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

November 4, 2024

Abstract

The 2024 U.S. Presidential Election on November 5 will feature a contest between Democratic Vice President Kamala Harris and former Republican President Donald Trump. This paper employs a poll-of-polls method and a Bayesian generalized linear model to forecast the popular vote share for each candidate. By focusing on polls conducted by reliable pollsters within the last 60 days, the results predict that 50.9% of voters supporting either Harris or Trump will favor Harris, with her support concentrated in the West Coast and Northeast regions. Forecasting election outcomes is important as voter preferences help shape the campaign strategies and understand U.S. citizens' priority issues.

Table of contents

[*] Code and data supporting 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 elect the country's president for the next four years. The current vice president, Kamala Harris, is the Democratic Party candidate, while Donald Trump, the former president from the Republican Party, is running for a nonconsecutive term. The United States elects its president through the Electoral College. Each state's electoral votes depend on its representation in the Senate and House of Representatives, which means it is possible to win the popular vote yet lose the election. Despite the importance of electoral votes and the presence of multiple candidates, this paper will focus on the popular votes of Kamala Harris and Donald Trump because of the dominance of the Democratic and Republican parties in the current U.S. political system. Electoral votes will be calculated only for states with available data due to limited information 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 predict the outcomes with the following predictors: pollster, pollster's numeric grade, recency of the poll, and the respondent groups' population. Using the poll-of-polls method, which aggregates several polls to achieve greater forecast accuracy, the predicted national popular vote for Harris, considering only the two main candidates, is 50.9%, which is 1.8% higher than Trump's support. This difference will likely be smaller when all candidates in the 2024 election are accounted for. Based on predicted state results, Harris shows higher support in 16 of 28 states, Trump in 9, and 3 states have an equal split of 50% support for each. Harris's supporters are concentrated on the West Coast and in the Northeast, while Trump's supporters tend to be in the central part of the U.S. Presidential election forecasts have always been significant as they provide insight into citizens' priorities 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 focus. 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), `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 are based on poll results updated as of October 25. The original dataset includes 3,352 polls from various pollsters asking participants about their support for candidates in the upcoming presidential election.

The key variables from the dataset used in this analysis and their meanings 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 2.5 and that ended no more than 60 days ago, as voting preferences can shift over time. The meaning of the numeric grade and the reason for choosing 2.5 will be explained in Section 2.2.

2.2 Methodology and Measurement

Different polls from various pollsters use different methods to translate real-world phenomena into numerical data, but their underlying logic is similar. Pollsters typically begin by selecting a sample of the population that reflects the broader electorate; this could involve a national survey, a state-specific survey, or a targeted group. Based on the population sample and budget, the pollster will release the survey on preferred platforms. For example, they might release the poll on social media for more general and diverse participation or use news media to achieve a higher response rate. Effective pollsters design questions that are clear and unbiased, often asking respondents to choose from a list of candidates or political parties. Surveys usually include additional questions on demographics such as age, gender, and education, as well as political preferences and the likelihood that the respondent will vote. Good pollsters may adjust the weight based on these factors to enhance accuracy. Each response, representing an individual’s voting intention, becomes an entry in a dataset that captures these preferences.

This analysis employs the poll-of-polls method, which aggregates results from multiple polls rather than relying on a single survey. In this method, each poll is weighted based on factors such as sample size, recency, and the pollster’s historical accuracy. Polls with larger weights are considered more reliable in the aggregation. While any individual poll from the aggregated set could be used to forecast presidential election results, this analysis uses the aggregated poll to enhance accuracy and stability in predictions.

The aggregated dataset from FiveThirtyEight that I use 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 reliable survey methods and that pollsters are genuinely committed to 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 each polling firm's historical track record and methodological transparency (Silver 2024a). Additionally, polls are given reduced weight if the pollster has released a large number of surveys in a short period. The weight ranges from 0.5 to 3.0, with 3.0 indicating the most reliable pollsters. The numeric grades of all pollsters from the original dataset are shown in Figure 1, showing that the median numeric grade of all polls is 1.9, and the mean is approximately 2.17. A low numeric grade indicates low reliability, while filtering for only the highest grades would leave too few polls. To balance data quantity and reliability, I choose to include the top 40% of polls, filtering for a numeric grade of 2.5 or higher and excluding grades below 2.5. 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 include the Public Policy Institute of California, Pew Research Center, and the University of Illinois Springfield Survey Research Office (Silver 2024a).

More detailed information on the measurements and methodology for YouGov surveys can be found in Section 6.1. For the analysis below, the dataset has been cleaned, with the cleaning process detailed in Section 6.3.

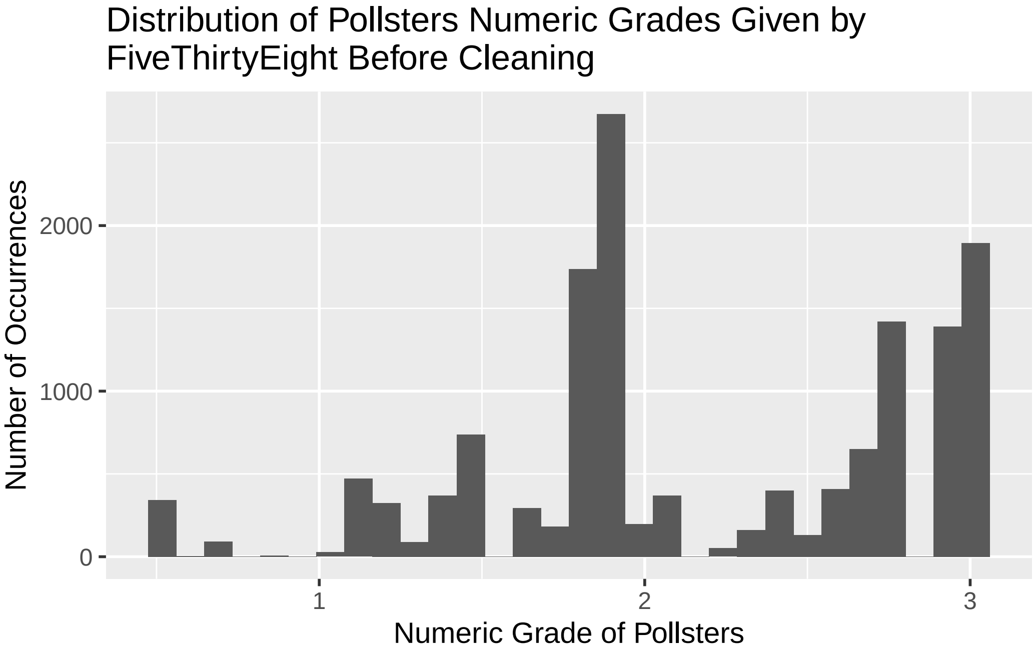
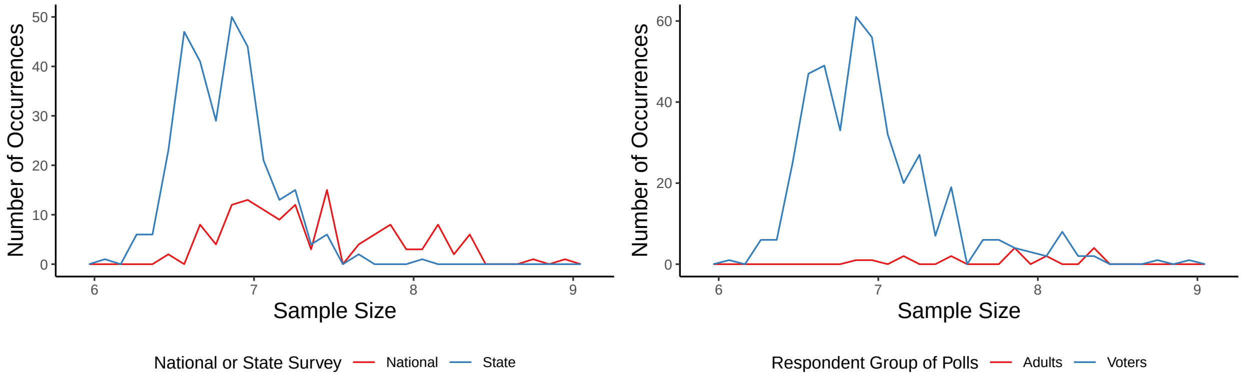


Figure 1: The numeric grades of all pollsters in the original dataset indicate a median grade of 1.9 and a mean of approximately 2.17

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 surveys have sample sizes under 3,000 (logarithm under 8), while the majority of state surveys fall below 1,100 (logarithm under 7). A few national surveys exceed 6,000, indicated by small red bumps around x = 9 in Figure 2 a. In Figure 2 b, most polls target adults as the respondent group, with only a few focusing on likely voters.



(a) Distribution of Log-Transformed Sample Sizes by Poll Type

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

Figure 2: The distribution of sample sizes, shown in logarithmic form, reveals that most surveys have sample sizes around 3,000, with state surveys generally below 1,100. Most polls target adults, while only a small portion focus specifically on voters.

Figure 3 displays the average support ratios for Kamala Harris and Donald Trump over the past 60 days from all the polls, with blue representing support for Harris and red representing support for Trump. The first presidential debate is marked, and the shaded ribbon areas indicate the range of support on different dates. From Figure 3, support levels fluctuate around 50%, showing that their support is closely aligned. Most of the time, Harris has slightly higher support than Trump, and there are three days when she has a significant lead. Right before the first presidential debate, Trump had a low support rate, which increased after the debate but remained below Harris's support for most of the time. The two numbers on the right represent their latest support ratios, with 50.19% supporting Harris and 49.81% supporting Trump.

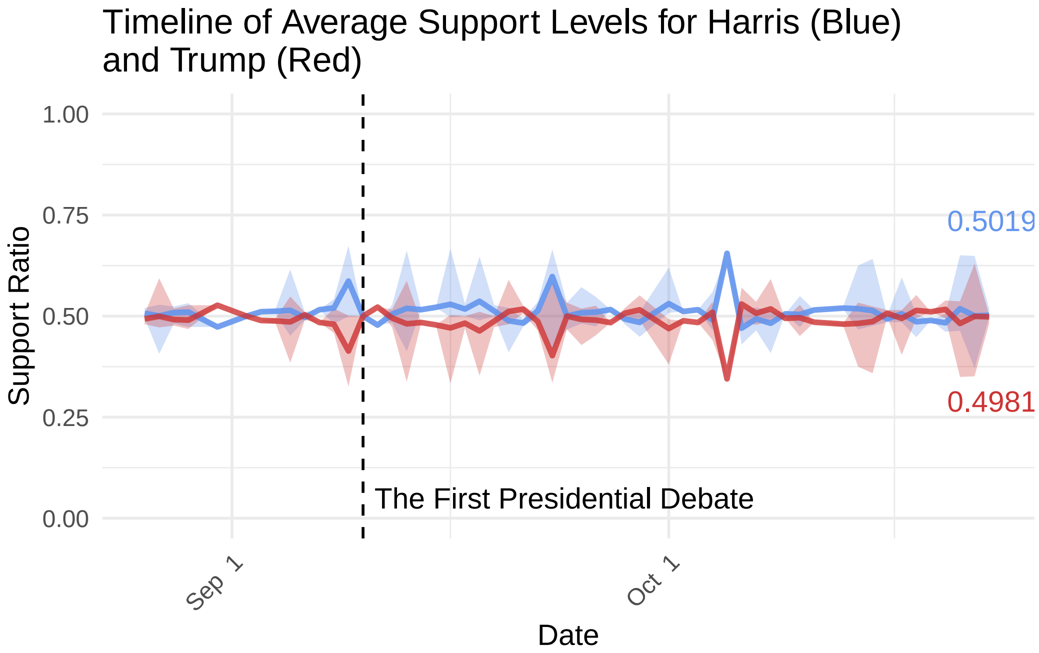


Figure 3: Timeline of average support levels for Harris and Trump from polls that ended within the last 60 days, with Harris generally having slightly higher support

From Table 1, the average proportion of support for Harris based on 131 national surveys is 51.1%, with a median support slightly higher at 51.2%. This indicates that just over half of the respondents favor her over Trump, who has an average support of 48.9%. Harris's support ranges from a minimum of 48.4% to a maximum of 53.5%. With a standard deviation of 0.012, 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 51.1%.

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

Avg Support for Harris	Median Support	Min Support	Max Support	SD of Support	Total Polls
0.511	0.512	0.484	0.535	0.012	131

Since the results of state polls are more skewed than those of national polls, I will calculate the weighted average proportion of support for Harris based on sample size for state polls. Table 2 shows the number of surveys conducted for each state, if applicable, along with the weighted average proportion of support. At least 30 polls target Arizona, Georgia, North Carolina, Pennsylvania, and Wisconsin, while Florida, Michigan, Nevada, and Texas each have at least 10 surveys. The remaining states have either a single-digit number of surveys or none at all. Out of the 28 states, 11 show more support for Trump, while 16 have more support for Harris, with North Carolina showing 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. South Dakota is the only state with less than 40% support for Harris, indicating that Trump is leading by more than 20% there.

Table 2: This table displays the weighted average proportion of support for Kamala Harris across various states, highlighting that 16 out of 28 states show more support for her, while 11 favor Trump. Notably, California, Maryland, Massachusetts, and Washington have over 60% support for Harris, whereas South Dakota has less than 40%.

Table 2: Weighted Average Support for Harris from State Surveys

State of Survey	Number of Survey	Weighted Average Proportion of Harris Support
Arizona	32	0.487
California	5	0.632
Connecticut	1	0.589
Florida	10	0.459
Georgia	32	0.495
Indiana	1	0.413
Iowa	1	0.478
Maryland	7	0.660
Massachusetts	4	0.653
Michigan	26	0.504
Minnesota	3	0.529
Missouri	1	0.440
Montana	5	0.409
Nebraska	8	0.503
Nevada	13	0.505
New Hampshire	3	0.545
New Mexico	1	0.543
New York	5	0.578
North Carolina	41	0.500
Ohio	9	0.468
Pennsylvania	42	0.506
Rhode Island	3	0.579
South Carolina	1	0.449
South Dakota	1	0.371
Texas	12	0.470
Virginia	6	0.543
Washington	1	0.620
Wisconsin	35	0.509

Figure 4 uses a gradient color 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; conversely, the redder a state appears, the higher the predicted proportion of support for Trump. States shown in grey have no available state surveys, so state-level data is lacking. 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 South Dakota, Montana, and Indiana, demonstrate strong support for Trump.

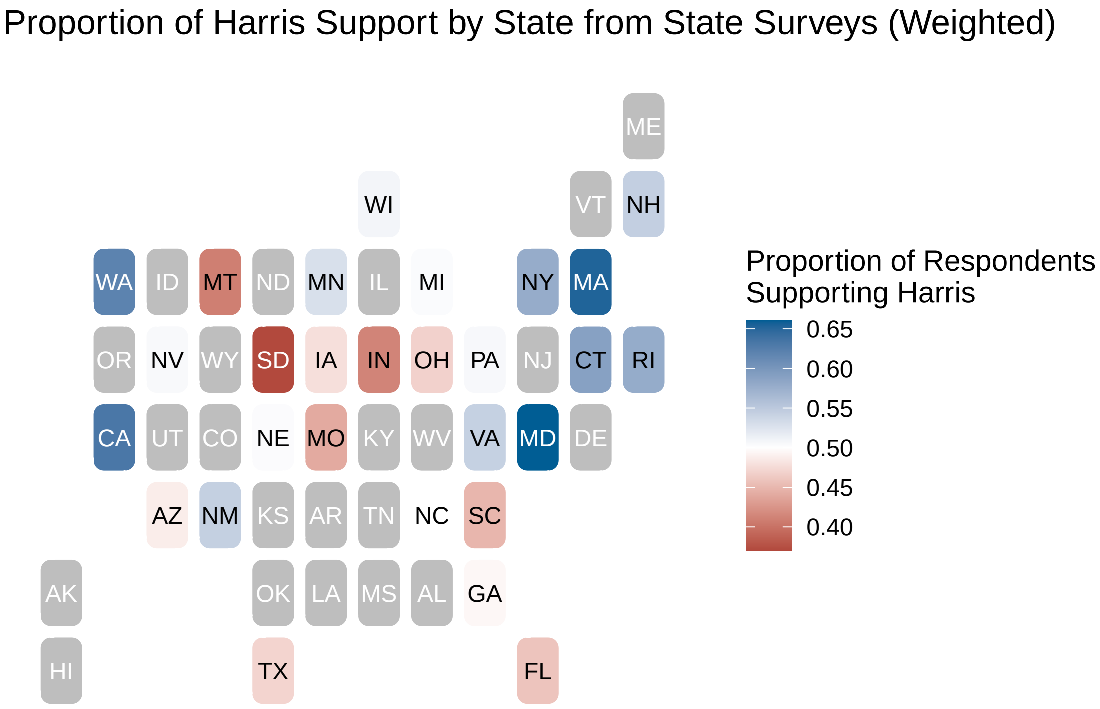


Figure 4: This map illustrates the weighted support for Kamala Harris and Donald Trump by state, with bluer states indicating higher support for Harris and redder states indicating higher support for Trump; states in grey lack available survey data. West Coast and Northeast states show strong support for Harris, while several Midwestern states favor Trump.

3 Model

3.1 Model Set-up

I use the `stan_glm` function from the `rstanarm` package (Goodrich et al. 2024) to create a Bayesian regression model with a Normal distribution that incorporates multiple predictors, using the programming language `R` (R Core Team 2022). 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, 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 proportion of support for Harris when all predictors are zero. It follows a normal prior distribution 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 normal prior distribution 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 both partisan and internal campaign polls, so I selected Pollster as one of the factors, since 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. Because 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, was selected as a factor. Additionally, I included population because exploring the voting preferences of individuals who cannot or will not vote is meaningless.

3.2 Model Justification and Limitations

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 also anticipate a positive relationship between Trump's support and pollsters that focus on states with higher Trump's support. 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.

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 that Harris's support is continuous, which may not be appropriate if the sample size is 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.

A model validation that shows the model is not overfit using out-of-sample testing is included in this repo, and can be found at `scripts/05-model_validation.R`. 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 is slightly smaller than 0.6, with a relatively larger confidence interval.

The coefficients for different pollsters vary, reflecting that they tend to attract audiences with differing voting preferences. As mentioned in Section 3.2, I expect to see a positive relationship between Harris's support and the pollsters that focus on states with high support for her, 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". Pollsters "UC Berkeley" and "UMass Amherst" also have positive coefficients, as they are located in California and Massachusetts, respectively, both of which show high support for Harris. Conversely, "Winthrop University" has a small negative coefficient because it is located in Southern Carolina, 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 from the pollster "GQR".

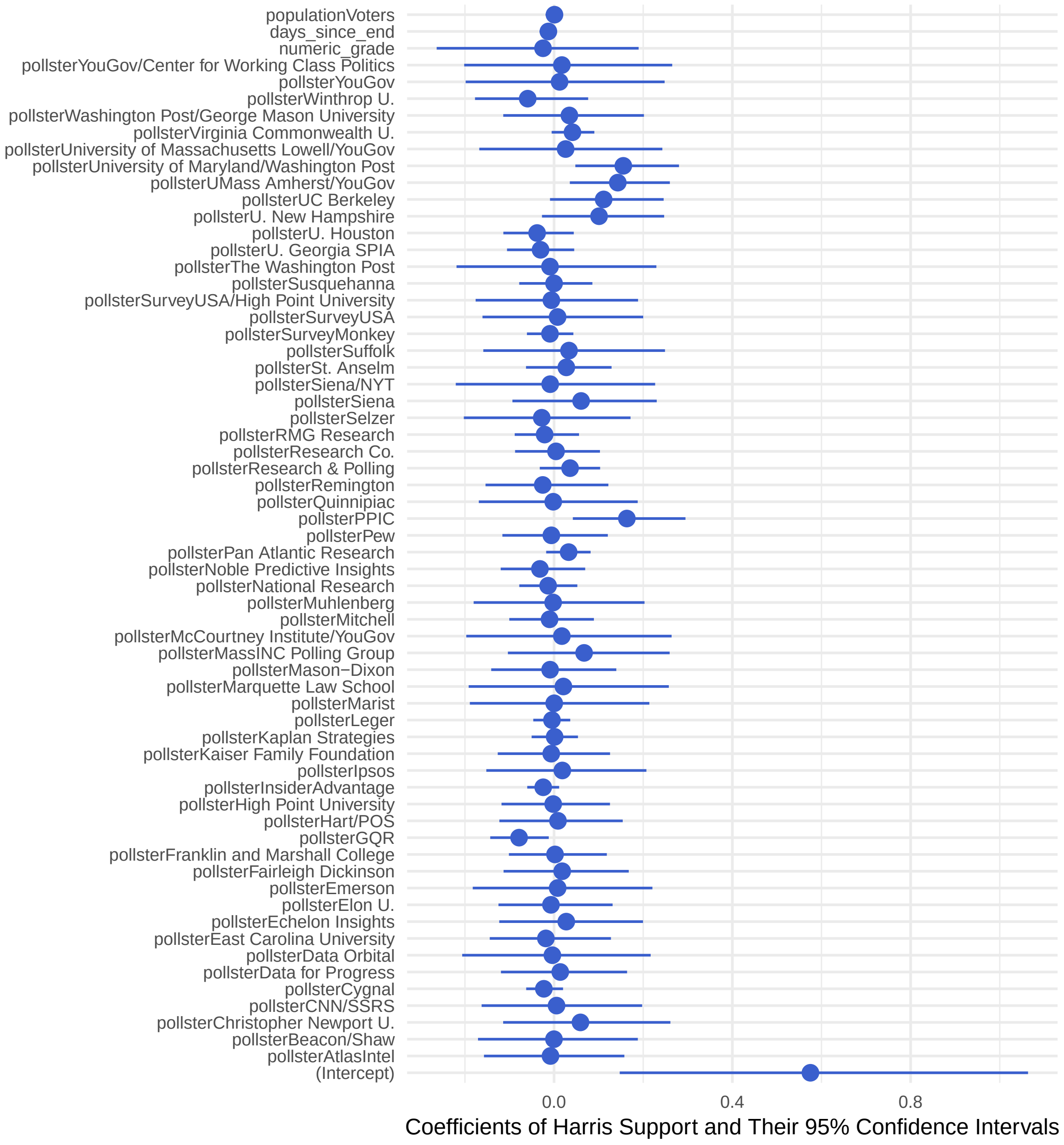


Figure 5: Model Results for Harris Support Based on Pollster, Pollster Quality, Survey Recency, and Survey Population

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 only considering voters selecting one of them. The proportion of Harris's support ranges from 48.2% to 53.1%, with a standard deviation of 0.011, which is slightly smaller than the standard deviation in Table 1, indicating less variation after prediction. The weighted average of Harris's support, considering sample size, is 50.8%, closely aligning with the unweighted proportion of 50.9% in Table 3.

Table 3: Summary Statistics for the Predicted National Average Support of Kamala Harris

Avg Support for Harris	Median Support	Min Support	Max Support	SD of Support	Total Polls
0.509	0.51	0.482	0.531	0.011	131

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 28 states with data, 16 show higher support for Harris, 9 show higher support for Trump, while 3 states have equal support for both candidates. California and Maryland maintain a support rate for Harris that exceeds 60%, and other states with high support for Harris include Rhode Island, New Hampshire, Massachusetts, New York, and Connecticut, all of which exceed 55%. South Carolina is the only state with less than 45% support for Harris.

Table 4: The predicted weighted support by state reveals a concentration around 50%, with 16 states favoring Kamala Harris, 9 favoring Trump, and 3 showing equal support. California and Maryland have over 60% support for Harris, while South Carolina is the only state with less than 45% support for her.

Table 4: Predicted Weighted Proportion of Harris Support by State

State	Predicted Proportion of Support For Harris (Weighted)
Arizona	0.498
California	0.601
Connecticut	0.570
Florida	0.489
Georgia	0.499
Indiana	0.505
Iowa	0.477
Maryland	0.621
Massachusetts	0.570
Michigan	0.500
Minnesota	0.506
Missouri	0.505
Montana	0.484
Nebraska	0.500
Nevada	0.501
New Hampshire	0.561
New Mexico	0.510
New York	0.565
North Carolina	0.499
Ohio	0.492
Pennsylvania	0.501
Rhode Island	0.582
South Carolina	0.448

State	Predicted Proportion of Support For Harris (Weighted)
South Dakota	0.499
Texas	0.500
Virginia	0.521
Washington	0.506
Wisconsin	0.507

Figure 6 uses a gradient color 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 appearing the bluest. In contrast, states in the central region show higher support for Trump, with South Carolina being the reddest. Compared to Figure 4, the overall color is lighter, revealing that the model has mitigated certain biases and preferences.

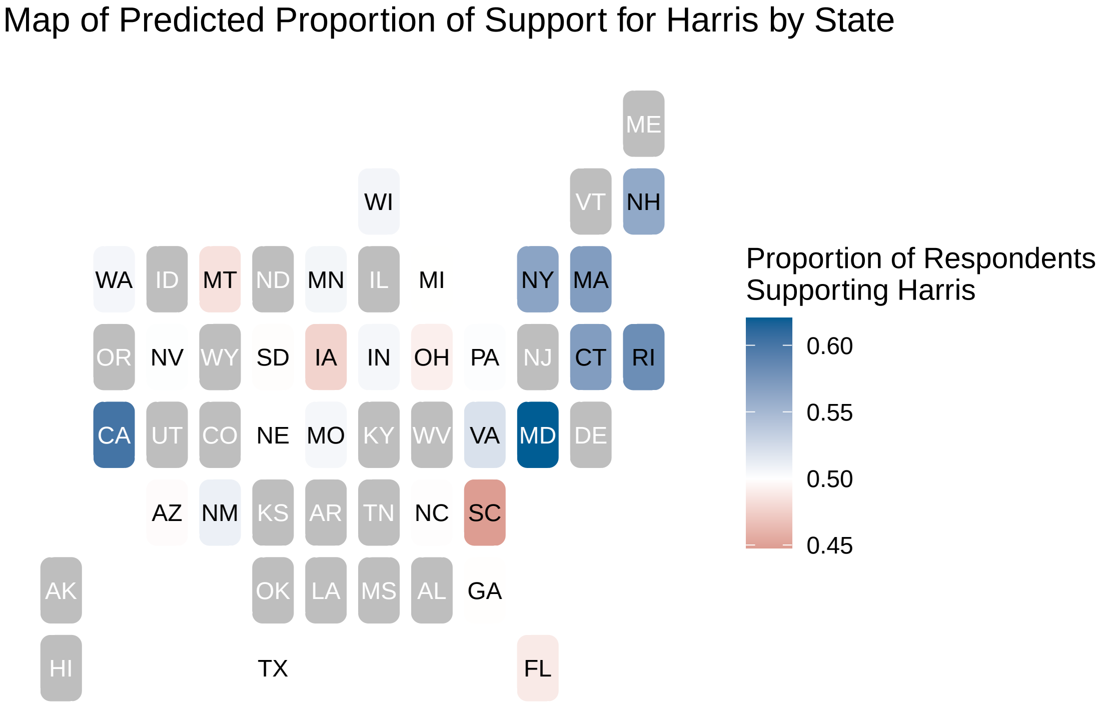


Figure 6: This map utilizes a gradient color scale to illustrate the predicted weighted support for Harris by state, with blue representing support for Harris and red representing support for Trump. Support for Harris is stronger in the West Coast and Northeast, while central states show higher support for Trump.

Considering the 28 states with data, the total electoral votes for Harris is 217, while Trump has 113 votes. Among 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 (270 to Win 2024b). Additionally, many of the excluded states, such as Wyoming, West Virginia, and Oklahoma, have a strong Republican inclination, indicating that Harris's lead in electoral votes among 28 states is not guaranteed.

5 Discussions

5.1 Each State Has Its Preference

The Democrats emphasize civil rights, social responsibility, and government-funded healthcare, while the Republicans focus on lower taxes, traditional values, and individual and family 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 central region lean Republican, as shown in Figure 6. This trend may also correlate with rural and suburban demographics, where urban voters tend to support the Democratic Party, while rural voters are more likely to favor Republicans (Kim Parker and Igielnik 2018). The Northeast is home to many prestigious universities, including Ivy League institutions. As illustrated in 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 favor 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 a higher number of protests from January 2023 to August 2024 align with those that support Harris (Statista 2024). Given that the Democratic Party's emphasis on social justice and diversity, this correlation between support for Harris and progressive values is reasonable. Regions with fewer protests tend to support Trump, where stability and tradition are more prevalent.

5.2 Pollster and Polling Methodology

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 others favor the Republican Party. Therefore, I hope that these effects will counterbalance each other in this analysis, reducing potential bias. This 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.

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 difference in support rates underscores how data collection methods can inadvertently privilege certain voices while marginalizing others, thereby revealing underlying social inequalities. Accurate representation in polling is necessary for reflecting the diverse views of the electorate.

5.3 Shy Trump Voters Theory and Non-Response Bias

Some people who believe Trump will win talk about the concept of "shy Trump voters". This theory suggests that some individuals don't openly admit they're voting for conservative candidates like Trump due to social pressure. This is similar to Britain's "shy Tories" phenomenon, where polls tend to underestimate Conservative support. However, there isn't much solid evidence to support the shy Trump voter theory, as many argue that a significant portion of Trump supporters are actually open and proud (Silver 2024b).

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)

Instead of "shy Trump voters", the real issue may be non-response bias. This occurs when pollsters fail to reach certain voter groups, rather than when individuals hide their true intentions. In both 2016 and 2020, pollsters struggled to connect with enough Trump supporters, likely because Trump's base often has lower civic engagement, making them less likely to participate in surveys (Silver 2024b).

Additionally, as a Black female candidate, Kamala Harris could face unique challenges. While the "Bradley effect" – where voters hesitate to admit they won't vote for a Black candidate – didn't affect Barack Obama in the past, Harris could still encounter similar obstacles. For instance, in 2016, undecided voters leaned away from Hillary Clinton, which might be a risk for Harris (Silver 2024b).

5.4 Electoral Votes and Swing States

The predicted support for Harris in each state, based on Table 4, shows that Harris leads with 217 electoral votes, while Trump holds 113. Among the three tied states, Michigan, Nebraska, and Texas, Texas is typically Republican-leaning and has a substantial number of electoral votes, suggesting that support for Trump may be underestimated. Although the difference between support for the Republican and Democratic parties has been narrowing over the years, with Democratic support rising from 38.0% of voters in 2000 to 46.5% in 2020 (270 to Win 2024c), there was still more than a 5% Republican lead in the last election. If Texas, with its 40 electoral votes, ultimately leans Republican, this would significantly boost Trump's standing in the electoral count. Additionally, certain states not included in this analysis, such as Wyoming, West Virginia, and Oklahoma, are known to have strong Republican inclinations, suggesting that Harris's lead in electoral votes among these 28 states may not guarantee an overall lead nationwide.

Regarding swing states, Florida, with 48.9% support for Harris, is the only state indicating at least a 2% lead. Among other states, Nevada, Pennsylvania, and Wisconsin show a slight preference for Harris, while Arizona, Georgia, and North Carolina show slight support for Trump (270 to Win 2024b). These small leads remain uncertain, with support close to evenly split.

Swing states or states without data that have more than 15 electoral votes include Pennsylvania (slightly prefers Harris), Illinois (no data), Georgia (slightly prefers Trump), Michigan (tied), and North Carolina (slightly prefers Trump).

5.5 Limitations and Weaknesses

I eliminated all polls conducted by pollsters with a grade lower than 2.5, as this grade is assigned by FiveThirtyEight based on historical accuracy and methodological transparency. However, one potential weakness of the model is that I also included the numeric grade as a predictor, which may duplicate its effect on the model.

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.

Another limitation is the aggregation of certain state-specific congressional districts with their respective states. For instance, Nebraska's 2nd Congressional District (CD-2), which awards an electoral college vote independently, is grouped with the rest of Nebraska in the dataset. Though the proportion of CD-2 is small in the dataset, this may ignore unique voting patterns in regions like CD-2 that have distinct electoral significance. 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 2024a). However, Table 2 indicates a 50.3% support for Harris in Nebraska, and Figure 7 shows that surveys conducted using the "live phone" or "live phone/text-to-web/email/mail-to-web/mail-to-phone" methodologies has 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. This raises questions about whether the polls are biased or if residents are genuinely changing their opinions this year.

5.6 Next Steps

This year's election is particularly unique due to Biden's withdrawal a few months before the election. For future studies, 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 behavior. Comparing the results of the previous election with current trends will demonstrate how unforeseen events can alter political preferences among citizens. 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.

6 Appendix

6.1 YouGov Surveys

YouGov is an international online research data and analytics technology group dedicated to providing unparalleled insights into global public opinion. As a leading platform for online market research, YouGov draws information from a continuously growing dataset of over 27 million registered panel members, which they refer to as "living data." Their innovative approach ensures the accuracy and actionability of consumer insights. Globally recognized for their data precision, YouGov is frequently cited by the media and is considered one of the most trusted sources of market research (YouGov 2023).

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

6.1.1 Methodology Overview

YouGov conducts its surveys using online polling, targeting all U.S. citizens eligible to vote. Respondents are selected through non-probability sampling, meaning that not everyone in the population has an equal chance of being chosen. However, 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 under study, including age, gender, race, education, and voting behavior. The data is then adjusted so that the survey results align with the actual distribution of these characteristics in the target population. This means that if a survey has a higher or lower proportion of individuals from a specific demographic group than is present in the population, the results are weighted to correct for this imbalance.

YouGov employs several verification and quality control measures. 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 data accuracy. 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 various 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 encompasses more than 95% of Americans. Respondents are incentivized through a points system that can be exchanged for small rewards. When determining whom to invite for participation in surveys, YouGov considers several factors, such as how recently a respondent has completed a survey, their preference for 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 balance reliability and efficiency.

YouGov employs a multilevel regression with post-stratification model for vote estimation. The 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 grants 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 posed is, "In November 2024, who would you vote for in the presidential election if these were the candidates?" The wording and response options are varied over time (YouGov 2024a). Respondents are selected using random sampling, stratified by gender, age, race, education, geographic region, and voter registration from the most recent American Community Survey. The sample is weighted according to gender, age, race, education, 2020 election turnout and presidential vote, baseline party identification, and current voter registration status.

6.1.2 Methodology Evaluation

YouGov employs a well-structured methodology for conducting its surveys, enabling rapid data collection and analysis. Although they use non-probability sampling, which means that not everyone in the population has an equal chance of being selected, the sample is adjusted to better reflect the target population. The fact that they recruit participants from a variety of sources and implement several verification and quality control measures also enhances the quality of the survey results.

One of the strengths of YouGov's methodology is its use of statistical weighting to adjust for demographic imbalances in the sample. By matching respondents to the target population based on key characteristics such as age, gender, race, education, and voting behavior, YouGov mitigates biases and ensures that its polls are representative. Another strength is its focus on data accuracy and quality control. For example, monitoring the consistency of responses can significantly enhance the reliability of the data collected.

A limitation is that the census data used for weighting purposes comes from the 2019 American Community Survey. This data from five years ago may not accurately reflect current demographic information when used for adjustments. Utilizing more recent census data could enhance the accuracy of their weight adjustments.

6.2 Idealized Survey

There are four key questions for pollsters (Clinton 2024):

- Do respondents match the electorate demographically in terms of sex, age, education, race, etc.? (This was the problem in 2016)
- Do respondents match the electorate politically after the sample is adjusted by demographic factors? (This was the problem in 2020.)
- Which respondents will vote?
- Should the pollster trust the data?

With these four questions in mind, I would create a idealized survey as follows.

6.2.1 Methodology

If I had a budget of $100K to forecast the U.S. presidential election, I would conduct a survey with a sample size of 5,000 to 8,000 respondents, consisting of 15 questions. I would aim to keep the survey duration between 5 and 10 minutes, as shorter surveys tend to produce higher response and completion rates than longer surveys (Anaesth 2022). To ensure that the sample accurately reflects the U.S. population, I would use stratified sampling based on key demographics such as age, gender, race, education, and income. Respondents would be asked whom they voted for in the last election and which political party they lean toward, allowing for adjustments based on demographic and political factors.

The main body of respondents would be built through an online opt-in panel. This panel would be recruited using a mix of traditional advertising, partnerships with news websites like The New York Times, and social media outreach to ensure diverse representation. I would allocate a larger portion of the budget to social media platforms like Instagram and YouTube, as these are widely used, and the expected participation rate on social media may be lower than on news media. Special attention would be given to recruit underrepresented groups, including rural populations, non-college-educated voters, and minority communities. To incentivize participation, I would allocate 2,500 dollars for rewards, distributed as five $500 cash rewards.

Upon signing up, participants would undergo IP verification to prevent fraudulent responses, and all participants would be required to verify their identity via email activation. Given the short time required to fill out the form and the relatively high rewards, participants are unlikely to be annoyed by the email activation step. To ensure data quality, I would implement measures such as time checks and answer consistency checks. If participants complete the survey too quickly or provide answers that are predominantly “prefer not to say” or “other,” those responses would be discarded, helping to ensure the reliability of the data.

Merely asking, “Who did you vote for in the last election?” with the option “I did not vote” is insufficient, as it assumes that past voting behavior will replicate in the 2024 election (Clinton 2024). Therefore, I would include three additional questions regarding participants’ attitudes toward voting to assess their likelihood of voting. The survey questions are provided at the end of this section.

After collecting all survey responses, I would adjust the weights based on demographics using IPUMS 2024 U.S. Census data, political preferences based on previous election results, and likelihood to vote, ensuring a more representative sample. I would then employ a multilevel regression with post-stratification model to forecast the results.

Budget Allocation:

- 60K for recruitment and advertising
- 5K for Incentives for respondents
- 25K for data processing, weighting, and modelling
- 10K for data security

6.2.2 Survey Questions

The link to the survey is here:

- https://docs.google.com/forms/d/e/1FAIpQLSeCFvjTdktOWxJnHqrWZkQbJth7xLXj3YUuTkUWn0zvGGEbfw/viewform?usp=sf_link

The survey introduction and questions are listed below:

This survey is designed to understand voter preferences and demographic information. Your responses will remain anonymous and will be used to analyze and forecast the outcomes of the 2024 presidential election. Completing the survey will give you a chance to win a $500 cash reward. This survey should take approximately 5-10 minutes to complete. If you have any concerns or questions, please contact the survey coordinator, Boxuan Yi, at boxuan.yi@mail.utoronto.ca.

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?**

Thank you for your time and response. Again, your responses will remain anonymous and will be used to analyze and forecast the outcomes of the 2024 presidential election.

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

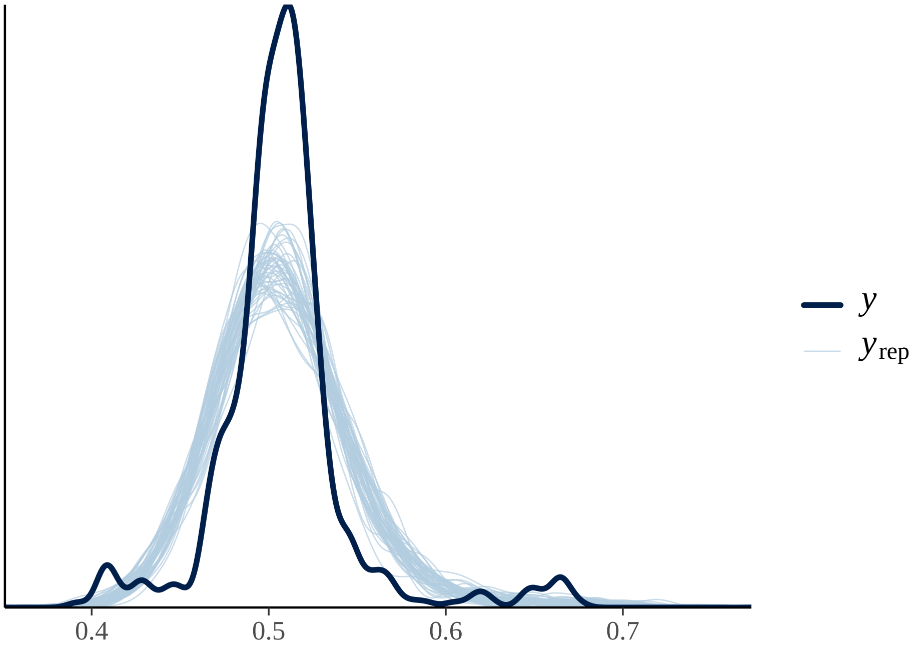


Figure 8: Posterior Prediction Check for the Forecast Model

References

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

———. 2024b. "State Electoral Vote History: 1900 to Present." https://www.270towin.com/state-electoral-vote-history/.

———. 2024c. "Texas Presidential Election Voting History - 270toWin — 270towin.com." https://www.270towin.com/states/Texas.

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

Anaesth, Saudi J. 2022. "How Short or Long Should Be a Questionnaire for Any Research? Researchers Dilemma in Deciding the Appropriate Questionnaire Length." https://pmc.ncbi.nlm.nih.gov/articles/PMC8846243/.

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.

Clinton, Josh. 2024. "Poll Results Depend on Pollster Choices as Much as Voters' Decisions." https://goodauthority.org/news/election-poll-vote2024-data-pollster-choices-weighting/.

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." 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. "Presidential General Election 2024." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Silver, Nate. 2024a. "538's Pollster Ratings." https://projects.fivethirtyeight.com/pollster-ratings/.

———. 2024b. "Opinion | Nate Silver: Here's What My Gut Says About the Election, but Don't Trust Anyone's Gut, Even Mine." https://www.nytimes.com/2024/10/23/opinion/election-polls-results-trump-harris.html?unlocked_article_code=1.UU4.pFkQ.F2hD-woxmiEj&smid=url-share.

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. "Presidential Elections and the American Political System." https://dk.usembassy.gov/usa-i-skolen/presidential-elections-and-the-american-political-system/.

Wickham, Hadley. 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. 2023. "About YouGov." https://today.yougov.com/about.

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

———. 2024b. "Methodology." https://today.yougov.com/about/panel-methodology.