Trump vs. Harris: A 2024 US Presidential Elections Modelling Forecast*

Trump trumps Harris to become the next President of the United States

Sophia Brothers   Deyi Kong   Rayan Awad Alim

November 4, 2024

This paper predicts the 2024 U.S. Presidential election. We utilizes linear regression model deploying polling data and demographic factors to predict both popular vote and Electoral College outcomes. The findings indicate a close race between Donald Trump and Kamala Harris, and we predict Trump will secure the Electoral College and Popular Vote due to key wins in tight races for the Battleground States. The results emphasize the role of state-level dynamics, particularly in Battleground states, in determining election outcomes.

Table of contents

---

1 Introduction

The U.S. presidential election is a pivotal moment not only shapes domestic political landscape but also far-reaching its consequences on a global scale. Republicans and Democrats held primary elections across the country earlier in 2024 that eventually filtered down to the main candidates: Donald Trump and Kamala Harris. The polarized political climate and shifting public opinion have made predicting outcomes increasingly challenging, where winning key swing states is important to securing victory.

The main estimand is the projected support percentage and each candidate's winning probability in terms of state-level victories and Electoral College votes. This paper utilizes linear regression model incorporates numeric value of the credibility of pollsters, number of respondents, poll's recency to estimate the probability of each candidate winning in individual states. The objective is to translate these probabilities into a forecast the overall winner of the election. Our findings show a close race, with Trump holding a slight edge in both the popular vote and Electoral College due to projected victories in key swing states.

Trump's path to victory hinges on vital wins in states such as Pennsylvania, Georgia, and Wisconsin, where his projected margins are narrow but statistically significant. Meanwhile, Harris secures strong performances in populous Democratic strongholds like California and New York, but these do not offset Trump's advantage in less populous yet vital battleground states. This result underscores the significant influence of swing states and the Electoral College system in determining the outcome, particularly in closely contested elections. This contributes to a broader understanding of how statistical modeling can inform political strategy and election analysis.

The structure of the paper is as follows: Section 2 outlines the data sources and variables considered, followed by the model setup in Section 3.1 and justification in Section 3.2. The results in Section 4 presents the key findings of the analysis, with a discussion on the implications.

2

Section 5 then discusses potential limitations and suggestions for future research. Section 6 provides additional detailed information about the data, model and methodology.

2 Data

2.1 Overview

The data used in this analysis comes from a combination of publicly available polling data for the 2024 U.S. Presidential election (*Presidential Election Polls-2024* 2024). The analysis uses the statistical programming language R (R Core Team 2023) and several libraries, including `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `knitr` (Xie 2024), `dplyr` (Wickham et al. 2023), `arrow` (Richardson et al. 2024), `purrr` (Wickham and Henry 2023), and `here` (Müller 2020), for data manipulation. `ggplot2` (Wickham 2016) and `kableExtra` (Zhu 2024) for visualization. The dataset covers various polls conducted across multiple states, capturing the support for each major candidate—Donald Trump and Kamala Harris—along with detailed attributes of the polls.

2.2 Measurement

The primary measurement process reflects the transformation of raw polling data into structured entries to actionable analysis for election forecasting. We first select and rename key variables from raw data to focus on relevant information, such as pollster, pollscore, state, polling dates, sample size, and candidate support percentage. These variables represent essential poll characteristics that capture public sentiment in response to real-world campaign dynamics and demographic differences across states.

To address instances where the original data referenced Joe Biden instead of Kamala Harris, we assumed the entries for Biden will goes to Harris in consequences of Democratic Party. This approach is supported by "Harris's truncated presidential campaign relies heavily upon President Biden's policy framework and she eaddress the campaign's weaknesses and win over relevant voter demographics (Sigmon 2024). Thus, substituting Biden's polling data with Harris's ensures continuity in demographic appeal without altering established polling trends. Then we filter to include only polls relevant to the two candidates Trump and Harris, and features like PollRecency were constructed to weigh recent polls higher. The data was cleaned by removing irrelevant or incomplete observations, and new variables were constructed where necessary.

3

2.3 Outcome Variables

The main outcome variable of interest is the Percentage of support each candidate has in a given poll, which represents the proportion of respondents who favor one candidate over the other. This variable is needed to estimate both the popular vote and Electoral College outcomes. We see a bimodal distribution in Figure 1. This indicates that candidates are either recieveing very high percentages of support or very low percentages of support. When we pull out Harris and Trump and view their distributions versus all candidates in Figure 2, we see that these two candidates see statistically high percentages of support, indicating they are leading the polls.
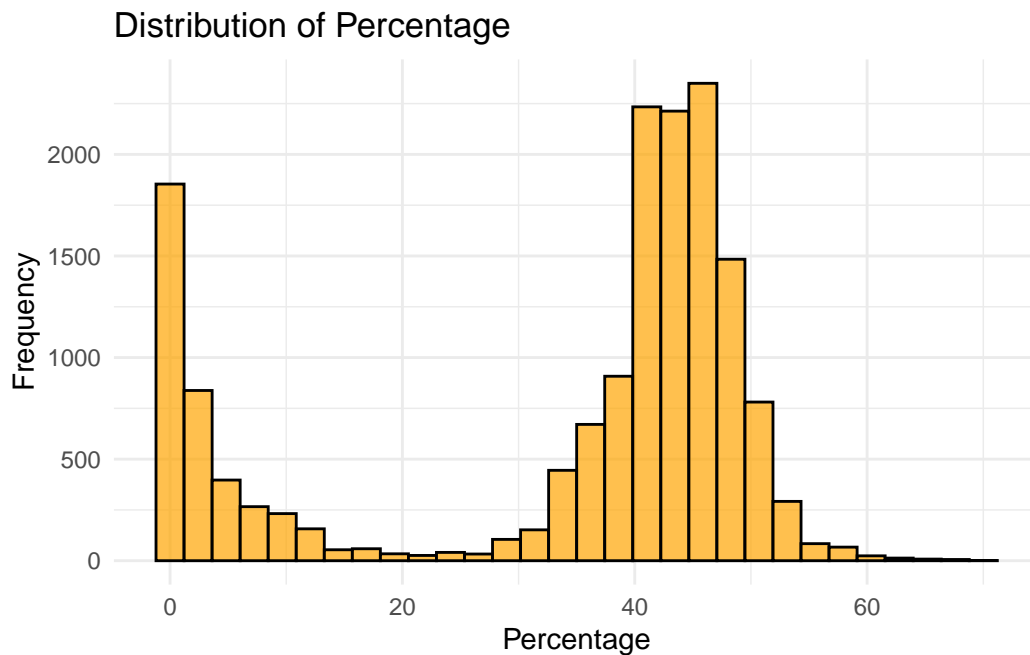

Figure 1: The distribution of what percentage of support a candidate recieves in a poll

2.4 Predictor Variables

In this analysis, several key predictors were identified to evaluate polling data effectively. The first variable, **PollScore**, is a numeric representation of the pollster's credibility, which stands for "Predictive Optimization of Latent Skill Level in Surveys, Considering Overall Record, Empirically." This score helps assess the error and bias associated with a specific pollster, where lower values (including negative numbers) indicate better credibility (FiveThirtyEight 2024). Figure 3 shows whether the majority of pollsters tend to have high or low credibility scores. We see a higher concentration of negative scores, which suggests a trend toward more reputable pollsters in the dataset.
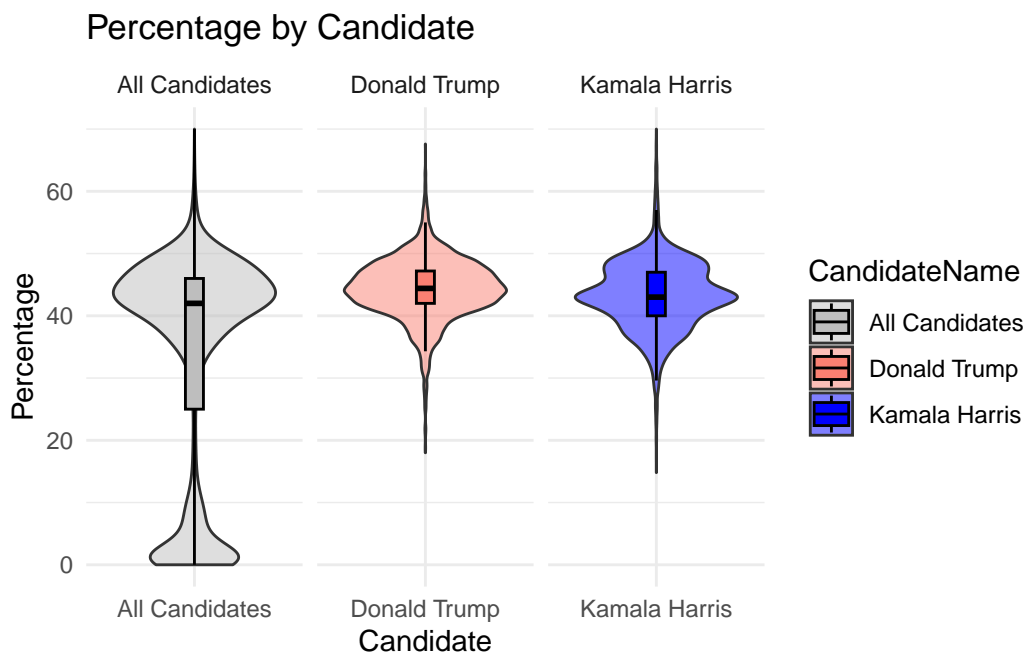
Percentage by Candidate

Figure 2: violin plot displaying the percentage of support Harris and Trump receive compared to all candidates
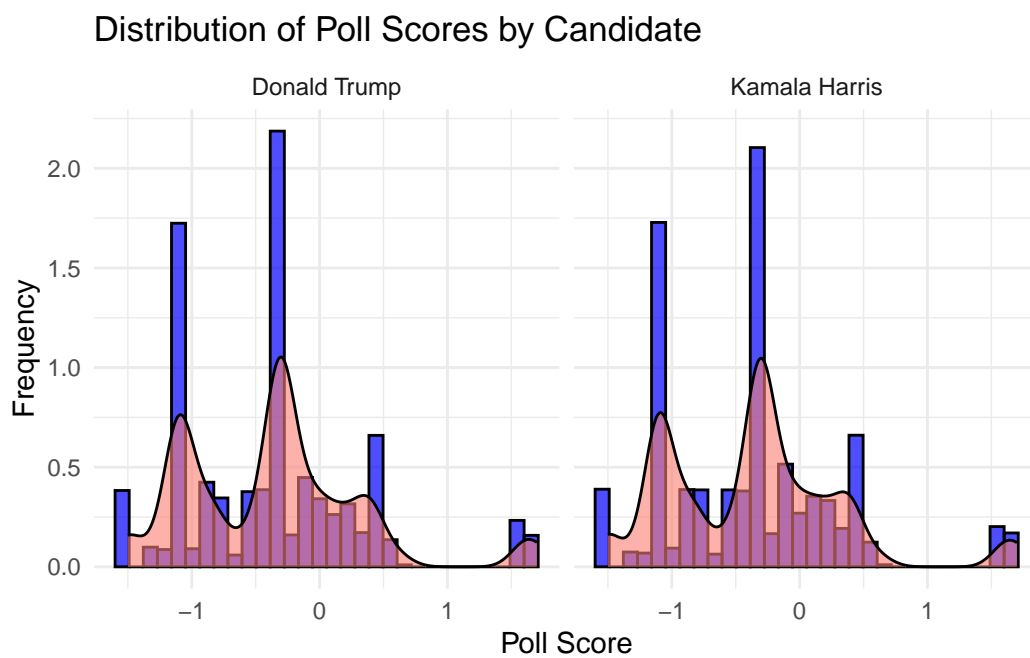
Distribution of Poll Scores by Candidate

Figure 3: Distribution of Poll Scores by Candidate with denisty overlayed

Next, **SampleSize** refers to the number of respondents involved in each poll. A larger sample size generally enhances the reliability of poll results, as it is more likely to represent the broader population accurately. In Figure 4 we see that the vast majority of polls in the dataset have a sample size of less than 2500. As discussed in Section 6, this number maximizes efficiency and accuracy while minimizing the margin of error. The **State** variable denotes the U.S. state where the poll was conducted, in order to give geographical context. All 50 states are represented in this analysis, and Nebraska and Maine are broken up into electoral districts. Washington DC is not included in this dataset. Figure 5 shows the number of polls conducted in each state. Battleground states notably have a higher number of polls conducted.
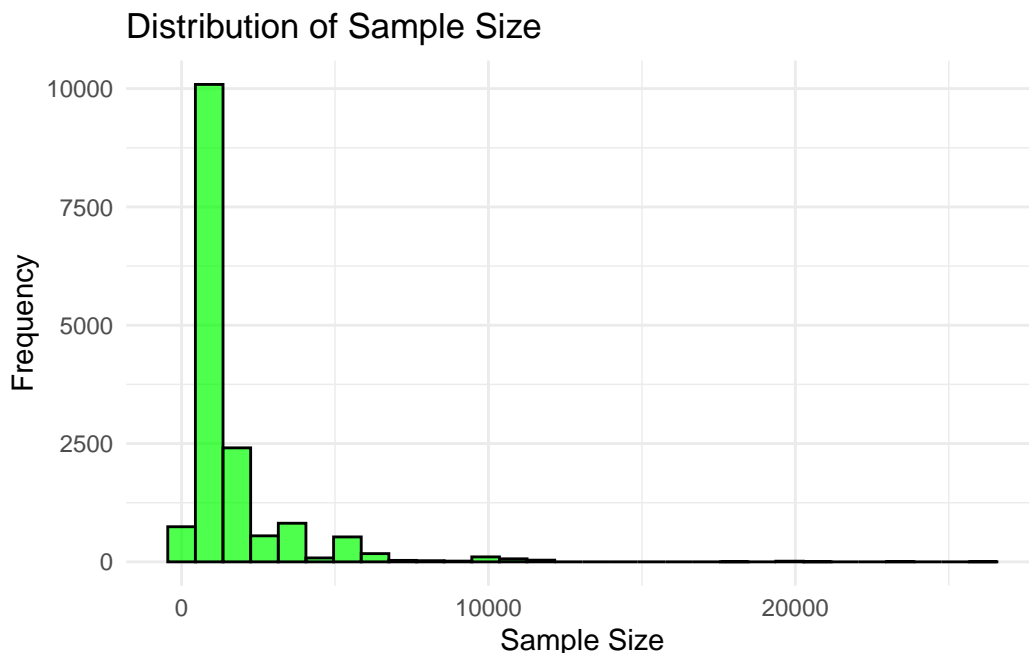
Distribution of Sample Size

Figure 4: The distribution of sample sizes used in each poll

PollRecency is a variable we created to measure the number of days between the start date of the poll and the present. This is important as more recent polls tend to reflect the current political climate and voter sentiment more accurately than polls that were conducted a year ago, or more. In Figure 6 we see that the majority of polls were conducted more recently than not, however polls date back until the last presidential election in 2020. Lastly, **NumericGrade** is a numerical assessment of the pollster's reliability. Our dataset displays polls that tend to have a higher numeric grade as depicted in Figure 7.

While **CandidateName** is not a predictor variable within the model, it serves as a filter for the predictions being analyzed. This allows us to compare between the two primary candidates—Donald Trump and Kamala Harris.

Figure 8 displays relationships among the key variables in the dataset. The strong negative
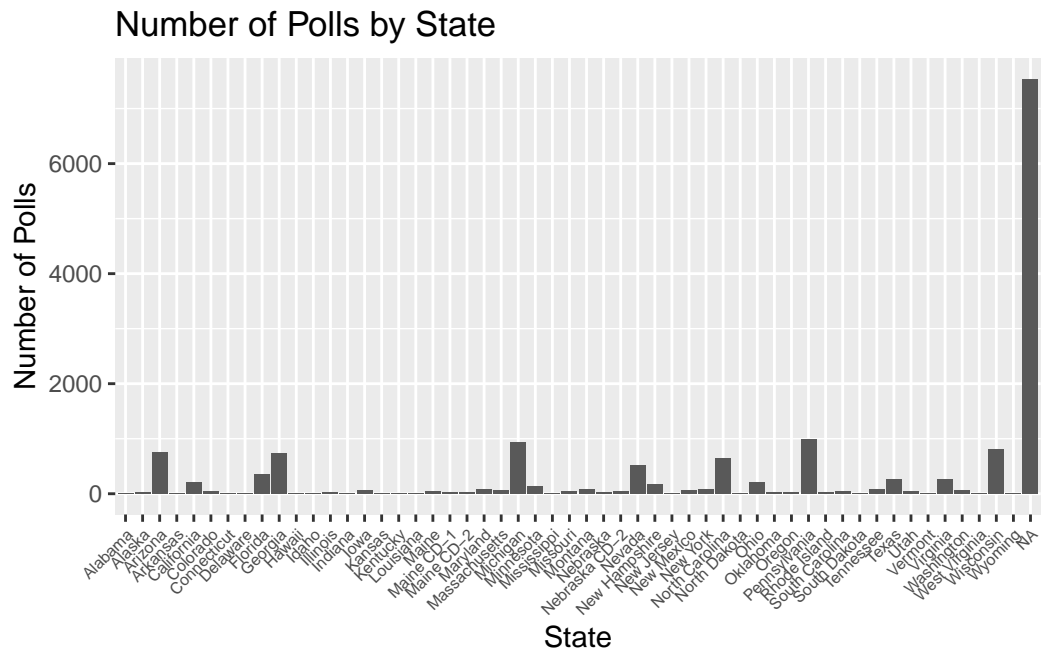
Number of Polls by State

Figure 5: The number of polls conducted in each state.
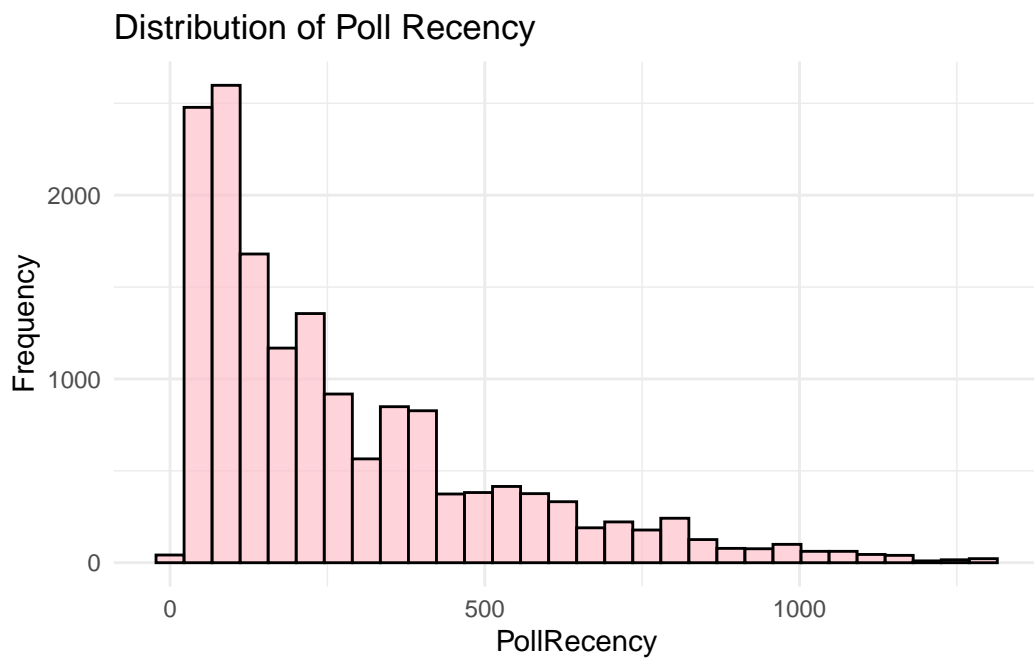
Distribution of Poll Recency

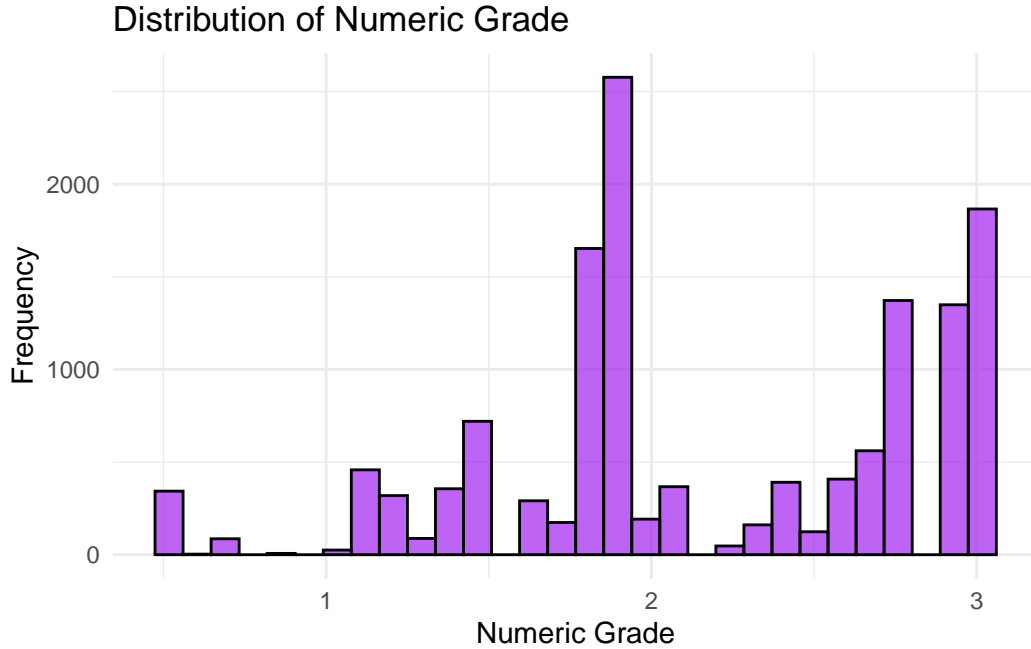Figure 6: The distribution of how recent a poll was conducted

Figure 7: The distribution of numeric grades assigned to each poll

correlation between PollScore and NumericGrade (-0.89) indicates that as the PollScore of the pollster decreases (lower PollScore values are better), the reliability of the pollster increases (higher NumericGrade values are preferable). This shows that more credible pollsters are likely to be more reliable, which makes sense given the context.

PollRecency exhibits a weak positive correlation with SampleSize (0.07) and Percentage (0.23), suggesting that more recent polls are slightly associated with a larger number of respondents and may result in higher outcome percentages (which may be due to a reduction in the number of candidates as the election grows closer). However, the correlation with PollScore is very weak (-0.05), indicating that the recency of polls does not significantly impact their credibility or reliability.

Percentage, as the outcome variable, shows a weak positive correlation with SampleSize (0.16), suggesting that larger sample sizes may be associated with higher percentages, although the effect is modest. There is a slight negative correlation with PollScore (-0.02), reinforcing the notion that higher percentages do not consistently relate to the reliability of the pollsters.

3 Model

To predict the outcome of the 2024 U.S. Presidential election, we developed two linear regression models: one for Kamala Harris and one for Donald Trump. These models aim to
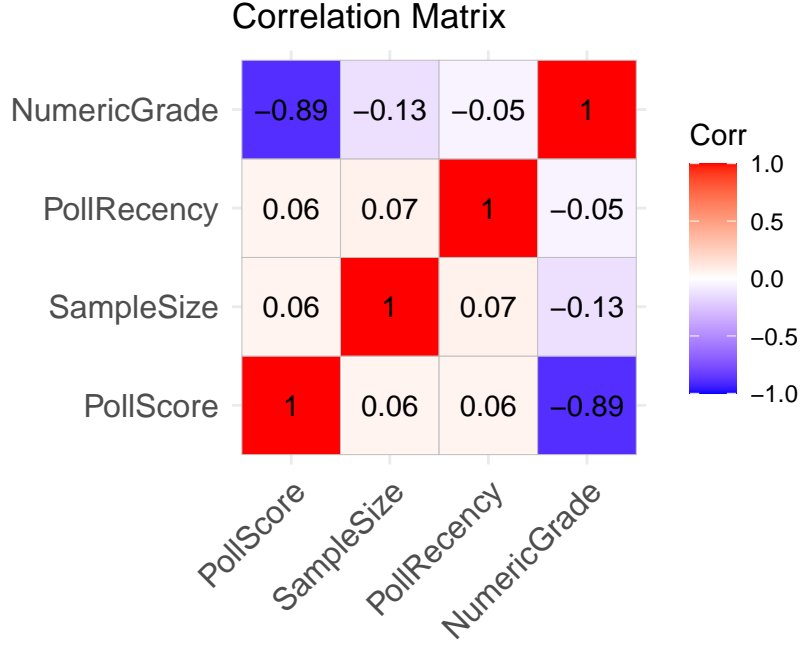
Figure 8: correlation between PollScore, the sample size of the poll, and the percentage of support for a candidate

estimate the percentage of support for each candidate based on the aforementioned predictor variables.

3.1 Model Set-Up

This paper utilizes linear regression models:

$$Percentage_H = \beta_0 + \beta_1 Pollscore_i + \beta_2 SampleSize + \beta_3 NumericGrade_i + \beta_4 State_i + \beta_5 PollRecency + \epsilon_i$$

$$Percentage_T = \beta_0 + \beta_1 Pollscore_i + \beta_2 SampleSize + \beta_3 NumericGrade_i + \beta_4 State_i + \beta_5 PollRecency + \epsilon_i$$

In this model, $Percentage$ represents the predicted support percentage for a candidate in a given poll. $Percentage_H$ and $Percentage_T$ represent the predicted support for Harris and Trump respectively. Each predictor variable is carefully selected to capture factors influencing voter support, with coefficients β_i representing the strength of each factor's impact. The error term ϵ_i accounts for random variability not explained by the model.

For a more granular analysis, separate linear regression models were constructed for each of the two primary candidates, Harris and Trump, to predict their state-by-state poll percentages. These models allow for a detailed view of each candidate's performance across states. Based on these percentages, we forecast the likely winner of the popular vote as well as the state's electoral votes. These models filter for *CandidateName* since the model is candidate-specific.

As defined in Section 2, each predictor variable has been chosen to reflect characteristics that influence election polling.

3.2 Model Justification

While other complex models could be appropriate in this instance, these models were chosen for its balance between simplicity and predictive power. Linear regression allows us to capture relationships between polling quality, sample size, and recency, while still providing interpretable coefficients. Given the nature of polling data, this approach is both appropriate and commonly used in election forecasting. A linear model allows us to directly interpret the influence of each predictor on voter support.

3.3 Assumptions and Limitations

This model operates under several assumptions, including that relationships between predictors and outcomes are linear, errors are normally distributed, and predictors are not excessively collinear. Potential limitations include the fact that linear regression may oversimplify relationships in polling data, especially if interactions between predictors exist. Furthermore, as the model relies on historical polling data, it assumes that past trends are indicative of future behavior, which may not hold true if significant political events alter voter preferences unexpectedly, as discussed in Section 5.

3.4 Model Validation

The model was implemented using R, with out-of-sample testing and Root Mean Square Error (RMSE) calculations applied to assess predictive accuracy as predicted in Table 1. By splitting the data into training and test sets, we evaluated how well the model performs on unseen data. This approach helps ensure that the model generalizes effectively to new data and avoids overfitting.

Table 1: RMSE Results for Candidate Models

Model	Root Mean Squared Error (RMSE)
Donald Trump	3.47

Kamala Harris	4.79

4 Results

The national averages for both candidates, as shown in Table 2, show that Donald Trump is predicted to receive an average of 45.26% of the popular vote, which, when normalized, represents 50.96% of the total forecasted support, while Kamala Harris is predicted to receive 43.56% with a normalized value of 49.04%. In terms of electoral votes, Table 3 shows that Donald Trump is projected to secure 302 votes, while Kamala Harris is expected to gain 233 votes.

Table 2: Predicted National Popular Vote for Donald Trump and Kamala Harris

Trump Average Predicted %	Trump Normalized %	Harris Average Predicted %	Harris Normalized %
45.26	50.96	43.56	49.04

Table 3: Predicted Number of Electoral Votes for Donald Trump and Kamala Harris

Trumps Total Electoral Votes	Harris Total Electoral Votes
302	233

The predicted outcomes for states where the margin is 4% or less are summarized in Table 4. These states are known as our Battleground states (aka swing states). Donald Trump leads in several key battleground states, including Pennsylvania (45.68% for Trump vs. 44.83% for Harris), Georgia (46.94% vs. 43.49%), and Michigan (44.86% vs. 43.87%), while Kamala Harris secures leads in states such as Wisconsin (45.50% for Harris vs. 45.33% for Trump) and Minnesota (44.73% vs. 41.64%). Figure 9 is a visual representation of the predicted outcomes for all states and how it would appear as an electoral map.

Table 4: Predicted Percentage of Electoral Votes Per State (Margin 4%)

State	Trump Predicted %	Harris Predicted %	Winner
Arizona	46.29	42.80	Donald Trump
Georgia	46.94	43.49	Donald Trump
Maine CD-2	46.10	43.18	Donald Trump
Michigan	44.86	43.87	Donald Trump

Minnesota	41.64	44.73	Kamala Harris
Nevada	45.80	43.17	Donald Trump
North Carolina	46.58	43.55	Donald Trump
Pennsylvania	45.68	44.83	Donald Trump
Virginia	41.52	44.67	Kamala Harris
Wisconsin	45.33	45.50	Kamala Harris

2024 US Presidential Election Predictions by State
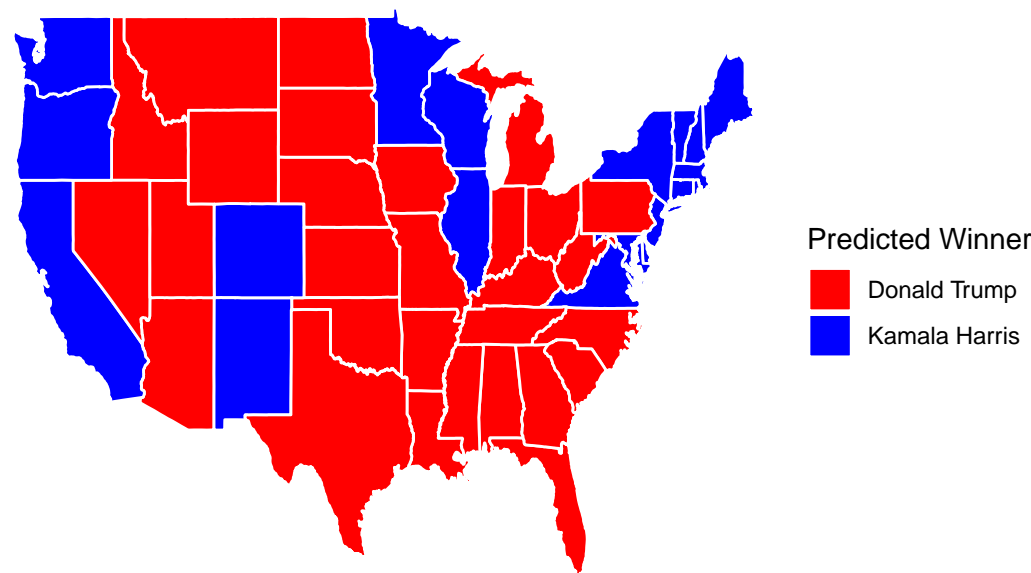
Figure 9: Mapped Model-predicted outcomes for 2024 elections by State

The polling data includes an average poll score of -0.379, with an average sample size of approximately 1,605 individuals per poll. The average percentage of support across all states for the candidates is 33.68%. This dataset contains a total of 15,829 individual polls, allowing for the model to be trained off a significant number of datapoints. The summary statistics are displayed in Table 5.

Table 5: Summary Statistics of Model Results

Average Poll Score	Average Sample Size	Average Percentage	Total Polls
-0.379	1605.742	33.677	15829

5 Discussion

5.1 What did we do?

This paper analyzes and predicts the 2024 U.S. presidential election outcomes through regression models based on polling data for two main candidates, Donald Trump and Kamala Harris. Using a state-by-state approach, it evaluates each candidate's predicted percentage of support and calculates the anticipated electoral votes to provide an overall electoral projection. The model incorporates the predictor factors: PollScore, NumericGrade, SampleSize, State, and PollRecency.

5.2 What did we learn?

This paper shows how regional polarization in the U.S. and strong state-based support can determine election outcomes, especially in battleground states. From this analysis, we observe that regions such as the Midwest and the South tend to favor conservative candidates, while the Northeast and West Coast are typically more supportive of liberal candidates. Trump's predicted success in key battleground states, such as Pennsylvania and Michigan, leads to a victory in the Electoral College. Meanwhile, the difference in the predicted percentages for the popular vote is less than 2%. Although the popular vote is predicted to be nearly balanced, with a margin of less than 2% separating the candidates, the narrow margins in key battleground states demonstrate why every vote is so important. Small shifts in turnout or voter sentiment can truly change the overall outcome of the election, and thus the future of the United States, dramatically. A prime example of this took place in Georgia.

In 2020, Georgia turned blue in a presidential election for the first time since 1992 ("Georgia" 2024). A major factor in this shift was the influx of young, diverse individuals in the Atlanta suburban area. In Gwinnett County, the non-Hispanic white population decreased from 67% in 2000 to around 35% by 2020 (Bacon Jr. 2020). Mitt Romney, a Republican, had won Gwinnett County by 9 points in 2012. In 2020, it shifted to an 18-point win for Joe Biden. Additionally, there was a major campaign by Stacey Abrams to mobilize previously disengaged voters, particularly people of color (Bacon Jr. 2020). All of this led to Georgia becoming a swing state after 28 years of voting red. Biden received only 0.2% more of the vote than Trump in Georgia.

5.3 What are some weaknesses?

One limitation is its reliance on polling data, which can be inherently biased due to sampling issues, response rates, and potential biases in who chooses to respond to polls. Moreover, as the election date approaches, polling data may fail to capture recent shifts in voter sentiment

resulting from unforeseen events or major campaign developments. For example, last-minute political controversies or debates can alter public opinion quickly.

For instance, a recent incident involving derogatory comments about Puerto Rico made at a Trump rally has quickly influenced the views of Puerto Rican voters in Pennsylvania. Puerto Rican voters, who make up nearly half a million people in the state, have voiced anger and disapproval, with some even reversing their voting plans (Hill, McCarthy, and Otterbein 2024). WhatsApp groups, morning radio shows, and community discussions have amplified these sentiments, illustrating how quickly backlash can spread through influential networks, particularly in swing states (Hill, McCarthy, and Otterbein 2024). Our model predicts Trump will win Pennsylvania by a margin of 0.8%, which is less than the 3.7% of Pennsylvania's Population that Puerto Ricans make up (United States Census Bureau 2020). Situations like these could easily cause Pennsylvania's 19 electoral votes to swing the other way.

Another limitation is the assumption that past trends will remain consistent across elections, which does not account for the ongoing demographic and ideological shifts within states. For instance, the influx of highly educated individuals into certain states can shift voting patterns over time, as research indicates that Americans with a college degree or higher are more likely to vote for liberal candidates (Pew Research Center 2016). Should a state's population of college-educated voters grow significantly, this change could lead to unexpected voting outcomes that our model might not fully capture.

Additionally, the model excludes third-party candidates, who, while unlikely to win, often garner a small but meaningful percentage of votes. Independent candidates historically attract around 2-5% of the popular vote, diverting votes from major party candidates, which can be especially impactful in closely contested states. This factor may lead to underestimations in one candidate's projected vote share if independents disproportionately draw support from one side.

These limitations point to a larger issue: that forecasting an election relies on data subject to change and that unexpected events and demographic trends can disrupt otherwise reliable patterns.

5.4 What is left to learn? How should we proceed in the future?

Future research could better the model by incorporating live data updates and adjusting for factors such as voter turnout projections and demographic shifts at the state level. For example, as states experience growth in college-educated residents or younger populations, these demographic factors may gradually influence overall voting patterns. A model that can adapt ro reflect these trends could improve predictive accuracy over time. Additionally, including third-party candidates in the model would provide a more well-rounded view of the vote breakdown, as these candidates can meaningfully impact the final distribution of votes between the two primary candidates. As previously mentioned, third-party candidates can capture a small but notable portion of votes, which can tilt tight races in battleground states.

Future models could potentially account for this by assessing third-party candidates' historical performance to estimate the effect on major candidates' vote shares.

6 Appendix

6.1 Data Cleaning Notes

We began by importing the raw dataset using the read_csv function from the tidyverse package. To focus our analysis on more relevant variables, we selected specific columns, such as poll identifiers and polling metrics, omitting any unnecessary columns.

We then renaming columns for clarity. For example, we changed pct to Percentage, making it easier for anyone working with the data to understand what each variable represents.

We converted date data into date format since they were represented as strings in the raw dataset. This was done by using the mdy() function, after which we removed any rows with missing values for StartDate, EndDate, or ElectionDate to avoid issues later in our analysis.

We also modified the CandidateName variable to attribute polls for Joe Biden to Kamala Harris, as Joe Biden dropped out of the race and Harris replaced him as the democratic nominee.

We also created a new variable called PollRecency, which indicates how many days have passed since each poll started. This was done because as we get closer to election date, polls become more meaningful as they represent the most recent viewpoints of voters.

We also wrapped the cleaning process in a tryCatch block in order to mitigate any errors that arose throughout the cleaning process.

After completing the cleaning, we saved the final dataset in both Parquet and CSV formats for later analysis.

6.2 Pollster Methodology

The New York Times/Siena College Poll was ranked the most accurate political pollster in the US following the 2022 midterm elections (FiveThirtyEight 2024). They received the max 538 rating of 3 stars, one of only four pollsters to do so. They also had a pollscore of -1.5, meaning they had the least amount of error and bias attributed to them of the 282 pollsters ranked by 538 (FiveThirtyEight 2024).

The Times/Siena Poll targets registered voters in the United States, emphasizing those in battleground states during presidential elections. Since the US decides it's President based on the Electoral College rather than the national popular vote, this is an important focus in order to correctly determine election outcomes. The sampling frame is derived from a voter file which containing lists of 200 million registered voters as well as their demographic information and prior voting behavior data (Times 2024).

Each poll typically surveys around 1,000 respondents, which is deemed sufficient for achieving a margin of sampling error of approximately 3-4 percentage points. The Times found that

tripling the sample size would only reduce the margin of error by a percent or two. This sample size allows for a balance between representation and efficiency (Times 2024).

Respondents are recruited through live phone interviews conducted by trained interviewers from call centers in various states. The polling effort includes both landline and cellphone calls, with over 90% of respondents being reached via cellphones (Times 2024).

Times/Siena uses a stratified random sampling technique, where the sample is divided into strata based on demographics such as age, race, gender, and party affiliation. This helps to ensure proportional representation from each group. Stratification helps in achieving a sample that mirrors the voting population, minimizing biases related to demographic variables (i.e. uneducated voters are less likely to respond to polls). However, it faces limitations as stratification increases the complexity of the sampling process, and if strata are not adequately represented in the sample, the results may still skew. Times/Siena combats this by applying weighting adjustments post-survey to correct for any demographic discrepancies (Times 2024).

Given the declining response rates typical of phone surveys, the Times/Siena Poll faces challenges in reaching respondents. Approximately 2% of those contacted ultimately participate. This pollster uses follow-up calls to make multiple attempts to reach individuals who initially do not respond (Times 2024). In 2022, they conducted an experiment to find out if responses were skewed by non-response bias (Cohn 2022). This was done via a mail survey that offered a reward of up to $25 for a response. This saw a 30% response rate, which is significantly larger than the 2% response rate Times/Siena gets over phone. Although there were differences between the types of respondents you get between the two methodologies, there was not a meaningful difference in who these respondants claim they would vote for (Cohn 2022).

The questionnaires used by Times/Siena are designed to capture a broad range of opinions while remaining concise. Interviews are kept under 15 minutes to reduce survey fatigue and maximize response rates (Times 2024). This pollster prioritizes making sure every viewpoint is accurately represented in order to make the questions fair and ensure that respondents are not being swayed in a particular direction. This means the questions must remain objective and clear. However, a survey with limited response options and questions may mean a respondent's view is not captured in its full breadth. It also means that more granular opinions may be overlooked in favor of broader trends.

The New York Times/Siena College Poll is rated the most accurate pollster in the US for a reason - it has a well-defined methodology with a fleshed out population frame, sampling approach, and contingencies for non-response handling.

6.3 Idealized Methodology

The idealized methodology for forecasting the US presidential election targets eligible registered voters from all 50 states and Washington D.C., ensuring representation from each state in

order to create an Electoral College forecast, with particular emphasis on swing (battleground) states.

The sampling method will utilize stratified random sampling, which stratifies the sample by state. Voters will be further stratified based on factors such as age, gender, race/ethnicity, income, and political affiliation. In swing states, we will ensure oversampling to capture more granular data and reduce our margin of error.

To maintain a national margin of error of $\pm 4\%$, based on the experiment conducted by Times/Siena, the total sample size will be 1,000 respondents per state (Times 2024). Swing states will be oversampled, with over 2,000 respondents to achieve a margin of error of $\pm 2\%$. Given an estimated response rate of 2%, we aim to recruit approximately 50,000 potential respondents per state, with a target of 100,000 in battleground states.

To recruit respondents, we will utilize online panel providers like YouGov, Ipsos, and Dynata to obtain geographically and demographically targeted participants. Additionally, we will conduct state-specific phone surveys, employing interactive voice response (IVR) systems and live calls to reach older and less tech-savvy voters, particularly in states with a higher proportion of rural populations. To incentivize participation, all respondents will be entered into state-level sweepstakes, with one winner selected per state.

Data collection will be conducted on platforms such as Google Forms. Telephone surveys will follow the layout and workflow of the Google Form survey. The survey is designed to take approximately five to ten minutes to complete in order to minimize drop-off rates. It will include national-level questions, covering voting intent, candidate preference, key issues, and voter perceptions. After collection, the data will be weighted by state and demographic factors to ensure accuracy, utilizing the latest Census data and state voting records.

For data validation, we will implement several measures, including completion time checks to identify speeding responses, IP tracking to prevent duplicate entries, and reCAPTCHA to guard against bots. Respondents will provide their state of residence and ZIP code for additional verification. Furthermore, the data will be weighted according to state demographics and previous election turnout data, utilizing information from Census and state voting records. Special attention will be paid to swing states to ensure no key voter segments are underrepresented.

In terms of budget allocation, we will designate funds as follows:

1. Survey Development: $7,000

   - The $7,000 covers the costs associated with designing the survey, including the creation of questions, structuring the survey for clarity and flow, and ensuring that it meets methodological standards. It may also include hiring experts in survey design or consulting with statisticians to validate the survey framework and ensure questions are unbiased. Additionally, while Google Forms is free, this budget may account for platform costs if needed.

2. Accurate Sampling: $10,000

   - The sampling budget is allocated for the costs associated with selecting a representative sample of voters. This includes expenses related to stratified random sampling, where the sample is segmented by various demographics and state representation. In order to ensure each demographic share is accurately represented, we may need to purchase access to demographic data or employ statisticians to ensure that the sampling aligns with the population of each state.

3. Recruiting Respondents: $45,000

   - The majority of the budget is dedicated to recruiting respondents, which is necessary for achieving a large and demographically representative sample. This cost will cover payments to online panel providers which charge fees for access to their respondent pools. Additionally, this budget will cover costs related to conducting state-specific phone surveys and IVR calls, such as hiring call center staff or IVR technology.

4. Respondent Incentives: $28,000

   - To encourage participation, we plan to run state-level sweepstakes, where respondents have a chance to win prizes for completing the survey. A $500 prize will be awarded to one winner in each state (+ Washington DC). The remaining budget will cover administrative expenses associated with running the sweepstakes.

5. Data Analysis and Reporting: $10,000

   - This portion of the budget is reserved for the costs related to analyzing the collected data and reporting the findings. This may include hiring data analysts or consultants who specialize in statistical analysis and survey data interpretation. The funds will also cover software tools necessary for data analysis, as well as the creation of visualizations and reports to communicate the results effectively.

An example of the survey can be found at this link: https://forms.gle/igzK683kHLYytGvu8

References

Bacon Jr., Perry. 2020. "How Georgia Turned Blue." *FiveThirtyEight*, November. https://fivethirtyeight.com/features/how-georgia-turned-blue/.

Cohn, Nate. 2022. "Are the Polls Still Missing 'Hidden' Republicans? Here's What We're Doing to Find Out." *The New York Times*, November. https://www.nytimes.com/2022/11/08/upshot/poll-experiment-wisconsin-trump.html.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. "538's Pollster Ratings." ABC News; ABC News Internet Ventures. https://www.fivethirtyeight.com/methodology/pollster-ratings/.

"Georgia." 2024. Electoral Ventures LLC. https://www.270towin.com/states/Georgia.

Hill, Meredith Lee, Mia McCarthy, and Holly Otterbein. 2024. "Trump's Puerto Rico Fallout Is 'Spreading Like Wildfire' in Pennsylvania." *E&E News*. https://www.politico.com/news/2024/10/28/trump-rally-puerto-rico-pennsylvania-fallout-00185935.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

Pew Research Center. 2016. "A Wider Ideological Gap Between More and Less Educated Adults." https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/.

Presidential Election Polls-2024. 2024. New York, USA: FiveThirtyEight. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

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://CRAN.R-project.org/package=arrow.

Sigmon, Eric. 2024. "The New Democratic Presidential Ticket: What's Changed and What Remains the Same." Elcano Royal Institute. https://www.realinstitutoelcano.org/en/analyses/the-new-democratic-presidential-ticket-whats-changed-and-what-remains-the-same/.

Times, The New York. 2024. "You Ask, We Answer: How the Times/Siena Poll Is Conducted." *The New York Times*, October. https://www.nytimes.com/article/times-siena-poll-methodology.html.

United States Census Bureau. 2020. "2020 Census Data." https://data.census.gov/table/ACSDT5YSPT2021.B05001?t=402&g=040XX00US42.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. 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/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2023. *Purrr: Functional Programming Tools.* https://CRAN.R-project.org/package=purrr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.