Predicting the 2024 US Presidential Election: A Polling-Based Forecast*

Trump will get support about 47.55% by pollster and may lose the election

Tianrui Fu Yiyue Deng Jianing Li

November 4, 2024

This paper uses polling data to predict the outcome of the 2024 U.S. presidential election. By applying generalized linear regression and Bayesian models, we analyze the factors related to the polls, including polling organizations and the ratings associated with those organizations, on Donald Trump's support rate. The research results indicate a significant relationship between the variables, with Donald Trump's support rate showing an upward trend. Additionally, we further discuss the advantages and disadvantages of the two models, as well as potential methods for improvement.

Table of contents

1 Introduction **1**

2 Data **2**
2.1 Overview . . . 2
2.2 Cleaning Data . . . 2
2.3 Measurement . . . 2
2.4 Outcome variables . . . 3
2.5 Predictor variables . . . 5
2.5.1 Pollster . . . 5
2.5.2 Poll Score . . . 5
2.5.3 Transparency Score . . . 6
2.5.4 Numeric Grade . . . 7
2.5.5 End date . . . 8
2.5.6 Relationship between pollster, pollscore and end date . . . 9

*Code and data are available at: https://github.com/FrankFU323/U.S.-presidential-election.git

2.5.7 Relationship between transparency score and numeric grade 10

3 Model **11**
3.1 Model Selection 11
3.2 Model set-up 11
3.2.1 Model 1 – GLM 12
3.2.2 Model 2 – Bayesian model for Trump 12
3.3 Model justification 12

4 Results **13**
4.1 Result of model for analysis data 13
4.2 Result for model prediction 15

5 Discussion **17**
5.1 First discussion point 17
5.2 Second discussion point 18
5.3 Third discussion point 18
5.4 Weaknesses and next steps 18

A Appendix **19**
A.1 Methodology Overview and Evaluation of YouGov Polling 19
A.2 Idealized Survey Methodology 20
A.3 Clean data 23
A.4 Variable details 24
A.5 Results of model 24
A.5.1 Table for summary of model results 24
A.5.2 Bayesian model for Harris data set 26
A.5.3 Predictions for both Trump and Harris 27

B Model details **29**

References **30**

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

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

2 Data

2.1 Overview

We use the statistical programming language R Core Team (2023), the templete from Alexander (2023) completed by following packages tidyverse Wickham et al. (2019), dplyr Wickham et al. (2023), rstanarm Goodrich et al. (2022), arrow Richardson et al. (2024), modelr Wickham (2023), modelsummary Arel-Bundock (2022), ggplot2 Wickham (2016), here Müller (2020), kableExtra Zhu (2024) and knitr Xie (2023) to complete this analysis. Our data about the latest polling outcomes is from the website FiveThirtyEight (2024). It has 52 variables and 15891 observations in this data set, including pollster, poll score, etc.

Table 1: The example of chosen variable datset

pollster	pollscore	numeric_grade	transparency_score	end_date	pct
Quinnipiac	-0.5	2.8	9.0	2024-09-29	50.0
YouGov	-1.1	3.0	9.0	2024-09-24	44.0
Siena/NYT	-1.5	3.0	9.0	2024-10-26	51.0
SurveyUSA	-1.3	2.8	8.0	2024-09-18	42.0
CES / YouGov	-1.1	3.0	10.0	2024-10-25	51.0
The Washington Post	-1.2	3.0	9.0	2024-10-30	47.0
YouGov	-1.1	3.0	9.0	2024-10-17	38.0
Siena/NYT	-1.5	3.0	9.0	2024-10-23	47.0
Ipsos	-0.9	2.8	9.5	2024-10-22	47.0
AtlasIntel	-0.8	2.7	6.0	2024-11-02	50.4

2.2 Cleaning Data

We have cleaned the data set and the detailed procedure see from the appendix Section A.3.

2.3 Measurement

In the realm of political polling, measurement is critical as it transforms abstract voter sentiments into quantifiable estimates that can influence electoral outcomes. The fundamental challenge lies in converting individual opinions—such as a voter's intention to support a particular candidate—into structured data that can effectively forecast the number of electoral college votes that candidate might secure. This transformation begins with the design of the polling methodology, where specific phenomena in the real world, such as voter preferences, are captured through surveys.

This dataset originates from FiveThirtyEight (2024), a trusted source known for its rigorous standards in polling data collection, ensuring the widest possible coverage of voters and comprehensive collection and disclosure of all relevant information. This includes details like the pollster's name, identification number, survey dates, and associated trust levels. In our analysis dataset, which consists of 492 observations related to Donald Trump's polling outcomes, we have identified 38 relevant variables that play a significant role in understanding electoral support. Key among these are the numeric grade, transparency score, poll score, pollster, end date, and percentage of support for Trump. Each variable is meticulously measured to ensure that they accurately reflect the sentiments of likely voters.

The dataset presents results as the proportion of votes each candidate receives according to each pollster, with this approach adjusting over time to account for changes in voter sentiment, which can shift rapidly due to the approach of Election Day, candidate speeches, and other factors. By applying these measurement principles, we systematically convert nuanced public opinion into a structured dataset that not only captures voter intentions but also provides a reliable foundation for predictive modeling.

Ultimately, this rigorous measurement approach allows us to build a Bayesian model to assess how different pollsters might influence the final outcome of the U.S. presidential election, thus bridging the gap between individual voter opinions and electoral predictions.

The explain of each used variable please see in the Section A.4.

2.4 Outcome variables

The Figure 1 shows the distribution of polling data over time, with each color representing a different polling organization. We can see that as time progresses from August to October, there is a steady increase in the amount of polling data, peaking around early October. This suggests that multiple organizations have contributed to polling data on a consistent basis over this period.

In the Figure 2, it reveals a central peak around the 45-50% range, indicating that most of the values for this variable are concentrated in this area. The shape of the plot suggests a right-skewed distribution, with relatively few data points falling below 40% or above 50%. This gives an overview of the central tendency and variability of the polling percentages.

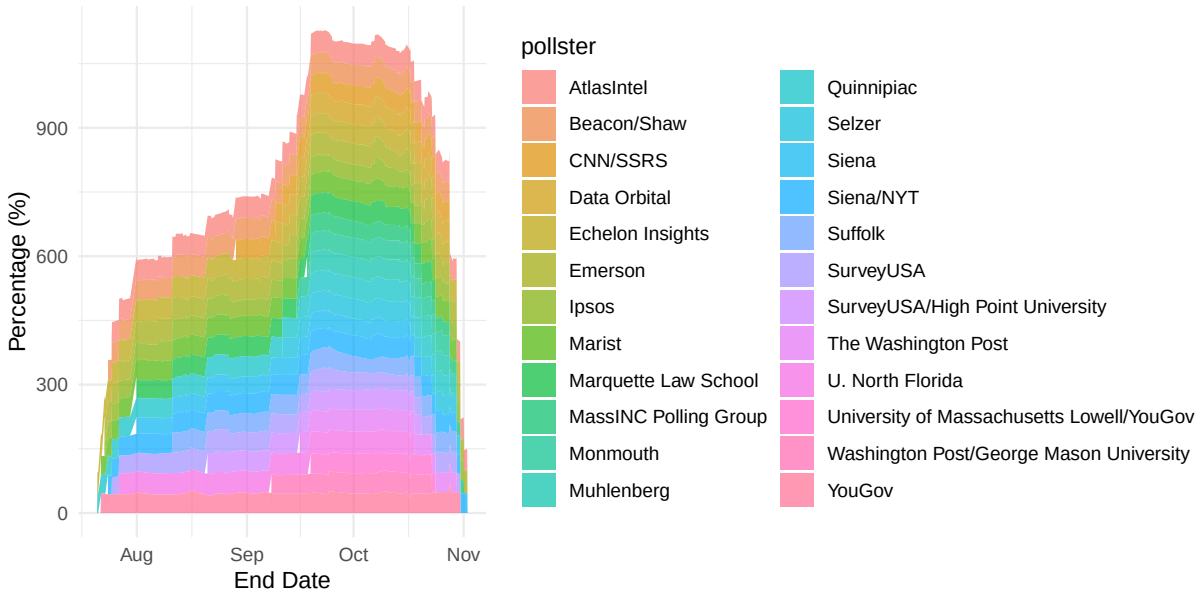

Figure 1: Percentage of Each Pollster Over Time

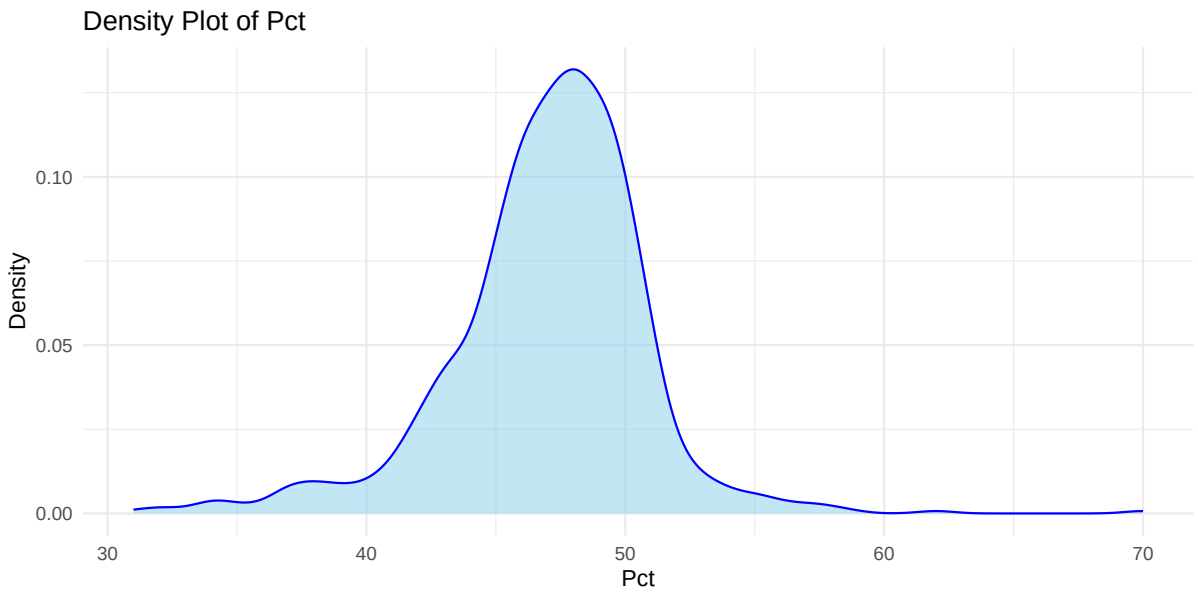

Figure 2: Density plot of pct

2.5 Predictor variables

2.5.1 Pollster

In Figure 3, the predictor variable is the count of polls conducted by each pollster. The bar chart displays the number of polls released by different polling organizations, ordered from highest to lowest frequency. The bars range in color from light blue to dark blue, indicating the frequency of polls conducted by each organization. Siena/NYT and YouGov are the most active pollsters, with 136 and 121 polls conducted, respectively. Emerson and AtlasIntel also have high polling frequencies, with 73 and 82 polls. In contrast, several pollsters, like Christopher Newport University and McCourtney Institute/YouGov, conducted only one poll, indicating their minimal activity in comparison.

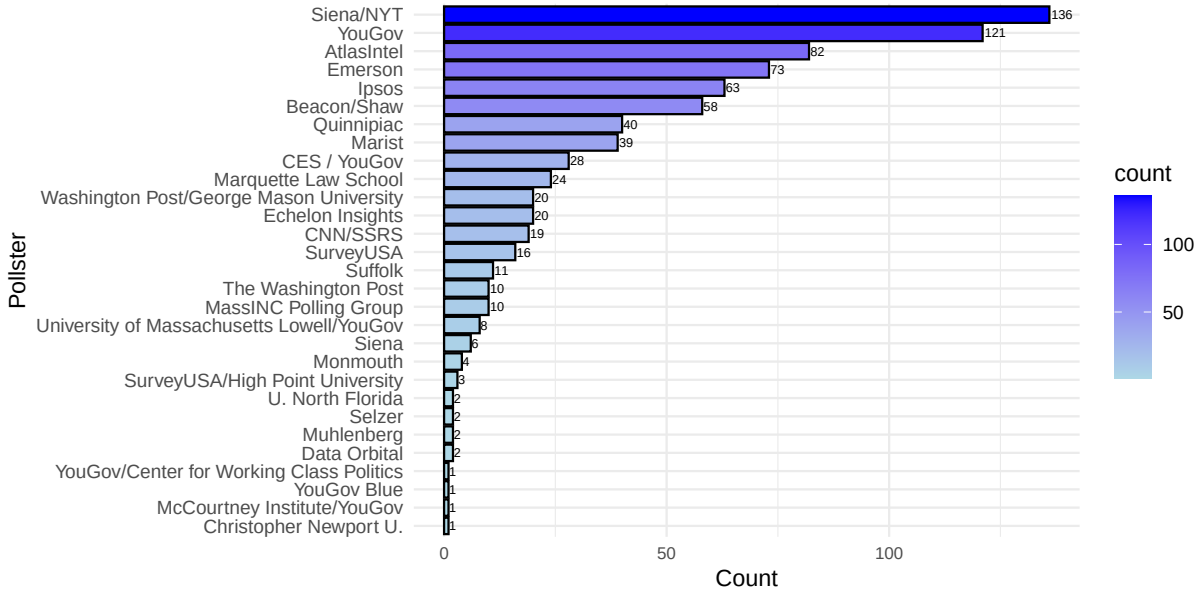

Figure 3: Frequency of each pollster

2.5.2 Poll Score

Figure 4 is a vertical bar chart displaying the distribution of poll scores, ranging from -0.5 to -1.5. The most frequent poll score is -1.1, indicating a high concentration around this value. Most pollsters' poll scores are below -0.8, which suggests that overall scores are relatively low, reflecting a higher level of trust in the polls. Further analysis shows that the lower the poll score, the higher the trust level. Therefore, the distribution is notably skewed, with most data concentrated in the lower range (around -1.1 and below). Higher negative scores (closer to -0.5) are relatively rare, indicating that most polling agencies demonstrated high credibility in this survey, earning broad approval from the public or voters.

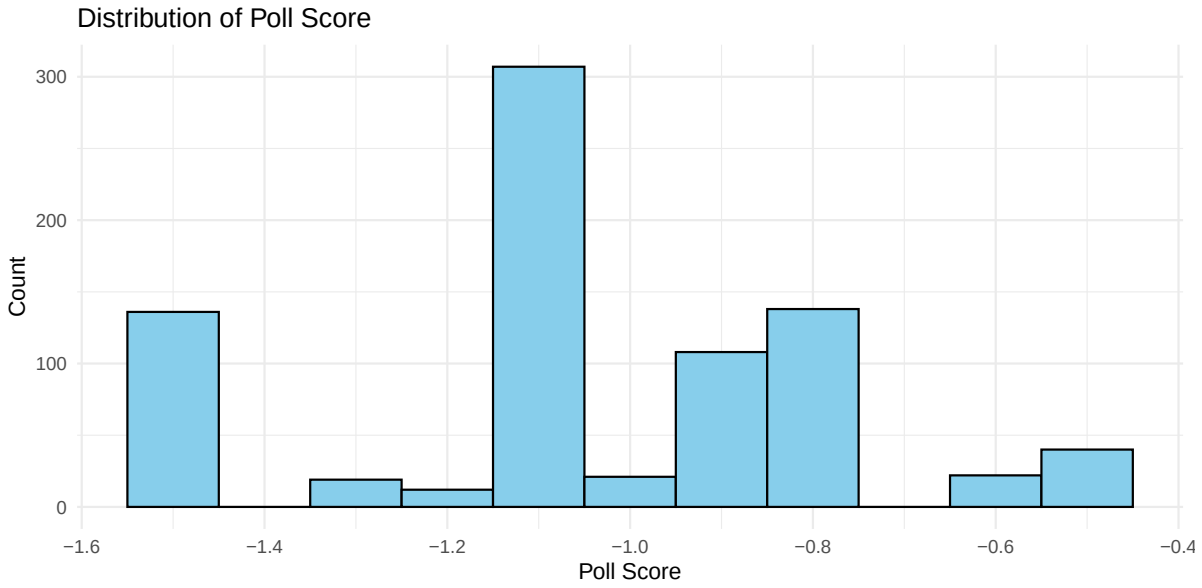

Figure 4: Distribution of Poll Score

2.5.3 Transparency Score

Figure 5 illustrates the distribution of "Transparency Scores" related to the 2024 election, possibly representing public or media ratings of candidate transparency, ranging from 5 to 10. The majority of scores are clustered around 8, showing that most respondents rate transparency fairly high. Scores of 6 and 7 are less frequent, while extreme scores of 5 and 10 are rare. This indicates that most voters or media sources perceive candidates as moderately to highly transparent, with scores skewed toward the higher end of the scale.

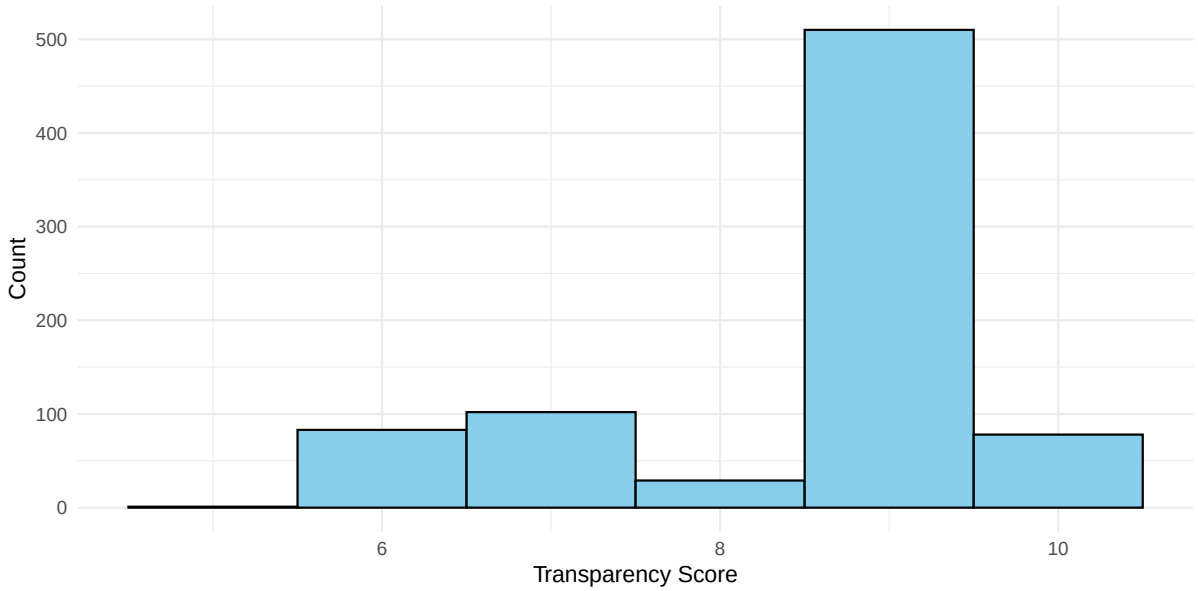

Figure 5: Distribution of Transparency Score

2.5.4 Numeric Grade

Figure 6 displays the distribution of a "Numeric Grade" associated with the 2024 election, which could represent candidate approval ratings, public opinion scores, or voter satisfaction. The grades are primarily concentrated between 2.7 and 3.0, with a noticeable peak close to 3.0, indicating a higher frequency of favorable ratings. This distribution suggests strong public support or positive sentiment, with the majority of respondents giving relatively high scores on this variable.

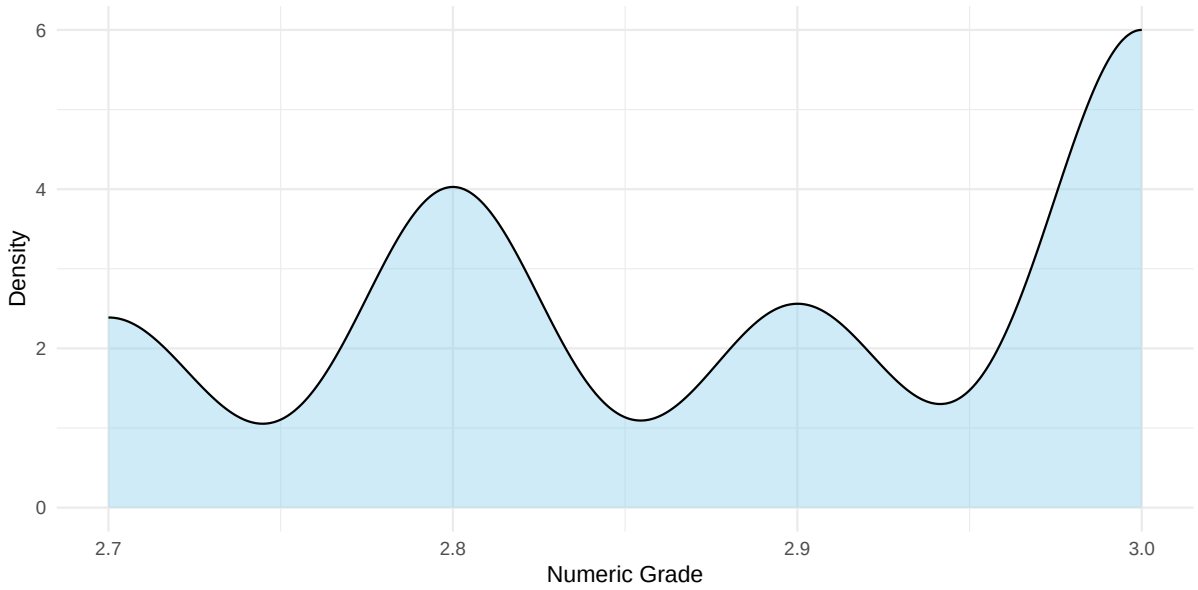

Figure 6: Density Plot of Numeric Grade

2.5.5 End date

Figure 7 shows the distribution of a particular variable over time related to the 2024 U.S. election. The x-axis represents dates from August to November, and the y-axis shows the count. This variable might represent the frequency of certain events, such as news coverage, voter registrations, or polling responses. The data is relatively sparse from August to October, with small peaks in some areas. In November, however, there is a significant increase in counts, particularly in early November, possibly linked to the heightened election activity or voter engagement as Election Day approaches.

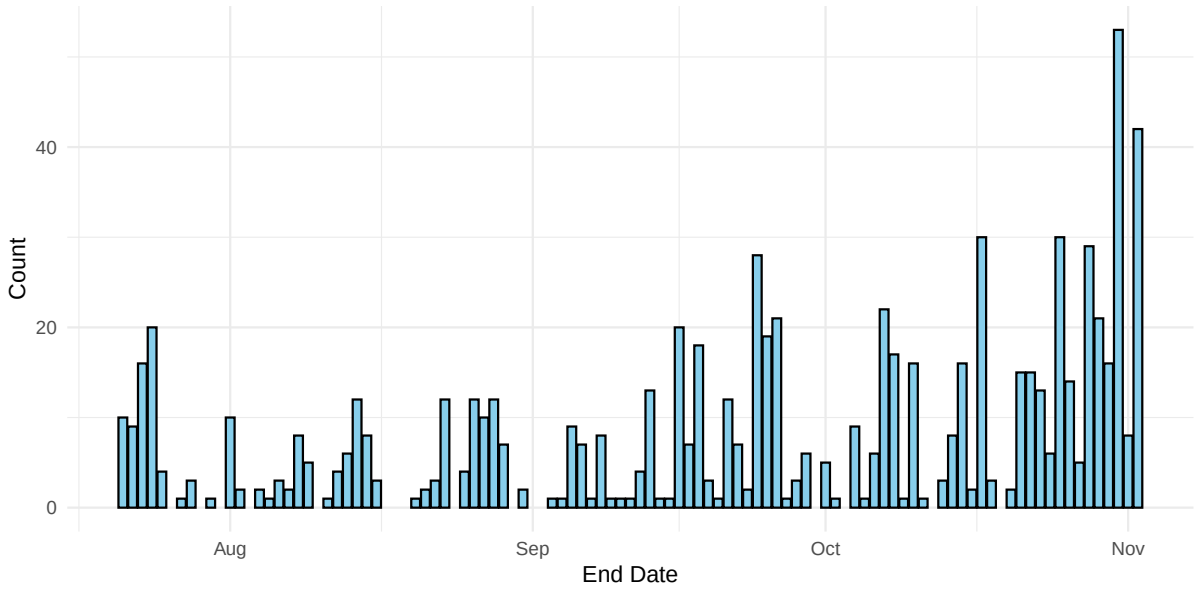

Figure 7: Frequency of Records by End Date

2.5.6 Relationship between pollster, pollscore and end date

In Figure 8, the predictor variable is the average poll score for each pollster over time, from August to October. This heat map visualizes the poll scores of each organization on different dates, with color intensity indicating the score levels—darker colors represent lower scores. The pollsters with the most consistently low average poll scores (darker colors) include Selzer, Siena/NYT, and Marquette Law School. These pollsters frequently show results lower than others over time, suggesting they may consistently lean toward one direction in their results. In contrast, pollsters like YouGov Center for Working Class Politics and YouGov Blue display lighter colors, indicating relatively higher scores across their polls.

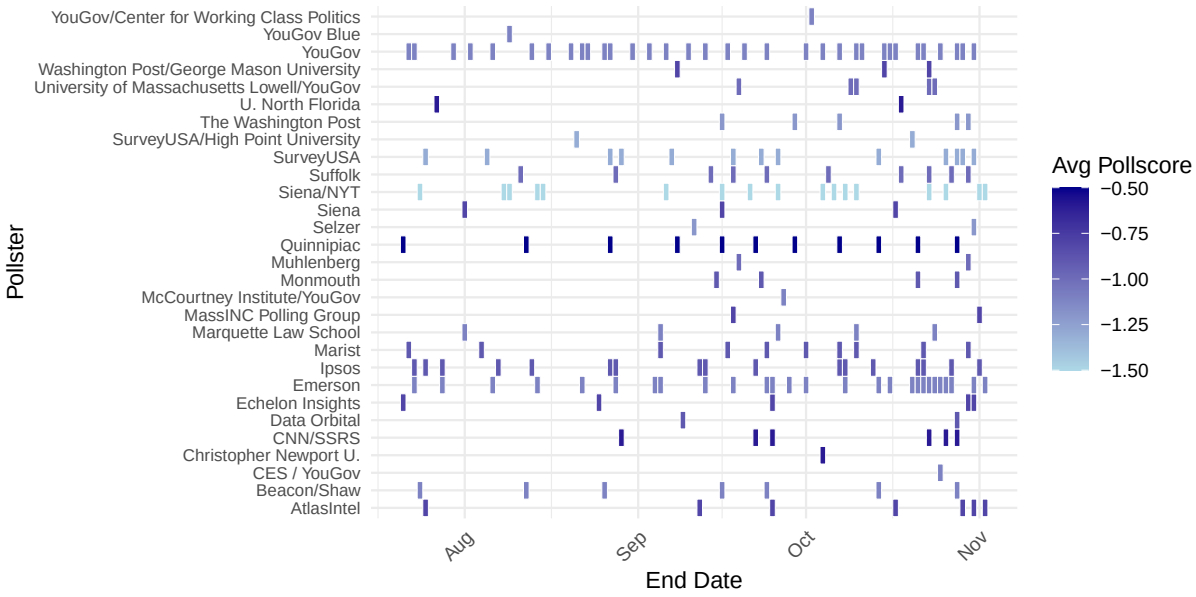

Figure 8: Average Pollscore by Pollster and End Date

2.5.7 Relationship between transparency score and numeric grade

In Figure 9, there is a clear positive relationship between transparency score and numeric grade. Pollsters with higher transparency scores, such as those scoring around 10, tend to achieve higher numeric grades, close to 3.0. Conversely, pollsters with lower transparency scores tend to have lower numeric grades, closer to 2.7. This pattern suggests that greater transparency is associated with better overall pollster ratings.

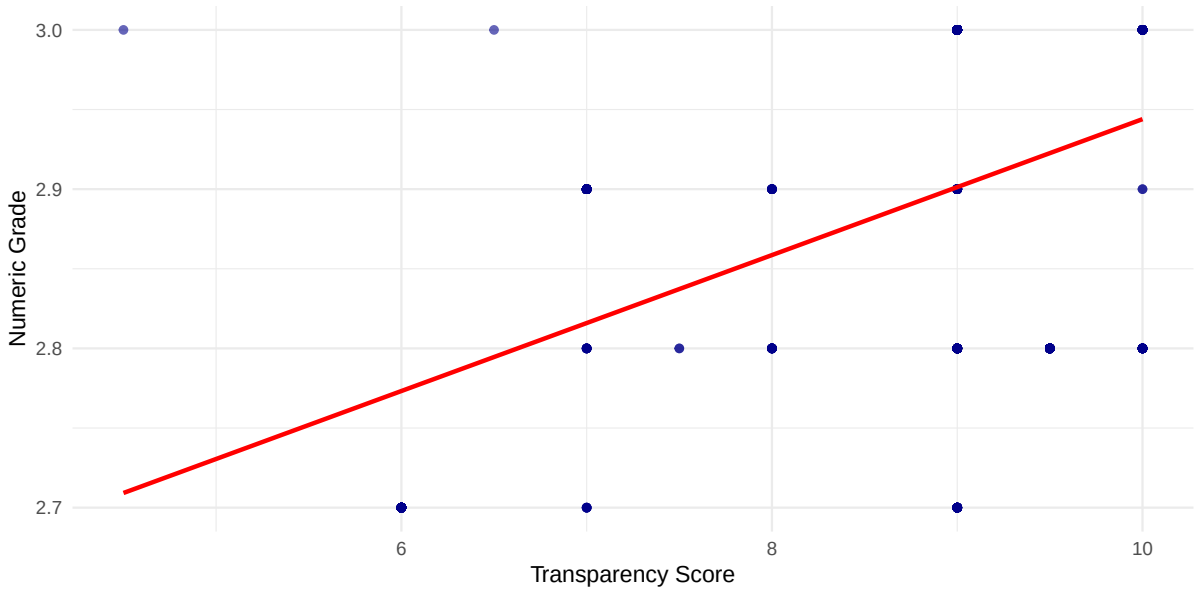

Figure 9: Scatter Plot of transparency score and numeric grade

3 Model

3.1 Model Selection

Our model aims to predict the changes in Donald Trump's probability of winning the 2024 U.S. election over time while examining the influence of key factors, such as pollster reliability scores, transparency, and specific pollster effects, on polling results. For model selection, we initially considered a generalized linear regression to provide a straightforward analysis of the relationships between these factors. However, given the limitations of generalized linear regression in handling uncertainty and dynamic data, we ultimately chose a Bayesian model. The Bayesian approach enables the integration of prior information and dynamically updates predictions, offering greater stability and accuracy under high uncertainty. Below is a brief overview of our model.

Background details and diagnostics are included in Appendix B.

3.2 Model set-up

In this Bayesian framework, we assume a normal distribution of poll results around a mean affected by key predictors: numeric grade, transparency score, and pollscore. Define y_i as the percentage of Donald Trump. Then β_i represents the numeric grade, γ_i represents the transparency score, and δ_i represents the pollscore.

3.2.1 Model 1 – GLM

$$y_i = \alpha + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{transparency\_score}_i + \beta_3 \cdot \text{pollscore}_i \tag{1}$$

$$+ \sum_{j=1}^{N} \gamma_j \cdot \text{pollster}_j + \delta \cdot \text{end\_date}_i + \epsilon_i \tag{2}$$

where: - α is the intercept, representing the average poll level. - β_1, β_2, and β_3 are the regression coefficients for numeric grade, transparency score, and pollscore, respectively. - γ_j indicates the fixed effect of each pollster. - δ is the regression coefficient for the end date. - ϵ_i is the error term.

3.2.2 Model 2 – Bayesian model for Trump

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{3}$$

$$\mu_i = \alpha + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{transparency\_score}_i + \beta_3 \cdot \text{pollscore}_i \tag{4}$$

$$+ \sum_{j=1}^{N} \gamma_j \cdot \text{pollster}_j + \delta_i \cdot \text{end\_date}_i \tag{5}$$

$$\alpha \sim \text{Normal}(50, 10) \tag{6}$$

$$\beta_1, \beta_2, \beta_3 \sim \text{Normal}(0, 5) \tag{7}$$

$$\gamma_j \sim \text{Normal}(0, 5) \tag{8}$$

$$\sigma \sim \text{Exponential}(1) \tag{9}$$

where: - α represents an average poll result. - β_1, β_2, and β_3 capture the unique effect of their respective predictors on the poll percentage. - γ_j denotes a specific pollster effect. - δ_i accounts for time trends.

We run the model in R Core Team (2023) using the package of rstanarm Goodrich et al. (2022).

3.3 Model justification

The choice of a Bayesian model to analyze Donald Trump's voting support rate stems from the need to effectively integrate uncertainty and leverage prior knowledge in estimating outcomes. Given that the dataset includes key predictors related to polling organizations—such as numeric_grade, transparency_score, pollster, poll_score, end_date, and pct—this approach allows for a comprehensive understanding of how these variables influence voting results.

The numeric_grade and transparency_score are crucial in assessing the reliability of polling organizations. By integrating these factors, the Bayesian model can provide a nuanced analysis of how the quality and transparency of polling organizations impact Trump's support rate. This is particularly important in the context of public opinion polling, as perceptions of reliability can significantly affect the interpretation of results.

Moreover, the Bayesian framework enables the incorporation of prior distributions for model parameters, reflecting beliefs about their possible values before observing the data. This is especially beneficial in a domain where historical data and expert opinion can provide valuable insights.

Additionally, by including the pollster variable, the model accounts for the fixed effects of different polling organizations, allowing for a more targeted analysis that acknowledges the unique characteristics of each organization. The end_date variable helps to consider time trends, ensuring that the analysis remains relevant to the evolving political landscape, especially as elections approach.

However, there are areas for improvement within the model. First, refining the selection of variables related to voting support could enhance the model's predictive power. For instance, incorporating socio-economic factors, demographic data, or shifts in voter behavior may provide a more comprehensive perspective. Additionally, the model's predictive performance could be assessed through cross-validation or other model evaluation techniques to ensure the robustness and reliability of the results.

By employing this Bayesian model, our aim is to provide robust estimates of Trump's support rate, considering both the statistical characteristics of the data and the inherent uncertainty related to polling, while exploring avenues for further model improvement.

4 Results

The results part is combined with the result for models of analysis data and the result of predictions by Bayesian model on November 5, 2024. In order to get to know whether the votes of Trump will win the election, we combined the data of Harris who have the most competitive candidate other than Trump to do predictions. It would help us be faster and more accurate to do the predict judgment.

4.1 Result of model for analysis data

In Figure 10, we applied a generalized linear regression model to predict the percentage of polls Trump according each pollster. Specifically, the support rate data points range from 40% to 55%, with most clustered between 45% and 50%. The orange trend line shows a subtle upward trend, indicating a gradual increase in Trump's average support rate over time. For example, in early August, the support rate is around 45%, while by mid-October, the average

support rate shown by the trend line is close to 50%. This suggests that Trump's support rate has risen by about 5% over these three months. For the model summary of Generalized Linear Regression and Bayesian Model, the table can see at Section A.5.

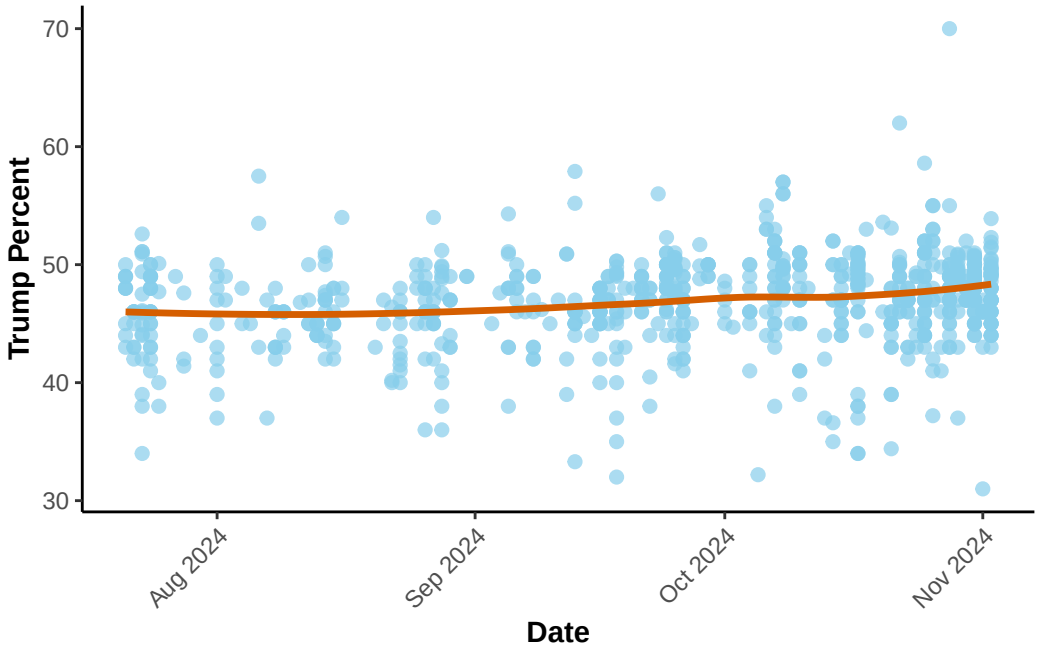

Figure 10: Donald Trump Support Over Time by Generalized Linear Regression

In Figure 11, we applied a bayesian model to predict the percentage of polls Trump according each pollster. Each data point is color-coded by pollster, with support rates ranging from 35% to 60%. The shaded gray area around the trend line represents the confidence interval, indicating the uncertainty in support rate variations. The overall trend line (blue) shows a slight upward trend, increasing from about 45% in August to nearly 50% in October. In particular, data points in early August are more dispersed, with some pollsters reporting low support rates (around 40%) and others reporting higher rates (over 50%). By November, most poll results have converged, with support rates centered between 45% and 50%. Notably, certain pollsters, like YouGov and Marist, provide either consistently higher or lower support rate estimates compared to others, indicating some bias or variation among polling organizations.

We also applied a Bayesian model to predict the percentage of polls Harris according each pollster. The figure can see it in Section A.5.

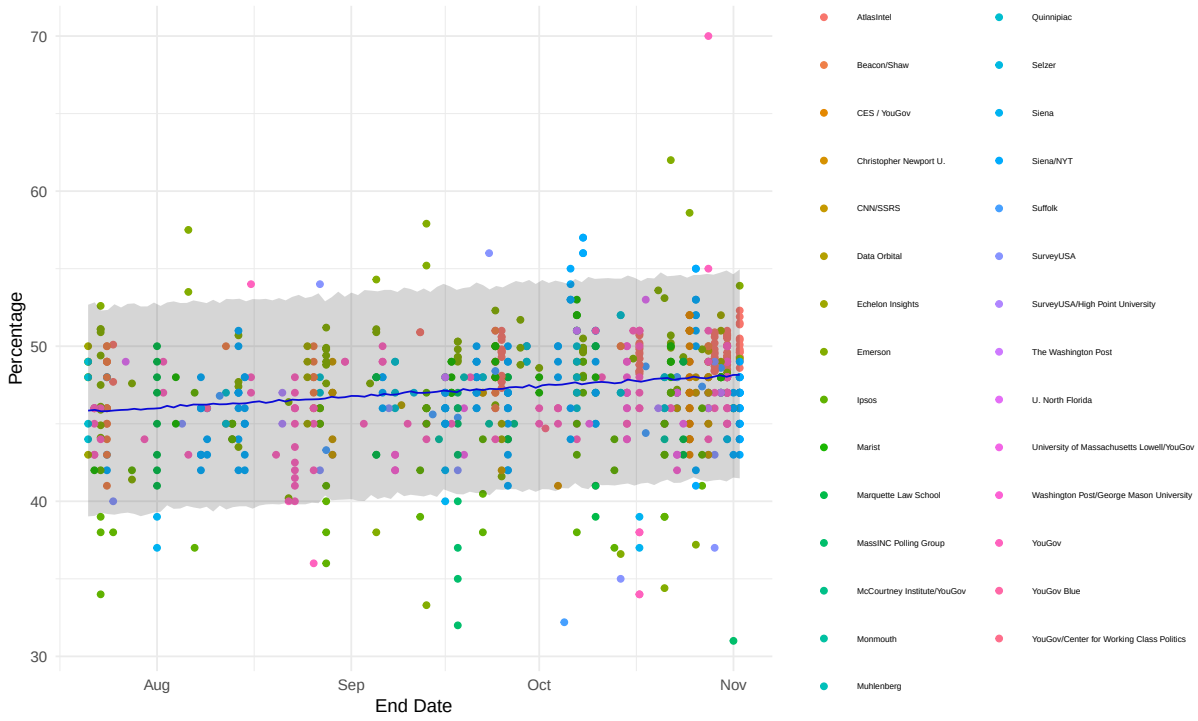

Figure 11: Poll Percentage over Time with Bayesian Fit for Trump

4.2 Result for model prediction

Figure 12 compares the predicted mean support rates for Trump and Harris across different polling organizations in 5th November of 2024. The vertical axis shows the predicted support rate (in percentage), and the horizontal axis lists the various pollsters. The red line represents Harris's predicted support rate, while the blue line represents Trump's.

For example, in predictions from Christopher Newport U. and MassINC Polling Group, Harris has a noticeably higher average support rate than Trump, exceeding 50%, while Trump's support rate is below 50%. In Siena's predictions, Trump's support rate is slightly lower than Harris's, reaching around 41%, while Harris's support is close to 53%. In Christopher Newport U.'s, MassINC Polling Group's, Siena's and Washington Post/George Mason University's predictions, Trump's support rate is all lower than Harris's and have a big gap.

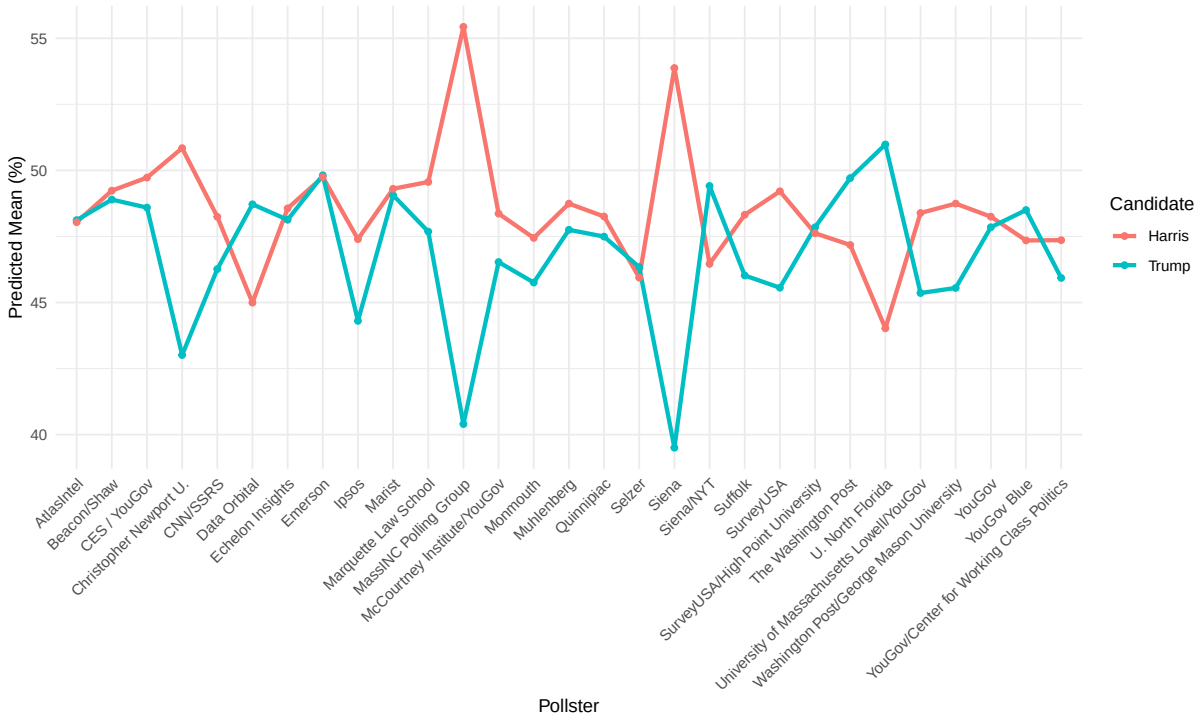

Figure 12: Predicted Mean Percentage by Pollster for Trump and Harris

In Table 2, we merged and compared the Bayesian model predictions of the vote shares for Trump and Harris on November 5, 2024, by pollster. The green background color represents the same pollster, indicating that the side with the larger predicted vote share is expected to win based on the data collected and summarized by that pollster. Conversely, red indicates the losing side. In Table 2, a total of 27 pollsters participated in the polling, with 9 pollsters predicting a Trump victory and 18 pollsters predicting a Harris victory. In Table 3, we calculated the mean support for Trump and Harris, estimating Trump's probability of leading probability to be 31% and the mean support be 46.87%. For the further more predictions about the Bayesian model, we can see at Section A.5.

Table 2: Predictions for both Trump and Harris by pollster

	Predictions Summary		
	pollster	**pred_mean.Trump**	**pred_mean.Harris**
2.5%	**Emerson**	49.81549	49.74995
2.5%1	**AtlasIntel**	48.11692	48.04041
2.5%2	**Siena/NYT**	49.41257	46.46172
2.5%3	**MassINC Polling Group**	40.40029	55.43243
2.5%4	**Ipsos**	44.30453	47.39779
2.5%5	**YouGov**	47.84888	48.24803
2.5%6	**SurveyUSA**	45.56023	49.20986

2.5%7	**Echelon Insights**	48.13166	48.56242
2.5%8	**Selzer**	46.33986	45.94006
2.5%9	**Suffolk**	46.02405	48.32055
2.5%10	**Marist**	49.06094	49.30425
2.5%11	**Muhlenberg**	47.75231	48.74332
2.5%12	**The Washington Post**	49.70756	47.17822
2.5%13	**Data Orbital**	48.71777	44.99492
2.5%14	**Monmouth**	45.75619	47.44451
2.5%15	**Quinnipiac**	47.49468	48.25494
2.5%16	**Beacon/Shaw**	48.89540	49.23308
2.5%17	**CNN/SSRS**	46.26501	48.24482
2.5%18	**CES / YouGov**	48.59251	49.72693
2.5%19	**Marquette Law School**	47.68378	49.56380
2.5%20	**University of Massachusetts Lowell/YouGov**	45.35861	48.38879
2.5%21	**Washington Post/George Mason University**	45.55097	48.74419
2.5%22	**SurveyUSA/High Point University**	47.84759	47.62245
2.5%23	**U. North Florida**	50.98170	44.02510
2.5%24	**Siena**	39.50460	53.87453
2.5%25	**Christopher Newport U.**	43.01220	50.84202
2.5%26	**YouGov/Center for Working Class Politics**	45.92983	47.35936
2.5%27	**McCourtney Institute/YouGov**	46.53130	48.36614
2.5%28	**YouGov Blue**	48.49834	47.34990

Table 3: Summary of Predictions by mean and lead probability

Metric	Value
Trump Mean Support	46.87
Harris Mean Support	48.50
Trump Lead Probability	0.31

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

A.1 Methodology Overview and Evaluation of YouGov Polling

Introduction

YouGov is an online polling company that uses a combination of non-probability sampling and a model known as multilevel regression with post-stratification (MRP) to estimate voter intentions. This appendix reviews YouGov's approach, including sample recruitment, data collection, response handling, questionnaire design, and MRP modeling. It also assesses the strengths and limitations of YouGov's methods.

1. Population, Frame, and Sample Composition

YouGov's target population for its election polling includes American adults, with a focus on registered voters. It uses an online panel as its frame, drawing participants from a pool of volunteers who sign up and provide demographic information. This setup allows YouGov to adjust samples to align with the population it seeks to represent. Additionally, YouGov uses the TargetSmart voter file to verify that the sample aligns with national voter demographics. During election periods, YouGov increases its sample size, beginning with nearly 100,000 responses and adding another 20,000 in September and October to update its model.

2. Sample Recruitment and Representativeness

YouGov recruits panel members online through ads and partnerships with other websites. This approach allows any American adult with internet access to join. Members provide basic demographic information, which YouGov uses to select participants and apply statistical weighting to reflect the population accurately. Panelists earn points for participating, redeemable for small rewards, which encourages engagement and enhances data quality. However, since recruitment is internet-based, some groups without reliable internet access, such as rural or low-income populations, may be underrepresented.

3. Sampling Approach and Methodological Trade-offs

YouGov's approach involves non-probability sampling, meaning that not everyone in the population has an equal chance of being selected. To address this, YouGov applies the MRP model, dividing respondents into subgroups based on characteristics like age, gender, race, education, region, and political affiliation. Each subgroup is weighted to match its share in the population. This method is efficient but has trade-offs, as MRP may not fully capture smaller states or hard-to-reach groups.

4. Non-response Handling and Quality Control Measures

YouGov uses strict quality control measures to manage non-response. Surveys include checks for speed and consistency, removing low-quality responses. Panelists who repeatedly fail these checks are excluded, helping to reduce bias and improve data quality. These quality controls are essential for the MRP model, which relies on accurate data to produce reliable estimates.

5. Questionnaire Design

YouGov's questionnaires are neutral and straightforward, with randomized question order to reduce bias. They include various question types, such as text, images, and audio, helping respondents understand the context. Options like "prefer not to say" are included for sensitive questions, supporting respondent privacy. However, since the surveys are conducted online, some groups without internet access may be excluded, and shorter questionnaires may limit the depth of data collected.

6. MRP Model for Vote Estimation

YouGov's MRP model estimates voter intentions through three stages: estimating likelihood to vote, predicting support for a candidate among likely voters, and aggregating results to calculate overall support. By matching responses to the TargetSmart voter file, the model generates estimates at both national and regional levels. The MRP model also helps YouGov track changes in voter intentions over time.

7. Strengths and Limitations of the YouGov Approach

YouGov's approach has several strengths. The MRP model allows for accurate predictions at regional and national levels, and the use of voter files enhances sample representativeness. Repeated interviews with panelists allow YouGov to observe shifts in voter intentions over time. However, limitations include potential representativeness issues due to non-probability sampling, particularly in small states and among underrepresented groups. The internet-based survey method also may not fully capture populations without reliable internet access.

Conclusion

Through a combination of MRP modeling, quality control, and a large online panel, YouGov provides a structured approach for election polling. Although challenges remain in achieving full representativeness and internet coverage, YouGov's approach offers useful information on U.S. voter intentions. The MRP model has proven effective in past elections and serves as a practical tool for future polling and tracking trends.

A.2 Idealized Survey Methodology

The budget for this survey is $100,000, aimed at predicting the outcome of the 2024 United States Presidential Election. This methodology includes stratified sampling, multi-platform recruitment, data validation, and multi-wave data aggregation to ensure representative and reliable data.

Sampling Approach: Stratified Random Sampling

To obtain a representative sample, we use stratified random sampling with a sample size of 5,000 respondents. Sampling is stratified by age, gender, education, and geographic region to ensure broad coverage across voter demographics. The age groups include 18-29, 30-44, 45-64,

and 65 and above. Gender is categorized as male, female, and other, while education is classified as high school or below, bachelor's degree, and master's degree or above. The geographic region includes all U.S. states. Stratified sampling ensures that the sample accurately reflects voter characteristics, providing a solid foundation for subsequent analysis.

Recruitment: Multi-Platform and Interactive Engagement

Recruitment is conducted through a multi-platform strategy to ensure wide coverage among voters. Targeted ads are deployed on Google, Facebook, and Twitter to attract respondents with specific demographic characteristics. The ad content is concise and highlights anonymity and the research purpose to encourage participation. In addition, we use Random Digit Dialing (RDD) to contact older adults and residents in remote areas who may be less accessible online, ensuring their participation in the survey. To increase the completion rate, a small reward system is implemented, with one out of every 100 participants receiving a $5 to $10 incentive.

Survey Platform: Google Forms and Phone Outreach

Google Forms is used as the primary data collection platform, enabling respondents to complete the survey on a computer or mobile device with ease. We have set Google Forms to restrict submissions to one per Google account to prevent duplicate responses. For those who may not have easy online access, telephone outreach will be conducted, especially targeting older adults and rural voters, to ensure the diversity and inclusivity of the sample.

Data Validation: Post-Stratification Weighting

To ensure data quality, post-stratification weighting will be applied during the analysis phase based on demographic characteristics. This adjustment aligns the sample structure with the national voter distribution, reducing bias and increasing the accuracy of our findings. Multi-layered data validation measures enhance the reliability and representativeness of the results.

Poll Aggregation: Multi-Wave Polling and Adjustments

To track changes in voter sentiment over time, we will conduct multiple waves of data collection and aggregation. The survey will be administered in multiple rounds throughout the election cycle, with each round spaced 3-4 weeks apart. Each wave of data will be collected and analyzed independently, then aggregated using weighted averages to smooth out single-instance fluctuations and help identify long-term trends, providing a reliable basis for predictions.

Budget Allocation: Phased Spending and Testing

To maximize budget efficiency, we will use a phased spending strategy. An initial 30% of the advertising budget will be allocated to testing across platforms to determine effectiveness, with the remaining 70% directed to the most successful channels. The budget breakdown is as follows: social media advertising ($40,000) for broad and targeted outreach, phone outreach ($20,000) for RDD calls to less accessible populations, small rewards ($15,000) to encourage completion, and data cleaning and analysis ($25,000) for validation, weighting, and aggregation across multiple waves.

Survey Content and Link

Survey Title: 2024 United States Presidential Election Survey

Survey Link: https://forms.gle/EzyHp3zuX8Cu6Ep8A

Survey Questions:

1. What is your age?

o 18-29

o 30-44

o 45-64

o 65 and above

2. What is your gender?

o Male

o Female

o Other

o Prefer not to say

3. What is your highest level of education?

o High school or below

o Bachelor's degree

o Master's degree or above

4. Which state do you currently reside in?

(Dropdown menu with state names)

5. Do you plan to vote in the 2024 election?

o Yes

o No

o Not sure

6. If the election were held today, which candidate would you be more likely to support?

o Kamala Harris (Democratic Party)

o Donald Trump (Republican Party)

o Other

o Undecided

7. Which of the following issues would have the greatest impact on your decision in the 2024 election? (Select one)

o Economic stability and growth

o Access to quality healthcare

o Education reform and funding

o Environmental sustainability and climate action

o Addressing social and racial inequalities

o Other (please specify): _____

8. If a future candidate proposed a policy that aligns perfectly with your concerns, would you consider changing your voting intention?

o Yes

o No

o Not sure

9. If you would like to participate in the reward draw, please provide your email address below. Your contact information will only be used to notify winners. We will reach out to winners via email.

Email: _____

A.3 Clean data

It starts by reading the dataset and standardizing column names. Next, it removes irrelevant columns like sponsor_ids, sponsors, and etc and filters for high-quality polls (with a rating of 2.7 or higher) that specifically track Trump's support. National polls missing state information are labeled as "National," and dates are formatted and filtered to include only those on or after July 21, 2024 (when Trump declared his candidacy). Finally, the percentage supporting Trump is converted to an actual count for further modeling.

A.4 Variable details

We have used the variable numeric_grade, transparency_score, pollster, poll score, end_date and pct in the building of Bayesian model. Among these variables in the data set after analysis,

- The numeric_grade is the numeric rating given to the pollster to indicate their quality or reliability from 2.7 to 3.0.
- The transparency_score is the grade for how transparent a pollster is, calculated based on how much information it discloses about its polls and weighted by recency from 4.5 to 10.0.
- The pollster is the name of the polling organization that conducted the poll included Marquette Law School, CNN/SSRS and etc.
- The poll score is the numeric value representing the score or reliability of the pollster in question from -1.5 to -0.5 and negative numbers of poll score are better.
- The end_date is the date of polling ends.
- The pct is the percentage of the vote or support that the candidate received in the poll and keep integer.

A.5 Results of model

A.5.1 Table for summary of model results

In Table 4, we conducted a summary of three models, including a generalized linear regression and a Bayesian model for Trump's data, as well as a Bayesian model for Harris's data. This summary includes the γ_j values for each pollster and monitoring data for the models, such as AIC, BIC, etc.

Table 4: Explanatory models of flight time based on wing width and wing length

[!h]

	Model Summary		
	Model by glm	Model by bayes with Trump	Model by bayes with Harris
(Intercept)	−352.728	−390.051	−225.798
	(86.299)		
numeric_grade	−9.323	−3.038	0.533
	(12.348)		
transparency_score	0.679	−0.218	−0.025
	(0.588)		
pollscore	2.490	1.958	0.444
	(12.352)		
Beacon/Shaw	−1.368	0.780	1.149
	(4.578)		
pollsterCES / YouGov	−1.243	0.495	1.576
	(4.632)		
Christopher Newport U.	−9.906	−5.230	3.021
	(4.050)		
CNN/SSRS	−5.206	−1.836	0.270
	(2.432)		
Data Orbital	−1.945	0.433	−3.135
	(3.668)		
Echelon Insights	−2.996	−0.052	0.486
	(1.968)		
Emerson	1.849	1.562	1.544
	(3.405)		
Ipsos	−6.559	−3.830	−0.712
	(2.651)		
Marist	−0.185	0.924	1.138
	(2.336)		
Marquette Law School	−2.316	−0.493	1.448
	(4.657)		
MassINC Polling Group	−8.604	−7.745	7.321
	(1.489)		
McCourtney Institute/YouGov	−0.951	−1.681	0.266
	(4.771)		
pollsterMonmouth	−4.285	−2.453	−0.635
	(2.951)		
Muhlenberg	−1.848	−0.451	0.641
	(3.706)		
Quinnipiac	−3.155	−0.698	0.107
	(3.482)		
Selzer	−2.937	−1.824	−2.187
	(5.524)		
Siena	−10.494	−8.715	5.756
	(1.518)		
Siena/NYT	0.539	1.199	−1.647
	(8.401)		
Suffolk	−2.876	−2.133	0.209
	(2.977)		
SurveyUSA	−3.914	−2.714	1.190
	(6.513)		
SurveyUSA/High Point University	−3.331	−0.351	−0.387
	(7.661)		
The Washington Post	0.725	1.523	−0.871
	(5.228)		
University of Massachusetts Lowell/YouGov	−4.507	−2.828	0.228
	(3.410)		
Washington Post/George Mason University	−5.634	−2.650	0.754
	(1.950)		
YouGov	−1.250	−0.349	0.190
	(4.219)		
YouGov Blue	−0.606	0.258	−0.665
	(5.347)		
End Date	0.021	0.023	0.014
	(0.004)		
U. North Florida		2.760	−4.044
YouGov/Center for Working Class Politics		−2.217	−0.571
Num.Obs.	803	803	782
R2	0.287	0.280	0.201
R2 Adj.		0.240	0.146
AIC	4213.4		
BIC	4363.4		
Log.Lik.	−2074.682	−2080.530	−2025.723
ELPD		−2106.5	−2050.7
ELPD s.e.		40.8	41.8
LOOIC		4213.0	4101.4
LOOIC s.e.		81.6	83.5
WAIC		4211.5	4099.6
RMSE	3.21	3.52	3.22

[a] This table shows the regression models with custom variable names.

A.5.2 Bayesian model for Harris data set

In Figure 13, we applied a bayesian model to predict the percentage of polls Harris according each pollster. Each data point is color-coded by pollster, with support rates ranging from 35% to 65%. The shaded gray area around the trend line represents the confidence interval, indicating the uncertainty in support rate variations. The overall trend line (blue) shows a slight upward trend, increasing from about 46% in August to nearly 48% in October. Compared to Trump's model, the increasing trend of Harris is lower than Trump's. By October, most poll results have converged, with support rates centered between 45% and 50%.

Figure 13: Poll Percentage over Time with Bayesian Fit for Harris

A.5.3 Predictions for both Trump and Harris

In Table 5, we summarized the Bayesian model predictions for Trump and Harris on November 5, 2024, based on 27 pollsters, including the mean, lower bound, and upper bound.

Table 5: predictions for trump and harris by pollster

	Predictions Summary				
	Pollster	**Candidate**	**Predicted Mean**	**Lower Bound**	**Upper Bound**
2.5%	Emerson	Trump	49.74807	43.28219	56.35324
2.5%1	AtlasIntel	Trump	48.17328	41.60123	54.64031
2.5%2	Siena/NYT	Trump	49.42633	42.38901	56.25740
2.5%3	MassINC Polling Group	Trump	40.43736	33.54492	47.35021
2.5%4	Ipsos	Trump	44.35723	37.80983	51.06026
2.5%5	YouGov	Trump	47.88230	41.24154	54.41102
2.5%6	SurveyUSA	Trump	45.38837	38.59699	52.14075
2.5%7	Echelon Insights	Trump	48.12700	41.49321	55.09593
2.5%8	Selzer	Trump	46.31983	38.51049	53.96826
2.5%9	Suffolk	Trump	46.01116	39.33450	52.61313
2.5%10	Marist	Trump	49.15435	42.84796	55.67257
2.5%11	Muhlenberg	Trump	47.67240	40.26288	55.23404
2.5%12	The Washington Post	Trump	49.76048	43.08948	56.59855
2.5%13	Data Orbital	Trump	48.70637	41.12452	56.35085
2.5%14	Monmouth	Trump	45.76013	38.79269	53.12481
2.5%15	Quinnipiac	Trump	47.63194	40.11602	55.05057
2.5%16	Beacon/Shaw	Trump	49.05579	42.57426	55.50880
2.5%17	CNN/SSRS	Trump	46.30395	39.08137	53.75664
2.5%18	CES / YouGov	Trump	48.66837	41.86771	55.45890
2.5%19	Marquette Law School	Trump	47.75041	41.35784	54.47066
2.5%20	University of Massachusetts Lowell/YouGov	Trump	45.37743	38.74817	51.94390
2.5%21	Washington Post/George Mason University	Trump	45.50342	38.62663	52.25494

2.5%22	SurveyUSA/High Point University	Trump	47.87389	40.44454	55.27375
2.5%23	U. North Florida	Trump	50.92362	42.50054	59.16042
2.5%24	Siena	Trump	39.40856	32.12704	46.55104
2.5%25	Christopher Newport U.	Trump	42.89052	34.43363	51.61160
2.5%26	YouGov/Center for Working Class Politics	Trump	46.03093	37.41851	55.04925
2.5%27	McCourtney Institute/YouGov	Trump	46.45987	37.98113	54.93659
2.5%28	YouGov Blue	Trump	48.56749	40.20211	57.01238
2.5%29	Emerson	Harris	49.63418	42.87829	56.14878
2.5%30	AtlasIntel	Harris	48.00457	41.41379	54.68073
2.5%31	Siena/NYT	Harris	46.38911	39.39285	53.52612
2.5%32	MassINC Polling Group	Harris	55.34354	48.59579	62.24625
2.5%33	Ipsos	Harris	47.35787	40.63884	54.03892
2.5%34	YouGov	Harris	48.31199	41.74463	54.82360
2.5%35	SurveyUSA	Harris	49.29559	42.52625	56.07497
2.5%36	Echelon Insights	Harris	48.57784	41.82573	55.68554
2.5%37	Selzer	Harris	45.91051	38.09348	53.55376
2.5%38	Suffolk	Harris	48.22747	41.38776	55.15722
2.5%39	Marist	Harris	49.18951	42.66968	55.59954
2.5%40	Muhlenberg	Harris	48.76800	40.89858	56.70053
2.5%41	The Washington Post	Harris	47.27206	40.58815	54.17179
2.5%42	Data Orbital	Harris	44.88530	37.20791	52.65914
2.5%43	Monmouth	Harris	47.47059	40.31200	54.69014
2.5%44	Quinnipiac	Harris	48.12331	40.51441	55.35946
2.5%45	Beacon/Shaw	Harris	49.22415	42.71583	55.75335
2.5%46	CNN/SSRS	Harris	48.42867	41.17889	55.81444
2.5%47	CES / YouGov	Harris	49.64200	43.16074	56.14841
2.5%48	Marquette Law School	Harris	49.51876	43.05076	56.09031
2.5%49	University of Massachusetts Lowell/YouGov	Harris	48.31432	41.49899	55.01765

2.5%50	Washington Post/George Mason University	Harris	48.79504	41.93451	55.37136
2.5%51	SurveyUSA/High Point University	Harris	47.61006	40.40853	55.18262
2.5%52	U. North Florida	Harris	44.08972	35.93581	52.16615
2.5%53	Siena	Harris	53.71982	46.50436	60.95860
2.5%54	Christopher Newport U.	Harris	50.93937	41.98553	59.89053
2.5%55	YouGov/Center for Working Class Politics	Harris	47.46588	38.68029	55.91440
2.5%56	McCourtney Institute/YouGov	Harris	48.32295	39.86215	56.84292
2.5%57	YouGov Blue	Harris	47.40733	39.10670	55.71866

B Model details

To maintain readability while demonstrating model robustness, we include the following in the appendix:

- **Prior Justification and Sensitivity Analyses:** Alternative priors and their justifications are provided, alongside a sensitivity analysis to examine the impact of these priors on the posterior distributions.
- **Model Validation and Out-of-Sample Testing:** Validation metrics, such as RMSE calculations, out-of-sample testing, and test-train splits, offer evidence of the model's predictive accuracy, complementing in-sample performance.
- **Alternative Models and Comparison:** An analysis of simpler and more complex models is included, explaining the rationale for selecting this Bayesian model based on performance metrics and interpretability.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

FiveThirtyEight. 2024. "2024 National Presidential Poll Results." 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/.

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

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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

———. 2023. *Modelr: Modelling Functions That Work with the Pipe*. https://CRAN.R-project.org/package=modelr.

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

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.