Forecasting the 2024 U.S. Presidential Election: An Analysis of Battleground States*

Trump Favored Nationally, While Harris Leads in 4 of 7 Battleground States

Shamayla Durrin

November 11, 2024

This study employs a poll-of-polls approach to predict support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election, focusing on key battleground states to forecast the likely winner of the electoral college. By aggregating multiple polls and applying a weighted linear regression model with predictors like pollster reliability, sample size, state, and recency, we estimate higher national support for Trump than Harris. Our analysis also shows a close competition in battleground states, with Trump holding a slight lead in Arizona, Georgia, and North Carolina, while Harris leads in Michigan, Nevada, Pennsylvania, and Wisconsin. These findings show the value of aggregating polls over relying on individual surveys, offering a robust forecast of electoral outcomes.

1 Introduction

While individual polls provide snapshots of public opinion, they are often subject to biases and methodological differences. In this paper, we aim to forecast voter support for Kamala Harris and Donald Trump by using data from multiple polling sources, reducing individual poll biases, and improving overall prediction accuracy. Our analysis estimates the levels of support for each candidate nationally while focusing on swing states, which are likely to be decisive in determining the Electoral College outcome.

The estimand of this study is the level of voter support for each candidate, Kamala Harris and Donald Trump, as reported across multiple polls. To estimate this, we developed a regression model incorporating variables such as pollster, sample size, state, and recency of the poll, with an emphasis on accurately capturing state-level dynamics. Our findings indicate that, on a national level, support between Kamala Harris and Donald Trump is closely balanced. Harris's

[*] Code and data are available at: https://github.com/krishnak30/US_elections.

estimated national support is around 48%, while Trump's support is slightly higher, around 49%. We found that Trump holds a lead in battleground states like Arizona and Georgia, with a support margin of over 1%. In contrast, Harris shows small leads in Michigan, Nevada, Pennsylvania, and Wisconsin, with her support margin in Wisconsin over 2%, making it her strongest battleground. North Carolina is one of the tightest races, with Trump leading by only 0.26%. This analysis is useful for political strategists, media analysts, and the general public by offering a view of the electoral landscape.

This paper is organized as follows: In Section Section 2, we show summary statistics, plot distributions of key variables, and examine relationships between variables. In Section Section 3, we discuss our forecasting approach, model selection, justification for the chosen model, and the mechanism of deriving poll weights based on pollster reliability, ultimately presenting our predictions. In Section Section 4, we address the broader implications of our findings, acknowledge limitations, and suggest directions for future work. Appendix A contains details of the data cleaning process and model diagnostics.

2 Data

2.1 Variables of Interest

2.1.1 Summary Statistics of Key Variables and Measurement Method

We measure voters' intentions by surveying a sample of individuals at various points leading up to the election, asking about their current opinions on candidates, issues, or likelihood to vote. Each response is recorded with a timestamp, allowing us to analyze changes over time (recency) and connect shifting opinions to projected behaviors. We then aggregate and analyze these responses to predict patterns, identifying trends that could indicate likely voting actions on Election Day.

For this analysis, a subset of variables was selected from the raw data set. Two additional variables, 'national' and 'state', were created using the existing 'state' and 'end date' variables. A short description of each variable of interest and their respective measurement methods is given below.

- *Pollster*: The name of the polling organization conducting the poll (e.g., YouGov, RMG Research). This variable helps adjust for poll-specific biases. Every pollster that publicly published a scientific poll about the 2024 U.S. Presidential Election is included in the data set. The value measured for this variable is the name that appears in the scientific poll (Morris 2024b).
- *Numeric Grade*: A numeric rating given to the pollster, representing the accuracy and methodological transparency of the organization (e.g., 3.0), with higher grades indicating more reliable pollsters. The value of this variable is calculated by the 538 team. All

national and state-level polls that were conducted in 1998 or later are considered to determine and assign a numeric grade to each pollster (Morris 2024a).

- *Pollscore*: A score that reflects the reliability of each pollster, capturing their historical accuracy and biases. Negative values indicate better predictive accuracy, and rewards pollsters who are accurate and precise. This variable is calculated by aggregating predictive error and predictive bias together (Morris 2024a). Predictive error is the difference between expectations and reality (Renewable Energy 2012) and predictive bias is the systematic bias introduced in the methodology (Aguinis and Culpepper 2024).
- *State*: The U.S. state where the poll was conducted, allowing for analysis of regional differences in candidate support. The value measured for this variable is the primary state that appears in the scientific poll's methodology (Morris 2024b).
- *National*: A binary variable indicating whether the poll is national (1 for national polls, 0 for state polls).
- *End Date*: The date the poll ended, reflecting the currency of the data. This variable is taken from the publication and is the last date for which the survey was active (Morris 2024b).
- *Sample Size*: The total number of respondents in the poll. This is measured by the number of completed surveys (Morris 2024b).
- *Candidate Name*: The name of the candidate being polled (e.g., Kamala Harris or Donald Trump), identifying the focus of each poll result.
- *Pct (Percentage)*: The percentage of respondents in the poll who support the specified candidate.
- *Recency*: This variable measures how recent each poll is. This variable is calculated by subtracting the end date of the poll from the end date of the most recent poll.

Table 1: Summary Statistics of Numerical Variable

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
numeric_grade	21	0	2.2	0.6	1.0	2.1	3.0
pollscore	21	0	-0.5	0.6	-1.5	-0.4	1.7
sample_size	594	0	1908.8	2602.0	147.0	999.0	20762.0
pct	216	0	46.9	4.0	25.0	47.0	70.0
recency	91	0	43.6	25.9	0.0	41.0	90.0

Table 1 shows the average poll in our data set has a numeric grade of 2.2. This suggests most polls are of moderate to high quality in terms of reliability. The mean poll score of -0.5 suggests these polls are accurate, as negative values imply reduced bias. Also, the large average sample size of 1908 respondents across polls reduces variability and ensures reliable predictions for the candidate support model.

2.1.2 Variation of Poll Quality and Support for Candiates by Pollster

Table 2: Top 5 most frequent pollsters, with count of polls, average pollscore (lower scores indicate less bias), and average numeric grade (higher values indicate greater reliability).

Pollster	Count	Average Pollscore	Average Numeric Grade
Morning Consult	470	-0.3	1.9
Siena/NYT	188	-1.5	3.0
Redfield & Wilton Strategies	176	0.4	1.8
Emerson	116	-1.1	2.9
YouGov	113	-1.1	3.0

Table 2 shows the top 5 most frequent pollsters in the dataset, with each pollster's total poll count, average poll score (measuring reliability and historical accuracy), and average numeric grade (indicating overall quality). The most frequent pollster, Morning Consult, has an average poll score and a slightly above-average numeric grade, suggesting the data provided is reliable. However, there is variability in scores and numeric grades across the top pollsters, showing differences in quality and potential biases among them.

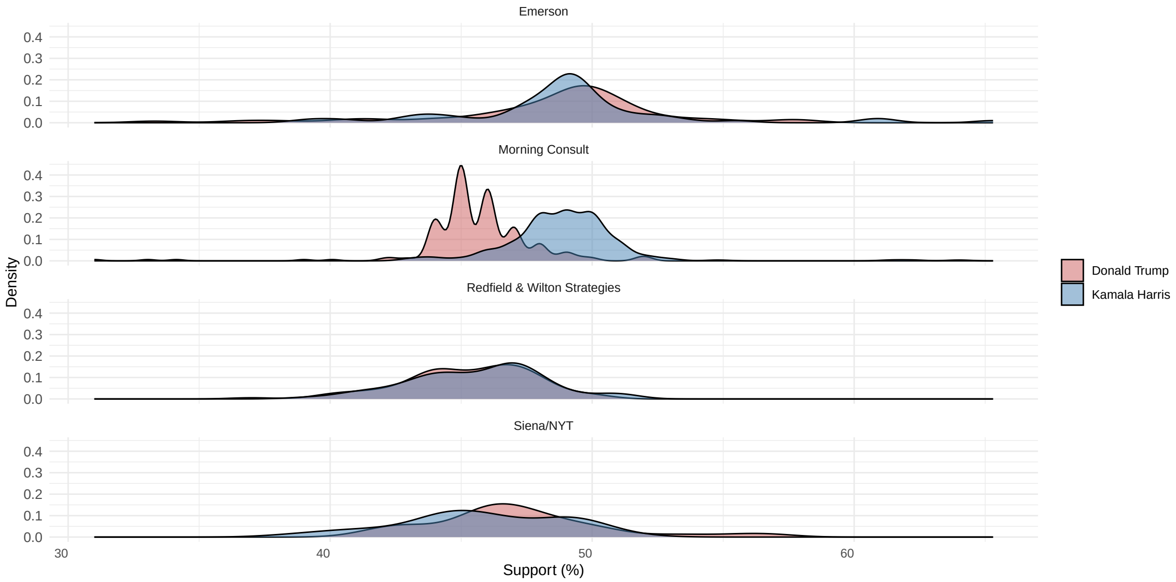

Figure 1: Support distribution for Kamala Harris and Donald Trump by major pollsters, highlighting variability in reported support across organizations and the value of aggregating multiple polls.

Figure 1 shows the distribution of support for Kamala Harris and Donald Trump by pollsters, highlighting variability across different polling organizations. Morning Consult, for example, demonstrates a wide range of reported support, with more spread in Trump's support. In contrast, Siena/NYT shows less variation, with Harris consistently leading. This variability across pollsters shows the importance of aggregating multiple polls to account for organization-specific biases and ensure a more balanced view of candidate support.

2.1.3 Sample Size of Polls

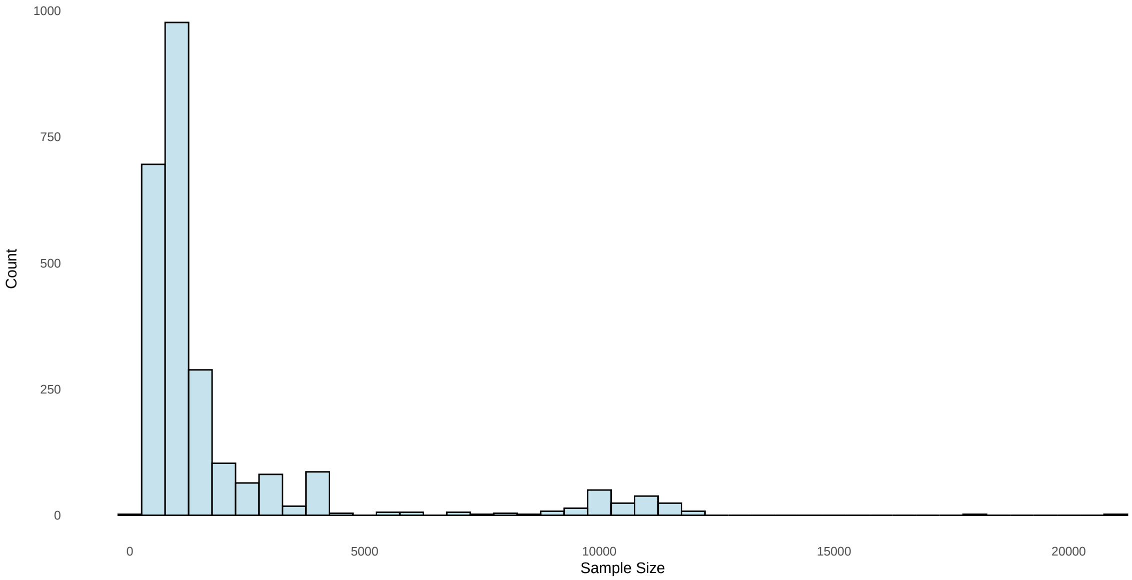

Figure 2: Distribution of Sample Sizes Across Polls: Majority of polls have sample sizes under 5,000, with a few outliers at larger sizes.

Figure 2 shows a clear right-skewed distribution. Most of the sample sizes are clustered between 0 and 3000 respondents, with a sharp peak around 1000-1500 respondents. This indicates that the majority of polls have smaller sample sizes. As sample size increases, the frequency decreases, with few polls conducted with sample sizes larger than 5000, though there are a few outliers with sizes approaching 10,000 or more. This wide range in sample sizes can affect the precision of estimates across different polls.

2.1.4 Distribution of Numeric Grade and Pollscore of Polls

Figure 3 shows the distribution of Numeric Grade (left) and Pollscore (right) across polling organizations. The numerical grade distribution shows that most pollsters are rated between 1.5 and 3, with peaks around 2.0 and 2.5, suggesting a concentration of pollsters with moderate

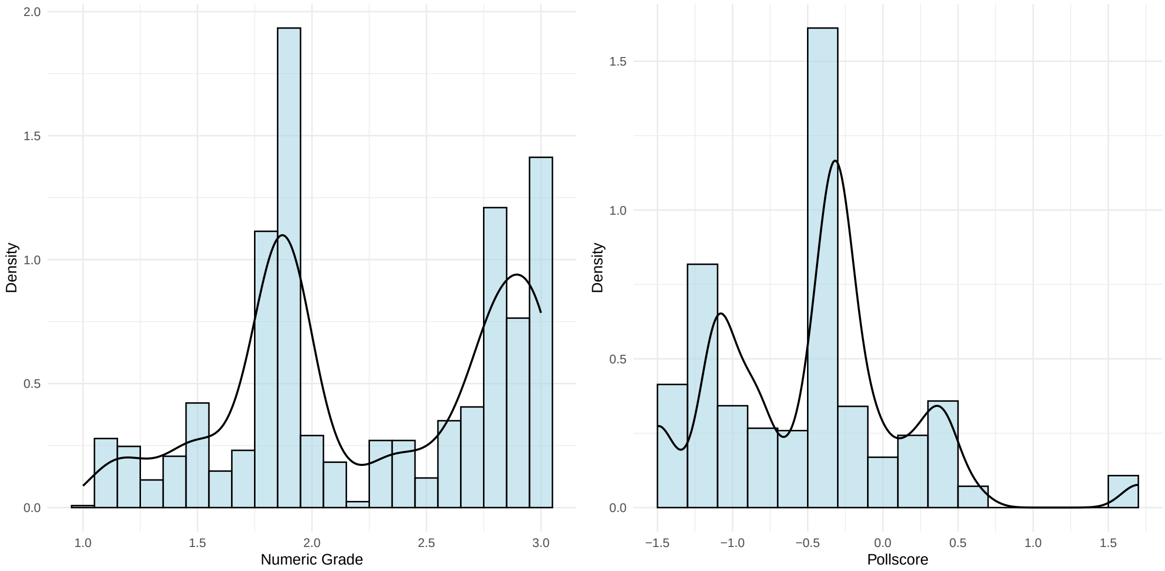

Figure 3: Distribution of Numeric Grade and Pollscore among polling organizations, highlighting variability in pollster reliability and potential bias across polls.

to high reliability scores. In contrast, the Pollscore distribution, where lower values indicate higher reliability, shows a range primarily between -1.5 and 0, with a notable peak around -0.5. This suggests that while many polls demonstrate relatively low bias, there is still variability in reliability across organizations. The distinction between these two metrics emphasizes the need to consider both quality (Numeric Grade) and potential systematic bias (Pollscore) when weighing polls in the model.

2.1.5 Distribution of Polls by Poll Type and Candidate

Figure 4 illustrates the distribution of polls between candidates and poll types. The left chart shows that the majority of polls are conducted at the state level, with a smaller portion being national. The right chart shows that polling is almost evenly split between Kamala Harris and Donald Trump. This distribution underscores the model's balanced approach to capturing state-level differences as well as broader national trends, providing a view of candidate support across different contexts.

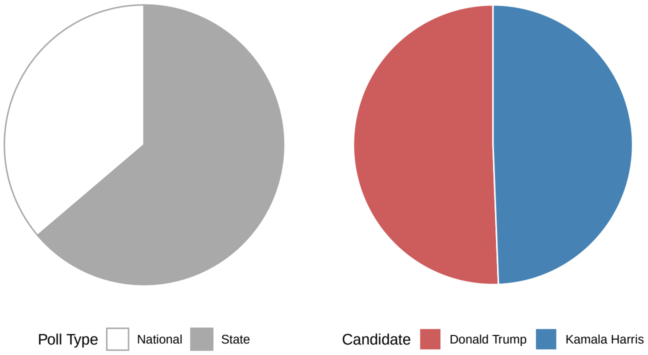

Figure 4: Poll distribution by type (state vs. national) and candidate (Trump vs. Harris), showing a majority of state polls and near-equal coverage for each candidate.

2.1.6 Support Trend For Candidates

Figure 5 shows the support trends for Kamala Harris and Donald Trump from August to October. Each point represents a poll result, color-coded by candidates, with the trend lines highlighting the overall changes in support over time. Kamala Harris's support remains relatively steady but shows slight fluctuations around mid-September, while Donald Trump's support appears to have a small upward trend toward October. The distribution of points is dense throughout, reflecting consistent polling activity, though some dates show more concentrated polling. This visualization indicates that while both candidates maintain stable support levels, minor shifts occur as the election approaches, underscoring the importance of tracking trends over time rather than relying on individual polls.

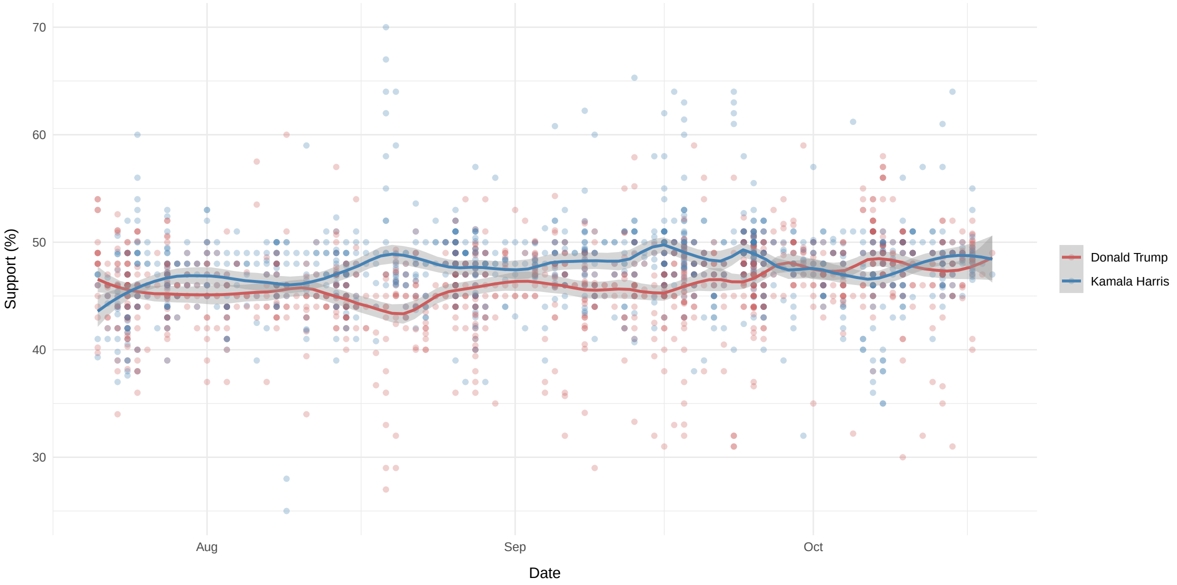

Figure 5: Support trends for Kamala Harris and Donald Trump over time. Trend lines indicate slight shifts in support as the election nears, with consistent polling frequency throughout the period.

2.1.7 Relationship Between Sample Size and Suppoert for Candidates

Figure 6 shows the relationship between sample size and support percentage for Donald Trump and Kamala Harris. For both candidates, the majority of polls have a sample size below 2,500, but there are some larger polls exceeding 10,000 respondents. In Trump's chart, there's a slight downward trend, suggesting that larger sample sizes may show marginally lower support. In contrast, Harris's chart shows a slight upward trend with larger sample sizes, indicating a marginal increase in support with larger poll samples. This highlights the variability in polling support depending on the sample size, emphasizing the importance of including sample size in our model.

2.1.8 Variability of Candidate support Across US States

Figure 7 shows the average support for Kamala Harris as a proportion relative to both Kamala Harris and Donald Trump across the continental U.S. states. States shaded in blue indicate a higher proportion of support for Harris, while red shades represent stronger support for Trump, with color intensity reflecting the support margin. Gray-shaded states lack sufficient polling data.

In swing states, such as Pennsylvania, Michigan, and Arizona, there is a balanced distribution of support, suggesting close competition in these battleground areas. Meanwhile, traditional

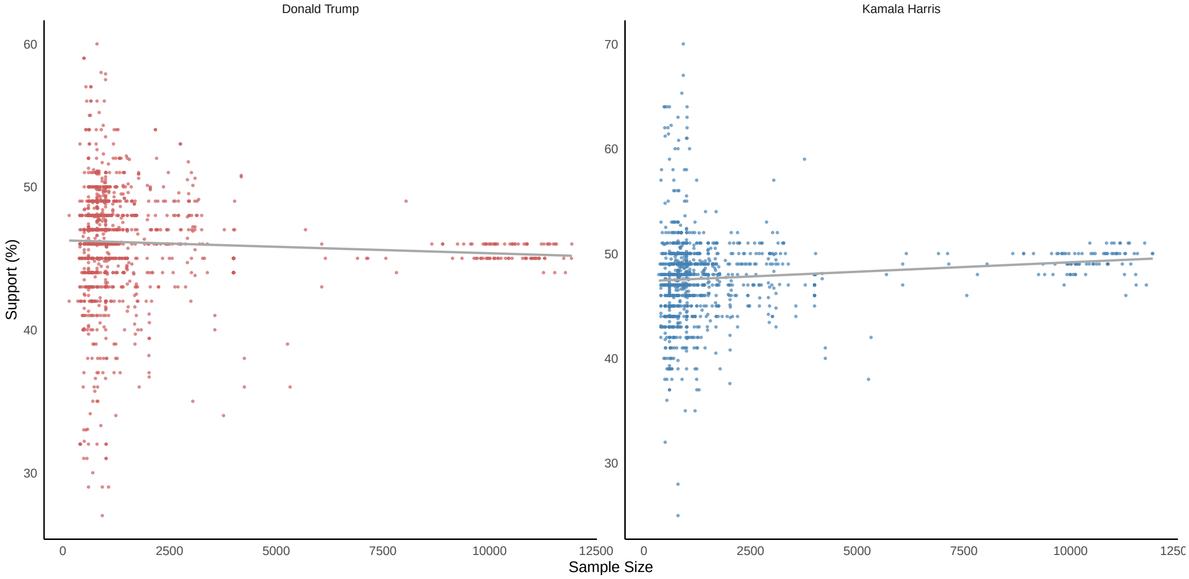

Figure 6: Relationship between Sample Size and Support Percentage for Donald Trump and Kamala Harris, showing slight downward and upward trends in support with increasing sample size for Trump and Harris, respectively.

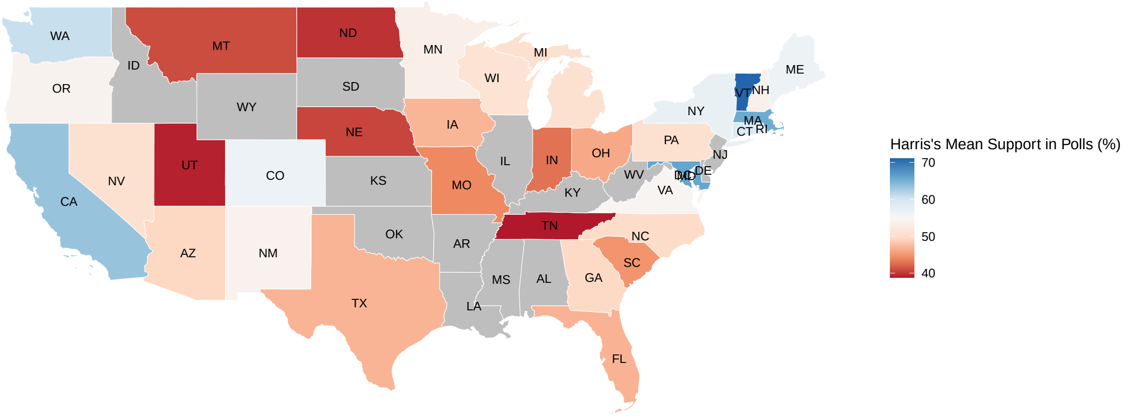

Figure 7: Proportion of support for Kamala Harris relative to Donald Trump across U.S. states. Blue indicates states where Harris has higher support, while red indicates states where Trump leads. Color intensity reflects the magnitude of support difference, with gray indicating insufficient polling data.

Democratic strongholds, like California and New York, show a preference for Harris, with deeper blue shades. Conversely, traditionally Republican states, such as Texas and Tennessee, exhibit higher support for Trump, with intense red shades highlighting his advantage. This visualization emphasizes the varied regional dynamics of candidate support, with distinct patterns in both competitive and historically partisan states.

3 Forecasting Election Outcome through Pooling Polls

3.1 Forecasting Approach

The polls of polls methodology is widely used in election prediction as it aggregates multiple polls to provide a more reliable estimate of voter support, rather than relying on any single poll. The goal is to reduce errors and biases present in individual polls by using a weighted average of many different polls.

In our approach, we will employ linear modeling of voter support percentage (pct) on pollster and other independent variables such as sample size, poll recency, and poll scope (state vs. national). This will allow us to smooth out the inherent noise, biases, and variability across different pollsters. Once we obtain the predicted values from our model, we will weight these predictions based on the numeric grade (quality score) of each pollster to calculate an overall national estimate of the outcome. Additionally, we will separately compute estimates for key battleground states, where voter behavior can be more volatile and pivotal in deciding the outcome of the election. This approach helps us capture both national trends and state-level dynamics.

3.2 Model

In this section, we aim to address the inherent biases and differences present in various polling data to arrive at a robust prediction model. The core challenge lies in selecting a model with an optimal balance between complexity and fit, ensuring it accurately captures the dynamics of polling data while avoiding overfitting. To this end, we carefully evaluated different model specifications to determine the most appropriate one for our forecasting purpose.

Given that variables like numeric grade and poll score are perfectly collinear with the pollster, they were excluded from the regression analysis to avoid multicollinearity issues. These variables, however, remain integral to our weighting strategy, where they will be used to adjust for differences in polling accuracy and reliability. Instead, we focus on key features such as pollster, sample size, state, and recency, gradually adding complexity to the model.

By comparing model specifications that incorporate these variables, we aim to select the model with the right balance between predictive accuracy and generalizability, ultimately providing the best possible forecast.

3.3 Model Set Up

We aim to model the percentage of support for Kamala Harris and Donald Trump in each poll as a function of the pollster the sample size, the state, and the recency of the poll.

$$y_i = \alpha + \beta_1 \cdot \text{pollster}_i + \beta_2 \cdot \text{sample\_size}_i + \beta_3 \cdot \text{recency}_i + \beta_4 \cdot \text{state}_i + \epsilon_i$$

Where

- y_i is the percentage of support for candidate in poll i,
- α is the intercept,
- β_1 captures the effect of the polling organization,
- β_2 captures the effect of the sample size,
- β_3 captures the effect of recency (how recent the poll is),
- β_4 capture the effects of the different states
- ϵ_i represents the error term, assumed to follow a normal distribution with mean 0.

3.4 Model Justification

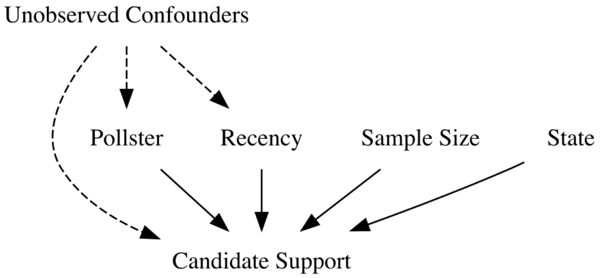

Figure 8: Factors influencing candidate support, with poll attributes (Pollster, Sample Size, State, Recency) as direct contributors and unobserved confounders representing potential biases in polling results.

In our model, we aim to smooth out discrepancies and biases across various polling organizations using a polls-of-polls approach. Given the potential for individual pollsters to introduce systematic differences—due to variations in sampling methods, question phrasing, and historical leanings—our model includes a pollster variable to adjust for these organization-specific biases. This allows us to capture an aggregated view of public support that is less susceptible to the idiosyncrasies of any single poll. Furthermore, we incorporate sample size as a predictor, as polls with larger samples tend to yield more stable and reliable results, reducing random fluctuations caused by smaller samples. The state variable accounts for regional political differences, ensuring that the model captures varying levels of support across geographic and demographic groups, which is important for understanding the political landscape. Additionally, recency is included to prioritize more recent polls, as public opinion can shift in response to political events, and recent data is generally more reflective of current sentiment. By integrating these factors, our model seeks to produce a more stable measure of candidate support, reducing noise from individual poll discrepancies and focusing on a balanced, up-to-date aggregation of polling data.

We opted for a linear model due to its capacity to quantify the marginal effects of each predictor (pollster, sample size, state, and recency) on candidate support. This structure is well-suited to our polls-of-polls approach, as it allows for the estimation of fixed effects that can control for systematic biases across pollsters, while accommodating the influence of sample size and recency as continuous variables. All modeling was conducted using the base R package (R Core Team 2024), specifically utilizing the lm() function from the stats package for linear regression analysis.

3.5 Model Results

Table 3: Model performance summary showing improved fit and accuracy as State and Recency are added, with R^2 increasing from 0.375 in Model 1 to 0.774 in Model 3, and RMSE decreasing from 3.053 to 1.836.

Table 3: Model Summary with Included Variables

Model	Variables	R^2	Adjusted R^2	AIC	BIC	RMSE
Model 1	Pollster, Sample Size	0.375	0.320	6498	7026	3.053
Model 2	Pollster, Sample Size, State	0.720	0.685	5575	6292	2.043
Model 3	Pollster, Sample Size, State, Recency	0.774	0.746	5312	6034	1.836

Table 3 summarizes the performance metrics for three models with progressively added variables. Model 1, which includes only Pollster and Sample Size, achieves an R^2 of 0.375, indicating that these variables alone explain about 37.5% of the variance in candidate support. Model 2 incorporates State as an additional predictor, resulting in a substantial improvement, with an R^2 of 0.720 and a reduction in both AIC and RMSE, showing better model fit and predictive accuracy. Model 3 further adds Recency, which increases the R^2 to 0.774 and decreases the RMSE to 1.836, indicating enhanced explanatory power and prediction accuracy.

This progression highlights the benefit of adding contextual variables like State and Recency to better capture the complexities of polling data.Figure 9 helps us the visualise the model results.

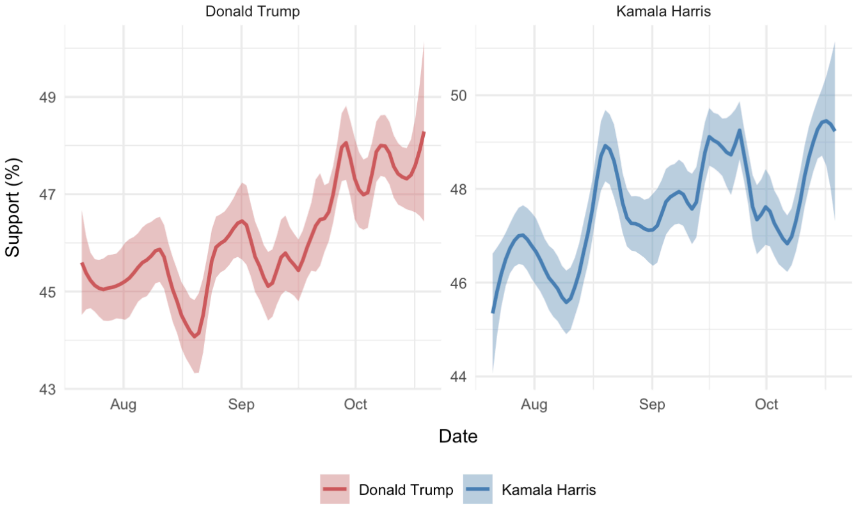

Figure 9: Model summaries for Kamala Harris and Donald Trump at the national level using polling data.

3.6 Prediction

To predict Kamala Harris' overall support, we used a weighted average approach based on the quality of each poll. The weights are calculated using each poll's `numeric_grade`, which reflects the reliability and transparency of the polling methodology.

We define the weight for each pollster w_i as follows:

$$w_i = \frac{\text{numeric\_grade}_i \times (\text{maxPollscore} - \text{pollscore}_i)}{\sum_{i=1}^{n} \text{numeric\_grade}_i \times (\text{maxPollscore} - \text{pollscore}_i)}$$

where:

- w_i represents the weight assigned to poll i,
- $numericgrade_i$ is the numeric grade of poll i, and
- n is the total number of polls used in the analysis.
- $pollscore_i$ is the is the pollscore of poll i which reflects the estimated bias of the poll (with more negative values indicating less bias)

- $maxPollscore$ is the maximum pollscore across all polls.

The weight assigned to each poll combines both its quality (as represented by the numeric grade) and its level of bias (as indicated by the poll score). Since a more negative poll score reflects a lower level of bias, the formula uses the difference between the maximum poll score and each poll's specific poll score. This approach gives more weight to polls that are both highly graded (indicating higher reliability and transparency) and less biased. By combining these two factors, the weighting system emphasizes polls that are reliable and minimally biased, ensuring that they have a stronger influence on the overall calculation. Additionally, the total of all weights is normalized to sum to one, so each poll's weight is proportionate to its quality and relative lack of bias, resulting in a more balanced and accurate average of public support.

Using these weights, the overall weighted prediction of candidate's support is calculated by summing the weighted predicted values from our regression model:

$$\text{Overall Weighted Support} = \sum_{i=1}^{n} w_i \cdot \hat{y}_i$$

- $\hat{y}_i$ is the predicted percentage of support for Kamala Harris from poll i.

3.6.1 Comparing Overall Weighted Support Across All Polls

Aggregating support across all polls and applying our weighted approach, we estimate the overall support for each candidate. The weights, calculated based on both the poll's numeric grade and poll score, ensure that higher-quality polls with less bias contribute more to our estimates. Based on this approach, Kamala Harris has an estimated overall weighted support of around 47.77%, while Donald Trump stands at around 46.31%. This aggregation takes into account the varied methodologies and sampling qualities of different polling organizations.

3.6.2 State-Level Predictions

In this section, we examine the predicted support for Kamala Harris and Donald Trump within each state, based on our weighted approach. By aggregating poll results within each state and applying weights that adjust for poll quality and bias, we can estimate candidate support with a more localized perspective. This allows us to identify state-by-state competition and highlight key battleground states where the margins are close.

For each state, we compare the predicted support percentages and determine the projected winner based on which candidate has higher weighted support. We also calculate the margin of difference between the candidates, which shows how close each state's race is.

Figure 10 presents the predicted winner of the 2024 U.S. Presidential Election by state based on aggregated polling data, with red representing states projected to favor Kamala Harris and

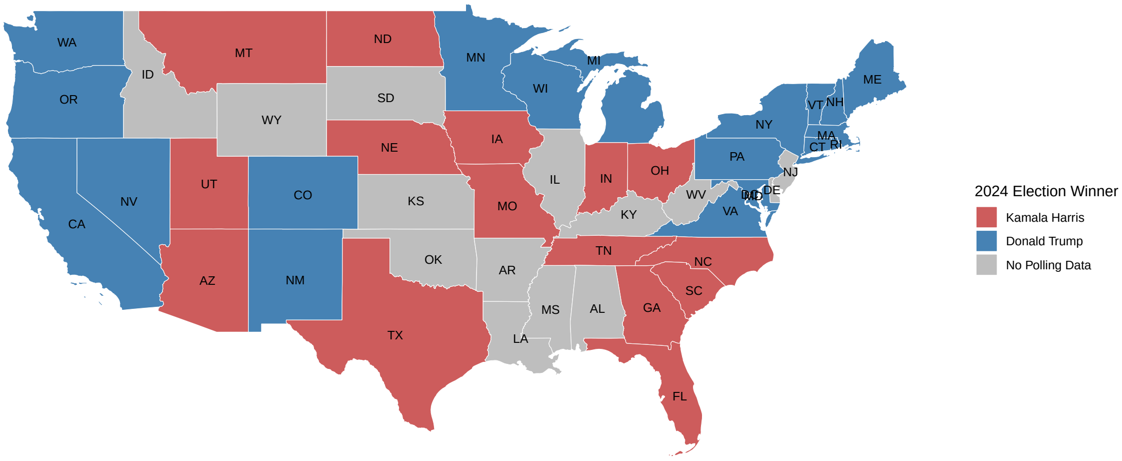

Figure 10: Predicted 2024 U.S. Presidential Election winner by state, with Kamala Harris-leaning states in blue, Donald Trump-leaning states in red, and states lacking sufficient polling data in gray. The map highlights traditional Democratic and Republican strongholds as well as key battleground states.

blue indicating those favoring Donald Trump. Notably, traditional Democratic strongholds in the Northeast and West Coast, such as California, New York, and Washington, show solid support for Harris. Conversely, traditional Republican states, including Texas, Florida, and much of the South, show strong support for Trump.

Swing states, like Pennsylvania, Michigan, and Wisconsin, are predominantly leaning toward Harris, indicating a possible advantage for her in these key battlegrounds, although other typical swing states like Arizona and Georgia lean toward Trump. Gray-shaded states lack sufficient polling data to make a prediction, reflecting areas of data scarcity in the model's projections. This visualization highlights the geographical and political divides across the U.S., as well as the importance of swing states in election outcomes.

Table 4: Predicted support for Kamala Harris and Donald Trump across key battleground states, indicating the winner and the support margin (%) in each state.

State	Kamala Support (%)	Trump Support (%)	Predicted Winner	Support Margin (%)
Arizona	46.89	48.22	Donald Trump	1.33
Georgia	47.10	48.22	Donald Trump	1.12
Michigan	47.45	46.34	Kamala Harris	1.11

Table 4: Predicted support for Kamala Harris and Donald Trump across key battleground states, indicating the winner and the support margin (%) in each state.

State	Kamala Support (%)	Trump Support (%)	Predicted Winner	Support Margin (%)
Nevada	47.49	46.48	Kamala Harris	1.01
North Carolina	47.61	47.87	Donald Trump	0.27
Pennsylvania	48.01	46.97	Kamala Harris	1.04
Wisconsin	48.67	46.51	Kamala Harris	2.17

Many polling websites such as USAFacts (USA Facts 2024) and FiveThirtyEight (FiveThirtyEight 2024) are calling Arizona, Georgia, Michigan, Pennsylvania, Wisconsin, North Carolina and Nevada the 'swing states' of the 2024 presidential election. Swing states are important for election predictions because these states, with their historically close voting margins, often determine the overall outcome by tipping the electoral balance toward one candidate. Table 4 summarizes the predicted support percentages for Kamala Harris and Donald Trump in key battleground states, along with the predicted winner and the support margin. The data shows close margins in these states, with some like Arizona and Georgia leaning towards Trump, while others such as Michigan and Pennsylvania favor Harris by narrow margins. The support margin column highlights the competitiveness of these states, with all margins below 2.2%.

4 Discussion

4.1 Key Findings

This paper predicts the support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election using a "poll-of-polls" approach. By aggregating multiple polls, we reduce the influence of individual surveys' biases and create a more accurate and balanced forecast across candidates. Our model uses predictors, such as pollster, sample size, state, and recency of the polls. We applied a weighting scheme based on each pollster's numeric grade and poll score which takes into consideration reliability and bias. This weighted approach produces a prediction of candidate support that reflects variations across different states and polling organizations.

One main finding from our analysis is the importance of poll recency. Including recency notably increased the model's explanatory power, with improvements in R^2 and reductions in RMSE compared to models without it. This underscores the dynamic nature of public opinion, where recent events can influence voter perceptions and candidate support.

Our state-level analysis highlights close competition observed in key battleground states, emphasizing their pivotal role in the 2024 election. Our model shows that in states like Pennsylvania, Michigan, and Wisconsin, support levels for Harris and Trump are nearly tied, reflecting the intense contest for these electoral votes. The predictions show that even minor shifts in support within these swing states could decisively impact the overall election outcome. This analysis shows the importance of regional polling in capturing voter sentiment and the heightened influence that battleground states hold in shaping the final results.

4.2 Polling Uncertainty in 2024 Election

Polling provides a general sense of public opinion, but it shouldn't be viewed as a guaranteed predictor of election outcomes. Past U.S. elections show why polls can be misleading. For example, in 2016, polls showed Hillary Clinton in the lead. According to national polls, Clinton would have won the popular vote by about 3.2% (Durand et al. 2017). However, Donald Trump won battleground states, surprising many who relied on those predictions (Silver 2024). Trump won the electoral vote with 306 votes, while Clinton had 232 (Durand et al. 2017).

In 2020, 93% of national polls overestimated support for President Joe Biden (Kennedy et al. 2021). Biden's lead was especially overestimated in states like Florida and Wisconsin, where Trump performed better than expected (Edwards-Levy 2020). These examples highlight the limitations of polls, especially when they can't capture late shifts in voter preferences, changes in turnout, and the difficulty of reaching a representative sample (Pew Research Center 2024). While polls can show trends, they are not always reliable for predicting what will happen on election day. ## Trump and Harris' Path to Victory

Both Donald Trump and Kamala Harris have potential paths to victory in the 2024 election. For Trump, one factor is the "shy voter" effect, where some supporters may be hesitant to share their views with pollsters due to social pressures. This can lead to polls underestimating his actual support (Woodie 2024). Additionally, Trump's base includes groups that are harder to reach through standard polling, such as rural voters and individuals who avoid mainstream media (Brown 2024). These voters may be underrepresented in polls but can play an important role in election outcomes.

On the other hand, Kamala Harris could secure a win if her campaign mobilizes specific demographics, including young people, women, and minority voters. As the first Black and South Asian woman on a presidential ticket, Harris's candidacy may energize these groups and boost turnout among historically Democratic-leaning voters (Pew Research Center 2024). The growing diversity and youth of the electorate could also benefit her, especially in urban areas and battleground states (Blow 2022). With strong efforts in voter registration and mobilization, Harris could turn polling support into actual votes, providing her with a strong chance at victory if turnout aligns with her campaign's goals.

4.3 Weaknesses and Future Directions

Our model assumes linear relationships between predictors and candidate support, which might not capture more complex or interaction-based trends. Additionally, while we weighted polls based on their numeric grades, these scores may not fully reflect each pollster's accuracy, leaving room for improvement. Future research could explore non-linear methods, such as machine learning, to capture potential interactions between variables. Refining the weighting mechanism by considering pollster performance history or state-specific polling differences could further enhance predictive accuracy.

In conclusion, this paper contributes to election forecasting by providing a weighted estimate of candidate support across both national and state levels. The evolving nature of voter sentiment and poll accuracy suggests the need for adaptive models that take into consideration ongoing changes in public opinion.

Appendix

A Data Cleaning and Modeling

A.1 Data Cleaning

The raw data for this project, sourced from FiveThirtyEight, (FiveThirtyEight 2024) underwent a series of cleaning steps to prepare it for analysis. Initially, duplicate rows were removed to ensure that unique observations remained, facilitated by the `janitor` package (Firke 2023). A new binary variable, 'national', was created to indicate whether a poll was conducted at the national or state level. Missing values in the 'state' column were replaced with "Not Applicable," and numeric grades were evaluated to filter out low-quality pollsters, keeping only those with a numeric grade above 1. This cutoff was selected to retain mid to high-level pollsters for more reliable results. These steps were performed using functions from the `tidyverse` package (Wickham et al. 2019). Dates were also standardized and converted into a proper format for analysis using the same package. Polls related to Kamala Harris were retained for further analysis, and percentage support values were transformed into actual numbers of supporters based on sample size. Pollster counts below five were excluded to focus on more reliable data sources. Polls regarding Kamala Harris were filtered to include only those conducted after her official candidacy announcement on July 21, 2024, ensuring the data reflects post-announcement public sentiment. The cleaned dataset was saved in Parquet format for efficient storage and retrieval, using the `arrow` package (Richardson et al. 2024).

A.2 Model Diagnostics

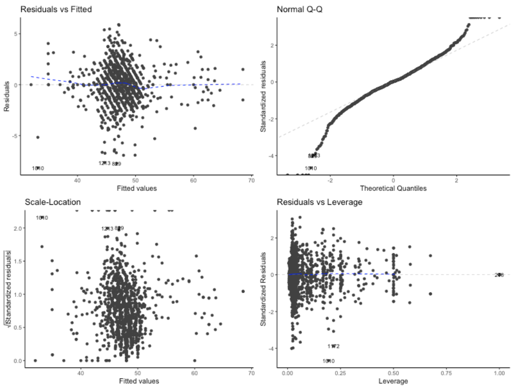

Figure 11: Diagnostic plots for Model 3

The diagnostic plots for Model 3 are given in Figure 11. We notice the following:

1. **Residuals vs. Fitted Plot**: This plot checks for non-linearity and heteroscedasticity. The residuals are scattered around the horizontal axis with no clear pattern, indicating that the linearity assumption is reasonable. However, there is a slight spread around the center, suggesting some mild heteroscedasticity.
2. **Normal Q-Q Plot**: This plot assesses the normality of residuals. Most points align along the diagonal line, although there are some deviations at the tails. This suggests that the residuals are approximately normally distributed, with minor deviations in the extremes.
3. **Scale-Location Plot**: This plot further checks for homoscedasticity. The residuals appear to be evenly spread across the fitted values, supporting the homoscedasticity assumption, although slight deviations are present in certain regions.
4. **Residuals vs. Leverage Plot**: This plot identifies potential influential points. While most points have low leverage, a few points exhibit higher leverage, as indicated by their distance from the center. However, no points exceed Cook's distance threshold, indicating no extreme outliers that would unduly influence the model.

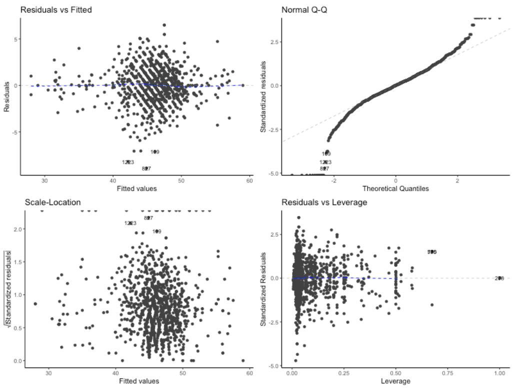

Figure 12: Diagnostic plots for Model 6

The diagnostic plots for Model 6 are given in Figure 12. We notice the following:

1. **Residuals vs Fitted**: This plot suggests a fairly random spread of residuals around zero, indicating that the linearity assumption holds reasonably well. However, there is a slight clustering of points in the center, which could indicate minor heteroscedasticity but is not severe.
2. **Normal Q-Q Plot**: The Q-Q plot shows that the residuals generally follow a normal distribution, with only a few deviations at the tails. This suggests that the normality assumption is mostly met, though some minor deviations in the upper tail indicate possible outliers.

3. **Scale-Location (Spread-Location) Plot**: The residuals appear randomly dispersed with no clear pattern, supporting the homoscedasticity (constant variance) assumption. However, some observations near the upper range might indicate slight variance inconsistency, though it is minimal.

4. **Residuals vs Leverage**: This plot does not show any influential outliers with high leverage that might impact the model unduly. The Cook's distance lines show that no data points exceed these thresholds, indicating that no individual observation is disproportionately influencing the model.

Overall, the diagnostic plots for both models suggest that they meet the assumptions for linear regression fairly well, with minor deviations that are not expected to severely impact the model's validity.

B Code Styling

The code in this paper was reviewed and formatted for consistency using lintr (Hester et al. 2024) and styler (Müller and Walthert 2024), ensuring readability and adherence to style standards.

C Reproducibility

To replicate the findings presented in this paper, users should execute the scripts available in the GitHub repository. Begin by running the 00-install_packages.R script, which installs all required packages for the analysis.

D Acknowledgments

We extend our gratitude to (Alexander 2023), which provided invaluable guidance in establishing a reproducible workflow and inspired many of the code structures used in this paper.

References

Aguinis, Herman, and Steven Culpepper. 2024. *Improving Our Understanding of Predictive Bias in Testing*. https://psycnet.apa.org/fulltext/2024-15734-001.html.

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

Blow, Charles. 2022. "Young Voters Are Frustrated. They're Staying Engaged 'Out of Sheer Self-Defense." *The New York Times*. https://www.nytimes.com/2023/09/20/opinion/young-voters-2024.html.

Brown, Justin. 2024. "Why the Polls Might Be Wrong — in Kamala Harris' Favor." *Politico*. https://www.politico.com/news/magazine/2024/11/01/shy-kamala-harris-voters-polling-00186653.

Durand, Claire, Courtney Kennedy, Mark Blumenthal, Scott Clement, Joshua Clinton, Charles Franklin, Kyley McGeeney, and Lee Miringoff. 2017. "An Evaluation of 2016 Election Polls in the United States." https://www.researchgate.net/publicatio/317170048_AN_EVALUATION_OF_2016_ELECTION_POLLS_IN_THE_U_NITED_STATES.

Edwards-Levy, Ariel. 2020. "How Well Can Polls Predict Who'll Win the Election?" *HuffPost*. https://www.huffpost.com/entry/2020-polls-predict-election_n_5f453ff1c5b60c7ec416c6e8.

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

FiveThirtyEight. 2024. "FiveThirtyEight U.S. Election Polls." https://projects.fivethirtyeight.com/polls/.

Hester, Jim, Florent Angly, Russ Hyde, Michael Chirico, Kun Ren, Alexander Rosenstock, and Indrajeet Patil. 2024. *lintr: A 'Linter' for r Code*. https://CRAN.R-project.org/package=lintr.

Kennedy, Courtney, Jesse Lopez, Scott Keeter, Arnold Lau, Nick Hatley, and Nick Bertoni. 2021. "Confronting 2016 and 2020 Polling Limitations." https://www.pewresearch.org/methods/2021/04/08/confronting-2016-and-2020-polling-limitations/.

Morris, Elliott. 2024a. *How 538's Pollster Ratings Work*. https://abcnews.go.com/538/538s-pollster-ratings-work/story?id=105398138.

———. 2024b. *Trump Leads in Swing-State Polls and Is Tied with Biden Nationally*. https://abcnews.go.com/538/trump-leads-swing-state-polls-tied-biden-nationally/story?id=109506070.

Müller, Kirill, and Lorenz Walthert. 2024. *styler: Non-Invasive Pretty Printing of r Code*. https://CRAN.R-project.org/package=styler.

Pew Research Center. 2024. "Harris Energizes Democrats in Transformed Presidential Race." https://www.pewresearch.org/politics/2024/08/14/harris-energizes-democrats-in-transformed-presidential-race/.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Renewable Energy. 2012. *Prediction Error*. https://www.sciencedirect.com/topics/engineering/prediction-error.

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.

Silver, Nate. 2024. "Nate Silver: Here's What My Gut Says about the Election, but Don't Trust Anyone's Gut, Even Mine." *The New York Times*. https://www.nytimes.com/2024/10/23/opinion/election-polls-results-trump-harris.html.

USA Facts. 2024. "What Are the Current Swing States and How Have They Changed Over Time?" https://usafacts.org/articles/what-are-the-current-swing-states-and-how-have-they-changed-over-time/.

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.

Woodie, Alex. 2024. "After Misses in 2016 and 2020, Have We Fixed Political Polling in 2024?" *Big Data Wire*. https://www.bigdatawire.com/2024/10/31/after-misses-in-2016-and-2020-have-we-fixed-political-polling-in-2024/.