# Forecasting the 2024 U.S. Presidential Election*

## Bayesian Models Predict Kamala Harris's Path to Victory

Abdullah Motasim Elizabeth Luong

November 4, 2024

This paper presents a predictive analysis of the 2024 U.S. Presidential Election, focusing on estimating support for Kamala Harris and Donald Trump through aggregated polling data. Using Bayesian Generalized Linear Models (GLMs) and Monte Carlo simulations, we analyze polling trends across various states to assess each candidate's projected support and overall probability of winning. Results indicate a 63.2% probability of victory for Harris and a 36.8% probability for Trump, with notable regional variations. Harris shows strong support in traditional Democratic strongholds, while Trump leads in Republican-leaning areas, underscoring the importance of state-specific dynamics in shaping election outcomes. This study highlights the influence of pollster reliability, sample sizes, and regional factors, offering knowledge into the potential direction of the 2024 election.

## Table of contents

*Code and data are available at: https://github.com/abdullah-motasim/Forecasting-2024-US-Presidential-Election.

# 1 Introduction

On November 5 2024 the U.S. Presidential Elections will be held to determine the 47th President of the United States. Since the U.S. is a powerhouse within the world, often affecting many international affairs, the result of this election will have a global impact on all countries and as such there is much interest in which of the top two candidates Donald Trump or Kamala Harris will end up winning the election. As such, leading up to the election time there are many polls conducted by different pollsters to determine the winner of the election, but these polls often vary significantly between each other due to biases and sampling limitations. Poll aggregation is a method which aims to mitigate these differences by combining or aggregating the results of multiple polls to provide a more robust estimate of the election outcome.

This study seeks to forecast the 2024 U.S. Presidential election outcome by developing a generalized linear model based on poll aggregation data. Specifically, we evaluate predictors including demographics, polling trends, and historical voting patterns. While the popular vote does not directly determine the presidential outcome due to the electoral college system, our model's findings offer a clearer picture of voter preferences and potential trends in key states. This paper outlines our modeling approach, data processing techniques, and results, with a final section discussing model limitations and implications for future election forecasting.

The primary estimand in this study is the probability of a Democratic or Republican victory in the popular vote, based on current polling trends and demographic predictors.

Utilizing Bayesian Generalized Linear Models (GLM) and Monte Carlo simulations to estimate the predicted support percentages for Kamala Harris and Donald Trump. Harris has a higher support percentage over Trump with a win probability of 63.2%, while Trump has a 36.8% win probability. The model suggest that although there is close competition between both candidates, Harris has an advantage based on the polling data.

Running the simulations for this data showcases the importance of state-level effects, support for each candidate varies significantly by region. The models indicate that Harris has strong support in traditionally Democratic states, whereas Trump maintains a lead in Republican-leading regions. Support for each candidate indicates that regional factors are pivotal in determining the final outcome for American elections.

The remainder of this paper is structured as follows. Section 2 discusses the data types included in the raw data, the cleaning process for the data, and the reason for selecting the data set we did. Section 3 discusses model specification and justification for Bayesian GLMs and Monte Carlo simulation. Section 4 analyzes the trends and correlations between different variables utilizing tabular and graphical means. Section 5 discusses the results of Section 4 going into detail on what the simulation results can tell us about the upcoming election, as well as assumptions and limitations in data.

# 2 Data

## 2.1 Overview

The aggregated dataset containing the polling data for the 2024 U.S. election was sourced from FiveThrityEight (FiveThirtyEight 2024) and all analysis was performed using R R Core Team (2023) and the following packages: (R Core Team 2023) and the following packages: tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2021), janitor (Firke 2021), testthat (Wickham and Hester 2021), arrow (Team 2021), here (Müller 2020), rstanarm (Goodrich et al. 2022), bayesplot (Gabry and Bürkner 2021), gt (Iannone and RStudio 2021), kableExtra (Zhu 2021), knitr (Xie 2021), and webshot2 (Keitt and Chang 2021).

The FiveThirtyEight dataset was chosen over others as it is valued for its commitment to transparency, advanced modeling techniques, rigorous poll aggregation, and inclusion of broader socio-economic factors. The initial raw data contained about 17000 observations of 52 variables from multiple pollsters and after cleaning the data we were left with about 9000 observations of 7 variables. The data cleaning involved removing missing values and only selecting variables which were of importance to us, these variables were:

- **pollster_name** - The name of the polling organization that conducted the poll (e.g., YouGov, RMG Research).
- **candidate** - The name of the candidate in the poll (e.g., Kamala Harris, Donald Trump).
- **percentage** - The percentage of the vote or support that the candidate received in the poll (e.g., 51.0 for Kamala Harris, 48.0 for Donald Trump).
- **party** - The political party of the candidate in the poll (e.g., DEM for Democrats, REP for Republicans).
- **sample_size** - The total number of respondents participating in the poll (e.g., 2712).
- **state** - The U.S. state where the poll was conducted or focused, if applicable.
- **numeric_grade** - A numeric rating given to the pollster to indicate their quality or reliability (e.g., 3.0).

These variables were selected due to their direct relevance to assessing candidate support, polling reliability, and geographic voting trends. Specifically, pollster_name and numeric_grade are used to track potential biases and variations across polling organizations, as different methodologies can impact results. Candidate/Party and percentage provide direct indicators of support levels for each contender, which is essential for making accurate projections of election outcomes. Also, it is worth noting that there were more Sample_size is key to assessing the reliability of each poll, as larger samples tend to yield more accurate representations of public opinion. Lastly, state enables us to account for regional variations, key in a system like the U.S. where state-level results play a determining role in the election.

Collectively, these variables offer a focused view of the polling landscape, supporting a more reliable aggregation and interpretation of the data.

## 2.2 Measurement

Since different pollsters utilize various methods to collect their data and convert real-world phenomena into entries in their datasets, we have chosen to focus on the method utilized by Emerson, as it is the most frequent pollster within our cleaned data, accounting for 823 of the observations. Analyzing Emerson's methodology provides valuable knowledge on how polling results are generated, thereby enhancing our understanding of the dataset.

Emerson College Polling employs a combination of Interactive Voice Response (IVR) and online panel methodologies for data collection. The recent survey of Michigan sampled 1,000 likely voters. Emerson developed a series of survey questions to ask these potential candidates such as do you prove or disapprove of the job Joe Biden is doing as president? What is your party registration? etc. After this the results of the survey are analyzed to get the percentages of votes for each option and these are than presented within the aggregated data set.

Note this is a general overview of the measurement method utilized by Emerson, a more thorough breakdown of measurement can be found in Section A.

## 2.3 Outcome variable

The primary outcome variable in our analysis is the percentage of support each candidate receives in various polls, as reported by different polling organizations. This variable captures the proportion of respondents who indicate support for a specific candidate (e.g., Kamala Harris, Donald Trump) in a given poll. By aggregating the reported percentages across pollsters and accounting for factors such as sample size and pollster reliability, we can obtain a view of each candidate's current standing in the 2024 U.S. election. This outcome variable directly reflects voter preference trends and provides a baseline for modeling anticipated election results. The robustness of this measure is influenced by the accuracy of individual poll data and the methodological rigor of each pollster, such as Emerson's use of IVR and online panels, ensuring it represents a reliable indicator of the broader electorate's sentiments.

To illustrate the predicted support for the two primary candidates in our analysis, Kamala Harris and Donald Trump, we included a density plot Figure 1 that visualizes the simulated support percentages for each candidate. This plot provides a comparison of predicted support, with Harris shown in blue and Trump in red. The density plot highlights the distribution of predicted support percentages based on our model simulations, allowing us to observe both the central tendency and spread of each candidate's support.

In this visual, we can see that Harris and Trump exhibit overlapping distributions, though with slightly different peaks. This overlapping region reflects a competitive election landscape, where variations in poll results or sampling could impact the projected outcome. The density peaks suggest that Harris has a slightly higher central tendency in predicted support, yet the close proximity of both distributions emphasizes the uncertainty and competitiveness of the race. By examining the density of predicted percentages, this plot contributes to a clearer understanding of the variability in support for each candidate as forecasted by our model.

## 2.4 Predictor variables

To model the outcome variable effectively, we selected 3 predictor variables based on their relevance to candidate support and regional trends:

Pollster Name: Identifying the organization conducting each poll (e.g., YouGov, Emerson) allows us to assess potential biases and variation in methodologies that could impact polling accuracy. Given that certain pollsters (like Emerson) have more frequent entries in our cleaned dataset, this predictor provides valuable findingsinto consistency and reliability across poll sources.

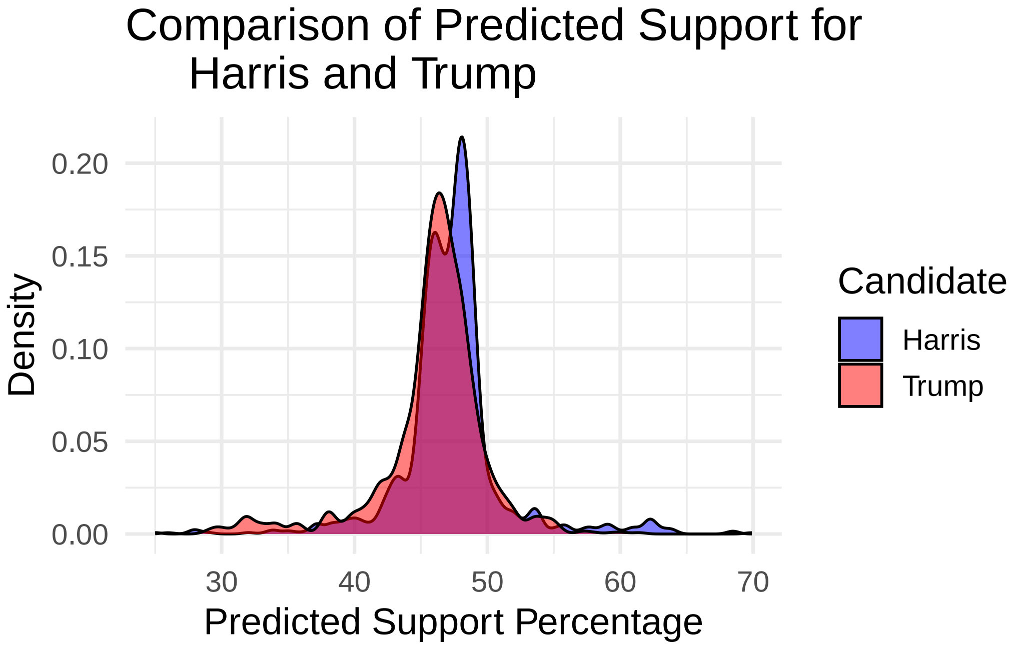

Figure 1: This density plot compares the predicted support percentages for Kamala Harris and Donald Trump based on model simulations."

Sample Size: Poll sample size directly influences the reliability of the reported percentages. Polls with larger sample sizes tend to yield more accurate estimates, reducing margin of error and providing a more representative snapshot of voter opinion. Incorporating this variable enhances the robustness of our predictions by weighting polls based on their sample adequacy.

State: Including the state where each poll was conducted allows us to account for regional voting patterns, which are essential in a federal system like the U.S.

Each predictor variable was selected to provide a balanced perspective on polling results, helping us account for various factors that influence voter behavior and polling reliability. These variables offer a structured approach to analyzing the 2024 U.S. election, supporting a model that is both statistically grounded and contextually aware.

Figure 2 illustrates the distribution of numeric reliability grades for the five most frequent pollsters in the dataset: Beacon/Shaw, Emerson, Morning Consult, Redfield & Wilton Strategies, and Siena/NYT. Each boxplot represents the variation in numeric reliability grades for individual polls from these pollsters, with jittered points indicating individual poll grades. This visualization helps highlight the consistency of reliability scores across these pollsters, with Emerson and Morning Consult exhibiting particularly stable distributions around higher reliability grades compared to others.

This reliability assessment is needed in weighting pollster data appropriately in our analysis. For example, Emerson, one of the most frequent pollsters in the dataset, shows a narrow spread in its reliability score, suggesting consistent methodology and accuracy. By understanding the reliability of each pollster, we can adjust our model to account for potential biases or methodological weaknesses that might otherwise skew the results. This reliability variable acts as a quality measure, allowing us to place greater confidence in polls from consistently high-rated pollsters.

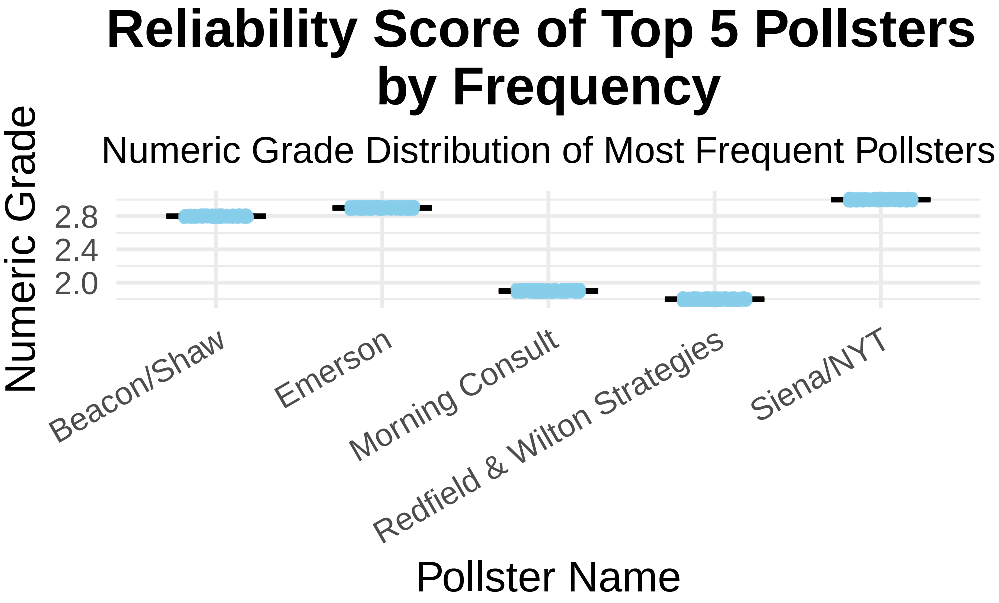

Figure 2: Each boxplot shows the distribution of numeric reliability grades for the top 5 pollsters, with individual poll grades represented by jittered points.

Figure 3 provides a breakdown of sample sizes across different U.S. states, with Pennsylvania, Michigan, and Wisconsin leading in terms of sample volume. These sample sizes are ket predictors, as polls with larger sample sizes generally yield more reliable data. This visualization underscores the emphasis placed on certain swing states and battleground regions, reflecting where polling efforts are most concentrated and, by extension, where voter sentiment may be considered more closely in analyses.

Incorporating sample size as a predictor allows the model to adjust the weight of each poll based on its reliability. For instance, a poll with a large sample size from a key state such as Pennsylvania is likely to be more indicative of statewide voter sentiment than a poll with a smaller sample size. This adjustment based on sample size helps to create a more representative model, giving us a stronger basis for forecasting outcomes in the election.

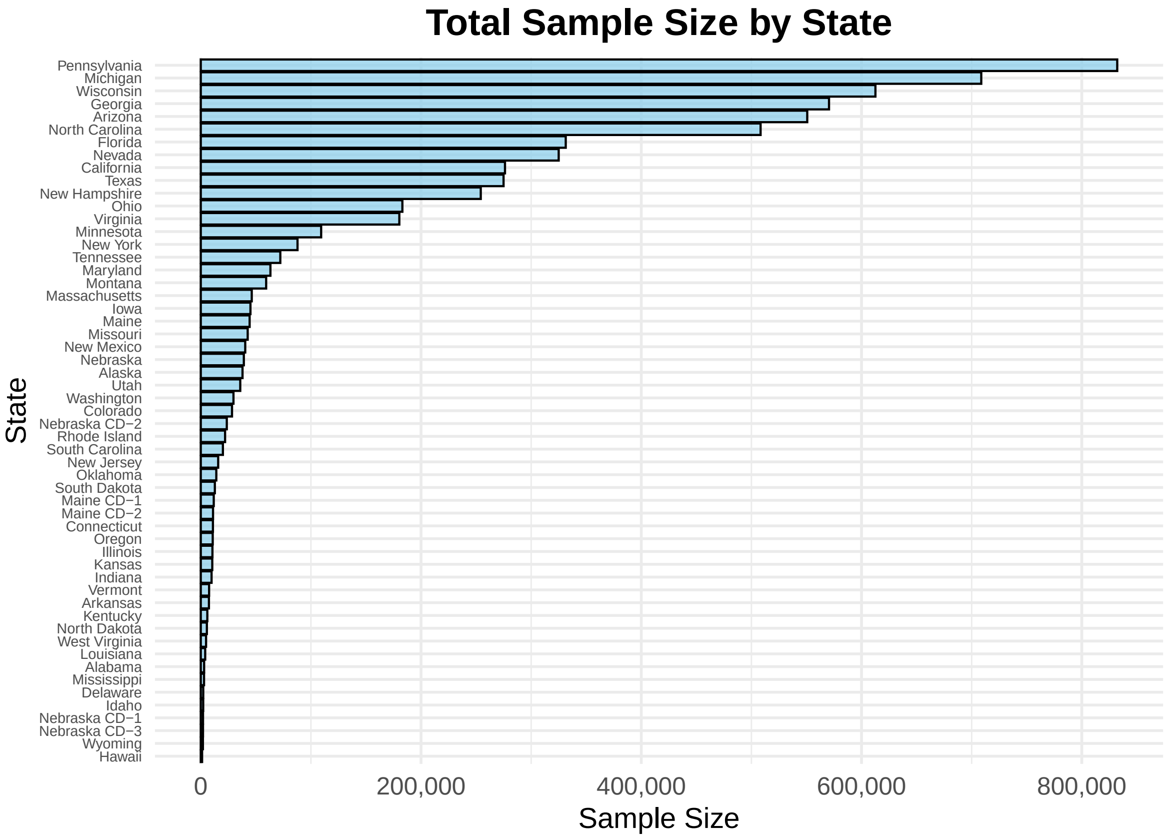

Fig 3: This chart displays the total sample size per state for survey data, ordered from highest to lowest sample size.

Figure 3: This chart displays the total sample size per state for survey data, ordered from highest to lowest sample size.

# 3 Model

This analysis uses Bayesian Generalized Linear Models (GLM) with Gaussian distributions to predict the support percentages for Kamala Harris and Donald Trump across multiple polls. The continuous outcome variable, representing the predicted support percentage, is modeled as a function of key predictors to capture poll-specific and demographic variations.

## 3.1 Model set-up

Given the continuous nature of support percentages, a Gaussian family with an **identity link function** was selected. This choice allows the model to accurately estimate continuous variations in support percentages, adjusting for effects from pollster, sample size, and state. The model can be expressed as follows:

$$\text{Support Percentage} = \beta_0 + \beta_1 \cdot \text{pollster} + \beta_2 \cdot \text{sample size} + \beta_3 \cdot \text{state}$$

Where:

- $\beta_0$ is the intercept term
- $\beta_1$ represents the effect of the pollster source.
- $\beta_2$ captures the influence of the sample size on support percentage.
- $\beta_3$ accounts for state-level variations.

## 3.2 Explaination of Variables and Inclusions

Each predictor in the model is chosen for its relevance in reflecting the polling data's characteristics:

1. **Pollster (`pollster`)**: Captures potential systematic differences among organizations, ensuring that support estimates are not bias by any single pollster's methodology.
2. **Sample Size (`sample_size`)**: Accounts for the reliability of each poll, as larger samples tend to reduce margin of error.
3. **State (`state`)**: Controls for geographic and political variations, allowing the model to adjust for differences in support based on regional preferences.

The summary statistics below Table 1 outline key variables in our analysis, focusing on Kamala Harris and Donald Trump. These tables display the distribution of support percentages and sample sizes, with metrics including mean, median, standard deviation, minimum, and maximum.

Table 1: Summary Statistics for Key Variables Comparing Kamala Harris and Donald Trump

| | Percentage Statistics | | | | | Sample Size Statistics | |
|---|---|---|---|---|---|---|---|
| Candidate | Mean | Median | SD | Min | Max | Sample Size Mean | Sample Size SD |
| Kamala Harris | 47.32 | 47.6 | 4.46 | 25 | 70 | 910.86 | 544.67 |
| Donald Trump | 45.56 | 46.0 | 5.37 | 21 | 70 | 879.21 | 524.25 |

## 3.3 Model justification

### 3.3.1 Bayesian Framework

The models utilize a Bayesian approach with normal priors on the coefficients, implemented through the `rstanarm` package of Goodrich et al. (2022) in R (R Core Team 2023). The Bayesian framework offers two primary advantages for our analysis:

1. **Uncertainty Representation**: Posterior distributions for each coefficient provide a clear indication of uncertainty around the estimates, allowing for more informative predictions.
2. **Regularization**: Normal priors help stabilize estimates, especially for smaller sample sizes or polls with limited data points, reducing overfitting.

### 3.3.2 Monte Carlo Simulation for Win Probability Estimation

To assess the probability of each candidate winning, we employed Monte Carlo simulations using posterior predictions from the GLMs:

1. **Generate Posterior Predictions**: Posterior distributions from each candidate's model produce a range of possible support percentages for Harris and Trump.
2. **Simulate Election Outcomes**: Each simulation iteration compares Harris' and Trump's predicted support. Harris is marked as the winner if her support surpasses Trump's, and vice versa.
3. **Estimate Win Probabilities**: By repeating this comparison over thousands of iterations, we estimate win probabilities for each candidate based on the frequency of simulated wins.

## 3.4 Model Assumptions and Limitations

The Bayesian Generalized Linear Models in this analysis have key assumptions and limitations that may impact predictive accuracy.

First, the assumption of a Gaussian distribution for support percentages may not fully capture the true nature of polling data, particularly if the data include skewness or outliers. While a Gaussian distribution efficiently models continuous support, it may lead to biased estimates in smaller states or in polls with high variance if non-normal patterns are present.

Second, the model assumes that each pollster's effect on support estimates is consistent across all states, disregarding potential variations in polling methodologies and accuracy by region. This simplification may overlook regional variations, especially if certain pollsters perform differently in specific states, potentially reducing accuracy in cases of state-specific polling practices.

Finally, the model treats state-level effects as static, assuming that state-specific support remains stable over time. This limits the model's ability to adapt to changing voter sentiment within states as the election nears. Significant state-level shifts that occur after the latest data collection may not be reflected in the predictions, leading to outdated estimates.

While these assumptions balance interpretability and reliability, they also indicate areas for future improvement, such as dynamic state effects and region-specific pollster adjustments, to further enhance predictive accuracy.

# 4 Results

The results of the Bayesian Generalized Linear Model (GLM) analysis, which aimed to predict support percentages for Kamala Harris and Donald Trump across various polls. The findings, represented through probability estimates, predicted support distributions, and comparative density plots, provide information about each candidate's likely performance in hypothetical head-to-head polling scenarios.

## 4.1 Predicted Support Percentages

### 4.1.1 Kamala Harris

The distribution of predicted support for Kamala Harris is displayed in Figure 4. The model estimates her average support at approximately 47.32%, with the majority of values clustering around this mean. This relatively tight distribution indicates consistent support for Harris across polls, with limited variability. The standard deviation of her predicted support suggests that her polling results remain stable, even in states with more diverse polling outcomes.

### 4.1.2 Donald Trump

Figure 5 illustrates the predicted support for Donald Trump, centered around a mean of 45.56%. Compared to Harris, Trump's predicted support distribution shows slightly greater variability, indicating a broader spread in his support across polls. The higher standard deviation in Trump's support highlights increased variability in polling outcomes, suggesting a wider range of voter responses depending on the region and poll characteristics.

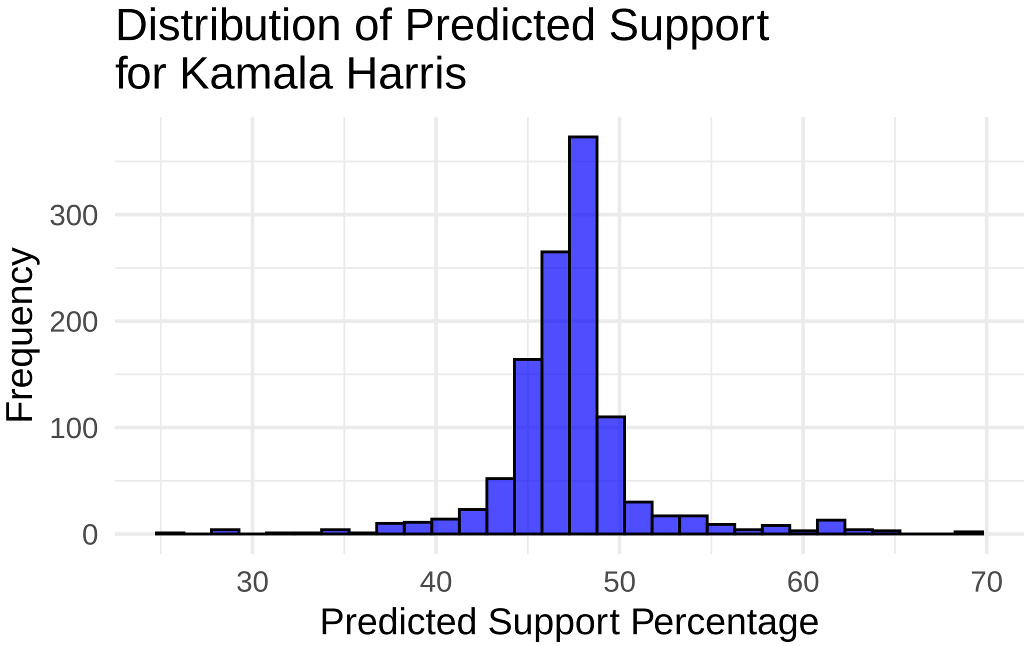

Figure 4: This histogram shows the distribution of predicted support percentages for Kamala Harris based on model simulations.

#### 4.1.3 Comparison of Predicted Support

The combined density plot was Figure 1 previously shown Section 2.3 provides a direct comparison of Harris and Trump's predicted support distributions. Harris's distribution (in blue) peaks at a higher level than Trump's (in red), indicating her slight edge in predicted support. Both distributions overlap significantly, particularly between 45% and 50%, underscoring the close nature of their polling outcomes. This overlap suggests that while Harris holds a predicted lead, the margins are slim, indicating a competitive scenario in many states.

### 4.2 Probability of Winning

Using Monte Carlo simulation, we estimated the likelihood of each candidate securing a win based on their predicted support. According to the model, Harris has a 63.2% probability of winning, while Trump has a 36.8% probability. These probabilities were calculated by comparing predicted support across thousands of iterations, with each iteration counting a win for the candidate who garnered higher support. This probabilistic approach captures the uncertainty inherent in election predictions and emphasizes Harris' modest advantage in a hypothetical head-to-head match-up.

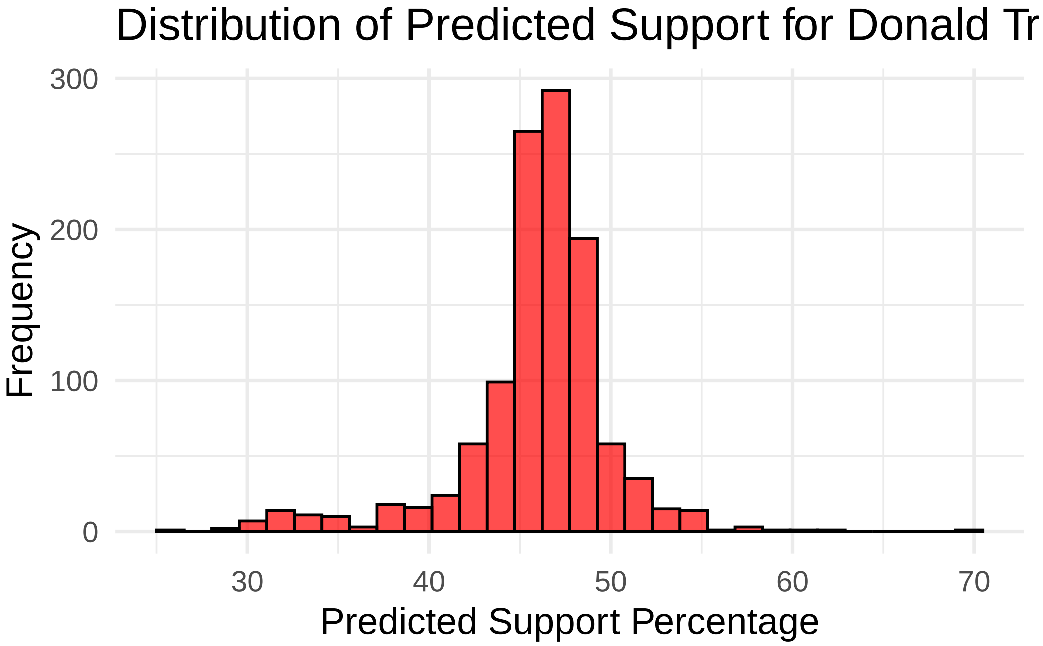

Figure 5: This histogram shows the distribution of predicted support percentages for Donald Trump based on model simulations.

## 4.3 State-Specific Trends

### 4.3.1 Regional Variation in Support

The model showcases notable differences in support by state, aligning with known regional preferences. For instance, Harris' support is consistently higher in traditionally Democratic states like California and New York, where her predicted vote share surpasses 60%. Conversely, Trump shows strong predicted support in states like Texas and Florida, with anticipated leads of 5-10 percentage points, reflecting historical Republican preferences.

### 4.3.2 Battleground States

In key states like Pennsylvania and Ohio, the predicted support for both candidates is nearly evenly split, signifying highly competitive races. The model suggests that Trump may hold a slight edge in Ohio, while Harris appears to lead narrowly in Pennsylvania. These projections highlight the importance of minor polling shifts in these states, where even small changes in support could significantly influence the final outcome.

### 4.4 Summary of Predicted Insights

Overall, the model's results showcase the impact of state-level factors, polling characteristics, and candidate-specific trends on predicted outcomes. The Bayesian GLM, enhanced with Monte Carlo simulations, enables the estimation of win probabilities and predicted support with a quantifiable degree of uncertainty. Harris' predicted advantage, while modest, is supported by stable polling across states, whereas Trump's broader distribution points to more variable support, especially in key states. This analysis provides a foundation for understanding competitive dynamics and regional differences in the 2024 U.S. presidential race, with implications for candidate strategies in battleground states.

## 5 Discussion - Predictive Modeling of 2024 Election Outcomes

### 5.1 Key Findings

The analysis utilizes Bayesian Generalized Linear Models (GLMs) and Monte Carlo simulations to estimate predicted support percentages for Kamala Harris and Donald Trump. With a win probability of 63.2% for Harris and 36.8% for Trump, the model suggests a modest advantage for Harris, albeit within a competitive margin. The density plot Figure 1 previously introduce in Section 2.3, highlights the distributions of predicted support for both candidates, showing a slight overlap, which indicates close competition and shows the uncertainty inherent in forecasting elections based on polling data.

Our findings emphasize the roles of pollster reliability, sample size, and state-level effects in influencing predicted support. Specifically, the distributions of predicted support Figure 5 and Figure 4 reveal that both Harris and Trump have a strong central tendency around similar support percentages, with Harris showing a slightly higher mean. This slight advantage for Harris in the distribution of predicted support aligns with her higher win probability. However, the narrowness of this advantage indicates that slight variations in support could easily alter the projected outcome.

### 5.2 Influence of Sample Size and Pollster Effects

The results suggest that larger sample sizes generally produce more stable support estimates, reducing the influence of random fluctuations in polling data. Polls with larger sample sizes are less likely to exhibit extreme support values, which improves the reliability of the predictions. However, smaller sample sizes tend to introduce variability, often exaggerating support levels for one candidate. For instance, in battleground states with smaller sample sizes, our model indicates higher variability in support percentages, which could lead to overestimated support in certain cases. Thus, integrating sample size as a predictor helps moderate these extremes and improves the precision of predictions.

Pollster effects are also significant in this analysis. Different pollsters have unique methodologies, sampling techniques, and weighting schemes, which can lead to systematic differences in reported support. By including pollster as a predictor in the model, we control for these variations, allowing for a more standardized comparison of support across polls. This inclusion reduces the likelihood that any single pollster's methodology disproportionately skews the predictions, resulting in more balanced estimates across different polls.

## 5.3 State-Level Variations in Support

The model highlights the importance of state-level effects, revealing that support for each candidate varies significantly by region. For example, Harris is shown to have strong support in traditionally Democratic states, whereas Trump maintains a lead in Republican-leaning regions. In battleground states such as Florida and Pennsylvania, the predicted support percentages are closer, reflecting the competitive nature of these regions. The density plot in Figure 1 introduced in section Section 2.3 illustrates that while Harris has a slight overall advantage, support for both candidates closely overlaps in key states, indicating that regional factors are pivotal in determining final outcomes.

Our results underscore the necessity of including state-level predictors to capture these regional variations. Ignoring state-level effects could oversimplify the model, potentially misrepresenting the strength of support for each candidate in different parts of the country. By accounting for these variations, our model can provide a more detailed prediction that aligns with observed state-specific voting patterns.

## 5.4 Weaknesses

While the models presented provide a thorough approach to forecasting the 2024 election, both the data and models have inherent weaknesses that may impact the accuracy of the predictions.

### 5.4.1 Data-Related Weaknesses

The polling data used in this analysis, while thorough, presents challenges in terms of representativeness and variability. Polls differ widely in sample size, methodology, and quality, which can introduce biases that are difficult to adjust for consistently across states. For instance, smaller sample sizes, particularly in less-populated or battleground states, may lead to exaggerated or unstable estimates of support. Additionally, polling firms often adopt different weighting schemes or sampling frames, which may not fully capture the demographic and geographic diversity within each state. This variability can result in inconsistencies when aggregating polls across different regions, potentially skewing the predictions for states with sparse polling data.

Another challenge is the limited frequency of polling in certain regions. Many states receive minimal polling attention, leading to sparse data that may not accurately reflect shifts in voter sentiment over time. This lack of consistent polling can cause the model to rely on outdated or unrepresentative data points, especially in states where political dynamics are changing swiftly. Consequently, the model may overestimate or underestimate support in these areas, reducing the accuracy of the overall predictions.

### 5.4.2 Model-Related Weaknesses

The model's reliance on a Gaussian distribution to predict support percentages may not fully account for the non-normality often observed in polling data. Polling data can exhibit skewed distributions or heavy tails, particularly in states where support for a candidate is highly polarized. By assuming a Gaussian distribution, the model may miss these variations, potentially leading to biased estimates in states where the distribution of support is not symmetric.

Additionally, the model assumes that the effect of each pollster on support estimates is consistent across all states, an assumption that may oversimplify regional differences in polling accuracy. Different pollsters have varying levels of reliability in specific regions, influenced by factors like local sampling techniques, demographic targeting, and survey timing. By treating pollster effects as uniform, the model may fail to capture these subtleties, potentially resulting in inaccurate support estimates for states where polling practices differ significantly.

The static nature of the state-level effects in the model also presents a limitation. Political dynamics within states are fluid, particularly in the lead-up to an election, when voter sentiment may shift swiftly in response to events, campaigns, or emerging issues. By treating state-level effects as constant, the model lacks the flexibility to account for these temporal changes. Consequently, predictions based on older polls may not accurately reflect current support levels, especially in states with fast-changing political landscapes.

### 5.4.3 Impact on Prediction Accuracy

These data and model weaknesses underscore the challenges of election forecasting and highlight the potential for misestimations in certain regions. The variability in polling data quality and the assumption of static state effects may reduce the model's predictive accuracy, particularly in battleground states where small shifts in voter sentiment could have a significant impact on the election outcome. Furthermore, the lack of temporal adaptation in the model may lead to outdated predictions as the election approaches, particularly if there are last-minute shifts in public opinion.

## 5.5 Future Directions

To enhance the predictive accuracy of election forecasting models, future research could focus on several key improvements. One promising avenue is the integration of time-series elements to account for temporal changes in voter sentiment as election day approaches. A time-series approach would allow the model to weigh recent polls more heavily than older data, reflecting shifts in public opinion as new events, debates, and campaign efforts unfold. This would result in a more dynamic model that adapts to real-time changes in voter preferences, increasing prediction accuracy closer to election day.

Incorporating voter turnout predictions based on historical data and demographic trends could further improve the model's reliability. Turnout often varies by demographic group, driven by factors such as candidate appeal, mobilization efforts, and election-specific conditions like weather. Adjusting predictions based on turnout likelihoods by group would create a more detailed view of potential election outcomes, especially in swing states where voter turnout can significantly impact results.

Expanding the Bayesian framework to incorporate historical voting patterns as informative priors would be particularly valuable in states with sparse or inconsistent polling. Bayesian methods can update predictions as new polling data becomes available, grounding predictions in past election trends while allowing flexibility for current shifts. This approach would mitigate the volatility often seen in states with limited data, leading to more stable estimates.

Finally, integrating alternative data sources, such as social media sentiment analysis or economic indicators, could enrich the model by capturing real-time shifts in voter sentiment not reflected in traditional polling. For instance, social media activity following major events like debates or policy announcements could serve as an indicator of changes in support, supplementing polling data to create a more responsive forecasting model. Economic indicators, such as unemployment rates or inflation, could also help determine voter priorities and potential support patterns, enhancing the model's predictive capacity in economically sensitive regions.

Implementing these advancements would address current data and model limitations, creating a more adaptive and robust forecasting tool capable of accurately capturing the complexities of voter behavior in a dynamic election landscape.

## 5.6 Conclusion

The analysis demonstrates that both pollster reliability and sample size are key factors in accurately predicting support percentages, while state-specific effects capture essential regional variations. However, the model's assumptions and limitations suggest avenues for further refinement. Incorporating dynamic, turnout-adjusted, and multi-source data into future models could yield more precise election forecasts. In sum, while this model provides a solid foundation for understanding potential outcomes in the 2024 election, continuous refinement and adaptation to new data sources and methods will enhance the reliability of election prediction models.

# Appendix

## A Pollster Methodology Overview and Evaluation

The Emerson College Polling method utilizes a hybrid approach, combining Interactive Voice Response (IVR) with an online panel to gather data from likely voters in Michigan. This methodology is used in translating individual opinions into quantifiable data that can inform electoral forecasts.

**Population, Frame, and Sample**
The target population is defined as the collection of all items about which we would like to speak, in this case it would be all voters in the state of Michigan.

The sampling frame defined as a list of all the items from the target population that we could get data about, in this case it would be all voters who are able to respond to the IVR or online panel survey

A sample is defined as the items from the sampling frame that we get data about, in this case that would be the 1000 people that responded to the ICR or online panel.

Non-response refers to missing data in a dataset when respondents either skip specific questions, groups of questions, or even refuse to participate entirely. Non-response from individuals introduces sampling error, as it may lead to biases in the results if certain population segments are underrepresented. Government statistics are generally reliable but must be viewed critically, especially considering the potential for political influences, methodological limitations, or errors. Despite this, when methodologies are transparent and data is subjected to independent audits, government statistics can still serve as valuable sources of information for policy research, but users should always be aware of potential biases and limitations.

**Sample Recruitment and Sampling Approach**
Emerson employs both IVR and online panel methodologies to recruit participants. The IVR method facilitates automated phone interviews, reaching demographics that may be less likely to participate in online surveys, such as older individuals or those with limited internet access. In contrast, the online panel is sourced from CINT, engaging participants who are more willing to complete surveys digitally. This dual approach captures a broader spectrum of voter sentiments and demographics, balancing the advantages of both methods. However, while IVR can access hard-to-reach populations, it may also yield lower response rates compared to the more engaged online panel.

**Non-response Handling**
While the IVR method can reach various demographics, it is essential to consider potential non-response bias. Respondents who opt not to participate may share similar characteristics, leading to skewed results. Emerson may implement follow-up strategies or incentives to encourage participation, but these measures are not explicitly detailed in the methodology.

**Questionnaire Evaluation**
Emerson crafted a thorough questionnaire that included a range of questions, such as "Do you approve or disapprove of the job Joe Biden is doing as President?" and "What is your party registration?" Respondents provided their answers using multiple-choice formats, offering options like "Approve," "Disapprove," "Neutral or no opinion," as well as party affiliations such as "Democrat," "Republican," and "Independent/other." The full questionnaire and its results can be accessed at this link. After collecting the responses, the results were tallied to calculate the percentage for each option, enabling pollsters to convert individual opinions into broader trends. While the questions are generally clear and allow for straightforward responses, careful consideration should be given to the wording to avoid any potential bias. Questions that lead respondents toward a particular answer could impact the validity of the data collected.

**Overview of Methodology**

- **Data Collection Techniques:** Emerson employs both IVR and online panel methodologies. IVR allows for automated phone interviews, which can reach a diverse demographic, including those less likely to participate in online surveys. The online panel, sourced from CINT, complements this by engaging participants who are more accessible and willing to complete surveys digitally. This dual approach is intended to capture a broad spectrum of voter sentiments and demographics.
- **Sample Size and Credibility:** The survey includes a sample of 1,000 likely voters, providing a reasonable basis for statistical inference. Emerson states a credibility interval of ±3 percentage points, which indicates a moderate level of precision in their estimates. This credibility interval should be considered when interpreting results, as it reflects the potential variability in public opinion.
- **Weighting Adjustments:** To ensure that the sample accurately reflects the population of likely voters, the data is weighted by gender, education, race, age, party registration, and region. This weighting is key, as it aligns the sample with demographic characteristics of the actual voter population, enhancing the reliability of the findings.

**Strengths of the Methodology**

- **Diverse Reach:** The combination of IVR and online panel methods allows Emerson to capture data from a wide range of demographics, minimizing the risk of bias that may arise from using a single method. This is particularly important in a politically diverse state like Michigan.
- **Rigorous Weighting:** The approach to weighting the data helps to correct any imbalances in the sample, ensuring a more accurate reflection of voter sentiments across different demographic groups.

- **Clarity in Reporting:** Emerson provides detailed demographic breakdowns and cross-tabulations, allowing for a accurate understanding of how different groups perceive key issues and candidates.

**Limitations of the Methodology**

- **Potential Non-response Bias:** While the IVR method can reach various demographics, those who opt not to participate may share similar characteristics, potentially leading to non-response bias. Additionally, the online panel may attract a skewed sample if certain demographics are overrepresented or underrepresented in the panel.
- **Higher Credibility Intervals for Subgroups:** Subsets based on demographics (e.g., age or party registration) often come with larger credibility intervals due to smaller sample sizes. This can diminish the reliability of conclusions drawn from those specific groups.
- **Temporal Limitations:** The survey results reflect a snapshot of opinions at a specific time (October 25-27, 2024), which may not account for swiftly changing political landscapes, particularly as Election Day approaches.

In summary, Emerson College Polling employs a robust methodology that effectively combines IVR and online panel methodologies to gauge voter sentiment. While there are inherent limitations, such as potential biases and the credibility of subgroup analysis, the strengths of this approach lie in its thorough reach and careful weighting adjustments. Overall, this methodology offers a valuable tool for understanding the current political climate in Michigan.

## B Idealized Survey

In this appendix, I outline an idealized methodology and survey for forecasting the U.S. presidential election, utilizing a budget of $100,000. This approach combines robust sampling techniques, diverse recruitment strategies, and thorough data validation to ensure accurate and reliable information on voter sentiment.

**Sampling Approach**

1. **Target Population**: The target population is defined as the collection of all items about which we would like to speak, in this case it would consists of all registered voters in the United States. The survey will focus on likely voters, including individuals aged 18 and older across various demographic groups, including age, gender, race, and political affiliation.
2. **Sample Size**: A sample size of approximately 5,000 likely voters will be utilized. This sample size is designed to provide a high level of statistical confidence, with a credibility interval of ±2 percentage points.

3. **Sampling Technique**:

A **stratified random sampling** method which involves dividing subjects into subgroups on characteristics that they share and than these subgroups are randomly sampled. In this case we will choose the subgroups based on demographic variables, such as age, gender, race, and region, we will than randomly sample participants from these subgroups.

**Recruitment of Respondents**

1. **Data Sources**:

   - Voter registration databases will be utilized to identify potential respondents. These databases will provide contact information for registered voters across the United States.
   - Partnerships with organizations that specialize in voter engagement will also be established to enhance the recruitment process.

2. **Recruitment Methods**:

   - **Online Panel**: An online panel sourced from reputable survey platforms (e.g., SurveyMonkey, Qualtrics) will be used to recruit respondents who opt-in to participate in surveys.
   - **IVR and SMS Outreach**: Interactive Voice Response (IVR) and SMS outreach will be implemented to reach demographics that may be less likely to engage in online surveys. This will ensure inclusivity and broaden the reach of the survey.

3. **Incentives**: Participants will be incentivized to complete the survey with small monetary rewards (e.g., $5) or entry into a raffle for a larger prize, thereby increasing response rates and engagement.

**Data Validation and Quality Assurance**

1. **Pre-Survey Testing**: A pilot test of the survey will be conducted with a smaller group of respondents (approximately 200) to identify any ambiguities or biases in the questions. Feedback will be incorporated to refine the survey instrument.

2. **Quality Checks**:

   - During data collection, real-time monitoring will be implemented to identify and address any inconsistencies in responses.
   - Automated checks will flag any duplicate responses, incomplete surveys, or inconsistent answer patterns.

3. **Post-Survey Validation**: After data collection, a validation process will be employed to cross-reference responses with demographic data to ensure the sample reflects the broader population accurately.

4. **Importance of Non-Response**: Non-response refers to missing data in a dataset when respondents either skip specific questions, groups of questions, or even refuse to participate entirely. Non-response from individuals introduces sampling error, as it may lead to biases in the results if certain population segments are underrepresented. To address this issue we will aim to follow up with non-respondents, using reminders or incentives to encourage participation. Additionally, we will utilize simulations to determine if there are any groups which are underrepresented within out data and than use weighted adjustments to compensate for these underrepresented groups.

### Poll Aggregation and Analysis

1. **Data Aggregation**: The data will be aggregated and analyzed using statistical software (e.g., R or Python) to compute margins of error and establish confidence intervals for key questions, such as candidate approval ratings and voting intentions.

2. **Cross-Tabulation**: The survey results will be cross-tabulated by demographic variables to provide information on how different groups perceive candidates and issues. This will enable analysis of voter sentiment across demographics.

3. **Reporting**: Findings will be compiled into a thorough report summarizing key observations, trends, and forecasts for the upcoming presidential election. The report will be shared with stakeholders, including political analysts and campaign teams.

### Survey Implementation

The survey will be implemented using Google Forms. Below is a link to the actual survey, which includes an introductory section, well-constructed questions, and a closing section thanking respondents for their participation:

**Survey Link:** U.S. Presidential Election Forecasting Survey

**Survey Copy:**

### Introduction

Thank you for participating in our survey! Your input is invaluable for understanding voter sentiment leading up to the upcoming US presidential election. This survey will take approximately 5-10 minutes to complete. Your responses are anonymous and will be used solely for research purposes.

If you have any questions or concerns, please contact Elizabeth Luong at elizabethh.luong@mail.utoronto.ca.

### Section 1: Demographic Information

1. **What is your age?**

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

2. **What is your gender?**
   - Male
   - Female
   - Non-binary/Third gender
   - Prefer not to say
3. **What is your race or ethnicity:**
   - White
   - Black or African American
   - Asian
   - Hispanic or Latino
   - Other (please specify): [Open text field]
4. **In which state do you reside?**
   - [Open text field]
5. **What is your highest level of education completed?**
   - Some high school
   - High school diploma or equivalent
   - Some college
   - Bachelor's degree
   - Graduate degree
6. **What is your party affiliation?**

- Democrat
- Republican
- Independent
- Other (please specify): [Open text field]

**Section 2: Political Opinions**

6. **How likely are you to vote in the upcoming presidential election?**
    - Very likely
    - Somewhat likely
    - Unsure
    - Somewhat unlikely
    - Very unlikely
7. **Which candidate do you plan to vote for in the upcoming presidential election?**
    - Donald Trump (Republican)
    - Kamala Harris (Democrat)
    - Robert F. Kennedy Jr. (Independent)
    - Undecided
    - Other (please specify): [Open text field]
8. **What are the most important issues influencing your vote?** (Select up to three)
    - Economy
    - Healthcare
    - Education
    - Climate Change
    - Social Justice
    - National Security
    - Immigration
    - Other (please specify): [Open text field]
9. **How do you feel about the job performance of the current President?**

- Strongly approve
- Approve
- Neutral
- Disapprove
- Strongly disapprove

**Section 3: Media Consumption**

10. **Which sources do you primarily use to get news about the election?** (Select all that apply)

    - Television
    - Online news websites
    - Social media
    - Radio
    - Newspapers
    - Podcasts
    - Other (please specify): [Open text field]

11. **How often do you discuss politics with friends and family?**

    - Daily
    - Weekly
    - Monthly
    - Rarely
    - Never

**Conclusion**

Thank you for taking the time to complete this survey! Your feedback is important for understanding voter sentiment as we approach the election. If you have any additional comments or thoughts you'd like to share, please reach out to Elizabeth Luong at elizabethh.luong@mail.utoronto.ca.

**Thank you again for your participation!**

## References

Firke, Sam. 2021. *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.

Gabry, Jonah, and Månnel Bürkner. 2021. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

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

Iannone, Richard, and RStudio. 2021. *Gt: Easily Create Presentation-Ready Display Tables*. https://CRAN.R-project.org/package=gt.

Keitt, Timothy H., and Winston Chang. 2021. *Webshot2: Take Screenshots of Web Pages*. https://CRAN.R-project.org/package=webshot2.

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/.

Team, Apache Arrow Development. 2021. *Arrow: Integration to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

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. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2021. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.