Forecasting Presidential Candidate Support during the Final Week before Election Day*

In the final week before the election, Harris is expected to lead the polls with 48.03% of support with Trump following close behind with 47.98% of support.

Tianning He Julia Lee Shuangyuan Yang

November 4, 2024

With the Presidential General Election Polls data provided by 538, the present analysis builds a model to forecast the percentage of support Harris and Trump will receive during the week before election day. By pooling poll data by week through the use of seasonal indexes and a linear model, this analysis forecasts that prior to election day, Harris will receive 48.03% of support and Trump will receive 47.98% of support from the polls. This suggests that the election will be a very tight race that may come down to the outcomes of the swing states. Through this, this analysis further examines the impacts of pooling poll data and discusses how improvements may be made to its model.

1 Introduction

The recent US presidential election has been a hot topic of discussion, and as the election approaches, speculation about the final outcome has begun. Accurately predicting the outcome has become a focus of attention for political analysts, the media, and the public. Polling organizations such as YouGov have conducted surveys that provide a deeper look into voters' intentions and preferences, which can help to forecast potential election results. However, people's intentions, preferences, and opinions can be influenced by different social contexts (Moussaïd et al. 2013), meaning that voters' opinions on who should be the next president of their country can change over time.

With this, the objective of the present analysis is to build a model that forecasts the percentage of support Kamala Harris and Donald Trump will receive in the final week before the election, while accounting for variations across time. The opinions and preferences of voters' are what

*Code and data are available at: https://github.com/JuliaJLee/Forecasting_US_Election_2024.git

this analysis is interested in. By using seasonal indexes along with a regression model, this analysis seeks to measure and approximate voters' opinions to generate more precise forecasts that reflect their preferences.

This analysis first aims to understand how the percentage of support for both candidates has changed over time along with how the percentage of support for one candidate has changed relative to the other. Then, as a means to observe how the model affects the data, this analysis examines how the percentages of support predicted by the linear model compare to the weighted averages over time. Finally, the model is used to predict the percentage of support Harris and Trump will receive in the final week before the election. Additionally, this analysis considers the accuracy of the linear model's predictions and investigates the impact of the "pooling polls" method on election predictions. This provides a deeper awareness of how the complexity that surrounds having an opinion can be translated into data that ultimately predicts real-world election results.

As a result, this analysis finds that Harris will lead the polls with 48.03% of support with Trump close behind with 47.98% of support, suggesting that this election will be a tight race that may need to be determined by the outcomes of the swing states. These findings further highlight important considerations for how to improve the ways in which poll data is interpreted and different prediction methods for future elections. The following expands on the ideas above by presenting a description of the poll data for this analysis (Section 2), a more in-depth explanation of model and how it is used (Section 3), an overview of the results (Section 4), and a discussion of the overall analysis (Section 5).

2 Data

To simulate, test, download, clean, and model data throughout this analysis, the statistical programming language R was used (R Core Team 2023). Specific libraries that assisted the analysis include `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023),`tinytex` (Xie 2019), `ggplot2` (Wickham 2016), `knitr` (Xie 2015), `arrow` (Richardson et al. 2024), `here` (Müller and Bryan 2020), `testthat` (Wickham 2011), and `modelsummary` (V 2022).

2.1 Presidential General Election Polls Data

The present analysis uses Presidential Election Polls data provided by 538 (FiveThirtyEight 2024). The data contains records of various pollsters, several aspects of each pollster (e.g. a poll identification number, poll quality characterized by a numeric grade, sample size, etc.), candidates for the presidential election, and the percentage of support (pct) each candidate received based on the polls. Poll results are periodically updated to the 538 website and can be downloaded by users as csv files. The types of pollsters that can be included in the data by 538 are determined by various methodological and ethical standards set by 538 (Radcliffe and Morris 2023).

As the original data acquired by 538 is an extensive dataset, the following will provide details for the variables that are considered within the present analysis.

- **Pollster**: This variable contains the names of polling organizations that conducted each poll (e.g. YouGov, AtlasIntel, Siena, etc.).
- **Numeric Grade**: This variable reflects the quality or reliability of a pollster. It ranges from 0.5 to 3.0, increasing by 0.1. A larger number indicates a higher-quality pollster.
- **Population (pop.)**: This variable represents the voting status of those who respond to a given pollster. Respondents for different pollsters can be likely voters (i.e. voters who are more likely to vote on election day) or registered voters (i.e. everyone who can vote).
- **State**: Polls can be state-specific (i.e. conducted within a specific state) or they can be national. In the original data, blank values indicate national polls.
- **Sample Size**: This variable reflects the total number of respondents for each poll within the data.
- **Hypothetical (hypo.)**: Polls can ask respondents about a hypothetical match-up between different candidates. With this, polls that do not consider hypothetical match-ups are denoted as "False" in the data.
- **Candidate Name**: This variable contains the names of all the candidates that were asked about in the poll. For this analysis, the presidential candidates, Kamala Harris and Donald Trump, are considered.
- **Poll Start and End Dates**: The start and end dates indicate when a poll was launched and when that same poll was closed.
- **Percentage of Support (pct)**: The percentage of support (pct) reflects the percentage of support a candidate received in a given poll.

Table 1: Organized Presidential General Election Poll Data

Pollster	Numeric Grade	State	Start Date	End Date	Sample Size	pop.	hypo.	Candidate Name	pct
AtlasIntel	2.7	North Carolina	10/30/24	10/31/24	1373	lv	FALSE	Kamala Harris	46.7
AtlasIntel	2.7	North Carolina	10/30/24	10/31/24	1373	lv	FALSE	Donald Trump	50.7
AtlasIntel	2.7	North Carolina	10/30/24	10/31/24	1373	lv	FALSE	Kamala Harris	47.0

Table 1 displays the data that is used throughout the analysis and illustrates each of the variables described above. Summary statistics for these variables can be found in the Appendix (Section A.1).

Putting these variables together, the analysis is able to construct a model that forecasts the percentage of support (pct) for Harris and Trump by considering high-quality pollsters, a

population of voters who a more likely to vote on election day, polls from various states, poll sample sizes, and the times at which polls were conducted. An in-depth explanation as to why the variables above are selected is provided in (Section 3).

2.2 An Account on Measurement

From the day of the week to different survey methods (e.g. email, phone, mail, etc.), there are many things that can influence how voters' opinions or preferences are shaped and how those opinions are reduced down to a value. The percentage of support (pct) for a candidate is the value that is meant to reflect voters' opinions on who should be president. As a single percentage is the end result of a survey that was conducted by a pollster, voters' thoughts and opinions can be translated into an entry in a dataset by considering the ways in which pollsters build their surveys along with how they analyze and interpret the responses. Opinions may be converted into data by calculating average or percentages of responses to closed-ended questions. Preference towards a particular candidate or party could be measured through the coding and counting of themes or key words within responses to open-ended questions. Additionally, how responses are interpreted (e.g. how themes are generated for open-ended responses) can impact how voters' thoughts are measured. Thus, complex thoughts and opinions can be measured through careful consideration about what is being asked, how it is being asked, and how the responses are interpreted.

A detailed account about the methodology of a pollster found within the data obtained from 538 along with an idealized methodology and survey can be found in the Appendix (Section A.3, Section A.4).

3 Model

The objective of the present analysis is to forecast the percentage of support both presidential candidates, Kamala Harris and Donald Trump, will receive in the final week (October 27, 2024 to November 2, 2024) leading up to the election. The Presidential Poll data provided by 538 (FiveThirtyEight 2024) reflects voters' opinions and preferences about who should be the next President of the United States across time. As those opinions or preferences are subject to change as time goes on, this model seeks to account for this variability by building "seasonal indexes" and using them with a linear model to forecast the percentage of support for both presidential candidates.

In this model, a "season" is referred to as a 7-day week that is found between Sunday August 4, 2024 and Saturday October 26, 2024. The start date is August 4, 2024 because this date allows there to be enough data to observe the percentage of support for both candidates over several weeks. The end date is October 26, 2024 as this leaves roughly a week (October 27, 2024 to November 2, 2024) before the election on Tuesday November 5, 2024 to ensure that a forecast for this final week can be made.

Further, this model considers the following variables:

- **Pollsters with a numeric grade of 2.7 and above**: This model uses a cut-off of 2.7 for the numeric grade to strike a balance between the amount of data and the quality of the data. The design of this model requires that there is at least one poll that was conducted in each "season" or week, and this could not have been satisfied if only pollsters with a numeric grade of 3 are considered. A numeric grade of 2.7 and above allows this model to have sufficient data as well as data of high quality.
- **The likely voter (lv) population**: Likely voters are defined as voters who intend to vote on election day. With this, the model focuses on this particular voter population to generate a forecast that more closely resembles election day.
- **States**: This model looks at polls that were state-specific rather than national polls. Though each state is not considered individually in this analysis, this model considers these state-specific polls as they also allow for sufficient data to be used within the model.
- **Sample Size**: Poll sample size is used within the model to pool the poll data for each week within the time period defined by the model above.
- **Polls that did not ask about hypothetical match-ups**: As this model aims to forecast and compare support for the presidential candidates in the current 2024 election, hypothetical match-ups are not considered.
- **Presidential Candidates**: This model looks at the percentage of support for both Kamala Harris and Donald Trump throughout the analysis.
- **Poll Start and End Dates**: Start and end dates are important variables in the model as they are used to categorize poll data into the different "seasons" or weeks between August 4, 2024 and October 26, 2024.
- **Percentage of Support (pct)**: This is the target variable to be estimated by the model.

3.1 Model Process

Code that runs through the following steps can be found in the repository linked on page 1 (Section 1).

3.1.1 Step 1: Organize Poll Data for each Candidate by Week

For this model, a total of 12 weeks of poll data is analyzed. The exact start and end dates of each week (defined by the model) can be found below in Table 2. Every four weeks corresponds to a month. The first four weeks are in August 2024, the next four weeks are in September 2024, and the last four weeks are in October 2024.

Table 2: Weeks Defined by the Model

Week	Dates
1	Aug. 4-10
2	Aug. 11-17
3	Aug. 18-24
4	Aug. 25-31
5	Sept. 1-7
6	Sept. 8-14
7	Sept. 15-21
8	Sept. 22-28
9	Sept. 29 - Oct. 5
10	Oct. 6-12
11	Oct. 13-19
12	Oct. 20-26

Using the start and end dates of the polls, the model first filters on the polls that were conducted between August 4, 2024 and October 26, 2024 for each candidate. Then, it assigns each poll to a week as outlined in Table 2. An example outcome is shown below for Kamala Harris (Table 3).

Table 3: Poll Data Organized by Week For Harris

Sample Size	Candidate	Percentage of Support (pct)	Week
1,000	Kamala Harris	42.5	1
619	Kamala Harris	50.0	1
661	Kamala Harris	50.0	1
693	Kamala Harris	50.0	1
1,738	Kamala Harris	50.0	2
1,000	Kamala Harris	49.3	2

Each row in the outcome above (Table 3) represents a poll, and the "Week" column indicates the week in which that poll occurred. For example, the first four rows show polls that were conducted in the week of August 4 to 10, 2024. The last two rows show polls that occurred between August 11 to 17, 2024 (i.e. Week 2). Each poll's sample size and pct estimate are also included.

3.1.2 Step 2: Pool Poll Data by Week for both Candidates

With polls organized by week, the model now pools all the polls that occurred within a single week to generate a weighted average estimate of pct for that week.

To pool the polls for each week, this model first creates a weight for each poll by:

- (1) Finding the sum of the sample sizes of all polls in a given week
- (2) Dividing each sample size of each poll within that week by the sum found in the previous step

An example of these two steps would be to find the sum of the first four sample sizes for week 1 in Table 3, and then divide 1000, 619, 661, and 693 from the first four rows of Table 3 by the sum of the first four sample sizes.

Next, each weight for each poll within a given week is multiplied to the corresponding pct estimate. For example, by using Table 3, the model would multiply the quotient of (1000/sum of sample sizes) by 42.5, which is the pct estimate that corresponds to the poll with a sample size of 1000 in row 1 of Table 3.

Lastly, by taking the sum of the products of each weight and the corresponding pct estimate, the model produces a weighted average pct estimate for each week. An example outcome is shown below for Kamala Harris (Table 4).

Table 4: Weighted Average Percentage of Support for Harris By Week

Week	Weighted Average Percentage of Support (pct)
1	47.5
2	47.8
3	47.0
4	48.7

Table 4 shows the average percentage of support that Harris received in the first four weeks. For instance, in Week 1 (i.e. during the week of August 4 to 10, 2024), Harris received an average support of 47.5% across the polls that occurred in within this time period.

3.1.3 Step 3: Fit a Regression Model for Each Candidate using the Pooled Poll Data

Using the weighted average percentage of support (pct) for each week as the response variable, the model performs two regression analyses to predict the support for each candidate during each of the 12 weeks. The linear models are structured as follows:

$$\hat{y}_i = b_0 + b_1 \cdot w_i + \epsilon_i$$

where

- $\hat{y_i}$ represents the percentage of support (for Harris or Trump),
- b_0 represents the intercept of the linear models,
- b_1 represents the effect of each week,
- w_i represents the time period of a week (i = 1, 2,..., 12)
- ϵ_i captures the error within the linear models

Summary outputs for each model (one for Harris and another for Trump) along with model diagnostics to validate these models can be found in the Appendix (Section A.2).

3.1.4 Step 4: Find Seasonal (i.e. weekly) Indexes to Forecast Support for Harris and Trump

With the linear models fitted above, the model now creates a "seasonal" or weekly index for each week so that the week leading up to the election can be predicted while accounting for differences in voter opinions across different time periods.

A seasonal index for each week is calculated by first computing the ratio,

$$\frac{y}{\hat{y_i}}$$

.

For each week, the model takes the weighted average percentage of support (pct) that was found in Step 2 (Section 3.1.2) and divides it by the predicted value that is found using the linear model. This produces an outcome like the following in Table 5.

Table 5: Ratios for Each Week

Week	Weighted Average Percentage of Support (pct)	Predicted Average Percent of Support (pct)	Ratio
1	47.48	48.08	0.988
2	47.76	48.11	0.993
3	47.00	48.13	0.977
4	48.67	48.15	1.011

Now, the since the data is manipulated such that every four weeks corresponds to a month (August, September, and October), it follows that the final week (October 27, 2024 to November 2, 2024) that this model aims to forecast is Week 13 and the first week of the next month, November. So, this model computes the average of the ratios for the first, second, third, and

fourth weeks across each month to obtain a "seasonal" or weekly index that can be used to forecast Week 13. An example outcome is shown below Table 6.

Table 6: Seasonal (Weekly) Index for Each Week Across 3 Months

Week 1	Week 2	Week 3	Week 4
0.993	0.995	1.011	1.001

Using the seasonal index for Week 1 (0.993) presented in Table 6, this model can forecast the percentage of support that Harris will receive in Week 13 by:

- (1) Plugging in $w = 13$ to the linear model for Harris to predict her percentage of support
- (2) Multiplying the predicted value from the linear model by the seasonal index

The example outcomes provided throughout this section are for Kamala Harris only. It is important to note that the same process is also repeated for Donald Trump within the model.

3.2 Evaluating the Model

By pooling the polls for each of the defined weeks, the model assumes that the polls are unbiased – which is often not the case. While pooling polls that have occurred in a similar time period provides more precision than a single poll, a limitation of this model is that it overlooks the potential biases that can exist within the polls. Biases within polls can arise from their methodology, their audience, and the location in which the poll was conducted. As these variables are not explicitly considered by the model, it would not be appropriate to apply this model to forecast percentage of support as a function of different methodologies, voter populations, or states.

Despite these limitations, this model's strength lies in its ability to account for variations across time. This approach of using seasonal indexes and regression to forecast the percentage of support (pct) for both presidential candidates is able to capture seasonal (i.e. weekly) variation within the percentage of support candidates received and assess long-term trends. As such, this model can provide both a numerical outcome (i.e. a forecasted percentage of support) for each candidate along with a means to observe how the percentage of support for the presidential candidates has changed over time. As these strengths align with the objective of the analysis to forecast the percentage of support the presidential candidates will receive in the final week (October 27, 2024 to November 2, 2024) leading up to the election, this model is employed to obtain the findings presented in the next section (Section 4).

4 Results

4.1 Change in the Percentage of Support (pct) for Both Presidential Candidates Over Time

This initial analysis aims to understand how the percentage of support for both candidates has changed over the course of the 12 weeks defined by the model.

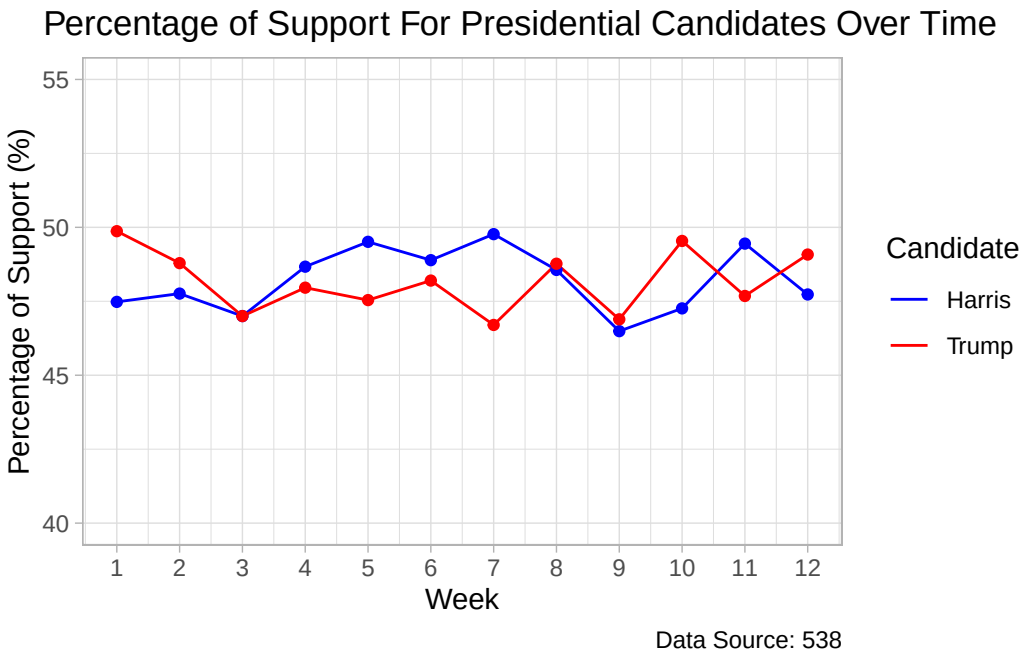

Figure 1: Percentage of Support for Both Candidates Over 12 Weeks (August 4 to October 26, 2024)

Figure 1 presents the weighted average percentage of support for each presidential candidate over time. While both candidates appear to have had fluctuations in the percentage of support they received, the difference between their percentage of support is smaller in Week 12 than it was at Week 1. In week 1, Trump received an additional 2.39% of support compared to Harris, and in Week 12, his percentage of support was 1.35% higher than Harris'. This pattern further indicates that as support for Trump has dropped, support for Harris has increased over the course of 12 weeks.

4.2 Predicting the Percentage of Support (pct) for Both Presidential Candidates Over Time

Figure 2 and Figure 3 provides a comparison of the predicted percentage of support from the linear model and the weighted average percentage of support that was computed with the poll data for Harris and Trump. Though both models do not seem to accurately capture the fluctuations in support each candidate received over time, they do appear to capture the overall increase in support for Harris along with the overall decrease in support for Trump over the 12 weeks. This further gives rise to the idea that while the percentage of support for Trump decreased, the percentage of support for Harris increased as the weeks went by.

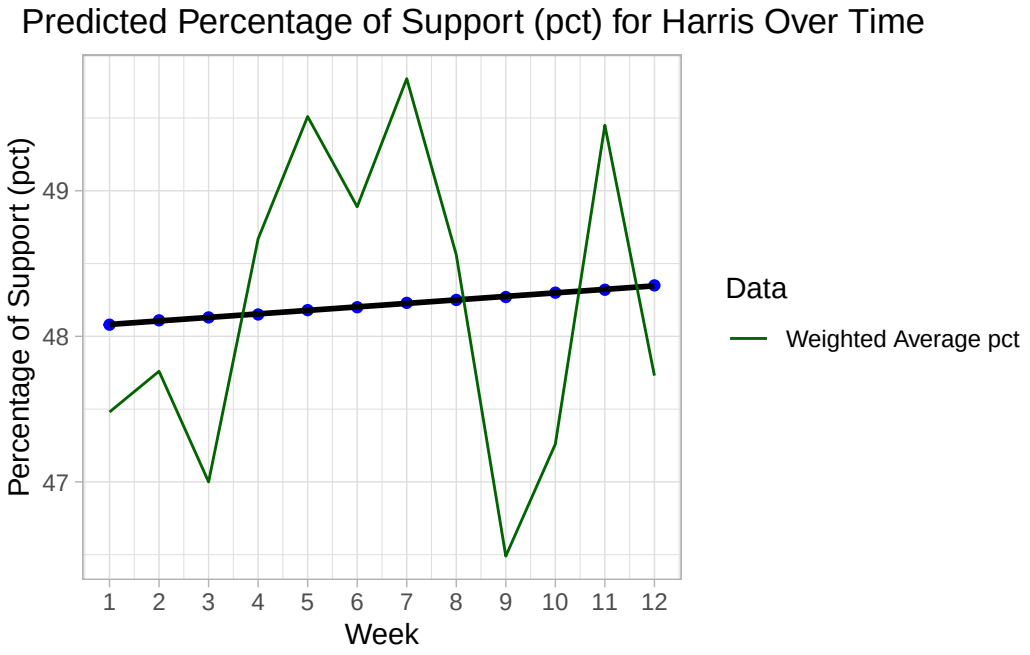

Figure 2: Predicted Percentage of Support for Harris Over 12 Weeks (August 4 to October 26, 2024)

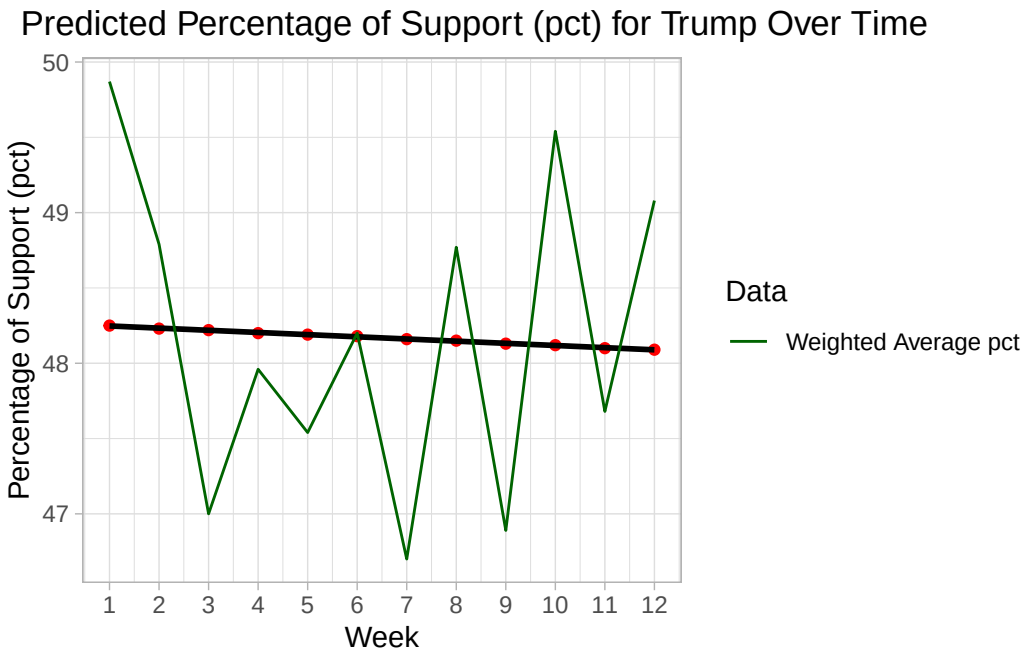

Figure 3: Predicted Percentage of Support for Trump Over 12 Weeks (August 4 to October 26, 2024)

4.3 Forecasting the Percentage of Support (pct) for Both Presidential Candidates For Week 13

The main objective of this model is to forecast the percentage of support Harris and Trump will receive in the final week before the election. This week - October 27, 2024 to November 2, 2024 - is Week 13, and the model uses a seasonal index along with a predicted value from the linear model to compute a forecast. The forecasted percentages of support for each candidate are shown in Figure 4 below.

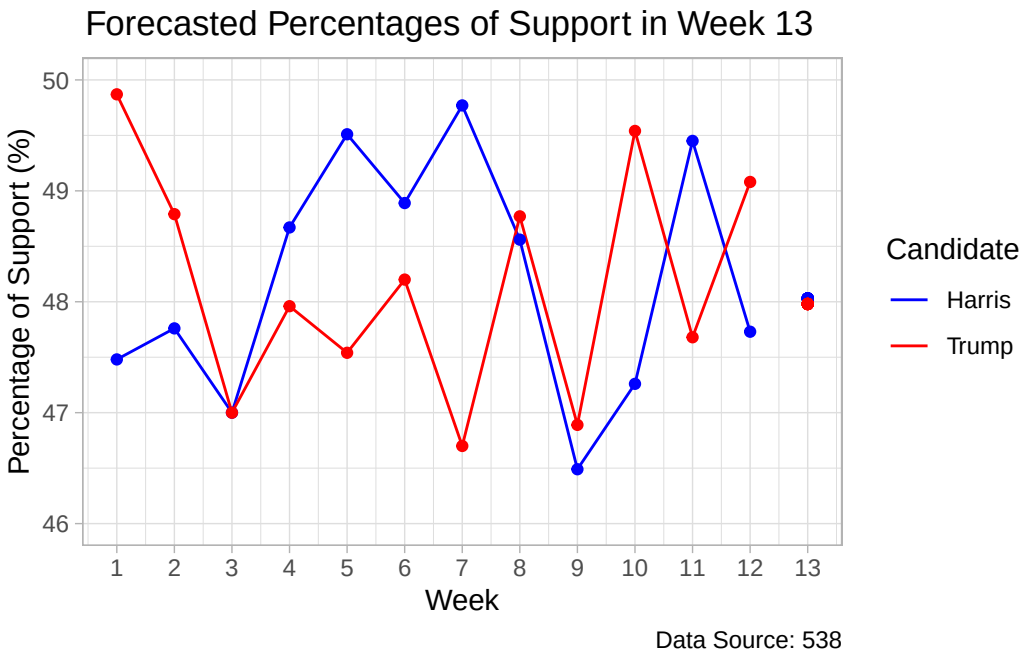

Figure 4: Presenting the Forecasted Percentages of Support for Harris and Trump in Week 13

Figure 4 shows two additional points (one for Harris and the other for Trump), and these points represent the forecasted percentage of support for each candidate in Week 13. With this, it appears as though the presidential election will be a tight race as the forecasts of percentage of support for Harris and Trump in Week 13 are very similar. In the final week leading up to the election (October 27, 2024 to November 2, 2024), Harris is expected to receive more support while Trump is expected to receive less support in Week 13. The exact forecasts of the model are as follows:

- This model forecasts that the percentage of support for Harris in the week leading up to the election will be 48.03%.
- On the other hand, the model forecasts that the percentage of support for Trump in the week leading up to the election will be 47.98%.

5 Discussion

As the 2024 US election approaches, the present analysis forecasts the percentage of support both presidential candidates will receive in the final week leading up to election day on Tuesday November 5, 2024. Through the use of seasonal indexes and a regression model, this analysis takes in to account how voter opinions or preferences change over time to generate more precise forecasts. With this, an understanding of how support for Harris and Trump has changed over time is provided in addition to the model's forecasts.

This analysis finds that while support for both Harris and Trump fluctuate over the course of the 12 weeks defined by the model, overall support for Harris appears to increase over time. In contrast, despite having a high percentage of support in Week 1, Trump is found to have received less support in Week 12. This pattern indicates that the percentages of support for Harris and Trump are converging as time goes on, implying that the election will be a tight race that may need to be determined by the swing states. Additionally, the forecasts computed by the model show that Harris will receive greater support than Trump during the week before the election (Week 13). However, as the difference in support that Harris and Trump will receive during this time is extremely small (0.05%), this result also gives rise to the notion that the 2024 election will continue to be a close race until election day.

As briefly mentioned in Section 3.2, a limitation of this analysis is that it assumes that polls are unbiased to pool them together. By not considering the effects of methodology, vote population (e.g. registered voters vs. likely voter), different states, and poll quality, the model may be amplifying the biases of different polls as it pools them together (Jackman 2005). Another limitation of this model is that it considers 12 weeks of data to forecast. Though this is a sufficient amount of data for the model, a stronger representation of the change in candidate support over time as well as more robust forecasts may be produced with a larger quantity of data.

Given these limitations, moving forward, this analysis should be adapted to incorporate a way to account for variation across polls. With a model that considers the effects of methodology, vote population, states, and poll quality along with seasonal indexes, more precise and less biased estimates of the percentage of support for candidates during a given time period could be made. As the findings of this analysis point to a very close race that can be determined by the swing states, a next step for the model could be to forecast the percentage of support for each candidate within each swing state to present a more detailed prediction of the election. Lastly, it may also be helpful to adopt a more sophisticated method to pool the polls, such as pooling the polls by considering the margin of error of each poll (Jackman 2005). A more enhanced method may allow for better precision, and in turn, provide a more reliable way to forecast.

A Appendix

A.1 Analysis Data Summary Statistics

Table 7: Presidential General Election Poll Analysis Data Summary Statistics

Numeric Grade	Sample Size	pct
Min. :2.700	Min. : 375	Min. :25.00
1st Qu.:2.800	1st Qu.: 730	1st Qu.:47.00
Median :2.900	Median : 941	Median :48.50
Mean :2.863	Mean :1020	Mean :48.19
3rd Qu.:3.000	3rd Qu.:1136	3rd Qu.:50.00
Max. :3.000	Max. :6473	Max. :70.00

Table 7 presents the summary statistics for the variables, "Numeric Grade", "Sample Size", and "pct". As the variables "Pollster", "State", "Start Date", "End Date", "Population", "Hypothetical", and "Candidate Name" are classfied as a "character" data type, they are not included in the summary above.

A.2 Model Validation

A.2.1 Model for Harris

Table 8 shows the summary output of the linear model that predicts the percentage of support for Harris across weeks. The low AIC, BIC, and RMSE values ensure that this model fits the data well. Here, the low R^2 value indicates that the time variable, "Week", does not explain much of the variability of pct for Harris.

Looking at the Residuals vs. Predictor, Residuals vs. Fitted Values, and Normal Q-Q plots with Figure 5, Figure 6, and Figure 7, there are no systematic patters (e.g. fanning) and the residuals appear to follow a normal distribution with little deviation from the QQ-line in the Q-Q plot. This suggests that the assumptions of linearity, homoscedasticity, independence, and normality are satisfied by this model.

Table 8: Summary Table of the Regression Results for Harris

	(1)
(Intercept)	48.057
	(0.693)
Week	0.024
	(0.094)
Num.Obs.	12
R2	0.007
R2 Adj.	−0.093
AIC	40.7
BIC	42.2
Log.Lik.	−17.351
RMSE	1.03

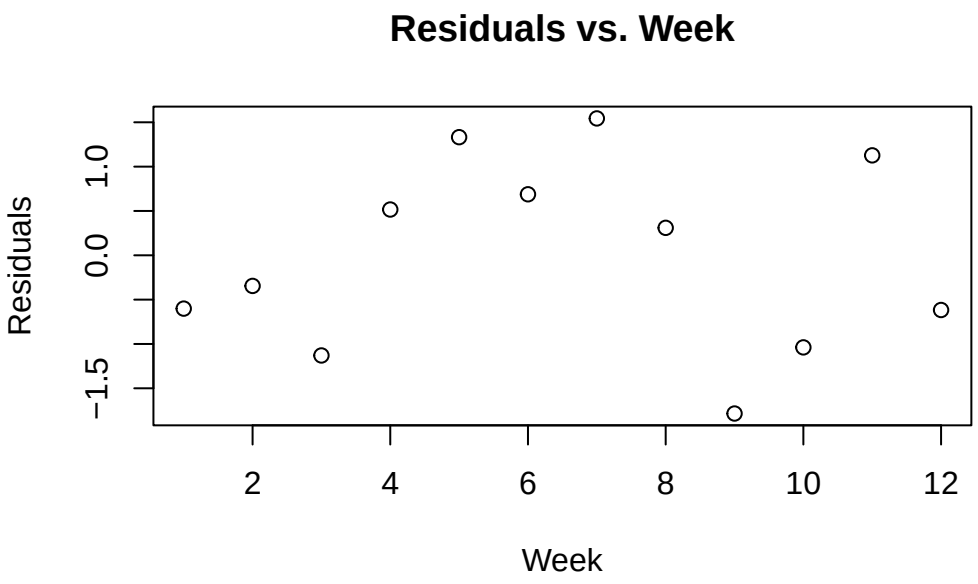

Figure 5: Showing the Residual vs. Predictor Plot of the Linear Model for Harris

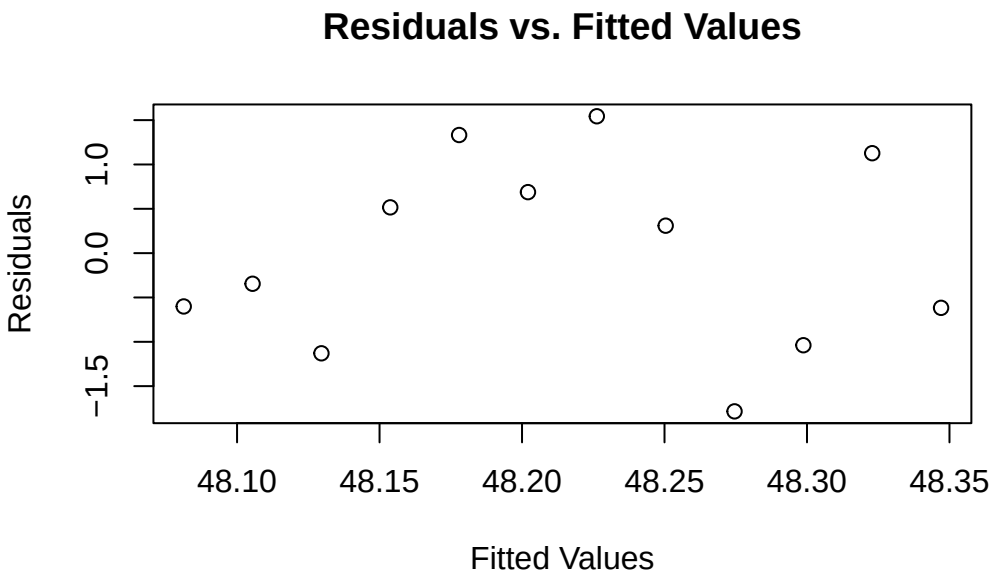

Figure 6: Showing the Residual vs. Fitted Values Plot of the Linear Model for Harris

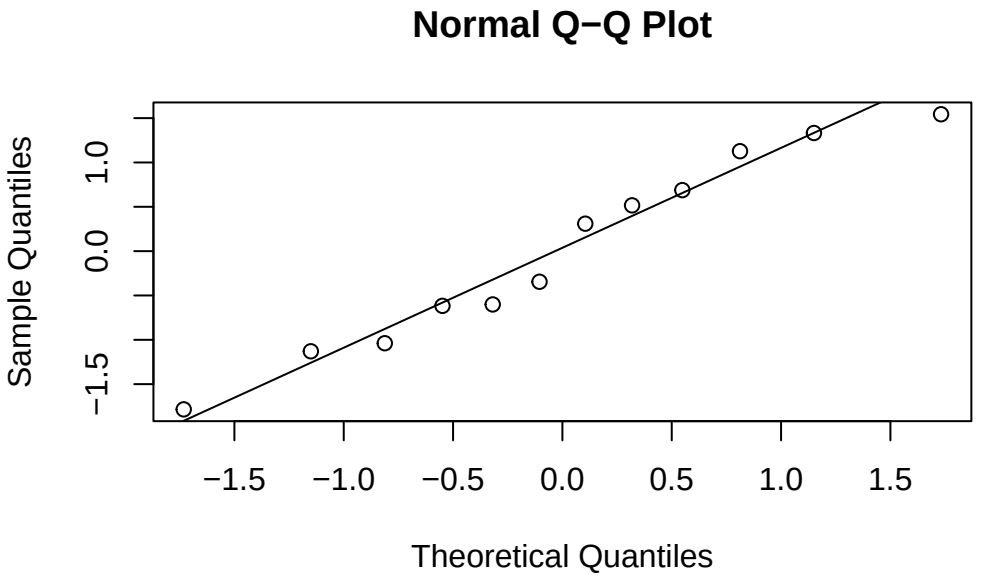

Figure 7: Showing the Normal Q-Q Plot of the Linear Model for Harris

Table 9: Summary Table of the Regression Results for Trump

	(1)
(Intercept)	48.261
	(0.679)
Week	−0.014
	(0.092)
Num.Obs.	12
R2	0.002
R2 Adj.	−0.097
AIC	40.2
BIC	41.7
Log.Lik.	−17.107
RMSE	1.01

A.2.2 Model for Trump

Table 9 shows the summary output of the linear model that predicts the percentage of support for Trump across weeks. Similar to Table 8, the model fits the data well and can provide more precise predictions as indicated by the low AIC, BIC, and RMSE values. The R^2 value is low for this model as well, implying that the time variable, "Week", does not explain much of the variability of pct for Trump.

Residual plots and a Normal Q-Q plot for this linear model are provided with Figure 8, Figure 9, and Figure 10. As there are no observable patterns within the residual plots (e.g. fanning) and the residuals closely follow the QQ-line in the Q-Q plot, this implies that the assumptions of linearity, homoscedasticity, independence, and normality are also satisfied by this model.

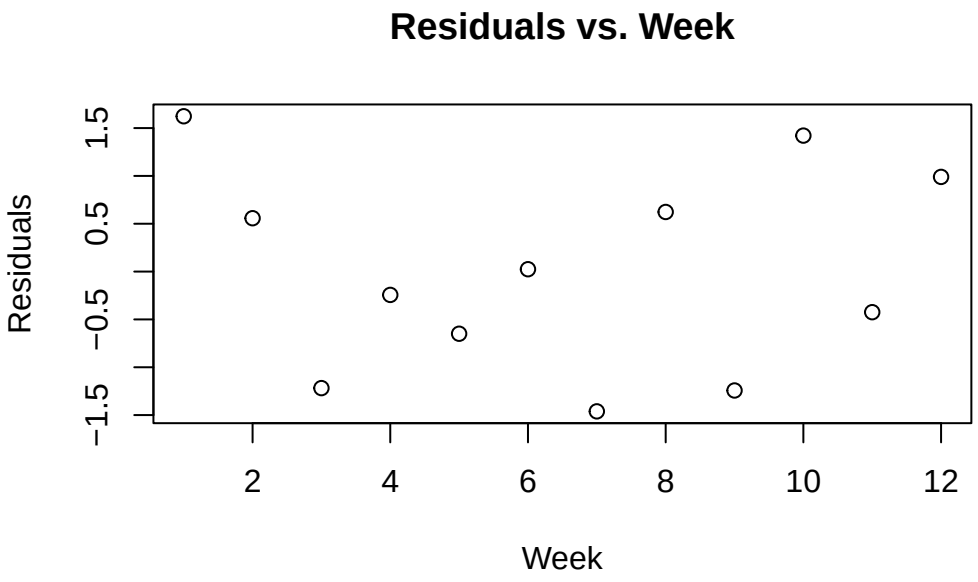

Figure 8: Showing the Residuals vs. Predictor Plot of the Linear Model for Trump

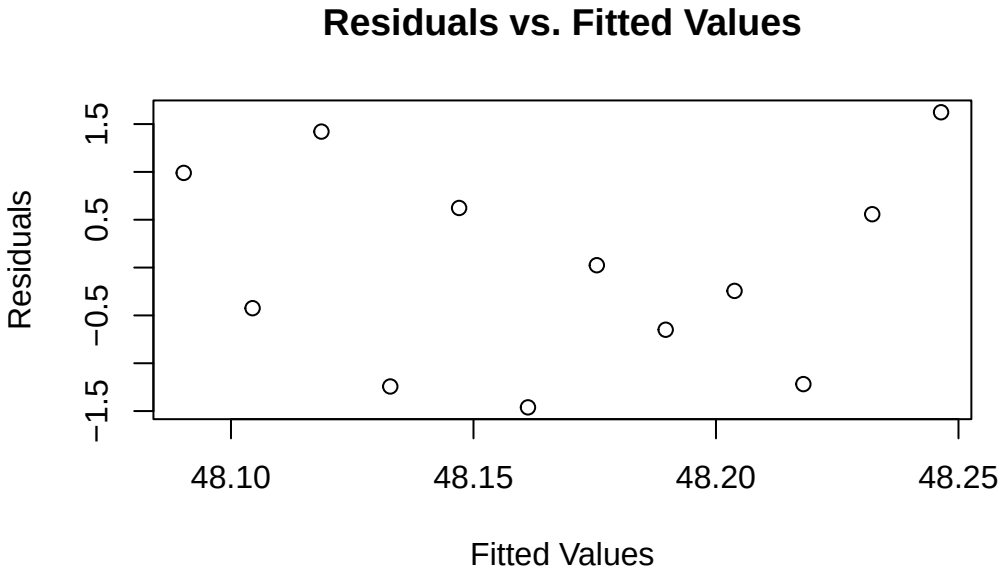

Figure 9: Showing the Residuals vs. Fitted Values Plot of the Linear Model for Trump

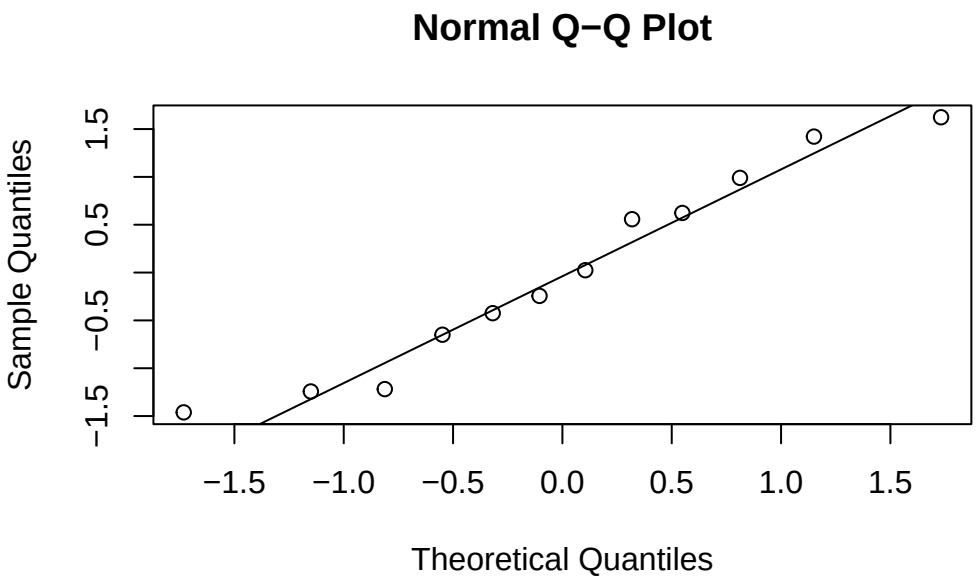

Figure 10: Showing the Normal Q-Q Plot of the Linear Model for Trump

A.3 Pollster Methodology

We selected the polling organization YouGov and discussed its survey methods as well as its main features, strengths and weaknesses. From the data obtained, the population surveyed by YouGov is American voters, especially citizens who are eligible to vote. YouGov's framework is usually participants who voluntarily register and regularly participate in surveys. These panelists express their views in the form of online questionnaires, and the sample is part of YouGov's online panel. In the data, we see some specific stratification information, such as political parties (DEM, REP, etc.), which indicates that YouGov may use stratified sampling to ensure the diversity of the sample. YouGov's sample is recruited through voluntary online panels. Users can actively register to become panel members and accept survey invitations at any time. This recruitment method is non-random, but the cost is relatively low and the number of people is large.

YouGov uses stratified sampling, that is, respondents are stratified according to demographic variables such as age, gender, political party, etc. to ensure that each subgroup in the sample is fully represented. Stratified sampling can ensure that different groups (such as party supporters, different age groups, etc.) are properly represented, avoid the dominance of a single group, improve the accuracy of estimates, and reduce sampling errors. However, stratified sampling may increase sampling errors if the stratification criteria are not properly selected or if there are large individual differences within the strata. Compared with simple random

sampling, the design and implementation of stratified sampling may be more time-consuming and require more resources to determine the stratification and sampling scheme, especially when the population size is large. In the article "On Stratified Sampling for Estimating Coalitional Values", the author A. Saavedra-Nieves explored how to use stratified sampling methods to estimate Owen values and Banzhaf–Owen values in TU-games and analyzed the stratified sampling method. The author divided the players' compatible coalitions into different strata according to size, and then used simple random sampling in each stratum to obtain the estimates. The method was verified and the results showed that the stratified sampling method can more effectively reduce errors and significantly reduce variance than traditional simple random sampling (Saavedra-Nieves 2023). When faced with large-scale data, stratified sampling can significantly improve computational efficiency and accuracy. These methods have wide applicability in practical applications.

YouGov uses weighting to adjust when dealing with non-response issues. When some people do not respond or the response rate of certain groups is low, YouGov will weight the responses of these groups based on demographic data to ensure that the final survey results can more accurately reflect the overall situation. This can help correct the bias caused by the low response rate of certain groups and make the results more representative.

Further, YouGov's questionnaire is answered online, which can quickly obtain a large amount of data, and the population (people who answer the questionnaire) is distributed in various places, which improves flexibility and efficiency. The questionnaire also covers a variety of candidates and parties, and distinguishes the support rate of different parties. The content of the questionnaire can be changed according to different groups to ensure that the survey questions are relevant to the background of the respondents. At the same time, there are some potential problems with the questionnaire. First, because the questionnaire is answered online, some people may not answer the questionnaire seriously, which will affect the accuracy of the questionnaire. Second, because participation in the questionnaire is voluntary, some groups may be under- or over-represented, which will also lead to biased survey results.

A.4 Idealized Pollster Methodology

Survey Form Link: https://forms.gle/S4cyiZNej46zfxq29

A copy of the survey questions can be found in Section A.5.

A.4.1 Introduction

This appendix presents an idealized methodology for forecasting the upcoming US. presidential election within a hypothetical $100,000 budget, minimizing challenges such as non-response bias, sampling difficulties, and response fatigue. This design is in consideration of representatives and accuracy by combining innovative sampling methods and a concise survey questions.

The goal is to accurately capture voter intentions in those swing states, which play an important role in the US. presidential election because these swing voters are more likely to be persuaded are influenced by both campaign efforts and candidate characteristics (Mayer 2007). In the survey, we set a series of screening questions to ensure that only eligible respondents are included in the data set. These questions are aimed at filtering out individuals who do not meet the necessary criteria, such as those under 18, non-U.S. citizens, non-swing-state residents, or people who won't participate in the upcoming election. By implementing these screening questions at the beginning of the survey, we can eliminate ineligible respondents early, which can conserve resources and ensure the quality of the data collected for subsequent questions.

Besides, for the survey questions, we design them focus on minimizing response burden and improving response quality by using a concise and well-structured questionnaire. Studies indicate that shorter surveys yield higher completion rates and better-quality data, which is essential for accurately capturing voter intentions and influence factors (Rolstad, Adler, and Rydén 2011). Thus, the survey is structured to be completed in under 10 minutes. And, the survey's structure places voting intention questions first, followed by key demographic and issue-related questions to avoid priming effects, which is consistent with best practices in polling methodology (Roscoe, Lang, and Sheth 1975).

Moreover, the survey includes a range of questions to analyze different influences on voter decisions in the U.S. presidential election. Firstly, the demographic questions such as age, gender, ethnicity, education, and income level provide insight into how population segments may align with certain candidates, allowing for analysis across key groups such as age and regional demographics, which can reveal trends like youth support or regional shifts. Moreover, questions of voter registration history examine respondents' previous voting habits, helping to distinguish between frequent voters and first-time voters, as past engagement can provide some information when predicting current voting intentions. Political affiliation questions assess respondents' party loyalty or attitudes to other political viewpoints, which is essential for identifying swing voters who might be persuaded by a candidate outside their usual party alignment. And questions of candidate preference and voting intention show respondents' current support for specific candidates and their likelihood of voting, offering a potential turnout, which is crucial for understanding committed versus uncertain voters. And the most important question is to ask respondents to choose some Potential factors that may influence their choice. These different factors are based on different ideas put forward by different candidates (BBC News 2024) such as the form of healthcare and education, immigration regulation, climate changing, trade, tax level, and so on, identifying the top concerns that influence voter behavior. These data provide insights into how different factors effect across demographics and influences candidate preference among specific voter groups.

Overall, this appendix provides a well-designed, evidence-based methodology for forecasting the US presidential election, balancing rigorous data collection with practical budgetary considerations. This methodology is not only adaptable for future elections but also emphasizes

transparency and reproducible, incorporating proven strategies to ensure high data quality and an accurate, representative electoral forecast.

A.4.2 Sampling Approach

A.4.2.1 Target Population

This survey targets the U.S. voting-age population, defined as U.S. citizens aged 18 and older who are residence in one of the U.S. swing states. Respondents are either planning to vote in the presidential election or have already vote. By including both likely voters and those who have already voted, the survey captures comprehensive results of voter preferences, covering both the intentions of those yet to vote and the decisions of early voters. This approach ensures that early voting trends, which has an increasingly significant factor in recent elections, are not overlooked, providing opinions to both anticipated and confirmed voting patterns. The survey focuses on likely voters by screening for high voting intention, enhancing data relevance by filtering out those with no intent to participate. Demographic diversity is also considerate, with stratified sampling used to ensure representation across age groups and demographics, from younger voters (18-24) to older adults (65+). Besides, the survey is for residences in swing states, enabling detailed understanding into regional and demographic factors that could influence the election. By including both new and established voters, the survey reflects a range of voting experiences and motivations, capturing the perspectives of first-time voters alongside those with consistent voting patterns. This approach to the target population produces a representative view of the voting-age population, providing a reliable election forecast.

A.4.2.2 Sample Size and Confidence

The survey will target a sample size of 3,000 respondents to achieve a $\pm 2\%$ margin of error at a 95% confidence level, a standard for ensuring robust national-level predictions. This sample size allows the survey to capture a reliable and comprehensive view of voter intentions and preferences within the broader U.S. voting-age population. The $\pm 2\%$ margin of error means that if the survey were conducted multiple times under the same conditions, the observed results would be expected to fall within 2 percentage points of the actual population values 95% of the time. This level of precision enhances the survey's capacity to detect meaningful differences across voter segments and to anticipate shifts in public opinion with confidence.

With a \$100,000 budget, the choice of 3,000 respondents represents an optimal balance between statistical reliability and cost-efficiency, enabling the inclusion of detailed demographic and geographic breakdowns without exceeding budget constraints. This sample size allows for adequate subgroup analysis, enabling examination of various demographics such as age, race, education, income of residences in swing states. For example, the sample allows for comparisons between younger and older voters, urban and rural residents, and high- and low-income voters, each of which may have unique voting behaviors. This granularity in data helps refine election forecasts and captures diverse perspectives within the electorate.

Increasing the sample size beyond 3,000 could further decrease the margin of error, improving precision in subgroup analyses and providing an even more detailed look at specific voter groups. However, larger samples would require additional resources, surpassing the set budget of $100,000. Given these budgetary constraints, a sample of 3,000 respondents provides statistically sound data that can guide reliable election predictions while maintaining cost-effectiveness and broad representative.

A.4.2.3 Stratified Sampling Approach

To ensure representative, the survey will divide the sample across key demographic factors. Stratifying across these variables allows the survey to more accurately reflect the diversity of the U.S. voting-age population, as each subgroup is proportionally represented. This approach minimizes biases by ensuring that each demographic group, particularly those that might have unique voting behaviors, is included in the sample in accordance with its actual share of the population. For example, by accounting for varying education levels and employment types, the survey can capture distinctions in political preferences that may emerge among blue-collar versus white-collar workers, or those with different levels of formal education.

To implement stratified sampling within swing states, which often experience highly competitive races because small shifts in voter behavior can significantly affect the overall election outcome. The survey divides the sample into key demographic and political subgroups, including age, gender, race, education level, income, and political affiliation. Each subgroup will be represented in proportion to its actual population within each swing state, based on recent census and voter registration data. For example, if 30% of a state's population is within the 18-24 age range, the sample will ensure that 30% of respondents from that state are also in this group, accurately reflecting voter intentions across age demographics. This proportional approach captures the unique preferences of each demographic, preventing skewed results that might arise if certain age groups were over or underrepresented. Additionally, smaller or historically underrepresented groups, such as certain racial minorities or lower-income populations, may be over sampled to guarantee adequate representation, which is particularly important in areas where shifts in these groups' voting behaviors could impact state outcomes. Geographic diversity is also addressed, with sampling from urban, suburban, and rural areas within each swing state to capture regional differences in voter preferences. Once data is collected, post-stratification weighting will adjust for any discrepancies, ensuring the final results remain representative of each state's population. This comprehensive stratification approach allows the survey to capture the full spectrum of voter perspectives across critical swing states, providing an accurate and detailed forecast of voter behavior in these pivotal regions.

A.4.2.4 Bias

Over-reliance on online surveys may get bias results toward younger respondents who are more Internet savvy, potentially under representing older and rural voters who are less likely to answer questions online. Similarly, only using mail surveys may also lead to bias. Younger

voters and urban residents may be less responsive to mailed invitations because they tend to rely more on digital communication and overlook physical mail, and frequent moves and high mail volume in urban areas can cause invitations being lost or disregarded. This potentially causes the sample skews toward older or more traditional respondents.

To address these limitations, some portion of the budget will be allocated to reaching diverse demographics through a mix of phone, in-person, and online methods, ensuring a more balanced and inclusive sample. Based on the study, the form of telephone survey will account for a larger proportion because this method attracts more attention and costs less than in person survey and mailing survey (Roscoe, Lang, and Sheth 1975). Additionally, to mitigate non-response bias across all modes, providing some incentives will be offered to encourage participation from harder-to-reach groups, improving response rates and enhancing data reliability, which also mentioned and proved by (Roscoe, Lang, and Sheth 1975).

A.4.3 Recruitment Strategy

A.4.3.1 Mixed-Mode Recruitment

The survey employs a mixed-mode recruitment strategy, integrating online, telephone, messages (SMS), and in-person surveys to maximize demographic and geographic inclusivity. This approach is informed by research showing that mixed-mode sampling can effectively address the limitations of single-mode surveys. Online surveys capture younger respondents, while telephone surveys engage older and rural respondents who may be less inclined to participate online. In-person surveys further enhance reach, particularly among individuals who may be challenging to contact through other methods. With a target sample size of 3,000 respondents, this approach ensures a low margin of error and supports detailed analysis across swing-state regions. The recruitment budget is allocated across four main methods: phone recruitment, online panel recruitment, message invitations, and in-person surveys.

Firstly, the recruitment strategy allocates $10,000 to phone recruitment, aiming to reach older voters, particularly those aged 65+ and rural residents who may be less likely to participate online. By including both fixed line and mobile, this method maximizes reach range, with an anticipated response rate of 10-15%. Invitation by making phone call doesn't cost so much budget.

Then, online panel recruitment, with a $36,000 budget, focuses on younger who is around 18-44 years old and urban respondents, which are more likely to engage through digital platforms. Besides, partnering with reputable providers like YouGov ensures access to large, pre-screened pools of eligible voters, with expected recruitment of around 1,000 respondents, consistent with typical online response rates. Although online panel itself doesn't so costly, we allocate the most budget to this method. This is because we need a lot of money to work with reliable and responsible partners.

Moreover, we allocate $10,000 budget to send invitations by messages method, target mobile-friendly users, especially younger voters (ages 18-34) who prefer quick, mobile access. By linking to an online survey, SMS invitations reach both urban and rural areas and are expected to yield about 500 respondents. Invitation by sending message also doesn't cost so much budget, so we assume that this cost as much as making phone for invitation.

Finally, in-person surveys are allocated $35,000 to increase inclusivity by engaging voters who may not participate through other method. These surveys are conducted at key locations such as public gatherings and events within swing states, providing access to hard-to-reach demographics, with an anticipated yield of 200-300 respondents. This approach ensures a diverse, representative sample, maximizing inclusivity and minimizing non-response bias. For in-person surveys, there is the high cost due to the need for trained interviewers, travel cost, and logistical assistance. Staffing expenses are also significant, as interviewers must be paid and may work evenings or weekends to reach respondents. Travel and setup costs, including transportation and equipment, increase the expenses, especially when covering multiple locations. Additionally, venue fees or permits for conducting surveys in public areas, as well as administrative support for planning and data management, further increase costs. Despite these highly expenses, in-person surveys are still essential for reaching underrepresented groups, improving inclusivity, and reducing biases in our survey.

Together, these four recruitment methods allow for a balanced and inclusive sample, minimizing non-response bias and ensuring representation across demographics. This comprehensive strategy strengthens data quality and support a highly accurate forecast of election outcomes.

A.4.3.2 Incentives for Participation

To improve response rates, especially among hard-to-reach groups, we will offer 3 dollars digital gift cards as an incentive for survey completion. With a budget of $9,000 allocated for approximately 3,000 respondents, these incentives are designed to encourage participation from groups that may be less likely to respond, such as low-income individuals and rural populations. This approach helps ensure a more inclusive sample by motivating participation across diverse demographics, ultimately enhancing the representative and reliability of the survey data. Providing incentives method is also proved in the (Roscoe, Lang, and Sheth 1975).

A.4.4 Data Validation

A.4.4.1 Weighting Adjustments

To ensure the reliability of collected data, data validation techniques and post-stratification weighting are employed. Automated screening tools such as CAPTCHA and IP address monitoring protect online surveys from multiple submissions or bot entries. In addition, attention-

check questions are incorporated to detect and filter out low-quality responses. Once data collection is complete, results are weighted to reflect known demographic characteristics of the likely voting population, addressing potential biases and aligning results with turnout patterns observed in past elections. This approach draws from the findings of (Kennedy et al. 2018), which underscore the importance of post-stratification adjustments for factors like education level to correct biases noted in recent election cycles.

A.4.4.2 Fraud Detection

To ensure high-quality responses and detect potential fraud, the survey employs several robust techniques. The first technique is consistency checks, which placed validation questions to cross-verify answers. For example, responses regarding party affiliation, voting intention, and specific issue preferences are cross-checked to identify inconsistencies that could indicate inattentive, unreliable, or fraudulent responses. If a respondent's answers show contradictions such as expressing support for a candidate but having inconsistent party alignment, their response may be flagged for further review.

Additionally, the survey utilizes a re-contacting method to verify data authenticity. A random subset of respondents will be selected for follow-up contact, where they are asked to reconfirm key responses from their original submission. This follow-up can help detect respondents who may have provided random or dishonest answers initially, as they may be unable to accurately repeat their responses. By confirming data accuracy through re-contacting, the survey strengthens the reliability of collected data, minimizing the impact of potential fraudulent responses. overall, these two methods enhance the survey's overall data integrity and ensure that the results are reliable and high-quality.

A.5 Copy of the survey questions

Voter Eligibility and Participation:

1. Are you a U.S. citizen eligible to vote in the ongoing U.S. presidential election? * Yes No
2. Are you at least 18 years old, making you eligible to vote in the ongoing U.S. presidential election? * Yes No
3. Do you currently reside in one of the following key U.S. swing states? (The order of the options is in alphabetic order) * Arizona Georgia Michigan Nevada North Carolina Pennsylvania Wisconsin No, I do not reside in any of these states
4. Have you already voted in the 2024 U.S. presidential election? * Yes, I have already voted No, but I plan to vote No, and I do not plan to vote

Voting Information:

1. Who did you vote for or plan to vote for in the ongoing U.S. presidential election? (The order of the options is in alphabetic order) * Donald Trump Kamala Harris Undecided
2. How did you vote or plan to vote in this election? * By mail Early voting in person Voting in person on Election Day Other:
3. What are the most important issues that will influence your vote? (Select up to 7) * Private Healthcare Public Healthcare Strictly control immigration Support Legal Immigration Reduce Environmental Regulation Strengthen Environmental Regulation Private Education Public Education Increasing Tax Reducing Tax Regulation of Guns Support to hold guns Anti-abortion Rights Pro-abortion Rights Other:
4. On a scale of 1 to 5, how motivated were you to vote in the ongoing election? * 1: Not motivated at all 2: Somewhat unmotivated 3: Neutral 4: Somewhat motivated 5: Very motivated
5. Have you voted in previous presidential elections? * Yes, I voted in the last election (2020) Yes, I voted in a previous presidential election but not in 2020 No, I have never voted in a presidential election

Demographics:

1. What is your age group? * 18-24 25-34 35-44 45-54 55-64 65+ Prefer not to say
2. What is your gender? * Female Male Non-binary Prefer not to say Other:
3. What is your race/ethnicity? (The order of the options is in alphabetic order) * American Indian or Alaska Native Asian Black or African American Hispanic, Latino or Spanish Origin Middle Eastern or North African Multiethnic Native Hawaiian or other Pacific Islander White Prefer not to say Other:

4. Are you married, widowed, divorced, separated, or never married? * Married Widowed Divorced Separated Never married Prefer not to say

5. What is the highest level of school you have completed or the highest degree you have received? * Less than high school degree High school degree or equivalent Some college but no degree Associate degree Bachelor degree Graduate degree (e.g., Masters, PhD, M.D) Prefer not to say Other:

6. Which of the following categories best describes your employment status? * Employed part time Employed full time Not employed, looking for work Not employed, not looking for work Retired Disabled, not able to work Prefer not to say Other:

7. What type of job or industry are you currently working in? (The order of the options is in alphabetic order) * Arts, Entertainment, or Media Construction or Trades (e.g., carpenter, electrician, plumber) Education (e.g., teacher, professor, administrator) Finance or Banking Government or Public Service Healthcare (e.g., doctor, nurse, healthcare assistant) Hospitality (e.g., hotel, restaurant, tourism) Manufacturing or Production Retail or Sales Retired Self-employed or Freelancer Student Technology (e.g., IT, software development, engineering) Transportation or Logistics (e.g., driver, warehouse, shipping) Unemployed Prefer not to say Other:

8. What is your household income? * Under $30,000 $30,000 - 60,000 $60,000 - 100,000 Over $100,000 Prefer not to say

9. Which of the following describes your current living situation? * Own my home Rent my home Prefer not to say Other:

10. How would you describe your political viewpoints? * Very liberal Slightly liberal Moderate Slightly conservative Very conservative Prefer not to say

11. Which political party do you currently identify with? (The order of the options is in alphabetic order) * Communist Party Constitution Party Democratic Party Green Party Independent Party Libertarian Party Pirate Party Republican Party I do not identify with any political party Prefer not to say Other:

References

BBC News. 2024. "What Are Harris and Trump's Policies?" 2024. https://www.bbc.com/news/articles/cwy343z53l1o.

FiveThirtyEight. 2024. "538 - Election Polls, Polictics, and Analysis." https://projects.fivethirtyeight.com/polls/.

Jackman, Simon. 2005. "Pooling the polls over an election campaign." *Australian Journal of Political Science* 40 (4): 499--517. https://doi.org/10.1080/10361140500302472.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, et al. 2018. "An Evaluation of the 2016 Election Polls in the United States." *Public Opinion Quarterly* 82 (1): 1–33.

Mayer, William G. 2007. "The Swing Voter in American Presidential Elections." *American Politics Research* 35 (3): 358–88.

Moussaïd, Mehdi, Juliane E. Kämmer, Pantelis P. Analytis, and Hansjörg Neth. 2013. "Social influence and the collective dynamics of opinion formation." *PLoS ONE* 8 (11). https://doi.org/10.1371/journal.pone.0078433.

Müller, Kirill, and Jennifer Bryan. 2020. *here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

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. Elliott Morris. 2023. "538's polls policy and FAQs." https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Rolstad, Sindre, John Adler, and Anna Rydén. 2011. "Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-Analysis." *Value in Health* 14 (8): 1101–8.

Roscoe, A Marvin, Dorothy Lang, and Jagdish N Sheth. 1975. "Follow-up Methods, Questionnaire Length, and Market Differences in Mail Surveys: In This Experimental Test, a Telephone Reminder Produced the Best Response Rate and Questionnaire Length Had No Effect on Rate of Return." *Journal of Marketing* 39 (2): 20–27.

Saavedra-Nieves, Alejandro. 2023. "On stratified sampling for estimating coalitional values." *Annals of Operations Research* 320 (1): 325--353. https://doi.org/10.1007/s10479-022-05044-0.

V, Arel-Bundock. 2022. *modelsummary: Data and Model Summaries in R. Journal of Statistical Software*. Vol. 103. https://doi.org/doi:10.18637/jss.v103.i01.

Wickham, Hadley. 2011. *testthat: Get Started with Testing. The R Journal*. Vol. 3. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. "ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "dplyr: A Grammar of Data Manipulation." https://cran.r-project.org/web/packages/dplyr/index.html.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2019. "TinyTeX: A Lightweight, Cross-Platform, and Easy-to-Maintain LaTeX Distribution Based on TeX Live." *TUGboat* 40 (1): 30–32. https://tug.org/TUGboat/Contents/contents40-1.html.