Forecasting the 2024 US Presidential Election*

Analyzing Demographic Patterns and Predicting Swing States

Tina Kim David Flores Kevin Shao

November 4, 2024

This paper analyzes recent swing state polling data to forecast outcomes for the 2024 U.S. Presidential Election. Our findings indicate a close race, with Republicans leading in key states such as Georgia, Nevada, North Carolina, New Hampshire, Wisconsin, and Pennsylvania, while Democrats hold a slight edge in Michigan and Arizona. The national polling average also leans towards Republican, suggesting a Republican win in the 2024 Election. These insights give us a clearer picture of voter preferences across states and highlight the importance of targeted campaign strategies.

1 Introduction

The 2024 United States Presidential Election will determine the next president of the United States of America, with the outcome likely to affect both domestic and international policy and economic trends. Former President Donald Trump, the 45th president of the United States, is representing the Republican Party. In contrast, Vice President Kamala Harris is now the Democratic Party's candidate in the 2024 U.S. Presidential Election, following President Joe Biden's decision to withdraw from the race (Jr (2024)). Democracy can be described as liberal and left-leaning, with the emphasis on economy intervention with implementation of policies such as minimum wage and progressive tax rates. Members of the Republican party tend to follow the free economy, and can be described as right-leaning ("Democrat Vs. Republican" (n.d.)).

The outcome of the 2024 Presidential Election will determine whether the Republican or Democratic Party will lead the United States for the next four years. Given the United States' significant global influence, accurately predicting the results of this election carries substantial implications for shaping policy direction, ensuring economic stability, and influencing international relations. Reliable predictions enable parties to refine their campaigns and mobilize

*Code and data are available at: https://github.com/DavidFJ207/USPresidentialForecast

voters, influencing turnout and perceptions of competitiveness. Additionally, these predictions provide stakeholders with valuable insights that help adjust strategies and facilitate discussions on pressing societal issues. Ultimately, effective election forecasting is essential for understanding governance and its impact on society.

This paper uses recent polling data from key swing states to provide a prediction of the 2024 election outcome. We aim to determine which candidate—Kamala Harris or Donald Trump—is more likely to win this election. Our response variable is in turn the probability of one candidate winning the election based on our chosen Emerson poll data. When one candidate's support hovers just below 50%, it becomes challenging to predict their election as the 47th President of the United States, especially considering the potential influence of non-participating voters. To increase the accuracy of our predictions, we will develop two models: one estimating the probability of Harris, representing the Democratic Party, winning the election, and the other estimating the probability of Trump, representing the Republican Party, winning. By comparing these predictions, we aim to arrive at a clearer conclusion about who is more likely to be elected as the next President of the United States.

In our models, we identified eight pivotal swing states—Michigan, Georgia, Nevada, New Hampshire, North Carolina, Pennsylvania, Wisconsin, and Arizona—crucial for predicting the 2024 election outcome. We constructed Bayesian generalized linear models to estimate the percentage likelihood of candidates Trump and Harris winning in these states, incorporating education level, gender distribution, ethnic demographics, historical voting patterns, and current leanings toward Trump or Harris. Our election forecast resulted in 6 out of 8 pivotal states, as well as the national polling average, leaning towards the Republican party.

The remainder of this paper is structured as follows: Section 2 details the data and measurement process; Section 3 covers the model; Section 4 provides results; and Section 5 discusses implications and future steps.

2 Data

2.1 Overview

We sourced the "Presidential General Election Polls" dataset from FiveThirtyEight (FiveThirtyEight 2024) and performed an in-depth analysis using the statistical programming language R (R Core Team 2023).

Our goal was to clean, organize, and analyze U.S. presidential election polling data to highlight voter preferences by state. We used `knitr` (Xie (2023)) for report generation, ensuring our analysis was clear and well-documented, and `kableExtra` (Zhu (2023)) to enhance table outputs for easy interpretation. `ggcorrplot` (Kassambara (2023)) helped us visualize correlations, revealing relationships between variables, while `zoo` (Zeileis and Grothendieck (2023)) efficiently handled time series data, such as shifts in voter preferences. The `here` (Müller

(2020)) package ensured consistent file paths, making our workflow reproducible. This streamlined approach provided a comprehensive, polished analysis of key trends and demographic distributions. Table 1 and Table 2 provide a snapshot of these findings, showcasing the results of our organized and thorough data analysis.

Table 1: Voter Party Preferences by State

State	Democrat Pct	Republican Pct	Voted Biden 2020	Voted Trump 2020
Arizona	48.2	50.5	46.5	46.2
Georgia	49.8	49.7	46.5	46.2
Michigan	49.3	49.8	44.9	44.5
Nevada	48.6	48.3	48.2	47.1
New Hampshire	50.9	47.2	44.9	44.5
North Carolina	48.2	50.2	44.9	44.5
Pennsylvania	49.0	50.7	44.9	44.5
Wisconsin	49.2	49.9	44.9	44.5

Table 2: Demographic Breakdowns by State

State	High Educated	Low Educated	Female	Male	Nonbinary	Caucasian	Minority Ethnicity
Arizona	23.8	24.1	52.2	47.6	0.2	71.2	28.8
Georgia	23.8	24.1	52.2	47.6	0.2	71.2	28.8
Michigan	23.8	24.1	51.9	47.2	0.9	71.2	28.8
Nevada	22.1	24.3	51.9	47.4	0.7	80.2	19.8
New Hampshire	23.8	24.1	51.9	47.2	0.9	71.2	28.8
North Carolina	23.8	24.1	51.9	47.2	0.9	71.2	28.8
Pennsylvania	23.8	24.1	51.9	47.2	0.9	71.2	28.8
Wisconsin	23.8	24.1	51.9	47.2	0.9	71.2	28.8

2.2 Measurement

We chose Emerson as our primary pollster, and the reasoning for this is detailed in Section A. Emerson's polls offered crucial information, such as polling dates, party affiliations, sample sizes, and the percentage of support for each political party.

Our analysis centered on voter preferences by party rather than individual candidates. This approach was more relevant for identifying overall trends and helped us forecast which states might swing in future elections, especially swing states. We will discuss this focus further in our results section, Section 4.

The Emerson dataset also included valuable demographic data, voting history, and approval ratings. To ensure a consistent analysis, we concentrated on questions that were uniformly asked across all states. This approach allowed us to aggregate and analyze data at the state level without being limited by state-specific polling variations.

2.3 Outcome Variables

Our initial analysis of the dataset revealed important trends in party preferences across states. In Figure 1, we show which states have the strongest leanings toward the Democratic or Republican parties. This visualization is crucial for identifying key swing states, which lie near the center of the graph and show nearly equal support for both parties. These states are especially important for predicting the 2024 election outcome.

From this analysis, we identified eight critical swing states: Michigan, Georgia, Nevada, New Hampshire, North Carolina, Pennsylvania, Wisconsin, and Arizona. These states will be central to our election prediction. Although Minnesota emerged as a potential swing state, we excluded it due to the wide range of polling data, as explained in Table 3. Additionally, we will consider the national average in our overall prediction.

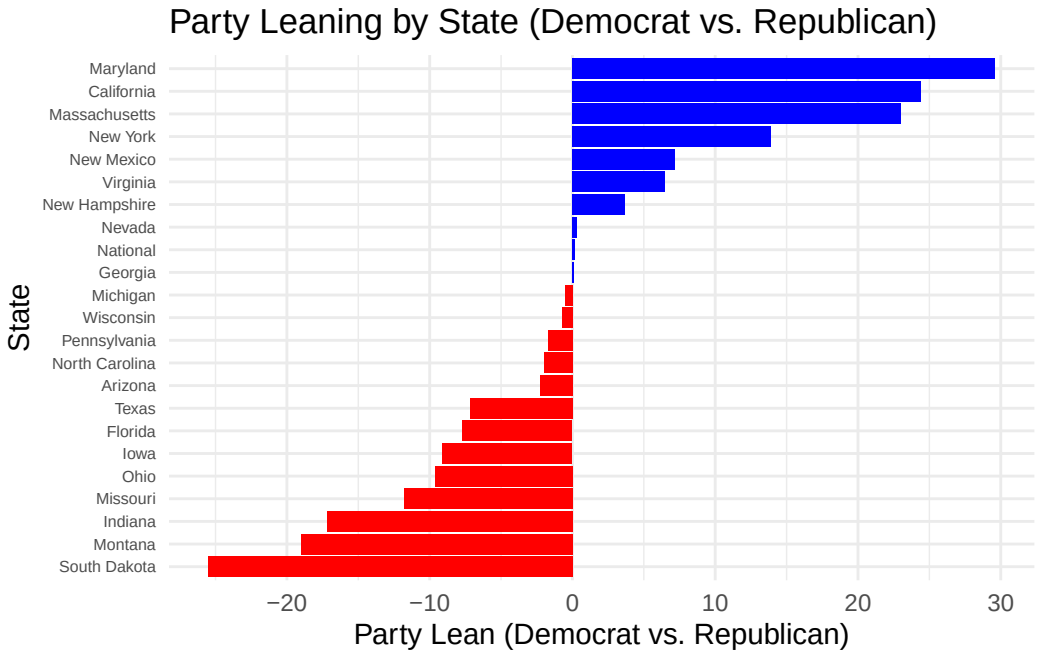

Figure 1: Example Predictor Variable Visualization

Table 3: Final Summary Table of 95% Confidence Intervals for Common States

State	DEM Mean (%)	DEM Lower 95%	DEM Upper 95%	REP Mean (%)	REP Lower 95%	REP Upper 95%
Georgia	48.1	44.6	51.5	49.2	46.7	51.7
Michigan	48.9	46.1	51.6	48.5	46.0	51.1
National	50.0	48.8	51.3	47.9	46.0	49.8

Table 3: Final Summary Table of 95% Confidence Intervals for Common States

State	DEM Mean (%)	DEM Lower 95%	DEM Upper 95%	REP Mean (%)	REP Lower 95%	REP Upper 95%
Nevada	49.0	48.4	49.6	48.5	47.9	49.2
New Hampshire	50.3	44.4	56.3	45.4	36.8	54.0
North Carolina	49.0	48.3	49.7	49.8	49.4	50.2
Pennsylvania	48.4	47.1	49.7	49.2	48.0	50.5
Wisconsin	48.7	46.9	50.5	49.0	46.9	51.2

Beyond identifying swing states, it's also important to understand the differences between the strongest Democratic ("blue") and Republican ("red") states. The bar graph in **?@fig-party-extremes** highlights these differences. It focuses on key factors that create the biggest contrasts between Democratic-leaning and Republican-leaning states, helping to show how these factors differ between the most extreme red and blue states.

To start, the graph features two bars, one for Democratic points and one for Republican points, reflecting the responses to questions asked by Emerson Pollster about voting intentions. These responses closely align with the percentages shown for each state, giving us a clear picture of voter preferences.

The **?@fig-party-extremes** visualization gives us valuable insight into how people's poll responses may influence their political leanings.

Through these visualizations, we not only identify which states are most likely to swing in future elections but also better understand the public opinions driving these partisan divides.

2.4 Predictor Variables

We then take a closer look at the demographic and social factors that shape voter preferences. The predictors we analyze include education levels, gender, and past voting behavior. These variables are important because they significantly impact political leanings and can influence how a state is likely to vote.

In Figure 3, we examine how these factors correlate with party lean across U.S. states. This analysis helps identify which factors have the greatest influence in each state, providing a clearer understanding of what drives voter behavior.

The correlation matrix in Figure 3 shows how different predictors relate to party preferences. We can see which factors, like education or gender, are more strongly tied to Democratic or Republican leanings. Blue shading indicates stronger Democratic support, while red shows

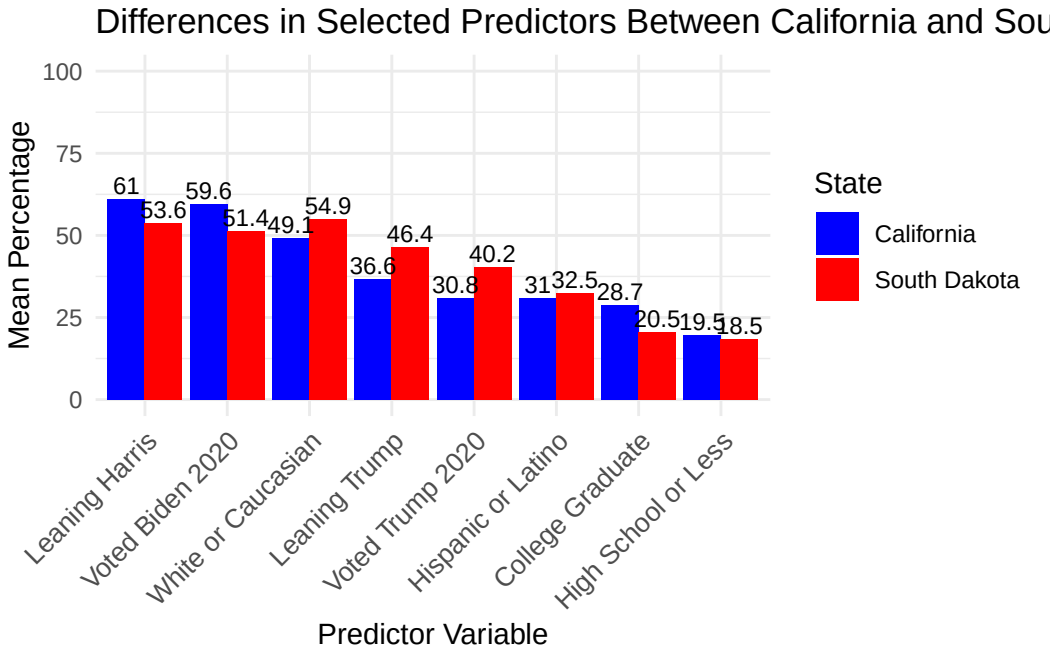

Figure 2: Differences in Selected Predictors Between California and South Dakota

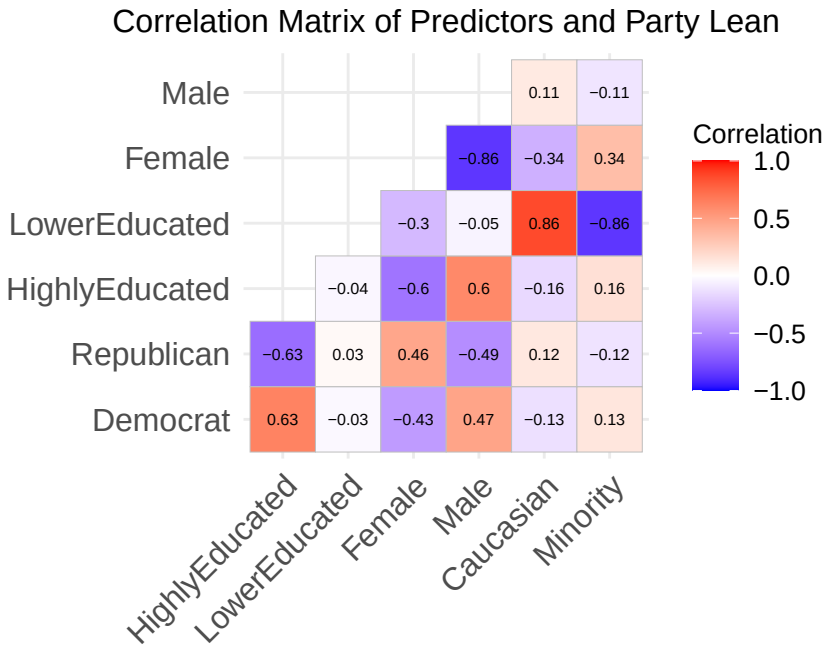

Figure 3: Example Predictor Variable Visualization

stronger Republican support. Key observations include higher education being linked to Democratic support and lower education favoring Republican. Minority groups generally lean Democratic, while Caucasian/White voters tend to support Trump. Additionally, women lean slightly more Democratic compared to men, who often lean Republican.

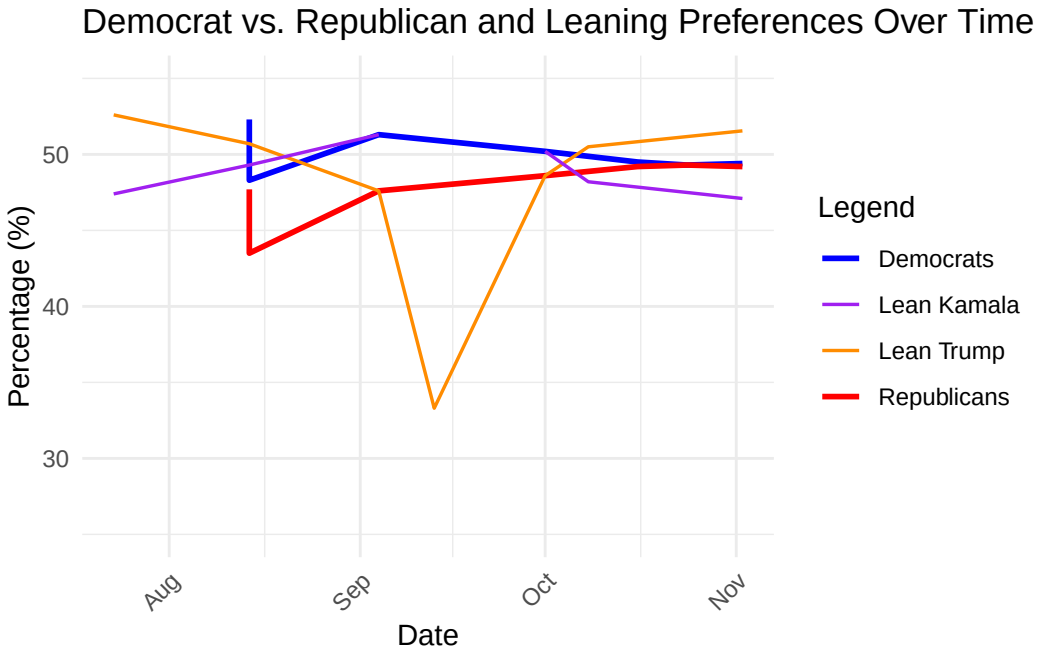

Figure 4: Parties Approval and Leaning Preferences

Lastly, we examine Democratic and Republican support, along with the population's leanings toward Kamala Harris or Donald Trump over time, starting from July 21st, when Joe Biden concluded his campaign, as reported by Rabson (2024). During this period, Democratic support has been decreasing steadily since September. Trump lost some support in September but gained it back, surpassing Kamala by early October. As the election approaches, both Democrats and Republicans have similar levels of overall support; however, the differences between support for Trump and Harris remain significant. By analyzing specific demographic factors and previous voting behavior, we aim to predict the 2024 election outcome.

Key predictor variables:

- **Education**: This measures education levels, with higher education often correlating with more Democratic support.
- **Race/Ethnicity**: This examines how different racial and ethnic groups tend to align politically, with minority groups often leaning Democratic and White/Caucasian voters leaning Republican.
- **Gender**: This captures how gender influences party preference, with women generally leaning more Democratic and men more Republican.

- **Priors (Past Voting Behavior)**: This looks at who people voted for in the past, helping us understand ongoing political support or disapproval.
- **Population Leaning Towards Harris**: This assesses how the population is feeling at the moment regarding who they may vote for, providing insight into shifting voter sentiments.
- **Population Leaning Towards Trump**: This assesses how the population is feeling at the moment regarding who they may vote for, providing insight into shifting voter sentiments.

3 Model

3.1 Forecasting Vote Outcomes in Key Swing States

Our model aims to forecast the voting outcome of the 2024 U.S. presidential election, focusing specifically on eight pivotal swing states. We carefully selected predictor variables based on a correlation matrix we generated previously (see Figure Figure 3).

3.2 Model Design and Approach

We constructed two separate Bayesian linear models to estimate the percentage likelihood of Trump and Harris winning in each swing state. Using `brms` (Bürkner (2023)) for Bayesian regression modeling, our models included demographic and historical data, allowing us to generate probabilistic predictions for both candidates. These models provide insights into which candidate is more likely to secure a win, even in closely contested states.

3.3 Key Components of Our Models

- **Outcome Variables**: The percentage of the popular vote each candidate is expected to receive.
- **Predictor Variables**: We include factors such as education level (e.g., college graduate vs. high school or less), gender distribution (male vs. female), and ethnic demographics (White or Caucasian vs. Hispanic or Latino). Additionally, historical voting patterns (e.g., 2020 voting preferences) and current leanings toward Trump or Harris are crucial predictors.
- **Model Structure**: Each model incorporates an intercept to represent baseline support and predictor coefficients to quantify the effect of each variable on the vote share. We use a Beta distribution to account for the bounded nature of percentage outcomes, ensuring our predictions remain between 0 and 100%.

3.4 Assumptions Underlying Our Models

1. **Linearity**: The relationship between each predictor and vote share is assumed to be approximately linear.
2. **Independence**: We assume that vote share predictions for each state are independent, conditional on the included predictors.
3. **Non-Normality**: Given the non-normal distribution of our response variables, using a generalized linear model is more appropriate. Our analysis revealed that the distribution of Democratic support resembles a Poisson-like distribution, while Republican support shows multimodal characteristics.
4. **Priors**: We applied weakly informative priors to regularize our predictions, with specific priors for those who voted for Trump and Biden in 2020, aligning with historical voting trends.

3.5 Bayesian Modeling Implementation

We implemented our models using the `brms` package in R, which allows for flexible Bayesian regression modeling. The models were fit with four chains, running 6,000 iterations each, to ensure convergence and reliable posterior estimates. Sensible default priors were adjusted to reflect our prior knowledge about voting behavior, and we conducted posterior predictive checks to assess the models' performance.

3.6 Results and Interpretation

The results, summarized in Table Table 4, highlight the predicted vote shares for each candidate in the swing states. For each state, we report the posterior mean, as well as 95% credible intervals, to capture the range of plausible outcomes. Our findings underscore the impact of demographic and historical factors on voting behavior, providing a nuanced understanding of which states are likely to be battlegrounds in the upcoming election.

By using a Bayesian approach, we not only make point estimates but also quantify the uncertainty surrounding these predictions. This is crucial for election forecasting, where even small shifts in voter behavior can have significant implications. Our analysis will inform campaign strategies and identify states where mobilizing specific demographics could be pivotal.

4 Results

Our results are summarized in Table 4.

Table 4: Prediction Summary for Swing States with Declared Winner

[!h]

State	Democrat Points	Republican Points	Winner
Michigan	0.49	0.48	Democrat
Georgia	0.49	0.50	Republican
Nevada	0.49	0.50	Republican
North Carolina	0.49	0.49	Republican
New Hampshire	0.49	0.49	Republican
Wisconsin	0.49	0.49	Republican
Pennsylvania	0.48	0.49	Republican
Arizona	0.49	0.48	Democrat
National	0.46	0.49	Republican

5 Discussion

5.1 Predictive Insights and Electoral Dynamics

Our model provides insightful predictions on the 2024 U.S. presidential election landscape, highlighting both the strengths and potential uncertainties within the swing states, and ultimately providing predictions of results of the 2024 United States of America Presidential Elections. By examining the demographic and historical predictors, we observed significant patterns in voter leanings and demographic shifts, particularly in critical swing states such as Georgia, Michigan, and Pennsylvania. The model suggests that minor shifts in demographics or campaign focus within these states could lead to substantial impacts on the election outcome. For example, Georgia, which exhibits an almost even split in party leanings, underscores the importance of targeted outreach to specific demographics, like college-educated voters and minority groups, which show slight inclinations towards the Democratic party.

5.2 The Role of Demographic Factors in Predicting Party Support

The analysis confirms that demographic variables such as education level, ethnicity, and gender hold substantial predictive value in determining party affiliation and voter preference. Our correlation matrix shows that higher education levels, for instance, are more closely associated with Democratic leanings, whereas lower education levels trend towards Republican support. Racial demographics also show a clear pattern, with minority groups leaning Democratic and White/Caucasian groups displaying stronger support for Republican candidates. Additionally, our results on gender differences show a modest trend, with women generally favoring Democratic candidates and men leaning Republican. This dynamic reinforces the importance of gender-specific appeals and messaging in campaign strategies. The model's findings imply that campaigns aiming to shift voter sentiment in swing states may benefit from tailored approaches that address the priorities and concerns specific to these demographic groups.

5.3 Weaknesses and Next Steps

While the model provides useful forecasts, there are inherent limitations and biases that may affect its predictions. One limitation is the assumption of linearity between predictor variables and vote share, which may not fully capture the complex dynamics of voter preferences in certain regions. Furthermore, the model's independence assumption may overlook interconnected factors among states or demographic groups, particularly as social and economic trends impact neighboring states in similar ways.

The use of historical voting data can also introduce potential biases, especially as the political landscape can shift unexpectedly due to national events or changing candidate appeal. The reliance on only Emerson polling data may also introduce biases from potential underrepresentation or overrepresentation of certain demographic groups. Future models should try to incorporate multiple polling sources and adjust for potential sampling biases across states.

Moving forward, enhancements could include non-linear models to capture more complex relationships between demographic variables and voter behavior. Additionally, incorporating data on emerging voter concerns, such as economic shifts and social issues, would allow the model to better adapt to the evolving political landscape. Expanding the dataset to include more recent and varied polling sources may help improve prediction accuracy.

Moreover, applying regularization techniques, such as LASSO or Ridge regression, might reduce overfitting and improve the generalizability of predictions across states. Incorporating voter sentiment analysis from social media or news coverage could provide real-time insights into voter sentiment, enabling the model to adapt to the ever-changing political dynamics. Lastly, building models that consider interactions among predictors, such as gender and education, could offer a more nuanced view of the factors influencing voter behavior, making predictions more responsive to shifts within pivotal swing states.

A Appendix: Pollster Methdology Overview - Weaknesses and Potential Improvements

Emerson College Polling initially developed from a classroom polling exercise and is now an "innovative, nationally-ranked" non-partisan polling center ("About Us" (n.d.)). For the 2024 USA Presidential Election pollster, the Emerson College Polling defined the population to be "likely voters" of 2024 United States Presidential Election. Emerson College Polling does indicate that the targeted population of "likely voters" is based on "2024 Likely Voter Modeling." However, the methodology for this modeling and the specific model is not indicated, so we can not determine the scope of the targeted population ("October 2024 National Poll: Harris 50" (2024)).

Emerson College Polling recruits people completing this survey for this national pollster by contacting mobile phones using "MMS-to-Online", "Online Opt-in Panel", and "IVR(Interactive Voice Response)" ("October 2024 National Poll: Harris 50" (2024)). In the MMS-to-Online approach, the target population were sent text messages with graphics that invite them to take a "screening questionnaire", and those who pass the questionnaire can move on to take the survey. The selected respondents were based on state voter files "provided by Aristotle." The Online Opt-in Panel approach involves respondents from the targeted population being invited to finish the survey by means of an online opt-in panel "provided by CINT" ("About Us" (n.d.)). Specifically, in the Online Panel approach, selected voters were pre-matched to "L2 voter file data provided by Rep Data" ("October 2024 National Poll: Harris 50" (2024)). Finally, the IVR approach involves making automatic telephone calls to selected "likely voters." Participants then answer the survey using their "touch-tone telephone". The email approach indicated in Emerson College Polling's official methodology ("About Us" (n.d.)) is not mentioned in the National Polling ("October 2024 National Poll: Harris 50" (2024)), so it can be assumed that this approach is not used in our chosen pollster data.

From the methodology of recruiting people indicated in the previous paragraph, we can see that while the targeted population of the national poll are likely voters, the sampling frame are likely voters that use phones, either mobile or landline. Here, note that the definition of likely voters is set by Emerson College Polling, and the datasets where each sample is chosen from is described in the three approaches by Emerson's methodologies. Emerson College Polling chooses 1000 samples from their targeted population of likely voters of the 2024 United States Presidential Election, and it is then "weighted by gender, education, race, age, party affiliation, and region based on 2024 likely voter modeling" for each national poll. As most people in the United States use phones by some means, the target population and sampling frame are highly similar, increasing the overall validity of the survey.

In the "MMS-to-Online" and "IVR(Interactive Voice Response)" approaches, random sampling is used, and in the "Online Opt-in Panel" approach, there is no indication of a specific sampling approach ("About Us" (n.d.)). We can conclude that the general approach used in this specific pollster is random sampling, since Emerson Polling first chooses a population frame, and then

each individual in this population frame has an equal chance to be chosen. Random sampling tends to reduce bias, simplify analysis, and is also easy to implement, but at the same time also requires much time and money. In consideration with the sample size of 1000 people compared to the targeted population of likely voters, which contains most of the population of the United States of America, there is a high possibility of existing selection bias occurring when the participants are not representative of the population.

There is also no specific indication of how non-response is treated, but the results section of national polls, such as the one from September 29 to October 1, 2024, implicates that non-response are eliminated from final recorded results of the survey. Therefore, we cannot ignore the possibility of non-response bias occurring when non-respondents differ significantly from respondents. For example, such non-response bias occurs when non-respondents are from California, while respondents are from other states except California.

The questionnaire set by Emerson College Polling for this survey, based on the national polls set from September 29 to October 1, 2024, compose mostly of questions with most common choices listed as choices, and also allowing participants to type in their own answer if it is not included in the list of most common choices. The questionnaire is generally written in an objective voice, and all questions only allow participants to make only one choice. All questions in the questionnaire are relatively-common, including ones about ethnicity, age range, region, and education which collects demographic data, and ones that are related to actual presidential-election predictions, including opinions about Joe Biden's performance, the two major parties (Democratic and Republican), whether s/he would vote, and voting inclinations. The only issue observed in the questionnaire is that it is too lengthy, containing more than 20 questions, which may decrease participants' motivations to complete the survey, even after starting it ("October 2024 National Poll: Harris 50" (2024)).

Therefore, after analyzing the advantages and disadvantages of this pollster about the 2024 United States Presidential Elections by Emerson College Polling, we can conclude that Emerson College Polling can improve their pollster by implementing the following approaches. Emerson can consider increasing the sample size to make their survey more representative of the targeted population, explicitly describing their definition of the targeted population "likely voters" and the model used to determine characteristics of the targeted population. Emerson can also consider changing from simple random to stratified sampling (dividing the population into independent 'subsets', then performing random sampling in each group), based on different states or other demographic features based on their targeted population model, making their sample more representative of the population, decrease the length of the questionnaire, and also explicitly explain their treatment of non-response, which will increase the validity of the study. Overall, this pollster is still highly valid, and generally reliable for predicting the 2024 United States Presidential Election (Alexander (2023)).

B Appendix: Idealized Survey Design for $100K Budget

Introduction

This appendix introduces an idealized methodology and survey to forecast the US presidential election with a budget of $100k. The goal is to use a sample survey that effectively captures public opinion and demographic characteristics.

Sampling Approach

In order to most accurately represent the sample population of US voters, we will be using a stratified random sampling approach. Stratified random sampling efficiently and fairly samples data from strata that are representative of a population compared to simple random sampling since it also reflects the population of interest's underrepresented subgroups, ultimately reducing potential bias.

The sample population will be taken from 10,000 eligible US voters represented in strata divided into different demographic factors such as age, gender, race, education level, income level, and political affiliation. This sample size can provide a statistically significant confidence interval for the strata.

Respondent Recruitment

We will be using 70% of our budget for respondent recruitment, and the remaining 30% for implementing the survey and validating results.

A proportion of our budget of $20,000 will be used for advertising on various social media platforms such as Instagram, Facebook, and Twitter to reach around 10,000 respondents. This online advertising is a good way to reach target respondents based on different demographics we are interested in, especially for younger or hard-to-reach populations.

Since digital ads may result in higher dropout rates and lower response quality, the majority of our budget will be allocated to make up for this limitation. A budget of $40,000 can be used to hire professional survey panels that work with companies that can provide large databases of target respondents that have been pre-screened to match our demographic of interest. Lastly, since pre-screened professional survey panels can introduce bias from being frequent survey responders, we will also use incentivized online survey platforms to recruit various participants that fit our demographics of interest. Participants will be offered small monetary incentives with our budget of $10,000, which can greatly increase completion rates and result in better response quality.

Data Validation

First, we need pre-screening questions to confirm the eligibility of respondents such as eligibility to vote to prevent bots and fake responses. For instance, asking whether they are a US citizen, and then asking if they are eligible to vote in the upcoming US presidential election. If their answers contradict these two questions, we know that their response lacks validity.

Additionally, we have attention checks to identify if respondents are paying attention to and understanding the survey questions. For example, we can ask them if they are paying attention to the question, and if they answer no, we know that they might be rushing or answering the survey carelessly.

Furthermore, we want consistency checks that include similar questions which are paraphrased in different parts of the survey to ensure that respondents are answering consistently.

Lastly, IP address monitoring can be used to prevent duplicate responses from the same person.

Poll Aggregation

We can collect poll data from various reputable polling sources that have good pollster reliability and are from more recent polls. This way, we can weigh the polls according to their quality. For instance, we would assign higher weights to pollsters that have a high rate of pollster reliability based on past results. We would also give higher weights to more recent polls that better showcase the current opinions of the upcoming US election. Lastly, we would emphasize the weight of larger sample sizes since these provide more accurate representation of the population of interest. We can also incorporate the Bayesian Inference method when updating new polls to adjust prior distributions with new information.

Survey Link: https://forms.gle/vm3gKbsMAZvzakFn8

Copy of Survey

A copy of the survey and its questions are provided below:

2024 US Presidential Election Survey

Thank you for considering participation in the 2024 US Presidential Election Survey. The purpose of this survey is to better understand and forecast election results. This survey gathers anonymous data on voter opinions and demographics for the 2024 US presidential election. You may stop the survey at any time without consequence. All responses will remain anonymous, and results will be reported in aggregate form only. A small monetary incentive will be offered upon completion.

If you have any questions or concerns about this survey, please reach out to any of the researchers below:

Tina Kim, University of Toronto: tinak.kim@mail.utoronto.ca David Flores, University of Toronto: davidgadiel.flores@mail.utoronto.ca Kevin Shao, University of Toronto: kevin.shao@mail.utoronto.ca

By proceeding, you confirm your consent to participate. - I consent to participate. - I do not consent to participate.

1. **Are you a US citizen?**

- Yes
- No

2. **Are you eligible to vote in the upcoming US presidential election?**

- Yes
- No

3. **What is your age group?** Under 18

- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 or older
- Prefer not to say

4. **What is your gender?**

- Male
- Female
- Non-binary
- Prefer not to say

5. **What is your ethnicity? (choose all that apply)**

- White / Caucasian
- Black or African American
- Hispanic or Latino
- Asian
- Native American or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Middle Eastern or North African

- Other
- Prefer not to say

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

- Less than high school
- High school diploma or equivalent (e.g., GED)
- Some college, no degree
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional degree (e.g., JD, MD)
- Doctorate (e.g., PhD, EdD)
- Prefer not to say

7. **Are you paying attention to this survey?**

- Yes
- No

8. **What is your income level?**

- Less than $20,000
- $20,000 - $39,999
- $40,000 - $59,999
- $60,000 - $79,999
- $80,000 - $99,999
- $100,000 - $149,999
- $150,000 - $199,999
- $200,000 or more
- Prefer not to say

9. **What is the highest level of education you have attained?**

- Less than high school
- High school diploma or equivalent (e.g., GED)
- Some college, no degree
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional degree (e.g., JD, MD)
- Doctorate (e.g., PhD, EdD)
- Prefer not to say

10. **What is your political affiliation?**

- Democrat
- Republican
- Independent
- Other
- Prefer not to say

11. **If the election were held today, which candidate would you vote for?**

- Candidate A
- Candidate B
- Candidate C
- Other
- Undecided
- Prefer not to say

C Appendix: Data Cleaning

The focus was cleaning up and organizing poll data from different states to create a clear and detailed timeline using data from Google Sheets and Excel files. To make this process easier and more efficient, we used several packages that helped us download, clean, and organize the information so it could be easily used later.

First, we used the **`googledrive`** package (Bryan and McGowan (2024)), which allowed our code to connect to Google Sheets and download the poll data. This made it possible for us to get the most up-to-date information. Once we had the data, we used another set of packages called **`tidyverse`** (Wickham (2024)), which includes **`dplyr`** (Wickham et al. (2023)) and **`readr`** (Wickham and Hester (2023a)), to clean up the information and organize it. For data coming from Excel files, we used the **`readxl`** package (Wickham and Bryan (2024)), which made it easy to read and work with the information from those files.

We also used **`lubridate`** (Grolemund and Wickham (2023)) to make sure dates were handled correctly, so the timeline made sense and everything was in order from oldest to newest.

The **`janitor`** package (Firke (2024)) was helpful for cleaning up column names. We also used **`testthat`** (Wickham and Hester (2023b)) to test our data, making sure everything was correct and worked as expected. Finally, we used a package called **`arrow`** (Richardson, Korn, and Dunnington (2024)) to save the data in a file format called Parquet, which makes the data smaller and faster to work with.

References

"About Us." n.d. Emerson College Polling. https://emersoncollegepolling.com/about/.

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman & Hall/CRC. https://tellingstorieswithdata.com/.

Bryan, Jennifer, and Lucy D'Agostino McGowan. 2024. *googledrive: An Interface to Google Drive*. https://CRAN.R-project.org/package=googledrive.

Bürkner, Paul-Christian. 2023. *brms: Bayesian Regression Models using 'Stan'*. https://CRAN.R-project.org/package=brms.

"Democrat Vs. Republican." n.d. Diffen. https://www.diffen.com/difference/Democrat_vs_Republican.

Firke, Sam. 2024. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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

Grolemund, Garrett, and Hadley Wickham. 2023. *lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Jr, Bernd Debusmann. 2024. "Biden Says He Quit US Presidential Race to 'Save Democracy'." BBC News. https://www.bbc.com/news/articles/c047281jj8do.

Kassambara, Alboukadel. 2023. *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*. https://CRAN.R-project.org/package=ggcorrplot.

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

"October 2024 National Poll: Harris 50." 2024. Emerson College Polling. https://emersoncollegepolling.com/october-2024-national-poll-harris-50-trump-48/.

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

Rabson, Mia. 2024. "Kamala Harris Would Be Democrats' 'Best Choice' If Biden Doesn't Run, Expert Says." https://globalnews.ca/news/10683018/kamala-harris-democratic-nomination-us-election-2/.

Richardson, Neal, Uwe Korn, and Dewey Dunnington. 2024. *arrow: Integration to the Apache Arrow Project*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2024. *tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Jennifer Bryan. 2024. *readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2023a. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

———. 2023b. *testthat: Unit Testing for R*. https://CRAN.R-project.org/package=testthat.

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

Zeileis, Achim, and Gabor Grothendieck. 2023. *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*. https://CRAN.R-project.org/package=zoo.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'kable()' + Pipe Syntax*. https:

//CRAN.R-project.org/package=kableExtra.