# US President Prediction*

**Predict the results of 2024 US President Election**

Yun Chu Felix Li Wen Han Zhao

November 4, 2024

In this study, we aim to study the 2024 U.S. Presidential Election with the methodology of polls of polls across various states. With the insightful analyses on polls provided by Redfiled & Wilton Strategies, the conducted survey on the population highlights various key point for the prediction of US president. Using Bayesian linear model, this report analyzes the prediction of presidential votes through data collected from pollsters. The prediction finding highlights the significance of candicate trustworthiness among the voter decision, providing insights for swing voters. This analysis offers valuable information into the potential electoral outcomes driving the 2024 US president election.

## Table of contents



*Code and data are available at: https://github.com/younazhao/US-President-Prediction/tree/main.

# 1 Introduction

As November approaches, the United State of America's presidential election will soon be determined. The tight competition between Kamala Harris representing the Democratic Party and Donald Trump representing the Republican Party will determine the next President of United State for a 4 year term beginning January 2025 Time (2024). The final result will be held on November 5th 2024, which will be decided by a favorable outcome biased toward one candidate. Before this final date, various pollsters will use the method of "polls by polls" to provide a high-level view of public sentiment by averaging results from multiple polling sources and methods FiveThirtyEight (2024a). From the dataset provided by Five Thirty Eight FiveThirtyEight (2024b), the paper could identify various results from polls of polls in order to predict the next president of America.

By performing a Bayesian linear model, which formulate linear regression using probability distributions rather than point estimates Science (n.d.), the analysis focus on the support rate

for the 2 main candidate, Kamala Harris and Donald Trump. This approach allows for a nuanced understanding of the likely range of outcomes by capturing variability in the polling data. By combining insights from various pollsters and applying a Bayesian linear model, this paper seeks to forecast the presidential election based on current trends and public opinion.

To do this, Five Thirty Eight data were analysed from 2022 to 2024 in Section 2. With our predicted model in Section 3, the results in Section 4 show a preference toward Harris with a positive intercept of 40.97, over the results of Trump. In addition, the overall trend of Harris support rate has a larger increments than Trumps, suggesting a higher probability of winning this election. A discussion is provided in results and future suggestion in Section 5. In the appendix, a pollster methodology analysis on Redfield & Wilton Strategies is performed in Section A.1. Lastly, an idealized methodology and survey is conducted in Section A.2.

# 2 Data

## 2.1 Overview

The data used in this paper is sourced from FiveThirtyEight (FiveThirtyEight 2024b). The polls dataset contains 52 columns and 16,408 rows. It has polls information from multiple pollsters with variables that evaluate each pollster's quality like numeric grade, pollscore, transparency and sample size, as well as other information about the pollster like start and end date of the poll. Similar datasets are available but this dataset has most precise date and is the reason that we choose this dataset.

We use the statistical programming language R (R Core Team 2023) to download, clean and analyze data, as well as in exploratory data analysis and building models. Besides, the following packages are utilized in model generation and result production process:

- tidyverse (Wickham et al. 2023)
- lubridate (Grolemund and Wickham 2011)
- dplyr (Wickham, François, et al. 2023a)
- knitr (Xie 2023)
- kableExtra (Zhu 2023)
- readr (Wickham, François, et al. 2023b)
- rstanarm (Andrew et al. 2023)
- bayesplot (Gabry et al. 2023)
- ggplot2 (Wickham 2016)
- ggpubr (Kassambara 2023)
- tidyr (Wickham 2023)
- janitor (Firke 2023)
- scales (Wickham and Seidel 2022)
- arrow (Richardson et al. 2023)

## 2.2 Measurement

Since the dataset is a collection of polls from various pollsters, the measurement is done differently. We're interested in how each pollster conduct the polls. In Section A.1, the methodology and measurement of Redfield & Wilton Strategies is explained in detail.

## 2.3 Data Cleaning

The data is cleaned by filtering out Harris and Trump respectively. Based on the average of the variables that evaluate the quality of the pollsters, we also filtered out the entries with insignificant and low numeric grade, transparency and pollscore.

## 2.4 Summary Statistics

Table 1: Poll Percentage Attributes

Table 1: Statistics for Poll Percentage

| Metric | Value |
|---|---|
| Mean | 33.63353 |
| Standard Deviation | 17.97572 |
| Minimum | 0.00000 |
| Maximum | 70.00000 |

Table 1 summarize the atribute of the poll percentages from the whole dataset. With maximum 70% and minimum 0%. We can tell that there is not extreme or abnormal values of the poll percentages.

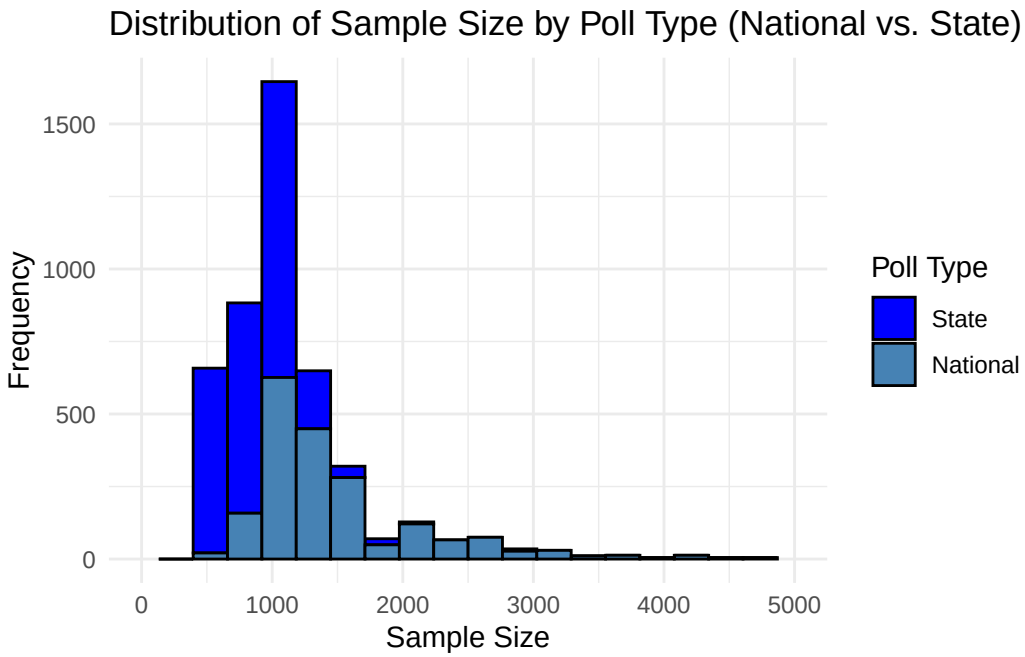


Figure 1: Distribution of Polls by Pollsters

Figure 1 shows the distribution of polls by sample size and state. National polls has larger sample size but lower frequency while State polls has the opposite trend.

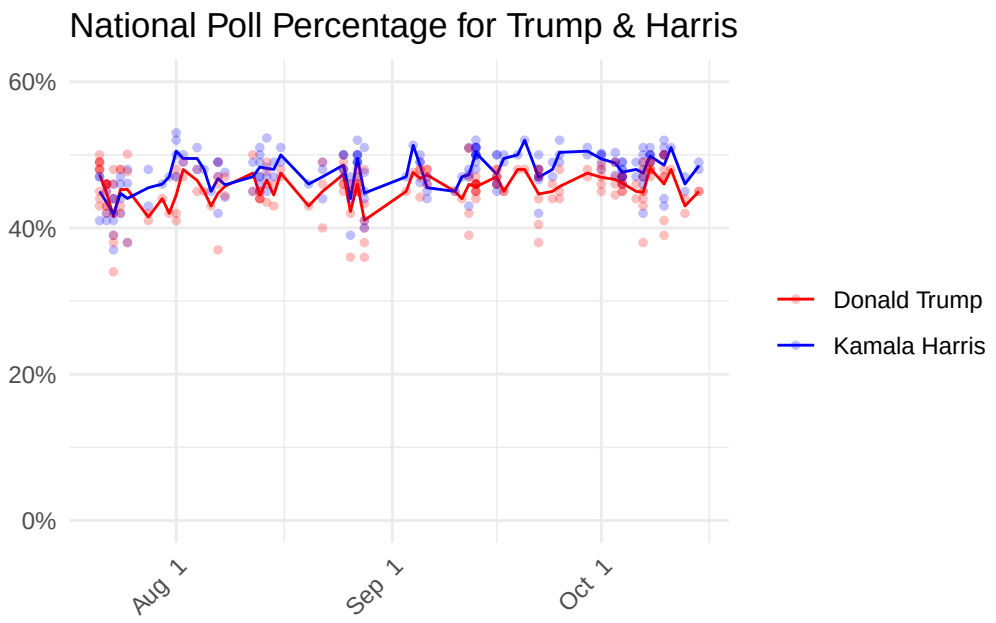


Figure 2: National Poll Percentage for Trump & Harris

Figure 2 shows the national poll percentage for the most competitive candidates: Donald Trump and Kamala Harris since Kamala Harris became the president candidate. As the date closet to election day, the poll percentages of the two candidates become closer and closer.

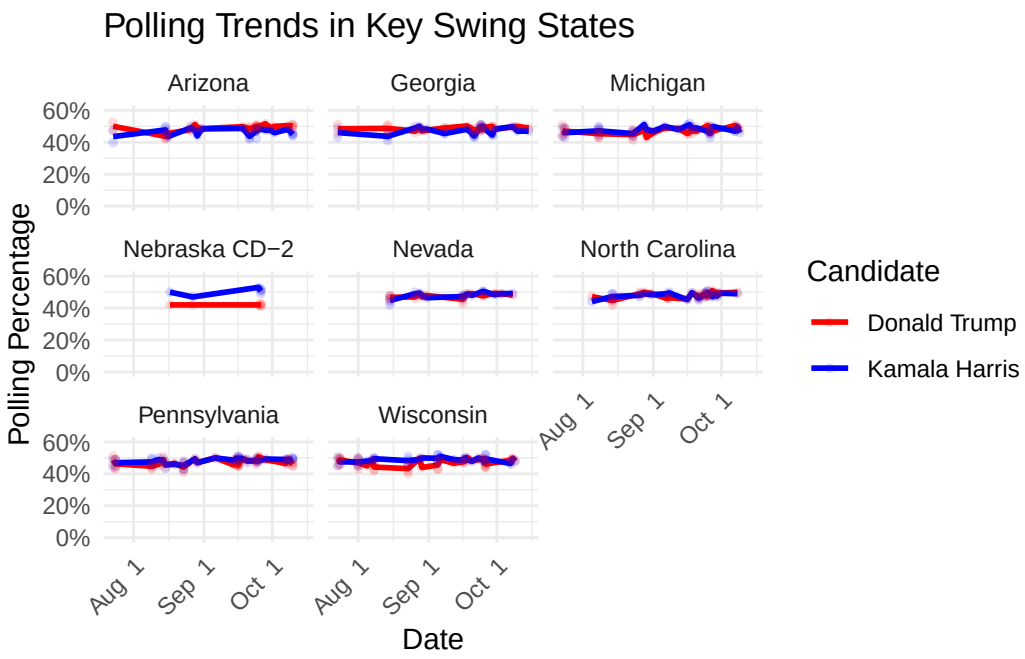


Figure 3: Swing States Poll Percentages

Figure 3 shows polling trends for Kamala Harris and Donald Trump in swing states. Most states show a close race with stable trends, except Nebraska CD-2, where Harris leads. This highlights the competitiveness of these states in the election.

# 3 Model

In this analysis, we aim to model the popularity trends for the two 2024 president nominates, Kamala Harris and Donald Trump, based on high-quality polling data collected at the national level after Harris's declaration on July 21, 2024. We used Bayesian linear regression models with Gaussian error structures to estimate changes in polling percentages over time for each candidate.

## 3.1 Model Check

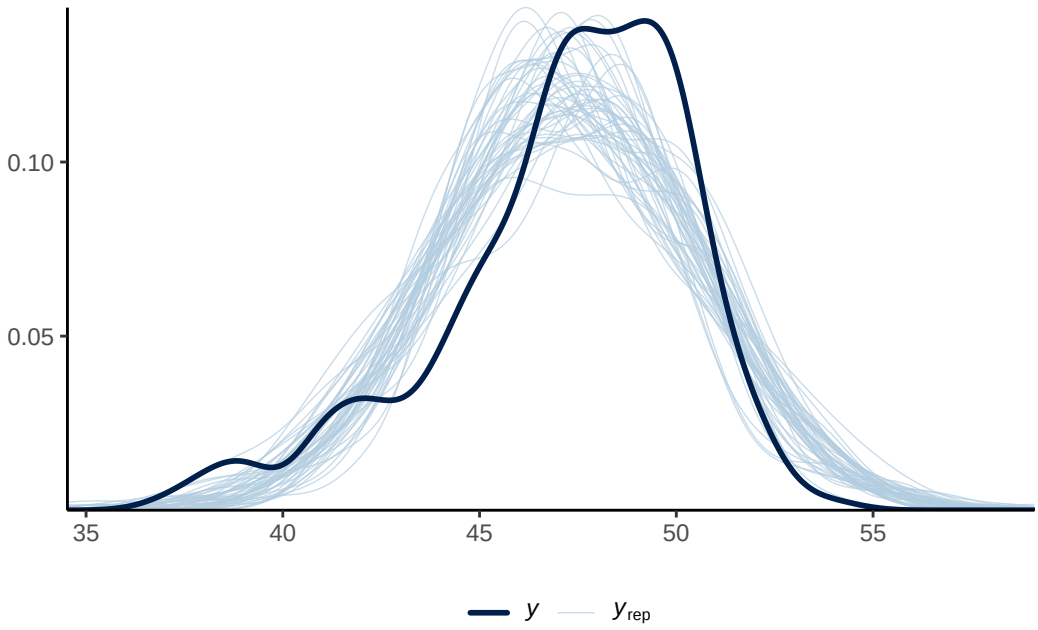


Figure 4: Posterior Predictive Check for Harris model

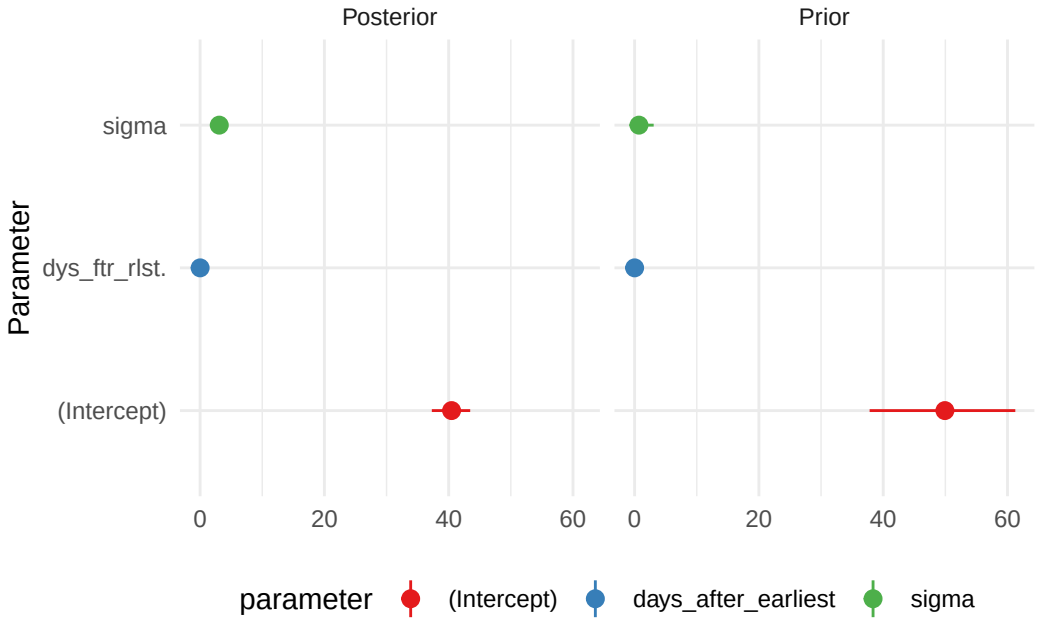


Figure 5: Comparing the posterior with the prior for Harris model

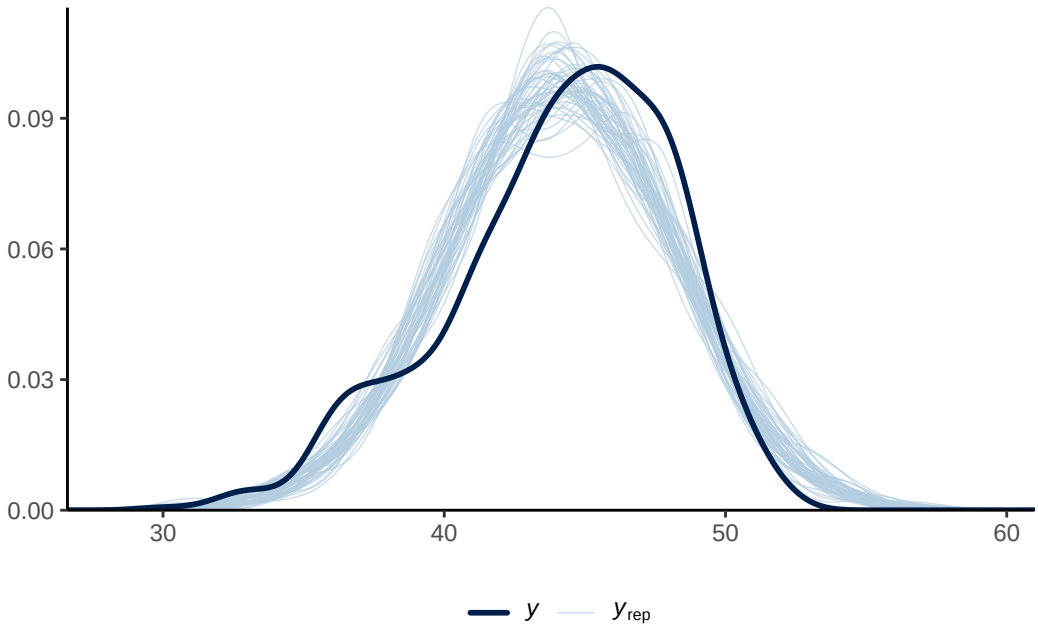


Figure 6: Posterior Predictive Check for Trump model

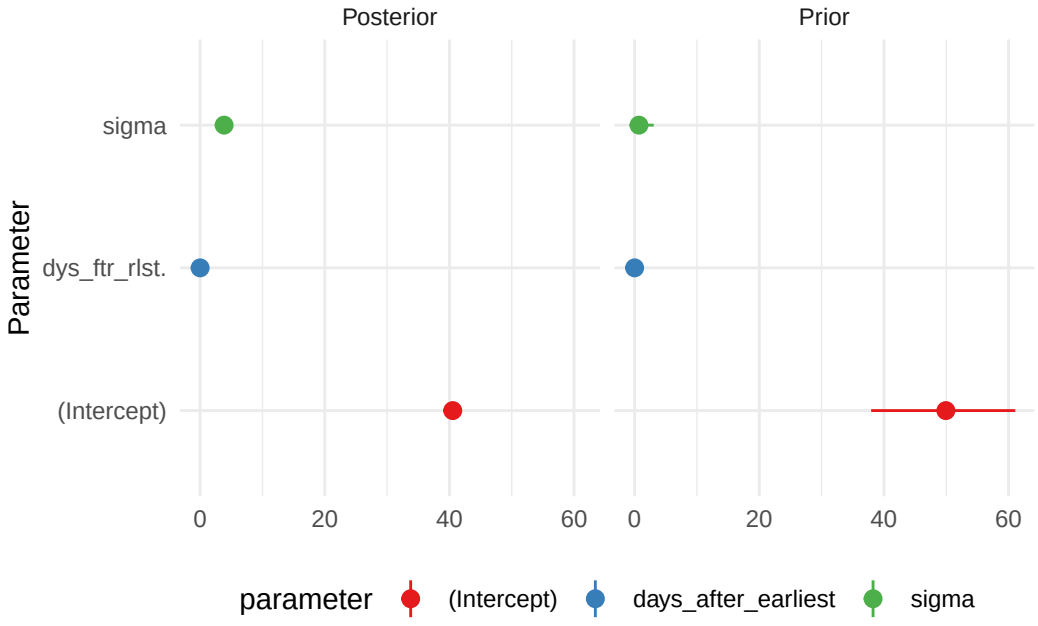


Figure 7: Comparing the posterior with the prior for Trump model

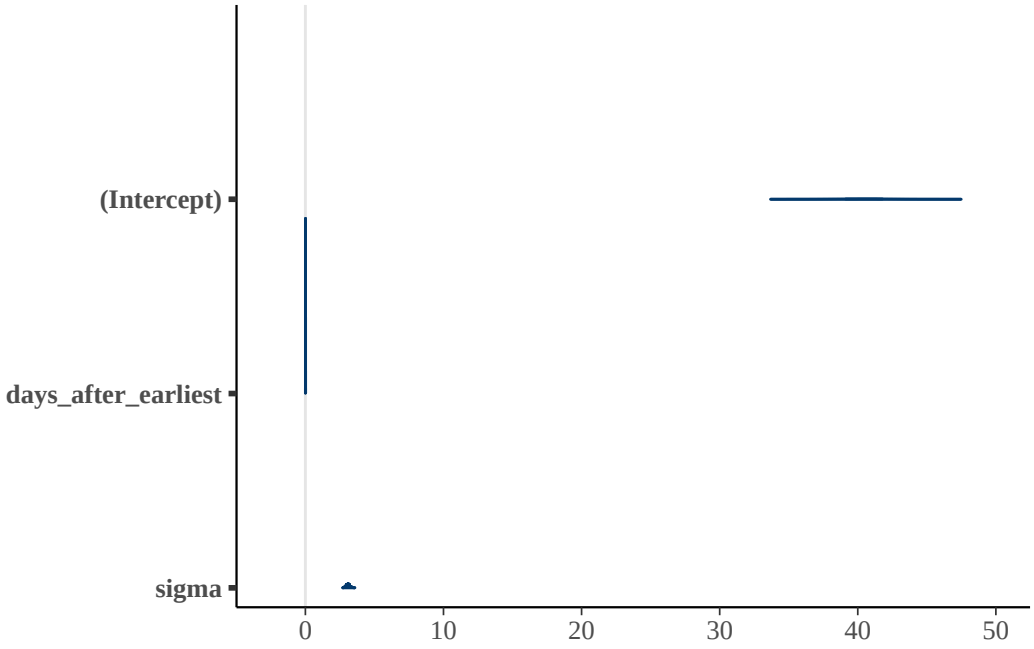


Figure 8: CI for predictors for Harris

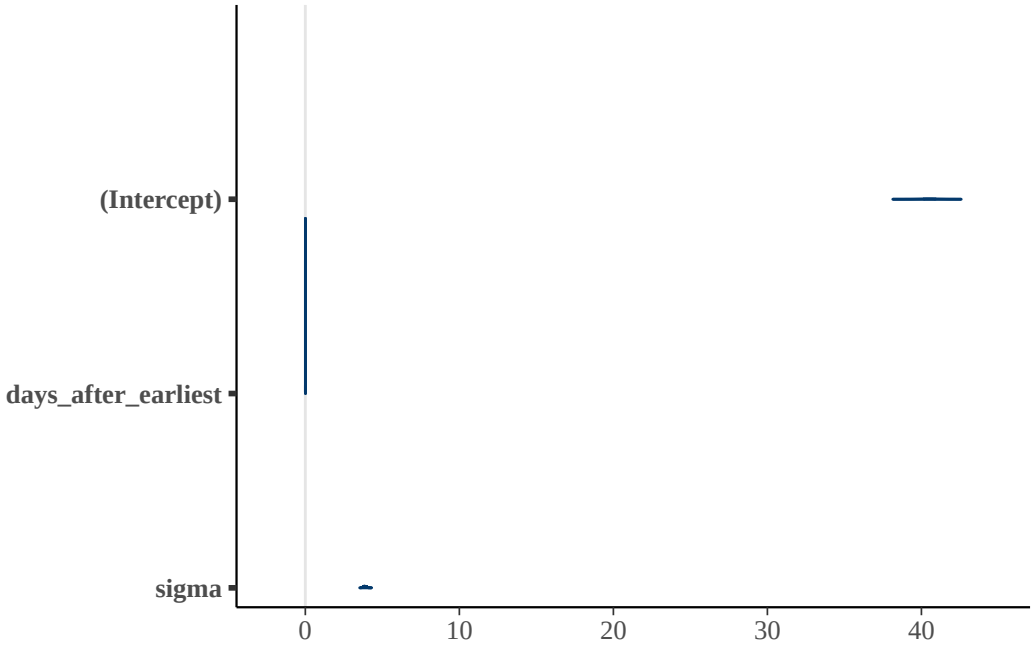


Figure 9: CI for predictors for Trump

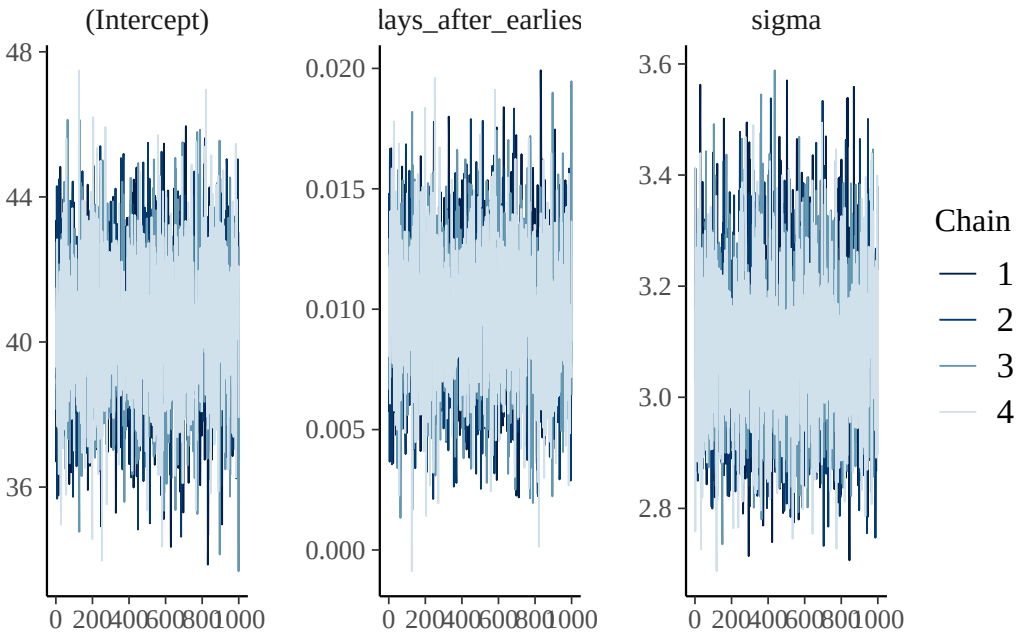


Figure 10: Trace plot for Harris model

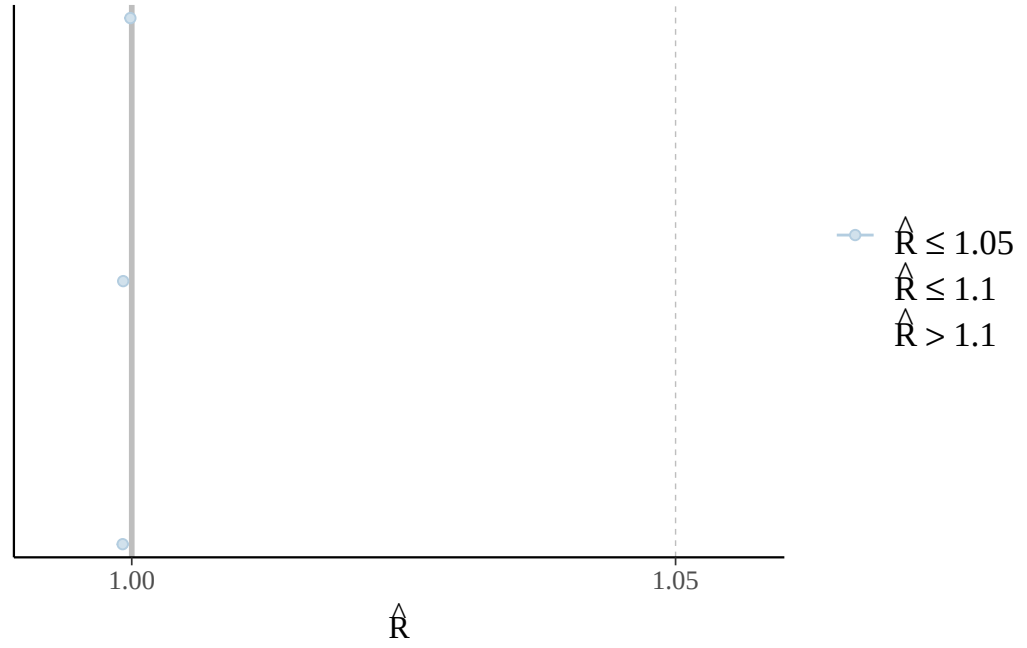


Figure 11: Rhat plot for Harris model

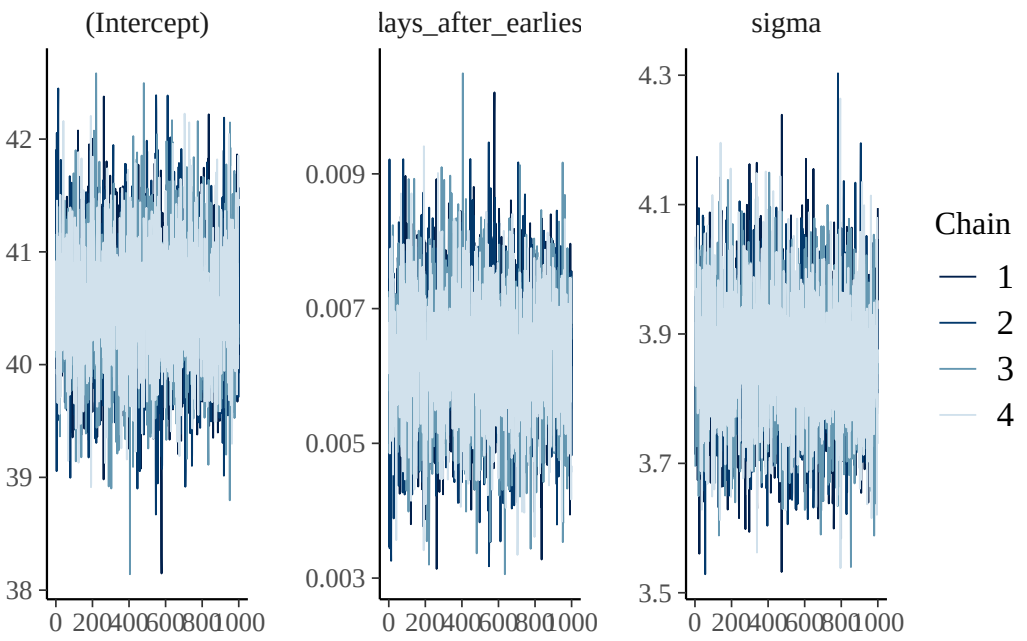


Figure 12: Trace plot for Trump model

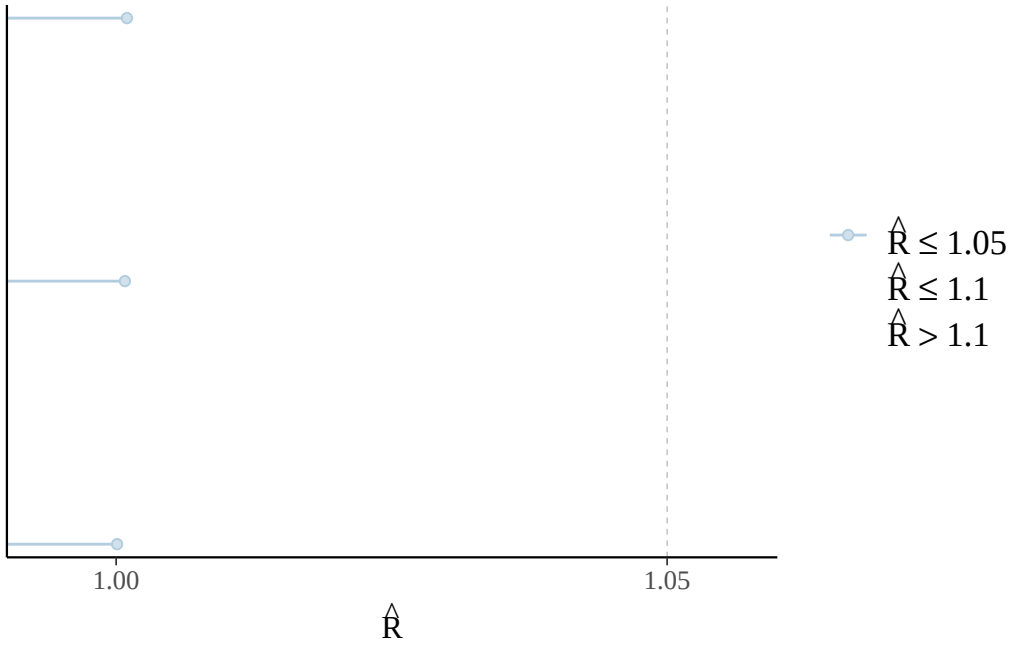


Figure 13: Rhat plot for Trump model

## 3.2 Data Filtering and Preparation

The data were first filtered to retain only high-quality polls, defined as those with a numeric grade of 2.0 or above and a transparency score of 4 or higher. We focused exclusively on polls where the *state* variable indicates national coverage, ensuring the poll represents nationwide opinions rather than a single state, which could introduce bias. Additionally, we filtered for polls conducted after July 21, 2024, the date when Kamala Harris announced her candidacy.

To prevent issues with missing data in our models, we excluded any rows where key variables (such as polling percentage or end date) were missing.

## 3.3 Model Specification

Two separate Bayesian linear regression models were fitted using the Brilleman et al. (2018) package. Both models specified the formula:

$$Y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \beta_0 + \beta_1 X_{i1} \tag{2}$$
$$\beta_0 \sim \text{Normal}(50, 5), \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 0.1) \tag{4}$$

Where $Y_i$ is the support rate for Harris or Trump, separately, and $X_{i1}$ is the number of days passed after the latest polling sample of Trump or Harris.

## 3.4 Model Justification and Limitations

We opted to use two separate models for Trump and Harris, as their supporter demographics differ significantly in ways that are challenging to capture within a single or even a small set of variables Center (2024). The polling dataset we obtained primarily reflects the characteristics and trends within individual polls rather than offering direct predictors of the presidential election outcome. To maximize the model's reliability, we refined the dataset by selecting polls with high credibility rather than attempting to predict outcomes directly from the available data.

Our model uses a simple linear approach with time as the sole predictor variable. While this may introduce potential bias by making a simplifying assumption about the trend over time, it also has the advantage of being straightforward to fit and interpret. This trade-off allows us to capture general trends while minimizing the risk of overfitting to limited or potentially inconsistent data sources.

# 4 Results

The results of the Bayesian linear regression models for Kamala Harris and Donald Trump are summarized in Table 2 and Table 3.

Table 2: Summary statistics of the Bayesian model for Harris

| Term | Estimate | Std. Error | 2.5% CI | 97.5% CI |
|---|---|---|---|---|
| (Intercept) | 40.466 | 1.988 | 37.287 | 43.462 |
| days_after_earliest | 0.010 | 0.003 | 0.005 | 0.014 |

Table 3: Summary statistics of the Bayesian model for Trump

| Term | Estimate | Std. Error | 2.5% CI | 97.5% CI |
|---|---|---|---|---|
| (Intercept) | 40.531 | 0.578 | 39.620 | 41.502 |
| days_after_earliest | 0.006 | 0.001 | 0.005 | 0.008 |

## 4.1 Overall Trend

The intercept of both model is positive, with Harris at 40.47 and Trump slightly lower than Harris at 40.36. The trend of support rate for both candidates with respect of time is positive but Harris's increment (0.10) is larger than Trump's (0.06).

## 4.2 Model Diagnostics

The model diagnostics indicate a strong fit and reliable predictive performance for both candidates, as detailed in the plots in the model section Section 3.1. Posterior predictive checks ( Figure 4, Figure 5, Figure 6, and Figure 7) show that the predicted values align well with the observed data, suggesting that the models effectively capture key data patterns. Furthermore, the comparison of posterior and prior variables indicates minimal deviation, affirming an appropriate choice of priors. The credible intervals for the predictors (Figure 8 and Figure 9) encompass the observed support rates, demonstrating that the models adequately account for uncertainty in their predictions.

Additional robustness checks (Figure 10, Figure 10, Figure 12, and Figure 13) confirm the stability of these models. The R-hat values for all parameters are below 1.1, indicating good convergence of the MCMC chains and reliable parameter estimates. This convergence supports the conclusion that the sampling process reached a stable distribution, reinforcing the credibility of our inferences.

In conclusion, these diagnostics verify that the models are both stable and effective in reflecting trends in the data, confirming their reliability for further analysis.

### 4.3 Prediction Visualization

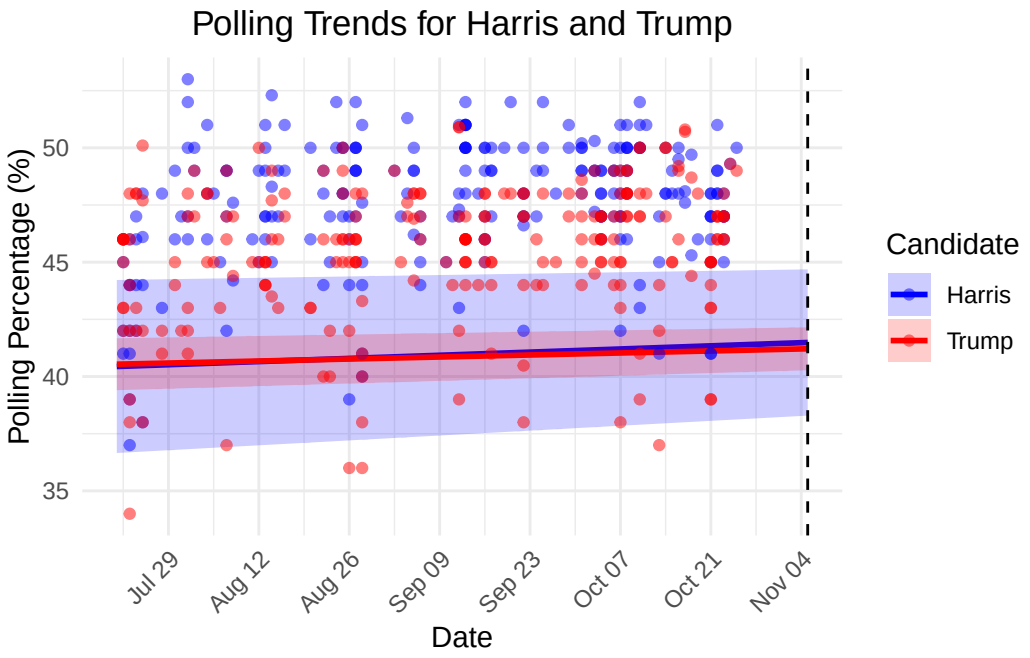


Figure 14: 2024 US Presidential Election Prediction Based Baysian Linear Model

## 5 Discussion

### 5.1 Limitation

Our current model relies solely on national polls with time as the predictor, which gives an overall trend but misses regional dynamics, especially in critical swing states. Including state-level polls and additional factors like state polls and weighting based on numeric grade of pollsters could improve accuracy by capturing localized trends.

The model's assumption of a constant trend over time introduces bias by ignoring natural fluctuations in voter sentiment due to campaign events or media cycles. A time series model, which accounts for these changes, would offer a more accurate prediction, as suggested in Section 5.2.

Additionally, our reliance on "polls of polls" introduces biases stemming from differences in polling methodologies and potential political leanings of individual pollsters. Excluding lower-graded polls further limits our data, potentially overlooking relevant insights. Thus, future improvements should focus on including state-specific data, additional predictors, and a time series approach to enhance predictive accuracy.

## 5.2 Suggestion for Future Step

In terms of suggestion for future step, we should considering using time series as a model for this research. As we use start time and end time for the dataset of pollster, the prediction is continuous on this timeline. As more polls of polls are performed by various pollster, more data points are been collected by Five Thirty Eight website FiveThirtyEight (2024b). By using a time series model, we would be able to use a line of best fit in between all the datapoint collected and using that line to predict the future results. We would also be able to adding the variable needed such as transparency score, pollscore, and numeric grade from the dataset in order to best predict our model.

In addition to the model selection, the dataset provided could also be improved for a future research. The dataset from Five Thirty Eight website FiveThirtyEight (2024b) mainly has variable related to pollsters. We should considering using characteristics about the presidential candidate in order to best predict the result of the election. For instance, the number of campaign done in each state, numbers of supporter in each campaign or various characteristics of candidate on political changes would be more accurate in the predicting support rate. Simply relying on the results of polls of polls will have various limitations since each pollsters' survey has pros and cons. The results of our prediction might be biased if any pollsters heavily rely on certain assumption, not counted in consideration in our model.

# A Appendix

## A.1 Appendix 1 - Pollster Methodology Analysis

This survey was conducted by Redfield & Wilton Strategies to assess the voting intentions of eligible voters in key U.S. swing states ahead of the 2024 Presidential Election. The primary goal of this poll is to provide an accurate and timely snapshot of public opinion in states where electoral outcomes are uncertain and could have a decisive impact on the overall result of the election. Swing states, due to their political volatility and diverse voter bases, are critical in determining the balance of power in the U.S. electoral system. Understanding voter preferences in these states is essential for political analysts, campaigns, and the general public.

### A.1.1 Population of Interest

The population of interest for this survey consists of all eligible voters residing in major U.S. swing states, specifically Arizona, Florida, Georgia, Michigan, North Carolina, and Pennsylvania. These states are known for their fluctuating political alignments and are expected to play a crucial role in the upcoming election.

### A.1.2 Sampling Frame

The population sampled includes eligible voters from Arizona, Florida, Georgia, Michigan, North Carolina, and Pennsylvania. Participants were selected via an online panel.

### A.1.3 Sample

The sample sizes for each state were as follows:

Arizona: 750 respondents

Florida: 1,350 respondents

Georgia: 927 respondents

Michigan: 970 respondents

North Carolina: 880 respondents

Pennsylvania: 1,070 respondents

### A.1.4 Weakness & Strength of the Methodology

In terms of strengths, Redfield & Wilton has a great reputation for producing reliable polling data. According to the dataset given, they have a high pollscore of 0.4, transparency score of 9 and a numeric grade of 1.8. Redfield & Wilton also has the highest amount of polls conducted which makes a great source to use. Redfield & Wilton often target swing states, which makes their polls results important to the US president election.

The weakness of their methodology would incorporate a certain potential bias based on their political leaning of their clients or media. This could cause a certain neutrality in their dataset. In addition, their only methodology is through online panel which do not create a variation in the dataset. Polls that heavily rely on online panels may miss some segments of the population that are less active online.

Overall, Redfield & Wilton has been considered as a competent, reputable and reliable pollster with high amount of polls conducted. Even though the pollster contained a potential bias in their methodology, it has been a reliable resource in prediction of US president election.

## A.2 Appendix 2 - Idealized Methodology and Survey

### A.2.1 Overview

This section introduces an idealized methodology and survey with $100K budget to predict the 2024 US presidential election. The goal is to maximize the accuracy of the prediction under the budget. The subsections that describe the details of the idealized methodology and survey include sampling approach, respondents recruiting method, data validation, poll aggregation and the survey questions.

### A.2.2 Sampling Approach

Cluster sampling is the sampling approach used. Specifically clusters are states. In our case, the swing states as they are the states that would affect the election result, i.e. Arizona, Florida, Gegorgia, Iowa, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. In each swing state, units are selected based on postal code. Using simple random sampling, 100 distinct postal codes are randomly selected. People whose living address have the postal codes selected are the target respondents. For the postal codes that lie in the non-residential area, the postal codes would be ignored.

### A.2.3 Recruitment Strategy

The strategy to be implemented to recruit respondents is a combination of physical and online recruitment.

If the building corresponds to the postal code is a condo building, a paper with QR code to the survey is going to be put in the lobby or elevator, or sent to residents through the property management. If the building is a residential house, then a letter with the QR code would be put in the mailbox.

All respondents are rewarded with a $5 deposit to their bank account or a gift card of their choice. Each IP address is limited to answer the survey once.

### A.2.4 Data Validation

To improve accuracy of the prediction results, the following data validation approaches will be done.

- Postal Code Validation

  - All the postal codes got from respondents are going to be validated to check if it is one of the randomly selected postal codes. If not, the response would not be considered when the prediction is done.

- Just-for-Rewards Prevention
  - To identify the responses that are not seriously answered, the last question of the survey is set test if the respondents are serious and careful when answering the questions.

- Age Validity
  - Responses with age under 18 but answered "Yes" to "Are you registered to Vote" are discarded.

### A.2.5 Poll Aggregation

For each of the swing states, the result would be calculated based on the votes to Trump versus Harris because they are the candidates with the greatest chances of winning. Based on the decisiveness of the respondents and the likelihood of voting from the responses, the individual votes will be weighted when calculating the results for each swing states.

To aggregate the polls for all states, the result of each states is multiplied by its electoral votes. After summing the electoral votes for Trump or Harris, the estimated electoral votes that Trump or Harris get is available. The candidate that has more than 270 electoral votes would be predicted to win the election {U.S. National Archives and Records Administration (2024)}.

### A.2.6 Survey

Google form of the survey is available here: Google Form

**US Presidential Election Survey**
**Purpose:** pre-election survey on public votes distribution.
**Estimated Completion Time:** Less than 5 minutes

All responses are anonymous.

1. What is the postal code of where you live?
2. Are you registered to vote?

- Yes
- No

3. If the election were held today, who would you most likely vote for?

- Donald Trump - Republican
- Kamala Harris - Democrat
- Other Candidates
- Not Decided Yet
- Prefer not to say

4. How decisive are you to vote for the option you chose in the last question?

- Very Decisive
- Pretty Decisive
- A Bit Indecisive
- Very Indecisive

5. How many characters are present in the word “President”?

- 6
- 8
- 10
- None of the Above

6. What age group are you in?

- 0 - 18
- 19 - 30
- 31 - 50
- 51 - 70
- 71+

7. What is your sex?

- Male
- Female
- Non-binary
- Prefer not to say

8. What is your Ethnicity /Race?

- White
- Black or African American
- Hispanic or Latino
- Asian
- Native American or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Middle Eastern or North African

- Prefer not to say

9. What is your Education Level?

- Less than high school
- High school diploma or equivalent
- Some college, no degree
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional or doctoral degree (JD, MD, PhD, etc.)
- Prefer not to say

10.How do you plan to vote?

- In-person on Election Day
- In-person early voting
- Mail-in/Absentee ballot
- Undecided

11. Is there anything else you'd like to share about your voting decision or concerns?

### A.2.7 Budget Specification

- Physical and Online Recruitment: $30,000
- Rewards for Respondents: $50,000
- Data Collection & Validation: $10,000
- Other: $10,000

Total: $100,000

### A.2.8 Conclusion

The survey uses cluster sampling to sample the swing states based on postal codes. After data collection and validation, weighting based on the likelihood of voting and decisiveness of the respondents, the total votes for Donald Trump and Kamala Harris are calulated. The candidate with more than 270 votes are predicted to win the election{U.S. National Archives and Records Administration (2024)}.

## References


Andrew, Gabriel, Jonah Gabry, Ben Goodrich, Sam Brilleman, et al. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Center, Pew Research. 2024. "The Harris-Trump Matchup." https://www.pewresearch.org/politics/2024/10/10/the-harris-trump-matchup/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024b. "Dataset: US Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

———. 2024a. "Dataset: US Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Gabry, Jonah et al. 2023. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

Grolemund, Garrett, and Hadley Wickham. 2011. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Kassambara, Alboukadel. 2023. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal et al. 2023. *Arrow: Integration to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

Science, Towards Data. n.d. "Introduction to Bayesian Linear Regression." https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7.

Time. 2024. "Key Dates for the 2024 u.s. Presidential Election." https://time.com/7010964/2024-election-calendar-dates/.

U.S. National Archives and Records Administration. 2024. "About the Electoral College." 2024. https://www.archives.gov/electoral-college/about.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley et al. 2023. *The Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

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

———. 2023b. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https:

//CRAN.R-project.org/package=scales.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.