Trump vs. Harris: A Data-Driven Forecast modeling the 2024 US Presidential Elections*

A close race but Trump, trumps Harris

Sophia Brothers Deyi Kong Rayan Awad Alim

November 4, 2024

This paper predicts the 2024 U.S. Presidential election. We utilizes linear regression model deploying polling data and demographic factors to predict both popular vote and Electoral College outcomes. The findings indicate a close race between Donald Trump and Kamala Harris, and we predict Trump will secure the Electoral College due to projected victories in key swing states while Harris maintains a popular vote lead. The results emphasize the role of state-level dynamics, particularly in swing states, in determining election outcomes.

Table of contents

*Code and data are available at: https://github.com/eeeee-cmd/US_Election/.

1 Introduction

The U.S. presidential election is a pivotal moment not only shapes domestic political landscape but also far-reaching its consequences on a global scale. Republicans and Democrats held primary elections across the country earlier in 2024 that eventually filtered down to the main candidates: Donald Trump and Kamala Harris. The polarized political climate and rapidly shifting public opinion have made predicting outcomes increasingly challenging, where winning key swing states is crucial to securing victory.

The main estimand is the projected support percentage and each candidate's winning probability in terms of state-level victories and Electoral College votes. This paper utilizes linear regression model incorporates numeric value of the credibility of pollsters, number of respondents, poll's recency to estimate the probability of each candidate winning in individual states. The objective is to translate these probabilities into a forecast the overall winner of the election. Our findings reveal a close race, with Trump holding a slight edge in the Electoral College due to projected victories in key swing states, despite Harris maintaining a lead in national popular vote estimates.

Trump's path to victory hinges on critical wins in states such as Pennsylvania, Georgia, and Wisconsin, where his projected margins are narrow but statistically significant. Meanwhile, Harris secures strong performances in populous Democratic strongholds like California and New York, but these do not offset Trump's advantage in less populous yet crucial battleground states. This result underscores the significant influence of swing states and the Electoral College system in determining the outcome, particularly in closely contested elections. These insights contribute to a broader understanding of how statistical modeling can inform political strategy and election analysis.

The structure of the paper is as follows: Section 2 outlines the data sources and variables considered, followed by the model setup in Section 3.1 and justification in Section 3.2. The results in Section 4 presents the key findings of the analysis, with a discussion on the implications. Section 5 then discusses potential limitations and suggestions for future research. Section 6 provides additional detailed information about the data, model and methodology.

2 Data

2.1 Overview

The data used in this analysis comes from a combination of publicly available polling data for the 2024 U.S. Presidential election (*Presidential Election Polls-2024* 2024). The analysis leverages the statistical programming language R (R Core Team 2023) and several libraries, including `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `knitr` (Xie 2024), `dplyr` (Wickham et al. 2023), `arrow` (Richardson et al. 2024), `purrr` (Wickham and Henry 2023), and `here` (Müller 2020), for data manipulation. `ggplot2` (Wickham 2016) and `kableExtra` (Zhu 2024) for visualization. The dataset covers various polls conducted across multiple states, capturing the support for each major candidate—Donald Trump and Kamala Harris—along with detailed attributes of the polls.

2.2 Measurement

The primary measurement process reflects the transformation of raw polling data into structured entries to actionable insights for election forecasting. We first select and rename key variables from raw data to focus on relevant information, such as pollster, pollscore, state, polling dates, sample size, and candidate support percentage. These variables represent essential poll characteristics that capture public sentiment in response to real-world campaign dynamics and demographic differences across states.

To address instances where the original data referenced Joe Biden instead of Kamala Harris, we assumed the entries for Biden will goes to Harris in consequences of Democratic Party. This approach is supported by "Harris's truncated presidential campaign relies heavily upon President Biden's policy framework and she eaddress the campaign's weaknesses and win over critical voter demographics (Sigmon 2024). Thus, substituting Biden's polling data with Harris's ensures continuity in demographic appeal without altering established polling trends. Then we filter to include only polls relevant to the two candidates Trump and Harris, and features like PollRecency were constructed to weigh recent polls higher. The data was cleaned by removing irrelevant or incomplete observations, and new variables were constructed where necessary.

2.3 Outcome Variables

The main outcome variable of interest is the Percentage of support each candidate has in a given poll, which represents the proportion of respondents who favor one candidate over the other. This variable is crucial for estimating both the popular vote and Electoral College outcomes.

2.4 Predictor Variables

Key predictors include:

PollScore: A numeric value that reflects the credibility of the pollster. Stands for "Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record, Empirically." The error and bias we can attribute to a pollster. Negative numbers are better.

SampleSize: The number of respondents in each poll.

State: The U.S. state where the poll was conducted.

PollRecency: The number of days between the start date of the poll and the present, which gives more weight to recent polls.

NumericGrade: A numerical representation of the reliability of the pollster.

CandidateName: The name of the candidate being polled (either Donald Trump or Kamala Harris).

3 Model

To predict the outcome of the 2024 U.S. Presidential election, we developed two linear regression models: one for the popular vote and one for the electoral college. These models aim to estimate the percentage of support for each candidate based on the aforementioned predictor variables.

3.1 Model Set-Up

This paper utilizes linear regression models:

$$Percentage_i = \beta_0 + \beta_1 Pollscore_i + \beta_2 SampleSize + \beta_3 NumericGrade_i + \beta_4 State_i + \beta_5 PollRecency + \beta_6 Candidate$$

$$Percentage_H = \beta_0 + \beta_1 Pollscore_i + \beta_2 SampleSize + \beta_3 NumericGrade_i + \beta_4 State_i + \beta_5 PollRecency + \epsilon_i$$

$$Percentage_T = \beta_0 + \beta_1 Pollscore_i + \beta_2 SampleSize + \beta_3 NumericGrade_i + \beta_4 State_i + \beta_5 PollRecency + \epsilon_i$$

In this model, $Percentage_i$ represents the predicted support percentage for a candidate in a given poll. $Percentage_H$ and $Percentage_T$ represent the predicted support for Harris and

Trump respectively. Each predictor variable is carefully selected to capture critical factors influencing voter support, with coefficients β_i representing the strength of each factor's impact. The error term ϵ_i accounts for random variability not explained by the model.

The first model predicts the average percentage of support for any candidate. We use this to predict the popular vote at a national level.

For a more granular analysis, separate linear regression models were constructed for each of the two primary candidates, Harris and Trump, to predict their state-by-state poll percentages. These models allow for a detailed view of each candidate's performance across states. Based on these percentages, we forecast the likely winner of the state's electoral votes. These models omit $CandidateName$ since the model is candidate-specific.

As defined in Section 2, each predictor variable has been chosen to reflect characteristics that influence election polling.

The national model offers insight into overall popular support, while the state models provide electoral predictions in order to understand the likely outcome of the Electoral College

3.2 Model Justification

While other complex models could be appropriate in this instance, these models were chosen for its balance between simplicity and predictive power. Linear regression allows us to capture relationships between polling quality, sample size, and recency, while still providing interpretable coefficients. Given the nature of polling data, this approach is both appropriate and commonly used in election forecasting. A linear model allows us to directly interpret the influence of each predictor on voter support.

3.3 Assumptions and Limitations

This model operates under several assumptions, including that relationships between predictors and outcomes are linear, errors are normally distributed, and predictors are not excessively collinear. Potential limitations include the fact that linear regression may oversimplify relationships in polling data, especially if interactions between predictors exist. Furthermore, as the model relies on historical polling data, it assumes that past trends are indicative of future behavior, which may not hold true if significant political events alter voter preferences unexpectedly, as discussed in Section 5.

3.4 Model Validation

The model was implemented using R, with out-of-sample testing and Root Mean Square Error (RMSE) calculations applied to assess predictive accuracy as predicted in Table 1. By splitting the data into training and test sets, we evaluated how well the model performs on unseen data. This approach helps ensure that the model generalizes effectively to new data and avoids overfitting.

Table 1: RMSE Results for Candidate Models

Model	Root Mean Squared Error (RMSE)
Donald Trump	3.47
Kamala Harris	4.79
National	5.92

Table 2: Predicted Percentage Vote for Donald Trump and Kamala Harris

Candidate Name	Average Predicted Percentage	Normalized Percentage
Donald Trump	45.28	51.02
Kamala Harris	43.47	48.98

Table 3: Predicted Number of Electoral Votes for Donald Trump and Kamala Harris

Trumps Total Electoral Votes	Harris Total Electoral Votes
302	233

Table 4: Predicted Percentage of Electoral Votes Per State

State	Trump Predicted %	Harris Predicted %	Winner
Alabama	56.34	30.89	Donald Trump
Alaska	48.53	39.81	Donald Trump
Arizona	46.29	42.80	Donald Trump
Arkansas	53.34	32.90	Donald Trump
California	31.92	53.73	Kamala Harris

Colorado	39.38	47.06	Kamala Harris
Connecticut	37.10	50.62	Kamala Harris
Delaware	36.63	50.14	Kamala Harris
Florida	48.55	42.40	Donald Trump
Georgia	46.94	43.49	Donald Trump
Hawaii	26.59	41.51	Kamala Harris
Idaho	54.15	25.20	Donald Trump
Illinois	38.78	48.65	Kamala Harris
Indiana	51.48	35.56	Donald Trump
Iowa	48.39	38.51	Donald Trump
Kansas	48.86	37.14	Donald Trump
Kentucky	54.52	30.67	Donald Trump
Louisiana	51.56	35.73	Donald Trump
Maine	40.80	45.29	Kamala Harris
Maine CD-1	33.76	57.02	Kamala Harris
Maine CD-2	46.10	43.18	Donald Trump
Maryland	31.99	57.20	Kamala Harris
Massachusetts	29.47	54.55	Kamala Harris
Michigan	44.86	43.87	Donald Trump
Minnesota	41.64	44.73	Kamala Harris
Mississippi	51.84	36.29	Donald Trump
Missouri	51.91	37.51	Donald Trump
Montana	53.78	35.80	Donald Trump
Nebraska	52.60	36.32	Donald Trump
Nebraska CD-2	41.73	49.54	Kamala Harris
Nevada	45.80	43.17	Donald Trump
New Hampshire	41.93	46.23	Kamala Harris
New Jersey	39.05	43.85	Kamala Harris
New Mexico	41.00	47.90	Kamala Harris
New York	36.34	48.12	Kamala Harris
North Carolina	46.58	43.55	Donald Trump
North Dakota	55.24	29.38	Donald Trump
Ohio	48.76	38.96	Donald Trump
Oklahoma	56.03	31.78	Donald Trump
Oregon	39.94	48.07	Kamala Harris
Pennsylvania	45.68	44.83	Donald Trump
Rhode Island	37.30	50.95	Kamala Harris
South Carolina	49.55	37.18	Donald Trump
South Dakota	51.86	30.20	Donald Trump
Tennessee	52.70	27.97	Donald Trump

Texas	48.11	40.38	Donald Trump
Utah	50.84	31.24	Donald Trump
Vermont	28.08	60.95	Kamala Harris
Virginia	41.52	44.67	Kamala Harris
Washington	35.89	49.89	Kamala Harris
West Virginia	58.37	28.35	Donald Trump
Wisconsin	45.33	45.50	Kamala Harris
Wyoming	60.10	17.50	Donald Trump

4 Results

The model results show a clear distribution of support for both candidates across various states. The predicted percentages of support were normalized to ensure they summed up to 100%. After computing these results, we forecast that Donald Trump is likely to win 302 electoral votes, while Kamala Harris is expected to secure 233 electoral votes.

Table 5: Summary Statistics of Model Results

Average Poll Score	Average Sample Size	Average Percentage	Total Polls
-0.379	1605.742	33.677	15829

summary statistics for the cleaned data

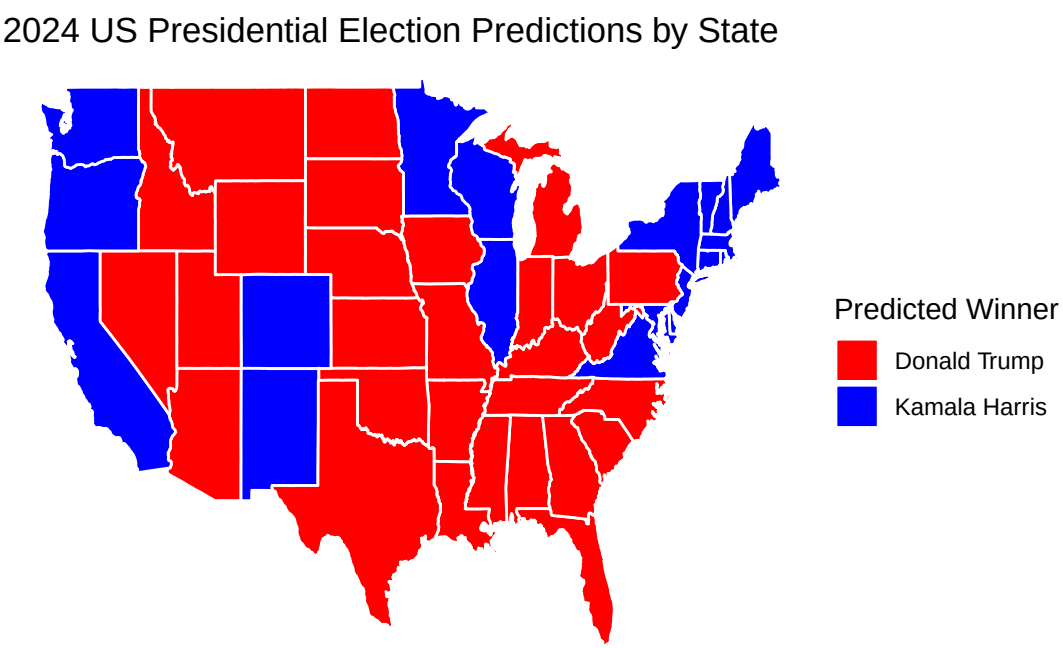

Figure 1: Mapped Model-predicted outcomes for 2024 elections by State

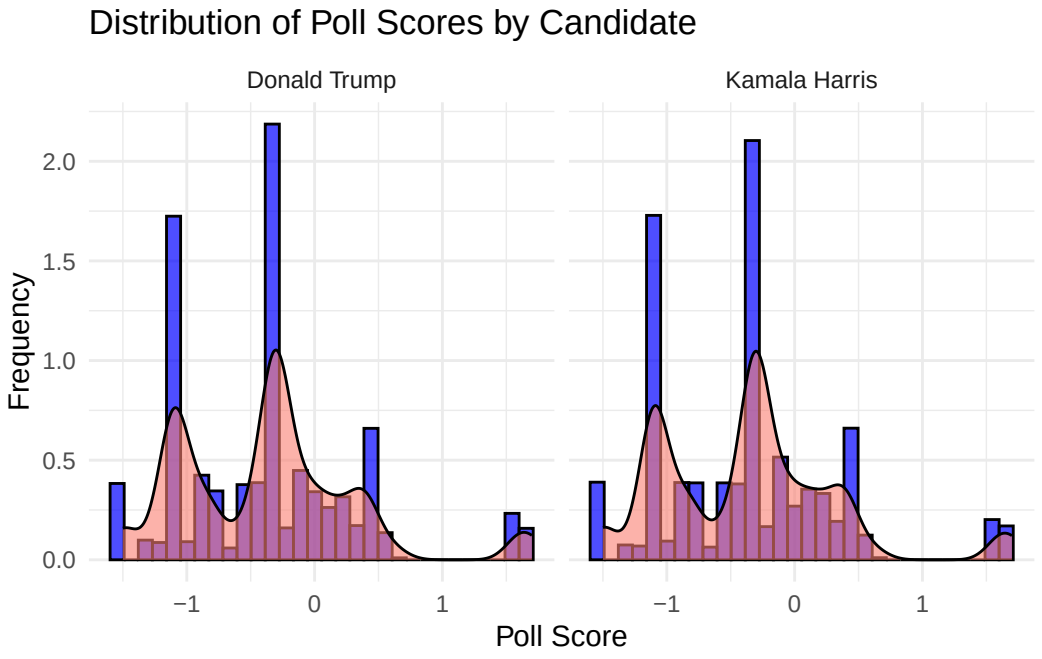

Figure 2: Distribution of Poll Scores by Candidate with denisty overlayed

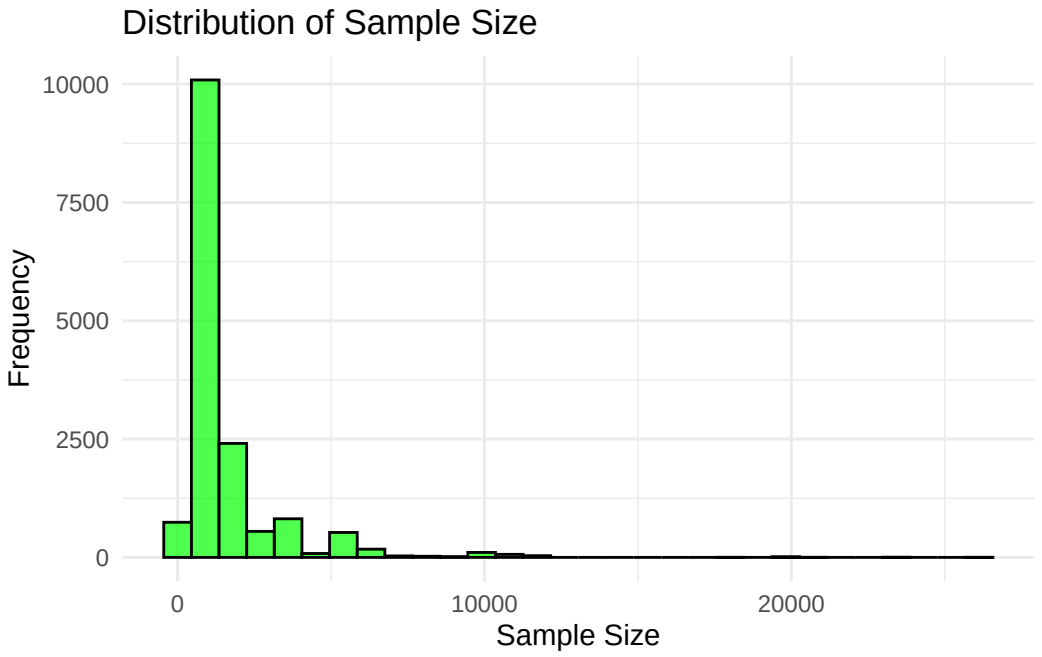

Figure 3: The distribution of sample sizes used in each poll

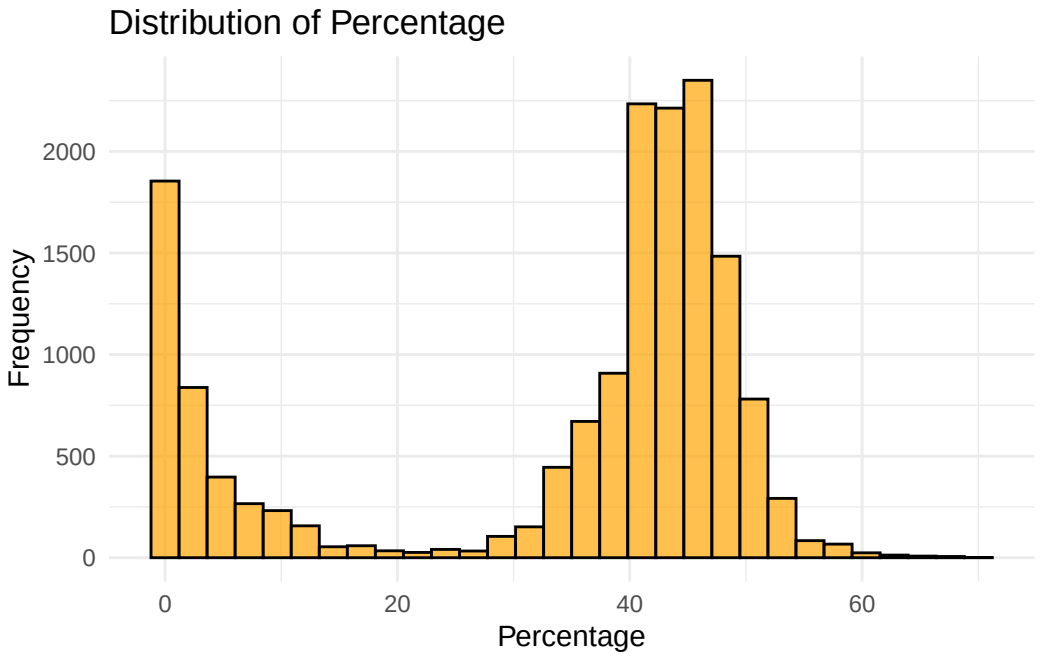

Figure 4: The distribution of what percentage of support a candidate recieves in a poll

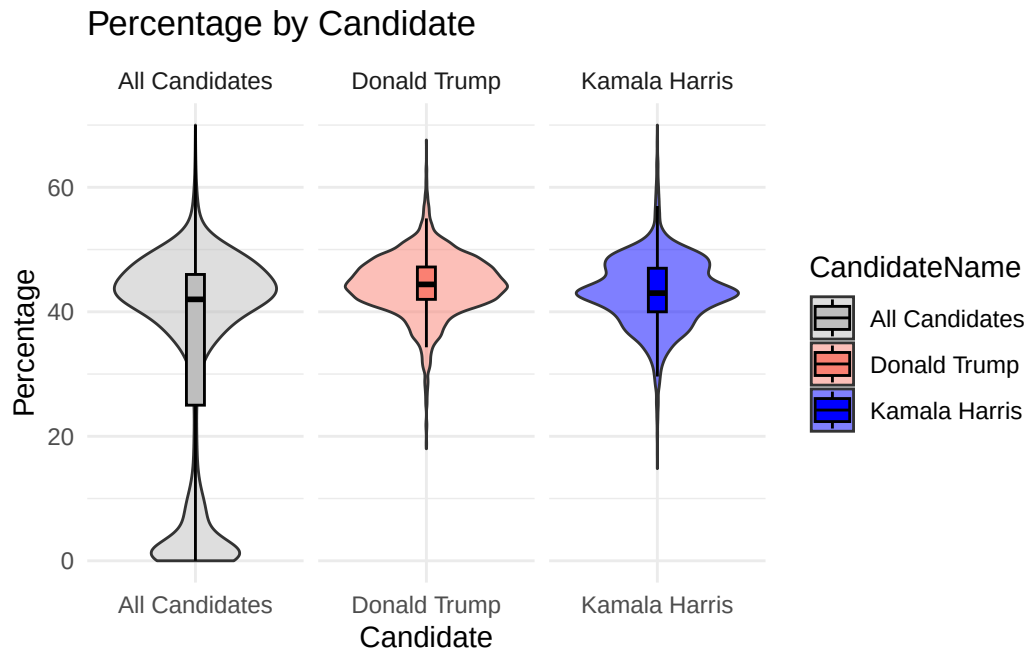

Figure 5: violin plot displaying the percentage of support Harris and Trump receive compared to all candidates

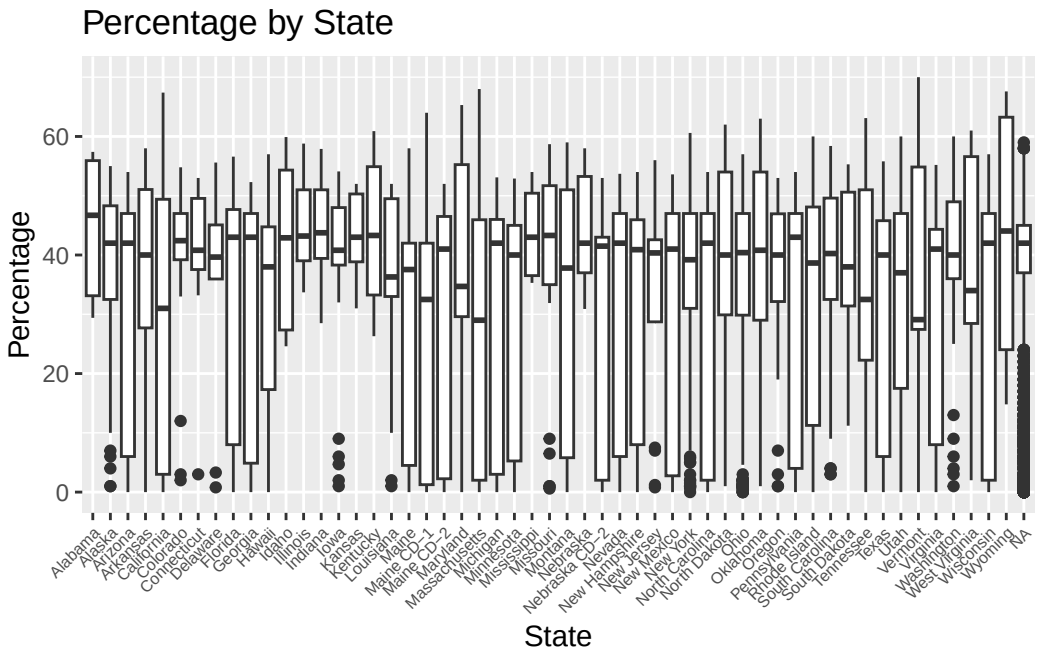

Figure 6: The percentage of support seen by state

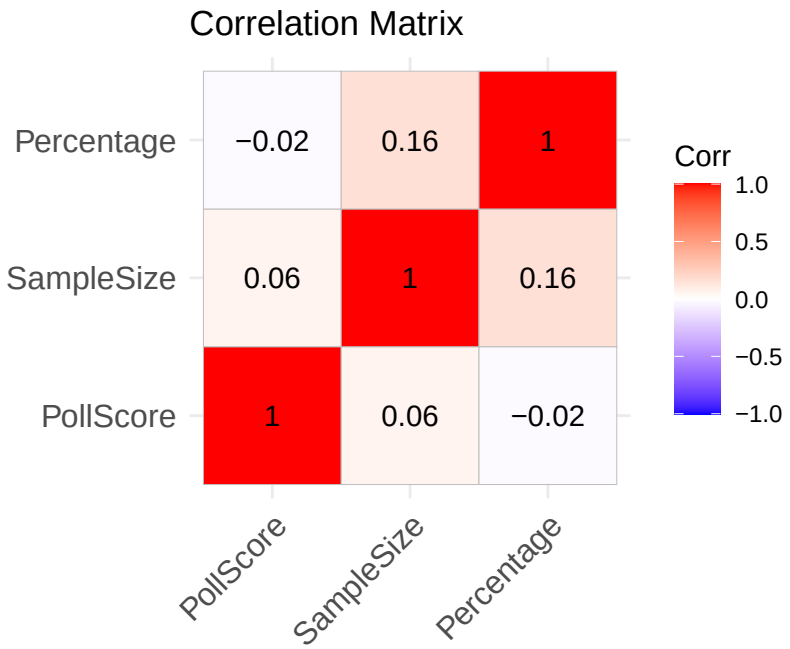

Figure 7: correlation between PollScore, the sample size of the poll, and the percentage of support for a candidate

5 Discussion

In this paper, we developed and implemented a predictive model to forecast the outcome of the 2024 U.S. Presidential election, focusing on both popular vote percentages and Electoral College results. The model incorporates polling data, demographic information, and state-specific effects.

The model teaches us that state-level dynamics play a crucial role in determining the winner of U.S. Presidential elections. Even though Harris is expected to lead in the popular vote, Trump's success in key battleground states suggests a victory in the Electoral College. An additional insight is that poll recency and pollster reliability significantly affect the accuracy of electoral predictions. By weighing recent and higher-quality polls more heavily, the model improves the robustness of its forecasts.

While the model adopts a linear regression framework for its simplicity and interpretability, this choice may limit its ability to capture more complex relationships among variables. Future iterations might benefit from exploring nonlinear models or advanced machine learning techniques, which could uncover patterns not readily observable through traditional methods.

One limitation of this model is that it depends entirely on polling data, which may not always capture late shifts in voter preferences or turnout variations. Furthermore, the model does

not account for external factors like campaign events or last-minute developments, which can significantly influence electoral outcomes.

Future research should aim to integrate alternative data sources, such as social media sentiment or economic indicators, to provide a more comprehensive view of voter preferences. More advanced machine learning techniques could also be explored to improve the accuracy of future election predictions.

6 Appendix

6.1 Pollster Methodology

The New York Times/Siena College Poll was ranked the most accurate political pollster in the US following the 2022 midterm elections (FiveThirtyEight 2024). They received the max 538 rating of 3 stars, one of only four pollsters to do so. They also had a pollscore of -1.5, meaning they had the least amount of error and bias attributed to them of the 282 pollsters ranked by 538 (FiveThirtyEight 2024).

The Times/Siena Poll targets registered voters in the United States, emphasizing those in battleground states during presidential elections. Since the US decides it's President based on the Electoral College rather than the national popular vote, this is an important focus in order to correctly determine election outcomes. The sampling frame is derived from a voter file which containing lists of 200 million registered voters as well as their demographic information and prior voting behavior data (Times 2024).

Each poll typically surveys around 1,000 respondents, which is deemed sufficient for achieving a margin of sampling error of approximately 3-4 percentage points. The Times found that tripling the sample size would only reduce the margin of error by a percent or two. This sample size allows for a balance between representation and efficiency (Times 2024).

Respondents are recruited through live phone interviews conducted by trained interviewers from call centers in various states. The polling effort includes both landline and cellphone calls, with over 90% of respondents being reached via cellphones (Times 2024).

Times/Siena uses a stratified random sampling technique, where the sample is divided into strata based on demographics such as age, race, gender, and party affiliation. This helps to ensure proportional representation from each group. Stratification helps in achieving a sample that mirrors the voting population, minimizing biases related to demographic variables (i.e. uneducated voters are less likely to respond to polls). However, it faces limitations as stratification increases the complexity of the sampling process, and if strata are not adequately represented in the sample, the results may still skew. Times/Siena combats this by applying weighting adjustments post-survey to correct for any demographic discrepancies (Times 2024).

Given the declining response rates typical of phone surveys, the Times/Siena Poll faces challenges in reaching respondents. Approximately 2% of those contacted ultimately participate. This pollster uses follow-up calls to make multiple attempts to reach individuals who initially do not respond (Times 2024). In 2022, they conducted an experiment to find out if responses were skewed by non-response bias (Cohn 2022). This was done via a mail survey that offered a reward of up to $25 for a response. This saw a 30% response rate, which is significantly larger than the 2% response rate Times/Siena gets over phone. Although there were differences between the types of respondents you get between the two methodologies, there was not a meaningful difference in who these respondants claim they would vote for (Cohn 2022).

The questionnaires used by Times/Siena are designed to capture a broad range of opinions while remaining concise. Interviews are kept under 15 minutes to reduce survey fatigue and maximize response rates (Times 2024). This pollster prioritizes making sure every viewpoint is accurately represented in order to make the questions fair and ensure that respondents are not being swayed in a particular direction. This means the questions must remain objective and clear. However, a survey with limited response options and questions may mean a respondent's view is not captured in its full breadth. It also means that more nuanced opinions may be overlooked in favor of broader trends.

The New York Times/Siena College Poll is rated the most accurate pollster in the US for a reason - it has a well-defined methodology with a fleshed out population frame, sampling approach, and contingencies for non-response handling.

6.2 Idealized Methodology

The idealized methodology for forecasting the US presidential election targets eligible registered voters from all 50 states and Washington D.C., ensuring representation from each state in order to create an Electoral College forecast, with particular emphasis on swing (battleground) states.

The sampling method will utilize stratified random sampling, which stratifies the sample by state. Voters will be further stratified based on factors such as age, gender, race/ethnicity, income, and political affiliation. In swing states, we will ensure oversampling to capture more granular data and reduce our margin of error.

To maintain a national margin of error of $\pm 4\%$, based on the experiment conducted by Times/Siena, the total sample size will be 1,000 respondents per state (Times 2024). Swing states will be oversampled, with over 2,000 respondents to achieve a margin of error of $\pm 2\%$. Given an estimated response rate of 2%, we aim to recruit approximately 50,000 potential respondents per state, with a target of 100,000 in battleground states.

To recruit respondents, we will utilize online panel providers like YouGov, Ipsos, and Dynata to obtain geographically and demographically targeted participants. Additionally, we will conduct state-specific phone surveys, employing interactive voice response (IVR) systems and live calls to reach older and less tech-savvy voters, particularly in states with a higher proportion of rural populations. To incentivize participation, all respondents will be entered into state-level sweepstakes, with one winner selected per state.

Data collection will be conducted on platforms such as Google Forms. Telephone surveys will follow the layout and workflow of the Google Form survey. The survey is designed to take approximately five to ten minutes to complete in order to minimize drop-off rates. It will include national-level questions, covering voting intent, candidate preference, key issues, and voter perceptions. After collection, the data will be weighted by state and demographic factors to ensure accuracy, utilizing the latest Census data and state voting records.

For data validation, we will implement several measures, including completion time checks to identify speeding responses, IP tracking to prevent duplicate entries, and reCAPTCHA to guard against bots. Respondents will provide their state of residence and ZIP code for additional verification. Furthermore, the data will be weighted according to state demographics and previous election turnout data, utilizing information from Census and state voting records. Special attention will be paid to swing states to ensure no key voter segments are underrepresented.

In terms of budget allocation, we will designate funds as follows:

1. Survey Development: $7,000
    - The $7,000 covers the costs associated with designing the survey, including the creation of questions, structuring the survey for clarity and flow, and ensuring that it meets methodological standards. It may also include hiring experts in survey design or consulting with statisticians to validate the survey framework and ensure questions are unbiased. Additionally, while Google Forms is free, this budget may account for platform costs if needed.
2. Accurate Sampling: $10,000
    - The sampling budget is allocated for the costs associated with selecting a representative sample of voters. This includes expenses related to stratified random sampling, where the sample is segmented by various demographics and state representation. In order to ensure each demographic share is accurately represented, we may need to purchase access to demographic data or employ statisticians to ensure that the sampling aligns with the population of each state.
3. Recruiting Respondents: $45,000
    - The majority of the budget is dedicated to recruiting respondents, which is necessary for achieving a large and demographically representative sample. This cost will cover payments to online panel providers which charge fees for access to their respondent pools. Additionally, this budget will cover costs related to conducting state-specific phone surveys and IVR calls, such as hiring call center staff or IVR technology.
4. Respondent Incentives: $28,000
    - To encourage participation, we plan to run state-level sweepstakes, where respondents have a chance to win prizes for completing the survey. A $500 prize will be awarded to one winner in each state (+ Washington DC). The remaining budget will cover administrative expenses associated with running the sweepstakes.
5. Data Analysis and Reporting: $10,000
    - This portion of the budget is reserved for the costs related to analyzing the collected data and reporting the findings. This may include hiring data analysts or consultants who specialize in statistical analysis and survey data interpretation.

The funds will also cover software tools necessary for data analysis, as well as the creation of visualizations and reports to communicate the results effectively.

An example of the survey can be found at this link: https://forms.gle/igzK683kHLYytGvu8

References

Cohn, Nate. 2022. "Are the Polls Still Missing 'Hidden' Republicans? Here's What We're Doing to Find Out." *The New York Times*, November. https://www.nytimes.com/2022/11/08/upshot/poll-experiment-wisconsin-trump.html.

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

FiveThirtyEight. 2024. "538's Pollster Ratings." ABC News; ABC News Internet Ventures. https://www.fivethirtyeight.com/methodology/pollster-ratings/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Presidential Election Polls-2024. 2024. New York, USA: FiveThirtyEight. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Sigmon, Eric. 2024. "The New Democratic Presidential Ticket: What's Changed and What Remains the Same." Elcano Royal Institute. https://www.realinstitutoelcano.org/en/analyses/the-new-democratic-presidential-ticket-whats-changed-and-what-remains-the-same/.

Times, The New York. 2024. "You Ask, We Answer: How the Times/Siena Poll Is Conducted." *The New York Times*, October. https://www.nytimes.com/article/times-siena-poll-methodology.html.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2023. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.