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 3, 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

1 Introduction **2**

2 Data **3**
2.1 Overview . . . 3
2.2 Measurement . . . 3
2.3 Outcome Variables . . . 3
2.4 Predictor Variables . . . 4

3 Model **4**
3.1 Model Set-Up . . . 4
3.2 Model Justification . . . 5

4 Results **7**

5 Discussion **11**

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

6 Appendix **12**
6.1 Additional Data Details 12
6.2 Model Details 12
6.3 Pollster Methodology 12
6.4 Idealized Methodology 12

References **15**

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 (**RealInstitutoElcano?**). 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.

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 several 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, filtered to Harris and Trump, based on the aforementioned predictor variables. This is done both nationally at a popular vote level, but also by state to determine the distribution of electoral votes.

3.1 Model Set-Up

This paper utilizes linear regression model:

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

where, Percentage$_i$ represents the predicted support percentage for a candidate in a given poll.

β_i are the coefficients of the model that quantify the impact of each predictor.

ϵ_i is the error term.

The results from these models are aggregated to compute the predicted percentage of support for each candidate in each state. Based on these percentages, we forecast the likely winner of the state's electoral votes.

3.2 Model Justification

This model was 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. The model was validated using out-of-sample testing and Root Mean Square Error to ensure predictive accuracy.

Table 1: 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 2: Predicted Number of Electoral Votes for Donald Trump and Kamala Harris

Trumps Total Electoral Votes	Harriss Total Electoral Votes
302	233

Table 3: 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 (see Appendix for details) 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.

[WILL CHANGE THE FORMAT OF THESE GRAPHS AND CAPTION/DESCRIPTION WILL BE ADDED]

Table 4: 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

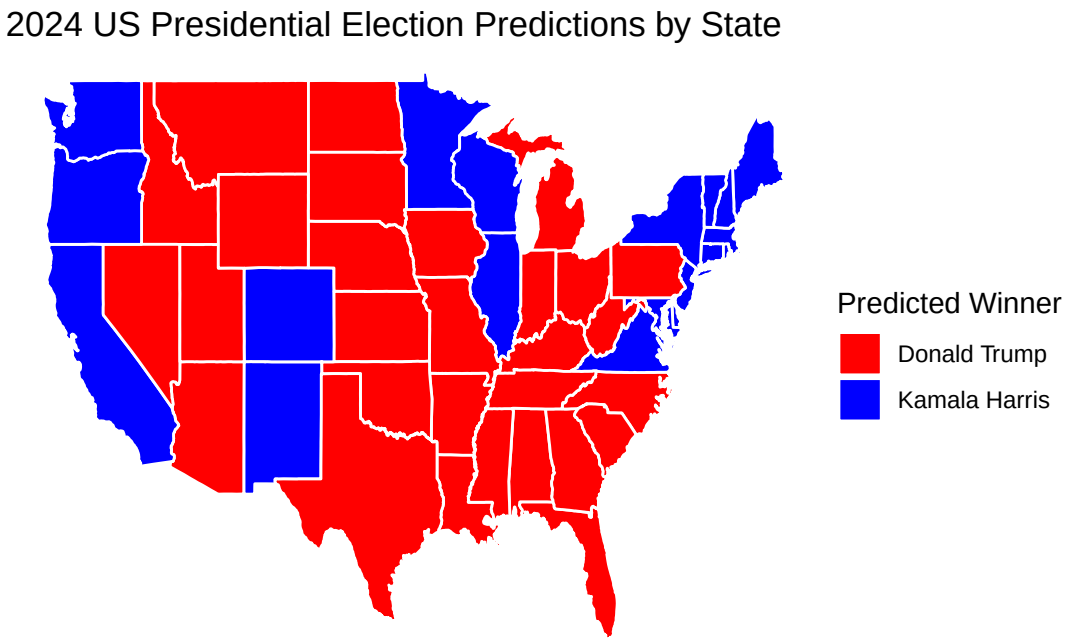

Figure 1: Mapped Model-predicted outcomes for 2024 elections by State

will update these box plots to different visuals due to box plots not showing the full dataset! (#Note to team: updated to violin plots to show distribution and selected only trump and harris and another one for all the candidates to show distribution of dataset)

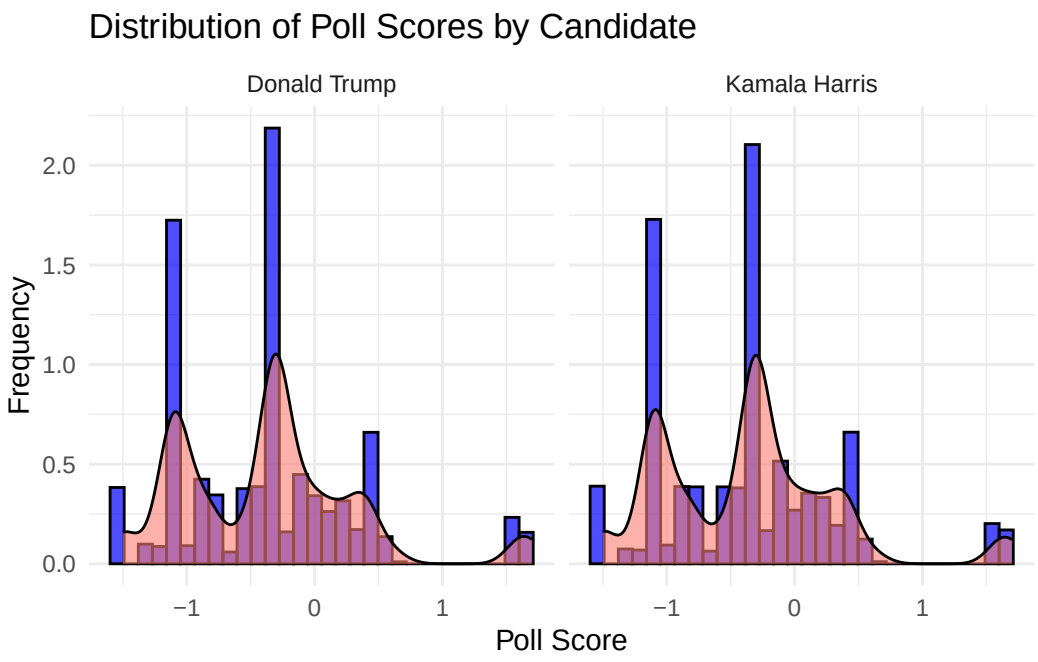

Figure 2: Distribution of Poll Scores by Candidate with denisty overlayed

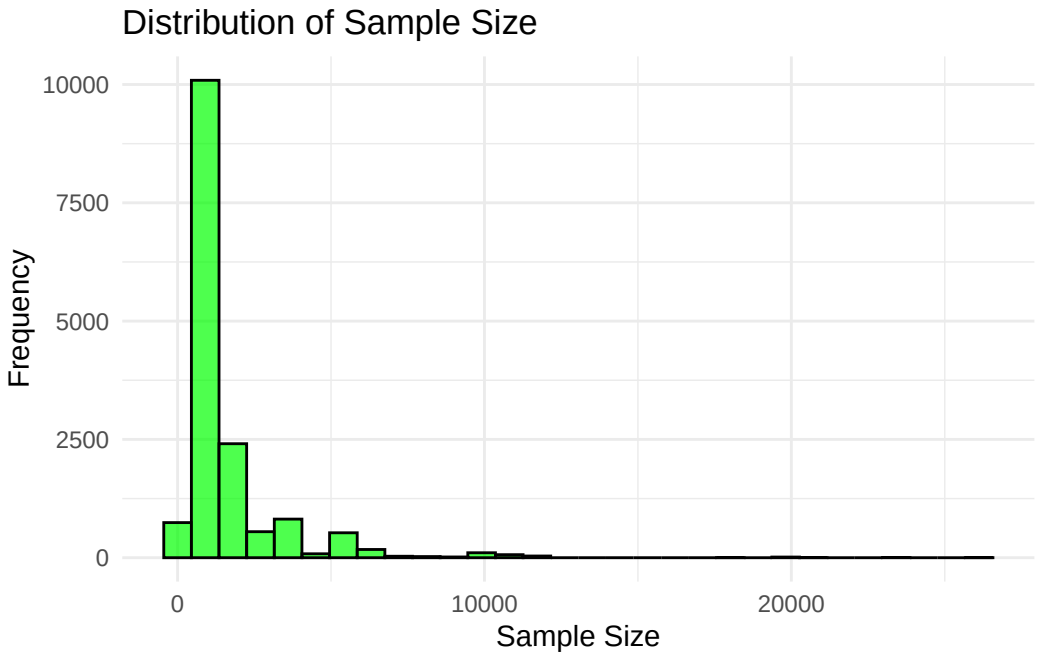

Figure 3: Need Edit

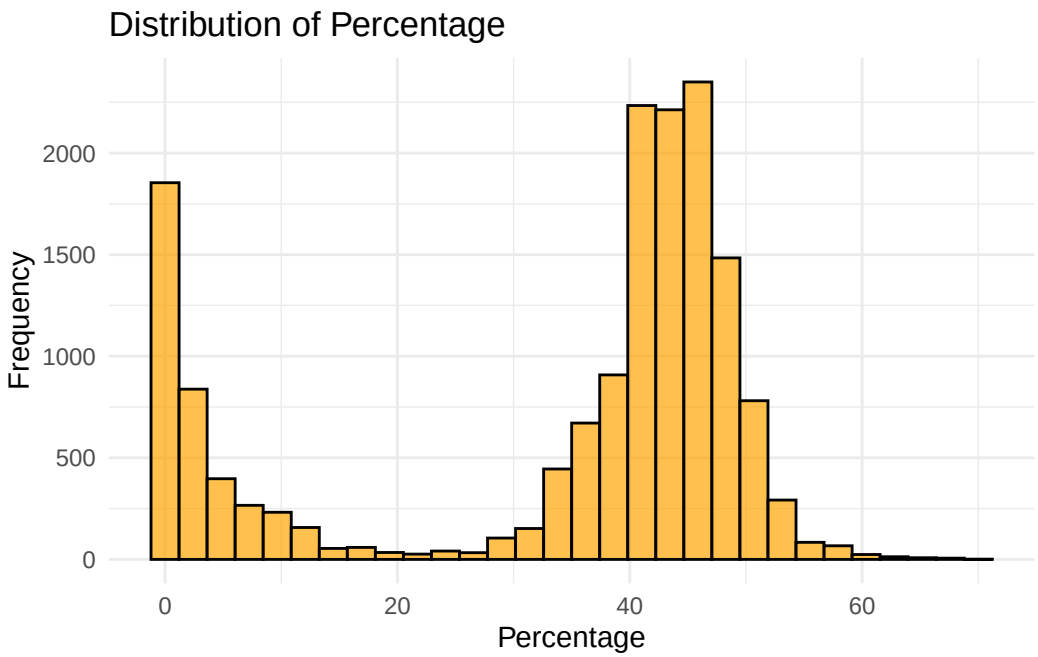

Figure 4: Need Edit

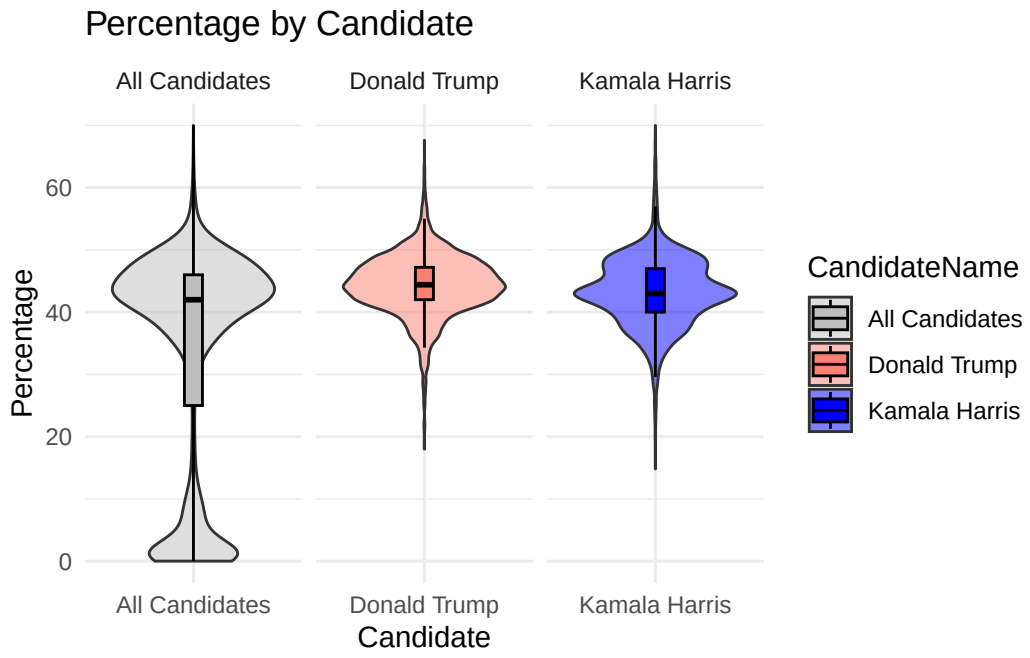

Figure 5: Need Edit

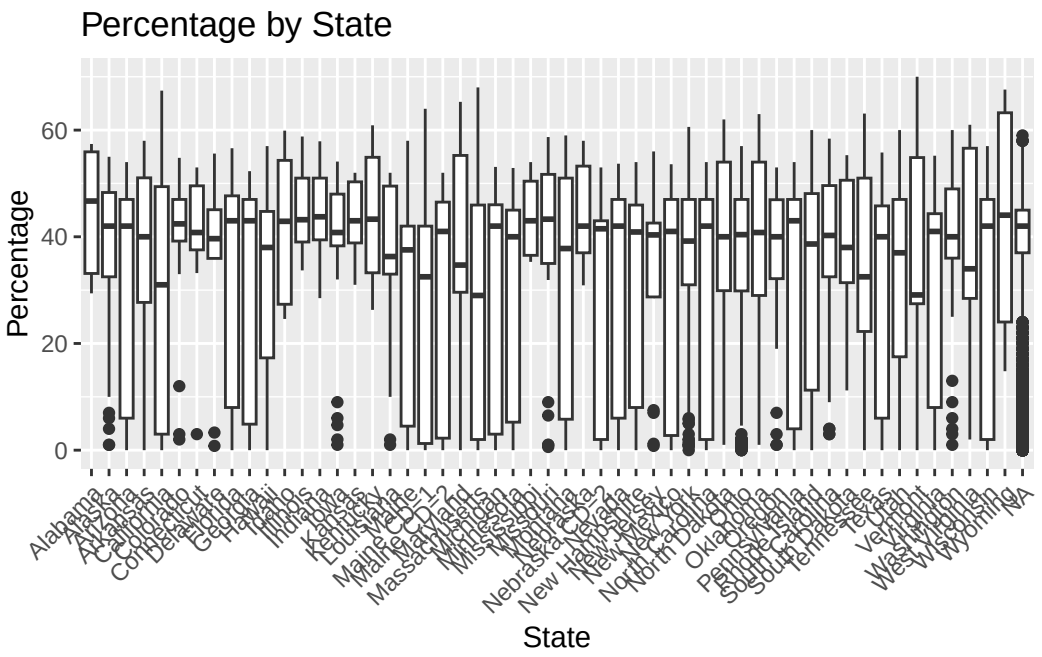

Figure 6: Need Edit

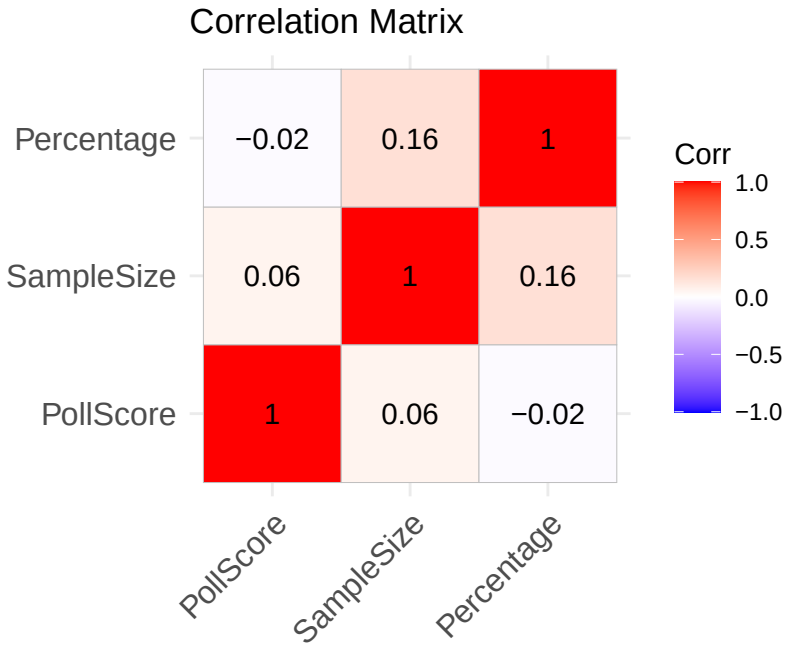

Figure 7: Need Edit

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 Additional Data Details

6.2 Model Details

6.3 Pollster Methodology

6.4 Idealized Methodology

This is an outline. I will turn this into paragraph style this week.

Target Population: The target population is eligible US voters from all 50 states and Washington D.C. –> representation from every state for an Electoral College forecast.

Sampling Method

- Stratified Random Sampling: The sample will be stratified by state, ensuring each state is proportionally represented in terms of population size and key voter demographics (age, gender, race/ethnicity, income, political affiliation).

- Key Stratification Variables: Within each state, voters will be further stratified based on party affiliation, urban/rural distribution, and voting history.

- For swing states, we will ensure oversampling to capture more granular data.

Sample Size

- The total sample size will be 15,000 respondents –> ensures state-level representation while maintaining a national margin of error of $\pm 1\%$.

- As mentioned above, swing states will have an additional focus, with 1,000+ respondents per state (depending on population size).

- Electoral vote-rich states (e.g., California, Texas, Florida) will be proportionally represented to reflect their impact on the election outcome.

- With an estimated response rate of 10%, we aim to recruit 150,000 potential respondents across states.

Recruiting Respondents

- Panel Providers: Utilize online panel providers (such as YouGov, Ipsos, or Dynata) for geographically and demographically targeted respondents by state.

- Social Media Ads: Conduct state-targeted ads on platforms like Facebook and Instagram*for broad geographic reach

- State-Specific Phone Surveys: To reach older and less tech-savvy voters, particularly in states with higher proportions of rural populations, we will use IVR phone surveys and live calls.

- Respondents will be entered into state-level sweepstakes, with one winner per state

Data Collection

- Surveys will be conducted on platforms like Qualtrics or SurveyMonkey, with branching logic for state-specific questions. Telephone surveys for older demographics will be integrated. The mock survey will be conducted using Google Forms for the purposes of accessibility.

- The survey will take approximately 5 minutes in order to minimize drop-off rates

- The survey will include state-specific questions and national-level questions, such as voting intent, candidate preference, key issues, and voter perception.

- Weighting by State: Data will be weighted post-collection to ensure state-level accuracy, considering the latest Census data and state voting records.

Data Validation

- Validate completion time (speeding checks), IP tracking (to prevent duplicate responses), and reCAPTCHA to prevent bots.

- Respondents will be asked their state of residence, and ZIP code for additional verification.

- Data will be weighted by state demographics and previous election turnout using Census and state voting data. Swing states will receive additional attention to ensure no underrepresentation of key voter segments.

include budget allocation?

Survey Structure (will be migrated to google form)

Introduction

- Brief description of the survey's purpose (US Presidential Election Forecast with a focus on state-level dynamics).

- Assurance of confidentiality and anonymity, with a clear outline of data use.

- Contact details for survey-related inquiries.

Questions

1. Demographics:

- State of Residence (Dropdown list of all 50 states and D.C.)

- Age, gender, race/ethnicity, income level, education, party affiliation (Republican, Democrat, Independent, Other).

2. Voter Registration:

- Are you registered to vote in the upcoming election? (Yes/No)

- How likely are you to vote in the upcoming election? (Likert scale)

3. Voting Intent:

- If the election were held today, who would you vote for? (List of candidates specific to the respondent's state, based on ballot access).

- How certain are you of your choice? (Likert scale)

4. Key Issues:

- What are the most important issues influencing your vote? (Rank the top 3 issues: Economy, Healthcare, Immigration, Climate Change, National Security, Social Issues, etc.)

5. State-Specific Issues:

- What state-level issues are important to you? (Open-ended question, with suggestions like state economy, healthcare access, local infrastructure, etc.)

6. Perceptions of Candidates:

- Rate the following candidates on trustworthiness, competence, leadership, etc. (Likert scale, customized for state-level candidates where applicable).

7. Previous Voting Behavior:

- Did you vote in the last presidential election? (Yes/No)

- If yes, who did you vote for? (Open response, with drop-down of major candidates for validation)

Closing

- Thank You: Thank the respondents for their time and participation, reminding them of the sweepstakes.

References

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

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.

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.