Tight battle between republicans and democrats with Kamala in the lead*

Pollsters in America are suggesting that Americans prefer Kamala as president

Chris Yong Hong Sen Veyasan Ragulan Prankit Bhardwaj

November 5, 2024

This paper forecasts the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump using linear regression models applied to recent polling data. The analysis projects a win for Harris, who leads Trump by an estimated 5 percentage points on average, with consistent support across most polling organizations and swing states. The models underscore the impact of pollster bias and regional variability on election outcomes. Future research should integrate non-linear modeling and Bayesian methods to improve predictive accuracy.

1 Introduction

In this paper we look at polling data from 538 on the 2024 US Presidential election, specifically on Demoncratic Candidate Kamala Harris. We will construct a model that will be used to make a prediction on whether Harris wins the US Presidency

Section 2 will outline the source of this data. Section 3 covers the model and it's parameters. Section 4 is where discussion will be made about the models predictions and how realistically they line up with current affairs. Finally, section 4 discusses any weakness and limitations that can be considered for another report.

This report was written with the assistance of R R Core Team (2023). Additionally, Wickham et al. (2019), and Goodrich et al. (2022) were used to clean the data, write models, and analyse results. The structure of this report heeds Rohan Alexander's example Alexander (2023).

*Code and data are available at: https://github.com/Monoji77/USA-pollster.

2 Data

The dataset provided by fivethirtyeight (Ryan Best 2024) contains over 15000 observations across 50 variables. Each observation is a poll conducted on the 2024 US Presidential Election. Fivethirtyeight's dataset was chosen due to it's comprehensive review of polls and pollsters, attributing grades in error and bias, as well as in transparency. Their thorough investigations ensure their dataset not only contains insightful polls, but also includes as much relevant information about the polls conducted as possible. This gives us plenty of predictors when building a model that will caluclate the percentage win of either Trump or Harris.

The dataset was read into R (R Core Team 2023), and cleaned using the R package tidyverse (Wickham et al. 2019). Due to Kamal's late entry into the race, some polls do not have any data on her specifically. This is why we set a cut-off date at 2024/07/21, the day Kamala replaced Joe Biden as the Democratic nominee.

Fivethirtyeight includes a metric called `numeric_grade`, which combines the transparency_score and transparency score. We only included polls with a `numeric_grade` at or above 2.5. This gave us the best balance between quantity and quality. Finally, we create 2 datasets, one with polls predicting a Harris win, the other a Trump win.

2.1 Predictors

2.1.1 `transparency_score`

Fivethirtyeight awards a score out of 10 to each pollster based on how informative their methodology is. Understanding the methodology of a pollster assits in understanding any errors or biases present in their findings.

According to Table 1 and Table 2, the difference between candidates is very minimal, with the minimum, maximum, and even average being the same (only in the media do we see a slight lead by Trump). Both Figure 1 & Figure 2 are similar as well, with most polls achieving a 9.0 transparency score.

Table 1: Summary of transparency_scores amongst polls predicting a Kamala Harris win. The closer the values are to 10, the more trustworthy the predictions.

Min	Mean	Median	Max
4	8.494875	9	10

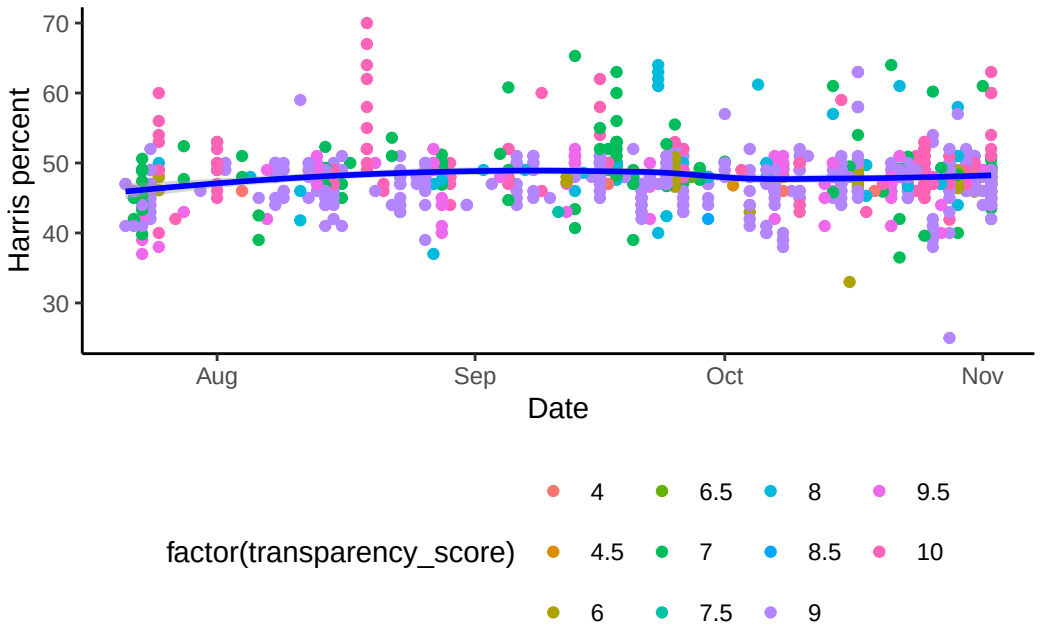


Figure 1: As election day draws close, pollsters are wrapping up their polls and posting their predictions on who will win. The polls in this graph predict Harris' win percentage, with the average in the blue line. The dots in the graph below represent individual polls conducted in the months leading up to the election. They are coloured based on their transparency_score, which is fivethirtyeight's metric for transpareny in the pollster's methodology.

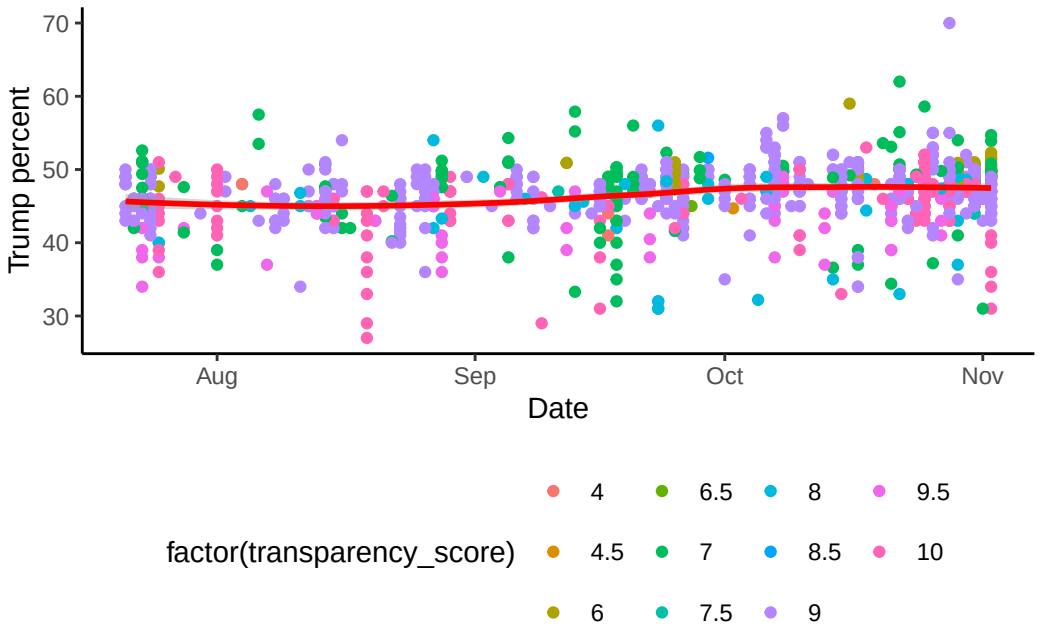


Figure 2: As election day draws close, pollsters are wrapping up their polls and posting their predictions on who will win. The polls in this graph predict Trump's win percentage, with the average in the blue line. The dots in the graph below represent individual polls conducted in the months leading up to the election. They are coloured based on their transparency_score, which is fivethirtyeight's metric for transpareny in the pollster's methodology.

Table 2: Summary of transparency_scores amongst polls predicting a Donald Trump win. The closer the values are to 10, the more trustworthy the predictions.

Min	Mean	Median	Max
4	8.506674	9	10

2.1.2 `numeric_grade`

A cumulative score given to pollster that combines fivethirtyeight's `pollscore` and `transparency_score`. Like `transparency_score`, we believe this metric can give us insight into potential biases in each pollster and poll, which is why it is included in our model.

Similar to `transparency_score`, Table 3 & Table 4 are identical except for the median, where Trump again holds a slight advantage. Similarly, Figure 3 & Figure 4 are almost similar, with a high number of 3's for numeric_grade. This makes sense, as `numeric_grade` is a combination of `transparency_score` and `pollscore`, so any patterns from the latter will apply to the former.

Table 3: Summary of numeric_grades amongst polls predicting a Kamala Harris win.

Min	Mean	Median	Max
2.5	2.853645	2.9	3

Table 4: Summary of transparency_scores amongst polls predicting a Donald Trump win. The closer the values are to 10, the more trustworthy the predictions.

Min	Mean	Median	Max
2.5	2.852836	2.9	3

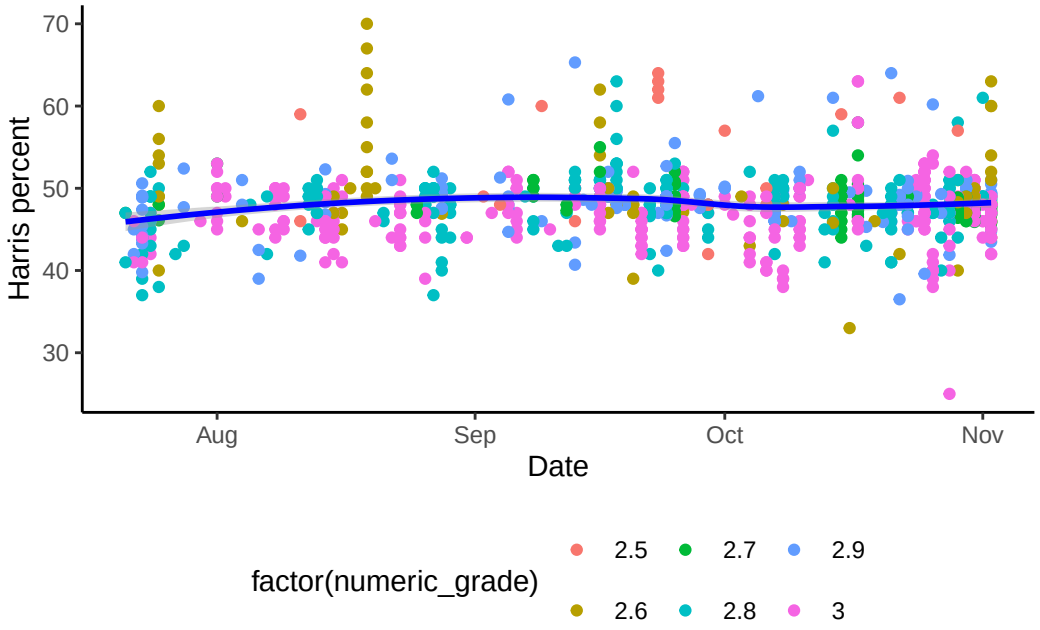


Figure 3: As election day draws close, pollsters are wrapping up their polls and posting their predictions on who will win. The polls in this graph predict Harris' win percentage, with the average in the blue line. The dots in the graph below represent individual polls conducted in the months leading up to the election. They are coloured based on their numeric_grade, which is fivethirtyeight's combined metric on transparency and bias/errors.

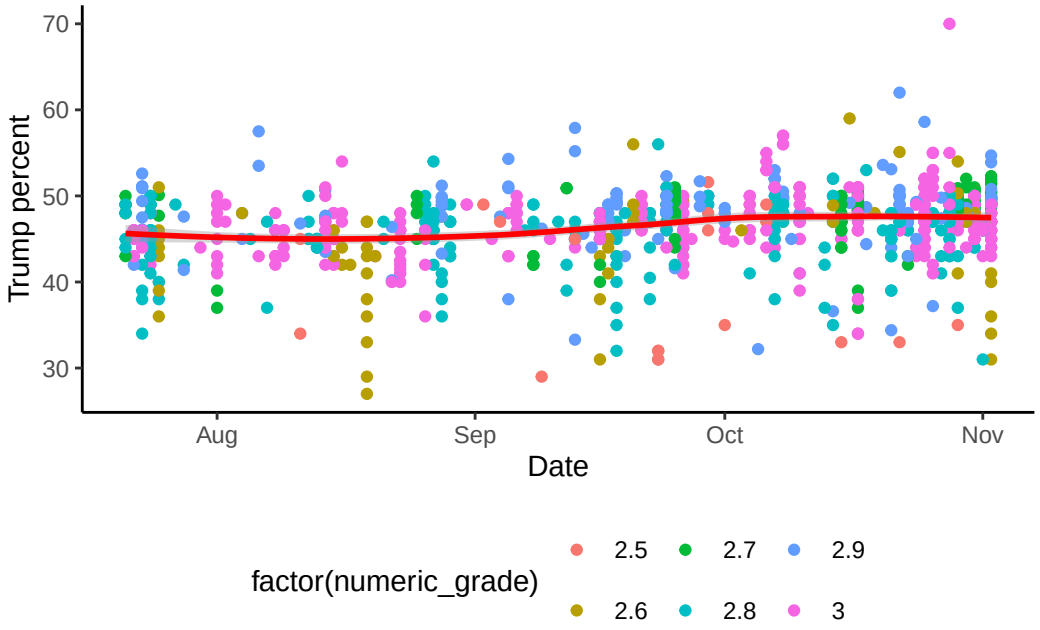


Figure 4: As election day draws close, pollsters are wrapping up their polls and posting their predictions on who will win. The polls in this graph predict Trumps' win percentage, with the average in the blue line. The dots in the graph below represent individual polls conducted in the months leading up to the election. They are coloured based on their numeric_grade, which is fivethirtyeight's combined metric on transparency and bias/errors.

2.1.3 `pollster`

Each poll is conducted by a pollster, who may make multiple polls leading up to the election. This may increase accuracy of their predictions, which is why it will be in our model.

Figure 5 & Figure 6 show an interesting bump in predictions by George Mass. This may have been an error on their part however, as it is almost immediately rectified.

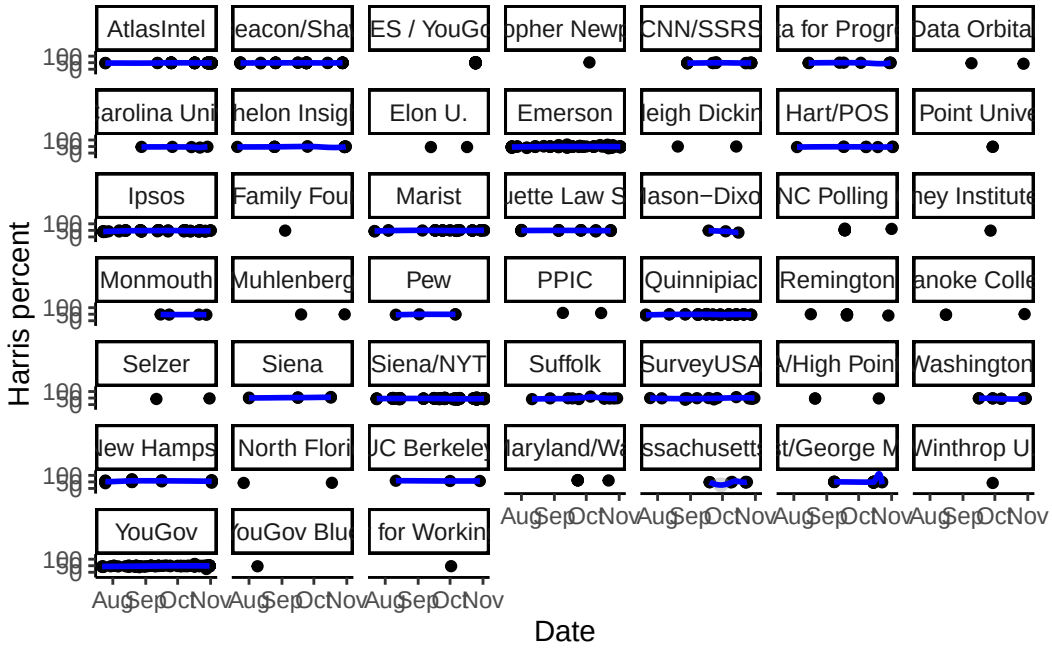


Figure 5: As election day draws close, pollsters are wrapping up their polls and posting their predictions on who will win. Each graph shows individual polls conducted in the months leading up to the election. They represent the change in prediciton for Kamala Harris winning the presidency.

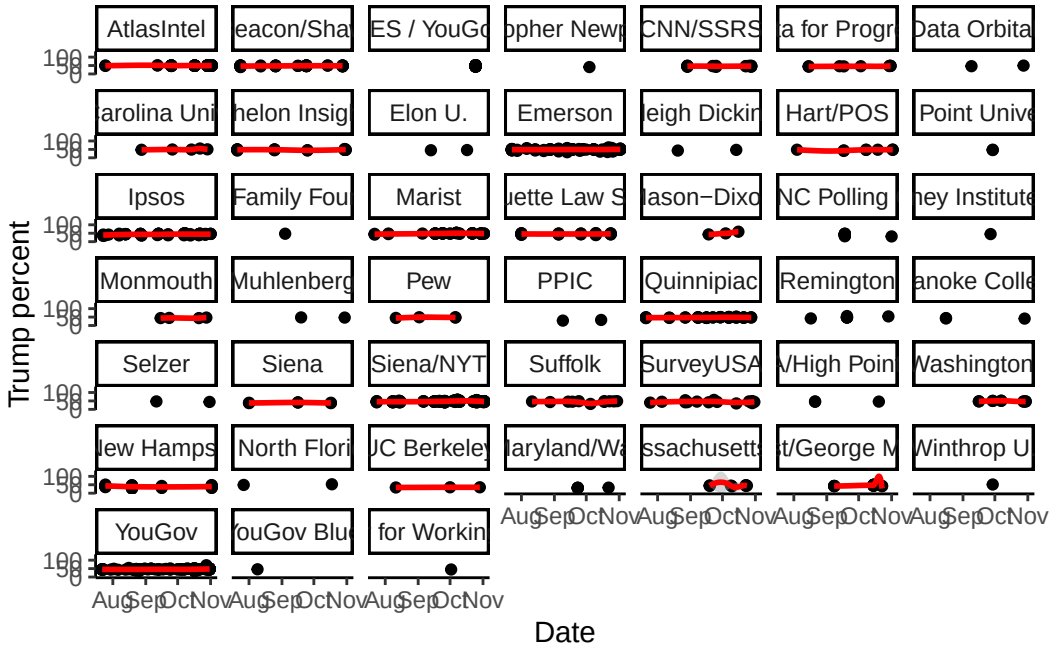


Figure 6: As election day draws close, pollsters are wrapping up their polls and posting their predictions on who will win. Each graph shows individual polls conducted in the months leading up to the election. They represent the change in prediciton for Donald Trump winning the presidency.

2.1.4 `states`

The states each poll was conducted in. This is not one of our predictors for our model, but it can provide insight into what the prediction will be.

Figure 7 has Harris win 4 t0 7 battleground states, but the margins are increadibly close and all it would take is for one of them to be incorrect to reverse the prediction.

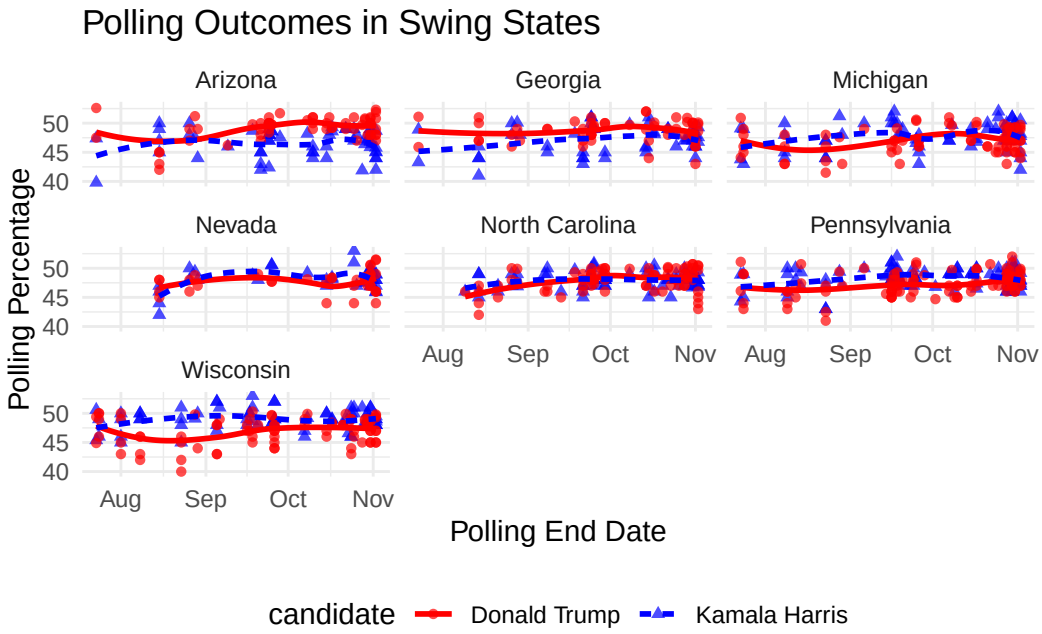


Figure 7: Battleground states are what usually determines the winner of a presidential election. This figure outlines the polling outcomes of each candidate from August to November, for each of the seven battleground states this election cycle.

2.1.5 `end_date`

Every poll has a start and end date for querying its participants. Polls that are conducted closer to election day may be more reflective of American opinions as opposed to polls conducted in August or September. This is why we will be using this variable as one of our predictors in the model.

Both Figure 8 & Figure 9 show a large volume of polls ending at the beginning of November, which would be the most accurate findiings leading up to the election.

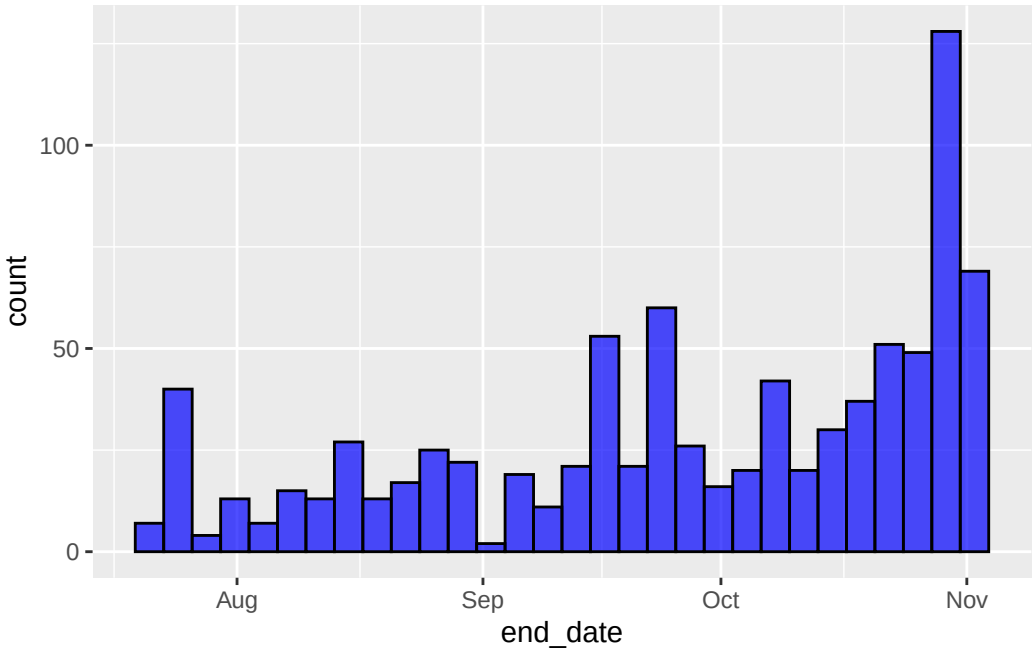


Figure 8: Polls are conducted from a start date to an end date. The closer the end date is to election day, the more insightful the results are to analysts and the general public. This distribution shows the end date of polls predicting a Kamala Harris win.

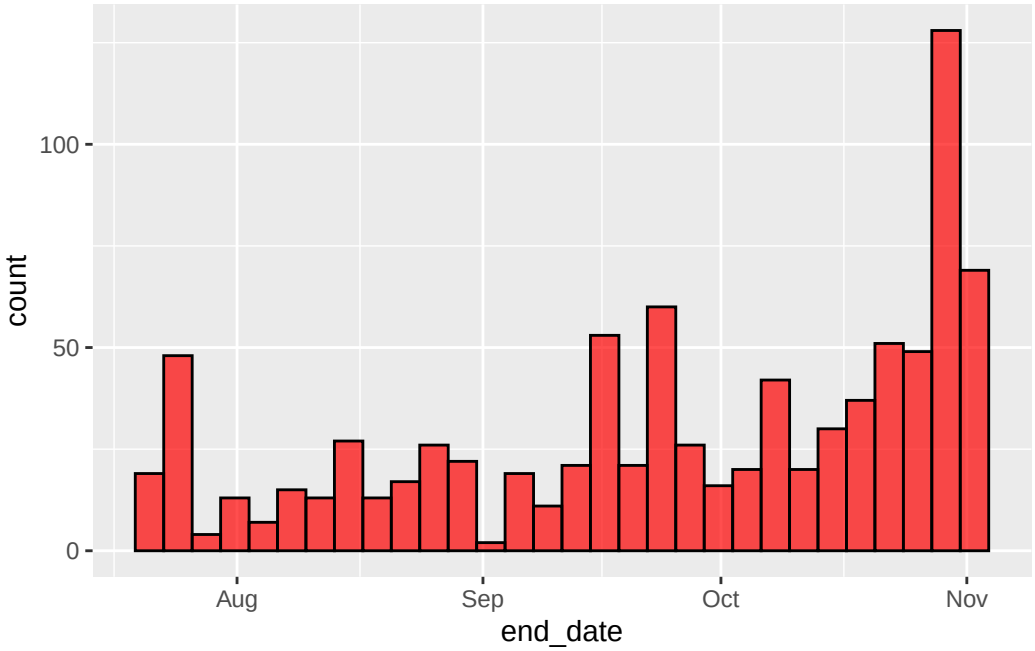


Figure 9: Polls are conducted from a start date to an end date. The closer the end date is to election day, the more insightful the results are to analysts and the general public. This distribution shows the end date of polls predicting a Donald Trump win.

2.2 Response

2.2.1 `pct`

This the the predicted percentage each candidate will get on election day. We have assumed a two-party system for our report, as any third-party candidates would have a negligible impact on election day.

Figure 10 & Figure 11 are centered just below 50% indicating a very close match. Table 5 & Table 6 show Harris leading in Mean and Median percentages however, ever so slightly.

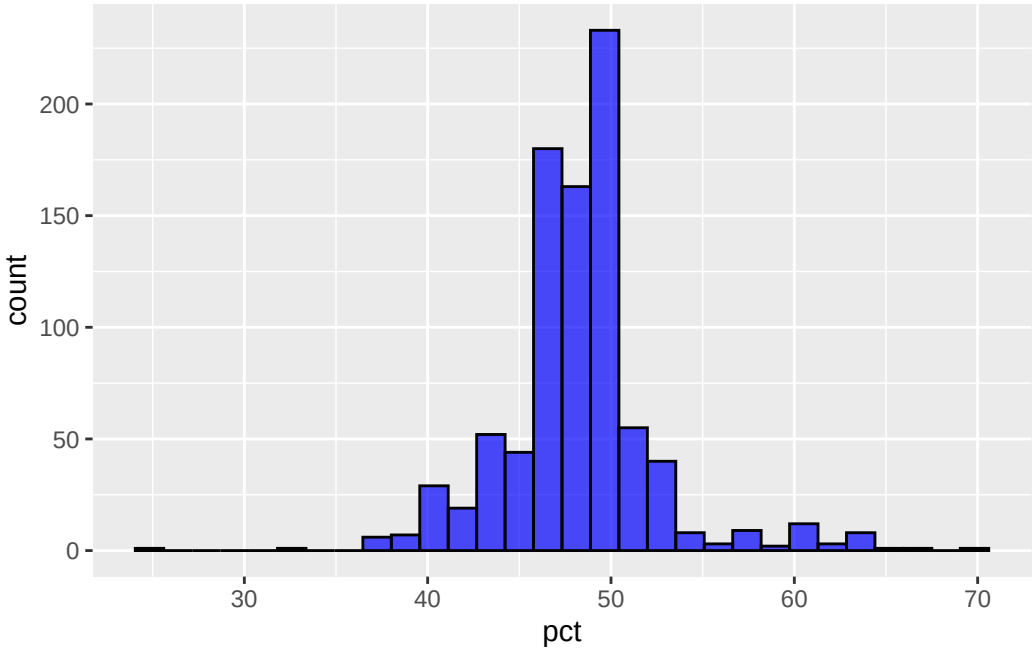


Figure 10: The distribution of predicted Harris vote percentage on election day.

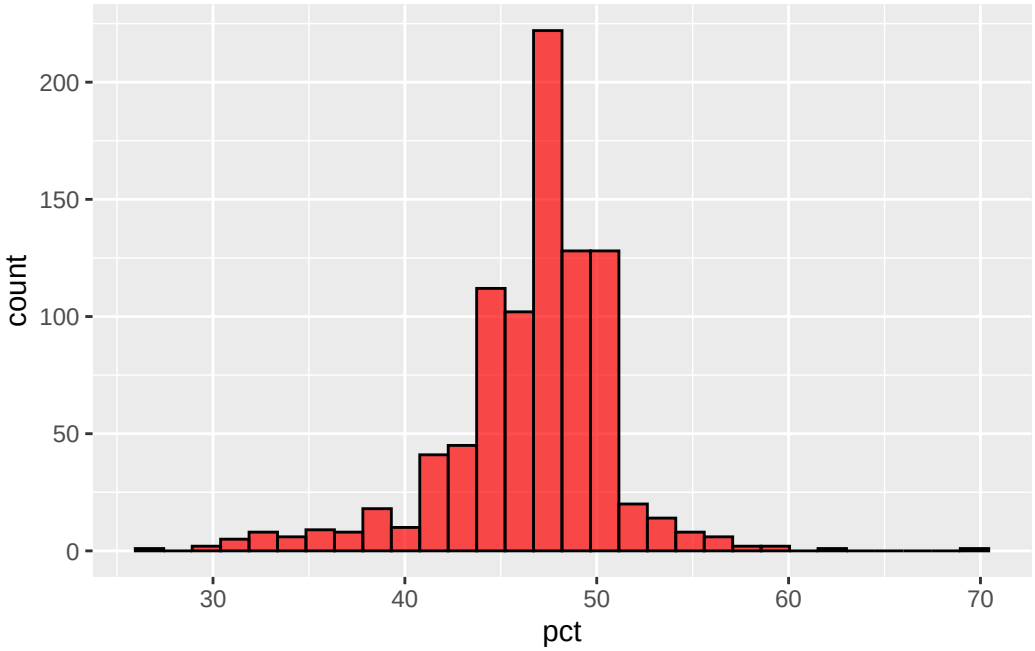


Figure 11: The distribution of predicted Trump vote percentage on election day.

Table 5: Summary of percentage vote predictions for Kamala Harris.

Min	Mean	Median	Max
25	48.14033	48	70

Table 6: Summary of percentage vote predictions for Donald Trump.

Min	Mean	Median	Max
27	46.6457	47	70

3 Measurement

The dataset comes from fivethirtyeight.com, a statistical analysis website dedicated to US sports and elections (Ryan Best 2024). Fivethirtyeight takes polls conducted on a particular topic (in this case the 2024 US presidency), and aggregates them in one dataset.

Election polls are used to gauge public opinion about candidates, political topics, voter engagement, and public opinion of the election process. Election polls also serve as key insight

to political organisations representing candidates, polls can show potential weak points across demographics.

Election polls are similarly to regular polls, they involve asking a sample population a series of questions via a delivery method. The key distinction is that election polls typically try to target *voters*, or people who are most likely to vote in the election cycle being investigated. This is a key challenge to pollsters, as there are many subgroups of voters, such as people who seldom vote or never vote. According to Pew Research, a third of US citizens did not vote in the 2020 election, despite the 2020 election cycle having high voter turnout (Keeter 2024). There may also be cases when respondents of a poll reconsider whether they will vote come election day.

Election polling has exploded in recent times according to Pew Research. They suggest that double there are double the amount of active pollsters in 2022 compared to 2000 (Keeter 2024). One reason mentioned is the advent of online polling, which is simple, cheap, and can easily reach US citizens across the country, compared to traditional methods such as poll by phone and mail. As a result, while the diversity of pollsters has increased dramatically, the variance in results methodology, and ethics have also increased

Errors in election polling have been an issue in the last two cycles of presidential elections. Between the 2016 and 2020 elections, there has been a trend of pollsters not fully capturing Trump's support. For example in 2020, despite Biden winning the election, there were reports from the American Associcate for Public Opinion Research that national polls overstated his victory over Trump by 3.9 percentage points (Keeter 2024).

Fivethirty eight mitigates the influx of pollsters and the potential for errors by routinely checking up on pollsters as they produce reports. Fivethirtyeight checks every pollster, regardless of experience or prestige. There are 2 types of checks fivethirtyeight uses: Methodology and Ethics.

Methodology refers to the scientific rigor exhibited by the pollster, when conducting the poll, and in the presentation of the results. Fivethirtyeight requires each poll to have or easily obtain the following (Radcliffe and Morris 2023):

- Pollster Identity
- Survey Dates
- Population Sample (Size and other distinct attributes)
- Polling Method (How was the poll conducted?)
- Sponsors (Identity and Amount

Even if these conditions are met, fivethirtyeight may refuse to include certain polls in their data, due to flaws in methodology. Some examples include (Radcliffe and Morris 2023):

- Polls with inappropriate sample for target population

- Polls based upon predictive models
- Polls that implement recontact with participants (potential for bias to creep into the sample)
- Polls done by amateur/non-professional pollsters
- Polls with an "informed ballot" (when information about a candidate is given to the participant before they are asked who they would vote for)

The second check fivethirtyeight implements is ethical standard, adapted from the American Association for Public Opinion Research's Code of Professional Ethics and Practices . Fivethirtyeight may refuse to include polls that (Radcliffe and Morris 2023):

- Fabricate or falsify data
- Are associated with the betting industry
- Refuse to disclose their methodology, either in their findings, or after being contacted for clarification
- Misrepresent their true purpose (being part of campaign analysis or done for a particular party or candidate)
- Disclose errors that come about from their work, and rectify them as best as possible
- Utilize methods that will give misleading results

Fivethirtyeight will conduct an initial check on pollsters, seeing if their methodology and ethics are publicly available, and are in agreement with fivethirtyeight's checks. If unclear, fivethirtyeight will do a more thorough investigation of the pollster, and ask the organizers questions about their methodology and ethics if none were visible in the public report. If there are any violations in ethics, the severity on consequences will vary on what the offence is. If there is evidence of falsifying data or engaging in betting markets, the pollster is blacklisted from their data. Otherwise, pollsters are given a chance to fix ethical issues presented by fivethirtyeight. Failure to do so results in the pollster and any of its polls being taken off the dataset, but fivethirtyeight can reverse this decision if the pollster demonstrates satifactory ethics and methodology practices(Radcliffe and Morris 2023).

Partisan polls are a small subset of polls that are made with the backing or organizations affiliated with one or more political organizations. Fivethirtyeight will allow these polls into their data, but they will be marked as "Partisan", and extra checks are put in place to ensure the polling data hasn't been influenced unethically by the pollster or it's affiliated organizations (Radcliffe and Morris 2023).

4 Model

Our modeling strategy estimates the polling support percentages for Kamala Harris and Donald Trump during the 2024 US election campaign. To achieve this, we constructed two linear models for each candidate, aimed at understanding how their support evolves over time while accounting for variations between pollsters. The goal is to balance simplicity and interpretability, while capturing key variations in polling trends.

Specifically, we consider two types of model for each candidate individually:

- **Model 1**: A multiple linear regression model predicting polling percentage (pct) as a function of end_date (the date when polling concluded), transparency_score (an indicator of how transparent and reliable a poll is), and numeric_grade (an independent quality rating for pollsters).
- **Model 2**: An extension of Model 1 that additionally includes the pollster information (pollster) to account for variability introduced by different polling organizations.

The models are built for both Kamala Harris and Donald Trump to facilitate a comparative analysis, aiming to understand their respective polling dynamics and how these vary by time, pollster quality, and polling organization.

Model 1: Linear Model by Date, Transparency Score, and Grade

$$y_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{transparency\_score}_i + \beta_3 \cdot \text{numeric\_grade}_i + \epsilon_i \quad (1)$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2) \quad (2)$$

where:

- y_i represents the polling percentage of support for the candidate in poll i.
- $beta_0$ is the intercept, representing the baseline support when all predictors are at their baseline levels.
- $beta_1$, β_2, β_3 represent the coefficients for the predictors: end_date, transparency_score, and numeric_grade, respectively.
- $epsilon_i$ is the error term, which is assumed to be normally distributed with variance σ^2.

This model helps capture the general trend in polling support over time while adjusting for the quality of the poll itself.

Model 2: Linear Model by Date and Pollster

$$y_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \gamma_{p[i]} + \epsilon_i \quad (3)$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2) \quad (4)$$

where:

- $\gamma_{p[i]}$ represents the fixed effect for pollster p conducting poll i, which captures variability across different polling organizations.
- All other terms are as defined in Model 1.

This model extends the first by considering pollster-specific differences in polling results. Different pollsters may have different methodologies or biases, which are accounted for by including pollster as a categorical variable.

Model Predictions and Augmentation

The models were then used to predict polling percentages, which were added to the original datasets as new columns: - For Kamala Harris, the predicted values (fitted_date and fitted_date_pollster) were added to the just_harris_high_quality dataset. - For Donald Trump, the analogous predicted values (fitted_date and fitted_date_pollster) were added to the just_trump_high_quality dataset.

These augmented datasets allowed for a comprehensive comparison of predicted versus observed polling percentages, and facilitated visualization of how support for each candidate evolved over time.

4.1 Model set-up

To model the polling percentages for Kamala Harris and Donald Trump, we used multiple linear regression to estimate the relationship between polling percentages (pct) and several predictors, including the poll's end date, pollster quality, and pollster organization. This section provides a detailed breakdown of how we constructed the models, including the selection of predictors and the data processing steps taken to ensure that the models capture meaningful patterns.

4.2 Data Preparation

The data used for modeling included high-quality election polling data for both Kamala Harris and Donald Trump. The data was filtered to include only polls that met a quality threshold (numeric_grade >= 2.5), ensuring that the analyses are based on reliable information. This threshold was chosen based on an examination of pollster ratings, and it helped exclude lower-quality polls that might introduce noise into the analysis.

Additionally, any missing values for state were replaced with "National" to indicate that those polls were not specific to any state. This allowed us to effectively categorize national polls alongside state-level polls for the purposes of modeling.

The key variables used in the models were:

- **end_date:** The date on which polling concluded. This variable was treated as a continuous measure of time, capturing trends in polling over the campaign period.

- **transparency_score:** A score reflecting the transparency of the pollster. This score was included to adjust for the quality of information provided by the polling organization.
- **numeric_grade:** A numeric score assigned to the pollster to indicate overall quality. Higher scores indicate higher pollster reliability.
- **pollster:** The polling organization conducting the poll. This categorical variable was included in Model 2 to account for differences between pollsters in methodology, sampling, or potential biases.

4.3 Feature Selection

We used the following predictors to model the polling percentage (pct) for both candidates:

1.Temporal Trend (end_date): - The end date of polling was included to capture changes in candidate support over time. Including time-based effects helped account for general campaign dynamics, such as events that might influence voter opinion.

2.Pollster Quality Indicators (transparency_score and numeric_grade): - These two variables represent measures of poll quality. The inclusion of transparency_score and numeric_grade allowed us to adjust for potential biases that could result from low-quality polling data. By incorporating these scores, the model effectively accounts for the fact that polls with higher transparency and quality are more reliable.

3.Pollster Organization (pollster): - In Model 2, the pollster variable was included as a categorical factor. This allows us to control for differences among polling organizations, which may have distinct methodologies, levels of accuracy, or biases. For instance, some pollsters might consistently report higher or lower percentages for specific candidates, and including this factor helps mitigate those effects.

4.4 Model Training and Prediction

- Model 1 was fitted to predict pct as a function of end_date, transparency_score, and numeric_grade. This model helps in understanding the general trends in candidate support, while adjusting for the quality of the polls included.
- Model 2 added pollster as an additional predictor, creating a richer model that accounts for differences between polling agencies. The inclusion of pollster captures variability related to different polling methodologies or inherent pollster biases.
- Both models were trained using the lm() function in R, which estimates the parameters by minimizing the sum of squared residuals. The models were fit separately for Kamala Harris and Donald Trump to allow for a detailed comparison of their respective polling trends.
- After fitting the models, predictions were made and added to the dataset for each candidate. This allowed us to compare actual polling percentages with those predicted by the models. Specifically:

- Model 1 Predictions (fitted_date) captured the predicted polling percentage based on date and poll quality.
- Model 2 Predictions (fitted_date_pollster) additionally included the effect of each pollster, providing a more nuanced estimate that adjusts for pollster-level effects.

4.5 Model Assumptions

The following assumptions underlie the multiple linear regression models used:

- **Linearity:** We assume that the relationship between the predictors (end_date, transparency_score, numeric_grade, and pollster) and the outcome (pct) is linear.
- **Independence:** Each poll is treated as an independent observation. We assume that polling percentages from different polls are not influenced by one another.
- **Homoscedasticity:** The variance of the residuals is assumed to be constant across all levels of the independent variables.
- **Normality of Residuals:** The error term (_i) is assumed to be normally distributed, which is required for the validity of hypothesis testing.
- **No Multicollinearity:** We checked that the predictors are not highly correlated, to ensure that the coefficients estimated by the models are reliable and that multicollinearity does not bias the results.

These assumptions are crucial for ensuring that the model coefficients are unbiased and that the predictions are meaningful.

4.6 Cross-Validation and Overfitting Prevention

To evaluate the robustness of our models, a train-test split was applied. The training data was used to fit the model, while the test data (which the model had not seen during training) was used to evaluate model performance. This approach helps in identifying overfitting, ensuring that the model can generalize beyond the specific dataset used for training.

For future improvements, we could explore the use of cross-validation for further robustness checks, ensuring that the model's performance is not overly dependent on a particular train-test split.

4.6.1 Model justification

The choice of multiple linear regression models for predicting the polling support of Kamala Harris and Donald Trump is driven by several key considerations regarding the nature of the data and the objectives of the analysis. Below, we provide a detailed justification for the chosen models and predictors, discuss the reasoning behind the modeling choices, and address limitations and alternatives considered during the process.

4.7 Why Multiple Linear Regression?

Multiple linear regression (MLR) was selected as the modeling framework due to its ability to estimate relationships between multiple predictors and a continuous outcome variable—in this case, the percentage of voter support (pct). MLR provides a straightforward approach to quantify the impact of each predictor on polling percentages, allowing us to make inferences about the strength and direction of these relationships.

The primary motivations for using multiple linear regression in this context include:

1.Simplicity and Interpretability: - Linear models offer a high degree of interpretability. Each coefficient in the model provides a clear indication of the expected change in voter support given a one-unit change in the corresponding predictor, holding all other predictors constant. This is crucial for understanding how time, poll quality, and pollster contribute to support dynamics.

2.Ability to Control for Multiple Predictors: - MLR allows us to control for several influential factors simultaneously, such as time (end_date), pollster quality (transparency_score and numeric_grade), and the polling organization (pollster). By accounting for these variables, we can isolate the individual impact of each predictor and better understand the underlying factors that affect polling outcomes.

3.Comparison Between Candidates: - We built identical models for both Kamala Harris and Donald Trump, allowing for a side-by-side comparison of their polling trends over time. This consistency enables a clearer understanding of differences in support and trends across the two candidates.

4.8 Why These Specific Predictors?

1.End Date (end_date): - The end date of polling is an essential predictor, as it captures temporal effects. Polling percentages often change over the course of a campaign due to events, news, and other dynamic factors. Including end_date helps us understand how support evolves over time.

2.Transparency Score (transparency_score) and Numeric Grade (numeric_grade): - These variables were included as indicators of pollster quality. The transparency_score reflects how openly pollsters report their methods, and numeric_grade represents an independent rating of the pollster's overall quality. - Including these predictors helps ensure that we account for potential biases introduced by lower-quality polls. For example, a positive coefficient on transparency_score would suggest that higher transparency is associated with higher or more reliable polling percentages for a given candidate.

3.Pollster (pollster): - Pollster was included as a categorical variable in Model 2 to capture differences across polling organizations. Different pollsters have varying methodologies and biases, which can lead to systematic differences in reported polling percentages. - Including

pollster allows the model to adjust for these variations, thereby improving the accuracy of the predictions.

4.9 Justification for Model Complexity

- **Model 1** is kept relatively simple, using end_date, transparency_score, and numeric_grade as predictors. This model is useful for understanding the general trend in polling support while accounting for poll quality.
- **Model 2** adds complexity by including the pollster (pollster) as a categorical predictor. This accounts for pollster-specific effects and helps to improve model accuracy by recognizing that different pollsters may report systematically different results. Given that different polling organizations use distinct methodologies, including this variable ensures that biases are appropriately adjusted.

4.10 Limitations of the Models

1.Assumptions of Linearity and Normality: - The linear models assume that the relationship between predictors and the outcome is linear, and that the residuals are normally distributed. In reality, polling data may not always satisfy these assumptions, especially if there are non-linear trends in voter support or heavy-tailed distributions in the errors.

2.Potential for Omitted Variable Bias: - While our models include several important predictors, there may still be unobserved factors influencing voter support that are not captured in the models. Examples include specific campaign events, candidate debates, or sudden shifts in voter sentiment.

3.Multicollinearity: - The predictors transparency_score and numeric_grade both relate to poll quality and may be correlated. This could introduce multicollinearity, which makes it difficult to determine the individual effect of each predictor. However, we inspected the variance inflation factors (VIFs) to ensure that multicollinearity was not excessively high.

4.Non-Independence of Observations: - Polls conducted by the same pollster over time might not be entirely independent, leading to autocorrelation in the data. While Model 2 partially addresses this by including pollster as a fixed effect, a more complex model (e.g., hierarchical or Bayesian) could better capture these dependencies.

4.11 Alternatives Considered

1.Logistic Regression: - Logistic regression was considered as an alternative, but it was ultimately deemed unsuitable because the outcome variable (pct) is continuous, rather than binary. Logistic regression is appropriate for classification tasks where the outcome is binary

(e.g., win or lose), but our goal is to predict support percentages, making multiple linear regression a more appropriate choice.

2.Bayesian Modeling: - A Bayesian approach could provide richer insights by incorporating prior beliefs and estimating posterior distributions for model parameters. This could be particularly useful for incorporating prior knowledge about pollster biases or expected trends. However, for this study, we focused on linear regression to prioritize interpretability and computational simplicity.

3.Inclusion of State-Level Effects: - An additional layer of complexity that could be added to the model is the inclusion of state-level effects to capture geographical variations in polling support. State-level data was available, but due to concerns over sample size in certain states, we opted not to include state as a random effect in this version of the models. Future iterations could explore adding a random effect for state to account for regional differences.

4.12 Evaluation Metrics

The models were evaluated based on:

- **Root Mean Squared Error (RMSE):** To assess the goodness-of-fit for both models. Lower RMSE values indicate that the model predictions are close to the observed polling percentages.
- **Cross-Validation:** We used a train-test split to ensure that the model generalizes well to unseen data. This helps prevent overfitting, where the model might otherwise perform well on training data but poorly on new observations.

4.13 Summary

Our modeling approach leverages multiple linear regression to capture the relationship between polling percentages and various time, quality, and methodological factors. By fitting separate models for Kamala Harris and Donald Trump, we are able to compare and contrast their polling trends while accounting for pollster quality and potential biases. The inclusion of pollster in Model 2 allows us to control for variability between polling organizations, enhancing the accuracy of the model's predictions.

The primary goal is to understand how support evolves over time and how it is influenced by the quality and methodology of the polling process. Despite some limitations, the models provide a useful framework for analyzing polling data, and future iterations could explore more sophisticated approaches such as Bayesian modeling or hierarchical random effects to improve robustness.

5 Results

5.1 Key Metrics for Model 1

Table 7 presents the key metrics for Model 1, which served as a baseline analysis for polling percentages for Kamala Harris and Donald Trump. The R-squared values are relatively low for both candidates, at 0.07 for Harris and 0.14 for Trump, indicating that this simple model explained only a small proportion of the variance in polling percentages. The AIC values were 4926.19 and 5053.99 for Harris and Trump respectively, with similar RMSE values around 4, suggesting limited predictive power. This baseline model was used to establish a simple trend before moving to a more comprehensive model that accounts for additional variability.

Table 7: Key Metrics for Harris's and Trump's Model 1

Candidate	R_squared	Adj_R_squared	AIC	RMSE
Kamala Harris	0.0693917	0.0661974	4926.186	3.977733
Donald Trump	0.1428927	0.1400197	5053.988	4.000399

5.2 Significant Predictors in Model 2

Table 8 and Table 9 present the summary statistics for Model 2 for Donald Trump and Kamala Harris respectively. The second model extends the baseline model by including pollster-specific effects, which significantly improves its predictive capabilities.

For Donald Trump, significant predictors include end_date (estimate = 0.0206, $p < 0.001$), which shows that his polling percentage tends to increase slightly over time. Specific pollsters like Ipsos (estimate = 5.39, $p < 0.001$) and MassINC Polling Group (estimate = 8.88, $p < 0.001$) show considerable effects, indicating variability in results due to the polling organization.

For Kamala Harris, end_date is also significant (estimate = 0.0129, $p = 0.0015$), suggesting a positive trend over time. Notable pollster effects include MassINC Polling Group (estimate = 7.70, $p < 0.001$) and University of Maryland/Washington Post (estimate = 14.62, $p < 0.001$), indicating significant variability in the polling results reported by different organizations.

Table 8: Significant Predictors in Trump's Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-362.0152590	76.1439120	-4.754356	0.0000023
end_date	0.0205665	0.0038044	5.405995	0.0000001

Table 8: Significant Predictors in Trump's Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
pollsterCES / YouGov	-2.0677729	0.7404316	-2.792659	0.0053447
pollsterChristopher Newport U.	-8.3144482	3.4019229	-2.444044	0.0147253
pollsterCNN/SSRS	-2.9380417	0.8633852	-3.402933	0.0006977
pollsterData for Progress	-2.9057044	1.1892279	-2.443354	0.0147532
pollsterHart/POS	-2.7784568	1.2538201	-2.215993	0.0269552
pollsterIpsos	-5.3860166	0.5777939	-9.321691	0.0000000
pollsterMarquette Law School	-3.1713246	0.7976517	-3.975827	0.0000761
pollsterMassINC Polling Group	-8.8758769	1.1368145	-7.807674	0.0000000
pollsterMonmouth	-4.3710060	1.7317807	-2.523995	0.0117835
pollsterPPIC	-18.1704828	2.4210375	-7.505246	0.0000000
pollsterQuinnipiac	-1.3244607	0.6610842	-2.003468	0.0454430
pollsterRoanoke College	-6.4806637	1.9928555	-3.251949	0.0011912
pollsterSiena	-9.8414190	1.4369741	-6.848710	0.0000000
pollsterSiena/NYT	-1.9808928	0.4806554	-4.121233	0.0000414
pollsterSuffolk	-3.8942064	1.0878076	-3.579867	0.0003632
pollsterSurveyUSA	-4.7504936	0.9284169	-5.116768	0.0000004
pollsterU. New Hampshire	-8.7759503	0.7603723	-11.541648	0.0000000
pollsterUC Berkeley	-14.4284057	1.9896731	-7.251646	0.0000000
pollsterUniversity of Maryland/Washington Post	-17.4075024	1.5592179	-11.164253	0.0000000

Table 8: Significant Predictors in Trump's Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
pollsterUniversity of Massachusetts Lowell/YouGov	-4.8385552	1.2525950	-3.862825	0.0001206
pollsterWashington Post/George Mason University	-3.6049407	0.8440475	-4.271016	0.0000216
pollsterYouGov	-2.7983112	0.4899299	-5.711657	0.0000000

Table 9: Significant Predictors in Harris's Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-209.3874946	80.6850714	-2.595121	0.0096223
end_date	0.0128517	0.0040313	3.188006	0.0014861
pollsterCES / YouGov	1.6185018	0.7478426	2.164228	0.0307307
pollsterEmerson	1.4617578	0.5490485	2.662347	0.0079096
pollsterMason-Dixon	-6.2968569	2.0081605	-3.135634	0.0017751
pollsterMassINC Polling Group	7.7017543	1.1485716	6.705506	0.0000000
pollsterPPIC	11.9426363	2.4452746	4.883965	0.0000012
pollsterSiena	6.1482641	1.4519983	4.234347	0.0000255
pollsterSiena/NYT	-1.8046820	0.4862417	-3.711492	0.0002198
pollsterU. New Hampshire	7.9704127	0.7699816	10.351432	0.0000000
pollsterUC Berkeley	10.1564260	2.0097111	5.053675	0.0000005
pollsterUniversity of Maryland/Washington Post	14.6195031	1.5749204	9.282693	0.0000000

5.3 Visual Representation of Model Predictions

Figure 12 and Figure 13 depict the predicted polling percentages for both candidates over time based on Model 2. The solid lines represent the smoothed predictions for Kamala Harris (blue) and Donald Trump (red). The visualizations reveal that both candidates maintain relatively stable trends, with some fluctuation, particularly influenced by pollsters and specific dates.

Figure 13 indicates that Kamala Harris tends to have a slightly higher polling percentage compared to Donald Trump, with less fluctuation over the observed period. This suggests

that her polling results are more consistent across different pollsters and dates.

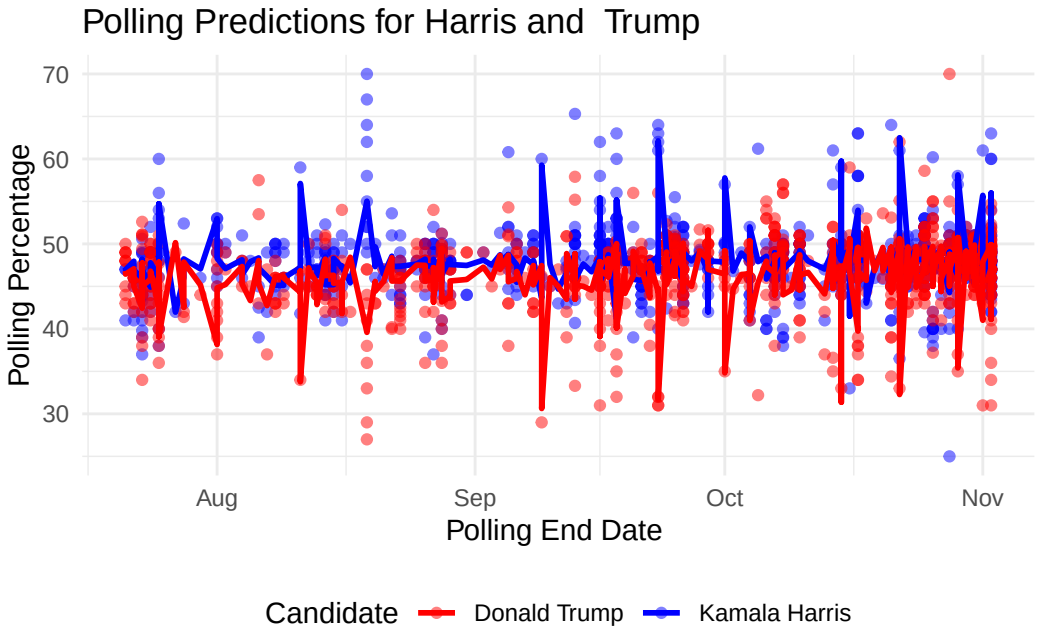


Figure 12

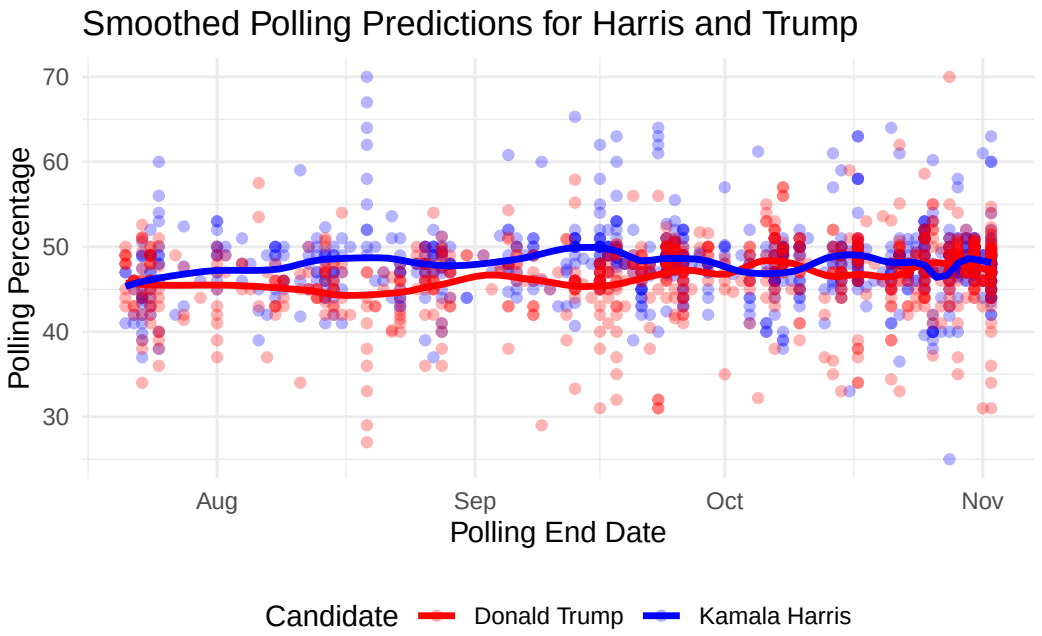


Figure 13

5.4 Summary

The results indicate that while both candidates' polling percentages are influenced by factors such as pollster-specific effects and end date, Kamala Harris shows more consistent support over time. The extended Model 2 provided better explanatory power compared to the baseline Model 1, as reflected in the inclusion of pollster effects that help explain the observed variability in polling data.

6 Discussion

6.1 Key Findings

This paper provides a forecast for the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump, leveraging linear regression models applied to recent polling data. The analysis projects Kamala Harris as the leading candidate, with consistent support across various polls and polling organizations, suggesting her potential to win the upcoming election.

Model 1, which included basic predictors like end date, transparency score, and numeric grade, demonstrated limited explanatory power, with R-squared values of 0.07 for Harris and 0.14 for

Trump. However, Model 2, which introduced pollster-specific effects, provided better insights into the variability in polling percentages.

The smoothed polling predictions in Figure 13 clearly showed that Harris is ahead of Trump throughout the analyzed period. This consistent lead, combined with higher polling percentages reported by major polling organizations such as MassINC Polling Group and University of Maryland/Washington Post, indicates Harris's current advantage heading toward the election.

6.2 Pollster Impact and Voter Sentiment

Pollster-specific effects played a substantial role in the reported polling outcomes. For Kamala Harris, significant pollster effects included organizations like MassINC Polling Group and University of Maryland/Washington Post, which reported notably higher polling percentages for Harris. This suggests that certain polling organizations may have a favorable sampling bias or different methodologies that influence their reported results.

Similarly, Donald Trump also saw significant variability across pollsters. Notably, Ipsos and MassINC Polling Group were among those reporting significantly higher polling percentages for Trump. These pollster-specific biases suggest the importance of accounting for the polling organization when analyzing election forecasts.

6.3 Regional Analysis and Swing States

The regional analysis highlighted in swing states shows varying degrees of support for both candidates, with some states showing Harris in the lead and others favoring Trump. Notable swing states like Arizona and Pennsylvania displayed significant variation, with Harris maintaining a narrow lead in some polls while Trump showed higher support in others.

The combined analysis of both Harris and Trump across the swing states suggests that Harris is predicted to outperform Trump, albeit with narrow margins in key regions. This is critical for electoral success, as even slight changes in voter sentiment in swing states can have a substantial impact on the final election outcome. The projections emphasize the importance of strategic campaigning in these battleground states, where the race remains close.

6.4 Polling and Model Uncertainty

The polling data used in this analysis, while indicative of current trends, introduces uncertainties due to incomplete coverage and pollster bias. The variability observed across different pollsters underscores the challenges in capturing a fully representative sample of the voting population. Certain demographic groups may be underrepresented, resulting in biased outcomes.

The models assumed linear relationships between polling percentages and predictors, which may have oversimplified the complex relationships at play in voter sentiment. This linearity assumption, while useful for initial analysis, may overlook potential non-linear dynamics that can significantly affect polling outcomes, especially during dynamic campaign events or sudden shifts in public opinion.

6.5 Limitations and Weaknesses

Despite the predictive capabilities of the models, several limitations should be acknowledged:

1.Assumption of Linearity: The assumption of linearity restricts the model's ability to capture non-linear shifts in voter sentiment, particularly during pivotal campaign moments. Incorporating non-linear modeling techniques could better reflect the complexity of voter behavior.

2.Pollster Variability: The differences in poll results across organizations, as observed in Model 2, point to systematic biases in polling methodologies. These biases may reflect differences in sampling techniques, question wording, or demographic targeting, which can skew reported outcomes.

3.Potential Polling Biases: Nonresponse bias remains a significant challenge, particularly for Trump's support, as seen in previous election cycles. The "Shy Tory Effect" may lead to an underreporting of support for Trump, making it difficult to capture the full extent of his voter base.

4.Historical Data and Recency Bias: The reliance on historical data risks the model being influenced by past trends that may not hold in the current electoral cycle. Unexpected shifts in voter sentiment, as seen in previous elections, could lead to discrepancies between the model predictions and actual outcomes.

6.6 Implications for Future Research

The results of this analysis highlight areas for improvement and further research:

- Non-Linear Modeling: Future research could explore the use of non-linear models or splines to capture the complexity of voter sentiment, particularly during pivotal campaign events.
- Bayesian Methods for Dynamic Updates: Implementing a Bayesian framework would allow the model to update its predictions in real time as new polling data becomes available. This could help better capture the evolving nature of voter preferences, especially in response to major campaign developments.
- Incorporation of Demographic Data: Including demographic information—such as age, gender, income, and education—could provide more granular insights into voter behavior and help identify which demographic groups are most influential in determining electoral outcomes.

- Weighting Polls by Credibility: Incorporating a weighting scheme based on pollster credibility and transparency scores could help adjust for systematic biases and provide a more balanced view of the polling landscape.

6.7 Conclusion

The analysis presented in this paper indicates a clear lead for Kamala Harris over Donald Trump in the 2024 U.S. Presidential Election, based on recent polling data. While the models capture broad trends in support and demonstrate the impact of polling methodology, they also emphasize the inherent uncertainty in electoral forecasting. The current projections suggest that Harris is positioned well to win, particularly if the trends observed in polling data continue. However, these results should be interpreted cautiously, acknowledging the dynamic nature of elections and the various limitations inherent in polling-based modeling.

Future studies should seek to incorporate additional data sources and more sophisticated modeling techniques to better capture the complex dynamics of voter behavior, particularly in an ever-evolving political environment.

Appendix

A Patriot Polling: Wisconsin Presidential Analysis Data Collection

A.1 Methodology: Overview

From 12 to 14 October 2024, Patriot Polling has conducted phone surveys in the state of Wisconsin of the United States of America, which borders Minnesota, Iowa, Michegan, and Illinois. Their target population are voters in Wisconsin. For sample of landlines which are characterised by households, an equal number of phone numbers are randomly selected across every landline block in Wisconsin using Random Digit Dialing (RDD). As for the sample of mobile numbers, which are characterised by an individual with access to a digital mobile device, the sample was purchased from a consumer contact number database. For this poll, a total of 803 respondents has completed the survey either from contact through the landline or their personal contact number. (Ruggieri 2024)

Ruggieri (2024) conducted polling in such a way where a randomly contacted respondent would hear pre-recorded voice messages and users are able to interact with the automated phone system. This system is called Interactive voice response (IVR) which can handle outbound calls more systematically since the same questions would be asked sequentially in the survey. There were 2 questions provided by Patriot Polling that were asked to respondents. Both questions begin with 'How will you vote for president?'. This is followed by the candidates name as it appears in the voting ballot. There was no indication whether this was all the questions that were asked but there were tables provided by the pollster showing the breakdown by sex, education, preferred political affiliation and income. Therefore, it would be safe to assume that there were questions related to the socio-demographic characteristics of the respondents even though we don't know the specific questions asked.

A.2 Methodology: Population

The targeted population of this poll was not explicitly stated. While the pollster did conduct phone surveys, they did not mention any inclusion or exclusion criteria. Furthermore, the use of IVR during phone surveys and lack of information provided regarding the questions asked as discussed in Section A.1, it will be difficult to guarantee that the polls are targeting all eligible voters in Wisconsin. Given that nature of the polling article focuses on Wisconsin, it would be safe to assume that the targeted population would be eligible voters above 18 years old in Wisconsin.

A.3 Methodology: Sampling Frame

Since phone surveys were conducted on both landlines and smartphones, the sampling frame of this poll would be eligible voters above 18 years old in Wisconsin with a working landline or possess a phone with a working contact number.

A.4 Strengths

This automated method of phone surveys is a lot more convenient and efficient than in person type of surveys since it removes the need for a surveyor to visit respondents face to face and also removes the need for a surveyor to contact multiple respondents over the phone. This decreases labour effort and cost.

Furthermore, with real-time feedback from phone surveys, the list of questions can be easily and quickly adjusted to ensure the ethical integrity of the survey being conducted. According to Ruggieri (2024), this survey took place in 2 days which could have taken a lot longer if other in-person type of surveying methods were used in this research.

Particularly, the polling company used a repertoire of RDD sampling and stratified sampling on landlines to ensure that every landline block (strata) has equal numbers of respondents to be included in the final sample. This helps to ensure diversity from the sample by including different geographical landline blocks, reducing sampling bias. The purchasing of mobile phones from phone number databases abates sampling bias in areas where a particular demographic - younger people who are fully dependant on mobile phones and not landlines - would otherwise be underrepresented if only RDD sampling was conducted on landlines.

A.5 Limitations and weakness

Firstly, there was a lack of information regarding how RDD was performed for the landline sample, how they handled non-working landlines, as well as the total sample size attributed from landline and mobile phone surveys.

The handling of no-response entities were also not explicitly mentioned by Patriot Polling. This could translate to respondents who were chosen but did not pick up their mobile phones/landlines as it was a busy period in the day or due to potential heightened awareness of phone based scams. This means that the final sample obtained might not be representative of the Wisconsin voter population.

Also, the initial phrasing to the second question can very easily confuse respondents. Although the initial phrasing is followed by the senate candidate's name, the initial phrasing 'How will you vote for president' still uses the wrong word 'president' instead of 'senate'. This does not reflect a high quality survey, which could lead to decreased trust from the respondent to

Patriot Polling, in turn potentially resulting in a premature end to the surveys, voiding the initial response.

Overall, the biggest weakness in the methodology would be the purchasing of phone numbers from a consumer database as it introduces selection bias. Individuals from middle to higher socio-economic status would have greater purchasing power than their lower income counterparts, which means individuals from lower socio-economic status may not have phones. In addition, elderlies within Wisconsin may not necessarily have mobile phones due to the lack of skills to use such technologies. Also, individuals would've opted for do-not-call registries, opted out of marketing databases or simply have prepaid phones which do not link to their personal information. This could mean an under representation of individuals from the said groups.

A.6 Simulation exploring weaknesses in Patriot's Polling's methodology

To illustrate the biggest concern raised in Section A.5, we will be simulating a Patriot's Polling's methodology by using US Census data to obtain random our pool of samples that are representive of the population of Wisconsin with information about their age, race and county. Since the raw data provided by IPUM (Ruggles et al. 2024) encoded states and counties, we combined the raw data with a supplementary excel file provided by IPUM to obtain the actual county name. We will be using the variable $CISMRTPHN$ from the census data which represents whether the particular respondent of the census data has a smartphone. This variable will be used to mirror a Patriot's Polling methodology of obtaining personal contact number by purchasing from an online database. We used this approach as only Wisconsin respondents with smartphones can possibly be included in a smartphone contact number database.

For the sake of discussion, we will be obtaining 2 samples. **Sample A will include respondents above the voting age of 18 in Wisconsin regardless of whether they possess a smartphone while sample B will only include respondents above the voting age of 18 in Wisconsin who have smartphones.** Sample B will mimic purchasing phone numbers from a database.

Table 10 shows us the proportion of white, black and asian races from sample A while Table 11 shows us that from sample B. We notice that black individuals from Wisconsin are underrepresented (8.7% vs 9.4%). According to Johnson (2018), black poverty rates in Wisconsin are 2.5 times higher than the overall Wisconsin poverty rate. This supports the finding that if we were to include only individuals with smartphones, certain racial groups might be underrepresented. Likewise, we are able to observe Asians being overrepresented (4.4% vs 3.8%). Ricketts and Kent (2024) exerts that for various educational attainment levels, Asians are earning more than individuals who belong to white, black or hispanic households. It is unsurprising that Asians are being overrepresented as higher wealth among the Asian population in Wisconsin would translate to higher purchasing power to purchase smartphones. Asians would naturally be more likely to be included in a sample where only smartphone users are considered, by

extension, Asians would be more likely to be included in a sample obtained from a consumer database containing user contact numbers.

Table 10: proportion of races from non-smartphone-discriminating sample

race	count	proportion
Asian	38	3.8%
Black	94	9.4%
White	768	76.8%
others	100	10%

Table 11: proportion of races from smartphone-only sample

race	count	proportion_chosen
Asian	44	4.4%
Black	87	8.7%
White	759	75.9%
others	110	11%

This becomes an issue because underrepresentation of black communities and overrepresentation of asian communities in Wisconsin would not reflect the actual voting preferences of a particular racial demographic, resulting in sampling bias, potentially misleading people who interpret a Patriot Polling's poll on Wisconsin or even their written articles.

Next, Table 12 shows us the the percentage difference in individuals chosen categorised by age groups from sample A and B. We are immediately alerted to a 20% decrease in the number of elderlies chosen from sampling method B compared to sampling method A. As of 2024, elderlies above the age of 65 are either from the baby boomer generation or silent generation, who are known to experience difficulty when using smartphones. Pew Research Center (2024) shows us that 97% of Americans between 18 to 49 have smartphones, 89% for that of age groups 50 to 64, and a lower 76% of the American elderlies above 65 year old who own smartphones. Therefore, by proceeding with sample B (parallel to purchasing phone numbers from consumer database) for which the inclusion critera selects individuals who own smartphones, we might be discounting the votes of elderly population, resulting in selection bias.

Table 12: percentage difference in individuals chosen from sample A and B by age groups

age bin	sample A	sample B	perc diff
above 65	238	191	-20%
18-30	191	206	+8%
30-65	571	603	+6%

B Idealised methodology for presidential election forecasting

B.1 Context of presidential election

US General elections runs every 4 years. The current state of the US electoral system is a two-party system where the republican and democratic parties dominate the political space (U.S. Embassy & Consulate in the Kingdom of Denmark, n.d.). During general elections, eligible voters in America are actually voting for a group of people called electors within their state. Elector candidates from 38 out of 50 states are usually pledged to support a specific political party. While there had been 'faithless electors' who cast a vote to the opposing party, they account for lesser than 1% of the elector population from the past 58 presidential elections. 'Faithless electors' had never affected the outcome of the elections (Otis 2024). If a elector candidate receives majority of votes by voters in their state, they gain the privilege of all electoral votes allocated for that state. All electors make up the electoral college. The electoral college would then casts their votes to their preferred presidential nominee, for which the majority would determine who becomes president.

B.2 Split electoral votes in Nebraska and Maine

In the states of Nebraska and Maine, individuals are voting not at the state level, but at the congressional district level. Nebraska has 3 districts while Maine has 2. One elector is chosen for each congressional district, and an additional 2 electors is given for the statewide majority. Unlike the other states which operate on a winner-takes-all, this difference in voting allows for greater variation in the outcome of the general elections (Encyclopædia Britannica, Inc 2024).

B.3 Idealised methodology: Pilot

The purpose of a pilot testing is to ensure a well-designed election forecasting survey. Due to the nature of the general election characterised by the votes of the electoral college, random sampling has to be performed across all 50 states of America. $5,000 should go towards a pilot test obtaining 200 respondents per state. This pilot test would target battleground states such as Wisconsin as well as Maine and Nebraska to test specific issues around a split vote between the two presidential candidate. Begin by using stratified sampling that divides each state population into strata by various socio demographic factors (e.g. gender, highest educational attainment, race, income), then conduct systematic random sampling within each strata. An optional $5 digital or mailed voucher can be used to incentivise respondents to complete the survey.

Respondents should be recruited quickly using by random digit dialing (RDD) to reduce selection bias. Firstly, obtain the list of possible area codes corresponding to the first 6 digits

of a contact number within a state. Then randomly sample the remaining 4 digits. Coupled with the use of an automated interactive voice response system, this would increase efficiency while taking note of non-response from either premature end to phone surveys or non-working landlines/phonelines. This also decreases response bias as unlisted contact numbers would be included in the sampling frame as well. Theoretically, a larger than expected sample has to be recruited in order to account for contact numbers that are not in use. Also, it is important to distinguish landline RDD sampling and cellphone RDD sampling as they represent different communties within the population as mentioned in Section A.4 and Section A.5

As for the contents of the phone survey, begin by providing detailed, succint information. We would introduce our company name, the purpose of the phone survey, and how only their polling preference is recorded and how their contact information would be used (but not saved) for the optional voucher redeeming purposes. If the phone survey is conducted in the first congressional district of Maine, explicitly mention that this survey is used to obtain polling preferences among voters within the first congressional district of Maine, and there are no political affiliations. It is paramount to obtain consent from the respondent. Therefore, have an option to let a respondent opt out of the survey without any form of discrimination. If a respondent chooses to continue the survey, we can ask them for their highest educational background, race, gender and/or other socio demographic information with the option to indicate non-response for any particular question. If a respondent successfully completes the phone survey, record the information into a secure, password protected excel sheet for further data analysis. It is also important to honor the incentives should the respondent complete the survey by using only the phone number. At the end, thank the respondent for their coorperative participation.

In order to expand the sampling frame to include voters who may not prefer to do a phone surveys, we can conduct a Google forms survey as well which is free of charge. The format of the form should mimic that of the RDD random sampling method. Refer to this google form for an example of the survey questions aksed.

After obtaining all pilot samples, validate the pilot data by checking for any patterns or outliers. In particular, we have to pay special attention to ensure that the samples received are balanced and representive of the voting population in that particular state by cross referencing to census data available from IPUMS US census data (Ruggles et al. 2024). If the data is not representive of a particular population, it might be necessary to conduct in person surveys in regions of the state where certain demographics are underrepresented.
These forms of recruiting respondents should mimic the actual sampling conducted on a larger scale after a successful pilot.

B.4 Idealised methodology: Large scale sampling

By refining our phone surveys and google form surveys, we can expand the surveys to all states in America. For this we budget out $80,000 to obtain between 2000 to 5000 samples across

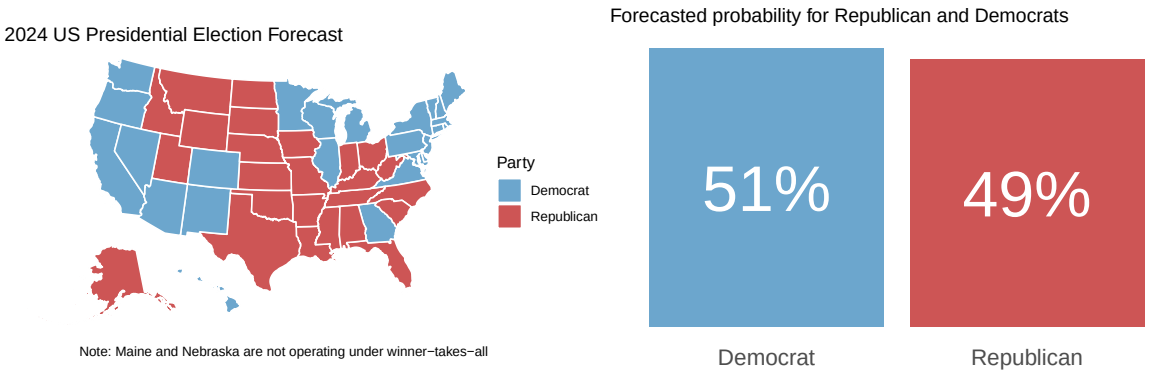
each state, depending on the state's population. For the states of Nebraska and Maine, we will be performing both analysis for individual congressional districts and statewide majority votes to determine the number of estimated electoral votes as described in Section B.2. Particularly for all battleground states as seen from the past 2020 elections like Arizona, Florida and Maine's second congressional district, we will be oversampling to ensure that the preferences of sampled voters are better representive of the overall voter population in their particular state or congressional district, leading to reduced sampling error and more precise forecasting of the presidential elections.

If the pilot discovers regions where certain demographics are underrepresented (e.g. poorer communities without digital communication devices nor landlines), surveyors would have to be employed in these regions to obtain samples directly.

Ultimately, the algorithm to forecast the presidential winner involves obtaining the majority vote based on samples obtained within each state (and districts in Nebraska and Maine). Then based on the number of electors assigned to that region, we will sum up the number of electors affiliated to the presidential candidates. This sum divided by the 538 total electors would be the predicted probability that a presidential candidate would win in the upcoming election. $10,000 will be spent on the cost of surveyors. The remaining $5,000 from our budget will be used as contigency funds.

B.5 Presentation of polling results

We can map out the results to give a visual representation of how the votes mapped out to each state. Figure 14a provides a quick overview to an estimated presidential candidate preference in each state in America from the obtained samples. The colour distinction makes it quickly identifiable the majority within each state. Figure 14b provides the final predicted probability of how likely a party (presidential candidate) would be elected president. The following map utilises data obtained from Cook Political Report (Wasserman et al., n.d.). Tigris library (Walker 2024) coupled with tidyverse library (Wickham et al. 2019) was used to generate the US map shown in Figure 14a. Tidyverse library was also used to generate the bar chart shown in Figure 14b.



(a) Map forecast

(b) Overall forecast

Figure 14: Forecast of US presidential elections

References

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

Encyclopædia Britannica, Inc. 2024. *Electoral College*. https://www.britannica.com/video/demystified-how-does-electoral-college-work/-250292.

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

Johnson, Deborah. 2018. *Study Finds Wisconsin's African American Poverty Rate Three to Four Times Higher Than White Poverty Rate*. University Of Wisconsin-Madison. https://news.wisc.edu/study-finds-wisconsins-african-american-poverty-rate-three-to-four-times-higher-than-white-poverty-rate/.

Keeter, Scott. 2024. "Public Opinion Polling Basics." *Public Opinion Polling Basics*. Pew Research Center. https://www.pewresearch.org/course/public-opinion-polling-basics/#how-does-polling-work.

Otis, Deb. 2024. *Do Faithless Electors Change Presidential Election Results?* https://fairvote.org/do-faithless-electors-change-presidential-election-results/.

Pew Research Center. 2024. *Mobile Fact Sheet*. https://www.pewresearch.org/internet/fact-sheet/mobile/?tabItem=5b319c90-7363-4881-8e6f-f98925683a2f.

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

Radcliffe, Mary, and G Elliot Morris. 2023. "538's polls policy and FAQs." *ABC News*. ABC News Network. https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

Ricketts, Lowell R., and Ana Hernández Kent. 2024. *Wealth and Its Distribution: A Look at Asian American Households in 2022*. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/on-the-economy/2024/aug/wealth-distribution-look-asian-american-households-2022#:~:text=Indeed%2C%20Asian%20household%20wealth%20is,

several%20factors%2C%20such%20as%20education.

Ruggieri, Lucca. 2024. *Patriot Polling Data Collection Methodology*. https://patriotpolling.com/our-polls/f/trump-and-baldwin-hold-narrow-leads-in-wisconsin.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rogers, and Megan Schouweiler. 2024. IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/https://doi.org/10.18128/D010.V15.0.

Ryan Best, Aaron Bycoffe. 2024. "FiveThirtyEight: Latest US 2024 General Election Polls." *FiveThirtyEight*. ABCNews. https://projects.fivethirtyeight.com/polls/.

U.S. Embassy & Consulate in the Kingdom of Denmark. n.d. *Presidential Elections and the American Political System*. https://dk.usembassy.gov/usa-i-skolen/presidential-elections-and-the-american-political-system/#:~:text=That%20means%20that%20two%20parties,Party%20and%20Natural%20Law%20Party.

Walker, Kyle. 2024. *Tigris: Load Census TIGER/Line Shapefiles*. https://CRAN.R-project.org/package=tigris.

Wasserman, David, Sophie Andrews, Leo Saenger, Lev Cohen, Ally Flinn, and Griff Tatarsky. n.d. Cook Political Report. https://www.cookpolitical.com/vote-tracker/2020/electoral-college.

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.