# Forecasting the 2024 US Presidential Election*

**My subtitle if needed**

Tina Kim David Flores Kevin Shao

October 27, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

## 1 Introduction

*Overview paragraph:* Provide a brief overview of the context of the upcoming US presidential election and the importance of forecasting its results.

*Estimand paragraph:* Define the estimand clearly (e.g., predicting the probability of a candidate winning the election based on poll data).

*Results paragraph:* Summarize the key findings of the model, highlighting its predictive accuracy and implications.

*Why it matters paragraph:* Explain the broader significance of accurately forecasting election results for politics, society, and policymaking.

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

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

# 2 Data

## 2.1 Overview

Latest Poll Points by Party

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 primary objective was to clean, organize, and analyze U.S. presidential election polling data to provide insights into voter preferences by state. This involved selecting data from reputable pollsters, organizing it by state, and addressing any missing or incomplete entries. Additional datasets were merged where necessary, and after thoroughly addressing missing values, we finalized a polished dataset ready for analysis. Below is a table summarizing the polling data by state, showcasing voter preferences by party and demographic breakdowns. The table highlights key trends and differences across states in terms of party lean and demographic distribution. **?@fig-predictors-pct-summary** provides a snapshot of these findings.

## 2.2 Measurement

For this analysis, we selected Emerson as our primary pollster. The rationale behind this choice is explained in detail in Section 3. Emerson's polls provided essential details, including polling date ranges, party affiliations, sample sizes, and the percentage breakdown of support for different political parties.

We focused on analyzing voter preferences based on party affiliation rather than individual candidate support, as this was more relevant to predicting overall trends. By capturing the broader party dynamics, our model aims to forecast how states are likely to swing in future elections, especially with respect to swing states. This focus will be elaborated in our results section, Section 5.

Additionally, the Emerson dataset included critical demographic data, such as respondents' political preferences, motivations for voting in the 2024 election, opinions on Donald Trump's legal battles, and perspectives on major political figures and issues. These details, coupled with information on party affiliation, sources of political news, and other personal attributes, allowed us to paint a richer picture of the voting landscape.

To ensure uniformity across all states, we concentrated on questions that were consistently asked across the entire dataset. This enabled us to perform state-level aggregation and analysis without being limited by state-specific polling questions, ensuring a standardized approach across the board.

| State | Democrat Pct | Republican Pct |
|---|---|---|
| Arizona | 48.2 | 50.5 |
| Arkansas | 24.4 | 56.6 |
| California | 61.0 | 36.6 |
| Colorado | 41.0 | 35.0 |
| Connecticut | 49.4 | 40.0 |
| Florida | 45.9 | 53.6 |
| Georgia | 49.8 | 49.7 |
| Idaho | 25.8 | 54.5 |
| Illinois | 43.4 | 33.7 |
| Indiana | 40.7 | 57.9 |
| Iowa | 39.6 | 48.2 |
| Kansas | 31.0 | 47.0 |
| Kentucky | 26.3 | 54.6 |
| Maine | 50.9 | 40.1 |
| Maryland | 64.0 | 34.4 |
| Massachusetts | 54.0 | 33.7 |
| Michigan | 49.7 | 49.6 |
| Minnesota | 51.0 | 49.0 |
| Missouri | 43.4 | 55.2 |
| Montana | 42.5 | 57.5 |
| National | 49.3 | 49.3 |
| Nebraska | 30.9 | 46.8 |
| Nevada | 49.2 | 48.1 |
| New Hampshire | 50.9 | 47.2 |
| New Jersey | 45.7 | 39.4 |
| New Mexico | 53.6 | 46.4 |
| New York | 55.5 | 41.6 |
| North Carolina | 48.2 | 50.2 |
| North Dakota | 17.0 | 54.0 |
| Ohio | 44.7 | 54.3 |
| Oklahoma | 27.0 | 55.0 |
| Oregon | 50.7 | 35.3 |
| Pennsylvania | 49.0 | 50.7 |
| South Carolina | 36.7 | 50.6 |
| South Dakota | 36.5 | 62.0 |
| Tennessee | 22.0 | 55.3 |
| Texas | 45.9 | 53.1 |
| Utah | 33.6 | 47.0 |
| Virginia | 52.7 | 46.2 |
| Washington | 49.4 | 38.7 |
| West Virginia | 22.5 | 59.3 |
| Wisconsin | 49.2 | 49.9 |
| Wyoming | 14.8 | 67.6 |

Figure 1

| State | High Educated | Low Educated | Female | Male | Nonbinary | Caucasian | Minority Ethnicity |
|---|---|---|---|---|---|---|---|
| Arizona | 23.8 | 24.1 | 51.8 | 47.0 | 1.2 | 71.2 | 28.8 |
| Arkansas | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| California | 23.8 | 24.1 | 51.9 | 47.2 | 0.9 | 71.2 | 28.8 |
| Colorado | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Connecticut | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Florida | 20.8 | 24.0 | 52.1 | 47.3 | 0.6 | 65.2 | 34.8 |
| Georgia | 23.8 | 24.1 | 52.0 | 47.3 | 0.7 | 71.2 | 28.8 |
| Idaho | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Illinois | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Indiana | 21.1 | 27.6 | 52.2 | 47.0 | 0.8 | 81.0 | 19.0 |
| Iowa | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Kansas | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Kentucky | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Maine | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Maryland | 20.8 | 24.0 | 52.1 | 47.3 | 0.6 | 65.2 | 34.8 |
| Massachusetts | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Michigan | 18.6 | 25.5 | 49.3 | 49.2 | 1.5 | 87.9 | 12.1 |
| Minnesota | 30.1 | 17.7 | 52.5 | 46.7 | 0.8 | 68.3 | 31.7 |
| Missouri | 18.6 | 25.5 | 49.3 | 49.2 | 1.5 | 87.9 | 12.1 |
| Montana | 23.8 | 24.1 | 51.8 | 47.0 | 1.2 | 71.2 | 28.8 |
| National | 23.6 | 26.3 | 51.1 | 47.2 | 1.7 | 64.8 | 35.2 |
| Nebraska | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Nevada | 18.6 | 25.5 | 49.3 | 49.2 | 1.5 | 87.9 | 12.1 |
| New Hampshire | 22.1 | 24.3 | 51.8 | 46.4 | 1.8 | 80.2 | 19.8 |
| New Jersey | 23.1 | 23.6 | 54.3 | 45.4 | 0.3 | 71.9 | 28.1 |
| New Mexico | 30.1 | 17.7 | 52.5 | 46.7 | 0.8 | 68.3 | 31.7 |
| New York | 18.6 | 25.5 | 49.3 | 49.2 | 1.5 | 87.9 | 12.1 |
| North Carolina | 25.8 | 21.8 | 53.7 | 44.6 | 1.7 | 60.9 | 39.1 |
| North Dakota | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Ohio | 23.8 | 24.1 | 51.8 | 47.0 | 1.2 | 71.2 | 28.8 |
| Oklahoma | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Oregon | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Pennsylvania | 23.8 | 24.1 | 52.1 | 47.2 | 0.7 | 71.2 | 28.8 |
| South Carolina | 21.1 | 24.7 | 50.8 | 48.6 | 0.0 | 77.7 | 22.3 |
| South Dakota | 20.8 | 24.0 | 52.1 | 47.3 | 0.6 | 65.2 | 34.8 |
| Tennessee | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Texas | 20.9 | 24.2 | 51.9 | 47.3 | 0.8 | 65.2 | 34.8 |
| Utah | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Virginia | 18.6 | 25.5 | 49.3 | 49.2 | 1.5 | 87.9 | 12.1 |
| Washington | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| West Virginia | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Wisconsin | 20.8 | 24.0 | 52.1 | 47.3 | 0.6 | 65.2 | 34.8 |
| Wyoming | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Figure 2

## 2.3 Outcome variables

The initial analysis of the dataset revealed some key trends in party preferences across different states. In Figure 3, we illustrate which states exhibit the strongest leanings toward either the Democratic or Republican party. This visualization plays a crucial role in identifying the pivotal swing states that lie close to the center of the graph. These swing states, which show near-equal support for both parties, are of particular interest for future election predictions, as they are likely to determine the overall election outcome.

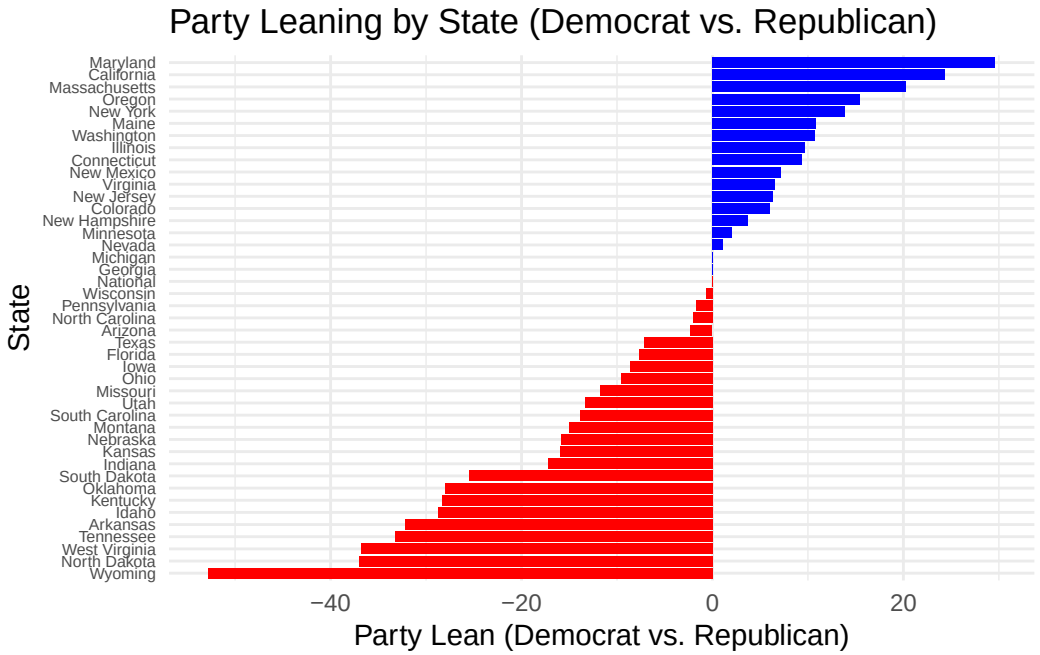



Figure 3: Example Predictor Variable Visualization

Beyond identifying swing states, it's also important to understand the differences in demographics between the most strongly Democratic ("blue") and Republican ("red") states. The bar graph in Figure 4 takes a closer look at these differences. It focuses on key factors that show the biggest contrasts between Democratic-leaning and Republican-leaning states, helping to highlight how these factors differ between the most extreme red and blue states.

The Figure 4 visualization provides insight into how people's responses to polls may influence their political preferences. This extra layer of analysis helps clarify the factors shaping the political leanings of different states and adds depth to our understanding of the electorate. In this case, we examine people's opinions on the current president and their voting choices in past elections.

Through these visualizations, we not only identify which states are most likely to swing in future elections but also gain a deeper understanding of the public's opinions that contribute


Figure 4: Demographic Differences Between Top Red and Blue States

to these partisan divides.

## 2.4 Predictor variables

Next, we take a deeper dive into the various demographic and social factors that influence voter preferences. The predictors we consider include education levels, gender, ethnicity, and age group. These variables are crucial as they shape political preferences and are likely to influence which way a state leans politically.

In Figure 5, we explore how these factors correlate with party lean across the U.S. states. For instance, higher levels of education may correlate with a greater likelihood of supporting the Democratic party, while other demographic factors such as ethnicity and age may play different roles in shaping political affiliation. This analysis helps us to identify which factors have the greatest influence in each state, offering a clearer understanding of what drives the electorate.

The correlation matrix in Figure 5 provides a detailed look at how various predictors are related to party preferences. From this, we can see which predictors—such as education or gender—are more strongly associated with Democratic or Republican leanings. The blue shading indicates stronger correlations with Democratic support, while red represents stronger correlations with Republican support.

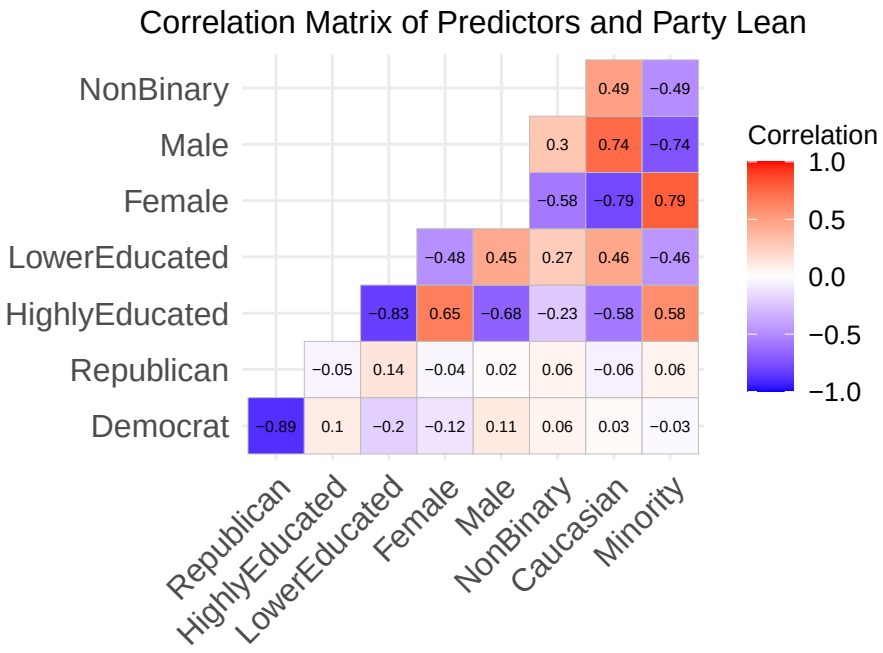



Figure 5: Example Predictor Variable Visualization

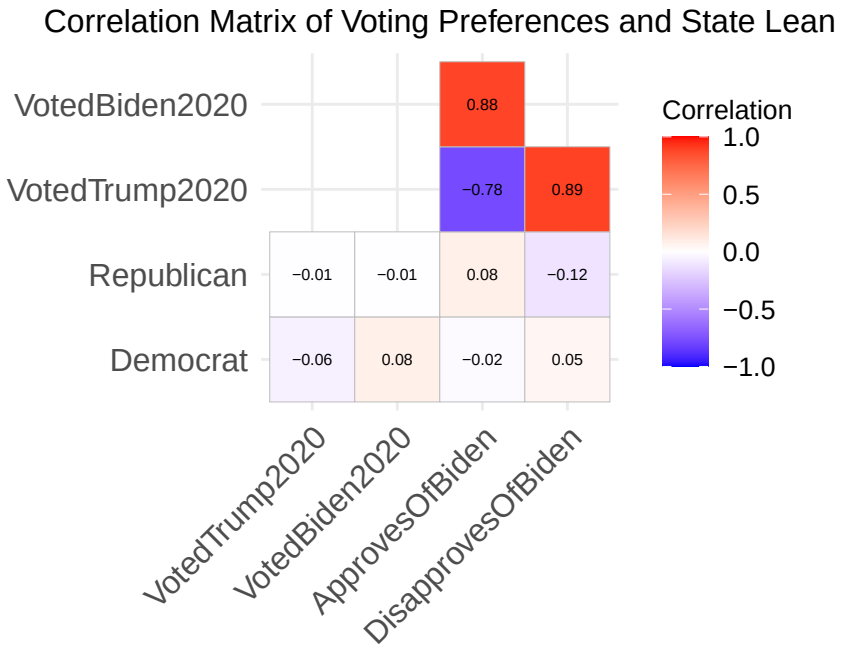



Figure 6: Example Voting and Approval Predictor Variable Visualization

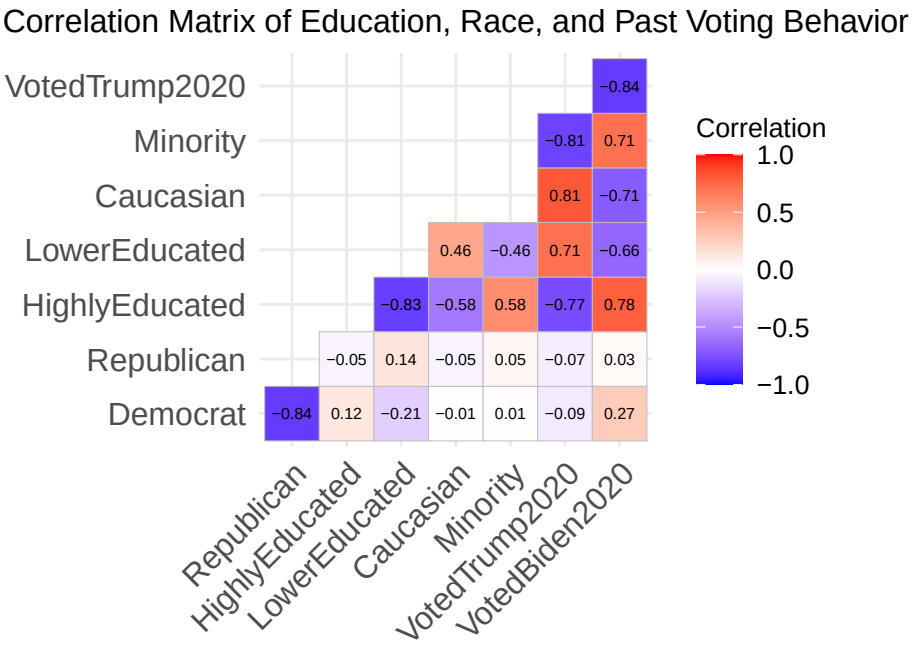



Figure 7: Correlation Matrix of Education, Race, and Past Voting Behavior

```
# Convert `end_date` to Date format and ensure numeric columns
time_data <- time_data %>%
  mutate(end_date = as.Date(end_date, format = "%m/%d/%y"),
         pct = as.numeric(pct),
         `Do you approve or disapprove of the job Joe Biden is doing as President? Approve` =
           as.numeric(`Do you approve or disapprove of the job Joe Biden is doing as Presider

# Average entries for the same date and party, then apply a 14-day rolling average for smoothi
time_data <- time_data %>%
  group_by(end_date, party) %>%
  summarise(
    pct = mean(pct, na.rm = TRUE),
    biden_approval = mean(`Do you approve or disapprove of the job Joe Biden is doing as Pres
  ) %>%
  ungroup() %>%
  group_by(party) %>%
  mutate(
    pct_smooth = rollapply(pct, width = 6, FUN = mean, align = "right", fill = NA)
  ) %>%
  ungroup() %>%
```

```
  mutate(
    biden_approval_smooth = rollapply(biden_approval, width = 26, FUN = mean, align = "right"
  )
```

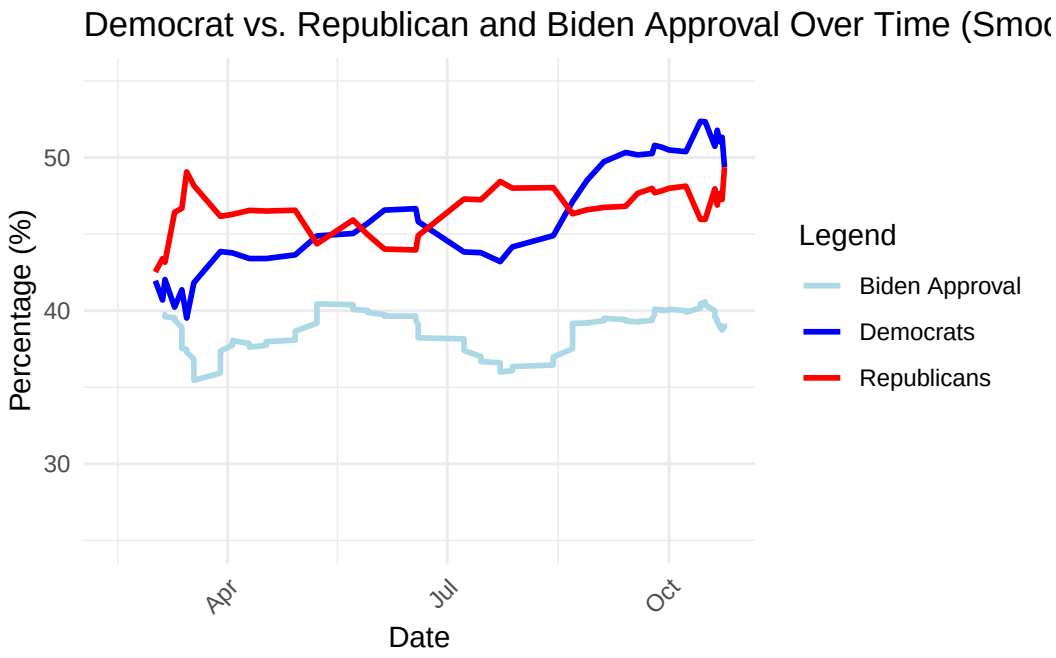
```
`summarise()` has grouped output by 'end_date'. You can override using the
`.groups` argument.
```

```
# Separate data for Democrats, Republicans, and Biden Approval
# Filter data for Democrats where state is NA
democrat_data <- time_data %>%
  filter(party == "DEM")

# Filter data for Republicans where state is NA
republican_data <- time_data %>%
  filter(party == "REP")

biden_approval <- time_data %>%
  filter(party == "DEM" | party == "REP") %>%
  select(end_date, biden_approval_smooth) %>%
  distinct()

# Create the time plot with ggplot2
ggplot() +
  # Plot Democrats' smoothed pct over time (blue)
  geom_line(data = democrat_data, aes(x = end_date, y = pct_smooth, color = "Democrats"), siz

  # Plot Republicans' smoothed pct over time (red)
  geom_line(data = republican_data, aes(x = end_date, y = pct_smooth, color = "Republicans")

  # Plot Biden's smoothed Approval over time (green)
  geom_line(data = biden_approval, aes(x = end_date, y = biden_approval_smooth, color = "Bide

  # Customize labels, theme, colors, and y-axis limits
  labs(title = "Democrat vs. Republican and Biden Approval Over Time (Smoothed)",
       x = "Date",
       y = "Percentage (%)",
       color = "Legend") +
  scale_color_manual(values = c("Democrats" = "blue", "Republicans" = "red", "Biden Approval"
  ylim(25, 55) + # Set y-axis range from 25 to 55
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: Removed 5 rows containing missing values or values outside the scale range
(`geom_line()`).
Removed 5 rows containing missing values or values outside the scale range
(`geom_line()`).
```

```
Warning: Removed 7 rows containing missing values or values outside the scale range
(`geom_line()`).
```




This analysis offers valuable insights into how specific demographic factors shape the political landscape in different states, allowing us to better predict electoral outcomes based on shifting demographic trends.

## 3 Pollster Methodology Overview (Appendix A)

Append a detailed review of a chosen pollster's methodology, survey techniques, strengths, and weaknesses.

# 4 Model

## 4.1 Model Development

The goal of our model is to forecast the popular vote outcome of the 2024 US presidential election.

*Person B:* Choose a linear or generalized linear model. Justify the selection based on your research goal and data structure.

## 4.2 Model Set-up

Define the model mathematically and contextually (e.g., linear model predicting vote share).

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i x_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\sigma \sim \text{Exponential}(1) \tag{5}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

*Person B*: Define all model components and assumptions.

## 4.3 Model Justification

Our results are summarized in Table **??**. Discuss the implications of the findings and their predictive accuracy.

# 5 Results

Our results are summarized in Table **??**.

*Person B*: Visualize the model's results and include any performance metrics (e.g., RMSE, test/train split).

# 6 Discussion

## 6.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

## 6.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

## 6.3 Third discussion point

Discuss what the model reveals about the election forecast and its potential impact on understanding voting behavior.

*Person C*: Discuss limitations of the model and areas for further improvement.

## 6.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

# Appendix

## .1 Appendix A: Pollster Methodology Overview

*Person A*: Provide a detailed methodology review for the chosen pollster, including survey design, population sampling, non-response handling, etc.

## .2 Appendix B: Idealized Survey Design for $100K Budget

*Person C*: Develop and describe an ideal survey design for forecasting the election with a $100K budget, and include survey link.

Sampling approach: The sampling approach we will use is stratified random sampling where we divide the population into strata based on…. (age, gender, education level, etc.) Recruit Respondents: We will recruit respondents using an online survey on Google Forms so that we can minimize the cost and maximize the range of respondents we can reach. We will spend a portion of our budget (specify here) to advertise these surveys and also send out emails, with an additional monetary incentive (specify here) to encourage more participation. Data validation: IP address tracking to prevent duplicate responses. Poll Aggregation: Incorporate Bayesian Inference.

Survey Link: (Short questionnaire asking for demographic questions to be added here) Copy of Survey:

## .3 Additional Data & Model Details

Include any technical details on data cleaning, model diagnostics, and posterior checks.

# A Additional data details

# B Model details

## B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected by, the data

## B.2 Diagnostics

**?@fig-stanareyouokay-1** is a trace plot. It shows... This suggests...

**?@fig-stanareyouokay-2** is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algorithm

## References

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

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

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