Forecasting the United States 2024 Presidential Election*

Kamala Harris wins 60th presidency with 265 Electoral College Votes

Robert Ford Michelle Ji Cher Ning

November 3, 2024

This report analyzes compiled polling surveys for the 2024 U.S. presidential election using FiveThirtyEight data (FiveThirtyEight 2024b) and the 'pool of polls' method. From the analysis, bayesian and generalized linear models were built to forecast the outcome of the presidential election based on popular vote and the electoral college system. The models forecast Kamala Harris as the winner of election, winning 265 electoral college votes versus the 172 votes Donald Trump acquires. Election forecasting is crucial for informing voters, guiding campaign strategies, and engaging the public in the electoral process. Additionally, it helps predict potential policy shifts and holds political institutions accountable, contributing to a more transparent and informed democracy.

1 Introduction

Every four years, the United States presidential election occurs and is one of the most significant political events internationally. The complex process of the election involves both the popular and electoral college system. The candidate that receives the popular vote by the voters wins the electoral votes for that state (Weber 2024). After the popular vote is accounted for, the candidate who receives the majority of the electoral votes from all states wins the presidential election. The number of electoral votes varies by state and is proportional to each state's respective population (Weber 2024).

Specifically for the 2024 presidential election, the top candidates include Donald Trump and Kamala Harris, both with differing opinions on key topics such as immigration, technology, abortion, and transatlantic affairs. For example, regarding technology, both parties are concerned with the rapid progression of artificial intelligence's capabilities but differ in the response and regulation of it. Republicans believe that the current practices are unfair and disproportionately target their voices, whereas democrats generally are in favor of tighter regulation (NBC 2024). Another key issue important to voters and the candidates is abortion. As former president, he supported restrictions on abortion access and advocated for policies that align with the conservative movement's opposition to abortion rights (McKenzie Beard and Abbie Cheeseman and Justine McDaniel 2024). On the other hand, Harris has been a key supporter in the pro-choice movement and has been an advocate for protecting and expanding access to abortion services, as well as female autonomy (Chad de Guzman and Koh Ewe 2024). The outcome of this election will have profound consequences for various legislative policies and the future direction of the United States.

*Code and data are available at: https://github.com/Ford-Robert/us-presidential-election.

To forecast the 2024 United States presidential election, a generalized linear model and Bayesian model were built. These models will help us predict who will win the election based on the electoral college system. A key finding from our model is that we predict Kamala Harris to win the presidential election with 265 electoral college votes. This would be a historic win, as she would be the first woman to become US president, a significant milestone in American politics. One drawback of our model stems from representativeness of the dataset, as it does not include all 50 states in the polls, which may impact the electoral college votes.

This paper is broken down into various sections, including Data, Modeling, Results, Discussion, Conclusion, and Appendices. This paper uses data made available by FiveThirtyEight (FiveThirtyEight 2024b) which compiles individual polling surveys based on state and methodology. Section 2 explores the data, highlighting key aspects that may be useful to future policymakers or campaign strategists, as well as details the variables present in the dataset. Section 3 presents the models that were built and used to forecast the election. Section 4 details the conclusions of our model and Section 5 explores possible implications and insights of our conclusions. Section A include an idealized methodology and survey that we could hypothetically run, with the task of forecasting the US presidential election. This section also includes an in-depth analysis of one specific pollster's methodology found within the larger dataset.

2 Data

The dataset used was obtained through FiveThirtyEight Interactives (FiveThirtyEight 2024b) and is focused on the United States presidential general polls for 2024. The polling data extracted utilize the "pool of polls" method, which combines multiple polling results into a single estimate. By aggregating data from several polls, this method aims to create a more stable and reliable measure of public opinion by reducing the impact of outliers, sampling error, and individual poll biases. This dataset is crucial to help us build a model to forecast the 2024 US Presidential election and can help us analyze key features in voter habits for future political analysis. Note that polling data collection stopped on October 19. Our model's results are based solely on data available as of that date and do not incorporate any subsequent polling updates.

There were two other potential datasets that could have been used, titled "Presidential Polling Averages" and "President Primary Polls", both found through FiveThirtyEight (FiveThirtyEight 2024a). The "Presidential Polling Averages" dataset only included data from the Biden versus Trump election in 2020 and was not updated to include data for the upcoming 2024 election and therefore inadequate. The "Presidential Primary Polls" dataset included very similar information as our chosen dataset, however, it included candidates that have since dropped out or were not major running candidates, such as Nikki Haley and Michelle Obama. It includes observations that are unnecessary and creates noise, therefore we did not choose this dataset either.

The variables used in the dataset include pollster name, pollster weight, method, state, sample size, pollscore, collection/end date, transparency score, candidate, support percentage, election date, days to election, initial weight, weight, and poll region. Pollster Name indicates which pollster the individual poll came from and the weight is assigned by FiveThirtyEight depending on sample size and if they have multiple polls out in a short time to adjust for any bias. Method is the way the poll was deployed, including but not limited to live phone, online panel, and text-to-web. State indicates what state the poll was conducted and sample size is the size of the poll. The pollscore is the rating FiveThirtyEight gives each indidivual poll, on factors such on ethical and methodological elements.The candidate variable indicates the candidate a certain percentage of people in the poll voted for and "support percentage" represents that specific number. The election date and days to election variables indicate how far out the poll was conducted with respect with the election date of November 5, 2024. We also reorganized the dataset to include a poll region, which

divides the US into geographical regions such as rust belt, northeast, etc. The table in Section B.3 details the first few observations of the cleaned dataset, which we used to build our model.

For further analysis, the cleaned dataset was separated into two different datasets by beach name, for simplicity purposes. Data was analyzed through the R programming software (R Core Team 2023) and packages such as tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), knitr (Xie 2014), and dplyr (Wickham et al. 2023) were used to help download, clean, simulate, analyze, and test the data.

Some pollsters are included in the dataset more than others based on the frequency of their conducted poll surveys and if they align with FiveThirtyEight's guidelines, which can be seen in **?@fig-1**. **?@fig-1** highlights the distribution of the top 25 pollsters and how frequently they appear in the dataset, with Morning Consult being the most popular pollster with almost 500 appearances. Noting which pollsters are more trustworthy and provide reliable data is important for evaluating the quality of future analyses.

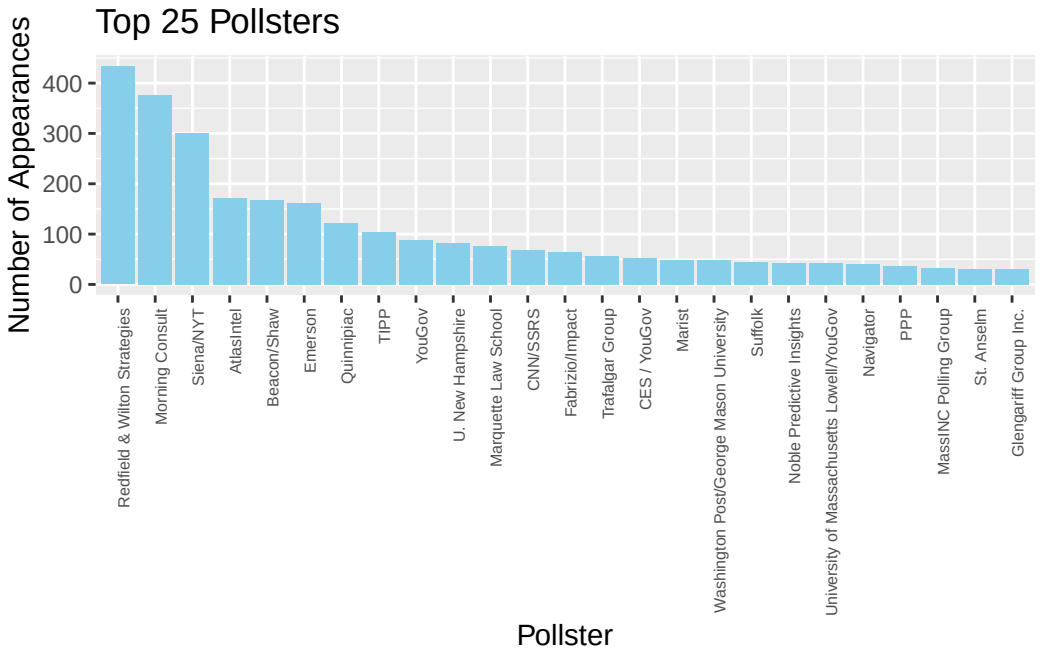

Figure 1: Top 25 pollsters in overall polling data

Figure 2 details the frequency of various polling methods used in the individual polling data, showing online panels are the most popular methodology used to deploy these polls, followed by live phone. The polling method plays a crucial role in reducing certain response biases. For instance, if the poll is conducted via an online panel with anonymous responses, participants may feel more comfortable expressing their true voting intentions, potentially reducing social desirability bias.

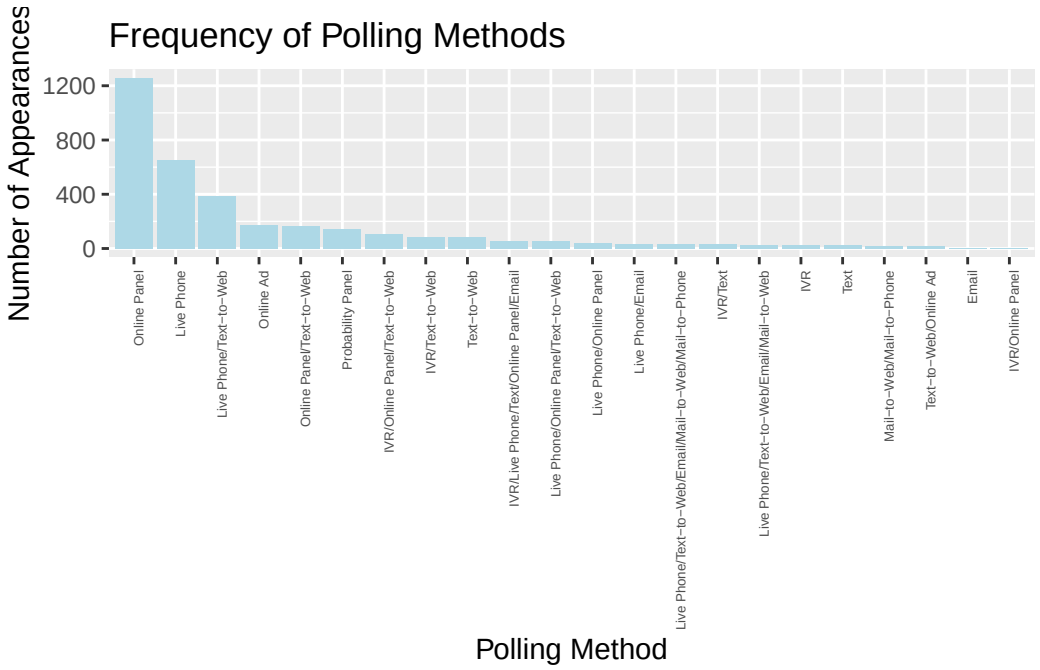

Figure 2: Frequency of Various Polling Methods used by FiveThirtyEight in their "pool of polls"

Finally, Figure 3 shows the number of votes per candidate. As expected, Harris and Trump have the highest number of votes, with Harris having slightly more votes based on the surveyed polls than Trump. Other candidates with a noticeable number of votes in the graph include Robert F. Kennedy and Jill Stein. However, their totals are significantly lower than those of Harris and Trump, making them likely insignificant in the election. It is also important to mention that Biden still appears in these polls, and therefore in Figure 3, due to the lag time in poll data processing by FiveThirtyEight. However, since he has dropped out of the race, he will not be included in our model.

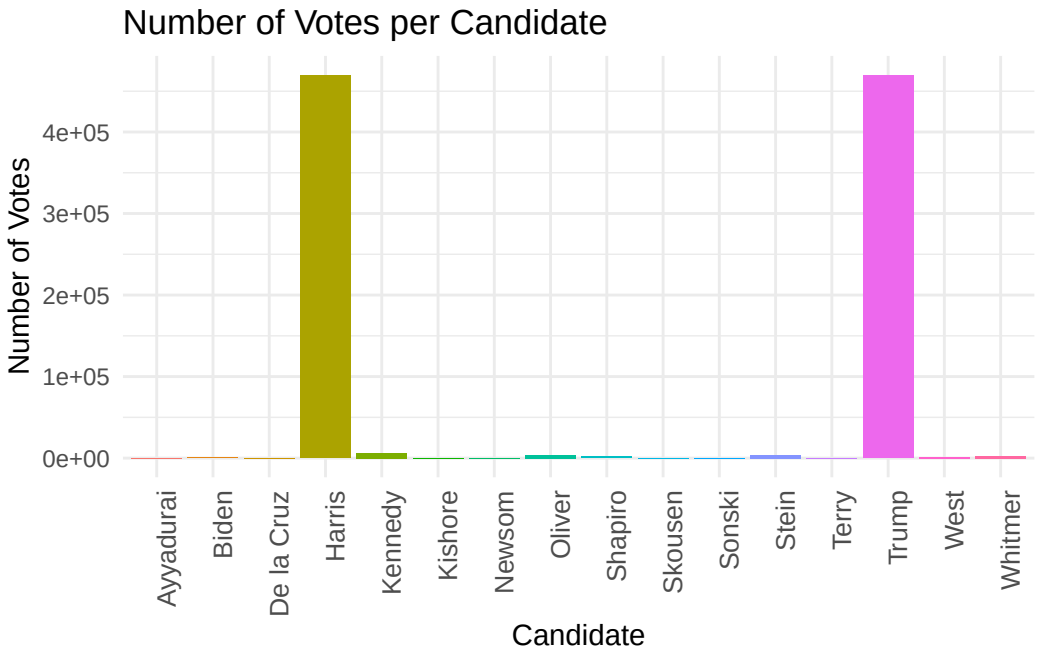

Figure 3: Number of Votes per Candidate. This reflects the popular vote for each candidate and does not account for the electoral college sytem.

3 Model

To predict state level support and calculate the expected Electoral Votes (EV) for each candidate in the upcoming US election, we employed a Bayesian linear regression model. The model is mathematically specified as follows:

$$\text{Support}_i = \beta_0 + \beta_1 \times \text{Sample Size}_i + \beta_2 \times \text{Days to election}_i + \beta_3 \times \text{Transparency Score}_i + \beta_4 \times \text{Pollscore}_i + \gamma \times \text{State}_i + \epsilon_i$$

- **supporti**
  Definition: The support level for candidate (i) in a given poll.
  Justification: Measures the proportion of respondents supporting candidate (i), providing insight into their current standing.
- **sample_sizei**
  Definition: Represents the number of respondents in poll (i).
  Justification: Larger sample sizes generally yield more accurate estimates of support, reducing sampling variability.
- **days_to_electioni**
  Definition: Denotes the number of days remaining until the election when poll (i) was conducted.
  Justification: Accounts for temporal proximity to the election, capturing potential fluctuations in support as the election approaches.
- **transparency_scorei**
  Definition: A numerical score reflecting the pollster's transparency regarding their methodology, ranging up to 10, with higher scores indicating greater transparency.
  Justification: Higher transparency scores may correlate with the reliability and credibility of poll results, influencing voter trust.
- **pollscorei**
  Definition: A numeric value representing the reliability and bias of the pollster, where negative values denote better reliability.
  Justification: Accounts for systematic deviations in poll results based on pollster performance, ensuring more accurate support estimates.
- **statei**
  Definition: Captures state-specific effects as fixed effects.
  Justification: Allows the model to account for regional variations in support that are not explained by other predictors, incorporating state-specific nuances.
- **i**
  Definition: The error term, assumed to follow a normal distribution.
  Justification: Represents unexplained variability in the model, ensuring that the residuals meet the assumptions of the statistical analysis.

Variable Definitions:

- **Sample Size ((sample_size))**: The size of the poll is included as larger samples generally provide more accurate estimates of support, reducing sampling variability.
- **Days to Election ((days_to_election))**: This variable accounts for the temporal proximity to the election, as support levels may fluctuate as the election approaches, capturing trends in voter sentiment over time.

- **Transparency Score ((transparency_score))**: A higher transparency score indicates greater disclosure of poll methodology, potentially correlating with the reliability and credibility of the poll results.
- **Pollscore ((pollscore))**: This score represents the inherent bias and error of the pollster, with lower (more negative) values indicating higher reliability. It accounts for systematic deviations in poll results based on historical pollster performance [REF 538 METHOD].
- **State ((state))**: Including state as a fixed effect allows the model to account for regional variations in support that are not captured by other predictors, ensuring state-specific nuances are incorporated into the support estimates.

An alternative approach considered was modeling State as a random effect to account for unobserved heterogeneity across states. However, given the focus on specific state-level predictions and the availability of sufficient data per state, fixed effects are chosen in this case. Ideally we would fit both models and use diagnostics to choose which has lower bais and variance, and thus better predictive quality. Additionally, simpler models excluding variables like transparency_score and pollscore were evaluated but found inadequate in capturing the state-by-state differences in the transparency and reliability of the polls, potentially leading to biased support estimates.

Process:

For each candidate and state, we aggregated polling data by calculating the mean values of each predictor. For example, here is the mathematical notation of Mean Days to Election, this value was calculated for each state for both candidates:

$\text{Mean Days to Election}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \text{Days to Election}_i$

where (N_s) is the number of polls for state (s).

Averaging predictors per state smooths out poll-to-poll variability and provides a representative set of predictors for each state, allowing use to make reliable state-level predictions. Instead of averaging, a training/testing split or regional aggregations based on 538's political regions could have been employed. However, a training/testing split is beyond the scope of this paper, and using political regions would have complicated the derivation of the model.

Using the averaged state level predictors, we generated posterior predictions by drawing 1,000 samples from the posterior distributions of the model parameters. These predictions simulate 1,000 possible election outcomes, and by taking the proportion of victories for Harris and Trump we estimate the probability each candidate has of winning the election. To generate an election map, we simply take the average outcome of each state across the 1,000 simulations. Alternative methods could include using point estimates from the posterior mean; however, simulating multiple outcomes offers a more comprehensive assessment of uncertainty and variability in election outcomes.

For states lacking recent polling data, we incorporated historical averages of Democratic and Republican support from elections since 2000. This decision is based on the assumption that certain states are very unlikely to change their voting patterns, reducing the necessity for current polling data. The ten states that do not have polling data [REF] have not changed parties in at least the last 20 plus years. This may be why they do not have polling data, as it would be a waste of resources for agencies to poll them. Alternatively, imputation methods or political regional averages could have been used, but historical averages are used for their simplicity and relevance.

The Bayesian models were implemented using the `rstanarm` package in R, which interfaces with Stan for efficient Markov Chain Monte Carlo (MCMC) sampling. Data manipulation and visualization were conducted using the `tidyverse` suite of packages, while model diagnostics leveraged `bayesplot` and other

related packages. This combination of tools facilitated a streamlined workflow for model fitting, prediction, and validation.

-Diagnostics

Comprehensive diagnostics were performed to assess model convergence, fit, and predictive performance. Graphs and more details can be found in the relevant appendices. We evaluated model convergence using the R-hat statistic and trace plots. All R-hat values for model parameters were below 1.1, indicating satisfactory convergence. Trace plots revealed well-mixed chains without discernible trends, further supporting the reliability of the MCMC sampling process [REF].

Posterior predictive checks were conducted to assess the model's ability to replicate observed data. Density overlay plots for both Trump and Harris models indicated reasonable fits, although some discrepancies were noted near the peaks of the distributions. Specifically, the Trump model's replicated data tended to be slightly smaller around the peak, while the Harris model exhibited more extreme differences at the peak. Residual plots appeared randomly scattered, suggesting no major systematic biases, although some clumping was observed, indicating areas where the model may fit less effectively [REF].

Residual plots, depicting residuals versus fitted values, showed notable clumping in the central region for both models, along with vertical lines. This pattern may indicate unmodeled heterogeneity or data limitations, suggesting areas where the model's fit could be improved. Despite these observations, the overall residual distribution did not reveal significant systematic errors, supporting the model's general adequacy.

Ideally several models would be created then using model selection techniques would be used to determine the best model for predicting the outcome of the US election. Due to time constraints the folling models were considered but not implemented. A set of Bayesian models with state as a random effect and models excluding specific predictors like transparency_score. More complex models with additional interaction terms or non-linear components were also considered

3.0.0.1 Assumptions and Limitations

The model operates under several key assumptions:

1. **Linearity:** The relationship between predictors and support is linear.
2. **Normality of Residuals:** The error terms follow a normal distribution.
3. **Independence:** Observations are independent given the predictors.
4. **Stable Partisan Leanings:** Historical averages adequately represent states without current polling data.

Limitations:

- **Model Misspecification:** If non-linear relationships exist between predictors and support, the linear model may not capture these dynamics effectively.
- **Reliance on Historical Data:** Using historical averages for certain states may not account for recent political shifts or emerging trends.
- **Residual Clumping:** Observed clumping in residual plots suggests potential areas for model refinement, such as incorporating additional predictors or interaction terms.

The final Bayesian linear regression model balances complexity and interpretability, incorporating relevant predictors to capture poll reliability and state-specific effects while maintaining computational efficiency. The inclusion of transparency_score and pollscore enhances the model's ability to account for poll quality,

while fixed state effects ensure accurate regional support estimates. Validation through convergence diagnostics, posterior predictive checks, and residual analysis confirms the model's robustness and reliability within its assumptions.

All analyses were conducted using R (version 4.2.1) with the following key packages:

- **tidyverse:** For data manipulation and visualization.
- **rstanarm:** For Bayesian model fitting using Stan.
- **brms:** Alternative package for Bayesian regression modeling.
- **bayesplot:** For comprehensive model diagnostics and visualization.

4 Results

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

FOR IDEALIZED SURVEY !TODO like the other appendix, need to talk about depth about the techniques. Need to create a google form that is the actual survey. Every decision made about the idealized survey must be justified with literature. Every detail of the survey must be accounted for. Need to write about how the survey was created, and identify and justify all of the tradeoff made. TODO!

A Appendix

!TODO need to add details about the techniques used by the pollster. For example, if stratified sampling was used, dive into what is stratified sampling. What is it good for why is works. Reference a academic paper about stratified sampling. TODO!

B One particular poll within our data sample is one conducted between October 11-14 by Beacon Research and Shaw & Company Research (cite Fox article). Their analysis found 48% of participants favoured Harris while 50% favoured Trump. Targeting the sample frame of registered American voters, this poll collected data on 1,110 participants through either live phonecall interviews or an online survey. The sample was recruited by applying the probability proportionate to size method on the nationwide voter file of registered voters' phone numbers, meaning that the contacted individuals would be proportionally representative of the number of voters per state (cite Fox article). This ensures that the voices represented by poll results would match the ones involved with the real election as much as possible.

B.1 Analysis into a Pollster's Methodology

One particular poll within our data sample is one conducted between October 11-14 by Beacon Research and Shaw & Company Research (cite Fox article). Targeting the sample frame of registered American voters, this poll collected data on 1,110 participants' responses on 51 questions through either live phonecall interviews or an online survey. The sample was recruited by applying the probability proportionate to size method on the nationwide voter file of registered voters' phone numbers, meaning that the contacted individuals would be proportionally representative of the number of voters per state (cite Fox article). This ensures that the voices represented by poll results would match the ones involved with the real election as much as possible.

The key finding from this poll is that 48% of participants favored Harris while 50% favored Trump, with the lead being consistent in the larger sample of registered voters and smaller subsample of likely voters. Despite this, responses also indicate that Harris is narrowly leading in seven swing states, meaning that Democrats could potentially win by electoral college while losing the popular vote (cite Fox article). This report also notes that current results are Trump's highest approval ratings since Biden dropped out of the race, while support for Harris is at its lowest.

One strength of this questionnaire is that information regarding the respondent's confidence level and commitment level was also collected. With questions such as "How often do you make a point to read or listen to the news?" and "Are you certain to support that candidate, or do you think you may change

your mind and support someone else?", this poll is able to gain a more indepth view of how easily these participants may be swayed in the time between the polling and the real election (cite Fox article). As well, probabilistic statistical models based on past voting history, interest in current election, age, education, race, ethnicity, church attendance, and marital status were used to predict the respondents' likeliness to vote; interestingly, results found that the results for the full sample vs subsample of 870 likely voters varied by a ±3%, which can be quite significant given how tight the race is currently (cite Fox article).

On the other hand, one potential weakness of this poll is that it was sponsored by Fox News, a news site that has historically been Republican-leaning. Though from the reported methodology alone, no evident biasing of the pollster can be observed, this potential source of biased reporting must be noted as it can affect the type of analyses that occurred and which key findings are the focus of news reports. As well, another potential issue is the lack of indication regarding how non-responses of "Don't know" were handled during analysis (cite Fox article). This is a critical area to pay attention to as it is an option offered on every question, selected by up to 4% of participants on key questions such as "If the presidential election were today, how would you vote?" (cite Fox article).

B.2 Idealized Methodology and Survey

B.3 Table Detailing First Few Observations of Dataset

Table 1: First few observations of cleaned data set

pollster	numeric_grade	pollscore	end_date	transparency_score	question_id	method
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215183	Online Ad

Table 2: First few observations of cleaned data set

state	sample_size	candidate	support	election_date	days_to_election	initial_weight	weight	pol_region
North Carolina	1373	Harris	46.7	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Trump	50.7	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Stein	0.7	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Oliver	0.3	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Harris	47.0	2024-11-05	5	0.9486833	1.094691	Southeast

References

Chad de Guzman and Koh Ewe. 2024. "A Guide to Kamala Harris' Views on Abortion, the Economy, and More." https://time.com/7001208/kamala-harris-views-abortion-economy-immigration-israel-gaza/.

FiveThirtyEight. 2024a. "FiveThirtyEight Latest Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

———. 2024b. *FiveThirtyEight Presidential General Polls Data*. https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

McKenzie Beard and Abbie Cheeseman and Justine McDaniel. 2024. "Harris Vs. Trump on Abortion: Where They Stand on the Issue." https://www.washingtonpost.com/politics/interactive/2024/trump-harris-abortion/.

NBC. 2024. "Harris and Trump: Compare Where They Stand on Key Issues." https://www.nbcnews.com/politics/2024-election/harris-trump-stance-issues-policies-president-race-rcna150570.

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

Weber. 2024. "Electoral College Information." https://www.sos.ca.gov/elections/electoral-college#:~:text=Under%20the%20%22Electoral%20College%22%20system,more%20%22votes%22%20it%20gets.

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.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.