Forecasting the 2024 U.S. Presidential Election*

Modeling State-Level Polling Data to Forecast the Electoral College Outcome

Tim Chen Steven Li Tommy Fu

November 4, 2024

We develop statistical models to forecast the 2024 U.S. presidential election outcome between Kamala Harris and Donald Trump using polling data from July through October 2024, incorporating pollster reliability metrics, sample sizes, and state-level effects to predict Electoral College results. Analyzing over 800 polls from diverse polling organizations, we find that pollster rating and state-level factors significantly influence reported candidate support, with our models predicting 306 electoral votes for Harris and 232 for Trump. Our analysis demonstrates that while national polling averages provide useful findings, state-specific predictions are essential for accurate Electoral College forecasting, particularly in battleground states where margins are historically narrow. These findings highlight the importance of incorporating methodological rigor in polling analysis and demonstrate how sophisticated statistical modeling can improve electoral predictions despite the challenges of modern polling.

Table of contents

*Code and data are available at: https://github.com/timchen0326/US_presidential_election_forecast_2024.git

1 Introduction

The United States presidential election of 2024 presents unique challenges for electoral forecasting, particularly following notable polling inaccuracies in recent elections. The 2016 and 2020 elections demonstrated significant underestimation of Republican support in key states, leading to decreased public confidence in traditional polling methods. Research by the American Association for Public Opinion Research has identified several contributing factors, including difficulties in obtaining representative samples and predicting voter turnout amid increasing political polarization (Keeter 2024).

Contemporary polling organizations have implemented methodological improvements to address these challenges. As documented by Keeter (2024), modern electoral polling requires sophisticated weighting procedures and enhanced quality control mechanisms to counteract declining response rates and partisan non-response bias. These refinements are especially important in swing states, where narrow margins typically determine electoral outcomes and where previous polling errors have had the most significant impact (Viala-Gaudefroy 2024). This methodological evolution reflects the polling industry's adaptation to changing communication patterns and voter behavior.

This paper aims to forecast the winner of the 2024 United States presidential election between Kamala Harris and Donald Trump by predicting each candidate's Electoral College vote total. Specifically, our estimand is the number of Electoral College votes each candidate will receive, calculated by converting state-level popular vote predictions into electoral votes through each state's winner-take-all system. Using polling data from FiveThirtyEight's 2024 Presidential Election Forecast Database collected between July and October 2024 (FiveThirtyEight 2024), we develop multi-linear regression models that incorporate pollster reliability metrics, sample sizes, and state-specific effects. Our analysis predicts that Kamala Harris will receive 306 electoral votes compared to Donald Trump's 232, surpassing the 270-vote threshold required for victory. We also provide an evaluation of YouGov's (YouGov 2024) polling methodology and present an idealized survey approach for future election forecasting.

The remainder of this paper is structured as follows. Section 2 examines our dataset and key variables. Section 3 details our statistical approach and model development. Section 4 presents electoral college projections. Section 5 explores implications and limitations. Section A provide an in-depth evaluation of polling methodologies and our proposed ideal survey design.

2 Data

2.1 Overview

This study analyzes polling data from FiveThirtyEight's 2024 Presidential Election Forecast Database (FiveThirtyEight 2024). The database, maintained by ABC News, provides stan-

dardized polling results from diverse polling organizations, with each poll receiving rigorous quality assessments and methodological evaluations. We examine both national and state-level polls conducted between July and October 2024, focusing on the presidential race between Kamala Harris and Donald Trump. Table 4 in the appendix presents a sample of our dataset.

2.2 Outcome Variable

Our primary dependent variable is candidate support percentage (`pct`), representing voter preferences between the two major candidates. Figure 3 illustrates the distribution of support levels, indicating a competitive race where both Harris and Trump consistently receive between 40-50% support across polls. The relatively narrow range of these percentages underscores the importance of methodological precision in detecting meaningful differences in voter preference.

2.3 Predictor Variables

We identify several important predictors that influence polling accuracy and electoral predictions. These variables fall into two categories: poll quality metrics and survey characteristics.

2.3.1 Poll Quality Metrics

Table 5 summarizes three fundamental quality metrics in our dataset:

- **Pollscore**: A measure of polling organization performance, where negative values indicate better accuracy. Our data shows a mean pollscore of -1.06 (SD = 0.28, range: -1.50 to -0.50), suggesting consistently high-quality polling organizations in our sample.
- **Numeric Grade**: An evaluation of pollster reliability on a standardized scale. With a mean of 2.89 (SD = 0.10, range: 2.70 to 3.00), our dataset maintains strong methodological standards. We established a minimum threshold of 2.7 to ensure rigor.
- **Transparency Score**: A quantification of methodological disclosure (maximum: 10). Our sample demonstrates strong transparency practices with a mean of 8.59 (SD = 1.04), indicating thorough methodological documentation.

2.3.2 Survey Characteristics

- **Methodology**: As shown in Figure 4, polling approaches vary significantly. Traditional live phone interviews and online panels dominate, while hybrid methods combining multiple approaches (e.g., IVR/Online Panel/Text-to-Web) reflect adaptation to changing communication patterns.

- **Sample Size**: Figure 5 illustrates substantial variation in respondent numbers, with polls averaging 1,114.76 respondents (SD = 583.72, range: 450-4,253). The right-skewed distribution reflects larger samples in national polls, with state-level polls typically using smaller, targeted samples.
- **State**: Table 6 shows the distribution of polls across states, with 261 national surveys complemented by intensive polling in battleground states: Pennsylvania (86 polls), Wisconsin (82), and North Carolina (64). This distribution reflects strategic focus on electorally competitive regions.
- **End Date**: Figure 6 tracks polling frequency over time, showing increased activity during September and early October 2024, with peak weekly frequencies in late September. This pattern captures intensified data collection as Election Day approaches.

2.4 Measurement and Limitations

There are several measurement and limitation considerations for our dataset:

- **Poll Quality:** While our pollscore and numeric grade filters help ensure data quality, these metrics are based on historical performance and may not fully capture current methodological improvements or deterioration.
- **Temporal Dynamics:** Our dataset provides discrete snapshots of voter preferences rather than continuous measurement. This limitation is particularly relevant given the fast evolution of political narratives and voter sentiment during presidential campaigns.
- **Geographic Coverage:** Although we have national and state-level polling data, coverage varies by state. Battleground states typically have more frequent polling, while safer states may have sparse data, potentially affecting our state-level predictions.
- **Response Bias:** Despite careful methodology by pollsters, self-selection bias in survey participation and social desirability bias in responses remain potential concerns.

2.5 Data Processing and Analysis

Our analysis employs R (R Core Team 2023) with specialized packages: tidyverse (Wickham et al. 2019) for data manipulation, dplyr (Wickham et al. 2023) for transformation, janitor (Firke 2023) for naming standardization, and lubridate (Grolemund and Wickham 2011) for date handling. The data cleaning process, detailed in Section B.1, involved filtering for high-quality polls (numeric_grade $\geq$ 2.7), standardizing geographic information, and creating consistent date formats and candidate indicators.

Our variable selection prioritized metrics that directly influence polling accuracy and electoral forecasting while excluding variables that could introduce noise or redundancy. We retained `pollscore` and `numeric_grade` as established measures of polling reliability, while excluding related but less precise metrics such as `pollster_rating_name`. Geographic and temporal variables (`state`, `end_date`) were essential for capturing regional variations and

time-series patterns. We deliberately omitted variables that could introduce partisan bias (e.g., `sponsor_candidate`, `endorsed_candidate_name`) or duplicate existing information (e.g., `population_full`, `display_name`). Administrative identifiers (`poll_id`, `race_id`) and URLs were excluded as they didn't serve analytical purposes. This focused selection maintains analytical rigor while avoiding the risks of over-parameterization.

3 Modeling Support for the Candidates

To investigate the factors influencing voter support for Kamala Harris and Donald Trump, we employed linear regression models using polling data. Our analysis aimed to assess how predictors such as sample size, pollster ratings, transparency scores, and state-level variables affect the reported percentage of support for each candidate.

We began with simple linear regression models to examine the relationship between sample size (`sample_size`) and the percentage of support (`pct`) for each candidate.

Figures Figure 7 and Figure 8 illustrate the relationship between sample size and support percentage for Kamala Harris and Donald Trump, respectively. These plots provide a visual representation of the trends observed in the linear regression models for each candidate.

3.1 Multiple Linear Regression Models

Recognizing the complex nature of voter support, we extended our analysis using multiple linear regression (MLR) models. The initial model incorporated several predictors believed to influence polling outcomes:

$$\text{pct} = \beta_0 + \beta_1 \times \text{numeric\_grade} + \beta_2 \times \text{pollscore} + \beta_3 \times \text{transparency\_score}$$

$$+\beta_4 \times \text{sample\_size} + \beta_5 \times \text{state} + \beta_6 \times \text{methodology} + \epsilon$$

Variables were selected based on their relevance to polling accuracy and potential influence on voter preferences. Pollster ratings and transparency scores were included to account for differences in pollster reliability. State-level data were incorporated to capture regional variations in support. Sample size and methodology were included to control for their potential impact on polling results (see Figure 9 and Figure 10 as it illustrates residual plots and Q-Q plots).

3.2 Multicollinearity Check Using Variance Inflation Factor (VIF)

To ensure that the predictors used in both models do not exhibit multi-collinearity, we checked the Variance Inflation Factor (VIF) for each predictor. High VIF values indicate multi-collinearity, which can affect the stability and reliability of the model coefficients.

Table 1: Harris MLR model Variance Inflation Factor (VIF)

	GVIF	Df	GVIF^(1/(2*Df))
numeric_grade	4.516	1	2.125
pollscore	3.050	1	1.746
transparency_score	7.164	1	2.677
sample_size	1.676	1	1.295
state	18.682	26	1.058
methodology	83.067	10	1.247

To assess multicollinearity among predictors, we calculated the Variance Inflation Factor (VIF) for each variable(Table 1). High VIF values suggest multicollinearity, which can compromise the stability of coefficient estimates.

The initial VIF analysis demonstrates multicollinearity concerns, particularly with transparency score and methodology. To address this, we refined the model by removing less significant predictors with high VIF values, resulting in:

$$\text{pct} = \beta_0 + \beta_1 \times \text{numeric\_grade} + \beta_2 \times \text{pollscore} + \beta_3 \times \text{sample\_size} + \beta_4 \times \text{state}$$

Table 2: Harris Refined MLR model Variance Inflation Factor (VIF)

	GVIF	Df	GVIF^(1/(2*Df))
numeric_grade	2.435	1	1.560
pollscore	2.353	1	1.534
sample_size	1.593	1	1.262
state	2.076	26	1.014

This refinement reduced VIF values to acceptable levels, improving the model's reliability and interpretability.

3.3 Stepwise Model Selection

To optimize the model further, we applied stepwise selection using the Akaike Information Criterion (AIC) as the criterion for adding or removing predictors. The final model obtained is:

$$\text{pct} = \beta_0 + \beta_1 \times \text{pollscore} + \beta_2 \times \log(\text{sample\_size}) + \beta_3 \times \text{state} + \epsilon$$

The logarithmic transformation of sample size was introduced to account for diminishing returns in effect size with increasing sample sizes.

The final model was selected based on statistical significance, model diagnostics, and theoretical considerations. Pollster rating and state remained significant predictors, consistent with expectations that pollster credibility and regional factors influence polling results. The logarithmic transformation improved model fit and addressed heteroscedasticity associated with sample size.

All statistical analyses were conducted using (R Core Team 2023). The lm() function was used for regression modeling, and the car package (Fox and Weisberg 2019) was employed to calculate VIF values. Stepwise selection was performed using the step() function. Diagnostic plots were generated to assess model assumptions.

3.4 National-Level Model Evaluation and Predictive Accuracy

After finalizing the model selection process, we evaluated the predictive performance of our models for Kamala Harris and Donald Trump using national-level polling data. Focusing on national polls, we split the dataset into training and test sets, maintaining consistency with a fixed random seed to ensure reproducibility. For each candidate, we developed a multiple linear regression model using pollscore and a log-transformed sample_size as predictors, capturing factors pertinent to national-level polling dynamics.

We then trained the models on the training subset and generated predictions for the test subset, evaluating model accuracy using the Root Mean Squared Error (RMSE). The RMSE metric quantifies the average prediction error in the test set, providing an indication of each model's reliability in predicting future national poll outcomes. The Harris model yielded an RMSE of 3.12, while the Trump model's RMSE was 2.40, indicating that both models provide reasonably accurate predictions, with the Trump model demonstrating slightly lower average error in predicting national polling support. These results reflect the models' robustness and highlight their utility for assessing national-level candidate support.

4 Electoral College Prediction

Table 3: Electoral Votes for Each Candidate

Candidate	Electoral Votes
Harris	306
Trump	232

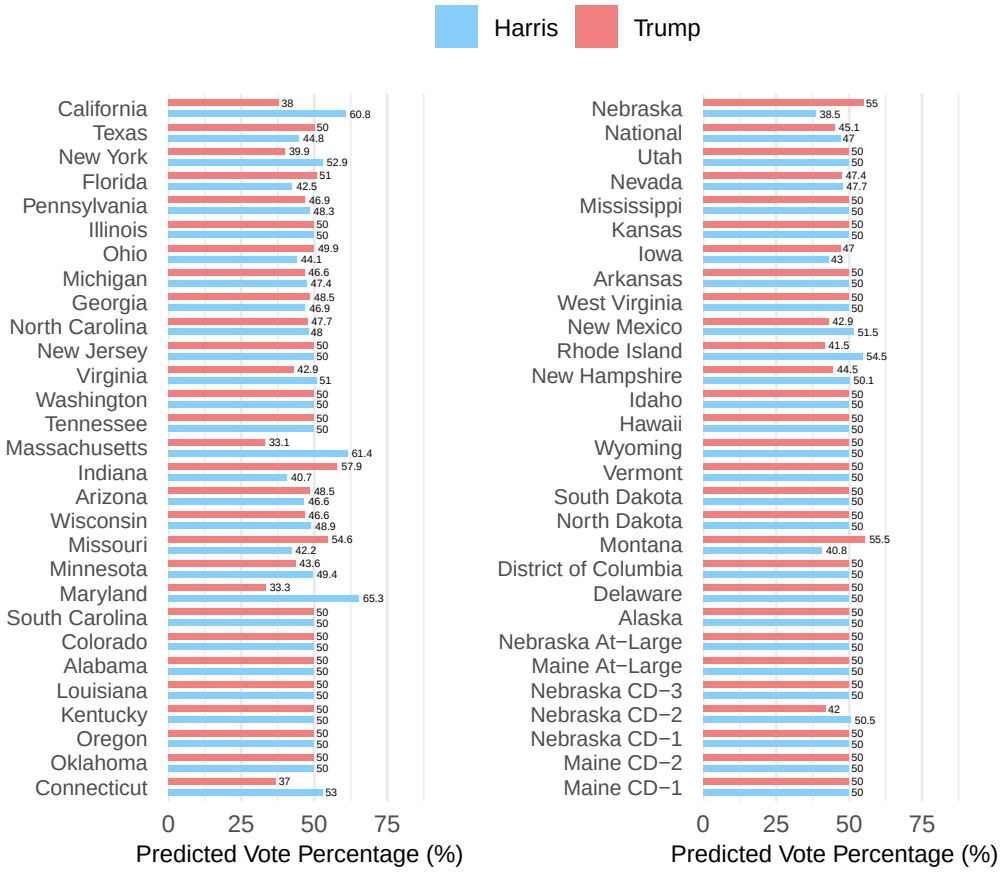

Figure 1: Predicted Vote Percentages by State

To project the outcome of the 2024 U.S. Presidential election between Kamala Harris and Donald Trump, we conducted a statistical analysis that integrates predicted state-level polling percentages with the allocation of electoral votes. This methodology employs multiple linear regression models to estimate each candidate's support and translates these predictions into potential Electoral College results, adhering to the structure of the U.S. electoral system.

1. **Prediction of State-Level Polling Percentages**: We utilized finalized multiple linear regression models for both Kamala Harris and Donald Trump. These models were applied to their respective datasets to generate predicted polling percentages for each candidate across various states. The predictions reflect the expected levels of support based on historical polling data and relevant covariates included in the models.

2. **Aggregation of Predicted Support by State**: For each candidate, we computed the average predicted polling percentage within each state. This aggregation provides a concise state-level summary of anticipated support, facilitating direct comparisons between the candidates in each state (see Figure 1).

3. **Integration of Electoral Vote Allocations**: We established a dataset detailing the electoral vote distribution for each state, including the distinct allocations for Maine and Nebraska, which can split their electoral votes by congressional district. This dataset was merged with the state-level predicted support data to align electoral votes with the corresponding predicted support.

4. **Addressing Missing Data**: We identified states lacking predicted polling data and assigned a neutral predicted support of 50% to both candidates in these states. This assumption accounts for the absence of data while ensuring that all states are included in the analysis.

5. **Determination of State Winners**: Based on the aggregated predicted percentages, we determined the projected winner in each state. A candidate was designated as the winner if their predicted polling percentage exceeded that of their opponent. In cases where the predicted percentages were equal, we randomly assigned the winner using a stochastic approach to reflect the uncertainty inherent in tied predictions.

6. **Calculation of Total Electoral Votes**: We summed the electoral votes from all states won by each candidate to compute their total electoral vote counts. This calculation is necessary, as securing at least 270 electoral votes is necessary to win the presidency under the U.S. electoral system.

7. **Interpretation of Results**: The analysis indicated that Kamala Harris is projected to receive a total of **306 electoral votes**, while Donald Trump is projected to receive **232 electoral votes**. Harris surpasses the 270-vote threshold required for victory, suggesting a strong position in the Electoral College based on our predictions. This outcome implies that Harris is likely to win the election, given her substantial lead in projected electoral votes (see Figure 2).

This statistical approach effectively connects state-level polling data to Electoral College projections. By employing multiple linear regression models to predict polling percentages and integrating these predictions with the electoral vote framework, we provide a detailed analysis of potential election results. This methodology reflects the structure of the U.S. electoral system and provides valuable results into how predicted voter support may translate into electoral success.

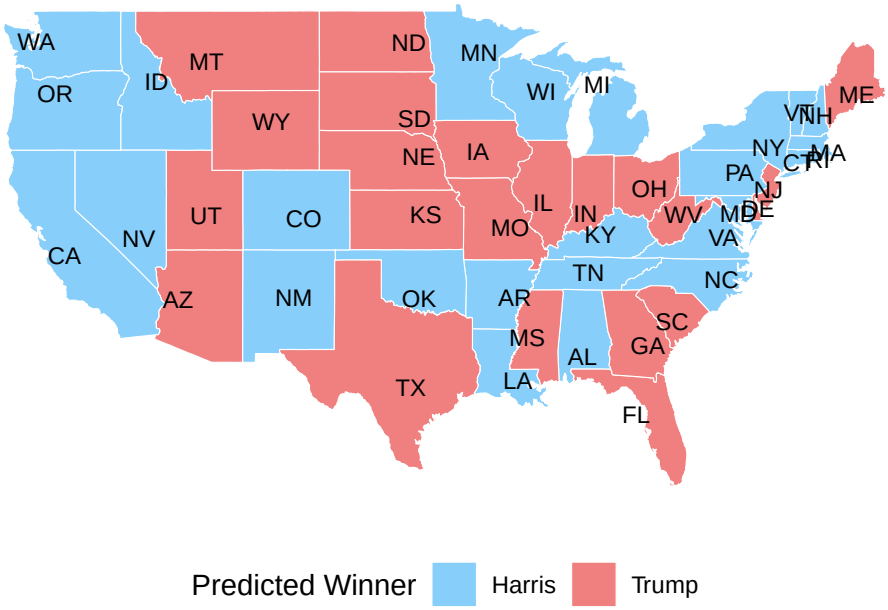

Figure 2: Electoral College Predition Map

Figure 2 displays a predicted U.S. Electoral College outcome in a hypothetical race between Harris and Trump, with blue and red indicating states projected to support each candidate, respectively. Harris is expected to win in traditional Democratic strongholds like the West Coast, Northeast, and parts of the Midwest, areas historically aligned with Democratic candidates due to factors such as urbanization, higher education levels, and diverse populations. In contrast, Trump's projected wins across the South and Great Plains reflect longstanding Republican preferences in these regions, shaped by conservative values and economic policies that resonate in rural and suburban areas. The distribution highlights both entrenched partisan divides and key battleground states, such as Florida and Pennsylvania, that could be decisive given their recent variability in party support.

5 Discussion

5.1 Summary of Findings

In this paper, we developed statistical models to forecast the outcome of the 2024 United States presidential election between Kamala Harris and Donald Trump. Utilizing polling data from FiveThirtyEight's 2024 Presidential Election Forecast Database, we analyzed factors influencing candidate support through linear and multiple linear regression models. We incorporated variables such as pollster rating, transparency score, sample size, and state-level data to predict the percentage of support for each candidate across different states. By aggregating these

state-level predictions and integrating electoral vote allocations, we projected that Kamala Harris would receive 306 electoral votes, surpassing the 270 required to win the presidency.

5.2 Polling Methodologies and Voter Support

Our analysis indicates that pollster rating and state-level factors significantly influence reported support for both candidates. For Kamala Harris, the multiple linear regression model shows that higher pollster ratings and certain states are associated with increased support percentages. This suggests that polls conducted by more reliable pollsters may report higher support for Harris, possibly due to methodological rigor or sampling techniques that capture her voter base more accurately. For Donald Trump, pollster transparency and state differences play a notable role, indicating that how openly pollsters disclose their methodologies can affect the reported support for Trump.

5.3 Implications for Electoral Forecasting

The projection of Kamala Harris receiving 306 electoral votes underscores the importance of incorporating state-level analyses in electoral forecasting. Our findings highlight that national polling averages may not sufficiently capture the complexities of the Electoral College system. By focusing on state-specific predictions and accounting for variables that influence voter support in different regions, forecasters can achieve more accurate predictions. This approach acknowledges the heterogeneity of the electorate and the pivotal role of battleground states in determining election outcomes.

5.4 Limitations

Despite the strengths of our modeling approach, several limitations must be acknowledged. First, the reliance on historical polling data and pollster ratings assumes that past performance is indicative of future accuracy, which may not hold true if polling methodologies or voter behaviors change significantly. Second, the underrepresentation of certain states due to sparse polling data may affect the reliability of our state-level predictions. Assigning a default 50% support in states with missing data introduces uncertainty and may not reflect actual voter preferences. Additionally, the models may not fully account for dynamic factors such as late-breaking events, shifts in voter sentiment, or turnout variations that can influence election results.

5.5 Future Research and Recommendations

To enhance the accuracy of electoral forecasts, future research should consider incorporating real-time data and alternative data sources such as social media trends, economic indicators, and demographic shifts. Employing more sophisticated modeling techniques, such as hierarchical Bayesian models or machine learning algorithms, may capture complex interactions and non-linear relationships among variables. Furthermore, improving polling methodologies to address issues of response bias and sample representativeness is essential. Collaborations between pollsters, statisticians, and political scientists can foster the development of more robust predictive models that adapt to the evolving electoral landscape.

A Appendix

A.1 YouGov Pollster Methodology Overview and Evaluation

YouGov(YouGov 2024) utilizes an online panel with non-probability sampling, adjusting for demographic imbalances through weighting techniques. This approach ensures that survey results are representative of the broader population, despite the limitations of non-random sampling.

The primary strength of YouGov's methodology lies in its efficiency and cost-effectiveness, enabling quick data collection. However, the reliance on non-probability sampling can introduce biases, which are partially mitigated by their rigorous weighting process. While effective in providing timely findings, this method requires careful consideration of its limitations in representativeness.

A.1.1 Survey Population and Sampling

YouGov's surveys target U.S. adults, utilizing an opt-in online panel that covers approximately 95% of the population. For general population surveys, they typically sample between 1,000 and 2,000 respondents. These participants are selected to reflect key demographic and political characteristics, ensuring the sample is representative of the broader population.

A.1.2 Panel Recruitment and Participation

YouGov(YouGov 2024) recruits panel members through targeted advertising and partnerships with various websites. Surveys are available in multiple languages, including Spanish, to ensure inclusivity and broad demographic representation. Participants are incentivized with points redeemable for small monetary rewards, though many join to contribute to research efforts.

A.1.3 Quality Control

YouGov(YouGov 2024) employs several measures to maintain data quality:

- Verification of panelist identity through email and IP checks
- Response quality surveys to gauge reliability
- Monitoring of response times and patterns
- Removal of respondents who fail quality checks
- Question randomization to reduce bias

A.1.4 Non-response and Weighting

To address potential biases, (YouGov 2024) applies statistical weighting based on demographics (age, gender, race, education) and political factors (voting behavior, party identification). Their weighting process considers multiple characteristics simultaneously to better reflect real-world demographic intersections.

A.1.5 Strengths and Limitations

The key strengths of YouGov's methodology are its efficient data collection, cost-effectiveness, and capacity to track public opinion trends over time. However, the reliance on non-probability sampling can introduce biases, and the online-only format may lead to underrepresentation of certain populations. While sophisticated weighting techniques mitigate some of these issues, they cannot fully eliminate all potential biases.

A.2 Idealized Survey Methodology

This idealized survey methodology presents a structured plan for forecasting the U.S. presidential election within a $100,000 budget. It balances statistical rigor with practicality to ensure accurate predictions of both the popular vote and Electoral College outcomes. The methodology includes developed sampling techniques, robust survey design, and detailed data analysis to capture key demographic and political variables. By addressing potential biases and focusing on cost-effective data collection, this plan ensures reliable and representative results into voter behavior.

A.2.1 Sampling Strategy

The target population for this survey is eligible voters across the United States who are likely to participate in the upcoming presidential election. To achieve a representative sample:

- **Sampling Frame:** Utilize a combination of registered voter lists and demographic data from reputable sources such as the US Census Bureau.
- **Sampling Method:** Implement stratified random sampling to ensure representation across key demographics, including age, gender, race, education level, and geographic location.
- **Sample Size Calculation:** Aim for a sample size of approximately 10,000 respondents to achieve a margin of error of $\pm 1\%$ at a 95% confidence level.
- **Geographical Distribution:** Allocate samples proportionally across all 50 states and the District of Columbia, with oversampling in swing states to better predict electoral college outcomes.
- **Addressing Sampling Biases:** Apply weighting adjustments to account for underrepresented groups and ensure that the sample mirrors the overall voter population.

A.2.2 Recruitment Plan

To recruit respondents effectively, we will utilize online panels, social media advertising, and partnerships with community organizations to reach a diverse audience. Offering modest incentives, such as $5 digital gift cards, encourages participation while managing costs. Quota sampling within strata maintains demographic balance, and follow-up reminders along with mobile-friendly survey formats help reduce non-response bias. The data collection will occur over a two-week period to capture timely opinions without introducing temporal biases.

A.2.3 Survey Design Elements

The survey is crafted to elicit accurate and meaningful responses:

- **Question Types and Formats:** Use a mix of closed-ended questions and multiple-choice options for clarity and ease of analysis.
- **Response Options:** Include balanced and neutral response choices, with options for "Undecided" or "Prefer not to say."
- **Question Order and Flow:** Begin with general questions to build rapport, followed by more specific vote intention queries, and conclude with demographic questions.
- **Demographic Information:** Collect data on age, gender, race, education, income, and geographic location.
- **Political Affiliation and History:** Ask about party affiliation, past voting behavior, and political engagement.
- **Likely Voter Screens:** Include questions to gauge voting likelihood, such as past voting frequency and intention to vote in the upcoming election.
- **Vote Intention Questions:** Directly ask which candidate the respondent intends to vote for, ensuring confidentiality and anonymity.

A.2.4 Quality Control

To maintain data integrity, we implement several quality control measures. Real-time validation checks within the survey prevent inconsistent or illogical responses. Attention-check questions identify disengaged respondents. We use unique survey links and track IP addresses to prevent duplicate submissions, while CAPTCHA verification deters automated responses. Incomplete or suspicious responses are excluded during data cleaning to ensure the final dataset is robust and reliable.

A.2.5 Data Processing

These data processing steps will be taken to ensure accurate analysis:

- **Weighting Methodology:** Adjust survey results using weighting factors based on demographic proportions in the voting population.
- **Handling Missing Data:** Employ imputation techniques or exclude cases with significant missing information.
- **Outlier Detection:** Identify and review outliers that may skew results, determining whether to retain or discard them.
- **Response Validation:** Cross-check responses for consistency and plausibility.
- **Poll Aggregation Approach:** Combine survey data with other reputable polls using meta-analytic techniques to enhance prediction accuracy.

A.2.6 Budget Allocation

A budget allocation of $100,000 ensures all aspects are adequately funded:

- **Recruitment Costs:** $40,000 for advertising and partnerships to reach potential respondents.
- **Incentive Payments:** $50,000 allocated for participant incentives ($5 x 10,000 respondents).
- **Survey Platform Fees:** $2,000 for premium features on a survey platform like Google Forms or an equivalent.
- **Data Analysis Tools:** $3,000 for statistical software licenses and data processing tools.
- **Quality Control Measures:** $3,000 for implementing validation systems and CAPTCHA services.
- **Administrative Costs:** $2,000 for project management and miscellaneous expenses.

A.2.7 Google Survey Link

https://forms.gle/81vCBvH8t7B9NfJ69

A.2.8 Google Survey

U.S. Presidential Election Survey 2024

-This survey is part of a research study aimed at forecasting the 2024 U.S. Presidential Election. Your responses are confidential and will be used solely for research purposes. This survey will take approximately 5 minutes to complete.

Are you eligible to vote in the 2024 U.S. Presidential Election?

-Yes -No

Are you planning to vote in the upcoming 2024 U.S. Presidential Election?

-Yes -No -Maybe

What is your current party affiliation? -Democratic Party -Republican Party -Independent -Other (please specify) -Prefer not to say

Did you vote in the 2020 U.S. Presidential Election? -Yes -No -Prefer not to say

If the presidential election were held today, which candidate would you most likely vote for? -Kamala Harris (Democratic Party) -Donald Trump (Republican Party) -Another candidate (please specify) -Undecided -Prefer not to say

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

Least likely | 1 | 2 | 3 | 4 | 5 | Most likely

What is your age?

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

What is your race/ethnicity? -White -Black or African American -Hispanic or Latino -Asian -Native American or Alaska Native -Native Hawaiian or Pacific Islander -Prefer not to say -Other (please specify)

What is your highest level of education? -Less than high school -High school graduate -Some college -Bachelor's degree -Postgraduate degree -Prefer not to say

What is your approximate annual household income? -Less than 25,000 -25,000 to 49,999 -50,000 to 74,999 -75,000 to 99,999 -100,000 or more -Prefer not to say

To ensure the quality of our data, please select "Blue" from the options below:

- Red
- Blue
- Green
- Yellow

Conclusion

Thank you for your participation!
Your responses have been recorded.

A.2.9 Conclusion

This methodology presents a feasible and thorough plan to forecast the US presidential election within the specified budget. By adhering to best practices in survey design and execution, and by carefully considering both the popular vote and electoral college implications, the survey aims to provide accurate and reliable results into voter intentions.

B Appendix

B.1 Data Processing Steps

Our cleaning process followed several key steps:

1. Filtered for high-quality polls using a minimum threshold ($numeric\_grade \geq 2.7$)
2. Limited temporal coverage to post-campaign announcement period (after July 21, 2024)
3. Standardized geographic data:
    i. Coded missing state information as "National" polls
    ii. Verified state names for consistency
4. Transformed key variables:
    i. Converted *End Date* values to standardized date format
    ii. Calculated absolute supporter numbers from percentages and sample sizes
    iii. Created binary candidate indicators (Harris = 1, Trump = 0)

B.2 Data Tables and Figures

Table 4: Sample Overview of Selected Variables in the Polling Dataset

	Row 1	Row 2	Row 3	Row 4	Row 5
pollster	Washington Post/George Mason University	CNN/SSRS	Siena/NYT	YouGov	CNN/SSRS
pollscore	-0.8	-0.6	-1.5	-1.1	-0.6
numeric_grade	2.7	2.8	3.0	3.0	2.8
transparency_score	9	10	9	9	10
sample_size	1005	931	1142	1033	708
methodology	Live Phone/Text-to-Web	Live Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	Live Phone	Online Panel	Probability Panel
pct	42	48	48	47	48
state	Virginia	North Carolina	National	National	Michigan
candidate_name	Donald Trump	Donald Trump	Donald Trump	Donald Trump	Kamala Harris
end_date	2024-09-08	2024-09-25	2024-07-24	2024-10-04	2024-08-29

Table 4 provides an overview of key variables from different polls conducted by various organizations, including Washington Post/George Mason University, CNN/SSRS, Siena/NYT, and YouGov. It highlights differences in methodologies, such as live phone interviews, online panels,

and mixed methods, as well as sample sizes ranging from 708 to 1,142 respondents. Pollscores, which reflect pollster reliability, vary from -1.5 (indicating higher reliability for Siena/NYT) to -0.6 (for CNN/SSRS). Transparency scores show that YouGov and CNN/SSRS exhibit full transparency with perfect scores of 10.

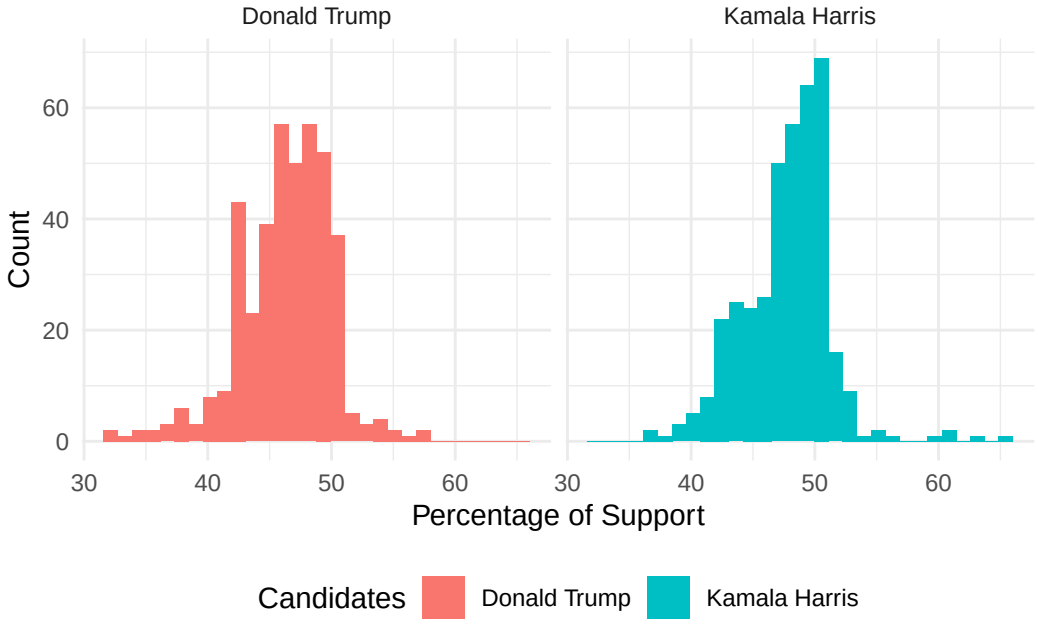

Figure 3: Distribution of Support Percentages by Candidate

Figure 3 shows the distribution of support percentages for Donald Trump and Kamala Harris. Both distributions are centered around the mid-40% to mid-50% range, reflecting a competitive race. Trump's support is more concentrated around the 45-50% range, indicating relatively consistent polling results. Harris's distribution is slightly more varied, suggesting greater fluctuation in her polling support. This visualization highlights the close contest between the two candidates, with neither consistently achieving overwhelming support across polls.

Table 5: Summary Statistics of Key Poll Metrics

Statistic	sample	pollscore	numeric	transparency
Mean	1114.76	-1.06	2.89	8.59
Median	1000.00	-1.10	2.90	9.00
SD	583.72	0.28	0.10	1.04
Min	450.00	-1.50	2.70	6.00
Max	4253.00	-0.50	3.00	10.00

Table 5 presents the summary statistics for key poll metrics, including sample size, pollscore (pollster reliability), numeric grade, and transparency score. The average sample size across

polls is 1,114.76 respondents, with a standard deviation of 583.72, indicating variability in the number of participants. The median sample size is 1,000, with a minimum of 450 and a maximum of 4,253 respondents.

The average pollscore is -1.06, with a median of -1.10, reflecting generally high reliability across pollsters (lower scores indicate greater reliability). The numeric grade, which measures methodological rigor, averages 2.89 with minimal variability (SD of 0.10). The transparency score averages 8.59, suggesting most polls are transparent about their methods, though scores range from 6.00 to 10.00. These statistics highlight the overall quality and variation in polling practices used in the analysis.

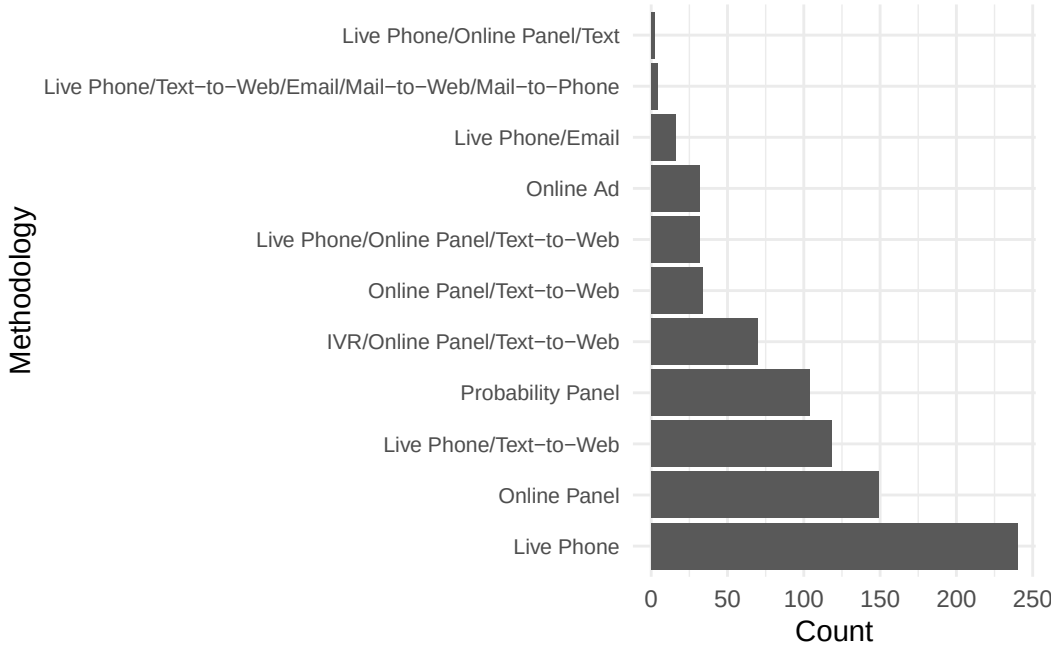

Figure 4: Distribution of Polling Methodologies

Figure 4 illustrates the distribution of various polling methodologies used in the dataset. The most frequently employed methods include "Live Phone," "Online Panel," and "Live Phone/Text-to-Web," highlighting a preference for direct communication methods combined with online components. Less common methodologies, such as "Live Phone/Online Panel/Text" and "Live Phone/Email," appear at the lower end of the distribution. This diversity in polling methods reflects the varied approaches taken by pollsters to gather representative data, balancing traditional techniques with modern, web-based options.

Figure 5 shows the distribution of sample sizes across the polls in the dataset. The majority of polls have sample sizes clustered around 1,000 respondents, which is a typical size for achieving a balance between precision and cost. A few polls have significantly larger sample sizes, extending beyond 2,000, with some even exceeding 4,000 respondents. This suggests

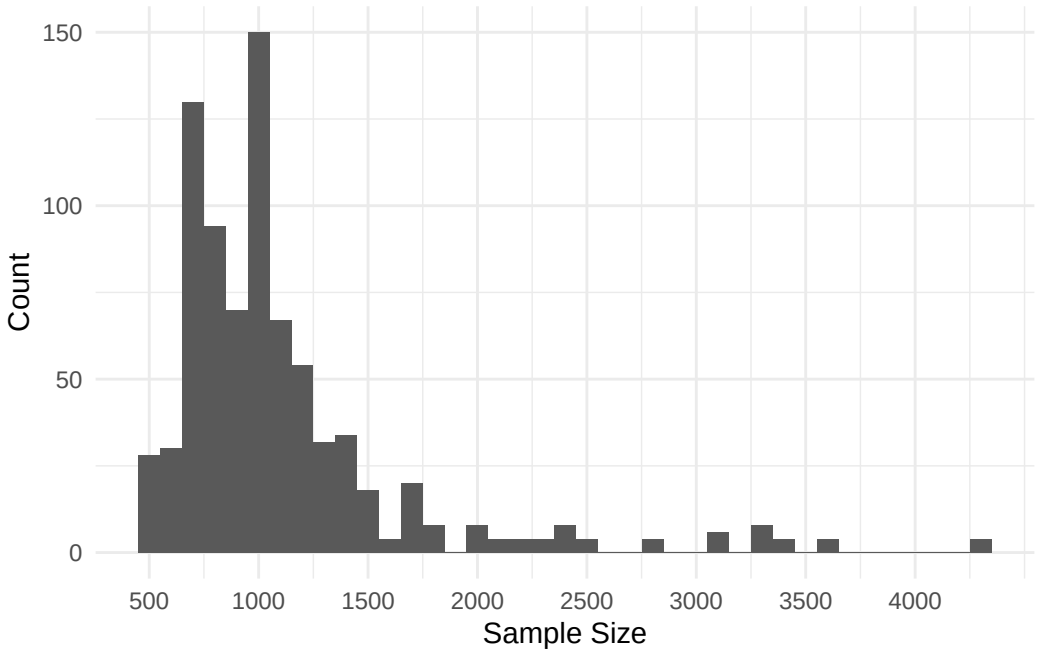

Figure 5: Distribution of Poll Sample Sizes

that while most polls aim for a moderate sample size, there are instances where more extensive sampling is conducted, possibly for greater accuracy or in significant regions. The distribution reflects the common practice of balancing statistical reliability with logistical constraints.

Table 6 illustrates the number of polls conducted over time, showing weekly fluctuations in polling activity. There is a noticeable increase in the number of polls around mid-September, peaking at over 160 polls in a week, followed by a decline as the election date approaches. This pattern suggests heightened polling activity during key points in the campaign, likely caused by major events or debates, before tapering off closer to the election. The variability highlights how polling frequency can change in response to the evolving political landscape.

Figure 6 provides the polling frequency by state, highlighting where polling efforts were concentrated. National polls dominate with 261 instances, while key battleground states like Pennsylvania (86 polls), Wisconsin (82 polls), and North Carolina (64 polls) also see significant attention. These states are important for determining election outcomes, explaining the higher frequency. Conversely, states like Missouri, Montana, and Nebraska show fewer polls (4 each), reflecting less competitive races or a perceived lower impact on the overall election result. This distribution underscores the strategic focus on swing states in polling efforts.

Figure 7 displays a scatter plot with a linear regression line showing the relationship between sample size and the percentage of support for Kamala Harris. The plot suggests a very weak positive correlation, as indicated by the near-flat regression line. This implies that variations in sample size have a minimal effect on the percentage of support for Harris. The spread

Table 6: Polling Frequency by State

State	Number of Polls
National	261
Pennsylvania	86
Wisconsin	82
North Carolina	64
Michigan	58
Georgia	56
Arizona	52
Nevada	22
Minnesota	12
Nebraska CD-2	12
Texas	12
Virginia	12
Florida	10
New York	10
Ohio	10
Massachusetts	6
New Mexico	6
Missouri	4
Montana	4
Nebraska	4
New Hampshire	4
Rhode Island	4
California	2
Connecticut	2
Indiana	2
Iowa	2
Maryland	2

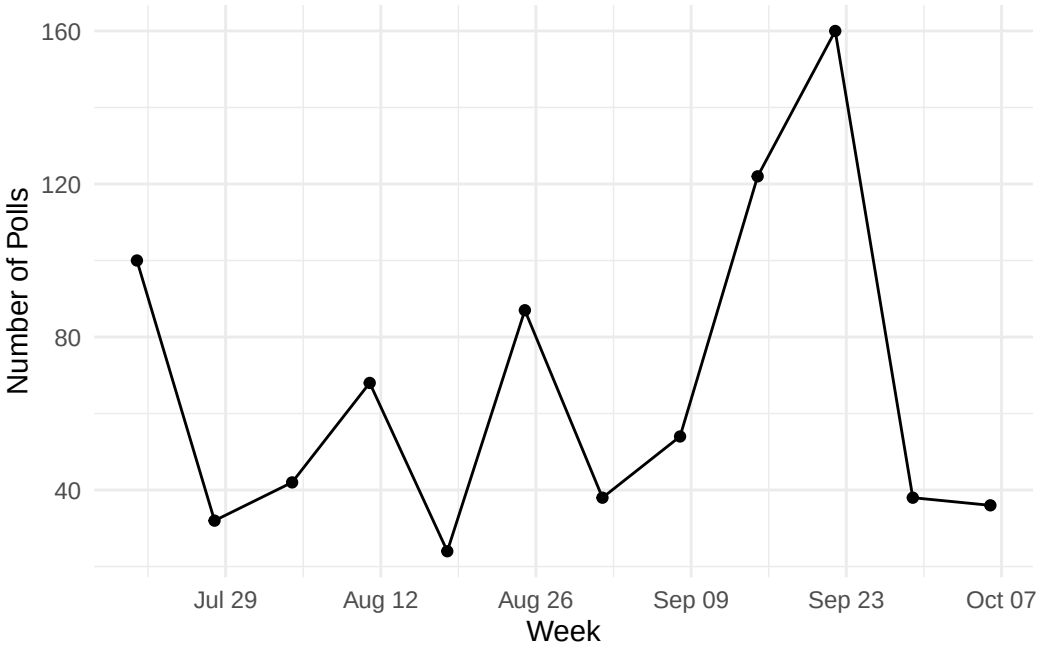

Figure 6: Number of Polls Over Time

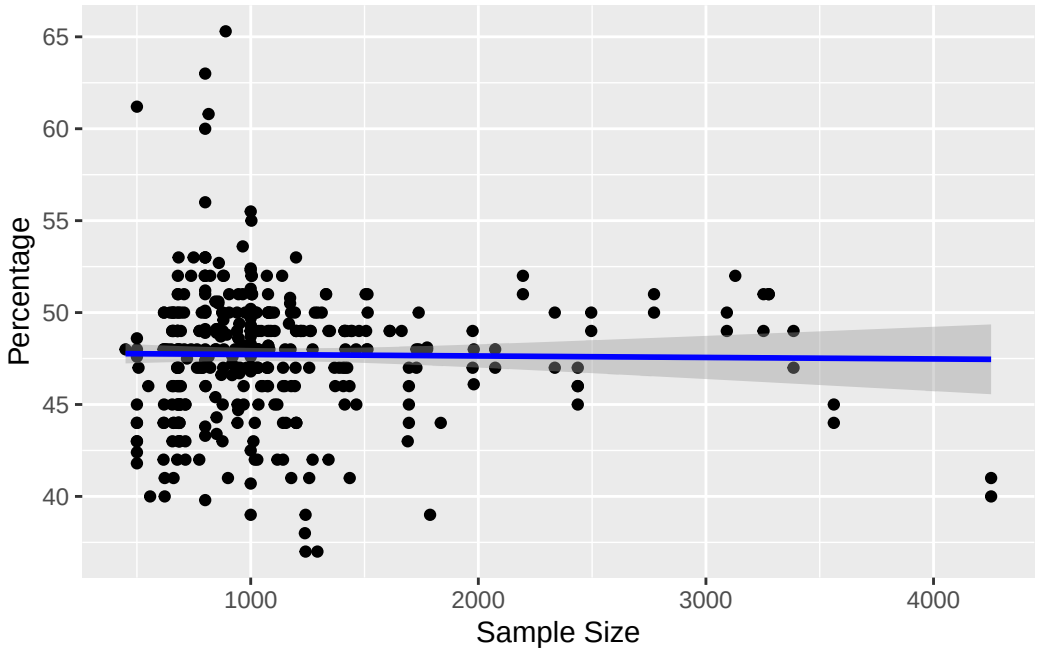

Figure 7: Linear Regression of Percentage vs Sample Size for Kamala Harris

of data points shows a consistent range of support percentages across different sample sizes, reinforcing the observation that larger sample sizes do not significantly alter her support levels in the polls.

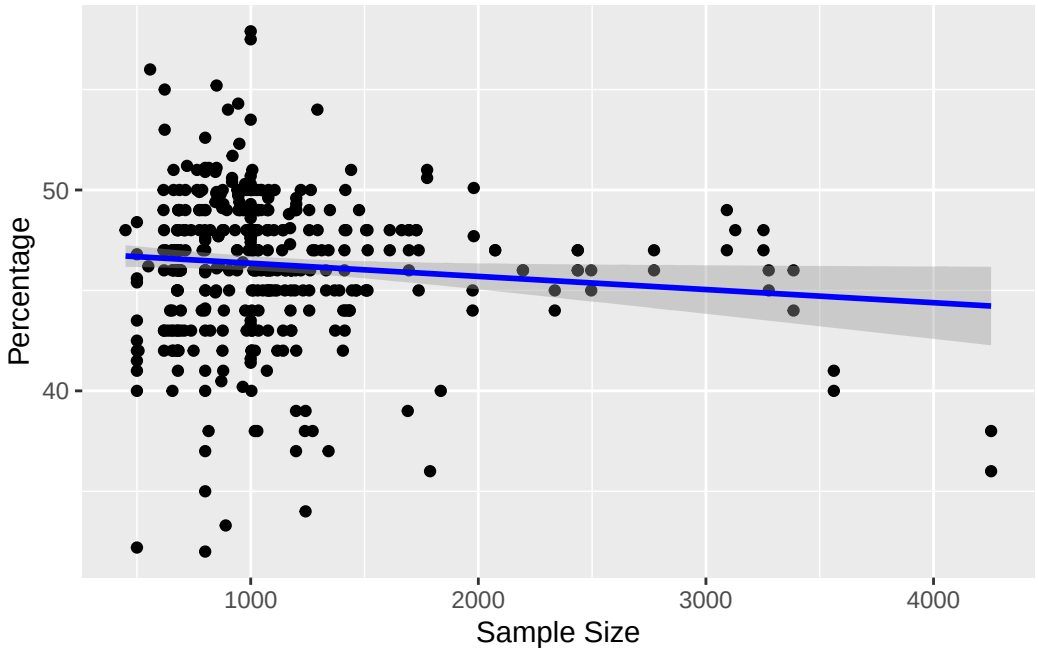

Figure 8: Linear Regression of Percentage vs Sample Size for Donald Trump

Figure 8 shows a weak negative relationship between poll sample size and Trump's support percentage, as indicated by the slight downward slope of the regression line. While larger sample sizes are associated with a marginal decrease in reported support, the scattered distribution of data points suggests considerable variability, implying that other factors like pollster methodology or regional differences significantly influence the results. The narrow confidence interval around the trend line reflects a modest level of certainty in this weak relationship, though the presence of outliers highlights exceptions where larger samples do not necessarily align with the general trend.

Figure 9 The **Residuals vs Fitted** plot (left) for Kamala Harris's multi-linear regression model indicates that most residuals are evenly distributed around the horizontal axis (residuals = 0), suggesting that the model's predictions are unbiased across different fitted values. However, the slight curve near the middle of the plot implies potential non-linearity or missing variables. Additionally, some points (e.g., labeled 263, 352, 370) are noticeable outliers, which may affect the model's accuracy.

The **Q-Q plot** (right) shows the standardized residuals against the theoretical quantiles. The residuals mostly follow the straight line, indicating that they are approximately normally distributed. However, deviations at both tails (particularly on the left) suggest the presence of

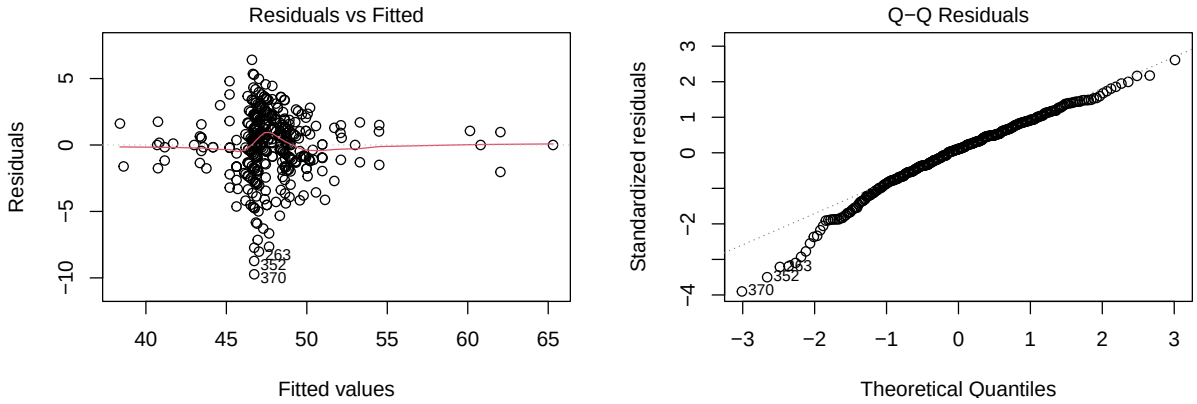

Figure 9: Multi-Linear Regression model for Kamala Harris

some extreme values, which could impact the normality assumption and model reliability.

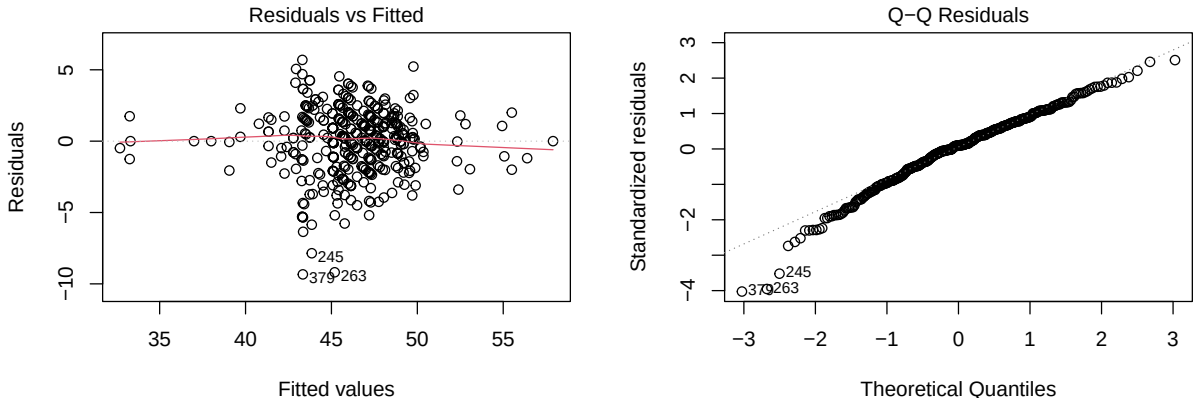

Figure 10: Multi-Linear Regression model for Donald Trump

Figure 10, we observe diagnostic plots for Donald Trump's multi-linear regression model.

The Residuals vs Fitted plot (left) shows that the residuals are scattered fairly evenly around the horizontal line at zero, indicating no major patterns of bias across the range of fitted values. However, the slight curve in the middle suggests potential mild non-linearity or missing variables in the model. A few outliers, such as observations 245, 379, and 263, stand out and may warrant further investigation to assess their influence on the model's performance.

The Q-Q plot (right) for standardized residuals displays a relatively straight line, implying that the residuals are approximately normally distributed. There are slight deviations at both extremes, indicating some outliers or deviations from normality at the tails. This could hint at potential issues in capturing extreme values accurately but generally suggests the model performs well in satisfying the normality assumption.

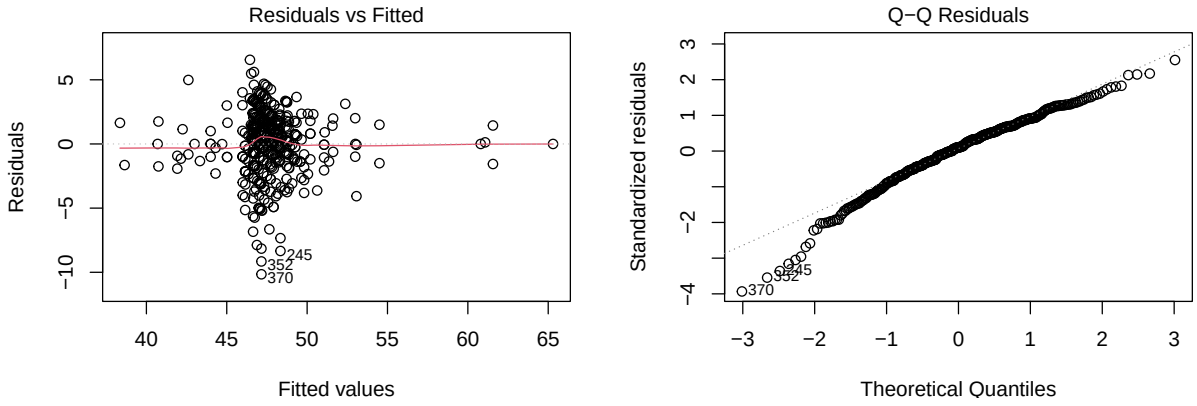

Figure 11: Harris Refined Multi-Linear Regression Model

Figure 11, we see diagnostic plots for the refined multi-linear regression model for Kamala Harris.

The Residuals vs Fitted plot (left) shows that the residuals are relatively evenly distributed around the horizontal line at zero, indicating that there are no major issues with non-linearity or heteroscedasticity. The slight curve in the red line suggests minimal deviation from linearity, which could be due to the data's inherent characteristics or model limitations. There are a few noticeable outliers, like observations 245, 352, and 370, which might need further investigation to ensure they do not disproportionately influence the model.

The Q-Q plot (right) demonstrates that the residuals generally follow a normal distribution, as indicated by the points aligning closely with the 45-degree line. However, some deviations at the tails suggest the presence of outliers or slight departures from normality. This generally supports the adequacy of the model but points to areas where model refinement or robust handling of outliers might improve fit.

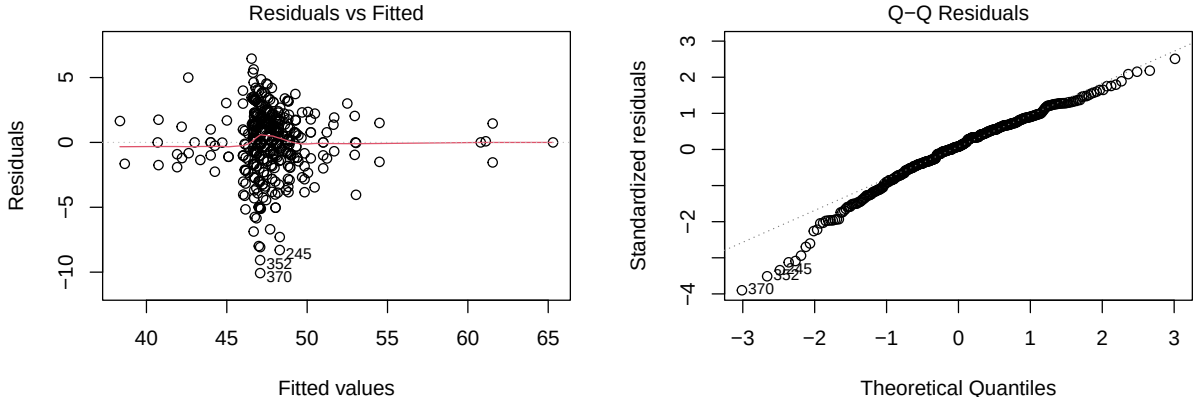

Figure 12: Harris Final Multi-Linear Regression Model

Figure 12, presents the diagnostic plots for the final multi-linear regression model for Kamala Harris.

The Residuals vs Fitted plot (left) reveals a fairly uniform spread of residuals around the zero line, suggesting that the model captures the linear relationship well without major systematic errors. The red line, representing the smoothed residuals, remains close to zero, with slight deviations indicating minimal non-linearity. A few outliers, notably points 245, 352, and 370, are present but do not seem to heavily distort the overall pattern, though their influence may need checking.

The Q-Q plot (right) shows that the residuals align closely with the theoretical quantiles of a normal distribution, confirming that the assumption of normally distributed errors holds reasonably well. The points at the extremes deviate slightly, indicating minor issues with the tails of the distribution, but these are not severe. This suggests that the model assumptions are largely satisfied, making the model robust for the intended analysis.

References

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

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

Fox, John, and Sanford Weisberg. 2019. *An r Companion to Applied Regression*. https://cran.r-project.org/package=car.

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/.

Keeter, Scott. 2024. "Key Things to Know about U.S. Election Polling in 2024." https://www.pewresearch.org/short-reads/2024/08/28/key-things-to-know-about-us-election-polling-in-2024/.

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

Viala-Gaudefroy, Jérôme. 2024. "2024 US Presidential Election: Can We Believe the Polls?" *The Conversation*. https://theconversation.com/2024-us-presidential-election-can-we-believe-the-polls-240834.

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

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

YouGov. 2024. "YouGov Polling Data." https://ca.yougov.com/en-ca/.