My title*

My subtitle if needed

Chenming Zhao

October 30, 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

Polling data for the 2024 U.S. presidential election is provided by FiveThirtyEight (FiveThirtyEight 2024). This dataset compiles information from numerous polls conducted by various pollsters, detailing public support levels for presidential candidates. This data is periodically updated on FiveThirtyEight's website to reflect the latest polling information for the 2024 election. Each record includes information we need like state in which the poll was conducted (if applicable), polling dates (start and end), the specific pollster, the sample size of each poll, the polling methodology (such as phone, online, etc.), and support percentages for each

*Code and data are available at: https://github.com/RohanAlexander/starter_folder. We gratefully acknowledge FiveThirtyEight for providing the polling data used in this analysis. The data is available at: https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

candidate. This dataset serves as the foundation for analyzing trends and building predictive models for the winner of the upcoming US presidential election. Table tbl-pollingdata below previews this information for polls conducted in 2023 and 2024.

To simulate, test, analyze, and clean the polling data, the statistical programming language R was employed (R Core Team 2023). Key libraries that supported the data analysis include `tidyverse` (**tidyverse?**), `readxl` (Wickham and Bryan 2023), `dplyr` (Wickham et al. 2023), `knitr` (Xie 2023), `ggplot2` (Wickham 2016), and `broom` (Robinson, Hayes, and Couch 2024) for tidying model outputs. Additionally, the `rstanarm` (Goodrich et al. 2022) library was utilized to implement multilevel modeling for prediction purposes. Also this analysis inherits the folder structure from Professor Rohan Alexander from University of Toronto (Alexander 2024).

2.2 Measurement

The process of transforming real-world phenomena into structured data typically involves several critical steps, from designing surveys to final data entry. First, we identify what needs to be measured. In this case, we focus on public opinion regarding the 2024 presidential election, particularly support rates for various candidates. It is essential to define the data scope (e.g., national or state level) and key variables—such as polling information (e.g., polling organization, start and end dates), sample details (e.g., sample size and demographics), and candidate details (e.g., candidate name and party affiliation).

Next, polling organizations develop survey questions based on the desired data, considering sample size, sampling methods, and survey modes to capture respondents' preferences and voting intentions. Different methodologies impact the results, as they reflect the views of distinct groups, which may be influenced by factors like access channels, demographics, and technology usage. For instance, face-to-face surveys tend to capture opinions from offline-inclined groups, whereas voluntary online surveys exclude those less active online cite .

Once the survey is designed, polling begins, and responses are aggregated by the polling organization. Respondents' answers are recorded and categorized, converting their opinions into structured data points, though these lack the nuances of human complexity. To protect privacy, data are often anonymized and aggregated, retaining key insights. After collection, raw data undergo cleaning and processing to minimize sampling biases and errors. Yet, any single poll can have random and systematic error sources, so combining multiple poll results helps mitigate these errors cite .

Processed data are entered into structured datasets. Ideally, well-formatted polling data are stored as individual records, typically including fields like polling organization, sample size, survey dates, margin of error, candidate support percentages, and methodology description. On platforms like FiveThirtyEight, these fields are standardized for ease of comparison and analysis. Finally, FiveThirtyEight and similar platforms aggregate these structured records

and present them to the public or allow non-commercial research use, with data usually updated as new polls are released. This enables analysts and the public to track changes in candidate support, analyze trends, and interpret results. An interesting phenomenon, however, is that as election day approaches, poll results often "converge" or "follow" the average, reducing polling error, with evidence suggesting some "final adjustments" in methodology may occur cite .

2.3 Analysis Data

This analysis is focused on Donald Trump, so we only collected observations where the candidate_name is Trump. At the same time, we prioritize the quality of the polls. In a political approval rate prediction model, which is often influenced by numerous complex factors, we encounter data that is typically complex and noisy. Low-quality poll results can introduce significant bias into our prediction model, further impacting the accuracy of inferences. Therefore, only polls with a numeric_grade above 2.7 were selected. Additionally, Trump's main opponent, Kamala Harris, only entered the race after Biden's withdrawal, a major political event that could have significant effects on the support rates for both sides. As such, we focused solely on polls conducted after Biden's withdrawal to enhance the model's prediction accuracy for the final outcome.

In researching the likelihood of Trump winning, we selected specific variables to focus on those most impactful for predicting approval rates. Here is an overview of data used for this analysis Table 1. Following are selected ourcome variables and predictor variables, and explainations respectively.

2.3.1 Outcome variables

In this analysis of Donald Trump's support rate in the 2024 U.S. presidential election, the outcome variable is Trump's support rate (pct) because it is directly related to the election results. The primary research goal of this analysis is to predict Trump's chances of winning in the 2024 election, so the support rate is the main metric we aim to model and forecast. The support rate represents the proportion of respondents in a specific poll who support Trump. We seek to understand how the support rate is influenced by different factors, such as polling methodology, time, or geographical variables. By modeling the fluctuations in Trump's support rate, we can infer his electoral performance, analyze the trends within his support base, and assess the impact of various factors on his popularity. Understanding these dynamics provides insights into Trump's standing relative to his opponents and supports our overall election prediction.

2.3.2 Predictor variables

In this analysis, the predictor variables are the factors that we believe can influence the support rate for Trump. These variables are chosen based on their potential to explain changes in support rate. Following are the predictor variables that will be included in our model, along with an explanation of their significance.

start_date: Public sentiment and approval rates fluctuate dynamically during an election period, influenced by the passage of time, major events, and campaign strategies. By recording the date of each poll, we can observe changes in support over time.

state/national: Political environments and voter preferences may vary significantly by state, and national polls cannot fully reflect conditions in individual states. The electoral college system in the U.S. makes state-level approval rates crucial for predicting the likelihood of winning.

pollster: Different polling organizations use varying survey methods, sampling strategies, and data-processing procedures, which impact the reliability and accuracy of the poll results. Knowing which polling organization published the data helps assess potential biases in the data.

methodology: Different survey methods (e.g., telephone interviews, online questionnaires) significantly impact the representativeness of the results, as they directly influence the characteristics of the sampled population. Some methods may introduce biases toward specific demographic groups; for example, telephone surveys may skew toward older demographics, while online surveys may skew toward younger ones.

is_national: This variable helps us distinguish between national and state-level polls, ensuring the model can adapt and adjust across different geographic levels.

days_since_Biden_Withdrawal: Trump's main opponent, Kamala Harris, entered the race only after Biden withdrew. Major political events like Biden's withdrawal can significantly impact voter approval rates, and tracking the time since this event helps assess both its short-term and long-term effects.

Other potential variables (such as voter demographics like gender, age, race, and income) could indeed influence voter approval rates, but they primarily pertain to a **micro-level** analysis. After incorporating GDP and per capita income data from various U.S. states and nationwide in 2023 into a generalized linear model, we found that their correlation with presidential election outcomes was not very strong, likely due to data limitations in capturing the effect of economic changes on the election. Thus, we did not include these factors. Additionally, individual voter characteristics (such as gender and education level) are not consistently available in most polls. In comparison, the variables we chose are more complete and comprehensively capture the macro nature of the data. Therefore, selecting these variables allows us to maintain data simplicity while maximizing the capture of key factors influencing approval rates, avoiding the noise introduced by excessive micro-level variables.

```
analysis_data <- read_csv(here::here("/cloud/project/data/02-analysis_data/analysis_trump_dat
```

Table 1: The First Few Lines of Cleaned Analysis Data

start_date	state/national	pollster	methodology	Trump Support Rate	is_national	days_since_Biden_Withdrawal
2024-10-23	Texas	Siena/NYT	Live Phone	52	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	50	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	52	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	51	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	55	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	53	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	55	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	53	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	42	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	41	0	94

2.3.3 Figures

The following figures, Figure 1, Figure 2 and Figure 3, display various aspects of data related to Trump's support rate, helping us understand how support rates vary with different factors. In Figure 1 and Figure 2, we have marked the occurrence count of each pollster or state on the x-axis, helping us identify potential sources of bias in the data. Low-frequency items (such as data points from certain pollsters or states) may reduce the representativeness of the model—low-frequency observations imply insufficient data for specific pollsters or states, which may not accurately reflect voter support rates in these areas. This incomplete data could lead to unreliable predictions for these pollsters or states, increasing prediction errors. Additionally, low-frequency data is more susceptible to extreme values. With fewer observations, certain outlier data points (e.g., a particularly high or low support rate in a single poll) can appear disproportionately significant, affecting the overall prediction. When fitting the model to

low-frequency data, it may be more influenced by the "noise" in these data points, potentially introducing bias into the overall results. Systematic biases in low-frequency observations could spread throughout the model, impacting the overall accuracy of the predictions.

Although low-frequency observations may introduce some bias, directly removing them would result in information loss, especially in election predictions where the representativeness of different states and pollsters is crucial. Different pollsters and states have their own resources and priorities, leading to uneven data coverage. Some pollsters may prefer to publish data at specific times, while certain states, due to smaller populations or stable voter behavior, may receive less attention. Thus, it is challenging to completely balance the data quantity across different items. Although some observations have low frequency, they still contain valuable information. For instance, even though smaller or remote states may have limited data, their voter preferences are still significant under the Electoral College system. Completely deleting these observations would diminish these states' representation in the overall analysis, weakening the model's accuracy in reflecting support rates. Additionally, we aim to cover as many pollsters and states as possible with the model, so we retain even low-frequency observations. Thus, we keep these observations and mitigate the bias they introduce by combining other variables to build a more comprehensive and robust predictive model.

Figure 1 shows the average support rates for the top and bottom ten pollsters, highlighting each pollster's most commonly used methodology. We can observe that the support rates between the top 5 and bottom 5 pollsters differ by about 10%, which helps us understand the bias introduced by certain pollsters' influence on support rates, as some pollsters may tend to show higher or lower support rates. This chart also examines the different methodologies used by these pollsters, as varying survey methods (e.g., phone surveys or online panels) may affect the results by reaching different groups of voters. Focusing on these differences aids in assessing the reliability and bias of the data.

Figure 2 compares the average support rates across states with the national average, showing regional differences in voter preferences. Due to the Electoral College system in the U.S. presidential election, variations in state support rates can significantly impact the election outcome. This analysis highlights the differences between state and national polls, helping us understand Trump's relative support in each state. Paying attention to geographic disparities improves the model's prediction accuracy.

Figure 3 displays the trend in support rates over time for the top five pollsters with the most observations. We chose these pollsters because a sufficient number of observations is needed to accurately capture trend changes and reduce the impact of extreme values. Each subplot represents a pollster and shows the change in Trump's support rate since Biden's withdrawal. This helps us observe the dynamic change in support rates following major events, and we can also see how different pollsters respond to the same event. Analyzing time trends aids in understanding shifts in public opinion, providing more time-sensitive support rate predictions.

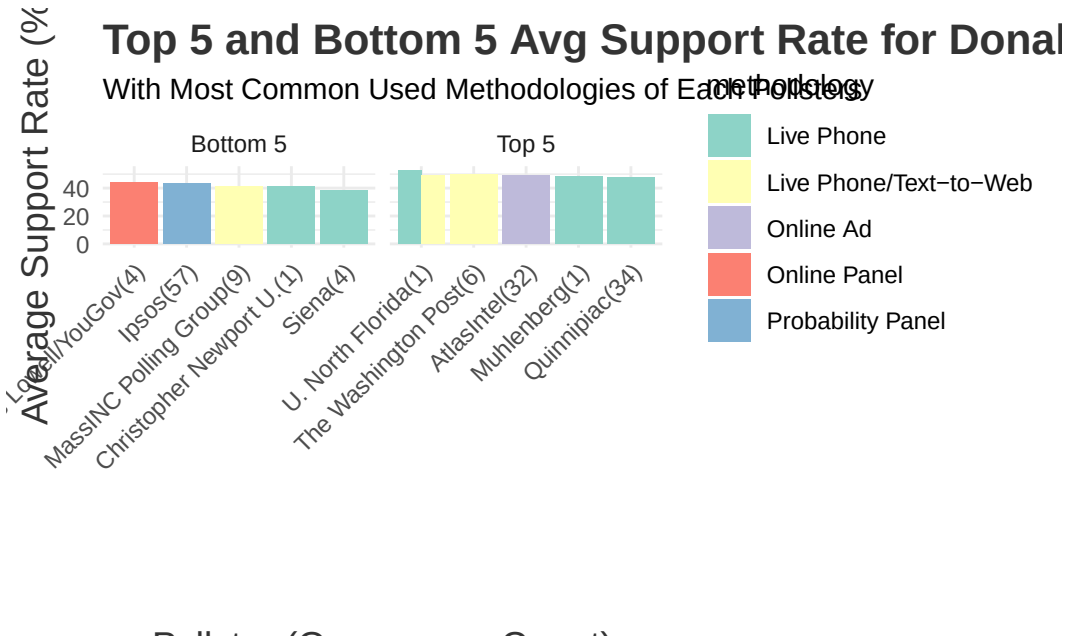

Figure 1: Support Rates for Donald Trump of Ten Pollsters with Their Most Common used Metholodogy

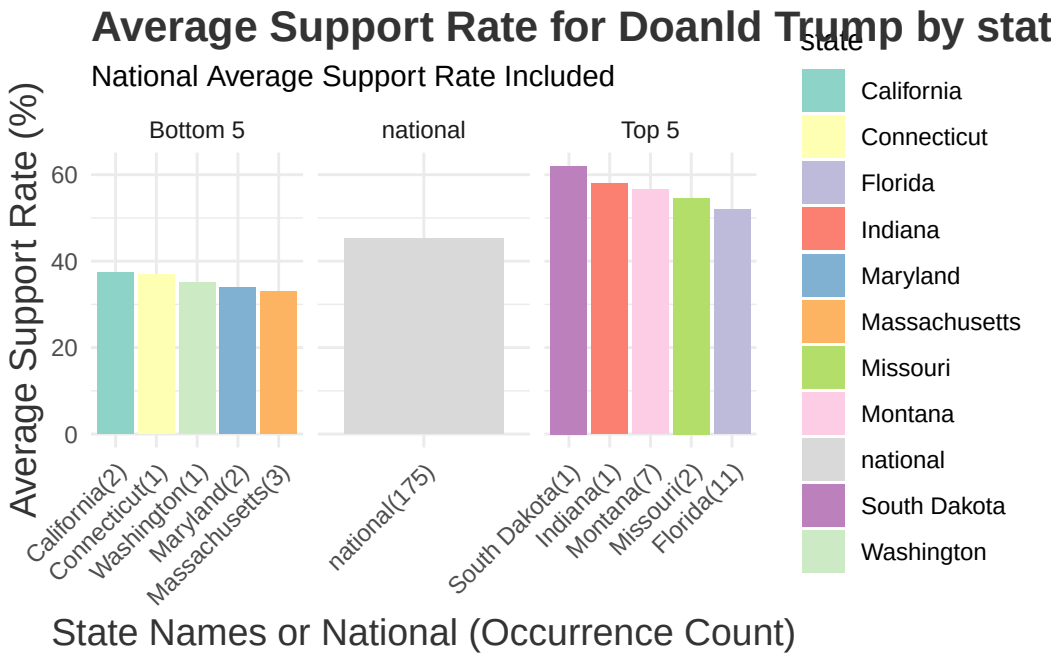

Figure 2: Average Support Rate for Trump of Each Category Based on State

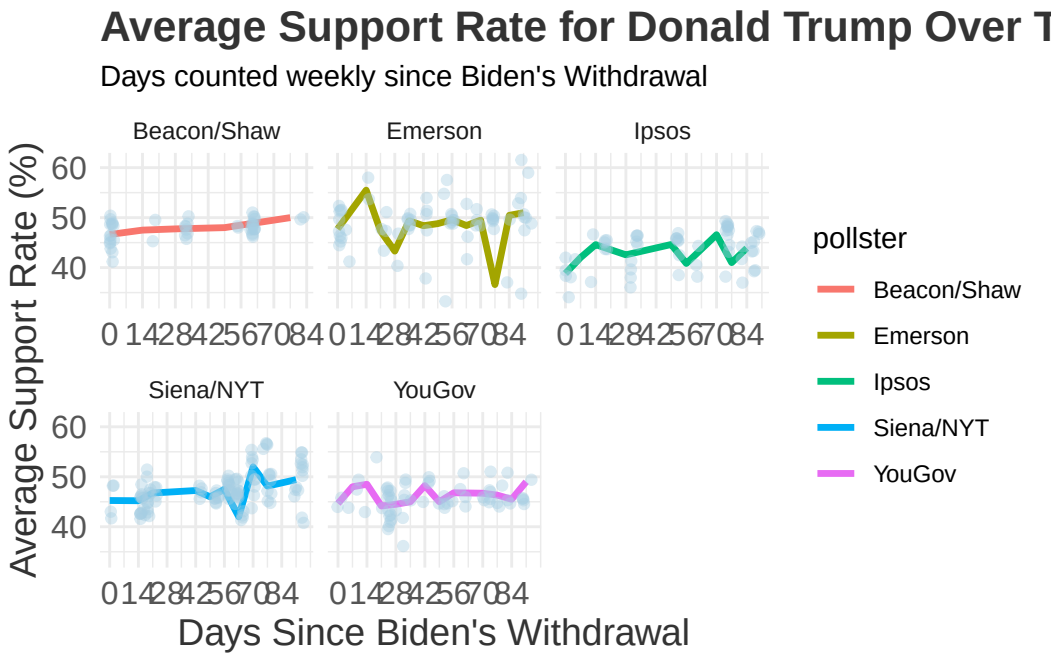

Figure 3: Trend for Average Support Rate for Donald Trump Over Time Since Joe Biden's Withdrawal on 2024 July 21

References

Alexander, Rohan. 2024. "Starter Folder." https://github.com/RohanAlexander/starter_folder.git.

FiveThirtyEight. 2024. "2024 Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

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

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://readxl.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://cran.r-project.org/web/packages/knitr/index.html.