Forecasting the 2024 U.S. Presidential Election: A Poll-Based Approach*

Kamala Harris Poised for Victory

Xinxiang Gao Ariel Xing John Zhang†

November 2, 2024

This paper presents a forecasting model for the 2024 U.S. Presidential Election, using polling data to analyze voter support for Kamala Harris and Donald Trump. Our approach combines a Baseline Model, which captures national polling trends, with a Primary Model focused on key swing states. Findings suggest a national advantage for Harris, with critical leads in swing states that could secure her a projected 292 electoral votes. This dual-model approach underscores the influence of state-level dynamics on national outcomes, offering insights into how polling data can more accurately capture voter behavior across a diverse electoral landscape.

Table of contents

*Code and data are available at: https://github.com/xgao28/election_forecast.
†The authors are listed in alphabetical order by last name.

1 Introduction

The U.S. Presidential Election is a critical event shaping the country's political direction, often influenced by a complex interplay of public opinion, socio-economic factors, and electoral mechanisms. In recent election cycles, the predictive accuracy of polling has been a topic of significant discussion, as polls not only gauge public sentiment but also influence campaign strategies, media narratives, and voter perceptions. However, challenges such as non-response bias, methodological inconsistencies, and a misalignment between the popular vote and electoral outcomes highlight the need for a robust forecasting model that captures both national and state-specific trends.

This paper aims to address these challenges by developing a polling-based forecasting model for the 2024 U.S. Presidential Election, with a focus on the two main candidates, Kamala Harris and Donald Trump. Using a dual-model approach, we employ a **Baseline Model** to provide a broad national outlook and a **Primary Model** to examine state-level dynamics, especially in key swing states. The Baseline Model aggregates national polling data to predict which candidate would win a majority of popular support, while the Primary Model applies a more nuanced, state-level analysis aligned with the electoral college structure. This combination seeks to address the gap in existing models, which often fail to account for the specific impact of swing states on electoral outcomes, despite their pivotal role in determining the presidency.

Our analysis indicates that while Kamala Harris shows a lead in national polling, as reflected in the Baseline Model, the decisive factors lie within swing states such as Pennsylvania, Michigan, and Wisconsin, where the Primary Model highlights Harris's slight advantages. The state-level predictions from the Primary Model suggest a possible electoral college win for Harris, projecting her to receive 292 electoral votes, above the 270-vote threshold needed to secure the presidency. This finding underscores the unique nature of U.S. elections, where winning the national popular vote does not necessarily translate to an electoral college victory. Our approach, by leveraging poll data with attention to poll quality, transparency, sample size, and geographic variation, seeks to provide a nuanced and accurate forecast that reflects both general sentiment and the regional intricacies of the election.

The structure of this paper is as follows: Section 2 provides an overview of the dataset, detailing the polling data sources, variables, and preprocessing steps. Section 3 outlines the modeling approach, presenting the rationale behind the Baseline and Primary Models and discussing their respective contributions to the analysis. Section 4 presents the results, including national and state-level predictions, followed by Section 5, which explores the implications of these findings, discusses limitations, and suggests future research directions. An appendix provides additional methodological details and diagnostics.

1.1 Estimand

The primary estimand in this study is the predicted percentage of support each candidate will receive on election day, both nationally and within critical swing states. This percentage serves as the foundation for forecasting which candidate is likely to win the election, by capturing trends in public sentiment over time and projecting how these trends may manifest in the final vote count.

2 Data

2.1 Dataset Overview

This study leverages a dataset of polling data for the 2024 U.S. Presidential Election, focusing on predicting support for the two main candidates, Kamala Harris and Donald Trump. Each entry in the dataset represents results from individual polls, capturing both the timing and geographical scope of polling, which is crucial for analyzing national trends and state-specific shifts, particularly in key swing states. By collating data from multiple polling sources, the dataset captures a broad snapshot of voter sentiment across the country, reflecting variations across time, poll methodologies, and geographic regions.

The outcome variable, **Polling Percentage ("pct")**, represents the percentage of respondents in each poll who indicate support for either Harris or Trump. This variable is central to our analysis, as it reflects the changing landscape of voter sentiment and serves as the basis for our forecasts.

2.2 Predictor and Additional Variables

- **Days Toward Election ("days_towards_election")**: The primary predictor in the model, this variable indicates the number of days remaining until the election at the time each poll was conducted. It is essential for capturing time-based trends, enabling the model to adjust for shifts in public opinion as the election approaches. As the election nears, this variable helps illustrate the patterns of support stabilization, often observed closer to election day.

Additional potential predictors include variables reflecting the quality, sample characteristics, and transparency of each poll, which help contextualize the primary predictor and provide an understanding of polling reliability:

- **Poll Quality Score ("pollscore")**: This variable captures the methodological reliability of each poll, including factors such as sampling technique and historical accuracy. Polls with higher scores typically offer more robust estimates, as they reflect the pollster's performance and adherence to reliable sampling practices.

- **Transparency Score ("transparency_score")**: This score measures the level of methodological detail each poll discloses. Higher transparency generally correlates with increased confidence in the data, as it indicates comprehensive disclosure of methods, sample details, and collection practices.
- **Sample Size ("sample_size")**: The number of respondents in each poll is crucial for assessing precision. Larger sample sizes tend to reduce the margin of error, making the estimates more representative of the general population.
- **State ("state")**: This geographic variable distinguishes between national and state-level polls, allowing the model to capture regional variations in voter support. State-specific polling data are particularly valuable for modeling outcomes in swing states, where election outcomes often hinge on narrow margins.

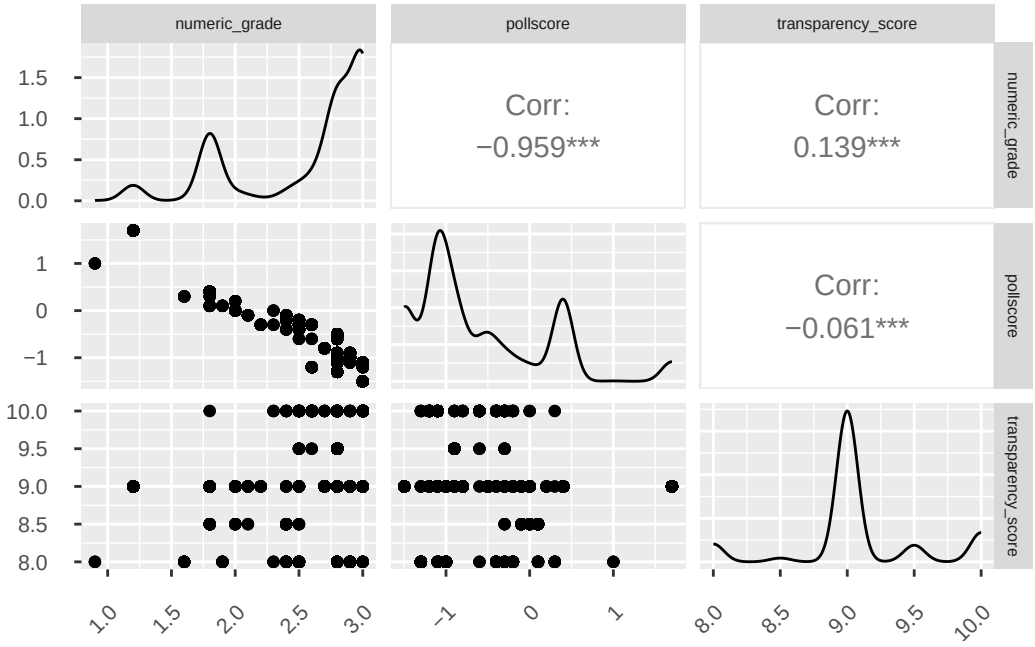
2.3 Measurement and Data Processing

Each variable in the dataset is constructed to accurately reflect polling characteristics across various dimensions of time, geography, and poll quality. Poll quality and transparency scores are derived based on historical pollster performance and disclosure levels, offering an indirect measure of data reliability. The primary predictor, **days toward election**, and the outcome variable, **pct**, are drawn directly from each poll entry. Polling percentages are scaled by sample size, which provides a weighted measure of candidate support, ensuring that polls with larger sample sizes have a proportionally greater impact on model predictions.

To enhance interpretability and prediction accuracy, we adjusted the dataset by scaling polling percentages to reflect each poll's sample size, multiplying average support by sample size, and normalizing by 0.01. This process ensures that larger samples contribute more meaningfully to the model's overall prediction, thus offering a balanced representation of likely voting intentions.

2.4 Exploratory Data Analysis and Summary Statistics

2.4.1 Multicollinearity Considerations



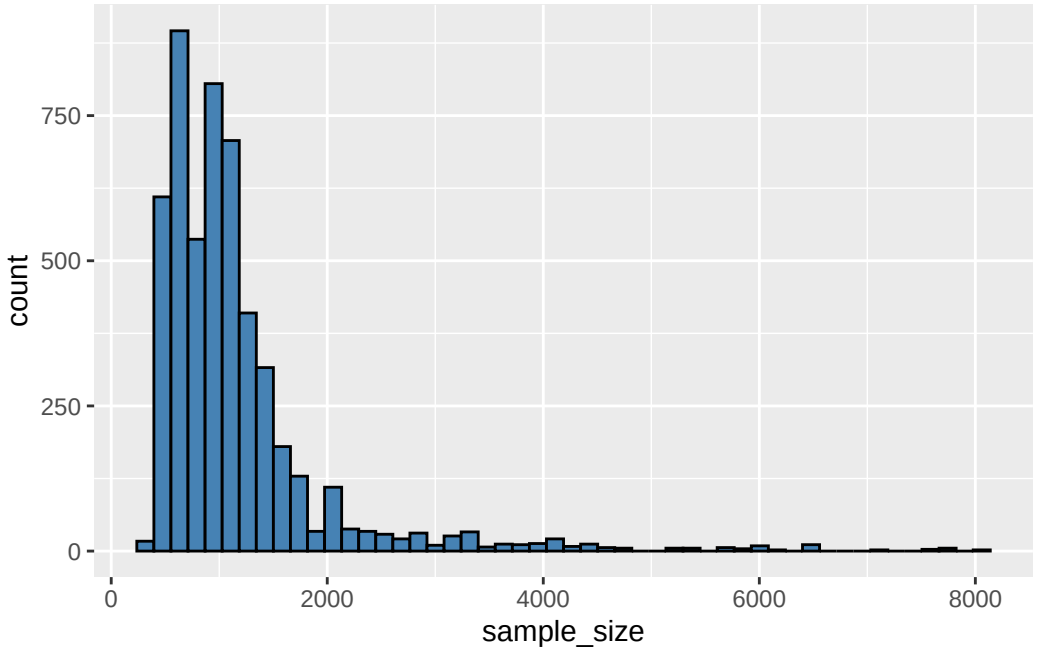
While selecting the predictors for the model, multicollinearity was a key concern. Multicollinearity occurs when two or more predictors are highly correlated, which can inflate the variance of the coefficient estimates and make the model less reliable. In our pair plot analysis, we observed a high correlation between "numeric_grade" and "pollscore," suggesting that they measure similar aspects of polling quality.

```
data_swing <- data %>% filter(state %in% c("Nevada", "Arizona", "Wisconsin", "Michigan", "Pe

data_swing %>%
  group_by(state) %>%
  summarise(count = n(),
            `mean pollscore` = round(mean(pollscore, na.rm = TRUE), 3),
            `mean numeric grade` = round(mean(numeric_grade, na.rm = TRUE), 3),
            `mean transparency score` = round(mean(transparency_score, na.rm = TRUE), 3),
            `mean sample size` = round(mean(sample_size, na.rm = TRUE))) %>%
  kable()
```

state	count	mean pollscore	mean numeric grade	mean transparency score	mean sample size
Arizona	265	-0.649	2.523	9.057	771
Georgia	279	-0.548	2.503	9.032	898
Michigan	318	-0.599	2.566	9.003	789
Nevada	156	-0.658	2.512	9.071	664
North Carolina	273	-0.413	2.401	8.908	836
Pennsylvania	422	-0.684	2.605	8.921	959
Wisconsin	383	-0.857	2.772	9.324	780

```
data %>%
  filter(sample_size < 10000) %>%
  ggplot(aes(x = sample_size)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "black")
```




```
data %>%
  filter(sample_size >= 10000) %>%
  arrange(sample_size)
```

```
  poll_id transparency_score numeric_grade pollscore state sample_size end_date
```

```
1   88104                  9             2       0.2  <NA>       18123   9/4/24
2   88104                  9             2       0.2  <NA>       18123   9/4/24
3   88104                  9             2       0.2  <NA>       20762   9/4/24
4   88104                  9             2       0.2  <NA>       20762   9/4/24
5   88989                 10             3      -1.1  <NA>       48732 10/25/24
6   88989                 10             3      -1.1  <NA>       48732 10/25/24
7   88989                 10             3      -1.1  <NA>       78247 10/25/24
8   88989                 10             3      -1.1  <NA>       78247 10/25/24
  party candidate_name days_towards_election pct
1   DEM  Kamala Harris                    62  44
2   REP   Donald Trump                    62  41
3   DEM  Kamala Harris                    62  39
4   REP   Donald Trump                    62  37
5   DEM  Kamala Harris                    11  51
6   REP   Donald Trump                    11  47
7   DEM  Kamala Harris                    11  51
8   REP   Donald Trump                    11  46
```

To understand the general trends in the data, we conducted exploratory data analysis using summary statistics and visualizations. Figure 1 (below) illustrates the polling percentages for each candidate over time, allowing us to observe fluctuations in support as the election day draws near.

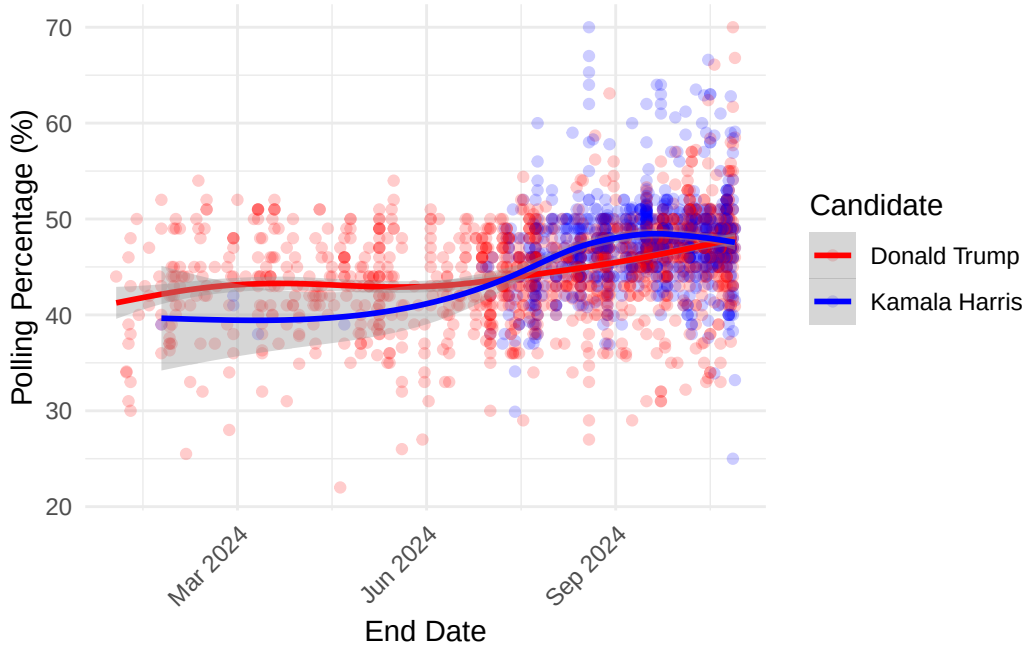


Figure 1: Polling Percentage over Time by Candidate

3 Model

In this study, we employ two main models to forecast the outcome of the 2024 U.S. Presidential Election using polling data for the candidates Kamala Harris and Donald Trump. The models, a Baseline Model (Winner-Take-All National Model) and a Swing State-Based Model, provide two distinct but complementary approaches to assessing candidate support. The Baseline Model aggregates national polling data to project which candidate is likely to win based on national polling trends and assigns all electoral votes to the candidate with the higher projected national percentage on election day. This approach offers a straightforward prediction of overall voter sentiment across the country and captures the aggregate trend of support for each candidate as the election approaches.

On the other hand, the Swing State-Based Model is designed to provide a more granular projection by focusing on polling data at the state level, particularly for key swing states. This model assumes that the winner in each swing state can be determined by state-specific polling trends, and it projects support for each candidate within these competitive regions. By combining outcomes from both swing states and solid states, this model allows for a more nuanced prediction that aligns with the actual mechanics of the Electoral College. Together, these models are well-suited to address both the general national polling trends and the crucial role of swing states in determining the final electoral outcome.

Table 2: Baseline Model

term	estimate	std.error	t.stat	p.value	candidate
(Intercept)	48.189	0.207	232.882	0	Kamala Harris
days_towards_election	-0.012	0.002	-5.119	0	Kamala Harris
(Intercept)	46.010	0.177	260.055	0	Donald Trump
days_towards_election	-0.009	0.001	-10.235	0	Donald Trump

3.1 1. Baseline Model (Winner-Take-All National Model)

The baseline model aggregates national polling data to predict a winner, with all electoral votes awarded to the candidate with the higher national polling percentage. Let: - $P_H(t)$: Harris's national polling percentage at t days towards the election. - $P_T(t)$: Trump's national polling percentage at t days towards the election.

We model each candidate's polling as:

$$P_H(t) = \alpha_H + \beta_H t$$

$$P_T(t) = \alpha_T + \beta_T t$$

where α_H and α_T are the intercepts (predicted percentages on election day, $t = 0$), and β_H and β_T are the slopes.

Then, our aim is the intercept, $P_H(0) = \alpha_H, \quad P_T(0) = \alpha_T$.

We then draw the prediction result as follows: Harris wins (all electoral votes) if $\alpha_H > \alpha_T$, and Trump wins (all electoral votes) otherwise.

Table 3: Primary Model

term	estimate	std.error	t.stat	p.value	state	candidate
(Intercept)	48.798	0.535	91.297	0.000	Arizona	Donald Trump
days_towards_election	-0.018	0.003	-5.298	0.000	Arizona	Donald Trump
(Intercept)	47.097	0.436	108.050	0.000	Arizona	Kamala Harris
days_towards_election	-0.011	0.005	-2.157	0.036	Arizona	Kamala Harris
(Intercept)	48.585	0.435	111.706	0.000	Georgia	Donald Trump
days_towards_election	-0.012	0.002	-5.089	0.000	Georgia	Donald Trump
(Intercept)	47.824	0.478	100.074	0.000	Georgia	Kamala Harris
days_towards_election	-0.014	0.005	-2.604	0.012	Georgia	Kamala Harris
(Intercept)	46.797	0.541	86.472	0.000	Michigan	Donald Trump
days_towards_election	-0.013	0.003	-4.271	0.000	Michigan	Donald Trump
(Intercept)	47.981	0.472	101.674	0.000	Michigan	Kamala Harris
days_towards_election	-0.010	0.005	-1.939	0.058	Michigan	Kamala Harris
(Intercept)	46.312	0.701	66.022	0.000	Nevada	Donald Trump
days_towards_election	0.002	0.004	0.377	0.708	Nevada	Donald Trump
(Intercept)	48.098	0.686	70.162	0.000	Nevada	Kamala Harris
days_towards_election	-0.020	0.006	-3.252	0.003	Nevada	Kamala Harris

Table 3: Primary Model

term	estimate	std.error	t.stat	p.value	state	candidate
(Intercept)	48.118	0.326	147.531	0.000	North Carolina	Donald Trump
days_towards_election	-0.015	0.002	-7.253	0.000	North Carolina	Donald Trump
(Intercept)	48.154	0.490	98.255	0.000	North Carolina	Kamala Harris
days_towards_election	-0.022	0.009	-2.336	0.023	North Carolina	Kamala Harris
(Intercept)	47.215	0.367	128.546	0.000	Pennsylvania	Donald Trump
days_towards_election	-0.015	0.002	-6.703	0.000	Pennsylvania	Donald Trump
(Intercept)	48.844	0.333	146.547	0.000	Pennsylvania	Kamala Harris
days_towards_election	-0.019	0.004	-4.785	0.000	Pennsylvania	Kamala Harris
(Intercept)	46.667	0.461	101.243	0.000	Wisconsin	Donald Trump
days_towards_election	-0.009	0.003	-2.935	0.004	Wisconsin	Donald Trump
(Intercept)	49.305	0.364	135.402	0.000	Wisconsin	Kamala Harris
days_towards_election	-0.010	0.004	-2.414	0.019	Wisconsin	Kamala Harris

3.2 2. Swing State-Based Model

This model forecasts electoral outcomes by projecting state-level polling results for each candidate, combining swing state outcomes with votes from solid states.

1. **Swing State Polling Projections**
   For each swing state $s \in S$, let:

   - $P_{H_s}(t)$: Harris's polling percentage in state s at t days before the election.
   - $P_{T_s}(t)$: Trump's polling percentage in state s at t days before the election.

   We model each swing state polling as:

$$P_{H_s}(t) = \alpha_{H_s} + \beta_{H_s} t$$

$$P_{T_s}(t) = \alpha_{T_s} + \beta_{T_s} t$$

where α_{H_s} and α_{T_s} are intercepts (predicted percentages at $t = 0$) and β_{H_s}, β_{T_s} are slopes.

Then, we obtain the predicted polling percentage for Harris in state s : $P_{H_s}(0) = \alpha_{H_s}$, and that for Trump in state s : $P_{T_s}(0) = \alpha_{T_s}$. The candidate mentioned above wins in state s if their predicted polling percentage is higher than the other candidate.

2. **Electoral Vote Aggregation**
   Based on the electoral vote policy in swing states, the winner of the state obtains all the vote from that state. Obtaining the predicted winner of the swing states, we calculate their total electoral votes as two components: their votes from the solid states as mentioned by Electoral Ventures LLC (2024), and the votes from the swing states that would advocate the candidate based on the predictions. Eventually, the candidate surpassing 270 electoral votes in our prediction wins.

3.3 Model justification

The two-model approach is justified by the need to capture both national and state-level dynamics in election forecasting. The Baseline Model's focus on national polling trends is appropriate for understanding the overall sentiment of the voting population, providing an aggregate view that reflects general support levels for each candidate. Since national support can indicate the broader trajectory of an election, the Winner-Take-All National Model captures this sentiment in a straightforward, interpretable way. The inclusion of a temporal component, represented by days toward the election, allows both models to track changes in support levels as the election day approaches. This time-based element is critical, as voter sentiment often fluctuates and solidifies in the lead-up to an election, especially with the influence of campaign events and media coverage.

The Primary Model, focusing on swing states, is particularly valuable because of the unique structure of the U.S. Electoral College, where specific states can disproportionately influence the final outcome. Including state-specific intercepts and slopes enables this model to account for localized variations in support, reflecting the critical importance of state-level outcomes in swing regions. By incorporating state indicators, this model can address the heterogeneity of voting behavior across different regions, capturing the specific dynamics of battleground states where candidate support may diverge significantly from national trends. Aggregating the electoral votes based on projected winners in both swing and solid states makes this model well-aligned with real-world electoral processes, enhancing its practical utility in forecasting the election outcome.

Additionally, the linear relationship assumed in these models is justified by the often gradual and linear shifts in voter preferences over short periods, especially as the election date nears. State projections are treated independently, consistent with the winner-take-all approach in

most U.S. states. Although linear models are generally effective for capturing large-scale polling trends, alternative approaches, such as Bayesian methods for quantifying uncertainty, could be considered for highly volatile polling scenarios. Nonetheless, the current models' simplicity and interpretability make them particularly suitable for capturing the aggregate and state-level trends necessary for this election forecast, providing a balanced approach that aligns well with the complexities of the U.S. electoral system.

4 Results

The results from our forecasting models provide insight into predicted support levels for Kamala Harris and Donald Trump, drawing from both national and state-level polling data. Each model presents a distinct view of candidate support and projected outcomes based on trends in polling data as election day approaches.

4.1 National Polling Trends and Baseline Model Results

The Baseline Model evaluates national polling data to project which candidate would win all electoral votes based on the higher national polling percentage on election day. The intercept term, representing predicted support on election day, shows that Kamala Harris is expected to receive approximately 48.19% of the national vote, while Donald Trump is projected to receive around 46.01%. This 2.18 percentage point lead for Harris in the Baseline Model suggests a slight advantage for her on a national scale. This model illustrates overall voter sentiment trends, capturing how each candidate's support changes as election day approaches and indicating a national-level lean toward Harris if these trends hold.

Table 4: Baseline Model Result

Kamala Harris	Donald Trump	difference
48.189	46.01	2.179

4.2 Swing State Analysis and Primary Model Results

In contrast, the Primary Model delves into state-specific polling data, focusing on swing states where vote margins are especially narrow and could decisively impact the electoral outcome. Intercept estimates for each candidate across key swing states show a mixed pattern, with Harris projected to lead in states like Wisconsin (49.30% vs. 46.67%) and Pennsylvania (48.84% vs. 47.21%), where her support surpasses Trump's by 2.64 and 1.63 percentage points, respectively. However, Trump maintains a lead in Arizona (48.80% vs. 47.10%) and Georgia (48.59% vs. 47.82%), reflecting the competitive nature of these swing states.

Table 1 provides a summary of intercept estimates for both candidates across swing states, showing how close these races remain. Harris's estimated advantage in states like Wisconsin and Pennsylvania signals potential strategic wins in the electoral college, while Trump's narrow leads in Arizona and Georgia underscore the contested nature of the 2024 election in these battleground areas.

Table 5: Primary Model Result

State	Donald Trump	Kamala Harris	Difference	Votes
Arizona	48.798	47.097	-1.701	6
Georgia	48.585	47.824	-0.761	11
Michigan	46.797	47.981	1.184	10
Nevada	46.312	48.098	1.786	15
North Carolina	48.118	48.154	0.036	19
Pennsylvania	47.215	48.844	1.629	16
Wisconsin	46.667	49.305	2.638	16

4.3 Electoral Vote Projections

Integrating the swing state projections with solid state predictions, the Primary Model estimates the electoral vote totals for each candidate. Kamala Harris is projected to win 292 electoral votes, surpassing the 270-vote threshold required to secure victory. Donald Trump, by comparison, is projected to receive 246 electoral votes. This projection, based on aggregating state-level outcomes, suggests that Harris has a favorable chance of achieving the necessary electoral college majority if current polling trends continue. The model indicates that Harris's projected wins in swing states like Wisconsin, Michigan, and Pennsylvania could be critical to her securing an electoral majority.

These results from both the Baseline and Primary Models provide distinct perspectives on the likely outcome of the 2024 U.S. Presidential Election. The Baseline Model projects a national-level popular vote advantage for Harris, while the Primary Model's state-based approach points to an electoral college path for her, albeit with highly competitive results in key swing states that remain essential to the final outcome.

Table 6: Predicted Electoral Votes for Each Candidate

Candidate	Solid State	Predicted Swing State	Total Predicted Votes
Harris	226	76	302
Trump	219	17	236

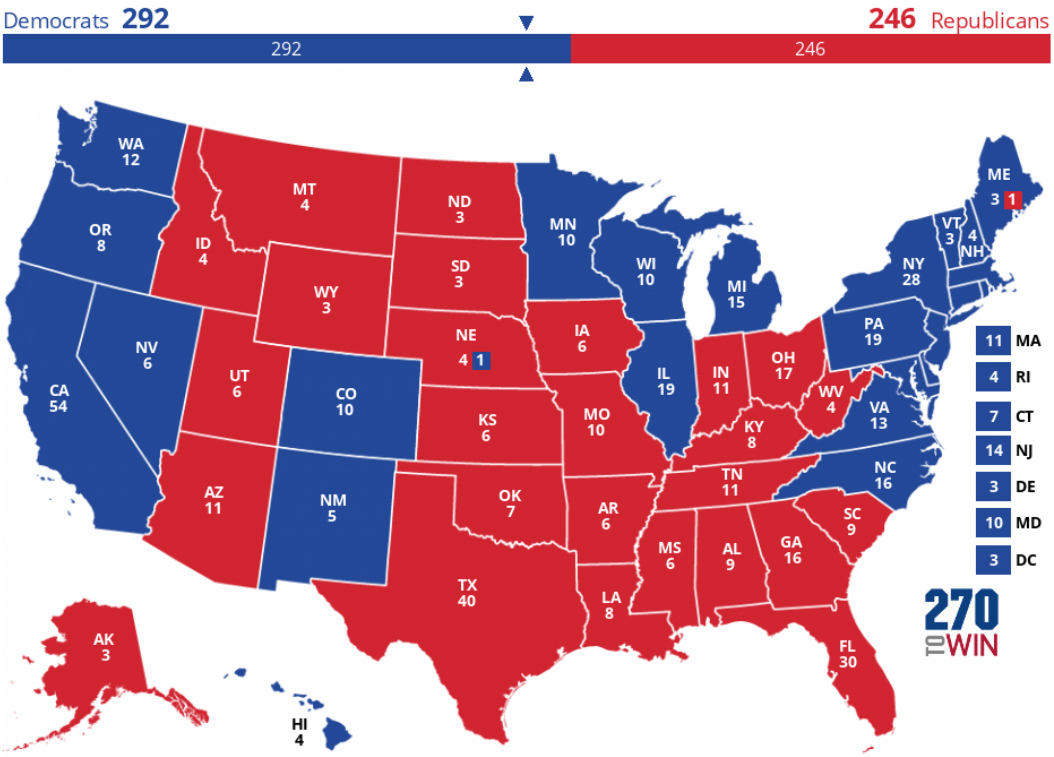


Figure 2: Result

5 Discussion

5.1 Summary of Contributions

This paper presents a polling-based forecasting model for the 2024 U.S. Presidential Election, utilizing two complementary approaches to predict candidate support for Kamala Harris and Donald Trump. The **Baseline Model** (Winner-Take-All National Model) offers a straightforward projection of overall national sentiment by aggregating national polling data, while the **Primary Model** (Swing State-Based Model) provides a more refined, state-specific analysis, focusing especially on swing states crucial to the electoral college outcome. The two models together balance broad sentiment tracking with the targeted detail needed for swing-state-specific electoral forecasting. Through this dual approach, the analysis captures both national-level trends and the granular dynamics within pivotal swing states, offering an informative look into the likely distribution of popular and electoral votes as election day approaches.

5.2 Insights into Election Dynamics

The results reveal critical insights into the nature of U.S. elections and the structure of the electoral college system. One key takeaway from this analysis is that winning the national popular vote does not guarantee a candidate's victory in the U.S. presidential election. The Baseline Model indicates a potential popular vote lead for Kamala Harris; however, the ultimate victory depends on success within specific swing states, as highlighted by the Primary Model. This reflects a unique feature of the U.S. election system, where a candidate can win the presidency without securing the majority of the popular vote by focusing on electoral votes, especially in key battleground regions. Therefore, while Harris's projected national lead is notable, the Primary Model's swing-state analysis underscores the importance of targeting resources and strategies in these competitive areas.

5.3 Comparison of Model Effectiveness

Each model brings distinct strengths and weaknesses, serving different forecasting needs. The **Baseline Model** provides a broad, data-rich perspective, aggregating a large volume of polling data to capture general national sentiment. This simplicity allows for a straightforward, easy-to-understand representation of national trends in support levels. However, this model does not account for the state-by-state dynamics critical to winning the electoral college and may oversimplify the complexities of an election where only the electoral votes, not the popular vote, decide the winner.

In contrast, the **Primary Model** offers more precise predictions by incorporating state-level polling data for swing states, which are decisive in the electoral college outcome. By focusing on these battleground areas, the model provides a nuanced picture that aligns more closely with

the mechanics of the U.S. election system. This model, however, requires more detailed data for each state and is therefore less broadly applicable to a general sense of popular sentiment but is far more insightful for understanding likely outcomes under the electoral college system.

5.4 Limitations and Areas for Improvement

Despite the insights provided by both models, there are limitations to this approach. The models rely on polling data, which is inherently subject to variability and potential biases, such as non-response bias, sampling errors, and the limitations of data collection methods across different polling organizations. Additionally, both models assume a linear trend in polling changes as election day approaches, which may not fully capture sudden shifts in voter sentiment that can occur due to unforeseen events, campaign dynamics, or emerging social issues.

Furthermore, the Baseline Model's national aggregate approach does not account for the unique political landscape within individual states, potentially underestimating the nuances in regional voting behaviors. Meanwhile, the Primary Model, while more granular, requires extensive state-specific data that may not be consistently available or reliable across all swing states.

5.5 Future Directions

Future research could benefit from integrating more sophisticated methods to address the limitations of linear assumptions and polling variability. For instance, incorporating time-series models that allow for non-linear trends could better capture sudden shifts in voter sentiment over time. Additionally, exploring Bayesian models to handle polling uncertainty more effectively may provide a way to quantify the degree of confidence in each state's predicted support levels.

Further improvements could also be achieved by incorporating demographic and socio-economic data to create more robust and representative models of each state's voter base. Finally, validating these models using historical data from previous elections could enhance their robustness and accuracy, offering a clearer picture of how these methods perform under different electoral conditions.

In conclusion, this study demonstrates the importance of integrating both national and state-level data for election forecasting and highlights how U.S. election outcomes hinge on both popular and electoral vote dynamics. By combining the strengths of a broad national model and a targeted swing-state approach, this paper provides a comprehensive framework for anticipating electoral outcomes in a system where popular sentiment and electoral mechanics are often misaligned.

Appendix

A Pollster Methodology Overview and Evaluation: YouGov

YouGov is a widely recognized survey and market research company that operates one of the largest global online panels, providing insights on public opinion across various topics, from politics to consumer behavior. As an online survey platform, YouGov employs a nonprobability sampling methodology, relying on a self-recruited panel and extensive demographic data to approximate representativeness. This methodology offers several advantages, including cost-efficiency, rapid data collection, and the ability to adapt to specific research needs. However, like all methodologies, it comes with its own set of trade-offs, particularly in terms of representativeness and potential selection biases.

This appendix takes a closer look at YouGov's sampling approach, panel recruitment methods, handling of non-response, and overall survey methodology. It also includes a reflection on the user experience, examining how elements like the onboarding process and incentive structure can shape participant engagement and data quality. By analyzing the strengths and weaknesses of YouGov's approach, this paper aims to provide a comprehensive understanding of how YouGov balances methodological rigor with the practical realities of online polling, highlighting both the reliability and the limitations of its data.

A.1 Population, Frame, and Sample

For YouGov surveys, the population often includes U.S. adults or specific subgroups, such as registered voters or other demographic or political segments. The frame from which YouGov draws its samples is its proprietary online panel, consisting of self-recruited individuals who provide extensive demographic information upon joining. This frame is nonprobabilistic and is designed to capture a wide demographic, with adjustments to enhance representativeness through post-survey weighting (YouGov).

The sample is selected from this online panel based on survey requirements. For general surveys representing the U.S. adult population, YouGov applies demographic matching and stratified sampling to recruit individuals whose characteristics resemble the target population. For more targeted surveys, such as those involving younger adults or specific voter segments, YouGov filters panelists based on demographic and political characteristics, ensuring that the survey sample aligns with the study's focus (YouGov).

A.2 Sample Recruitment

YouGov's panelists are recruited through online advertisements, partnerships with various websites, and organic sign-ups. This open online recruitment allows any adult within the

United States with internet access to join the panel, promoting inclusivity in terms of access and accessibility (YouGov). Once part of the panel, participants voluntarily complete surveys for points that can be exchanged for rewards, making monetary compensation a moderate incentive but not the sole motivator for participation. This approach is well-suited to engage a broad and diverse sample, although it does introduce limitations by excluding those without internet access.

Because the recruitment is continuous and participants often stay on the panel long-term, YouGov can maintain and monitor demographic balance, update panelist information over time, and reduce redundancy in questions for returning members. However, relying solely on online recruitment may introduce a bias toward individuals more inclined toward online engagement, who may not fully represent groups with lower internet usage, such as older adults or those from lower-income backgrounds (YouGov).

A.3 Sampling Approach and Trade-offs

YouGov employs a nonprobability sampling approach, meaning that not every individual in the population has an equal chance of selection. Within its panel, YouGov uses stratified sampling, drawing subsets of respondents based on demographic targets (e.g., age, race, gender, and region) to match the sample as closely as possible to the desired population structure.

A.3.1 Trade-offs of the Nonprobability Approach

- **Strengths**: Nonprobability sampling allows YouGov to gather data rapidly and cost-effectively. By using stratified sampling and post-survey weighting, the company aims to approximate the representativeness of probability-based samples while maintaining flexibility in addressing specific research goals and client needs (YouGov).
- **Limitations**: Because nonprobability samples lack random selection, they do not guarantee true representativeness, which can introduce unknown biases. YouGov addresses this partially by applying demographic weights, but nonprobability sampling's inherent limitations mean results should be interpreted as estimates rather than fully precise reflections of the target population (Huffington Post).

A.4 Handling Non-Response

Non-response is a common issue in survey research, particularly with online surveys where participants may decline to participate, partially complete surveys, or provide inconsistent answers. YouGov tackles non-response bias through quality control processes that exclude unreliable responses. These controls involve detecting respondents who complete surveys too quickly, answer inconsistently, or fail attention checks (YouGov).

In addition, YouGov adjusts for non-response by weighting the final sample to reflect the target population's demographic characteristics more accurately. This adjustment helps address imbalances resulting from non-response, enhancing the data's generalizability. However, non-response bias remains a concern, especially as weighting cannot entirely account for individuals not present in the sample or for differences in survey engagement across demographics (YouGov).

A.5 Strengths and Weaknesses of YouGov's Methodology

A.5.1 Strengths

- **Efficient Data Collection**: YouGov's nonprobability online sampling allows for efficient, cost-effective data collection, enabling timely survey responses that can address rapidly changing social and political topics.
- **Targeted Panel Management**: YouGov's proprietary panel enables consistent, direct contact with respondents, allowing for long-term engagement and improved tracking of demographic and opinion changes without needing repetitive data entry.
- **Rigorous Quality Controls**: Extensive quality checks, including IP verification, email authentication, and response validity checks, help YouGov maintain high data integrity. Respondents who fail quality checks are excluded, helping to reduce noise from unreliable data.
- **Adjustments for Representativeness**: While nonprobability sampling has limitations, YouGov partially mitigates these by applying sophisticated weighting based on benchmarks from sources like the U.S. Census. This enhances the sample's demographic alignment with the general population (YouGov).

A.5.2 Weaknesses

- **Nonprobability Sampling Limitations**: The lack of true random sampling introduces potential bias, as certain population segments may be underrepresented. For example, people without internet access or those less inclined toward online engagement are systematically excluded, which could affect representativeness.
- **Potential Selection Bias**: Although incentives are modest, they may still attract individuals who are already more engaged or motivated to participate in surveys, which could influence results. The weighting procedures help but may not fully eliminate biases if certain groups are consistently underrepresented.
- **Limitations of Online-Only Surveys**: Although internet penetration in the U.S. is high, reliance on online-only surveys may still leave out individuals less comfortable with or able to access online technology, skewing results slightly.

- **Questionnaire Design Constraints**: While YouGov's questionnaire design is generally sound, with randomized question orders and multimedia inclusion for clarity, online surveys can limit response depth compared to in-person or phone surveys. Responses are constrained by pre-set answer options, potentially oversimplifying complex views and reducing respondent nuance (Huffington Post).

A.6 Reflection

We created an account on YouGov to gain real experience with the platform, as illustrated in Figure 3 and here is my reflection on the onboarding process and user experience.

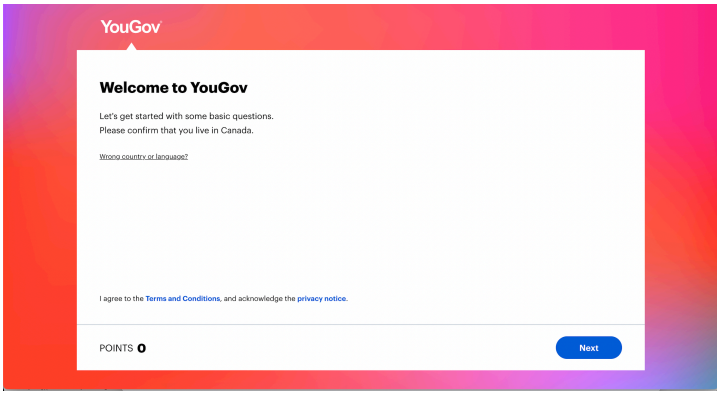


Figure 3: YouGov Welcome Page

The YouGov survey onboarding process is designed to be clear and straightforward, making it easy for new users to sign up and begin participating. The questions asked during onboarding are simple and direct, covering basic demographic details without overwhelming the participant. The first question of the first survey we took at YouGov is shown at Figure 4. This streamlined approach likely helps increase engagement and reduces drop-off rates, ensuring that users can quickly understand what is required of them. The layout and visuals are accessible, contributing to a positive first impression and making it more likely that new panelists will continue engaging with future surveys.

To enhance data integrity, YouGov includes verification steps such as confirming users' email addresses through a code. This extra step not only ensures that participants are real and unique individuals but also helps prevent fraudulent or duplicate accounts, as shown in Figure 5, which can otherwise skew survey results. By verifying email addresses, YouGov can maintain a higher quality of data and build a more reliable panel, knowing that each participant is committed enough to complete these security measures. This verification process is a straightforward yet effective method for improving data quality from the outset.

Data privacy is also well-considered in YouGov's onboarding, with clear options allowing users to control how their information is shared. For instance, participants can opt in or out of having

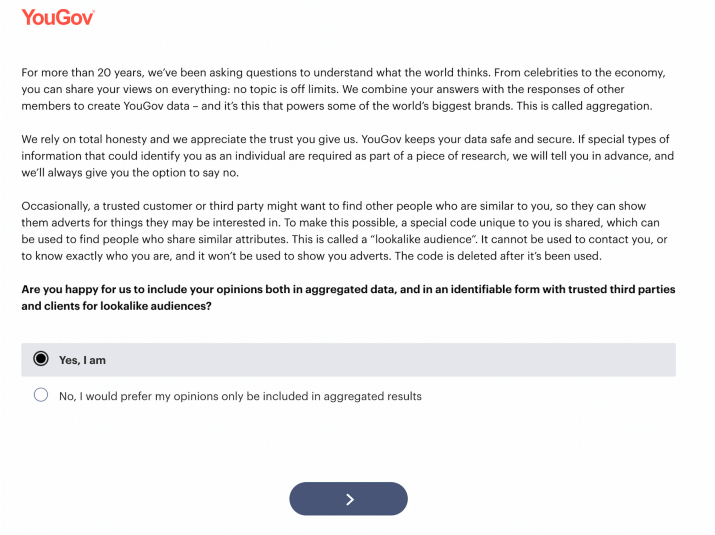


Figure 4: YouGov Survey

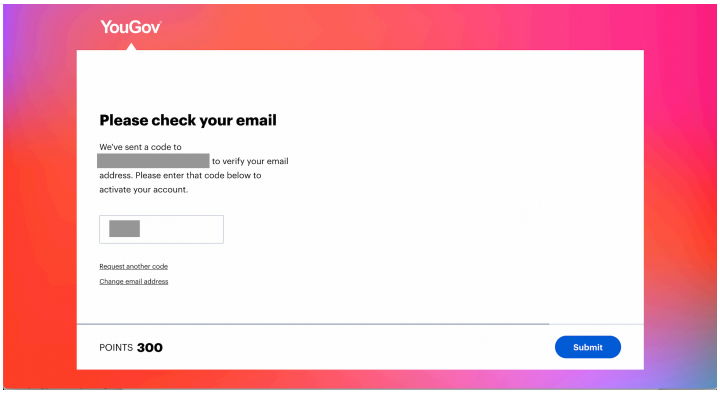


Figure 5: YouGov Email Verification

their responses shared in identifiable formats with trusted third parties. Figure 6 includes an example of how YouGov is letting users to control their information privacy. This transparency around data usage builds trust, showing participants that their privacy is valued and that they have agency over their data. The option to decide on data sharing likely encourages long-term participation, as users are reassured that their information will be handled responsibly.

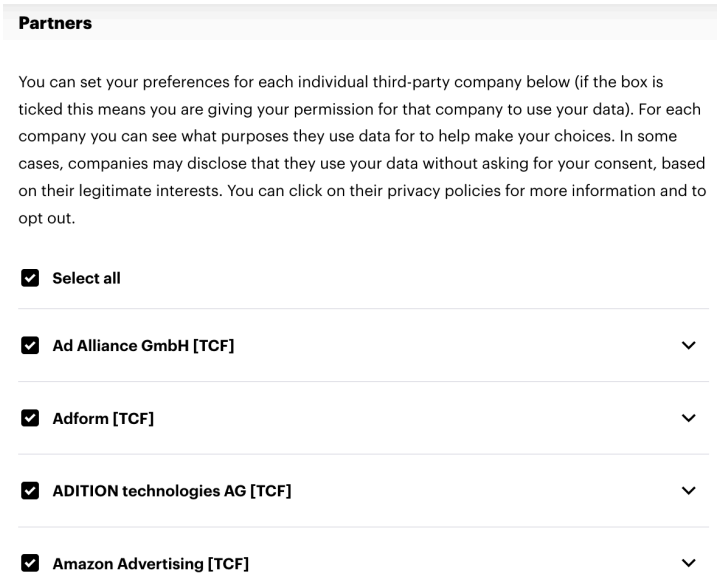


Figure 6: YouGov Data Privacy

However, the onboarding interface in Figure 7 has a gamified feel, awarding points for each step completed. This reward system, while motivating, could encourage participants to treat the surveys more like a task for monetary gain than an opportunity to provide genuine responses. The point-based incentive structure may attract users who are primarily interested in earning rewards, which risks impacting data quality if participants prioritize speed over thoughtful answers. This setup introduces the possibility that some users are engaging more for the rewards than for contributing to research, which could affect the reliability of the responses YouGov collects.

B Methodology and Survey Design for 2024 U.S. Presidential Election Forecast

B.1 Methodology

This part of the appendix provides a detailed methodology and survey design to forecast the 2024 U.S. presidential election. Our survey focuses on swing states: Arizona, Georgia, Michi-

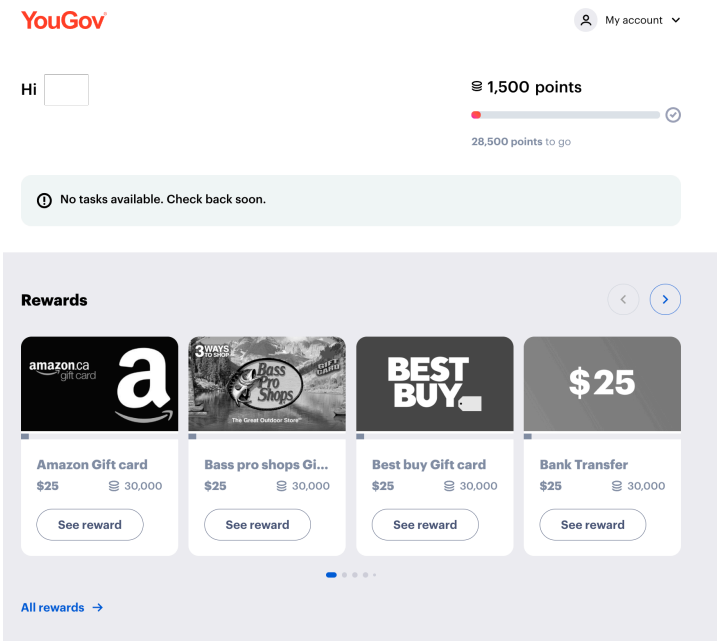


Figure 7: YouGov Interface

gan, Nevada, North Carolina, Pennsylvania, and Wisconsin due to the results of other states being certain. The survey targets different age groups using tailored recruitment strategies. Below, we break down the methodology, budget, recruitment strategies, and survey design to ensure representative participation.

B.1.1 Sampling Approach

- Total Sample Size: 7,000 respondents (1,000 participants per state).
- Target Population: Eligible voters from the following swing states: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, Wisconsin.

B.1.1.1 Sampling Breakdown by Age Group:

- 18-24 years: 150 participants per state (total 1,050) – reached through university students.
- 25-60 years: 700 participants across states (total 4,900) – reached via online ads and survey platforms.
- 60+ years: 150 participants per state (total 1,050) – reached through phone calls.
- Stratified Random Sampling:

  - Stratify by age, gender, race/ethnicity, education, income, and region.
  - Ensure proportional representation based on **U.S. Census data** to avoid bias.

- Weighting Strategy:
  - Apply **post-stratification weighting** to adjust for any sampling imbalance, aligning with national demographics.

B.1.2 Recruitment Plan

B.1.2.1 18-24 Years (University Students):

- University Collaboration: Partner with academic institutions in each swing state to send survey invitations via email on behalf of the project.
- Rationale: Universities provide direct access to students, improving recruitment efficiency and response rates.
- Incentive: Each respondent receives a $10 e-gift card for a Walmart or Costco purchase.

B.1.2.2 25-60 Years (General Population):

- Online Ads and Survey Platforms: Use Facebook, Instagram, and platforms like Prolific to recruit participants from swing states.
- Budget Allocation: Allocate $10,000 for online ads to reach residents aged 25-60.
- Incentive: Participants receive e-gift cards upon survey completion.

B.1.2.3 60+ Years (Senior Citizens):

- Phone Surveys: Conduct live calls to reach seniors, considering that older people are not comfortable using electronic devices. Phone numbers will be sourced from voter registration databases (where legally accessible) or senior community networks.
- Incentive: Seniors can choose a physical Walmart or Costco gift card, which will be sent by mail.
- Recruitment Channels:
  - Complement with **social media ads** on Facebook, Instagram, and LinkedIn to reach underrepresented groups.
  - **Incentives**: Offer **$10 gift cards** to increase participation and engagement.

B.1.3 Trade-off

B.1.3.1 Data Representativeness vs. Control over Data Collection

- Issue: Although age quotas ensure participation from different age groups (18-24, 25-60, 60+), it is not feasible to set strict limits for race and income. Without perfectly matched samples for race and income, bias may arise in the survey results.
- Solution:
  - Weighting Strategy: Calculate population proportions for race and income within each state using Census data. Assign weights to groups based on their representation to mitigate bias.

B.1.3.2 Efficient Recruitment vs. Potential Selection Bias

- Issue: Limiting the 18-24 age group to university students may introduce selection bias. University students often differ from their non-student peers in socioeconomic background and possibly political perspectives.
- Impact: This selection bias may affect the generalizability of findings, as university students' perspectives might differ from all young adults.

B.1.3.3 Effective Senior Outreach vs. Higher Costs and Response Bias

- Issue: Using live phone operators to reach seniors increases costs, and some seniors may remain unreachable by phone.
- Impact: Higher costs may reduce resources for other age groups. Response bias may arise if unreachable seniors have different views, affecting the survey's representativeness.

B.1.3.4 Incentivizing Participation vs. Economic Bias in Sample

- Issue: Offering $10 gift cards as an incentive could attract participants more likely to need the incentive, potentially introducing economic bias.
- Impact: Overrepresentation of lower-income participants could skew results, affecting insights if income level correlates with political perspectives.

B.1.4 Survey Implementation and Design

In designing this survey, I referenced insights from the article on survey methods from (**AnnualReviewofEconomics?**), particularly on the advantages of online surveys over in-person, phone, or mail surveys. The article highlighted that online surveys provide flexibility, allowing respondents to complete them at their convenience, which minimizes selection bias related to work schedules or availability. This flexibility is especially beneficial for students and working-age individuals, who may find it challenging to respond during traditional hours. The article also emphasized how mobile technology can increase participation by reaching populations that are otherwise hard to engage, such as younger respondents, frequent movers, or those in remote areas. In line with these insights, I designed this survey with an online format for younger and working-age participants, maximizing accessibility and reducing drop-out rates. Additionally, I applied the article's suggestion of keeping surveys concise to prevent fatigue and disengagement, which helps maintain completion rates and improve data quality. The use of digital incentives, as noted in the article, also broadens the appeal and ensures we reach a diverse participant base across income levels, providing flexibility in rewards to cater to different motivations. This approach helped shape a balanced methodology that optimizes participation across demographic groups and aligns with best practices in survey design.

- Survey Platform: Google Forms
  - **Link**: https://forms.gle/2MGYeZavDsCNuWZ1A
  - Accessible to respondents via **email, social media, and direct recruitment channels**.
- Survey Structure:
  1. Demographics Section: Age, gender, race/ethnicity, income, education, and state of residence.
  2. Voting Preferences Section: Candidate choice, likelihood of voting, and party affiliation.
  3. Key Issues Section: Identify priority issues (e.g., economy, healthcare, immigration).
  4. Thank You Message:
     - *"Thank you for completing the survey! Your input is greatly appreciated and will help provide insights into the upcoming 2024 election."*

B.1.5 Data Validation

- Techniques for Data Quality:
  1. Screening Questions: Confirm eligibility (e.g., 18+ years old, registered voter status).

2. Attention Checks: Include a question like *"Please select 'Agree' for this item"* to verify respondents are attentive.
3. IP Geolocation: Validate state residency based on reported location.
4. Duplicate Detection: Identify and remove duplicate responses.

B.1.6 Poll Aggregation and Reporting

- Poll Aggregation:
  - Use **weighted averages** to account for differences in sample size and demographics.
- Margin of Error**:
  - National Margin of Error: ±1% at the 95% confidence level.
  - State-Level Margins: ±5-10% depending on the sample size for each state.

B.1.7 Budget Allocation

Expense	**Estimated Cost**
Participant Incentives	$70,000
Online Ads and Panel Provider Fees	$10,000
Phone Survey Salaries	$8,000
Google Forms (Platform)	Free
Data Validation & Analysis	$10,000
Miscellaneous Expenses	$2,000
Total	**$100,000**

B.2 Survey Questions

Below is the full content of the survey to be implemented using Google Forms:

1. What is your age?
   - 18-24
   - 25-39
   - 40-60
   - 60+

2. What is your gender?
    - Male
    - Female
    - Other: _____
3. What is your race/ethnicity?
    - White
    - Black
    - Hispanic or Latino
    - Asian
    - Indigenous
    - Other: _____
4. What is your highest level of education?
    - Less than high school
    - High school diploma
    - Some college
    - Bachelor's degree
    - Graduate degree or higher
5. What is your annual household income?
    - Less than $25,000
    - $25,000 - $49,999
    - $50,000 - $99,999
    - $100,000 or more
6. Which state do you currently reside in?

7. Are you a registered voter?
    - Yes
    - No
    - Not sure
8. How likely are you to vote in the 2024 presidential election?
    - 1 (Definitely will not vote)
    - 2
    - 3
    - 4
    - 5 (Definitely will vote)
9. If the 2024 election were held today, who would you vote for?
    - Kamala Harris (Democrat)
    - Donald Trump (Republican)
    - Undecided
    - Other: _____
10. What is the most important issue for you in this election?

- The economy
- Healthcare
- Immigration
- Climate change
- Social Security and Medicare
- Foreign policy

11. In your opinion, how will the majority in your state vote in 2024?

- Democrat
- Republican
- Too close to predict

12. Do you have any additional comments or suggestions?

Thank you for completing the survey! Your input is greatly appreciated and will help provide insights into the upcoming 2024 election.

B.3 Google Forms Link

https://forms.gle/2MGYeZavDsCNuWZ1A

C References

Electoral Ventures LLC, (2024). 270toWin - 2024 Presidential Election Interactive Map. 270toWin.com. https://www.270towin.com/