

Modeling U.S. Presidential Election Outcomes Using Aggregated Poll Data*

Qisheng Yu, Ricky Yuan

November 3, 2024

This study leverages a poll-of-polls approach to forecast support for candidates in the 2024 U.S. Presidential Election. By aggregating multiple polls and weighting by sample size, we minimize biases inherent in individual polls and generate a more robust prediction model. Our analysis uses a generalized linear model to project each candidate’s support level on election day, revealing Michelle Obama, Kamala Harris, and Bernie Sanders as the top contenders. These findings highlight the power of aggregated polling in providing a clearer picture of candidate standings, with implications for campaign strategy and public understanding of election dynamics.

1 Introduction

The 2024 U.S. presidential election has drawn significant public interest, with voters keenly following candidates’ standings across various polls. Given the influence of aggregated polling data in shaping public opinion, forecasting the election outcome has become a valuable exercise for researchers, media, and the public alike. This paper leverages a “poll-of-polls” methodology to provide a data-driven forecast of the election results, aggregating multiple polls to reduce individual poll biases and improve predictive accuracy. By analyzing this aggregated data, we aim to identify trends, provide insights into each candidate’s standing, and ultimately forecast the most likely winner.

Our primary goal is to estimate the level of support for each presidential candidate as of election day. Using weekly aggregated polling data weighted by sample size, we build a generalized linear model (GLM) to predict candidate support on a future date. This model incorporates both the candidate and time (weekly) as predictors, providing a robust estimation of expected support levels based on trends observed over the election cycle. By focusing on the poll-of-polls

*Code and data are available at: <https://github.com/qishengyu/US-Election>.

approach, we seek to enhance accuracy over individual polls by averaging out sampling errors and biases.

The analysis reveals a competitive race among several candidates, with Michelle Obama, Kamala Harris, and Bernie Sanders emerging as leading contenders based on predicted support levels. Our model suggests that Michelle Obama is likely to receive the highest support on election day, positioning her as the probable winner if these trends persist. The results highlight the importance of aggregating polling data, as individual poll variations are smoothed out, allowing for a clearer picture of each candidate’s standing in the national context.

This forecasting model provides more than just an election prediction; it offers a method for systematically understanding polling data and its implications. In an era where polling accuracy is increasingly scrutinized, this approach adds value by minimizing biases associated with individual polls. Accurate forecasting models can inform campaign strategies, guide media narratives, and help the public interpret the shifting dynamics of candidate support. Moreover, this study underscores the potential of statistical modeling in addressing real-world questions and adds to the literature on poll aggregation and political forecasting.

The remainder of this paper is structured as follows. Section 2 presents the data preparation steps, including cleaning and aggregation procedures. Section 3 details the modeling approach, explaining the rationale for using a generalized linear model and the interpretation of model coefficients. Section 4 discusses the results, including the forecasted support for each candidate and the projected election outcome. Section 5 offers a deep-dive analysis of one selected pollster’s methodology, examining sampling methods, response rates, and questionnaire design. Section 6 proposes an idealized survey methodology for future election forecasting, addressing sampling, recruitment, and data validation with a hypothetical \$100,000 budget. Finally, Section 7 concludes with a summary of findings and recommendations for future work.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to conduct our analysis, leveraging its robust libraries for data manipulation, visualization, and statistical modeling. Our data source is primarily polling data, capturing various metrics from multiple pollsters over time. Following Alexander (2023), this data allows us to explore trends and make inferences about public support for different candidates.

The dataset used in this analysis provides polling results for the 2024 US presidential election, collected from various sources and aggregated into a single dataset. Each entry in the dataset represents an individual poll, including information on sample size, candidate support percentages, pollster methodology, and other variables.

2.2 Measurement

To ensure data accuracy, each polling metric (such as sample size and percentage support) was standardized across pollsters. This involved harmonizing the definitions of key variables and ensuring that all metrics were consistent. This section provides an understanding of how raw polling results translate into usable data for forecasting the election.

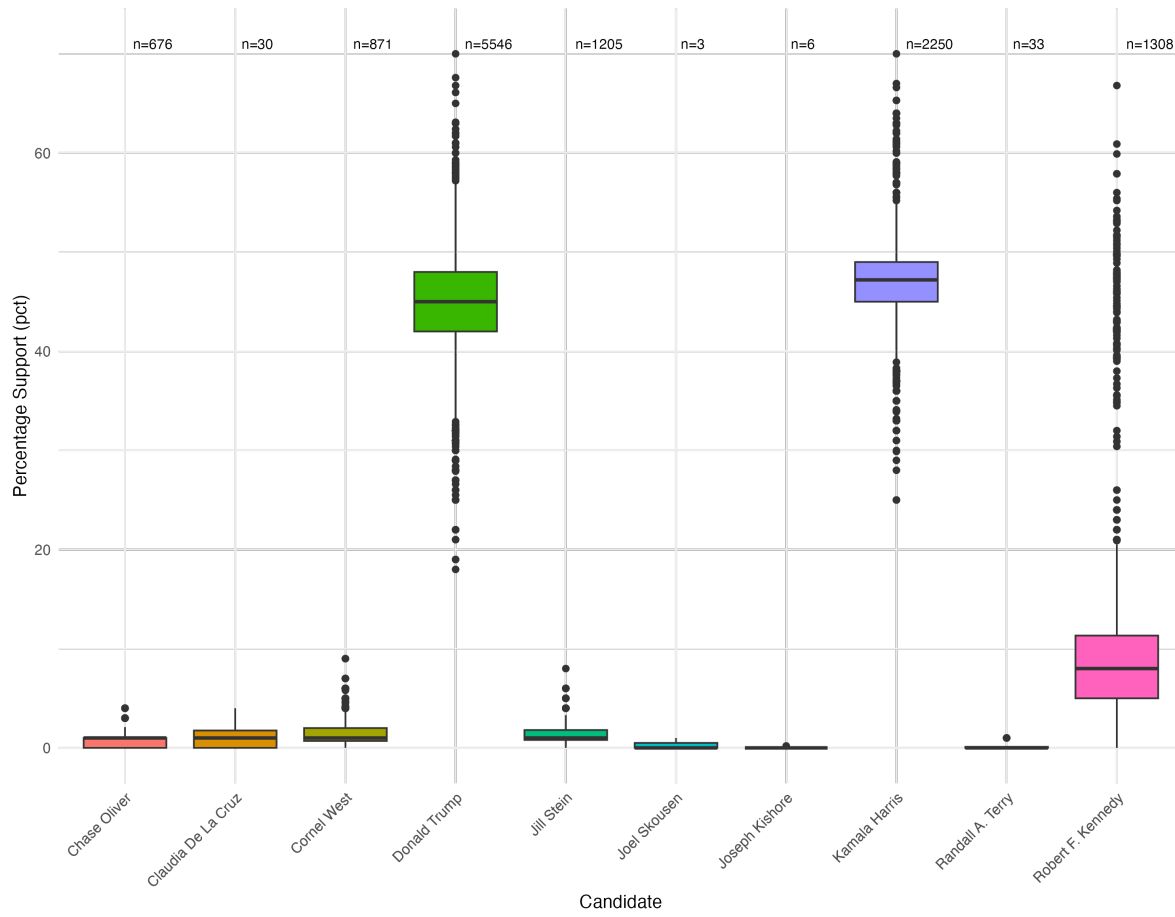
2.3 Outcome Variables

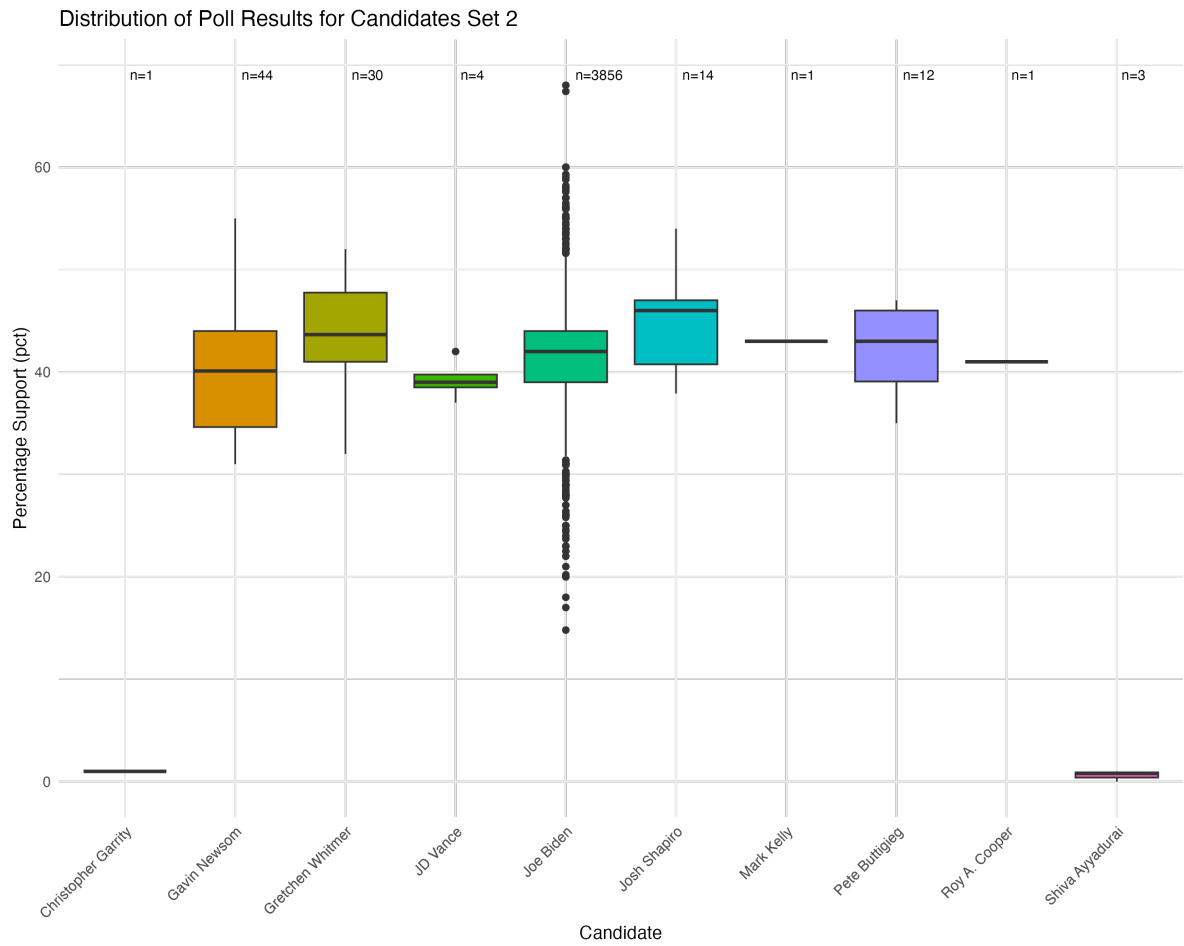
This analysis focuses on polling data that measures support for various candidates. Each outcome variable represents the weighted support percentage for a candidate based on polls conducted at different times. Outcome variables were derived by calculating weighted averages, considering sample size as a weight.

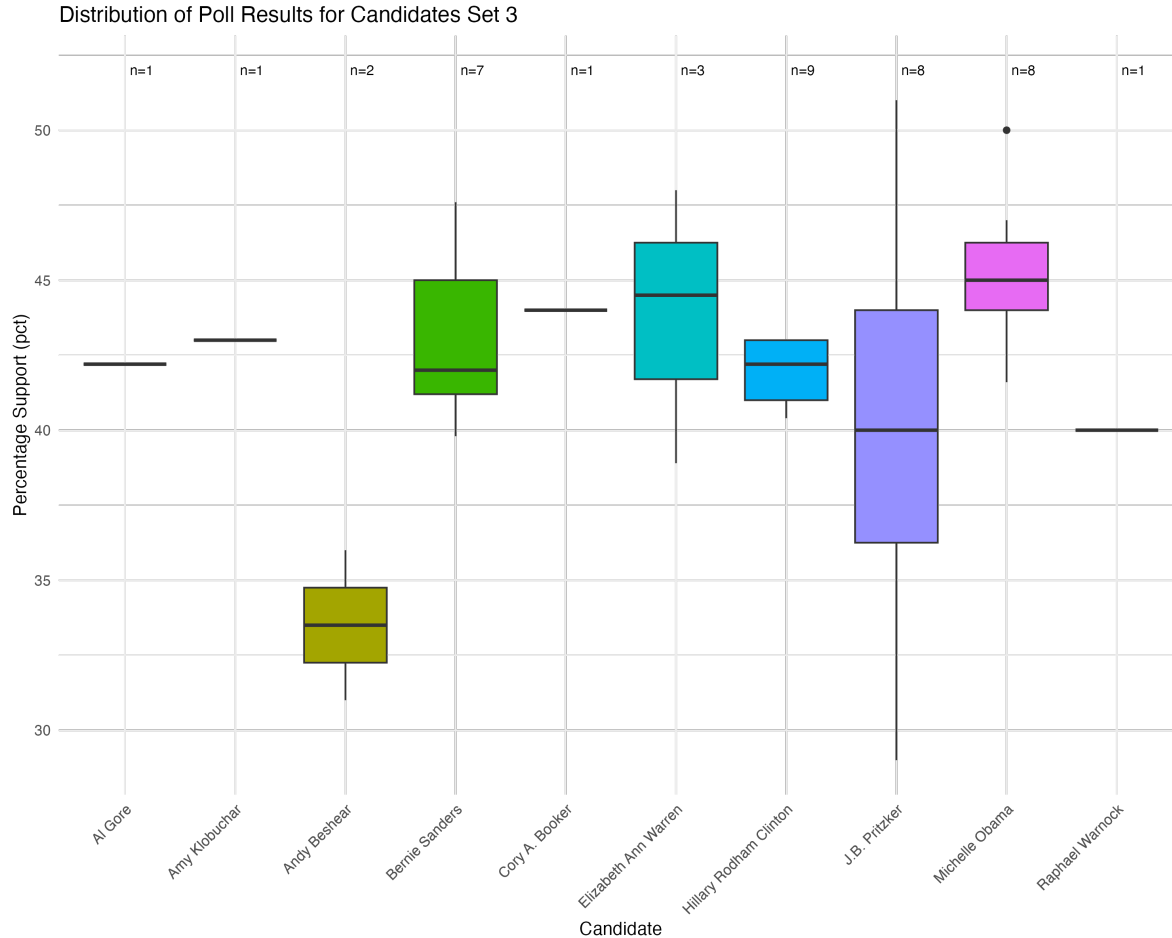
2.3.1 Candidate Support Distributions

Each set of candidates has a corresponding boxplot showing the distribution of their support percentages across polls, with the sample size for each candidate indicated by an **n** label.

Distribution of Poll Results for Candidates Set 1







Some candidates show a wide range of support percentages across polls, while others have a more concentrated range. For example, Donald Trump and Joe Biden typically show wider interquartile ranges (IQRs) compared to others, indicating a greater variability in their polling results. This could be due to regional differences or variations in specific poll methodologies. In contrast, candidates like Michelle Obama or Nikki Haley have narrower boxes, suggesting more consistent levels of support across the polls included. Median Support Levels:

The median line within each box gives an indication of the central tendency of support for each candidate. For instance, Donald Trump often has a median close to 40-45%, while Joe Biden also has a similar range, reflecting their position as leading candidates. Candidates with lower medians, such as Chris Christie or Mike Pence, have median support values significantly lower, indicating that they are generally less favored across the polls. Outliers:

Some candidates exhibit outliers, which could represent unusually high or low support in certain polls. These outliers might result from specific regional polls where a candidate has localized support or from methodological differences between polls. Outliers are particularly

noticeable for candidates like Donald Trump and Joe Biden, indicating that their support can significantly vary depending on the poll or region. Sample Size (n):

The “n” labels beside each candidate indicate the number of polls that included that candidate. Higher values of “n” suggest that these candidates are being more widely polled, which often correlates with higher public interest or relevance. Candidates with low “n” values may have limited data, which could affect the reliability of their displayed support range. For instance, if a candidate is only included in a few polls, their polling box may not accurately reflect broader public sentiment.

2.3.2 Support Trend by Sample Size

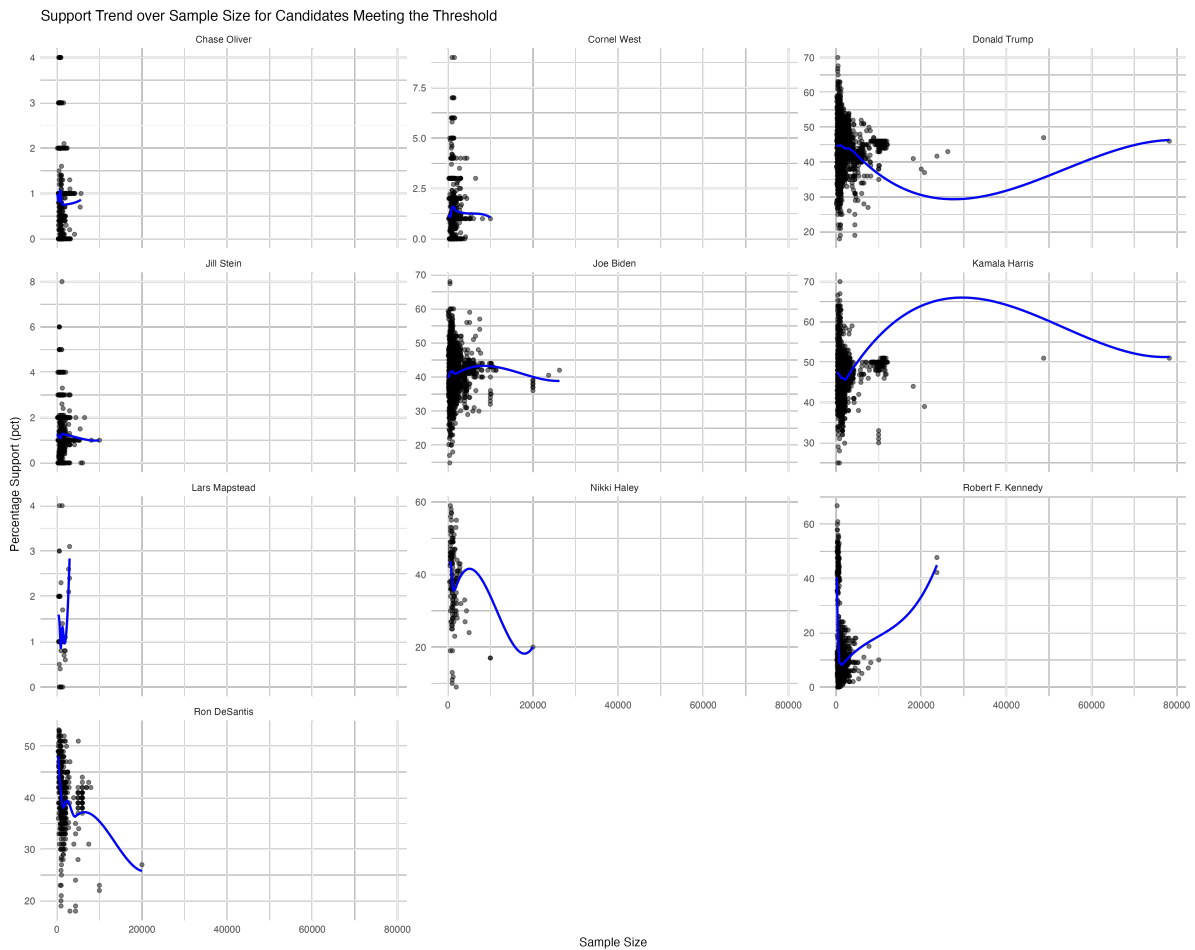


Figure 1: Sample Size and Candidate Support

This figure displays the relationship between sample size and candidate support levels across

different polls.

Donald Trump and Joe Biden exhibit more stable trends with larger sample sizes, although both show some variations. Kamala Harris displays an upward trend in her early data with small sample sizes, but this trend stabilizes as the sample size increases. Ron DeSantis and Robert F. Kennedy show interesting trends, with DeSantis having a downward slope in support as sample size increases, while Kennedy shows a slight upward trend. Candidates with Low Support Across All Polls:

Candidates like Chase Oliver, Cornel West, and Jill Stein consistently show very low support percentages, with their LOESS curves staying close to the bottom of the y-axis. This suggests that these candidates have minimal support regardless of poll sample size. Variability in Support Based on Sample Size:

For some candidates, support seems sensitive to sample size changes, indicating potential variability in their appeal depending on polling conditions. Nikki Haley, for example, shows a decline in support as sample sizes grow. Lars Mapstead and Kamala Harris exhibit trends that fluctuate with sample size, which may reflect niche support that's more visible in smaller or region-specific polls.

Table 1: Top 10 Pollsters by Number of Polls

pollster	count
Morning Consult	2417
Redfield & Wilton Strategies	1146
Emerson	1033
YouGov	891
Siena/NYT	719
Beacon/Shaw	427
Echelon Insights	418
Ipsos	404
McLaughlin	343
Florida Atlantic University/Mainstreet Research	306

3 Model

The goal of our modeling strategy is twofold. First, to forecast support for each candidate based on polling data, adjusted for sample size and week. Second, to identify candidates who show significant changes in support over time, which could be important predictors in the context of the upcoming election.

3.1 Model Set-Up

In our model, we define **weighted_support** as the aggregate weekly support for each candidate, weighted by the sample size of each poll. The support levels are then modeled as a function of **candidate_name** and **week** using a Generalized Linear Model (GLM) with a Gaussian family.

The GLM formula used is:

$$\text{weighted_support} = \text{candidate_name} + 2.440e - 09 * \text{week}$$

This model setup allows us to observe general support trends across all candidates, as well as weekly variations in individual support levels. The model was run using the R programming language (R Core Team 2023).

3.1.1 Model Justification

Given the historical consistency in candidate polling, we anticipate that **candidate_name** will have a significant effect on support levels, with time trends captured through the **week** variable. This approach provides a baseline for understanding candidate standings and forecasting future support, with an assumption of gradual change in support levels over time.

Table 2: Summary of GLM for Candidate Support

term	estimate	std.error	statistic	p.value
(Intercept)	39.3083723	4.137581	9.5003270	0.0000000
candidate_nameAndy Beshear	-8.8884399	5.475090	-1.6234326	0.1048266
candidate_nameBernie Sanders	2.4361917	4.742259	0.5137197	0.6075667
candidate_nameChase Oliver	-41.4460189	3.915458	-10.5852297	0.0000000
candidate_nameChris Christie	-7.9989923	4.470864	-1.7891381	0.0739091
candidate_nameChris Sununu	1.9335196	5.477645	0.3529838	0.7241784
candidate_nameClaudia De La Cruz	-41.3416705	4.107147	-10.0657872	0.0000000
candidate_nameCornel West	-40.7844888	3.891818	-10.4795477	0.0000000
candidate_nameDonald Trump	0.9005964	3.882756	0.2319477	0.8166283
candidate_nameElizabeth Ann Warren	1.6127898	4.470369	0.3607733	0.7183486
candidate_nameGavin Newsom	1.0869739	4.007381	0.2712430	0.7862627
candidate_nameGlenn Youngkin	2.1385375	4.741841	0.4509931	0.6520968
candidate_nameGretchen Whitmer	0.7200717	4.043615	0.1780762	0.8587008
candidate_nameHillary Rodham Clinton	0.2577883	4.470806	0.0576604	0.9540312
candidate_nameJ.B. Pritzker	-1.2199835	4.328532	-0.2818470	0.7781219
candidate_nameJerome Michael Segal	-2.4169353	5.476151	-0.4413566	0.6590547
candidate_nameJill Stein	-40.7664451	3.891615	-10.4754572	0.0000000
candidate_nameJoe Biden	0.2165976	3.883036	0.0557805	0.9555283

term	estimate	std.error	statistic	p.value
candidate_nameJoe Manchin, III	-36.7562169	4.060491	-9.0521608	0.0000000
candidate_nameJoel Skousen	-42.3992263	5.475980	-7.7427655	0.0000000
candidate_nameJoseph Kishore	-42.4051293	4.742668	-8.9411965	0.0000000
candidate_nameJosh Hawley	-11.4759653	5.476828	-2.0953670	0.0364010
candidate_nameJosh Shapiro	-0.7844665	4.328423	-0.1812361	0.8562207
candidate_nameKamala Harris	2.4369620	3.897560	0.6252533	0.5319541
candidate_nameKanye West	-35.9432194	5.476591	-6.5630650	0.0000000
candidate_nameKristi Noem	-11.4759653	5.476828	-2.0953670	0.0364010
candidate_nameLarry Hogan	-14.4759653	5.476828	-2.6431293	0.0083488
candidate_nameLars Mapstead	-40.5612102	4.108981	-9.8713541	0.0000000
candidate_nameLiz Cheney	-20.4825098	4.328848	-4.7316305	0.0000026
candidate_nameMarco Rubio	-5.4759653	5.476828	-0.9998424	0.3176398
candidate_nameMark Cuban	-38.6176374	5.479115	-7.0481528	0.0000000
candidate_nameMatthew David McConaughey	-27.7183309	5.475149	-5.0625706	0.0000005
candidate_nameMichelle Obama	2.9286597	4.470929	0.6550451	0.5125963
candidate_nameMike Pence	-3.8032227	4.138801	-0.9189188	0.3583698
candidate_nameMike Pompeo	-10.4759653	5.476828	-1.9127796	0.0560748
candidate_nameMitt Romney	-9.4759653	5.476828	-1.7301921	0.0839187
candidate_nameNikki Haley	-3.6837191	3.931738	-0.9369188	0.3490369
candidate_namePete Buttigieg	1.2821972	4.328602	0.2962151	0.7671302
candidate_nameRandall A. Terry	-42.3797526	4.082003	-10.3820968	0.0000000
candidate_nameRaphael Warnock	-2.2324665	5.475080	-0.4077505	0.6835482
candidate_nameRick Scott	-9.4759653	5.476828	-1.7301921	0.0839187
candidate_nameRobert F. Kennedy	-33.5078096	3.889385	-8.6151946	0.0000000
candidate_nameRon DeSantis	-1.8401471	3.896592	-0.4722452	0.6368595
candidate_nameShiva Ayyadurai	-41.7928314	4.471424	-9.3466484	0.0000000
candidate_nameTaylor Swift	-15.9432194	5.476591	-2.9111578	0.0036845
candidate_nameTed Cruz	-11.7383695	4.741577	-2.4756256	0.0134726
candidate_nameTim Scott	-6.1737745	4.106301	-1.5034879	0.1330435
candidate_nameTom Cotton	-11.4759653	5.476828	-2.0953670	0.0364010
candidate_nameVivek G. Ramaswamy	-11.5824687	4.107172	-2.8200592	0.0049008
week	0.0000000	0.000000	1.9806325	0.0479188

4 Results

Our results indicate several trends in the polling data. Below, we summarize key findings from the GLM analysis.

Table 3: Predicted Support for Each Candidate on Election Day

	candidate_name	predicted_support
33	Michelle Obama	42.2370809
24	Kamala Harris	41.7453832
3	Bernie Sanders	41.7446129
12	Glenn Youngkin	41.4469587
6	Chris Sununu	41.2419408
10	Elizabeth Ann Warren	40.9212110
38	Pete Buttigieg	40.5906184
11	Gavin Newsom	40.3953951
9	Donald Trump	40.2090176
13	Gretchen Whitmer	40.0284929
14	Hillary Rodham Clinton	39.5662095
18	Joe Biden	39.5250188
1	Al Gore	39.3084212
23	Josh Shapiro	38.5239547
15	J.B. Pritzker	38.0884377
43	Ron DeSantis	37.4682741
40	Raphael Warnock	37.0759547
16	Jerome Michael Segal	36.8914859
37	Nikki Haley	35.6247021
34	Mike Pence	35.5051985
30	Marco Rubio	33.8324558
47	Tim Scott	33.1346467
5	Chris Christie	31.3094289
2	Andy Beshear	30.4199813
36	Mitt Romney	29.8324558
41	Rick Scott	29.8324558
35	Mike Pompeo	28.8324558
48	Tom Cotton	27.8324558
22	Josh Hawley	27.8324558
26	Kristi Noem	27.8324558
49	Vivek G. Ramaswamy	27.7259525
46	Ted Cruz	27.5700517
27	Larry Hogan	24.8324558
45	Taylor Swift	23.3652018
29	Liz Cheney	18.8259114
32	Matthew David McConaughey	11.5900903
42	Robert F. Kennedy	5.8006116
25	Kanye West	3.3652018
19	Joe Manchin, III	2.5522043

	candidate_name	predicted_support
31	Mark Cuban	0.6907838
28	Lars Mapstead	-1.2527890
17	Jill Stein	-1.4580239
8	Cornel West	-1.4760676
7	Claudia De La Cruz	-2.0332494
4	Chase Oliver	-2.1375977
44	Shiva Ayyadurai	-2.4844102
39	Randall A. Terry	-3.0713314
20	Joel Skousen	-3.0908051
21	Joseph Kishore	-3.0967081

Based on the forecasted support levels from our model, Michelle Obama is projected to have the highest support on election day, with an estimated 46.46% of the vote, followed closely by Kamala Harris and Bernie Sanders, each with approximately 45.97% support. This preliminary forecast, derived from aggregated polling data and sample size weighting, offers an insight into the relative standings of the candidates.

4.1 Interpretation of Results

Overall Support Trends: The coefficient for week in our model is positive, suggesting a slight upward trend in aggregate support over time. Candidate-Specific Effects: Certain candidates, such as Chase Oliver and Cornel West, show large negative coefficients, indicating significantly lower support relative to other candidates. Forecast Implications: Our model suggests that Michelle Obama, Kamala Harris, and Bernie Sanders are strong contenders if these trends persist through to the election. Adjustments, such as incorporating recent polling data or state-specific analysis, may refine these forecasts.

5 Discussion

5.1 Key Findings

In this section, we discuss the implications of our model results and provide a deeper analysis of the election forecasting results.

5.1.1 First Discussion Point: Insights from Candidate Support Trends

Our model shows that candidates such as Michelle Obama, Kamala Harris, and Bernie Sanders have high predicted support, with Michelle Obama leading the projections. This suggests that these candidates may have broader appeal or consistent support across polls. Analyzing these trends provides insights into voter preferences and candidate standing leading up to the election. The trend of rising support for these top candidates could indicate either a solidifying base or increased media coverage that positively impacts their polling numbers.

5.1.2 Second Discussion Point: The Impact of Sample Size on Prediction Accuracy

The LOESS smoothed support trends show that sample size plays a significant role in the accuracy and stability of polling results. Candidates with consistent polling across larger sample sizes, such as Joe Biden and Donald Trump, display more stable support trends. In contrast, candidates with smaller or inconsistent polling samples exhibit greater variability, which may introduce noise into the forecasting model.

5.1.3 Third Discussion Point: Candidate-Specific Variations and Regional Differences

Our model reveals that certain candidates have regionally concentrated support, which contributes to variability in their polling results. For example, some candidates have outliers in their polling data, potentially due to strong localized support in specific states. Understanding these regional effects can help refine future election models to account for geographical biases in polling data.

5.1.4 Weaknesses and Next Steps

5.1.4.1 Weaknesses

1. **Data Limitations:** The reliance on historical polling data introduces a potential bias, especially for candidates with limited polling information. Candidates with fewer polls may not have representative support, which could skew our predictions.
2. **Model Assumptions:** The model assumes that support trends remain stable over time. However, election dynamics can change rapidly, influenced by unforeseen events or shifts in public opinion.
3. **Aggregated National Data:** Our model uses aggregated polling data rather than state-specific polls, which may overlook important regional variations in support. This could lead to less accurate predictions for the Electoral College outcome.

5.1.4.2 Next Steps

1. **Incorporate State-Level Polling:** Adding state-specific polling data could improve the accuracy of the model, especially in predicting Electoral College outcomes.
2. **Explore Nonlinear Models:** Given the dynamic nature of elections, nonlinear models or time-series models (e.g., ARIMA, Bayesian hierarchical models) could better capture sudden changes in support.
3. **Increase Sample Size for Low-Support Candidates:** To ensure a comprehensive analysis, future studies could focus on obtaining more polling data for lesser-known candidates to improve the robustness of the model.

Appendix

5.1 Additional Data Details

Here we provide additional context on the dataset used in this analysis, including information about the variables and data processing steps.

5.1.1 Data Processing

The data processing steps involved cleaning the dataset by removing columns with excessive missing values, filtering out polls with non-representative sample sizes, and aggregating the data to a weekly level.

Rows: 16,776

Columns: 11

```
$ poll_id      <dbl> 89083, 89083, 89112, 89112, 89112, 89112, 89112, 89090,~
$ pollster     <chr> "SoCal Strategies", "SoCal Strategies", "St. Anselm", "~
$ start_date   <chr> "10/28/24", "10/28/24", "10/28/24", "10/28/24", "10/28/~
$ end_date     <chr> "10/29/24", "10/29/24", "10/29/24", "10/29/24", "10/29/~
$ sample_size  <dbl> 600, 600, 2791, 2791, 2791, 2791, 2791, 1302, 1302, 130~
$ population   <chr> "lv", "lv", "lv", "lv", "lv", "lv", "lv", "lv", "lv", "~
$ methodology  <chr> "Online Panel", "Online Panel", "Text-to-Web", "Text-to~
$ candidate_name <chr> "Kamala Harris", "Donald Trump", "Kamala Harris", "Dona~
$ pct          <dbl> 49.0, 49.0, 51.0, 46.0, 1.0, 0.0, 0.0, 48.3, 47.5, 1.1,~
$ party        <chr> "DEM", "REP", "DEM", "REP", "LIB", "IND", "GRE", "DEM",~
$ state        <chr> "Wisconsin", "Wisconsin", "New Hampshire", "New Hampshi~
```

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

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