sequence of 20 days beyond the dataset's last recorded date. Because we want to predict the result in November 5. This expansion allows for forecasting near-term polling trends. Also, capturing potential shifts in candidate support leading up to election day. The model was trained on existing data and applied to this extended date range using predict(model, newdata = ., type = "response").

Our model setup and variable selection are as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1(\text{Date}) + \beta_2(\text{Candidate})$$

where: -

$$p_i \text{ represents the probability of support for candidate i at a specific date.}$$

-

$$\beta_0 \text{ is the intercept, representing baseline support.}$$

-

$$\beta_1 \text{ captures changes in support over time (date effect).}$$

-

$$\beta_2 \text{ is a categorical variable accounting for each candidate.}$$

3.1.1 Model justification

The choice of a logistic regression model was guided by the nature of polling data. Nature of polling data represents proportions that naturally bound between 0 and 1. The logistic function's S-shaped curve effectively models the relationship between time and polling percentage, accommodating non-linear trends that may arise as election day nears. Firstly, candidate names should be included as a categorical predictor in the model. It allows us to capture individual impact in support rates by candidate. In addition, a transformation from date to numeric is made on end_date in order to get a numerical format. It is called the end_date_numeric. It helps put date as a predictor in the model. By structuring the model this way, it can predict the percentage of the poll by both the impact of candidates and trends over time.

4 Results

The analysis of polling data across key swing states for the 2024 U.S. Presidential Election reveals insights into the levels of support for candidates Donald Trump and Kamala Harris. The findings are presented through three primary visualizations: the average polling percentage by state, the distribution of polling percentages, and the difference in average polling percentage between the two candidates.

4.1 Average Polling Percentage by State

From Figure 1, we can see that in the swing states of Arizona, Georgia, and North Carolina, Trump holds a narrow lead. In contrast, Harris leads in Pennsylvania, Wisconsin, Nevada, and Michigan. When looking at the overall average across these states, Harris slightly edges out Trump, indicating her overall advantage in these swing areas. In our assumed electoral map, this projects Kamala Harris over the edge with 276 electoral votes as opposed to Donald Trump's 262. This shows that our model predicts Kamala Harris to win the 2024 Presidential Election if she is able to maintain her lead over all of Pennsylvania, Wisconsin, Nevada, and Michigan.

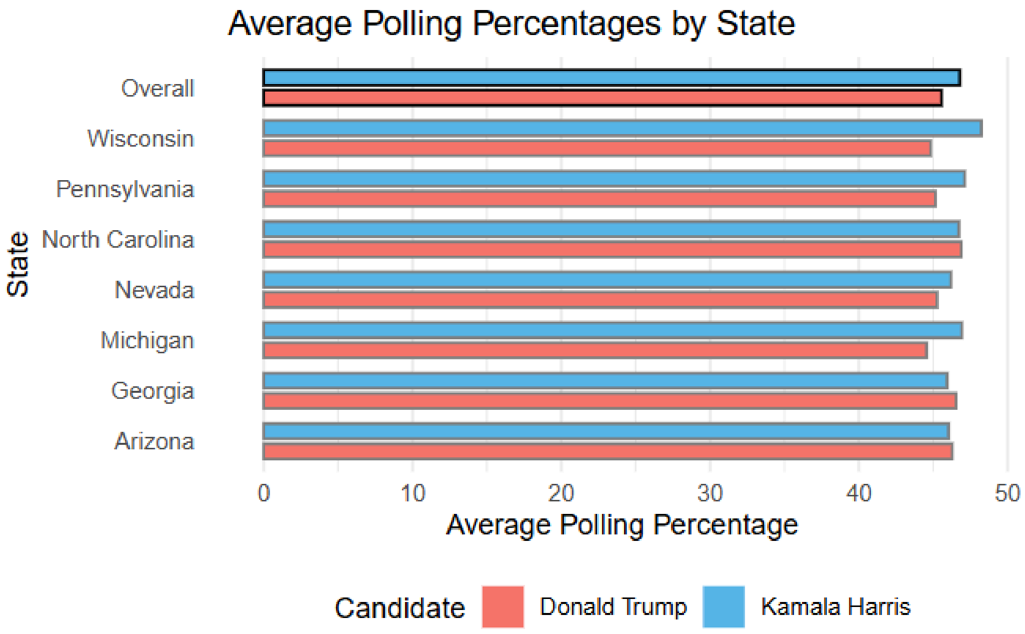

Figure 1: Average Polling Percentage by State

4.2 Polling Percentage Distribution

The box plot in Figure 2 demonstrates the range and consistency of support for both candidates across each state. Trump shows a wider range of support, reflecting variability, while Harris has relatively steady polling percentages across states, maintaining higher consistency. Harris's support is less variable and generally centers around higher polling percentages compared to Trump's.

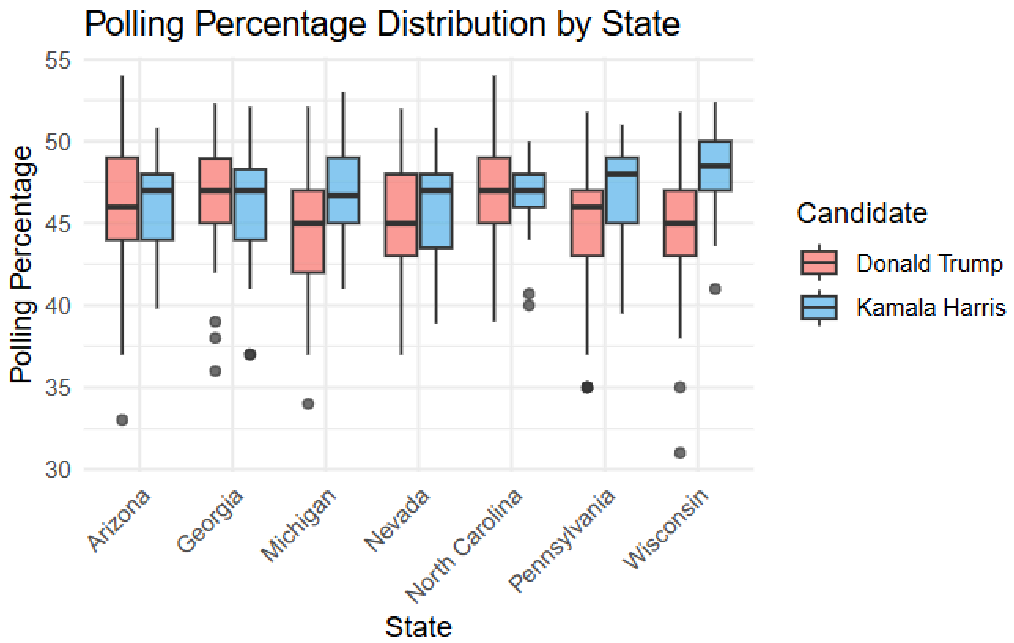

Figure 2: Polling Percentage Distribution

4.3 Difference in Average Polling Percentage

Figure 3 highlights the difference in average polling percentages by state, with positive values indicating Trump's lead and negative values indicating Harris's. Trump's lead in the three states (Pennsylvania, Wisconsin, Nevada, and Michigan) is relatively narrow, with margins of less than 1 percent. In contrast, Harris's lead in the states favoring her (Pennsylvania, Wisconsin, Nevada, and Michigan) ranges from 1 to 4 percent, indicating her stronger support in the states where she leads. Consequently, this difference in average polling percentages show that Kamala Harris' lead of 276 electoral votes is supported by a reliable margin and Donald Trump's lead in Arizona, Georgia, and North Carolina is turbulent.

4.4 Swing State Polling Prediction Model

Figure 4 shows the trend of support rate changes for each candidate before the upcoming election, as well as the model's predicted values for the next 20 days.We predicts that Kamala Harris may win based on our model. However, this shows the trend of support rate changes for each candidate before the upcoming election, as well as the model's predicted values for the days until the election. This support rate changes model exhibits a higher rate of growth in Donald Trump's share of support. Although Kamala Harris is in the lead, Donald Trump's

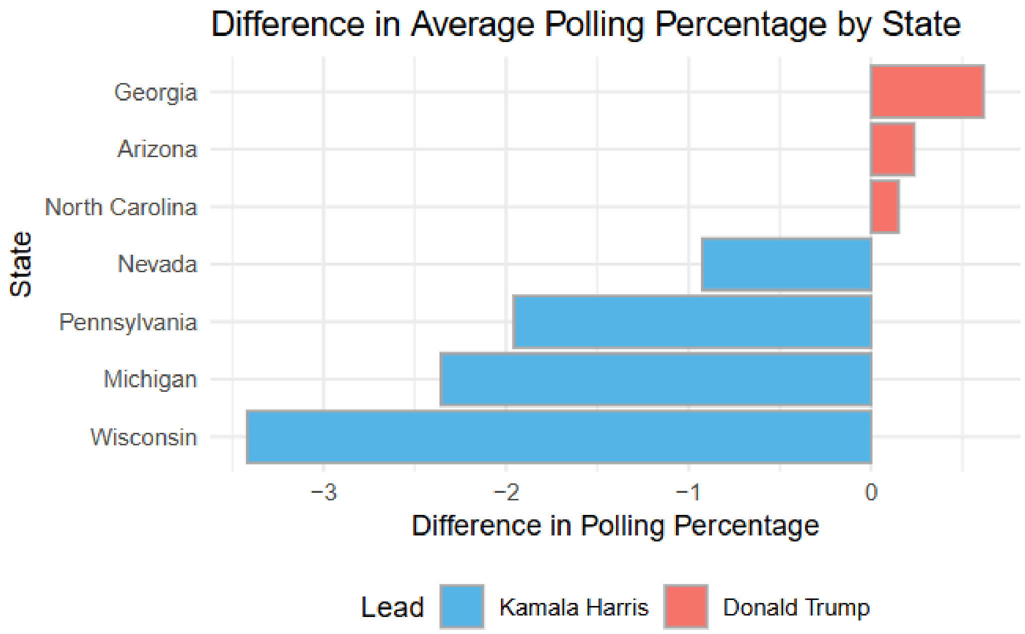

Figure 3: Difference in Average Polling Percentage

growth of support rate is an element that could stifle Harris' lead and drop her below her required 270 electoral votes.

5 Discussion

5.1 Turbulent Margins

Based on our results from Figure 1 predicted that Kamala Harris to win the presidential election with 276 electoral votes. Furthermore, Figure 2 showed that her lead in the states of Pennsylvania, Wisconsin, Nevada, and Michigan are stronger than Donald Trump's support in Pennsylvania, Wisconsin, Nevada, and Michigan. From Figure 2, we can also observe that Kamala Harris' lead in Nevada seems to be the weakest. In the framework of the electoral college system, losing Nevada to Donald Trump would not change the results as Nevada holds 6 electoral votes, which means that Kamala Harris would still possess the necessary 270 to win the election. However, this is not the case if Kamala Harris loses support in any of the three other states of Pennsylvania, Wisconsin, and Michigan as this would place her beneath the 270 seat threshold. In fact, losing any one of these states would result in Donald Trump's victory. Figure 3 showed that the support rate of Donald Trump is increasing faster than that of Kamala Harris', threatening Harris' lead over Pennsylvania, Wisconsin, and Michigan.

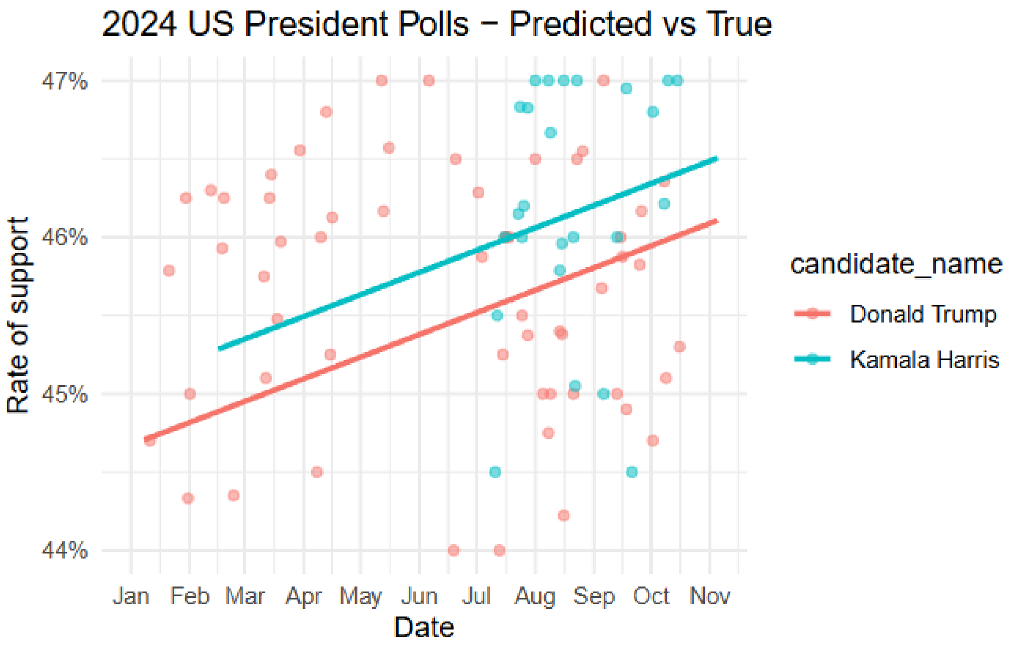

Figure 4: Swing State Polling Prediction Model

5.2 Impact of Key Demographics and Policy Appeals

Harris's lead in several swing states may be influenced by her campaign's appeal to key demographic groups, particularly immigrants. The Democratic platform has emphasized social and economic benefits for immigrant communities, which may have strengthened her base in states with significant immigrant populations. Policies that promise support for healthcare, education, and other public services are likely contributing to Harris's appeal among these groups, solidifying her lead in states like Nevada and Michigan. This demographic support provides Harris with a strategic advantage, as these communities may play a crucial role in voter turnout and loyalty.

5.3 Influence of Endorsements and High-Profile Backing

Recently, high-profile figures like Elon Musk have publicly thrown their support behind Trump, leveraging their influence to rally support for his campaign (Zahn 2024). Musk's endorsement could invigorate Trump's base, especially among Musk's followers and those engaged in the fields of business and technology. This visible backing from such influential figures has the potential to sway undecided voters and influence last-minute decisions, particularly in swing states where Trump's lead remains narrow.

Musk's endorsement might not only energize Trump's existing supporters but could also lead to a boost in campaign funds and social media engagement, further strengthening Trump's position. In closely contested states, where a few thousand votes can make a significant difference, this level of backing from a high-profile figure could be crucial. If Musk's involvement draws in additional resources and amplifies Trump's visibility, it may ultimately impact the final outcome of the election on November 5th.

5.4 Limitations of the Analysis: Focus on Swing States

While the analysis focuses on key swing states, it does not consider the broader national picture, which includes states with clear, one-sided advantages for either candidate. This narrowed focus on swing states provides insights into swing areas but does not give a complete picture of the national landscape. Therefore, while Harris appears to have an edge in these swing states, it is unclear whether she is leading overall. The situation in solidly red or blue states could still shift the electoral balance, meaning Trump could potentially gain ground through strong performances in traditionally Republican states.

5.5 Potential Influence of Late Campaign Efforts

With a few days remaining before the election, both campaigns are likely to intensify their outreach efforts, particularly in swing states. Trump's recent increase in targeted advertisements and campaign rallies in swing states could energize his supporters, narrowing Harris's lead. On the other hand, Harris's campaign may focus on mobilizing key demographics, such as young and minority voters, to secure her advantage. Late-breaking events, public debates, or controversies could sway undecided voters and ultimately impact the election results in these states.

5.6 Limitations and Next Steps

This analysis is limited to examining current polling data in swing states and does not account for unexpected shifts in voter sentiment or turnout on election day. Additionally, focusing exclusively on swing states omits the influence of non-swing states that may still play a role in the electoral vote tally. Future analyses could incorporate national polling data to provide a more comprehensive perspective and examine demographic breakdowns within each state to understand which groups are driving support for each candidate. Integrating data on recent campaign activities, endorsements, and voter sentiment on pressing issues could also improve the predictive accuracy of election outcomes.

Appendix A

A Sampling Methodology Employed by New York Times and Siena College:

Our analysis and prediction utilizes data collected by a single pollster and it is important to scrutinize the sampling methodology behind the sampling methods used to determine the extent of the external validity of our predictions. Primarily, the data gathered by New York Times was done so with the collaboration with Sienna College through live-interviewers reaching voters by telephone.

A.1 Target Population and Sampling Frame

The poll targets registered voters in the United States, with a particular focus on voters in swing states for presidential races. Consequently, the target population is registered voters of the United States of America and the sampling frame is the voter file, a national list of registered voters from which NYC employs random sampling to create their pollsters (The New York Times 2024). This voter file contains information on more than 200 million Americans that includes information about their party affiliation, voting history, and basic demographic information. From this sampling frame, respondents are randomly selected while balancing key demographic and political characteristics through both front-end sampling and post-survey weighting pursuant to appropriate representation of the entire target population (The New York Times 2024).

A.2 Recruitment Methods

The poll relies on random digit dialing to reach voters, predominantly via cell phones. While landlines are also called, the dominance of cellphone usage aligns with contemporary communication trends. This results in their rendition of random sampling, which contributes to the external validity of data. However, reaching voters by phone has become more difficult due to low response rates requiring many attempts to secure responses from certain individuals (response rates are as low as 2% at NYC). In fact, according to an article by Emerson College, telephone polling has become entirely obsolete (Kimball, Mumford, and Taglia 2024). Despite this, the phone methodology is still seen as effective for its ability to quickly reach a diverse voter pool across various states as well as seen as the golden standard for polling. Although this can be seen as random sampling, the concern of low response rates and reliance on dialing to reach voters raises concerns of selection bias. There are questions as to whether a certain demographic will be more responsive to being reached through a cell phone. Moreover, voters that are reached over the phone may feel the urge to provide more 'socially desirable' responses that are not reflective of their true opinions.

A.3 Non-Response Remedies

With a response rate of 2%, there is valid concern over non-response bias underrepresenting certain demographics. The poll addresses non-response through a combination of front-end sampling adjustments (calling harder-to-reach voters) and post-survey weighting. Weighting is applied to ensure that the final sample aligns with the demographic and political composition of the electorate, based on voter file data. Non-response bias is a persistent challenge, and even though weighting helps, it's impossible to be certain that respondents are fully representative of those who do not respond. Experiments with alternative methods, such as mail surveys with incentives of cash vouchers have shown to increase response rate up to 30% (The New York Times 2024).

B Sampling Approach: Stratified Sampling

B.1 Methodology

The Times/Siena Poll employs a stratified random sampling approach to ensure that the sample accurately represents the diversity within the target population of registered voters in the United States. This method involves dividing the population into distinct subgroups, or "strata," based on key characteristics such as state, age, gender, race, education level, and political affiliation. For example, in presidential elections, particular emphasis is placed on voters from swing states like Arizona, North Carolina, Georgia, and Pennsylvania. By focusing on these battleground states, the poll aims to capture the opinions of voters who are most likely to influence the election outcome. Once the strata are established, respondents are randomly selected within each group, ensuring that all demographic groups are appropriately represented in proportion to their actual distribution within the electorate. Furthermore, post-survey weighting is employed to more closely resemble this distribution.

B.2 Advantages

Stratified sampling offers the advantage of higher precision in representing the population's diversity, as each subgroup is proportionally represented. By breaking down the sample into distinct groups, the Times/Siena Poll can adjust for demographic variations between states or voter groups, which is particularly useful in a diverse electorate. This approach helps to minimize sampling error by ensuring that all significant voter subgroups are included, thus increasing the reliability of the poll's predictions. It is standard practice for them to sample 1,000 respondents for their polls (The New York Times 2024). Using stratified sampling to have this sample size closely resemble that of the entire target population cuts costs and provides efficient real-time data.

B.3 Stratified Sampling over Random Sampling

If the methodology randomly sampled from the target population without regard for the distribution of demographics in the state, significant increases to the sample sizes have to be made. Random sampling will have largely less precision in representing the distribution of different demographics within a given state unless a large enough sample size were to be drawn. Although it is ideal to have probability based sampling increasingly resemble the demographic distribution quota with growing sample sizes, it is unfeasible to conduct such a large scale poll as this approach would encounter a limitation of resources. With declining response rates and the difficulty of reaching voters, New York Times/Siena's nonprobability sampling is a more realistic and efficient method of polling than probability sampling is (Alexander 2023). As such, the stratified sampling approach allows New York Times/Siena to represent the entire target population with a sample size of 1,000.

B.4 Challenges and Trade-offs

Although stratified sampling improves representativeness, it presents certain challenges in execution. First, identifying and reaching a sufficient number of respondents from harder-to-reach groups (such as young voters, minority populations, and those without reliable phone access) can be difficult, especially given the 2% response rate. To counterbalance the low response rates, the poll requires multiple call attempts and careful post-survey weighting to adjust for any potential underrepresentation. Additionally, while phone polling offers broad geographic reach and a standardized mode of data collection, it limits the sample to voters with phone access who are willing to participate in phone interviews, potentially introducing selection bias if certain demographics are less reachable by phone. Although New York Times/Sienna College employs random digit dialing from their sampling frame, their rendition of sample recruitment requires them to sample individuals according to quotas, resulting in nonprobability sampling.

B.5 Evaluation of Questionnaire Design in the Times/Siena Poll

B.5.1 Strengths

The Times/Siena Poll's approach to questionnaire design reflects a careful effort to craft unbiased, relevant, and inclusive questions. By tailoring questions to current events and topics of widespread interest, the poll aims to capture public sentiment on issues that directly affect voters' attitudes, such as economic conditions or recent political debates. This responsiveness to current events strengthens the poll's relevance and allows it to track shifts in opinion over time. The poll's commitment to neutrality is a significant strength. Each question is designed to ensure that respondents, regardless of their political views, feel that their perspectives are fairly represented. The team behind the poll spends substantial time crafting response options to reflect a balanced spectrum of views, minimizing potential response bias and enhancing the questionnaire's reliability across different demographic and ideological groups (The New York Times 2024). Clarity is also a central focus in their questionnaire development. The team emphasizes ensuring that each question is clear and unambiguous, minimizing the risk of misinterpretation. This dedication to straightforward, widely understandable questions improves the validity of responses, as it reduces the chance that respondents interpret questions differently based on background or knowledge level.

B.5.2 Limitations

Despite the strengths of the Times/Siena Poll's questionnaire design, there are some limitations. One notable concern is that, although the poll tries to avoid leading questions, any standardized questionnaire may still unintentionally prime respondents or introduce subtle biases, especially when questions involve sensitive or polarizing topics. Although the pollster tries to avoid attrition bias by keeping each interview less than 15 minutes, in seeking to make questions clear and concise, the questionnaire may not capture the full nuance of voter opinions, especially on multifaceted issues like economic policy or social justice. While the poll strives to capture real views without introducing new ideas, respondents are still prone to a social desirability bias. This is especially possible in phone interviews, where respondents may be more inclined to provide answers they perceive as socially acceptable rather than expressing their true opinions. Although the design aims to be fair, the influence of an interviewer's presence may subtly affect responses (Roger Tourangeau and Rasinski 2000).