A Logistic Regression Prediction of 2024 Election Odds Gives Harris a 38.3% Winrate*

Andrew Goh Yisu Hou

November 4, 2024

The 2024 Presidential Election, run between the Democratic candidate Vice President Kamala Harris and Republican candidate Former President Donald Trump, is both widely predicted to be a close race (FiveThirtyEight 2024) and considered to be the "most important election of your life" by a majority of likely voters this election season, as estimated by Quinnipiac (Quinnipiac University Poll 2024b). To predict the results of this election, we use a FiveThirtyEight poll-of-polls dataset (FiveThirtyEight 2024), fitting multinomial logistic regression models with the data both at a national level and for each of the 7 swing states. Using these estimates for the probability of Harris winning the popular votes in each of those categories, we find that Vice President Harris has a 38.3% probability of winning the election in the context of the electoral college, a statistically significant result after considering margin of error and various biases.

1 Introduction

With most polls maintaining close (<1) odds at both the national and statewide level for the seven swing states (FiveThirtyEight 2024; The New York Times 2024b), the 2024 Presidential Elections are proving to be an extremely volatile one, with poll predictions regularly flipping and disagreeing with each other, even up until the day before Election Day. The volatility even extends beyond the polling itself – since the January 6 Capitol Attack, prominent figures supporting both parties have painted the 2024 elections specifically as one that underpins the very democratic integrity of the country. Elon Musk, for example, has openly expressed concerns of a Harris win leading to a one-party country (Musk 2024), and Zack Beauchamp from Vox has described a Trump re-election as an "extinction-level threat to democracy" (Beauchamp 2024). In short, the stakes are high, at least as perceived in the eyes of voters (Quinnipiac University Poll 2024b).

*Code and data are available at: https://github.com/YisuHou1/US_Election_Statistics.

In order to arrive at such a conclusion, we used poll-by-polls data from FiveThirtyEight (FiveThirtyEight 2024) to fit multinomial logistic regression models for both the nation as a whole and each of the seven swing states. Taking into consideration the mispredictions by pollsters for both the 2016 and 2020 elections, where Trump won the presidential race against Clinton in a 304-227 landslide despite predictions universally predicting a landslide win in the opposite direction (Courtney Kennedy 2022), we regress on the characteristics of the poll/pollster such as transparency rating and sponsors and correct the results for systematic polling biases based on past election data, an approach predicted to yield better results this year (Cohn 2024).

Our estimand, or what we seek to predict, is the probability that Kamala Harris wins the election on the 6th of November.

After obtaining the probabilities that Harris wins each swing state via randomised sampling using our fitted models, we then analysed these probabilities in the context of the electoral college, i.e. the probability that Harris can obtain 44 votes from the swing states, which would all but guarantee her step into office (The New York Times 2024b). We find that Harris has a 38.3% chance of winning the general election, suggesting a likely Trump advantage after accounting for the margin of error in all related processes and observations.

The body of the paper contains the data, model, results, and discussion sections. In order, Section 2 contains an overview of the data used, the variables involved and the tools used in our analysis and interpretation of the data, Section 3 describes the model in detail and other related miscellaneous musings, Section 4 contains a run-down and explanation of the results and findings from the model, and Section 5 contains a discussion of the implications and limitations of the result, among others. The appendices contain a sample of an idealised methodology and survey, as well as a case study of the sample and methodology of Quinnipiac's Pennsylvania poll (Quinnipiac University Poll 2024b, 2024a).

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) for the graphing, analysis and presentation of the project as a whole. Caret (Kuhn and Max 2008), glmnet (Friedman, Tibshirani, and Hastie 2010; Simon et al. 2011; Tay, Narasimhan, and Hastie 2023), nnet (Venables and Ripley 2002), lubridate (Grolemund and Wickham 2011) and pROC (Robin et al. 2011) were used directly in the analysis of the data. Tidyverse (Wickham et al. 2019), stringr (Wickham 2023), dplyr (Wickham et al. 2023), and styler (Müller and Walthert 2023) were used in the presentation and/or styling of the data, graphs, and paper as a whole.

In order to build a model that predicts the results of the 2024 election, we use survey/poll data on likely voters to inform our insights regarding likely election results, an approach

widely accepted in political contexts (AAPOR 2022). We use a poll-of-polls dataset from 538 (FiveThirtyEight 2024), consisting of aggregated polling data from FiveThirtyEight's 2024 presidential election polling database from various national and state-level polls from pollsters all around the country, recording information such as the date when the poll concluded, the entity conducting the poll, and the estimated support rate of the candidate as given by the poll. The data spans national and state-level polls, with particular focus on key battleground states that could determine the Electoral College outcome. This dataset was chosen over alternatives like RealClearPolitics because of 538's comprehensive methodological adjustments and transparent quality standards for included polls.

After cleaning the data to split results by candidate, removing polls with N/A values and ratings lower than 2.5, and further selecting polls polling likely voters as opposed to the whole population, we are left with 746 data points with the variables pollster, has_sponsor, numeric_grade, pollscore, transparency_score, sample_size, end_date, state, candidate_name, pct (support rate in percentage), and cycle (election cycle, i.e. 2020/2024).

2.2 Measurement

The "from-voter-to-data" process for the 538 dataset involves three processes: measurement, where voter responses are collected and adjusted to best provide insight into a true support percentage population parameter, data collection, where data from polls is obtained and cleaned by 538 to only include polls adhering to their set of standards, and aggregation, where additional scores/variables are given to each poll to reflect their accuracy, transparency and bias levels with regards to polling scores (Morris 2024b).

The measurement process involves several sequential steps, beginning with initial data collection where pollsters conduct surveys using various modes such as phone, online, or mixed-mode approaches to gather voter preferences. Sample processing follows, involving likely voter screens to identify probable voters, demographic weighting to match population benchmarks, and various adjustments to ensure representativeness (Quinnipiac University Poll 2024b; The New York Times 2024a).

The data collection process involves obtaining data from various election polls from around the country, given that they meet specific methodological criteria. Each included poll must provide clear documentation of pollster identity, survey dates, and sampled population, maintain a minimum sample size of 100 respondents, and demonstrate transparent methodology including polling mode, sample source, and weighting procedures. Scientific sampling methods attempting to achieve representative samples are mandatory, as is the disclosure of poll sponsorship and funding sources. The dataset explicitly excludes non-scientific polls lacking representative sampling, MRP-smoothed data, recontact surveys, DIY polls from nonprofessional sources, polls with leading questions or hypothetical matchups, and subsamples from multi-state polls without geographic verification (G. Elliot Morris 2023).

The aggregation phase, finally, involves rating polls by pollster polling quality (ranging from 0.5 to 3.0 stars), with polls with higher historical accuracy, lower consistent partisan biases, and higher transparency of methodology boasting higher ratings (Morris 2024a). Transparency is also a rating and is done through a 10-question yes/no checklist on how much of the methodology is available to the public - note how this is completely separate from the quality of the methodology itself. Each question is graded on a 0/0.5/1 scale, for a combined transparency score that ranges from 0 to 10. Most polls in our dataset miss the mark on one or two criteron, netting a final transparency score of 9 - this is partially due to the data cleaning process removing many relatively worsely conducted polls.

Several important measurement limitations to FiveThirtyEight's methodology must be acknowledged. These include temporal gaps between polls creating discontinuous measurement of voter sentiment, response rates and participation biases potentially skewing samples, and varying geographic coverage with swing states being overrepresented. Additionally, third-party candidate treatment varies across polls, and is difficult to standardise (with respect to other data points) consistently.

2.3 Outcome variable

The outcome variable is the level of support for the two leading US presidential candidates, Vice President Kamala Harris and Former President Donald Trump, recorded as a percentage of likely voters. For the national-level analysis, the support percentage represents the percentage of polled individuals who responded in favor of a specific candidate to a question that asked about their preferred presidential candidate in a national-level poll. For the regional datasets, the support percentage represents the percentage of likely voters who preferred a specific candidate in a poll that targets a specific state or congressional district of the United States. Figure 1 shows the distribution of support percentages in the aggregated polling data from 2020 and 2024 presidential elections.

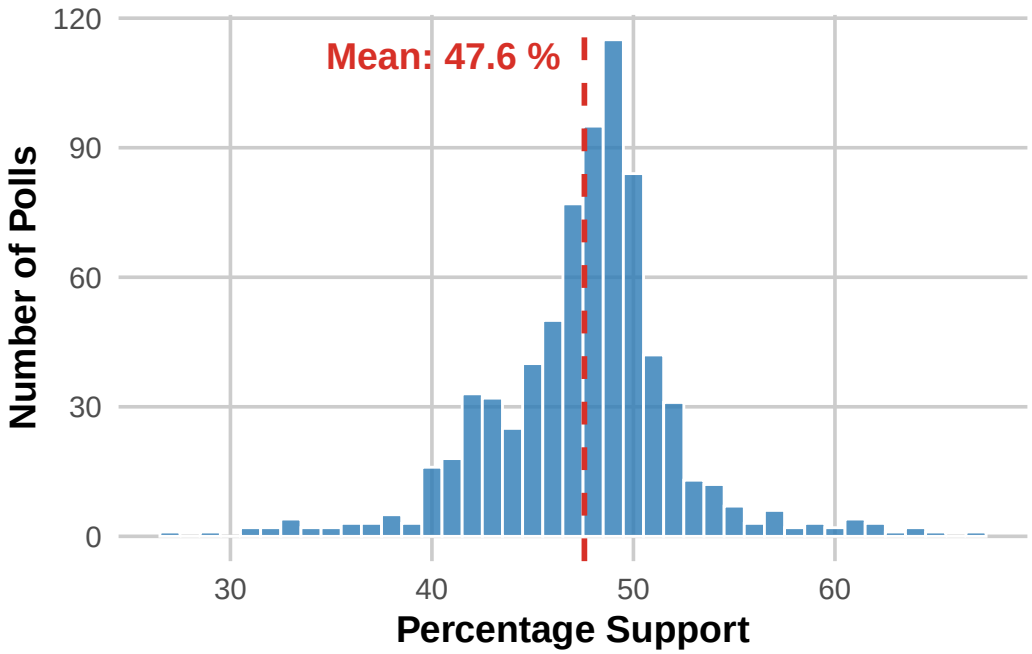

Figure 1: Percentage Support in US Presidential Polls

The aggregated polling data includes support percentages of different presidential candidates in polls with different FiveThirtyEight ratings (G. Elliot Morris 2023) across two elections.

As displayed by Figure 1, the distribution of support percentages is close to a normal distribution, with most data points between 45 and 52 percent.

2.4 Predictor variables

The predictor variables that we ended up using in our final logistic regression model were the poll end dates, transparency score, and sample sizes of the polls, as well as whether or not a poll was sponsored. The poll end dates were included to capture the temporal changes in candidates' support rate. Transparency score was selected as an identification of the polls' trustworthiness, which must be controlled for in the analysis. Other variables representing the poll quality in the raw dataset, such as the poll score and the FiveThirtyEight rating, were excluded from the models for two reasons. First, the documentation by FiveThirtyEight suggests that the numeric rating, poll score, and transparency score are highly related (G. Elliot Morris 2023). Second, the data cleaning process filtered data points with high FiveThirtyEight ratings and poll scores. The distributions of selected variables are as follows:

2.4.1 End Date

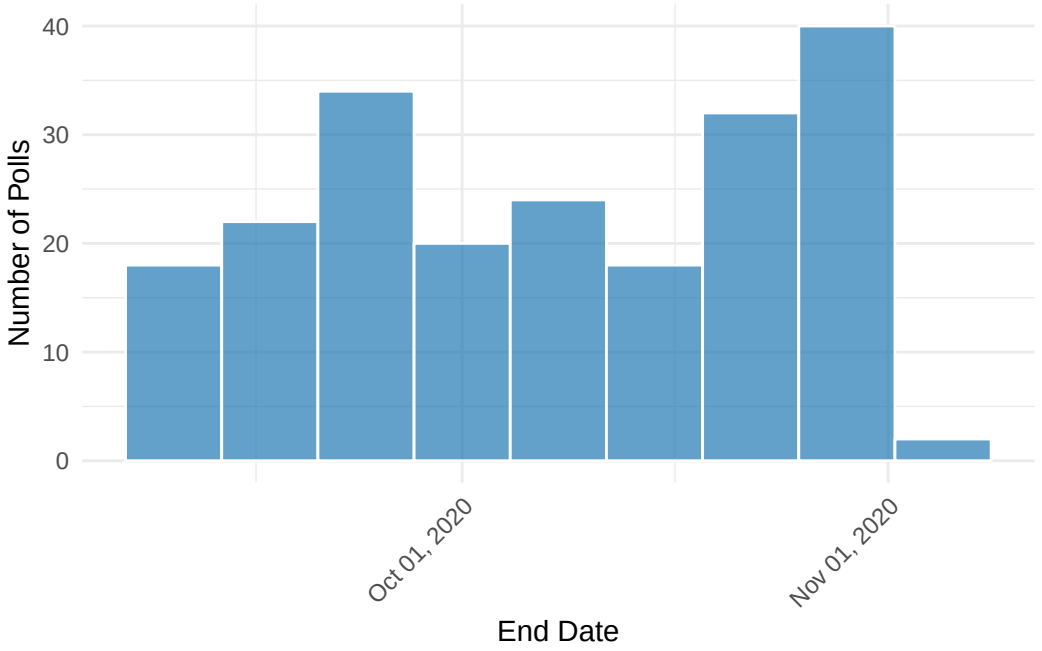

Figure 2: Distribution of Poll End Date, 2020 Cycle

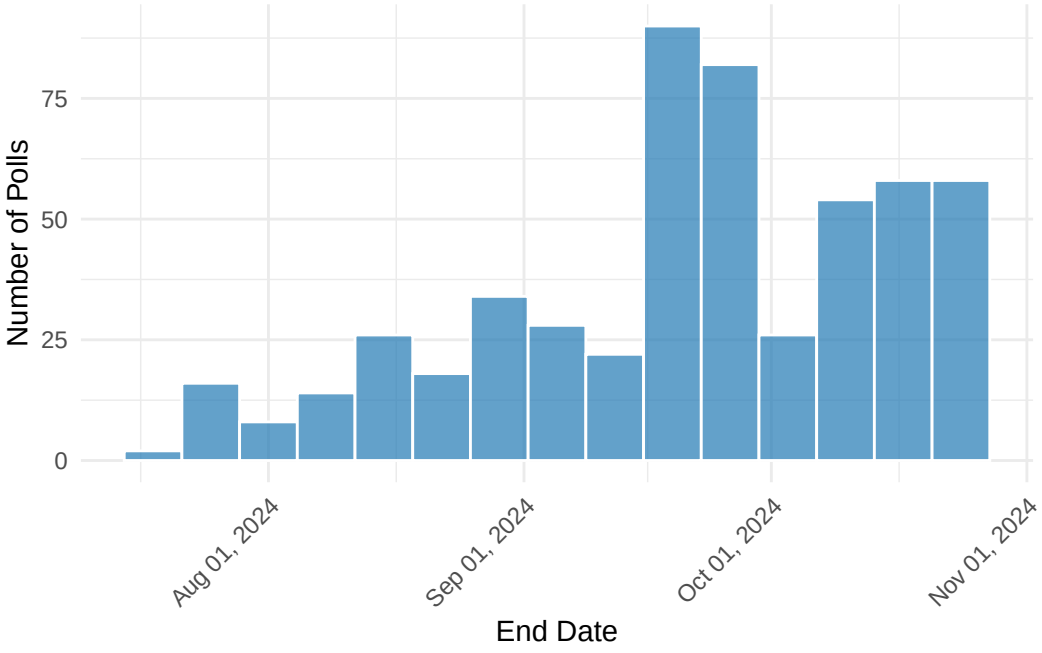

Figure 3: Distribution of Poll End Date, 2024 Cycle

The end date indicates the date that a poll ends on. The dates are graphed separately between the 2020 and 2024 cycles for ease of viewing. Notice how there is a disproportionately large uptick in the number of polls that end on the end of a month.

2.4.2 Sponsors

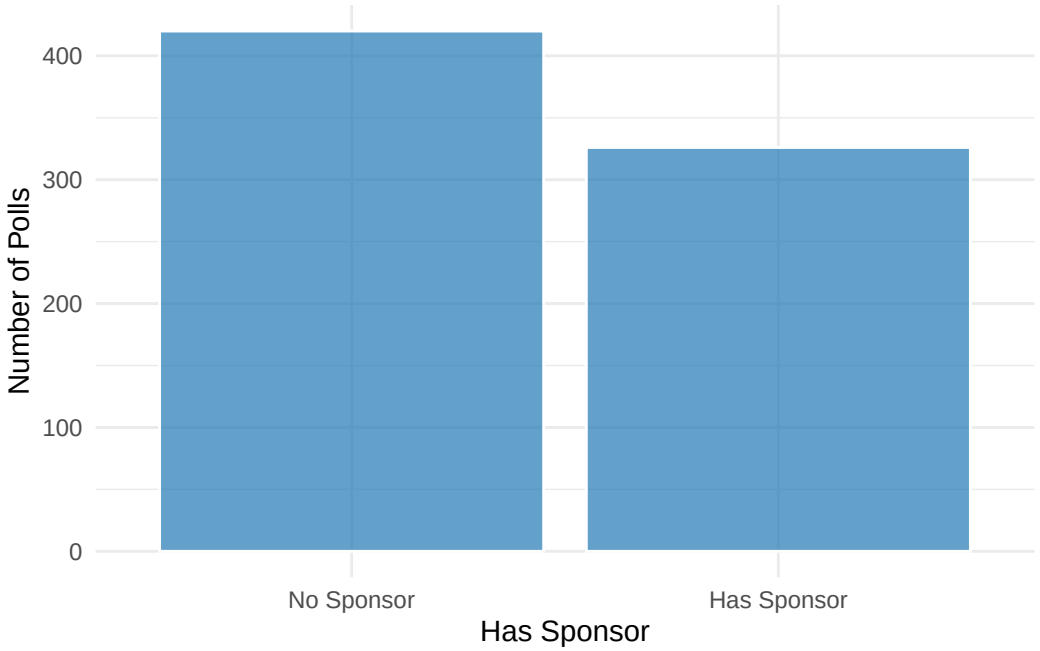

Figure 4: Number of Polls with and without Sponsors

Has_sponsor is a categorical variable that keeps track of whether the poll is sponsored by a third party other than the one conducting the poll itself. This is distributed roughly 60-40, with the larger portion of polls being independently funded (i.e. no sponsors).

2.4.3 Transparency Score

The transparency score is a metric given to a poll/pollster by 538 that measures the transparency of the methodology with which the poll is conducted. This is done through a 10-question yes/no checklist on how much of the methodology is available to the public - note how this is completely separate from the quality of the methodology itself. Each question is graded on a 0/0.5/1 scale, for a combined transparency score that ranges from 0 to 10 (Morris 2024a). Most polls in our dataset miss the mark on one or two criteron, netting a final transparency score of 9 - this is partially due to the data cleaning process removing many relatively worsely conducted polls.

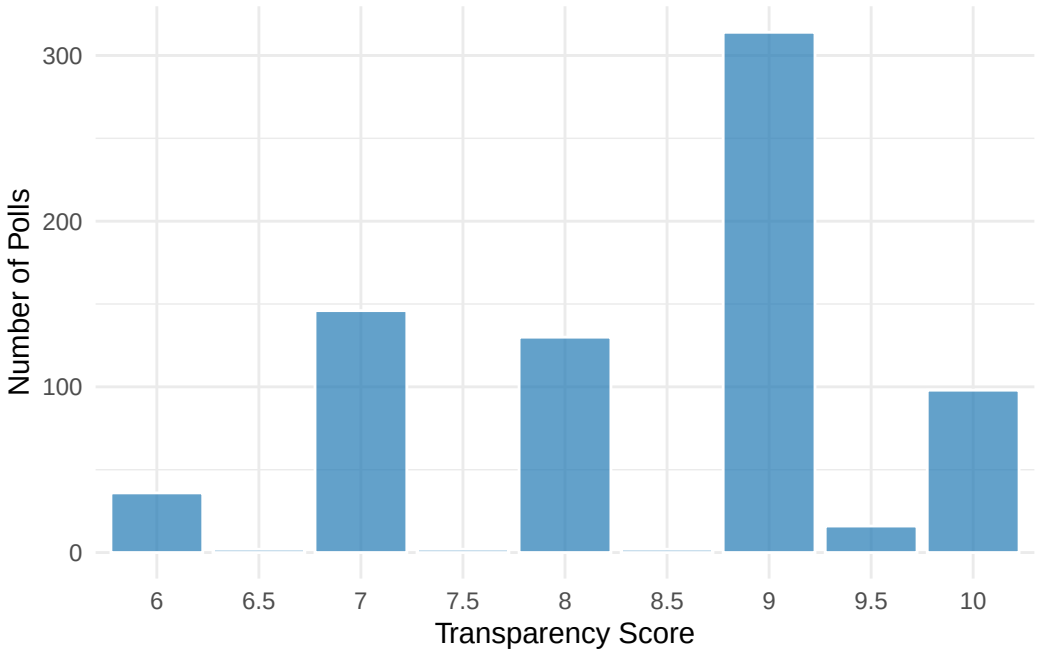

Figure 5: Distribution of Poll Transparency Scores

2.4.4 Sample Size

Sample size, finally, is a measure of the number of participants in each poll. Around 1000 participants seems to overwhelmingly be the final survey size that pollsters decide on, with only single-digit numbers of polls (with rating >= 2.5) boasting more than four thousand respondents.

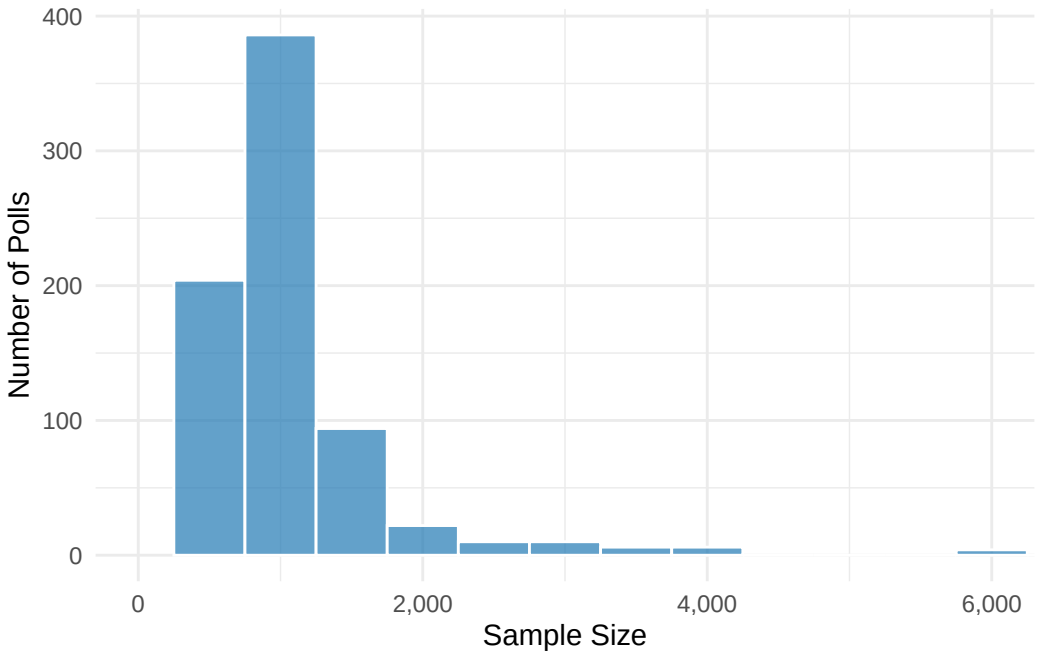

Figure 6: Distribution of Poll Sample Sizes

3 Model

The goal of our modelling strategy is twofold. Firstly, we seek to predict the winning candidate on election day by calculating the expected win rate of Vice President Kamala Harris and Former President Donald Trump in the seven swing states: Pennsylvania, Georgia, North Carolina, Michigan, Arizona, Wisconsin, and Nevada (The New York Times 2024b). Secondly, we hope to predict the candidate who has the popular vote on election day, controlling for the potential bias of national polls by accounting for the difference between polled popular vote and the actual popular vote on election day for the 2020 cycle.

To achieve the first goal, the polls dataset was divided by individual states, and only data points from the seven swing states were selected, forming seven smaller datasets. Then, for each state, the harris_win variable was constructed by comparing the support rate of the candidates in each poll. The variable was set to 1 if the support rate of Harris is greater than or equal to Trump's support rate and 0 when Trump's support rate is greater. A logistic model was constructed for each state, using harris_win as the outcome and previously stated predictors as inputs.

Then, after the models were constructed, we conducted 1000 simulations of election results with randomly-generated parameters. The date parameter was set to election day, and the other three predictor variables were simulated from a normal distribution with the mean and

standard deviation of the training dataset. We then averaged the values to obtain our final prediction for the probability of a Harris win in the state of interest.

Then, using the probabilities for each state and their respective electoral college votes, we calculated the probability that Vice President Harris successfully obtains the 44 required electoral college votes to secure the presidency. This is done via an R script that can be found in the end of /scripts/05-model_data.R.

For the popular vote, we utilized the data from 2020 to calculate each pollster's bias by identifying the average difference between the popular vote on election day and their polled national support rates of Trump and President Joe Biden. Next, we adjusted the polled national support rates of pollsters in the 2024 cycle by the calculated gap in 2020. By making this adjustment, we are assuming that the pollsters systematically bias 2024 candidates equally compared to 2020 candidates. We acknowledge that the assumption is most likely false. Ideally, incorporating poll data from earlier election cycles may produce a more accurate indicator of how pollsters systematically bias candidates from the two dominant political parties. However, comprehensive poll data before 2020 is unavailable. Since popular vote is unrelated to the winner of the election, we are using this analysis as an exploration of methods that incorporate historical data without emphasizing predictiveness. Using both adjusted and unadjusted national support levels, multiple linear regressions were created with support levels as the outcome and stated predictors as inputs to predict candidates' support rate over time. Adjusted and unadjusted logistic regressions were made using harris_win as the outcome and the same predictors to predict the candidate with the popular vote on election day. The attempt to control for historical bias was not applied to the swing states due to a lack of data.

3.1 Model set-up

3.1.1 Logistic models

Define y_i as the binary outcome variable indicating whether Harris wins (1) or not (0) for the i^{th} observation. x_{1i} is the end_date for the i^{th} observation. x_{2i} is the binary value of has_sponsor for the i^{th} observation. x_{3i} is the transparency_score for the i^{th} observation. x_{4i} is the sample_size for the i^{th} observation.

The logistic regression model can be expressed as follows:

$$y_i \sim \text{Bernoulli}(p_i) \tag{1}$$
$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} \tag{2}$$

Line (1) specifies that the outcome variable y_i follows a Bernoulli distribution with probability p_i of success (i.e., Harris winning). Line (2) links the linear combination of predictors to the

probability p_i, where β_0 is the intercept term, and $\beta_1, \beta_2, \beta_3, \beta_4$ are coefficients corresponding to each predictor variable above.

This logistic regression model was applied independently to the data on each swing state and the adjusted and unadjusted popular vote polls.

3.1.2 Multiple linear regression models

Define y_i as the percentage national support for Harris or Trump for the i^{th} observation. x_{1i} is the end_date for the i^{th} observation. x_{2i} is the binary value of has_sponsor for the i^{th} observation. x_{3i} is the transparency_score for the i^{th} observation. x_{4i} is the sample_size for the i^{th} observation.

The multiple linear regression model can be expressed as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} \tag{3}$$

Where β_0 is the intercept term, and $\beta_1, \beta_2, \beta_3, \beta_4$ are coefficients corresponding to each predictor variable.

Separate multiple linear regression models were applied to the polled national support rate of Harris and Trump, with and without adjustments from historical data.

3.1.3 Model justification

We decided to use a logistic model to predict the winner on election day because it directly outputs candidates' chance of winning each swing state, enabling us to use a probabilistic function to identify their overall chance of winning. Comparatively, a linear regression model merely predicts the level of support for candidates, which fails to identify a winner.

We identified the AUC values and ROC plots for each of the seven swing-state logistic regressions. The ROC plots are presented in the Additional Data Details section (Section B), from Figure 9 to Figure 15. The AUC values are: 0.90 for Pennsylvania, 0.88 for Georgia, 0.74 for North Carolina, 0.93 for Michigan, 1.00 for Arizona, 0.86 for Wisconsin, and 0.65 for Nevada. All states other than Nevada have acceptable or excellent AUC values, indicating that the respective logistic models are reliable when distinguishing cases where Harris wins. The low AUC value for Nevada is potentially caused by the limited sample size; there are only 9 polls conducted in Nevada.

Moreover, the sample size restrictions were the reason why we did not use training and testing data to validate our logistic models. The state with the greatest number of polls, Pennsylvania, only records 26 polls. Utilizing training and testing datasets implies that the datasets may

include only a couple of data points in some states, which is insufficient to construct or test the model. On top of that, according to (De Mulder, Bethard, and Moens 2015), achieving low training error, such as by using a training-MSE-minimising regression model such as our own, is sufficient to ensure the predictive capabilities of a model bounded by some constant variance term.

For the analysis of the national popular vote, the logistic model of adjusted polls possess an AUC value of 0.98, and the logistic model of unadjusted polls have an AUC of 0.94, indicating high reliability of categorization. Their respective ROC plots are presented as Figure 16 and Figure 17.

RMSE values and R squared values were calculated for the multiple linear regression models predicting popular vote. The RMSE values for all four models, adjusted and unadjusted support levels of the two candidates, are greater than 1, indicating that the fitted values and the support levels of actual polls are more than 1 percent apart on average, which is significant when analyzing popular vote. Moreover, all R squared values other than the model for the adjusted support rate of Trump, are smaller than 0.3, indicating the inability of the models to account for the variance of national support rates in the data. The model for the adjusted support rate of Trump has a R squared value of 0.69, indicating a relatively high proportion of variance in polled support rates being explained by the multiple linear regression model. This was expected, as the predictors such as a poll's end date and transparency does not necessarily correlate with the outcome of the polls. More predictive variables used by FiveThirtyEight (Morris 2024a) for their predictions are unavailable in the raw dataset, which is a key limitation of the data.

4 Results

4.1 Swing states

Table 1 shows the average likelihood percentage of Harris winning each swing state, calculated by using the logistic model on 1,000 hypothetical data points for each state, generated with the normal distribution of predictors and setting the end_date to election day. It indicates that Harris has a greater probability to win in Pennsylvania, Georgia, and Nevada, while Trump is more likely to win in North Carolina, Michigan, Arizona, and Wisconsin. The electoral votes in each state is also presented.

Table 1: Predicted Harris Win Percentage in Swing States

State	Harris Win Probability (%)	Electoral Votes
Pennsylvania	57.76	20
Georgia	71.47	16
North Carolina	10.72	15

Michigan	47.57	16
Arizona	0.00	11
Wisconsin	38.26	10
Nevada	76.26	6

Combining with the fact that Harris requires 44 electoral votes to win and Trump requires 51, we have found using the probability function that Harris has a 38.3 chance to acquire more than 44 electoral votes, according to the probabilities and electoral votes in Table 1. On the other hand, Trump has a 61.7 probability to acquire more than 51 votes. Thus, our conclusion is that while both candidates possess a significant chance of winning, Trump has the greater probability of acquiring the necessary electoral votes from the seven swing states on election day.

4.2 Popular vote

Figure 7 presents the fitted values of the candidates' adjusted national support rates as well as the data points of polled support rates adjusted for historical pollster bias. It indicates that Trump is expected to have higher national support levels, but the two converges as election day approaches. This result is counter intuitive, as most national polls indicate that Harris will have the popular vote. However, the adjusted values for Trump's support level are significantly greater than the polled values because historically, the actual percentage of the popular vote for Trump in 2020 was significantly greater than his polled percentages of the popular vote. Thus, assuming that the 2024 national polls are equally biased against Trump compared to the 2020 polls, it is difficult to tell who will get the popular vote on election day. However, as previously stated, this assumption is likely false, as the candidates have changed.

Figure 8 Presents the predicted unadjusted national support rates for Harris and Trump. The conclusions one can make from this plot is similar to those of most polls: Harris is expected to win the popular vote on any day, including the election day.

The regression coefficients for the logistic and linear models are presented in the Appendix because they are not the outcome of interest.

5 Discussion

5.1 The illusion of polarization

As mentioned in the introduction, one thing that stands out about the 2024 election in particular is just how polarized the two factions are. Instead of merely being a choice over which individual to helm the country over the next four years, this particular election is being framed more and more as a "fight to protect our democracy" from both sides of the debate stand.

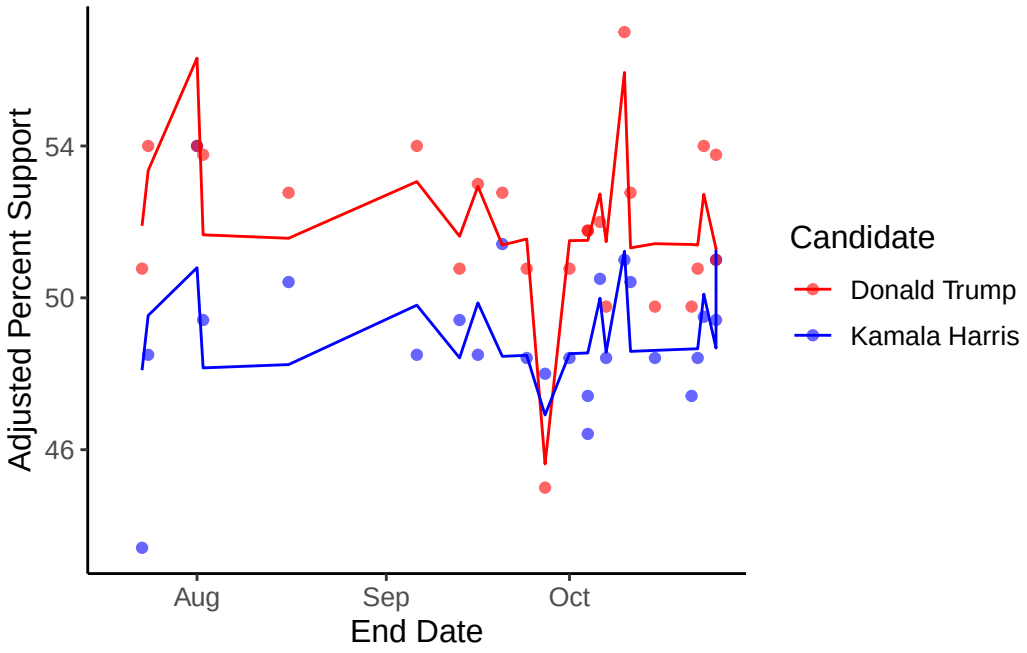

Figure 7: Adjusted Popular Vote Prediction, 2024

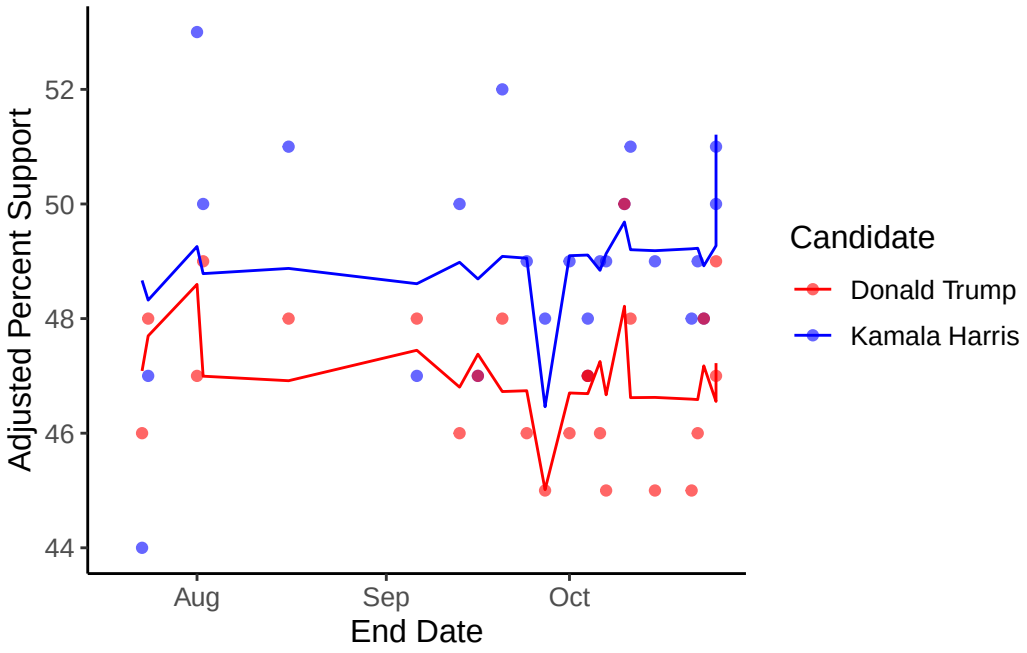

Figure 8: Unadjusted Popular Vote Prediction, 2024

This, while evidently a new level of polarization, merely mirrors the trend in recent years more than anything else: party loyalty has been on a steady uprise since 1972, with loyalty in weak partisan and leaning independent voters increasing by approximately 50% in the 40 years that followed (Abramowitz and Webster 2016). New records were further set in 2016, with Republican validated voters voting Trump a staggering 92% to 4% for Clinton, and Democratic validated voters voting Clinton an equally extreme 94% to 5% Trump support (Pew Research Center 2018).

New voting statistics, however, suggest that this polarization is likely only more prominent in the upper echelons of political representation, however: a Washington Post article found that approximately 1 in 8 women and 1 in 10 men are now voting differently from their spouses compared to before (Bump 2024), a development that suggests a decrease in long-increasing partisanship trends. This is especially pronounced since households in the States overwhelmingly vote together: at least 70% of them do, at least, which means that this value is projected to decrease from 70% to approximately 62% (Hersh and Ghitza 2018). In short: while it seems that ideologies are shifting more and more away from each other, the stats say otherwise. So does the Overton window (Raines 2024).

5.2 Win probabilities versus support ratings: why our predictions are as extreme as they are

Our model led to a win probability prediction with numbers much more skewed towards one side than most pollsters currently predict: under this model, we find that Harris has a 40-60 chance against Trump, whereas most pollsters nowadays hover somewhere between 50-50 to 45-55 (towards either side). This is mainly a consequence of using a multinomial logistic regression model as opposed than other models – instead of predicting the support rates of the two candidates, such a model outputs the probability that a random support rate value from the predicted distribution of Harris support rates is larger than that of the predicted distribution of Trump support rates. This means that in states such as Nevada, where 538 gives Trump a +3 advantage, factoring in for margin of error and standard error means that the probability of Harris winning the state is not in fact 47% but rather much closer to zero, which we decided is a more intuitive way of understanding election odds.

A final reason for this is how swing states are polled much more sample-size-wise than other states, leading to a much higher power value than polls from said other states – a +3 lead in a sample of 1000 people from California's 39 million might not mean too much, but a +3 lead in a sample of 3000 people from Nevada's 3 million represents the actual results for 0.3% of the state's population, holding much more relative weight when it comes to election predictions.

5.3 Weaknesses and next steps

The dataset used for this analysis has a number of clear limitations. First of all, there is a lack of high-quality polls. Though FiveThirtyEight distributes each pollster with a rating out of 3.0, almost all pollsters under the rating of 2.7 produces outcomes that are unreliable and highly volatile. For this research, it was only possible for us to use pollsters with a 3.0 rating when analyzing national polls; we had to lower the standards to 2.5 for the analysis of swing states, as the number of high-quality state-specific polls are very limited. The inclusion of unreliable data reduces the predictability of our logistic models and thus our final outcome. The lack of predictability of our Nevada model is a perfect example of how low-quality data and the lack of data points impacts our ability to create effective models. However, we acknowledge that this is beyond the control of FiveThirtyEight, as they do not conduct the polls.

The second major challenge presented by the dataset is the lack of predictors. FiveThirtyEight utilizes predictive variables that are unavailable to us, such as weighted values that represent the extent to which a pollster is partisan (Morris 2024a). The predictors used in this research, such as the transparency score of the pollster, are not related to their bias towards any candidate from a partisanship point of view, limiting the predictive power of our models.

Appendix

.1 Pollster methodology overview and evaluation

The Quinnipiac University Poll conducts independent polling in swing states. To analyze their methodology, we look specifically into their October 2024 Pennsylvania polls (Quinnipiac University Poll 2024b).

Their target population is likely voters aged 18 and older in Pennsylvania (Quinnipiac University Poll 2024a). To reach this population, they use likely voters aged 18 and older with phone numbers (both landline and cell) as their sampling frame, i.e. the frame of possible subjects that they sample observations from. Quinnipiac University employed Random Digit Dialing (RDD) to generate their sample of 2,186 respondents. This dual-frame approach reflects modern communication patterns, with 1,644 cell phone and 542 landline completions. However, using phones as a sampling frame means they cannot reach voters without phone access, introducing potential coverage bias.

Their sampling approach uses stratification by Census division according to area code, meaning that they divide Pennsylvania into geographic regions before using RDD to sample within each region. This strategy ensures even geographic representation but adds complexity to the sampling process (Keeter 2024). For each selected number, they attempt contact at least three times before marking it as non-responsive. For landline calls, they ask to speak with the household member who has the next birthday, a simple but effective randomization technique. Afterwards, a series of screening questions confirm that the subject is indeed a likely voter, after which the subject's responses are then formally taken as part of the sample.

After collecting responses, Quinnipiac adjusts their data through post-stratification weighting. In this weighting, they compare their sample's demographic composition to known population benchmarks from the Census (like age, gender, education, and race distributions in Pennsylvania) and adjust the weight given to each response to match these benchmarks. For example, if their sample has too few young voters compared to Census data, responses from young voters would be weighted more heavily. While this helps correct for sampling imbalances, it can increase the variance in their estimates if the weights vary substantially (CLYDE, HEMMERLE, and BANCROFT 1963).

The survey administration addresses measurement issues through its design. Live interviewers conduct all interviews, enabling question clarification and generating higher response rates compared to automated systems. However, live interviewers may introduce social desirability bias, where respondents might modify their answers to appear more socially acceptable. This becomes particularly relevant in political polling, where respondents might hesitate to express unpopular political views (AAPOR 2022).

Several types of bias affect the poll's results. Self-selection bias occurs because certain types of people (typically those more politically engaged or with stronger views) are more likely to agree to participate in the survey. Non-response bias arises when people who respond

differ systematically from those who don't - for instance, busier people might be less likely to answer calls, potentially underrepresenting certain occupational groups (Peytchev, Baxter, and Carley-Baxter 2009). Coverage bias means some groups (like those without phones) have no chance of being included in the sample.

Ethics-wise, the survey does a good job of informing individuals of the details of the survey, such as its purpose and how participants' data is to be used. Personally-identifiable data (i.e. name, phone number, etc.) is not collected during the survey other than to ensure that repeat numbers are not drawn, and the usage of live interviewers to conduct the survey ensures that concerns about the ethics of the survey can be voiced and answered on-the-spot. Combined with the lack of financial incentive for participants, which suggests that all participants participate of their complete free will and intention, this is an ethically sound methodology for conducting a political survey.

Finally, the poll also faces common challenges in political polls such as this one - the five-day field period (October 24-28) may miss opinion changes close to election day, and while weighting adjustments help correct for demographic imbalances, they may increase variance in the estimates if some groups need to be weighted heavily to match population benchmarks, for example.

Overall, Quinnipiac's methodology represents a balanced approach to managing practical constraints and statistical rigor in modern political polling - while some common biases are still likely to skew the poll results off the true support levels for Harris, for example, the poll uses methods such as post-stratification weighting to tradeoff biases at the cost of model variance (Cohn 2024). Using live interviewers unavoidably introduces social desirability bias, however, and significantly increases poll costs per quota. Modifying the methodology to remove this aspect of the survey would potentially reduce bias and allow for larger samples to be taken, in turn opening up possibilities for cross-validation and the such, which then reduces the effect of increased model variance on the final results.

.2 Idealised methodology

With a $100,000 budget, our approach focuses on producing accurate state-level estimates in key battleground states, which would then inform our national forecast. We prioritize Pennsylvania, Michigan, Wisconsin, Georgia, Arizona, and Nevada, allocating resources proportionally based on each state's electoral importance and expected margin of victory.

Our sampling strategy employs both probability and non-probability methods. Probability sampling (60% of budget, i.e. $60k) means every member of our target population has a known, non-zero chance of being selected - this allows us to calculate proper margins of error and make statistical inferences about the population. For this, we use dual-frame random digit dialing (RDD) for phone surveys and address-based sampling (ABS) for mail-to-web recruitment. RDD involves generating random phone numbers within active area codes, while ABS uses the U.S. Postal Service's delivery database as a sampling frame, both proven methods of sampling

(CUMMINGS 1979). The ABS approach helps reach households without reliable phone access. We stratify our sample by geography, demographics, and previous voting patterns to ensure representation across key subgroups - meaning we divide the population into these subgroups and sample from each independently.

For non-probability sampling (40% of budget, i.e. $40k), where respondents' selection probabilities are unknown and not everyone has a chance of being selected, we recruit through multiple online panel vendors and use targeted social media advertising to reach traditionally underrepresented groups. While this approach introduces potential selection bias because participants self-select into the sample, it helps reach younger voters who are less responsive to traditional survey methods (Reg Baker and Tourangeau 2013). We implement quota sampling within these non-probability samples to match key demographic targets - for example, stopping collection from certain demographic groups once their quota is filled.

Respondent recruitment uses multiple contact methods - mail, email, text, and phone - with attempts made at different times and days to maximize response rates. We offer a $10 gift card incentive for completed surveys and provide both English and Spanish language options. This mixed-mode contact strategy helps reduce non-response bias by providing multiple ways to participate.

Data validation is crucial for maintaining quality. We cross-reference responses with voter files where available - meaning we check if respondents' self-reported registration status matches official records. We screen for duplicate responses using IP addresses and phone numbers, and implement attention checks within the survey (questions with known correct answers to ensure respondents are reading carefully). Speed checks identify rushed responses that might indicate low-quality data by flagging completions that fall below a minimum reasonable completion time, while consistency checks across related questions help identify potentially fraudulent responses by looking for logical contradictions in answers.

Our weighting approach uses post-stratification to known population benchmarks - this means we adjust the weight given to each response so that our sample matches known population characteristics. For example, if our sample has 30% college graduates but the population has 40%, we would give more weight to responses from college graduates. We include demographics (age, race, education, gender), geographic location, past voting behavior, and party registration in our weighting scheme. We produce daily estimates using a 7-day rolling average, which helps smooth out daily fluctuations while remaining responsive to real changes in voter preferences (Courtney Kennedy 2022).

Finally, how ethics are handled is a crucial part of any survey methodology. In this idealised methodology, consent will be asked for at the beginning of the survey, full disclosure of how information is used will be given beforehand and no self-identifiable information will be recorded (so no names, phone numbers, etc.). The $10 incentive is enough to hopefully make it wortile for participants' time, but also not ideally not significant enough of an incentive to make individuals suppress otherwise deal-breaking concerns with the survey purely for the sake of the incentive.

The survey instrument itself focuses on six key areas: screening questions to identify likely voters, voting intentions (including direct questions about Trump vs. Harris preferences), political preferences, demographics, issue priorities, and media consumption patterns. We've implemented this survey design in Google Forms, which can be found here: https://forms.gle/pk7vDiMHwEGLMK849

This methodology balances statistical rigor with practical constraints, while acknowledging and attempting to address the key challenges in modern political polling: declining response rates, coverage bias, and the increasing difficulty of reaching a representative sample of likely voters.

.3 Idealised survey

The survey, made using Google Forms, is linked here: https://forms.gle/pk7vDiMHwEGLMK849 Note that the questions are identical for both the phone and online surveys bar q6. A copy of the survey that is identical to the one implemented in the Google Forms above is presented below: Thank you for participating in this survey about the 2024 U.S. Presidential Election. This survey is part of a research project at the University of Toronto studying voting intentions and political attitudes.

Estimated completion time: 8-10 minutes

Your responses will be kept confidential and used only for research purposes. Email information, and any other information that may personally identify you, is not gathered. You may skip any questions you prefer not to answer, though complete responses are most helpful for our research.

For questions or concerns about this survey, please contact: andrew.goh@mail.utoronto.ca

SCREENING SECTION: Q1. Are you 18 years of age or older?

Yes No [END SURVEY]

Q2. Are you a U.S. citizen?

Yes No [END SURVEY]

Q3. Are you registered to vote at your current address?

Yes No Not sure [If No or Not sure: Do you plan to register before the November 2024 election?]

VOTING INTENTION: Q4. How likely are you to vote in the 2024 presidential election?

Definitely will vote Probably will vote Might or might not vote Probably will not vote Definitely will not vote

Q5. If the 2024 presidential election were held today, and the candidates were Kamala Harris (Democrat) and Donald Trump (Republican), who would you vote for?

Kamala Harris Donald Trump Another candidate (please specify) Would not vote Not sure

ATTENTION CHECK: Q6. To ensure you're reading carefully, please select "Somewhat disagree" for this question: "I enjoy following political news."

Strongly agree Somewhat agree Somewhat disagree Strongly disagree No opinion

POLITICAL PREFERENCES: Q7. Generally speaking, you consider yourself a:

Democrat Republican Independent Something else (please specify)

Q8. How would you rate the current state of the U.S. economy?

Very poor Poor Fair Good Excellent

ISSUE PRIORITIES: Q9. Which ONE of the following issues is most important to you when deciding how to vote?

Economy and jobs Immigration Healthcare Climate change Crime and public safety Education National security Abortion rights Gun policy Something else (please specify)

For each of the following issues, please indicate whether you think Kamala Harris or Donald Trump would do a better job handling it:

Q10. Economy and jobs:

Kamala Harris would do better Donald Trump would do better No difference Not sure

Q11. Human rights and freedom of speech: [Same options] Q12. Abortion: [Same options] Q13. Healthcare: [Same options] Q14. Immigration: [Same options] Q15. National security: [Same options]

MEDIA CONSUMPTION: Q16. Where do you most often get your news about politics? (Select all that apply)

Network TV news (ABC, CBS, NBC) Cable TV news (CNN, Fox News, MSNBC) Local TV news Radio Print newspapers News websites Social media Friends and family Other (please specify)

Q17. How many hours per day do you typically spend following news about politics?

Less than 1 hour 1-2 hours 2-4 hours More than 4 hours

DEMOGRAPHICS: Q18. What is your age?

18-24 25-34 35-44 45-54 55-64 65 or older

Q19. What is your gender?

Male Female Non-binary/Other Prefer not to say

Q20. What is your race/ethnicity? (Select all that apply)

White Black or African American Hispanic or Latino Asian Native American Other (please specify) Prefer not to say

Q21. What is the highest level of education you have completed?

Less than high school High school graduate Some college Associate's degree Bachelor's degree Graduate degree Prefer not to say

Q22. What was your total household income before taxes in 2023?

Under $25,000 $25,000-$49,999 $50,000-$74,999 $75,000-$99,999 $100,000-$149,999 $150,000 or more Prefer not to say

CONSISTENCY CHECK: Q23. Looking ahead to November 2024, if Kamala Harris is the Democratic nominee and Donald Trump is the Republican nominee, how do you think you will vote?

Kamala Harris Donald Trump Another candidate (please specify) Would not vote Not sure

[END OF SURVEY] Thank you for completing this survey about the 2024 U.S. Presidential Election. Your responses will help us better understand voter preferences and political attitudes across the country. If you have any questions about this research or would like to be informed about the results, please contact andrew.goh@mail.utoronto.ca. Your time and participation is greatly appreciated.

A Additional data details

B Model details

B.1 Diagnostics

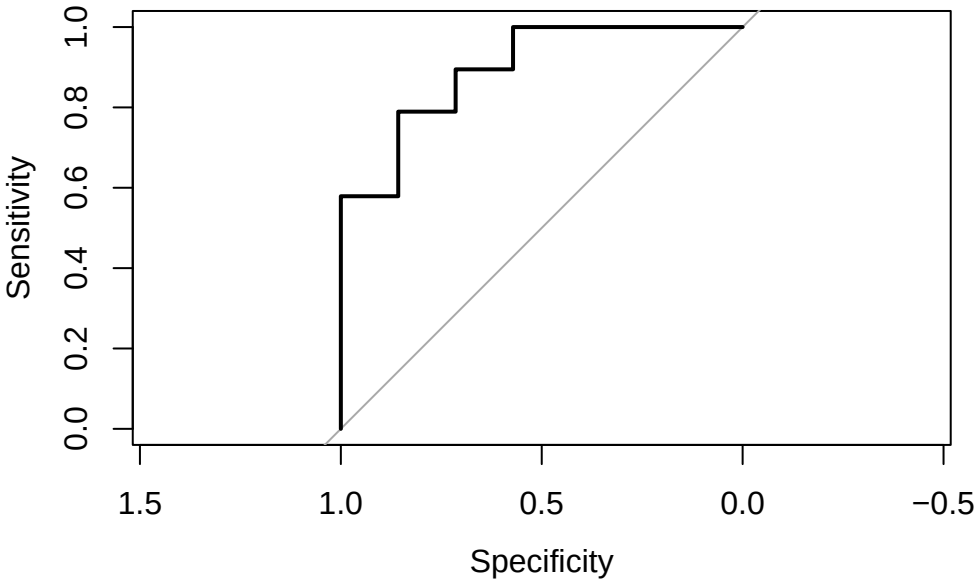

Figure 9: ROC Curve for Harris Win Prediction, Pennsylvania

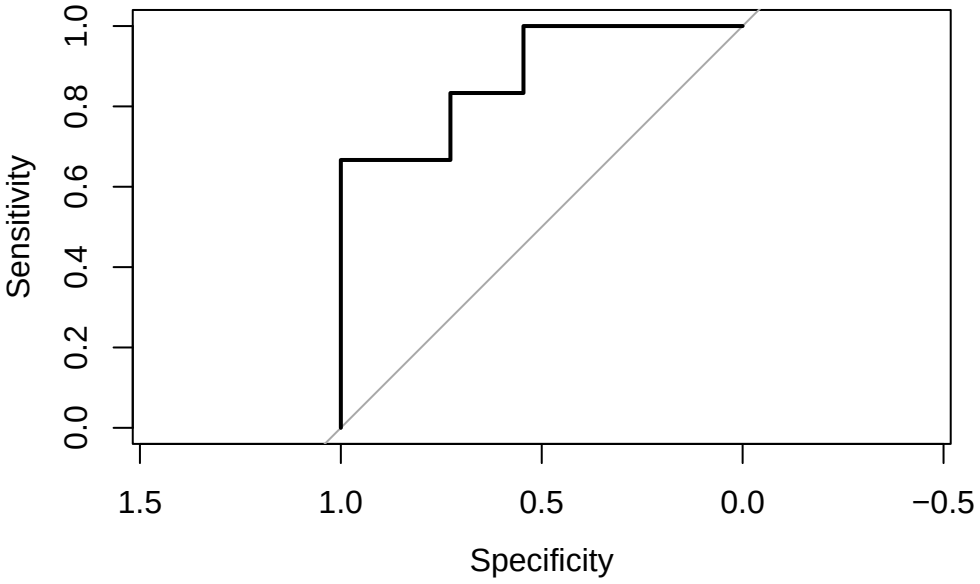

Figure 10: ROC Curve for Harris Win Prediction, Georgia

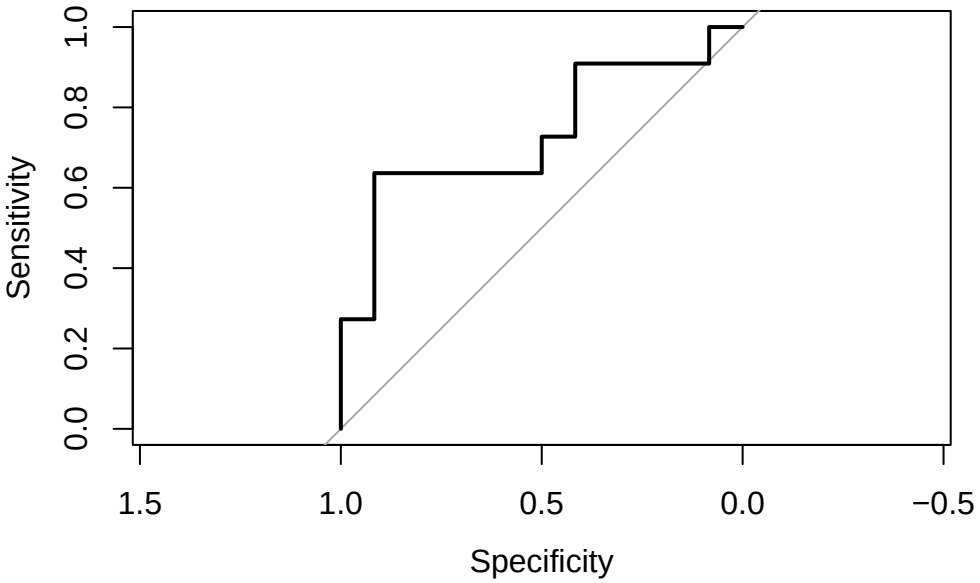

Figure 11: ROC Curve for Harris Win Prediction, North Carolina

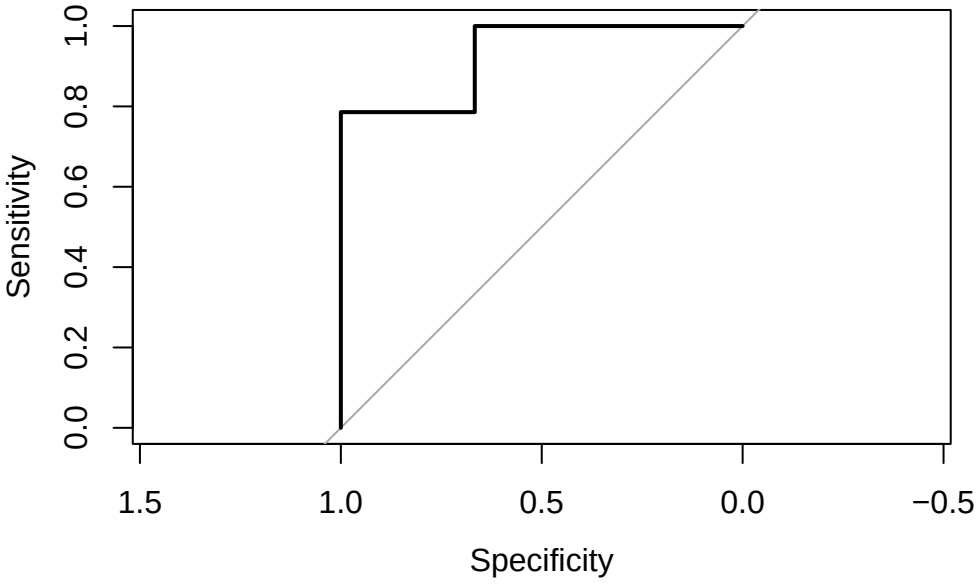

Figure 12: ROC Curve for Harris Win Prediction, Michigan

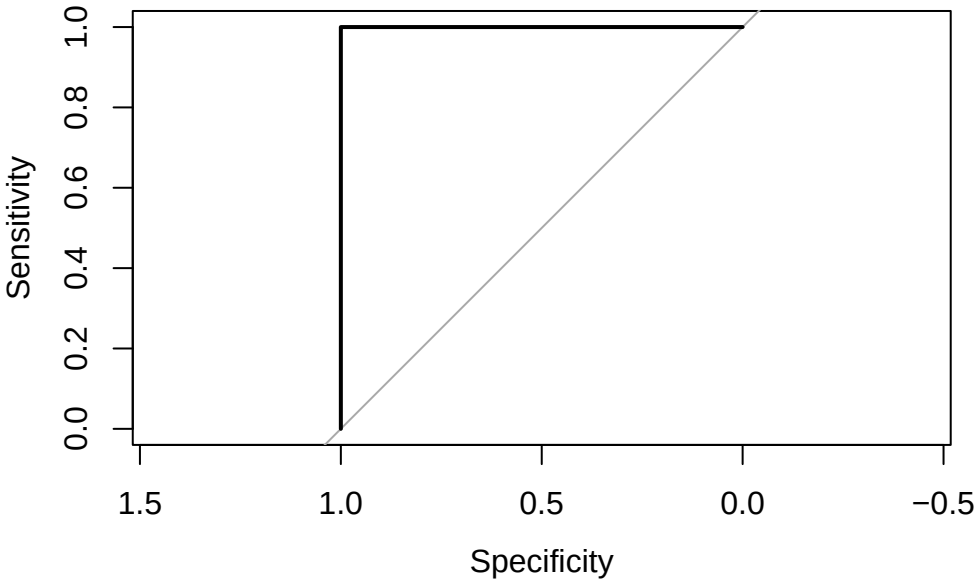

Figure 13: ROC Curve for Harris Win Prediction, Arizona

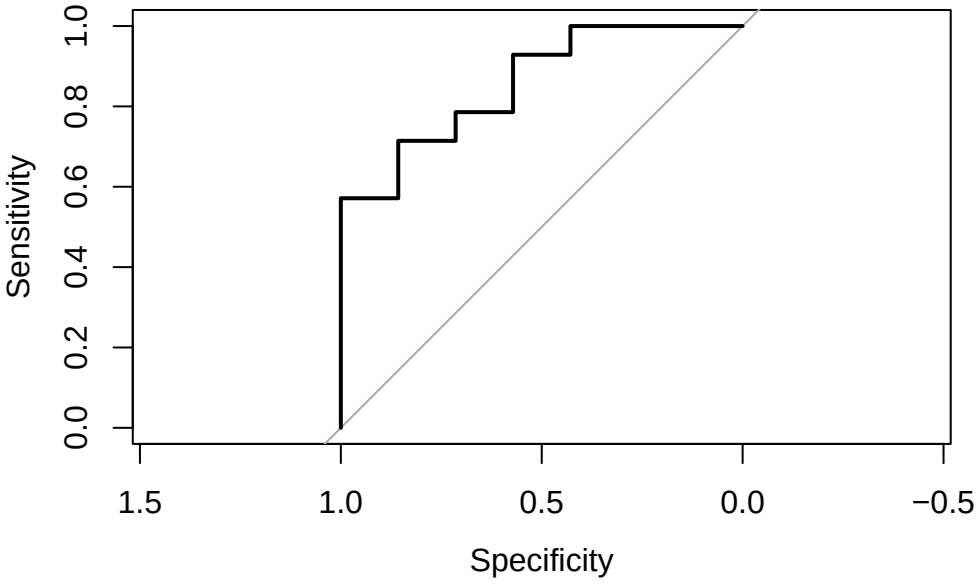

Figure 14: ROC Curve for Harris Win Prediction, Wisconsin

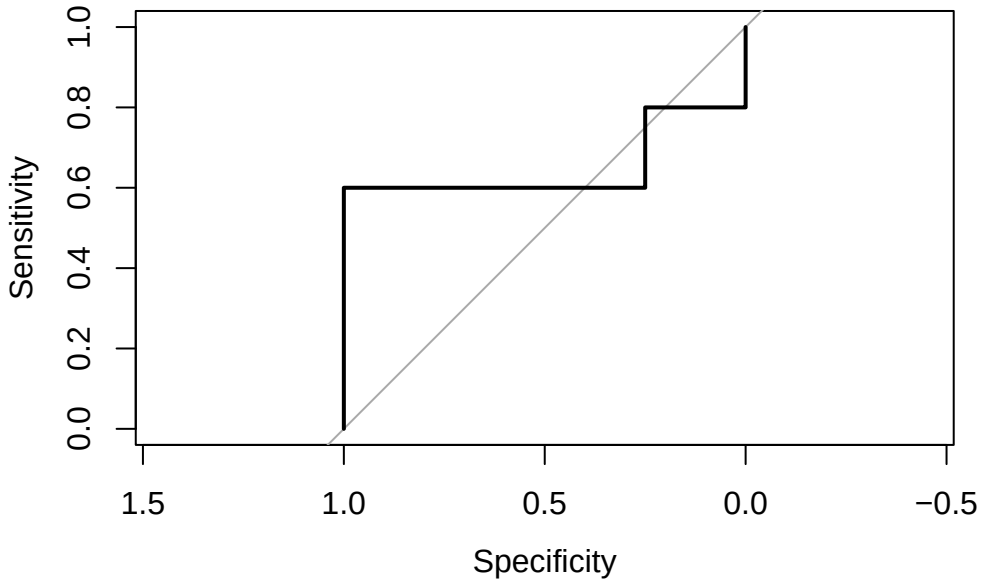

Figure 15: ROC Curve for Harris Win Prediction, Nevada

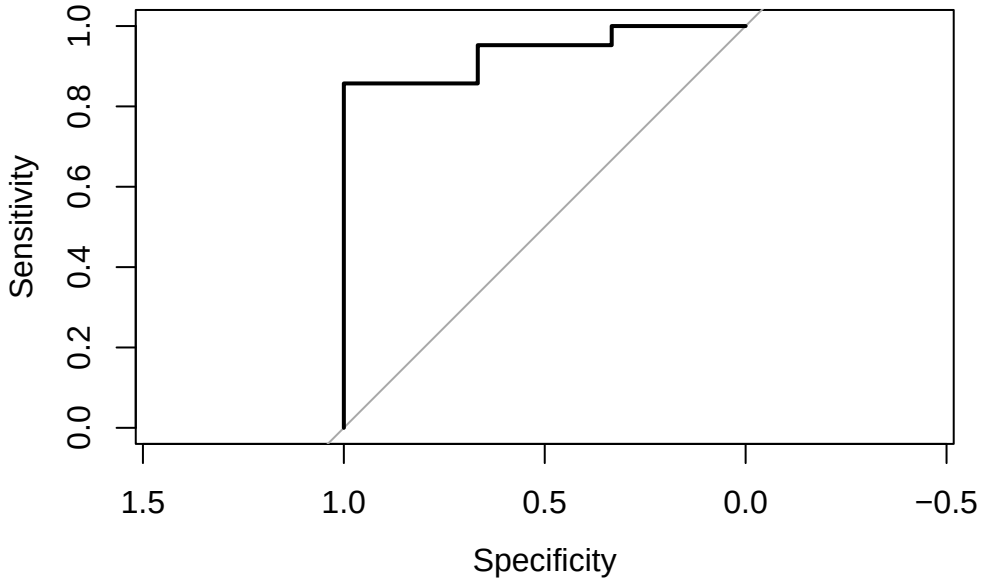

Figure 16: ROC Curve for Popular Vote Prediction

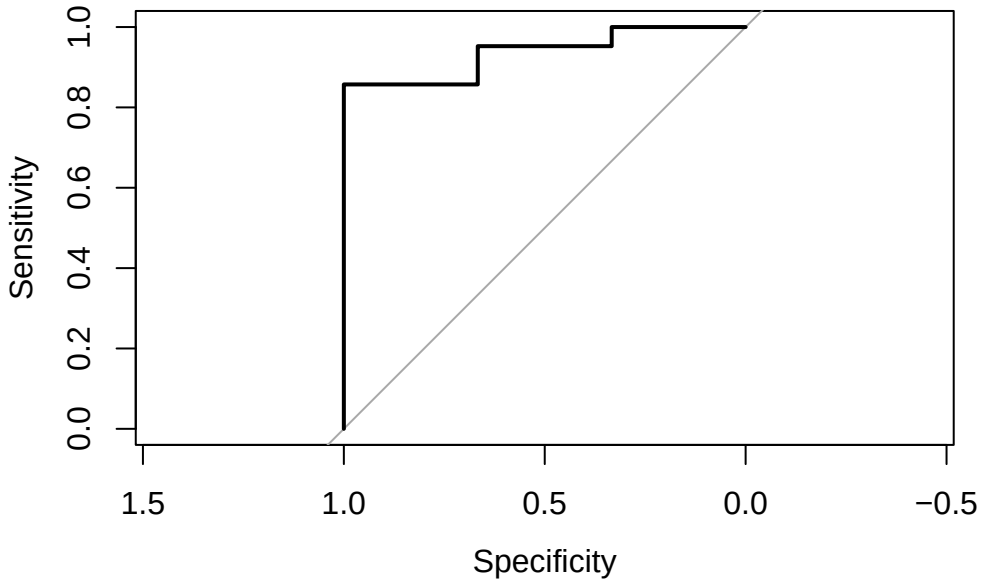

Figure 17: ROC Curve for Popular Vote Prediction, Adjusted

B.2 Coefficients

Table 2: Logistic Regression Coefficients for Pennsylvania

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	662.16	671.93	0.99	0.32
end_date	-0.03	0.03	-0.99	0.32
has_sponsor	-3.55	3.01	-1.18	0.24
transparency_score	1.00	0.66	1.52	0.13
sample_size	0.00	0.00	-1.05	0.29

Table 3: Logistic Regression Coefficients for Georgia

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	-2095.09	1448.99	-1.45	0.15
end_date	0.10	0.07	1.44	0.15
has_sponsor	6.38	3.81	1.68	0.09
transparency_score	1.96	1.29	1.51	0.13
sample_size	-0.01	0.01	-1.70	0.09

Table 4: Logistic Regression Coefficients for Michigan

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	1590.16	1828.03	0.87	0.38
end_date	-0.08	0.09	-0.87	0.38
has_sponsor	19.93	6387.34	0.00	1.00
transparency_score	1.33	1.06	1.26	0.21
sample_size	0.00	0.01	-0.14	0.89

Table 5: Logistic Regression Coefficients for North_Carolina

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	1090.80	657.79	1.66	0.10
end_date	-0.05	0.03	-1.66	0.10
has_sponsor	-0.61	1.26	-0.48	0.63
transparency_score	0.24	0.49	0.49	0.62
sample_size	0.00	0.00	1.23	0.22

Table 6: Logistic Regression Coefficients for Arizona

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	50523.04	92616479.19	0	1
end_date	-2.53	4650.19	0	1
has_sponsor	-39.09	126344.21	0	1
transparency_score	-10.73	38576.48	0	1
sample_size	0.09	477.60	0	1

Table 7: Logistic Regression Coefficients for Wisconsin

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	1056.70	825.99	1.28	0.20
end_date	-0.05	0.04	-1.28	0.20
has_sponsor	-2.85	1.83	-1.56	0.12
transparency_score	0.51	0.54	0.94	0.35
sample_size	0.00	0.00	-0.91	0.36

Table 8: Logistic Regression Coefficients for Nevada

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	-531.40	838.39	-0.63	0.53
end_date	0.03	0.04	0.63	0.53
has_sponsor	0.05	1.54	0.03	0.97
transparency_score	0.75	1.00	0.75	0.46
sample_size	0.00	0.01	0.33	0.74

References

AAPOR. 2022. *Sampling Methods for Political Polling*. American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/12/Sampling-Methods-for-Political-Polling-508.pdf.

Abramowitz, Alan, and Steven Webster. 2016. "All Politics Is National: The Rise of Negative Partisanship and the Nationalization of u.s. House and Senate Elections in the 21st Century." stevenwwebster.com/research/all_politics_is_national.pdf.

Beauchamp, Zack. 2024. "It's Not Alarmist: A Second Trump Term Really Is an Extinction-Level Threat to Democracy." *Vox*, January. https://www.vox.com/policy/381636/trump-2024-democracy-threat-orban-second-term.

Bump, Philip. 2024. "1 in 8 Women Say They've Secretly Voted Differently Than Partners." *The Washington Post*, November. https://www.washingtonpost.com/politics/2024/11/01/women-voting-secret-choice/.

CLYDE, ROBERT W., WILLIAM J. HEMMERLE, and T. A. BANCROFT. 1963. "AN APPLICATION OF 'POST STRATIFICATION' TECHNIQUE IN LOCAL TV ELECTION PREDICTIONS." *Public Opinion Quarterly* 27 (3): 467–72. https://doi.org/10.1086/267189.

Cohn, Nate. 2024. *How One Polling Decision Is Leading to Two Distinct Stories of the Election*. The New York Times. https://www.nytimes.com/2024/10/06/upshot/polling-methods-election.html.

Courtney Kennedy, Scott Clement, Mark Blumenthal. 2022. *AN EVALUATION OF 2016 ELECTION POLLS IN THE UNITED STATES*. American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/12/Sampling-Methods-for-Political-Polling-508.pdf.

CUMMINGS, K. MICHAEL. 1979. "Random Digit Dialing: A Sampling Technique for Telephone Surveys." *Public Opinion Quarterly* 43 (2): 233–44. https://doi.org/10.1086/268514.

De Mulder, Wim, Steven Bethard, and Marie-Francine Moens. 2015. "A Survey on the Application of Recurrent Neural Networks to Statistical Language Modeling." *Computer Speech & Language* 30 (1): 61–98. https://doi.org/https://doi.org/10.1016/j.csl.2014.09.005.

FiveThirtyEight. 2024. "Dataset: US Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hersh, Eitan, and Yair Ghitza. 2018. "Mixed Partisan Households and Electoral Participation in the United States." *PLoS ONE* 13 (10): e0203997. https://doi.org/10.1371/journal.

pone.0203997.

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

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Morris, G. Elliot. 2024a. *How 538's Pollster Ratings Work*. FiveThirtyEight. https://abcnews.go.com/538/538s-pollster-ratings-work/story?id=105398138.

———. 2024b. *Trump Leads in Swing-State Polls and Is Tied with Biden Nationally*. FiveThirtyEight. https://abcnews.go.com/538/trump-leads-swing-state-polls-tied-biden-nationally/story?id=109506070.

Müller, Kirill, and Lorenz Walthert. 2023. *Styler: Non-Invasive Pretty Printing of r Code*. https://github.com/r-lib/styler.

Musk, Elon. 2024. "This Is Actually Happening!" X Post. https://x.com/elonmusk/status/1770030227390914624.

Pew Research Center. 2018. "An Examination of the 2016 Electorate, Based on Validated Voters." Research Report. Pew Research Center. https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/.

Peytchev, Andy, Rodney K. Baxter, and Lisa R. Carley-Baxter. 2009. "Not All Survey Effort is Equal: Reduction of Nonresponse Bias and Nonresponse Error." *Public Opinion Quarterly* 73 (4): 785–806. https://doi.org/10.1093/poq/nfp037.

Quinnipiac University Poll. 2024a. "Pennsylvania October 30, 2024 Poll Sample and Methodology." Poll Results. Quinnipiac University. https://poll.qu.edu/images/polling/pa/pa10302024_demos_pece24.pdf.

———. 2024b. "Trump Leads Biden in Head-to-Head Matchup, Quinnipiac University National Poll Finds; Majority of Voters Say Trump Not Fit to Be President." Poll Release. Quinnipiac University. https://poll.qu.edu/poll-release?releaseid=3916.

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

Raines, Jack. 2024. "The Overton Window Has Shifted." Blog Post. *Medium*. https://medium.com/@jackraines/the-overton-window-has-shifted-68dbb3305cbc.

Reg Baker, Nancy A. Bates, J. Michael Brick, and Roger Tourangeau. 2013. *REPORT OF THE AAPOR TASK FORCE ON NONPROBABILITY SAMPLING*. American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/11/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77.

Simon, Noah, Jerome Friedman, Robert Tibshirani, and Trevor Hastie. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

Tay, J. Kenneth, Balasubramanian Narasimhan, and Trevor Hastie. 2023. "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software* 106 (1): 1–31. https://doi.org/10.18637/jss.v106.i01.

The New York Times. 2024a. "How the New York Times Calculates Polling Averages." Methodology Article. https://www.nytimes.com/article/election-polling-averages-methodology.html.

———. 2024b. "Latest 2024 Presidential Election Polls." Interactive News Article. The New York Times. https://www.nytimes.com/interactive/2024/us/elections/polls-president.html.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.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://dplyr.tidyverse.org.