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

October 31, 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 {#sec-data}, we clean the data, explore summary statistics, plot distributions of key variables, and examine relationships between variables. In {#sec-model}, 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 {#sec-discussion}, 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 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

From Table 1 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 2 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.

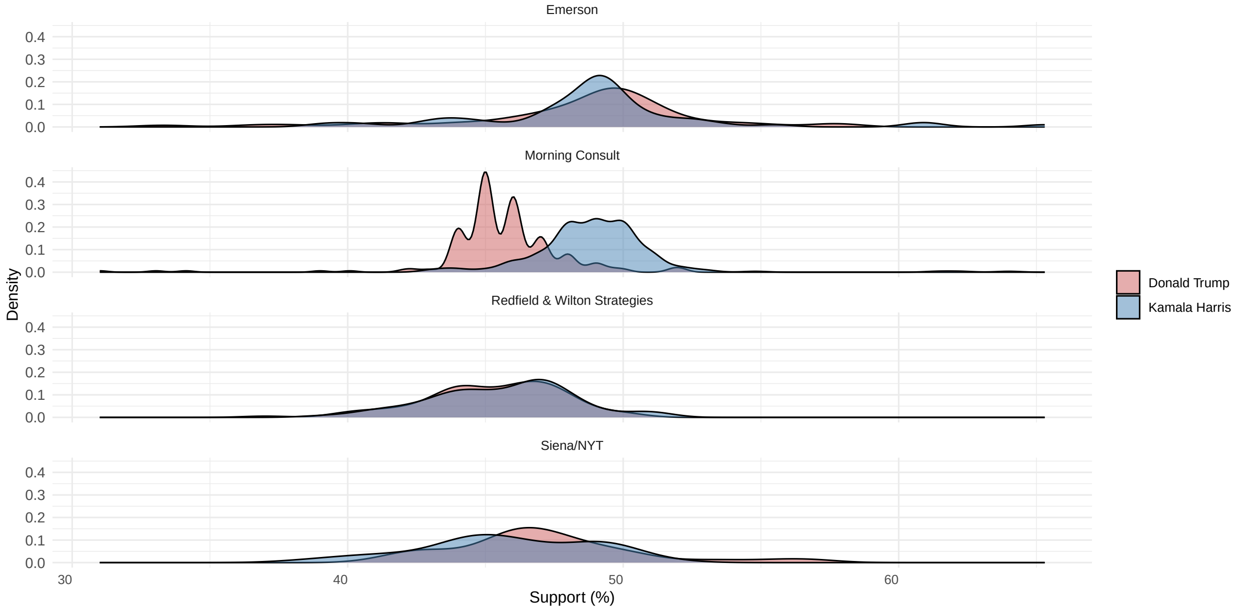

Figure 1: Support distribution for Kamala Harris and Donald Trump by major pollsters, highlighting variability in reported support across organizations and the value of aggregating multiple polls for balanced insights.

Figure 1 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

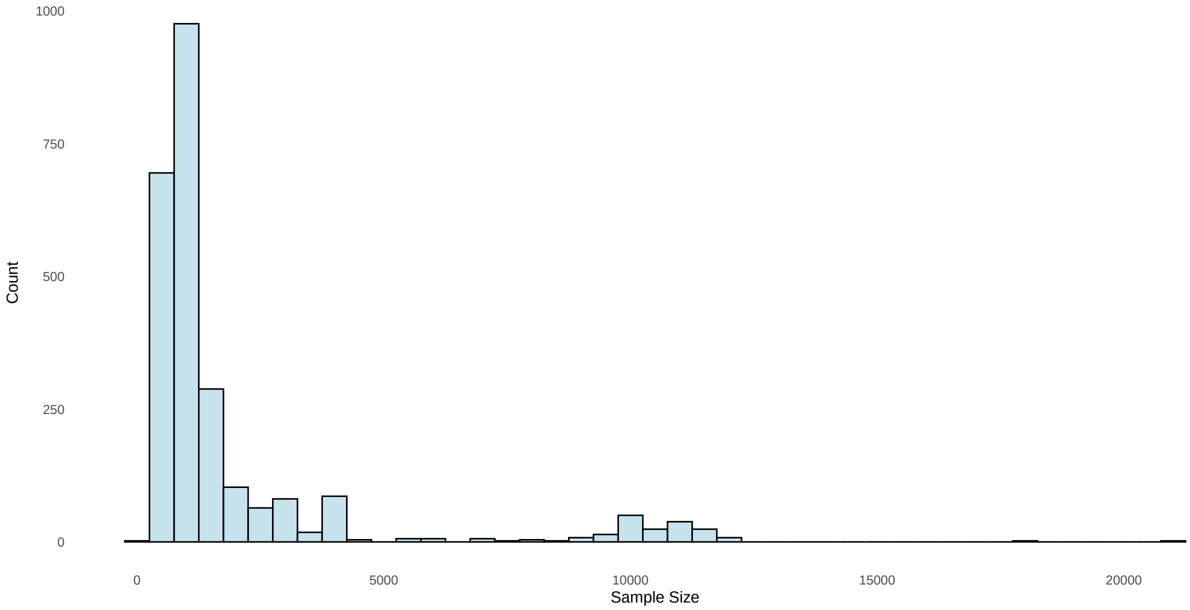

Figure 2: Distribution of Sample Sizes Across Polls: Majority of polls have sample sizes under 5,000, with a few outliers at larger sizes.

Figure 2 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 3 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.

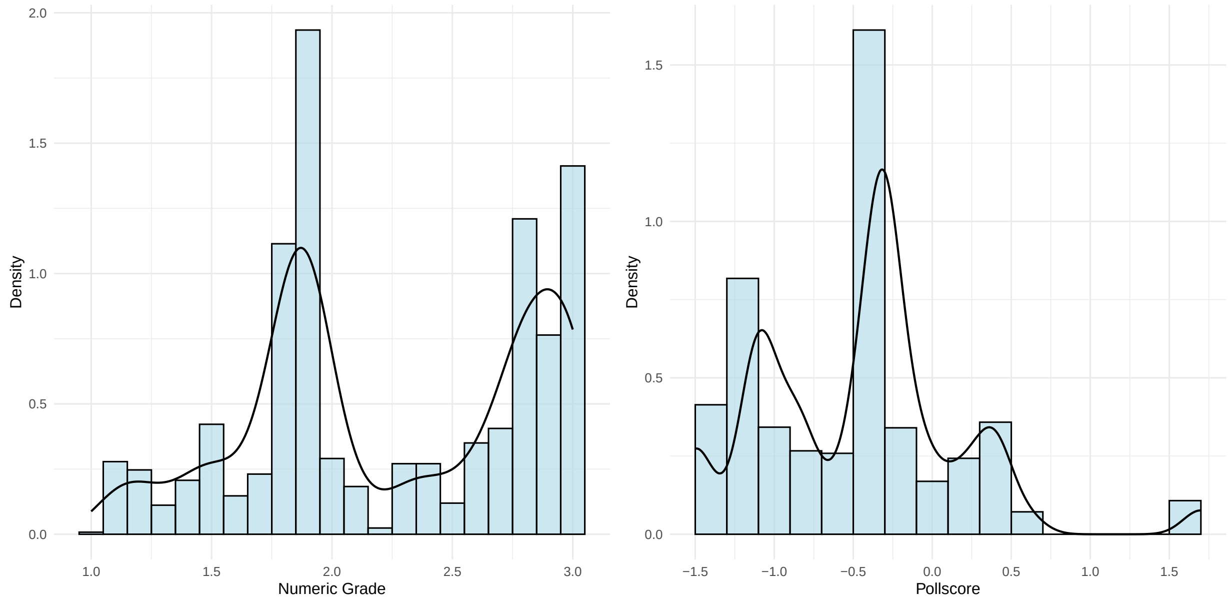

Figure 3: Distribution of Numeric Grade and Pollscore among polling organizations, highlighting variability in pollster reliability and potential bias across polls.

1.2.5 Distribution of Polls by Poll Type and Candidate

Figure 4 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

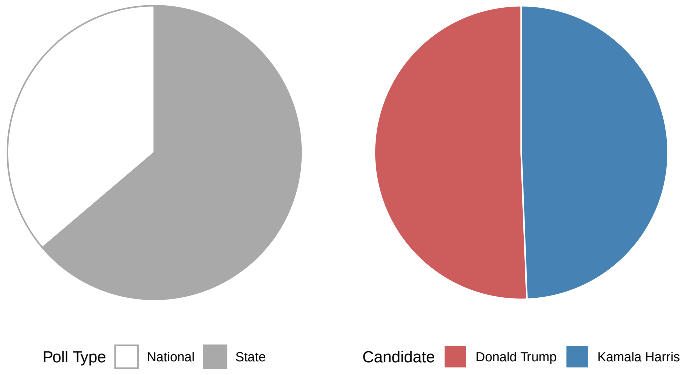

Figure 4: Poll distribution by type (state vs. national) and candidate (Trump vs. Harris), showing a majority of state polls and near-equal coverage for each candidate.

1.2.6 Support Trend For Candidates

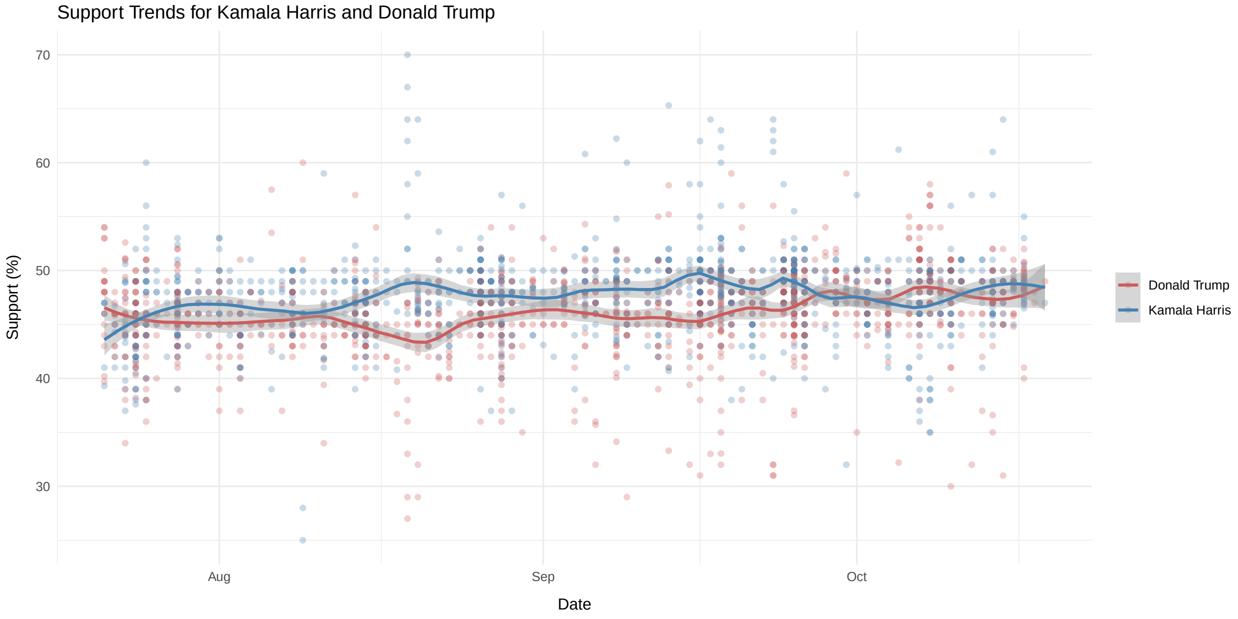

Figure 5: Figure showing support trends for Kamala Harris and Donald Trump over time. Trend lines indicate slight shifts in support as the election nears, with consistent polling frequency throughout the period.

Figure 5 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

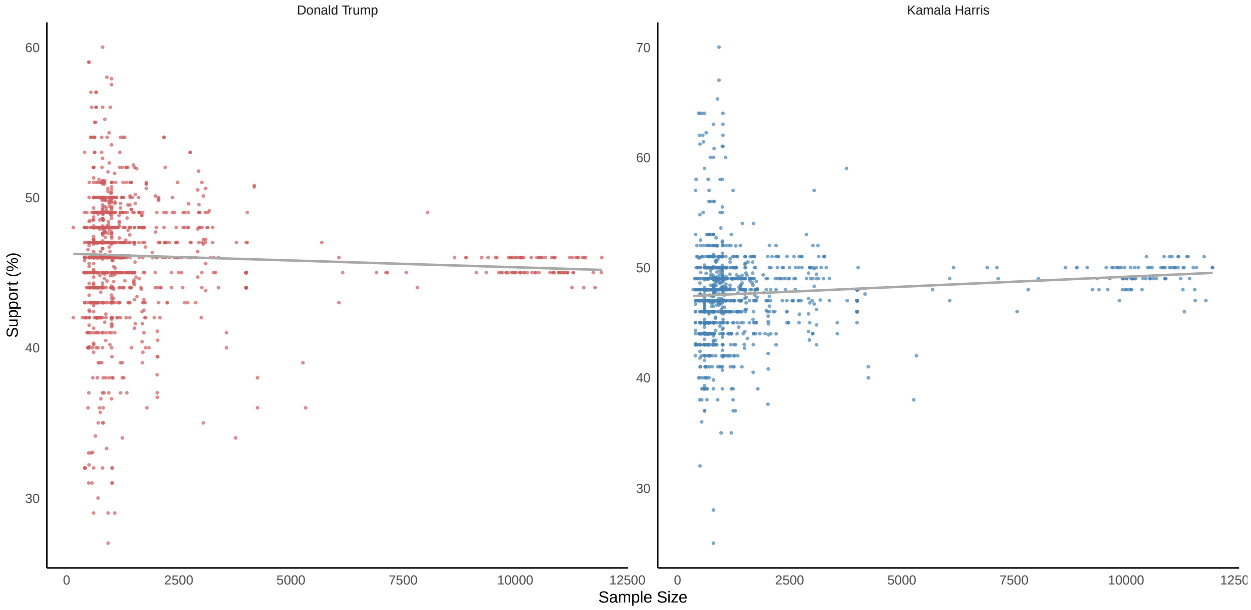

Figure 6: Relationship between Sample Size and Support Percentage for Donald Trump and Kamala Harris, showing slight downward and upward trends in support with increasing sample size for Trump and Harris, respectively.

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

1.2.8 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 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.

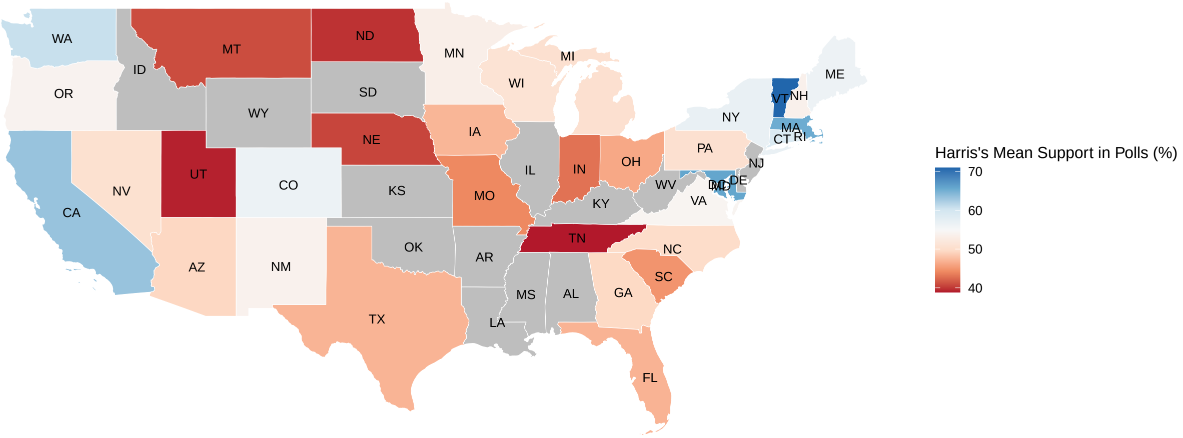

Figure 7: Map showing the proportion of support for Kamala Harris relative to Donald Trump across U.S. states. Blue indicates states where Harris has higher support, while red indicates states where Trump leads. Color intensity reflects the magnitude of support difference, with gray indicating insufficient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

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

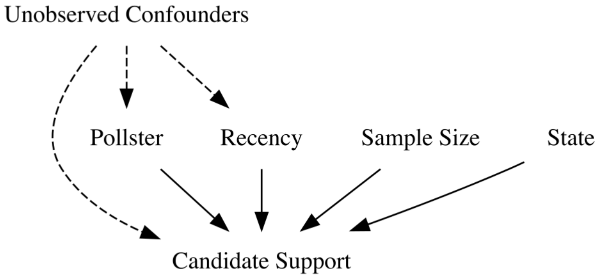

Figure 8: Directed Acyclic Graph illustrating factors influencing candidate support, with poll attributes (Pollster, Sample Size, State, Recency) as direct contributors and unobserved confounders representing potential biases in polling results.

to produce a more stable and comprehensive 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 in a straightforward manner. 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.

2.5 Model Results

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	6498.418	7026.156	3.053
Model 2	Pollster, Sample Size, State	0.720	0.685	5574.796	6292.110	2.043
Model 3	Pollster, Sample Size, State, Recency	0.774	0.746	5311.836	6034.274	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.

2.6 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 pollscore). Since a more negative pollscore reflects a lower level of bias, the formula uses the difference between the maximum pollscore and each poll's specific pollscore. 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 in 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.

2.6.1 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 provides a comprehensive view, taking into account the varied methodologies and sampling qualities of different polling organizations.

2.6.2 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. Additionally, we calculate the margin of difference between the candidates, which can provide insights into how close each state's race is.

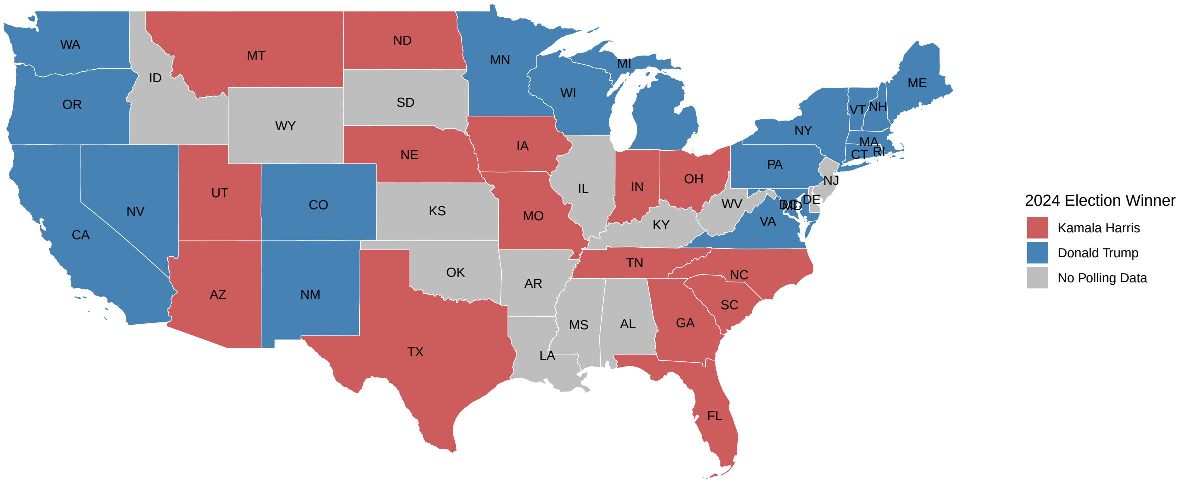

Figure 9: Predicted 2024 U.S. Presidential Election winner by state, with Kamala Harris-leaning states in blue, Donald Trump-leaning states in red, and states lacking sufficient polling data in gray. The map highlights traditional Democratic and Republican strongholds as well as key battleground states.

Figure 9 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 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 critical role swing states play in determining the election outcome.

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	46.89009	48.21778	Donald Trump	1.3276921
Georgia	47.09997	48.21968	Donald Trump	1.1197167
Michigan	47.44921	46.34289	Kamala Harris	1.1063207
Nevada	47.48644	46.47506	Kamala Harris	1.0113774
North Carolina	47.60615	47.87192	Donald Trump	0.2657668
Pennsylvania	48.01476	46.96981	Kamala Harris	1.0449520
Wisconsin	48.67186	46.50662	Kamala Harris	2.1652329

USAFacts(USA Facts 2024),FiveThirtyEight(FiveThirtyEight 2024) and many other polling websites are calling Arizona, Georgia, Michigan, Pennsylvania, Wisconsin, North Carolina, Nevada the 'swing states' of the 2024 presidential election.Polling in swing states is crucial 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 slightly 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%, indicating highly contested races.

3 Discussion

In this paper, we set out to predict the support for both Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election using a "polls-of-polls" approach. By aggregating multiple polls, we sought to mitigate the inherent biases present in individual surveys, creating a more accurate and balanced forecast across candidates. Our analysis was driven by a linear regression model using key predictors, such as pollster, sample size, state, and recency of the polls. After calculating predicted values from the model, we applied a weighting scheme based on each pollster's numeric grade and pollscore, incorporating both reliability and bias. This weighted approach enabled us to produce an overall prediction of candidate support, reflecting variations across different states and polling organizations.

One key insight 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 reveals that in states like Pennsylvania, Michigan, and Wisconsin, support levels for Harris and Trump are nearly tied, reflecting the intense contest for these critical 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 nuanced voter sentiment and the heightened influence that battleground states hold in shaping the final results.

However, there are limitations to our approach. 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 nuances could further enhance predictive accuracy.

In conclusion, this paper contributes to election forecasting by providing a comprehensive, weighted estimate of candidate support across both national and state levels. While our predictions provide valuable insights into potential election outcomes, the evolving nature of voter sentiment and poll accuracy reflects the need for adaptive models that account for ongoing changes in public opinion.

Appendix

A Emerson College Polling Methodology (October 14-16, 2024)

A.1 Overview

Emerson College Polling, known for its extensive online polling methods, conducted a survey from October 14 to 16, 2024, targeting 1,000 likely voters. The poll measured voter preferences between Kamala Harris and Donald Trump, revealing a near tie with Harris at 50% and Trump at 49%.

A.2 Population, Frame, and Sample

The target population is likely voters in the U.S. elections. These voters are determined by the voters' likeliness to vote in the upcoming elections and the voter's voting history. Both of these factors are self-reported.

The sampling frame is the voters that are on the Aristotle list and on the online panel provided by CINT. Aristotle is an online supplier of voter files and CINT is a software company that connects researchers, such as Emerson, to online panels. This approach helps ensure a representative sample by accurately reflecting the electorate.

The sample size is the 1,000 likely voters chosen arbitrarily from the sampling frame. Likely voters are determined by a combination of voter history, registration status, and demographic data. These factors are self-reported.

A.3 Sample Recruitment and Approach

Emerson College used a mixed-mode sampling approach for its polling. Specifically, three primary modes were used to collect data for its October 14-16, 2024 poll:

- MMS-to-web text surveys: Respondents were contacted via text messages sent to cell phones. These messages directed participants to complete the survey online. This method was conducted using voter lists provided by Aristotle.
- Interactive Voice Response (IVR): Landlines were targeted using automated phone calls where respondents could answer questions using their keypads. The contact information for this was also sourced from Aristotle's voter lists.
- Online Panel from CINT: Emerson utilized a pre-screened, opt-in online panel of voters provided by CINT.

A.4 Advantages and Trade-offs in Mixed-Mode Sampling

The multi-mode sampling approach used in this poll has several advantages:

- Wider demographic reach: Having a combination of data collection methods allows Emerson to gather responses from a broader demographic group. In this polling, MMS-to-web surveys mostly gather responses from younger voters and those who primarily use mobile devices. IVR target older and rural voters, who are more likely to respond via landlines. Online panels is more efficient in reaching tech-savvy individuals.
- Cost-efficient: Automated technologies reduce the cost compared to live interviews. In a polling context, this allows for a larger sample size.

However, the trade-offs include:

- Coverage bias: IVR only reaches individuals with landlines, which may skew results towards older demographics.
- Self-selection bias: Respondents in online panels tend to be more engaged, which can lead to over-representation of politically active individuals.

A.5 Non-response Handling

Emerson addresses non-response using a weighting system. The data is adjusted based on key demographic variables like age, gender, race, education, and party affiliation, making the sample more representative of the electorate. Weights are applied to compensate for under- or over-represented groups, ensuring balanced results.

A.6 Questionnaire Design

The Emerson survey features a straightforward, efficient design focused on key issues and voter preferences. Strengths of the questionnaire include:

- Simplicity: Clear, concise questions makes the survey easy for respondents to understand and to provide accurate answers.
- Topical relevance: The poll zeroes in on vote preference and demographic splits, delivering timely insights during election cycles.

However, weaknesses include:

- Lack of nuance: While the survey captures basic preferences, it does not delve deeply into voters' motivations or the factors driving their choices.
- Mode limitations: Different survey modes can affect how respondents engage with the questions. This can potentially influence the depth and accuracy of voter responses.

A.7 Conclusion

Emerson College Polling's mixed-mode methodology is effective in balancing cost with broad demographic reach. Its weighting techniques contribute to reliable and accurate results despite non-responses. While the use of multiple survey modes increases population representation, possible biases—such as lack of nuance and mode effects need to be taken into consideration when interpreting results.

In sum, this method is a strong predictor of election outcomes and continues to adapt to the evolving challenges of modern polling.

B Idealized Methodology

B.1 Overview

This methodology outlines an election forecasting plan with a budget of $100,000. Heavily inspired by TIPP Insights and Emerson College's hybrid approaches, the survey methodology balances cost-efficiency and data accuracy by using different modes of data collection and wide demographic reach.

This forecasting plan will predict the popular vote and Electoral College outcomes, emphasizing key swing states and underrepresented voter groups.

B.2 Sampling Approach

This idealized methodology uses a multi-mode hybrid sampling approach. This allows the use of both probability-based and non-probability-based methods:

- Probability-based methods:
  - Live phone interviews (landline and cell): As used by TIPP, live phone interviews is a direct and personal method of collecting responses. Voter lists would be sourced from Aristotle to ensure we contact likely voters. This mode reaches older voters and voters with less internet connection.
  - Interactive Voice Response (IVR): Similar to Emerson's approach, IVR for landline users balances costs and ensures demographic representation, particularly among older or less tech-savvy populations. This method captures quick responses and is less expensive than live interviews.
- Non-probability methods:
  - MMS-to-web (mobile): Following Emerson's MMS-to-web strategy, we would send multimedia messages to likely voters' cell phones, encouraging them to complete an online survey. This method works well for younger, mobile-first voters, helping fill in gaps where traditional methods might miss.

  - Online panel recruitment via CINT: To mirror TIPP's online panel approach, a portion of the survey would be conducted via an opt-in panel of pre-screened voters from CINT, capturing tech-savvy respondents. CINT's platform ensures that different voter groups, including hard-to-reach groups, are represented in the polling.

B.3 Recruitment Strategy

By using Aristotle's voter lists and CINT's platform as a sampling frame, the sample consists of likely voters based on registration status and past voting behavior. Quota sampling would be used to match the demographic profile of the U.S. electorate. This ensures that respondents in the sample reflect key demographic variables such as age, gender, race, and education level.

We would oversample key swing states like Pennsylvania, Wisconsin, and Florida. Key swing states are states where there is no overwhelming support for a particular candidate or political party. The outcome in these states often determine the overall result of the election, making them critical in securing a win in the Electoral College. Oversampling these states allows a more accurate forecasting in these battlegrounds, which are critical for predicting the Electoral College outcome.

B.4 Data Validation

Given that the data is collected in a hybrid format, we would use duplicate detection methods via IP addresses or phone numbers to eliminate multiple responses from the same individual. Also, we would implement attention checks to identify respondents who are not fully engaged. Both duplicate respondents and unengaged respondents bias results, and should be removed from the data set.

Following TIPP's iterative weighting process, the data will be weighted to ensure that the sample aligns with key demographics (race, gender, education) based on census data.

B.5 Conclusion

By combining TIPP's live phone/IVR hybrid and Emerson College's mixed-mode survey methods, this idealized methodology captures diverse voter demographics. It uses both probability-based and non-probability-based sampling methods to forecast the 2024 U.S. Presidential Election.

A link to the survey can be found at: https://forms.gle/haGLaQBPQhrXE6Xa7

C Copy of the Survey

2024 U.S. Presidential Election Survey

Thank you for your interest in the 2024 U.S. Presidential Election Survey. The 2024 U.S. Presidential Election is set to take place on November 5, 2024. This election will decide the next President of the United States, a position that holds significant influence over national and global policies.

Your participation in this survey is entirely voluntary, and you can choose to withdraw at any time without any consequences. The information collected will be used solely for research purposes to gain insights into voter preferences and trends as we approach this important election. All responses will remain confidential and will be reported only in aggregate form.

Contact Information:
For questions regarding this survey or the methodology used, please reach out to:

- **Denise Chang**, University of Toronto
  Email: dede.chang@mail.utoronto.ca
- **Shamayla Durrin,** University of Toronto
  Email: shamayla.islam@mail.utoronto.ca
- **Krishna Kumar,** University of Toronto
  Email: krishna.kumar@mail.utoronto.ca

* Indicates required question

1. By clicking the button below, I consent to participate in this survey, and I agree that my responses may be used for research purposes, including analysis and reporting on voter sentiment related to the 2024 U.S. Presidential Election. I understand that my participation is voluntary and that I can withdraw at any time without any consequences. My responses will be kept confidential and will be used only in aggregate form. *

   Mark only one oval.

   ○ I agree

Section 1: Voting Intention

2. How likely are you to vote in the 2024 U.S. Presidential Election? *

Mark only one oval.

○ Not likely at all

○ Not very likely

○ Somewhat likely

○ Extremely likely

○ Prefer not to say

3. If the 2024 U.S. Presidential Election were held today, who would you vote for? *

Mark only one oval.

○ Kamala Harris (Democratic Party)

○ Donald Trump (Republican Party)

○ Undecided

○ Prefer not to say

○ Other: ______________________________

4. If you answered "Undecided" in the previous question, who are you leaning towards?

Mark only one oval.

○ Kamala Harris (Democratic Party)

○ Donald Trump (Republican Party)

○ Prefer not to say

○ Other: ______________________________

Section 2: Key Issue Priorities

5. What are the three most important issues influencing your vote in this election? (Select up to 3) *

Check all that apply.

□ Economy
□ Immigration
□ Healthcare
□ National Security
□ Education
□ Climate Change
□ Other: ______________________

6. How has the state of the U.S. economy influenced your voting choice? *

Mark only one oval.

○ Not at all
○ Somewhat
○ Greatly
○ Unsure

Section 3: Voting History

7. Did you vote in the 2020 U.S. Presidential Election? *

Mark only one oval.

○ Yes
○ No
○ Prefer not to say

8. If you voted in 2020, for whom did you vote? *

Mark only one oval.

○ Joe Biden (Democratic party)

○ Donald Trump (Republican party)

○ Prefer not to say

○ Other: ______________________________

Section 4: Voter Engagement

9. How motivated are you to vote in the 2024 election compared to past elections? *

Mark only one oval.

○ Less motivated

○ About the same

○ More motivated

○ Unsure

10. Which of the following describes your voting method? *

Mark only one oval.

○ I plan to vote in person on Election Day

○ I plan to vote early in person

○ I plan to vote by mail

○ I have already voted (via early voting or mail-in)

○ I do not plan to vote

Section 5: Demographic Questions

11. What is your age range? (years) *

Mark only one oval.

○ 18-24
○ 25-34
○ 35-44
○ 45-54
○ 55-64
○ 65 or older
○ Prefer not to say

12. What gender do you identify as? *

Mark only one oval.

○ Male
○ Female
○ Non-binary
○ Prefer not to say
○ Other: ______________________________

13. Which of the following best describes your race or ethnicity? *

Mark only one oval.

○ White or Causasian
○ Black or African American
○ Hispanic or Latino of any race
○ Asian
○ Other or multiple races

14. What is the highest level of education you have completed? *

Mark only one oval.

○ Less than high school
○ High school diploma or GED
○ Some college, no degree
○ Associate degree
○ Bachelor's degree
○ Graduate or professional degree
○ Prefer not to say

15. What is your current employment status? *

Mark only one oval.

○ Employed full-time
○ Employed part-time
○ Self-employed
○ Unemployed
○ Student
○ Retired
○ Prefer not to say

16. Do you consider yourself a supporter of any of the following political parties? *

Mark only one oval.

○ Democratic Party

○ Republican Party

○ Independant

○ Prefer not to say

○ Other: ______________________

D Model Diagnostics

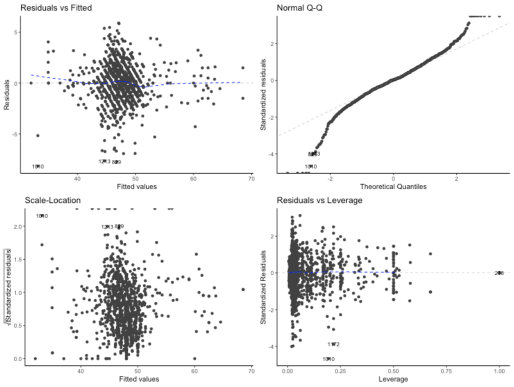

Figure 10

The diagnostic plots for Model 3ni sgiven in Figure 10. 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 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.

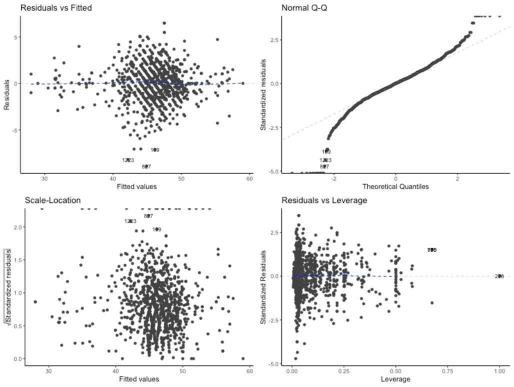

Figure 11

The diagnostic plots for Model 6 is given in Figure 11. 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 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 model suggests that they meet the assumptions for linear regression fairly well, with minor deviations that are not expected to severely impact the model's validity.

E 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.

F 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.

G 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

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

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/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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.

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

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

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

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.