Predicting the 2024 U.S. Presidential Election: Trump's Likely Victory Based on Polling Analysis*

A Bayesian Approach Reveals Strong Support for Donald Trump Across Key States, Highlighting Critical Trends in Voter Preferences Leading Up to Election Day

Jimin Lee Sarah Ding Xiyan Chen

November 2, 2024

This paper presents a predictive model for the 2024 United States Presidential Election using state-level polling data. By aggregating high-quality polls, we predict the likely winner between Donald Trump and Kamala Harris in each state. A logistic regression model estimates the probability of Trump winning, with polling percentages as predictor variables. The model's results highlight key battleground states and offer insights into voter dynamics, while also reflecting on the limitations of polling data in forecasting political outcomes. Our findings suggest that Trump holds an adantage in critical swing states, positinoning him as the likely winner of the election.

1 Introduction

This paper develops a predictive model for the 2024 US Presidential Election using state-level polling data (FiveThirtyEight 2024) to forecast the likely winner between Donald Trump and Kamala Harris. We aggregate high-quality poll data to create a logistic regression model that estimates the probability of a Trump or Harris victory in each state. By comparing polling percentages for both candidates, we aim to predict election outcomes based on voter support patterns across the states.

The estimand in this study is the probability that Donald Trump or Kamala Harris wins a given state, based on aggregated state-level polling averages. The binary outcome variable in

*Code and data are available at: https://github.com/jamiejiminlee/2024_US_Elections.git.

our model indicates whether Trump (1) or Harris (0) is predicted to win in each state, and the predictor variables are the average polling percentages for each candidate.

Our model utilizes the Bayesian logistic regression model to predict election outcomes by comparing the average polling percentages for Trump and Harris in each state. The results highlight the geographic distribution of support, with some states clearly favoring one candidate over the other. Swing states emerge as critical battlegrounds, where polling percentages are closely contested and could influence the final election result.

Accurate election predictions provide valuable insights into voter dynamics and help political analysts, campaigns, and the public anticipate electoral outcomes. By focusing on high-quality polling data, our model improves the reliability of predictions and highlights key regions where voter sentiment may shift, ultimately affecting the election.

The results of our model predict a victory for Donald Trump in the 2024 U.S. Presidential Election. Key battleground states, such as Florida, Ohio, and North Carolina, lean toward Trump, emphasizing his advantage in critical regions. Our analysis points to Trump as the probable winner, with swing states playing a decisive role in determining the final outcome.

The remainder of this paper is structured as follows. In Section 2, we describe the data and variables used in the analysis. Section 3 outlines the model setup and estimation strategy. Section 4 presents the results, including the predicted election outcomes and visualizations. Finally, Section 5 discusses the implications and limitations of our findings.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to perform data cleaning and analysis using the datasets we obtained from the Project 538 online database (FiveThirtyEight 2024). The database provides a wealth of information on the current cycle of presidential general election polls. It consists of the polling results from various pollsters such as the New York Times, ActiVote, Morning Consult, etc. Information about the transparency of methodology as well as the quality of the data is present as well, denoted as transparency score and numeric grades, which are significant to our data cleaning process to ensure the quality of data. Polling results are specific to different states in the US, mostly focusing on the six battleground states that are the likeliest to determine the outcome of the elections. Thus, we chose to use states, candidate names, and percentage of support to perform statistical analysis and gather our results.

2.2 Measurement

The measurement of voter support in the dataset is derived from raw polling data obtained from the Project 538 online database (FiveThirtyEight 2024), which serves as a representation of potential election results for the 2024 U.S. Presidential Election. Each entry in the data corresponds to a poll conducted by each pollster, capturing the percentage of respondents who express support for Donald Trump or Kamala Harris. Each pollster has their approach of sampling respondents as well as their methodology of approaching respondents, which affects each pollster's numeric grades and transparency scores.

Therefore, to ensure the quality of the data, only polls from reputable pollsters with a numeric grade of 3 or higher and transparency scores above 6 were used for our paper. These polls measure voter preferences through structured survey questions, and the results are aggregated to create state-level averages for each candidate. The dataset thus transforms general voter sentiment into individual data points that reflect the competitive dynamics between Trump and Harris across different states.

2.3 Outcome variables

The outcome variable in our analysis is the binary variable winner, which represents the predicted winner of the 2024 US Presidential Election in each state. The value of winner is set to 1 if Donald Trump is predicted to win the state, and 0 if Kamala Harris is predicted to win. This binary outcome is determined by comparing the polling support for both candidates within each state. By setting up this binary variable, we aim to forecast which candidate will secure more votes in each state based on the aggregated polling data.

Figure 1 provides an overview of the average support for Trump and Harris across all states, along with the predicted winner for each state based on our model.

Based on the table provided, Trump is predicted to win in the states of Arizona, Florida, Georgia, Missouri, Montana, Nevada, Ohio, and Texas, while Harris is predicted to win in Michigan, Nebraska, North Carolina, Pennsylvania, Virginia, and Wisconsin. However, this does not necessarily determine the overall winner of the election, as U.S. presidential elections are decided by the Electoral College rather than by a simple majority in each state. Each state carries a different number of electoral votes, and a candidate must secure at least 270 electoral votes to win the presidency. To predict the overall winner, it would be necessary to assign each state's electoral votes to the respective predicted winner and calculate the total for each candidate. With the electoral vote counts for each state, we could determine which candidate is projected to win the election.

State	Trump %	Harris %	Winner
Arizona	47.4	45.5	Trump
Florida	53.8	40.5	Trump
Georgia	47.8	45.2	Trump
Michigan	45.5	46.9	Harris
Missouri	54.0	41.0	Trump
Montana	56.5	39.0	Trump
Nebraska	42.0	50.8	Harris
Nevada	46.7	43.4	Trump
North Carolina	47.1	46.8	Trump
Ohio	49.9	44.2	Trump
Pennsylvania	45.7	47.0	Harris
Texas	49.8	44.2	Trump
Virginia	41.0	45.2	Harris
Wisconsin	45.3	48.8	Harris

Figure 1: Summary of 2024 U.S. Presidential Election Polling Averages by State for Trump and Harris, with Indicated Leading Candidate. States with missing data, such as Colorado, are labeled as "NA."

2.4 Predictor variables

The predictor variables used in the model are the aggregated polling percentages for Donald Trump (Trump_pct) and Kamala Harris (Harris_pct). These variables are calculated as the average support for each candidate, using polling data filtered to include only high-quality pollsters with a numeric grade of 3 or higher and transparency scores above 6. These averages reflect the level of support for each candidate across all polls in each state.

To further illustrate the distribution of support, a side-by-side bar graph is provided, comparing the average polling percentages for both candidates across all states. Figure 2 offers a clear comparison of the support levels, helping to visualize the competitive dynamics within each state. The side-by-side comparison represents the percentage of support each candidate has received, based on aggregated poll data from pollsters with high-quality scores. Trump's support is shown in red, while Harris's support is depicted in red.

?@fig-bar-plot illustrates the state-by-state polling data by comparing the average support percentages for Trump and Harris. In each state, the red bar represents Trump's average polling percentage, while the blue bar represents Harris's. Trump shows stronger support in states like **Montana**, **Missouri**, **Florida**, and **Nevada**, where his average polling percentage surpasses that of Harris. Conversely, Harris has a slight edge in states such as **Wisconsin**, **North Carolina**, **Nebraska**, **Pennsylvania**, and **Virginia**. Some states, like Ohio and **Georgia**, exhibit closely matched polling averages, indicating a tighter race. This visualization

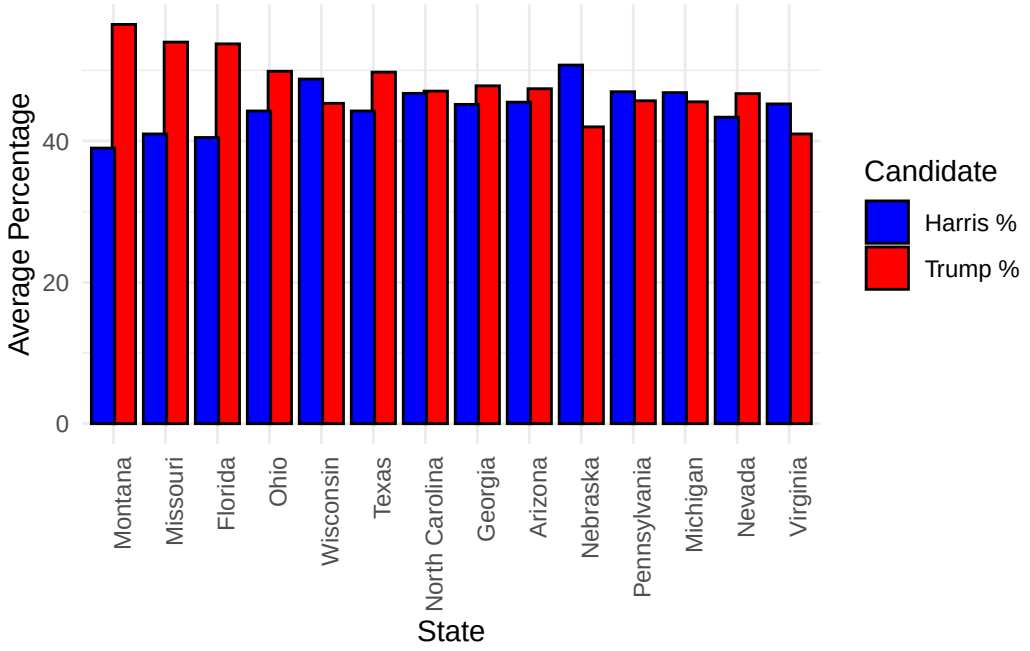


Figure 2: Comparison of 2024 U.S. Presidential Election Polling Averages for Trump and Harris by State. Bars represent the average polling percentages for each candidate, with states missing data, such as Colorado, labeled accordingly.

underscores the regional variation in candidate support across states, reflecting diverse voter preferences that could influence the final outcome based on the Electoral College distribution.

3 Model

The goal of our modeling strategy is to predict the winner of the 2024 US Presidential Election in each state based on polling data, while also assessing the likelihood of Donald Trump or Kamala Harris winning. We use a Bayesian hierarchical modeling approach with natural splines to capture time-dependent trends in support and random effects to account for state-level variability. This strategy enables us to model changes in candidate support over time across states, allowing for more nuanced, probabilistic predictions of each candidate's support levels and overall election outcomes.

Separate models are fit for each candidate's polling support, with state-specific random intercepts and a natural spline for time, allowing the model to capture state-level deviations from the national average as well as time-based changes in support. This hierarchical Bayesian approach enables partial pooling across states, stabilizing predictions for states with limited polling data. Further background details, model specifications, and diagnostics are included in Appendix B.

3.1 Model set-up

The binary outcome variable (y_i) represents whether Donald Trump is predicted to lead in state (i) based on polling averages for both candidates over time. Here, (y_i = 1) indicates a Trump lead in state (i), and (y_i = 0) indicates a lead for Kamala Harris. The model estimates the support levels for each candidate based on polling data using a Gaussian likelihood function, with time trends captured by a natural spline and state-level variation captured through random intercepts.

The predictor variables are (Trump_pct_i), the predicted percentage of support for Donald Trump in state (i), and (Harris_pct_i), the predicted percentage of support for Kamala Harris in state (i). A separate model is fit for each candidate's support over time, capturing the time-varying nature of support with a natural spline function and allowing for state-specific random effects.

The model is specified as follows:

[

$$
\begin{aligned}
\text{Trump\_pct}_i &\sim \text{Normal}(\mu_{\text{Trump},i}, \sigma_{\text{Trump}}) \\
\mu_{\text{Trump},i} &= \alpha_{\text{Trump}} + f(\text{end\_date\_num}_i) + u_{\text{state}}^{\text{Trump}} \\
u_{\text{state}}^{\text{Trump}} &\sim \text{Normal}(0, \tau_{\text{Trump}}) \\
\text{Harris\_pct}_i &\sim \text{Normal}(\mu_{\text{Harris},i}, \sigma_{\text{Harris}}) \\
\mu_{\text{Harris},i} &= \alpha_{\text{Harris}} + f(\text{end\_date\_num}_i) + u_{\text{state}}^{\text{Harris}} \\
u_{\text{state}}^{\text{Harris}} &\sim \text{Normal}(0, \tau_{\text{Harris}})
\end{aligned}
$$

]

where: - (Trump_pct*i*) *and (Harris_pct*i) are the polling support levels for Donald Trump and Kamala Harris in state (i). - (f(end_date_num*i)) is a natural spline function of time (measured as* ***end_date_num****), with 4 degrees of freedom, capturing the non-linear time trends in support. - (* {Trump}) and (*{Harris}) are intercepts for the models, representing the baseline support levels for each candidate. - (u*{state}^{Trump}) and (u_{state}^{Harris}) are state-level random effects, accounting for differences in baseline support across states.

For both models, we use weakly informative Normal priors. The intercepts (*{Trump}) and (* {Harris}) are given Normal priors with a mean of 0 and a standard deviation of 10, reflecting prior uncertainty about the baseline support levels. The time trend coefficients are also given Normal priors with mean 0 and standard deviation 5, allowing for flexibility in capturing the temporal changes in support. State-level random effects for each candidate are modeled with a Normal distribution centered at 0, with the variance estimated from the data.

The model was implemented in R using the `brms` package, which allows for Bayesian inference using `Stan` as the backend. This approach captures the dynamic nature of polling support over time, while accounting for state-specific variability and allowing us to estimate the probability of Trump or Harris leading in each state.

3.2 Model Justification

We expect a positive relationship between each candidate's polling support over time and their likelihood of winning a given state. As Donald Trump's polling percentage in a state increases, his probability of leading in that state rises, while higher support for Kamala Harris increases her probability of winning. By modeling each candidate's support trends over time, we capture shifts in voter sentiment that reflect campaign dynamics, national events, and other time-dependent factors influencing support levels.

A Bayesian hierarchical model was chosen for its ability to handle the complexity of election forecasting, where both temporal trends and state-specific variations play significant roles. Bayesian inference allows us to incorporate prior beliefs and uncertainties about candidate

support, while the hierarchical structure enables partial pooling across states. This is particularly useful for states with sparse data, as the model can "borrow strength" from the national trend, stabilizing predictions.

The inclusion of a natural spline for time allows us to model non-linear changes in support, capturing important shifts in polling trends as the election approaches. Random effects for each state account for baseline differences in support levels, reflecting the unique political climate of each state and reducing the risk of overfitting by allowing state-level deviations from the national average.

This model assumes that polling data reasonably reflects voter intentions and that each state's outcome can be modeled independently of others, though regional trends are partially addressed by state-level random effects. Potential biases inherent in polling data, such as underrepresented demographics or sampling error, could affect our predictions. The model was implemented in R using the brms package, which provides Bayesian inference via Stan, allowing for flexible, probabilistic predictions of each candidate's support across states.

3.3 Model Summary - Harris Model

Table 2: Fixed Effects Summary for Harris Model.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	42.8187095	1.161575	40.4427592	45.056818
nsenddatenumdfEQ41	3.1291491	1.154933	0.7670180	5.317951
nsenddatenumdfEQ42	2.9131143	1.326625	0.2874328	5.513211
nsenddatenumdfEQ43	-0.5551641	2.946899	-6.2604996	5.344961
nsenddatenumdfEQ44	5.0719098	2.564334	-0.2184008	10.087825

EDIT - explain model summary table

3.4 Model summary - Trump Model

Table 3: Fixed Effects Summary for Trump Model.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	44.427447	1.228804	41.8987075	46.798459
nsenddatenumdfEQ41	2.847019	1.193056	0.3373114	5.048196
nsenddatenumdfEQ42	4.288574	1.418572	1.4216286	6.943765
nsenddatenumdfEQ43	3.173093	3.116416	-2.9258070	9.330824
nsenddatenumdfEQ44	7.323299	2.501353	2.2431969	12.264437

EDIT - explain model summary table

4 Results

4.1 National Support Trends Over Time

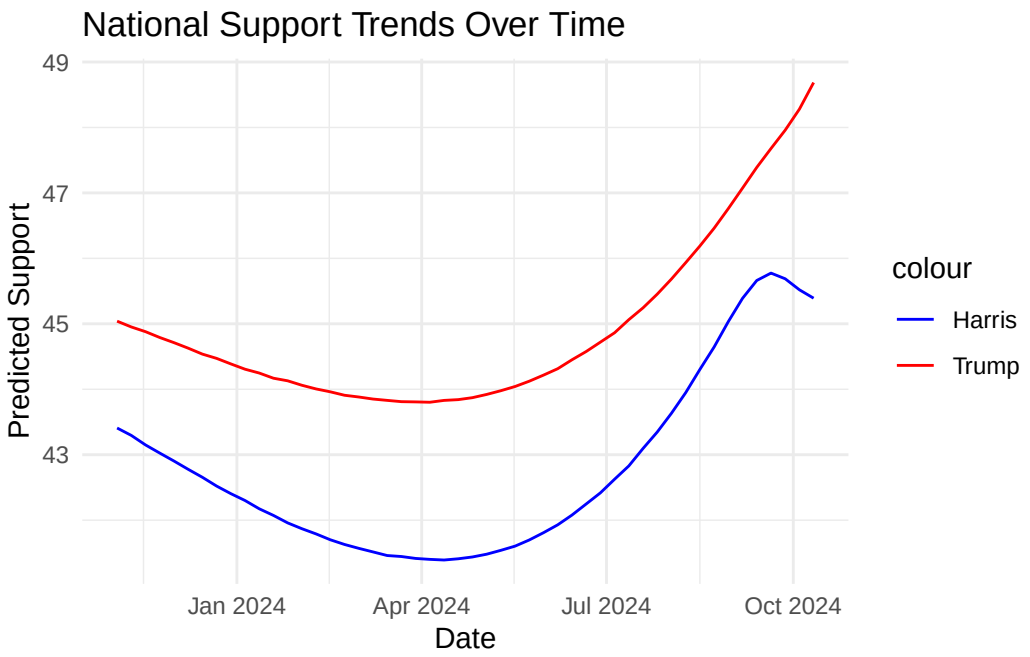


Figure 3: National Support Trends Over Time for the 2024 U.S. Presidential Election. The plot displays the predicted national average support for Donald Trump (red) and Kamala Harris (blue) over time, based on a Bayesian hierarchical model with a natural spline fit. The trends show changes in polling support for both candidates throughout 2024, highlighting key shifts in public opinion as the election approaches.

EDIT - explain results on graph

4.2 State Level Support Trends Over Time

EDIT - explain results on graph

4.3 Probability of Winning by State on Election Day

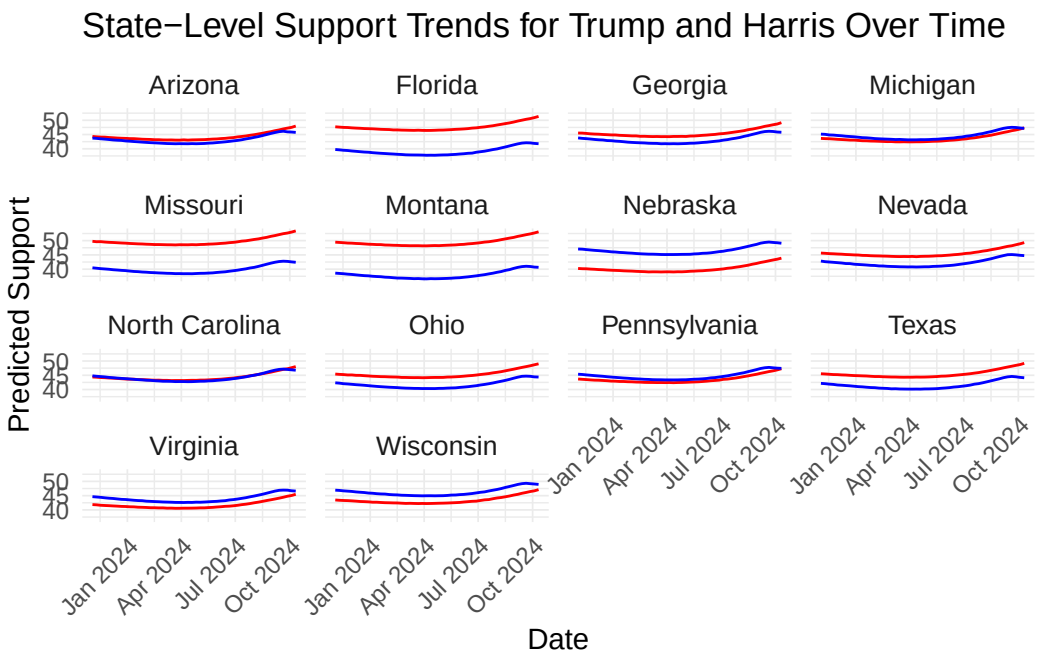


Figure 4: State-Level Support Trends for Trump and Harris Over Time.

Table 4

state	prob_trump_wins	prob_harris_wins
Arizona	86.38	12.07
Florida	99.65	0.38
Georgia	90.47	9.37
Michigan	67.68	32.33
Missouri	99.78	0.23
Montana	99.95	0.02
Nebraska	22.70	78.10
Nevada	94.75	5.23
North Carolina	77.62	22.47
Ohio	95.60	4.53
Pennsylvania	65.92	34.58
Texas	97.07	2.92
Virginia	53.43	46.72
Wisconsin	46.58	52.80

Probability of Trump Winning by State on Election Day.

EDIT - explain table results

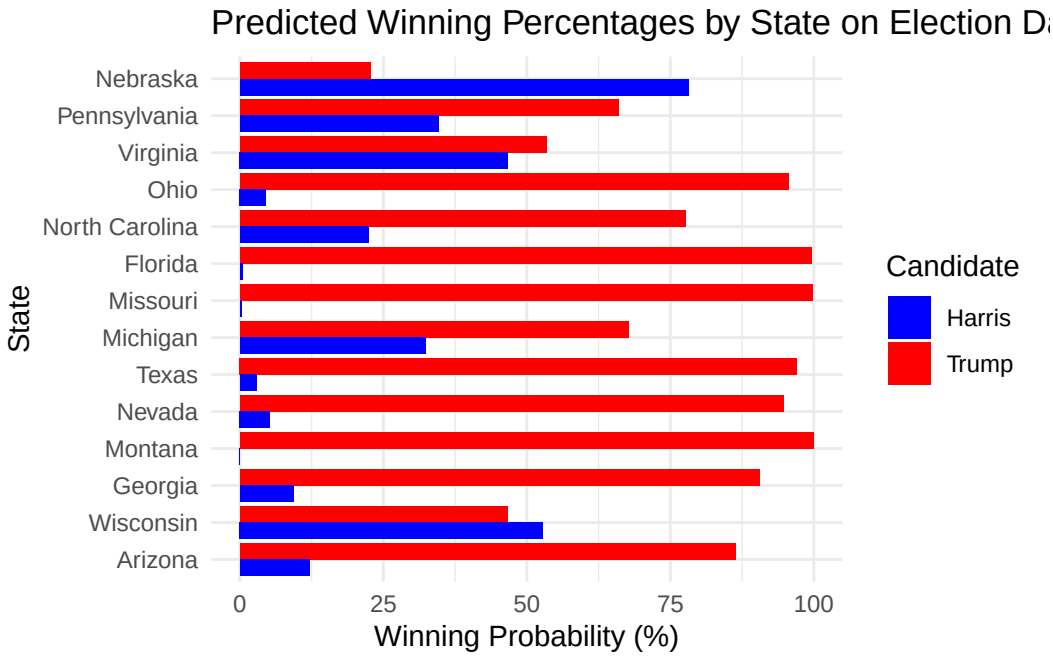


Figure 5: Probability of Trump Winning by State on Election Day.

EDIT - explain graph results

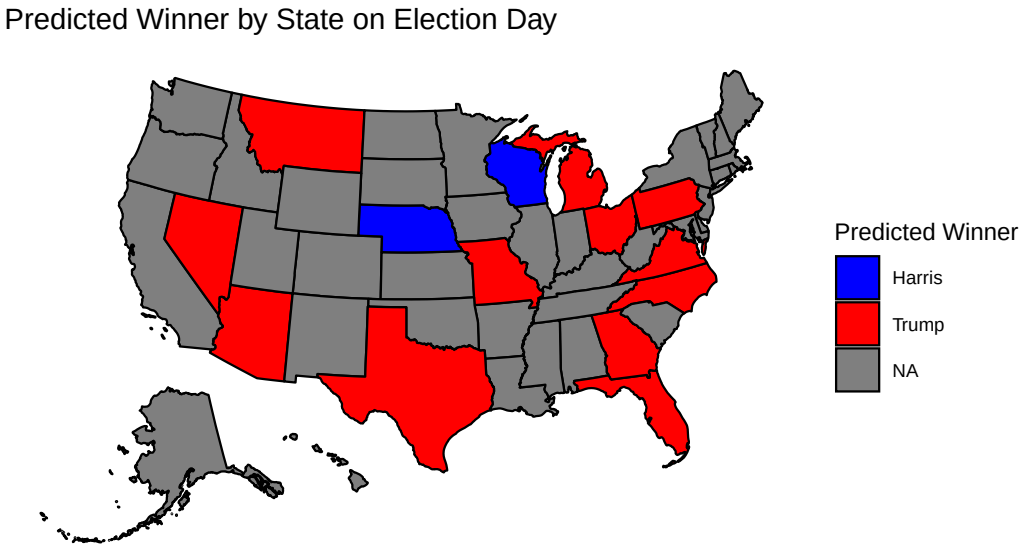
4.4 US Map Showing Predicted Winner by State on Election Day

```
#| echo: false
#| eval: true
#| label: fig-us-map
#| fig-cap: "Predicted Winner by State on Election Day."
#| warning: false

# Add predicted winner column based on `prob_trump_wins`
state_predictions <- state_predictions %>%
  mutate(predicted_winner = ifelse(prob_trump_wins > 0.5, "Trump", "Harris"))

# Plot the US map with predicted winner for each state
plot_usmap(data = state_predictions, values = "predicted_winner", regions = "states") +
  scale_fill_manual(values = c("Trump" = "red", "Harris" = "blue")) +
```

```
  labs(title = "Predicted Winner by State on Election Day", fill = "Predicted Winner") +
  theme(legend.position = "right")
```



EDIT - explain map results

5 Discussion

5.1 Interpretation of Results

Since the electoral college system makes state-level forecasts crucial, and the U.S. operates under a predominantly two-party system, the focus of election predictions should be on battleground states. While many states are strongly aligned with either the Democratic or Republican party, a handful of key battleground states—where neither party has overwhelming dominance—will likely determine the overall outcome of the election. These battleground states, often evenly split in voter sentiment, hold immense significance because their electoral votes can swing the election in favor of one candidate or the other.

The results of our model underscore the critical role that these battleground states will play in the 2024 election. Although many states are predictably stable due to their party loyalty, it is the six battleground states that will serve as the main arena for competition. In these states, where voter preferences are not as firmly entrenched, even minor shifts in public opinion or turnout could lead to vastly different outcomes. As a result, a comprehensive and precise

prediction of voting behavior in these battleground states is far more important than in states where the result is a foregone conclusion.

5.2 Model Performance

Out-of-sample testing results Appendix B.1 demonstrate the effectiveness and reliability of our Bayesian logistic regression model in predicting state-level outcomes for the 2024 U.S. presidential election. Our accuracy test generated a 75% score for the model, indicating our model's high reliability for making predictions about state-level election outcomes. Since presidential elections hinge on state-level wins, a reliable state-by-state prediction would imply a reliable national outcome prediction.

The RMSE value generated for our model is 0.5, which indicates that the model-predicted probabilities are close to the true outcomes. The confusion matrix breaks down the model's predictions into four categories: 1. orrectly predicted Trump wins, 2. Correctly predicted Harris wins, 3. Incorrectly predicted Trump wins, 4. Incorrectly predicted Harris wins. The generated confusion matrix for our model all had values of 0 to 2 for all four categories, indicating a balanced prediction between the two candidates, and suggesting the reliability of our model. Out-of-sample testing was conducted to test the reliability of our model, with further details provided in Apppendix B.1.

5.3 Weaknesses and Next Steps

While our model provides valuable insights into the electoral landscape, it faces limitations due to missing data for certain states, which results in some states being left blank in our predictions. Although our primary focus is on understanding the dynamics within battleground states, the lack of information from other states can lead to incomplete predictions which may obscure broader trends affecting voter sentiment.

To address these weaknesses and improve the prediction, we propose several next steps: actively seeking additional data sources to fill gaps, including polling data and historical records; conducting sensitivity analyses to understand the impact of missing data on predictions; and analyzing existing trends within battleground states for more insights.

Another weakness of our model is the lack of consideration for the national vote, which may impact our election predictions. The national vote refers to the total number of votes cast by citizens across the country during an election, representing the support for each presidential candidate. Although the national vote does not directly determine the outcome of the election due to the Electoral College system, it can still provide valuable insights into voter sentiment and trends. By not considering the national vote into our analysis, we risk missing potential influences on voting behavior and electoral outcomes.

Appendix

A Additional Data Details

Given the nature of the presidential election, the electoral college is more impactful and influential on the possible outcome compared to the nationwide popular votes. Therefore, the majority of the pollsters choose to focus on the battleground states to perform their polling, to gain more insight into the public sentiment specific to certain states. For instance, Siena/NYT mentions that they have in-state interviewers who call their respondents in states such as Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin.

This could explain the discrepancies in our data, where there are many missing values from states other than the battleground ones. Furthermore, many pollsters focus on state-wide polls instead of nationwide polls given the nature of the election, resulting in fewer data points or missing data for nationwide poll results. Even though the national popular vote and the results in the electoral college don't line up a lot of the time, the popular votes can still provide information about issues and opinions that are shaping the election as a whole. Therefore, even if we are focusing on selective states, results from other states and popular votes still remain relevant to outcome prediction.

During our data cleaning process, we added constraints of high numeric grades and high transparency scores, leading to fewer data points for us to build the model on. A numeric grade is a numeric rating given to the pollster indicating their quality or reliability, with a highest rating of 3.0. The transparency score is a score reflecting the pollster's transparency about their methodology where the highest score is 10. Therefore, to ensure high-quality and reliable data to build our model on, we filtered pollsters with numeric grades of 3.0 and transparency scores of 5 or higher. This is one of the reasons why there are no data points available for Kamala Harris for the state of Colorado, represented by the empty row Figure 1.

Furthermore, in the full raw data set from 538 (FiveThirtyEight 2024), for the state of Colorado, even without any constraints for numeric grade or transparency scores, there are only 3 data points for Kamala Harris. After reviewing the datasets, we found that except for the 3 available data points, all the rest of the polls ended in June 2024, which is a month earlier than when Biden dropped out of the race before Harris entered the presidential election race. It would not be possible that there are data points available for Harris if she was not participating in the race yet. The state of Colorado is a blue state, with the Democrats winning the state in every presidential election since 2008. Since our paper focuses on mainly the battleground states to predict election outcomes, limited data from a blue state like Colorado would not be too influential on our results.

B Model details

B.1 Diagnostics

Out-of-sample testing was conducted to test our model's reliability in predicting potential presidential election outcomes. We train our model on a training set of our data and then test its performance on the test set of our data, we evaluate how well the model generalizes to new data that it was not trained on.

This process helps reveal whether the model is over-fitting to the training data or truly capturing the underlying patterns that apply more broadly. The test produces an accuracy score, RMSE, and a confusion matrix, each measuring different aspects of the model's prediction. Accuracy is the proportion of correctly predicted outcomes compared to the total number of predictions made in the test set. The testing was implemented using the following packages in R (R Core Team 2023): caret (Kuhn et al. 2023), rstanarm (Gabry et al. 2023), ggplot2 (Wickham et al. 2023b), dplyr (Wickham et al. 2023a).

B.1.1 Model Diagnostics Result

?@fig-barplot-diagnostics displays the performance of the Bayesian logistic regression model used to predict the outcome of the 2024 presidential election. The model's accuracy is approximately 75%, indicating that 75% of the test set's state-level outcomes were correctly predicted. The RMSE (Root Mean Squared Error) is about 0.5, showing the average error in predicting the winning candidate across states.

This suggests the model performs reasonably well, but there is some error in predicting the exact outcome in certain states, which may be due to variations in polling data or other unaccounted factors.

B.1.2 Confusion Matrix

?@tbl-confusion-matrix displays the model's predictive performance in distinguishing between state-level wins for Trump (coded as 1) and Harris (coded as 0).

In this case, the model correctly predicted 12 Harris wins and 15 Trump wins (true positives and true negatives). However, there were 5 instances where the model incorrectly predicted a Trump win when Harris won (false positives), and 3 instances where it predicted a Harris win when Trump won (false negatives). This breakdown helps assess not only the overall accuracy but also how well the model distinguishes between close contests in different states, which is critical for predicting the outcome of the 2024 presidential election.

C Pollster Methodology Overview and Evaluation

The following information about the Siena College/New York Times poll methodology is based on the details provided in the article "How The Times/Siena Poll Is Conducted" by the New York Times (The New York Times 2023).

The Siena College/New York Times poll for the 2024 presidential election uses a robust sampling approach to ensure accuracy and relevance. The poll applies a stratified dual-frame sample, drawing from both land-lines and cell phones. Each poll is conducted by phone using live interviewers at call centers based in Florida, New York, South Carolina, Texas, and Virginia. The sampling population is all registered voters who live in the six battleground states: Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin.

Given the nature of the presidential elections, the decision is based on the electoral college instead of the popular vote, thus the pollster focuses on polling on the states that are likeliest to decide the outcome of the race. The voters' information is taken from the L-2 voter file, which includes details such as voter registration and history of participation in previous elections.

To refine their sample, Siena College adjusts the data by key demographic and political variables such as region, race/ethnicity, party affiliation, education, and voting patterns from the 2020 election. These adjustments help ensure the sample accurately represents the electorate. Additionally, they use a "likely voter screen," which combines self-reported likelihood of voting with historical voter behavior to estimate how likely respondents are to vote in the upcoming election.

At every step of the survey, Sienna/NYT uses the information in the data to try to ensure that they have the right number of Democrats and Republicans, young people and old people, the right ratio of people with different income levels, and a diverse mix of different races and regions. Once the survey is complete, they compare their respondents to the voter file and use a process known as weighting to ensure that the sample reflects the broader voting population. This combination of historical data and weighted adjustments ensures the poll is designed to predict election outcomes as accurately as possible.

The New York Times/Siena College pollsters handle non-response by using weighting and adjusting their sample to correct for any biases that might emerge due to individuals not responding. Specifically, they adjust on multiple demographic and political variables such as age, gender, education, race/ethnicity, and party affiliation. They also account for variations in voter likelihood and previous voting patterns, ensuring that the sample represents the likely electorate as accurately as possible.

This helps reduce the potential bias from non-response, especially since certain demographic groups or political affiliations may be less likely to respond to polls. By rebalancing the sample, they ensure that even if some groups have lower response rates, their representation in the final poll results aligns with what is expected based on historical trends and current voter enthusiasm.

D Idealized Methodology

Building on our discussion of the New York Times/Siena College Poll, we now present an idealized methodology that aims to enhance the accuracy and reliability of our predictions. This approach incorporates best practices and innovative techniques to ensure a comprehensive prediction of voter behavior in the upcoming 2024 U.S. presidential election. To demonstrate our idealized methodology, we generated a survey on the 2024 US Presidential election, which can be accessed via the the URL provided in (Google 2024) under **?@sec-references**.

D.1 Sampling Approach

Using data from voter files, which contain demographic information about registered voters, we aim to ensure that our sample accurately represents the population. To achieve this, we will use stratified random sampling. Given the importance of state-level forecasts due to the Electoral College system, we will first stratify by state.

Within each state, we will apply stratified sampling again, dividing the population into 6 groups and selecting 100 random samples from each stratum within those groups, ensuring our sample captures key demographic critical for predicting voter behavior. The groups and their respective strata include age groups (18-29, 30-44, 45-64, and 65+), gender (Male, Female, and Other), and race/ethnicity (White, Black, Latino, Asian, and Other). Additionally, household status will be categorized as either renting or owning, with home ownership serving as a proxy for wealth. Voting history will be classified into those who voted in the previous election and non-voters, while party registration will include Democrats, Republicans, and Independents.

For the recruitment of participants, we will begin by sending survey invitations to our selected voters via email, offering a $20 incentive to encourage participation. To improve response rates and reduce non-response bias, we will send a reminder email three days later. This approach will help minimize expenses to some extent. If we do not receive a response within a week (7 days), we will follow up by phone to reach those who did not respond to the email invitation.

D.2 Data Validation

After completing the survey, we will adjust the data through weighting to ensure the sample accurately reflects the broader population for predictive purposes. More weight will be assigned proportionally to each stratum based on its size, and additional weight will be given to respondents from strata that are less likely to participate in surveys. And to avoid duplicate responses, each voter will be assigned a unique ID with only the first response from each ID being retained. To ensure the selected voters are completing the survey correctly, we will track responses in real-time using a centralized system. Additionally, we will include validation questions in the survey to catch careless or fraudulent responses.

References {sec-references}

FiveThirtyEight. 2024. *FiveThirtyEight: 2024 US Presidential Election Polls*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Gabry, Jonah et al. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Google. 2024. *Google Form: [2024 US Presidential Election Survey]*. https://docs.google.com/forms/d/1S6aN5Q82orjKYJI4i7001sPyn5jhRkgNIkI9teiQq_o/edit.

Kuhn, Max et al. 2023. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

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

The New York Times. 2023. "How the Times and Siena College Poll Was Conducted." https://www.nytimes.com/article/times-siena-poll-methodology.html.

Wickham, Hadley et al. 2023a. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

——— et al. 2023b. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.