Predicting the 2024 US Presidential Election with a Model-Based Forecast*

Using Generalized Linear Models to Predict Election Outcomes

Yuanyi (Leo) Liu Dezhen Chen Ziyuan Shen

October 31, 2024

In this paper, we develop a linear model to predict the outcome of the 2024 U.S. presidential election using a "poll-of-polls" approach that combines data from various prominent pollsters. Our analysis indicates that Donald Trump's probability of winning varies between 45% and 55% across states, depending on regional polling trends. However, Kamala Harris is projected to secure key states and achieve a majority of electoral votes, with a projected count of 239 to Trump's 98 based on the testing dataset. These predictions are based on the testing dataset, which lacks data for some states, limiting our ability to produce results for all states. This paper underscore the effectiveness of aggregated polls in producing reliable election forecasts.

Table of contents

*Code and data are available at: Forecasting the 2024 US Presidential Election.

1 Introduction

The results of presidential elections influence national and international policies and determine governance and economic priorities. An accurate election forecast is a valuable tool for political strategists, the media, and the public. The 2024 U.S. presidential election is expected to be highly competitive, with Donald Trump and Kamala Harris as the leading candidates. Despite the abundance of polling data, individual polls are often subject to bias and short-term fluctuations, making it difficult to predict election outcomes with confidence. This paper addresses this issue by using aggregate polling data to provide more reliable predictions of election outcomes.

Our findings suggest that while Trump has a competitive chance, Harris is projected to win key states and receive a majority of electoral votes. Harris will receive 239 electoral votes to Trump's 98 votes. This is important because accurate election forecasts help reduce uncertainty, which allows political stakeholders to allocate resources efficiently and gives the public a better understanding of likely outcomes. Aggregate polling reduces uncertainty and provides a clearer picture of potential outcomes.

The structure of the paper is organized as follows: following Section 1, Section 2 presents the data collection and cleaning process, along with an overview of the variables used in the analysis. Section 3 introduces the forecasting models, explaining why the selected models are suitable for predicting election outcomes based on aggregated polling data. Then Section 4 presents the main findings, including detailed crime trends for each neighborhood and year. Finally, Section 5 provides the results, highlighting key trends and predictions. Eventually, Section 5 concludes with a discussion of the findings, evaluating the reliability of the forecasts and identifying potential limitations of the models.

1.1 Estimand

In this study, our estimand is Donald Trump's percentage of polling in each state ahead of the 2024 U.S. presidential election. The object of the estimation is Trump's percentage of polling data in each state based on aggregated information from various sources By using a linear regression model, we aim to capture changes in public opinion over time and provide a clearer understanding of voter preferences and the likely election outcome.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to process and analyze polling data for the 2024 U.S. Presidential election. The dataset used for this analysis was obtained from the FiveThirtyEight 2024 U.S. Presidential Election Polls (Ryan Best 2024). It consists of polling data for the 2024 general election, covering various polling organizations and methodologies. Following methodologies discussed by "Telling Stories with Data" (Alexander 2023), we forecast election outcomes using the "poll of polls" method, which aggregates results from multiple polls to reduce bias and provide a more accurate representation of voter sentiment. For key operations, please refer to Appendix C.

The dataset includes 15,891 rows and 52 columns, covering various pollster attributes such as pollster name, state, methodology, and polling results. To ensure the reliability of our analysis, we filter the data to include only polls with a numeric grade of 2.5 or higher, which represents high-quality, reputable pollsters. This filtering allows us to focus on polls that follow rigorous standards and have demonstrated transparency and accuracy.

Additionally, we limit the dataset to polls conducted after July 15, 2024, when Donald Trump officially announced his campaign, ensuring that the data reflects the most current public sentiment. Other similar datasets from prior elections, such as data from previous general election cycles, could have been used for comparison. However, given that this analysis focuses on the upcoming 2024 election and the shift in public opinion, the most relevant data is specific to the current cycle.

2.2 Measurement

Polling data is a measurement of public sentiment captured through various survey methods. Polling organizations collect responses from individuals representing different segments of the population and ask them about their voting preferences. These responses are then weighted to reflect a more accurate representation of the electorate, based on factors like age, gender, race, and geographic region.

For this dataset, the poll results reflect the percentage of respondents who support a particular candidate. Different polling methods—such as live phone interviews, online panels, and mixed methodologies—capture this information. Each polling organization applies its own methodology, which can influence the results. For example, polls that use live interviews may experience different respondent behavior than those that use online surveys. More details on polling methodologies can be found in the ABC News methodology explanation.

The key measurement process involves converting real-world voter preferences into a dataset of percentages that reflect support for a candidate at a specific point in time. Polling organizations typically provide this data at a state or national level, which allows for both localized and broad interpretations of voter sentiment.

2.3 Outcome variables

2.3.1 Percent Support for Donald Trump

The primary outcome variable is the percentage (`pct`) of respondents who indicated support for Donald Trump in the 2024 general election. This percentage is calculated based on the total number of respondents in each poll divide by `sample_size`. The data includes observations from various states, providing both national and state-specific measures of Trump's support.

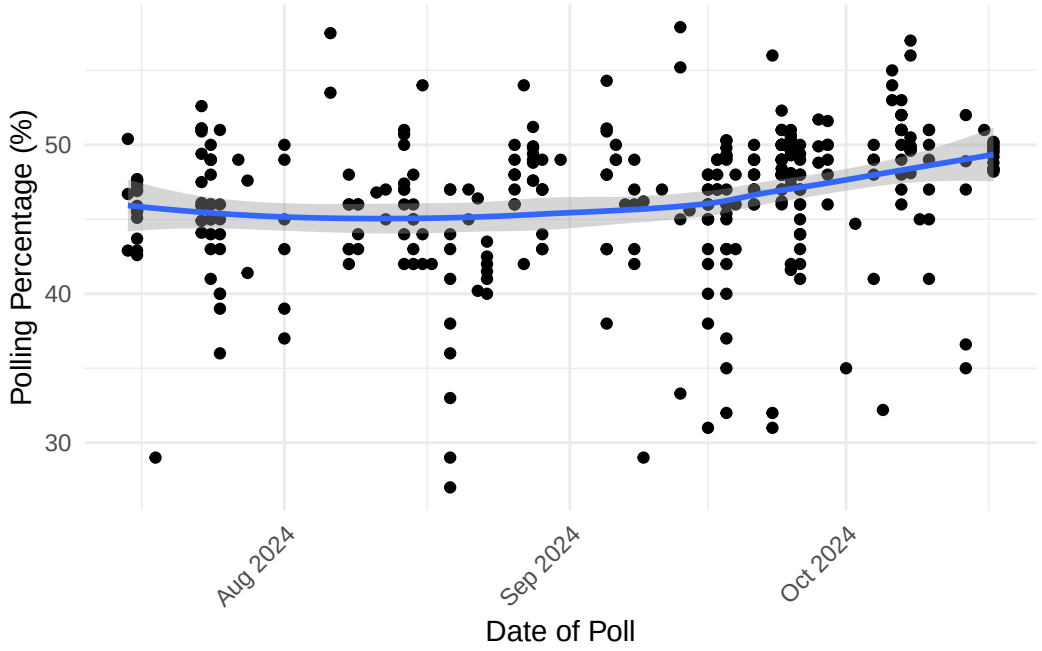

Figure 1: Donald Trump's polling percentages from July to October 2024. Each dot indicates the polling percentage for Trump on a specific date, while the blue line shows the trend in Trump's support.

Figure 1 shows Donald Trump's polling percentages over time from July to October 2024. Each black dot represents an individual poll conducted on a specific date, indicating the percentage of voters who supported Trump in that poll. The blue line represents a trendline, smoothing out the individual poll results to show the general trajectory of Trump's support over time.

Initially, Trump's polling percentage appears stable with small fluctuations around 45%. However, as the campaign progresses into October, there is a slight upward trend in Trump's polling percentage, suggesting that his support increased as the election neared.

The shaded region around the blue line represents the confidence interval, indicating the range within which the true polling percentage is likely to fall, accounting for sampling variation.

2.4 Predictor variables

2.4.1 Polling Methodology

The `methodology` variable describes the method each pollster used to collect data. It includes approaches such as live phone interviews, online surveys, and mixed methods. For polls that used multiple collection methods (e.g., phone and online), we labeled them as "Mixed Voting" for simplicity. The methodology is important because different approaches can introduce varying degrees of bias, influencing the final reported percentages.

Placeholder for a table showing the distribution of polling methodologies across different states.

2.4.2 State

The `state` variable identifies whether the poll is state-specific or national. In state-specific polls, the data reflects localized voter preferences, while national polls aggregate opinions across the entire country. For this analysis, we ignore national polls and focus entirely on state polls because electoral outcomes are determined on a state-by-state basis in the U.S. election system.

Table 1: The electoral votes allocated to each U.S. state, listed in two sets for better readability. Each state's electoral vote count reflects its representation in the Electoral College.

States	Electoral Votes	States	Electoral Votes
Alabama	9	Nebraska	10
Alaska	3	Nevada	6
Arizona	11	New Hampshire	4
Arkansas	6	New Jersey	14
California	55	New Mexico	5
Colorado	9	New York	29
Connecticut	7	North Carolina	16
Delaware	3	North Dakota	3
Florida	29	Ohio	18
Georgia	16	Oklahoma	7

States	Electoral Votes	States	Electoral Votes
Hawaii	4	Oregon	6
Idaho	4	Pennsylvania	20
Illinois	20	Rhode Island	4
Indiana	11	South Carolina	9
Iowa	6	South Dakota	3
Kansas	6	Tennessee	11
Kentucky	8	Texas	38
Louisiana	8	Utah	6
Maine	4	Vermont	3
Maryland	10	Virginia	13
Massachusetts	11	Washington	12
Michigan	15	West Virginia	5
Minnesota	10	Wisconsin	10
Mississippi	6	Wyoming	3
Missouri	10	District of Columbia	3
Montana	3		

Table 1 illustrates the number of electoral votes allocated to each U.S. state, based on data from the official government website (Nolan 2008). This information will be used to map the predicted results for each state later, allowing us to calculate the total number of electoral votes that Donald Trump is projected to receive.

2.4.3 Polling Score and Transparency

The `numeric_grade` variable is a rating assigned to each polling organization, indicating the reliability of the poll. Pollsters with higher numeric grades are considered more reliable and consistent in their methodology (e.g. higher than 2.5). Additionally, the `transparency_score` measures how openly a pollster shares their methodology and data, which further affects the trustworthiness of their results.

Figure 2 presents two bar charts illustrating the distribution of Polling Scores (Numeric Grades) and Transparency Scores across the polling organizations in the analysis:

- **Polling Scores**: Most organizations have numeric grades between 2.8 and 3.0, indicating that the majority of pollsters are considered highly reliable. Very few have grades below 2.7, which would suggest lower reliability.
- **Transparency Scores**: Transparency scores show more variability, with many pollsters achieving scores above 9, reflecting a high level of openness in their methodology. Fewer organizations have transparency scores below 6, indicating that while some pollsters are less open, the majority strive for transparency in their practices.

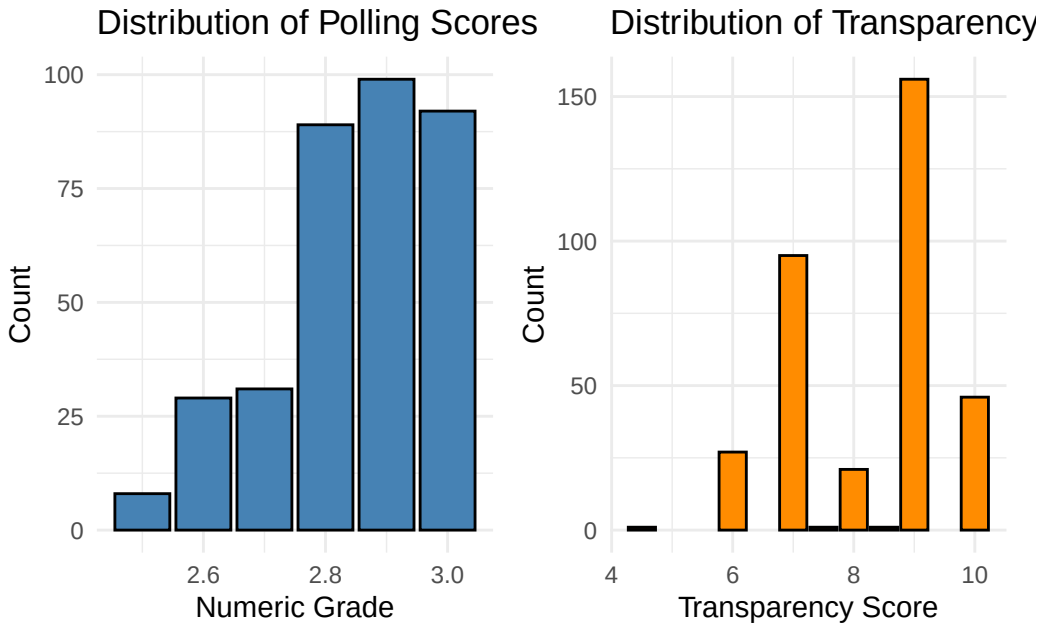

Figure 2: Distribution of Polling Scores and Transparency Scores of Polling Organizations in the 2024 U.S. Presidential Election

2.4.4 Sample Size

The `sample_size` variable represents the total number of respondents surveyed in each poll. A larger sample size generally leads to more reliable and precise estimates of voter sentiment, as it reduces the margin of error. In contrast, smaller sample sizes can result in greater variability and less confidence in the results.

2.4.5 Poll Duration

The `duration` variable is a constructed variable that represents how long Donald Trump has been in the 2024 presidential campaign, measured as the number of days from his official campaign announcement on July 15, 2024 until the `end_date` of each poll. This variable helps quantify how much time has passed since Trump officially entered the race and allows us to examine how his support has evolved over the course of his campaign.

3 Model

Our modeling approach aims to quantify the relationship between various polling metrics and the percentage support for a candidate. For this analysis, we use a linear model (LM) to exam-

ine how factors such as poll score, methodology, poll duration, sample size, transparency score, and state influence support percentages. The model is implemented using the lm function, with a Gaussian distribution to capture the variability in support rates.

In this analysis, we use predictors that reflect both the methodological quality and structural characteristics of each poll. Specifically, we include variables such as `numeric_grade`, which quantifies the poll's quality, `methodology`, `duration` (the length of the election period), `sample_size`, `transparency_score`, and `state` (the regional factor).

The model assumes that the distribution of support percentage, given these polling characteristics, follows a normal distribution. This Gaussian assumption facilitates parameter estimation, which is a standard approach in linear regression. To prevent overfitting and ensure interpretability, we assume moderate priors, maintaining balanced uncertainty across our predictors. This approach enables us to assess the impact of polling characteristics on candidate support while ensuring stability in our findings.

3.1 Model set-up

The model predicts Trump's support percentage using the following predictor variables:

- Poll Score (`numeric_grade`): Represents the quality rating of the poll.
- Methodology (`methodology`): Different methods used to conduct the poll, such as live phone, mixed voting, or online ads.
- Poll Duration (`duration`): The the length of the election period was conducted.
- Sample Size (`sample_size`): The number of respondents in the poll.
- Transparency Score (`transparency_score`): A measure of how transparent the polling data and methodology are.
- State (`state`): A categorical variable representing the U.S. state where the poll was conducted.

The model takes the form:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{Poll Score}_i + \beta_2 \cdot \text{Methodology}_i \\
&\quad + \beta_3 \cdot \text{Poll Duration}_i + \beta_4 \cdot \text{Sample Size}_i \\
&\quad + \beta_5 \cdot \text{Rransparency Score}_i + \beta_6 \cdot \text{State}_i + \epsilon_i \\
\epsilon_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Where:

- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are the coefficients for each predictor.
- σ^2 is the variance of the error term.

The model is executed in `R` (R Core Team 2023) using the `rstanarm` package (Goodrich et al. 2022). Default priors from `rstanarm` (Goodrich et al. 2022) are used, with the priors set to have a mean of zero and a moderate standard deviation to ensure a reasonable level of regularization.

3.1.1 Model justification

Existing research and political science theories suggest that factors such as poll quality, sample size, transparency, methodology, poll duration, and the state where the poll takes place can influence Trump's support percentage. Higher quality polls and larger sample sizes are believed to yield more reliable estimates, while transparency scores can affect levels of public trust, potentially altering support. Moreover, different polling methodologies (e.g., online surveys, telephone interviews) may affect how representative the sample is, while longer polling durations can reflect shifts in public sentiment over time. State-specific factors, such as local political events or population characteristics, can also contribute to the variation in support. The timing of the poll, particularly in relation to key political events, might further influence public opinion.

A linear regression model was chosen to predict Trump's support percentage because the dependent variable is continuous and tends to follow a normal distribution. Linear regression is a straightforward method to analyze how multiple factors contribute to an outcome, and it offers easy-to-interpret coefficients for each predictor. The model helps us determine the extent to which different aspects of polling influence the level of public support for Trump.

Further justification for using this model comes from its alignment with the central limit theorem, which suggests that with sufficient sample size, the distribution of poll percentages should approximate normality. Additionally, the relationships between the predictors (e.g., poll quality, methodology type) and support align well with established theories in political behavior, giving a solid theoretical underpinning to our model. The simplicity of a linear regression approach also helps mitigate overfitting, ensuring our results are generalizable to a broader set of polling data.

4 Results

Section 4 examines the relationship between polling score, methodology, poll duration, transparency score, sample size, and state, with respect to Donald Trump's polling performance in the 2024 U.S. Presidential Election. Using a dataset that captures polling data across various

states and polling organizations, we apply a linear regression model to identify key factors that influence Trump's support. Below, we present the results of our model and its implications.

We predicted the outcome based on our test dataset. However, due to missing data for certain states in our original dataset, we were unable to make predictions for all states. This limitation arises because the states with NA values in the dataset are not represented in the model, preventing accurate forecasts for those regions.

4.1 Model Results and Interpretation

The linear regression model built by our training dataset, using 384 data points, estimated the factors influencing Trump's support levels. As Table 2 shows, the intercept, estimated at 43.39, represents the level of support when all other predictors are at baseline. The model achieved an R-squared value of 0.85, indicating that 85% of the variance in Trump's support percentage can be explained by the predictors included in the model. The adjusted R-squared value of 0.82 also shows that the model effectively captures the relationships between variables without overfitting. The Root Mean Square Error (RMSE) was 2.03, indicating the average difference between predicted and actual values. The relatively low RMSE value suggests that the model provides reasonably accurate predictions.

The results highlight that `numeric_grade` and `sample_size` are particularly important in predicting support levels, with larger and higher-quality polls offering more reliable estimates. The transparency score and state-level coefficients further provide understanding into regional differences and the role of trust in polling, helping us better understand the dynamics behind Trump's public support.The coefficients for `numeric_grade` and `sample_size` underline the importance of poll quality and sample size in predicting support levels. Larger, higher-quality polls offer more estimates. Additionally, transparency and state effects provide insight into regional differences and trust factors affecting polling results, giving a clearer understanding of the dynamics influencing Trump's public support.

4.2 Predicted Electoral Outcomes

We employed a regression model to estimate the percentage of votes Trump is likely to receive in each state. Using these predicted results and the electoral vote allocations, we identified the winner for each state and calculated the total electoral votes for both Trump and Harris.

Table 3 presents a summary of the predictions, including Trump's estimated percentage, the number of electoral votes in each state, and the projected winner. For instance:

- In Arizona, Trump is expected to receive 48.94% of the vote, resulting in a win for Harris with 11 electoral votes.
- In Missouri, Trump is projected to receive 51.64%, thereby winning all 10 electoral votes for that state.

Table 2: Summary of the Linear Regression Model for Predicting Polling Outcomes

	Linear Model
(Intercept)	43.387
numeric_grade	1.859
as.factor(methodology)Live Phone	−1.280
as.factor(methodology)Mixed Voting	0.012
as.factor(methodology)Online Ad	−0.948
as.factor(methodology)Online Panel	−2.321
as.factor(methodology)Probability Panel	−1.492
duration	0.030
sample_size	0.002
transparency_score	−0.323
stateCalifornia	−16.090
stateConnecticut	−12.659
stateFlorida	2.113
stateGeorgia	0.021
stateIowa	−0.751
stateMaine	−5.214
stateMaine CD-1	−7.540
stateMaine CD-2	1.640
stateMaryland	−17.786
stateMassachusetts	−16.475
stateMichigan	−1.478
stateMinnesota	−4.143
stateMissouri	5.423
stateMontana	6.082
stateNebraska CD-2	−6.064
stateNevada	−1.356
stateNew Hampshire	−4.282
stateNew Mexico	−4.697
stateNew York	−9.303
stateNorth Carolina	−0.800
stateOhio	1.317
statePennsylvania	−1.547
stateRhode Island	−9.107
stateSouth Carolina	2.199
stateTexas	1.535
stateVermont	−18.116
stateVirginia	−4.335
stateWashington	−12.625
stateWisconsin	−1.571
Num.Obs.	242
R2	0.846
R2 Adj.	0.817
AIC	1065.7
BIC	1205.2
Log.Lik.	−492.846
RMSE	1.85

- In Ohio, with an estimated 51.98% of the vote, Trump secures Ohio's 18 electoral votes.

Electoral Vote Count: Trump Electoral Votes: 98 - Harris Electoral Votes: 239. This indicates that Harris is projected to win the 2024 U.S. Presidential Election with 239 electoral votes, while Trump is projected to gather 98. These results hinge on the outcomes in various battleground states, with several being extremely close.

Table 3: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
Arizona	48.94	11	Harris
California	33.80	55	Harris
Florida	50.53	29	Trump
Georgia	49.58	16	Harris
Maine	40.95	4	Harris
Maryland	32.06	10	Harris
Massachusetts	33.18	11	Harris
Michigan	46.74	15	Harris
Minnesota	42.84	10	Harris
Missouri	51.64	10	Trump
Montana	54.75	3	Trump
Nevada	48.07	6	Harris
New Hampshire	45.87	4	Harris
New Mexico	44.61	5	Harris
New York	39.59	29	Harris
North Carolina	46.83	16	Harris
Ohio	51.98	18	Trump
Pennsylvania	47.48	20	Harris
Rhode Island	40.55	4	Harris
Texas	50.88	38	Trump
Virginia	44.21	13	Harris
Wisconsin	46.78	10	Harris

4.3 Prediction map

Figure 3 displays the predicted outcome for each state in the 2024 U.S. Presidential Election, with green representing states where Trump is expected to win and blue for those where Harris is predicted to prevail.

The map illustrates Trump gaining traction across central and southern states, including Texas and Missouri, which have traditionally leaned Republican. Meanwhile, Harris holds a solid

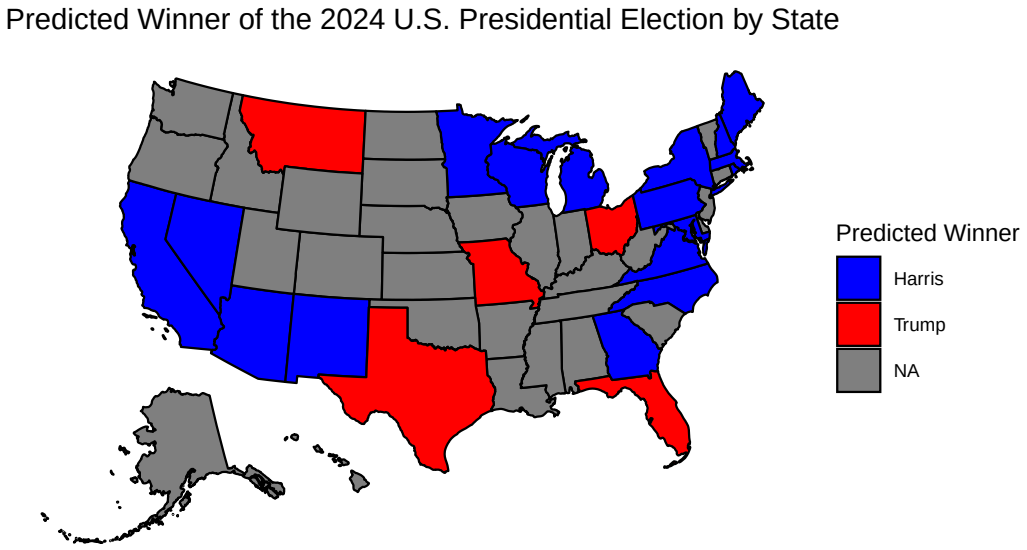

Figure 3: Predicted Winner of 2024 US Presidential Election by State

lead in states along the West Coast and in the Northeast, such as California and New York, regions that have a history of Democratic support.

Several battleground states, including Florida and Ohio, are projected to favor Trump, with a slight advantage that marks a contrast to Harris's stronger leads in places like California. The map also shows a number of central states in gray, reflecting either a lack of definitive prediction or insufficient data. Overall, the map reveals distinct regional patterns, with Trump expected to perform well in the heartland and Harris finding support along the coasts. The outcome may hinge on the final results in the competitive states like Pennsylvania, Arizona, and North Carolina.

5 Discussion

5.1 Interpretation of Polling Data and Electoral Implications

This paper examines the factors affecting Donald Trump's polling performance in the 2024 U.S. Presidential Election. Our linear regression model shows that poll duration and state-specific factors are the strongest predictors of Trump's support. Longer poll durations tend to produce more stable data by capturing changes in public sentiment over time rather than short-term reactions. The state variable underscores the importance of local dynamics, as individual states bring unique policy priorities and demographic trends to their voter bases. These regional distinctions are essential in the Electoral College system, where state-level support directly impacts the final electoral count.

Transparency and sample size play secondary roles in data reliability. Polls with high transparency scores provide clear information about methodology, making them more reliable for analysis. Larger sample sizes reduce margins of error, adding stability to projections, especially in closely contested states. While both transparency and sample size support reliable predictions, they have less influence on voter preferences than duration and state-specific factors.

In comparison, polling methodology (e.g., live phone, online) and numeric grade (a general reliability rating) have minimal impact on Trump's support in our model. Although these factors contribute to overall data quality, they do not significantly shape voter sentiment. This finding suggests that regional context and polling consistency over time are more important in understanding Trump's polling performance in this election than the specific techniques or reputations of polling organizations.

5.2 Influence of State Policies on Voter Preferences

Beyond the mechanics of polling, our analysis suggests that state-specific policies and regional cultural factors could influence voter preferences and ultimately shape electoral outcomes. For instance, certain policies, such as the legalization of recreational marijuana, reflect broader ideological values that may correlate with support for particular candidates. Although our model does not directly analyze the impact of marijuana policies, including these cultural aspects could offer a more comprehensive view of voter alignment.

States with legal recreational marijuana, for example, might lean more progressive or libertarian, potentially aligning with candidates like Harris who advocate for policies seen as liberal or socially progressive. This policy environment might signal broader state values that align with a particular party's stance, beyond just the specific candidate. Conversely, states that prohibit recreational marijuana use may lean more conservative, suggesting they may have stronger support for Trump. Although this relationship is not absolute, understanding such cultural factors helps contextualize the state-level support each candidate may receive and could be valuable in identifying battleground states. Additionally, factors like local economic policies, educational investment, or healthcare reform could influence voter preferences, making these areas ripe for future analysis.

5.3 Limitations of the Data and Model

One limitation of this study is the incomplete dataset, with some states missing from the polling data. This limits our ability to produce a fully accurate prediction, as certain areas are excluded from the model. Additionally, the focus on state-level data may overlook finer voting patterns within states, particularly in districts with unique political identities.

5.4 Future Research Directions

Future studies could improve by incorporating demographic variables such as age, income, and party affiliation, which might offer a clearer understanding of voter preferences. Examining additional state-level policies, like healthcare or education, may also provide insights into factors shaping voter choices.

Another possible enhancement involves increasing the geographic specificity of the model to include congressional districts or smaller regions, which may lead to more precise forecasts, especially in competitive states. As polling data collection methods evolve, combining traditional regression with machine learning techniques could allow for models that adapt to ongoing changes in public opinion, resulting in more timely and adaptable forecasts.

Appendix

A Pollster Methodology Overview and Evaluation

A.1 Overview of Siena/NYT

Siena College Research Institute (SCRI) and The New York Times launched a collaborative political poll in July 2013. Siena/NYT was known for its rigorous methodology and accuracy, which provided its first real-time midterm election poll in 2018 (Nagourney and Igielnik 2024). In 2024, FiveThirtyEight named it the most accurate pollster in the United States. The polling methodology emphasizes live phone interviews and representative sampling and focuses on transparency and statistical accuracy to provide a deep understanding of voter sentiment.

A.2 Target Population, Frame, and Sample

- Target Population: Registered voters or likely voters in the United States, especially voters in major battleground states during an election in 2024.
- Sample Frame: U.S. registered voters based on a national voter file maintained by L2, with a cell phone number that matches the voter file.
- Sample Size: Sample sizes vary across different polls. The poll in late October 2024 surveyed over 2,500 voters, and 2,097 of them completed the full survey. The margin of error for likely voters is about ± 2.2 percentage points.

A.3 Sample Recruitment

The sample was recruited by telephone. Registered voters were selected from the voter files held by L2 and contacted via their listed telephone numbers including landlines and mobile phone numbers. More than 260,000 calls were made to more than 80,000 voters, with over 98% of respondents using mobile phones. Interviews were conducted by live telephone interviews in both English and Spanish.

A.4 Sampling Approach and Trade-offs

Siena/NYT used a stratified sampling method (Nagourney and Igielnik 2024). Polls were stratified by demographic characteristics such as political party, race and region. This improves accuracy but adds cost and complexity. It also makes it challenging to reach certain populations such as young voters.

Advantages of stratified sampling:

- Generates a representative sample that captures the diverse opinions of different voter populations.
- Helps to ensure that under-represented groups are given appropriate weighting, and reduces potential bias.
- Provides a more accurate reflection of the wider population by stratifying on key characteristics.

Disadvantages of stratified sampling:

- Increased cost and time to complete polls, especially for live telephone interviews.
- There are challenges in reaching specific demographic groups, such as younger voters who may be less likely to participate in the poll.
- Reliance on weighting adjustments to correct for response discrepancies, which does not fully eliminate residual bias.

A.5 Non-response Handling

Weighting adjustment: The Times weighted the sample using the survey software package in R and adjusted for unequal probability of selection in each stratum. The sample was also split by political party and weighted according to parameters of registered voter characteristics from the voter file. This weighting step ensures that respondents from underrepresented demographic groups such as those without a college degree are given more weight so that the results reflect the voting population overall.

Non-Response Model: To adjust for non-response bias, the L2 voter file was stratified by 'state precinct', 'political party', 'race', 'gender', 'marital status', and 'voting history'. The average expected response rate is based on a single-place non-response rate model from previous Siena/NYT surveys, which can be adjusted appropriately to account for different response rates in different groups.

Multiple calls and language adaptation: Interviewers made more than 260,000 calls to more than 80,000 voters, resulting in a sample of 2,516 respondents. Interviewers conducted the interviews in English and Spanish, and bilingual interviewers adapted the language of the interviews to respondent preferences, which helped to increase the participation of Spanish-speaking respondents. Re-contacts by interviewers reduced the non-response rate.

A.6 Questionnaire Evaluation

Strengths:

- Minimize bias through the use of straightforward and non-leading questions.
- Transparency was maintained by publishing the questions verbatim to allow for public scrutiny.

- Bilingual interviewers ensured that respondents could participate in their preferred language (English or Spanish), which increased the participation of Spanish-speaking voters and improved the quality of respondents' answers.

Weaknesses:

- The length of the poll can lead to respondent fatigue, which affects the quality of the data and increases dropout rates.
- It can be challenging to limit calls to less than 15 minutes, which can lead to incomplete or rushed responses, thus affecting the overall reliability of the data.

A.7 Conclusion

The Siena/NYT poll uses stratified sampling and live phone interviews to reach a representative sample of voters. While the use of composite weighting and bilingual interviews provides advantages such as transparency and reduced bias, challenges remain including high costs and the potential for bias due to non-response. Despite these drawbacks, Siena/NYT's commitment to transparency and rigorous methodology makes it a reliable source for understanding voter sentiment and opinion.

B Idealized Methodology and Survey

B.1 Overview

The objective of this survey is to predict voter sentiment in the upcoming U.S. presidential election, by using a budget of $100,000 to collect accurate and varied data on voting intentions, candidate favourability, and key issues from a representative sample of registered or likely voters across the United States. The methodology is designed to maximize accuracy, minimize bias and take into account a variety of demographic, geographic and political factors that influence voting behaviour.

B.2 Sampling Approach

The sampling approach is designed to increase coverage and ensure that a representative voting population is covered. This will be achieved by using a combination of stratified random sampling and quota sampling of key demographics.

B.2.1 Stratification Variables

- Age: 18-29, 30-44, 45-64, 65+
- Gender: Male, Female, Non-binary
- Race: White, Black /African American, Hispanic/Latino, Asian, Native, Other
- Education level: High School, Some College, Bachelor's Degree, Graduate Degree, Other
- Geographic region: Northeast, Midwest, South, West
- Political Party: Democratic, Republican, Independent, Other

B.2.2 Sample Size

The target sample size was 5,000 respondents, which are stratified by the Stratification Variables described above. This stratification ensures that all major voter groups are proportionately represented. 5,000 samples have a margin of error of about $\pm 1\%$ and a confidence level of 95%, which is appropriate for predicting elections.

B.3 Recruitment Strategy

The recruitment strategy was separated into two main approaches to ensure a diverse and representative sample:

Online Survey panels:

- Work with established survey panels such as Lucid or YouGov.
- These panels have access to large representative databases of voters.
- Recruiting through panels ensures quality and reliability, as they have built-in validation and quality control processes.

Social Media Advertising:

- Use social media platforms such as Facebook, Instagram and LinkedIn to reach potential respondents.
- Target groups that are underrepresented in traditional panels, including young voters and minority groups.
- Social media ads will be customized with demographic information to increase reach and participation.

Incentives: Offer gift card lottery prizes to all participants to encourage participation and ensure a high response rate.

B.4 Data validation and quality control

To maintain the integrity of the data, we will take the following measures:

- **Attention Checks:** At least two 'attention check' questions will be included in the survey to identify inattention or robots.
- **Duplicate Detection:** Use email or IP-based verification to prevent duplicate responses. Ensure that each respondent provides only one response set to keep the dataset unique.
- **Response Time Monitoring:** Track the time it takes respondents to complete surveys. Responses with apparently faster-than-average completion times are labelled as potentially low-quality.
- **Data Consistency Check:** Analyses the consistency of responses, especially between related questions. Inconsistent answers will be labelled for review.
- **Panel partner validation:** For respondents recruited through an online survey panel, rely on the panel provider's in-built validation systems. These systems use various quality checks, such as identity verification and response consistency, which ensure high-quality data from panel participants.

B.5 Poll aggregation and forecasting

B.5.1 Poll aggregation

- 'Poll of polls' approach: Survey data will be combined with data from some reliable polling sources such as Lucid, YouGov and others. This approach aggregates the results of multiple reputable polls to improve the forecast reliability by reducing bias and individual polling errors.
- Weighting adjustments: Polls are weighted according to sample quality, size, repeatability, and demographic match. Demographic mismatches are adjusted using raking to ensure that the aggregated data accurately represents the U.S. population.

B.5.2 Forecasting Method

- Bayesian modelling: Aggregate polling data is combined with historical electoral trends and demographic information using Bayesian statistical models. This approach allows for dynamic updating as new data becomes available, which improves accuracy.
- Uncertainty Quantification: Includes confidence intervals to express uncertainty in forecasts for a more detailed interpretation.

B.6 Budget Allocation

- Survey Panel Recruitment: $50,000
- Social Media Advertising: $50,000
- Incentives(lottery prizes): $15,000
- Data Cleaning and Validation: $5,000
- Poll Aggregation and Modeling: $10,000

B.7 Survey Implementation

The survey will be available via Google Forms to ensure broad accessibility and easy data collection. The link to the survey will be distributed via email invitations to panellists and targeted social media adverts. To ensure maximum reach and diversity, multiple channels will be used, including Facebook, Instagram, and other social media platforms. The link to the survey is here.

Distribution Plan:

- **Panel invitations:** Members of established survey panels (e.g. Lucid, YouGov) will receive email invitations to participate in the survey.
- **Social media activities:** Targeted adverts will be placed on social media platforms to reach under-represented groups, especially young voters and national minority groups.
- **Follow-up reminders:** An automated reminder email will be sent to participants who have not completed the survey to maximize response rates.
- **Survey Monitoring:** Response rates will be closely monitored throughout the data collection period and sampling methods will be adjusted if certain population quotas are not reached.
- **Accessibility:** The survey will be mobile-friendly, which will allow respondents to complete the survey on any device.

B.8 Survey Structure

Survey Introduction: Welcome! We are surveying to better understand voter preferences and priorities for the upcoming U.S. presidential election. Your participation is important and your answers will help us accurately predict the outcome of the election. It will take approximately 10 minutes to complete the survey. All responses are confidential and will be used for research purposes only.

If you have any questions or concerns, please feel free to contact our research team at winniekeai23@gmail.com (Ziyuan Shen, Yuanyi (Leo) Liu, Dezhen Chen).

As a thank you for your participation, you will be entered into a raffle to win a grand prize of a gift card. We are deeply grateful for your participation and appreciate the time and effort that you put in.

Survey Question:

Part 1: Screening for eligibility

1.Are you an eligible U.S. citizen?
* Yes / No
2.Are you currently 18 years of age or older?
* Yes / No
3.Have you completed your registration to become a voter?
* Yes / No / Plan to register

Part 2: Demographics

4.What is your age?
* 18-24 / 25-34 / 35-44 / 45-54 / 55+
5.What is your gender?
* Male / Female / Non-binary / Prefer not to say
6.Which of the following best describes your education level?
* High School / Some College / Bachelor's Degree / Graduate Degree / Other
7.Which region of the U.S. do you currently reside in?
* Northeast / Midwest / South / West
8.Do you generally identify with a political party?
* Democratic / Republican / Independent / None

Part 3: Voting Intentions

9.How likely are you to participate in the upcoming presidential election?
* Very Likely / Somewhat Likely / Not Sure / Unlikely / Very Unlikely
10.If the presidential election were held today, for which candidate would you vote?
* Kamala Harris (Democrat) / Donald Trump (Republican) / Undecided / Prefer not to say
11.What are the top three issues that will influence your vote?
* Economy / Healthcare / Education / Gun Policy / Taxes / Foreign Policy / Social Security / Civil Rights / Technology / Privacy / Other[text]

Part 4: Information Sources and Engagement

12.Which of the following sources do you rely on most for political news and information?
* Television News / Newspapers / Online News Websites / Social Media / Radio / Friends and Family / Other
13.How often do you engage in discussions about political topics with friends or family?
* Daily / Weekly / Monthly / Rarely / Never

Part 5: Verify

14.Please select 'Agree' to verify that you are paying attention.
* Agree / Disagree
15.Do you agree to participate in this survey? Your responses will be kept confidential and used only for research purposes.
* Yes / No

End Part:

Thank you for your participation!

Your responses are valuable to us in anticipating the upcoming election and understanding voters' priorities. If you have any questions or would like more information about our research, please feel free to contact us at winniekeai23@gmail.com(Ziyuan Shen, Yuanyi (Leo) Liu, Dezhen Chen).

C Data Manipulation and Cleaning

In the data cleaning process, we prepared the raw election data for analysis by applying transformations, filtering, and restructuring using several R packages, including `dplyr` (Wickham et al. 2023), `janitor` (Firke 2024), `tidyverse` (Wickham et al. 2019), `arrow` (Richardson et al. 2024), and `rsample` (Frick et al. 2024).

1. **Standardizing Column Names**: Using the `clean_names()` function, we converted column names to a consistent lowercase format, making them easier to reference and manipulate.
2. **Filtering Polls by Quality and Candidate**: With the `dplyr` package from `tidyverse`, we filtered the dataset to include only higher-quality polls with a `numeric_grade` of 2.5 or above, ensuring that only credible polls were considered. We further filtered to focus on polls related specifically to Donald Trump (`candidate_name == "Donald Trump"`), aligning the dataset with our analysis objective.
3. **Date Conversion and Filtering**: We converted the `end_date` column to a Date format using `mdy()` from `lubridate` (part of `tidyverse`), ensuring date consistency. We then filtered the data to include only polls conducted after July 15, 2024—the date of Trump's campaign announcement—allowing us to focus on his active campaign period.
4. **Creating Duration and Support Columns**:
    - **Duration**: We created a `duration` variable using the `difftime()` function from base `R`, which calculates the number of days between each poll's end date and Trump's campaign start date. This allows us to observe how polling data shifts over time.

- **Number of Trump Supporters**: We calculated `num_trump`, the estimated number of respondents supporting Trump in each poll, by applying the support percentage (`pct`) to the sample size and rounding the result with `dplyr` functions.

5. **Methodology Update**: We used `str_detect()` from `stringr` (in `tidyverse`) to identify polls that listed multiple polling methods (e.g., those containing "/") and standardized them to "Mixed Voting" to clarify the `methodology` column.
6. **Selecting and Renaming Columns**: We used `dplyr` functions to select the relevant columns (`poll_id`, `numeric_grade`, `methodology`, `transparency_score`, `end_date`, `sample_size`, `pollster_name` (renamed from `display_name`), `state`, `pct`, `duration`, and `num_trump`), streamlining the dataset for analysis.
7. **Removing Duplicates and Missing Values**: With `dplyr`, we removed duplicate rows using `distinct()` and dropped rows with missing values using `drop_na()`, ensuring that only complete and unique data entries were retained.
8. **Splitting the Data**: We utilized `rsample` to perform a stratified split on the cleaned dataset, using the `state` variable to ensure balanced representation across states. The data was divided into training (70%) and testing (30%) sets, preparing it for modeling.

Finally, we saved the cleaned dataset and the training/testing splits in CSV and Parquet formats using `arrow`, facilitating efficient storage and future accessibility.

D Model details

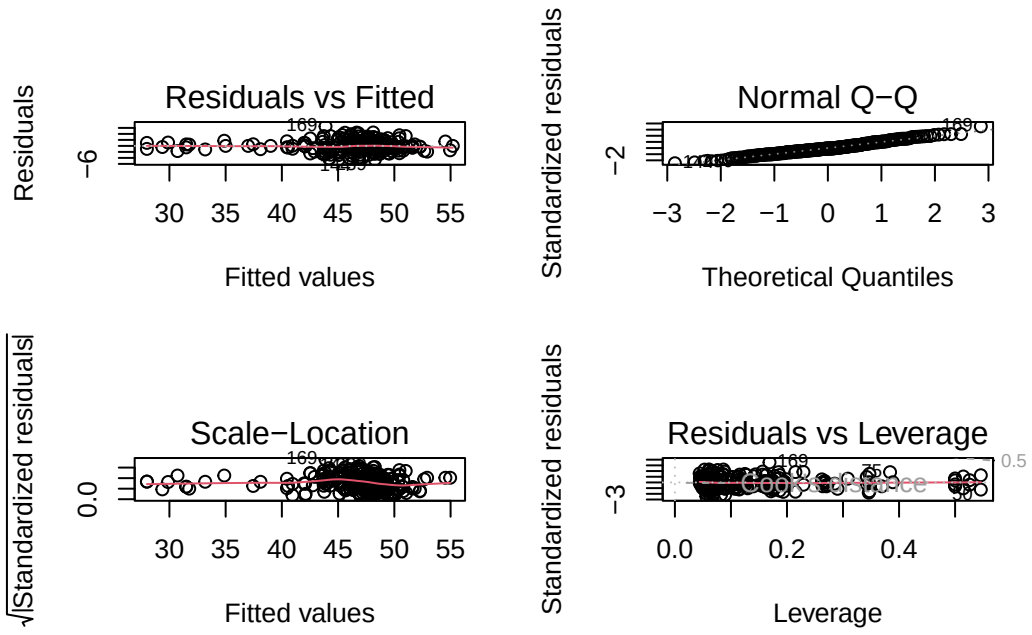

Figure 4: Model Diagnosis

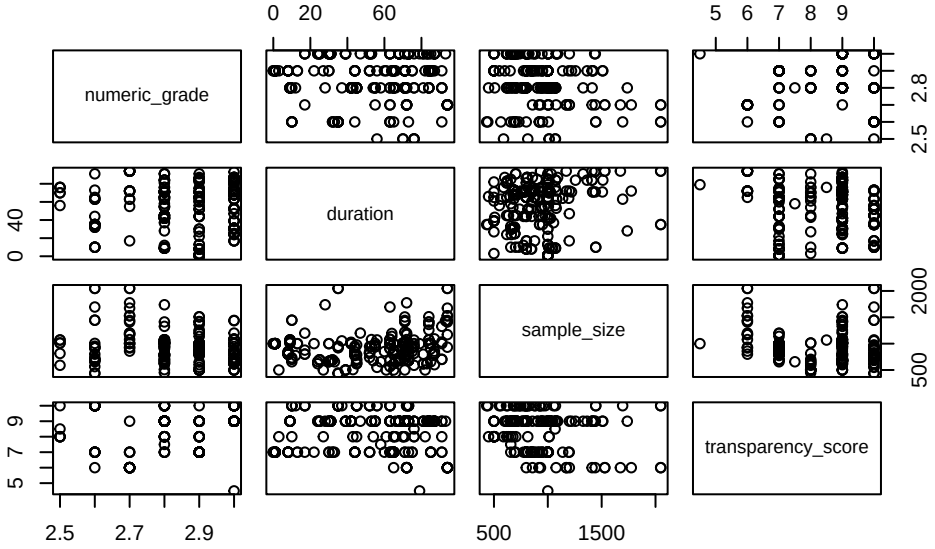

Figure 5: Model Diagnosis

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Firke, Sam. 2024. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Frick, Hannah, Fanny Chow, Max Kuhn, Michael Mahoney, Julia Silge, and Hadley Wickham. 2024. *Rsample: General Resampling Infrastructure*. https://rsample.tidymodels.org.

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

Nagourney, Adam, and Ruth Igielnik. 2024. "Harris and Trump Deadlocked to the End, Final Times/Siena National Poll Finds." *The New York Times*. The New York Times. https://www.nytimes.com/2024/10/25/us/politics/poll-harris-trump-times-siena.html.

Nolan, Jacqueline V. 2008. "United States Electoral College, Votes by State." *Library of Congress*. https://www.loc.gov/resource/g3701f.cp000001/.

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://github.com/apache/arrow/.

Ryan Best, Aaron Bycoffe. 2024. "National: President: General Election: 2024 Polls." *FiveThirtyEight*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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.