My title*

My subtitle if needed

First author   Another author

November 1, 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.

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. 2019). The Here package simplified file path management, ensuring smooth data access across systems (**citehere?**).

We utilized Janitor for comprehensive data cleaning, which helped us identify and correct inconsistencies (**citejanitor?**), while Lubridate supported the handling of time-related variables (**citelubridate?**). 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 ratings above 2, using data collected from November 15, 2022, to October 27, 2024.

The main useful variables that the data contains are: - **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 - **Pollscore**: A measure of the pollster's reliability, with lower (often negative) values indicating higher predictive accuracy - **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 - **End Date**: The completion date of the poll, aiding in temporal alignment for trend analyses

2.2 Measurement

In this section, we describe the process of transforming raw polling data into a structured dataset for analysis. The data is sourced from actual polls conducted by various organizations across the United States, each using different methodologies to record public support for Donald Trump. This diversity of sources enhances the range and credibility of the data.

Once collected, the poll results are aggregated into comprehensive datasets, such as those provided by FiveThirtyEight (FiveThirtyEight 2024). Key variables include the poll's end date, pollster identity, state, and a quality score assessing each poll's reliability. These variables play a role in shaping Trump's approval ratings and can improve our ability to predict his likelihood of becoming the next U.S. President.

This structured dataset enables us to analyze patterns and trends in Trump's support over time and across regions, allowing us to explore how factors like state, poll score, and pollster influence public sentiment. By grounding our analysis in actual polling data, we ensure greater credibility in our results, ultimately supporting more accurate predictions for the upcoming 2024 U.S. election.

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.
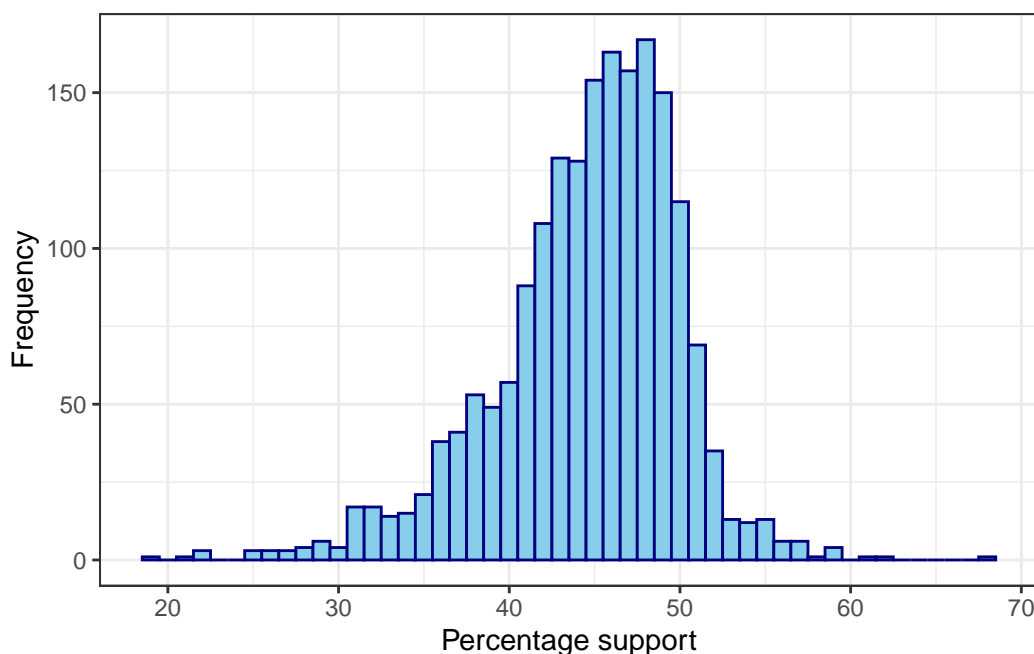
Figure 1: Distribution of percentage support for Trump

2.3.2 Overview of election data in different states

Here we looked at how each state polled in the popular vote. What we found for each state had to do with its party affiliation. the REP partisan states would have a higher Trump vote, above 50 per cent. then there are the swing states where the number of votes generally swings around the 50 mark. And states that are predominantly DEM partisan will have much lower support for Trump generally below 50 per cent

Figure 2: state and poll

2.3.3 Overview of voting method

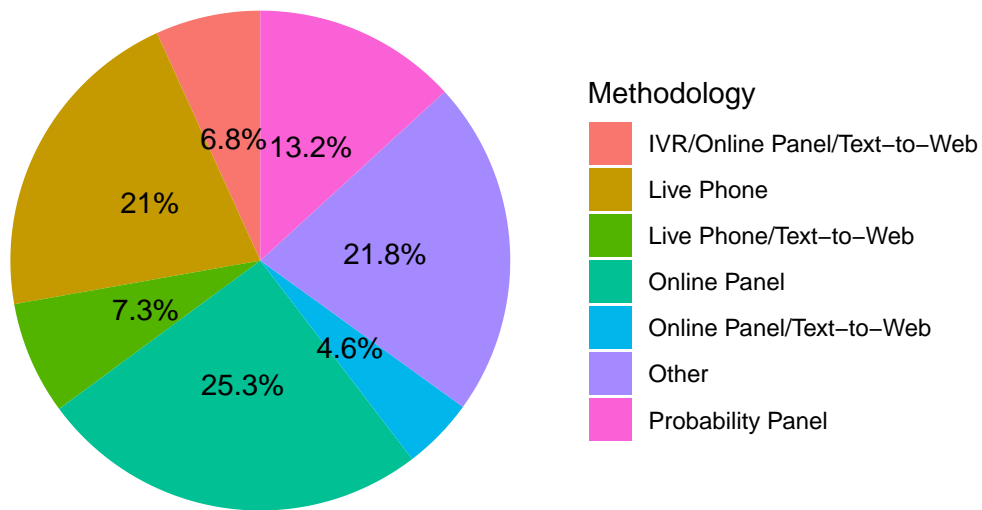The primary data collection methods include online panels, live phone interviews, automated phone surveys, and mail questionnaires. These methods are widely used due to their efficiency in reaching diverse respondent groups. Online panels are suitable for internet users, live phone interviews allow for detailed feedback and access to those less active online, automated phone surveys are cost-effective with broad reach, and mail questionnaires are ideal for older individuals or those not using the internet. The variety of these methods ensures broad representation and accuracy in the data.

2.4 Predictor variables

Our predictor variables include candidate_name, sample_size, pollscore, numeric_grade, recency_weight, and state. We intend to leverage these variables to predict Trump's likelihood of being elected. - **numeric_grade**: A numeric rating indicating each pollster's reliability - **state**: The US state where the poll was conducted, if applicable. - **sample_size**: The total number of respondents participating in the poll - **Pollscore**: A quantitative measure of the pollster's reliability, where lower (negative) values suggest higher predictive accuracy - **recency_weight**: Indicates how recent each poll is, calculated based on the time difference from the most recent poll.

4

Proportion of Methodology Types

Figure 3: method
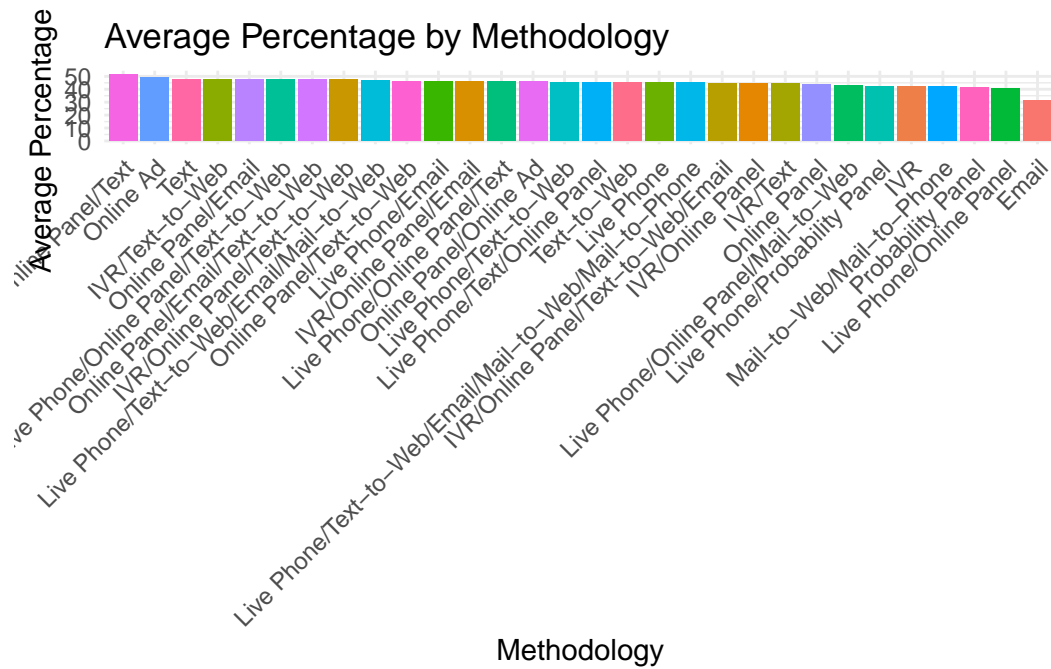
Figure 4: method-pct

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

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

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, 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.