My title*

My subtitle if needed

First author Another author

November 2, 2024

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

1 Introduction

2 Data

2.1 Overview

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

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

Table 1: Varibales of raw data

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

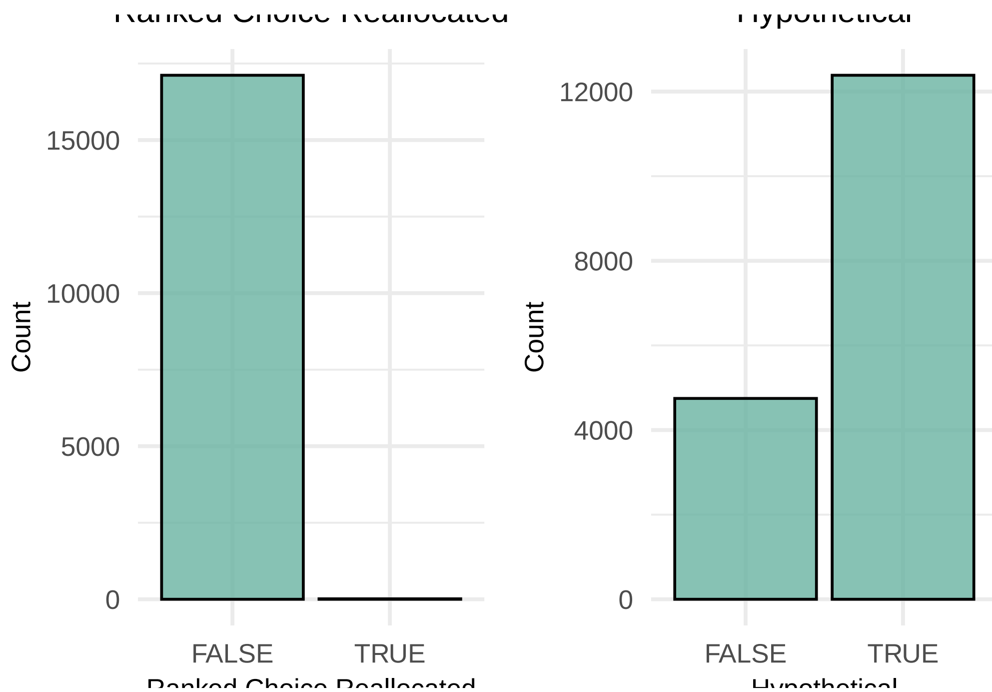
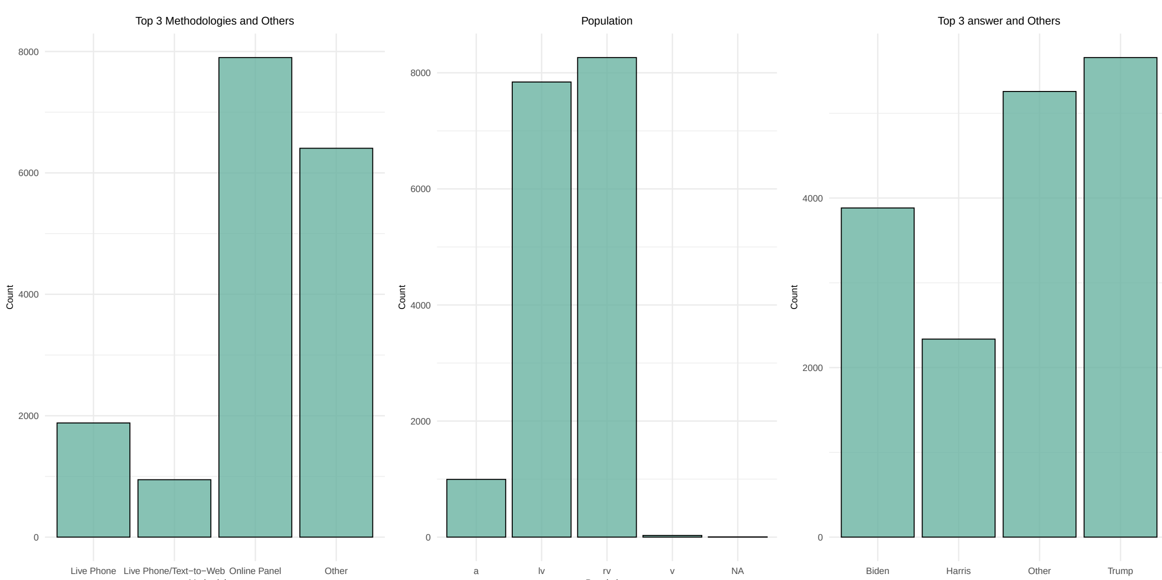


Figure 1: Boolean variables



numerical variables "numeric_grade" "pollscore" "transparency_score" "start_date" "end_date" "sample_size" "pct"

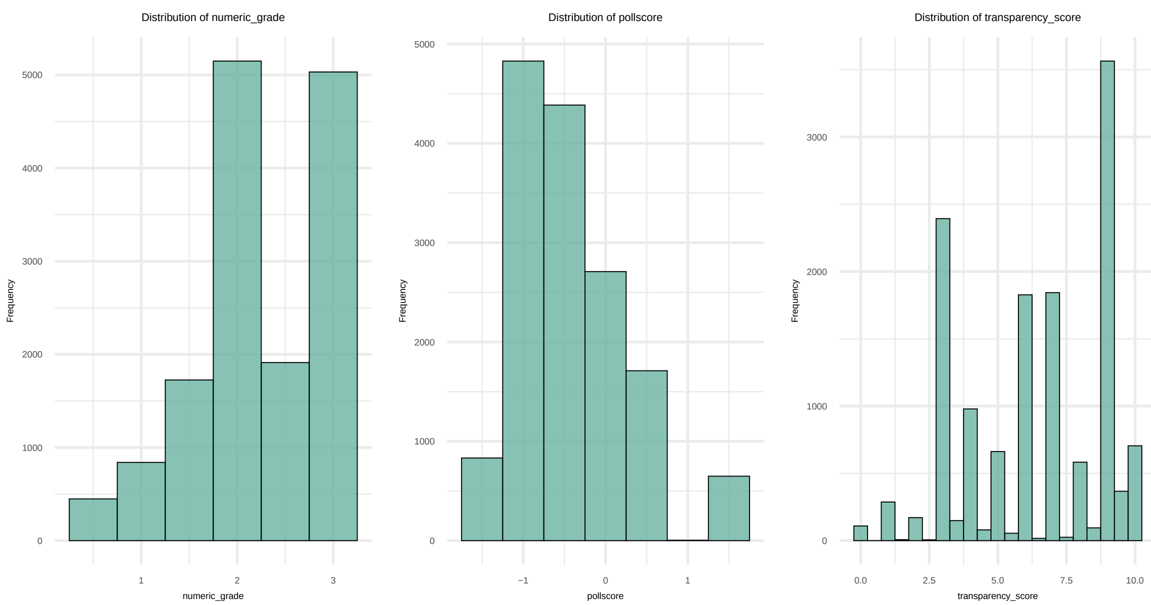


Figure 3: Distribution of numerical varibales part 1

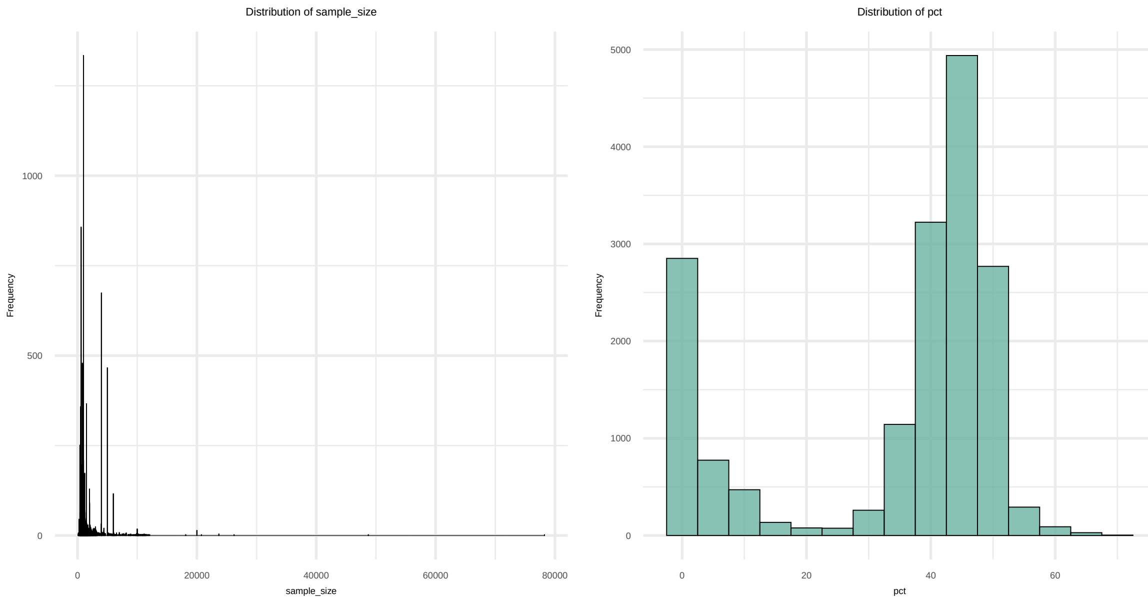


Figure 4: Distribution of numerical varibales part 2

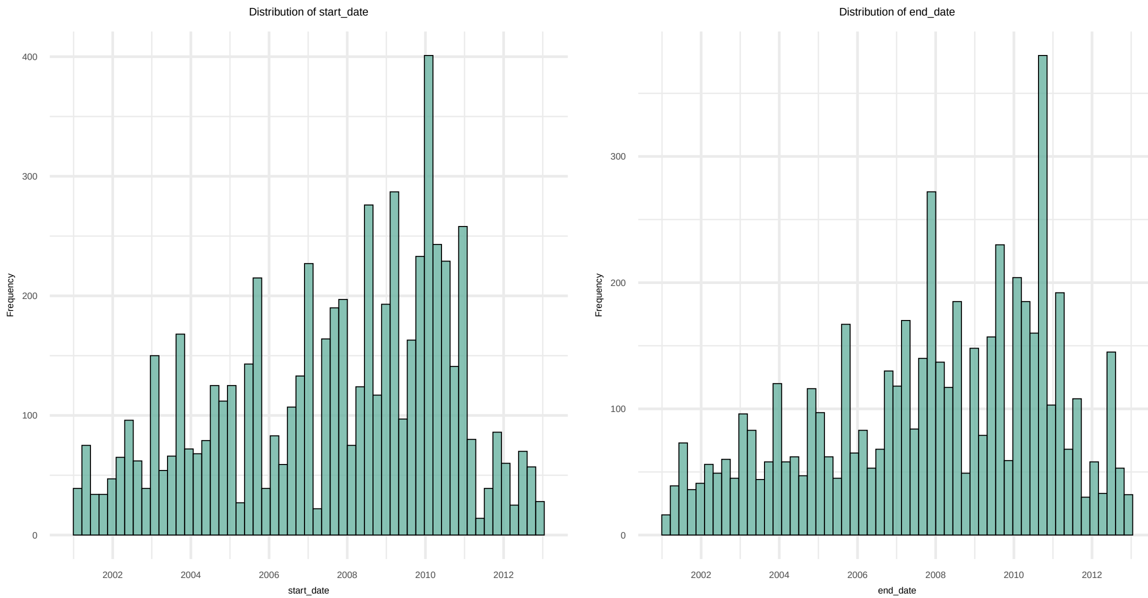


Figure 5: Distribution of date varibales

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	score	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

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 Table 10, 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:

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

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.

```
# Define the directory path containing the parquet files
p <- "data/03-cleaned_data"
parquet_files <- list.files(path = here::here(p), full.names = TRUE)

# Initialize an empty dataframe to store the weighted score for each file
weighted_scores <- data.frame(Candidate = character(), Weighted_Score = numeric(), stringsAs

# Iterate over each parquet file
for (file in parquet_files) {
  # Read the parquet file
  df <- read_parquet(here::here(file))

  # Calculate the mean of score and sample_size
  score_mean <- mean(df$score, na.rm = TRUE)
  sample_size_mean <- mean(df$sample_size, na.rm = TRUE)

  # Calculate the weighted score (multiply by total number of rows)
  weighted_score <- score_mean * sample_size_mean * nrow(df)

  # Append the result to the weighted_scores dataframe
  weighted_scores <- rbind(weighted_scores, data.frame(Candidate = sub('_cleaned_data.parque
}

# Get the top 5 weighted scores
top_5_weighted_scores <- weighted_scores %>%
```

```
  arrange(desc(Weighted_Score)) %>%
  head(5)

# Print the top 5 weighted scores using kable
top_5_weighted_scores %>% kable()
```

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

Table 10: Top 5 Candidates

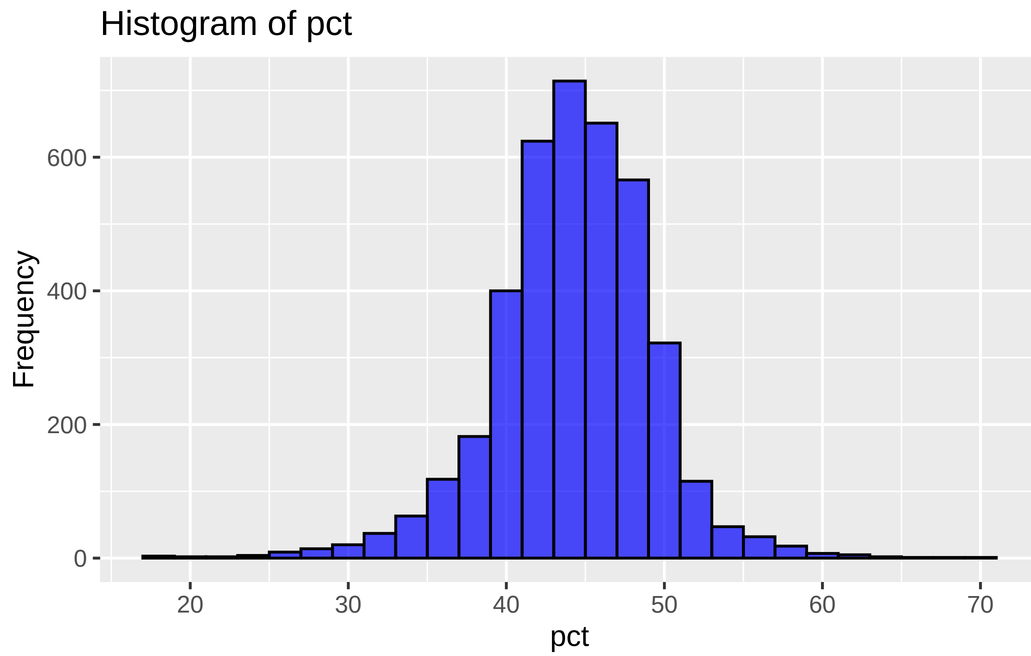
candidate	number of polls	weighted score
Donald Trump	5657	252424.47
Joe Biden	3883	161611.44
Kamala Harris	2336	109501.54
Ron DeSantis	466	18822.81
Robert F. Kennedy	1330	14749.50

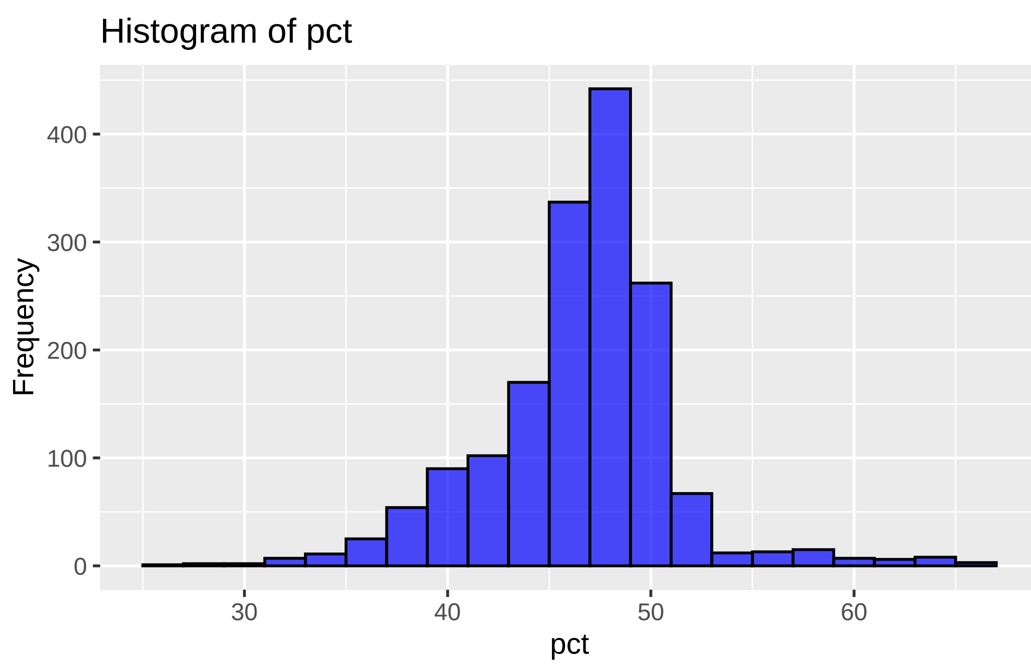
3.2 Model set-up

3.2.1 response variable

score response variable
score score score 50 score 44.63 24.68 score score

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



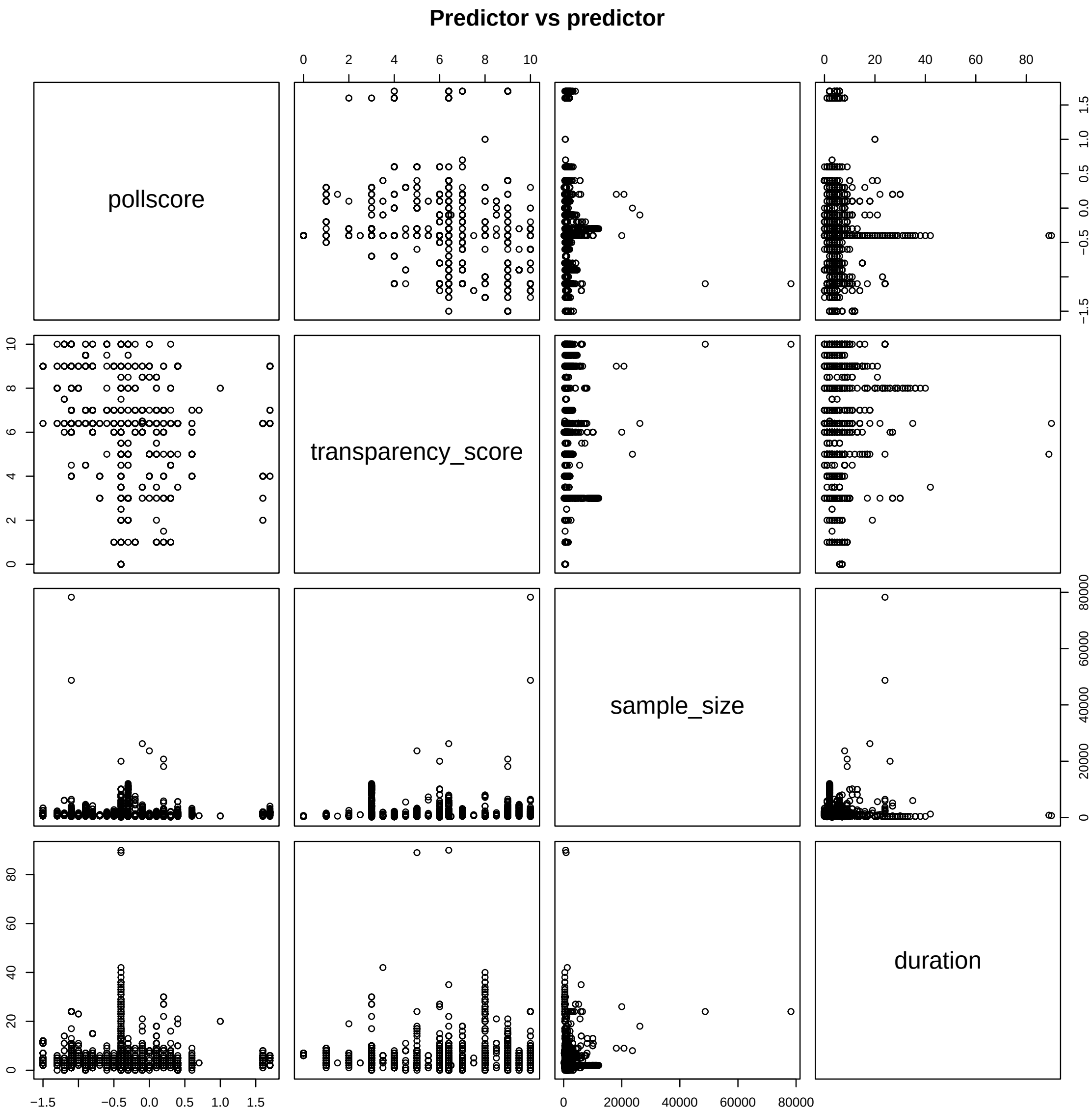


3.2.2 Predictor

predictors

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

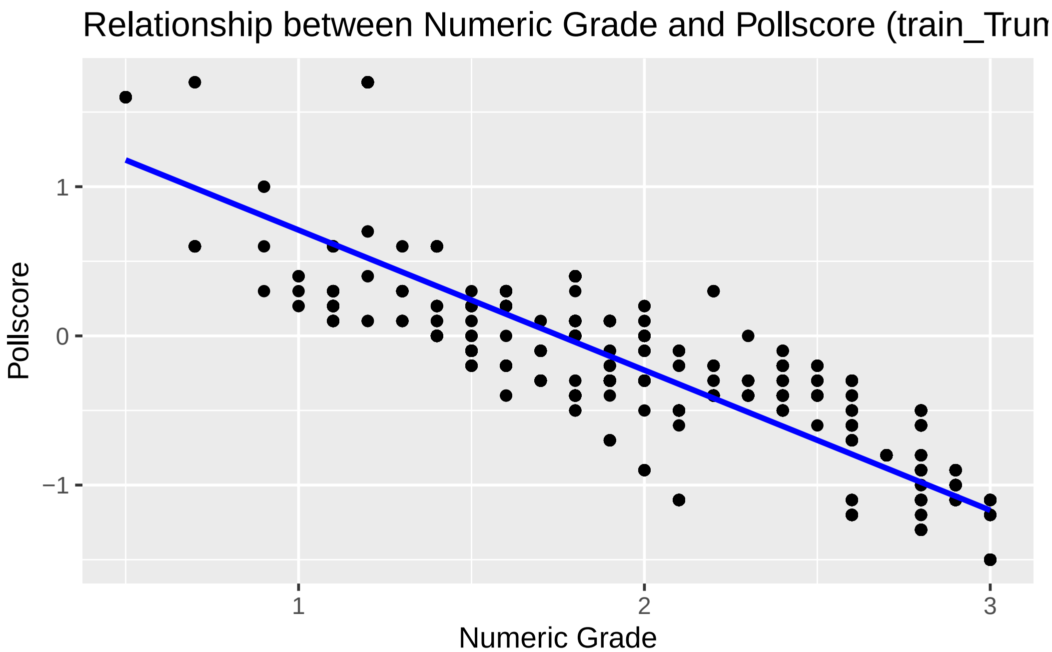
numerical predictor correlation

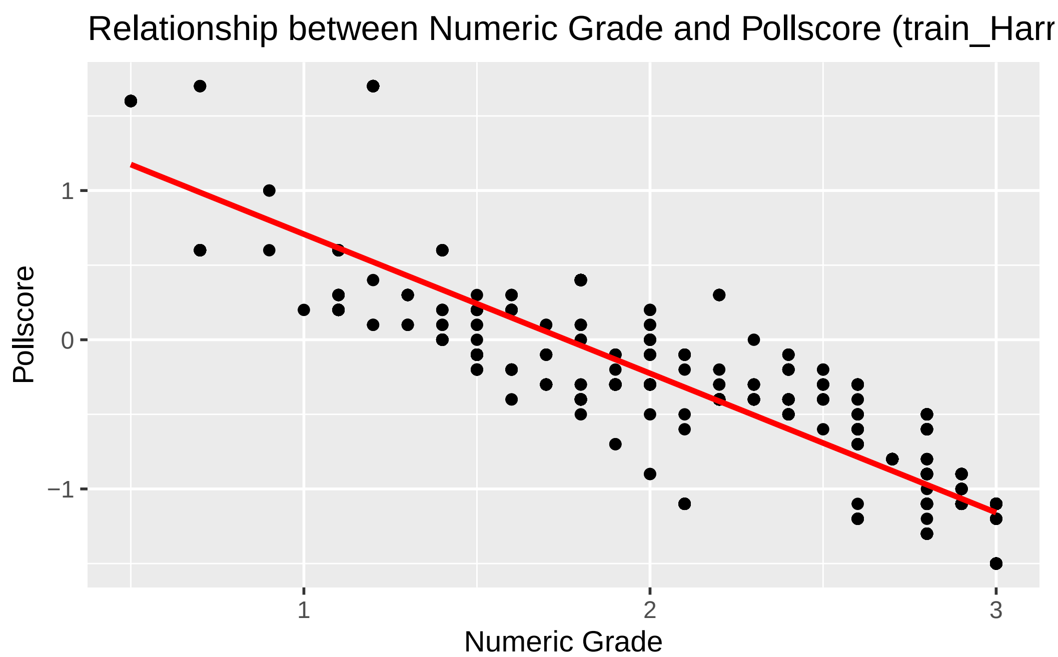


categorical

3.2.3 alternative models

numeric_grade pollscore





predictors methodology ranked_choice_reallocated statistical signifi- cant p value 0.05.

|Variable	|P-	|Variable	|P−value |
----------------------		|:------------------------------|:-----	
(Intercept)	|0.00	|(Intercept)	|0.000000e+00
pollscore	|1.92	|pollscore	|1.571566e−17
nsparency_score	|2	|transparency_score	|1.374640e−(
duration	|3.16	|duration	|1.782720e−15
ample_size	|3.8	|sample_size	|2.782493e−0:
populationlv	|5.3	|populationlv	|1.861643e−09
populationrv	|2.1	|populationrv	|2.207540e−02
populationv	|6.3	|populationv	|1.123945e−06
potheticalTRUE	|1.	|hypotheticalTRUE	|3.610001e−(
d_choice_reallocatedTRUE		|ranked_choice_reallocatedTRUE	|6.050213
ethodologylevel2	|1	|methodologylevel2	|3.326043e−(

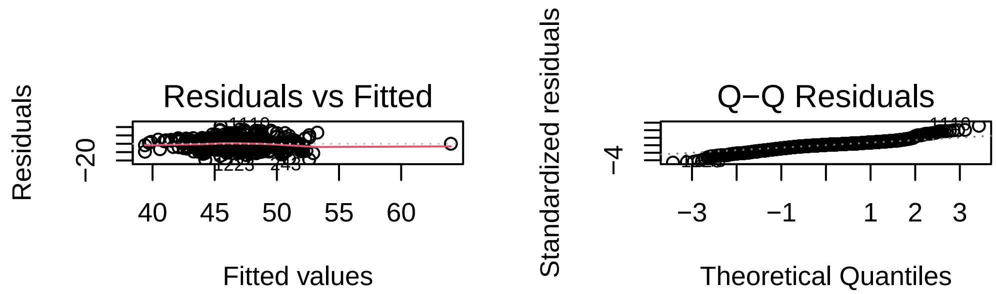
```
[1] "\nTrump_model <- lm(\n  score ~ pollscore + transparency_score + duration + sample_size


Call:
lm(formula = score ~ pollscore + transparency_score + duration +
    sample_size + population + hypothetical, data = Trump)

Residuals:
     Min       1Q   Median       3Q      Max
-26.0485  -1.8860   0.0366   2.1279  23.6316

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.293e+01  4.370e-01  98.253  < 2e-16 ***
pollscore          -6.757e-01  1.175e-01  -5.749 9.66e-09 ***
transparency_score -1.337e-01  3.280e-02  -4.076 4.68e-05 ***
duration            5.821e-02  1.579e-02   3.687  0.00023 ***
sample_size        -1.783e-04  3.003e-05  -5.937 3.16e-09 ***
populationlv        5.059e+00  3.353e-01  15.091  < 2e-16 ***
populationrv        4.106e+00  3.309e-01  12.410  < 2e-16 ***
populationv         2.201e+00  1.463e+00   1.505  0.13239
hypotheticalTRUE   -2.969e+00  1.600e-01 -18.551  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.502 on 3951 degrees of freedom
Multiple R-squared:  0.1803,    Adjusted R-squared:  0.1787
F-statistic: 108.7 on 8 and 3951 DF,  p-value: < 2.2e-16



Call:
lm(formula = score ~ pollscore + transparency_score + duration +
    sample_size + population + hypothetical, data = Harris)

Residuals:
    Min      1Q  Median      3Q     Max
-19.242  -1.769   0.435   2.008  21.526

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.397e+01  6.566e-01  66.968  < 2e-16 ***
pollscore          -1.438e+00  1.720e-01  -8.364  < 2e-16 ***
```

```
transparency_score -1.318e-01  4.341e-02  -3.036 0.002433 **
duration            1.625e-01  2.008e-02   8.092 1.14e-15 ***
sample_size         1.689e-04  5.065e-05   3.334 0.000877 ***
populationlv        3.201e+00  5.247e-01   6.100 1.32e-09 ***
populationrv        1.263e+00  5.437e-01   2.323 0.020316 *
populationv         2.031e+01  4.216e+00   4.817 1.59e-06 ***
hypotheticalTRUE   -1.614e+00  2.686e-01  -6.007 2.32e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.177 on 1627 degrees of freedom
Multiple R-squared:  0.1715,    Adjusted R-squared:  0.1674
F-statistic:  42.1 on 8 and 1627 DF,  p-value: < 2.2e-16
```

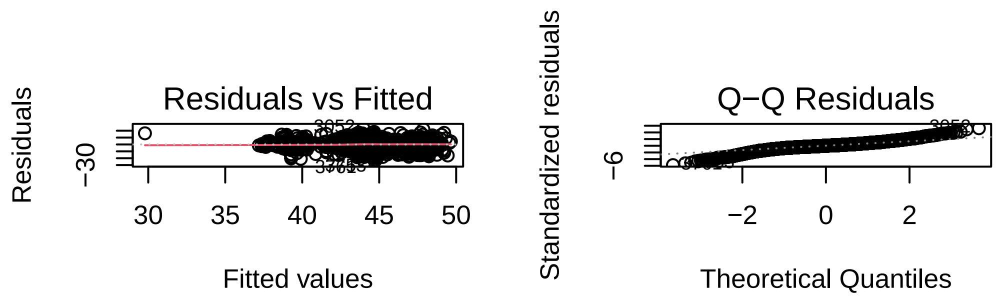
3.3 validation

GVIF 1 1.3

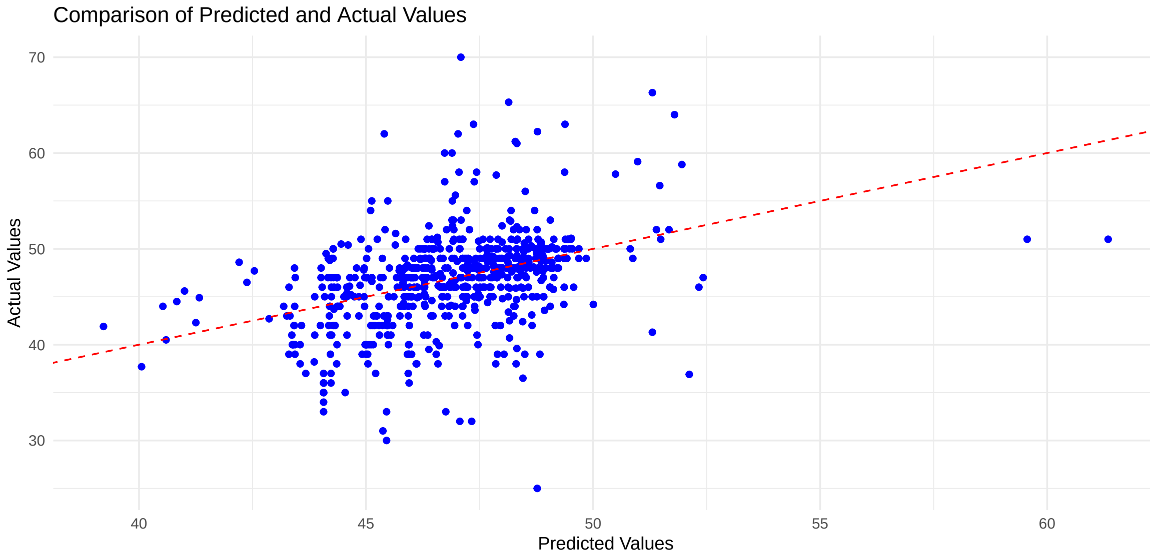
```
                       GVIF Df GVIF^(1/(2*Df))
pollscore          1.101322  1        1.049439
transparency_score 1.126838  1        1.061526
duration           1.037154  1        1.018408
sample_size        1.051550  1        1.025451
population         1.203265  3        1.031320
hypothetical       1.112077  1        1.054550

                       GVIF Df GVIF^(1/(2*Df))
pollscore          1.102432  1        1.049968
transparency_score 1.205052  1        1.097748
duration           1.087927  1        1.043037
sample_size        1.090086  1        1.044072
population         1.111258  3        1.017738
hypothetical       1.028432  1        1.014116
```

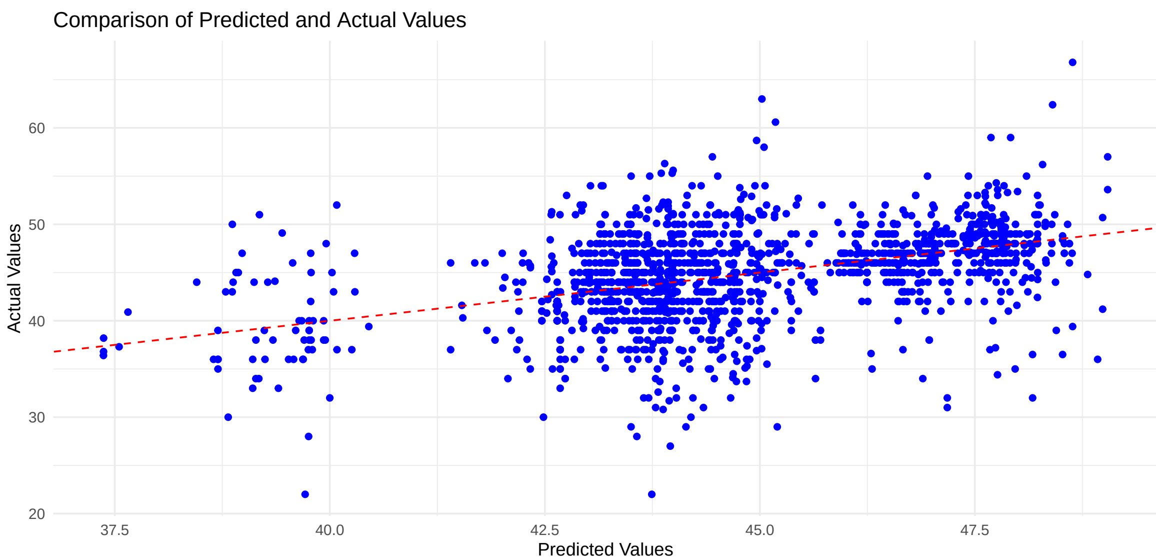



```r
par(mfrow=c(2,2))
plot(Trump_model,1)
plot(Trump_model,2)
```



response variable normal numerical variable MLR





4 Result

4.1 featured values used in prediction

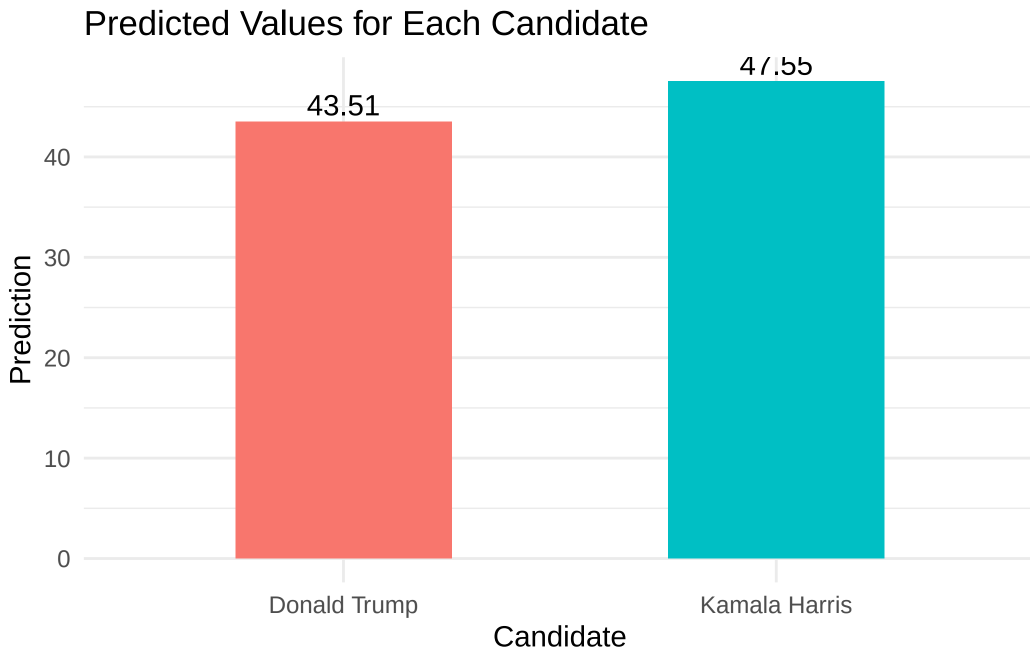
In our analysis, we designated specific poll-related features as "featured values" to serve as representative indicators within each candidate's dataset. These featured values were chosen to highlight the most impactful aspects of the polling data that consistently influenced the support scores for Donald 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.

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