Prediction and analysis of 2024 US presidential election*

Forecasting Election Outcomes: A Logistic Regression Approach to Voter Support

Zilin Liu

November 4, 2024

This paper uses the data of the 2024 American election to make a prediction, and establishes a logistic regression model to discuss the election probability of Donald Trump. In this paper, states, sample size, start date and transparency are used as predictive variables to predict whether Donald Trump will be elected president in the future. This paper also expounds the advantages and limitations of statistical modeling in election data, and fills the gap of the influence of different characteristics on election voting.

Table of contents

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

1 Introduction

The American election, as one of the most remarkable political events in the world, not only determines the future direction of the United States, but also has a far-reaching impact on international relations and global policies. In this election, Donald Trump and Kamala Harris are the two main candidates, and their political ideas, policy propositions and personal charm will all become important factors affecting voters' decision-making. Through in-depth analysis of the 2024 presidential election in the United States, this paper aims to provide readers with a clear election picture. This paper tries to fill this gap in the academic field by collecting and analyzing the latest poll data and analyzing the characteristics of the States, samples size, start date and transparency score. Through this study, I hope readers can judge from the characteristics of states and so on. This estimate is the difference in the voting rate of Donald Trump in eight different States and other predictors.The estimated values in our logistics regression analysis are the regression coefficient of each predictor-voting state, the regression coefficient of logarithmic sample size, the regression coefficient of survey start date and the regression coefficient of transparency score. These coefficients represent the estimated influence of each predictor on the voted of successful election, while other predictors keep fixed.

The remainder of this paper is structured as follows. Section 2 introduces the data sources, preprocessing procedures, variable selection standards, and possible correlations between independent and dependent variables. Section 3 succinctly the hierarchical Bayesian model explain. Section 4 along with an overview of the model's output to bolster its accuracy and

forecast Trump's support rate in various geographic areas on November 5, 2024 and shows the result of state's support rate. Section 5 use the established logistics regression model to predict Trump's support rate. Section 6 talks about mitations of the data and polls, the difficulty of converting our real world into data, and future directions.

2 Data

2.1 Overview

The data used in this paper comes from the survey data of FiveThirtyEight (FiveThirtyEight (2024)) Presidential general election polls (2024). Statistical programming language R is used to retrieve (R Core Team 2023), clean and process data. In particular, the following R packages are used in this report analysis: `tidyverse` (Wickham et al. 2019) is used for data cleaning and processing, `ggplot2` (Wickham 2016) is used for data visualization, `modelsummary` (Arel-Bundock 2022) is used for predicting the output of model data, and `rstanarm` (Goodrich et al. 2022) is used for constructing Bayesian prediction model. Other libraries that supported the data analysis include `knitr` (Xie 2023), `tibble`(Müller and Wickham 2023), `readr`(Wickham, Hester, and Bryan 2024), `collapse` (Krantz 2024), `arrow`(Richardson et al. 2024), `dplyr`(Wickham et al. 2023) and `marginaleffects`(Arel-Bundock, Greifer, and Heiss Forthcoming). Additionally, Professor Rohan Alexander of the University of Toronto (Alexander 2024) provided the folder structure for this analysis.

2.2 Measurement

The data set of this analysis is the data about polls from FiveThirtyEight website (FiveThirtyEight (2024)), which is a source with reliable data quality and good reputation.

This data focus on the importance of different variables used to measure public opinions and voting quality, especially in the context of political polls. Pct (Support Rate): This is the main variable in the data set, representing the support rate of each candidate. It is used as an indicator of public opinion in every poll. The assumption here is that the sample population can accurately reflect the wider voting population, and the higher the percentage, the more popular the candidate is. Generally, more than 50% of the candidates are supported. Sample size: The data set includes the variable sample size, which is very important for the representation of the polls. Generally speaking, the larger the sample size, the more reliable the percentage support rate is, thus reducing the error range. Transparency_score: This variable measures the transparency of each poll when its methodology is made public. The higher the score, the greater the reliability, which means the less the possibility of bias when the poll is introduced into the analysis.

Although data sets provide a comprehensive view of public opinion trends, there are some potential limitations that may affect the validity and reliability of variables. These limitations

include: Differences in survey methods: Different survey methods may have different effects on the results. Regional deviation in the sample: If the sample is not evenly distributed geographically, it may lead to regional deviation and affect the representation of the results. Variability of public opinions: public opinions change with time, which may affect the stability and forecasting ability of data. Small measurement errors are acceptable and have little impact on this analysis.

2.3 Analysis Data

This analysis data set is mainly an analysis of Donald Trump's vote rate in the 2024 US presidential election. The primary research goal of this project analysis is to predict whether Trump is likely to win in the 2024 general election, so the vote rate is our main indicator for modeling and forecasting. The vote rate represents whether the proportion of respondents who support Trump in a specific poll exceeds 50%. This analysis data set uses only polls with a numeric_grade higher than 2.5 and data with a starting date greater than July 1, 2024, which ensures the reliability and timeliness of the analysis data. We will explore how the vote rate is affected by different factors, such as state, time or transparency, and sample size. By simulating the fluctuation of Trump's support rate, we can infer his election performance, analyze the trend within his support base and analyze the influence of various factors on his popularity.

2.4 Outcome variables

There are many variables in the original data. This case, some variables will be selected for prediction analysis. The outcome variable of this analysis is vote, and only the date of candidate Donald Trump is kept. Binary logistics regression analysis is used to predict the support rate of Donald Trump. A new variable vote is derived from this report. When pct is greater than or equal to 50%, it is defined as yes (is_vote=1), otherwise it is NO (is_vote=0). In this study, it is considered that Trump will win the election if the vote rate is greater than 50%.

2.5 Predictor variables

This paper also selected the following variables as predictor variables for analysis.

State: for the variable state, eight states with more data are selected, namely "Georgia","Michigan","Pennsylvania","Arizona","North Carolina","Wisconsin","Texas" and "Florida".As we all know, some swing States have not yet determined the support rate of candidates, but most States have almost locked in candidates.

sample size: The larger the sample size, the more reliable the data is. In order to solve the variance problem and meet the assumptions of the regression model, the sample size is logarithmic transformed in the analysis.

Starting time, this data analysis believes that Trump's support rate will change with the growth of the survey time, and the survey date will start from July 1, 2024.

Transparency grade: Research shows that the transparency of polling methods is very important for the credibility and reliability of polling results. The higher the score, the greater the reliability, which means the less the possibility of bias when the poll is introduced into the analysis.

Table 1: Top 10 Records of Trump's Poll Data

pct	state	start_date	transparency_score	vote	log_sample_size
50.6	Arizona	2024-10-25	6	yes	7.3
50.8	Arizona	2024-10-25	6	yes	7.3
50.2	Georgia	2024-10-25	6	yes	7.3
50.9	Georgia	2024-10-25	6	yes	7.3
49.2	Michigan	2024-10-25	6	no	6.9
49.3	Michigan	2024-10-25	6	no	6.9
48.4	North Carolina	2024-10-25	6	no	7.4
48.4	North Carolina	2024-10-25	6	no	7.4
48.7	Pennsylvania	2024-10-25	6	no	7.2
49.6	Pennsylvania	2024-10-25	6	no	7.2

2.6 Visual analysis

The following figures, Figure 1, Figure 2, display various aspects of data related to Trump's support rate.

Figure 1 shows the data of American elections in eight States in 2024. It can be seen that Texas and Florida support Trump more, while Pennsylvania and Wisconsin support Trump less.

Figure 2 shows the changes in the support rate of Trump's election in eight American states from July to November, 2024. From the four figures, we can see that in Texas and Florida, the support rate of most surveys is above 50%, while in other States, the support rate has obviously increased with time, all of which have increased to about 50%. Through the support rate trend chart of Trump's eight States, it is found that Trump's support rate is rising near the US election.

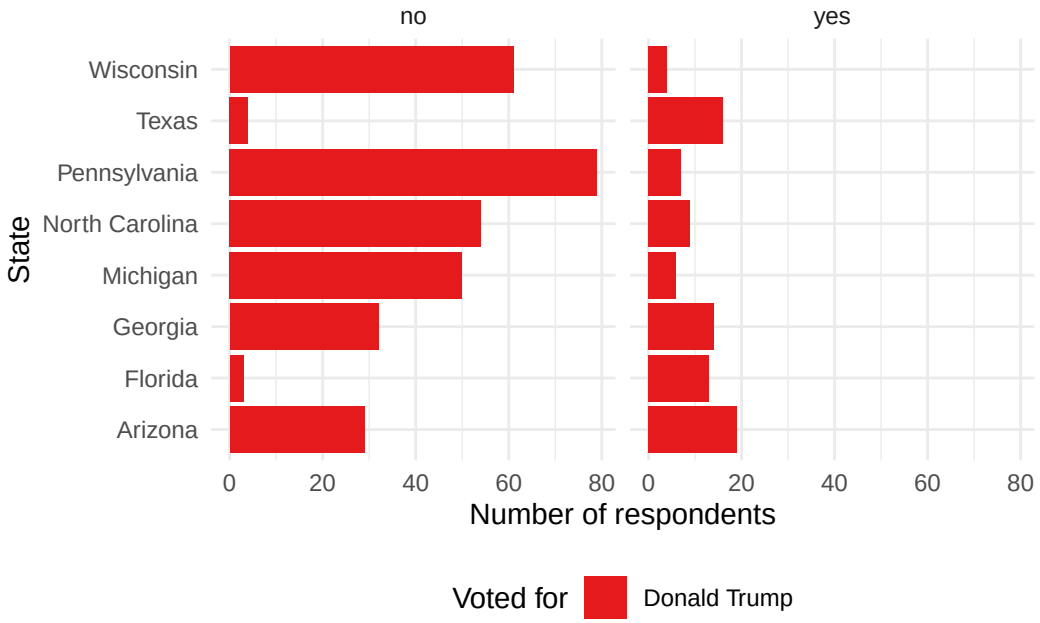

Figure 1: Support Rates for Donald Trump of Eight-state

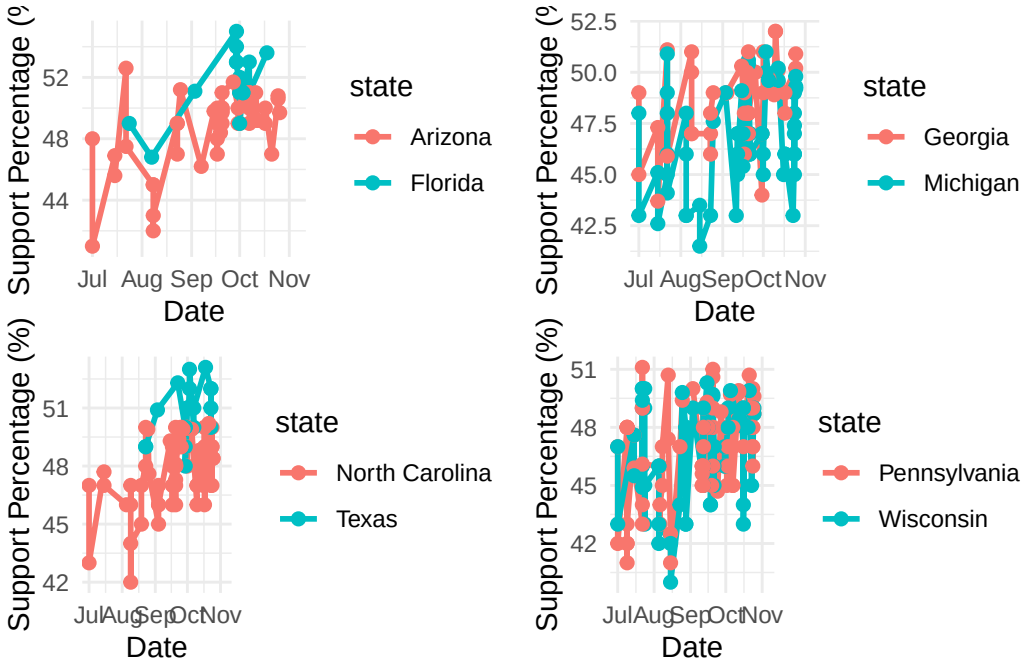

Figure 2: Support Percentage Over Time by State

3 Model

My modeling goal is to establish a logistic regression model by using logistic distribution, because the dependent variable is a binary variable, which is used to predict the successful election rate of Trump.

3.1 Model set-up

Define yi as the whether to support Trump as president (pct >=50) and the explanatory variable as the state, sample size, start date and transparency.

$$y_i|\pi_i \sim \text{Bern}(\pi_i)$$

Formula:

$$\text{logit}(y_i) = \beta_0 + \beta_1 \times \text{state} + \beta_2 \times \text{start date} + \beta_3 \times \log \text{sample size} + \beta_4 \times \text{transparency score}$$

where:

- y_i is a binary variable, whether the election is successful or not.
- $\beta_0 \sim \text{Normal}(0, 2.5)$ is the global intercept for Trump support.
- $\beta_1 \sim \text{Normal}(0, 2.5)$ represents the effect of state on Trump support.
- $\beta_2 \sim \text{Normal}(0, 2.5)$ represents the effect of start date.
- $\beta_3 \sim \text{Normal}(0, 2.5)$ represents the effect of sample size.
- $\beta_4 \sim \text{Normal}(0, 2.5)$ represents the effect of transparency score.

We run the model in R (R Core Team 2023) using the `rstanarm` package of (Goodrich et al. 2022). We use the default priors from `rstanarm` (Goodrich et al. 2022).

3.1.1 Model justification

According to the state data and the voting rates of Harris and Trump, this analysis shows that Michigan and Pennsylvania support Harris more, while Georgia and Arizona have lower voting rates for Harris.

Table 2

Table 3: Logistic regression analysis Voted for Donald Trump

	Donald Trump
(Intercept)	−206.689
	(117.609)
stateFlorida	2.040
	(0.765)
stateGeorgia	−0.540
	(0.461)
stateMichigan	−1.863
	(0.549)
stateNorth Carolina	−1.630
	(0.491)
statePennsylvania	−2.334
	(0.523)
stateTexas	1.632
	(0.678)
stateWisconsin	−2.354
	(0.604)
start_date	0.010
	(0.006)
log_sample_size	1.295
	(0.446)
transparency_score	−0.124
	(0.118)
Num.Obs.	400
R2	0.338
Log.Lik.	−152.315
ELPD	−164.2
ELPD s.e.	13.2
LOOIC	328.4
LOOIC s.e.	26.4
WAIC	328.2
RMSE	0.35

4 Results

Logistics regression results Table 2 show that Florida and Texas are Trump's staunch support states, while the support of the other six states is relatively low. The influence of the start time of the survey on Trump's support rate is positive. As the time of the US election approaches, Trump's support rate gradually increases. The influence of sample size on Trump's support rate is positive. With the increase of sample size, Trump's support rate gradually increases. The transparency score of the survey has no significant impact on Trump's support rate. The goodness of fit of this logistics regression model is 36.6% in R square and 0.34 in RMSE, which shows that the model has the possibility of further improvement.

5 Prediction

Predicting Trump's support rate, use the established logistics regression model to predict, create predictive variables, select eight states as the predictive variables of states, select November May, 2024 as the starting date, use the average sample size of the analysis data set as simple size, and select the average transparency score of the analysis data set as transparency score. Through the prediction analysis (Table 3), it is concluded that the support rate of three States is more than 50%, namely Arizona, Florida and Texas. The support rate of other States is less than 0.5. This project can use the same logistic model to predict all States in the United States and get the final prediction probability.

Table 4: Prediction for Trump's support rate

Estimate	2.5% CI	97.5% CI	State
0.529	0.326	0.723	Arizona
0.395	0.212	0.614	Georgia
0.150	0.053	0.311	Michigan
0.181	0.081	0.339	North Carolina
0.099	0.035	0.231	Pennsylvania
0.096	0.027	0.242	Wisconsin
0.850	0.636	0.957	Texas
0.896	0.688	0.980	Florida

6 Discussion

6.1 Variations in State-Level Support Rates

In this study, we analyzed Donald Trump's election situation in the 2024 U.S. presidential election, and the results show significant differences in support rates across various states. Specifically, in Florida and Texas, Trump's support rate is noticeably higher than in other states. These differences may be influenced by a variety of factors, including local economic conditions, social issues, and the intensity of campaign activities by the candidates. By discussing these factors, we gain a better understanding of how state-level support rate differences are formed. This finding highlights the importance of regional variation in elections and suggests that state-level data are essential for improving the accuracy of election prediction models.

6.2 Impact of Poll Sample Size and Timing on Support Rates

The study also found that the timing and sample size of polls have a significant impact on Trump's support rate. The analysis indicates that the later a poll begins, the higher Trump's support rate, possibly because voter positions become more definite as the election approaches. Additionally, larger sample sizes correlate with higher support rates for Trump, which might be because polls with larger samples can more accurately reflect public opinion. This finding emphasizes the need to consider the effects of sample size and timing when using polling data to ensure the accuracy of predictions. By controlling for these variables, we can enhance the robustness and predictive power of the model.

6.3 Model Limitations and Directions for Future Improvement

While the linear regression model in this study provides valuable insights into predicting Trump's support rate, it still has limitations. For example, the model does not fully cover data from all states, which may affect the comprehensiveness of the predictions. Additionally, the model lacks some potential explanatory variables, which could impact the thoroughness of voter behavior analysis. Future research could improve the model's predictive accuracy by incorporating more representative explanatory variables, such as voter age and income levels. Further model improvements could also involve incorporating nonlinear or machine learning algorithms to better capture complex voter preference changes, providing a stronger foundation for future political analysis and prediction.

A Appendix: TIPP's methodology

A.1 Trend In Politics and Policy

TIPP (TechnoMetrica Institute of Policy and Politics) specializes in investigating and analyzing political, social, and economic trends in the United States. Known for its in-depth polling on public sentiment, TIPP frequently utilizes online panels and mixed-mode survey methods, including phone and internet-based surveys, to gather data that reflects the current opinions of the American populace (Blumenthal et al. 2010). By focusing on public opinions, policy shifts, and election dynamics, TIPP aims to provide comprehensive insights that capture the pulse of the nation. Transparency is a central tenet of TIPP's methodology; the organization emphasizes clear reporting on sample selection, survey design, and result interpretation, which enhances the reliability and validity of its findings (Pew Research Center 2022).

TIPP typically targets the U.S. adult population or registered voters, recruiting participants through online advertisements, email invitations, and social media platforms. This approach broadens its reach, enabling it to access a diverse pool of respondents. To ensure representation, TIPP often employs probabilistic sampling methods, such as random sampling, which provides each respondent with an equal chance of being selected. This probabilistic approach mitigates selection bias, making TIPP's findings more generalizable to the broader population.

The use of online surveys also allows TIPP to distribute and collect questionnaires swiftly, providing researchers with large datasets in a short time frame. This flexibility is particularly valuable in rapidly changing political and social climates, where timely data can reveal shifts in public opinion or emerging trends. Moreover, online surveys are cost-effective and scalable, allowing TIPP to conduct frequent polls without incurring the high costs associated with traditional survey methods (Couper, 2000). This adaptability makes TIPP's data collection approach well-suited for analyzing evolving political and social landscapes, offering policymakers and analysts timely insights into public opinion.

A.2 Population, Frame, and Sample of the Poll

Target population: -Registered voters in the United States.

Sampling method: Stratified random sampling. -Divide the population into different layers (for example, according to geographical area, age, gender, race and political orientation), and then randomly select samples from each layer.

Sample size: -According to the budget and the required error range, determine the appropriate sample size. Generally speaking, the larger the sample, the higher the reliability of the results.

Recruitment method: Email invitation. -Use the voter registration database to send email invitations to potential interviewees. Social media advertising: Put targeted advertisements on Facebook, Twitter and other platforms to attract respondents to participate in the survey.

Online market research platform: Cooperate with online market research platforms such as Google Forms to recruit respondents.

Incentive measures: Provide small cash rewards or gift cards to improve the response rate.

Survey tools: Use online survey software (such as Google Forms and Qualtrics) to collect data.

A.3 Advantages and Trade-offs in TIPP's Online Sampling Approach

TIPP's use of online and digital recruitment channels has significant advantages. Primarily, this approach allows for cost-effective and rapid data collection, enabling TIPP to produce timely insights on public opinion as social and political landscapes shift. By leveraging online panels and digital recruitment, TIPP reduces coverage bias, as individuals with access to the internet, regardless of their geographical location, can participate. This inclusive approach captures the perspectives of tech-savvy individuals and mobile-first users who may otherwise be overlooked by traditional survey methods.

However, there are trade-offs associated with this method. The reliance on online recruitment can introduce selection bias, as those who participate are generally more comfortable with technology and may not be representative of the entire electorate. Additionally, online panel participants may exhibit "professional respondent" behavior, where frequent survey-takers may skew results due to their familiarity with survey formats. Such biases can impact the representativeness of the results and require careful adjustment through weighting techniques.

A.4 Questionnaire Design

TIPP's questionnaire design emphasizes clarity and brevity. Questions are crafted to be straightforward and free from technical jargon, enhancing respondent understanding and reducing the likelihood of response errors. This design strategy is particularly effective in retaining respondents' engagement throughout the survey. Moreover, TIPP's focus on current, high-interest topics, such as political preferences and key social issues, ensures that the survey content remains relevant to respondents' lives.

However, one limitation of TIPP's questionnaire design is its relatively narrow focus, as it often captures only the surface-level opinions of respondents without delving into the underlying motivations for their views. Additionally, since TIPP employs multiple survey channels, respondent engagement levels may vary across modes. For instance, respondents recruited through social media may interact differently with the survey compared to those from online panels, potentially influencing the depth and accuracy of their responses.

A.5 Conclusion

TIPP's online sampling and recruitment approach enables the organization to balance cost-effectiveness with comprehensive demographic reach. By combining online panels, social media outreach, and robust weighting systems, TIPP enhances the representativeness and reliability of its data. While this methodology effectively reduces coverage bias, potential challenges—such as selection bias from frequent online survey participants and limited engagement depth due to mode-specific interactions—must be taken into account during analysis. These considerations help ensure that TIPP's survey findings remain relevant and accurate in capturing the dynamics of public opinion.

B Methodology and Survey Design

To investigate voter support patterns in the 2024 U.S. presidential election, we designed a methodology that integrates both probabilistic sampling and precise demographic stratification. This approach allows us to capture reliable insights across diverse voter segments and geographic distributions. Utilizing stratified random sampling, our survey divides the sample into key demographic segments—such as age, education, and regional representation—enhancing accuracy and reducing sampling bias, as recommended by election research literatured on both registered voters and those categorized as likely voters, providing a balance between broad societal attitudes and a focus on election outcomes. The survey was structured using digital platforms like Google Forms, ensuring accessibility and ease of use for respondents. The choice of an online platform supports efficient data collection and offers scalability within the constraints of our $100,000 budget.

The survey was carefully designed to mitigate non-response bias, with respondent recruitment achieved through targeted social media ads and email outreach. Small monetary incentives were provided to participants to encourage a higher response rate and mitigate potential attrition issues, reflecting best practices for survey engagement in electoral studies .

Data quaorously maintained through a multi-stage validation process. Responses were cross-verified with public voter files, and the dataset was cleansed to exclude incomplete entries. Following data collection, we applied weighting adjustments to align the sample structure with national voter demographics, ensuring representativeness in critical variables such as region and age distribution .

Budget allocation to support the core research objectives: respondent incentives and recruitment accounted for approximately half of the total budget, with the remainder distributed across data processing, modeling efforts, and operational costs including software and survey platform fees.

Budget Allocation:

-**Respondent Recruitment and Incentives**: $50,000

-**Data Processing and Validation**: $15,000

-**Survey Platform and Operational Costs**: $10,000

-**Modeling and Data Analysis**: $20,000

-**Contingency and Miscellaneous Expenses**: $5,000

A link to the survey can be found at: https://forms.gle/SXX3jEZw7ivrC7Sg8

References

Alexander, Rohan. 2024. "Starter Folder." https://github.com/RohanAlexander/starter_folder.git.

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.

Arel-Bundock, Vincent, Noah Greifer, and Andrew Heiss. Forthcoming. "How to Interpret Statistical Models Using `marginaleffects` in `R` and `Python`." *Journal of Statistical Software*, Forthcoming. https://marginaleffects.com.

Blumenthal, Mark, Charles Franklin, Nate Silver, Scott Keeter, and Andrew Gelman. 2010. "Symposium: Polls, Forecasts, and Aggregators." *Public Opinion Quarterly* 74 (5): 845–65. https://doi.org/10.1093/poq/nfq076.

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

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Krantz, Sebastian. 2024. *Collapse: Advanced and Fast Data Transformation in r*. https://doi.org/10.5281/zenodo.8433090.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Pew Research Center. 2022. "Data from Pew Research Center Advance Knowledge in Public Opinion Research (Strategies for Detecting Insincere Respondents in Online Polling)." *NewsRX LLC*, 80–80. https://global-factiva-com.myaccess.library.utoronto.ca/sb/default.aspx.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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

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

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://cran.r-project.org/web/packages/knitr/index.html.