My title*

My subtitle if needed

First author Another author

November 2, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

The U.S. presidential election is a globally significant event, drawing substantial international attention due to its potential impact on economies, international relations, and pressing social issues, including climate change and human rights. As the world anticipates the 2024 U.S. electoral contest, forecasting the possible outcomes can provide critical insights for policymakers, businesses, and civil society groups. Predicting election outcomes has long been one of the most challenging tasks for political scientists, statisticians, and analysts. This study aims to predict the level of support for leading candidates, specifically Kamala Harris and Donald Trump, using aggregated poll data and statistical modeling.

To provide a robust prediction, we utilize a "poll-of-polls" approach, combining polling data from multiple organizations at both national and state levels to enhance accuracy and stability in forecasted support. Our analysis applies a multilinear regression model, focusing on key variables like pollster reliability, sample size, and duration, which contribute to the prediction of each candidate's support level. we aim to predict the percentage of support for the main candidates Kamala Harris and Donald Trump.

The results of our linear model indicate closely matched levels of voter support for both candidates. Donald Trump's predicted support is approximately 45.2%, while Kamala Harris's predicted support is slightly higher at 45.9%. These findings suggest that, while both candidates maintain stable support, Harris holds a slight advantage. The consistency across both models underlines the importance of accounting for variability in polling methodologies and regional voter bases. result

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

This study is important because it not only forecasts the immediate electoral outcomes but also provides critical insights into the potential global impact of the U.S. presidential election. Predicting election results helps global stakeholders—such as policymakers, investors, and international organizations—prepare for possible changes in U.S. policy, which could significantly affect trade, climate commitments, and foreign relations, thereby influencing global economic stability and strategic alliances. Furthermore, our findings highlight the need for transparency in polling methodologies and emphasize the importance of developing more precise and comprehensive forecasts.

This paper is organized as follows: The next section outlines the data collection and filtering processes, detailing the key variables and sources. Following that, we discuss our methodological framework, including the construction of the linear and Bayesian models. In the results section, we present the outcomes of both models, and a discussion explores the broader implications of these findings. Finally, the paper concludes with recommendations for future research and potential applications of these forecasting models in electoral studies. cross-reference

2 Data

2.1 Overview

The dataset comes from FiveThirtyEight's 'Presidential Election Polls (Current Cycle)' (Ryan Best and Wiederkehr, 2024). 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.

The analysis and visualizations in this paper are based on polling results as of October 22. The dataset includes 52 variables and 17,133 samples from various polling sources, asking participants who they support in the upcoming presidential election. To ensure accuracy and consistency, the data has been carefully processed and cleaned to remove biases and guarantee data integrity and comparability.

This project leverages several R packages, including tidyverse(**tidyverse?**), rstanarm(**rstanarm?**), testthat(**testthat?**), readr(**readr?**), broom(**broom?**), ggplot2(**ggplot2?**), and posterior(**posterior?**), to clean, analyze, and visualize polling data for the 2024 U.S. presidential election forecast. These packages facilitate a reproducible approach to data handling, statistical modeling, and result presentation in this study.

2.2 Raw data

Raw data 52 variable 17133 sample.

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		

variables appdendix

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).
start_date	The date the poll began (e.g., 10/8/24).
end_date	The date the poll ended (e.g., 10/11/24).
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).

52 varibles project "notes", "url", "url_article", "url_topline", "url_crosstab", "source"

variables "pollster", "sponsors", "display_name", "pollster_rating_name", "sponsor_candidate", "endorsed_candidate_name", "population_full", "candidate_id", "candidate_name"

Table 3: Constant variables

Variable	Value
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
nationwide_batch	FALSE

categorical "poll_id", "pollster_id", "sponsor_ids", "pollster_rating_id", "methodology", "state", "sponsor_candidate_id", "sponsor_candidate_party", "question_id", "population", "tracking", "created_at", "internal", "partisan","race_id", "ranked_choice_reallocated", "ranked_choice_round", "hypothetical","party","answer"

categorical appendix "poll_id", "methodology", "population", "ranked_choice_reallocated", "hypothetical","answer"

3530 poll

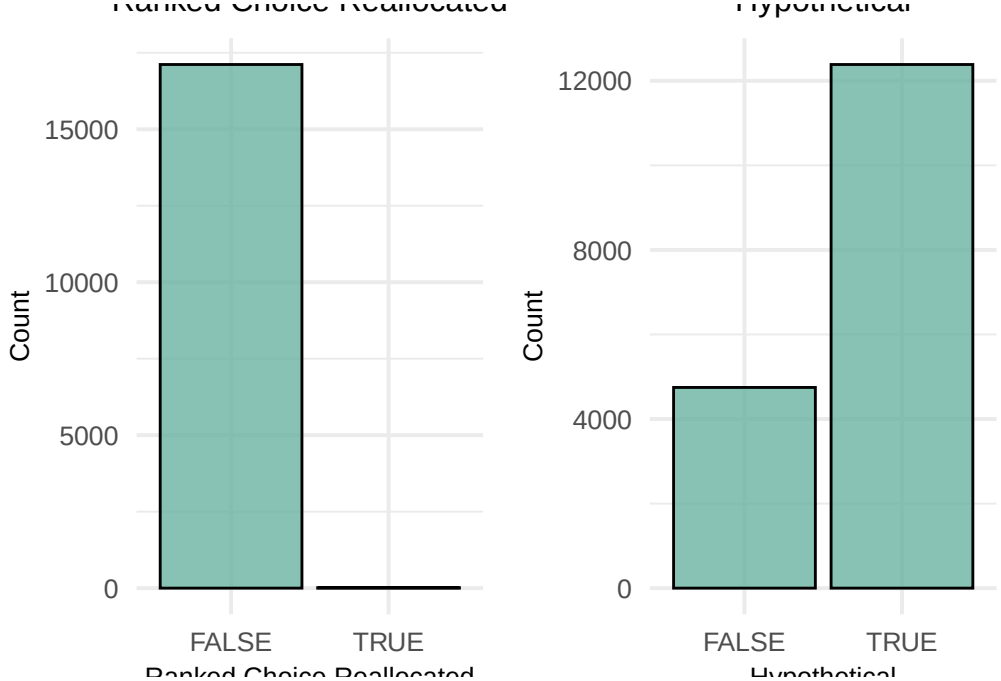


Figure 1: Boolean variables

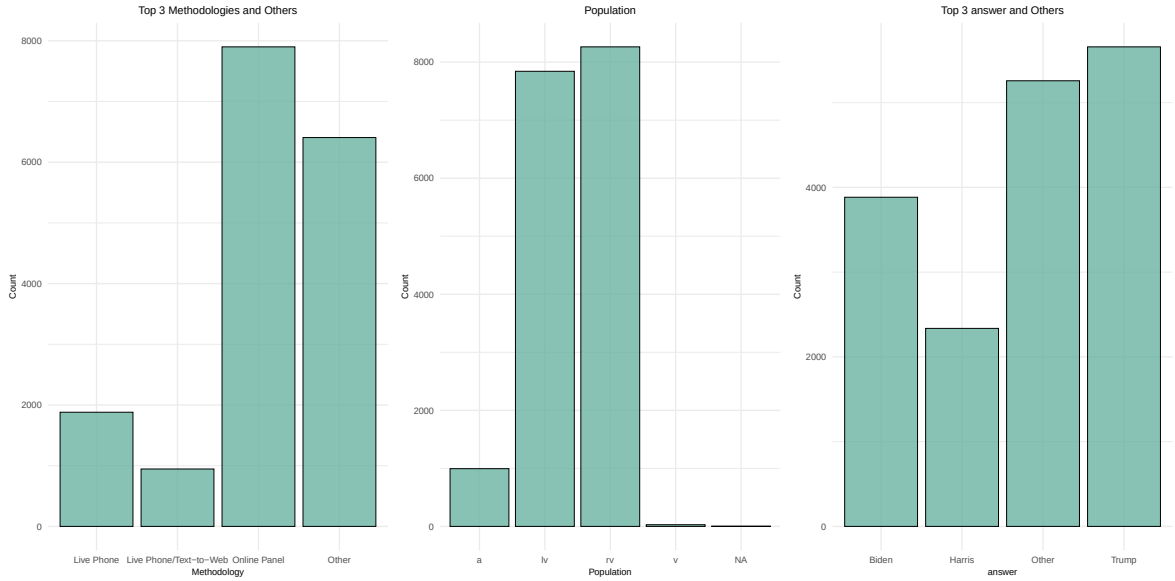


Figure 2: Catogorical variables

numerical variables “numeric_grade” “pollscore” “transparency_score” “start_date” “end_date” “sample_size” “pct”

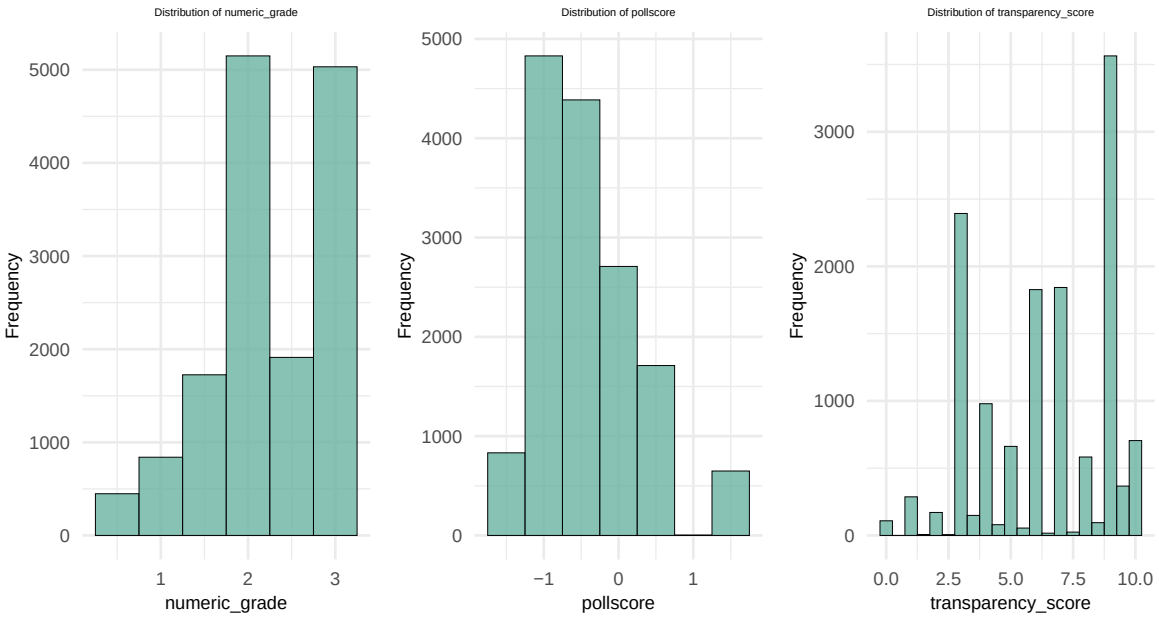


Figure 3: Distribution of numerical varibales part 1

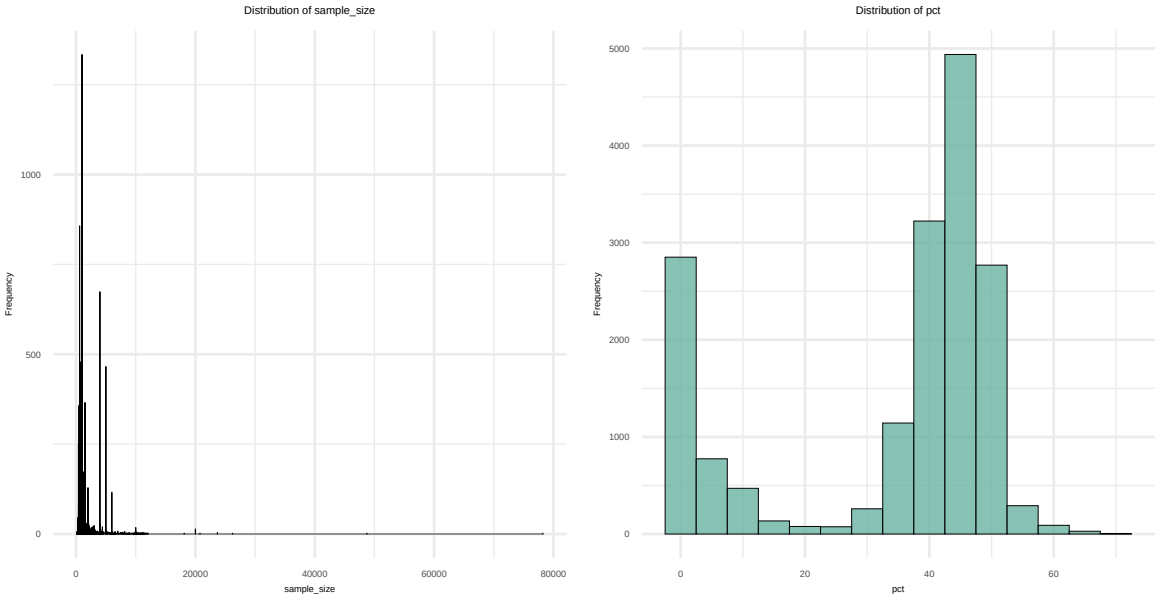


Figure 4: Distribution of numerical varibales part 2

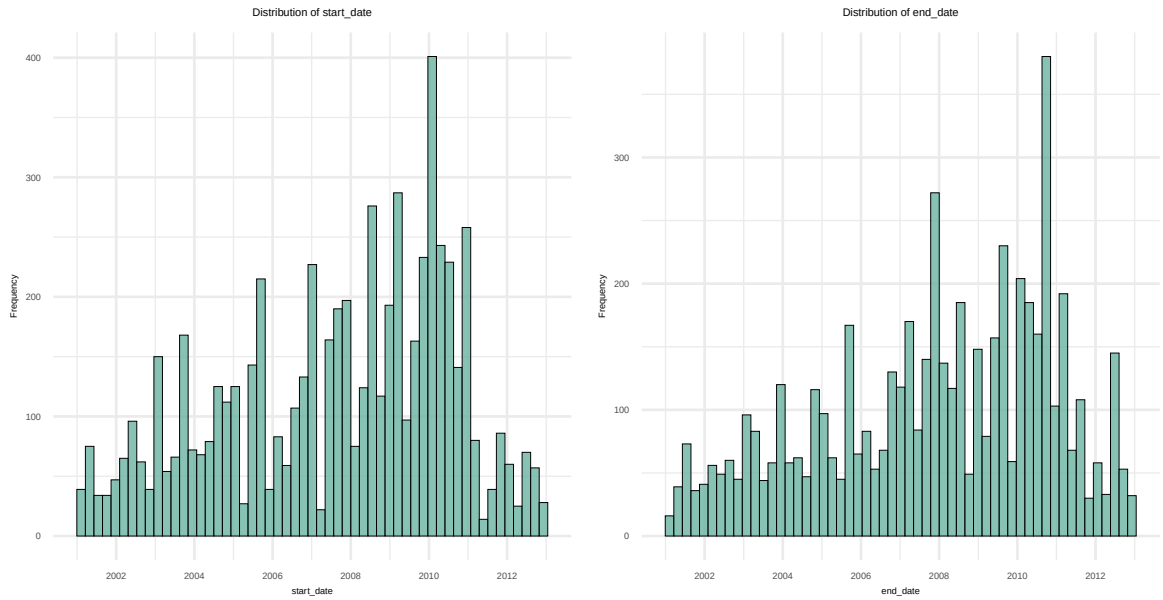


Figure 5: Distribution of date varibales

2.3 Cleaned data

In the raw data, we initially identified a total of 52 variables. Some of these variables, such as 'url', are clearly unrelated to the objectives of this project. There are also constant variables, such as 'election_date', which consistently contains the value '11/5/24'. Additionally, we found duplicate variables conveying the same information, like 'pollster_id' and 'pollster'.

Therefore, we first removed these irrelevant and redundant variables. The remaining variables are as follows:

Table 4: Remained variables

poll_id	pollster_id	sponsor_ids
pollster_rating_id	numeric_grade	pollscore
methodology	transparency_score	state
start_date	end_date	sponsor_candidate_id
sponsor_candidate_party	question_id	sample_size
population	tracking	created_at
internal	partisan	race_id
ranked_choice_reallocated	ranked_choice_round	hypothetical
party	answer	pct

Next, we calculated the percentage of missing values for each variable across the entire dataset. We then removed all variables with more than 40% missing values. These variables, along with their respective proportions of missing values, are as follows:

Table 5: Variables with big porpotion of missing values

Variable	NA Proportion
sponsor_ids	0.52
state	0.46
start_date	0.63
end_date	0.68
sponsor_candidate_id	0.98
sponsor_candidate_party	0.98
tracking	0.91
internal	0.85
partisan	0.92
ranked_choice_round	1.00

Since the influence of pollsters can be quantified using their ratings, such as 'numeric_grade', 'pollscore', and 'transparency_score', we removed these variables to simplify the dataset and the model. Similarly, 'created_at' was also removed due to its strong correlation with 'start_date'.

Finally, due to the limitations of our model, we removed 'race_id', 'party', and 'question_id'. The reason for this is that we will extract and analyze the data for each candidate individually, which makes 'race_id' and 'party' constant within the corresponding dataset. Additionally,

'question_id' contains 6,421 unique values, making it unsuitable for categorization, and we removed it to avoid overfitting the model.

Finally, the remaining variables are as follows:

Table 6: Final variables

poll_id	numeric_grade	pollscore
methodology	transparency_score	sample_size
population	ranked_choice_reallocated	hypothetical
answer	pct	start_date
end_date		

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.

Next, we extracted the data 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.

Next, we 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:

Table 7: Example of analysis data

numeric_grade	pollscore	methodology	transparency_score	sample_size	population	ranked_choice_reallocated	hypothetical	pct	duration
2.7	-0.8	level1	6	1373	lv	FALSE	FALSE	50.7	1
2.7	-0.8	level1	6	1373	lv	FALSE	FALSE	50.7	1
2.7	-0.8	level1	6	1005	lv	FALSE	FALSE	51.0	1
2.7	-0.8	level1	6	1212	lv	FALSE	FALSE	48.8	1
2.7	-0.8	level1	6	1212	lv	FALSE	FALSE	50.1	1
2.7	-0.8	level1	6	1136	lv	FALSE	FALSE	49.2	1

2.4 Measurement

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.

2.5 Similar dataset

3 Model

3.1 Model overview

2 model 2024 11 5

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

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

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

The core assumption of multiple linear regression is that there is a linear relationship between the dependent variable (outcome) and the independent variables. This linear relationship can be visually checked using scatter plots, which ideally should display a straight-line pattern rather than a curve. The residuals (the differences between observed and predicted values) should be normally distributed. This assumption can be assessed by examining a Q-Q plot. The correlation between independent variables should not be too high, meaning multicollinearity should be avoided. This can be checked using the Variance Inflation Factor (VIF). Homoscedasticity: The variance of the error terms (residuals) should be consistent across all levels of the independent variables. Residuals plotted against predicted values should not display any obvious pattern. Furthermore, in our model, the variable "duration" is derived from the difference between "start_date" and "end_date" in the original data. This means our model cannot account for the effects brought by time series, but considering the linear relationship between these two, we simplified it to meet the assumptions of using MLR.

Additionally, there are 51 different categories for the "methodology" variable in the original dataset. Having too many categories for a categorical variable would increase the complexity of our model, so we simplified it into four levels. Level 1 represents the lowest reliability of the methodology, while level 4 represents the highest. The specific classifications are as follows:

Table 8: Methodology classfication

Level	Methodologies
level 1	Email, Email/Online Ad, Live Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone, Mail-to-Web/Mail-to-Phone, Online Ad
level 2	App Panel, IVR, IVR/Live Phone/Text/Online Panel/Email, IVR/Online Panel, IVR/Online Panel/Email, IVR/Online Panel/Text-to-Web, IVR/Online Panel/Text-to-Web/Email, IVR/Text, IVR/Text-to-Web, IVR/Text-to-Web/Email, Live Phone/Email, Live Phone/Online Panel/Mail-to-Web, Live Phone/Text/Online Ad, Live Phone/Text-to-Web/Email, Live Phone/Text-to-Web/Email/Mail-to-Web, Live Phone/Text-to-Web/Online Ad, Online Panel/Email, Online Panel/Email/Text-to-Web, Online Panel/Online Ad, Text-to-Web/Email, Text-to-Web/Online Ad

level 3	IVR/Live Phone/Online Panel, IVR/Live Phone/Online Panel/Text-to-Web, IVR/Live Phone/Text, IVR/Live Phone/Text-to-Web, Live Phone/Online Panel/App Panel, Live Phone/Online Panel/Text, Live Phone/Online Panel/Text-to-Web, Live Phone/Online Panel/Text-to-Web/Text, Live Phone/Text, Live Phone/Text/Online Panel, Live Phone/Text-to-Web, Live Phone/Text-to-Web/App Panel, Online Panel, Online Panel/Text, Online Panel/Text-to-Web, Online Panel/Text-to-Web/Text, Text, Text-to-Web
level 4	Live Phone, Live Phone/Online Panel, Live Phone/Probability Panel, Online Panel/Probability Panel, Probability Panel

The different methodologies were evaluated based on reliability scores ranging from 1 to 10. The scores of 51 combinations were calculated by averaging the individual scores of each methodology in the combination. The results were classified into four levels: high reliability (8.5-10), medium-high reliability (7-8.49), medium reliability (5-6.99), and low reliability (below 5). Methodologies were scored based on several criteria, including statistical rigor, representativeness, response rate, interaction quality, and cost efficiency. High-scoring methodologies, such as those employing strict statistical sampling methods like Probability Panels or those using Live Phone surveys with broad coverage and low refusal rates, received high scores due to their strong representativeness and reliability. Medium-high scoring methodologies included Online Panels, which offer good coverage and low cost but are susceptible to self-selection bias, and Text-to-Web methods, which improve response rates but may have limited representativeness depending on the target demographics. Medium-scoring methodologies, such as App Panels and IVR (Interactive Voice Response), tend to lack broad representativeness or have limitations in interaction quality, making them suitable for niche audiences but not generalizable to a wider population. Low-scoring methodologies, including Email Surveys and methods relying on Online Ads, often suffer from low response rates and significant selection bias, which negatively impact their reliability. These scoring criteria ensure that the methodologies are evaluated in a consistent manner, with higher scores reflecting stronger statistical foundations, broader representativeness, and better data quality.

Additionally, the reason we only compared the data of Trump and Harris to predict which of them would win the election is that, after comparing the weighted competitiveness scores of all candidates, we found that the top three — Trump, Harris, and Biden — had significantly higher scores than the others. As shown in **?@tbl-t5c**, the third-place candidate, Harris, had a weighted competitiveness score of 180714535, which is 5.15 times higher than the fourth-place score of 34987031. Considering that Biden has withdrawn from the race, we ultimately decided to only predict between the most competitive candidates, Trump and Harris. The specific weighting formula is as follows:

$$weighted\_score = \frac{\sum_{i=1}^{n} score_i}{n} \times \sum_{i=1}^{n} Sample\_Size_i \quad (3)$$

Where:

n represents the number of polls nominating this candidate.

$Score_i$ represents the score obtained by the candidate in the i^{th} poll.

$Sample\_Size_i$ represents the sample size of the i^{th} poll.

Table 9: Weighted score for top 5 candidates

Candidate	Weighted_Score
Trump	440292227
Biden	299557346
Harris	180714535
DeSantis	34987031
Kennedy	16789972

3.2 Model set-up

3.2.1 response variable

score response variable

score score score 50 score 44.63 24.68 score score

Table 10: Sumarry of score in training datasets

dataset	mean	variance	sample_size
train_Trump	44.63	24.68	3960
train_Harris	46.92	20.96	1636

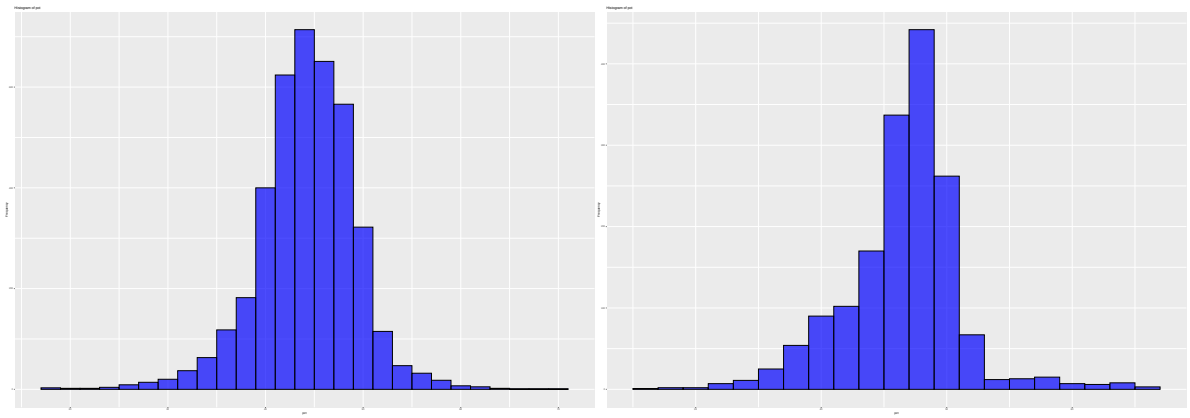


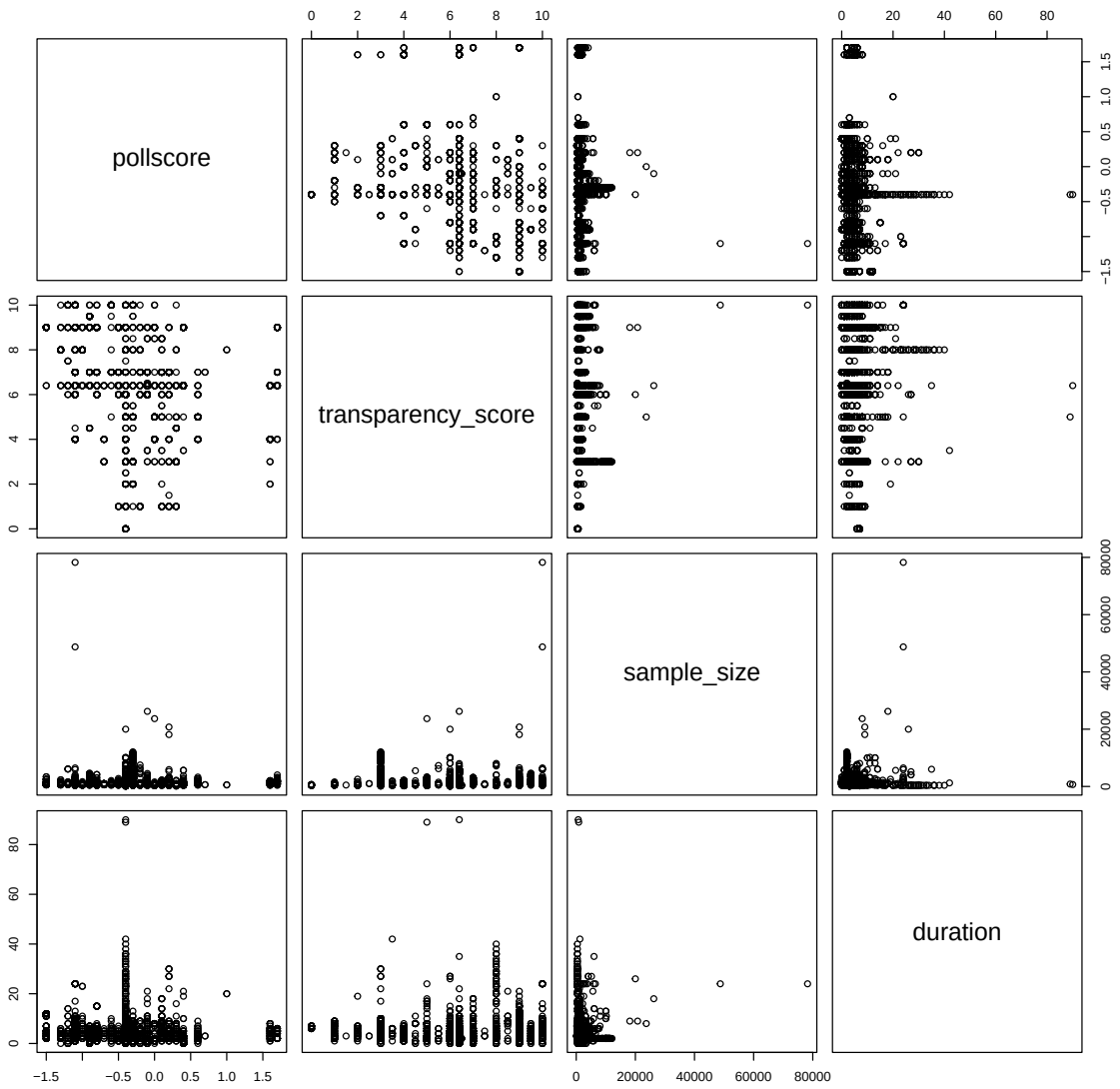
Figure 6: Distribution of scores in training dataset

3.2.2 Predictor

predictors

```
[1] "pollscore"                 "methodology"
[3] "transparency_score"        "sample_size"
[5] "population"                "ranked_choice_reallocated"
[7] "hypothetical"              "duration"
```

numerical predictor correlation


Predictor vs predictor
pollscore
transparency_score
sample_size
duration

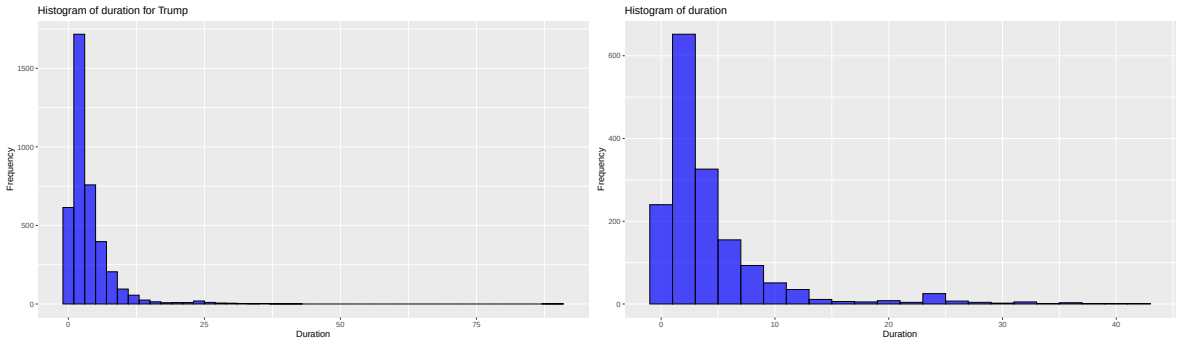


Figure 7: Distribution of durations in training dataset

categorical

3.2.3 alternative models

numeric_grade pollscore

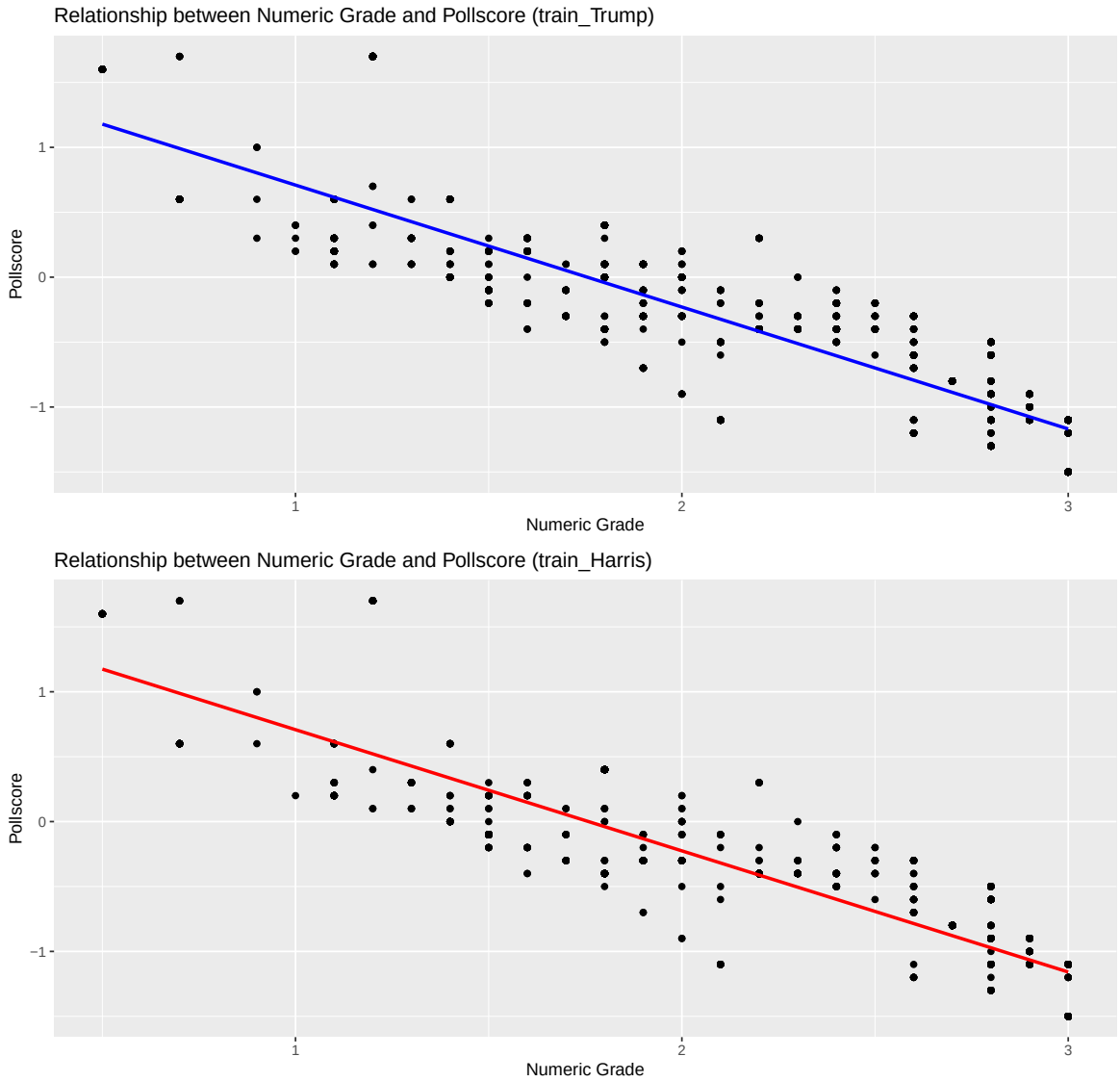


Figure 8: Relationship between numeric_grade and pollscore

predictors methodology ranked_choice_reallocated statistical significant p value 0.05.

Table 11: Significant level of variblaes in the aboanded Harris's model

Table 11: 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 12: Significant level of variblaes in the aboanded Trump's model

Table 12: Model for Trump

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

3.3 validation

GVIF 1 1.3

Table 13: VIF of Harris's final model

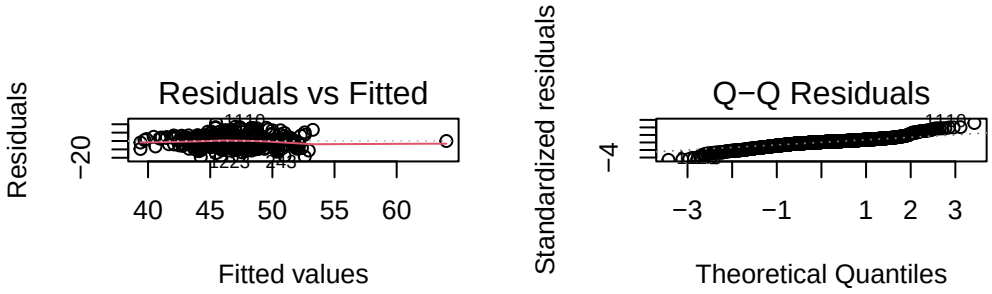
Table 13: VIF for Harris Model

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

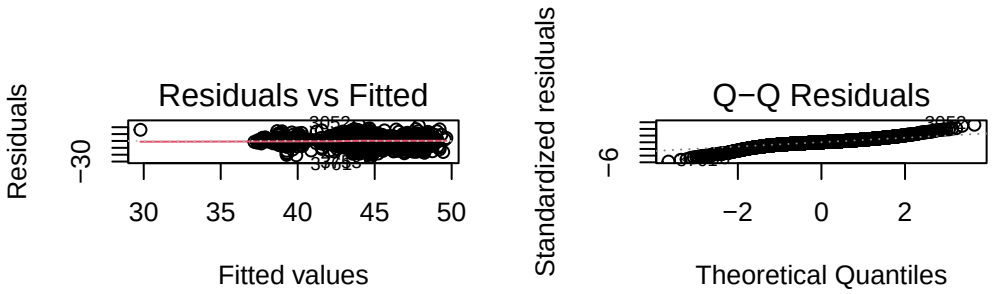
Table 14: VIF of Trump's final model

Table 14: VIF for Harris Model

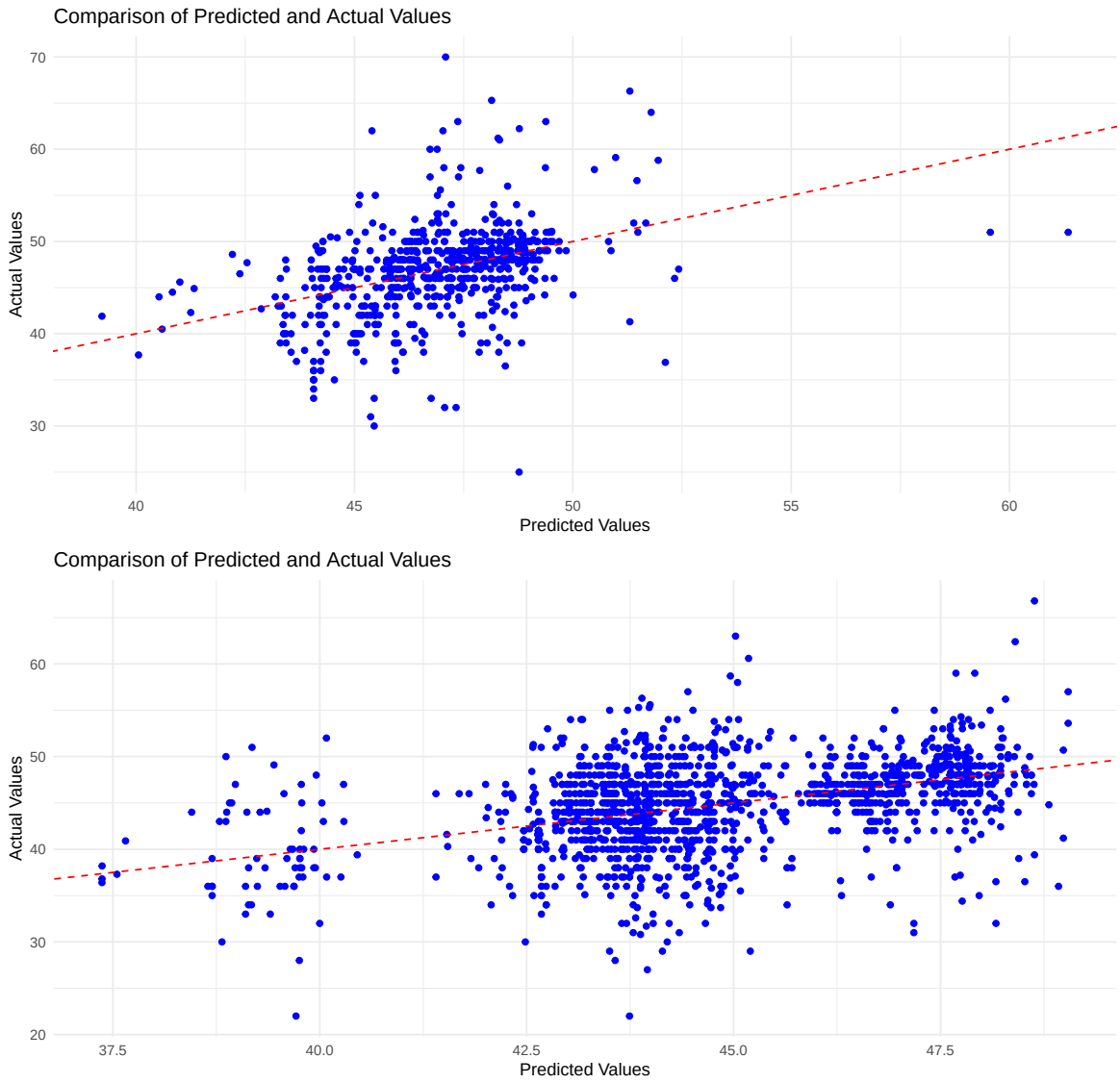
	GVIF
pollscore	1.102432
transparency_score	1.205052
duration	1.087927
sample_size	1.090086
population	1.111258
hypothetical	1.028432




```
par(mfrow=c(2,2))
plot(Trump_model,1)
plot(Trump_model,2)
```



response variable normal numerical variable MLR



4 Result

4.1 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.

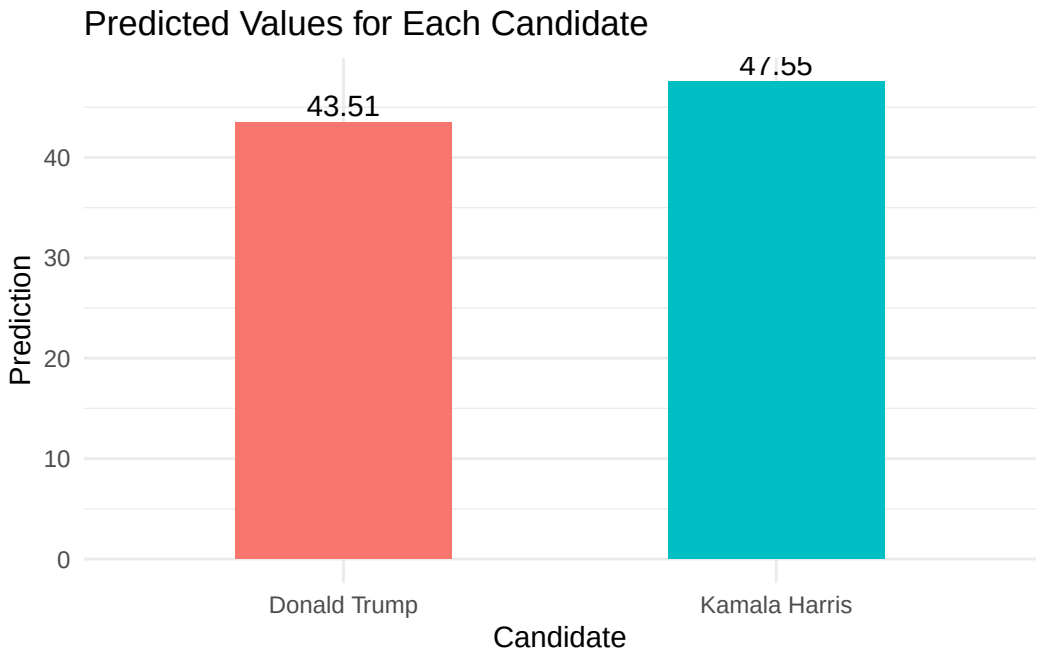
We selected representative feature values for each candidate's dataset by processing each variable based on its type. For numeric variables (e.g., numeric_grade, pollscore, transparency_score, duration, sample_size), we determined whether to use the mean or median by evaluating skewness; variables with 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 15: 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

4.2 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.



4.3 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.

5 Discussion

6 Appendix

6.1 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.

6.1.1 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.

6.1.2 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.

6.1.3 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.

6.1.4 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.

6.1.5 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.

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

6.2.1 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.

6.2.2 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.

6.2.3 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.

6.2.4 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.

6.2.5 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.

6.2.6 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.

6.2.7 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

6.2.8 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)*

6.3 Raw data full descriptions

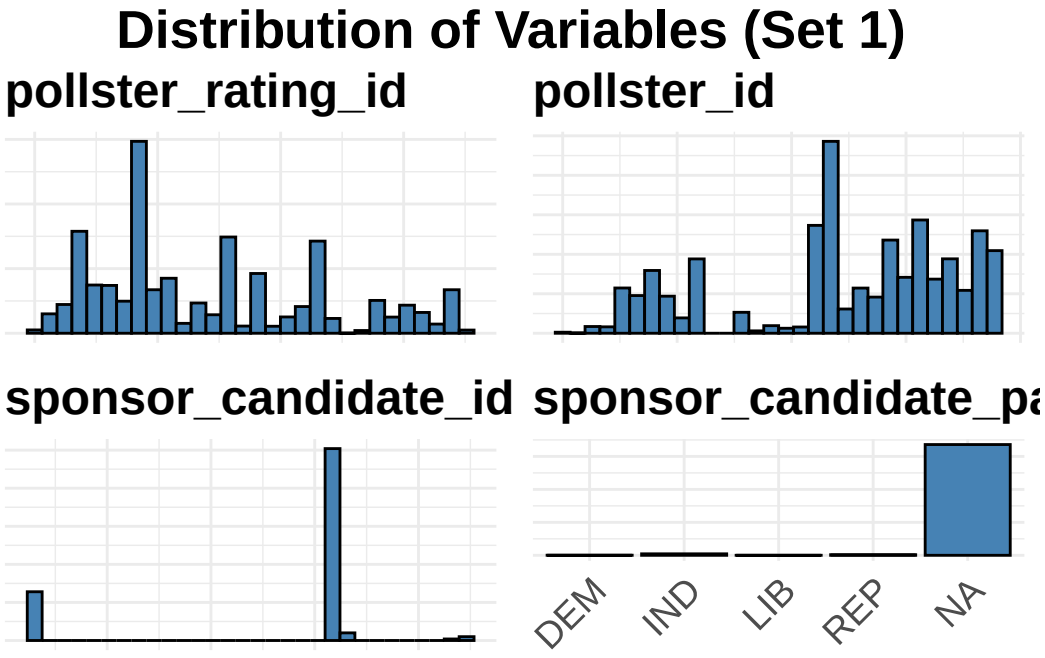


Figure 9: Raw Data Distribution

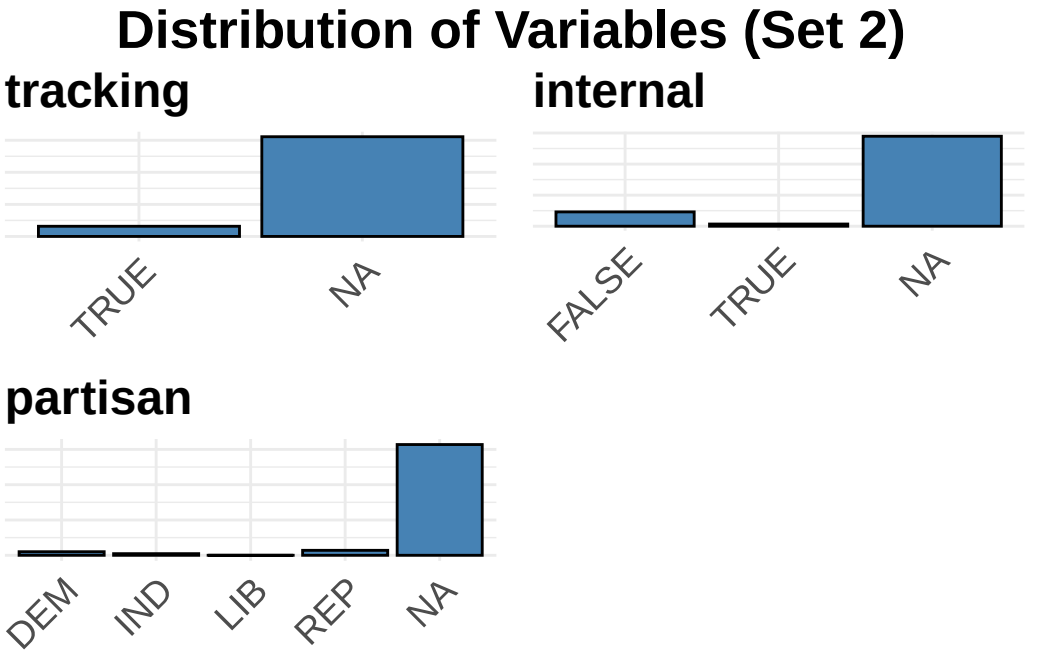


Figure 10: Raw Data Distribution

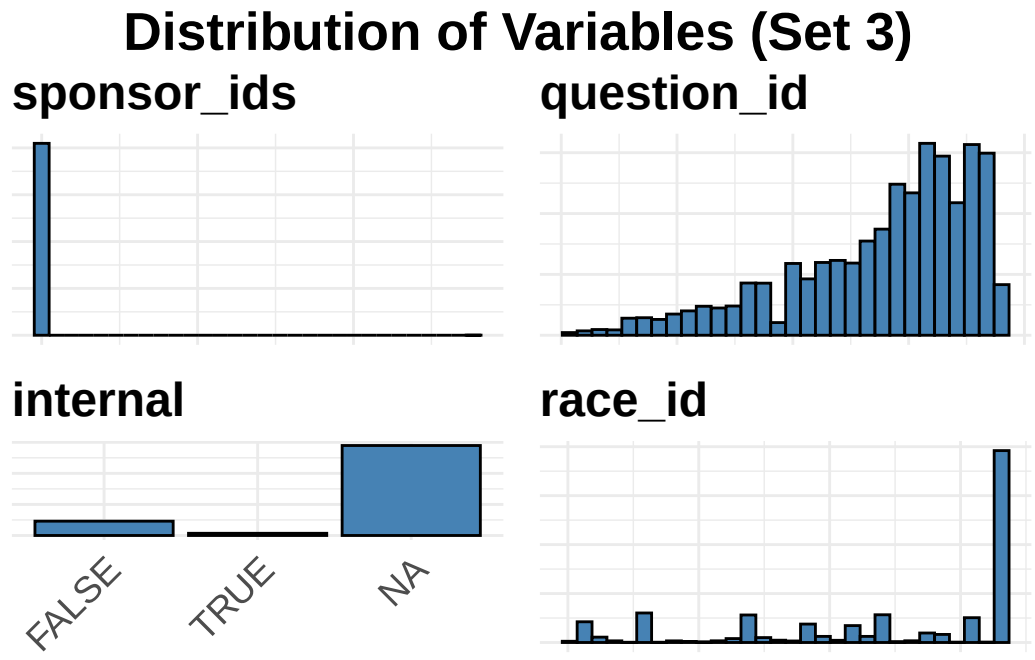


Figure 11: Raw Data Distribution

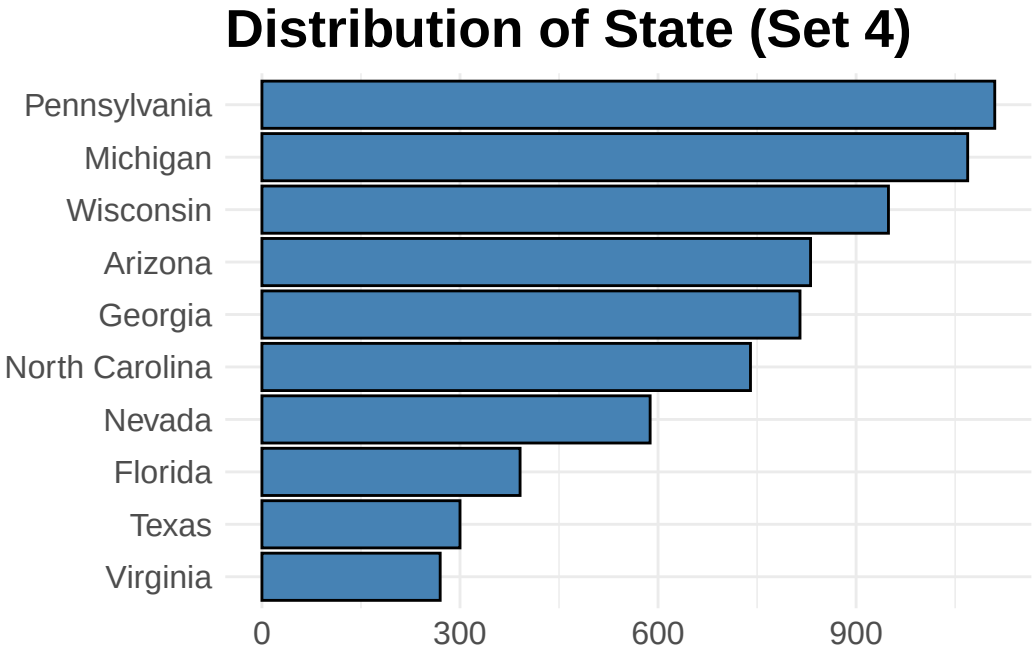


Figure 12: Raw Data Distribution

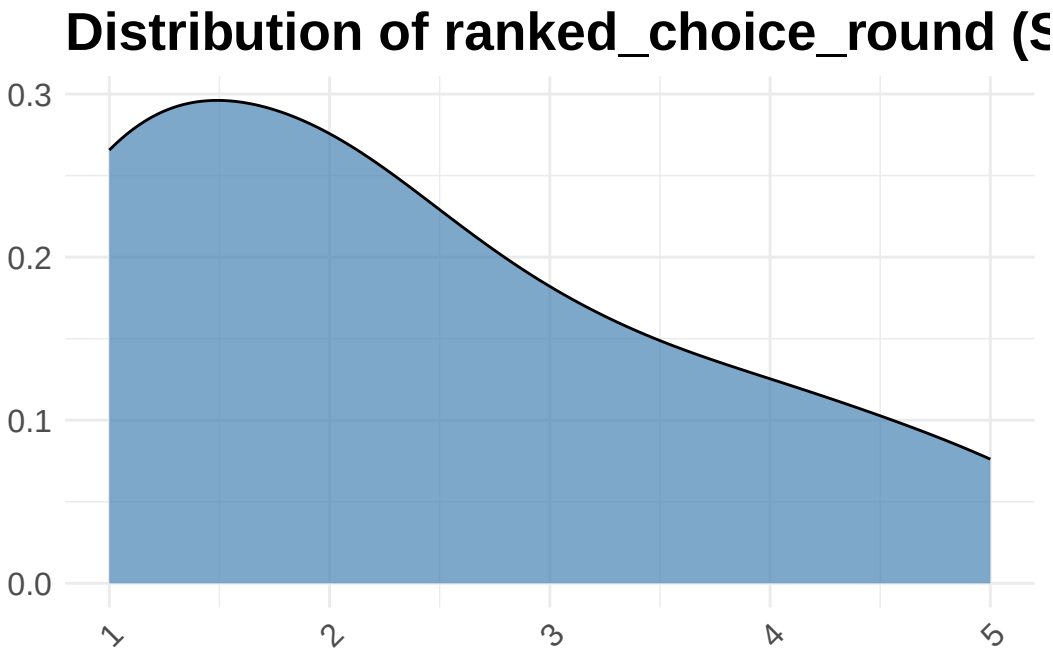


Figure 13: Raw Data Distribution

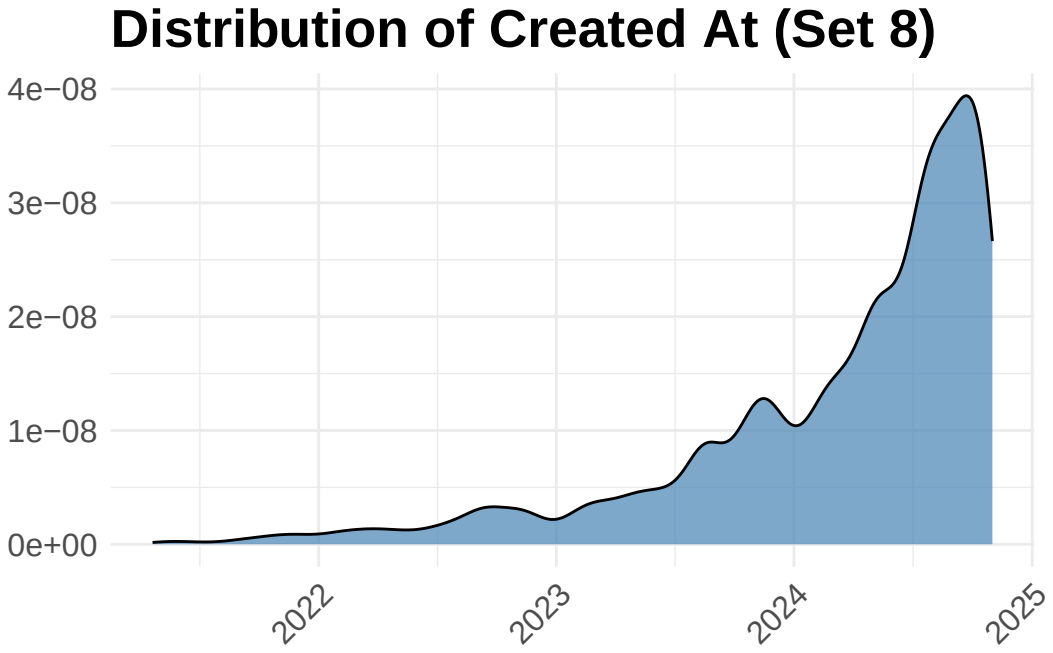


Figure 14: Raw Data Distribution