Forecasting the 2024 U.S. Presidential Election: Kamala Harris' Projected Victory*

A Bayesian Analysis of Polling Data from Key Battleground States"

Jimin Lee Sarah Ding Xiyan Chen

November 3, 2024

This paper presents a predictive model for the 2024 United States Presidential Election, focusing on critical battleground states. Utilizing a Bayesian approach with state-level polling data, we estimate the winning probabilities for candidates Donald Trump and Kamala Harris. Our findings reveal that Harris is projected to win in six out of seven key battleground states, including Michigan, Nevada, and North Carolina, with predicted winning percentages surpassing 50%. Given this significant advantage, we predict Kamala Harris to be the likely winner of the election. This analysis underscores the importance of battleground states in shaping electoral outcomes and provides valuable insights into voter sentiment as Election Day approaches.

Table of contents

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

1 Introduction

The 2024 United States Presidential Election is shaping up to be a critical event, particularly as battleground states emerge as pivotal areas influencing the electoral outcome. As candidates Donald Trump and Kamala Harris intensify their campaigns, understanding voter preferences in these key regions becomes essential for predicting the election results. This paper utilizes state-level polling data from Project 538 (FiveThirtyEight 2024) and the statistical programming language R (R Core Team 2023) to build a predictive model that estimates the likelihood of victory for each candidate in several key battleground states, including Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin.

The primary estimand in this study is the probability that either Donald Trump or Kamala Harris will win each battleground state, based on aggregated polling averages. Our model employs a Bayesian hierarchical approach with natural splines to capture the nuances of voter sentiment over time, allowing for the estimation of winning probabilities across these crucial states. By comparing the predicted winning probabilities for both candidates, we can ascertain the likely winner based on the prevailing voter support patterns.

Our analysis reveals that Kamala Harris is projected to win in six out of the seven battleground states examined, indicating a favorable voter sentiment in states such as Michigan and Nevada. Conversely, Donald Trump shows a competitive stance in Arizona, where he is also predicted to secure a slight edge. The calculation of the predicted winner is determined by comparing the estimated probabilities for each candidate, which underscores Harris's overall advantage in the election.

Understanding these predictions is vital as it offers insights into the dynamics of voter behavior and the potential implications for election outcomes. This information can help political analysts, campaign strategists, and the public anticipate how these battleground states may influence the overall results. Accurate predictions not only inform campaign strategies but also highlight regions where voter sentiment may shift in the lead-up to Election Day.

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

The dataset of the 2024 Presidential Election cycle is obtained from Project 538 (FiveThirtyEight 2024). The dataset is periodically updated throughout the presidential election campaign to reflect current and most up-to-date polling results. It consists of polling information from various pollsters such as the New York Times, ActiVote, Morning Consult, etc. For every poll conducted, relevant variables such as state, start_date, end_date, transparency_score, candidate_name, and percentage of support (pct) for the appropriate candidates are included. The dataset compiles data from all major pollsters in the US and provides us with an understanding of the current state of the presidential election.

The present analysis focuses on Donald Trump and Kamala Harris, the two leading candidates in the current presidential election race. For all the following variables mentioned, we collected observations for both of the candidates. We collect observations from the dataset where the 'candidate_name' is Trump and Harris. We are interested in the support for both candidates over time, for each state, therefore, we collect observations of 'end_date', and 'state', as well as the electorate support for each candidate is denoted as 'pct'. Furthermore, since there are many pollsters present in the dataset, we extracted pollsters with high-quality data that is reliable, which is denoted as having a high 'numeric_grade' and 'transparency_score', details of data cleaning are further explained in Appendix A.

2.2 Measurement

In the current presidential election cycle, Americans' opinions on voting for their preferred candidate are transformed into data points through a series of steps that involve surveying, data processing, and structuring responses. Pollsters then store these responses in structured databases for analysis. In the dataset obtained from Project 538 (FiveThirtyEight 2024), each entry in the dataset corresponds to a poll conducted by each pollster, capturing the percentage of respondents who express support for each candidate. Each pollster has different approaches to sampling respondents, recruiting respondents, as well as questionnaire types, which affects each pollster's reliability and data quality.

Pollsters begin by selecting a sample from the registered voter population, often using voter files with demographic and contact information. They then apply methods of stratified random sampling to ensure the sample reflects the electorate's diversity in terms of age, gender, race, education, voter behavior, etc.

Then, pollsters reach out to voters through various channels, such as phone calls, emails, text messages, or online panels. These are called recruiting methods and usually, each of them has its trade-offs. If the respondents are connected through any of these channels, they are asked a series of questions either by a live interviewer or a preset of questions.

The types of questions often include:

- "If the election were held today, would you vote for [Candidate A] or [Candidate B]?"
- "How strongly do you support [Candidate]?" (e.g., on a scale of 1 to 10, or strongly, somewhat, undecided)
- "Do you support [Policy A] proposed by [Candidate A] or [Policy B] proposed by [Candidate B]?"

Responses are recorded digitally, either by survey software or directly into a database by interviewers, capturing both the answers and relevant demographic data. Each response becomes a data point in the database. For example, if a respondent prefers Candidate A, their answer may be recorded as a binary variable (e.g., "1" for Candidate A, "0" for Candidate B). Along with candidate preference, pollsters capture demographic and behavioral details, such as age, gender, race, education, income, voter history, and party registration.

Each respondent's answers and demographic details are grouped and saved in a structured database. Pollsters then clean the data by checking for inconsistencies, removing incomplete responses, and addressing non-responses. Pollsters also perform weighting to ensure the representativeness of data collected, details of weighting are further explained in {appendix 1}.

Once cleaned and weighted, data points are aggregated to determine overall candidate support, often producing metrics such as the percentage of respondents supporting each candidate by state or other specific demographics. Each pollster would have a database with all the weighted and aggregated data, and Project 538 presents a compilation of the databases obtained from various pollsters.

2.3 Variables

The collected dataset contains several key variables relevant to the analysis:

- **State**: Indicates the state where the polling took place, critical for understanding regional support dynamics.
- **Polling Date**: The date when the poll was conducted, which helps in analyzing trends over time.
- **Pollster**: The organization conducting the poll, providing insight into the reliability and methodology of the polling data.
- **Numeric Grade**: A score assigned to each poll based on its methodology and historical accuracy, which assists in filtering out less reliable polls.
- **Transparency Score**: A measure indicating how transparent the pollster is regarding their methodology, further enhancing data credibility.
- **Candidate Name**: Identifies the candidate being polled, specifically focusing on Donald Trump and Kamala Harris.
- **Percentage of Support (pct)**: Represents the percentage of respondents supporting each candidate, which is the primary outcome variable of interest.

To ensure the quality of the analysis, only those polls that meet specific thresholds were included - for more details, refer to Appendix A.

2.3.1 Outcome Variable

As our focus is to analyze and predict voter support for Kamala Harris and Donald Trump during the current presidential election cycle, the primary outcome variables of interest are the percentages of support for each candidate, denoted as 'pct'.The pct is categorized by candidate and scaled to a proportion by diving by 100. This variable represents the proportion of voter support for each candidate based on polling data. Through understanding the percentage of support (pct) for each candidate, we can observe and infer the candidate that has a leading advantage in winning the presidential election race.

To give an overview of the outcome variable, divided per candidate, we provide summary statistics of this key variable, highlighting the characteristics of the polling support for each candidate across the battleground states.

Table 1: Summary Statistics for Trump and Harris Polling Data - Presents key summary statistics, including total polls, average support percentages, and the maximum and minimum polling levels for both candidates across the specified battleground states. All values are rounded to three decimal places, providing an overview of the polling landscape as of the selected election date.

Statistic	Value
Total Polls	107.000
Average Trump Support	47.801
Max Trump Support	51.700
Min Trump Support	43.000
Average Harris Support	47.999
Max Harris Support	51.200
Min Harris Support	43.600

For instance, Table 1 reveals a total of 107 polls conducted, calculated as the total count of rows in our cleaned dataset. 'Total polls' reflects the total number of unique poll entries after filtering for only battleground states, polls with sufficient data quality, and polls that report support percentages for both Trump and Harris. The 107 polls conducted display an average support level of 47.801% for Trump and 47.999% for Harris. This suggests that based on the dataset, Trump has a leading advantage in winning the presidential election race by having a slighty higher average level of support than Harris. The highest percentage of support for Trump in the 107 polls was 51.700% while the lowest recorded support was 43%. For Harris, the highest percentage of support also reached 51.200%, with a minimum of 43.600%. These

statistics illustrate the competitive landscape of the presidential election race, without a clear leading candidate in the current race.

It is important to note that since 'pct' stands for percentage, there may be a misconception that the values for Trump and Harris will always add up to 100%. For instance, in the summary statistics presented in Table 1, the average support for Donald Trump is 47.801%, while the average support for Kamala Harris is 47.999%. This results in a combined average support of 95.800%, indicating that there is a significant portion of respondents who are either undecided or supporting other candidates. This highlights the competitive nature of the election and emphasizes the necessity of examining not only the percentages attributed to the main candidates but also the context of voter preferences that could influence the final outcome.

2.3.2 Predictive Variables

In this section, we outline the two key predictor variables used in our analysis: 'state', which captures the geographical context of the polling data, and 'end_date', which reflects the timing of the polls relative to the election, allowing us to assess trends in voter support over time.

2.3.2.1 State

The state ('state') variable indicates the U.S. state where the poll was conducted or targeted. In the context of the U.S. presidential election, the outcome is frequently determined by several key battlegroud states. In the current 2024 election cycle, there are seven battleground states that are pivotal in the contest between Kamala Harris, representing the Democratic Party, and Donald Trump, representing the Republican Party. To enhance our model and refine our predictions, we focused exclusively on these seven battleground states. The polling results from each of these states play a crucial role in predicting electoral outcomes, as the percentage of support is evaluated on a state-by-state basis.

The distribution of the state variable is visualized in Figure 1 and indicates that some states, such as Pennsylvania, North Carolina and Wisconsin have more polling observations compared to others, like Nevada. This variability reflects differences in polling frequency and interest among these battleground states. States with more polling observations provide a larger sample size, potentially leading to more stable and reliable predictions for those states in the model.

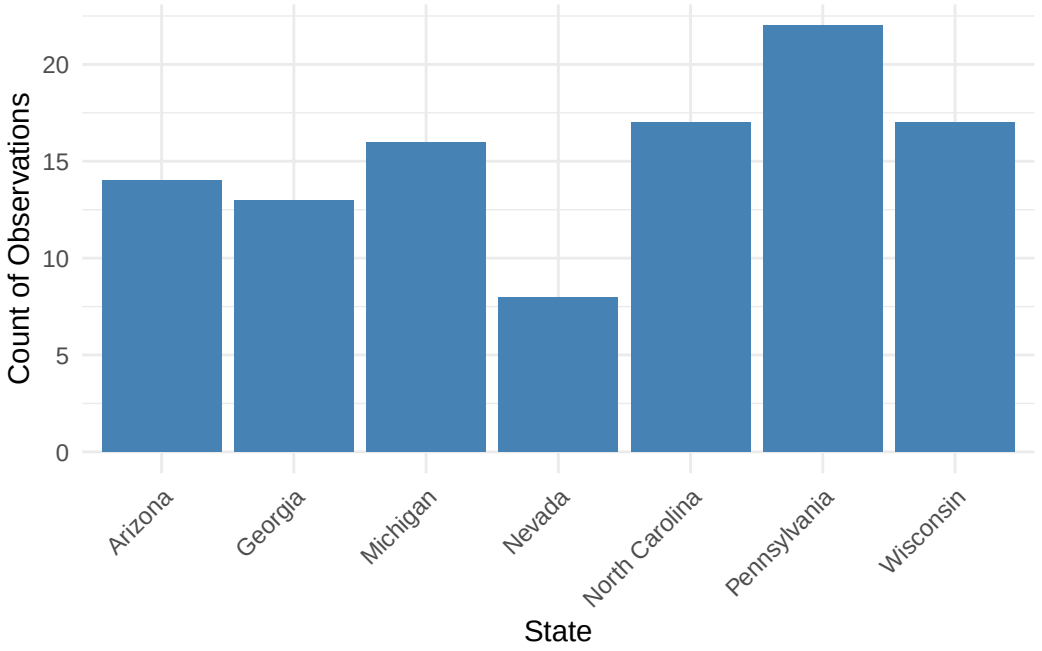

Figure 1: Distribution of Observations for Each State in Polling Data. Illustrates the number of polling observations across key battleground states in the lead-up to the 2024 United States Presidential Election. Notably, Pennsylvania has the highest count of observations, suggesting a more robust polling presence, while Nevada shows fewer observations.

2.3.2.2 End Date

End date ('end_date') represents the date the poll ended, essentially the date the poll stopped collecting data from electorates. In the context of our paper, the end date of each poll serves as one of the predictive variables and is used to assess changes in voter support over time. Including end date as a variable in our model can help us identify patterns in voter support as Election Day approaches, potentially reflecting the impact of campaign events, debates, or other political developments. Figure 2 displays the distribution of polling end dates from July 2024 to October 2024.

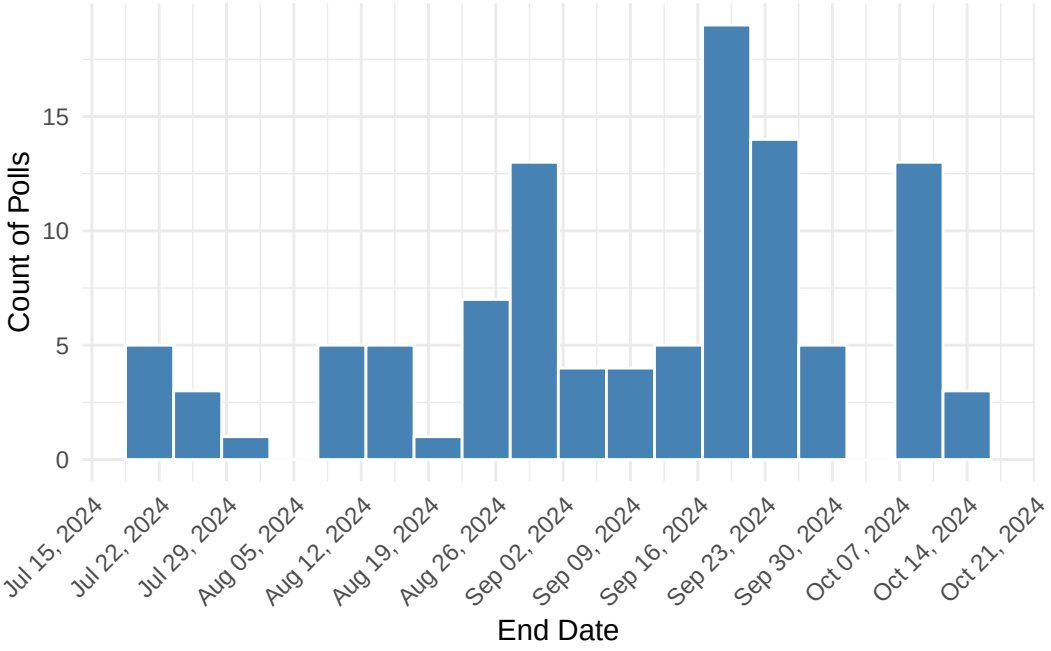

Figure 2: Distribution of Polling End Dates. This histogram displays the frequency of polling end dates for key battleground states leading up to the 2024 United States Presidential Election, highlighting peaks in polling activity around late September.

We filtered for the earliest end date as July 2024 since Harris officially announced her presidential campaign on July 21, 2024. Therefore, any polls that has an end date before July would not contain polling results for Harris. The data from (FiveThirtyEight 2024) is last updated on October 15, 2024, thus the latest end date being October 21, 2024. Furthermore, Figure 2 also reveals that polling activity is not evenly distributed over time. There are clear peaks where the number of polls conducted is higher, notably in early September, late September, as well as early October. This concentration of polls likely corresponds to period of heightened political interest or significant campaign events that prompt increased polling.

2.3.3 Other Variables

This section details additional variables that provide essential context for our analysis, including numeric grade, which evaluates the quality of the polls; transparency score, indicating the openness of poll methodologies; pollster, identifying the organizations conducting the polls; and start date, which denotes the initiation of the polling period for each survey.

2.3.3.1 Numeric Grade

Numeric Grade ('numeric_grade') represents the quality of each poll, as rated by (FiveThirtyEight 2024), based on factors such as methodology, transparency, and historical accuracy. In the context of our paper, numeric grade allows us to weigh the reliability of different polls. By including numeric grade as a variable in our model, we aim to account for variations in poll quality, ensuring that polls with higher ratings (indicating higher reliability) contribute more significantly to our predictions of voter support. Figure 3 displays the distribution of numeric grades for the polls included in our analysis, ranging from 2.5 to 3.0. Polls with higher grades are generally seen as more methodologically rigorous and transparent, which helps improve the reliability of our model's predictions. The distribution shows several peaks, indicating that certain grade levels (around 2.7, 2.8, and 3.0) are more common among the polling data we collected.

This variability in numeric grade highlights the diversity in polling quality within our dataset. The presence of peaks suggests that some pollsters consistently receive higher ratings, likely due to their use of more rigorous methods or transparent reporting standards. Conversely, the valleys in the distribution may indicate fewer polls from sources with lower numeric grades, reflecting a potential tendency for lower-quality polls to be less prevalent or less frequently conducted. By accounting for numeric grade, we can improve the accuracy of predictions by emphasizing data from more reliable sources. This consideration is crucial, as polling quality can impact the stability of the model's forecasts, especially in closely contested states where polling accuracy is essential for reliable predictions.

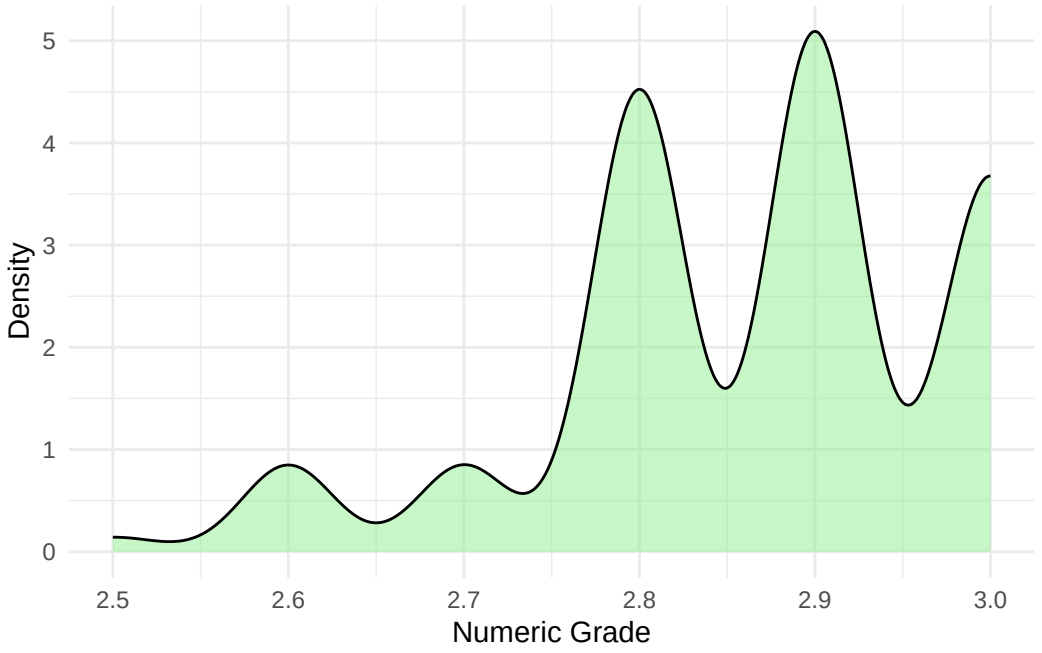

Figure 3: Density plot illustrating the distribution of numeric grades assigned to pollsters within the polling data as the 2024 United States Presidential Election approaches. The graph highlights the frequency of numeric grades, indicating key concentrations that suggest variability in pollster quality.

2.3.3.2 Transparency Score

Transparency Score ('transparency_score') represents the degree of methodological transparency provided by each pollster, rated on a scale from 0 to 10 by (FiveThirtyEight 2024). This score reflects the extent to which pollsters disclose details about their data collection and analysis methods, such as sample size, weighting, and margin of error. In the context of our paper, transparency score is an essential variable for evaluating the reliability of polling data, as higher transparency scores generally indicate a greater level of methodological rigor and openness. Figure 4 displays the distribution of transparency scores across the polling data for key battleground states. The plot shows that transparency scores cluster heavily around ratings of 7 and 9, suggesting that a majority of polls in our dataset come from sources that maintain a relatively high level of transparency. There are smaller peaks at lower scores (5 and 10), indicating that while some pollsters offer limited transparency, a few polls also achieve perfect transparency scores.

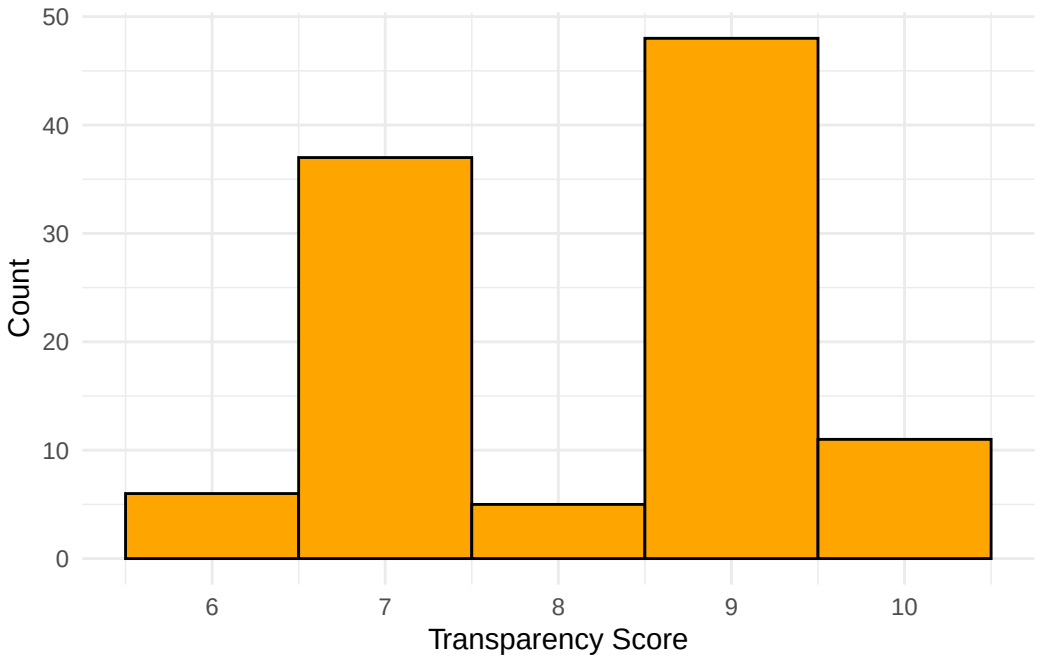

Figure 4: Histogram displaying the distribution of transparency scores among pollsters in the polling data leading up to the 2024 United States Presidential Election. The plot indicates the frequency of various transparency scores, highlighting a concentration of scores around 7 and 9, which suggests a predominance of high transparency among the assessed pollsters.

This distribution is important for understanding the credibility of our data sources. Polls with higher transparency scores are more likely to provide consistent and reliable results, as they openly disclose their methodologies, allowing us to assess potential biases or limitations in

their data. In contrast, polls with lower transparency scores may lack sufficient methodological detail, making it challenging to gauge their accuracy. Incorporating transparency scores into our model helps differentiate between high- and low-transparency polls. This adjustment is crucial as transparency can influence the stability of the model's predictions, especially in closely contested states where the reliability of polling data may impact the accuracy of our forecasted election outcomes.

2.3.3.3 Pollster

Pollster ('pollster') represents the various organizations responsible for conducting polls across key battleground states in the lead-up to the 2024 United States Presidential Election. In the context of our paper, the diversity and frequency of polling by different pollsters are essential for understanding the reliability and variance in the data. Each pollster may use different methodologies, sample populations, and data collection techniques, which can affect the results and introduce variability in polling outcomes.

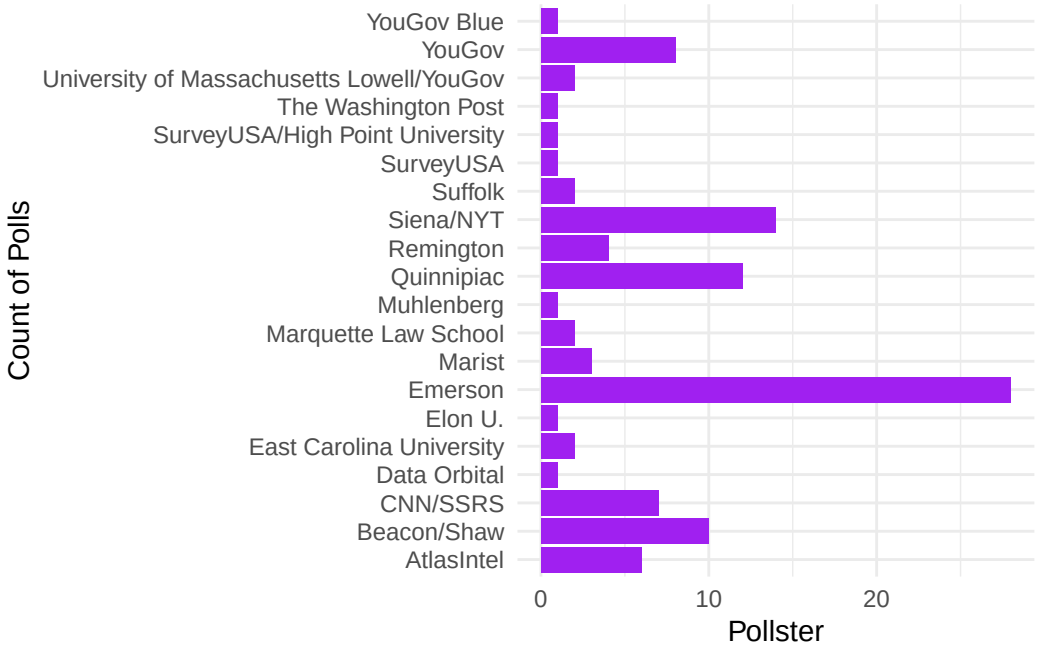

Figure 5: Horizontal bar chart displaying the distribution of polls conducted by various pollsters in key battleground states for the 2024 United States Presidential Election. The chart highlights the frequency of polling activities across different organizations, with Marist and Emerson standing out as the most prolific pollsters.

Figure 5 displays the count of polls conducted by each pollster, illustrating the distribution of polling efforts across multiple organizations. Notably, Emerson is the most frequent pollster, conducting the highest number of polls, followed by Quinnipiac and Siena/NYT. Other

pollsters, such as YouGov, CNN/SSRS, and Marist, also contribute a significant number of polls, whereas some organizations like The Washington Post and SurveyUSA have comparatively fewer entries in the dataset.This distribution is crucial for evaluating the consistency and breadth of the polling data. Pollsters with a higher count of polls, like Emerson, contribute substantially to the dataset and can have a greater influence on the model's predictions, particularly if their methodology or sample selection differs from others. On the other hand, pollsters with fewer entries may provide valuable but limited data points, potentially adding unique perspectives but with less influence on overall trends.

2.3.3.4 Start Date

Start date ('start_date') represents the date when each poll began collecting responses. In the context of our paper, start date is a key variable as it allows us to focus on polling data collected after July 21, 2024—the date Kamala Harris officially announced her presidential campaign. By filtering the dataset to include only polls with a start date after this announcement, we ensure that each poll reflects voter sentiment with Harris as an active candidate, enabling a more accurate comparison between her and Donald Trump. Figure 6 displays the distribution of polling start dates from late July to mid-October 2024, showing when polling efforts began across key battleground states. The histogram reveals peaks in polling activity, particularly in mid-August, early September, and late September, indicating periods of heightened interest or campaign events that likely prompted increased polling.

This distribution provides insights into the timing and intensity of polling efforts throughout the campaign season. By including only polls initiated after Harris's campaign launch, we capture voter preferences in a context where both major candidates are actively campaigning, thus making the polling data more relevant for direct comparisons. This filtered timeline ensures our model focuses on data that reflects the competitive dynamics between Harris and Trump, enhancing the validity of our predictive analysis as we approach Election Day.

3 Model

The Bayesian Beta regression models constructed in this paper aim to predict the probability of victory for Donald Trump and Kamala Harris in the 2024 U.S. Presidential Election by analyzing polling data of the 7 battleground states. One Bayesian Beta regression model is constructed for each candidate, and each of them estimates the proportion of support for each candidate based on polling data, accounting for both polling trends over time, and state-level differences in voter sentiment. These models allow us to evaluate how support for each candidate fluctuates over time in key states, and predict potential electoral outcomes.

One of the goals of our model is to analyze changes in voter support for each candidate over time, as Election Day approaches. By using end_date_num (days the earliest poll date) as a predictor with natural splines, the model captures non-linear trends in support. The

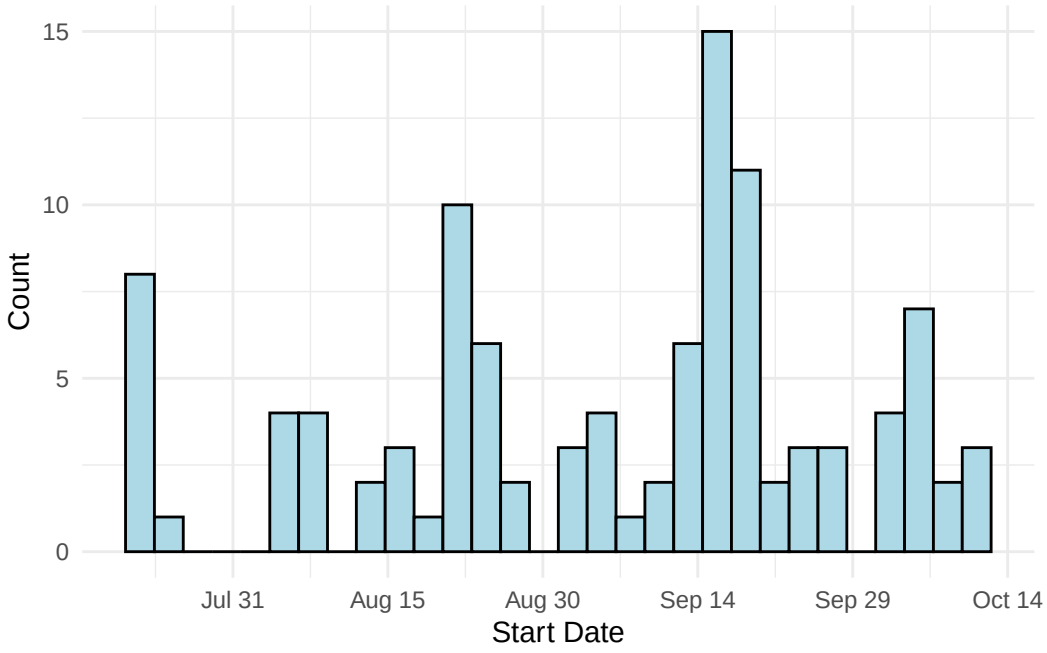

Figure 6: Histogram illustrating the distribution of polling start dates for key battleground states leading up to the 2024 United States Presidential Election. The chart highlights the frequency of polling activities initiated over time, with a notable peak on September 14, indicating a surge in polling efforts as the election approached.

time-based analysis provides us with information about how voter preference may shift due to campaign events, news cycles, and other influential factors in the months leading up to the election.

Another goal of our model is to predict the potential winner of the Presidential Election. We recognize that battleground states have unique political landscapes, so we include state as a random effect in each model. This allows each of the 7 battleground states to have a unique baseline level of support for each candidate, capturing regional differences in voter preferences. By incorporating state-level effects, the model can show variations in support across different states, revealing the potential winning candidate for the presidential election. Further background details, model specifications, and diagnostics are included in Appendix B.

3.1 Model Assumptions

In order to ensure the validity and reliability of our Bayesian models for predicting voter support percentages for Donald Trump and Kamala Harris, we establish several key assumptions that guide the modeling process:

- **Linearity in the Log-Odds**: The relationship between predictor variables (end_date and state) and the log-odds of the outcome variable (percentage of support) is linear. To account for non-linear trends over time, nature splines are used.
- **Independence**: The outcome for each polling observation is assumed to be independent of others. Each poll represents a distinct sample of voter sentiment at a specific time and location.
- **Appropriate Distribution for Proportions**: Since the model predicts support percentages (bounded between 0 and 100), a Beta distribution with a logit link is applied, which is ideal for data that are restricted to this range.
- **Random Effects**: To capture variation across states, state-level random effects are included, allowing each state to have its unique baseline level of support, acknowledging regional political differences
- **Prior Distributions**: The model uses weakly informative priors for the intercept and coefficients, which helps in generating stable estimates and reducing the influence of outliers or unusual polling data.

3.2 Model Setup

Our models were ran in R (R Core Team 2023) using the 'rstanarm' package of Gabry et al. (2023).

3.2.1 Trump Model

The primary estimand for Donald Trump's model is the probability of victory in the selected battleground states, derived from a logistic regression framework. The model is structured as follows:

$$P(\text{Victory}_{Trump} = 1) = \frac{e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}{1 + e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}$$

Where:

- $P(\text{Victory}_{Trump})$ is the predicted probability of Donald Trump winning.
- β_0 represents the intercept, indicating the baseline log-odds of winning.
- β_1 is the coefficient for the natural spline transformation of the numeric end date (end_date_num), capturing nonlinear trends in voter support over time.
- u_{state} accounts for random effects specific to each state, reflecting variations in voter preferences.

The dataset for the Trump model was prepared by extracting the relevant polling data, including the state, end date, and corresponding percentage of support for Trump ($Trump\_pct$). The end date was converted into a numeric format, denoting the number of days since the earliest polling date in the dataset. This transformation allows the model to effectively capture the evolving nature of voter sentiment as the election approaches.

The model was implemented using the 'brms' package (**brms?**) in R (R Core Team 2023), applying a Beta distribution with a logit link function to accurately model the proportions of support:

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

Here, y_i is the proportion of support for Trump in a given state, μ_i is the mean of the Beta distribution, and ϕ is the precision parameter reflecting variability.

We specified priors for the model parameters as follows:

- For the intercept β_0, we set a prior of $\beta_0 \sim \text{Normal}(0, 10)$, allowing the data to inform the estimation of baseline log-odds without imposing strong assumptions.
- For the slope β_1, we applied a prior of $\beta_1 \sim \text{Normal}(0, 5)$, reflecting reasonable expectations regarding the influence of polling trends.

Natural splines were used to model the effect of the end date, accommodating observed nonlinearity in voter support. Key assumptions include the independence of observations within states and the suitability of the Beta distribution for modeling proportions. However, limitations may arise from the accuracy of polling data and the assumption that historical voting patterns will predict future behavior.

3.2.2 Harris Model

Similar to the Trump model, the estimand for Kamala Harris's model is the probability of her victory in the selected battleground states, modeled within a logistic regression framework. The formulation is analogous:

$$P(\text{Victory}_{Harris} = 1) = \frac{e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}{1 + e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}$$

Where the components mirror those described for the Trump model, tailored to reflect the probability of victory for Harris.

For this model, the dataset was constructed similarly, extracting polling data specifically for Harris ($Harris\_pct$). The same transformations were applied, converting end dates into a numeric format to facilitate the modeling of trends over time.

The Harris model was also implemented using the `brms` package, employing a **Beta distribution** with a **logit link function** to represent the proportions of support accurately. The same mathematical representation and priors were applied as in the Trump model:

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

With the same definitions for y_i, μ_i, and ϕ.

Natural splines were employed to adapt to the non-linear trends in voter support over time. The assumptions and limitations discussed for the Trump model apply equally to the Harris model.

By utilizing Bayesian beta regression models for both candidates, we leverage their flexibility and capacity to incorporate prior beliefs, which is particularly advantageous in the context of electoral forecasting where uncertainty plays a critical role.

3.3 Model Justification

The Bayesian Beta Regression model with a logit link function is the most suitable model for our prediction of the 2024 U.S. Presidential Election because it aligns closely with both the nature of the data and the goals of our paper, which include understanding voter support trends over time and accounting for regional differences across battleground states.

Our primary outcome variable is the percentage of support for each candidate in 7 key battleground states, scaled to a proportion between 0 and 1. Support percentages are naturally bounded between 0% and 100%, and this bounded nature poses challenges for traditional linear regression, which assumes an unbounded continuous outcome. The Beta distribution is ideal for modeling proportion data because it is inherently defined on the interval (0, 1). This

allows us to model the proportions of support without the risk of generating unrealistic values outside this range. Unlike a linear model, which could predict support percentages exceeding 100% or falling below 0%, the Beta distribution respects the boundaries of our data. The logit link function further ensures that the model predicts in terms of log odds, which is suitable for interpreting support as probabilities—an essential consideration given our goal of forecasting potential victory in each state.

One of our key goals is to examine how support for each candidate evolves over time as Election Day approaches. Polling data often show non-linear trends, with fluctuations in support that may correspond to campaign events, debates, or shifts in the political landscape. Using natural splines with end_date_num (the number of days since the earliest poll date) allows the model to capture these non-linear trends in a flexible way. By including natural splines with 4 degrees of freedom, we allow the model to capture gradual changes in voter sentiment while avoiding overfitting to noise or abrupt fluctuations in the data. This is particularly important in the context of election polling, where voter opinions can change either slowly or rapidly depending on unpredictable external events. Alternative models, such as standard logistic regression without splines, would fail to capture these changes, potentially leading to a less accurate depiction of how support varies over time.

A unique aspect of our analysis is the focus on battleground states, which have distinct political dynamics that can heavily influence election outcomes. Incorporating state as a random effect provides flexibility in estimating state-specific baseline support levels for each candidate, while still allowing us to identify overall trends across all states. This approach is more effective than treating state as a fixed effect because it avoids overparameterizing the model with individual state coefficients, which could lead to overfitting in cases with limited data for some states. The random effect structure also helps generalize the model's predictions, making it more robust when predicting support in battleground states with varying polling frequencies.

The Bayesian approach offers advantages for our model by allowing us to incorporate prior information, enhancing the stability and reliability of our estimates. Given that polling data can be sparse for certain states or dates, Bayesian inference helps stabilize parameter estimates, especially when there are limited data points for certain states or specific dates. By using weakly informative priors, we let the data primarily determine the model's outcomes, while still providing some structure to prevent overfitting or extreme parameter values. The Bayesian framework also produces credible intervals, which provide a clear interpretation of uncertainty around our predictions. In the context of election forecasting, this is valuable as it enables us to quantify the range of plausible outcomes for each candidate's support, adding a layer of transparency to our predictions. This model provides interpretability that aligns well with the goals of our paper. By structuring the model with a logit link function, we can interpret the results in terms of odds or probability of victory, which is highly relevant in an election context. This interpretability allows us to understand the likelihood of each candidate winning in a specific state at a given time.

Additionally, the Bayesian Beta regression model offers flexibility in how we examine each candidate's trends individually. By constructing separate models for Trump and Harris, we

can directly compare their temporal trends and state-level variations without confounding effects.

Overall, the Bayesian Beta regression model with a logit link function is the most suitable model for our study. It allows for flexible non-linear trends over time, accounts for regional variation across battleground states, and provides interpretable results in terms of victory probabilities. The Bayesian framework stabilizes estimates in the presence of sparse data and provides meaningful uncertainty estimates. The model's structure aligns well with our goal of forecasting each candidate's probability of victory, justifying the choice for achieving a comprehensive analysis and prediction of the 2024 U.S. Presidential Election.

3.4 Model Summary

3.4.1 Trump Model Summary

Table 2: Fixed Effects Summary for Trump Model.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.0922852	0.0305620	-0.1520050	-0.0333538
nsenddatenumdfEQ41	-0.0071493	0.0355882	-0.0793200	0.0623891
nsenddatenumdfEQ42	0.0802768	0.0317675	0.0179060	0.1426297
nsenddatenumdfEQ43	0.0025949	0.0676124	-0.1341383	0.1336160
nsenddatenumdfEQ44	0.0789177	0.0400326	0.0013945	0.1589037

3.4.2 Harris Model Summary

Table 3: Fixed Effects Summary for Harris Model.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.1508299	0.0279264	-0.2041163	-0.0954951
nsenddatenumdfEQ41	0.1225217	0.0341326	0.0552911	0.1903117
nsenddatenumdfEQ42	0.0248781	0.0291738	-0.0321764	0.0833965
nsenddatenumdfEQ43	0.1191128	0.0630871	-0.0070147	0.2416542
nsenddatenumdfEQ44	0.0669873	0.0382015	-0.0071597	0.1426828

4 Results

In this section, we present the results of our analysis, focusing on the electoral support for Donald Trump and Kamala Harris across key battleground states during the 2024 Presidential

Election cycle. We begin by examining the state-level polling averages for both candidates, which provide a snapshot of voter preferences as captured by various pollsters. Next, we explore how polling averages have evolved over time, highlighting trends that may influence election outcomes. We then assess the predicted probabilities of Trump winning on Election Day, followed by a detailed look at the likelihood of winning by state. Finally, we visualize the predicted winner by state on a US map, summarizing our findings and their implications for the electoral landscape.

4.1 State-Level Polling Averages for Trump and Harris

Table 1 displays overview of the percentage of support for Trump and Harris by each battle-ground state. It provides us with a high level understanding of the current voter dynamic. The pct results are aggregated averages of all pct data points for Trump and Harris in each state. By comparing the percentage of support for T.rump and Harris, we can determine the candidate that has a leading advantage in winning each state. For example,Trump has 49.2% support in Arizona while Harris has 46.8% support, which means Trump has won the state of Arizona.

Table 4: Table displaying the percentage of support for Donald Trump and Kamala Harris across key battleground states in the lead-up to the 2024 United States Presidential Election. The table lists states with their respective support percentages, highlighting the competitive landscape as voters prepare to make their decisions on Election Day. Notably, Trump maintains a slight lead in states such as Arizona and Georgia, while Harris shows stronger support in states like Michigan and Nevada. This data provides insight into the current polling dynamics and potential electoral outcomes in these critical states.

State	Trump %	Harris %
Arizona	49.2	46.8
Georgia	48.9	47.2
Michigan	46.9	48.1
Nevada	47.9	48.5
North Carolina	47.6	48.1
Pennsylvania	47.4	48.3
Wisconsin	47.2	48.9

This distribution is crucial for evaluating the consistency and breadth of the polling data. Pollsters with a higher count of polls, like Emerson, contribute substantially to the dataset and can have a greater influence on the model's predictions, particularly if their methodology or sample selection differs from others. On the other hand, pollsters with fewer entries may provide valuable but limited data points, potentially adding unique perspectives but with less influence on overall trends.

4.2 Polling Averages for Trump and Harris Over Time in Battleground States

Figure 7 provides a visual comparison of the current polling support for Donald Trump and Kamala Harris across key battleground states, tracked over time from August 2024 to October 2024. Each line plot represents the average polling percentages for both candidates in a specific state, with red lines denoting Trump's support and blue lines representing Harris's support. Across several states, there are noticeably increased fluctuations in polling averages for both candidates closer to the election particularly from late September to October. This pattern may indicate heightened voter interest and engagement as the election date approaches which is November 4th, 2024, it could also reflect changes in public opinion due to campaign events, political debates, or other external factors.

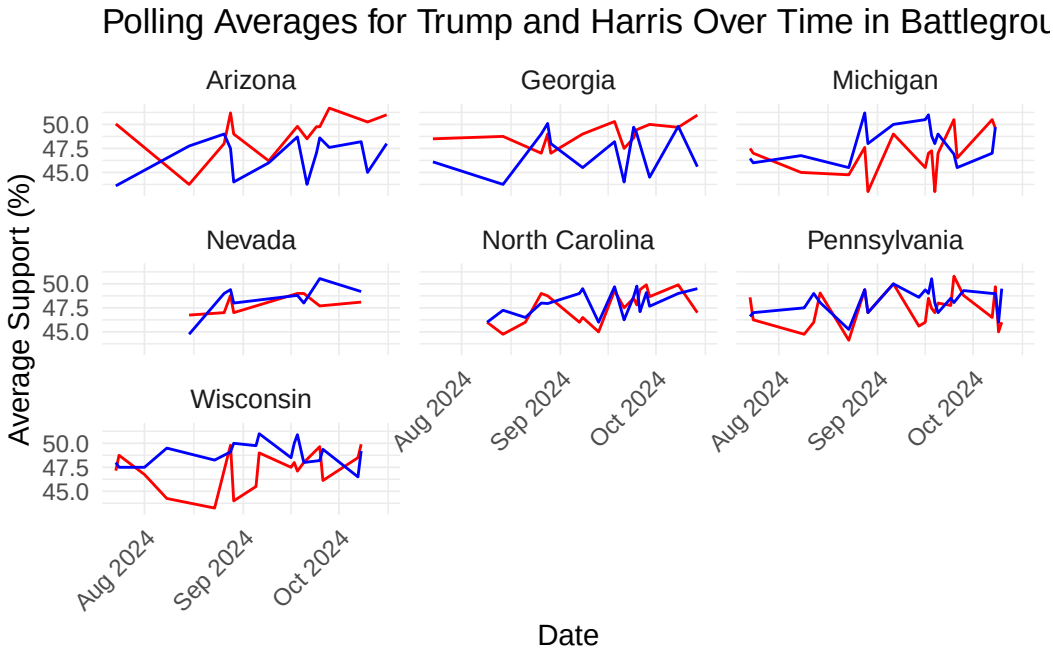

Figure 7: Polling Averages for Trump and Harris Over Time in Battleground States.

Out of the seven battleground states, Arizona and Georgia display a clear leading win by Trump while Nevada and North Carolina display a clear leading win by Harris. Michigan, Pennsylvania, and Wisconsin are the three states that don't have a clear leading winner. Therefore, these three states are the key states that could determine the election outcome as it approaches election day. Election outcomes could be easily flipped if either of the candidates can sweep more than one out of the three states. Candidates should focus on strategizing their campaigns to increase their voters' support at critical moments. The frequent fluctuations of greater margins also indicate the uncertainty and instability of voters' attitudes towards both candidates.

Since it is difficult to conclude a definitive predictive outcome by observing the current data, we utilize our models, described in Section 3 to predict the possible outcomes by incorporating the given data. By using the current data, the model can identify the state-specific trends thus generating more accurate predictions of the outcome. Relevant prediction results are provided below.

4.3 State-Level Probability of Trump Winning on Election Day

By incorporating our Bayesian models, we generate the predicted probabilities of winner for both Donald Trump and Kamala Harris in key battleground states, on Election Day (November 5th, 2024). Figure 8 provides a visual summary of these predicted probabilities.

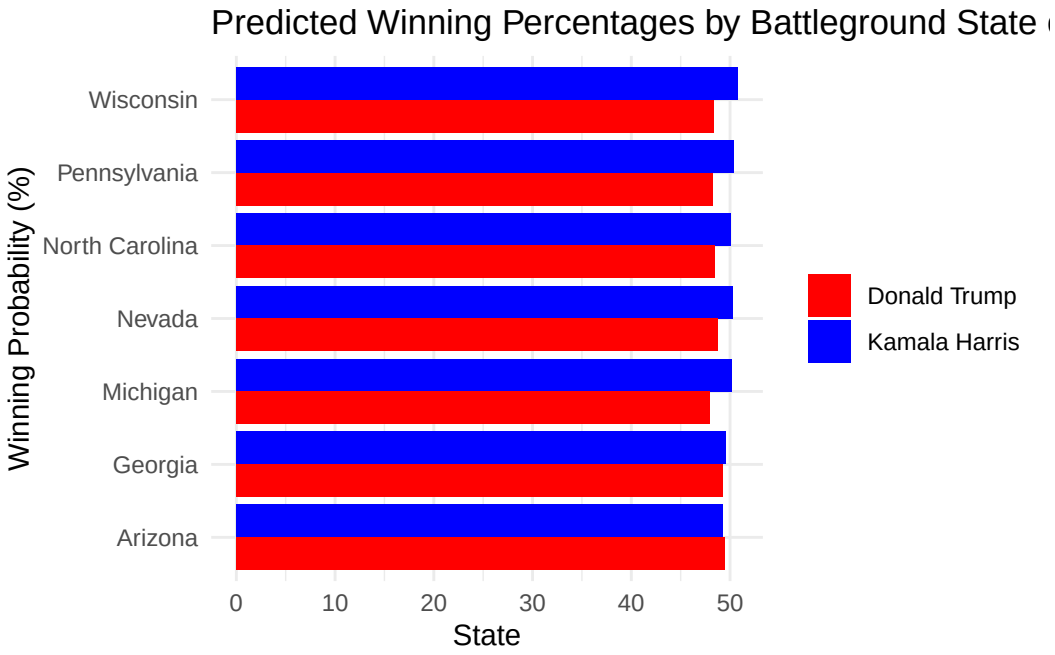

Figure 8: Probability of Trump Winning by State on Election Day.

These predictions are based on polling data collected up to October 19th, 2024 to forecast each candidate's support within these critical states. Each state's political dynamics are unique, and this is reflected in the model's predictions. States like Arizona and Georgia show relatively close probabilities for both candidates, marking them as key states to focus on as they are the most influential on the election outcomes. The five other states all show relatively distinct leading support for Harris, suggesting that Harris's support base in those states has been consistently strong in the polls leading up to Election Day.

4.4 Probability of Winning by State on Election Day

Table 5 displays the numeric percentage of the probabilities of winning for Trump and Harris. By comparing the predicted probabilities, the predicted winner for each state is assigned accordingly. For instance, the table displays that Trump has a 49.48% predicted probability of winning Arizona compared to 49.24% of Harris's, making Trump the predicted winner in Arizona. For all six other states, Harris has a higher predicted probability of winning than Trump.

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

State	Predicted Trump Win (%)	Predicted Harris Win (%)	Predicted Winner
Arizona	49.48	49.24	Donald Trump
Georgia	49.27	49.54	Kamala Harris
Michigan	47.97	50.22	Kamala Harris
Nevada	48.79	50.28	Kamala Harris
North Carolina	48.45	50.05	Kamala Harris
Pennsylvania	48.23	50.33	Kamala Harris
Wisconsin	48.32	50.78	Kamala Harris

4.5 US Map Showing Predicted Winner by State on Election Day

To visualize the colour distribution of the predicted results, we emloy the US map from (Albers et al. 2023) to display the results from Table 5.

- 'Predicted Winner' column is Donald Trump: State filled red
- 'Predicted Winner' column is Kamala Harris: State filled blue

According to Figure 9, Arizona is the only state where Trump has a leading win while all the other six states are led by Harris. Since Harris has a six-to-one lead in the predicted percentage of support for all battleground states, we conclude that Harris is the predicted winner for the 2024 cycle of Presidential Election.

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

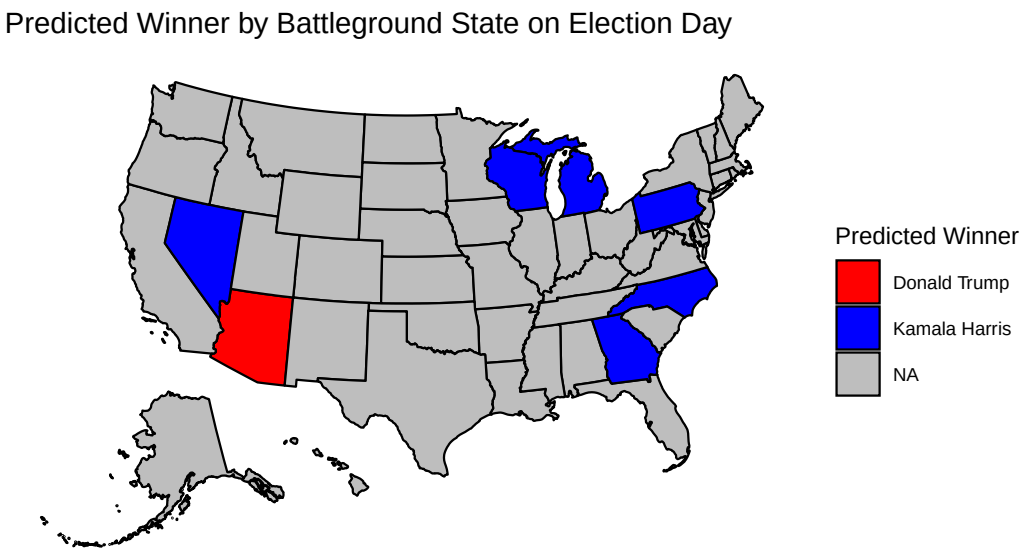

Figure 9: Predicted Winner by Battleground State on Election Day.

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.

5.2.1 Model Performance - Trump Model

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.

5.2.2 Model Performance - Harris Model

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, Sienna/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 **?@fig-averagetable**.

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 Model Diagnostics

In order to assess the validity and reliability of our Bayesian models for predicting voter support for Kamala Harris and Donald Trump, we conducted posterior predictive checks for both models. Posterior predictive checks are a crucial diagnostic tool that allows us to evaluate how well our models can replicate observed data based on the parameters estimated during the modeling process. By comparing the distribution of the observed support percentages with the predicted support values, we can identify potential discrepancies and gauge the overall fit of our models.

These checks not only help us verify whether the model assumptions hold but also provide insights into how well the models capture the underlying data-generating processes. A strong alignment between observed and predicted values would suggest that our models are robust and capable of accurately reflecting the dynamics of voter support in the current electoral context. Conversely, significant deviations may indicate areas for model refinement or highlight limitations in our current approach. Through these diagnostics, we aim to ensure the credibility of our findings and enhance our understanding of the electoral landscape as we approach the 2024 Presidential Election.

B.2 Posterior Predictive Check for the Trump Model

The posterior predictive check for the Trump model is illustrated Figure 10, showcasing the relationship between the observed support percentages y and the predicted support values y_{rep}. The plot displays a density estimate for both the actual observed data (in dark blue) and the predicted values generated by the model (in light blue). The primary aim of this diagnostic check is to assess how well the model captures the distribution of observed support for Donald Trump.

From the visualization, we can observe that the predicted values closely align with the actual observed support percentages. The dark blue curve representing the observed data indicates a concentrated support range around 48% to 52%, suggesting that the model effectively captures the central tendency of voter support for Trump. The light blue curves depict a variety of predicted densities, indicating the model's ability to simulate a range of possible outcomes around the observed values. The alignment of the y and y_{rep} distributions indicates that the model is successfully reflecting the dynamics of voter support, thereby validating the choice of model and its structure. However, the presence of some variability in the predicted values suggests that while the model fits the observed data well, there may still be unexplained fluctuations in support levels that could warrant further investigation. These limitations are further discussed in Section B.1. This examination enhances our confidence in the model's utility for forecasting electoral outcomes, ensuring it is robust enough to inform strategic decisions as the election approaches.

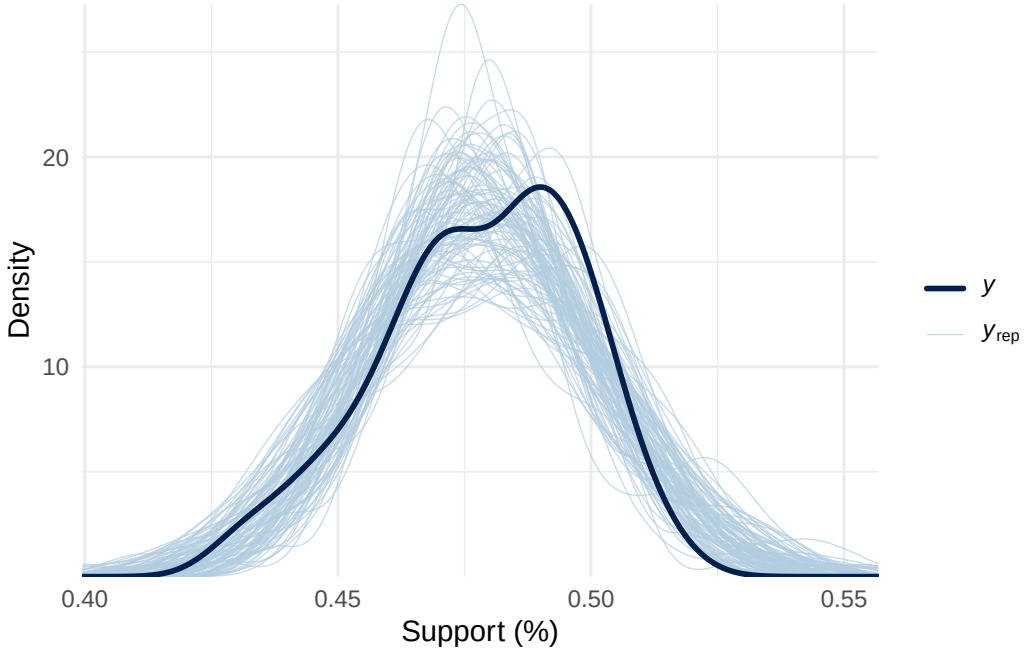

Figure 10: Posterior predictive check for Donald Trump's model, illustrating the distribution of observed support percentages y alongside the predicted support values y_{extrep}. The graph shows a strong overlap between the observed and predicted distributions, indicating that the model accurately captures the dynamics of voter support for Trump.

B.2.1 Model Diagnostics - Harris Model

The posterior predictive check for the Harris model is presented in Figure X, illustrating the relationship between the observed support percentages y and the predicted support values y_{rep}. Figure 11 displays a density estimate for both the actual observed data (depicted in dark blue) and the predicted values generated by the model (shown in light blue). The purpose of this diagnostic check is to evaluate how well the model captures the distribution of observed support for Kamala Harris.

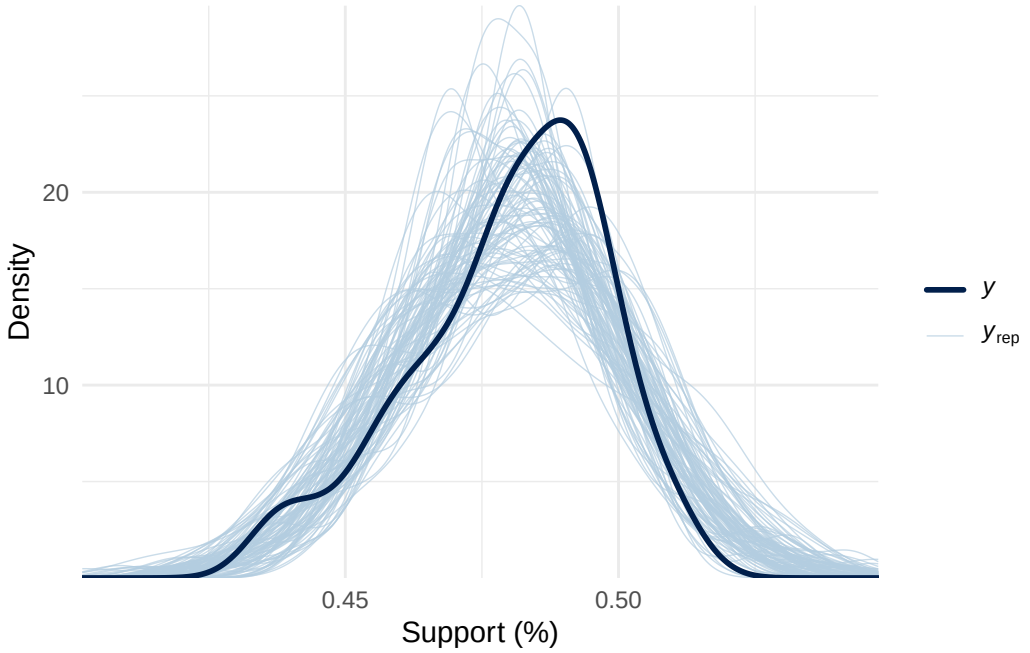

Figure 11: Posterior predictive check for Kamala Harris's model, illustrating the distribution of observed support percentages y against the predicted support values y_{extrep}. The graph shows a strong correspondence between the observed and predicted distributions, indicating that the model accurately captures the dynamics of voter support for Harris in the given dataset.

B.3 Posterior Predictive Check for the Harris Model

The posterior predictive check for the Harris model is presented in Figure X, illustrating the relationship between the observed support percentages y and the predicted support values y_{rep}. This visualization features a density estimate for both the actual observed data (depicted in dark blue) and the predicted values generated by the model (shown in light blue). The purpose of this diagnostic check is to evaluate how well the model captures the distribution of observed support for Kamala Harris.

In the plot, we observe a strong overlap between the y and y_{rep} distributions, with the dark blue curve representing the observed support percentages primarily concentrated around 48% to 52%. This indicates that the model successfully approximates the actual voting support dynamics for Harris. The light blue curves signify the distribution of predicted values, demonstrating the model's ability to replicate the observed trends effectively. The close alignment of the observed and predicted densities suggests that the model adequately reflects the underlying support for Harris among voters, thereby affirming the appropriateness of the model structure. However, similar to the Trump model, the presence of some variability in the predicted values indicates that while the model performs well, there may still be unexplained variations in support levels that could benefit from further analysis. Such limitations are also discussed in Section B.1.

C Pollster methodology overview and evaluation: NYT/Siena College

C.1 Background of NYT/Siena College Polling

The NYT/Siena College Polling is a U.S. polling organization where the New York Times and Siena College collaborate to deliver precise and timely poll results for the presidential elections (The New York Times 2023). For the 2024 election cycle, The NYT/Siena College poll uses a robust sampling approach to ensure accuracy and representativeness. The poll applies a stratified dual-frame sample, drawing from both landlines and cell phones.

C.2 Target population, Sampling frame, and Sample

- Target population: Target population refers to "the collection of all items about which we would like to speak" (Rohan Alexander 2024). In this case, NYT/Siena College's polling target population is all registered American voters who live in the six battleground states: Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin. Given the nature of the presidential elections, the election result is based on the electoral college instead of the popular vote, thus the pollster focuses on polling on the states that are the likeliest to decide the outcome of the race, which are called battleground states. The battleground states are states that have similar voter support for both the Democratic or the Republican party, a slight difference in support could determine which party wins a state.
- Frame: Sampling frame refers to "a list of all items from the target population that we could get data about" (Rohan Alexander 2024). NYT/Siena College's sampling frame is a comprehensive list of registered American voters obtained from the L-2 voter file, it includes details such as demographic characteristics, contact information, and voting history. By using the L-2 voter file as their sampling frame, the pollster can

ensure representativeness by allowing them to draw samples that accurately reflect the demographics of the registered voter population.

- Samples: Sample refers to "the items from the sampling frame that we get data about" (Rohan Alexander 2024). In this case, the samples of NYT/Sienna College are individuals who are registered American voters obtained from the L-2 voter file from each stratum from each state who have also answered and completed the questionnaires.

C.3 Sampling Methodology

NYT/Siena College uses stratified sampling as their methodology where the population is divided into distinct strata based on shared characteristics. These strata include age, gender, race, geographic region, and other various relevant variables. Once the population is divided, random samples are taken from each stratum, in proportion to their size in the population. For each poll, NYT/Siena College takes a sample size of 1000 registered American voters. To refine their sample, NYT/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, Siena/NYT uses the information in the datasets 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.

C.4 Trade-off of Sampling Methodology

Although stratified sampling has many strengths such as appropriate representation of the target population, reduction of sampling bias, and sampling efficiency, it also has some weaknesses. First, its implementation can be complex, stratified sampling requires detailed knowledge of the population to define the strata appropriately. This can be challenging to implement if accurate, up-to-date information on demographic or geographic characteristics is unavailable or difficult to access. For example, if there are significant changes in the L-2 voter file since the last election, it is difficult to take account into the unknown changes. Second, there could be potential for mis-stratification. If the chosen strata do not capture meaningful differences within the population, or if important subgroups are overlooked, the results might be biased or inaccurate. For example, if new voting trends emerge that were not captured in the strata definitions, the poll might miss important shifts in voter opinion. Third, stratified sampling can be time-consuming and resource-intensive. Diving the population into strata and ensuring

proportionate sampling within each group can be more time-consuming and resource-intensive compared to simple random sampling. Pollsters may need to gather detailed data on the target population, which adds to the cost and complexity of the survey. Lastly, there is the risk of over-emphasis on stratification. If the population is stratified too finely, the sample size in each stratum may become too small to yield meaningful results, increasing the margin of error for each subgroup.

C.5 Sample Recruitment Methods

Each poll is conducted by phone using live interviewers at call centers based in Florida, New York, South Carolina, Texas, and Virginia. Most people that the live interviewers call do not answer the call, and some don't finish the survey, leaving only about two percent both answer their phone and complete the poll. To mitigate this: First, the pollster keeps calling numbers until someone picks up, even though it can take hours–a reason why a high-quality poll is time-consuming and expensive to conduct. Second, voter files are used to help make sure the pollster is reaching a representative sample; for example, they may reach out to a higher percentage of people who are less likely to respond. Third, pollster mathematically adjusts the responses they do receive so they are representative of the wider population, it includes adjusting by age, race, gender, voting history and etc.

C.6 Non-response Handling

Non-responses are critical factors in polling because they affect the representativeness of the sample and, consequently, the accuracy of the poll results. When a significant portion of selected individuals doesn't respond, the poll risks becoming unrepresentative of the broader population, especially if certain groups are systematically more or less likely to respond. It is also important to know how pollsters handle non-response since it provides transparency about the polling methodology. It also gives confidence that the results are not skewed by an unrepresentative sample when pollsters clearly explain their non-response handling. Therefore, polls with thorough non-response strategies are generally more reliable than those without. The NYT/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.

C.7 Questionnaire Pros and Cons

Based on the NYT/Siena College's past practices on presidential election polls, they often ask questions about candidate preference, approval ratings, voter opinions on key issues, economic perceptions, as well as social and cultural issues. They also ask about the respondents' demographic information and voting behavior to categorize respondents for analysis.

C.7.1 Strengths of the NYT/Siena College questionnaires

First, they have a comprehensive question design. NYT/Sienna College polls ask a wide range of questions that provides a detailed view of the respondent's preferences and opinions. It helps produce multi-dimensional data that allows for in-depth analysis and segmentation of voter groups. Second, the questionnaires are tailored to likely voters, as identified through the L-2 voter file. This means that the voters are more likely to be up-to-date with relevant policies proposed by different parties, and have formed their own opinions about them, this improves the predictive reliability of the poll. Third, the questionnaires contain detailed demographic questions about the respondents which allow NYT/Siena College to stratify responses by voter subgroups, identifying varied voter bases and diverse issues.

C.7.2 Weaknesses of the NYT/Siena College questionnaires

First, the length and complexity of the questionnaires might cause the respondents to be less likely to respond or finish the polls. Respondents may rush through questions or drop out before completion. Respondents may also be prone to moderacy response bias where they have the tendency to choose middle options regardless of question content, as well as response order bias where respondents choose answers based on their order. Second, questionnaires have a dependence on self-reported likelihood to vote. Self-reported likelihood can be unreliable, especially for individuals whose intentions to vote may change close to the actual election date. This can reduce the accuracy of predictions based on likely voter models. Third, there are limited open-ended questions. Many voters may have mixed opinions on specific issues or policies that cannot be explained in a single-choice answer. However, answers to open-ended questions are often difficult to convert to actionable data, so questionnaires tend to rely on closed-ended questions for scalability. This might restrict the ability to capture complex reasonings behind their choices, which could add extra strain on the time and resources used to interpret the answers.

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 Section D.2.

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

Albers, L. E. et al. 2023. *Usmap: US Maps Including Alaska and Hawaii.* https://CRAN.R-project.org/package=usmap.

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.

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

Rohan Alexander. 2024. "Telling Stories with Data." https://tellingstorieswithdata.com/.

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