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

A Bayesian Analysis of Polling Data from Key Battleground States - Estimating Voter Support Dynamics for Donald Trump and Kamala Harris in the 2024 U.S. Presidential Election

Jimin Lee Sarah Ding Xiyan Chen

November 4, 2024

This paper presents a predictive model for the 2024 United States Presidential Election, focusing on essential battleground states. Utilizing a Bayesian framework and state-level polling data, we estimate the probabilities of victory for candidates Donald Trump and Kamala Harris. Our findings indicate that Harris is projected to secure a win in six out of seven key battleground states, with predicted support exceeding 50%. This analysis not only forecasts the likely winner of the U.S. election but also equips society and the economy to better undertstand the current U.S. political climate.

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 set to be a crucial moment in American political history, with battleground states serving as pivotal determinants of the electoral outcome. As candidates Donald Trump and Kamala Harris intensify their campaigns, understanding voter preferences in these key regions is essential for accurate election forecasting. This paper utilizes state-level polling data from Project 538 (FiveThirtyEight 2024) and the statistical programming language R (R Core Team 2023) to develop a predictive model estimating each candidate's likelihood of victory in critical battleground states, including Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin.

The primary focus of this study is the probability of victory for Trump or Harris in these battleground states, derived from aggregated polling averages. Employing a Bayesian hierarchical model combined with natural splines, our analysis effectively captures the temporal dynamics of voter sentiment, allowing us to estimate winning probabilities in these decisive areas. Our findings reveal a competitive landscape, with Harris projected to win in six of the seven battleground states analyzed, notably in Michigan and Nevada, where support leans in her favor. However, Trump remains competitive in Arizona, illustrating the variability of voter preferences across these regions.

These predictions are critical not only for forecasting the likely outcome of the election but also for enhancing societal and economic understanding of the current political climate in the U.S. The implications of our findings extend beyond mere predictions, offering a framework for analyzing potential shifts in voter sentiment as Election Day approaches. The remainder of this paper is organized as follows: Section 2 details the data and variables utilized in the analysis; Section 3 outlines the model setup and estimation strategies employed; Section 4

presents the results, including visualizations of the predicted election outcomes; and Section 5 discusses the implications and limitations of our findings.

2 Data

2.1 Overview

The data for this paper, obtained from Project 538 (FiveThirtyEight 2024) on October 19, 2024, encompasses polling information for the 2024 Presidential Election cycle, reflecting the most current polling results from various pollsters, including the New York Times, ActiVote, and Morning Consult. Key variables include **'state'**, **'start_date'**, **'end_date'**, **'transparency_score'**, **'candidate_name'**, and 'percentage of support (**'pct'**) for each candidate. Our analysis focuses on Donald Trump and Kamala Harris, the leading candidates, collecting observations where 'candidate_name' matches either Trump or Harris. We aim to track support over time for each candidate by extracting relevant data on **'end_date'**, **'state'**, and **'pct'**. To ensure reliability, we included only polls from pollsters with high **'numeric_grade'** and **'transparency_score'**. Details on the data cleaning process are provided in Appendix A.

Data cleaning and analysis are conducted by employing the following packages: `tidyverse` package of Wickham et al. (2023), `janitor` package of Firke (2023), `lubridate` package of Grolemund and Wickham (2011), `arrow` package of Richardson et al. (2023)

2.2 Measurement

Voter opinions are influenced by various factors, such as media, political events, personal beliefs, and social pressures, which can change frequently, especially near an election. Pollsters simplify these complex opinions into basic polling responses like "Trump," "Harris," or "Undecided," which can obscure nuanced voter sentiment. This initial conversion may not capture voters' internal conflicts or reservations, as respondents must choose a definitive option. These individual responses are aggregated into percentages, treating each "Trump" or "Harris" response equally, regardless of confidence level, further smoothing out complexities. The model uses these aggregated support percentages to predict election outcomes by estimating the likelihood of each candidate winning. This step assumes that poll data accurately reflects voter behavior on Election Day, which may not always hold true, as it simplifies support into binary outcomes (predicting Trump or Harris as the winner based on higher support).

The use of "state" as a predictor assumes that all responses within a state are comparable, despite internal diversity, such as urban-rural divides or cultural differences, which may not be fully captured. Similarly, the "end_date" variable tracks changes in voter sentiment over time to capture patterns influenced by campaign events or news. However, this variable simplifies

continuous shifts into discrete snapshots, assuming stability between poll dates. This may overlook rapid opinion changes following significant events. Additionally, the model presumes time's influence on voter sentiment is consistent across states, despite varying local contexts and demographics that may respond differently to events.

2.3 Variables

The collected data from Project538 (FiveThirtyEight 2024) contains several key variables relevant to the analysis:

- **'State'**: Indicates the state where the polling took place, critical for understanding regional support dynamics.
- **'End Date'**: The date when the poll ended, which helps in analyzing trends over time.
- **'Start Date'**: The date when the poll started, to filter for Kamala Harris's late entry into the presidential race
- **'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 running candidate, 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 on the data cleaning process, refer to Appendix A.1.

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 of polling data, showing a total of 107 polls conducted. Average support is 47.801% for Donald Trump and 47.999% for Kamala Harris, with similar maximum and minimum support levels, indicating a competitive electoral landscape.

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

Table 1 reveals a total of 107 polls conducted, representing unique poll entries filtered for battleground states, sufficient data quality, and reported support percentages for both Trump and Harris. Among these, the average support level is 47.801% for Trump and 47.999% for Harris, indicating a slight advantage for Trump in the presidential race. The highest recorded support for Trump is 51.700%, while the lowest is 43%; for Harris, the highest is 51.200% and the lowest is 43.600%. These figures illustrate a competitive landscape without a clear leading candidate. It is essential to note that the percentages for Trump and Harris do not sum to 100%, as the combined average of 95.800% highlights a significant portion of respondents who are 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.

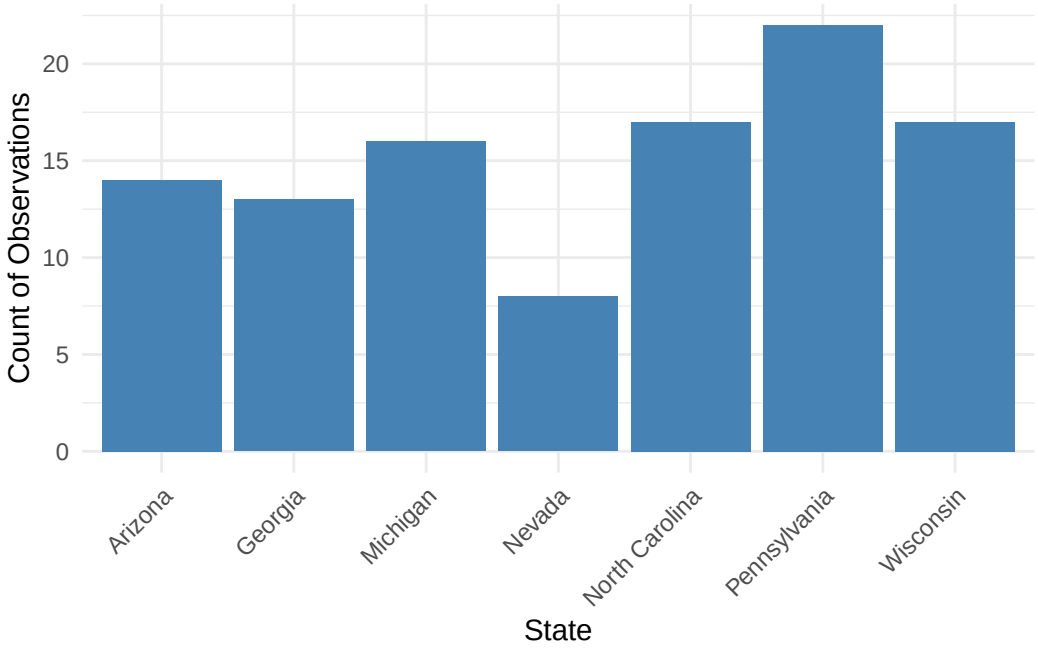

Figure 1: Distribution of Observations for Each State in Polling Data. Illustrates the number of polling observations across key battleground statesn. Pennsylvania has the highest count of observations, suggesting a more robust polling presence, while Nevada shows fewer observations.

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.

2.3.2.2 End Date

End date (**'end_date'**) marks the conclusion of data collection for each poll, serving as a predictive variable in our analysis of changes in voter support over time. This inclusion helps identify patterns as Election Day approaches, potentially reflecting the impact of campaign events, debates, or political developments. Figure 2 shows polling end dates from July 2024 to October 2024, with the earliest date set to July 2024, coinciding with Kamala Harris's official presidential campaign announcement on July 21, 2024. The latest end date, October 14, 2024, reflects the last data update from Project 538 (FiveThirtyEight 2024) on October 15, 2024.

The distribution reveals that polling activity is not evenly spread, with notable peaks in early September, late September, and early October, likely corresponding to periods of heightened political interest or significant campaign events that drive increased polling activity.

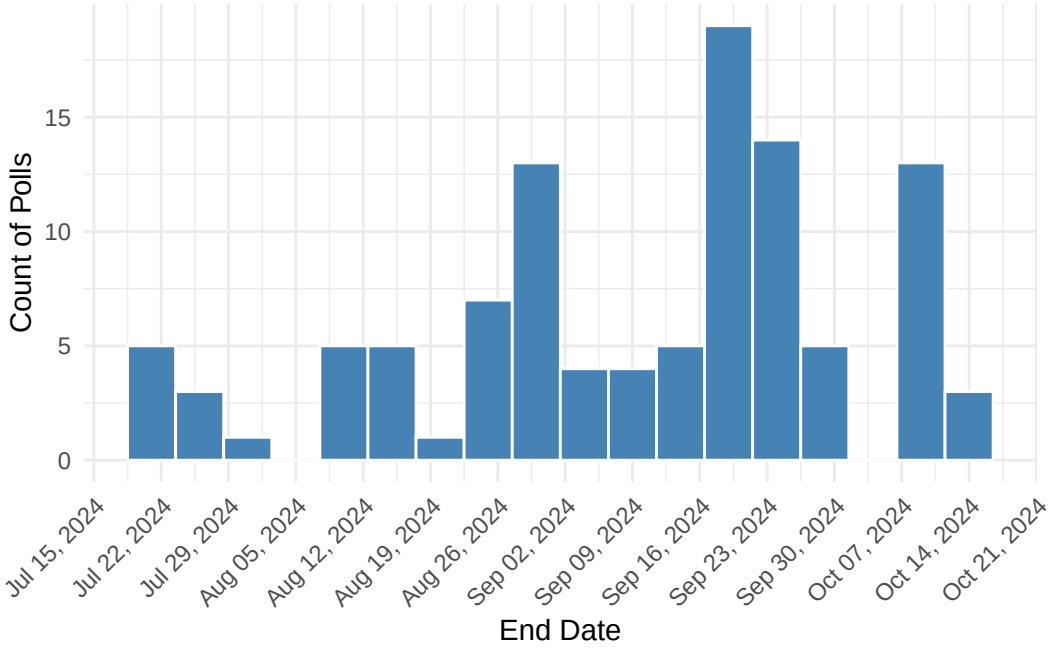

Figure 2: Distribution of Polling End Dates. This histogram displays the frequency of polling end dates for key battleground states, highlighting peaks in polling activity around late September.

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'**) assesses the quality of each poll, as rated by Project 538 (FiveThirtyEight 2024), based on factors like methodology, transparency, and historical accuracy. In our paper, this variable allows us to weigh the reliability of different polls, ensuring that those with higher ratings contribute more significantly to our voter support predictions.

Figure 3 illustrates the distribution of numeric grades in our analysis, which ranges from 2.5 to 3.0. Higher grades indicate more methodologically rigorous and transparent polls, enhancing the reliability of our model's predictions. The distribution shows several peaks around grades 2.7, 2.8, and 3.0, suggesting that certain pollsters consistently achieve higher ratings due to better methodologies or transparency. Conversely, valleys in the distribution may reflect a scarcity of polls from lower-rated sources. By incorporating numeric grade, we enhance prediction accuracy by prioritizing data from more reliable sources, which is particularly crucial in closely contested states where polling accuracy is essential for reliable forecasts.

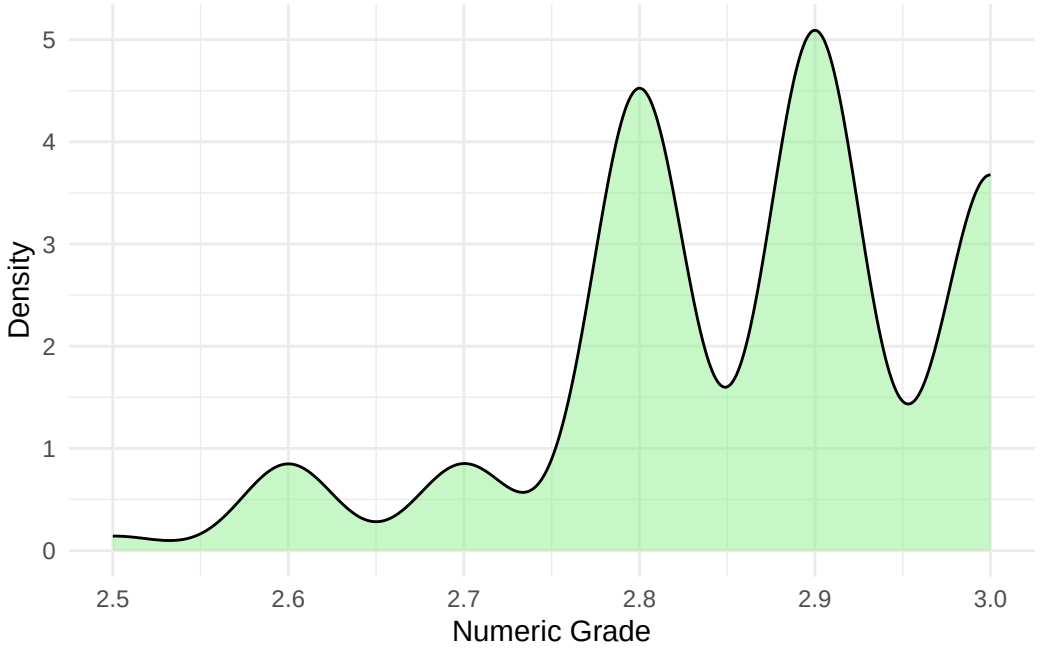

Figure 3: Density plot illustrating the distribution of numeric grades assigned to pollsters within the polling data. Highlights the frequency of numeric grades, indicating key concentrations that suggest variability in pollster quality.

2.3.3.2 Transparency Score

Transparency Score (**'transparency_score'**) quantifies the methodological transparency of each pollster on a scale from 0 to 10, as rated by Project 538 (FiveThirtyEight 2024). This score reflects the extent of disclosure regarding data collection and analysis methods, including sample size, weighting, and margin of error. In our paper, the transparency score is crucial for assessing the reliability of polling data; higher scores typically indicate greater methodological rigor.

Figure 4 illustrates the distribution of transparency scores in our dataset for key battleground states, showing a concentration around scores of 7 and 9, indicating that most polls come

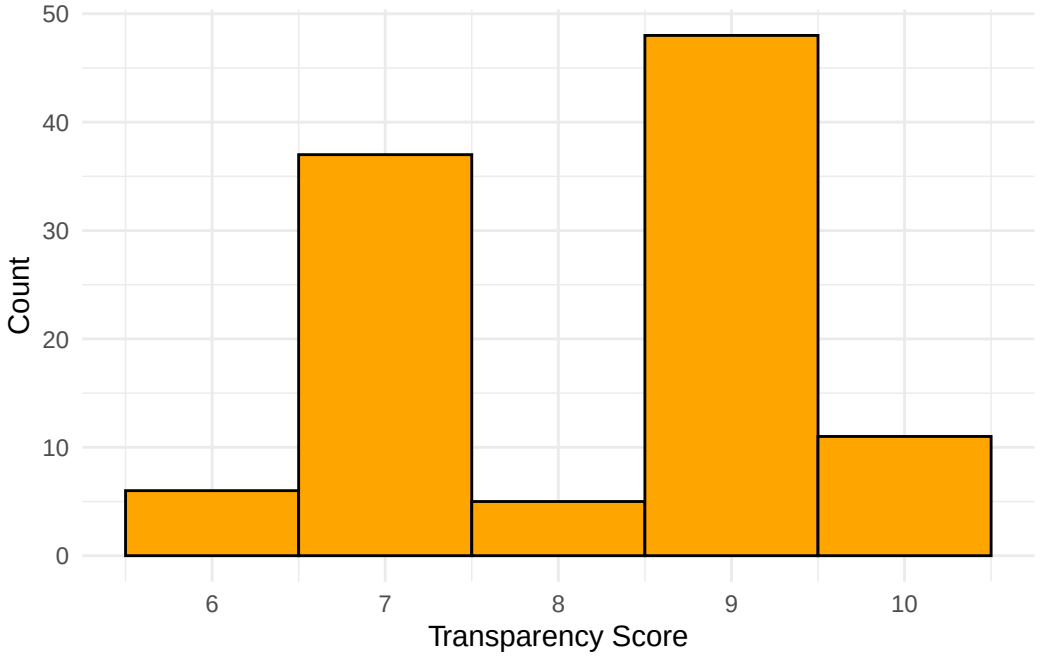

Figure 4: Histogram displaying the distribution of transparency scores among pollsters in the polling data. Indicates the frequency of various transparency scores, highlighting a concentration of scores around 7 and 9, which suggests a predominance of high transparency among the assessed pollsters.

from sources with relatively high transparency. Smaller peaks at lower scores (5) and a perfect score (10) suggest that while some pollsters provide limited transparency, a few achieve full disclosure. This distribution is vital for evaluating data credibility, as polls with higher transparency scores are more likely to yield consistent and reliable results, enabling us to identify potential biases. Incorporating transparency scores into our model helps differentiate between high- and low-transparency polls, which is essential for ensuring the stability of our predictions, particularly in closely contested states where polling accuracy is critical.

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.

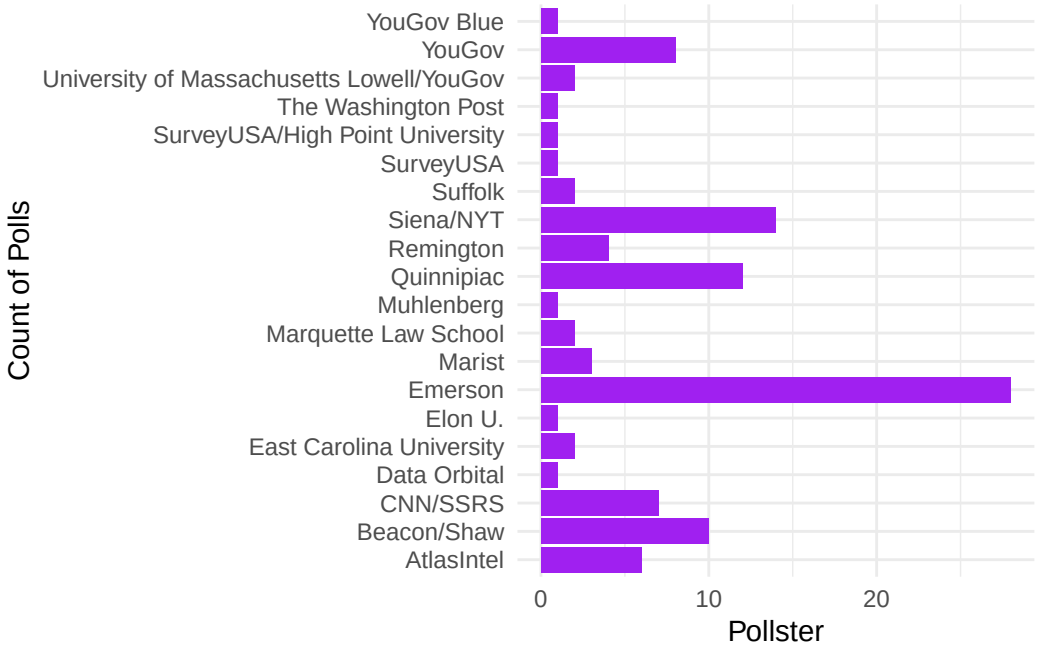

Figure 5: Horizontal bar chart displaying the distribution of polls conducted by various pollsters in key battleground states. Highlights the frequency of polling activities across different organizations, with Marist and Emerson standing out as the most prolific pollsters.

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.

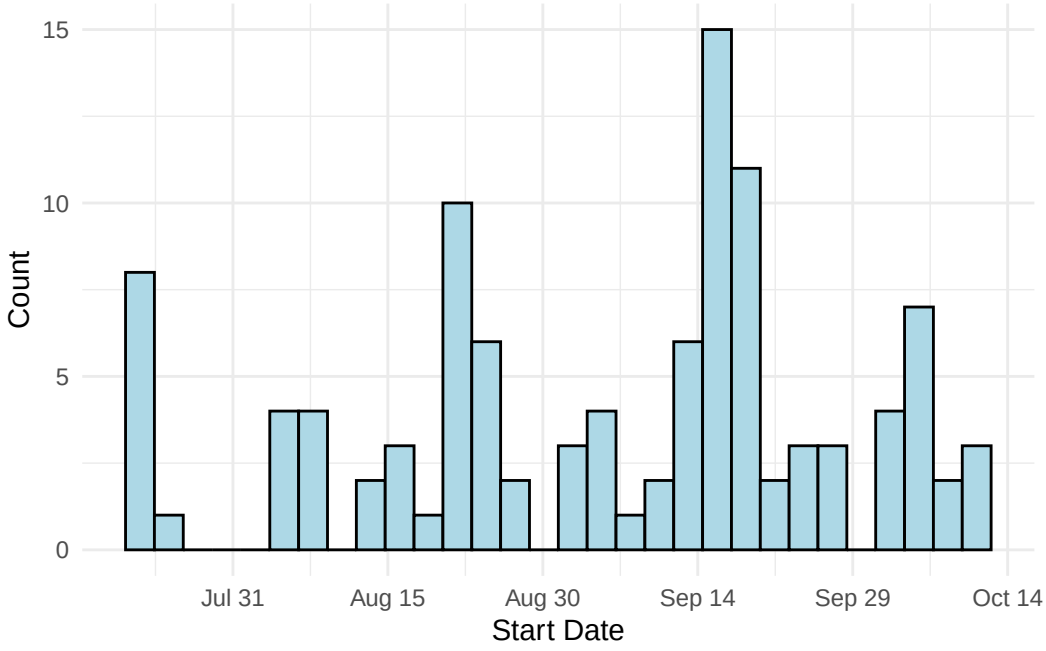

Figure 6: Histogram illustrating the distribution of polling start dates for key battleground states. Displays the frequency of polling activities initiated over time, with a notable peak on September 14, indicating a surge in polling efforts as the election approached.

Figure 6 illustrates the distribution of polling start dates from late July to mid-October 2024, highlighting when polling efforts commenced in key battleground states. The histogram shows peaks in activity during mid-August, early September, and late September, suggesting increased polling during significant campaign events. By including only polls initiated after Harris's campaign launch, we capture voter preferences in a context where both major candidates are actively campaigning. This approach ensures our model focuses on data that accurately reflects the competitive dynamics between Harris and Trump, enhancing the validity of our predictive analysis as Election Day approaches.

3 Model

3.1 Model Overview

The Bayesian Beta regression models in this paper estimate the probability of victory for Donald Trump and Kamala Harris in the 2024 U.S. Presidential Election by analyzing polling data from seven battleground states. Separate models are constructed for each candidate to estimate their support, considering polling trends over time and state-level differences in voter sentiment. This approach captures how support fluctuates leading up to Election Day and predicts potential electoral outcomes. To analyze changes in voter support over time, the models use natural splines with the **'end_date_num'** (days from the earliest **'end_date'** to the latest **'end_date'**) as a predictor, allowing for the identification of non-linear trends influenced by campaign events and news cycles. Additionally, state-level effects are incorporated as random factors, enabling the models to account for unique political landscapes across the battleground states. This regional adjustment helps highlight variations in candidate support and enhances the accuracy of predictions. Further detailed model specifications and diagnostics are provided in Appendix B.

3.2 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.

3.3 Model Setup

Modeling processes are conducted by employing the following packages: `rstanarm` package of Gabry et al. (2023), `brms` package of Bürkner (2023) for Bayesian regression modeling, `Matrix` package of Bates and Maechler (2023) for handling sparse and dense matrices, `splines` package of Team (2023) for regression splines, `caret` package of Kuhn et al. (2023) for machine learning support, and `broom.mixed` package of Bolker et al. (2023) for tidying mixed models.

3.3.1 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 (Bürkner 2023) 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.

3.3.2 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.

3.4 Model Justification

The Bayesian Beta Regression model with a logit link function is the most suitable model for predicting voter support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election. Our primary outcome variable is the percentage of support for each candidate in seven key battleground states, which is naturally bounded between 0% and 100%. The Beta distribution is ideal for modeling this proportion data, as it allows us to avoid unrealistic predictions outside the (0, 1) range. To capture non-linear trends in voter support as Election Day approaches, we employe natural splines with **'end_date_num'**, which represents the number of days from earliest **'end_date'** to latest **'end_date'**. Additionally, incorporating state-specific random effects acknowledges the variations in voter preferences across different regions, allowing for more accurate modeling of baseline support levels. The Bayesian framework enhances our model by enabling the incorporation of prior information, leading to more stable estimates, especially in the context of sparse polling data. Furthermore, the model's structure allows for clear interpretation in terms of victory probabilities, aligning well with our goals of forecasting electoral outcomes. Details are further discussed in Appendix B.

In addition to the Bayesian Beta regression model, other models such as generalized linear models (GLMs) with logistic regression could be employed to predict voter support; however, these may not adequately capture the non-linear trends inherent in polling data. Alternatively, machine learning approaches like random forests or gradient boosting could provide robust predictions by leveraging complex interactions in the data, but they may sacrifice interpretability compared to the Bayesian framework. Ultimately, the choice of model depends on the trade-off between flexibility, interpretability, and the specific characteristics of the data at hand.

3.5 Model Summary

3.5.1 Trump Model Summary

The summary in Table 2 presents fixed effects estimates reflecting the influence of time on Trump's support in battleground states. The intercept, estimated at -0.0923, indicates a slightly lower baseline log-odds of support when other predictors are held constant. The term

nsenddatenumdfEQ42 shows an estimate of 0.0803 with a statistically significant confidence interval (0.0179 to 0.1426), suggesting an increase in support during this period. Similarly, nsenddatenumdfEQ44 has an estimate of 0.0789 and a confidence interval above zero (0.0014 to 0.1589), also indicating a positive effect. In contrast, nsenddatenumdfEQ41 and nsenddatenumdfEQ43 exhibit less significant impacts, suggesting variability in time periods' influence on support. Standard errors for the fixed effects range from 0.0306 to 0.0676, with the lower standard error of 0.0306 for the intercept indicating a stable baseline estimate. However, the higher standard error of 0.0676 for nsenddatenumdfEQ43 reflects less precision in that estimate. Overall, these results imply that Trump's support fluctuated over time, likely in response to specific campaign events, with significant terms indicating changes in voter preference effectively captured by the model. Standard errors further inform the reliability of these estimates.

Table 2: Coefficient estimates from the Bayesian Beta regression model for Donald Trump, showing the impact of the natural spline transformation of end date on voter support. The intercept indicates baseline log-odds, while spline terms reflect nonlinear trends in voter sentiment, with positive coefficients suggesting increased support as Election Day approaches.

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

Overall, these results imply that Trump's support fluctuated over time, likely in response to specific campaign events, with significant terms indicating changes in voter preference effectively captured by the model. Standard errors further inform the reliability of these estimates.

3.5.2 Harris Model Summary

Table 3 similarly shows the impact of time on her support in the battleground states. The intercept is estimated at -0.1508, indicating a lower baseline support level compared to Trump's intercept of -0.0923. The term nsenddatenumdfEQ41 shows an estimate of 0.1225 with a confidence interval of 0.0553 to 0.1903, reflecting a significant increase in support during this period. Another notable effect is observed in nsenddatenumdfEQ43, which has an estimate of 0.1191 and a confidence interval from -0.0071 to 0.2417, suggesting a trend toward increased support. While other terms remain positive, they have less impact. The standard errors for fixed effects range from 0.0279 for the intercept to 0.0631 for nsenddatenumdfEQ43, with the

larger standard error for nsenddatenumdfEQ43 indicating some uncertainty about its influence on support, despite the positive estimate suggesting an upward trend. Compared to the Trump model, the Harris model's standard errors are generally smaller, indicating more consistent support trends over time.

Table 3: Coefficient estimates from the Bayesian Beta regression model for Kamala Harris, illustrating the effects of the natural spline transformation of end date on voter support. The intercept represents the baseline log-odds, while the spline terms indicate nonlinear trends in voter sentiment, with positive coefficients suggesting an increase in support as Election Day nears.

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

These results reveal fluctuations in Harris's support, particularly during the significant periods marked by nsenddatenumdfEQ41 and nsenddatenumdfEQ43, which aid in predicting her probability of victory and potential shifts in support as the election approaches.

4 Results

In this section, we present the analysis of electoral support for Donald Trump and Kamala Harris in key battleground states during the 2024 Presidential Election cycle. We start by examining state-level polling averages for both candidates, providing insights into voter preferences as reported by various pollsters. We then analyze the evolution of these polling averages over time, identifying trends that could affect election outcomes. Following this, we assess the predicted probabilities of Trump winning on Election Day and examine 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. Visualizations of results are generated by employing the following packages: `usmap` package of Albers et al. (2023) for mapping U.S. states, `kableExtra` package of Zhu (2023) for dynamic report generation and table formatting, `modelsummary` package of Arel-Bundock (2023) for summarizing models.

4.1 State-Level Polling Averages for Trump and Harris

Table 4 displays overview of the percentage of support for Trump and Harris by each battleground state. It provides us with a high level understanding of the current voter dynamic. The

percentages of support for Trump and Harris are aggregated averages of all **'pct'** data points for Trump and Harris in each state. By comparing the percentage of support for Trump 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: Percentage of support for Donald Trump and Kamala Harris across key battleground states. Trump maintains a slight lead in states such as Arizona and Georgia, while Harris shows stronger support in 5 of 7 key battleground 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

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 5th, 2024, it could also reflect changes in public opinion due to campaign events, political debates, or other external factors.

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

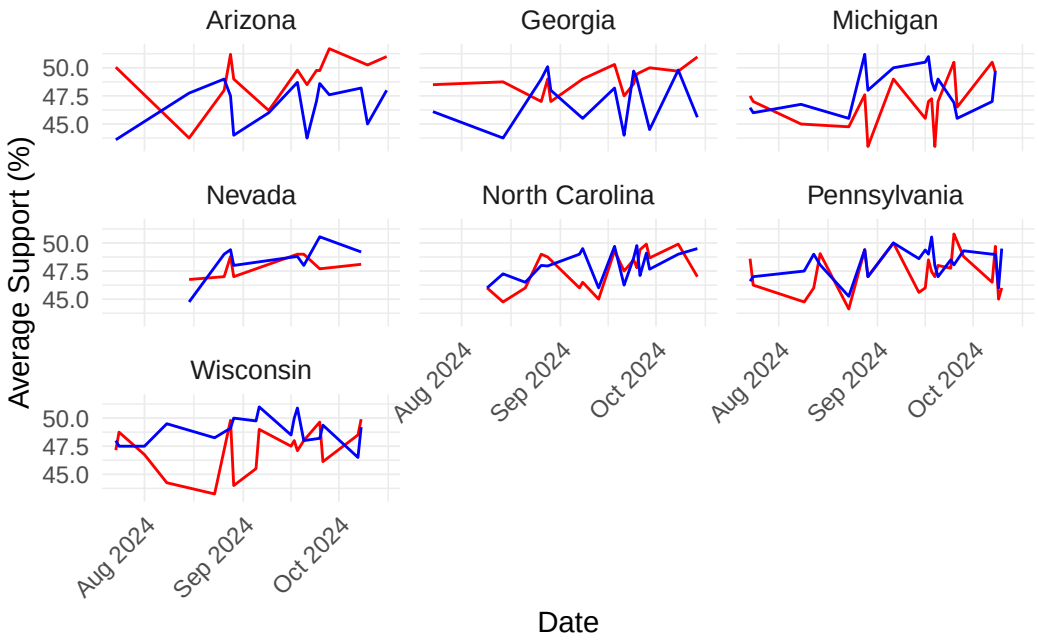

Figure 7: Polling averages for Donald Trump (red) and Kamala Harris (blue) in battleground states from August to October 2024. Captures the varying fluctuations in average support percentages across Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, indicating the competitive dynamics of the election.

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 in Figure 8.

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.

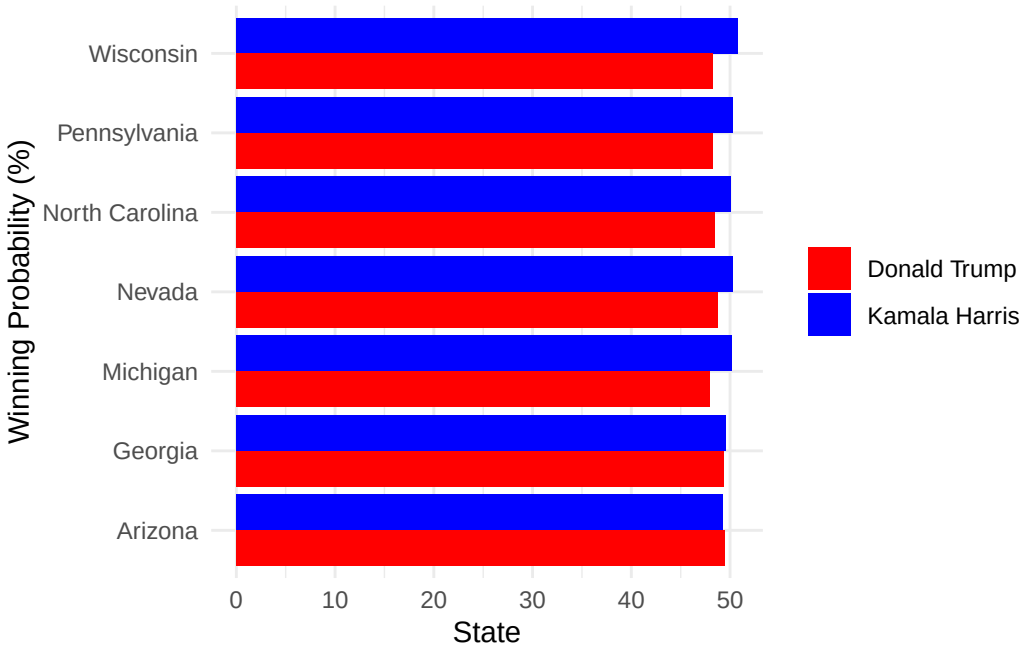

Figure 8: Predicted winning percentages for Donald Trump (red) and Kamala Harris (blue) in key battleground states on Election Day. Illustrates the estimated probability of victory for each candidate across Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, highlighting Harris's projected advantages in the majority of these states.

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: Predicted winning probabilities for Donald Trump and Kamala Harris in key battleground states, with percentages representing the likelihood of each candidate's victory. Summarized predicted support for each candidate in Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, indicating a competitive race, particularly with Harris expected to win in most states by narrow margins.

State	Predicted Trump Win (%)	Predicted Harris Win (%)	Predicted Winner
Arizona	49.51	49.22	Donald Trump
Georgia	49.32	49.56	Kamala Harris
Michigan	47.99	50.19	Kamala Harris
Nevada	48.80	50.25	Kamala Harris
North Carolina	48.48	50.04	Kamala Harris
Pennsylvania	48.24	50.32	Kamala Harris
Wisconsin	48.30	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 (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.

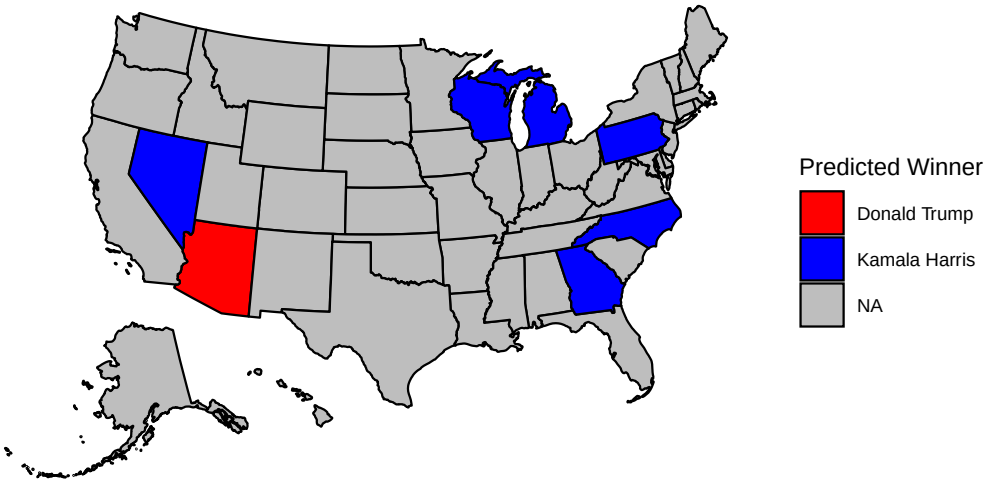

Figure 9: Predicted winners by battleground state on Election Day, illustrating the anticipated electoral outcomes for Donald Trump and Kamala Harris across key states. Highlights states where Trump is expected to win in red and states favoring Harris in blue, with gray indicating states where data is not applicable (NA). Underscores the competitive nature of the election and the distribution of predicted support across the country.

5 Discussion

5.1 Key Findings

This paper presents a forecast of the 2024 U.S, Presidential Election, leveraging Bayesian Beta regression models trained on polling data. We were able to model both the electoral support trend over time and the different political dynamics across the battleground states. The model's predictions reveal that, as Election Day approaches, Kamala Harris holds a lead in six of the seven battleground states–Pennsylvania, Michigan, Wisconsin, Nevada, North Carolina, and Georgia. However, her margins are consistently narrow, with predicted probabilities only slightly above 50% in each of these states. Harris shows minimal leads over Trump, with differences between the probability of winning of no more than 3 percent across all six states. Meanwhile, Trump also holds a narrow advantage in Arizona, suggesting a highly competitive electoral dynamic in this election.

The small margins in the predicted outcomes emphasize the potential volatility of the race; a slight shift in voter sentiment or late-breaking events could easily flip the outcome in these battleground states. This narrow margin of victory suggests a level of electoral fragility for Harris, as her support base may not be solidly locked in. The close race in each state implies that both candidates must maintain their voter engagement and strategic focus, as any shift in these states could have significant implications for the election's overall outcome. Thus, while Harris is favored in the majority of battleground states, the slight leads indicate a need for cautious interpretation, with both campaigns facing high stakes in these key states.

5.2 Real World Implications

If Kamala Harris wins the election by a narrow margin, the results could have significant real-world implications, reflecting the deeply polarized state of the nation. A close race indicates not only a tight political divide but also underscores broader cultural, economic, and social differences among Americans, often manifested through political affiliations. One major challenge for Harris would be governing with such a divided mandate; she could face resistance from nearly half the population, complicating her efforts to implement her agenda. This suggests that any significant policy shifts may encounter strong opposition, potentially deepening divisions rather than bridging them (Yourish and Gamio 2024).

Moreover, a narrow victory for Harris would reinforce national polarization, revealing the extent to which the country is split. The nearly equal support for both candidates highlights entrenched ideological divides, with voters committed to opposing perspectives on critical issues such as healthcare, immigration, climate policy, and the role of government (Heltzel and Laurin 2020). Economic concerns, including job security, wage equality, and tax policy, alongside cultural values related to family, education, and identity, play significant roles in voter alignment. Thus, a close election result suggests that these differences are sharply felt

across the electorate, with each side viewing the outcome as an opportunity to uphold their values and address their unique challenges.

5.3 Uncertainty Beyond Data and Modeling

Beyond data and modeling, external factors such as recency bias, social desirability bias, and non-response bias can influence the election outcome. Recency bias significantly shapes public perception, with many expecting Trump to win due to his 2016 victory and narrow 2020 loss, despite lagging in polls. This bias highlights a tendency to focus on recent events while overlooking broader trends. For example, in 2012, Barack Obama outperformed polls, a fact often overshadowed by recent election narratives (Nate Silver 2024). This can skew how people respond to polls and how pollsters interpret results, introducing potential polling errors that our model cannot fully account for.

Social desirability bias and non-response bias add further uncertainty, especially regarding Trump's support base. Social stigma may deter some voters from expressing conservative support openly, leading to under-reported Trump support in polls. Additionally, Trump's supporters often have lower social trust, making them less likely to participate in polls altogether (Nate Silver 2024). If certain groups systematically avoid participation, polling data may under-represent these voters, resulting in an underestimation of Trump's support in our model. On the other hand, Harris faces potential challenges such as the Bradley effect and gender-related bias. As a Black female candidate, Harris may encounter undecided voters leaning against her due to underlying biases, similar to what Hillary Clinton faced in 2016 (Nate Silver 2024). If undecided voters ultimately vote against Harris, our model may overestimate her support based on polling data alone, affecting prediction accuracy.

These biases are difficult to measure or correct, reflecting deeper societal attitudes and complexities in voter behavior. Although pollsters may attempt adjustments, these are often based on assumptions that cannot fully capture the dynamics at play. Therefore, our model's predictions should be interpreted cautiously, acknowledging that biases may lead to unexpected shifts beyond the range of data-driven forecasts.

5.4 Limitations and Weaknesses

Although the model is suitable and performs well on achieving the goal of predicting the outcome, it has several limitations that could be addressed in future studies. Here are some key areas:

- **Pollster Ratings and Credibility**: Polls from different organizations often vary in reliability, as some pollsters have more accurate or less biased methodologies. By integrating pollster ratings, the model could assign weights to different polls based on the historical accuracy and reputation of each organization. This adjustment would allow

the model to downplay the influence of potentially biased or less credible polls, particularly in competitive battleground states, where small differences in predicted support can significantly affect the outcome. This approach would lead to a more calibrated model that accounts for poll variability and thus provides a more reliable prediction.

- **Timing and Recency of Polls**: Voter sentiment is not static, and the timing of polls can significantly impact their relevance. As the election date nears, polls generally become more reflective of the final outcome. Incorporating a temporal weighting scheme that prioritizes recent polls could enhance the model's responsiveness to last-minute shifts in voter preferences, which often occur due to campaign events, debates, or emergent political issues. This weighting approach could also help mitigate the risk of using outdated data that no longer accurately reflects current voter sentiment.
- **Economic and Political Context**: The political landscape is shaped by various external factors, such as economic indicators (e.g., unemployment rates, inflation) and prominent political events. By incorporating economic data or major political issues, such as health care or immigration policies, the model could capture contextual variables that influence voter sentiment. Such contextual factors could be especially useful in a Bayesian framework, allowing us to adjust prior distributions based on known historical influences on voter behavior.
- **Inclusive of National/Popular votes**: Our model focuses exclusively on the seven battleground states—Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin—omitting data from solidly Democratic or Republican states and the national popular vote. This exclusion limits the model's ability to capture broader voter sentiment across the country, restricting its capacity for a holistic prediction of the 2024 Presidential Election. While battleground states are crucial due to the Electoral College system, popular vote trends from other states can indicate national shifts that may indirectly influence battleground dynamics. For example, a nationwide trend favoring or opposing a candidate could signal political momentum or reflect impactful events resonating with voters in battleground states. Without national data, our model may overlook these broader trends, potentially affecting the accuracy of its predictions in key states.

5.5 Implications for Future Modeling

For future modeling, incorporating data from a broader range of states or using a weighted approach that balances battleground states with national polling data could address these limitations. For example, incorporating popular vote trends and data from states with similar demographics or political contexts could help the model detect and account for national sentiment shifts. Additionally, adding more factors, such as ratings for pollsters, demographic details, and economic conditions, could help the model capture different influences on how people vote. Giving more weight to recent polls would also make the model more responsive to last-minute changes in public opinion as Election Day gets closer. By expanding the data

and including a wider range of details, future models could better balance the unique characteristics of battleground states with trends across the nation, resulting in a stronger and more reliable prediction tool for elections.

Appendix

A Additional Data Details

A.1 Data Cleaning

The data cleaning process for our analysis focused on refining the raw polling data obtained from Project 538's polling project (FiveThirtyEight 2024). We concentrated on seven key battleground states—Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin—recognized as pivotal for this year's election outcome. Key steps included filtering relevant columns such as state, polling dates, pollster information, numeric grades, transparency scores, candidate names, and support percentages to eliminate unnecessary data. We set thresholds for numeric grades (minimum of 2.5) and transparency scores (greater than 5) to ensure the inclusion of high-quality polling data. Polling dates were standardized, and we created a numeric variable representing the number of days since the earliest polling date. Additionally, we addressed data inconsistencies, like replacing "Nebraska CD-2" with "Nebraska," and aggregated polling percentages by averaging candidate support while preserving essential attributes. The cleaned dataset was saved in parquet format for efficient processing in subsequent analyses.

To ensure the integrity of our cleaned data, we implemented a comprehensive data testing process. This involved verifying the presence of required columns, checking data types for consistency, and ensuring no missing values in critical fields. We validated that candidate support percentages fell within the logical range of 0 to 100 and confirmed that numeric grades and transparency scores met our predefined thresholds. By rigorously applying these tests, we established that the data is reliable and suitable for modeling and analysis, providing a solid foundation for exploring voter behavior in the selected battleground states.

A.2 Limitations and Future Directions

To improve data selection for future studies, data from non-battleground states could add valuable context and enhance model robustness by reflecting a range of voter behaviors and regional influences. While these states may not directly impact the electoral outcome due to their predictability, understanding regional dynamics across all states could help the model detect shifts in voter sentiment that cut across state lines. For instance, trends in nearby states or those with similar demographic or economic characteristics might mirror those in battleground states, providing additional insights into potential voting patterns. Including these states in the model could also reveal hidden correlations and strengthen the model's predictive power by capturing a wider range of political and demographic factors. The focus on battleground states also means that the model may amplify the idiosyncrasies of these seven states without balancing them against broader national trends. Each battleground

state has unique political, economic, and demographic characteristics, and relying solely on these states could cause the model to disproportionately weigh regional dynamics that may not be representative of the country as a whole. As a result, predictions based only on these states might overestimate or underestimate certain voting behaviors, particularly if the election outcome depends on trends that are not captured within these seven states alone.

B Additional Model details

B.1 Model Justification - Further Details

The Bayesian Beta regression model with a logit link function is well-suited for predicting the 2024 U.S. Presidential Election due to the nature of our outcome data and analysis goals. Our primary outcome variable is the proportion of support for each candidate in seven battleground states, scaled between 0 and 1. The Beta distribution is ideal for modeling such bounded data, unlike linear regression, which assumes an unbounded outcome and risks unrealistic predictions outside the 0-100% range.

To capture non-linear trends in voter support over time, we use natural splines with the variable **'end_date_num'** (days since the earliest poll date), allowing the model to adapt to shifts due to campaign events or debates. Specifying four degrees of freedom for the splines balances flexibility in modeling gradual changes with reducing noise or overfitting. Standard logistic regression lacks this adaptability and would be less effective in capturing the evolving nature of voter sentiment. Additionally, incorporating state as a random effect enables the model to estimate state-specific baseline support levels, reflecting the distinct political dynamics of each battleground state while preserving overall trend analysis. This approach prevents overfitting by avoiding an overparameterized model, particularly useful when polling data for some states is sparse. The random effect structure thus enhances model robustness and generalizability across states with varying polling frequencies.

The Bayesian approach adds value by incorporating prior information, stabilizing parameter estimates where data are limited. We applied weakly informative priors to ensure data-driven estimation: the intercept ($\beta_0 \sim \text{Normal}(0, 10)$) and slope ($\beta_1 \sim \text{Normal}(0, 5)$). This setup supports stable inference, crucial in states or periods with sparse polling data. This model structure provides clear interpretability, essential for comparing the probability of victory for Trump and Harris. The logit link function facilitates results interpretation in terms of odds or victory probabilities, aligning with our election analysis goals. Constructing separate models for each candidate allows for direct comparisons of trends and state-level variations without confounding. Thus, the Bayesian Beta regression model with a logit link function is justified as the optimal choice. It captures non-linear temporal trends, accounts for regional variations, and delivers interpretable victory probabilities, supported by Bayesian inference that enhances reliability and reflects uncertainty in forecasts.

B.2 Model Diagnostics

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. These checks are essential for evaluating the models' ability to replicate observed polling data based on the estimated parameters. By comparing the distributions of observed support percentages with the predicted values, we can identify discrepancies and assess the overall fit of our models. Posterior predictive checks help verify model assumptions and provide insights into how well the models capture the underlying voter dynamics. A close alignment between observed and predicted values suggests that our models effectively reflect voter support trends in the electoral context. In contrast, significant deviations may reveal areas for model improvement or highlight limitations in our approach. This diagnostic process aims to enhance the credibility of our findings and deepen our understanding of the electoral landscape as the 2024 Presidential Election approaches.

B.3 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.2. 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.

B.3.1 Posterior Predictive Check for the Harris Model

The posterior predictive check for the Harris model illustrates the relationship between the observed support percentages y and the predicted support values y_{rep}. Figure 11 displays a

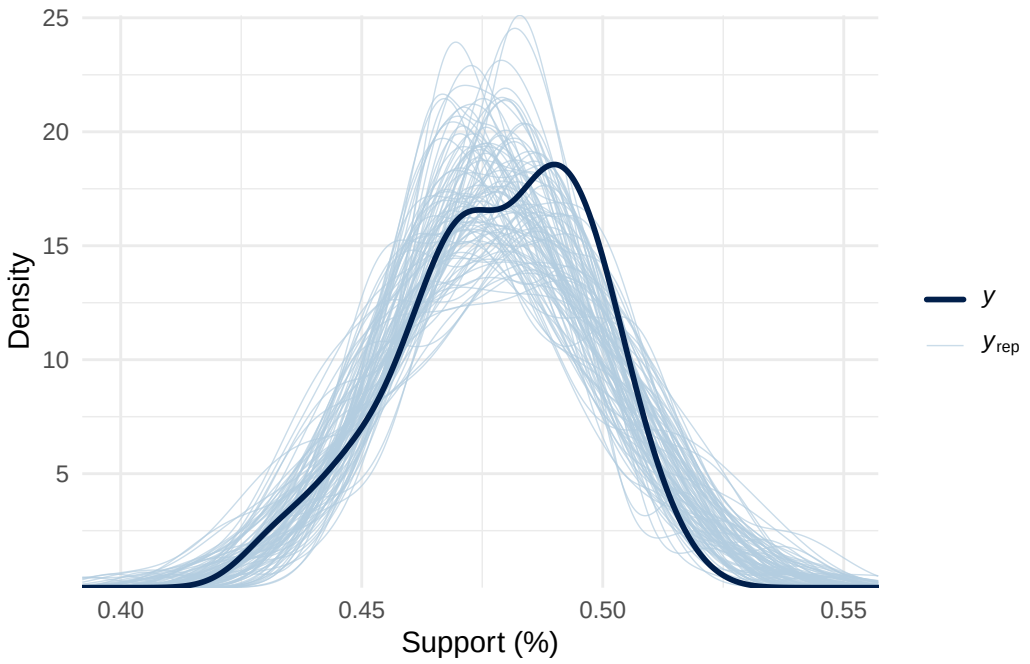

Figure 10: Posterior predictive check for Donald Trump's model, illustrating the distribution of observed support percentages y alongside the predicted support values y_{extrep}. The graph shows a strong overlap between the observed and predicted distributions, indicating that the model accurately captures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

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.

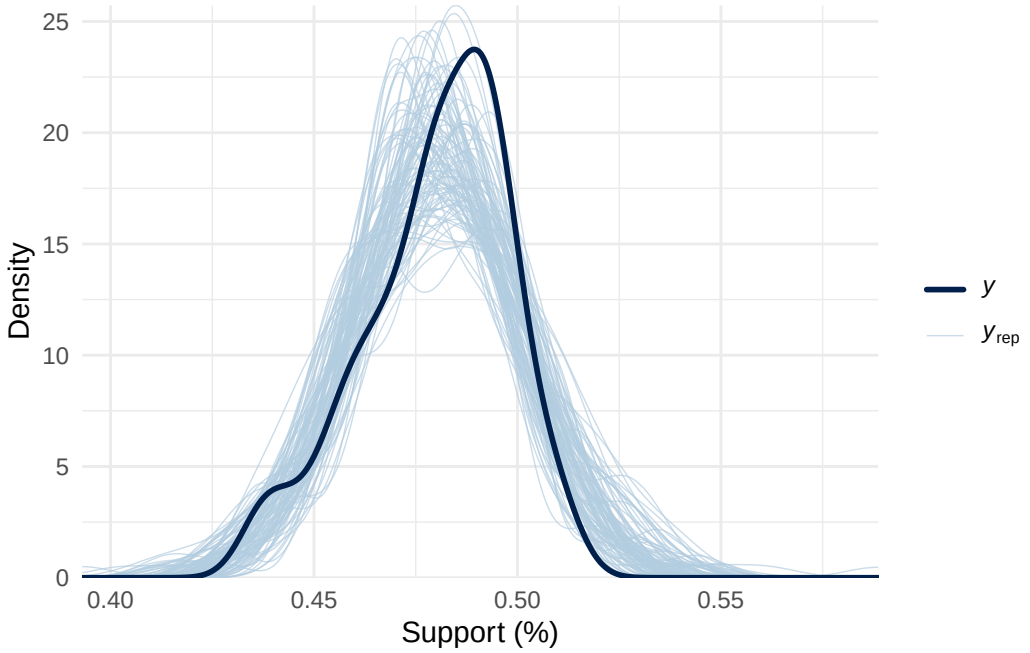

Figure 11: Posterior predictive check for Kamala Harris's model, illustrating the distribution of observed support percentages y against the predicted support values y_{extrep}. The graph shows a strong correspondence between the observed and predicted distributions, indicating that the model accurately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

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.2.

B.3.2 Model Performance

Based on the posterior predictive checks performed for both Trump and Harris models, we are able to make the following conclusions on the performance and reliability of our Bayesian regression models. The strong overlap of the observed distribution of support percentages and the predicted distribution from posterior samples suggests that the models are well-calibrated, meaning they accurately reflect the variability and central tendency of voter support for Trump and Harris, respectively. This alignment is particularly important given the fluctuating nature of polling data and the importance of accurately capturing voter sentiment as we approach Election Day.

In the context of our paper, these predictive checks provide evidence of the model's robustness and reliability in forecasting. Given our goal of predicting the likelihood of each candidate's victory in key battleground states, it is important that our models can not only predict the general trend in support but also approximate the changes in voter sentiment captured by the polling data. The successful predictive checks validate our choice of a Bayesian Beta regression model with a logit link function, demonstrating that this approach is indeed suitable for analyzing proportion data constrained within the 0-1 range.

Furthermore, these results support the use of state-level random effects and time-based splines, as the models are shown to accommodate both state-specific variations and temporal shifts in support. By accurately modeling these elements, our predictions provide more reliable insights into which candidate holds an advantage in each state. This model performance enhances confidence in our forecasts, suggesting that our predictions can serve as a meaningful tool for understanding electoral dynamics in these critical regions. These findings suggest the model's capacity to provide robust, data-driven projections of the election outcome, while also accounting for the inherent uncertainty and volatility within polling data.

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

We will initiate our survey by sending invitations to the 2,000 selected voters via email. To enhance response rates and reduce non-response bias, we will send a reminder email three days later. 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. To effectively manage the recruitment process, we will allocate `$100,000` of the budget for staffing. We will employ a small team of around 30 people responsible for sending invitations during the first three days and following up with non-respondents over the entire two-week process. Each employee will be paid `$24` per hour for 40 hours per week for two weeks, resulting in a total of 40 hours worked per employee and `$960` for each, `$28,800` in total. Additionally, we will allocate `$86,000` for the phone follow-up process as well as the data cleaning and analysis process. We will employ a team of 20 data scientists to perform data engineering and analysis, with them being paid $3560 each for the duration of their work term, which will be around three to five weeks.

Overall, this recruitment method is designed to be both efficient and effective. By reaching out to a diverse group of voters and encouraging participation through personalized follow-ups, we can collect high-quality data that accurately reflects the opinions and behaviors of the broader

population. Additionally, this recruitment method minimizes early-stage labor costs by using email as the initial contact method, reserving more follow-up efforts for those who have not yet responded after a week. This approach not only helps a higher participation rate but also improves the reliability of the data collected.

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

- 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.

- 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.
- 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

- 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.
- 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.
- 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. We aim to align these methods closely with real-world considerations to ensure it is efficient and represents the views of a broad population well, within a budget of $100,000. To demonstrate our idealized methodology, the following survey is generated, which can be accessed via the the URL provided in (Google 2024) under References D.6.2.

D.1 Target Population, Sampling Frame, and Sample

- **Target population**: The target population consists of all registered voters across the United States who are eligible to vote in the upcoming election. Since the focus is on achieving state-level accuracy, this population includes individuals across all states, spanning various demographic groups, household statuses, and voting behaviors.
- **Sampling frame**: The sampling frame is the voter file containing demographic and historical voting information about registered voters. This voter file is likely provided by a third-party data vendor (such as L-2 or a similar organization) and includes essential information like age, gender, race/ethnicity, household status, party registration, and voting history. The voter file ensures access to a comprehensive, up-to-date list that reflects the characteristics of the registered voting population.
- **Sample**: The sample consists of individuals selected from each stratum within each state. We will select 100 respondents from each stratum identified in the sampling frame, ensuring representation across key categories such as age groups, gender, race/ethnicity, household status, and party affiliation. By carefully selecting individuals from each stratum, we ensure that the sample reflects the voting intention of the target population, enabling us to make accurate predictions about voter behavior.

D.2 Sampling Methodology & Justification

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 demographics 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.

This stratified sampling approach ensures that our sample accurately reflects key demographics, essential for making reliable state-level predictions. By segmenting the sample by state and further by demographic groups—such as age, gender, race/ethnicity, household status, voting history, and party registration—we ensure that our sample reflects the key characteristics of the population. This method not only enhances the reliability of our state-level predictions, which is crucial due to the Electoral College system but also optimizes resources to be cost-effective to capture all significant voter demographics within budget.

D.3 Sample Recruitment Method

D.4 Survey Design

The survey will be designed to capture critical insights into voter preferences, opinions, and behaviors relevant to the 2024 election. It contains questions on candidates, key political issues, and demographic information, enabling a detailed analysis of trends within each stratum. To reduce survey fatigue and encourage high-quality responses, we prioritized concise, clear questions. To ensure inclusivity and ease of response, the survey will be accessible online and formatted for both desktop and mobile use.

The survey questions capture both quantitative and qualitative data that reflect voter intentions and behavior. By including a range of topics—from candidate preference to issue prioritization and engagement with political news. Each question is intended to collect actionable information that contributes to understanding voter stability and sentiment, which are all key in forming reliable predictions for the election. The scale-based approach allows respondents to express the intensity of their preferences, providing more nuanced data than simple yes-or-no answers. The questions are direct and structured to maintain respondent engagement, minimizing the risk of survey fatigue. Additionally, by setting a hypothetical match-up between prominent candidates- Kamala Harris and Donald Trump, we introduce a realistic context that grounds responses in potential voting scenarios, which adds to the predictive power of our data. Also, a validation question is asked to ensure our data remains reliable and accurate. It measures if the responder maintains integrity and attentiveness.

D.5 Data Validation

After completing the survey, we will adjust the data through weighting to ensure that our samples accurately reflect the broader population for predictive purposes. Each stratum will be assigned a weight proportionally based on its size. For each group, we will calculate the weight using the formula:

$$\text{Weight} = \frac{\text{Population Proportion}}{\text{Sample Proportion}}$$

This approach assigns higher weights to underrepresented groups and lower weights to overrepresented groups, ensuring that each group's influence in the sample aligns with its actual population (Kalton and Flores-Cervantes 2003). For instance, if young voters (ages 18-29) constitute 30% of the population but only make up 20% of our survey sample, this demographic is considered underrepresented. In this case, we would assign a larger weight to young voters, increasing their influence to better reflect their true proportion in the population. Additionally, extra weight will be given to respondents from strata who are less likely to participate in surveys. To avoid duplicate responses, each voter will be assigned a unique ID, retaining only the first response from each ID. To ensure that 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.

D.6 Survey Demo

Title: "2024 U.S. Presidential Election Survey"

D.6.1 Survey Introduction

We are conducting an important survey to understand voter opinions and predict the outcome of the upcoming 2024 U.S. Presidential Election. Your participation will help ensure that our findings accurately reflect the views of voters like you.

This survey will take about 5–10 minutes to complete. Your responses will remain completely anonymous and confidential. We are committed to protecting your privacy, and no personal information will be shared or used for any purpose beyond this research.

Thank you for your time and contribution.

D.6.2 Survey Questions

1. **Thinking ahead to the presidential general election, are you going to vote?**
   On a scale from 1 (Not at all likely) to 10 (Almost certain).
   - 1 (Not at all likely)
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7
   - 8
   - 9
   - 10 (Almost certain)
2. **Who do you want to vote for?**
   - Kamala Harris (Democratic Party)
   - Donald Trump (Republican Party)
   - Prefer not to answer
3. **How likely are you to change your mind before the election?**
   On a scale from 1 (Very unlikely) to 10 (Almost certain).

- 1 (Very unlikely)
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10 (Almost certain)

4. **Which of the following issues is most important to you in this election?**

   - Economy
   - Healthcare
   - Immigration
   - Climate Change
   - Safety
   - Education
   - Other (please specify)

5. **If you have to decide between Kamala Harris and Donald Trump, who will you vote for?**

   - Kamala Harris
   - Donald Trump
   - Prefer not to answer

6. **Who do you expect to win the upcoming elections?**

   - Kamala Harris
   - Donald Trump
   - Prefer not to answer

7. **How often do you follow political news?**
   On a scale from 1 (Rarely) to 10 (Daily).

   - 1 (Rarely)
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7
   - 8
   - 9
   - 10 (Daily)

8. **Would you be open to commenting on the issues in this survey and be interested in being contacted by a reporter?**

    - Yes
    - No

9. **For validation purposes, please select "Agree" from the options below.**

    - Agree
    - Disagree

References

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

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data*. https://CRAN.R-project.org/package=modelsummary.

Bates, Douglas, and Martin Maechler. 2023. *Matrix: Sparse and Dense Matrix Classes and Methods*. https://CRAN.R-project.org/package=Matrix.

Bolker, Ben et al. 2023. *broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Bürkner, Paul-Christian. 2023. *brms: Bayesian Regression Models using 'Stan'*. https://CRAN.R-project.org/package=brms.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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.

Grolemund, Garrett, and Hadley Wickham. 2011. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Heltzel, Gordon, and Kristin Laurin. 2020. "Polarization in America: Two Different Futures." 2020. https://www.sciencedirect.com/science/article/pii/S2352154620300450.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. "Weighting Methods." Statistics Sweden. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/weighting-methods.pdf.

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

Nate Silver. 2024. "Opinion | Election Polls and Results for Trump and Harris." 2024. https://www.nytimes.com/2024/10/23/opinion/election-polls-results-trump-harris.html.

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

Richardson, Neal et al. 2023. *Arrow: Integration to Apache Arrow for Data Frames*. https://CRAN.R-project.org/package=arrow.

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

Team, R Core. 2023. *splines: Regression Spline Functions and Classes*. https://stat.ethz.ch/R-manual/R-devel/library/splines/html/00Index.html.

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. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Yourish, Karen, and Lazaro Gamio. 2024. "How Republicans Reacted to Harris and Biden's Campaign Messages." 2024. https://www.nytimes.com/interactive/2024/07/24/us/politics/harris-biden-republican-reactions.html.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.