Predicting Kamala Harris's Victory in the 2024 US Election*

Poll Data Predicts Kamala Harris Win with 387 Electoral Votes

Xuanle Zhou Yongqi Liu Yuxuan Wei

November 2, 2024

This study provides a prediction for the 2024 U.S. presidential election, presenting potential outcomes based on a state-by-state electoral vote analysis. These projections can guide campaign strategies and influence public discourse. The results suggest Kamala Harris will secure a win with 387 electoral votes, significantly surpassing the 270 needed for victory. In contrast, Donald Trump is predicted to receive 151 electoral votes.

Table of contents

*Code and data are available at: https://github.com/wyx827/2024USpresidentialelection.git

1 Introduction

The 2024 U.S. Presidential Election will occur in a context of intensified political polarization, economic challenges, and demographic shifts, positioning it as a significant event in shaping America's future. Accurately forecasting this election's outcome is essential for understanding potential shifts in political leadership and policy direction. Polling data has traditionally been a primary tool for measuring voter sentiment; however, recent election cycles have revealed limitations in polling accuracy due to biases, non-response, and variations in poll quality. This study proposes a model that addresses these issues to improve the reliability of election forecasts.

This paper develops a predictive model to forecast the electoral outcomes between Kamala Harris and Donald Trump in the 2024 election. Using a multiple linear regression framework, we analyze aggregated polling data across U.S. states, incorporating factors such as poll quality, transparency, sample size, geographic indicators, and poll timing. By integrating these variables, the model aims to produce a detailed forecast that accounts for regional differences and methodological variations, offering a thorough view of projected support for each candidate.

The findings indicate a competitive election landscape, with Harris showing strong prospects in key battleground states, while Trump retains support in traditional Republican strongholds. These projections highlight the persistent influence of geographic voting patterns and the importance of polling methodology in electoral forecasting. This study offers valuable information for political analysts, campaign strategists, and the public, enhancing the understanding of how polling data can inform electoral outcome predictions.

The remainder of this paper is organized as follows: Section 2 describes the dataset and the preprocessing steps for model preparation. Section 3 explains the modeling approach, including predictor selection and the regression model structure. Section 4 presents the primary findings, detailing state-by-state predictions and overall electoral projections. Section 5 discusses the implications, limitations, and potential directions for future research. The Appendix provides supplementary information on methodology, diagnostics, technical details of the model setup, acknowledgments, and references.

In this analysis, the estimand is the true effect of polling factors on projected support for Trump and Harris in the 2024 U.S. election. Specifically, it reflects how variables such as poll quality, transparency, sample size, and timing influence the predicted vote percentages for Kamala Harris and Donald Trump across states. Defining the estimand clearly ensures consistency in the analysis by guiding model design and interpretation, preventing unintended shifts in focus due to minor methodological changes.

2 Data

2.1 Overview

This study uses R packages (R Core Team 2023) to clean and analyz the dataset, including libraries from tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), knitr (Xie 2023),usmap (Di Lorenzo 2024), corrplot (Wei and Simko 2024), arrow (Richardson et al. 2024).

After cleaning the data, which included grouping and removing missing values, the analysis dataset consists of 1,683 observations, focusing on the following 11 variables: pollster name, methodology, numeric grade, start date, end date, sample size, candidate name, percentage, transparency score, and population group.

2.2 Data Measurement and Considerations

In this analysis, the goal is to convert individual polling data, which reflects voter sentiment before Election Day, into an estimate of the overall electoral support for each candidate. This process involves systematically capturing public opinion about voting intentions and translating it into structured data that can reliably forecast electoral college outcomes.

The dataset originates from FiveThirtyEight (2024), a trusted source known for its rigorous standards in polling data collection. Only polls meeting specific quality criteria are included, ensuring that data represents a broad cross-section of likely voters across the U.S. Each poll must disclose essential details, including the pollster's name, survey dates, sample size, and methodological information. This transparency in methodology includes aspects like the polling medium (phone, online), use of voter files, demographic weighting, and any adjustments made to reflect a representative sample. These steps help ensure that the recorded data accurately captures respondents' voting intentions.

Once these polls pass quality checks, they are integrated into the FiveThirtyEight database, where they contribute to aggregated polling averages and forecast models. Poll results are recorded as percentages reflecting the level of support for each candidate, essentially quantifying voter sentiment. However, moving from these individual opinions to a reliable state-by-state and national forecast requires addressing inherent challenges, such as demographic imbalances and geographic biases. To mitigate these, demographic weights are applied, aligning sample composition with actual population distributions.

The measurement approach also includes recency adjustments to account for the time since each poll was conducted, as opinions can shift rapidly closer to Election Day. This approach, combined with selective inclusion criteria, ensures that the dataset reflects a consistent and structured view of voter sentiment across various demographic and regional groups, bridging the gap between initial polling data and final election predictions.

2.3 Outcome Variable

The outcome variable of interest for this research is the percentage representing the level of public support for Donald Trump. The distribution is shown in Figure 1, where the x-axis represents the support percentage, ranging from approximately 20% to 70%, and the y-axis displays the count of poll surveys reporting each support level. The red dashed line at the 50% mark acts as a benchmark, indicating the threshold needed for majority support.

The distribution shows that most observations cluster around a support level of approximately 48%, indicating significant backing from the electorate. A smaller portion of polls report support above 55%, suggesting that while Trump has a core base, many voters remain either undecided or opposed. This visualization highlights both the concentration and variability of support within the surveyed population.

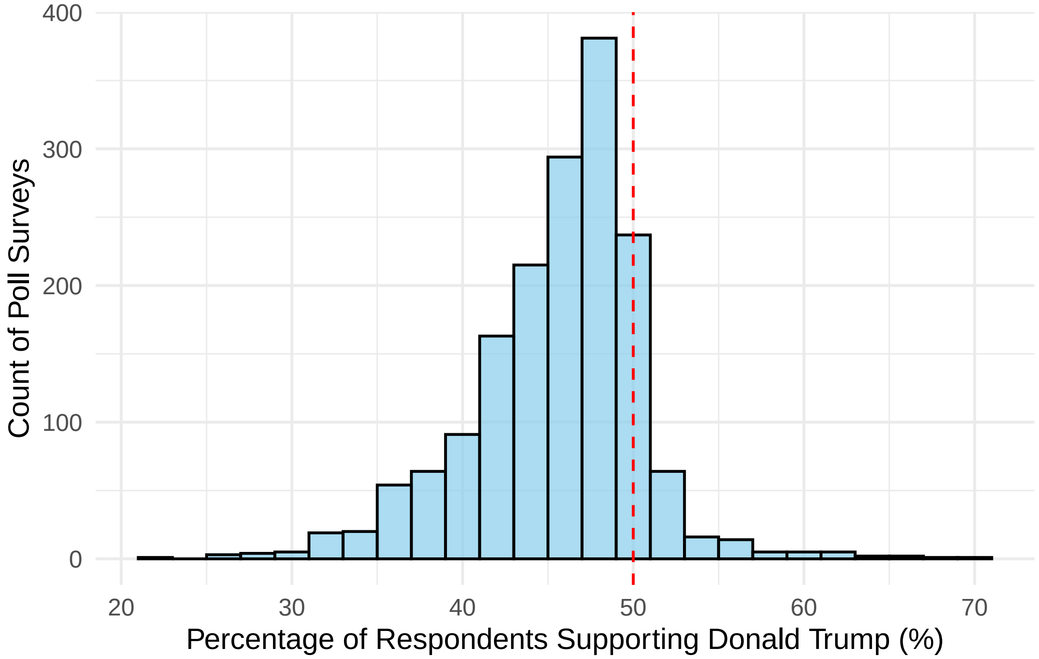

Figure 1: Distribution of Support for Donald Trump in Percentage Across Poll Survey Responses. The red dashed line at the 50% mark serves as a reference point, highlighting the threshold for majority support.

2.4 Predictor Variables

2.4.1 Numeric Grade

The numeric grade reflects the quality of the pollster, with FiveThirtyEight defining a scale from 0 to 3. A grade of 0 indicates a low-quality poll, while a grade of 3 signifies a high-quality pollster. After filtering for pollsters with a numeric grade higher than 2.5, we identified a total of 30 distinct pollsters. The distribution shows that grades cluster around 2.6, 2.8, and 2.9, with fewer polls receiving grades of 2.7 and 3.0, as shown in Figure 2 .

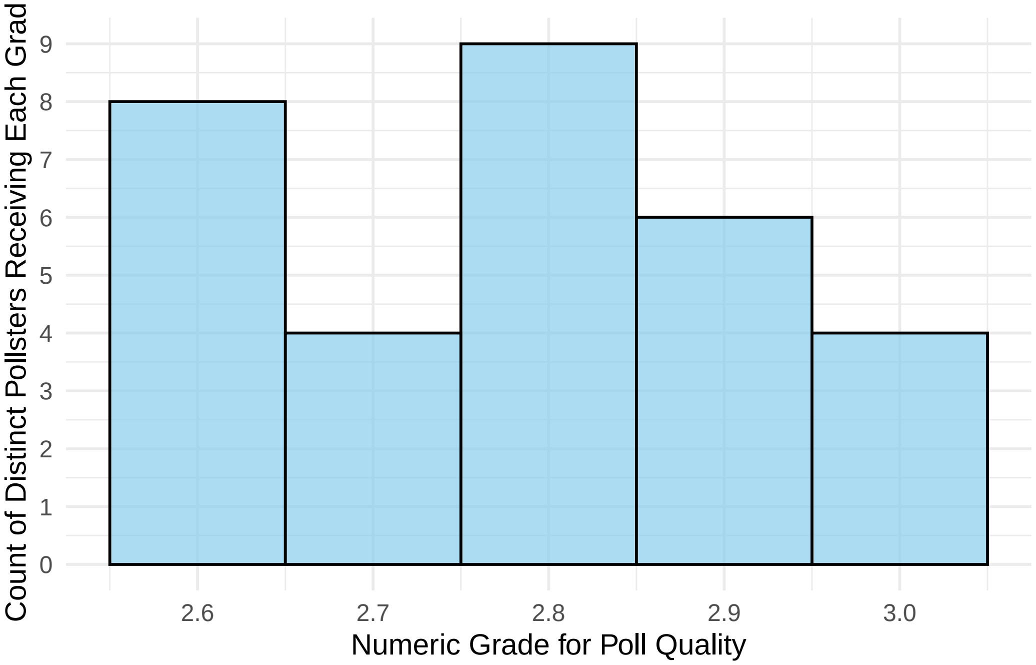

Figure 2: Distribution of Numeric Grades Across Polls (2.6 to 3.0)

2.4.2 Sample Size

The sample size represents the number of respondents in each poll. The distribution in Figure 3 is right-skewed, indicating a higher frequency of polls with smaller sample sizes. The peak occurs around 1,000 respondents, marking this as the most common sample size among the polls. Overall, the data suggests that each poll includes enough respondents to yield reliable information.

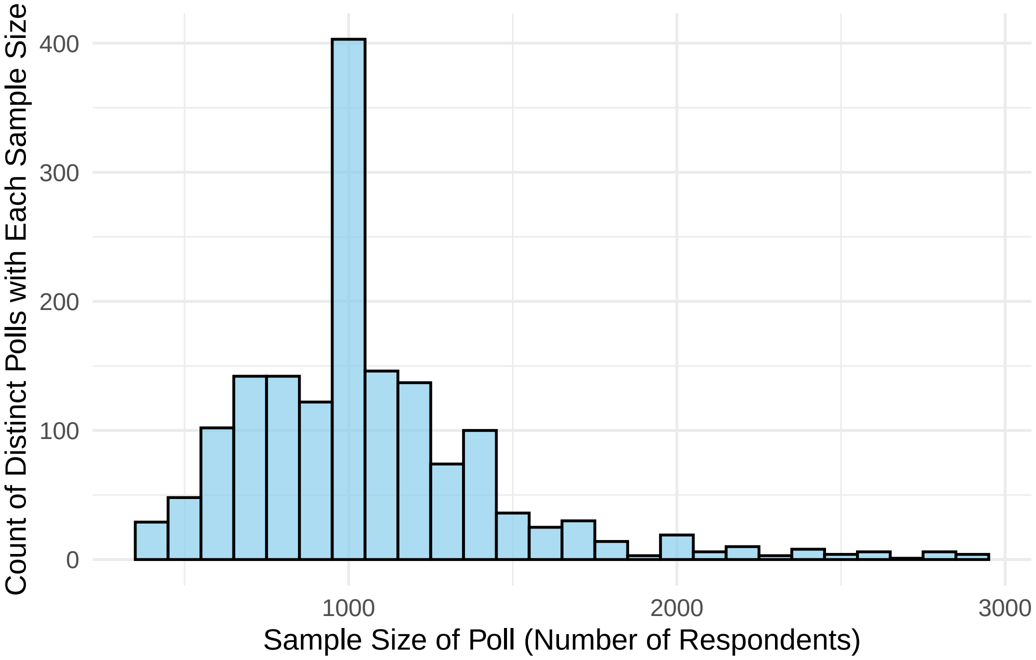

Figure 3: Distribution of Sample Sizes Across Poll Surveys

2.4.3 Transparency Score

The Transparency Score measures how transparent a pollster is, calculated based on the amount of information disclosed about its polls, weighted by recency. The highest possible score is 10, while the lowest is 0. The distribution of Transparency Scores for the filtered pollsters shows a peak around 9 as presented in Figure 4, indicating that this is the most common score. This suggests that among the selected pollsters, there is a predominance of high transparency scores.

2.4.4 Days Until Election

The days until election represents the remaining time leading up to the final election day. This variable is calculated by finding the difference between the end date of each poll survey and the final election date, which is November 5, 2024.

As shown in the Figure 5, the support percentage for Donald Trump fluctuated over time, with a slight upward trend as the election approached, indicated by the dashed trend line. This suggests that public opinion varied throughout the period, and Trump's support percentage experienced an increase as election day drew closer. However, it still remained below 50%. This trend highlights the importance of monitoring real-time polling data, as sentiment may shift in these months leading up to a closely contested election.

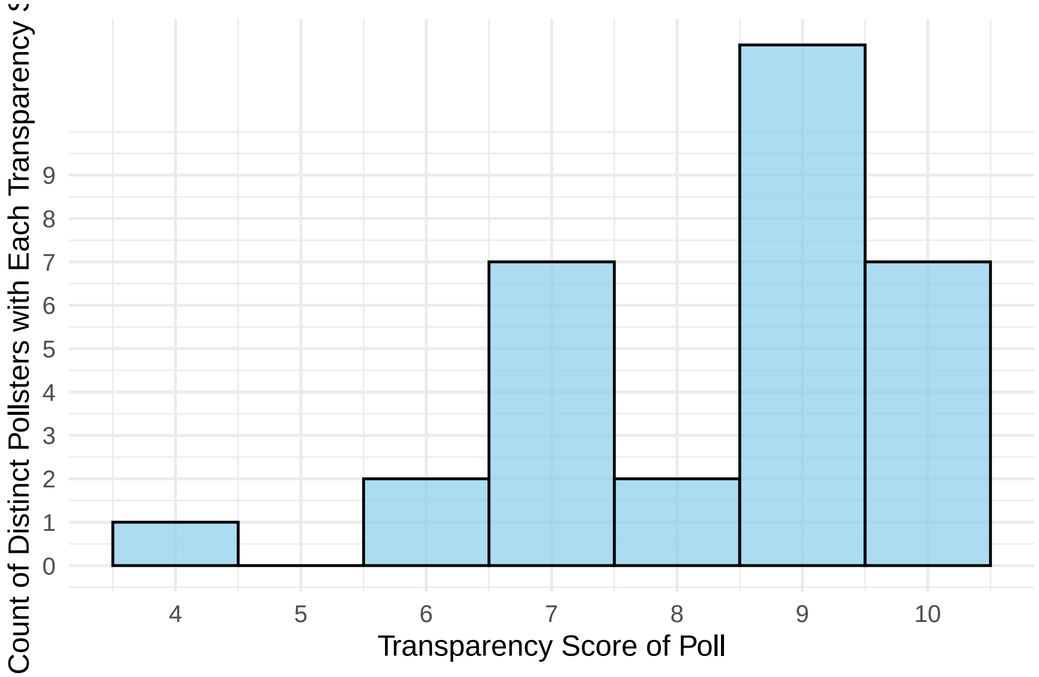

Figure 4: Distribution of Transparency Scores Across Polls

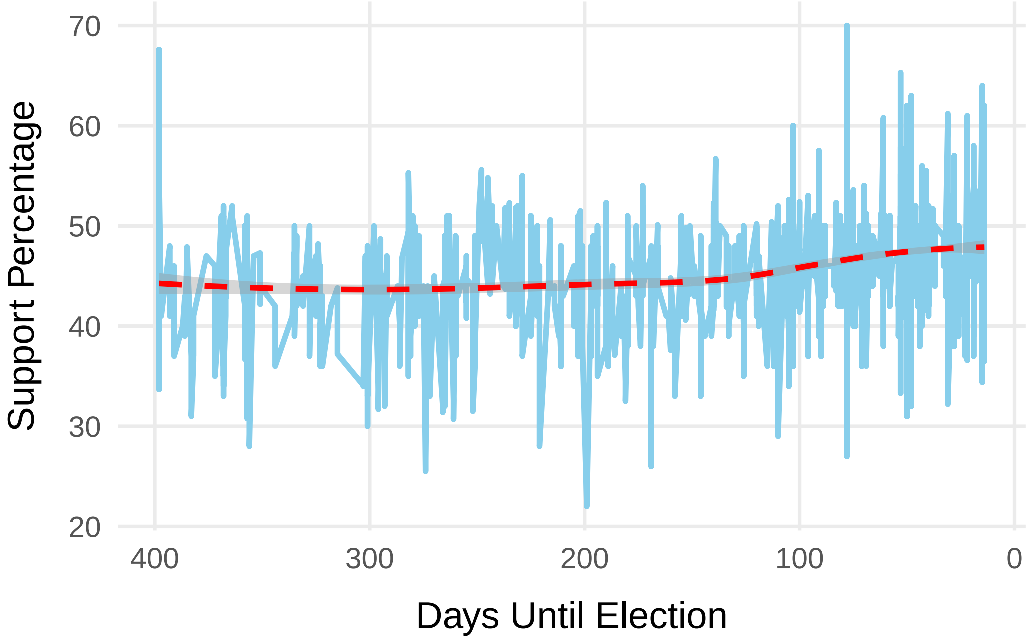

Figure 5: Trend of Support Percentage for Donald Trump Over Days Until Election. The red dashed line indicates the average support rate for Donald Trump at that point in time. Renew the data to the October 24.2024

3 Model

3.1 Model Selection

To model Donald Trump's polling percentages over time, we used a multiple linear regression framework, which estimates the relationship between polling percentages and various predictors by fitting a linear equation to the data. Analyzing the coefficients allows us to quantify the impact of each predictor on Trump's polling percentages, while also assessing the overall fit of the model for reliable predictions.

To prevent overfitting, we applied a train-test data split. The training data is used to build the model, enabling it to learn patterns and relationships within the data. The test data, which the model has not seen before, serves to evaluate its performance on new, unseen data. This separation ensures that the model captures patterns that will generalize beyond the training dataset, rather than just memorizing it.

Below, we briefly describe the multiple linear regression model used to examine Trump's winning probability. Additional details and model diagnostics are provided in Appendix C.

3.2 Multiple Linear Regression Model Overview

The model now predicts Trump's polling percentage (percent) using the following predictors:

- Numeric Grade (numeric_grade): Reflects the quality rating of the pollster.
- Sample Size (sample_size): The number of respondents in the poll.
- State (state): A categorical variable for different U.S. states.
- Transparency Score (transparency_score): A measure of how transparent the polling data and methodology are.
- Days Until Election (days_until_election): The left days until the US election.

The model takes the form:

$$\begin{aligned} \text{pct}_i &= \beta_0 + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{transparency\_score}_i && (1) \\ &\quad + \beta_3 \cdot \text{sample\_size}_i + \beta_4 \cdot \text{state}_i + \beta_5 \cdot \text{days\_until\_election}_i + \epsilon_i && (2) \\ \epsilon_i &\sim \text{Normal}(0, \sigma^2) && (3) \end{aligned}$$

Where:

$$\begin{aligned} \beta_0 &\text{ is the intercept term} && (4) \\ \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 &\text{ are the coefficients for each predictor} && (5) \\ \sigma^2 &\text{ is the variance of the error term} && (6) \end{aligned}$$

3.3 Interpretation of Coefficients

- Intercept (β_0): This is the predicted Trump polling percentage when all predictors (numeric grade, sample size, state, transparency score, and end date) are at their baseline or zero value.
- Numeric Grade (β_1): This coefficient measures how much Trump's polling percentage changes as the pollster's numeric grade increases. A positive and significant coefficient would indicate that higher-rated pollsters report better polling numbers for Trump, while a negative coefficient would suggest the opposite.
- Sample Size (β_2): This measures the impact of the number of respondents on Trump's polling percentage. A positive coefficient would indicate that larger sample sizes are associated with higher polling percentages for Trump.
- State (β_3): The coefficients for the state variable represent differences in Trump's polling percentage in each state compared to the reference state (baseline category). For example, if the coefficient for Florida is negative, it means Trump polls lower in Florida compared to the reference state. The state-level effects account for regional differences in Trump's support. Some states may show significantly higher or lower levels of support, even after adjusting for the time of the poll and pollster quality.
- Transparency Score (β_4): This coefficient shows how much Trump's polling percentage is affected by the transparency of the poll. A positive coefficient would indicate that polls with higher transparency tend to report higher polling percentages for Trump, whereas a negative coefficient would imply the opposite.
- Days Until Election (β_5): The counting down days is a time-related variable. A positive and significant coefficient would suggest that Trump's polling percentage has decreased as the election date approaches, while a negative coefficient would suggest an increase in his polling percent over time.

3.4 Model Justification

Based on the model summary shown in Table 1, we observe that several state-level coefficients are statistically significant, indicating regional variations in support for Trump. Additionally, the coefficients for `transparency_score` and `days_until_election` are both highly significant, suggesting that pollster quality and the timing of the polls have notable effects on Trump's polling percentages. A negative coefficient for `days_until_election` suggests a slight decline in support as the election approaches.

When evaluated on the test set, the model appears to struggle with accurately identifying Trump supporters, possibly due to unobserved factors or limitations in capturing the complexity of voter behavior.

The rationale for applying multiple regression in this context is to control for various influential factors simultaneously—such as time trends (`days_until_election`), pollster quality (`transparency_score`), and regional differences (`state-level effects`). This approach allows us to isolate the individual impact of each variable, providing a more detailed understanding of how each factor contributes to Trump's polling outcomes. The full coefficient output is shown in the Appendix C

Table 1: Regression Model Summary Table Display the first 10 rows

	Estimate	Std. Error	t value
(Intercept)	56.74	3.91	14.51
numeric_grade	1.55	1.16	1.34
sample_size	0.00	0.00	-0.57
stateArizona	-5.48	2.61	-2.10
stateCalifornia	-18.15	2.80	-6.49
stateColorado	-13.81	3.61	-3.83
stateConnecticut	-17.94	4.40	-4.08
stateFlorida	-4.24	2.86	-1.48
stateGeorgia	-5.33	2.61	-2.04
stateIdaho	4.29	4.41	0.97

4 Results

4.1 Predicted Electoral Outcomes

We applied a regression model to predict the percentage of votes Trump is expected to receive in each state. The model results, combined with each state's electoral vote allocation, allowed us to predict the winner in each state. Based on this, we calculated the total number of electoral votes for both Trump and Harris.

The table below (Table 2) summarizes the predicted results, showing Trump's predicted percentage, the number of electoral votes in each state, and the predicted winner (either Trump or Harris). For instance: - Alabama: Trump is predicted to win 53% of the vote, securing all 9 electoral votes. - California: Trump is predicted to receive 34.39% of the vote, resulting in a victory for Harris, who takes California's 55 electoral votes. - Florida: The model predicts a close race, with Trump at 48.89% of the vote, resulting in a Harris win in this battleground state.

Table 2: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
Arizona	47.64	11	Harris
California	34.50	54	Harris

Table 2: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
Colorado	36.41	10	Harris
Florida	50.04	30	Trump
Georgia	47.38	16	Harris
Iowa	46.35	6	Harris
Maine	41.84	4	Harris
Maryland	34.50	10	Harris
Massachusetts	28.24	11	Harris
Michigan	45.78	15	Harris
Minnesota	43.92	10	Harris
Missouri	54.11	10	Trump
Montana	53.93	4	Trump
Nebraska	45.44	5	Harris
Nevada	46.40	6	Harris
New Hampshire	42.57	4	Harris
New Jersey	39.40	14	Harris
New Mexico	42.57	5	Harris
New York	37.15	28	Harris
North Carolina	48.45	16	Harris
Ohio	48.59	17	Harris
Pennsylvania	46.18	19	Harris
Rhode Island	36.13	4	Harris
South Dakota	57.17	3	Trump
Texas	49.20	40	Harris
Vermont	27.00	3	Harris
Virginia	42.72	13	Harris
Wisconsin	45.62	10	Harris

4.2 Predicted Electoral Outcomes by State

The table in the Appendix 7 supplements the predicted electoral outcomes by providing likely results for states where the model could not generate predictions. Using historical data to estimate outcomes for these missing states ensures a complete view of electoral projections and accounts for every state in the final electoral vote totals.

The following map (Figure 6) shows the predicted winner for each state in the 2024 U.S. Presidential Election, based on the regression model's predicted vote percentages for Trump and Harris. The predicted outcome in the regression model reflects the geographic voting patterns, with Trump winning in traditionally Republican-leaning states like Alabama, Missouri, and Wyoming, while Harris dominates in Democratic strongholds such as California, New York, and Illinois. However, key battleground states such as Florida and Arizona are predicted to favor Harris, potentially determining the overall election outcome.

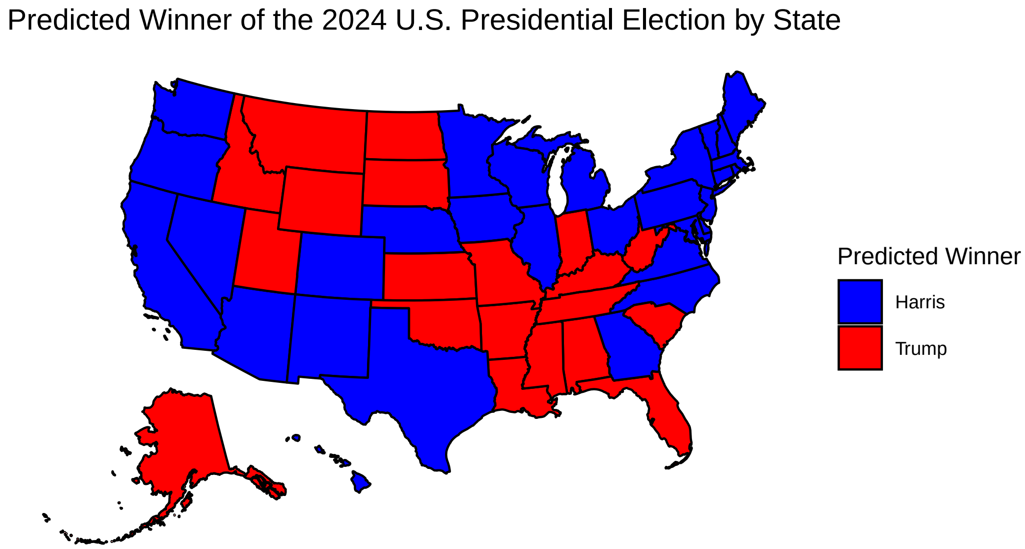
Predicted Winner of the 2024 U.S. Presidential Election by State
Predicted Winner
🟦 Harris
🟥 Trump

Figure 6

The table below (Table 3) shows the numerical predicted winner of the 2024 US Presidential Election by state. Harris is expected to win the 2024 U.S. Presidential Election, securing 387 electoral votes, compared to Trump's 151 electoral votes. The election outcome hinges on several battleground states, where the vote margins are predicted to be narrow.

Table 3: Numerical Predicted Winner of 2024 US Presidential Election by State

Candidate	Total Electoral Votes
Trump	151
Harris	387

5 Discussion

5.1 Key Findings

This paper presents a forecast for the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump, leveraging a linear regression model trained on historical state-level voting data and recent polling information. Our model projects a decisive victory for Harris, with an expected 387 electoral votes compared to Trump's 151, reinforcing the impact of demographic and geographic factors on U.S. electoral outcomes.

The analysis shows that state-specific factors, such as historical voting patterns and days until election day, significantly influence each candidate's polling outcomes. Traditional Democratic strongholds like California and New York favor Harris, while Republican-leaning states such as Alabama and Wyoming align with Trump. This distribution reflects persistent geographic voting patterns and highlights the impact of regional party loyalty.

Notably, the model also highlights the influence of battleground states, where even slight shifts in voter sentiment could alter the election's outcome. For example, Florida, a key swing state, is predicted to lean narrowly towards Trump with an estimated 50.04% of the vote, emphasizing the potential volatility in these regions.

5.2 Regional Voting Patterns and Their Impact on Electoral Outcomes

Our analysis demonstrates the persistence of geographic voting patterns in U.S. elections. As expected, traditional Democratic strongholds such as California, New York, and Illinois favor Harris, while Republican-dominated states like Alabama and Wyoming show support for Trump. These regional preferences align with historical voting behaviors, suggesting that party loyalty and demographic factors continue to shape election results.

The model also highlights the significant role of battleground states, with states such as Florida and Arizona predicted to lean toward Harris. The influence of these states, where margins are often narrow, emphasizes the importance of campaign efforts and voter mobilization in these regions. The projections illustrate that minor shifts in voter sentiment within these states could decisively impact the overall election outcome because of the high number of electoral votes.

5.3 Polling and Electoral Uncertainty Beyond the Data

The model's reliance on historical polling data and state-level predictions reflects certain limitations inherent in election forecasting. While the model effectively captures overall trends, it cannot account for unforeseen factors, such as shifts in voter turnout, the influence of third-party candidates, or sudden political or social events that could impact voter behavior. Additionally, the uneven distribution of survey data across states introduces further uncertainty.Sstates with infrequent polling or underrepresented demographic groups may lead to biased predictions. This variation in polling frequency and sample diversity presents a challenge in achieving a fully representative model, and these uncertainties beyond the data are much more complex.

5.4 Weakness

Despite the model estimating the voter's choice trend, it has limitations. For instance, assuming linear relationships between predictors may oversimplify complex voter behaviors and other external social influences. The dataset simplifies the potential influence behind the US election.

Furthermore, polling is susceptible to biases, which can skew results. For example, factors like nonresponse bias—where certain demographic groups, such as less civically engaged individuals, are less likely to participate—may influence results more heavily in this election cycle. Additionally, reliance on past election data introduces a risk of recency bias, as seen in public expectations shaped by Trump's unexpected win in 2016 and close finish in 2020. As Silver (2024) notes, intuition or prior election outcomes can be misleading, given the unique dynamics of each election. Polling errors, such as those seen in 2016 and 2020, further emphasize the need to carefully interpret our model's findings.

Silver (2024) also points out that the voters for Trump are less likely to finish the polls because of "Shy Tories". This may lead to bias in the survey sample collection. The deficiency of education background and sample credibility both influence the prediction.

With limited data for predicting outcomes in every state, some projections are based on historical voting patterns, assuming these trends will continue. However, this approach may be affected by the changing political landscape in the U.S., where evolving demographics and shifting public sentiment could lead to unexpected results.

5.5 Implications for Future Electoral Modeling

While historical voting patterns provide a solid foundation, they may not fully capture the complexities of modern elections, where factors such as social media influence and voter sentiment play increasingly prominent roles. Incorporating these additional variables, particularly in swing states, could refine the model's predictive power and provide a more comprehensive understanding of the electoral landscape.

For future research, expanding the model to include demographic data, such as age, education level, and income could reveal a deeper analysis of voter behavior on a multi-sided scale. Additionally, real-time polling updates could offer a dynamic perspective, allowing the model to adapt to changes in public opinion as the election date approaches. This variability suggests that further studies could benefit from integrating real-time data and employing Bayesian methodologies to account for evolving voter preferences. Such an approach could improve the accuracy of predictions, especially in regions with historically unpredictable outcomes. These adjustments could improve model's reliability, particularly in closely contested states where the margin of error can significantly impact the final prediction.

Appendix

A Pollster Methodology Overview and Evaluation

A.1 Overview of SurveyUSA

SurveyUSA is a privately held opinion research company that operates nationwide, across all 50 U.S. states. Since its founding, the company has conducted over 40,000 research projects, serving a client base of 400 organizations, including media outlets, corporations, non-profits, government agencies, and academic institutions. Known for its expertise in localized opinion research, SurveyUSA focuses on gathering data at the city, county, and regional levels. The company offers timely, cost-effective surveys tailored to meet specific client needs, distinguishing itself from larger global firms (SurveyUSA 2024b).

A.2 Population, Frame, and Sample

- Target Population: U.S. citizens eligible to vote in the 2024 presidential election.
- Sample Frame: U.S. households with either home telephones or access to devices such as phones or tablets.
- Sample Size: Sample sizes vary across different polls. For the 2024 U.S. presidential election cycle, SurveyUSA conducted 49 polls, with sample sizes ranging from 507 to 2,330 for registered voters or likely voters. The average sample size for these polls is approximately 1,045 households.

A.3 Recruitment

SurveyUSA employs a mixed-method approach to recruitment, including online panels, telephone calls, and a text-to-web method. Some respondents are recruited through Random Digit Dialing (RDD) using telephone samples purchased from Aristotle. RDD allows them to generate phone numbers at random, helping avoid biases that can result from using pre-existing contact lists. For those who do not use home telephones, SurveyUSA invite them to complete the survey on an electronic device such as a phone or tablet. Respondents from non-probability online panels are selected randomly by Cint Lucid Holdings LLC (SurveyUSA 2024a).

A.4 Sampling approach and Trade-offs

SurveyUSA uses a blend of probability and non-probability sampling methods. Some respondents are drawn from non-probability online panels, while others are recruited using probability-based telephone sampling. Responses are weighted based on the latest U.S. Census estimates for age, gender, ethnicity, and region, ensuring alignment with the target population. Questions and answer choices are rotated to reduce order bias, recency effects, and latency effects (SurveyUSA 2024a).

- **Advantages:** The diverse sampling approach not only ensures a broad range of opinions is captured but also complements probability-based sampling, which accurately reflects the overall population. Furthermore, reweighting the data according to U.S. Census demographics strengthens the credibility of the results by ensuring demographic accuracy. Additionally, rotating questions and answer choices helps mitigate bias, further improving the reliability of the data. Finally, the use of online surveys offers a cost-effective solution for efficient data collection.

- **Disadvantages:** Phone-based data collection tends to be time-consuming and can be affected by interviewer effects during telephone interviews. Additionally, challenges like non-response issues, such as busy signals or refusals to participate, can hinder the effectiveness of the data collection process.

A.5 Non-response Handling

Non-response in survey research is important to address because it introduces potential bias, impacting the accuracy and representativeness of the results.As mentioned by Groves and Couper (2012), when certain groups are underrepresented due to non-response, it can skew findings, making them less reflective of the actual population.For example, individuals who with demanding schedules, lower income or limited access to technology might be less likely to respond to a poll. Meanwhile, people with strong opinions or from specific demographic groups could be overrepresented in the results. This imbalance can create biased outcomes, where the poll data disproportionately reflects the views of a more vocal subset rather than an accurate cross-section of the entire population.

To mitigate non-response, SurveyUSA employs follow-up calls when interviews are interrupted by answering machines or busy signals. This additional effort helps reach individuals who may have been initially unavailable or hesitant, thereby improving the sample's representativeness. This method aligns with one of the practical strategies discussed by Groves and Couper (2012) to reduce non-response. However, SurveyUSA need to be caution that excessive or overly persistent follow-ups can lead to respondent discomfort, raise ethical concerns, and introduce potential biases if respondents feel obligated or pressured to participate.

Additionally, SurveyUSA applies weighting adjustments to balance demographic or behavioral tendencies that emerge from non-response patterns. As discussed by Peytchev (2013), while weighting can help realign the sample to more closely match population characteristics, it doesn't fully eliminate the issue, as certain individuals or groups may remain unreachable or unwilling to participate, especially in cases where key variables are missing.

Other approaches that SurveyUSA could consider include imputation, as discussed in the work by Bethlehem (2010). This method estimates missing responses based on existing data, helping to address gaps and reduce bias in cases of partial non-response.

A.6 Questionnaire Evaluation

- **Positive Aspects:** A logical flow between questions facilitates easy navigation for respondents throughout the survey, while simple wording promotes inclusivity by enabling individuals from diverse backgrounds to comprehend the questions. Furthermore, all questions are directly relevant to analyzing the 2024 U.S. presidential election, and providing predefined response options simplifies the choices for participants.

- **Negative Aspects:** Static options for party affiliation and ideology may overlook the subtleties of respondents' political beliefs. These fixed categories risk oversimplifying complex political identities.

A.7 Summary Evaluation

SurveyUSA's methodology reflects a balanced approach, leveraging various sampling approach and method to reach a representative sample. While its blend of probability and non-probability methods has strengths, such as cost-effectiveness and broad reach, it faces challenges related to telephone interview logistics, potential interviewer bias, and the limitations of fixed questionnaire options. Nevertheless, the inclusion of data weighting and question rotation adds credibility to its results, making SurveyUSA a reliable pollster for localized opinion research.

B Idealized Methodology and Survey

B.1 Objective and Overview

The primary goal of this survey methodology is to forecast the U.S. presidential election outcome by collecting representative data from a diverse cross-section of American voters (the target population). The target population refers to the entire group about which we want to make predictions—in this case, eligible U.S. voters across all demographics. With a $100,000 budget, this methodology employs probability sampling techniques, effective recruitment strategies, and data validation protocols to ensure accuracy and minimize bias. Probability sampling ensures that every individual in the target population has a known, non-zero chance of selection, making the results statistically valid and generalizable to the broader population. This approach is designed to account for demographic, geographic, and political factors that can influence voting behavior, thereby enhancing the reliability of predictions (FiveThirtyEight 2024; Pew Research Center 2024).

B.2 Core Objectives

- Obtain a probability sample that represents the U.S. electorate, with demographic diversity reflecting national voting patterns. This helps ensure that conclusions drawn from the sample are applicable to the target population.
- Ensure data quality through validation processes, including screening and cross-verification, to address potential errors and improve the accuracy of the data.
- Utilize statistical modeling and poll aggregation techniques to provide accurate predictions, as outlined in prior studies of electoral polling.

B.3 Sampling Strategy

Our sampling strategy combines **stratified** random sampling with **quota sampling** to ensure broad representation across key demographics. Stratified random sampling involves dividing the target population into subgroups (strata) based on demographic characteristics, then randomly sampling from each stratum. This technique enhances the representativeness of the sample and improves prediction accuracy by reducing demographic and geographic sampling bias (FiveThirtyEight 2024).

B.3.1 Stratification Variables

- **Age Groups**: 18-29, 30-44, 45-64, 65+
- **Gender**: Male, Female, Non-binary/Other
- **Race/Ethnicity**: White, Black, Hispanic/Latino, Asian, Indigenous, Other

- **Education Level**: No high school, High school graduate, College graduate, Post-graduate
- **Income Bracket**: <$30,000, $30,000-$60,000, $60,000-$100,000, >$100,000
- **Geographic Region**: Northeast, Midwest, South, West

This stratification aligns with best practices in survey sampling, allowing for improved accuracy and generalizability in election forecasting by ensuring that subgroups are proportionately represented (Pew Research Center 2024).

B.3.2 Sample Size

A total of **10,000 respondents** will be surveyed, providing a margin of error of approximately ±1% at a 95% confidence level. This sample size allows for detailed subgroup analysis (e.g., by state or demographic group), yielding statistically robust predictions that represent the target population.

B.3.3 Weighting

Post-stratification weights will be applied to adjust for any oversampling or undersampling of specific demographic groups. For instance, younger voters or underrepresented minorities will be weighted to reflect their true proportions in the voting population. This weighting ensures that the final sample aligns closely with the demographic makeup of the target population.

B.4 Recruitment Strategy

Recruitment is executed through **multi-channel outreach**, combining digital advertisements, email outreach, and partnerships with civic organizations. This strategy is designed to capture a probability sample and follows best practices in survey methodologies to minimize non-response bias—errors that occur when certain groups in the target population are less likely to respond to the survey. Non-response matters because it can distort the results if certain demographics are underrepresented in the final data (Pew Research Center 2024).

- **Digital Advertisements**: Targeted ads on platforms like Facebook, Instagram, and Google will recruit respondents based on their demographic profiles (age, gender, location, political interest).
- **Email Outreach**: If permissible, voter registration databases will be accessed to send email invitations to registered voters.
- **Partnerships with Civic Organizations**: Partnering with non-profits and civic organizations that engage diverse communities will help improve respondent diversity and reduce non-response bias.

- **Incentives**: Each participant will be entered into a lottery with a chance to win a $100 gift card to encourage participation and mitigate non-response.

B.5 Data Validation and Quality Assurance

Maintaining data integrity and ensuring high-quality responses are essential to the accuracy of the election forecast. Several measures will be implemented to validate responses and reduce noise in the dataset.

B.5.1 Data Validation Protocols

- **Real-time Captcha Verification**: This will prevent automated bots from submitting responses.
- **Email/Phone Verification**: Respondents will verify their email or phone number to ensure authenticity and prevent duplicate submissions.
- **Time on Task Monitoring**: The survey platform will monitor the time respondents spend on each question. Responses completed unusually quickly will be flagged for review.
- **Voter Registration Cross-Check**: If feasible, respondents will be cross-referenced with voter registration records to ensure eligibility.
- **Response Audits**: Randomly selected respondents will be contacted to verify the accuracy of their responses, ensuring dataset integrity.

These practices align with standards for data integrity and quality control in electoral polling (FiveThirtyEight 2024; Pew Research Center 2019).

B.6 Poll Aggregation and Data Analysis

B.6.1 Poll Aggregation

This survey will be combined with results from reputable polling firms (e.g., YouGov, Ipsos, Gallup) using a poll-of-polls approach to strengthen the forecast (Pew Research Center 2019).

- **Weighting by Methodology and Recency**: Poll results will be weighted based on the rigor of their methodology and the recency of the poll.
- **Handling Bias and Variability**: Aggregated results will adjust for pollster biases and variability between polls to ensure that no single poll dominates the prediction.

B.6.2 Modeling Approach

Bayesian hierarchical models will account for variability across different states, demographics, and regions. This approach allows us to model the popular vote and potentially translate it into **Electoral College predictions**.

B.7 Budget Allocation

- **Respondent Recruitment (Targeted ads, outreach)**: $70,000
- **Incentives (e.g., lottery prizes)**: $10,000
- **Survey Platform (Google Forms, Qualtrics subscription)**: $5,000
- **Data Validation Tools**: $5,000
- **Poll Aggregation & Analysis Software**: $10,000

B.8 Survey Implementation

The survey will be implemented via **Google Forms**, which offers a cost-effective platform for data collection. You can access the survey at the following link: Google Form Survey

B.8.1 Survey Structure

Introduction:

Thank you for taking part in this survey aimed at predicting the outcome of the 2024 US Presidential election. Your voices are valuable to our research.

Please note:

- **All responses will be kept strictly confidential.**
- **Your participation is entirely voluntary.**
- **We kindly request that you answer all questions honestly and to the best of your knowledge.**
- **The survey is estimated to take approximately 10 minutes to complete.**

If you have any inquiries or concerns regarding this survey, please don't hesitate to contact the research team at shaw.wei@mail.utoronto.ca.(Yuxuan Wei, Xuanle Zhou, Yongqi Liu)

Your contribution to this study is greatly appreciated! Each participant will be entered into a lottery with a chance to win a $100 gift card!

Section 1: Eligibility Screening:

Are you a U.S. citizen? - Yes - No [If No, end survey]

Will you be 18 or older by Election Day (November 5, 2024)? - Yes - No [If No, end survey]

Are you registered to vote in the United States? - Yes - No - Not sure - Plan to register before the election

Section 2: Demographic Information:

What is your age group? - 18-29 - 30-44 - 45-64 - 65 or older - Prefer not to say

What is your gender? - Male - Female - Non-binary/Other - Prefer not to say

What is your race/ethnicity? (Select all that apply) - White - Black or African American - Hispanic or Latino - Asian - American Indian or Alaska Native - Native Hawaiian or Pacific Islander - Prefer not to say - Other: [Short text answer]

What is your highest level of education completed? - No high school - High school graduate or equivalent - Some college, no degree - Bachelor's degree - Graduate or professional degree - Prefer not to say

What was your total household income in 2023? - Less than $30,000 - $30,000 - $59,999 - $60,000 - $99,999 - $100,000 - $149,999 - $150,000 or more - Prefer not to say

In which region of the United States do you currently reside? - Northeast (ME, NH, VT, MA, RI, CT, NY, NJ, PA) - Midwest (OH, IN, IL, MI, WI, MN, IA, MO, ND, SD, NE, KS) - South (DE, MD, DC, VA, WV, NC, SC, GA, FL, KY, TN, AL, MS, AR, LA, OK, TX) - West (MT, ID, WY, CO, NM, AZ, UT, NV, WA, OR, CA, AK, HI)

Section 3: Political Views and Voting Intentions:

How likely are you to vote in the 2024 Presidential election? - Definitely will vote - Probably will vote - Might or might not vote - Probably will not vote - Definitely will not vote

Generally speaking, do you usually think of yourself as a: - Democrat - Republican - Independent - Prefer not to say - Other: [Short text answer]

If the 2024 Presidential election were held today, who would you vote for? - Kamala Harris (Democrat) - Donald Trump (Republican) - Undecided - Prefer not to say - Other: [Short text answer]

How certain are you about your choice? - Very certain - Somewhat certain - Not very certain - Not at all certain - Prefer not to say

Which THREE issues are most important to you in deciding your vote? (Select exactly three) - Economy and jobs - Healthcare - Immigration - Climate change - National security - Education - Gun policy - Social justice/racial equality - Taxes - Crime and public safety - Foreign policy - Other: [Short text answer]

Section 4: Information Sources and Engagement:

What is your primary source of political news? (Select all that apply) - Network TV news (ABC, CBS, NBC) - Cable TV news (CNN, Fox News, MSNBC) - News websites - Social media - Radio - Print newspapers - Friends and family - Other: [Short text answer]

How closely have you been following news about the 2024 Presidential election? - Very closely - Somewhat closely - Not too closely - Not at all - Not sure

Section 5: Validation and Consent:

Please verify that you are a human by selecting “Blue” from the following options: - Red - Green - Blue - Yellow

Consent Statement: “I understand that my participation in this survey is voluntary and that my responses will be kept confidential. I agree that my responses may be used for research purposes.” - Yes, I agree - No, I do not agree

Email Address: [Email field]

End Message:

“Thank you for completing this survey. Your response has been recorded. If you have any questions about this survey or would like to be informed about the results, please contact at shaw.wei@mail.utoronto.ca.”

B.9 Survey Design Considerations

- **Question Wording**: All questions are designed to avoid bias or leading responses.
- **Neutrality**: Political questions are framed neutrally to avoid influencing respondents’ answers.
- **Pilot Testing**: The survey will undergo a pilot test to identify and resolve any issues before full deployment.

C Model details

C.1 Diagnostics

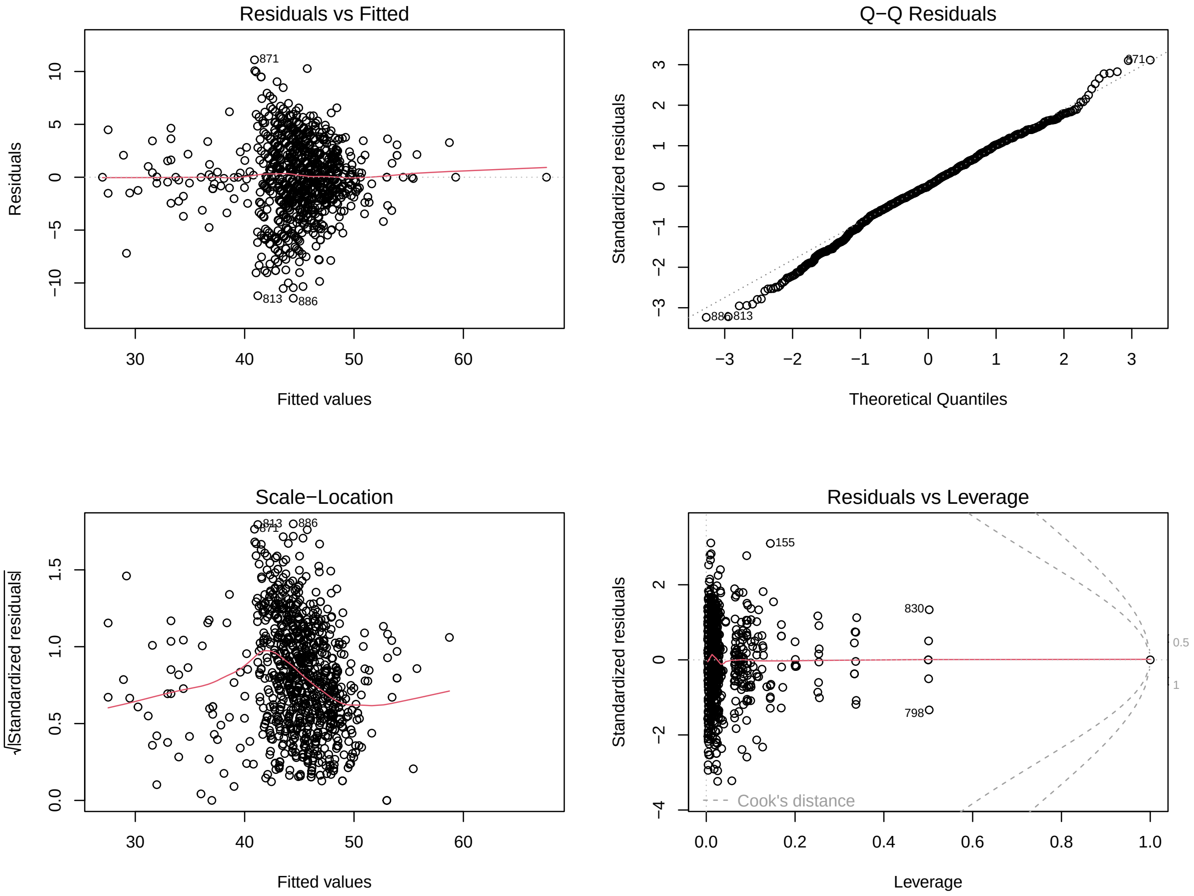

Figure 7

C.2 Mean Squared Error and Mean Absolute Error on Test Data

Table 4: Calculate Mean Squared Error (MSE) on Test Data

Metric	Value
Mean Squared Error (MSE)	11.30
Mean Absolute Error (MAE)	2.65
R-squared	0.56

C.3 Model Summary

	Estimate	Std. Error	t value
(Intercept)	56.74	3.91	14.51
numeric_grade	1.55	1.16	1.34
sample_size	0.00	0.00	-0.57
stateArizona	-5.48	2.61	-2.10
stateCalifornia	-18.15	2.80	-6.49
stateColorado	-13.81	3.61	-3.83
stateConnecticut	-17.94	4.40	-4.08
stateFlorida	-4.24	2.86	-1.48
stateGeorgia	-5.33	2.61	-2.04
stateIdaho	4.29	4.41	0.97
stateIllinois	-16.52	4.41	-3.75
stateIndiana	0.75	3.29	0.23
stateIowa	-5.44	3.60	-1.51
stateKansas	-3.21	4.41	-0.73
stateMaine	-10.13	2.79	-3.63
stateMaryland	-20.52	3.02	-6.80
stateMassachusetts	-23.37	2.82	-8.29
stateMichigan	-7.10	2.61	-2.72
stateMinnesota	-9.94	2.77	-3.59
stateMissouri	-0.92	3.12	-0.29
stateMontana	-0.01	2.96	0.00
stateNational	-8.26	2.57	-3.22
stateNebraska	-8.53	2.91	-2.93
stateNevada	-6.40	2.62	-2.44
stateNew Hampshire	-8.96	2.74	-3.27
stateNew Jersey	-13.62	4.40	-3.09
stateNew Mexico	-12.18	4.40	-2.77
stateNew York	-15.13	2.72	-5.57
stateNorth Carolina	-5.61	2.61	-2.15
stateOhio	-4.07	2.76	-1.48
statePennsylvania	-7.15	2.60	-2.75
stateRhode Island	-14.76	3.12	-4.73
stateSouth Carolina	-1.90	4.40	-0.43
stateSouth Dakota	3.23	3.29	0.98
stateTennessee	4.29	4.41	0.97
stateTexas	-4.48	2.72	-1.64
stateVermont	-24.91	4.41	-5.65
stateVirginia	-9.85	2.70	-3.66
stateWashington	-16.29	3.29	-4.96

	Estimate	Std. Error	t value
stateWest Virginia	9.09	4.41	2.06
stateWisconsin	-6.89	2.60	-2.65
stateWyoming	16.60	4.41	3.77
transparency_score	-0.78	0.10	-7.65
days_until_election	-0.01	0.00	-10.44
R-squared	0.51	NA	NA
Adjusted R-squared	0.49	NA	NA
F-statistic	21.95	NA	NA
Residual Std. Error	3.58	NA	NA

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	56.74	3.91	14.51	0.00
numeric_grade	1.55	1.16	1.34	0.18
sample_size	0.00	0.00	-0.57	0.57
stateArizona	-5.48	2.61	-2.10	0.04
stateCalifornia	-18.15	2.80	-6.49	0.00
stateColorado	-13.81	3.61	-3.83	0.00
stateConnecticut	-17.94	4.40	-4.08	0.00
stateFlorida	-4.24	2.86	-1.48	0.14
stateGeorgia	-5.33	2.61	-2.04	0.04
stateIdaho	4.29	4.41	0.97	0.33
stateIllinois	-16.52	4.41	-3.75	0.00
stateIndiana	0.75	3.29	0.23	0.82
stateIowa	-5.44	3.60	-1.51	0.13
stateKansas	-3.21	4.41	-0.73	0.47
stateMaine	-10.13	2.79	-3.63	0.00
stateMaryland	-20.52	3.02	-6.80	0.00
stateMassachusetts	-23.37	2.82	-8.29	0.00
stateMichigan	-7.10	2.61	-2.72	0.01
stateMinnesota	-9.94	2.77	-3.59	0.00
stateMissouri	-0.92	3.12	-0.29	0.77
stateMontana	-0.01	2.96	0.00	1.00
stateNational	-8.26	2.57	-3.22	0.00
stateNebraska	-8.53	2.91	-2.93	0.00
stateNevada	-6.40	2.62	-2.44	0.01
stateNew Hampshire	-8.96	2.74	-3.27	0.00
stateNew Jersey	-13.62	4.40	-3.09	0.00
stateNew Mexico	-12.18	4.40	-2.77	0.01
stateNew York	-15.13	2.72	-5.57	0.00

	Estimate	Std. Error	t value	Pr(>|t|)
stateNorth Carolina	-5.61	2.61	-2.15	0.03
stateOhio	-4.07	2.76	-1.48	0.14
statePennsylvania	-7.15	2.60	-2.75	0.01
stateRhode Island	-14.76	3.12	-4.73	0.00
stateSouth Carolina	-1.90	4.40	-0.43	0.67
stateSouth Dakota	3.23	3.29	0.98	0.33
stateTennessee	4.29	4.41	0.97	0.33
stateTexas	-4.48	2.72	-1.64	0.10
stateVermont	-24.91	4.41	-5.65	0.00
stateVirginia	-9.85	2.70	-3.66	0.00
stateWashington	-16.29	3.29	-4.96	0.00
stateWest Virginia	9.09	4.41	2.06	0.04
stateWisconsin	-6.89	2.60	-2.65	0.01
stateWyoming	16.60	4.41	3.77	0.00
transparency_score	-0.78	0.10	-7.65	0.00
days_until_election	-0.01	0.00	-10.44	0.00
R-squared	0.51	NA	NA	NA
Adjusted R-squared	0.49	NA	NA	NA
F-statistic	21.95	NA	NA	NA
F-statistic p-value	0.00	NA	NA	NA
Residual Std. Error	3.58	NA	NA	NA

Complete Model Coefficient Summary Table

C.4 Multicollinearity Check on the Training Data

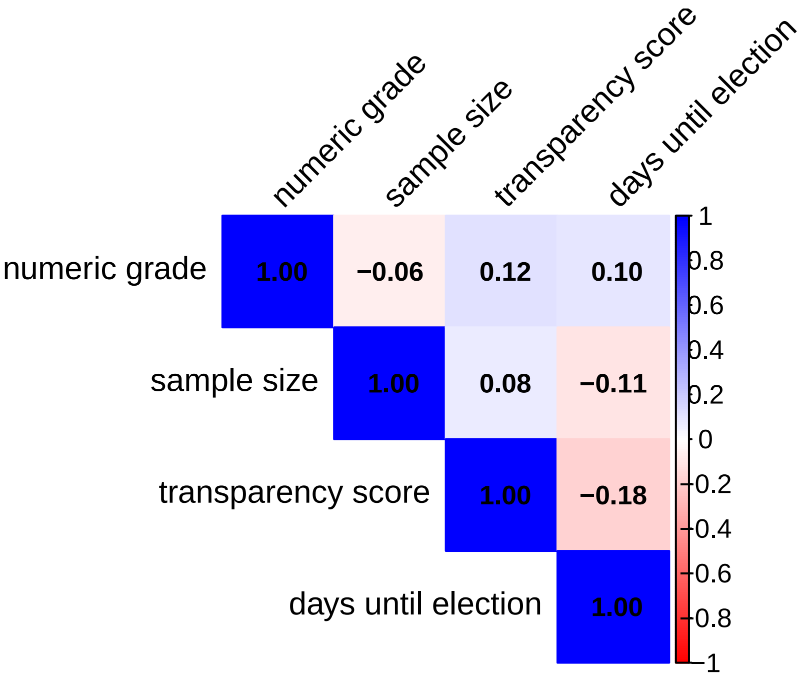

Figure 8: Correlation Matrix of Numeric Grade, Sample Size, Transparency Score and `Days Until Election`

C.5 Missing States with Predicted Winner Based on Historical Lean

Table 7: Missing States with Predicted Winner Based on Historical Lean

State	Electoral Votes	Prediction Based on Historical Lean
Alabama	9	Trump
Alaska	3	Trump
Arkansas	6	Trump
Connecticut	7	Harris
Delaware	3	Harris
Hawaii	4	Harris
Idaho	4	Trump
Illinois	19	Harris
Indiana	11	Trump
Kansas	6	Trump
Kentucky	8	Trump
Louisiana	8	Trump

Table 7: Missing States with Predicted Winner Based on Historical Lean

State	Electoral Votes	Prediction Based on Historical Lean
Mississippi	6	Trump
North Dakota	3	Trump
Oklahoma	7	Trump
Oregon	8	Harris
South Carolina	9	Trump
Tennessee	11	Trump
Utah	6	Trump
Washington	12	Harris
West Virginia	4	Trump
Wyoming	3	Trump
District of Columbia	3	Harris

D Acknowledgements

Thanks to Open AI and ChatGPT 4.0 is used to write the paper. This project utilized R Core Team (2023) and a variety of R packages that provided essential tools and functionality. We thank the Wickham et al. (2019), Wickham (2016) and Xie (2023) teams for their extensive suite of tools for data manipulation and visualization, which were instrumental in data cleaning and graphing. We also acknowledge Di Lorenzo (2024) for its support in creating U.S. geographic visualizations, and Wei and Simko (2024) for facilitating clear, informative correlation plots. Finally, Richardson et al. (2024) was vital for efficient data storage and handling with Parquet files. We are grateful to the developers and maintainers of these packages for their contributions to open-source software.

References

Bethlehem, Jelke. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78 (2): 161–88.

Di Lorenzo, Paolo. 2024. *usmap: US Maps Including Alaska and Hawaii*. https://CRAN.R-project.org/package=usmap.

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

Groves, Robert M, and Mick P Couper. 2012. *Nonresponse in Household Interview Surveys*. John Wiley & Sons.

Pew Research Center. 2019. "Methods 101: How Polls Work." Pew Research Center. https://www.pewresearch.org/methods/2019/07/16/video-how-is-polling-done-around-the-world/.

———. 2024. "Survey Methodology for Political Polling." Pew Research Center. https://www.pewresearch.org/methodology.

Peytchev, Andy. 2013. "Consequences of Survey Nonresponse." *The ANNALS of the American Academy of Political and Social Science* 645 (1): 88–111.

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

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

Silver, Nate. 2024. "My Gut Says Trump, 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?unlocked_article_code=1.UU4.pFkQ.F2hD-woxmiEj&smid=url-share.

SurveyUSA. 2024a. "SurveyUSA Methodology." https://www.surveyusa.net/methodology/.

———. 2024b. "SurveyUSA: America's Polling Agency." https://www.surveyusa.net/.

Wei, Taiyun, and Viliam Simko. 2024. *R package 'corrplot': Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

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

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

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.