



# PREDICTING 2024

**US ELECTION**



ZUBAIR ATHA | EESHA BARUA | SIBI MARAPPAN | JUSTIN MATHEW | PRANITHA NATARAJEN

A horizontal row of 15 light gray stars.




# PROJECT OVERVIEW

This project aims to predict the outcome of the 2024 U.S. presidential election by applying Bayesian modeling techniques to recent polling data and historical election results. This data-driven approach leverages statistical models to combine insights from both current and past data, generating reliable predictions of each candidate's chances. The presentation covers four key areas:

1. **Data Exploration:** An initial analysis of the 2020 election and 2024 polling data, identifying essential trends and observations.
2. **Data Cleaning and Sampling:** Steps taken to ensure data consistency and quality, focusing on critical variables and states.
3. **Key Insights:** Observations that guide the predictive model and define important variables.
4. **Machine Learning Approach:** Implementation of a Bayesian model, a probabilistic approach that dynamically updates with new polling data, providing flexible and adaptive predictions.

## Data Sources:

- **2020 Election Results:** State-by-state votes for Trump and Biden from the 2020 election serve as a baseline for understanding recent voting patterns.
  - **FiveThirtyEight Polling Data:** Includes active polling data for 2024 and historical data from past elections dating back to 1968, allowing comparisons across time and providing an understanding of current voter sentiment.
- 

# INITIAL DATA EXPLORATION – 2020

## ELECTION RESULTS LINE DATA

The 2020 election dataset provides crucial insights into voter behavior in the previous presidential election, offering a foundation for analyzing the 2024 race. Key variables include:

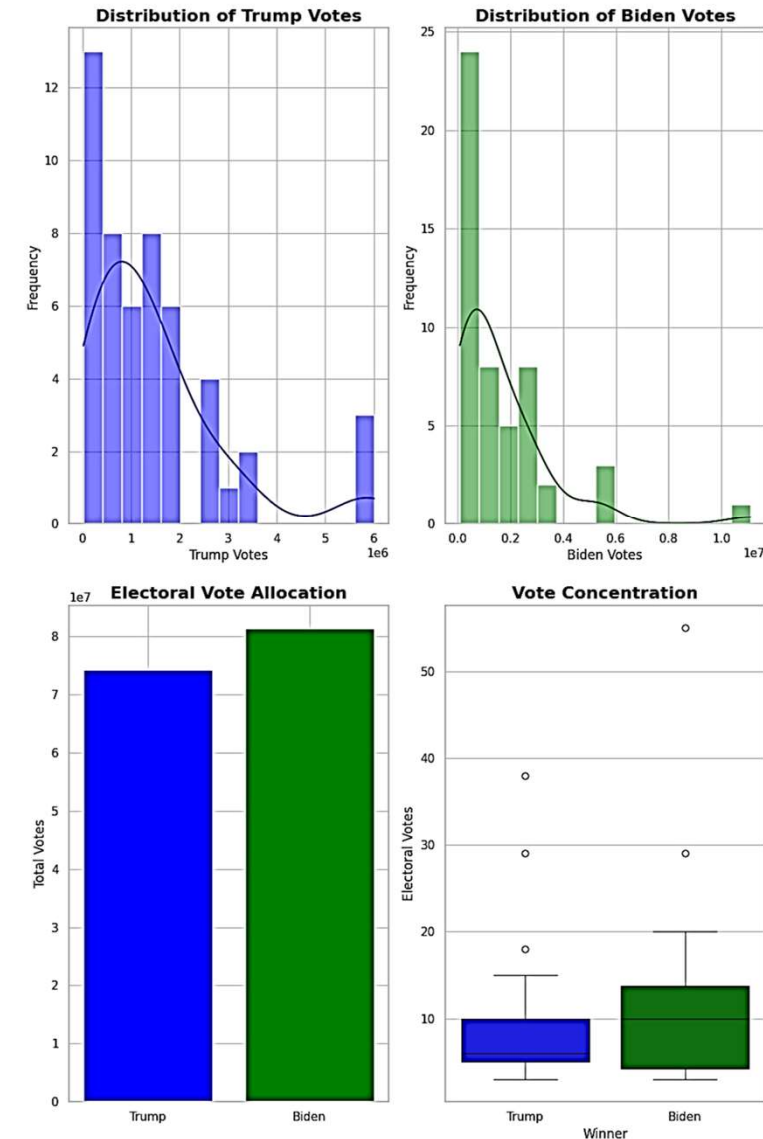
- **State:** Identifies each U.S. state.
- **Electoral Votes:** The number of electoral votes per state.
- **Total Votes for Trump and Biden:** The raw vote count for each candidate.
- **Winner:** Indicates the candidate who won the state in 2020.

### Visualizations:

1. **Vote Distribution Histogram:** A histogram visualizes the distribution of votes each candidate received across states, highlighting regional differences in support.
2. **Box Plot of Electoral Votes by Winner:** This plot shows the concentration of electoral votes won by each candidate, with Biden carrying states with higher electoral totals on average.

**Key Takeaways:** The 2020 results indicate strong regional trends, with Biden winning in states with higher populations and Trump performing well in traditionally red states. This data provides a baseline understanding of voter behavior, which informs the 2024 prediction model by highlighting historical voting patterns.

### 2020 Election Voting Data Analysis



# UNDERSTANDING CURRENT POLLING DATA FOR 2024 ELECTIONS

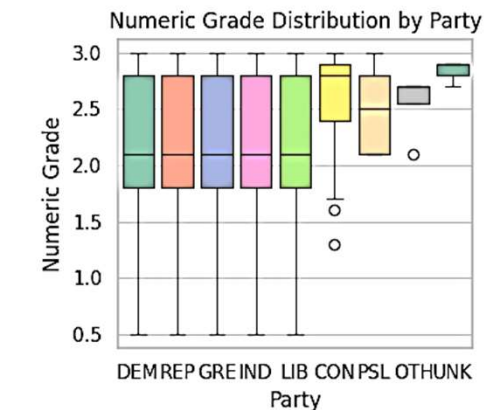
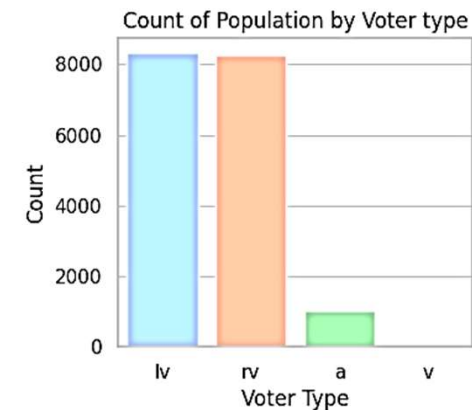
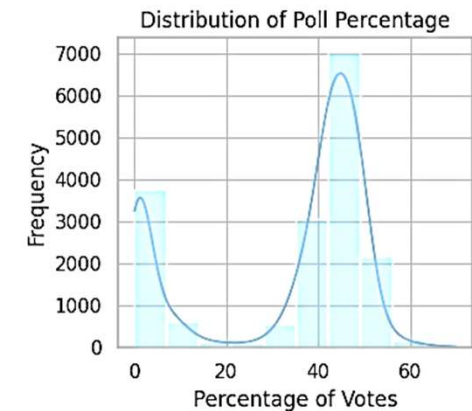
FiveThirtyEight's active polling dataset offers a snapshot of real-time public opinion in the 2024 election cycle. Key columns in this dataset include:

- **Pollster:** Organization conducting the poll.
- **Date:** Poll date, capturing the recency of the data.
- **Sample Size:** Number of respondents in each poll, impacting the reliability of the results.
- **Polling Methodology:** Method used (e.g., online, phone), affecting the sample's representativeness.
- **Population Type:** Differentiates between “likely voters” and “registered voters,” which affects data reliability.
- **Candidate Support Percentages:** Percentages of respondents supporting each candidate.

## Visualizations:

1. **Distribution of Poll Percentages:** This histogram highlights the variability of candidate support percentages across states.
2. **Population by Voter Type:** Visualizes the distribution of voter types, which affects poll reliability and accuracy. (lv=likely voters, rv=registered voters, a=adults, v=voters)
3. **Pollster Quality by Party:** A box plot comparing pollster reliability ratings across parties.

**Key Insight:** Polling data shows considerable variability, with polls from high-rated pollsters being more reliable. High sample sizes and reputable polling methods are critical, as they provide data with lower margins of error, essential for building accurate predictions.



# HISTORICAL POLLING TRENDS (1968-2020)

The historical polling dataset spans from 1968 to 2020, offering a longitudinal perspective on political trends across different categories of states.

## Visualization:

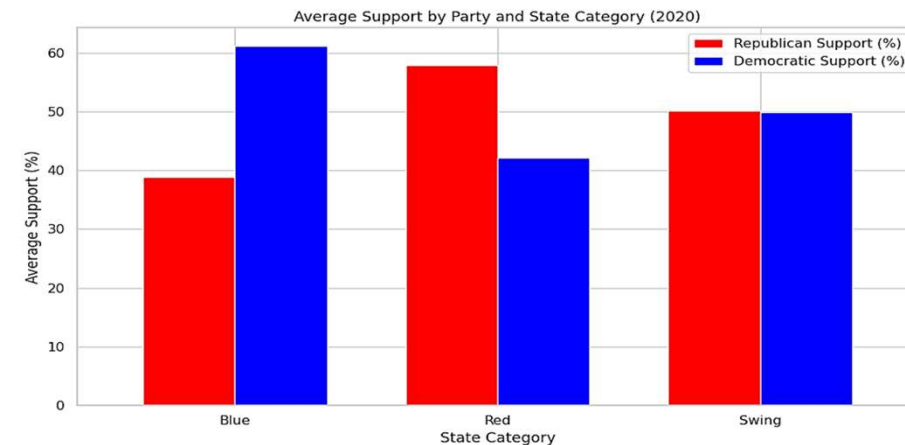
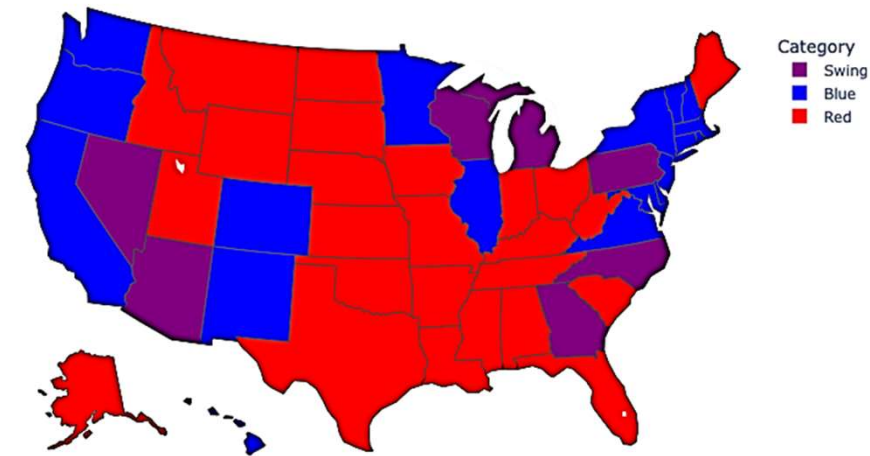
- **U.S. State Map with Colour Coding by Political Category:** Swing states in purple, red states in red, blue states in blue.
- **Bar Chart of Historical Trends by State Category:** This chart shows average polling support for each party within swing, blue, and red states across recent elections, providing a historical context for current trends.

## Observations:

- Blue (traditionally Democratic) and red states (traditionally Republican) are consistent, holding steady for their respective parties in 2024.
- Swing states show substantial variability in voting patterns and frequent shifts in party support, which means we need for a model that adapts dynamically to recent polling data.

**Key Takeaway:** While blue and red states generally maintain historical voting trends, swing states' variability makes them central to our prediction model. This insight helps determine weight distributions between historical and recent data in the Bayesian model.

2020 US Election State Categories



# DATA CLEANING AND FEATURE TRANSFORMATION

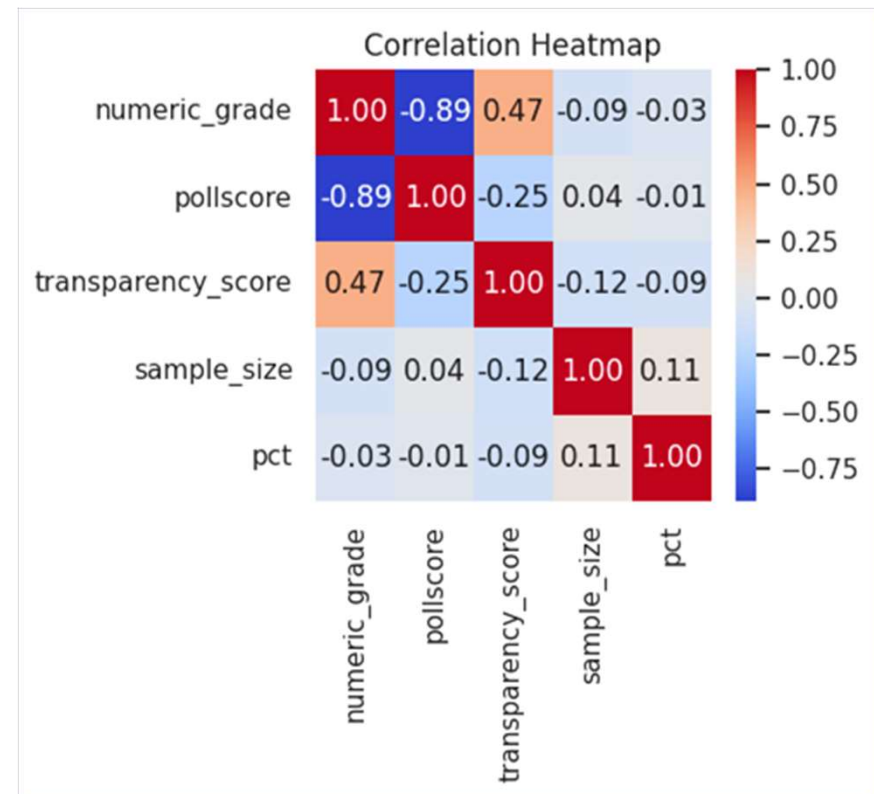
To ensure a reliable predictive model, data cleaning and sampling were undertaken to enhance data quality:

- **Feature Reduction:** Irrelevant or redundant columns (e.g., URLs, identifiers) were removed, focusing on features with a direct impact on predictions.
- **Handling Missing Data:** Missing numeric values (e.g., poll score, sample size) were filled with median values to maintain dataset integrity.
- **Sampling Strategy:** High-quality polling data from swing states received priority in sampling, as these states are likely to impact the final outcome more significantly.

## Visualization:

- **Correlation Heatmap of Key Numeric Features:** This heatmap reveals strong correlations between variables such as polling scores, transparency scores, and sample size, helping to inform which variables are prioritized in the model underscoring the importance of high-quality polls for reliability

**Key Takeaway:** By cleaning and sampling the data to focus on high-quality, relevant features, we reduce noise and improve the model's ability to make accurate predictions.



# POLLING TRENDS IN SWING STATES

Swing states are crucial to this analysis due to their history of variable election outcomes. For 2024, the focus is on states that are expected to be highly competitive: Pennsylvania, Wisconsin, Michigan, Georgia, Arizona, Nevada, and North Carolina.

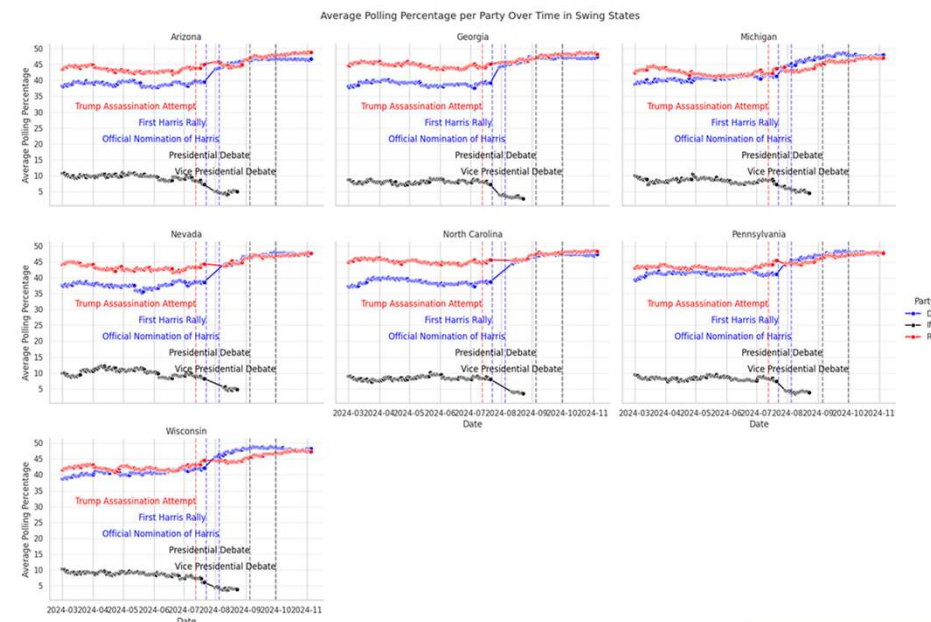
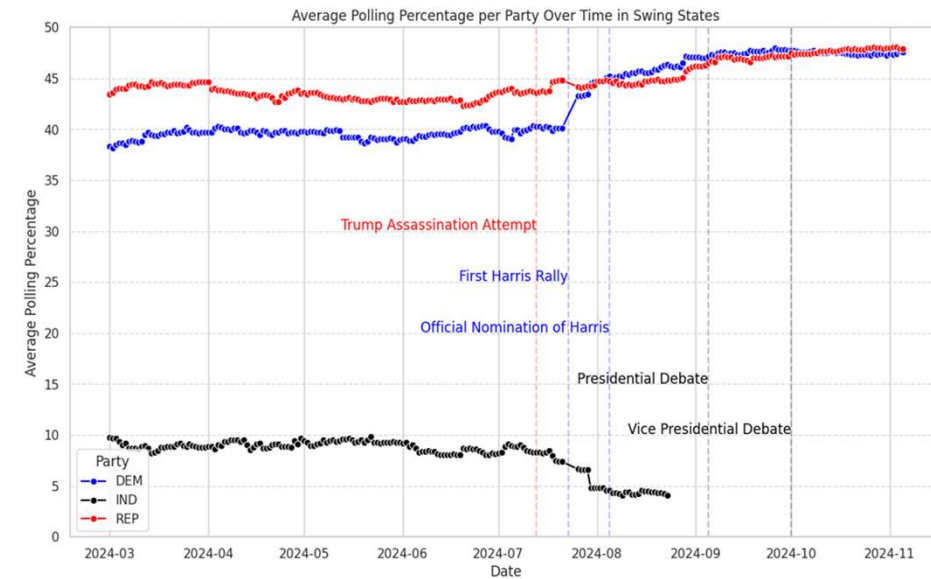
## Findings:

- Swing states show closely contested polling, with slight fluctuations reflecting high voter engagement and event sensitivity.
- The correlation between estimated and adjusted poll scores in swing states is strong at 0.995, suggesting that adjusted scores are highly reliable and help counter polling biases.

## Visualization:

- **Time Series Line Chart of Polling Trends Over Time in Swing States:** This line chart tracks each candidate's polling percentages over time, with markers for significant events like rallies and debates that influence voter sentiment.

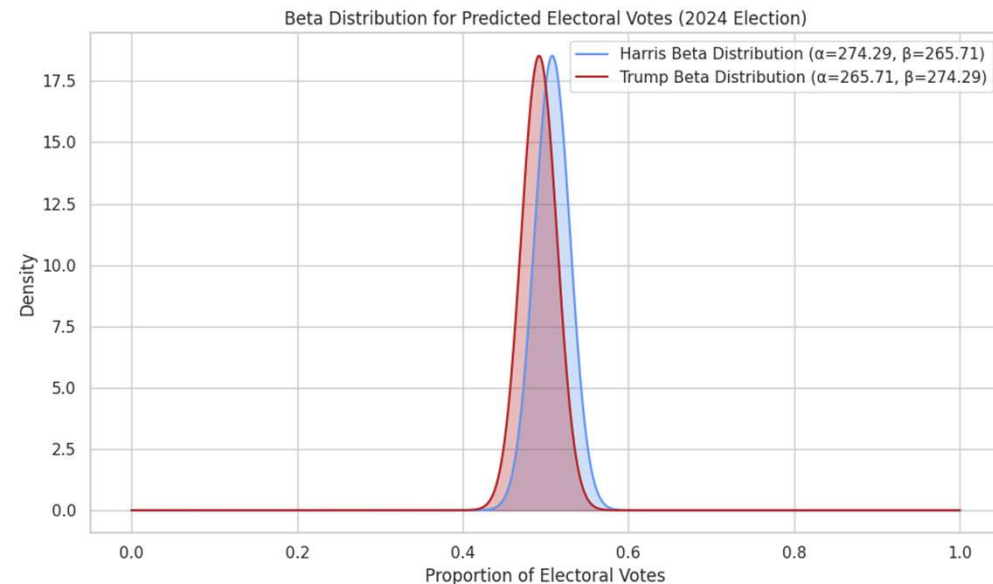
**Key Takeaway:** The minor differences in trend-adjusted polling scores suggest high polarization in swing states. This informs model confidence in these areas and highlights the importance of recent events in shaping public opinion.





# BAYESIAN MODELING TO PREDICT 2024 OUTCOMES

- **Why Bayesian Modeling?**
  - Bayesian models are well-suited for election forecasting because they allow us to combine prior data (2020 results) with current evidence (2024 polling), adjusting predictions as new data emerges.
- **Model Approach:**
  - We use a **Beta distribution** to simulate the probability of each candidate winning based on weighted averages of historical votes and recent polling data.
- **Visualization:**
  - **Beta Distribution Visualization:** This chart displays probability distributions for each candidate's chances of winning, demonstrating how different weights (for polling vs. historical data) affect these probabilities.
- **Key Takeaway:**
  - The Bayesian approach provides flexibility in adjusting for recent polling and makes probabilistic estimates of each candidate's success in the 2024 election.

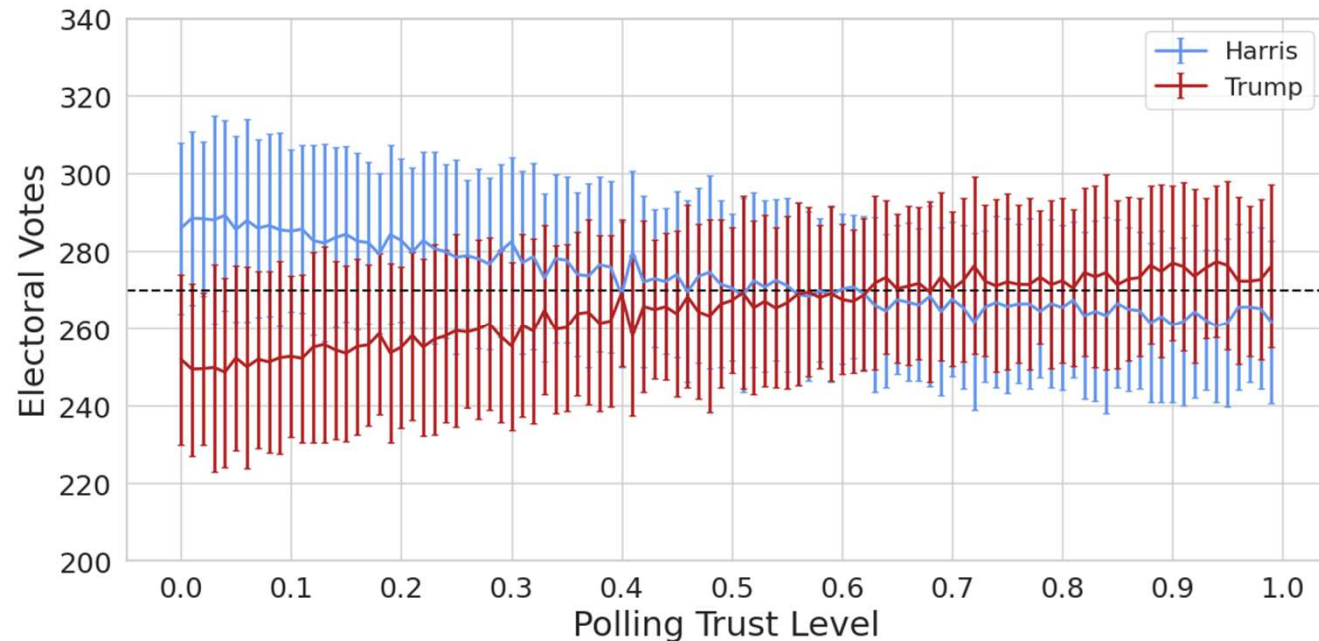




# RESULTS OF MODEL SIMULATIONS

The model conducts multiple simulations with varying weights on historical and polling data to capture possible outcomes under different scenarios:

- **Simulation Details:** Each scenario varies the emphasis on 2020 election results vs. current polling, reflecting the uncertainty of swing states.
- **Scenario Analysis:** Scenarios that heavily weight recent polling data favour candidates currently leading in the polls, while those focusing on historical data predict outcomes closer to 2020 results.



The goal is to explore how different levels of trust in historical voting patterns vs. current polls might affect predictions about who would win the electoral vote count in a future election scenario.



# CONCLUSION AND FUTURE WORK

## Summary:

- Bayesian model effectively combines historical and recent polling data for U.S. election forecasting.
- Swing states remain highly competitive and are crucial to the model's projections.

## Next Steps:

- **Further Refinement:** As new polling data emerges, the model can be re-run to provide updated probabilities.
- **Event-based Polling:** Incorporate adjustments for events such as debates and significant political announcements.
- **Ongoing Visualization:** Map results dynamically as more polls are added.
- **Ensemble Model:** We plan to run an ensemble of models, such as linear regression, random forests, CatBoost, and XGBoost, particularly for swing states. By tuning hyperparameters across these models, we aim to enhance the final prediction of vote share and improve accuracy.

**Final Observation:** Bayesian modeling with a dynamic weighting system provides an adaptable and robust framework for predicting election outcomes. Its reliance on polling adjustments enables real-time reflection of changing voter sentiments. By simulating various scenarios, the study aims to offer a comprehensive view of potential electoral outcomes, highlighting the importance of both historical context and current voter sentiment in shaping election forecasts.

