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

An Investigation into Poll Data Using Multiple Linear Regression

Xuanle Zhou Yongqi Liu Yuxuan Wei

October 31, 2024

The study predicting the 2024 U.S. presidential election is important as it provides the predicted outcomes for the US election. Analysis is based on state-by-state electoral votes, and these outcomes can influence campaign strategies and shape public discourse. The result indicates Kamala Harris will win by over 270 votes in this election cycle.

Table of contents

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

1 Introduction

The 2024 U.S. Presidential Election is set to unfold in a climate of heightened political polarization, economic challenges, and demographic shifts, making it a pivotal event in shaping America's future. Predicting the outcome of this election is crucial for gaining insights into potential changes in political leadership and policy directions. Polling data has traditionally served as a primary tool for gauging voter sentiment; however, recent election cycles have exposed limitations in polling accuracy due to biases, non-response issues, and variations in poll quality. To address these challenges, this study proposes a model that incorporates these factors to enhance the reliability of election forecasts.

This paper develops a prediction model aimed at forecasting 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, our model seeks to provide a nuanced forecast that accounts for regional differences and methodological variations across polls, offering a comprehensive view of the projected support for each candidate.

Our findings suggest a competitive election landscape, with Harris showing strong prospects in key battleground states while Trump retains support in traditional Republican strongholds. These projections underscore the enduring influence of geographic voting patterns and the impact of polling methodology on electoral forecasting. The study provides valuable insights for political analysts, campaign strategists, and the public, contributing to a better understanding of how polling data can be utilized to anticipate electoral outcomes.

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

2 Data

2.1 Overview

This study uses R packages (R Core Team 2023) to clean and analyz the dataset , including libraries from tidyverse [], ggplot2[].

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

The dataset used in this analysis originates from FiveThirtyEight, a trusted source known for its rigorous methodology in collecting and aggregating polling data to represent public opinion accurately. To maintain data integrity and ensure a high standard of reliability, only polls meeting stringent quality criteria are included in the dataset. Specifically, each poll must disclose crucial information, including the pollster's name, the exact dates of the survey, the sample size, and the full methodological details. These details encompass the polling medium (e.g., phone, online), use of verified voter files, demographic weighting criteria, and any adjustments applied to reflect a representative sample.

Polls are rigorously verified, and those that do not meet these requirements are excluded from the dataset. For instance, polls categorized as nonscientific, those that merge data from multiple unrelated sources, and those conducted by hobbyists or unverified entities are filtered out to prevent data quality compromise.

Once deemed credible, polls are incorporated into the FiveThirtyEight database. These polls then contribute to polling averages, forecast models, and political coverage, which depend heavily on the quality and reliability of public opinion data. This meticulous selection process ensures that the dataset captures a comprehensive and nuanced view of public sentiment, providing an accurate reflection of behavioral trends and preferences over time.

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 in Figure 1 indicates that most observations cluster around a support percentage of approximately 48%, suggesting moderate backing from the electorate. Additionally, a smaller proportion of polls show support exceeding 55%, indicating that while Trump has a core base, many voters remain either indifferent or opposed to him.

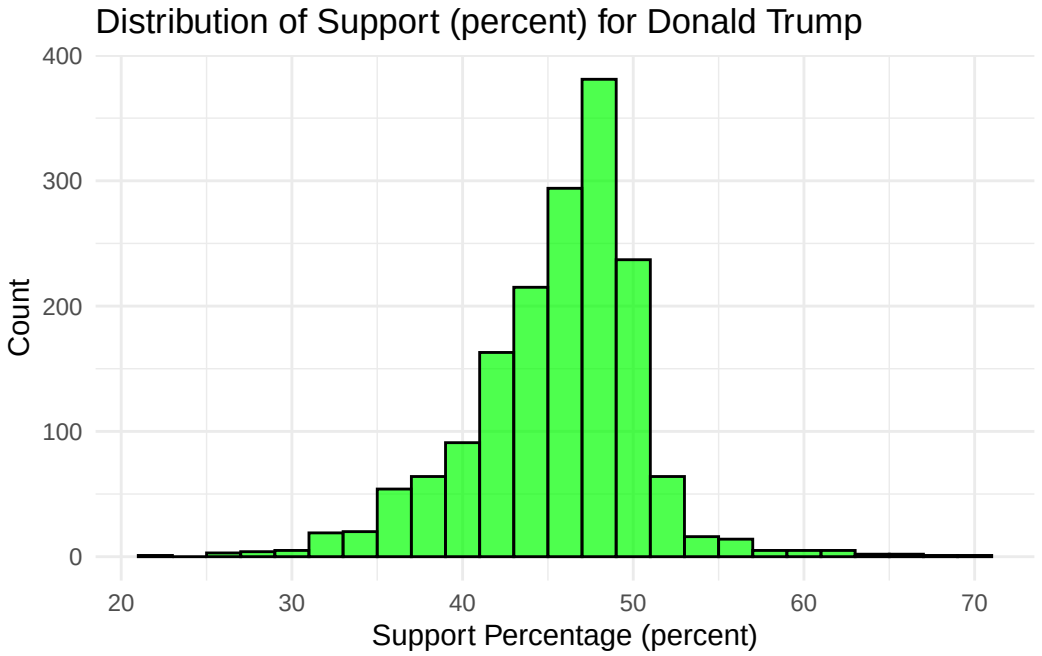

Figure 1: Distribution of Support (percent) for Donald Trump

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, with half of them scoring between 2.6 and 2.8, as shown in Figure 2 .

2.4.2 Sample Size

The sample size indicates the number of respondents in each poll. The distribution in Figure 3 exhibits a right-skewed shape, suggesting that there are more observations with smaller sample sizes compared to larger ones. The peak of the distribution is around 1,000, indicating that this is the most common sample size used in the polls. Overall, the sample size data highlights that each poll contains a sufficient number of respondents to provide reliable insights.

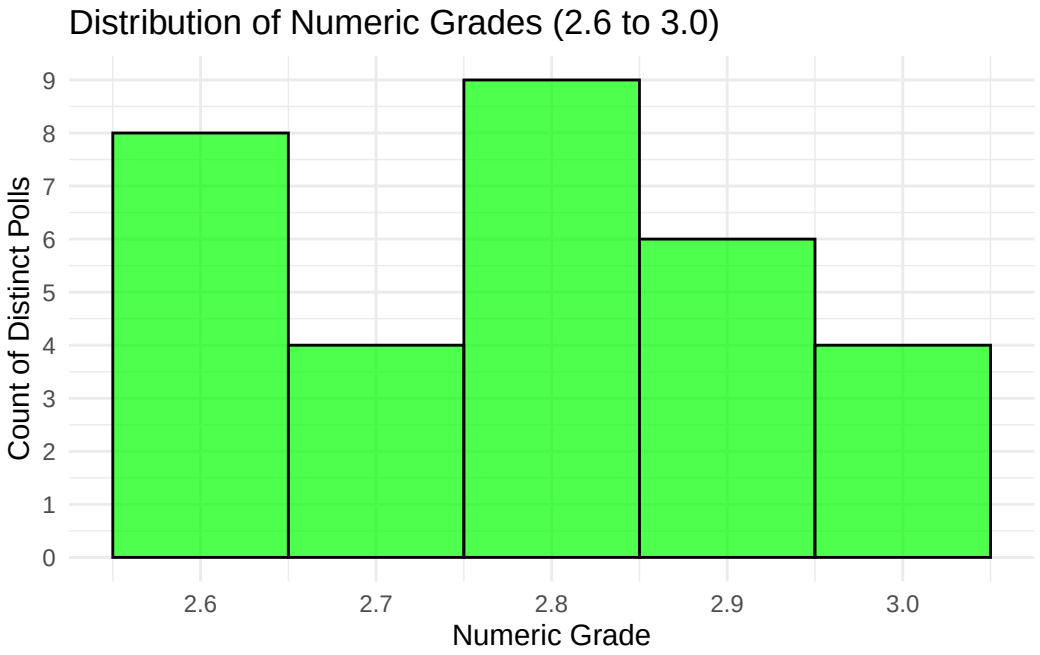

Figure 2: Distribution of Numeric Grades (2.6 to 3.0)

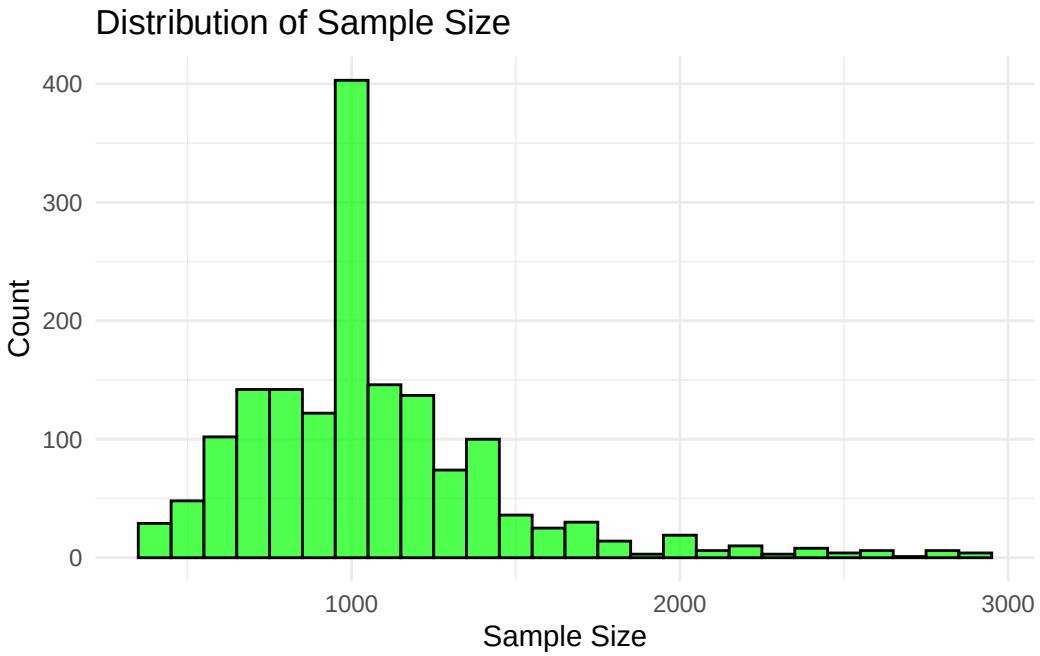

Figure 3: Distribution of Sample Size

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.

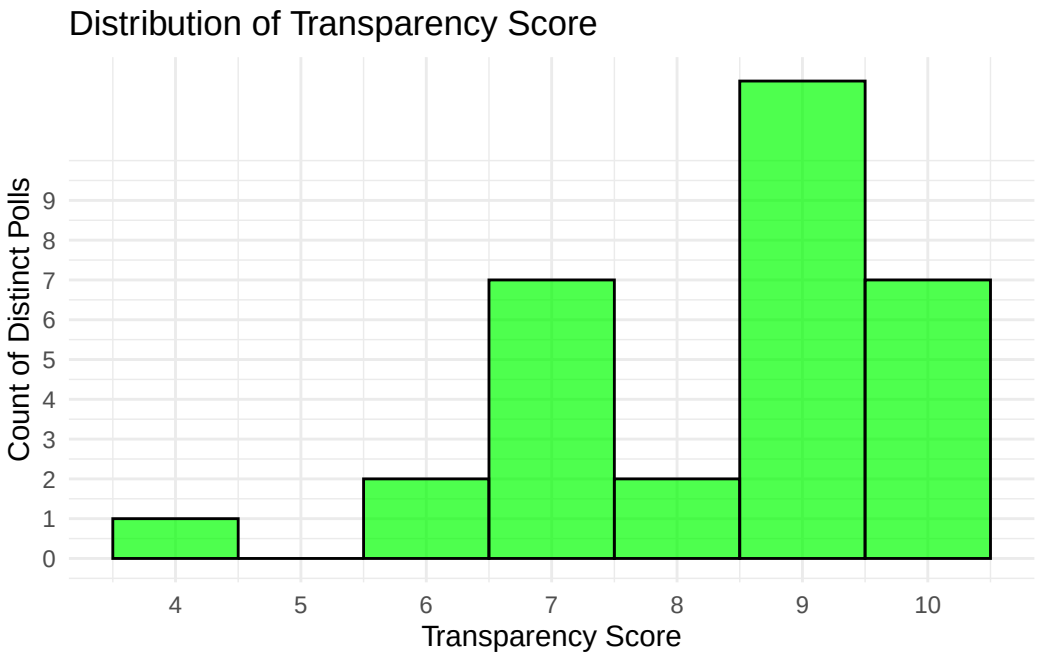

Figure 4: Distribution of Transparency Score

2.4.4 Days Until Election

The **days until election** represents the remaining time leading up to the final election day. As shown in the figure, the support percentage for Donald Trump fluctuated over time, with a slightly downward trend as the election approached, as indicated by the dashed trend line. This suggests that while public opinion varied significantly, there was a decrease in support as election day approached. This trend highlights the importance of monitoring real-time polling data, as sentiment may shift in the critical months leading up to a closely contested election.

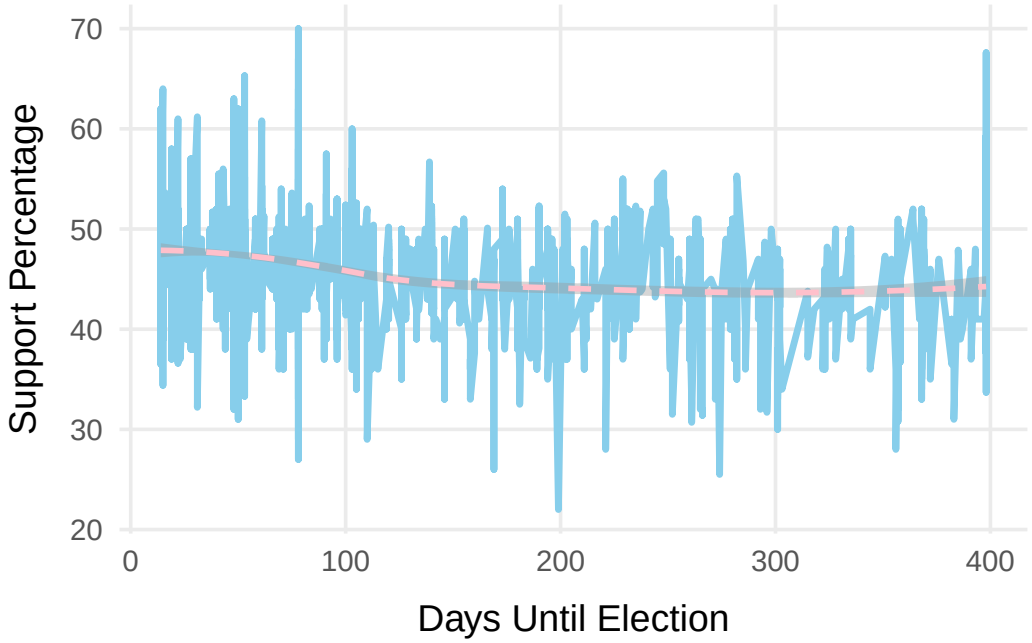

Figure 5: Trend of Support Percentage for Donald Trump Over Days Until Election. Fluctuations in support levels are observed as the election approaches, with a slight downward trend indicated by the dashed curve.

3 Model

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.

3.1 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:

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

3.2 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 increased as the election date approaches, while a negative coefficient would suggest a decrease in his polling numbers over time.The coefficient for end_date informs us about how Trump's polling percentages have evolved over time. A positive coefficient would suggest an upward trend, while a negative coefficient would indicate a decline.

3.3 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

	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

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 critical 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
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

Table 2: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
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 following table (Table 3) serves as a supplement to the predicted outcomes, filling in the gaps for states where the model couldn't provide predictions. It allows for a complete view of electoral projections by assigning a likely outcome in each missing state based on historical data. This approach ensures every state is accounted for in the final electoral vote totals.

Table 3: 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
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

Table 3: Missing States with Predicted Winner Based on Historical Lean

State	Electoral Votes	Prediction Based on Historical Lean
Wyoming	3	Trump
District of Columbia	3	Harris

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.

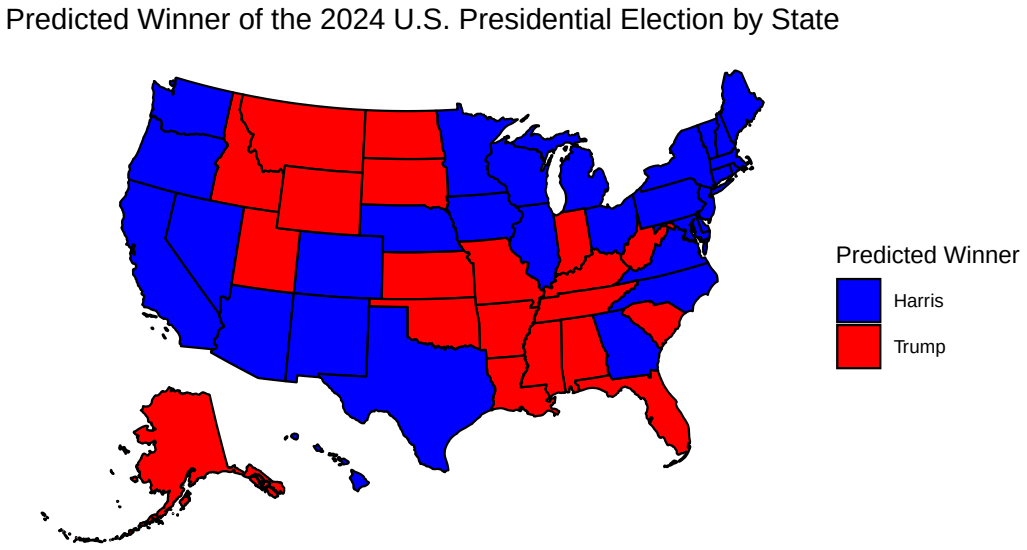

Figure 6

The table below (Table 4) 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 4: Numerical Predicted Winner of 2024 US Presidential Election by State

Candidate	Total Electoral Votes
Trump	151
Harris	387

Table 4: Numerical Predicted Winner of 2024 US Presidential Election by State

Candidate	Total Electoral Votes

5 Discussion

5.1 Overview of Predictive Modeling for the 2024 U.S. Election

This paper employs a regression model to forecast the 2024 U.S. Presidential Election outcomes between Kamala Harris and Donald Trump. By analyzing historical state-level voting data, we estimated Trump's vote percentages across states and calculated each candidate's projected electoral votes. Our model predicts a significant victory for Harris, who is projected to secure 387 electoral votes compared to Trump's 151. This finding underscores the impact of demographic and geographic factors on the U.S. electoral landscape. A geographic visualization further emphasizes the anticipated state-by-state outcomes and reveals enduring regional voting patterns.

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 critical 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.

5.3 Polling and Electoral Uncertainty

The model's reliance on historical polling data and state-level predictions reflects certain limitations inherent in election forecasting. For example, while the model captures general trends, it does not account for potential turnout variations, third-party candidates, or unexpected political events that might affect voter behavior. Furthermore, the use of polling data presents a challenge due to its variability; certain states with less frequent polling data show broader confidence intervals, indicating greater uncertainty in those regions.

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. Future research may also consider factors like demographic shifts and economic conditions, which could significantly influence voter preferences over time.

5.4 Implications for Future Electoral Modeling

The findings from this analysis underscore the need for a nuanced approach to election forecasting. 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 battleground states, could refine the model's predictive power and provide a more comprehensive understanding of the electoral landscape.

For future iterations, we recommend expanding the model to include demographic data, such as age, education level, and income, which could reveal deeper insights into voter behavior. 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. Employing these enhancements would improve model reliability, particularly in closely contested states where the margin of error can significantly impact the final prediction.

5.5 Weakness

Despite the model's strengths, it has limitations that requires further exploration. The assumption of linear relationships between predictors may oversimplify complex voter behaviors and social influences. The dataset simplifies the potentioanl influecne behind the US election. Additionally, the reliance on past data assumes that historical trends will persist, which may not hold true in the rapidly shifting political climate of the U.S. To address these issues, future models could benefit from incorporating voter turnout projections and real-time sentiment analysis, potentially drawing data from social media platforms and news sources to gauge public opinion. Moreover, a comparative post-election analysis could offer valuable insights, revealing areas where the model succeeded and where it may need refinement. This iterative approach would help create a more resilient and adaptable electoral forecasting model.

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.

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, while others, who do not use home telephones, are invited 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.

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.

- **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

In cases of non-response, SurveyUSA attempts follow-up calls if interviews are interrupted by answering machines or busy signals. Weighting is applied to adjust for non-response bias, although this doesn't completely eliminate challenges posed by unreachable or unwilling participants.

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 fail to capture the nuances of respondents' political beliefs. These rigid categories could oversimplify 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 goal of this survey methodology is to accurately forecast the outcome of the U.S. presidential election by collecting high-quality, representative data from a diverse set of respondents across the country. With a budget of $100,000, this methodology incorporates sophisticated sampling techniques, robust respondent recruitment strategies, and rigorous data validation protocols. The approach is designed to maximize accuracy, reduce bias, and account for various demographic, geographic, and political factors that influence voting behavior.

B.2 Core Objectives

- Obtain a representative sample of the U.S. electorate.
- Ensure data quality through rigorous validation.
- Leverage statistical modeling and poll aggregation for an accurate prediction.

B.3 Sampling Strategy

The sampling strategy is designed to ensure that the survey reaches a broad, representative section of the voting population. To achieve this, we will use **stratified random sampling** combined with **quota sampling** for key demographics. This ensures that each important subgroup within the population is adequately represented.

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

B.3.2 Sample Size

A total of **10,000 respondents** will be surveyed, providing a margin of error of approximately $\pm 1\%$ at a 95% confidence level. This sample size will allow for detailed subgroup analysis (e.g., by state, demographic group), yielding statistically robust predictions.

B.3.3 Weighting

Post-stratification weights will be applied to adjust for any oversampling or undersampling of specific demographic groups. For example, younger voters or underrepresented minorities will be weighted to reflect their true proportions in the voting population.

B.4 Recruitment Strategy

To maximize respondent diversity and ensure accurate sampling, the survey will employ **multi-channel recruitment**:

- **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 further boost respondent diversity.
- **Incentives**: Each participant will be entered into a lottery with a chance to win a $100 gift card to encourage participation.

B.5 Data Validation and Quality Assurance

Maintaining data integrity and ensuring high-quality responses are critical to the accuracy of the election forecast. Several measures will be put in place 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 suspiciously 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.

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) to strengthen the forecast through a **poll-of-polls** approach.

- **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 will allow for modeling the popular vote and potentially translating 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 insights 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

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

Table 5: Calculate Mean Squared Error (MSE) on Test Data

Metric	Value
Mean Squared Error (MSE)	11.3039329
Mean Absolute Error (MAE)	2.6492988
R-squared	0.5598326

C.3 Model Summary

Table 6: Complete Model Coefficient Summary Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	56.7394083	3.9106412	14.5089784	0.0000000
numeric_grade	1.5526128	1.1565512	1.3424506	0.1797873
sample_size	-0.0001363	0.0002407	-0.5660165	0.5715232
stateArizona	-5.4799358	2.6104787	-2.0992073	0.0360757

	Estimate	Std. Error	t value	Pr(>|t|)
stateCalifornia	-18.1491651	2.7965409	-6.4898621	0.0000000
stateColorado	-13.8071306	3.6086714	-3.8260981	0.0001392
stateConnecticut	-17.9388779	4.4020241	-4.0751431	0.0000500
stateFlorida	-4.2379822	2.8560281	-1.4838727	0.1381916
stateGeorgia	-5.3259895	2.6121595	-2.0389220	0.0417486
stateIdaho	4.2862357	4.4082928	0.9723120	0.3311558
stateIllinois	-16.5167618	4.4084278	-3.7466332	0.0001906
stateIndiana	0.7462074	3.2901264	0.2268020	0.8206290
stateIowa	-5.4426972	3.5992334	-1.5121823	0.1308374
stateKansas	-3.2141730	4.4083108	-0.7291167	0.4661193
stateMaine	-10.1282590	2.7871502	-3.6339121	0.0002949
stateMaryland	-20.5212347	3.0189643	-6.7974420	0.0000000
stateMassachusetts	-23.3727282	2.8206445	-8.2863078	0.0000000
stateMichigan	-7.0980095	2.6067966	-2.7228858	0.0065962
stateMinnesota	-9.9405995	2.7727731	-3.5850750	0.0003550
stateMissouri	-0.9198169	3.1224887	-0.2945781	0.7683839
stateMontana	-0.0108739	2.9574649	-0.0036768	0.9970672
stateNational	-8.2576446	2.5671930	-3.2166046	0.0013433
stateNebraska	-8.5264419	2.9084791	-2.9315809	0.0034575
stateNevada	-6.3984036	2.6242257	-2.4382063	0.0149522
stateNew Hampshire	-8.9611168	2.7378195	-3.2730853	0.0011042
stateNew Jersey	-13.6200257	4.4021016	-3.0939826	0.0020357
stateNew Mexico	-12.1819786	4.4048824	-2.7655627	0.0057984
stateNew York	-15.1328554	2.7165680	-5.5705786	0.0000000
stateNorth Carolina	-5.6070724	2.6130073	-2.1458311	0.0321525
stateOhio	-4.0705020	2.7568831	-1.4764870	0.1401618
statePennsylvania	-7.1452874	2.5998539	-2.7483419	0.0061092
stateRhode Island	-14.7603639	3.1202414	-4.7305198	0.0000026
stateSouth Carolina	-1.8962575	4.4019057	-0.4307810	0.6667304
stateSouth Dakota	3.2309652	3.2896339	0.9821656	0.3262813
stateTennessee	4.2937743	4.4060395	0.9745202	0.3300593
stateTexas	-4.4751542	2.7222211	-1.6439348	0.1005377
stateVermont	-24.9118292	4.4089996	-5.6502226	0.0000000
stateVirginia	-9.8539879	2.6954198	-3.6558268	0.0002712
stateWashington	-16.2868340	3.2858047	-4.9567261	0.0000009
stateWest Virginia	9.0929121	4.4080149	2.0628134	0.0394158
stateWisconsin	-6.8930538	2.6012097	-2.6499417	0.0081913
stateWyoming	16.6030395	4.4055893	3.7686308	0.0001748
transparency_score	-0.7815612	0.1021377	-7.6520359	0.0000000
days_until_election	-0.0118316	0.0011329	-10.4437625	0.0000000
R-squared	0.5110921	NA	NA	NA

	Estimate	Std. Error	t value	Pr(>|t|)
Adjusted R-squared	0.4878108	NA	NA	NA
F-statistic	21.9528769	NA	NA	NA
F-statistic p-value	0.0000000	NA	NA	NA
Residual Std. Error	3.5831229	NA	NA	NA

C.4 Multicollinearity Check on the Training Data

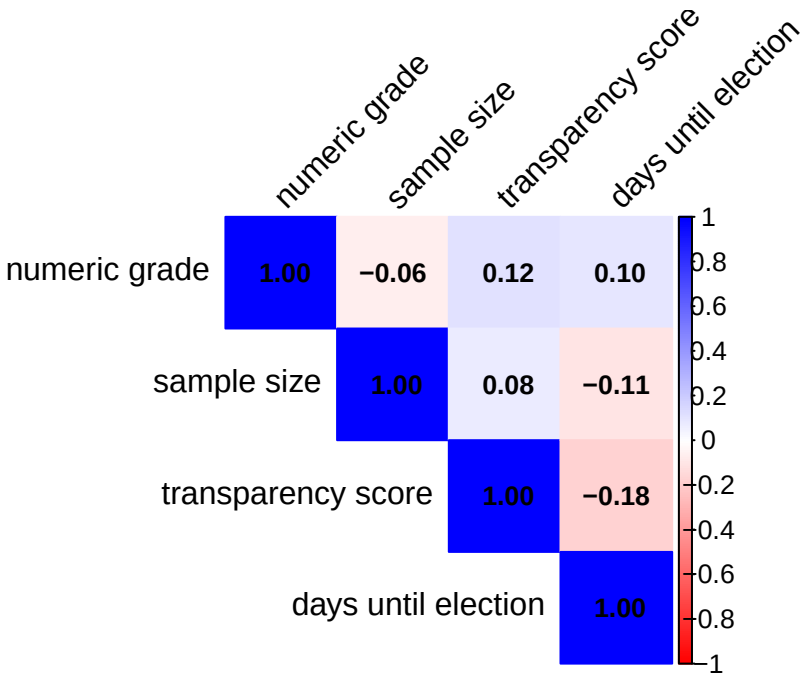

Figure 8: Correlation Matrix of "Numeric Grade", "Sample Size", "Transparency Score"and "Days Until Election"

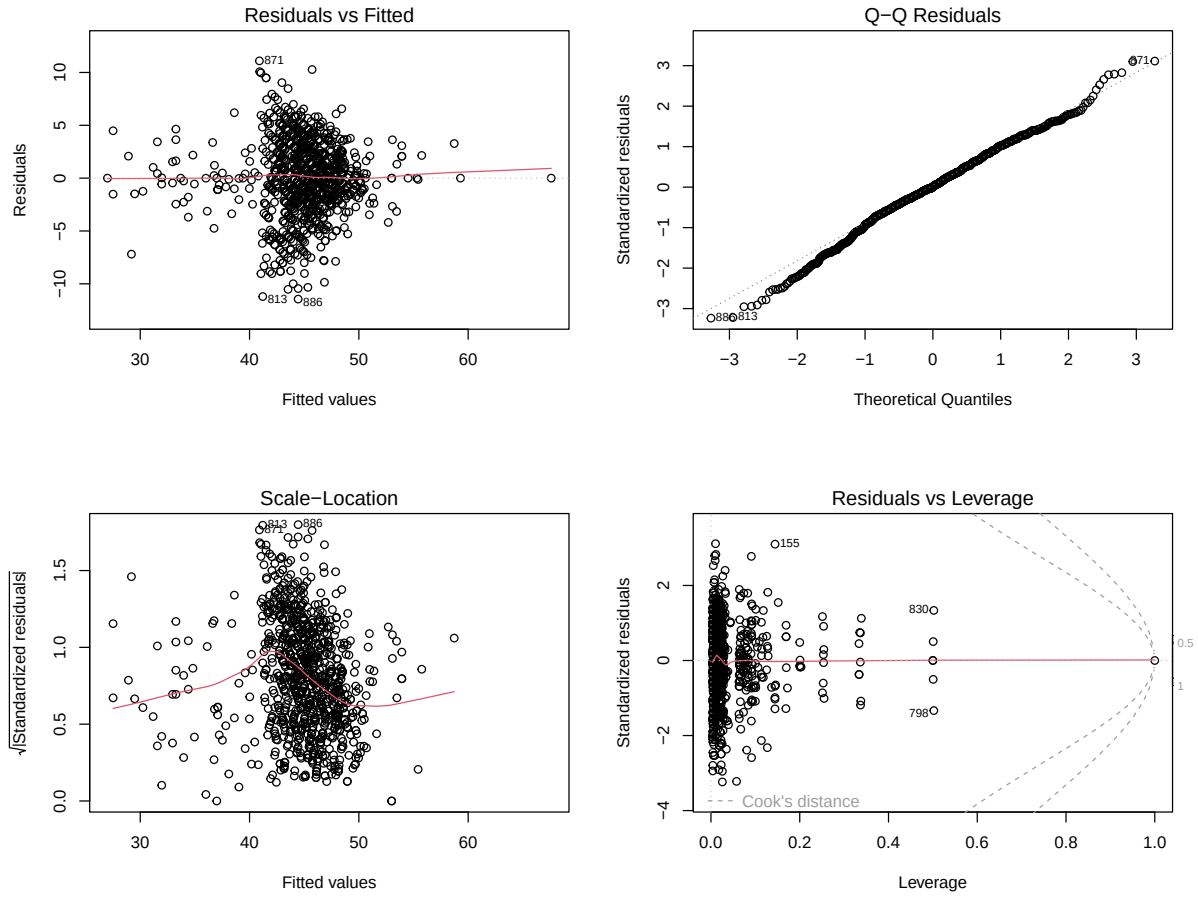
Residuals vs Fitted
Residuals
Fitted values
871
813
886
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
871
886
813
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
155
830
798
Cook's distance

Figure 7

D Acknowledgements

R Core Team (2023) Thanks to Open AI and ChatGPT 4.0 is used to write the analysis of the paper.

FiveThirtyEight (2024) provides the data. Wickham et al. (2019), Wickham (2016), Xie (2023)

References

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

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

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/.