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

Sophia Brothers Deyi Kong Rayan Awad Alim

October 22, 2024

This study models the 2024 U.S. Presidential election outcomes using advanced statistical techniques and predictive modeling. The analysis incorporates data from various polling sources and considers key demographic and political factors to forecast both the popular vote and Electoral College results. The findings suggest a close race between the two main candidates, Donald Trump and Kamala Harris, with a predicted victory for Trump in the Electoral College. This analysis contributes to understanding how public opinion data and electoral systems interact, providing insights into electoral dynamics in contemporary U.S. politics.

1 Introduction

Elections in the United States, particularly presidential elections, are pivotal moments that shape not only the nation's political landscape but also its international relations. The 2024 U.S. Presidential election, like its predecessors, promises to be contentious and closely watched. As electoral outcomes have become increasingly difficult to predict due to polarized political climates and fluctuating public opinion, sophisticated statistical models are essential tools for forecasting results and understanding the dynamics at play.

The core estimand of this study is the projected electoral outcome in terms of state-level victories and Electoral College votes for each candidate. The model leverages polling averages, historical voting patterns, and demographic predictors to estimate the probability of each candidate winning in individual states. The objective is to translate these probabilities into a forecast of Electoral College votes, thus determining the likely overall winner of the election. Given the significant influence of swing states, the model focuses on capturing their dynamics more accurately than national polling averages typically allow.

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

The results suggest 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.

Understanding these projected outcomes is essential for several reasons. First, the model highlights the limitations of relying solely on national polling to predict election results, emphasizing the continued relevance of the Electoral College system in determining the presidency. Additionally, it sheds light on the evolving political landscape in swing states, where demographic changes and shifting voter allegiances can have outsized impacts. 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: In Section 2, we provide outlines the data sources and variables considered, followed by the model setup in Section 3 and justification. The results Section 4 presents the key findings of the analysis, with a discussion on the implications of these results for political forecasting and public opinion analysis. Finally, the paper concludes with an overview of potential limitations and suggestions for future research in the discussion Section 5.

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. The analysis leverages R and several libraries, including tidyverse (Wickham et al. 2019), janitor (Firke 2023), knitr (Xie 2024), dplyr (Wickham et al. 2023), purrr (Wickham and Henry 2023), and here (Müller 2020), for data manipulation, and ggplot2 (Wickham 2016) 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 measurements in the dataset include polling scores, sample sizes, and predicted percentages for each candidate. The data is filtered to include only polls relevant to the two primary candidates, 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 two linear regression models: one for Donald Trump and one for Kamala Harris. 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 utilized a linear 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.

```
# A tibble: 2 x 3
  CandidateName AveragePredictedPercentage NormalizedPercentage
  <fct>                              <dbl>                <dbl>
1 Donald Trump                        45.3                 51.0
2 Kamala Harris                       43.5                 49.0
```

```
# A tibble: 1 x 2
  TrumpTotalElectoralVotes HarrisTotalElectoralVotes
                     <dbl>                     <dbl>
1                      302                       233
```

```
# A tibble: 54 x 4
   State          TrumpPredicted HarrisPredicted Winner
   <chr>                   <dbl>           <dbl> <chr>
 1 Alabama                  56.3            30.9 Donald Trump
 2 Alaska                   48.5            39.8 Donald Trump
 3 Arizona                  46.3            42.8 Donald Trump
 4 Arkansas                 53.3            32.9 Donald Trump
 5 California               31.9            53.7 Kamala Harris
 6 Colorado                 39.4            47.1 Kamala Harris
 7 Connecticut              37.1            50.6 Kamala Harris
 8 Delaware                 36.6            50.1 Kamala Harris
 9 Florida                  48.6            42.4 Donald Trump
10 Georgia                  46.9            43.5 Donald Trump
11 Hawaii                   26.6            41.5 Kamala Harris
12 Idaho                    54.2            25.2 Donald Trump
13 Illinois                 38.8            48.7 Kamala Harris
14 Indiana                  51.5            35.6 Donald Trump
15 Iowa                     48.4            38.5 Donald Trump
```

```
16 Kansas                   48.9            37.1 Donald Trump
17 Kentucky                 54.5            30.7 Donald Trump
18 Louisiana                51.6            35.7 Donald Trump
19 Maine                    40.8            45.3 Kamala Harris
20 Maine CD-1               33.8            57.0 Kamala Harris
21 Maine CD-2               46.1            43.2 Donald Trump
22 Maryland                 32.0            57.2 Kamala Harris
23 Massachusetts            29.5            54.6 Kamala Harris
24 Michigan                 44.9            43.9 Donald Trump
25 Minnesota                41.6            44.7 Kamala Harris
26 Mississippi              51.8            36.3 Donald Trump
27 Missouri                 51.9            37.5 Donald Trump
28 Montana                  53.8            35.8 Donald Trump
29 Nebraska                 52.6            36.3 Donald Trump
30 Nebraska CD-2            41.7            49.5 Kamala Harris
31 Nevada                   45.8            43.2 Donald Trump
32 New Hampshire            41.9            46.2 Kamala Harris
33 New Jersey               39.0            43.8 Kamala Harris
34 New Mexico               41.0            47.9 Kamala Harris
35 New York                 36.3            48.1 Kamala Harris
36 North Carolina           46.6            43.6 Donald Trump
37 North Dakota             55.2            29.4 Donald Trump
38 Ohio                     48.8            39.0 Donald Trump
39 Oklahoma                 56.0            31.8 Donald Trump
40 Oregon                   39.9            48.1 Kamala Harris
41 Pennsylvania             45.7            44.8 Donald Trump
42 Rhode Island             37.3            50.9 Kamala Harris
43 South Carolina           49.6            37.2 Donald Trump
44 South Dakota             51.9            30.2 Donald Trump
45 Tennessee                52.7            28.0 Donald Trump
46 Texas                    48.1            40.4 Donald Trump
47 Utah                     50.8            31.2 Donald Trump
48 Vermont                  28.1            60.9 Kamala Harris
49 Virginia                 41.5            44.7 Kamala Harris
50 Washington               35.9            49.9 Kamala Harris
51 West Virginia            58.4            28.3 Donald Trump
52 Wisconsin                45.3            45.5 Kamala Harris
53 Wyoming                  60.1            17.5 Donald Trump
54 <NA>                     NA              NA   Tie
```

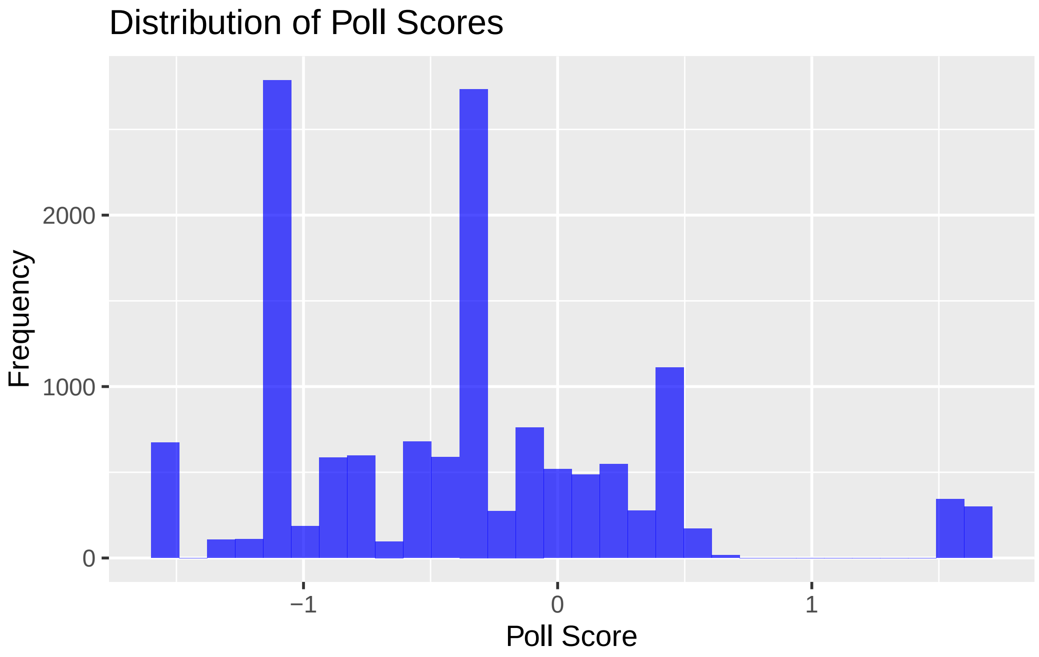
4 Results

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]

```
# A tibble: 1 x 4
  AveragePollScore AverageSampleSize AveragePercentage TotalPolls
             <dbl>             <dbl>             <dbl>      <int>
1           -0.379             1606.              33.7      15829
```

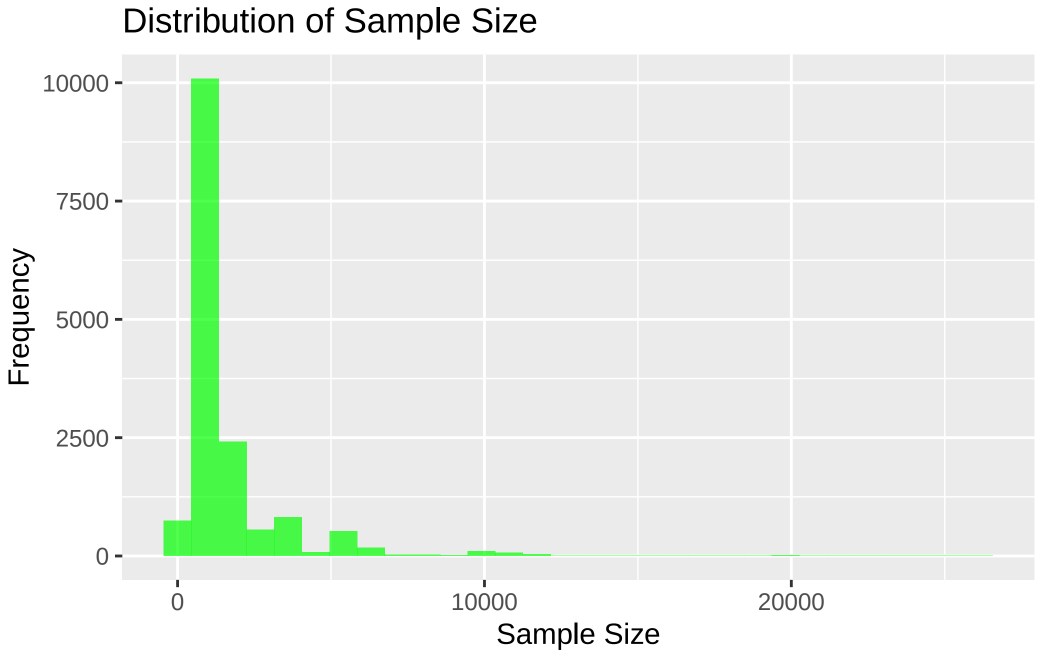
Figure 1: Need Edit



(a)

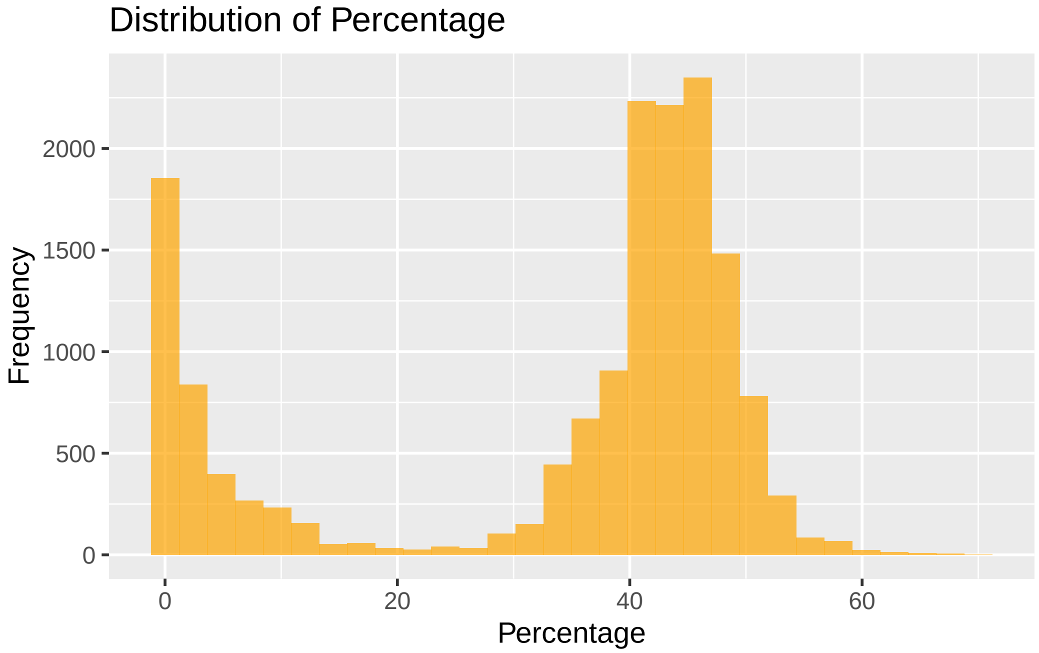
Figure 2: Need Edit

will update these box plots to different visuals due to box plots not showing the full dataset!



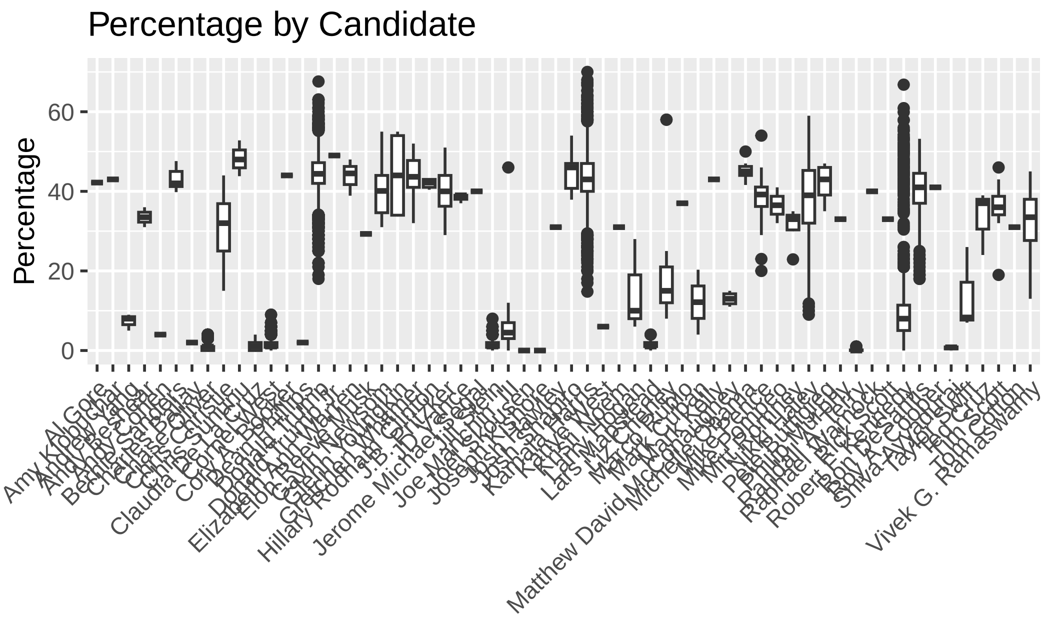
(a)

Figure 3: Need Edit



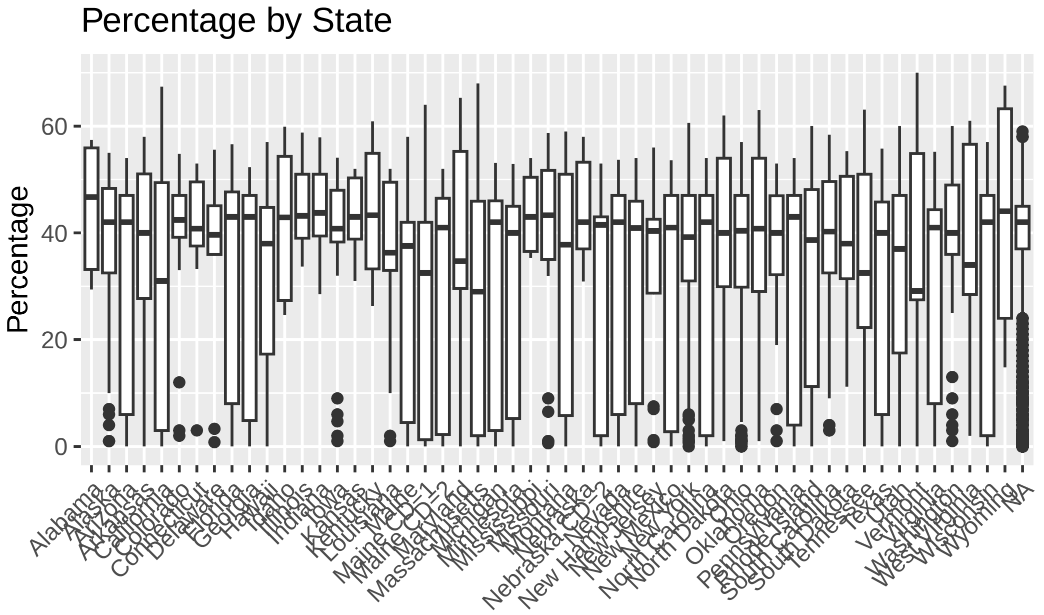
(a)

Figure 4: Need Edit



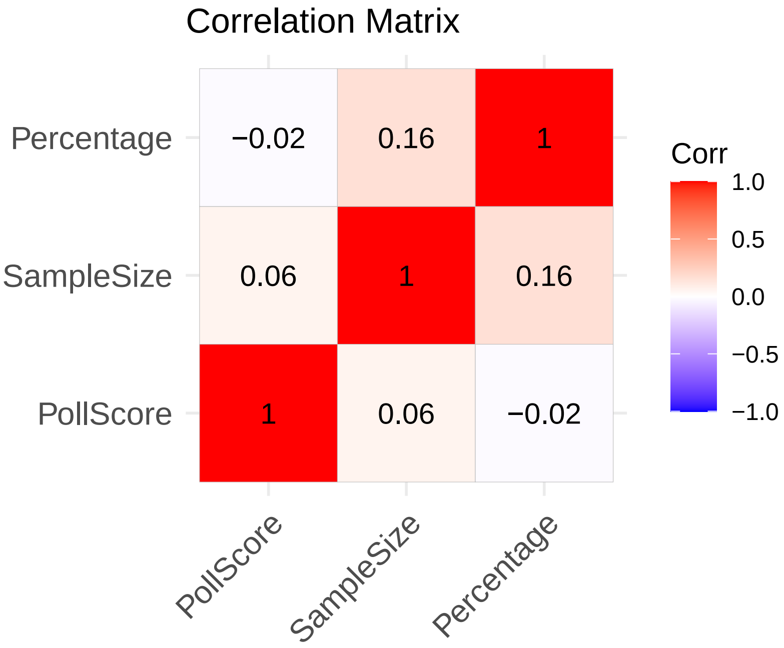
(a)

Figure 5: Need Edit



(a)

Figure 6: Need Edit



(a)

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 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 ±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.

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/.