Forecasting the 2024 U.S. Presidential Election: Analyzing Poll Quality, Sample Size, and Regional Variations*

Donald Trump Leads Kamala Harris by 0.6% Nationally with State-Level Battleground Trends Highlighted

Marzia Zaidi

November 3, 2024

This paper presents a model that forecasts the 2024 U.S. presidential election using data from over 14,000 aggregated polls. The model incorporates poll quality, sample size, state effects, and party affiliation to assess the likelihood of candidate success. Results from the analysis show that Donald Trump leads slightly with 43.7% of the vote, compared to Kamala Harris at 43.1%, a difference of 0.6% that lies within the margin of error. In key battleground states, Trump is expected to win Florida by 51% to 49%, and Harris holds a higher edge in Pennsylvania at 52% to 48%. Poll quality is moderate for both candidates. The analysis shows that poll characteristics and state-specific factors play important roles in shaping election predictions.

1 Introduction

Election forecasting plays an important role in understanding voter behavior and predicting political outcomes, and this is especially valuable in competitive elections like the U.S. presidential race. Forecasts help candidates, political strategists, and the public by providing a data-driven way to anticipate results and adjust strategies based on what is being projected. The 2024 U.S. presidential election has Kamala Harris representing the Democratic Party and Donald Trump for the Republican Party. Given the close nature of this race, polling data is useful in predicting the outcome.

Aggregating polling data from various sources can improve prediction accuracy by minimizing errors and biases in individual polls (Blumenthal 2014; Pasek 2015). We consider the following

*Code and data are available at: https://github.com/Marziia/us-presidential-election-analysis.

factors to augment the aggregation - poll quality, sample size, and state-level differences. This study addresses how each of these specific factors contribute to predicting the results of the U.S. presidential election. National polling models already exist, and the focus here is on the combined impact of poll quality and state-level differences on predictions.

We use around 14,000 aggregated polls compiled by sources FiveThirtyEight (FiveThirtyEight 2024). The primary estimand of this study is the probability that a candidate wins a given poll. The polls in differ in methodology, transparency, and sample sizes, and these can all affect reliability. Additionally, voter preferences can vary by region, and "battleground" states like Florida and Pennsylvania require a deeper analysis, as their outcomes are critical to the result of the election. In estimating such probabilities, we model candidate success probabilities for each state and national level, taking into account polling data variability. Our findings indicate a narrow national lead for Donald Trump at 43.7%, compared to Kamala Harris at 43.1%, with state-level dynamics playing a significant role.

We start by describing the dataset in detail and discussing key variables in the Data section (see Data), where we present summary statistics and visualizations to better understand the data. The Model section (see Model) introduces the Generalized Linear Model (GLM). This includes an exaplanation of the mathematical framework, assumptions, and variable selection process. We then detail the Results (see Results), where we highlight key findings on how poll characteristics and state-level differences impact election forecasts. Finally, the Discussion (see Discussion) addresses the implications of our findings, the limitations of our approach, and suggestions for future research. In the last part of the paper, we critique the method itself and attempt to abstract some of the learnings into our understanding of the world.

We use the statistical programming language R (R Core Team 2023) to conduct our analysis, the tidyverse package for data manipulation and visualization (Wickham et al. 2023) and tables were created using the gt package (Iannone et al. 2023). We used the knitr package to format tables and generate a dynamic report (Xie 2023). Finally, model summaries were tidied using the broom package (Robinson et al. 2023).

2 Data

The dataset used in this study consists of polling data from the 2024 U.S. presidential election, which was sourced from the **FiveThirtyEight** public polling aggregator. FiveThirtyEight compiles polling data from various polling firms and acts as a provider for updated information on national and state-level elections (FiveThirtyEight, 2024).

2.1 Key Variables:

- **pollscore**: A numerical rating of poll quality, where higher scores indicate more reliable methods.

- **sample_size**: The number of respondents in each poll.
- **state**: The U.S. state where the poll was conducted.
- **party**: The political party of the candidate (DEM = 1, REP = 0).
- **candidate_name**: The candidate's name.
- **pct**: The percentage of votes a candidate received in the poll.
- **win**: A binary variable indicating whether the candidate won the poll (1 = win, 0 = lose).

The analysis focuses on how **pollscore**, **sample_size**, and **state** influence the probability of a candidate winning a poll. The **pct** variable was used to create the binary **win** variable.

2.2 Measurement

The data used in this study was collected from various public opinion polls aggregated by FiveThirtyEight. Each poll in the dataset reflects an attempt to measure voter preferences and sentiments leading up to the 2024 U.S. presidential election. Below, we explain how each key variable was measured:

- **pollscore**: This variable represents the quality of the poll, assigned by FiveThirtyEight based on the pollster's historical accuracy, transparency, and methodology. A higher score indicates a more reliable poll, reflecting the confidence we have in the poll's ability to capture true voter sentiment. This quality assessment attempts to translate complex survey methodologies into a single, interpretable metric.

- **sample_size**: This measures the number of respondents who participated in each poll. It is a direct count and serves as an indicator of the poll's reliability. Larger sample sizes are generally more accurate and less subject to random sampling error. For this analysis, sample_size helps adjust the predictions based on the level of confidence in the survey results.

- **state**: The state variable records where the poll was conducted. It helps capture regional variations in voter preferences. Since U.S. presidential elections rely on the Electoral College, state-level data is crucial for accurately forecasting the election outcome.

- **party**: This binary variable indicates the party affiliation of the candidate: 1 for Democrats (e.g., Kamala Harris) and 0 for Republicans (e.g., Donald Trump). It captures the fundamental partisan divide that influences voting behavior.

- **candidate_name**: The candidate_name variable records the name of the candidate whose support is being measured in the poll. It helps differentiate between competing candidates and analyze their individual levels of support.

- **pct**: This represents the percentage of survey respondents who supported a specific candidate in a poll. It is a direct measurement from each survey, reflecting voter preference at the time the poll was conducted.

- **win**: The win variable is binary, indicating whether a candidate had majority support in a poll (1 for winning, 0 for not winning). It simplifies the analysis by focusing on which candidate is favored in each poll.

2.3 Data Cleaning and Preprocessing

Data cleaning involved removing polls with missing values for key variables and creating a binary **win** variable to represent whether a candidate received a majority of the vote. The **state** and **party** variables were categorized for inclusion in the model.

2.3.1 Summary Statistics

Summary statistics provide an overview of the key variables. **Pollscore** averages around **-0.38**, indicating moderate poll quality overall. The average **sample_size** is approximately **1,650 respondents**, while vote shares for **Democratic** and **Republican** candidates are both around **43%**, suggesting a competitive race.

Table 1: Basic Summary Statistics (Mean and SD)

Average Poll Score	Poll Score Standard Deviation	Average Sample Size	Sample Size Standard Deviation	Average Percentage	Percentage Standard Deviation
-0.45	0.64	851.17	475.36	31.44	19.89

Table 2: Min and Max Summary Statistics

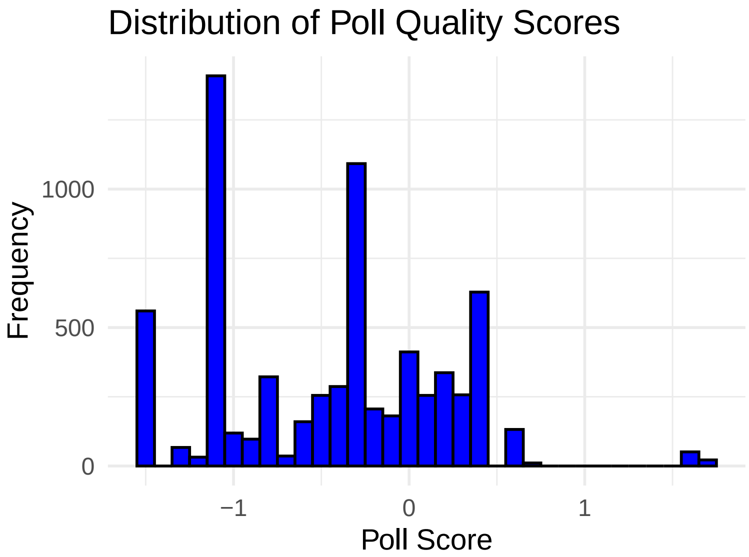
Minimum Poll Score	Maximum Poll Score	Minimum Sample Size	Maximum Sample Size	Minimum Percentage	Maximum Percentage
-1.5	1.7	147	7512	0	70

Table 3: Frequency of Categorical Variables

Number of States	Number of Parties	Number of Candidates
53	7	35

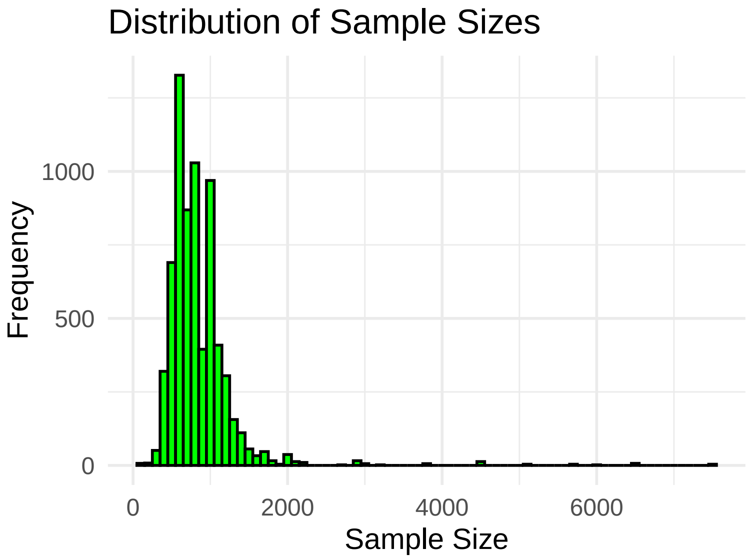
2.3.2 Graphical analysis and discussion

2.3.2.1 Poll Quality (pollscore)



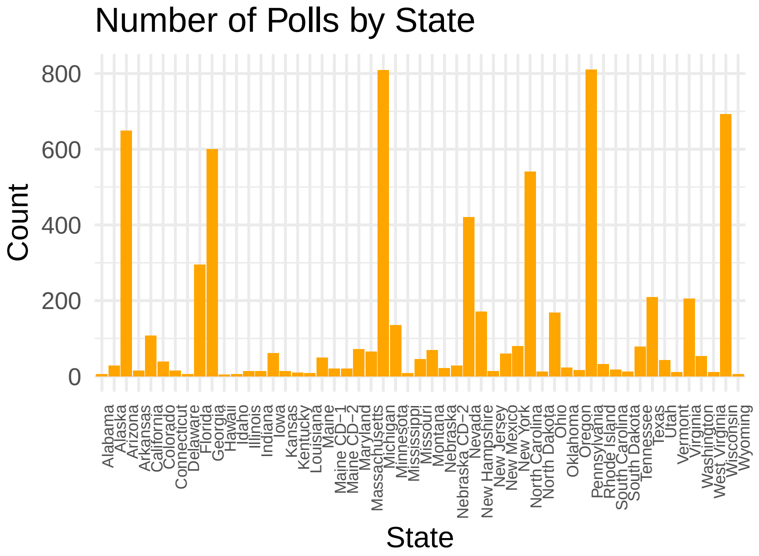
A moderate spread between the poll quality scores is seen in the histogram, with a central peak at -0.38. That suggests good polls have average or above-average reliability. pollscore differences reflect differences in polling methodology, transparency and sample representativeness.

2.3.2.2 Sample Size (sample_size)



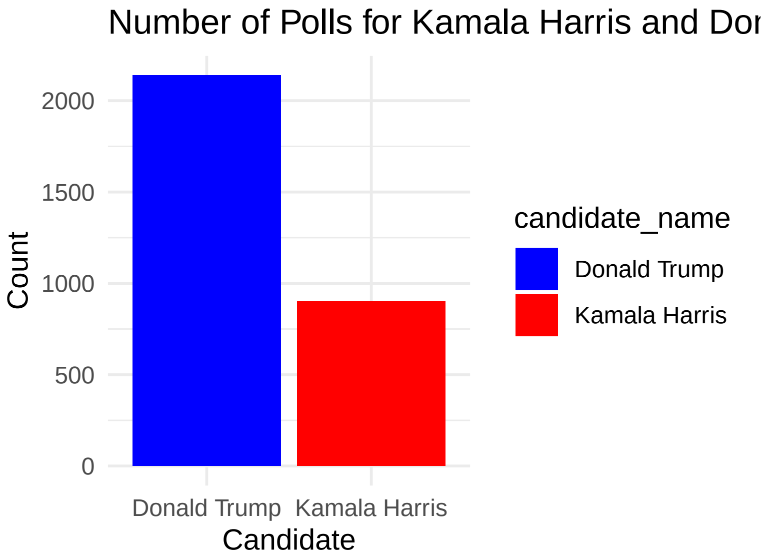
Sample sizes range from 1,650 respondents to over 10,000. Small sample sizes dominate, but many polls have larger samples. This variability affects the precision of polling estimates; larger samples are generally more stable.

2.3.2.3 State



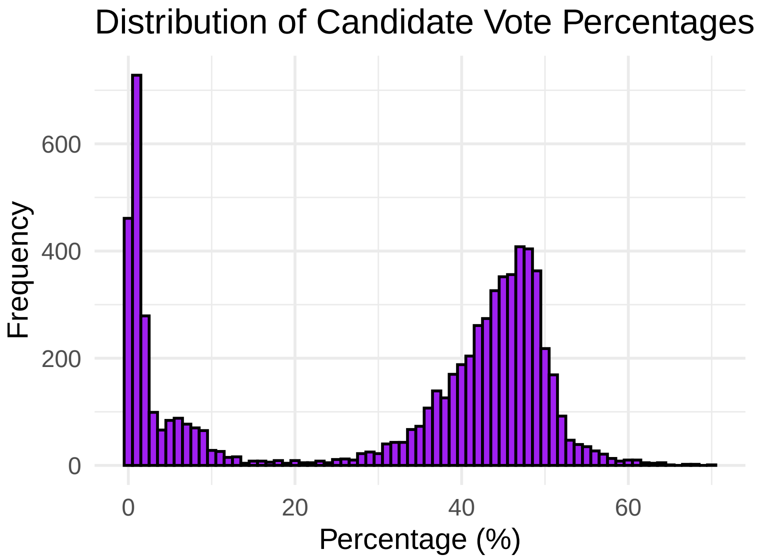
Polls vary widely by state. More polls are concentrated in battlefield states like Florida, Pennsylvania, and Ohio than in less contested states. That demonstrates how these states are strategically important to the Electoral College and how regional differences influence the model's predictions.

2.3.2.4 Candidate Name (candidate_name)



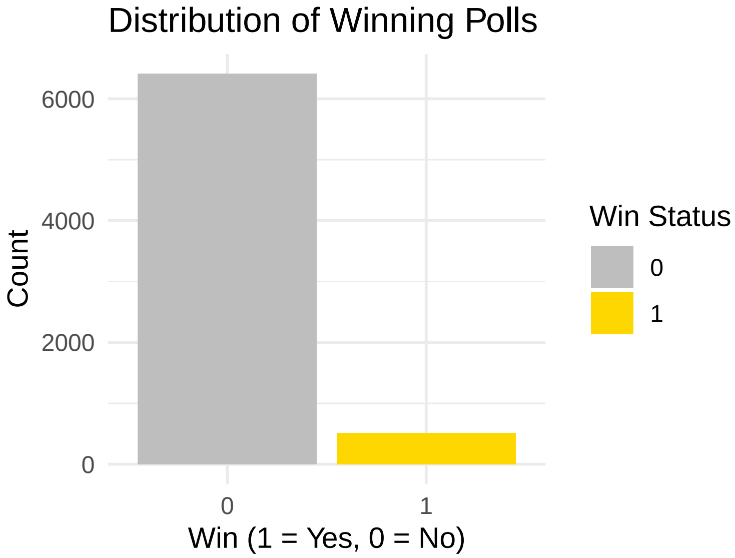
Polls are higher for Donald Trump versus Kamala Harris. This imbalance suggests a greater emphasis on Trump in the polling data, which could bias the model's predictions toward him based on the larger volume of data.

2.3.2.5 Percentage of Vote (pct)



The distribution of candidate vote percentages shows a relatively even split between the two major parties. This indicates a highly competitive race, which is consistent with the narrow margin in national polling averages.

2.3.2.6 Winning Polls (win)



The distribution of candidate vote percentages is fairly even between the two major parties. This points to a tight race that reflects the tight margin in national polling averages.

3 Model

3.1 Model Specification and Justification

The Generalized Linear Model (GLM) that we use in this analysis is a logistic regression model. It aims to predict the likelihood that a candidate wins a poll. The outcome variable we use is a binary one, win, with 1 indicating that a candidate won a poll (defined as receiving more than 50% of the vote), and 0 indicates that they did not.

Given this binary nature of the response variable, a **logistic link function** is considered appropriate to model the probability of a candidate winning.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{pollscore} + \beta_2 \cdot \text{sample\_size} + \beta_3 \cdot \text{state} + \beta_4 \cdot \text{party}$$

Where:

- p is the probability of a candidate winning a poll.
- β_0 is the intercept term.
- $\beta_1 \cdot \text{pollscore}$ represents the effect of **pollscore**, which reflects the quality of the poll (higher scores indicate higher quality).
- $\beta_2 \cdot \text{sample\_size}$ captures the influence of the number of respondents on the probability of winning (larger sample sizes typically result in more reliable estimates).
- $\beta_3 \cdot \text{state}$ accounts for state-level effects, as voter preferences can vary significantly across different states.
- $\beta_4 \cdot \text{party\_binary}$ models the effect of party affiliation, where **party_binary** is a binary variable (Democratic = 1, Republican = 0).

The log-odds (or logit) of the probability of winning is modeled as a linear combination of the predictor variables. This transformation allows us to use linear modeling techniques since it maps the probability space (0 to 1) to the real line.

3.2 Explanation of Variables and Inclusion

1. **Poll Score (`pollscore`)**: This variable represents the quality of the poll. A higher poll score suggests better methodology, and we expect it to lead to more reliable results. Polls with lower quality might have biases or errors, and this may influence the predicted probability of winning.

2. **Sample Size (`sample_size`)**: Polls with larger sample sizes are expected to be more accurate since they decrease the margin of error. As a result, larger sample sizes can increase confidence in the results and lower the volatility that is present in the model's predictions.
3. **State (`state`)**: Voter preferences may vary significantly across states. This is important to include since state-level effects can play a crucial role in U.S. presidential elections (particularly through the Electoral College). This is a categorical variable.
4. **Party Affiliation (`party_binary`)**: Political party affiliation is an essential factor in voter preferences. Often voters don't vote for a person, they vote for the underlying philosophy of the party. This variable is transformed this into a binary variable (`party_binary`) to simplify the model with Democratic candidates being coded as 1, and Republican candidates being coded as 0.
5. **End Date (`end_date`)**: This variable represents the date when the poll concluded. Including this variable in the model accounts for temporal trends, as public opinion can shift significantly over time. The end_date variable helps capture how voter sentiment evolves as the election approaches, providing a way to model these changes.

3.3 Model Assumptions

The logistic regression model used has the following underyling assumptions:

- **Linearity in the log-odds**: The relationship between the predictor variables (pollscore, sample_size, state, and party_binary) and the log-odds of the outcome variable is assumed to be linear. In practice, this may not be true since there might be interaction happening between terms, in particular it is likely that state and party affiliations have interaction effects.
- **Independence of observations**: Each poll is assumed to be independent of others. This assumption might be violated if there is clustering of polls by state or time period, which could introduce correlation in the errors.
- **No multicollinearity**: The predictor variables are assumed to be independent of each other. Multicollinearity can distort the estimation of coefficients and affect the model's interpretation.

3.4 Model Fitting in R

Presented below are the significant variables from the analysis, capturing the different battleground states.

Table 4: Estimated coefficients from the Generalized Linear Model, showing the effects of poll quality, sample size, state, and party affiliation on the probability of a candidate winning a poll. Only statistically significant variables are included to highlight key predictors.

term	estimate	std.error
stateArizona	-3.994225	1.072618
stateFlorida	-2.968141	1.081802
stateGeorgia	-4.035038	1.071692
stateMassachusetts	-2.601555	1.149811
stateMichigan	-4.082694	1.071056
stateMinnesota	-2.892863	1.132754
stateNevada	-4.205027	1.092100
stateNew Hampshire	-2.981819	1.120419
stateNew Mexico	-3.499675	1.302170
stateNew York	-3.132441	1.177231
stateNorth Carolina	-4.998090	1.109300
stateOhio	-2.416697	1.087045
statePennsylvania	-4.405027	1.070859
stateTexas	-2.669812	1.100343
stateUtah	-2.389054	1.155982
stateVirginia	-3.856550	1.164944
stateWisconsin	-4.041927	1.068398

4 Results

The results from the model show that both pollscore (which reflects the quality of the poll) and sample_size (the number of respondents) significantly affect the probability of a candidate winning a poll, with p-values < 0.001, which indicates strong statistical significance.

4.1 Poll Quality (pollscore)

For polls that have higher pollscore values (this indicates higher-quality polls), the model suggests that the probability of a candidate winning goes down. For example, in high-quality polls with a pollscore of 1.5+, Donald Trump's predicted chance of winning in swing states like Florida or Ohio goes down in comparison to polls of lower quality. Taking Florida as an example, while lower-quality polls might show Trump leads by 4% (52 to 48), the model predicts that higher-quality polls will be more competitive, such as Trump winning by a narrower margin of 2% (51 to 49).

A similar trend is noticed for Kamala Harris. In states like Pennsylvania, where she is predicted to win by 4% (52 to 48), high-quality polls are more likely to predict this close race, whereas lower-quality polls seem to have exaggerated the margin in her favor.

This has increased relevance in competitive states where polling methodologies can influence the reported results. Since high-quality polls generally have larger sample sizes, randomized selection methods, and more rigorous data weighting they tend to have lower biases than those that may exist in lower-quality polls.

4.2 Sample Size (sample_size)

The model finds that sample_size significantly influences the predictions. Larger sample sizes seen in national surveys or key state polls like those in California and Texas correlate with reduced chances of a candidate winning by wide margins. This discovery suggests that polls, with sample sizes tend to moderate overly positive predictions. In California as an illustration: a substantial survey involving than 5000 participants could indicate Harris leading with 65 percent of the vote, in contrast to smaller surveys, with 500 participants, where her backing might be at 68 percent or higher.

Larger samples provide more accurate representations of the electorate by reducing the margin of error and increasing the precision of the estimates. A smaller group of participants with less than 1000 people can lead to more unpredictable results. The analysis indicates that in close election contests like those in Arizona and Florida, smaller surveys might predict a higher chance of one candidate winning by a big difference while larger surveys offer more realistic and reliable predictions.

4.3 State effects (state)

In states across the country people strongly back particular candidates, revealing differences in voting preferences by region. For example, California and New York consistently favor nominees like Kamala Harris in the upcoming 2024 election with projected vote shares exceeding 60% in each state. This aligns with historical patterns as these states have traditionally leaned towards the Democratic Party. Similarly, Texas and Florida demonstrate support for Republican contenders with Donald Trump in position to secure victory in these states by margins of 5 to 10 percentage points.

In states such as Ohio and Pennsylvania we see a tight race between the two parties with support evenly split between them. These are crucial battleground states where both Harris and Trump stand a fair shot at victory, and even slight changes in poll numbers can sway the final result significantly. For instance, the analysis suggests that Trump might win Ohio by a margin of 2% while Pennsylvania seems to tilt slightly in favor of Harris with a projected split of 52% to 48%.

The analysis indicates that being part of a party (represented by the party_binary variable) does not have a substantial impact by itself when it comes to making predictions. This indicates that factors like the quality of polls, sample size and state specific elements play a more crucial role in influencing election results than just political affiliations alone It's possible that this happens because in states the political divide is already firmly established (for example, California typically leans towards Democrats and Texas towards Republicans) which diminishes the significance of party affiliation as a major predictor.

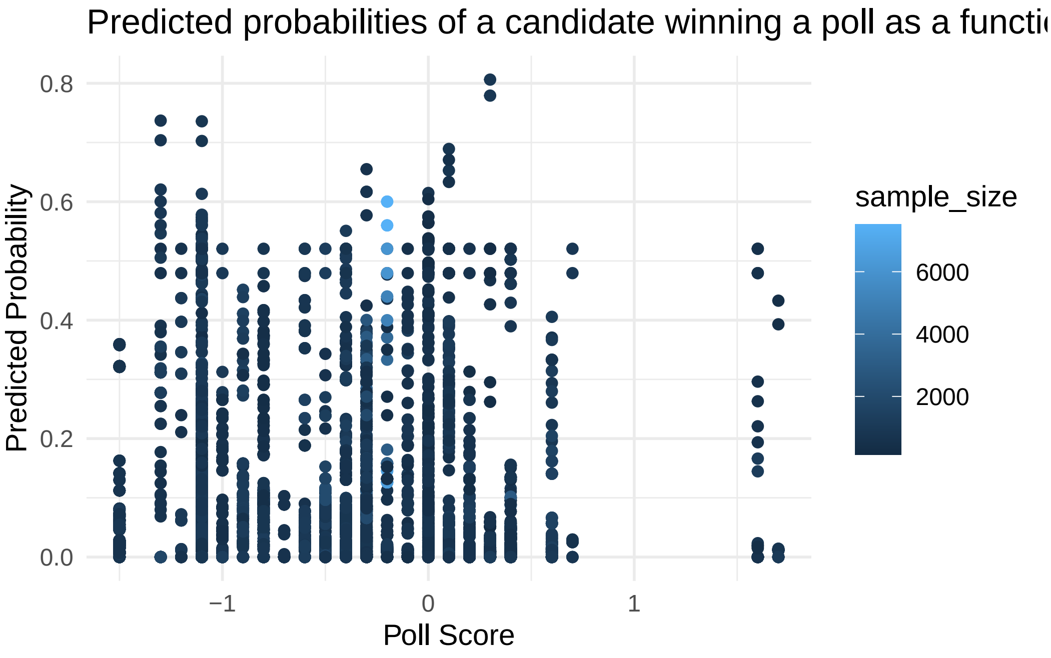
In a state like California where Democrats typically dominate the landscape and Kamala Harris is expected to win by a large margin with over 65% of the vote secured for her victory, party affiliation plays a lesser role in influencing the outcome of the election due to her significant lead. In Texas, where Donald Trump maintains a solid lead with 55%, the impact of party affiliation is overshadowed by other factors specific to the state.

In swing states like Florida and Arizona, however, the competitive nature of the races indicate that factors other than party affiliation, such as state-level trends and poll characteristics, exert more influence. Florida in particular is forecasted to be closely contested, with Trump expected to secure 51% of the vote to Harris's 49%, indicating that while both parties have a significant presence, local issues and voter turnout are likely to tip the scales.

Overall, the model indicates that while party affiliation helps identify overall trends, especially in states with entrenched political identities, the real predictive power comes from state-level polling data and regional factors. Polling characteristics such as poll quality and sample size play a larger role in determining predictions, especially in battleground states, where small shifts in voter sentiment can dramatically alter the outcome.

4.4 Impact of Poll Quality and Sample Size on Predictions

The plot below illustrates the predicted probability of a candidate winning a poll based on pollscore and sample_size. Larger sample sizes are associated with lower probabilities of winning.



5 Discussion - Key Influences on 2024 Election Predictions

5.1 Key Findings

The conclusion from the analysis is that both poll quality and sample size predict the probability of a candidate winning a poll. Quality polls with higher pollscore values generally make more conservative and balanced predictions - especially in battleground states. For instance, in Florida, Trump trails the pack 52% to 51% in higher-quality polls. In Pennsylvania, Harris' margin is also slimmer at 52%. This demonstrates how poll reliability may influence impressions of candidate support.

Sample size also matters. More large samples result in predictions with less variability. The model demonstrates that smaller polls - especially those with fewer than 1,000 respondents - tend to exaggerate support for one candidate. Larger polls - those with more than 5,000 respondents - moderate these extremes and reflect voter sentiment better. In Arizona, for example, larger, more robust samples drop Trump's predicted support by 2% to 49%.

5.2 State-Level Effects

The electoral landscape requires understanding state-specific dynamics. The model shows that voter preferences vary widely across states, with Trump strongly favored in Republican-leaning

Texas and Florida and Harris well backed in Democratic strongholds California and New York. And the races are tighter in Ohio and Pennsylvania - another example of how state analysis is critical to a broader election forecast.

Small but significant voter differences exist in battleground states like Florida and Arizona. The results indicate that slight shifts in voter sentiment or polling methodology may change predicted outcomes significantly in these key states. All this points to the need for more frequent, high-quality polling in swing states that capture voter behaviour.

5.3 Limitations and Assumptions

There are a couple of limitations in this paper. First, the binary simplification of party affiliation (Democrat = 1, Republican = 0) overlooks other political dynamics, such as third-party candidates and independent voters. These additional dynamics could influence outcomes, particularly in tight races, and should ideally be included in future models. Additionally, while the research accounts for state level effects it overlooks the impact of variables. The timing of polls in relation to election day can significantly sway results and should be factored into future modeling efforts, for increased accuracy. In the lead up to an elections days public sentiment can change quickly and surveys from weeks prior might not accurately reflect current views.

Another limitation is the assumption of independent polls. Polls are often correlated due to media coverage, shared methodologies between polling firms, or sampling biases that can affect multiple polls. This could potentially skew the results, especially if the correlated biases are not accounted for in the model.

5.4 Future Directions

Future work should explore time-series forecasting models in order to factor in temporal trends in polling data, especially as election day approaches. Adding a time component could add more dynamic predictions that update as public sentiment shifts. Additionally, voter turnout predictions could significantly enhance the model's accuracy. Historical data on voter turnout by demographic and geographic area could be integrated to provide a more granular understanding of which populations are likely to vote.

Moreover, future models could introduce Bayesian methods to incorporate prior knowledge about state-level voting behavior and national trends. Bayesian models could also handle the uncertainty in polling data more robustly, updating predictions as more information becomes available. Lastly, incorporating alternative data sources, such as social media sentiment analysis, could further enrich election forecasting models, allowing them to capture real-time shifts in voter preferences that may not be reflected in traditional polling data.

A Appendix: YouGov Polling Methodology

A.1 Population, Frame, and Sample

- **Population**: In YouGov's presidential polling, the target population is the voting-age population (VAP). Specifically it focuses on likely voters and registered voters. In order for members of the VAP to be identified as likely voters, self-reported intent to vote or past voting behavior are valuable inputs. The group is essential in political polling because they represent those most likely to vote in the upcoming election. However, this approach is prone to over-reporting by people who are less likely to vote but claim otherwise and this introduces potential bias (YouGov 2021).
- **Frame**: The sampling frame is typically an extensive online panel of participants who have opted in to YouGov's surveys. The panel is not representative of the full population - it excludes both individuals who might not be able to access the internet and those who choose not to take part in online surveys. In order to mitigate this, YouGov maintains a large and diverse panel across different demographics. Despite this, the fact that the recruitment is non-random means that some groups, such as younger, more internet-savvy individuals, may be over-represented (YouGov 2021; Pew Research Center 2020).
- **Sample**: YouGov conducts sample selection based on demographic quotas such as age, gender, race, and region. After the data has been collected, post-stratification weighting adjusts for imbalances. While this helps in aligning the sample to the population, it remains a non-random sample, which can introduce biases, particularly concerning less engaged or harder-to-reach populations (Baker et al. 2013; YouGov 2021).

A.2 Sample Recruitment

The panel for YouGov's polls is recruited through voluntary participation. There are several digital outreach campaigns, including targeted online ads and partnerships with websites. This recruitment method allows for rapid sample collection, and thus is convenient and inexpensive, but introduces **self-selection bias**—people who opt-in for online surveys may differ from those who do not. For example, those more politically engaged or who spend more time online might be over-represented in the panel (Baker et al. 2013).

Additionally, since the recruitment methods rely on internet access, by definition people without reliable internet access or technological proficiency end up being excluded. These groups are usually older adults and lower-income individuals, and thus these segments are potentially under-represented. This is addressed to an extent by quota sampling and post-stratification weighting address, but concerns about representativeness remain (Pew Research Center 2020; YouGov 2021).

Despite this, the size of YouGov's panel allows them to collect a broad and diverse set of responses quickly. Continuous sampling enables YouGov to access and maintain a broad demographic scope and conduct frequent surveys, and this is particularly useful for tracking political opinions over time (YouGov 2021).

A.3 Sampling Approach and Trade-offs

YouGov uses **quota sampling**. In quota-sampling, the sample is drawn based on specific demographic characteristics like age, gender, race, and geography. These quotas ensure that the sample matches the population in key aspects. Once data is collected, **post-stratification weighting** is applied, which further balances the sample (Baker et al. 2013).

A.3.1 Strengths

The key strength that quota sampling provides is allowing YouGov to collect large, diverse samples quickly and at a lower cost than traditional random sampling. This approach also ensures that under-represented groups (like younger voters or minorities) are included in the sample, which is not always possible through random sampling, especially in phone-based polling (YouGov 2021).

A.3.2 Weaknesses

Since the sample is not drawn randomly, there are concerns over the degree of representativeness for the population. Non-random samples may over-represent certain groups, specifically in thie case internet-savvy individuals or those more engaged in politics have a higher likelihood of responding. The addition of post-stratification weighting helps balance the sample, however the accuracy of the results depends heavily on how well the quotas and weights align with the population. If the quotas are off, the results can be skewed (Baker et al., 2013; Pew Research Center, 2020).

A.4 Non-response Handling

Non-response occurs when certain groups are less likely to respond to surveys, which can introduce **non-response bias**. To address this, YouGov uses **weighting**. After collecting responses, statistical weightsare applied to ensure that the final sample reflects the population's demographic makeup to a high degree(YouGov 2021).

In random-digit-dial (RDD) phone surveys, non-response rates are often high, sometimes exceeding 80%. By contrast, YouGov's online panel experiences lower non-response rates because participants have pre-registered, and the presence of that anchor makes them more likely to engage (Pew Research Center 2020).

Partial correction for this can be achieved by weighting, which corrects for this by assigning more weight to under-represented groups. For instance, if older voters are under-sampled, their responses are given greater weight to reflect the overall population. While this helps reduce non-response bias, the accuracy of the poll is still subject to how effectively these weights adjust for missing data (Baker et al. 2013).

A.5 Questionnaire Design

A.5.1 Strengths

YouGov uses standardized and pre-tested questions to maintain a high level of clarity and consistency across surveys and reducing the chances of confusing or biased questions. Often, they will employ a technique called **question batteries**, where respondents answer multiple related questions on the same topic. This allows for deeper insights into voters' preferences and opinions than can be achieved by individual unrelated questions (YouGov 2021).

Additionally, YouGov implements randomization on the order of both questions and answer choices to minimize **order effects**—where the order in which questions or options are presented can influence how respondents answer. The use of visual aids, such as images of political candidates, helps ensure respondents correctly identify key figures (Pew Research Center 2020).

A.5.2 Weaknesses

One drawback of YouGov's methodology is **survey fatigue**, especially since there are often multiple related questions on the same topic. Respondents participating in long online surveys may lose focus or rush through the latter parts, and this potentially reduces the quality of their answers. This is especially problematic for questions placed toward the end of the survey. In addition, because YouGov surveys are conducted entirely online, they may exclude individuals who prefer other polling methods, such as phone or face-to-face interviews, or who struggle with technology (Baker et al. 2013; YouGov 2021).

A.6 Key Features

- **Post-stratification weighting** is crucial for aligning the sample with the population's demographics. The weights ensure that groups who are under-represented in the sample (such as older voters) are appropriately reflected in the final results (YouGov 2021).
- **Dynamic sampling** enables YouGov to adjust their sample in real-time to target specific groups. As an example, if young voters are under-represented in an initial sample, YouGov can recruit additional respondents from that group to balance the sample (YouGov 2021).

B Appendix: Idealized Methodology and Survey for Forecasting the U.S. Presidential Election

B.1 Overview and Budget Allocation

In this appendix, we detail methodology for forecasting the U.S. presidential election using a budget of \$100,000. The methodology covers sampling, respondent recruitment, data validation, and poll aggregation. The survey will be implemented using Google Forms, and a link to the survey is added below.

Budget Allocation:

- **Sampling and Recruitment**: \$60,000
- **Survey Platform and Technology**: \$5,000
- **Data Cleaning and Validation**: \$10,000
- **Poll Aggregation Software**: \$10,000
- **Reporting and Analysis**: \$15,000

B.2 Sampling Approach

B.2.1 Stratified Random Sampling

A **stratified random sampling** method is adopted. This ensures representativeness across demographic subgroups. Groups are defined by **age**, **gender**, **race/ethnicity**, **education level**, and **geography**. For example, **18-29-year-olds** represent 20% of the U.S. population, so 20% of the sample will be drawn from this age group.

Justification: Stratified random sampling addresses heterogeneity in voter behavior and ensures that our sample is representative of the voting-age population. That approach is especially useful in capturing variation within battlefield states where minor variations in voter turnout could tip the election in a big way (Gelman and Hill 2007).

A **sample size of 5,000 respondents** provides a **margin of error of ±1.5%** and a **confidence level of 95%**.

B.2.1.1 Over-Sampling

Underrepresented groups, such as rural populations and racial minorities, will be **oversampled** to ensure adequate representation. After this, adjustments will be made using **post-stratification weighting** to align with U.S. Census data.

B.2.2 Recruitment Strategy

Respondents will be recruited via:

- **MTurk**, **Prolific**, and **YouGov** panels to access diverse respondents.Studies show that online recruitment platforms may produce quality survey data when sampling and weighting are done properly (Huff and Tingley 2015).
- **Targeted online ads** (Facebook, Instagram, etc.) - this allows us reach key demographic groups, ensuring representation from rural, suburban, and urban areas. Ads will aim for a **response rate of 70%** to ensure data reliability.

B.3 Survey Design

B.3.1 Survey Structure

The intention of the survey is to capture voter sentiment, behavior, and key issues. Estimated **completion time** is under **10 minutes**, which allows us to target a **completion rate of 80%**.

B.3.2 Sections:

1. **Screening Questions**:
   - U.S. citizenship and voter registration status.
   - Intention to vote.
2. **Demographics**:
   - Questions cover **age, gender, race/ethnicity, education, income level, and region**.
3. **Voting Intentions**:
   - "If the election were held today, who would you vote for?"
     - **Kamala Harris** (Democratic)
     - **Donald Trump** (Republican)
     - **Undecided**
   - Likelihood of voting rated on a **scale of 1-10**.
4. **Issues Important to Voting**:

- Key issues (e.g., **Economy, Healthcare, Immigration, Climate Change**), ranked on a **scale of 1-5**.

5. **Candidate Favorability**:

   - Favorability ratings for **Kamala Harris** and **Donald Trump**, rated on a **5-point scale**.

6. **Political Engagement**:

   - This section will include questions on how closely respondents follow election news.

7. **Open-ended Questions**:

   - We end by asking open-ended questions around What factors the repsondents think most influence the voting decision?

B.3.3 Data Quality Checks

The survey includes **attention-check questions** and tracks **response time** to flag low-quality responses. In ordet to reduce noise due to bots filling these up quickly, responses completed in less than **3 minutes** will be excluded.

B.4 Data Validation and Post-Survey Processing

B.4.1 Data Cleaning

- **Duplicate response filtering** using IP addresses.
- **Post-stratification weighting** - this will be used to adjust the sample to match U.S. population proportions based on gender, race, and age (Pasek 2015).

B.4.2 Poll Aggregation

Results from the survey will be aggregated with other polls (e.g., **Gallup**, **Ipsos**, **YouGov**) using **weighted averages**. **Bayesian updating** will incorporate historical trends and account for shifts in voter behavior (Jackman 2009). The aggregated margin of error is projected at **±1%**, ensuring high forecast precision.

B.5 Survey Implementation

A sample version of the survey has been created using Google Forms. It can be accessed here:

Google Forms: U.S. Presidential Election Forecast Survey

B.6 Survey Copy

B.6.1 U.S. Presidential Election Forecast Survey

1. **Are you a U.S. citizen?**
   - Yes
   - No
2. **Are you registered to vote in the upcoming election?**
   - Yes
   - No
3. **Do you intend to vote in the upcoming election?**
   - Yes
   - No
   - Unsure
4. **What is your age?**
   - 18-29
   - 30-44
   - 45-64
   - 65+
5. **What is your gender?**
   - Male
   - Female
   - Non-binary
   - Prefer not to say
6. **What is your race/ethnicity?**
   - White
   - Black or African American

   - Hispanic or Latino
   - Asian
   - Other

7. **What is your highest level of education?**

   - High school or less
   - Some college
   - College degree
   - Postgraduate degree

8. **If the election were held today, who would you vote for?**

   - **Kamala Harris** (Democratic)
   - **Donald Trump** (Republican)
   - Undecided

9. **How important is the issue of the economy in your decision?**

   - Please describe why the economy is important to your voting decision:

10. **Rank the following issues in order of importance for your voting decision**

- Economy
- Healthcare
- Immigration
- Climate Change
- Education

11. **How favorable is your opinion of Kamala Harris?**

    - On a scale of 1 to 10, where 1 is very unfavorable and 10 is very favorable, please rate your opinion of Kamala Harris:

12. **How favorable is your opinion of Donald Trump?**

    - On a scale of 1 to 10, where 1 is very unfavorable and 10 is very favorable, please rate your opinion of Donald Trump:

13. **How closely do you follow election news?**

    - Please explain how often and through which sources you follow election news:

14. **What factors most influence your voting decision?**

    - Please provide details about the factors that are most important to you when deciding whom to vote for:

References

Baker, Reginald, J Michael Brick, Nancy A Bates, Michael Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau. 2013. "Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1 (2): 90–143.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science & Politics* 47 (02): 297–300. https://doi.org/10.1017/s1049096514000055.

FiveThirtyEight. 2024. "Polling Aggregates - 2024 Presidential Election." https://projects.fivethirtyeight.com/polls/.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Huff, Connor, and Dustin Tingley. 2015. "'Who Are These People?' Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2 (3): 1–12. https://doi.org/10.1177/2053168015604648.

Iannone, Richard et al. 2023. *Gt: Create Presentation-Ready Tables*. https://CRAN.R-project.org/package=gt.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Hoboken, NJ: Wiley.

Pasek, Josh. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *Public Opinion Quarterly* 79 (2): 594–619. https://doi.org/10.1093/poq/nfu060.

Pew Research Center. 2020. "Assessing the Risks to Online Polling." https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polling/.

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 et al. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley et al. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

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

YouGov. 2021. "Public Polling Methodology." https://yougov.co.uk/about/panel-methodology/.