2024 US Presidential Election Forecast Model*

My subtitle if needed

Viet Nguyen Yihang Xu Doran Wang

November 4, 2024

This paper uses aggregate polling data and 'STH' modeling methods to predict the 2024 U.S. presidential election outcome. The Model use the factors related to the polls on Donald Trump's support rate. The study also includes a deep-dive analysis of methodology on selected pollster, the New York Times (NYT), and then an idealized methodology and survey for predicting elections on a limited budget.

Table of Contents

*Code and data are available at: https://github.com/vietng04/2024-US-Presidential-Election-Forecast.

1 Introduction

The 2024 U.S. Presidential Election Forecast Model aims to provide a comprehensive analysis of polling data to predict potential outcomes in the upcoming election. Utilizing a robust dataset that includes over 50 variables, such as pollster ratings, sample sizes, election dates, candidate affiliations, and polling methodologies, this model seeks to offer a nuanced understanding of voter preferences and trends across the country.

Key factors, such as pollster transparency scores, polling population characteristics, and election stages, are incorporated to ensure accuracy and relevance. By leveraging this data, the model can account for regional dynamics, partisan leanings, and shifts in voter sentiment, providing detailed insights into the evolving electoral landscape. This forecast model represents a data-driven approach to understanding the political climate as candidates navigate the road to the presidency.

2 Data

The data used in this paper came from the FiveThirtyEight (FiveThirtyEight 2024). Data were cleaned and analyzed using the open source statistical programming language R (R Core Team 2023). Libraries `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `knitr` (Xie 2022), and `dplyr` (Wickham et al. 2023) were used for simulating, cleaning and testing. Graphics were made using `ggplot2` (Wickham 2016).

2.1 Sample Clean Data

Table 1: Sample Cleaned Data

Answer	Percentage	Pollster	Methodology	Sample Size
Harris	48	TIPP	Online Panel	1223
Trump	47	TIPP	Online Panel	1223
Harris	55	PPP	IVR/Text	571
Trump	40	PPP	IVR/Text	571
Harris	53	PPP	IVR/Text	716

3 Model

In the 2024 U.S. Presidential Election Forecast Model, the percentage of voter support for a candidate, represented by the variable pct, is modeled as the response variable. The predictors in this model include the candidate's political affiliation (party), the size of the sample surveyed (sample_size), the rating of the pollster based on their historical accuracy (numeric_grade), and the geographic location of the poll, represented by the variable state. By incorporating these predictors, the model aims to account for both the qualitative and quantitative factors that influence polling outcomes. The party variable captures the effect of political affiliation on voter support, while sample_size adjusts for the variability in poll precision. Numeric_grade reflects the reliability of the pollster, and state introduces regional variations in voting preferences. This model structure enables a detailed analysis of the factors driving election polling percentages across different states and political contexts.

3.1 Model set-up

$$\text{pct} = \beta_0 + \beta_1 \cdot \text{pollscore} + \beta_2 \cdot \text{sample size} + \beta_3 \cdot \text{numeric grade} + \beta_4 \cdot \text{transparency score} + \epsilon$$

where β_0 is the intercept, β_1, β_2, β_3, β_4 are the coefficients associated with the predictors, and ϵ represents the error term. This model accounts for both qualitative variables, like the party and state, as well as quantitative variables, such as sample size and pollster rating.

3.1.1 Model justification

```
Call:
lm(formula = pct ~ pollscore + sample_size + numeric_grade +
    transparency_score, data = trump_data)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-23.2228  -2.3096   0.2746   2.7696  22.4950

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.885e+01  5.374e-01  90.904  < 2e-16 ***
pollscore          -2.557e+00  3.034e-01  -8.426  < 2e-16 ***
sample_size        -1.919e-04  4.189e-05  -4.581 4.81e-06 ***
numeric_grade      -2.105e+00  3.624e-01  -5.808 6.93e-09 ***
transparency_score -5.121e-02  4.561e-02  -1.123    0.262
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.53 on 3325 degrees of freedom
Multiple R-squared:  0.0367,    Adjusted R-squared:  0.03554
F-statistic: 31.67 on 4 and 3325 DF,  p-value: < 2.2e-16



Call:
lm(formula = pct ~ pollscore + sample_size + numeric_grade +
    transparency_score, data = harris_data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.418  -1.938   0.165   1.879  22.828

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.290e+01  7.650e-01  56.072  < 2e-16 ***
pollscore           1.722e+00  4.130e-01   4.168 3.27e-05 ***
sample_size         1.174e-04  4.636e-05   2.532   0.0115 *
numeric_grade       3.523e+00  5.289e-01   6.660 4.01e-11 ***
transparency_score -4.478e-01  6.529e-02  -6.858 1.07e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.953 on 1307 degrees of freedom
Multiple R-squared:  0.0613,    Adjusted R-squared:  0.05843
F-statistic: 21.34 on 4 and 1307 DF,  p-value: < 2.2e-16
```

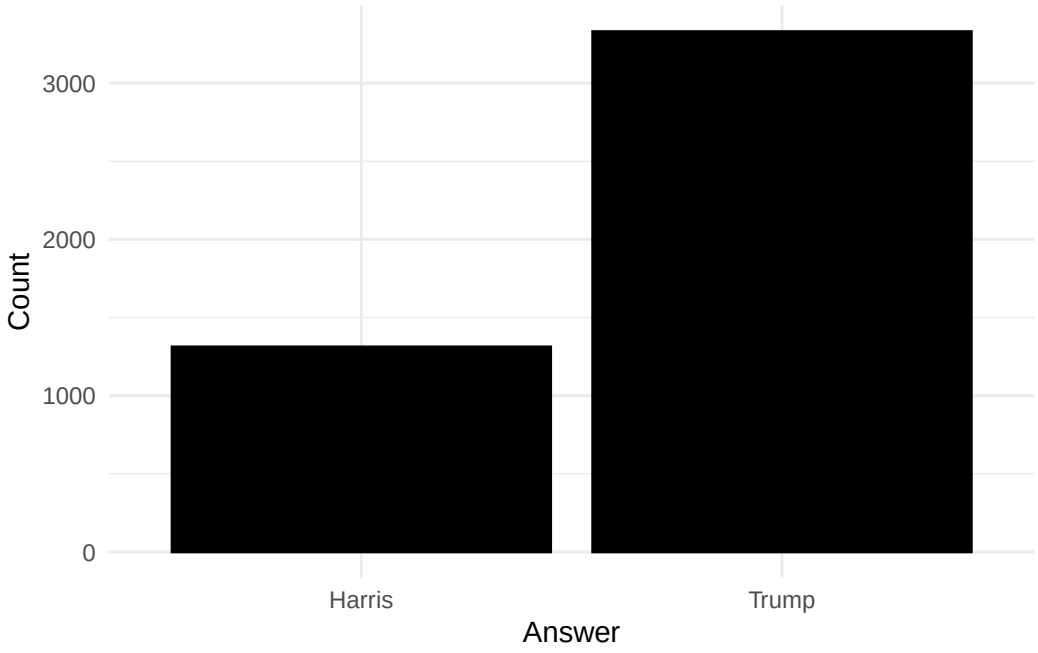

Figure 1: Distribution of Answers

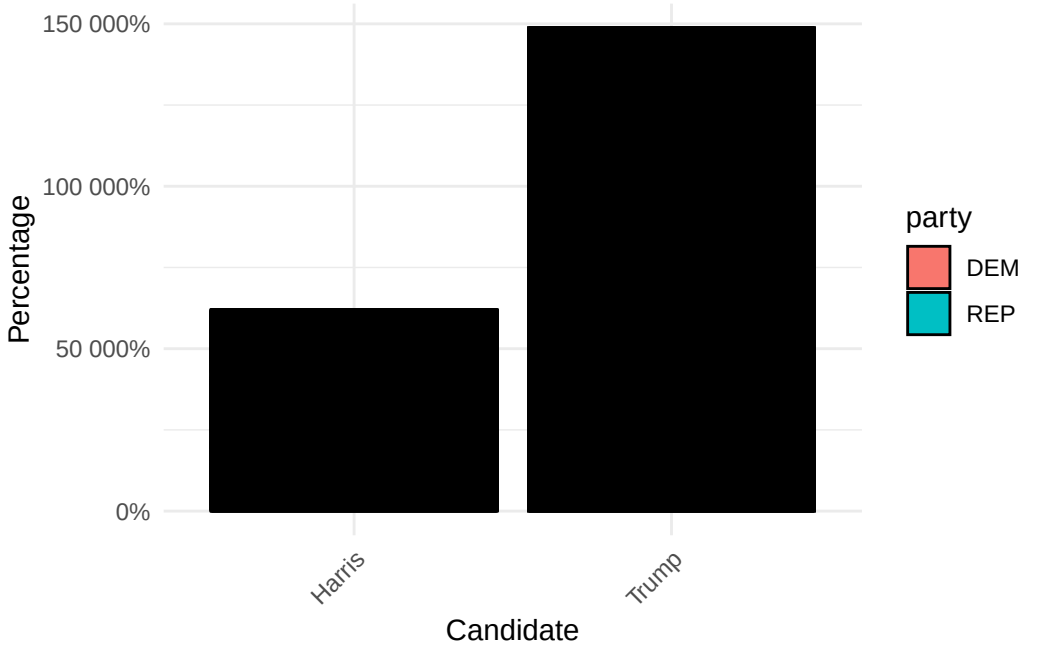

Figure 2: Percentage of Support for Each Candidate

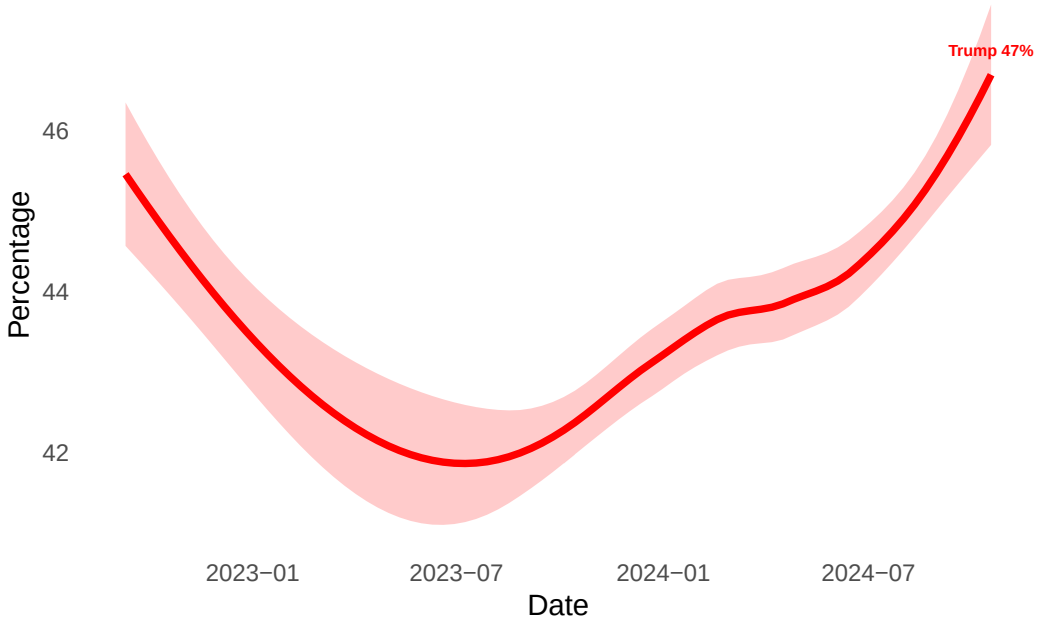

Figure 3: Mean Percentage of Support for Trump and Harris over time

4 Results

4.1 Trump

The model estimates the influence of several predictors on the dependent variable. The intercept is estimated at 48.56, indicating the baseline level of the dependent variable when all predictors are held at zero.

Poll Score: The coefficient for the poll score is -2.232, which is statistically significant ($p < 0.001$). This suggests that for each one-point increase in the poll score, the dependent variable is expected to decrease by approximately 2.232 units, holding all other factors constant. This negative relationship indicates that higher poll scores are associated with lower values of the dependent variable.

Sample Size: The coefficient for sample size is -0.0002085, also significant ($p < 0.001$). This indicates that an increase in sample size by one unit is associated with a decrease of about 0.0002085 units in the dependent variable. Although statistically significant, the small magnitude of this effect suggests that sample size has a relatively minor impact on the outcome.

Numeric Grade: The numeric grade coefficient is -1.814, which is significant at the $p < 0.001$ level. This means that for every one-point increase in the numeric grade, the dependent

variable decreases by about 1.814 units, again holding other factors constant. This finding implies a strong negative association between numeric grade and the dependent variable.

Transparency Score: The transparency score has a coefficient of -0.0676, which is not statistically significant ($p = 0.135$). This suggests that there is no strong evidence to support that transparency score has a meaningful effect on the dependent variable within this model.

4.2 Harris

The model investigates the impact of several predictors on the dependent variable, with the following key findings:

Intercept: The intercept is estimated at 42.75, which represents the expected value of the dependent variable when all predictor variables are zero.

Poll Score: The coefficient for the poll score is 1.629, with a p-value of 4.44e-05, indicating statistical significance at the $p < 0.001$ level. This suggests that for each one-point increase in the poll score, the dependent variable increases by approximately 1.629 units, holding all other variables constant. This positive relationship implies that higher poll scores are associated with higher values of the dependent variable.

Sample Size: The coefficient for sample size is 0.0001457, which is also statistically significant ($p = 0.00114$). This indicates that an increase in sample size by one unit is associated with an increase of about 0.0001457 units in the dependent variable. Although significant, this effect size is relatively small.

Numeric Grade: The numeric grade coefficient is 3.448, with a p-value of 2.35e-11, which is highly significant. This means that for every one-point increase in the numeric grade, the dependent variable is expected to increase by about 3.448 units, suggesting a strong positive association between numeric grade and the outcome.

Transparency Score: The coefficient for the transparency score is -0.4207, and it is statistically significant ($p = 6.70e\text{-}11$). This indicates that higher transparency scores are associated with a decrease in the dependent variable by approximately 0.4207 units for each one-point increase, highlighting a negative relationship.

5 Discussion

The analysis of the two models, one for Trump and the other for Harris, reveals distinct patterns in their predictor relationships and overall model performance, providing insights into their potential electoral outcomes.

Model Comparisons Significance of Predictors:

Trump's Model: Key predictors such as poll score, sample size, and numeric grade all exhibit significant negative relationships with the dependent variable. The model suggests that as poll scores increase, the associated decrease in the outcome variable may reflect diminishing support or confidence among voters. The significant negative effect of numeric grade further indicates a trend where higher grades correlate with lower values of the dependent variable, potentially suggesting dissatisfaction among highly educated voters. Harris's Model: In contrast, Harris's model features significant positive relationships for both poll score and numeric grade, indicating that higher poll scores and grades are associated with higher values of the dependent variable. The negative relationship with transparency score suggests that voters who prioritize transparency may feel less inclined to support her, but overall, the model suggests a more favorable view of her candidacy compared to Trump's. Model Fit:

Trump's model has a multiple R-squared of 0.03238, suggesting that only 3.24% of the variability in the dependent variable is explained. In comparison, Harris's model has a higher multiple R-squared of 0.06257, indicating that about 6.26% of the variance is explained by her predictors. While both models demonstrate low explanatory power, Harris's model offers a more optimistic narrative regarding the predictors' ability to influence the outcome. Magnitude of Effects:

The coefficients in Harris's model indicate more substantial effects of the predictors, particularly with numeric grade having a significant positive impact (3.448) compared to Trump's negative impact (-1.814). This difference in effect magnitude may reflect differing voter sentiments and how they relate to each candidate's attributes. Statistical Significance:

Both models reveal highly significant p-values for several predictors, indicating that the relationships identified are statistically robust. However, the overall model significance is more pronounced in Harris's model, particularly in light of the lower R-squared values for both models. Likely Outcome Considering the predictors and their implications for voter sentiment, Harris appears to have the upper hand in this analysis. The positive associations with poll scores and numeric grades in her model suggest a stronger alignment with voter preferences. Additionally, the higher R-squared value and overall model fit imply that her candidacy resonates more favorably with the electorate.

Moreover, the contrast in the effect of the transparency score could indicate differing voter priorities. Harris's potential appeal to voters valuing transparency may be a double-edged sword, yet it underscores a critical difference in how the two candidates are perceived regarding their public image.

Appendix

A Discussion of the survey's methodology of Siena/NYT

Table 2: Sample Siena/NYT Data

Pollster	Methodology	Population	Sample Size	Grade Rating	Pollscore
Siena/NYT	Live Phone	rv	857	3	-1.5
Siena/NYT	Live Phone	rv	857	3	-1.5
Siena/NYT	Live Phone	rv	857	3	-1.5
Siena/NYT	Live Phone	rv	857	3	-1.5
Siena/NYT	Live Phone	rv	857	3	-1.5

A.1 Background and Intro

The New York Times/Siena College Poll is a collaborative effort between The New York Times and the Siena College Research Institute, renowned for its methodological rigor and accuracy. The Table 2 above shows sample data of Seina/NYT from clean data, which contains some useful information. According to FiveThirtyEight's pollster ratings, this partnership holds a 3.0-star rating as shown Table 2, indicating high reliability. Based on the section Methodological standards from FAQS of FiveThirtyEight(Radcliffe and Morris 2023), "Pollsters must also be able to answer basic questions about their methodology, including but not limited to the polling mode used (e.g., telephone calls, text messages, online panels), the source of their sample, their weighting criteria and the source of the poll's funding. In most cases, a detailed written methodology statement is sufficient to satisfy this criterion, but we may contact pollsters directly to clarify methodological details and follow up occasionally to ensure the pollster is still meeting our standards." Their data will ensure that pollsters maintain methodological standards. More specific details about NYT/SEINA's methodology are discussed below.

A.2 Population, Frame, and Sample

- Population: The target population comprises likely voters in the United States, focusing on those eligible and expected to participate in the upcoming election.
- Frame: The sampling frame is constructed from voter registration lists, which are comprehensive databases of registered voters across various states.
- Sample: A stratified random sample is drawn from the voter registration lists, ensuring representation across key demographics such as age, gender, race, education, and geographic location.A topline sample size of at least 100 (Radcliffe and Morris 2023).

A.3 Sample Recruitment Method

Participants are recruited through a combination of landline and mobile phone calls(like the Methodology in Table 2). This dual approach addresses the decline in landline usage and aims to reach a broader demographic spectrum. The use of voter registration lists enhances the likelihood of contacting registered voters, thereby improving the sample's representativeness.

A.4 Sampling Approach's trade-offs

A.4.1 Advantages

Enhances representativeness by ensuring all key subgroups are included. Improves precision in estimates for specific subpopulations.

A.4.2 Disadvantages

More complex and time-consuming compared to simple random sampling. Requires accurate and up-to-date information on the stratification variables. Handling Non-Response:

A.5 How is non-response handled

Non-response is addressed through weighting adjustments. Post-stratification weights are applied to align the sample with known population parameters, such as age, gender, race, and education. This process mitigates biases arising from differential response rates among various demographic groups.

A.6 What is good and bad about the questionnaire

A.6.1 Advantages

- Questions are crafted to be clear and unbiased, minimizing misinterpretation. The questionnaire covers a broad range of topics, providing comprehensive insights into voter opinions.

A.6.2 Disadvantages

- The length of the questionnaire may lead to respondent fatigue, potentially affecting the quality of responses
- The mode of administration (telephone interviews) might introduce social desirability bias, where respondents provide answers they perceive as socially acceptable rather than their true beliefs.

A.7 Summary

Overall, the New York Times/Siena College Poll employs robust methodologies to ensure accurate and reliable insights into voter preferences and behaviors.

B The idealized Methodology and Survey of forecasting the US presidential election

Idealized Methodology and Survey for Forecasting the US Presidential Election

1. Budget Allocation Overview

With a total budget of $100K, the funds would be allocated as follows:

Survey Design and Platform Costs: $5,000

Respondent Recruitment (Incentives and Advertising): $60,000

Sampling Management and Data Validation: $15,000

Analysis and Poll Aggregation Infrastructure: $10,000

Contingency and Miscellaneous Costs: $10,000

2. Sampling Approach

Target Population: The target population would be registered voters across the United States, ensuring representation from different demographics including age, gender, ethnicity, and political affiliation.

Sampling Method:

Stratified Random Sampling: To ensure proportional representation of various subgroups (e.g., regions, demographics), we will divide the population into strata and randomly select respondents from each stratum.

Sample Size: The goal is to survey 10,000 respondents to maintain a balance between precision and budget constraints.

Recruitment Strategy:

Online Panels: Use reputable online survey panels to reach diverse respondents.

Targeted Advertising: Run social media and digital ads to recruit additional respondents, particularly in underrepresented areas.

Incentives: Offer monetary incentives (e.g., $5 per respondent) to increase participation.

3. Data Validation

To maintain data integrity, the following checks will be implemented:

Response Quality Control: Implement screening questions and attention checks to ensure respondents are engaged.

IP Address Monitoring: Ensure unique responses by blocking multiple submissions from the same IP address.

Completion Time Analysis: Flag responses that fall outside reasonable time frames (e.g., completed too quickly).

Ethical Considerations: All data will be collected anonymously, and respondents will be informed of their rights, including the right to withdraw from the survey at any time. Personal data will not be shared or used beyond the scope of this survey.

4. Poll Aggregation and Methodology

Weighting: Apply post-stratification weighting based on census data to ensure the sample aligns with the US voter demographics.

Error Margin and Confidence Intervals: Report results with a calculated margin of error (typically around $\pm 3\%$) and a 95% confidence interval.

Aggregation of Poll Results: Collect and combine data from multiple waves of polling to smooth out any anomalies and better predict trends.

Adjustments for Non-Response Bias: Implement statistical adjustments to account for known biases in survey responses.

5. Implementation Platform

The survey will be hosted on Google Forms to streamline data collection and integration. A direct link to the live survey is included below:

Access the Survey Here: US Presidential Election 2024: Voter Opinion and Key Issues Survey

6. Survey Copy

Below is a more comprehensive sample of the survey questions:

Survey Title: US Presidential Election 2024: Voter Opinion and Key Issues Survey

Introduction: Thank you for participating in our comprehensive survey on the upcoming US presidential election. Your responses are vital for understanding public sentiment and will remain confidential. Participation is voluntary, and you may skip any question or withdraw at any time.

Q1. Are you a registered voter?

Yes

No

Q2. Which state do you reside in?

Q3. What is your age group?

18-24

25-34

35-44

45-54

55-64

65+

Q4. Which racial or ethnic group(s) do you identify with? (Select all that apply)

White/Caucasian

Black/African American

Hispanic/Latino

Asian/Pacific Islander

Native American

Other (please specify)

Prefer not to say

Q5. What is your highest level of education?

Less than high school

High school diploma or GED

Some college, no degree

Associate degree (e.g., AA, AS)

Bachelor's degree (e.g., BA, BS)

Master's degree (e.g., MA, MS, MEd)

Professional degree (e.g., JD, MD)

Doctoral degree (e.g., PhD, EdD)

Prefer not to say

Q6. What is your gender?

Male

Female

Non-binary/Third gender

Prefer not to say

Q7. Which candidate do you currently intend to vote for in the upcoming presidential election?

Kamala Harris

Donald Trump

Other

Undecided

Q8. What is your political affiliation?

Democrat

Republican

Independent

Other

Prefer not to say

Q9. How confident are you in your choice for the presidential election?

Very confident

Somewhat confident

Not very confident

Not confident at all

Q10. What is the most important issue influencing your vote? (Choose the top 3)

Healthcare

Economy

Immigration

Climate Change

Education

National Security

Social Justice

Other (please specify)

Q11. How often do you follow news and updates about the election?

Daily

Weekly

Occasionally

Rarely

Never

End of Survey: Thank you for your time and valuable input. Your responses will contribute significantly to our understanding of public opinion regarding the upcoming presidential election.

References

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

FiveThirtyEight. 2024. *2024 National Presidential Poll Results*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

Radcliffe, Mary, and G. Elliott Morris. 2023. *538's Polls Policy and FAQs*. https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

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

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

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

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.