Poll-Based Forecasting of the 2024 U.S. Presidential Election*

Valid Multiple Linear Regression Explains Why Harris Will Win

Haowei Fan Fangning Zhang Shaotong Li

November 3, 2024

This study aggregates polling data to forecast support for 2024 U.S. presidential candidates Kamala Harris and Donald Trump. By applying linear models with adjustments, we identify a modest lead for Harris at 47.6% over Trump's 43.5%. The findings highlight that carefully weighted polling data provides a clearer view of electoral dynamics, showing how rigorous methodology in forecasting can yield more accurate insights into public opinion trends, aiding informed decision-making in political and economic spheres.

Table of Contents

*Code and data are available at: [https://github.com/HaoweiFan0912/US_Election-Forecast/tree/main].

Introduction

The U.S. presidential election has profound implications on a global scale, shaping international relations, economic policy, and key social issues. As the 2024 election approaches, the ability to accurately forecast potential outcomes is essential for policymakers, businesses, and organizations seeking to anticipate shifts in U.S. policy that may affect trade, climate commitments, and strategic alliances (**citeBBC?**). Additionally, recent analyses suggest that the election could introduce volatility into financial markets, emphasizing the importance of reliable predictions for strategic planning (**citeEuronews?**). However, predicting election outcomes presents significant challenges due to a multitude of influencing factors, such as media influence, public sentiment, and socio-political dynamics, which collectively add complexity to election forecasting (**citeOregon?**). Past elections further illustrate that targeted messaging and political events can profoundly impact voter behavior, complicating predictions even further (**citeYale?**).

This study seeks to address these complexities by examining voter support for the 2024 presidential candidates Kamala Harris and Donald Trump. While substantial polling data exists, current aggregation methods often lack consistency and fail to adequately consider critical factors such as poll reliability and sample size, leading to unreliable predictions. To address this gap, the present study adopts a "poll-of-polls" methodology, drawing from multiple data sources at both national and state levels. Through the application of multiple linear regression models, the study integrates essential variables, including pollster reliability, sample size, and polling duration, to produce a forecast that is both stable and transparent.

The estimand in this study represents the expected level of voter support for each primary candidate, Kamala Harris and Donald Trump, based on aggregated polling data across diverse demographics and polling methodologies. This measure aims to capture the central tendency of public opinion, adjusted for polling reliability and sample characteristics, to provide a stable estimate of each candidate's projected support under current conditions. By centering on this estimand, the analysis offers a robust and interpretable forecast applicable for strategic decision-making in both political and economic contexts.

The findings suggest a slight advantage for Harris, with a predicted support level of 47.6% compared to Trump's 43.5%. This marginal lead underscores the importance of methodological rigor, as systematically weighting data by reliability yields more consistent and interpretable predictions. The results indicate that while both candidates retain substantial support, polling methods and demographic representation can subtly shift the support dynamics, providing a deeper understanding of the electoral landscape beyond basic polling figures.

This paper is structured as follows: first, the data collection and filtering processes are outlined, followed by a description of the methodological framework, introducing linear modeling approaches. The results are then presented and analyzed, with a discussion on broader implications. The paper concludes with recommendations for future research and potential applications of these forecasting models in electoral studies.

This project leverages several R packages, including tidyverse(**citeTidyverse?**), janitor(**citeJanitor?**), tidyr(**citeTidyr?**), dplyr(**citeDplyr?**), lubridate(**citeLubridate?**), arrow(**citeArrow?**), patchwork(**citePatchwork?**), car(**citeCar?**), kableExtra(**citeKableExtra?**), gridExtra(**citeGridExtra?**), moments(**citeMoments?**), grid(**citeGrid?**), rstanarm(**citeRstanarm?**), and testthat(**citeTestthat?**), to clean, organize, analyze, and visualize data in forecasting voter support for the 2024 U.S. presidential election. These packages support a reproducible and rigorous approach to data management, statistical modeling, and result presentation, ensuring transparency and accuracy throughout the analysis process. cross-reference

Data

Overview

The dataset comes from FiveThirtyEight's 'Presidential Election Polls (Current Cycle)' ((**citeRawData?**)). FiveThirtyEight is a well-known website recognized for its political, economic, and sports analyses. Its polling aggregation methodology is highly regarded in the field, aiming to provide readers with transparent, scientific, and as accurate as possible predictions. This polling data is compiled from various polling agencies, encompassing a wide range of demographic information, which serves as an essential basis for analyzing public voting preferences in the upcoming presidential election.

In this section, we detail our selected variables, discuss key measurements, outline important limitations of our data, and our data cleaning process.

Raw data

The analysis and visualizations in this paper are based on polling results as of October 22. The dataset includes 52 variables, 17,133 samples and 3530 polls from various polling sources. These variables are shown in the below table(Table 1).

Table 1: Varibales of raw data

poll_id	pollster_id	pollster
sponsor_ids	sponsors	display_name
pollster_rating_id	pollster_rating_name	numeric_grade
pollscore	methodology	transparency_score
state	start_date	end_date
sponsor_candidate_id	sponsor_candidate	sponsor_candidate_party
endorsed_candidate_id	endorsed_candidate_name	endorsed_candidate_party
question_id	sample_size	population
subpopulation	population_full	tracking
created_at	notes	url
url_article	url_topline	url_crosstab
source	internal	partisan
race_id	cycle	office_type
seat_number	seat_name	election_date
stage	nationwide_batch	ranked_choice_reallocated
ranked_choice_round	hypothetical	party
answer	candidate_id	candidate_name
pct		

Cleaning Process

Firstly, there are several variables clearly irrelevant to the project and will not be discussed further: notes, url, url_article, url_topline, url_crosstab, and source.

Additionally, there are some duplicate variables, and we will retain only one of each, ignoring the rest: pollster, sponsors, display_name, pollster_rating_name, sponsor_candidate, endorsed_candidate_name, population_full, candidate_id, and candidate_name.

Constant variables, which cannot impact our predictions, will also be excluded from further discussion. These include: endorsed_candidate_id (NA), endorsed_candidate_party (NA), subpopulation (NA), cycle (2024), office_type (U.S. President), seat_number (0), seat_name (NA), election_date (11/5/24), stage (general), and nationwide_batch (FALSE).

After cleaning, 27 out of the 52 variables remained potentially relevant to our research. We selected 10 variables (Table 2) of interest and plotted their distributions in the data visualization section , while the plots of the remaining variables are included in the appendix .

Table 2: Important variables and their descriptions

Variable	Description
poll_id	Unique identifier for each poll conducted.
methodology	The method used to conduct the poll (e.g., Online Panel).
population	The abbreviated description of the respondent group, typically indicating their voting status (e.g., 'lv' for likely voters).
ranked_choice_reallocated	Indicates if ranked-choice voting reallocations have been applied in the results.
hypothetical	Indicates whether the poll is about a hypothetical match-up.
answer	The response or answer choice given in the poll (e.g., the candidate's party).
numeric_grade	A numeric rating given to the pollster to indicate their quality or reliability (e.g., 3.0).
pollscore	A numeric value representing the score or reliability of the pollster in question (e.g., -1.1).
transparency_score	A score reflecting the pollster's transparency about their methodology (e.g., 9.0).
sample_size	The total number of respondents participating in the poll (e.g., 2712).
pct	The percentage of the vote or support that the candidate received in the poll (e.g., 51.0 for Kamala Harris).

After finalizing the variables, we first created a new variable named 'duration', which replaced 'start_date' and 'end_date'. This new variable represents the number of days between 'start_date' and 'end_date'.Next, we categorized the 51 different methodologies into four levels, ranging from the least reliable and accurate (level_1) to the most reliable (level_4).

Subsequently, we handled the missing values by imputing numerical variables with their mean values and categorical variables with their mode. Since our results are not exact percentages, we used 'score' to name what would typically be called 'pct'. We then finalize and tidy up the variable names.

Then, the data is extracted for each candidate individually. We calculated a weighted score by weighting according to the number of times each candidate was mentioned in the polls. After comparison, we observed that the top three candidates—Trump, Harris, and Biden—had significantly higher scores than the remaining candidates. Given that Biden has withdrawn from the race, we are now focusing only on the datasets for Trump and Harris for further analysis.

We also split the data for Trump and Harris into a training set (70%) and a test set (30%). These four datasets form our analysis data. Below is a portion of the Trump training set for reference:

Data Visualization

The left bar chart in (Figure 1) shows the distribution of the `ranked_choice_reallocated` variable. The chart indicates that the majority of the data points are marked as `FALSE`, meaning ranked-choice voting reallocations have not been applied in most cases. Only a very small number of instances are marked as `TRUE`.

The right bar chart in (Figure 1) illustrates the distribution of the `hypothetical` variable. It shows that a larger proportion of the data is marked as `TRUE`, indicating that the poll results are often based on hypothetical match-ups. There are fewer instances marked as `FALSE`, where the poll is not hypothetical.

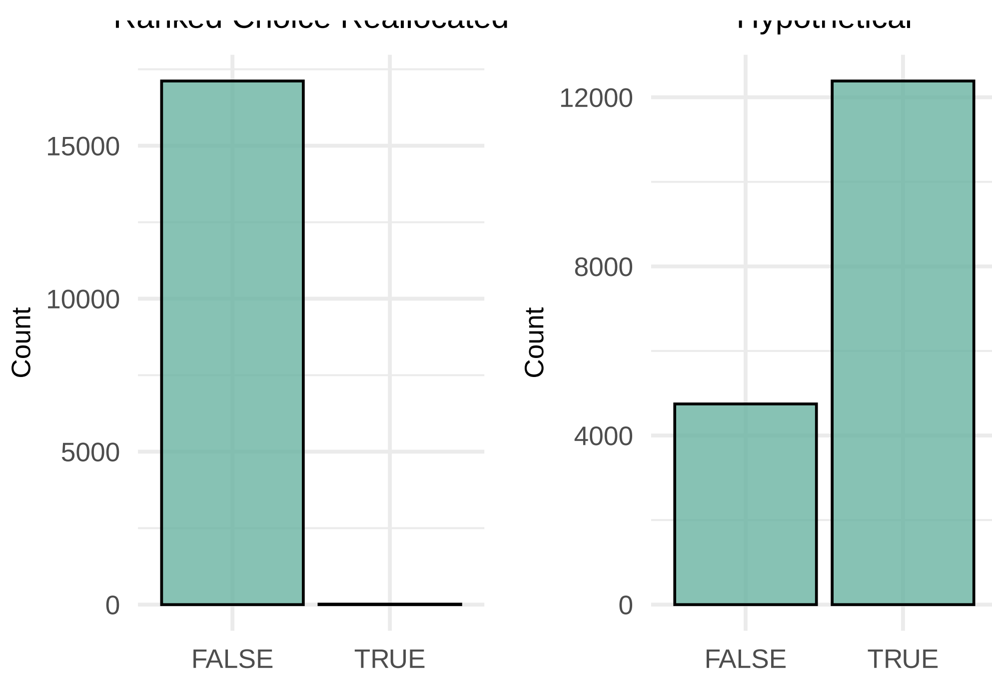


Figure 1: Boolean variables

The three charts depict the distribution of key variables in the dataset: `methodology`, `population`, and `answer`. In the most left chart of (Figure 2), we see that the most frequently used polling methodology is “Online Panel,” followed by “Live Phone” and “Live Phone/Text-to-Web.” The “Online Panel” category significantly outnumbers the others, while the “Other” category also includes a notable count, representing various methodologies grouped together.

The middle chart of (Figure 2) shows the distribution of different respondent groups. “lv” (likely voters) and “rv” (registered voters) dominate, with “rv” showing a slightly higher count, indicating that these two groups make up the majority of the sample. Other groups, such as “a” (all adults), “v,” and those with missing values (“NA”), represent much smaller portions of the respondent pool.

The most right chart of (Figure 2), `Top 3 Answer and Others`, illustrates the responses given in the polls. “Biden” and “Trump” have similar counts, with “Trump” being slightly higher, while the “Other” category also shows a significant proportion. The response for “Harris” is noticeably lower compared to the others. Overall, these charts provide a visual representation of the polling data, highlighting the dominant methodologies, respondent groups, and response preferences in the dataset.

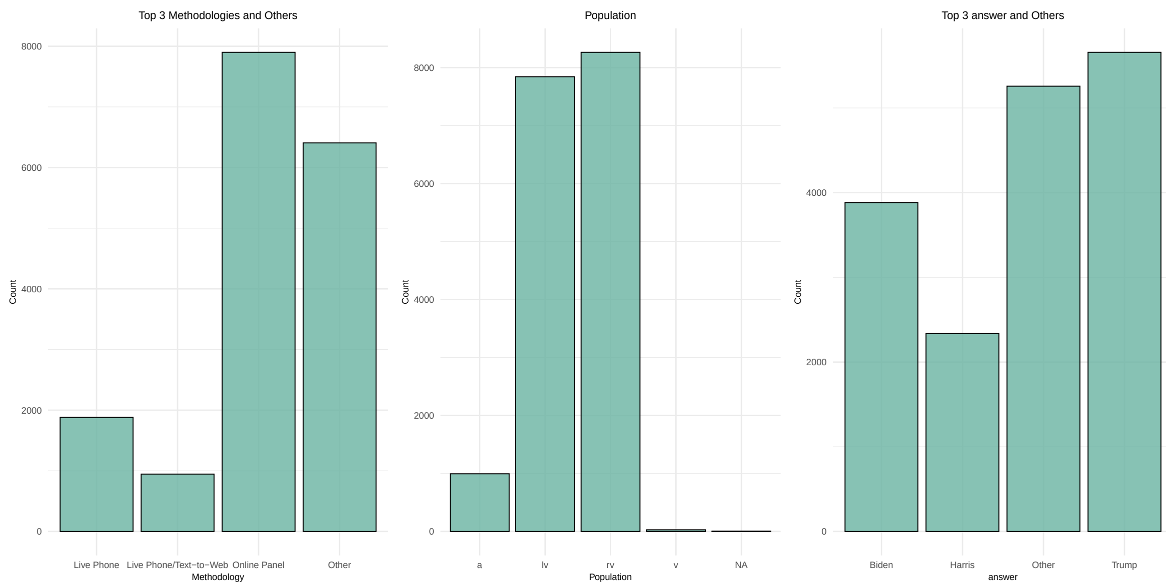


Figure 2: Catogorical variables

The `numeric_grade` histogram in (Figure 3) shows the distribution of a numeric rating assigned to pollsters, representing their overall quality or reliability. The data appear to cluster heavily around scores of 2 and 3, suggesting that a large number of pollsters fall within these quality ranges. The distribution is fairly symmetric, with a noticeable concentration at these higher scores, indicating that most pollsters are considered to be of moderate to good quality.

The `pollscore` histogram in (Figure 3) indicates the reliability of each pollster, with lower (more negative) values being better. The distribution shows a peak around zero, with a significant number of pollsters having scores close to zero or slightly negative, and most values being negative, indicating relatively high reliability overall. This implies that most pollsters have moderate to high reliability, with fewer pollsters achieving highly negative scores, which indicate better performance. The tail towards positive values suggests that a small subset of pollsters may have issues with reliability.

The distribution of `transparency_score` in (Figure 3) shows a wide spread, with notable peaks at several discrete points, but no clear pattern overall. Higher scores, such as 7.5 and 10, have high frequencies, indicating that some pollsters tend to achieve relatively high transparency. On the other hand, lower scores, such as 2.5 and 5, also show some clustering, but with less consistency.

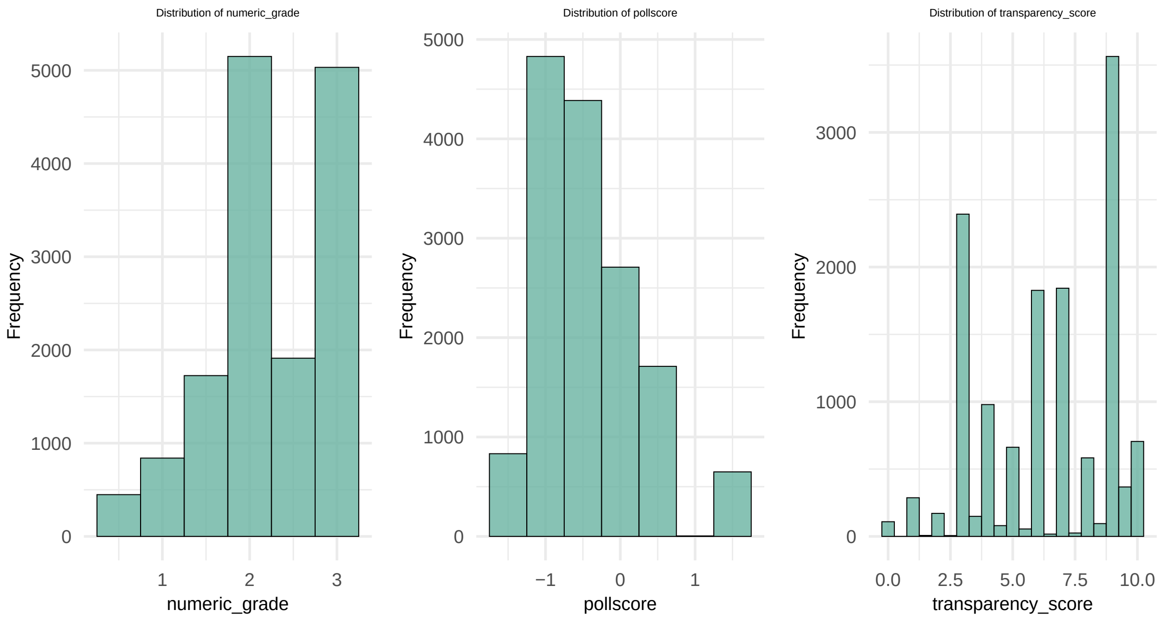


Figure 3: Distribution of numerical varibales part 1

The `Distribution of sample_size` histogram in (Figure 4) illustrates the frequency of poll sample sizes. The majority of polls have relatively small sample sizes, with the frequency decreasing sharply as the sample size increases. The distribution appears highly right-skewed, suggesting that larger sample sizes are much less common than smaller ones.

The `Distribution of pct` histogram in (Figure 4) represents the frequency distribution of vote percentages received by candidates in various polls. Most polls have percentage values concentrated around the 30-40% range, with visible peaks at around 0% and 40%. The distribution shows some variation across a wide range but seems to have a higher frequency in the middle range (30-40%) compared to the lower and higher extremes.

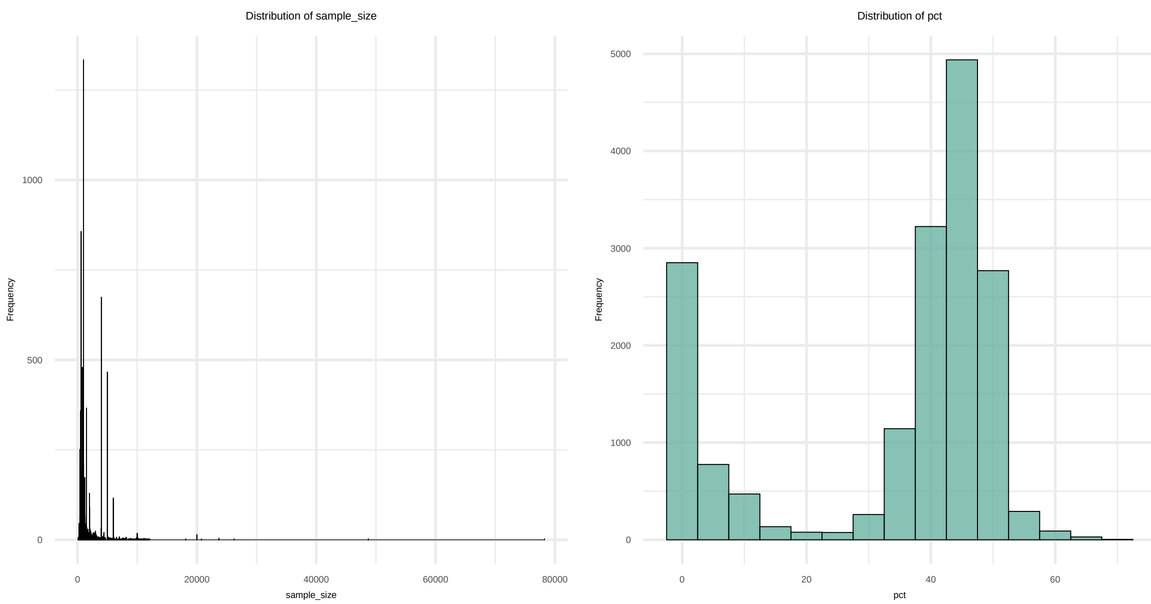


Figure 4: Distribution of numerical varibales part 2

appendix Both charts in (Figure 5) show the normal distribution of poll start and end dates over time. The left chart represents the frequency of polls by their start date, while the right chart represents the frequency of polls by their end date, with both distributions primarily concentrated around 2010.The frequency drops after 2010 in both charts.

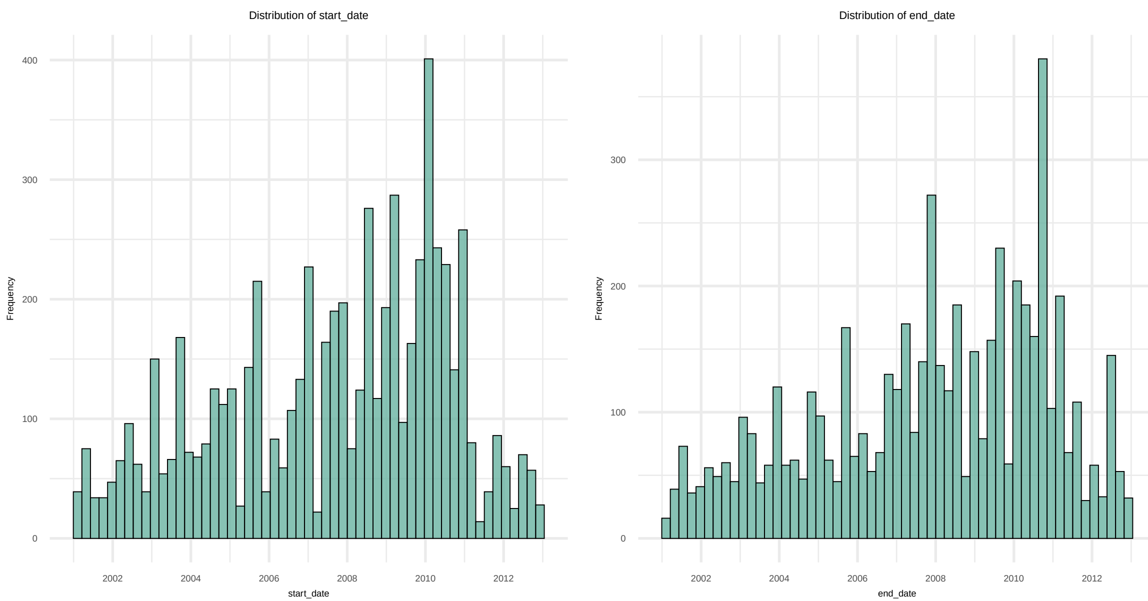


Figure 5: Distribution of date varibales

Measurement and Limitations

The method used to forecast the presidential election results is the poll-of-polls, which aggregates results from multiple polls instead of relying on a single survey, aiming to make the results more accurate and stable. In this method, each poll is assigned a weight based on factors such as sample size, recency, and the pollster's historical accuracy.

The dataset used for this prediction is from FiveThirtyEight, which includes scientifically sound public polls that meet methodological standards. Polling organizations are rated based on accuracy, transparency, and sample quality, represented by a numeric_grade (ranging from 0.5 to 3.0). Higher scores indicate greater reliability. The histogram of numeric_grade values shows a concentration around scores of 2 and 3, suggesting most pollsters are of moderate to good quality.

Polling organizations use different survey methods but follow similar principles. They select representative samples, publish surveys through chosen platforms, and aim to ask clear, unbiased questions. YouGov, discussed in the appendix, is one such example.

Survey data accuracy is limited by several factors. Sampling bias can lead to an unrepresentative sample, underrepresenting certain demographics. Response bias may occur if participants are not truthful or are influenced by question phrasing. Platform differences also impact reliability, as social media polls may attract different audiences compared to phone or in-person surveys. Pollscore and numeric grade filters help ensure quality, but they are based on historical data and may not reflect current survey quality. Additionally, the rapidly changing political narrative and voter sentiment during campaigns can affect polling accuracy. These factors contribute to inaccuracies in survey results, affecting the reliability of aggregated data.

Similar dataset

A dataset similar to ours titled 2024 National Polls ((**citeNYTData?**)) for the U.S. Presidential Election is found. It was Published by The New York Times, this dataset aggregates survey results from multiple polling organizations, focusing on the support levels for major presidential candidates and aiming to reflect voters' preferences and election trends. However, compared to our dataset, this one has fewer variables, which might reduce its predictive accuracy.

Model

Overview

In the modeling process, we used the `arrow` package to handle parquet files. The model parameters were examined using the `car` and `moments` packages. Additionally, `tidyverse`, `knitr`, `dplyr`, `patchwork`, `kableExtra`, `gridExtra`, and `grid` were employed to visualize and evaluate the model's performance and assumptions.

We developed two models to predict the final competitiveness of Trump and Harris in the November 5, 2024, U.S. presidential election. Both models were trained using a training set (70%) for each candidate, while the remaining 30% served as the test set.

The model formula is as follows:

$$\begin{aligned} Score_{Trump} = &\beta_1 Pollscore + \beta_2 Transparency\_score + \beta_3 Duration + \\ &\beta_4 Sample\_size + \beta_5 Population + \beta_6 Hypothetical + \beta_0 \end{aligned} \tag{1}$$

$$\begin{aligned} Score_{Harris} = &\alpha_1 Pollscore + \alpha_2 Transparency\_score + \alpha_3 Duration + \\ &\alpha_4 Sample\_size + \alpha_5 Population + \alpha_6 Hypothetical + \alpha_0 \end{aligned} \tag{2}$$

The estimands, Score_Trump and Score_Harris, represent the competitiveness of each candidate. A higher score indicates stronger competitiveness. If the predicted score for one candidate is higher than the other, we consider that candidate to be the likely winner of the election.

"pollscore", "transparency_score", "duration", "sample_size", "population", and "hypothetical" are our estimators. "duration" represents the length of a poll in days. Detailed descriptions of the other estimators are provided in the data section.

Notably, we used a Multiple Linear Regression (MLR) model, which implies the following assumptions:

1. **Linear Relationship**: A linear relationship exists between the estimand and the estimators.
2. **Multivariate Normality**: The residuals (differences between observed and predicted values) are normally distributed.
3. **No Multicollinearity**: The correlations between independent variables are not too high.
4. **Homoscedasticity**: The variance of residuals remains consistent across all values of the independent variables.

Estimand and Estimators

Our estimand, “score,” represents the support rate of a candidate in a particular poll, corresponding to the “pct” in the raw data. However, due to our methodology, the final result cannot be expressed as a proportion, and thus we named it “score.” The figures below show the distribution of scores for Trump and Harris in their respective training sets. It can be observed that both distributions are approximately normal. Different form raw…….

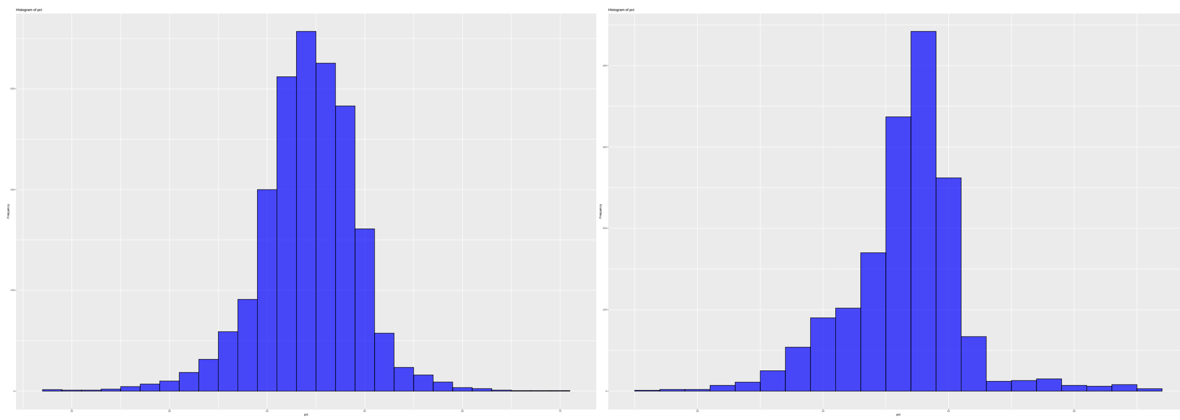


Figure 6: Distribution of scores in training dataset

Since our training set was obtained via simple random sampling, the distributions of “pollscore”, “transparency_score”, “sample_size”, “population”, and “hypothetical” are similar to those in the raw data, which will not be further elaborated here. The variable “duration” is derived from the difference between “start_date” and “end_date” in the original data. This means that our model cannot account for time series effects, but we simplified this to meet the assumptions of using MLR given the linear relationship between these two factors. Below are histograms that display the duration of polls nominating Trump and Harris. It can be observed that they follow a highly skewed distribution, with most values clustered around 1-2 days.

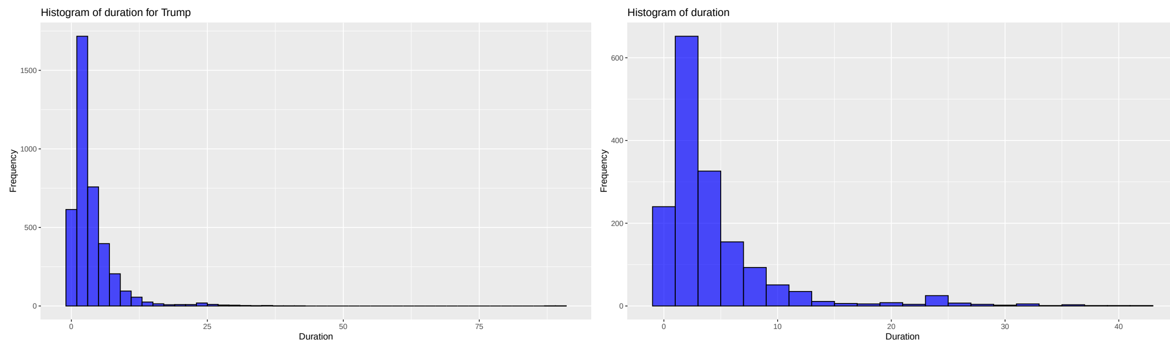


Figure 7: Distribution of durations in training dataset

It is worth noting that although our final model does not consider “methodology,” in the initial model exploration phase, we simplified the 51 different categories of the “methodology” variable into four levels to maintain simplicity. Level 1 represents the lowest reliability, while level 4 represents the highest. The specific classifications are as follows:

Methodologies were scored based on several criteria, including statistical rigor, representativeness, response rate, interaction quality, and cost efficiency.

- **High-scoring methodologies**: Methods such as Probability Panels or Live Phone surveys with broad coverage and low refusal rates received high scores due to strong representativeness and reliability.
- **Medium-high scoring methodologies**: Online Panels and Text-to-Web methods were rated in this category. Online Panels are cost-effective but susceptible to self-selection bias, whereas Text-to-Web improves response rates but may lack representativeness depending on the target demographics.
- **Medium-scoring methodologies**: App Panels and IVR (Interactive Voice Response) tend to lack broad representativeness or interaction quality, making them suitable only for niche audiences.
- **Low-scoring methodologies**: Email Surveys and methods relying on Online Ads often have low response rates and significant selection bias, which undermines their reliability.

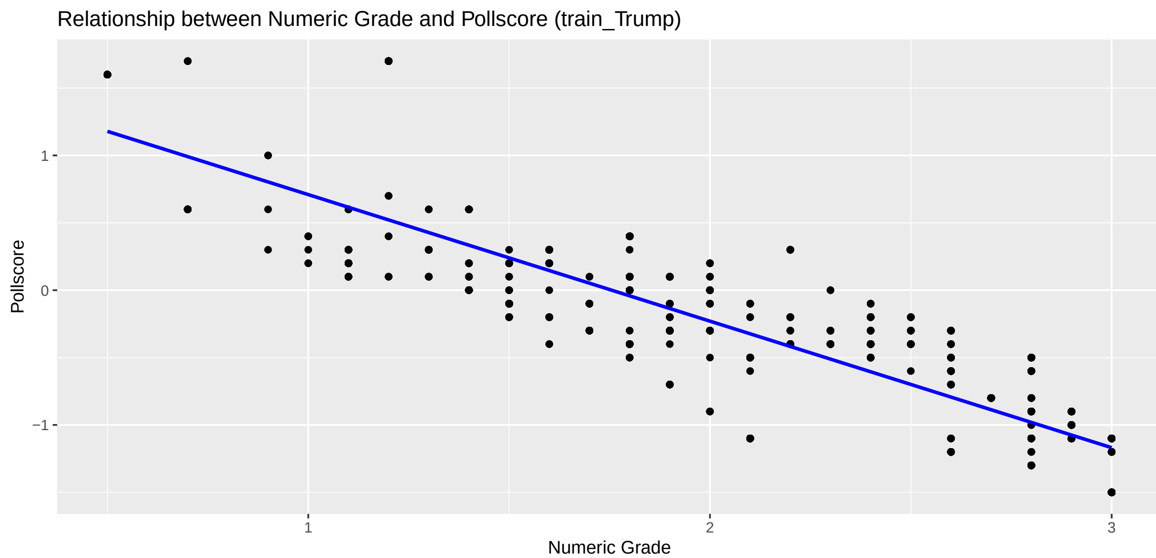
Alternative Models

Below are our initial models. Their estimators included all variables from the analysis datasets.

$$
\begin{aligned}
Score_{Trump} = &\beta_1 Pollscore + \beta_2 Transparency\_score + \beta_3 Duration + \\
&\beta_4 Sample\_size + \beta_5 Population + \beta_6 Hypothetical + \beta_0
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
Score_{Harris} = &\alpha_1 Pollscore + \alpha_2 Transparency\_score + \alpha_3 Duration + \\
&\alpha_4 Sample\_size + \alpha_5 Population + \alpha_6 Hypothetical + \alpha_0
\end{aligned} \tag{4}
$$

Upon comparing the relationship between "numeric_grade" and "pollscore," we observed a significant linear relationship, as shown in the following figure. Therefore, we removed "numeric_grade" from the model to satisfy MLR assumptions, resulting in our second model.



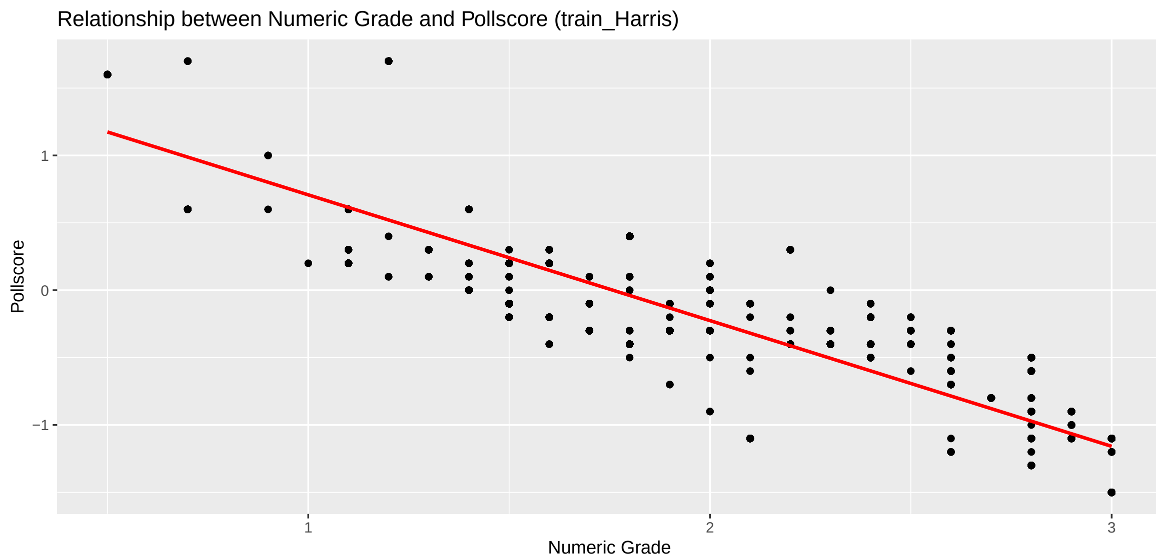


Figure 8: Relationship between numeric_grade and pollscore

The second model is as follows:

$$Score_{Trump} = \beta_1 Pollscore + \beta_2 Transparency\_score + \beta_3 Duration + \beta_4 Sample\_size + \beta_5 Population + \beta_6 Hypothetical + \beta_0 \quad (5)$$

$$Score_{Harris} = \alpha_1 Pollscore + \alpha_2 Transparency\_score + \alpha_3 Duration + \alpha_4 Sample\_size + \alpha_5 Population + \alpha_6 Hypothetical + \alpha_0 \quad (6)$$

We determined the significance of each predictor in the model by checking if the p-value was less than 0.5. As shown in the figure below, both "methodology" and "ranked_choice_reallocated" were found to be insignificant predictors in both models. Thus, they were excluded to reduce model complexity.

Table 3: Significant level of variblaes in the aboanded Harris's model

Table 3: Model for Harris

Variable	P-value
(Intercept)	0.000000e+00
pollscore	1.571566e-17
transparency_score	1.374640e-02
duration	1.782720e-15
sample_size	2.782493e-03
populationlv	1.861643e-09
populationrv	2.207540e-02
populationv	1.123945e-06
hypotheticalTRUE	3.610001e-09
ranked_choice_reallocatedTRUE	6.050213e-01
methodologylevel2	3.326043e-01
methodologylevel3	7.685775e-01
methodologylevel4	1.001406e-01

Table 4: Significant level of variblaes in the aboanded Trump's model

Table 4: Model for Trump

Variable	P-value
(Intercept)	0.000000e+00
pollscore	1.923424e-09
transparency_score	2.694363e-03
duration	3.164509e-04

Variable	P-value
sample_size	3.889702e-07
populationlv	5.329369e-45
populationrv	2.162684e-30
populationv	6.301089e-01
hypotheticalTRUE	1.969949e-77
ranked_choice_reallocatedTRUE	3.545196e-01
methodologylevel2	1.460160e-01
methodologylevel3	1.095258e-01
methodologylevel4	5.411974e-04

Our final model is as follows:

$$\begin{aligned} Score_{Trump} = &\beta_1 Pollscore + \beta_2 Transparency\_score + \beta_3 Duration + \\ &\beta_4 Sample\_size + \beta_5 Population + \beta_6 Hypothetical + \beta_0 \end{aligned} \tag{7}$$

$$\begin{aligned} Score_{Harris} = &\alpha_1 Pollscore + \alpha_2 Transparency\_score + \alpha_3 Duration + \\ &\alpha_4 Sample\_size + \alpha_5 Population + \alpha_6 Hypothetical + \alpha_0 \end{aligned} \tag{8}$$

Validation

First, we verified the assumptions mentioned in the overview. The table below shows the General Variance Inflation Factor (GVIF) for both models, which indicates that all predictors have a GVIF less than 1.3, suggesting no significant multicollinearity.

Table 5: VIF of Harris's final model

Table 5: VIF for Harris Model

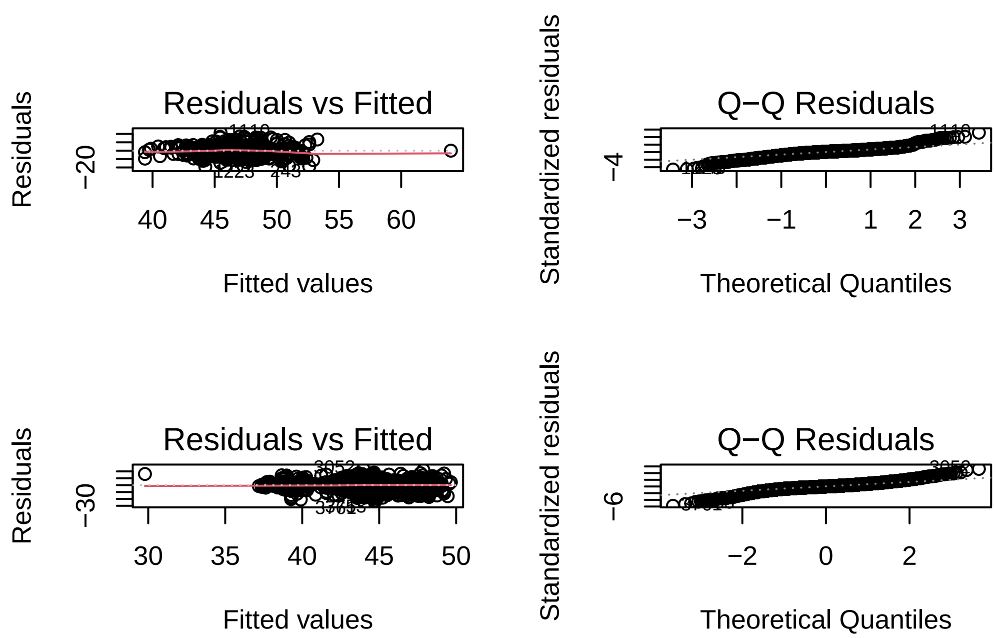
	GVIF
pollscore	1.102432
transparency_score	1.205052
duration	1.087927
sample_size	1.090086
population	1.111258
hypothetical	1.028432

Table 6: VIF of Trump's final model

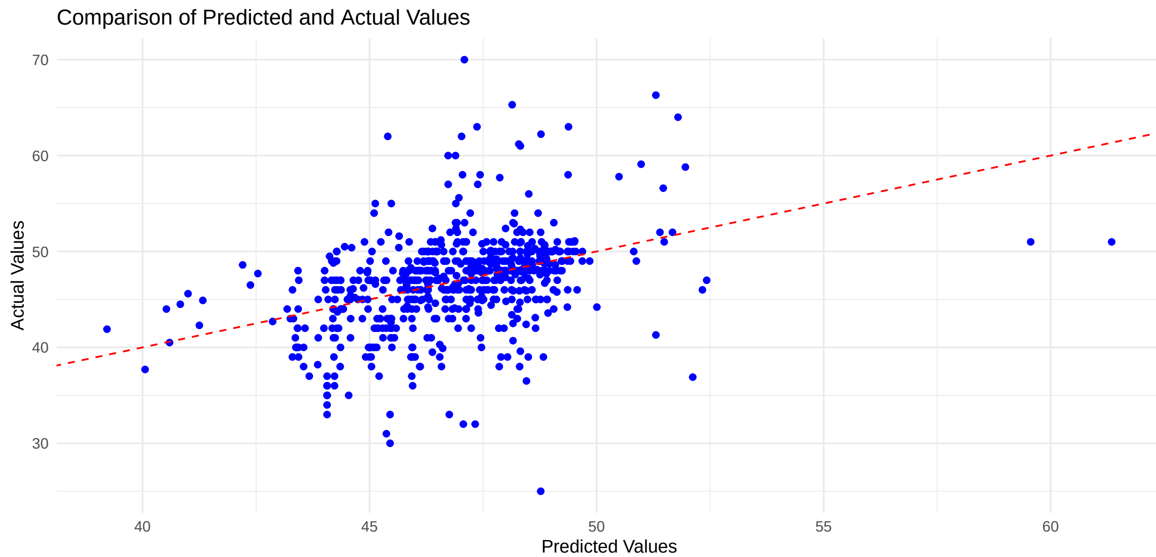
Table 6: VIF for Harris Model

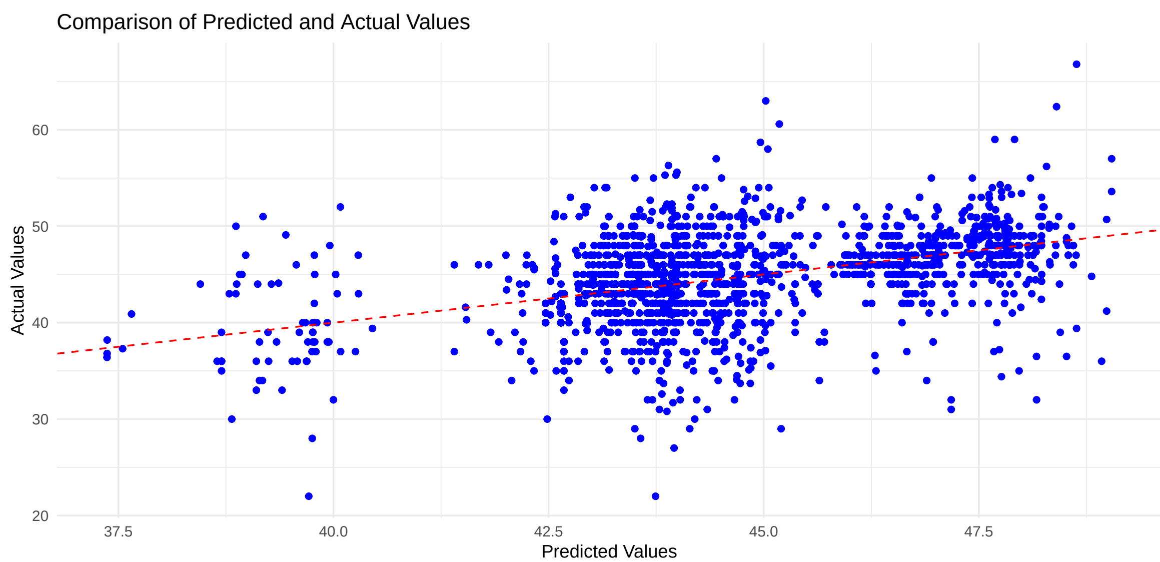
	GVIF
pollscore	1.102432
transparency_score	1.205052
duration	1.087927
sample_size	1.090086
population	1.111258
hypothetical	1.028432

The following figure presents the diagnostic plots for the Harris model. The plot on the left shows the relationship between residuals and predicted values, and since no obvious pattern is observed, our model satisfies the Homoscedasticity assumption. The right-hand plot is a Q-Q plot, and the points align along the diagonal, indicating that residuals are normally distributed and satisfy the Multivariate Normality condition.



The next figure compares predicted and actual values from the test data. For both the Trump and Harris models, the trend between predicted and actual values is roughly linear, indicating that our models are effective.





Model Discussion

As noted above, our model combines the start and end dates of a poll into the "duration" variable, which prevents it from effectively capturing the time series impact on the estimand. Additionally, due to the weak linear relationship between the estimators and the estimand, the explanatory power of our model, as reflected by the adjusted R-Square, is relatively low. Polynomial Regression or Generalized Linear Models might be better choices in this context. Moreover, when significant relationships exist between estimators, our model might fail.

Result

featured values used in prediction

In our analysis, we designated specific poll-related features as "featured values" to serve as representative indicators within each candidate's dataset. These featured values were chosen to highlight the most impactful aspects of the polling data that consistently influenced the support scores for Donald Trump and Kamala Harris. By selecting these representative values, we aimed to streamline the analysis and focus on the factors that most strongly characterized each candidate's data.

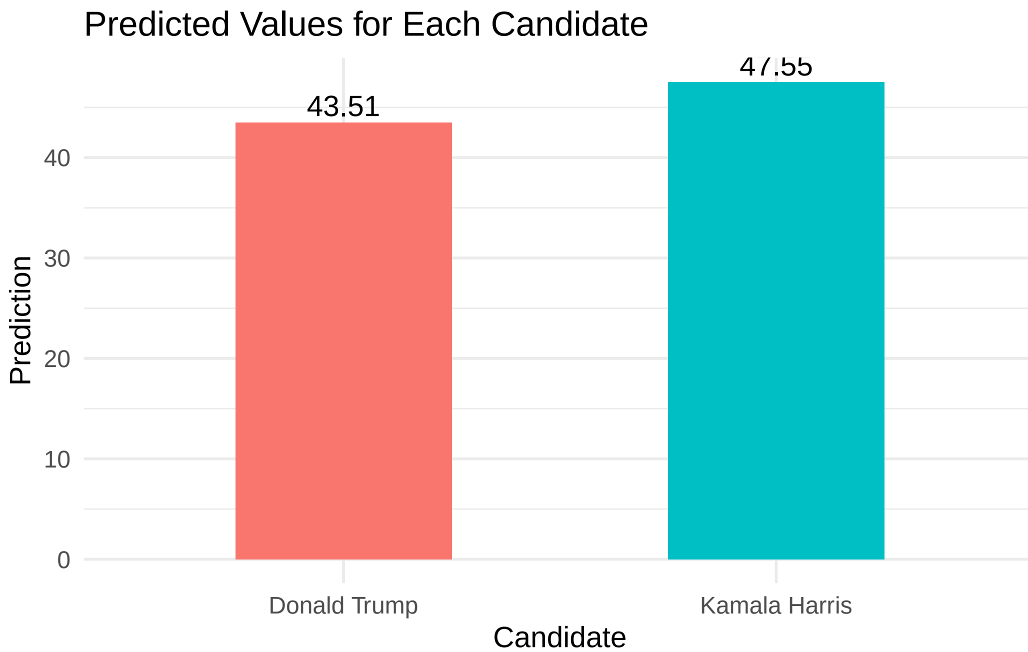
We selected representative feature values for each candidate's dataset by processing each variable based on its type. For numeric variables (e.g., numeric_grade, pollscore, transparency_score, duration, sample_size), we determined whether to use the mean or median by evaluating skewness; variables with low skewness used the mean, while more skewed variables used the median to represent typical values. Categorical variables (e.g., methodology, population) were represented by the most frequent category, while Boolean variables (e.g., ranked_choice_reallocated, hypothetical) were set to TRUE or FALSE based on the most common value. This approach allowed us to capture the key characteristics of each candidate's data in a summarized form.

Table 7: Datas for final prediction

variables	trump	harris
numeric_grade	2.17	2.19
pollscore	-0.4	-0.39
methodology	level3	level3
transparency_score	6.19	6.36
sample_size	1014	1000
population	rv	lv
ranked_choice_reallocated	FALSE	FALSE
hypothetical	TRUE	FALSE
score	44.62	46.88
duration	3	3

Prediction result of the linear model

Using the selected feature values for each candidate, we applied our trained predictive models to estimate the support levels for Kamala Harris and Donald Trump. The bar plot above illustrates the predicted values derived from our analysis. According to the model, Kamala Harris has a predicted support value of approximately 47.65, while Donald Trump is predicted to receive a support value of around 43.51. These predictions suggest an advantage for Kamala Harris over Donald Trump in terms of expected support within the context of the data used.



Conclusion

Overall, our approach demonstrates how feature engineering and predictive modeling can offer insights into candidate support based on the available data. However, it is important to interpret these results cautiously, as they rely on specific variables and assumptions embedded within the dataset. In this analysis, we aggregated representative feature values for each candidate by using the mean or median for numeric variables, the mode for categorical variables, and the most frequent occurrence for Boolean values. Further refinement and additional data could enhance the robustness of these predictions, contributing to a more comprehensive forecast in future studies.

Discussion

This paper presents a predictive analysis of voter support for 2024 U.S. presidential candidates Kamala Harris and Donald Trump. Using a "poll-of-polls" approach, it integrates polling data from diverse sources, emphasizing reliability, sample size, and methodology. Linear and Bayesian models were applied to improve prediction accuracy.

The analysis provides two key insights. First, our model predicts a slight advantage for Kamala Harris, with a support estimate of 47.6% compared to Donald Trump's 43.5%, indicating a competitive but narrow lead. Second, it underscores the impact of poll reliability on predictive accuracy. Traditional methods like live phone surveys show greater accuracy than online methods, reinforcing the importance of reliability-weighted data in improving forecast precision.

This analysis is limited by its use of static variables, which may not capture shifts in public opinion. For instance, major events like debates can influence voter sentiment, yet the model lacks real-time responsiveness. Additionally, aggregating data from different poll methodologies may introduce bias; for example, online panels may skew younger while live phone surveys often favor older demographics. Future work could include real-time data and refined weighting to enhance adaptability and consistency.

Future research could incorporate time series analysis to track evolving voter sentiment and add demographic or economic factors for greater precision. Integrating social media sentiment and machine learning could provide deeper insights into regional voting patterns.

Appendix

Analysis of YouGov Pollster Methodology

In this appendix, we provide a deep-dive analysis of the methodology employed by YouGov, one of the pollsters included in our sample. YouGov is an international online research data and analytics technology group. It is a leading platform for online survey, which has a continuously growing dataset of over 27 million registered members. This pollster has a 3.0 grade according to FiveThirtyEight, which is the highest score. This analysis covers key aspects of YouGov's survey methodology, highlighting its strengths, weaknesses, and the unique features of its approach.

Population, Frame, and Sample

YouGov utilizes an online panel to collect survey responses, with participants drawn from a broad population base, which typically comprises all U.S. adults citizens. Respondents are chosen based on a non-probability sampling, which means not everyone in the population has an equal chance of being selected. However, the sample is adjusted using statistical weighting to better represent the target population. The sampling frame consists of individuals who have signed up to participate in surveys, representing a range of demographic characteristics. However, as an online panel, there may be limitations regarding coverage bias, particularly for individuals with limited internet access.

Sample Recruitment

YouGov recruits participants through online advertisements and other digital marketing techniques, with surveys offers surveys in multiple languages. The recruitment process is designed to ensure that the panel is as representative as possible. For instance, YouGov collects information such as email addresses and IP addresses when new members join the panel. Additionally, YouGov monitors survey completion time and answer consistency to ensure the data is accurate. Respondents who fail quality checks are removed.

Sampling Approach and Trade-offs

YouGov employs a form of quota sampling combined with weighting adjustments to make the sample representative of the target population. To ensure representativeness, YouGov selects respondents based on key demographic characteristics such as age, gender, race, education, and voting behavior. These characteristics are used to set quotas, and the sample is adjusted with statistical weighting to align with the distribution of these characteristics in the target population. For example, if a particular demographic group is underrepresented in the sample, their responses are given greater weight to correct the imbalance. One trade-off of this method is that, although it helps improve representativeness, it may not fully eliminate selection bias due to the reliance on an online panel, which can lead to overrepresentation or underrepresentation of certain groups. Additionally, the process of weighting adjustments may introduce additional errors if the weights are inaccurate or if certain groups are given disproportionately high weights, leading to increased variability and potential bias in the final results.

Handling Non-response

Non-response is managed by using statistical weighting to adjust the sample to more closely reflect the demographic makeup of the target population. While this helps mitigate some of the biases associated with non-response, it cannot fully account for differences between respondents and non-respondents, especially when non-response is correlated with key survey variables.

Strengths and Weaknesses of the Questionnaire

The YouGov questionnaire is well-designed to capture a wide range of attitudes and behaviors. The use of standardized questions ensures consistency across surveys, allowing for longitudinal analysis. However, as an online survey, there is the risk of respondents providing socially desirable answers or rushing through the survey without providing thoughtful responses. Additionally, the format may limit the depth of responses compared to in-person interviews.

Overall, YouGov's methodology provides a cost-effective and timely approach to data collection, particularly useful for understanding trends across large populations. However, the use of an online panel introduces certain limitations that must be acknowledged when interpreting the results.

Ideal Methodology and Survey for Predicting the U.S. Presidential Election

Budget Overview

With a budget of $100,000, the goal is to design an efficient and representative method for predicting the U.S. presidential election. This methodology will include sampling strategies, respondent recruitment, data validation, poll aggregation, and survey implementation details.

Sampling Methodology

A stratified sampling approach will be used to ensure diversity and representation. The population will be divided into relevant strata such as age, gender, geographic region, race, and political affiliation. This approach ensures that each subgroup is adequately represented, thereby reducing sampling bias.

Respondent Recruitment

Respondents will be recruited through online panels. Partnerships with established survey platforms and third-party providers will help reach a broad and representative group of participants, such as through platforms like Instagram, YouTube, and various news websites. Small monetary compensation or gift cards will be offered as incentives to encourage participation. Additional incentives will be provided to underrepresented groups, such as individuals with lower educational attainment or residents of rural areas, to ensure more inclusive recruitment. The aim is for a sample size of approximately 10,000 respondents, which would achieve a margin of error of $\pm 1\%$ at a 95% confidence level.

Data Validation

Data validation will involve cross-referencing respondent demographic information with census data to confirm representativeness. Additionally, responses will be reviewed for accuracy, and suspicious or incomplete answers will be flagged for further inspection. Responses completed too quickly or that include repeated answers such as "prefer not to say" or "other" will be discarded. IP addresses will be tracked to prevent duplicate submissions.

Poll Aggregation and Methodology Features

Once all responses are collected, weights will be applied according to electoral demographics and voter turnout to ensure the sample represents the U.S. population. Poll aggregation will also involve adjustments for known biases, such as overreporting in certain demographic groups or historical voting trends. Bayesian updating will be used to refine predictions continuously as more data becomes available.

Survey Implementation

The survey will be implemented using Google Forms, which allows for easy distribution and data collection. The survey will include questions related to voter preferences, key issues, and demographic information. Questions will be designed to minimize leading language and provide a range of response options to avoid bias. Keeping the survey brief (approximately 5 minutes with 12 questions) will help maintain respondent focus.

Budget Allocation

- $60K for Recruitment Costs and Survey Platform Fees, including advertising
- $10K for respondent incentives
- $20K for data processing, weighting, and modeling
- $10K for data security and administrative costs

Survey Link and Copy

The Google Forms survey link will be included here: https://docs.google.com/forms/d/e/1FAIpQLSdcd_neJf83l R1vPk98gPIDsD4KC_X6T8tQ/viewform.

The survey questions are listed below: 1. **What is your age group?** - 18-24 - 25-34 - 35-44 - 45-54 - 55+

2. **What is your gender?**

   - Male
   - Female
   - Non-binary
   - Prefer not to say

3. **What is your ethnicity?**

   - White
   - Black or African American

- Asian
- Hispanic or Latino
- Native American or Alaska Native
- Two or more races
- Other
- Prefer not to say

4. **In which state do you currently reside?** *(Open-ended response)*

5. **What is your highest level of education completed?**

   - High school
   - Associate degree
   - Bachelor's degree
   - Other/Prefer not to say

6. **What is your political affiliation?**

   - Democrat
   - Republican
   - Independent
   - Other/Prefer not to say

7. **How likely are you to vote in the upcoming presidential election?** *(Scale of 1-5)*

8. **Which candidate do you currently support for president?**

   - Kamala Harris
   - Donald Trump
   - Other

9. **What is the most important issue to you in the upcoming election?**

   - Economy
   - Healthcare
   - Education
   - Climate change
   - Other/Prefer not to say

10. **What do you consider your economic status?**

    - Lower class
    - Lower-middle class
    - Middle class
    - Upper-middle class

- Upper class
- Prefer not to say

11. **How would you describe your household's financial situation compared to last year?**

    - Better
    - Worse
    - About the same
    - Prefer not to say

12. **How satisfied are you with the current administration's handling of key issues?** *(Scale of 1-5)*

Raw data full descriptions

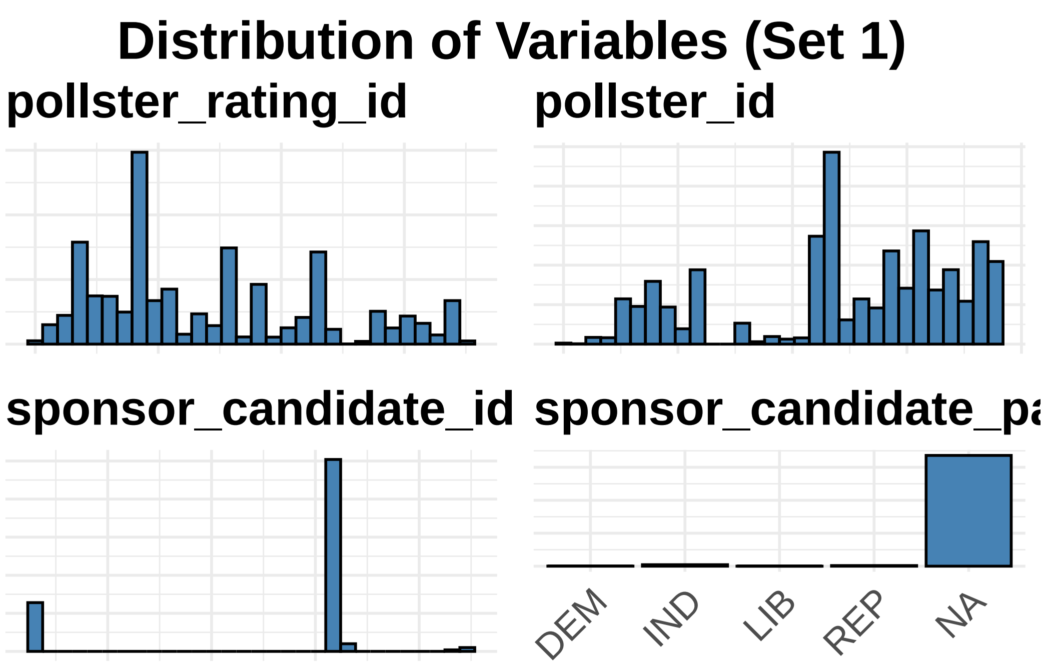


Figure 9: Raw Data Distribution

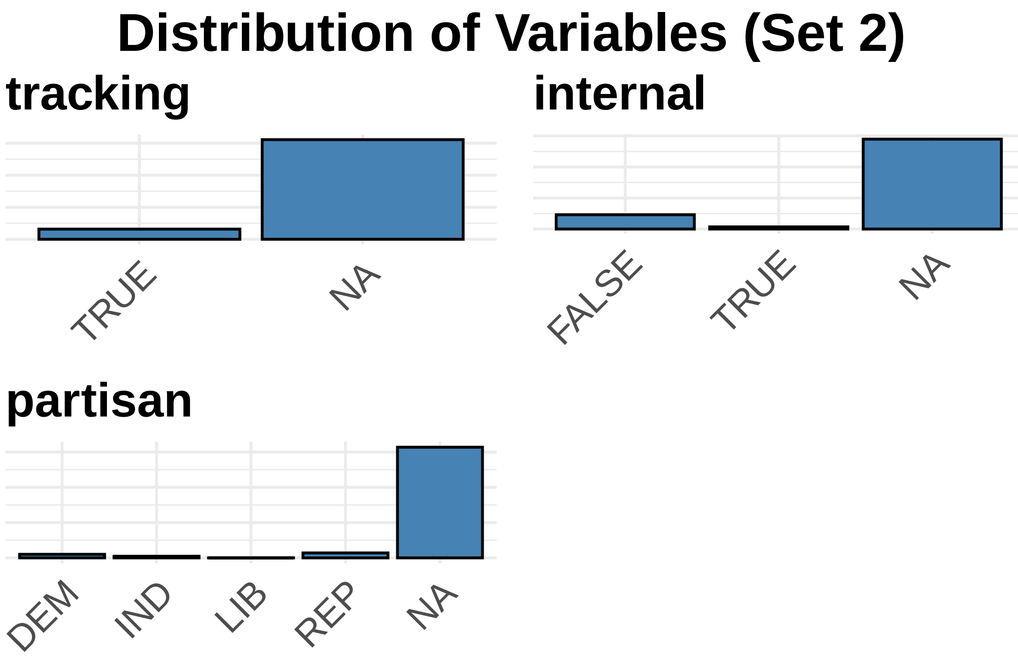


Figure 10: Raw Data Distribution

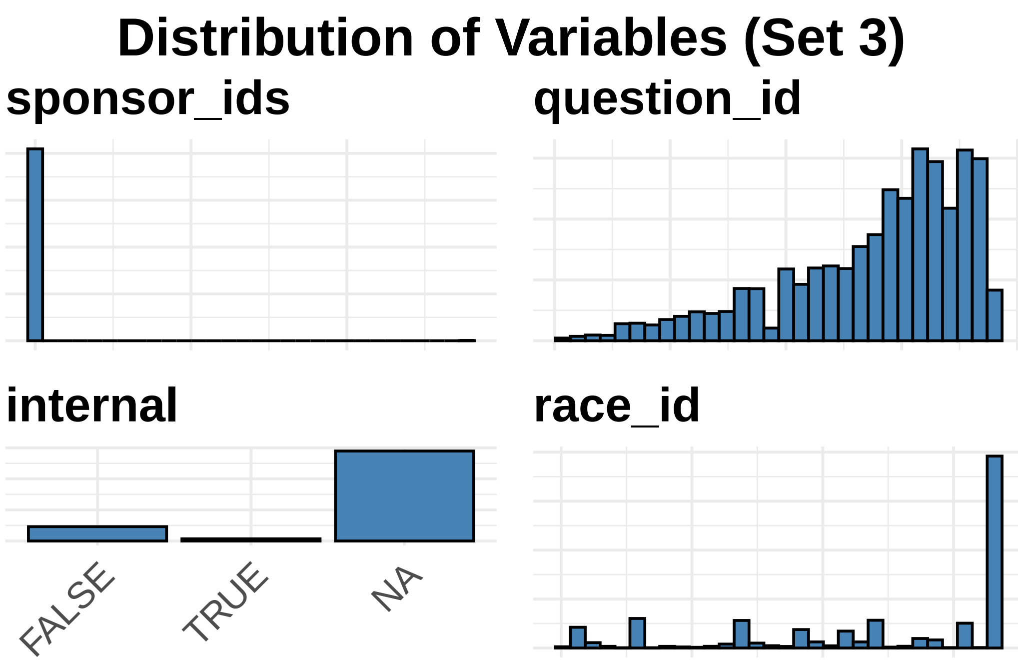


Figure 11: Raw Data Distribution

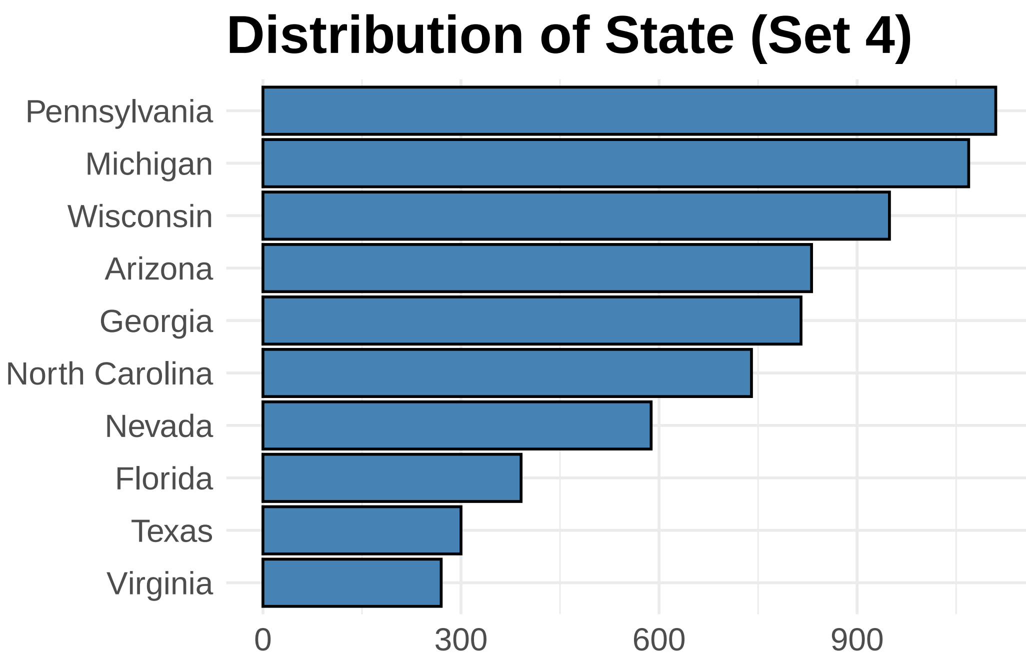


Figure 12: Raw Data Distribution

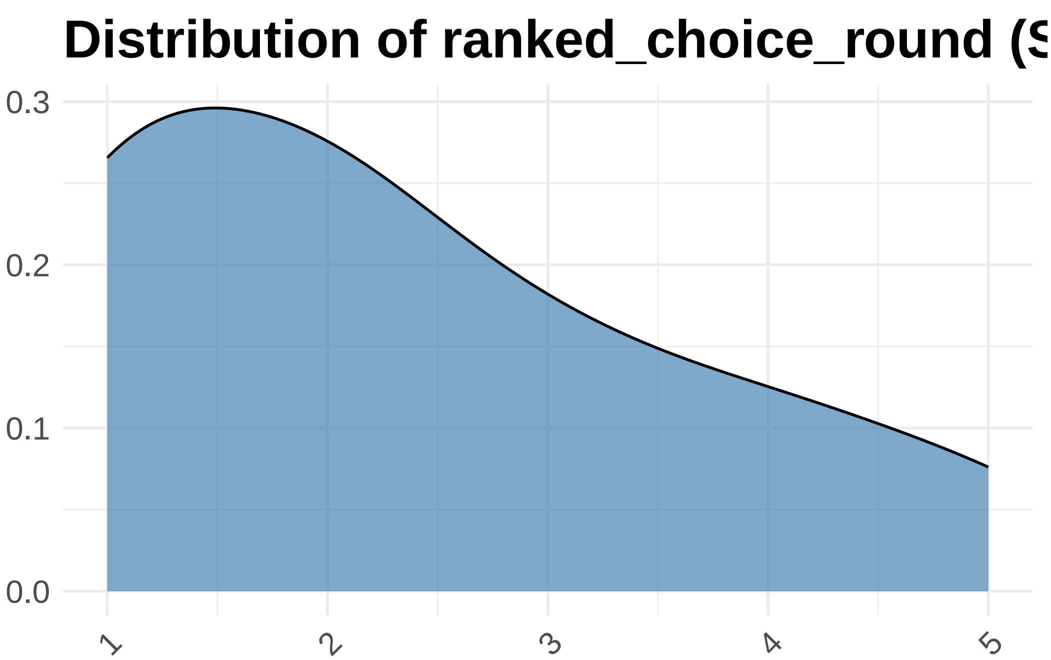


Figure 13: Raw Data Distribution

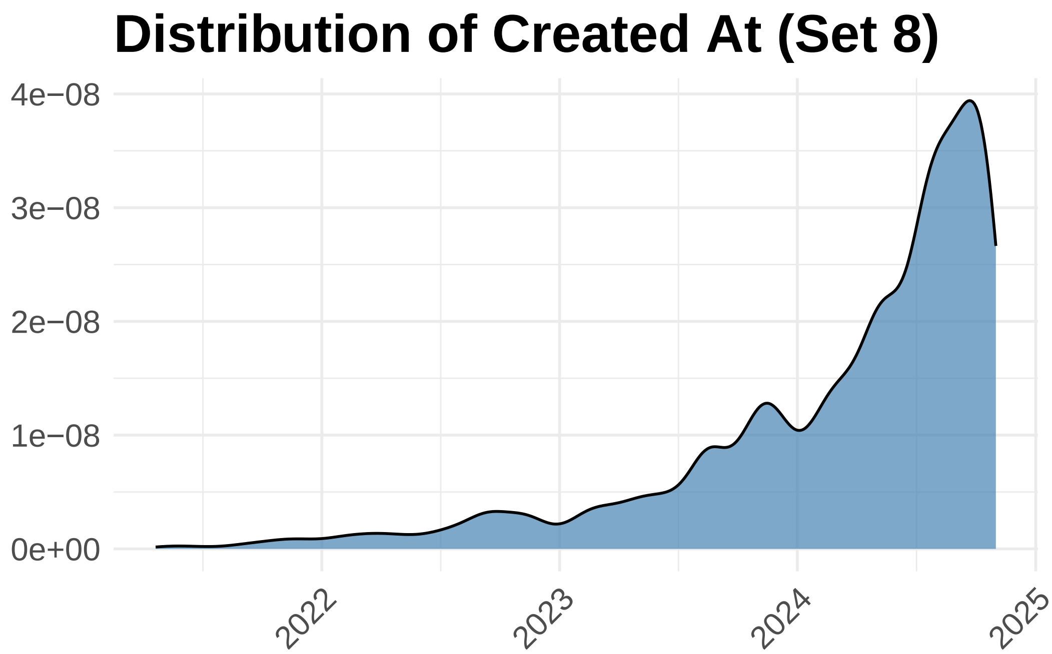


Figure 14: Raw Data Distribution