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

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

Shamayla Durrin Denise Chang Krishna Kumar

November 1, 2024

This study employs a polls-of-polls approach to predict support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election, with a focus 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 a higher national support for Trump relative to Harris. Our analysis also reveals a close competition in battleground states, with Trump holding a narrow lead in Arizona, Georgia, and North Carolina, while Harris leads in Michigan, Nevada, Pennsylvania, and Wisconsin. These findings underscore the value of aggregating polls over relying on individual surveys, offering a more comprehensive and 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 aggregating data from multiple polling sources, reducing individual poll biases, and improving overall prediction accuracy. Our analysis focuses on estimating the levels of support for each candidate not only nationally but also in key swing states that 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 estimated national support stands close to 48%, while Trump's support is slightly higher, around 49%. We found that Trump holds a marginal lead in battleground states like Arizona and Georgia, with a support margin slightly above 1%. In contrast, Harris shows narrow leads in Michigan, Nevada, Pennsylvania,

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

and Wisconsin, with her support margin in Wisconsin reaching over 2%, making it her strongest battleground. North Carolina emerges as one of the tightest races, with Trump leading by only 0.26%, underscoring the competitive nature of these regions.

This paper contributes to the field of election forecasting by emphasizing the importance of poll quality and recency in prediction models. By focusing on battleground states where the support margins are particularly close, our approach highlights the areas where campaign strategies and voter turnout efforts may have the greatest impact. This could provides insights for political strategists, media analysts, and the general public by offering a nuanced view of the electoral landscape.

This paper is organized as follows: In **??**, we clean the data, explore summary statistics, plot distributions of key variables, and examine relationships between variables. In **??**, 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. Finally, in **??**, we address the broader implications of our findings, acknowledge limitations, and suggest directions for future work. # Data {#sec-data} ## Measurement

In the dataset of our analysis, the process of measurement begins by capturing real-world public opinion through surveys. Polling organizations transform this observed phenomenon into structured data by recording support percentages, pollster details, and methodological choices, thus turning public sentiment into quantifiable entries in the dataset.

Public opinion polls are essential for understanding voter preferences and electoral dynamics, providing data that informs policy decisions and campaign strategies. Our data set captures public support for candidates at specific times, along with information on methodologies, sample sizes, and pollster ratings.

Polling organizations use various methods, such as App panels, IVR (Interactive Voice Response), Live phone calls, Text-to-web systems etc. These methods can influence the validity and reliability of results, with larger sample sizes generally yielding more reliable estimates. Some methods, like live phone interviews, may be seen as more trustworthy than others.

The dataset also includes important measures of pollster performance, such as numeric grades, poll scores, and transparency scores. These metrics are calculated by considering bias, race difficulty, predictive error, and transparency, allowing us to assess the accuracy of each poll. The "pollscore," for instance, reflects a pollster's past performance and reliability, adjusted for potential biases.

## 1.1 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 only 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). Furthermore, dates were standardized and converted into a proper format for analysis using the `lubridate` package (Grolemund and Wickham 2011). 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. Additionally, 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).

## 1.2 Explorations of Variables of Interest

### 1.2.1 Summary Statistics of Key Variables

For the purposes of analysis in this paper, only a subset of variables was selected. Two additional variable, national and state, was created using the existing 'state' and 'end date' variable. A short description of each variable of interest 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.
- *Numeric Grade*: A numeric rating given to the pollster, representing the quality or reliability of the organization (e.g., 3.0), with higher grades indicating more reliable pollsters.
- *Pollscore*: A score that reflects the reliability of each pollster, capturing their historical accuracy and any systematic bias. Negative values indicate better predictive accuracy.
- *State*: The U.S. state where the poll was conducted, allowing for analysis of regional differences in candidate support.
- *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.
- *Sample Size*: The total number of respondents in the poll.
- *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. It was 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 (%) | 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 | 20 762.0 |
| 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 |

From Table **??** we see that the average poll in our dataset has a numeric grade of 2.2, suggesting that most polls are of moderate to high quality in terms of reliability. The mean pollscore of -0.5 indicates a general trend toward negative values, which suggests these polls are relatively accurate, as more negative values imply reduced bias. Additionally, the large average sample size of 1908 respondents across polls provides a stable basis for our model, reducing variability and enhancing the reliability of predictions for candidate support.

### 1.2.2 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 **??** displays the top 5 most frequent pollsters in the dataset, with each pollster's total poll count, average pollscore (measuring reliability and historical accuracy), and average numeric grade (indicating overall quality). We observe that the most frequent pollster, Morning Consult, has a moderate pollscore and a slightly above-average numeric grade, indicating it provides fairly reliable data. However, there is variability in pollscore and numeric grade across the top pollsters, reflecting differences in quality and potential biases among them, which the model needs to adjust for to achieve accurate predictions.

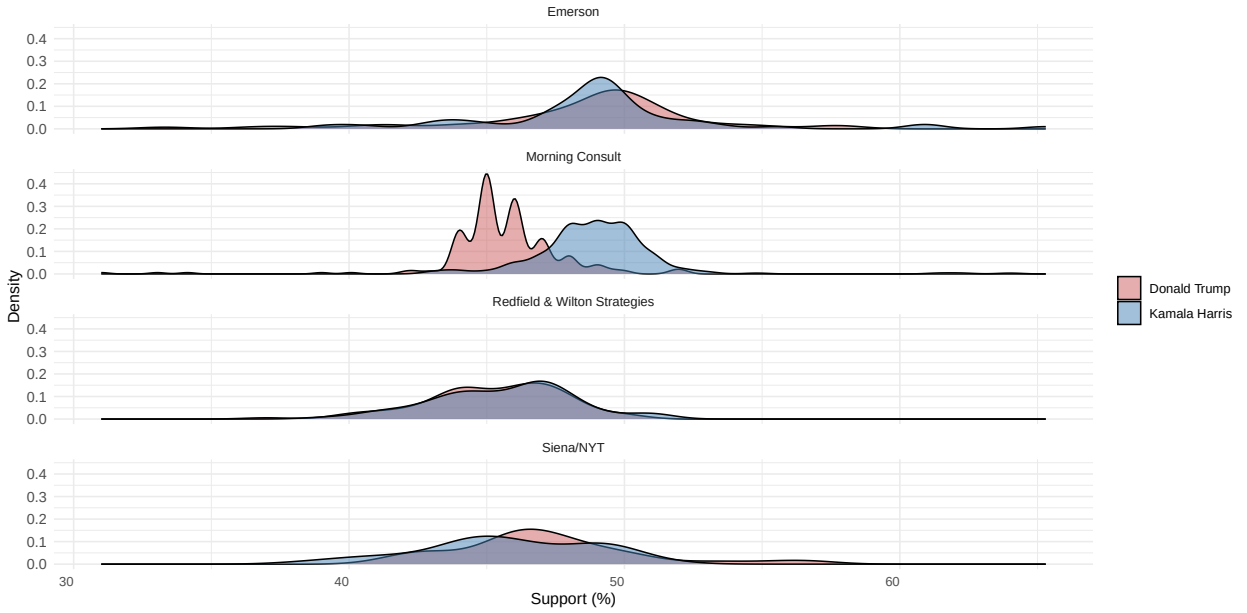


Figure 1: Support distribution for Kamala Harris and Donald Trump by major pollsters, highlighting variability in reported support across organizations and the value of aggregating multiple polls for balanced insights.

Figure **??** shows the distribution of support for Kamala Harris and Donald Trump by pollster, highlighting variability across different polling organizations. Morning Consult, for example, demonstrates a wide range in reported support, with more spread in Trump's support. In contrast,

Siena/NYT shows less variation, with Harris consistently leading. This variability across pollsters emphasizes the importance of aggregating multiple polls to account for organization-specific biases and ensure a more balanced view of candidate support.

### 1.2.3 Sample Size of Polls

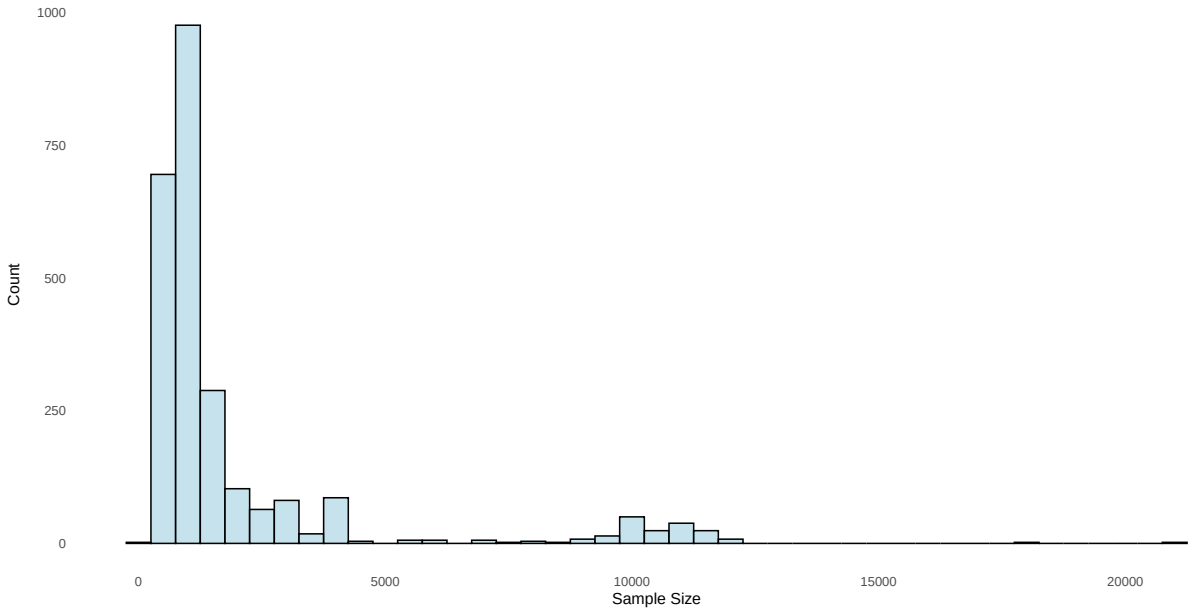


Figure 2: Distribution of Sample Sizes Across Polls: Majority of polls have sample sizes under 5,000, with a few outliers at larger sizes.

Figure **??** reveals a clear right-skewed distribution. Most of the sample sizes are clustered between 0 and 3000, with a sharp peak around 1000-1500. This indicates that the majority of polls are conducted with smaller sample sizes. As sample size increases, the frequency significantly drops, with very 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 might affect the precision of estimates in different polls.

### 1.2.4 Distribution of Numeric Grade and Pollscore of Polls

Figure **??** displays the distribution of Numeric Grade (left) and Pollscore (right) across polling organizations. The Numeric Grade distribution shows that most pollsters are rated between 1.5 and 3, with peaks around 2.0 and 2.5, indicating a concentration of pollsters with moderate 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 weighting polls in the model.

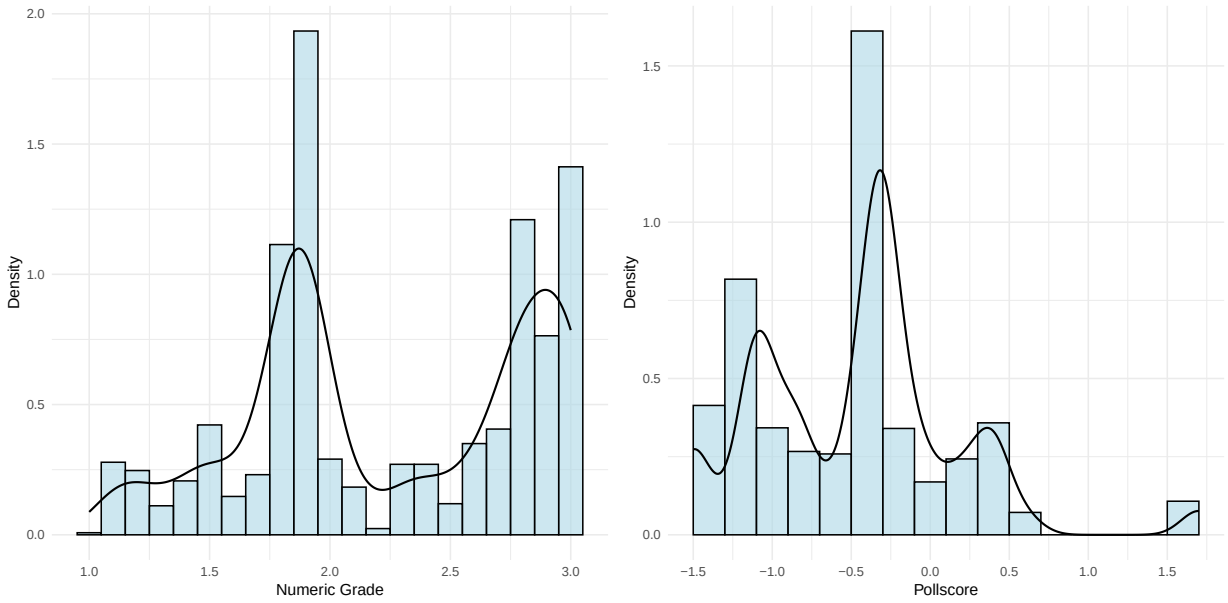


Figure 3: Distribution of Numeric Grade and Pollscore among polling organizations, highlighting variability in pollster reliability and potential bias across polls.

### 1.2.5 Distribution of Polls by Poll Type and Candidate

Figure **??** charts illustrate 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 nuances as well as broader national trends, providing a comprehensive view of candidate support across different contexts

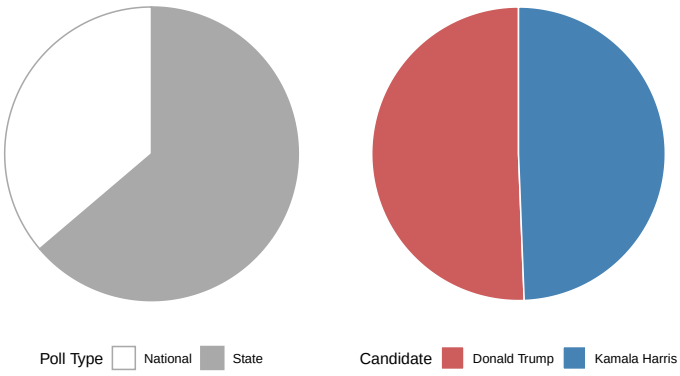


Figure 4: Poll distribution by type (state vs. national) and candidate (Trump vs. Harris), showing a majority of state polls and near-equal coverage for each candidate.

### 1.2.6 Support Trend For Candidates

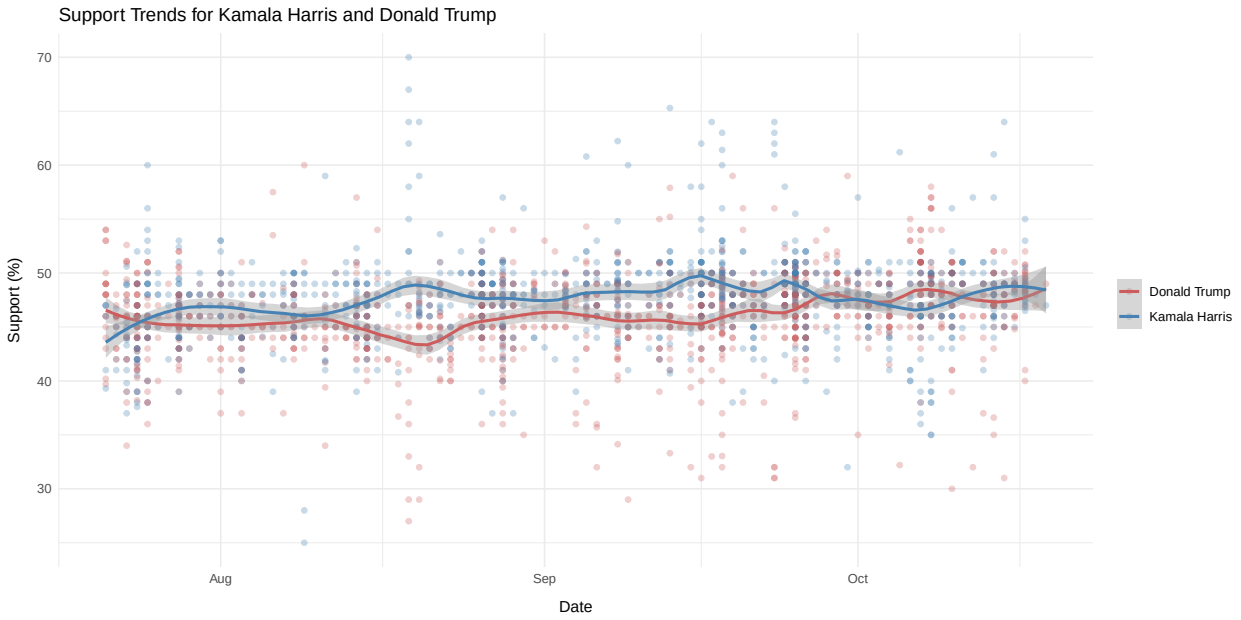


Figure 5: Figure showing support trends for Kamala Harris and Donald Trump over time. Trend lines indicate slight shifts in support as the election nears, with consistent polling frequency throughout the period.

Figure **??** shows the support trends for Kamala Harris and Donald Trump from August to October. Each point represents a poll result, color-coded by candidate, 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.

### 1.2.7 Relationship Between Sample Size and Suppoert for Candidates

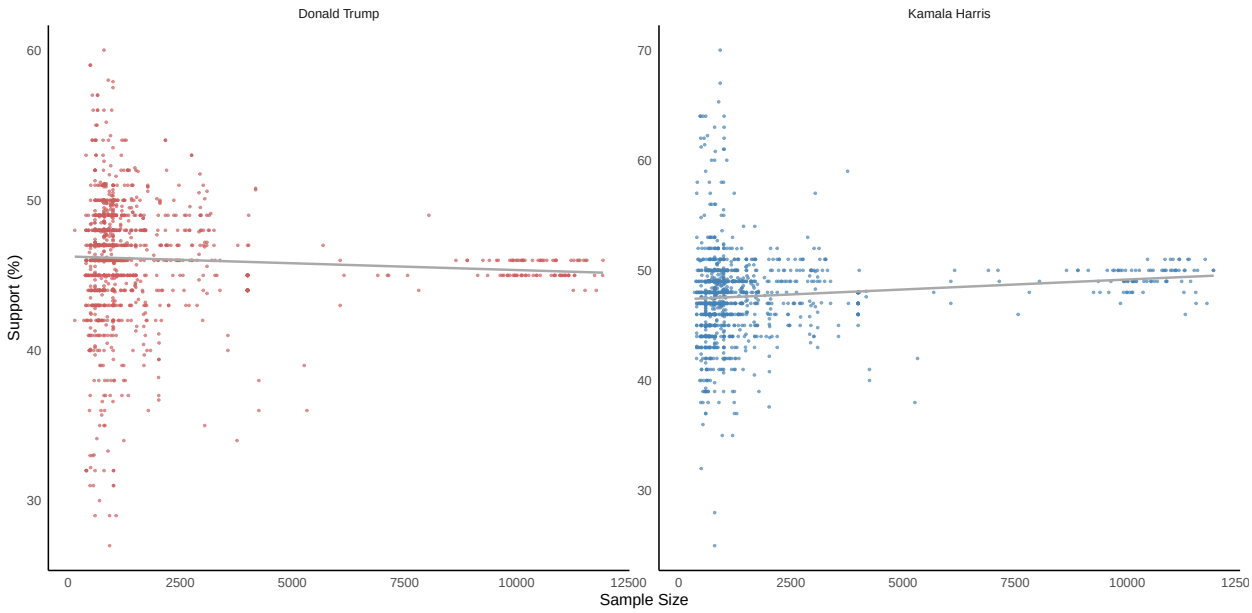


Figure 6: Relationship between Sample Size and Support Percentage for Donald Trump and Kamala Harris, showing slight downward and upward trends in support with increasing sample size for Trump and Harris, respectively.

Figure **??** 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, emphasiszing the importance of including sampel size in our model.

### 1.2.8 Variability of Candidate support Across US States

Figure **??** 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 key swing states, such as Pennsylvania, Michigan, and Arizona, we observe a balanced distribution of support, suggesting close competition in these battleground areas. Meanwhile, traditional Democratic strongholds, like California and New York, show a clear 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.

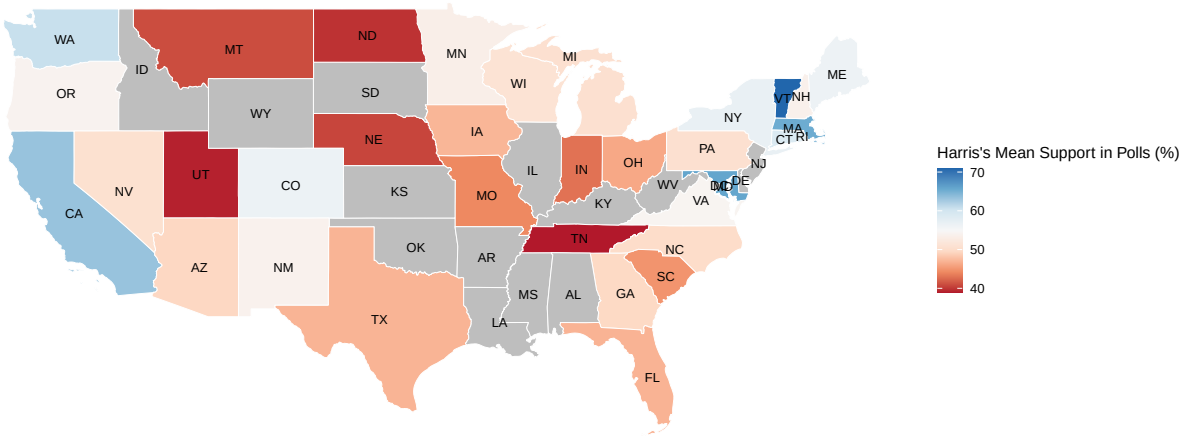


Figure 7: Map showing the proportion of support for Kamala Harris relative to Donald Trump across U.S. states. Blue indicates states where Harris has higher support, while red indicates states where Trump leads. Color intensity reflects the magnitude of support difference, with gray indicating insufficient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

# 2 Forecasting Election Outcome through Pooling Polls

## 2.1 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 final outcome of the election. This approach helps us capture both national trends and critical state-level dynamics.

## 2.2 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 pollscore are perfectly collinear with 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 systematically 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.

## 2.3 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 iii,
- $\alpha$ is the intercept,
- $\beta_1$ captures the effect of the polling organization,
- $\beta_2$ captures the effect of the sample size,
- $\beta_3$ captures the effect of recency (how recent the poll is),
- $\beta_4$ capture the effects of the different states
- $\epsilon_i$ represents the error term, assumed to follow a normal distribution with mean 0.

## 2.4 Model Justification

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 crucial for understanding the nuanced political landscape. Additionally, recency is included to prioritize more recent polls, as public opinion can shift rapidly in response to political events, and recent data is generally more reflective of current sentiment. By integrating these factors, our model seeks