# Forecasting the 2024 U.S. Presidential Election*

**My subtitle if needed**

Tim Chen Steven Li Tommy Fu

October 20, 2024

This paper presents a statistical model for forecasting the outcome of the 2024 U.S. Presidential Election. Using polling data from various sources, we develop multiple linear regression models to predict the percentage of support for the main candidates, Kamala Harris and Donald Trump, in different states. We aggregate these predictions to simulate the Electoral College vote and estimate the likelihood of either candidate winning. Our results suggest that neither candidate currently secures the required 270 electoral votes to win the presidency.

## 1 Introduction

Forecasting elections has been a challenging task for political scientists and statisticians, especially given the complexity and volatility of electoral dynamics in the U.S. The 2024 U.S. Presidential Election is no exception, with polling data playing a critical role in shaping public and expert expectations. Pollsters use different methodologies and sample populations, which can significantly affect the predictions.

This paper aims to forecast the 2024 U.S. Presidential Election using multiple linear regression models based on polling data. We use various predictors such as sample size, pollster ratings, and state-level data to predict support for Kamala Harris and Donald Trump across the United States. We then aggregate these predictions to simulate the Electoral College outcome.

The rest of this paper is structured as follows: Section 2 describes the data used for this analysis. Section 3 outlines the models developed for each candidate and their corresponding results. Section 4 discusses the aggregated Electoral College predictions. Finally, Section 5 provides conclusions and suggestions for future research.

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

# 2 Data

## 2.1 Overview

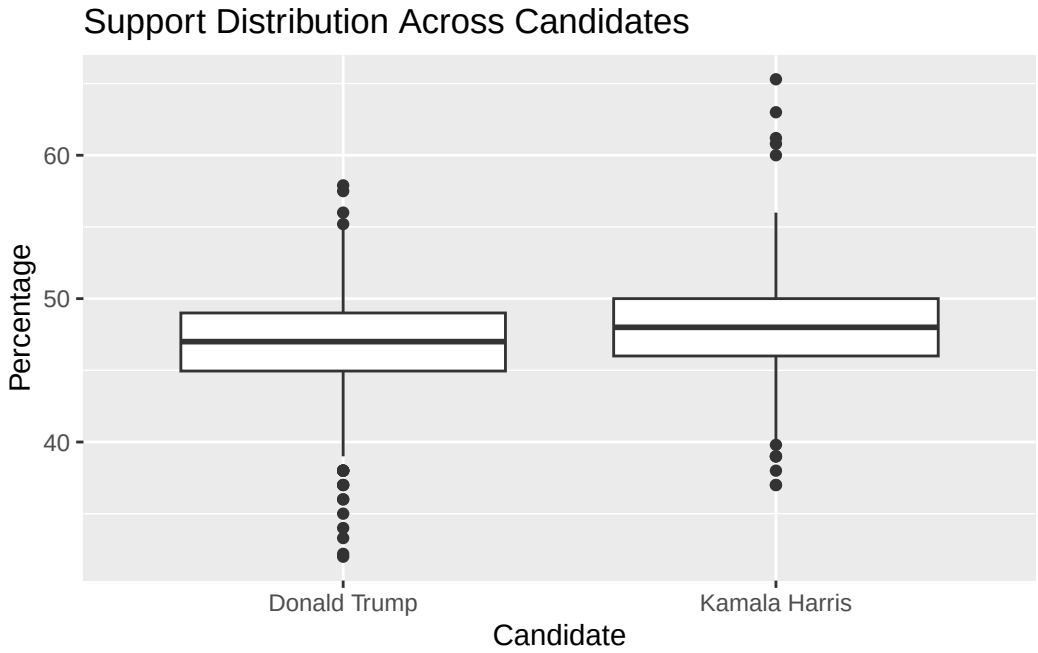
The dataset used in this analysis contains poll data from various sources, including information about polling organizations, sample sizes, methodologies, and state-level data. It includes several key variables, such as:

Pollster rating: A score that represents the reliability of the polling organization. Sample size: The number of respondents in each poll. Support percentage: The percentage of respondents supporting each candidate. State: The U.S. state where the poll was conducted or, in some cases, national polling data. The data has been processed and cleaned to ensure that all variables are correctly aligned for regression analysis. The variables were transformed into numeric forms where necessary (e.g., percentage of support), and missing data was handled appropriately.

## 2.2 Outcome variables

Our primary outcome variable is the percentage of support for Kamala Harris and Donald Trump in each poll. Figure 1 below shows the distribution of support for both candidates across the polls included in the dataset.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   45.00   47.00   46.98   49.00   65.30
```




Multiple Linear Regression (MLR) for Kamala Harris and Donald Trump

Electoral College Votes Prediction In this step, we aggregate the predictions for each state and determine the winner based on the electoral votes.

### 2.3 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

## 3 Modeling Support for the Candidates

### 3.1 Kamala Harris

We begin by modeling the percentage of support for Kamala Harris using a linear regression model. The predictors include the sample size, pollster ratings (e.g., pollscore and transparency score), and state. This model aims to quantify how these variables influence her support across different polls.

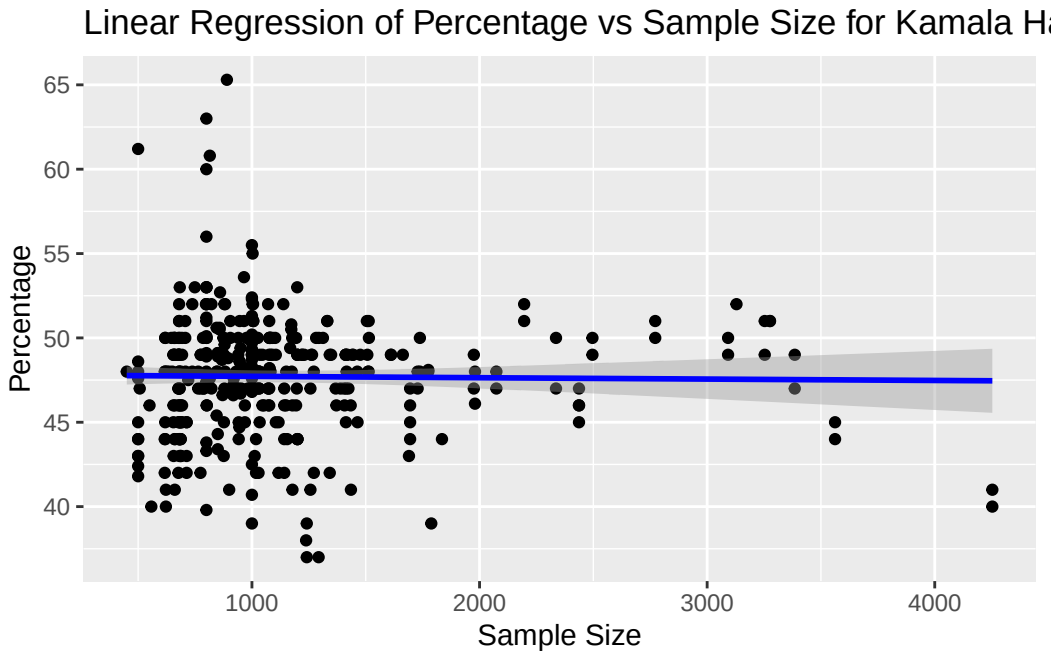
The results of the linear regression for Kamala Harris show that sample size has a statistically insignificant effect on her support. This suggests that factors other than sample size, such as

pollster methodology or regional biases, may play a more significant role in determining the level of support she receives.

Figure 2 illustrates the relationship between the sample size and support percentage for Kamala Harris.

```
# Plot the relationship for Kamala Harris
ggplot(harris_data, aes(x = sample_size, y = pct)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Linear Regression of Percentage vs Sample Size for Kamala Harris",
       x = "Sample Size",
       y = "Percentage")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




## 3.2 Donald Trump

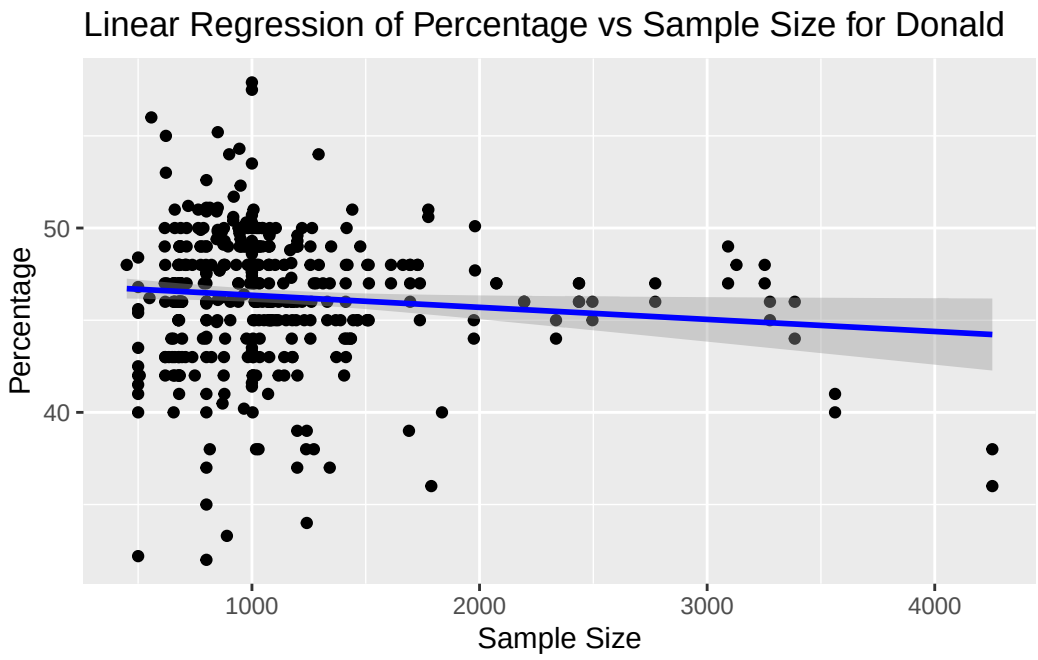
A similar linear regression model was applied to Donald Trump's data. The predictors remain the same, and the goal is to determine the factors that drive support for him.

The results indicate that sample size has a weak but statistically significant negative effect on support for Trump. This suggests that larger polls tend to show slightly lower support for Trump, although the effect size is small.

Figure 3 shows the relationship between the sample size and support percentage for Donald Trump.

```
# Plot the relationship for Kamala Harris
ggplot(trump_data, aes(x = sample_size, y = pct)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Linear Regression of Percentage vs Sample Size for Donald",
       x = "Sample Size",
       y = "Percentage")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

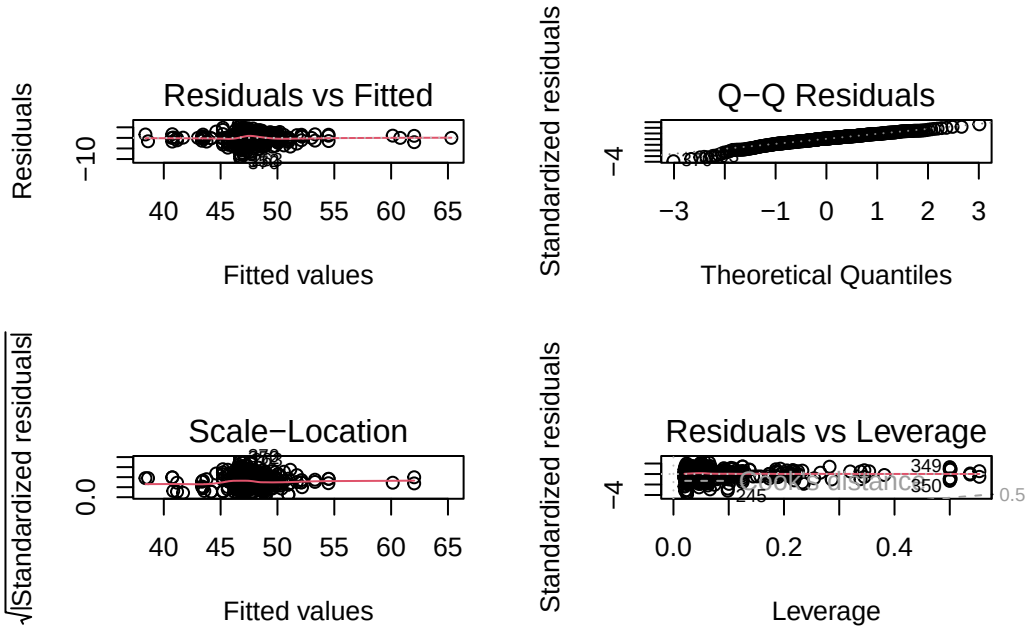



### 3.3 Multiple Linear Regression Models

To better capture the complexity of voter support, we constructed multiple linear regression (MLR) models for both Kamala Harris and Donald Trump. These models incorporate several predictors, including pollster rating (represented by pollscore), transparency score, sample size, and state-level data. By accounting for these factors, the MLR models allow us to control for more variables that influence voter support and provide more accurate predictions.

### 3.3.1 Kamala Harris Model

The MLR model for Kamala Harris takes into account multiple factors that may impact her support across different polls. This includes not just the sample size of the poll, but also how pollster reliability (numeric_grade), transparency of poll data (transparency_score), and state-level polling contribute to predicting voter preferences.

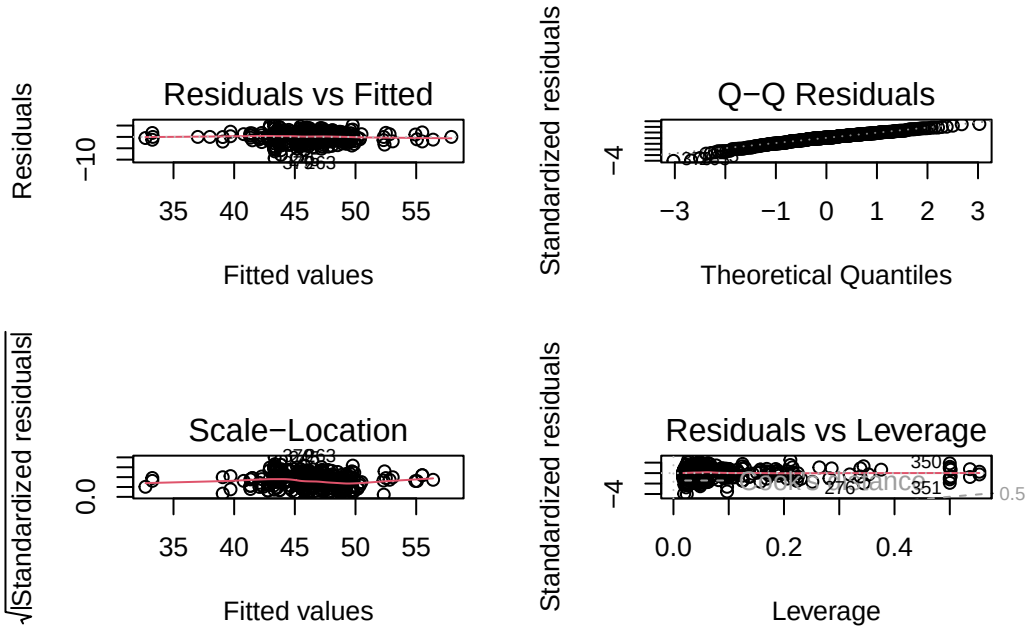



The results of the MLR for Kamala Harris show that several predictors, including state and pollster rating, significantly impact her predicted percentage of voter support. These results suggest that voter preferences for Harris vary widely depending on the state and the credibility of the pollster.

The model diagnostics (such as residual plots, QQ plots) were evaluated to ensure the assumptions of linear regression hold. Figure X shows diagnostic plots, which indicate that the model performs reasonably well in terms of residual behavior and normality. Similarly, we build an MLR model for Donald Trump.

### 3.3.2 Donald Trump Model

Similarly, we constructed an MLR model for Donald Trump using the same set of predictors to assess the factors that influence his support across the country.




The results for Donald Trump reveal that state-level variations and pollster transparency play a significant role in explaining his level of support. The model for Trump also passes diagnostic checks, as shown in Figure Y, indicating that the assumptions of linearity, independence, and homoscedasticity are reasonably met. Check Multicollinearity using Variance Inflation Factor (VIF)

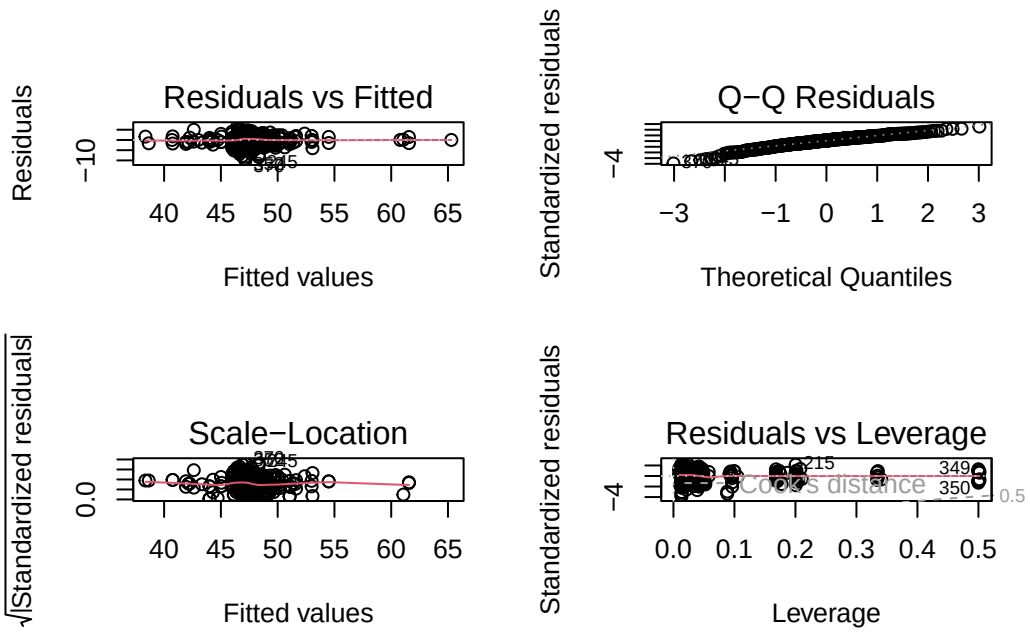
### 3.4 Multicollinearity Check Using Variance Inflation Factor (VIF)

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

```
                      GVIF Df GVIF^(1/(2*Df))
numeric_grade     4.516076  1        2.125106
pollscore         3.049542  1        1.746294
transparency_score 7.163704  1        2.676510
sample_size       1.676196  1        1.294680
state            18.682289 26        1.057915
methodology      83.067087 10        1.247302
```

```
                GVIF Df GVIF^(1/(2*Df))
numeric_grade 2.435023  1        1.560456
pollscore     2.352921  1        1.533923
```

```
sample_size    1.593471  1        1.262328
state          2.075947 26        1.014146
```




Based on the VIF results, we refined the model by removing less significant predictors (such as methodology and transparency_score) to reduce multicollinearity. This improves the model's accuracy and interpretability.

### 3.5 Stepwise Model Selection

To further optimize the MLR models, we performed stepwise model selection, which systematically adds or removes predictors to minimize the Akaike Information Criterion (AIC) and improve model fit.

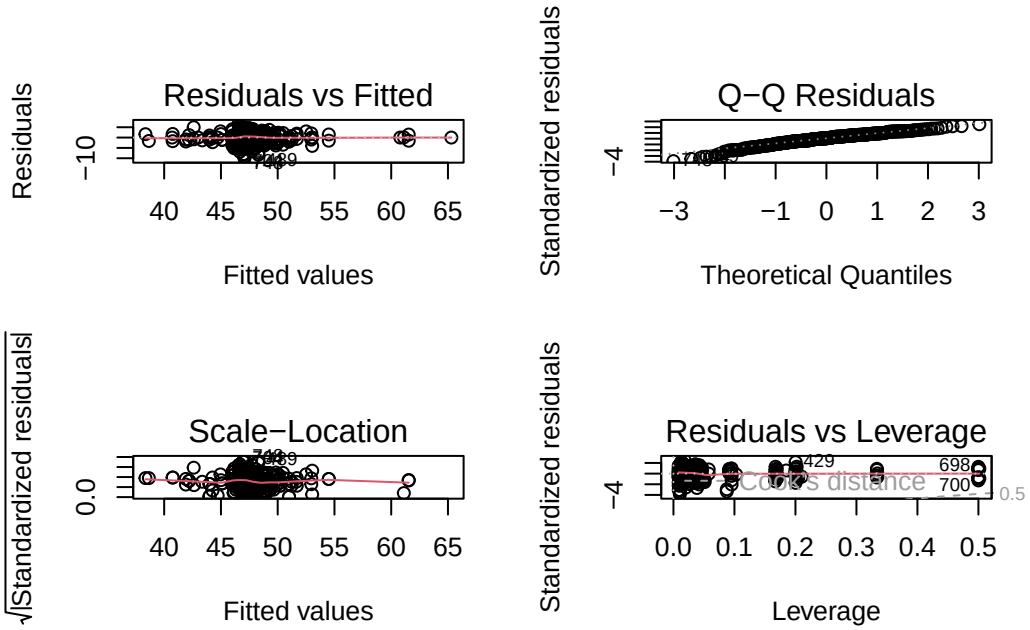
```
Start:  AIC=773.97
pct ~ numeric_grade + pollscore + sample_size + state

                Df Sum of Sq    RSS    AIC
- numeric_grade  1      2.55 2435.5 772.38
<none>                      2432.9 773.97
- sample_size    1     13.21 2446.1 774.08
- pollscore      1     19.93 2452.8 775.15
- state         26   2171.97 4604.9 970.80

Step:  AIC=772.38
```

```
pct ~ pollscore + sample_size + state

                Df Sum of Sq    RSS    AIC
<none>                       2435.5 772.38
- sample_size    1     14.10 2449.6 772.63
+ numeric_grade  1      2.55 2432.9 773.97
- pollscore      1     65.63 2501.1 780.75
- state         26   2188.00 4623.5 970.37
```




The stepwise model selection improved the model by retaining the most significant predictors and eliminating those with little explanatory power.

### 3.6 Final Predictions for Electoral College Votes

Using the final models for both candidates, we predicted the percentage of support in each state and aggregated the results to simulate the Electoral College outcome. The predicted percentage of support for each candidate is used to determine the likely winner in each state.

## 4 Electoral College Prediction

To forecast the winner of the 2024 election, we aggregate the predicted percentages of support from our models for each state and calculate the Electoral College votes. The candidate with 270 or more Electoral College votes is predicted to win the election.

```
[1] "Harris Electoral Votes: 216"
```

```
[1] "Trump Electoral Votes: 147"
```

```
[1] "No candidate reached 270 electoral votes."
```

According to our models, neither candidate secures the 270 Electoral College votes needed to win. Kamala Harris is projected to receive 216 electoral votes, while Donald Trump is predicted to win 147 electoral votes. However, due to the lack of 270 electoral votes for either candidate, the election remains highly competitive, and further developments may shift the balance.

# 5 Discussion

## 5.1 Key Findings

The multiple linear regression (MLR) models developed in this paper provide valuable insights into the factors that influence voter support for Kamala Harris and Donald Trump. Through the analysis of polling data, several key predictors were identified as significant, including state, pollster rating (pollscore), and transparency score.

Kamala Harris The MLR model for Kamala Harris highlights the importance of state-level factors in determining her support. The model shows that voter preferences vary significantly across different states, with certain states (e.g., California and Maryland) showing higher support levels, while others (e.g., Indiana and Missouri) show lower support. Pollster reliability, captured through numeric grades and transparency scores, also plays an important role, suggesting that voters may respond differently based on the credibility of the pollster.

Donald Trump Similarly, the MLR model for Donald Trump reveals that state-level variations are critical to understanding his level of support. The model indicates that Trump's support is more stable across certain states, but there are also notable outliers where his support fluctuates. Pollster characteristics, such as pollscore and transparency score, significantly impact Trump's predicted support, reflecting the importance of poll quality in predicting election outcomes.

Electoral College Forecast Aggregating state-level predictions into Electoral College votes demonstrates the potential competitiveness of the 2024 U.S. Presidential Election. According to the predictions, neither Kamala Harris nor Donald Trump currently secures the necessary 270 electoral votes to win. Harris is projected to receive 216 electoral votes, while Trump is predicted to receive 147 electoral votes. This forecast suggests that the election remains highly uncertain, with several key battleground states likely determining the final outcome.

## 5.2 Model Strengths

One of the main strengths of this analysis is the incorporation of multiple predictors that allow us to account for various factors influencing voter support. By considering state-level data, pollster ratings, and transparency scores, the models provide a more nuanced prediction of voter behavior compared to models that rely solely on national-level polling.

The use of stepwise model selection and multicollinearity checks further enhanced the robustness of the models by optimizing the choice of predictors and ensuring that the models do not suffer from unstable estimates caused by correlated predictors.

Additionally, the aggregation of state-level predictions into Electoral College outcomes presents a more realistic forecast of the election, as the U.S. Presidential election is ultimately decided by electoral votes, not popular votes.

## 5.3 Limitations

Despite the strengths of the models, there are several limitations that should be addressed:

Poll Reliability and Sampling Bias: The accuracy of the predictions depends heavily on the quality of the polling data. Although the model accounts for pollster ratings, polling methodologies can still introduce biases, particularly in states with limited polling data. Sampling errors and non-response biases could skew the results, especially in smaller states or regions with inconsistent polling coverage.

Static Prediction: The model provides a static snapshot of voter support based on current polling data, which may not capture the dynamic nature of voter preferences over time. As election day approaches, voter preferences may shift due to campaign events, debates, or other external factors. Without time-series data, the model may fail to account for these trends.

Unaccounted Variables: Although the model includes important predictors like state, pollscore, and transparency score, other potentially influential factors, such as economic conditions, campaign spending, and voter turnout, are not included in the analysis. These unaccounted variables may introduce inaccuracies in the final predictions. ## Future Research Directions Future work could improve upon this analysis by addressing some of the limitations mentioned above. For instance, incorporating time-series data could allow the model to capture how voter preferences evolve in response to external factors such as campaign events, economic developments, and political endorsements. A dynamic forecasting model that updates predictions as new polls are released would provide more timely and accurate forecasts.

Moreover, integrating other influential variables, such as economic indicators (e.g., unemployment rates, inflation), voter turnout models, and campaign spending data, could enhance the predictive power of the model. Including demographic data (e.g., age, education, income) could also improve the granularity of predictions, especially in battleground states where demographic shifts are critical to election outcomes.

Finally, expanding the model to account for ranked-choice voting in certain states could provide a more accurate forecast in scenarios where third-party candidates or run-off elections play a significant role.

# Appendix

# A Additional data details

# B Model details

## B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

## B.2 Diagnostics

**?@fig-stanareyouokay-1** is a trace plot. It shows… This suggests…

**?@fig-stanareyouokay-2** is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

## C References