgbt-population.aspx.

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

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. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wikipedia. 2024b. *Wikipedia*. Wikimedia Foundation. https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population.

———. 2024a. *Wikipedia*. Wikimedia Foundation. https://en.wikipedia.org/wiki/Red_states_and_blue_states.

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. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.