Predictive Modeling for Forcasting the 2024 US Presidential Election*

Trump's Narrow Victory Over Harris

First author Another author

November 2, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

2 Data

2.1 Overview

In this analysis, we used R (R Core Team 2023) to investigate polling data on public sentiment leading up to the election. Our dataset, sourced from FiveThirtyEight (FiveThirtyEight 2024), provides a detailed snapshot of shifting public opinion over time. We examined key factors influencing support percentages, including poll timing, pollster characteristics, and state-specific trends.

*Code and data are available at: https://github.com/kqlqkqlqF/Insights-and-Predictions-for-the-U.S.-Election.git.

Several R packages were instrumental in facilitating data manipulation, modeling, and visualization. Tidyverse served as the foundation for organizing and efficiently analyzing the data, seamlessly integrating multiple analytical tasks (Wickham et al. 2019a). The Here package simplified file path management, ensuring smooth data access across systems (Müller 2020). We utilized Janitor for comprehensive data cleaning, which helped us identify and correct inconsistencies (Firke 2023), while Lubridate supported the handling of time-related variables (Grolemund and Wickham 2020). Finally, Arrow enabled fast, memory-efficient access to large datasets, a crucial asset when working with extensive polling data (**citearrow?**). Our codebase and workflow adhered closely to best practices, as outlined in Alexander (2023).

Our group focused on Trump's approval ratings, aiming to ensure the credibility of the data. To achieve this, we selected only pollsters with numeric grade above 2.0, using data collected from November 15, 2022, to October 27, 2024.

2.2 Measurement

In this section, we will describe the process of converting raw poll data into a structured dataset for analysis. In this process, because this study focuses on studying the changes in Trump's support rate and predicting whether Trump can be successfully elected, all data collection and analysis will be carried out around Trump and his main opponent Harris. Raw poll data comes from actual polls conducted by various organizations across the United States. Each pollster uses different methods, such as online panels and Live Phone surveys, to record whether the public supports Donald Trump. After the poll results are collected, they are aggregated into comprehensive datasets, such as the dataset provided by FiveThirtyEight (FiveThirtyEight 2024). In this dataset, key factors include the start and end dates of the poll, the identity of the pollster, the state, the pollscore, and the numeric grade, which is an indicator to evaluate the reliability of each poll. These parameters will be explained in detail below. This structured dataset allows us to analyze Trump's support patterns and trends over time and across regions. We will explore how these factors affect public sentiment and predict the likelihood of Trump becoming the next US president.

- **Support Percentage (pct):** The percentage of respondents supporting each candidate, acting as the primary outcome variable for analysis.
- **State:** The geographical area covered by the poll, either state-specific or nationwide.
- **Poll ID:** A unique identifier for each poll, enabling easy tracking and management of entries.
- **Pollster:**The organization that conducted the poll, providing insight into the methodological quality.
- **Poll Score:** A measure of the pollster's reliability, with lower (often negative) values indicating higher predictive accuracy.

- **Numeric Grade:** A measure of the credibility or quality of the poll. To ensure higher credibility of the results, we removed all the original poll data with numeric grade lower than 2.0.
- **Sample Size:** The total number of respondents in each poll, which impacts the poll's statistical precision and margin of error.
- **Candidate Name:** The name of the candidate evaluated in the poll, allowing for candidate-specific analysis.
- **Start Date:** The starting date of the poll, aiding in temporal alignment for trend analyses.
- **End Date:** The completion date of the poll, aiding in temporal alignment for trend analyses.

Data analysis was enhanced by various packages. The tidyverse (Wickham et al. 2019b) suite facilitated efficient data manipulation and visualization, while ggplot2 (Wickham 2016) allowed for compelling visualizations. We used ggmap (McGlinn and Wickham 2023), built on ggplot2, to generate a map of shelter distribution in Toronto via the Google API. The here (Müller 2020) package simplified file management in our project directory. We utilized kableExtra (Zhu 2021) for visually appealing and customizable tables. For Bayesian analysis, we employed rstanarm (Goodrich et al. 2022), providing an elegant interface to Stan for estimating data relationships within a Bayesian framework. Report generation was managed with knitr (Xie 2023), enabling seamless integration of R code into our document. Other essential packages included tibble (Müller and Wickham 2022), stringr (Wickham 2020), lubridate (Grolemund and Wickham 2020), janitor (Firke 2023), and testthat (Wickham and RStudio 2020), contributing to various aspects of data analysis, from manipulation to quality assurance.

2.3 Outcome variables

2.3.1 Overview of Trump's Electoral Support

The Figure 1 illustrates the distribution of approval ratings for Trump. The majority of the approval ratings fall between 40% and 55%, forming a shape that resembles a normal distribution, with a peak around the 45% to 50% range. This suggests that, within the analyzed sample, most of the approval ratings cluster in this middle range, with relatively few instances of extremely high or low ratings.

The lower frequency of approval ratings below 30% and above 60% indicates that these extremes are relatively uncommon in the dataset. Overall, the concentration of support in this central range suggests a fairly consistent level of public support for Trump.

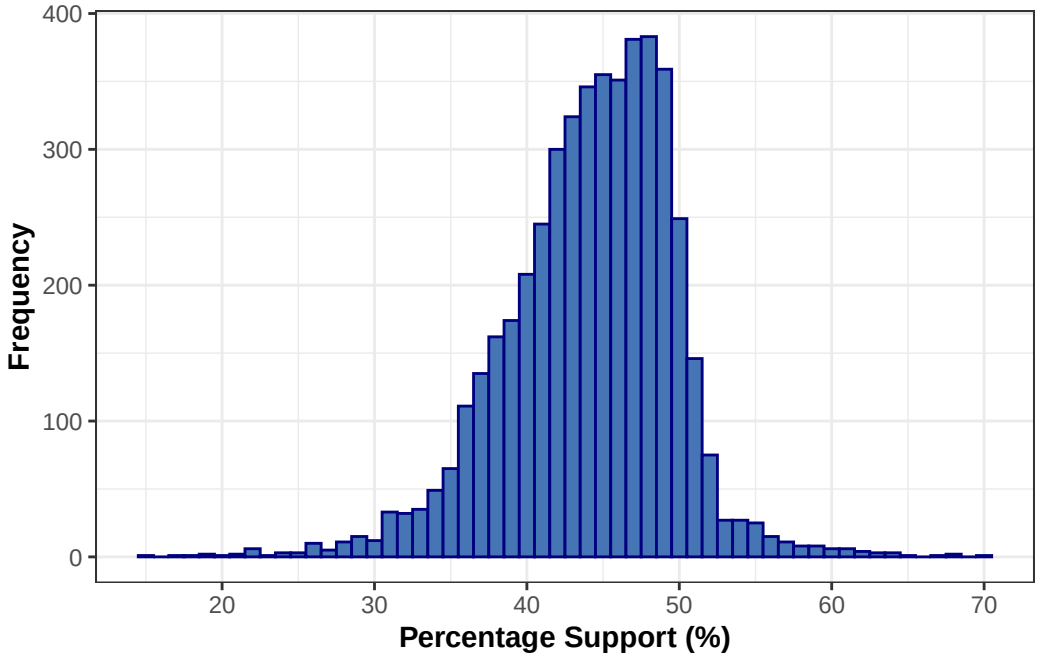

Figure 1: Distribution of percentage support for Trump

2.4 Predictor variables

2.4.1 Summary of Predictor Variables

- **State:** The U.S. state where the poll was conducted, if applicable.
- **Numeric Grade:** A numeric rating from 0.0 to 3.0 indicating each pollster's reliability.
- **Sample Size:** The total number of respondents participating in the poll.
- **Poll Score:** A quantitative measure of the pollster's reliability, where lower values suggest higher predictive accuracy.
- **Recency Weight:** A metric used to assess the relevance of polling data based on how close the polling dates are to an upcoming election. It is calculated by evaluating the number of days until the election from both the start and end dates of the poll, normalizing these values against the maximum days from other polls. The resulting weight gives more importance to more recent polling data, reflecting its greater influence on understanding current public sentiment..
- **Candidate Name:** Indicate the corresponding presidential candidate. Trump was represented by 0, while Harris was represented by 1.

2.4.2 State

In Figure 2, we analyzed the distribution of Trump's popular support rate across states. Trump's support rate aligns closely with each state's party preference: Republican states show high support for Trump, exceeding 50%, followed by swing states, where support typically hovers around 50%. In Democratic-leaning states, Trump's support is significantly lower, generally below 50%. The "national" category in the chart represents data covering the entire country without specifying a particular state. Trump's national support approaches but does not reach 50%.

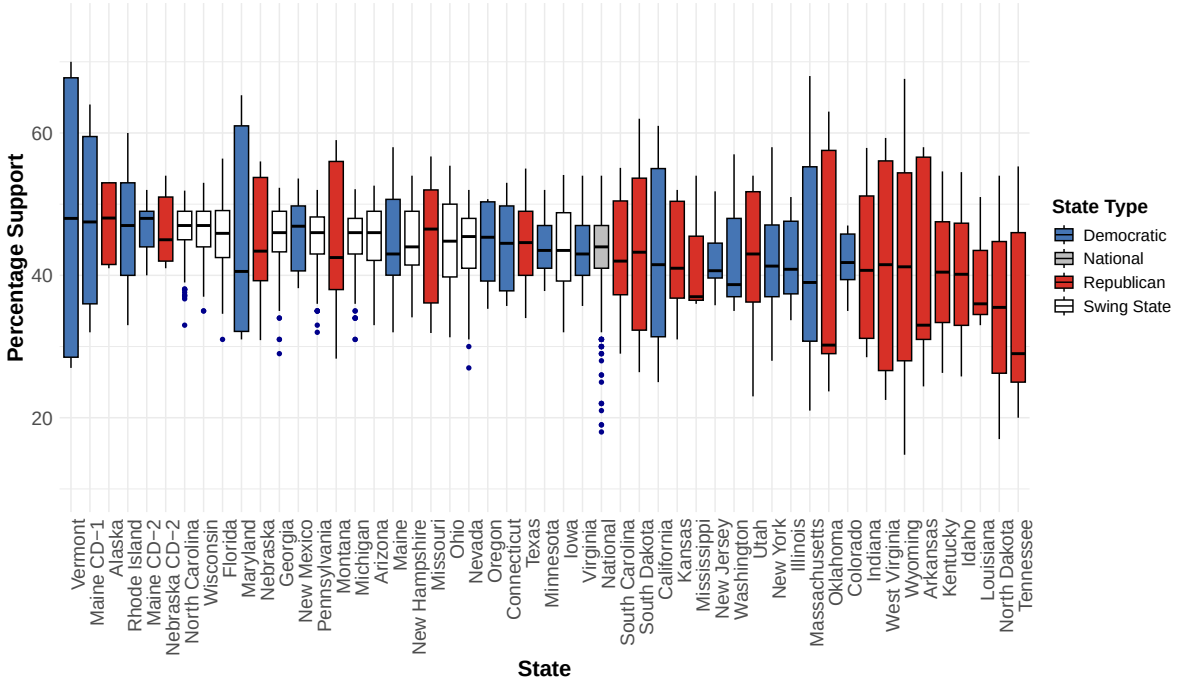

Figure 2: Overview of the Percentage Support of Trump Across Different States

2.4.3 Numeric Grade

In Figure 3, we analyzed the relationship between numeric grade and Trump's support rate. Each point in the chart represents a poll, with its numeric grade on the x-axis and Trump's support rate on the y-axis. The nearly flat trend line suggests that numeric grade has no clear relationship with Trump's support rate. However, this is a basic analysis and does not rule out the possibility that numeric grade could impact Trump's support rate under different variable conditions.

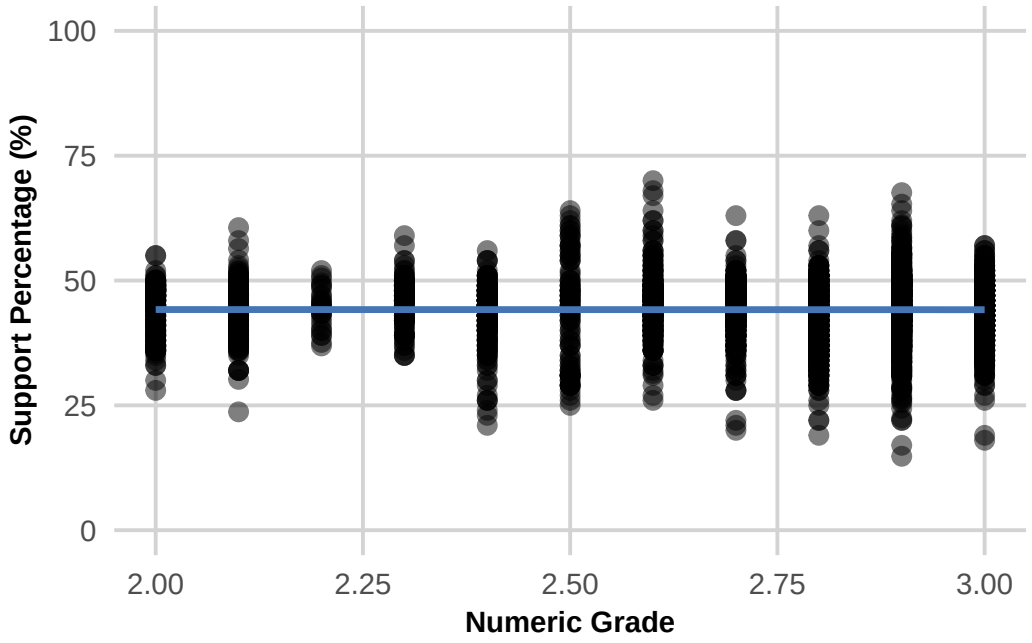

Figure 3: Relationship between Numeric Grade and Support Percentage of Trump

2.4.4 Sample Size

In Figure 4, we examined the relationship between sample size and Trump's support rate. The results show a slight downward trend in Trump's support as sample size increases. However, since most data points are concentrated in the 0–4000 range, with fewer data points above 4000, this may not accurately reflect the true relationship between sample size and support rate.

2.4.5 Poll Score

Analysis of Figure 5 shows no clear proportional or inverse relationship between poll scores and Trump's support rate. This suggests that while a higher poll score may indicate greater reliability, it does not directly translate into changes in candidate support. This also supports the data's credibility, as Trump's support rate remains consistent regardless of pollster ratings.

2.4.6 Recency Weight

Figure 6 shows the relationship between poll recency and Trump's support rate. It indicates a slight upward trend in Trump's support as polls get closer to Election Day, though the increase

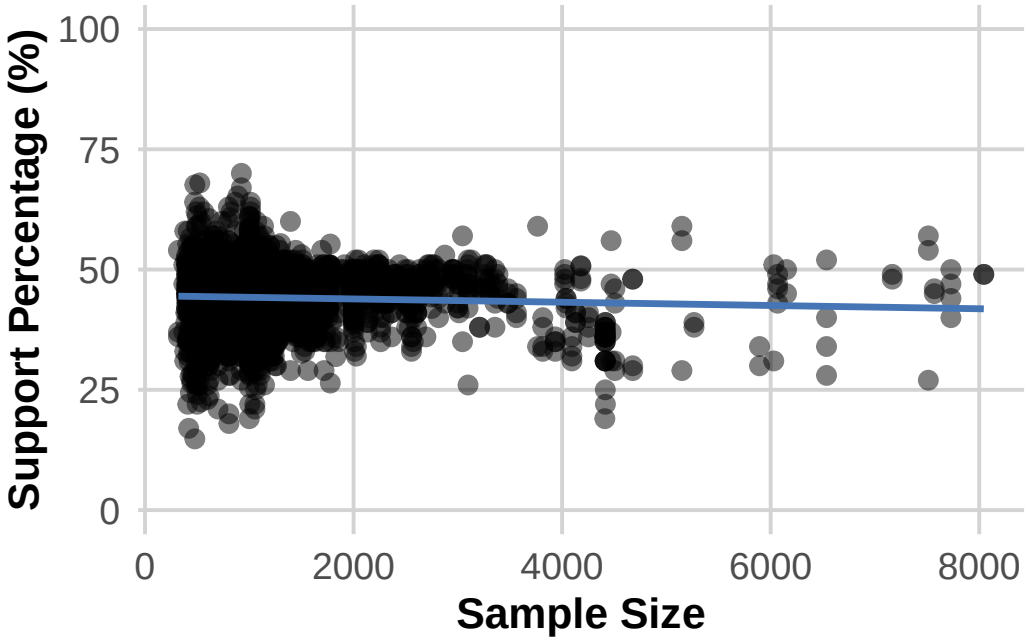

Figure 4: Relationship between Sample Size and Support Percentage for Trump

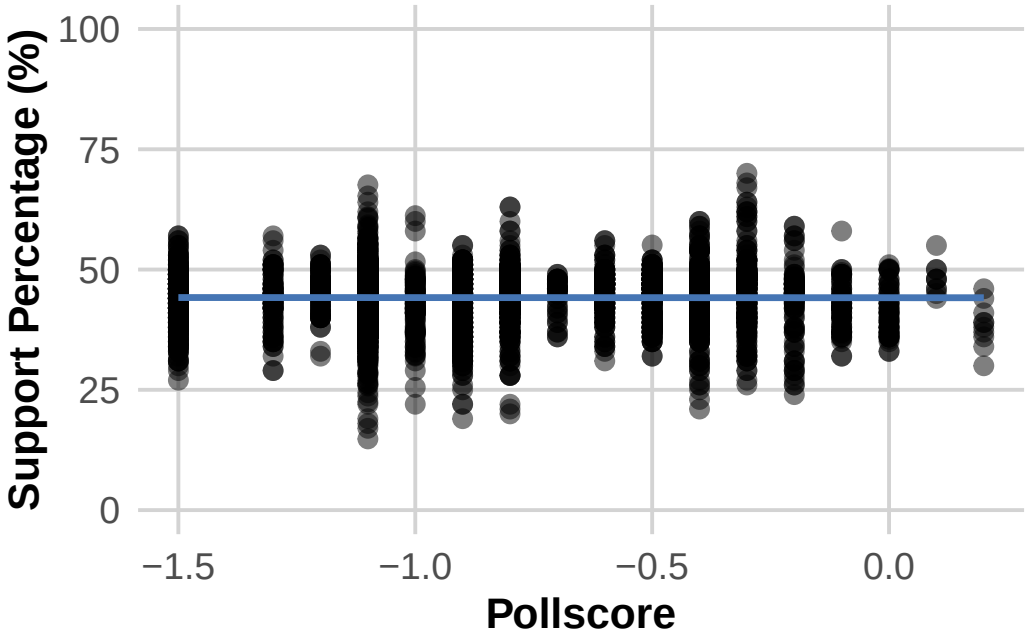

Figure 5: Relationship between Pollscore and Support Percentage of Trump

is not significant.

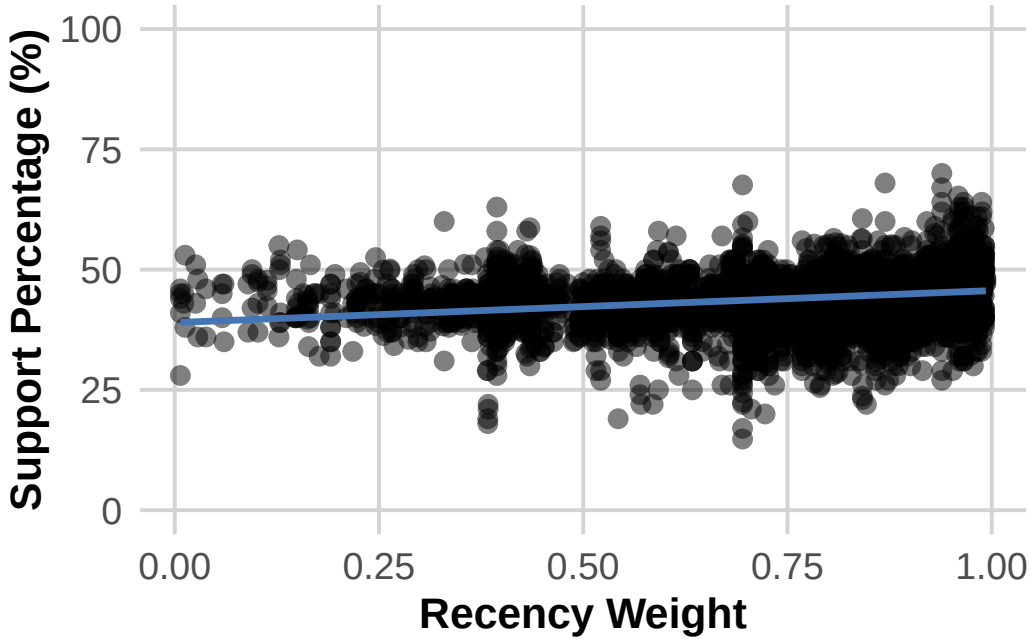

Figure 6: Relationship between Recency Weight and Support Percentage of Trump

2.4.7 Candidate Name

To predict Trump's probability of winning the election, we will create a model containing the aforementioned predictor variables to forecast the support rates for both Trump and his main opponent, Harris. To ensure fairness in the results, we will use the same model to predict the support rates for Trump and Harris. Thus, we have created a variable called candidate_name, where Trump is represented by a value of 0 and Harris by a value of 1.

3 Discussion

Appendix a.

3.1 Overview of Emerson College Polling Methodology (October 23-25, 2024)

The Emerson College Polling conducted a survey from October 23 to 25, 2024, targeting 1,000 likely voters to investigate the differences in support for various candidates. In this presidential election, 58% support former President Donald Trump, while 39% support Vice President Kamala Harris.

3.2 Population, Frame, and Sample

In this context, the target population consists of likely voters in the U.S. elections, defined by their likelihood to vote in the upcoming elections and their voting history, both of which are self-reported in the survey. The sampling frame specifically focuses on likely voters in Montana, who were reached through a combination of cell phone contacts provided by Aristotle and an online voter panel from CINT. The sample consists of 1,000 likely voters randomly selected from the sampling frame, with their status determined by a combination of voter history, registration status, and demographic data, all of which are self-reported. This methodology provides a balanced overview of Montana voters' priorities, with a credibility interval of +/- 3%.

3.3 Sampling Approach and its trade-offs

Emerson College utilized a mixed-mode sampling approach for its poll of likely voters in Montana. This strategy involved two main methods: sending MMS text messages linked to an online survey using Aristotle's voter lists and accessing a pre-screened, opt-in online panel from CINT. The MMS method is efficient and cost-effective, allowing participants to complete the survey at their convenience, which can enhance response rates. The online panel broadens coverage to include voters not reachable through text, capturing a wider demographic range across the state. Together, these methods create a diverse sample while reducing costs compared to traditional phone or in-person interviews.

However, this approach has trade-offs. The MMS survey requires recipients to have active cell phones and internet, potentially excluding older or less tech-savvy voters. Additionally, the online panel consists of self-selected participants, which may not fully reflect the general voter population. Mixing data from both sources can introduce inconsistencies, as each method may attract different respondent types, necessitating careful weighting to maintain balance and accuracy. Smaller demographic subsets, such as age, race, or education, carry higher credibility intervals due to reduced sample sizes, limiting precision in analysis. Overall, the mixed approach optimizes reach, reduces costs, and captures a representative snapshot of Montana's voter priorities, albeit with some limitations.

3.4 Non-response Handling

Emerson College does not provide specific details regarding its non-response management. While it mentions that data were weighted by demographics such as gender, education, race, age, party registration, and region to align with the 2024 likely voter model, this weighting primarily addresses demographic imbalances and does not directly mitigate non-response bias. The survey lacks information on common non-response strategies, such as follow-up attempts, participation incentives, or specific adjustments for non-responders. This absence raises concerns about potential non-response bias, particularly if certain demographic groups were less likely to engage with the survey.

3.5 Questionnaire Design

This questionnaire has strong points. Its straightforward and clear wording makes questions easy for respondents to follow and reduces potential confusion. By focusing on core issues like the economy, housing, and voter approval for specific candidates, it captures key voter concerns in Montana, offering a concise yet comprehensive view. The use of multiple questions around candidate approval, voter issues, and demographics adds depth and helps validate responses across topics. However, the questionnaire also has limitations. While demographic questions enhance the survey's representativeness, smaller groups (e.g., nonbinary individuals) may carry higher credibility intervals, reducing precision for those subgroups. The mixed-mode approach (online panel and mobile) improves access but still risks non-response bias, as certain demographics might be less likely to participate. Overall, the design achieves clarity and breadth, though response biases and sample variations should be considered in interpreting the findings. For example, in this survey of 1,000 Montana voters, only 5 respondents identified as nonbinary or other genders. Since statistical reliability depends on the number of responses, small groups have higher variability, meaning their responses can swing widely due to each individual answer carrying greater weight.

Appendix b.

3.1 Idealized Methodology for Forecasting the U.S. Presidential Election

We aim to develop a methodology for forecasting U.S. presidential election outcomes by conducting a survey with a $100,000 budget. Using stratified random sampling and multi-mode recruitment, the survey targets 10,000 likely voters across demographic and regional lines. Key measures include data validation checks, weighted analysis, and predictive modeling to ensure accuracy. Results will be enriched by aggregating reputable data sources like FiveThirtyEight for a comprehensive, reliable forecast.

3.2 Budget Allocation

Funding allocations will focus on ensuring thorough and effective sampling, recruitment, data validation, and analysis methods are employed with a total budget of no more than $10,000. The proposed budget breakdown is as follows:

Survey platform costs: $10,000 (subscription fees for online survey tools such as Google Forms or Qualtrics) Respondent incentives: $10,000 (gift cards or other incentives to encourage participation) Recruitment and staffing: $35,000 (staffing costs for survey distribution and data collection) Data analysis tools: $20,000 (statistical software licenses, data cleaning and analysis) Marketing and promotion: $10,000 (awareness and engagement campaigns) Contingency fund: $5,000 (for unexpected expenses)

3.3 Sampling Approach

The sampling approach will employ a stratified random sampling method to ensure representation across various demographic groups, including age, gender, race, education level, geographical location, and party affiliation. The target population consists of likely voters in the U.S., defined by historical voting behavior and self-reported intentions to vote. A sample size of approximately 10,000 respondents will be aimed at ensuring statistical robustness and a credibility interval of +/- 1% at a 95% confidence level. This will be achieved through a combination of national voter registration databases to identify potential respondents. For example, we can utilize the National Voter Registration Act (NVRA) data from the National Association of Secretaries of State (NASS) to access information on registered voters. This database will allow us to filter for likely voters based on their registration status and historical voting behavior, ensuring that our sampling frame is comprehensive and representative of the electorate. By leveraging such reliable sources, we can create a sampling framework that enhances the accuracy of our election forecasts.

3.4 Respondent Recruitment

To reach the target population, a multi-mode recruitment strategy will be implemented: Online Surveys: Use Google Forms to distribute the survey electronically. Telephone Surveys: Conduct live telephone interviews to capture demographics that might not engage online. Text Messaging Surveys: Implement SMS surveys to reach younger demographics and those without regular internet access. Incentives: Offer gift cards or other small incentives for participation, particularly for online respondents. This can enhance response rates and engagement.

3.5 Data Validation

To effectively reach our target population and capture a diverse range of perspectives, we will implement a multi-mode recruitment strategy tailored to different demographic groups and communication preferences. This approach includes online surveys using a Google Forms questionnaire, which will be widely distributed through social media, email lists, and community networks to maximize reach among individuals who frequently engage online. The user-friendly platform allows participants to complete the survey quickly and anonymously on any internet-enabled device. The survey can be accessed through the following link: https://forms.gle/oSbad52Vuw9Z9Wf46. Additionally, we will conduct live telephone interviews to include participants who may not be reachable through online channels, ensuring we capture responses from populations that might otherwise be underrepresented. To further engage younger demographics and individuals with limited internet access, we will implement SMS-based surveys, allowing participants to respond quickly via text. To encourage participation and improve response rates, small incentives such as digital gift cards will be offered, particularly for online respondents, with details communicated at the survey's start and awarded upon completion to ensure transparency. This comprehensive approach will enable us to gather a robust and representative dataset, providing valuable insights into the preferences and priorities of voters across multiple demographics.

3.6 Poll Aggregation and Modeling

Poll results will be aggregated using statistical methods to identify trends and analyze historical voting patterns. We will implement a weighted analysis to ensure demographic representation, applying specific weights based on factors such as age, gender, race, and education level, as well as state significance to reflect regional variations in voter behavior. Predictive analytics will primarily involve logistic regression to model voting preferences and forecast election outcomes, supplemented by time-series analysis to track changes in voter sentiment over the campaign period. To enhance our findings, we will combine our data with reputable sources like FiveThirtyEight, utilizing their aggregation techniques to enrich our analysis with broader insights.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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

FiveThirtyEight. 2024. "Dataset: US Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2020. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

McGlinn, Dale S., and Hadley Wickham. 2023. *Ggmap: Spatial Visualization in r*. https://cran.r-project.org/web/packages/ggmap/index.html.

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

Müller, Kirill, and Hadley Wickham. 2022. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org.

———. 2020. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

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

———, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and RStudio. 2020. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.