A Poll-of-Polls Forecast for the 2024 U.S. Presidential Election[*]

Harris Predicted to Win the Popular Vote With 50.9% Over Trump

Boxuan Yi

October 25, 2024

Abstract

The 2024 U.S. Presidential Election on November 5th will feature a contest between Democratic Vice President Kamala Harris and former Republican President Donald Trump. In this paper, I employ a poll-of-polls method and a Bayesian generalized linear model to predict the popular votes for both candidates. By filtering polls that ended within the last 60 days and were conducted by reliable pollsters, the results indicate that 50.9% of voters will support Harris, with her base concentrated in the West Coast and Northeast regions. The forecast of election results is important as it helps shape the campaign strategies of the two parties and understand U.S. citizens' priority issues.

Table of contents

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

1 Introduction

The upcoming 2024 United States presidential election, scheduled for Tuesday, November 5, will see U.S. citizens electing the country's president and vice president for the next four years. The current vice president, Kamala Harris, is the candidate from the Democratic Party, while Donald Trump, the former president from the Republican Party, is running for re-election for a nonconsecutive term. The United States elects its president through the Electoral College. The number of electoral votes assigned to each state depends on its representation in the Senate and House of Representatives, meaning it is possible to win the popular vote yet lose the election. Despite the significance of electoral votes and the presence of multiple candidates, this paper will focus on the popular votes of Kamala Harris and Donald Trump due to the dominance of the Democratic and Republican parties in the current U.S. political system. The electoral votes will be calculated only for states with data due to the limited information available for each state.

In this paper, the estimand being explored is the popular vote for Kamala Harris and Donald Trump. I will use a Bayesian generalized linear model to fill this gap, with the following predictors: pollster, numeric grade of the pollster, recency of the poll, and the population of the respondent groups. Using the poll-of-polls method, which aggregates several polls to forecast results aiming for greater accuracy, the predicted national popular vote for Harris only considering the two candidates is 50.9%, which is 1.8% higher than Trump's support. This difference will be smaller when accounting for all candidates in the 2024 presidential election. Based on state polls, 15 out of 30 states show higher support for Harris, 12 states show higher support for Trump, while 3 states have 50% support for each. It is also found that Harris's supporters are concentrated in the West Coast and Northeast regions, whereas Trump's supporters tend to be in the central part of the U.S. Presidential election forecasts have always been important because they help understand U.S. citizens' priority issues and assist parties in shaping campaign strategies for better decision-making.

The data utilized for analysis is presented in Section 2. Following that, Section 3 introduces the model created to forecast election outcomes. I will then explain the predicted results derived from the model in Section 4. Lastly, Section 5 discusses the results in a broader context and address weaknesses of this paper as well as the future explorations. This paper uses the programming language `R` (R Core Team 2022). The analysis, the model and all the visualizations use the following packages: `dplyr` (Wickham et al. 2023), `tidyverse` (Wickham et al. 2019), `here` (Müller 2020), `janitor` (Firke 2023), `knitr` (Xie 2014), `readr` (Wickham, Hester, and Bryan 2024), `modelsummary` (Arel-Bundock 2022), `ggplot2` (Wickham 2016), `statebins` (Rudis 2020), `arrow` (Richardson et al. 2024), `stringr` (Wickham 2023), `tidyr` (Wickham, Vaughan, and Girlich 2024), `lubridate` (Grolemund and Wickham 2011), `rstanarm` (Goodrich et al. 2024), `bayesplot` (Gabry et al. 2018).

2 Data

2.1 Overview

The data I use is from the 'Presidential General Election Polls (Current Cycle)' by FiveThirtyEight (Ryan Best and Wiederkehr 2024). The analysis and visualizations in this paper are based on poll results updated until October 22th. The dataset includes 3,227 polls from different pollsters asking participants who they support for the upcoming presidential election.

The key variables and their meanings from the dataset that I use are as follows:

- poll_id: Unique identifier for each poll conducted
- pollster: The name of the polling organization that conducted the poll
- numeric_grade: A numeric rating indicating each pollster's reliability
- state: The US state where the poll was conducted, if applicable.
- start_date: The date the poll began
- end_date: The date the poll ended
- sample_size: The total number of respondents participating in the poll
- population: The abbreviation of the respondent group indicating their voting status, such as "likely voters" or "adults."
- candidate_name: The name of the candidate in the poll
- pct: The percentage of support each candidate received in the poll

I create three more variables:

- state_or_national: Indicates whether the poll is a national poll or conducted in a specific state.
- days_since_end: The number of days since the survey ended.
- harris_support_ratio: The percentage of respondents supporting Harris among those who indicate support for either Harris or Trump.

I will only consider polls with a numeric grade of at least 2.0 that ended no more than 60 days ago, as people's voting preferences can change over time. The meaning of the numeric grade and the rationale for choosing 2.0 will be explained in Section 2.2. The cleaning process can be found in Section 6.3.

2.2 Methodology and Measurement

The method used to forecast the presidential election results is the poll-of-polls, which aggregates results from multiple polls instead of relying on a single survey, aiming to make the results more accurate and stable. In this method, each poll is assigned a weight based on factors such as sample size, recency, and the pollster's historical accuracy. Therefore, the larger the weight, the more reliable the polls are in the aggregation.

The dataset from FiveThirtyEight that I use for this forecast includes all publicly available scientific polls that meet methodological and ethical standards, including public partisan and internal campaign polls. This ensures that the polls included in their forecasts and models are based on sound survey methods and that pollsters are genuinely engaged in the pursuit of truth and knowledge (ABC News 2023). The weight provided by FiveThirtyEight, referred to as `numeric_grade` in the dataset, is based on the historical track record and methodological transparency of each polling firm's polls (Silver 2024). Additionally, polls receive reduced weight if the pollster who conducted them has released a large number of surveys in a short period of time. The weight ranges from 0.5 to 3.0, with 3.0 being the most reliable pollster. Out of 282 pollsters, 95 received a grade of no less than 2.0 (Silver 2024). As shown in Figure 1, the median numeric grade of all polls in the dataset falls between 1.5 and 2.0. To ensure the quantity and reliability of the data, I decided to exclude polls with a numeric grade lower than 2.0. Examples of pollsters with a 3.0 grade include The New York Times/Siena College, ABC News/The Washington Post, and YouGov. Examples with a 2.5 grade are the Public Policy Institute of California, Pew Research Center, and the University of Illinois Springfield Survey Research Office (Silver 2024).

Different polls from various pollsters have different methods for conducting surveys that translate real-world phenomena into numerical data. However, the underlying logic is similar. Pollsters typically start by selecting a sample of the population that reflects the broader electorate; this could be a national survey, a state survey, or a target group. Based on the population sample and budget, the pollster will release the survey on a

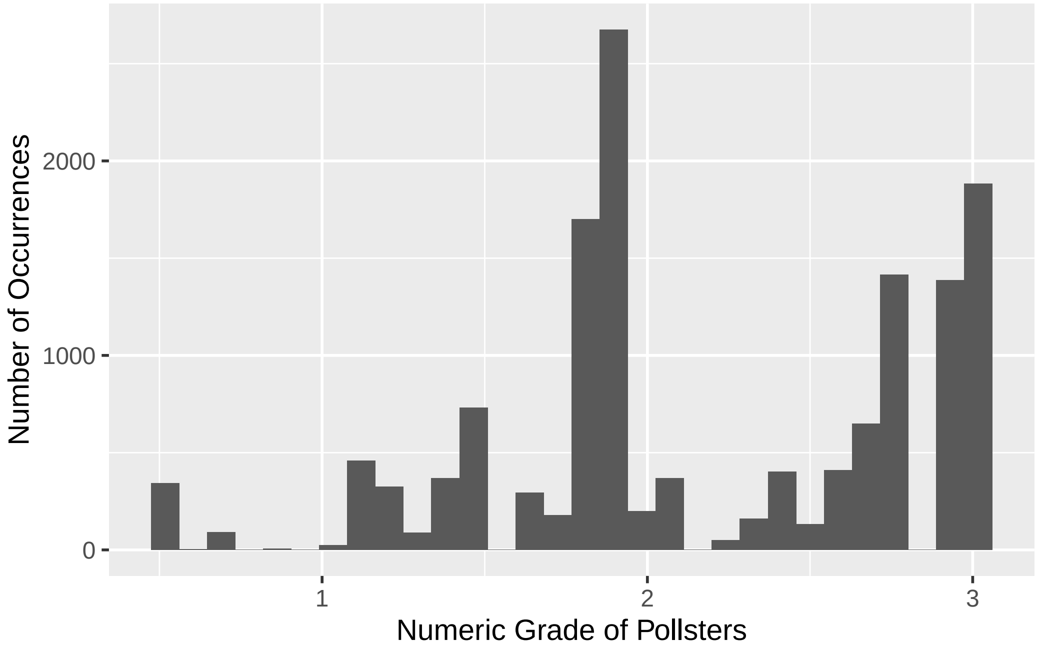


Figure 1: Distribution of Pollsters Numeric Grades Given by FiveThirtyEight Before Cleaning

preferred platforms. For example, they might release the poll on social media if they want more general and diverse participants, or use news media to achieve a higher response rate. Good pollsters design questions that are clear and unbiased, often asking respondents to choose from a list of candidates or political parties. The survey usually includes additional questions, such as the age, gender, and education of the participants. Each response, representing a real-world decision, becomes an entry in a dataset that captures individual voting intentions.

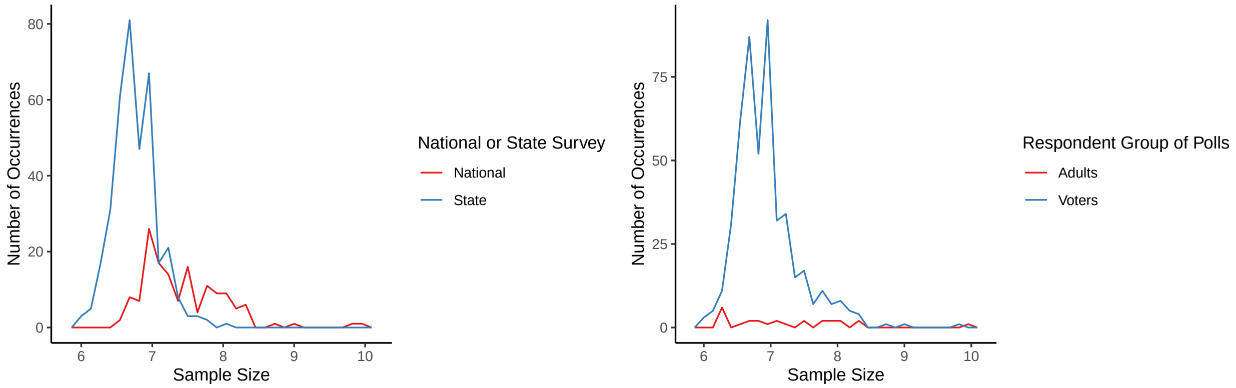
More detailed measurements regarding YouGov can be found in Section 6.1. For the analysis below, the dataset used is cleaned.

2.3 Data Visualization and Analysis

The distribution of sample sizes is presented in logarithmic form in Figure 2 for clearer visualization. As seen in Figure 2 a, there are significantly more state surveys than national surveys overall. Most national surveys have sample sizes under 2,500, while the majority of state surveys fall below 2,000. A few national surveys exceed 7,000, indicated by small red bumps in the distribution. In Figure 2 b, most polls target adults as the respondent group, with only a few focusing on voters.

As shown in Figure 3, state polls exhibit a wider range of support proportions, from below 0.4 to nearly 0.7, reflecting the diverse political landscape across different regions, whereas national polls tend to cluster more closely around a support ratio of 0.5. Interestingly, the mode of support for state polls is slightly below 0.5, while national polls demonstrate a mode and majority support slightly above 0.5.

From Table 1, the average proportion of support for Harris using 145 national surveys is 51.3%, and the median support is slightly higher at 51.4%, indicating that just over half of the respondents favor her over Trump, who has an average support of 48.7%. The support for Harris ranges from a minimum of 0.484 to a maximum of 53.6%. With a standard deviation of 0.011, the support levels exhibit low variability, signifying a consistent response pattern among the polls. The weighted average of Harris's support considering sample size is also 51.3%.



(a) Distribution of Log-Transformed Sample Sizes by Poll Type (State vs. National)

(b) Distribution of Log-Transformed Sample Sizes by Respondent Group

Figure 2: Frequency Distribution of Log-Transformed Sample Sizes

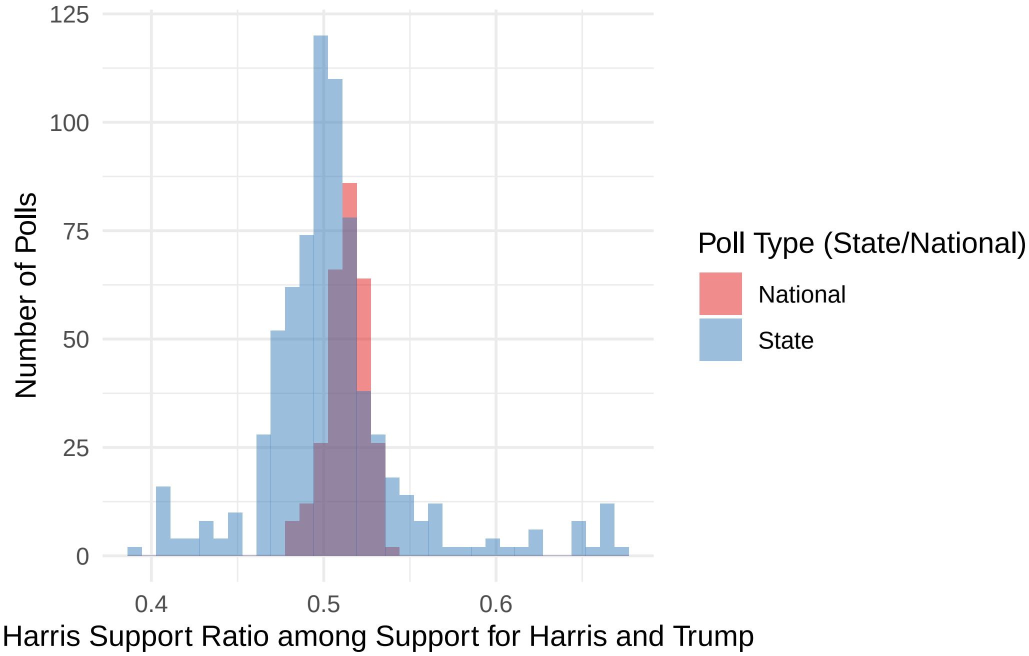


Figure 3: Distribution of Harris Support Ratios in State and National Polls

Table 1: Summary Statistics for the Proportion of Support for Kamala Harris in National Polls

Avg Support for Harris	Median Support	Min Support	Max Support	SD of Support	Total Polls
0.513	0.514	0.484	0.536	0.011	145

Table 2 shows the number of surveys conducted for each state, if applicable, and the weighted average proportion of support for Harris based on these state surveys using sample size. At least 30 polls target Arizona, Georgia, Michigan, North Carolina, Pennsylvania, and Wisconsin. Florida, Nevada, Ohio, Texas, and Virginia each have at least 10 surveys. The remaining states either have a single-digit number of surveys or none at all. Out of the 30 states, 12 show more support for Trump, while 17 have more support for Harris, with North Carolina having exactly equal support for both. In Table 2, California, Maryland, Massachusetts, and Washington all have more than 60% support for Harris, representing a 20% lead.

Table 2: Weighted Average Proportion of Harris Support from State Surveys

State of Survey	Number of Survey	Weighted Average Proportion of Harris Support
Alaska	1	0.448
Arizona	35	0.487
California	4	0.630
Connecticut	1	0.589
Florida	11	0.462
Georgia	36	0.493
Indiana	1	0.413
Iowa	2	0.472
Maine	3	0.552
Maryland	5	0.664
Massachusetts	5	0.653
Michigan	38	0.502
Minnesota	3	0.529
Missouri	2	0.442
Montana	6	0.407
Nebraska	8	0.503
Nevada	21	0.505
New Hampshire	5	0.544
New Mexico	3	0.550
New York	5	0.578
North Carolina	43	0.500
Ohio	10	0.466
Pennsylvania	48	0.506
Rhode Island	3	0.579
South Carolina	1	0.449
Texas	13	0.471
Utah	4	0.416
Virginia	10	0.549
Washington	1	0.620
Wisconsin	39	0.510

Figure 4 uses a gradient colour scale to present a U.S. map of weighted support by state. The bluer a state is, the higher the predicted proportion of support for Harris; the redder a state appears, the higher the predicted proportion of support for Trump. States shown in grey have no state surveys, so no state-level data is available. States on the West Coast and in the Northeast, such as Massachusetts, California, and Maryland, show high support for Harris, while states in the Midwest, including Utah, Montana, and Indiana, show strong support for Trump.

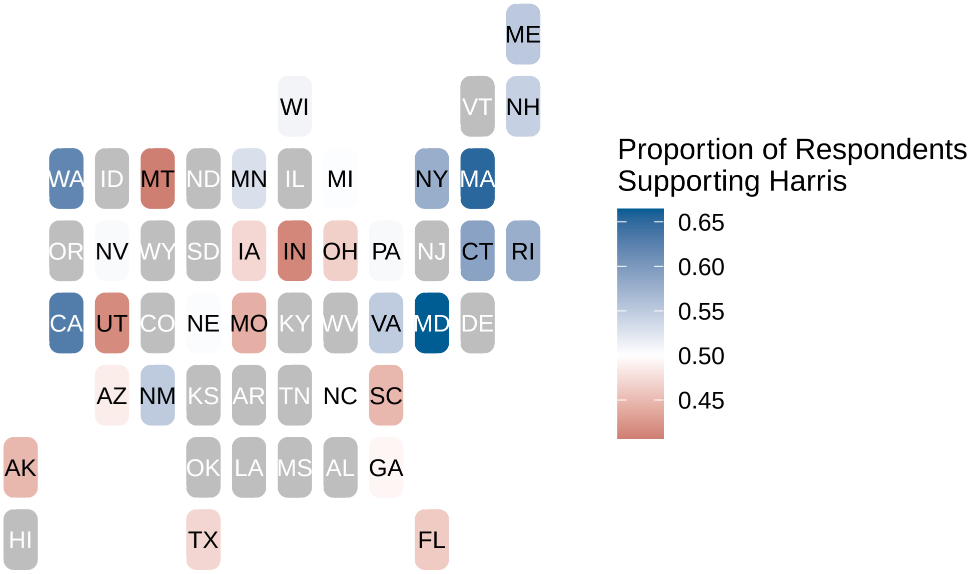


Figure 4: Proportion of Harris Support by State from State Surveys (Weighted)

3 Model

3.1 Model Set-up

To predict the 2024 U.S. presidential election results, I employ a Bayesian regression model with a Normal distribution that incorporates multiple predictors using the programming language `R` (R Core Team 2022) and the package `rstanarm` (Goodrich et al. 2024). In this model, the dependent variable is the proportion of respondents who support Kamala Harris, which is assumed to be continuous when the sample size is sufficiently large. The goal of the model is to estimate how Harris's support is influenced by four factors, which can be expressed as follows:

$$
\begin{aligned}
y_i|\mu_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \times \text{Pollster}_i + \beta_2 \times \text{Numeric Grade}_i + \beta_3 \times \text{Days Since End}_i + \beta_4 \times \text{Population}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5) \\
\beta_4 &\sim \text{Normal}(0, 2.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

where:

- y_i is the dependent variable, representing the proportion of respondents who support Harris.
- β_0 is the intercept term, representing the expected Harris support ratio when all predictors are zero. It follows a prior distribution that is normal with a mean of 0 and a standard deviation of 2.5.
- β_1, β_2, β_3, and β_4 are the coefficients corresponding to the predictor variables **Pollster**, **Numeric Grade**, **Days Since End**, and **Population**, respectively. **Days Since End** ranges from 0 to 1,

showing how recent the survey ended, with 0 meaning the most recent. Each of these coefficients follows a prior distribution that is normal with a mean of 0 and a standard deviation of 2.5.
- The residual standard deviation, σ, follows an exponential prior with a rate of 1.

As mentioned in Section 2.2, the dataset includes pollsters with partisan and internal campaign polls, so I selected Pollster as one of the factors, as different pollsters may attract respondents with varying voting preferences. The numeric grade is also included, as it reflects the historical accuracy and transparency of the pollsters. Since people's voting preferences can change over time and become more stable as the election approaches, recency, represented by the number of days since the survey ended, is selected as a factor. Additionally, I also contain population because exploring the voting preferences of individuals who can and will vote is meaningful.

The model assumes that the priors, particularly the normal priors on the coefficients, are appropriate. One potential limitation is that the model may underperform for highly skewed data or when extreme outliers significantly influence poll results. The model also assumes Harris's support to be continuous, which may not be appropriate if the sample sizes are too small. A logistic distribution could be considered if the dependent variable is whether Harris or Trump will win the election, as it is useful for dealing with binary outcome variables.

3.2 Model Justification

Section 2.3 shows that some states have a high preference for Harris, so I expect to see a positive relationship between Harris's support and pollsters that focus on these states. For instance, both Maryland and Washington have a high rate of support for Harris, so the pollster 'University of Maryland/Washington Post' should have a relatively large coefficient. I anticipate a positive relationship between Trump's support and pollsters that focus on these states as well. Additionally, I expect the intercept to be close to 0.5, with the coefficients for recency and population not far from 0, as there are no definitive relationships between the recency and population of polls and support for different candidates.

A Posterior Predictive Check is included in Section 6.4.

3.3 Model Results

The coefficients and their 95% confidence intervals are presented in Figure 5. The coefficients for `population`, `days_since_end`, and `numeric_grade` are all close to 0. The intercept falls between 0.5 and 0.6, with a relatively larger confidence interval.

The coefficients for different pollsters vary, reflecting that different pollsters tend to attract audiences with varying voting preferences. As mentioned in Section 3.2, I expect to see a positive relationship between Harris's support and pollsters that focus on these states, and the same for Trump's support. This is reflected in Figure 5. Both Maryland and Washington show a high rate of support for Harris, resulting in a large positive coefficient for the pollster University of Maryland/Washington Post. UC Berkeley also has a positive coefficient, as California has high support for Harris. Conversely, Winthrop U has a small negative coefficient because it is located in Southern California, which has a higher level of support for Trump. The largest positive coefficients, aside from the intercept, come from the pollsters PPIC and University of Maryland/Washington Post, while the smallest negative coefficient is associated with the pollster GQR.

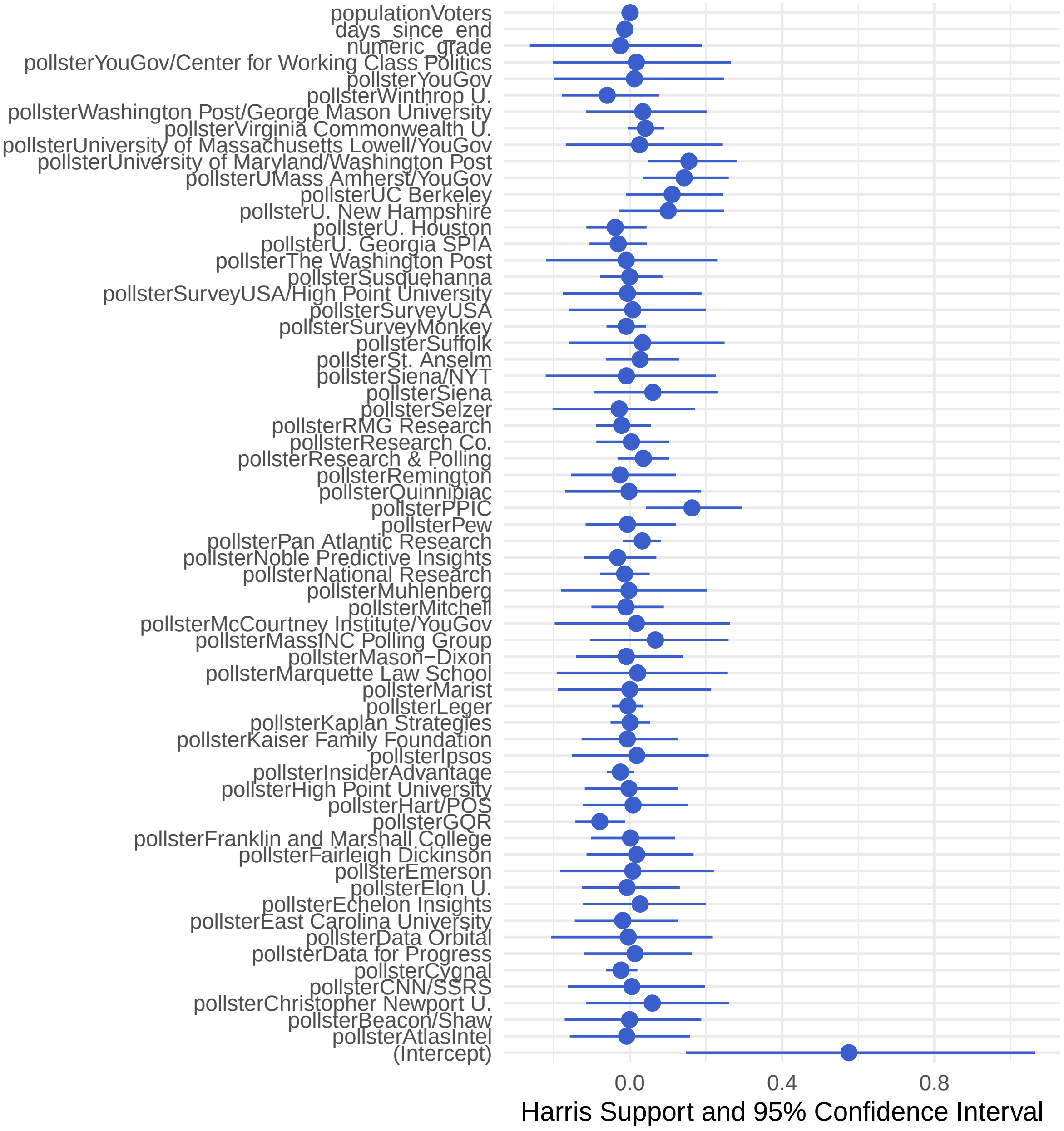


Figure 5: Model Results for Harris Support Based on Pollster, Pollster Quality, Survey Recency, and Survey Population

4 Results

The predicted national support is shown in Table 3. The forecast indicates that 50.9% of voters would choose Harris, while 49.1% would choose Trump if we only consider voters selecting one of them. Harris's support ranges from 47.7% to 58%, with a standard deviation of 0.014, which is slightly larger than the standard deviation in Table 1, indicating greater variation after prediction. The weighted average of Harris's support, considering sample size, is 50.8%, very close to the unweighted 50.9%.

Table 3: Summary Statistics for the Predicted Proportion of Support for Kamala Harris in National Polls

Avg Support for Harris	Median Support	Min Support	Max Support	SD of Support	Total Polls
0.509	0.512	0.477	0.58	0.014	145

The predicted weighted support, using sample size, by state is displayed in Table 4. Compared to Table 2, the proportion of support is overall more concentrated around 50%. Out of the 30 states with data, 15 states show higher support for Harris, 12 states have higher support for Trump, while 3 states have equal support for both candidates. Maryland still has a support rate for Harris that exceeds 60%, and other states with high support for Harris include Rhode Island, New Hampshire, Massachusetts, Maine, California, and Connecticut, all exceeding 55%. South Carolina is the only state with less than 45% support for Harris.

Table 4: Predicted Weighted Proportion of Harris Support Based on State Surveys

State	Predicted Proportion of Support For Harris (Weighted)
Alaska	0.500
Arizona	0.498
California	0.594
Connecticut	0.570
Florida	0.487
Georgia	0.498
Indiana	0.507
Iowa	0.486
Maine	0.552
Maryland	0.636
Massachusetts	0.572
Michigan	0.500
Minnesota	0.507
Missouri	0.480
Montana	0.484
Nebraska	0.500
Nevada	0.497
New Hampshire	0.552
New Mexico	0.540
New York	0.565
North Carolina	0.499
Ohio	0.491
Pennsylvania	0.501
Rhode Island	0.582
South Carolina	0.449
Texas	0.497
Utah	0.477
Virginia	0.536
Washington	0.504

Table 4: Predicted Weighted Proportion of Harris Support Based on State Surveys

State	Predicted Proportion of Support For Harris (Weighted)
Wisconsin	0.506

Figure 6 uses a gradient colour scale to indicate the predicted weighted support rate by state. States on the West Coast and in the Northeast show a higher rate of support for Harris, with Maryland and California being the bluest. In contrast, states in the Midwest have higher support for Trump, with South Carolina appearing the reddest. Compared to Figure 4, the overall colour is lighter, indicating that the model has mitigated certain biases and preferences.

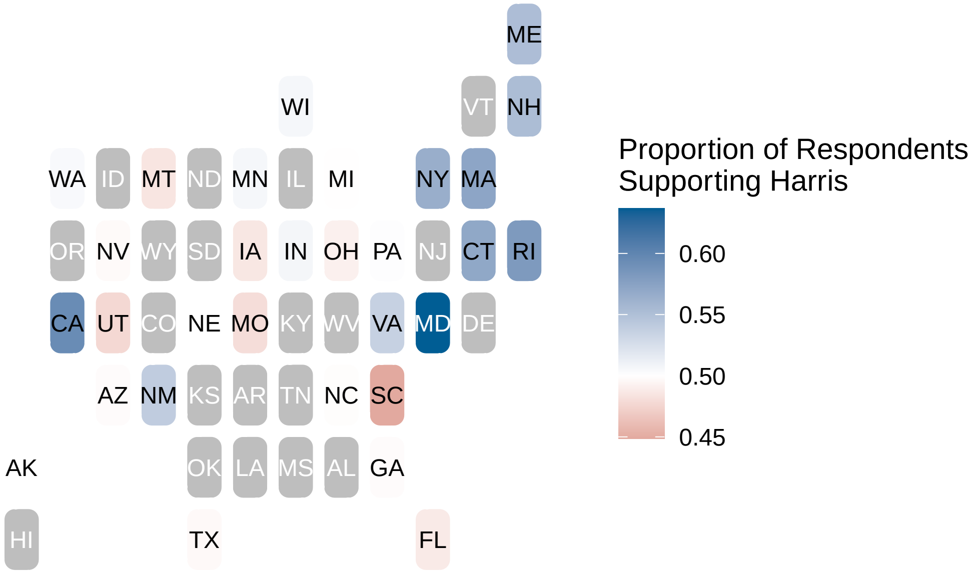


Figure 6: Predicted Proportion of Harris Support by State from State Surveys

Considering the 30 states with data, the total electoral votes for Harris is 202, while Trump has 174 votes. For the other states without data, Illinois, with 19 votes, has historically been a strong Democratic state, whereas Tennessee, with 11 votes, has been a strong Republican state ("State Electoral Vote History: 1900 to Present - 270toWin — 270towin.com"). Additionally, many of the excluded states, such as Wyoming, West Virginia, and Oklahoma, have a strong Republican inclination, meaning that Harris's lead in electoral votes is not guaranteed.

5 Discussions

5.1 Each State has it preference

The Democrats emphasize civil rights, social responsibility, and government-funded healthcare, while the Republicans focus on lower taxes, traditional values, and family and individual freedom (US Embassy & Consulate). The two parties typically maintain distinct yet steadfast stances, attracting a significant number of supporters and creating stable voting patterns in many states.

States on the West Coast and in the Northeast, which contain many large cities and urban areas, are generally wealthier than those in the central part of the country. The Republicans' and Trump's emphasis on economic improvement may explain why states in the Midwest lean Republican, as shown in Figure 6. This trend may also correlate with the rural and suburban demographics, where urban voters tend to support the Democratic Party while rural voters are more likely to favour Republicans (Kim Parker and Igielnik 2018). The Northeast contains many prestigious universities, including Ivy League institutions. From Figure 6, the strong support for Harris in the Northeast could be linked to the higher average education level in this region, as individuals with higher education levels tend to favour the Democratic Party (YouGov 2024a).

In addition to economic and educational factors, social and cultural values also shape voting patterns across states. Progressive states with the higher numbers of protests from January 2023 to August 2024 align with those that support Harris (Statista 2024). This is reasonable, given the Democratic Party's prioritization of social justice and diversity. Conversely, regions with fewer protests tend to support Trump, where stability and tradition are more prevalent.

5.2 Pollsters Attract Different Audiences

From Figure 5, different pollsters may have their own preferences, attracting audiences with similar inclinations and leading to distinct polling results, even when the populations surveyed are the same. Some pollsters are known for their strong partisan alignment with the Democratic Party, while some favour the Republican Party. Therefore, I hope that these effects will counterbalance each other in this analysis, reducing potential bias. This again highlights the advantages of the poll-of-polls method: while individual polls may exhibit bias, a large collection of polls tends to produce more stable and reliable results.

5.3 Polling Methodology on Voter Preferences

Polling methodology can introduce sampling biases, as different data collection methods can lead to distinct distributions of demographic groups, ultimately affecting the representativeness of poll results. For instance, relying solely on online surveys may exclude individuals without internet access, while phone surveys might miss those who are at work and unwilling to participate during working hours.

Figure 7 illustrates the polls conducted in Nebraska, grouped by methodology. Surveys collected through "live phone" or "live phone/text-to-web/email/mail-to-web/mail-to-phone" methods show around 55% support for Harris. In contrast, surveys using 'IVR/Online Panel/Text-to-Web' yield only 43.5% support for her. This significant difference in support rates underscores how data collection methods can inadvertently privilege certain voices while marginalizing others, thereby revealing underlying social inequalities. An accurate representation in polling is crucial for reflecting the diverse views of the electorate.

State	Methodology of Survey	Weighted Average of Harris Support
Nebraska	IVR/Online Panel/Text-to-Web	0.435
Nebraska	Live Phone	0.547
Nebraska	Live Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	0.558

Figure 7: Harris's Support in Nebraska by Polling Methodology (Weighted)

5.4 Limitations and Weaknesses

I eliminated all polls conducted by pollsters with a grade lower than 2.0, as this grade is provided by FiveThirtyEight based on historical accuracy and methodological transparency. One potential weakness of the model is that I also included numeric grade as a predictor, which may duplicate the effects of filtering the raw dataset.

The dataset used for this analysis contains a poll that ended on August 25, which includes Joe Biden as one of the candidates, even though he announced his withdrawal in July. While unreliable polls like this were filtered out, the reliance on the original aggregated dataset may require further verification to enhance the accuracy of the analysis.

In the past six elections, Nebraska has consistently been a strong red state, typically showing about 60% support for Republicans and 40% for Democrats (270 to Win 2024). However, as shown in Figure 7, surveys conducted using the "live phone" or "live phone/text-to-web/email/mail-to-web/mail-to-phone" methodologies indicate approximately 55% support for Harris. Similarly, North Carolina, which has traditionally been considered a safe red state, appears very neutral in both Figure 4 and Figure 6. There are fair number of 40 polls conducted in North Carolina and 8 in Nebraska. This raises questions about whether the polls are biased or if residents are genuinely changing their opinions this year.

5.5 Next Steps

The predictive model generated in this analysis, along with its strengths and limitations, serves as a foundation for ongoing discussions on campaign and policy strategies.

This year's election is particularly unique due to Biden's withdrawal a few months before the election. For future exploration, I recommend that survey companies analyze how support for the two parties shifts in response to Biden's announcement. This could help identify the number of voters who support the Democratic Party regardless of Biden's candidacy versus those who specifically supported him.

Furthermore, since the last election occurred during the COVID-19 pandemic, it would be beneficial for future research to examine the impacts of the pandemic on electoral behaviour. Comparing the results of the previous election with current trends will provide valuable insights into how unforeseen events can alter political preferences among citizens.

6 Appendix

6.1 Methodology of YouGov Surveys

YouGov is an international online research data and analytics technology group, whose goal is to offer unparalleled insight into what the world thinks. YouGov is a leading platform for online market research, drawing insights from a continuously growing dataset of over 27 million registered panel members, which they refer to as "living data." Their innovative approach ensures accurate and actionable consumer insights. Recognized globally for their data accuracy, YouGov is frequently cited by the press and is considered one of the most trusted sources of market research (YouGov).

The following methodology is from the YouGov Methodology website (YouGov 2024c) This pollster has a 3.0 grade according to FiveThirtyEight.

YouGov's surveys are conducted using online polling. The population is all the US citizens who can vote. Respondents are chosen based on a non-probability sampling, in which not everyone in the population has an equal chance of being selected, but the sample is adjusted using statistical weighting to better reflect the target population. To ensure representativeness, YouGov selects respondents who match the demographic characteristics of the population they are studying, including age, gender, race, education, and voting behaviour. The data will be then adjusted so that the survey results align with the actual distribution of these characteristics in the target population, meaning that if a survey has a higher or lower proportion of people from a certain demographic group than is present in the population, the results are weighted to correct for this imbalance.

YouGov employs several verification and quality control steps. For instance, when new members join the panel, YouGov collects demographic information and verifies respondents' email addresses and IP addresses. Additionally, YouGov monitors survey completion time and answer consistency to ensure the data is accurate. Panelists who provide unreliable data are either excluded from the final results or removed from the panel altogether.

To recruit a diverse panel, YouGov draws participants from many sources, including advertising and partnerships with other websites, and offers surveys in multiple languages. Although participation is limited to those with internet access, this still includes more than 95% of Americans. Respondents are also incentivized through a points system that can be exchanged for small rewards. When determining who to invite to participate in surveys, YouGov considers several factors, such as how recently a respondent has completed a survey, whether they prefer frequent participation, and their past response rates. For general population surveys, YouGov typically aims for sample sizes of 1,000 to 2,000 respondents to strike a balance between reliability and efficiency. YouGov uses a multilevel regression with post-stratification model for vote estimation. Margin of error is calculated for each survey to indicate the range within which the true population value is expected to fall. To ensure data security and privacy, YouGov gives respondents control over their personal information. Respondents can request a copy of their data, or ask for corrections or deletions. When findings are reported, the data is aggregated to prevent the identification of individual respondents.

For the 2024 Presidential Election trackers on the YouGov website, the data comes from regular tracking surveys conducted by YouGov. The Question is "In November 2024, who would you vote for in the presidential election if these were the candidates?", with question wording and response options varied over time (YouGov 2024b). Respondents were selected using random sampling, stratified by gender, age, race, education, geographic region, and voter registration from the most recent American Community Survey. The sample was weighted according to gender, age, race, education, 2020 election turnout and presidential vote, baseline party identification, and current voter registration status. One weakness is that the census data used for weighting is 2019 American Community Survey, which was five years ago. The results may be more accurate if they use newest census data.

6.2 Idealized Survey and Its Methodology

If I had a budget of $100K to forecast the U.S. presidential election, I would conduct one survey now and another one after a week, both aiming for a sample size of 5,000 to 8,000 respondents for a 5 to 10 minutes survey with 15 questions. I believe fewer people have the patience to finish a long survey. I would use stratified sampling to ensure that the sample reflects the U.S. population in terms of key demographics such as age, gender, race, education, region, and political affiliation.

I would build the main body of respondents through an online opt-in panel. This online panel would be recruited using a mix of traditional advertising, partnerships with news websites such as The New York Times, and social media outreach to ensure diverse representation. I would allocate more of the budget to social media platforms like Instagram and YouTube because most people use these platforms, and the expected participation rate would be lower on social media compared to news media. Special attention would be paid to recruiting underrepresented groups, such as rural populations, non-college-educated voters, and minority communities. I would allocate $2,500 as five 500 dollars rewards to incentivize participants.

When participants sign up, they would undergo IP verification to avoid fraudulent responses. All participants would need to verify their identity via email activation. Since the time required to fill out the form is short and the rewards are relatively high, participants would likely not be annoyed by the email activation step. To ensure data quality, I would also implement methods such as time checks and answer checks. If participants complete the survey too quickly or provide answers that are all "prefer not to say" or "other," those responses will be discarded.

After receiving all the surveys, I would adjust the weights based on demographics and votes from the last election to ensure the sample is representative of the U.S. population, using IPUMS 2024 U.S. Census data. I will then use a multilevel regression with post-stratification model to forecast the results using predictors such as state, gender, age, ethnicity, education, economic status, and religious affiliation.

Budget Allocation: 70K for recruitment and advertising; 5K for Incentives for respondents; 20K for data processing, weighting, and modelling; 5K for data security.

The link to the survey is here:

- https://docs.google.com/forms/d/e/1FAIpQLSeCFvjTdktOWxJnHqrWZkQbJth7xLXj3\YUuTkUWn0zvGGEbfw/vie

The survey questions are listed below:

1. **Which state do you currently live in?**

2. **What is your gender?**

   - a. Female
   - b. Male
   - c. Non-binary
   - d. Prefer not to say

3. **What is your age?**

4. **Who did you vote for in the last election?**

   - a. Joe Biden
   - b. Donald Trump
   - c. Other
   - d. I did not vote

5. **Who will you vote for in this election?**

   - a. Kamala Harris
   - b. Donald Trump
   - c. Other

6. **Do you consider yourself a:**

   - a. Strong Democrat
   - b. Democrat
   - c. Strong Republican
   - d. Republican
   - e. Independent

7. **What is your ethnicity?**

   - a. African American
   - b. Asian
   - c. Hispanic
   - d. White
   - e. Other

8. **What is your highest level of education?**

   - a. High school or less
   - b. College or University
   - c. Postgraduate
   - d. Doctoral
   - e. Prefer not to say

9. **What do you consider your economic status?**

   - a. Lower class
   - b. Lower-middle class
   - c. Middle class
   - d. Upper-middle class
   - e. Upper class

- f. Prefer not to say

10. **Are you religious?**

    - a. Yes
    - b. No
    - c. Prefer not to say

11. **How important are the following issues to you when deciding who to vote for?**

    - Economy and Taxes
    - Social Justice
    - Healthcare
    - Climate
    - Gun Control
    - Immigration

12. **On a scale 1 to 5, how likely are you going to vote?**

13. **How important do you think voting is?**

14. **What factors motivate you to vote? (Select all that apply)**

    - Concerns about specific issues
    - Media Coverage
    - Candidate personality
    - Influence of family and friends
    - Campaign promises
    - I always vote

15. **Any other thoughts or comments?**

6.3 Data Cleaning

I began by using the `clean_names()` function to standardize column names and then selected the most relevant variables as indicated in Section 2.1. To ensure data quality, I dropped any rows with missing values in the `numeric_grade` column and filtered out polls with a numeric grade less than 2, focusing on more reliable data. I also modified the `state` column to replace any missing state entries with "National" and standardized state names.

Next, I created a new column to indicate how recent each survey ended, using the `end_date` to calculate the number of days since the survey ends from now. I filtered for polls conducted within the last 60 days to ensure the data reflected current public sentiment. Additionally, I include only the candidates of interest, Donald Trump and Kamala Harris. Finally, I created a new column `harris_support_ratio` representing Kamala Harris's support ratio relative to the combined support for both her and Donald Trump.

6.4 Posterior Predictive Checks

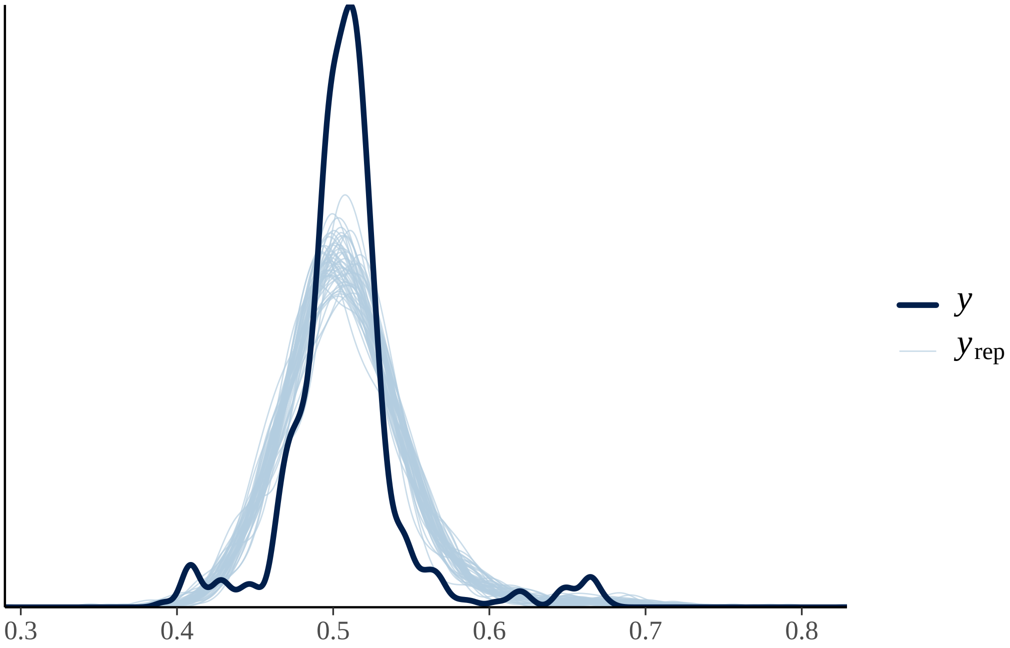


Figure 8: Posterior Prediction Check for the Forecast Model

References

270 to Win. 2024. "Nebraska Presidential Election Voting History - 270toWin — 270towin.com." https://www.270towin.com/states/Nebraska.

ABC News. 2023. "538's Polls Policy and FAQs — Abcnews.go.com." https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

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.

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

Gabry, Jonah, Michael Betancourt, Gabriel Laskey, Aki Vehtari, Mans Magnusson, and Paul-Christian Bürkner. 2018. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Kim Parker, Anna Brown, Juliana Menasce Horowitz, and Ruth Igielnik. 2018. "Urban, Suburban and Rural Residents' Views on Key Social and Political Issues — Pewresearch.org." https://www.pewresearch.org/social-trends/2018/05/22/urban-suburban-and-rural-residents-views-on-key-social-and-political-issues/.

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

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/.

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

Ryan Best, Ritchie King, Aaron Bycoffe, and Anna Wiederkehr. 2024. "National : President: General Election : 2024 Polls — Projects.fivethirtyeight.com." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Silver, Nate. 2024. "Pollster Ratings — Projects.fivethirtyeight.com." https://projects.fivethirtyeight.com/pollster-ratings/.

"State Electoral Vote History: 1900 to Present - 270toWin — 270towin.com." https://www.270towin.com/state-electoral-vote-history/.

Statista. 2024. "Number of Demonstrations, Including Riots and Protests, in the United States from January 2023 to August 2024, by State." https://www.statista.com/statistics/1484654/us-riots-and-protests-by-state/.

US Embassy & Consulate. "Technical Difficulties — Dk.usembassy.gov." https://dk.usembassy.gov/usa-i-skolen/presidential-elections-and-the-american-political-system/.

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

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

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://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.

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.

YouGov. "About YouGov | YouGov — Today.yougov.com." https://today.yougov.com/about.

———. 2024a. "2024 Presidential Vote Intent: Harris v. Trump — Today.yougov.com." https://today.yougov.com/topics/politics/trackers/2024-presidential-vote-intent-harris-trump?crossBreak=postgrad.

———. 2024b. "2024 Presidential Vote Intent: Harris v. Trump — Today.yougov.com." https://today.yougov.com/topics/politics/trackers/2024-presidential-vote-intent-harris-trump?period=3m.

———. 2024c. "Methodology | YouGov — Today.yougov.com." https://today.yougov.com/about/panel-methodology.