Forecasting the 2024 U.S. Presidential Election: Analyzing National Polling Trends, Sample Size Impact, and State-Level Variations[*]

National Polls Indicate a 46.3% Average Support for Trump, with Sample Size and State Differences Showing Significant Influence in Arizona, California, and Colorado

Jason Yang Jerry Xia Peter Fan

November 4, 2024

This paper develops a forecast model for the 2024 U.S. Presidential Election, analyzing over 1,000 aggregated national and state-level polls. The model uses variables including candidate support percentages, pollster reliability, sample size, and geographic distribution, with recency weighting applied to enhance the relevance of current voter sentiment. Results show Donald Trump leading nationally with an average of 46.3% support. In battleground states, Trump is projected to edge out Harris by 3.5% in Arizona, while Harris narrows the gap to within 1% in Pennsylvania. Monte Carlo simulations estimate the distribution of Electoral College outcomes, factoring in polling uncertainty and state-level variations. The analysis indicates that sample size and state-specific factors significantly influence prediction accuracy, though non-response bias and polling methodology may limit forecast reliability.

1 Introduction

The 2024 U.S. Presidential Election stands as a important moment in American politics, with widespread implications for both domestic policy and international relations. Accurately forecasting the election outcome is crucial for stakeholders ranging from policymakers to the general public. This study leverages national and state-level polling data to analyze voter

[*] Code and data are available at: https://github.com/Jerryx2020/US_election_prediction/tree/main.

sentiment toward the two main presidential candidates, Kamala Harris and Donald Trump. By focusing on polling trends since July 2024, we aim to provide an up-to-date assessment that reflects the most recent shifts in the political landscape.

To conduct our analysis, we utilize the statistical programming language R (R Core Team (2023)), along with packages that enhance data manipulation and visualization. Key libraries include the tidyverse collection (Wickham et al. (2019)), which offers a set of tools for data science tasks; janitor (Firke (2023)), which simplifies data cleaning processes; and ggplot2 (Wickham (2016)), for data visualization. These tools enable us to efficiently manage complex datasets and generate inferential visualizations that indicate polling trends across different regions and demographics.

Our methodology involves developing a multiple linear regression model to predict candidate support percentages based on variables such as pollster reliability, sample size, poll end dates, and geographic distribution. Additionally, we implement Monte Carlo simulations to estimate the distribution of possible Electoral College outcomes, accounting for state-level variations and uncertainties inherent in polling data. By integrating these statistical techniques, we aim to provide a robust framework for election forecasting that contributes valuable insights into the 2024 electoral race.

2 Data

2.1 Data Overview

This study employs various R packages (R Core Team (2023)) for data cleaning and analysis, including libraries from tidyverse (Wickham et al. (2019)), ggplot2 (Wickham (2016)), rstanarm (Goodrich et al. (2022)), testthat (Wickham (2011)), here (Müller (2020)), janitor (Firke (2023)), lubridate (Grolemund and Wickham (2011)), broom (Robinson, Hayes, and Couch (2024)), dplyr (Wickham et al. (2023)), knitr (Xie (2023)), and modelsummary (Arel-Bundock (2022)). Lastly, arrow (Richardson, McKinney, et al. (2024)) was used to store data efficiently.

2.1.1 Scope and Relevance of the Dataset

The dataset used in this analysis is sourced from the FiveThirtyEight poll aggregation for the 2024 U.S. presidential election (FiveThirtyEight 2024). It focuses on polling data collected at both the national and state levels, with an emphasis on the two main presidential candidates: Kamala Harris and Donald Trump. The dataset's relevance is tied to its ability to capture voter sentiment during a highly dynamic election period, specifically from July 2024 onwards. The aim of leveraging this dataset is to develop a robust model that forecasts the election outcome by integrating trends from both the popular vote and electoral college perspectives.

By concentrating on data from the official campaign period, we ensure that the dataset reflects the most current and meaningful voter preferences.

2.1.2 Characteristics of the Dataset

The dataset contains a variety of features that provide a nuanced understanding of voter sentiment across different demographics and regions. Key variables include candidate name, pollster, sample size, poll end date, state (or national), and percentage support for each candidate. With approximately 3,667 observations from multiple polling organizations, the dataset is both diverse and representative, allowing for an analysis that incorporates various polling methodologies. This diversity enhances the robustness of the forecasting model, as discussed in Section 4, where the modeling approach is detailed. Including data from multiple pollsters also allows us to account for biases that may arise from differing methodologies, thereby making the model results more reliable.

2.1.3 Data Collection Methodology

2.1.3.1 Polling Organizations and Methodologies

The polling data is aggregated from numerous organizations, each employing distinct methodologies to gather voter preferences. These pollsters are assigned numeric grades by FiveThirtyEight based on their reliability, transparency, and historical performance. By selecting only high-quality polling data, we aim to enhance the predictive power of our election model. The inclusion of a variety of polling methodologies is crucial, as it allows us to capture the nuances in voter sentiment across different regions and demographic groups, as elaborated in Section 5.2.

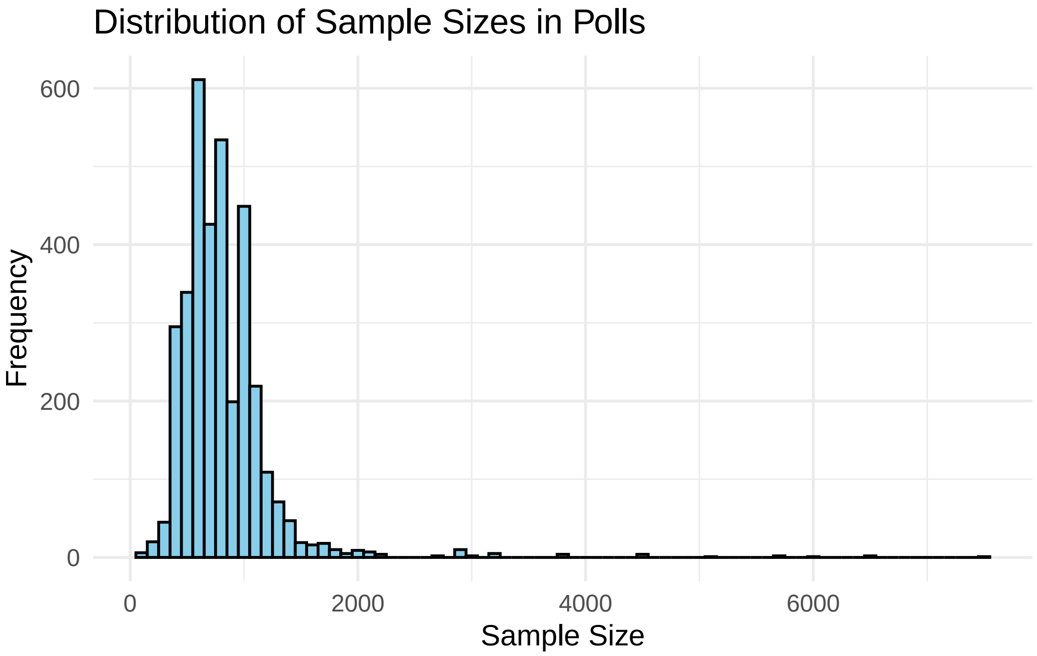
2.1.3.2 Ensuring Data Quality

To ensure data quality, a filtering process was applied to remove polls conducted by organizations with low numeric grades. Specifically, only pollsters with a numeric grade of 1 or higher were included. This threshold helps to minimize biases and inaccuracies in the data, thereby enhancing the robustness of the subsequent analysis. Additionally, pollsters with fewer than five polls were excluded to avoid unreliable estimates, ensuring that only consistent and representative polling data are used.

2.1.4 Data Cleaning Process

2.1.4.1 Initial Data Cleaning and Transformations

The raw data was subjected to multiple cleaning steps to prepare it for analysis. First, duplicate entries were removed, and column names were standardized for consistency. Key variables, such as candidate name, pollster, sample size, and poll end date, were retained, while irrelevant or redundant information was discarded. Additionally, the state variable was adjusted to aggregate specific districts—such as Nebraska CD-2 and Maine CD-1—into their respective states to facilitate streamlined geographic analysis. This transformation is essential for creating a consistent dataset that aligns with the electoral college framework discussed in Section 6.



This histogram (Figure 2.1) illustrates the distribution of sample sizes across all polls, highlighting the variability in poll sizes. Larger sample sizes are generally associated with more reliable polls, which is critical for making robust predictions.

2.1.4.2 Addressing Missing Data

Handling missing data was a critical step to ensure dataset completeness. Missing values were primarily found in the state column, which were replaced with “National” to denote national-level polls. Rows with missing information in key variables (e.g., candidate name or sample size) were removed. This approach ensured that the dataset was both complete and reliable for further analysis, as incomplete data could introduce biases into the model.

2.1.5 Key Variables and Derived Features

2.1.5.1 Description of Key Variables

The key variables retained for analysis include: - **Candidate Name**: Indicates the candidate (either Kamala Harris or Donald Trump) for whom polling data is reported. - **Pollster**: The organization conducting the poll, which is used to assess the quality and reliability of the poll. - **Sample Size**: Represents the number of respondents surveyed in each poll, which impacts the poll's margin of error. - **End Date**: The date when polling ended, used to calculate recency and weigh polls accordingly. - **State**: Specifies whether the poll is conducted at the state level or is national in scope. - **Percentage Support (pct)**: The percentage of respondents supporting each candidate, which serves as the primary measure of voter preference.

Figure 1: Candidate Support by State

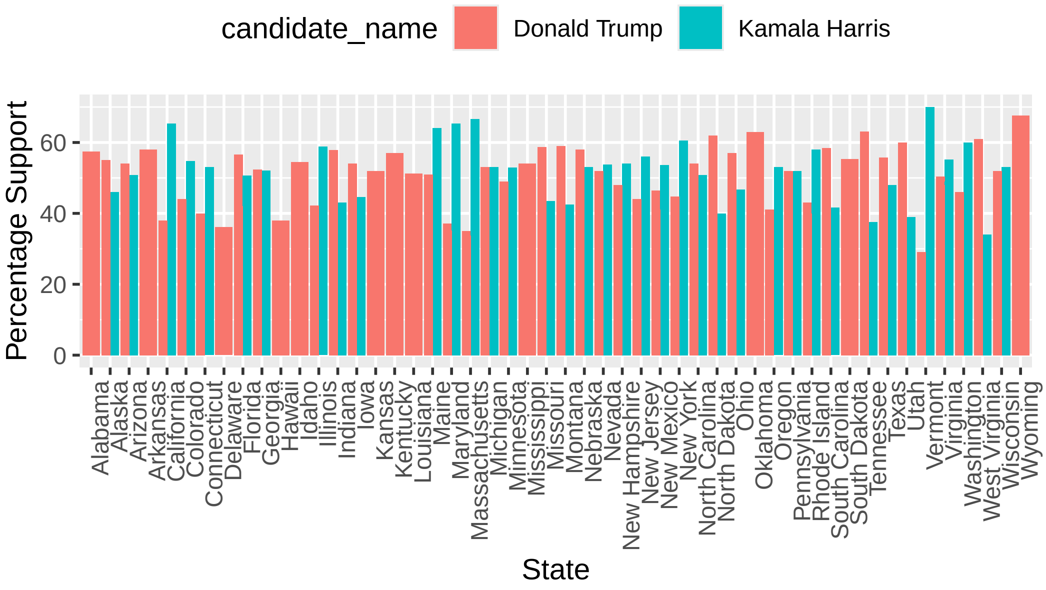


Figure 2: Distribution of Support for Each Candidate

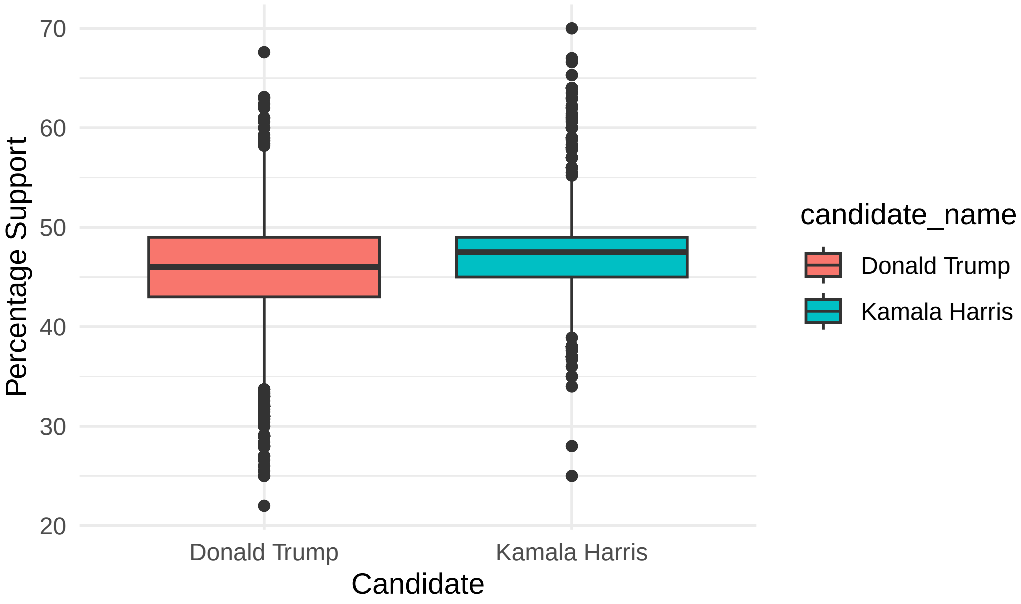


Figure 3: Sample Size vs Percentage Support

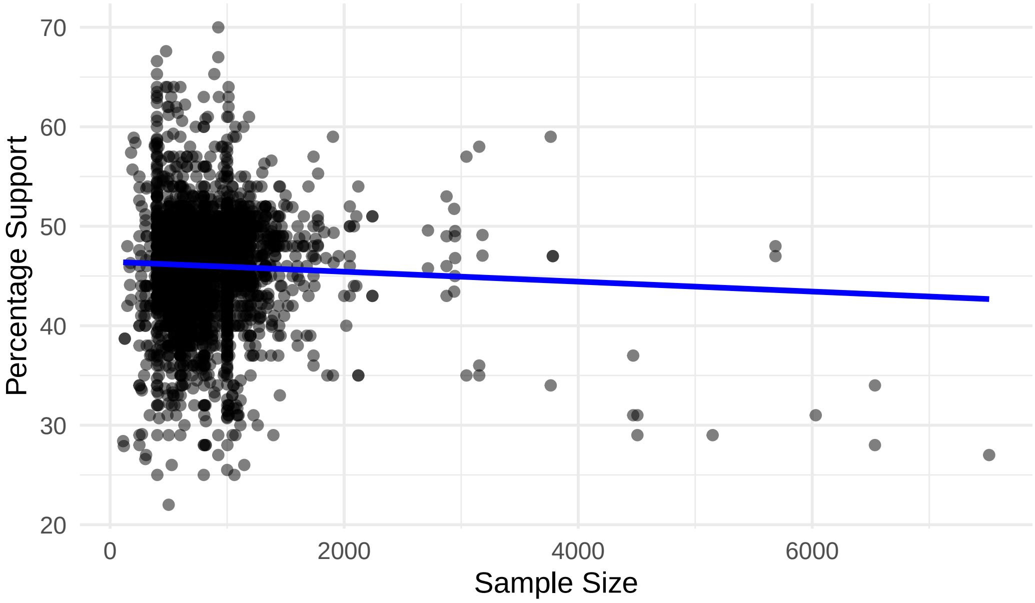


Figure 4: Polling End Date vs Percentage Support

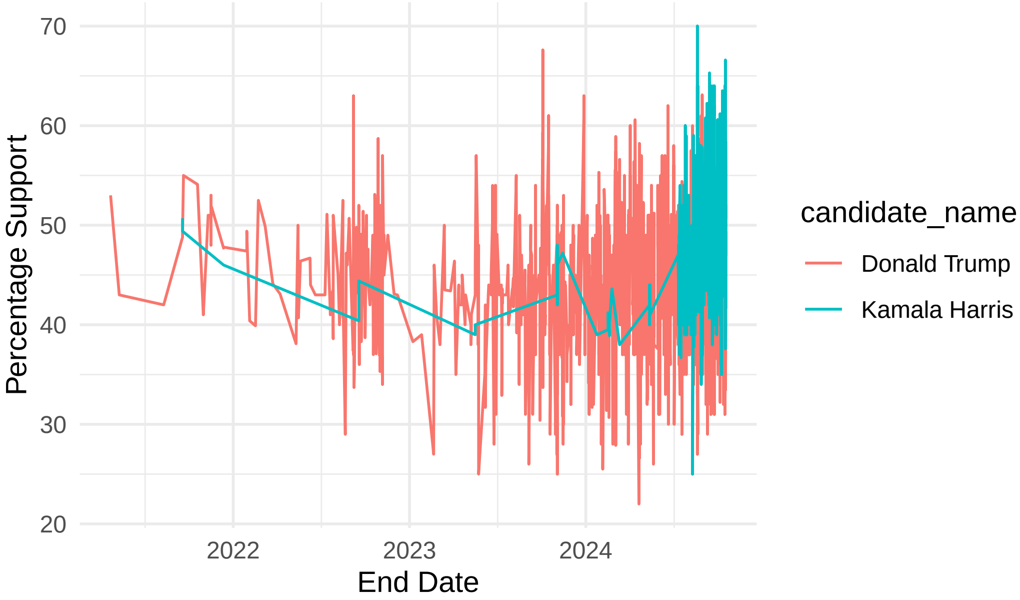


Figure 5: Distribution of Percentage Support

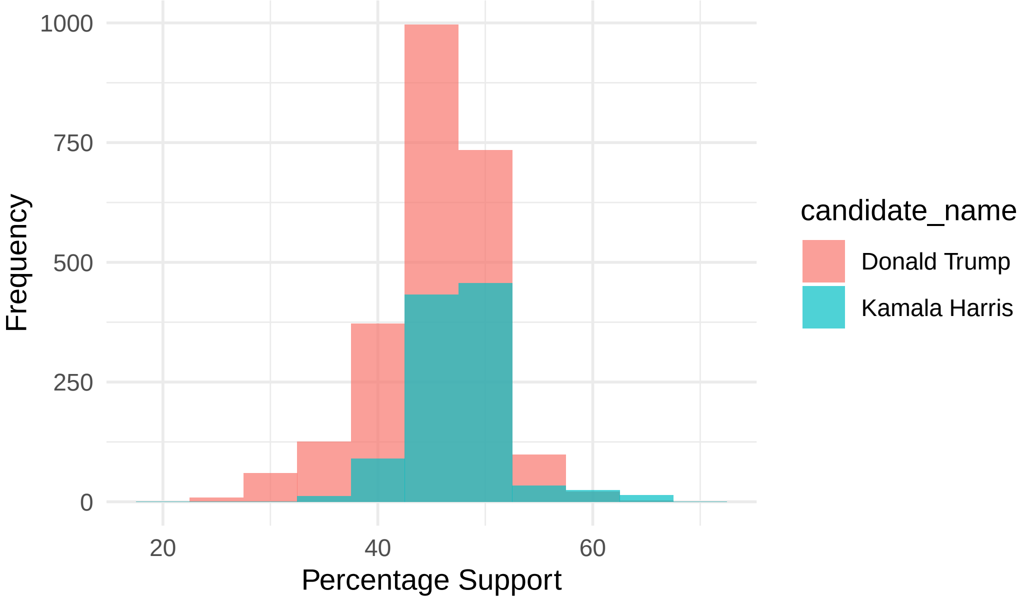


Figure 1 displays the percentage support for each candidate across different states. This visualization helps identify regional strengths and weaknesses for both candidates, providing a more granular view of voter sentiment.

Figure 2 displays a boxplot of distribution of support for Donald Trump and Kamala Harris, which is similar, with both candidates having median support around 50% and a comparable range, though there are outliers indicating some extreme opinions. Therefore, we can conclude that the 2024 U.S. presidential election will be quite close.

Figure 3 display a scatter plot that shows the relationship between sample size and percentage support, with a trend line indicating a slight negative slope. As sample size increases, the percentage support tends to decrease slightly, suggesting that larger sample sizes may slightly reduce support percentages, though the overall effect is minimal.

Figure 4 displays a line plot that shows the percentage support trends for Donald Trump and Kamala Harris over time. Trump's support fluctuates more widely with a slight downward trend, while Harris's support shows an upward trend, particularly toward the end of the time frame, when Joe Biden dropped the race. This suggests increasing support for Harris over time, as compared to a more variable and slightly declining trend for Trump.

Figure 5 displays a histogram that shows the distribution of percentage support for Donald Trump and Kamala Harris. Trump's support tends to cluster around the 45-50% range, while Harris's support is more spread but peaks around the 40-45% range. Both candidates have a similar central range of support, but from our simulated data set, Trump has a higher frequency of support values in the upper part of the range.

2.1.5.2 Variable Transformations and Feature Engineering

Several transformations were applied to prepare the data for modeling. Dates were standardized using the `lubridate` package to ensure uniformity across observations. Geographic transformations were also performed by aggregating specific state districts to their respective states, which simplified the geographic analysis and aligned with the electoral college modeling approach. Additionally, a new feature called **Recency** was engineered, representing the number of days since the most recent poll. This feature allows the model to weigh more recent polls more heavily, as explained in Section 4.3.

2.1.6 Filtering Criteria and Data Selection

2.1.6.1 Timeframe Filtering

To maintain the dataset's relevance to the current election cycle, polling data collected before July 21, 2024 (the official declaration of Kamala Harris's candidacy), was excluded. This filtering step was crucial for ensuring that the dataset reflects voter sentiment during the official campaign period. By focusing on this timeframe, the analysis becomes more representative of the dynamics that are most likely to influence the election outcome.

2.1.6.2 Pollster Reliability Filtering

As previously mentioned in Section 2.3.2, only pollsters with a numeric grade of 1 or higher were included. This criterion was set to enhance the reliability of the dataset and to focus on data sources that have historically shown accuracy in predicting election outcomes. Furthermore, pollsters with fewer than five conducted polls were excluded to avoid potential biases from underrepresented data sources.

2.1.7 Recency and Its Importance in Poll Weighting

2.1.7.1 Calculating Recency

Recency was calculated as the number of days between the reference date (the most recent poll in the dataset) and the poll end date. This calculation was used to create a recency weight for each poll, with more recent polls receiving higher weights. This weighting mechanism is essential for accurately forecasting voter preferences, as it allows the model to place greater emphasis on current data.

2.1.7.2 Application of Recency Weights

The recency weights were applied during the modeling process to account for the evolving nature of voter sentiment. By incorporating recency, the model becomes more adaptive to recent changes in voter preferences, which is particularly important given the rapidly changing political landscape. The use of recency weighting is further discussed in Section 4.4, where we elaborate on the modeling approach and its implications for forecast accuracy.

2.1.8 Summary and Visualization of Cleaned Data

2.1.8.1 Cleaned Dataset Overview

The final cleaned dataset contains approximately 3,000 observations, each representing a unique poll. The dataset includes details such as sample sizes, pollster information, candidate support percentages, and geographic coverage. This refined dataset serves as a robust foundation for the subsequent modeling and analysis steps, as described in Section 4.

2.1.8.2 Data Visualization

To gain insights into the distribution of key variables, visualizations such as histograms and bar charts were created. For instance, a histogram of **Sample Size** helps illustrate the variability in poll sizes, while bar charts displaying **Percentage Support** for each candidate across different states provide a clearer understanding of regional dynamics. These visualizations are included in the exploratory data analysis (Section 3) to help identify underlying trends and patterns in the dataset. The visualizations were created using `ggplot2` (Wickham 2016), ensuring that they are both informative and visually appealing, similar to the standards set by

Overview text

2.2 Measurement

In our study predicting the 2024 U.S. Presidential Election outcome, we focused on understanding what influences the percentage of support each candidate gets in the polls. The main thing we measured was the candidate support percentage—that's just the percentage of people in each poll who said they support either Kamala Harris or Donald Trump.

To figure out what might affect this support percentage, we looked at several factors. We considered who conducted the poll (the pollster) because different polling organizations might have different methods that can influence results. We also looked at the sample size, since polls with more people might be more accurate. Where the poll was conducted (which state) was important, too, because voter preferences can vary by location. We noted the date the poll ended, thinking that more recent polls might better reflect current opinions. Lastly, we identified which candidate the support percentage was for.

Polling data have been used for many years to predict which candidate will win the presidential race. The results are utilized by policymakers and corporations to refine and continue their strategies.

2.3 Outcome variables

In this analysis, the outcome variable of interest is the percentage of support a candidate receives in each poll, represented by pct. This variable indicates the proportion of respondents within a poll who favor a particular candidate, measured as a percentage. Given that pct serves as a continuous variable bounded between 0 and 100, it is well-suited for modeling the levels of support across different polls and over time.

2.4 Predictor variables

To accurately predict the 2024 Presidential Election, we have identified several key variables that capture a range of factors influencing voter behavior. These predictor variables include:

- Polling Percentage (PCT): Represents the percentage of respondents in a poll who support a specific candidate
- Pollster: Represents the name of the polling organization that conducted the poll.
- Candidate: The name of the candidate in the poll

- Sample Size (sample_size): size of the respondents participating in the poll
- State (state): the state of the poll was taken in
- End_date: the day the poll ended

3 Model

This analysis employs predictive models to estimate the anticipated support for the two major candidates in the 2024 U.S. Presidential Election: a multiple linear regression model. These models use high-quality polling data from reputable sources, with relevant predictors to provide insights into public opinion trends.

3.1 Prediction Model

3.1.1 Monte Carlo Simulation

We used a Monte Carlo simulation to create a distribution of potential support levels for Kamala Harris. This simulation can explain the uncertainty in polling estimates. For each poll, we generated a simulated support level, denoted as $\hat{y}i, j$, by adding a random error term $\epsilon i, j$ to the model's predicted support $\hat{y}i$. The error term, $\epsilon i, j$, is the sampling variability which is drawn from a normal distribution $N(0, \sigma^2)$:

$$\hat{y}_{i,j} = \hat{y}_i + \epsilon_{i,j}$$

In each simulation j, we calculated the average support level for Harris, $\bar{y}_j$, by aggregating the simulated support values across all polls:

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{i,j}$$

Repeating the process 1,000 times will produce a distribution of $\bar{y}_j$ values that captures the range of plausible support levels. This distribution enables us to estimate the likelihood of various support thresholds for Harris, taking into account the inherent uncertainty in polling data.

3.2 Model set-up

The multiple linear regression model is designed to predict the candidate's percentage of support (pct) as a linear function of various predictors. Formally, we define the model as:

$$\text{pct}_i = \beta_0 + \beta_1 \cdot \text{sample\_size}_i + \sum_{k=1}^{K} \delta_k \cdot \text{state}_{ik} + \beta_2 \cdot \text{end\_date}_i + \epsilon_i,$$

Where

- pct_i is the the predicted percentage of support for Kamala Harris in each poll
- β_0 is the intercept or expected value of PCT when all predictor variables are zero
- β_1 is the coefficient for sample size, showing the effect of the number of respondents in each poll on support
- δ_k is the coefficients representing the effect of each state
- β_2 is the coefficient for end_date, shoing how support changes over time
- ϵ_i is the error term, accounting for unexplained variability in support

3.2.1 Model justification

Predictors Selection: Variables such as sample_size, pollster, and state are included to capture systematic differences in polling methods and geographic variations in support. The date (end_date) of each poll controls for temporal effects on candidate popularity.

Linear regression was chosen for this analysis due to several key advantages. First, its simplicity and interpret ability make it highly suitable for examining the relationship between predictor variables and the response variable, specifically the candidate support percentage. Each coefficient in a linear regression model clearly represents the influence of a predictor on the outcome, which facilitates straightforward interpretation and communication of results. Moreover, linear regression is computationally efficient, making it ideal for rapid model estimation, tuning, and testing, even on relatively large datasets.

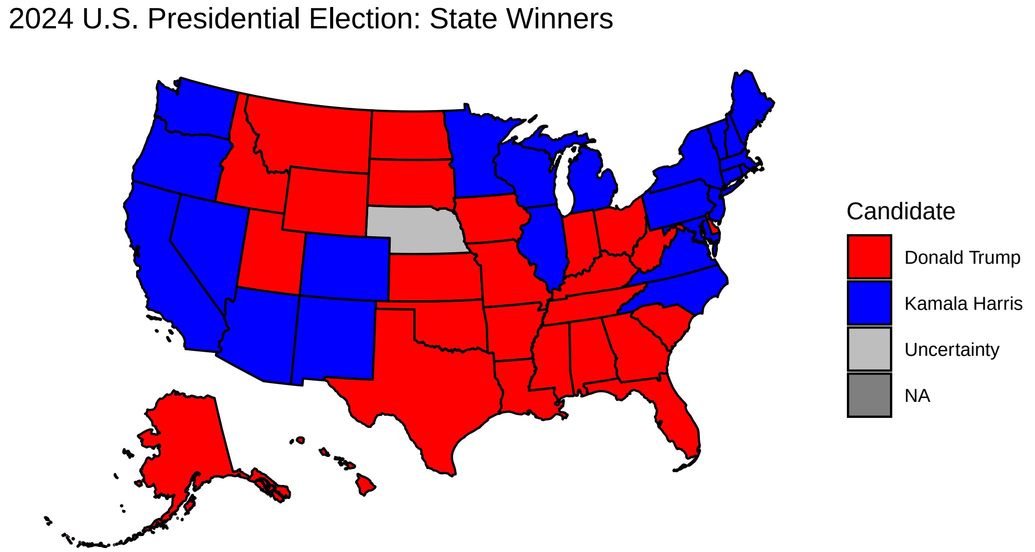
3.3 Regression Assumptions

- Linearity: The relationship between each predictor and the response (pct) is assumed to be linear.
- Normality: Errors are assumed to follow a normal distribution, which is tested during model diagnostics.

3.3.1 Checking Assumptions

In the Residuals vs. Fitted, there is no clear curved pattern, which suggests that the linear relationship might be appropriate. However, there are some clusters and a slight spread, which may indicate minor issues in capturing the relationship fully. Secondly, in the qqnorm plot, the points closely follow the diagonal line, indicating that residuals are approximately normal. This supports the normality assumption of the residuals, which is essential for reliable inference in linear regression.

3.4 Results



This map shows the predicted winners by state in the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump based on polling data.

- Red States: These states are predicted to lean toward Donald Trump. His support is concentrated in parts of the South, Midwest, and Mountain West regions, where he has historically performed well and appears to maintain a strong base of support.
- Blue States: These states are predicted to lean toward Kamala Harris. Her support is more concentrated in coastal regions like the West Coast and the Northeast, where more progressive policies and diverse populations tend to favor Democratic candidates.
- Gray States (Uncertainty): These states have uncertain outcomes, meaning that the polls do not show a clear leader. These could be considered potential swing states, where voter preferences may shift or are closely contested. These states will likely be highly targeted by both campaigns due to their potential to swing the overall election result.

4 Discussion

4.1 Purpose

In this study, we conducted a comprehensive analysis to forecast the outcome of the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump. Utilizing national and state-level polling data aggregated from FiveThirtyEight (FiveThirtyEight 2024), we aimed to capture the most recent shifts in voter sentiment since July 2024. Our methodology involved developing a multiple linear regression model that incorporated critical variables such as candidate support percentages, pollster reliability (as graded by FiveThirtyEight), sample size, poll end dates, and geographic distribution across states.

To enhance the predictive accuracy of our model, we applied recency weighting, emphasizing more recent polls to reflect current voter preferences more accurately. Recognizing that voter sentiment can change rapidly, especially in the months leading up to an election, this approach allowed us to give greater importance to the most current data. Additionally, we implemented Monte Carlo simulations to estimate the distribution of possible Electoral College outcomes. By accounting for state-level variations and uncertainties inherent in polling data, these simulations provided a probabilistic forecast rather than a deterministic one, offering a more nuanced understanding of potential election scenarios.

Our analysis not only focused on national trends but also delved into state trends, acknowledging the critical role of the Electoral College in determining the election outcome.

4.2 Limitation

Despite the robustness of our methodology, several limitations affect the generalizability and accuracy of our findings. A primary concern is the reliance on polling data, which is inherently subject to various biases and uncertainties. Polling errors can arise from sampling biases, non-response biases, and inaccuracies in weighting methodologies. For example, certain demographic groups may be underrepresented in polls due to lower participation rates, leading to skewed results. The phenomenon of "herding," where pollsters adjust their results to align with others to avoid being outliers (Silver 2024), can further distort the true range of voter sentiment

Since our projections suggest that Trump would win, it's important to address the role of standard deviations in our estimates and consider the possibility that Harris's support may be underestimated, similar to how Trump's support was in 2020 (Times (2024)). The idea of Harris being underestimated, much like Trump was in 2020, hinges on the possibility that polling errors or biases could be systematically missing certain groups of voters who support her. In 2020, Trump's support was underestimated in some key swing states due to factors such as non response bias, where Trump-leaning voters were less likely to participate in polls, and demographic weighting errors.

Appendix

A Pollster Analysis

The Marquette Law School Poll utilizes a rigorous probability-based methodology to conduct a national question survey of adults across the United States. This method combines address-based sampling (ABS), which uses the U.S. Postal Service's Computerized Delivery Sequence (CDS), and stratified random sampling, which covers virtually all residential addresses and is able to reach the offline population, thereby reducing the bias associated with surveys conducted solely via the Internet (David Rodbart 2024). The stratified random sampling method divides the population into subgroups (e.g., age, gender, education), to make sure that each subgroup is proportionately represented in the sample. This structure improves representativeness and accuracy, thereby supporting valid inferences about the total U.S. adult population.

A.1 What is the population, frame, and sample?

The population surveyed by the Marquette Law School Poll includes all U.S. adults aged 18 and older, living in the 50 states and D.C. The sampling frame was developed using Address-Based Sampling (ABS) by using the U.S. Postal Service's Computerized Delivery Sequence (CDS) file—a regularly updated, near-complete listing of all residential addresses in the U.S., excluding business addresses. This frame can cover almost every U.S. household, including non-internet and traditionally hard-to-reach areas (David Rodbart 2024).

The sample itself was drawn from the SSRS Opinion Panel, a probability-based panel that combines participants from ABS and a bilingual telephone survey conducted through the SSRS Omnibus. This dual recruitment strategy, which includes participants from both internet and non-internet households, improves demographic inclusivity. To achieve accurate representation, the survey sampled 1,005 adults, including registered and likely voters, and applied post-stratification weighting to adjust for demographic imbalances. This weighting aligns the sample with key population parameters, ensuring that the poll's findings are as representative as possible of the broader adult population.

A.2 How is the sample recruited?

Marquette Law School Poll using the SSRS Opinion Panel for the recruitment, a probability-based panel integrating Address-Based Sampling (ABS) and telephone survey data from the SSRS Omnibus, a nationally representative bilingual survey platform. This dual recruitment approach ensures that both internet and non-internet users are included, addressing the online-only bias often seen in web-based surveys and thereby enhancing sample inclusivity (David Rodbart 2024).

The Marquette Law School Poll uses stratified sampling to obtain a balanced and representative sample by differentiating participants based on key demographic variables such as age, gender, race, education, region, and political affiliation. The poll then ensures proportional representation of key demographic groups, which reduces the risk of sampling bias and increases the generalizability of the results.

Selected participants received a unique survey link through email invitations. Non-responders were sent up to two reminder emails, and, for those who opted in, a text notification reminder. To encourage completion and maintain high response rates, participants were offered an e-gift card as an incentive. This use of reminders and incentives is critical in minimizing non-response bias, thereby strengthening the reliability of the poll's findings.

A.3 What sampling approach is taken, and what are some of the trade-offs of this?

The Marquette Law School Poll uses both sampling probability sampling methods, address-based sampling (ABS) and stratified random sampling, this can increase representativeness and inclusiveness. ABS uses the U.S. Postal Service's Computerized Delivery Sequence (CDS), which covers nearly all residential addresses in the U.S., including homes with no Internet access. This approach is critical to minimizing the coverage bias associated with web-only surveys (David Rodbart 2024).

The poll also applies stratified random sampling to ensure that key demographic groups—such as age, gender, race, education, region, and political affiliation—are proportionally represented within the sample. Samples are taken from each subgroup once the population is divided into subgroups using stratified sampling. Consequently, sampling error is reduced and the sample findings accurately represent the views of each subgroup.

While this approach increases the reliability and representativeness of the findings, it also demands significant resources. ABS requires comprehensive address data and ongoing maintenance, while stratification and post-stratification weighting add complexity to both survey design and data processing. These resource-intensive processes make ABS and stratified random sampling more costly than simpler sampling methods but are justified by the improved accuracy and generalizability of the poll's results.

A.4 How is non-response handled?

The Marquette Law School Poll applies weighting adjustments across demographic factors, including age, gender, race, education, region, and internet access, aligning the sample's characteristics with those of the broader adult population (David Rodbart 2024). This can represent the under-represented groups more accurately and increase the weight of these respondents. This process enhances the survey's representativeness, ensuring findings reflect the national population reliably despite disparities in response rates.

A.5 What is good and bad about the questionnaire?

The Marquette Law School Poll is designed to boost respondent engagement and minimize survey fatigue. Using skip patterns that adapt questions based on previous answers, the survey reduces cognitive load, streamlines flow and lowers dropout rates by customizing the survey experience (David Rodbart 2024). This probability-based approach, supplemented by comprehensive weighting and stratification, further makes sure that the results accurately represent the U.S. adult population, with clear and straightforward questions that enhance accessibility across demographic groups.

However, despite ABS and telephone recruitment efforts, the reliance on web-based responses can cause some internet bias possibly underrepresenting households with no internet access. Additionally, the survey's length and complexity could add to respondent fatigue, potentially impacting response quality as participants progress. The response rate of approximately 47.4% reflects moderate attrition, typical in public opinion research, highlighting the necessity of post-stratification adjustments to address any demographic imbalances from non-response.

B Idealized Methodology

In designing a forecast for the upcoming US presidential election with a $100,000 budget, our methodology focuses on creating a representative survey that captures the voting intentions of likely voters across the United States. The objective is to gather reliable, high-quality data that supports a well-informed forecast through effective sampling, data validation, and aggregation techniques.

Our target population is voters aged 18 and above, with a sample size of approximately 10,000 responses, 200 responses from each state, to ensure statistical significance at the national level and within key demographic groups. To achieve balanced representation, we will employ stratified random sampling based on age, gender, race/ethnicity, education, region, and urban/rural status. This approach will involve weighting responses using recent census data and voter turnout statistics to address any demographic discrepancies.

We will combine online survey panels, social media outreach, and targeted telephone recruitment. These methods are chosen to increase accessibility for harder-to-reach demographics, such as older adults and rural residents. To encourage participation, we will offer modest incentives, such as a small monetary compensation or entry into a prize draw. All data collection will be conducted through a reliable survey platform—Google Forms.

The survey itself will be concise and straightforward, focusing on voting intentions, key issues, and demographic information. For data validation, the survey will include attention-check questions as respondents are very likely to complete the form for an entry into a prize draw. We will also conduct multiple waves of the survey, we will adopt a "poll-of-polls" approach, aggregating responses from different waves while applying weights based on relevance.

To mitigate biases such as non-response bias, social desirability bias, and order bias, our team will implement several strategies. We will send reminders to non-respondents to encourage participation, emphasizing the prize draw incentives to increase response rates. Survey questions will be carefully crafted to maintain neutral wording and avoid sensitive topics that might discourage honest responses. Additionally, within each wave of polls, we will randomize the order of questions to minimize the impact of question order on responses. These measures aim to enhance the reliability and accuracy of our survey results.

To ensure data validity, IP addresses would be recorded to prevent multiple submissions from the same source. Additionally, cookies would be implemented to prevent users from retaking the survey on the same device. Finally, we will verify respondents' email addresses to ensure uniqueness.

Ethically, our approach prioritizes respondent privacy and data security. We will ensure that participants' responses remain anonymous and that all data is handled securely. Transparency is also essential; respondents will be informed about the survey's sponsorship, methodology, and data use.

The survey will be implemented on Google Forms, with a link provided in the appendix. We also included the questions in the appendix for better understandings.

B.1 Budget Allocation

- Sampling and Data Collection (Advertisements and Recruitment Platforms): $40,000
- Participant Incentives (Prize Pool): $20,000
- Survey Design and Testing: $20,000
- Data Analysis and Validation: $20,000

C Idealized Survey

The proposed survey is designed using Google Forms, https://forms.gle/Vhmm1b1XnqTSoBWq8

C.1 Survey Copy

2024 U.S. Presidential Election Voter Intention Survey

Welcome to our survey! We are conducting research to understand public perceptions and opinions related to the upcoming 2024 Presidential Election. Your responses are valuable in helping us gain insight into voter perspectives and preferences.
By participating, you will gain the chance to enter a prize draw. Prizes include a chance to win gift cards, exclusive merchandise, and special deals with our sponsors.

The survey should take about 10-15 minutes to complete.

Please note:

- Your Privacy are our top concern. All responses are confidential and will be used solely for research purposes.
- Participation is entirely voluntary.

Contact Information

If you have any inquiries or need assistance with this survey, please contact us at:

- jzc.yang@mail.utoronto.ca
- 123-456-7891

* Indicates required question

1. Email *

2. Are you eligible to vote in the upcoming 2024 U.S. Presidential Election? *

Mark only one oval.

○ Yes

○ No

○ Other: ____________________

3. What is your age?

Mark only one oval.

○ 18-24

○ 25-34

○ 35-44

○ 45-54

○ 55-64

○ 65 or older

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

Check all that apply.

□ White
□ Black or African American
□ Hispanic or Latino
□ Asian
□ Native American or Alaska Native
□ Native Hawaiian or Other Pacific Islander
□ Other: ______________________________

5. What is your gender?

Mark only one oval.

○ Female

○ Male

○ Non-binary/Other

○ Prefer not to say

6. In which U.S. state do you currently reside?

Mark only one oval.

○ Alabama

○ Alaska

○ Arizona

○ Arkansas

○ California

○ Colorado

○ Connecticut

○ Delaware

○ Florida

○ Georgia

○ Hawaii

○ Idaho

○ Illinois

○ Indiana

○ Iowa

○ Kansas

○ Kentucky

○ Louisiana

○ Maine

○ Maryland

○ Massachusetts

○ Michigan

○ Minnesota

○ Mississippi

○ Missouri

○ Montana

○ Nebraska

○ Nevada

○ New Hampshire

○ New Jersey

○ New Mexico

○ New York

○ North Carolina

○ North Dakota

○ Ohio

○ Oklahoma

○ Oregon

○ Pennsylvania

○ Rhode Island

○ South Carolina

○ South Dakota

○ Tennessee

○ Texas

○ Utah

○ Vermont

○ Virginia

○ Washington

○ West Virginia

○ Wisconsin

○ Wyoming

7. What is your highest level of education?

Mark only one oval.

○ High school graduate, equivalent, or less

○ College Degree

○ Bachelor's Degree

○ Associate Degree

○ Graduate Degree

8. Who do you intend to vote for in the 2024 U.S. Presidential Election?

Mark only one oval.

○ Kamala Harris (Democratic Party)

○ Donald Trump (Republican Party)

○ Undecided

○ Other: ______________________________

9. How certain are you about your choice?

Mark only one oval.

○ Very certain

○ Somewhat certain

○ Unsure

10. Which of the following issues are most important in deciding your vote? (Select up to 3)

Check all that apply.

□ Economy and jobs
□ Healthcare
□ Climate change
□ Education
□ Immigration
□ National security
□ Social justice and equality
□ Abortion rights
□ Gun control
□ Other: ______________________________

11. If you voted in 2020, who did you vote for?

Mark only one oval.

○ Joe Biden

○ Donald Trump

○ Other: ______________________________

12. What are your primary sources for political news?

Mark only one oval.

○ Television

○ Newspapers/Magazines

○ Online News Websites

○ Social Media

○ Radio

○ Friends and Family

○ Other: ______________________________

13. Would you be willing to participate in follow-up surveys regarding the election?

Mark only one oval.

○ Yes

○ No

14. Do you have any additional comments or concerns about the 2024 election?

Thank You for Your Participation!

We sincerely appreciate you taking the time to complete our survey. Your valuable input is essential in helping us understand voter perspectives for the upcoming U.S. presidential election. Thank you for contributing to our research!

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

David Rodbart, Julia Dalagan. 2024. *Marquette Law School Poll Methodology Statement*. 1215 W Michigan St, Milwaukee, WI 53233, United States: Marquette University Law School. https://law.marquette.edu/poll/wp-content/uploads/2024/10/MLSPSC22Methodology.pdf.

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

FiveThirtyEight. 2024. *FiveThirtyEight 2024 u.s. Presidential Election Poll Aggregation*. New York, USA: FiveThirtyEight. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

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/.

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

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, Wes McKinney, et al. 2024. *arrow: Integration to Access and Manipulate Data in 'Apache Arrow' Format*. https://cran.r-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Silver, Nate. 2024. "Are the Polls Wrong Again? Why Experts Are Worried about 'Herding'." *New York Magazine: Intelligencer*. https://nymag.com/intelligencer/article/trump-harris-polls-herding-error-nate-silver.html.

Times, Financial. 2024. "What the Polls Can't Tell Us about America's Election." *Financial Times*. https://www.ft.com/content/870dabd1-93d2-4f88-9d10-b576ec5713c9.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 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, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.