Will Kamala Harris will win the 2024 Presidential election*

Analysing the polls using baysian modeling

Yan Mezhiborsky Cristina Burca Shreya Sakura Noskor

November 4, 2024

Abstract

We analyse the probability of Kamala Harris winning in the 7 swing states of USA using baysian modeling; Arizona, Michigan, Pennsylvania, Nevada, Georgia, Wisconsin, North Carolina. USA is a large country with its national economy affecting the global economic conditions, which is why predicting the future President of USA will help understand the future economic condiotion of the world. Through baysian modelling we found that North Carolina (47.26% support for Harris), Nevada(46.43% support for Harris), Wisconsin(48.38% support for Harris), Michigan(47.30% support for Harris), Pennsylvania(48.02% support for Harris) will vote Kamala Harris, while the rest of the swing states might vote Donald Trump, meaning that majority of the swing states will vote Kamala Harris. Provided this we predict that Kamal Harris will win the election. By creating a baysian model based on "polls of polls" where we compare the results from different polls, we are able to make this prediction.

Table of contents

*Code and data are available at: https://github.com/Mezhi18/US_Election2024.git.

1 Introduction

The 2024 United States presidential election represents a pivotal moment in the country's political landscape. As in previous elections, swing states are projected to play a critical role in determining the outcome. Swing states, which are characterized by their shifting voting patterns and balanced support for both major political parties, have historically been the focus of intense campaign efforts and polling analyses. Understanding the dynamics and voter preferences in these states is crucial to gaining insight into the broader electoral trends that could shape the nation's future.

In this paper we aim to analyze polling data related to the 2024 election, with a particular emphasis on the swing states. By examining various polls and identifying patterns in voter sentiment, we seek to uncover the factors that may influence voter behavior in these highly contested regions. Our analysis will explore demographic shifts, the impact of key issues, and the level of voter engagement across different swing states. Through a comprehensive statistical approach, we aim to contribute to the understanding of the evolving electoral landscape and provide meaningful insights into the forces shaping the 2024 presidential election. We use Bayesian regression model to create a model to predict the percent of population that will support in the swing states based on pollster, the number of days since Kamala Harris joined the race. More on the model and the variables, in the Data and Model sections.

We noticed in most predictions of the presidential candidate they are usually based on one survey or pollster and they try to gather data from the one poll to predict the the future President. However, through looking at the results from different polls we are able to get a better, more accurate prediction. We also look at over state polls which allows us to focus on the swing states. Swing states are states that change sides every election. The 2 major parties that fight for power are the Republicans and the Democrats. Some states are always voting for one party regardless of the candidate so, the swing states are the once that tell us more about who might win the election. The gap we are trying to cover is predicting the future President of USA by looking at different polls, while being as effecient as possible by looking only at the swing states.

To go a little more in-depth into what we actually did, we have to first discuss our estimand. Our estimand, the parameter we aim to study, is the predicted support for Kamala Harris in the swing states (measured in %). We do this by creating Baysian model with 3 parameters; state, pollster, and says since Kamala

Harris was announced as candidate. We have the priors set with a normal distribution centered at 0, with a standard deviation of 2.5. We expect that because the swing states are states that don't have a party that they dominantly vote for, the expected support will be around 50%. We also make other graphs in the Data section, which helps us understand the values themselves. From the result section we find that the average predicted support for Kamala Harris for the 7 swing states were as follows: Arizona at 46.26%, Nevada at 46.43%, Georgia at 46.59%, North Carolina at 47.26%, Michigan at 47.30%, Pennsylvania at 48.02%, Wisconsin at 48.38%. Based on this we predict that Arizona, and Georgia will vote Donald trump and the rest of the states will vote Kamala Harris.

Predicting who will become president is important because it helps people understand and prepare for future changes in the country's policies and priorities. A new president can shape laws, affect the economy, influence international relationships, and address key issues like healthcare, education, and climate change. Accurate predictions allow citizens, businesses, and governments to plan and adapt. For example, if polls indicate a particular candidate might win, groups can start thinking about how their plans might align with the candidate's proposed policies. It also helps keep people informed and engaged in the political process, making them more likely to vote and participate in democracy.

The paper is structured with the introduction in the beginning followed by the Data section wich takes a look at the methodology of the paper and how the data was gathered. We explain the variables we isolated in order for this analysis to be complete. Then it is followed by the Model section where we introduce the model and explain the graphs of related models. The result sections will cover the model summary and summary table of the predictions for each of the swing states. The last section is the discussion section where we talk about what the results means and we give our final conclusion and weaknesses. We also give a final word about who will win the upcoming election. We also have additional appendices, where we talk about the methodology of the polls and how they compare. The second appendix talks about a survey we designed with reasons and disadvantage. The third appendix just shows the models in more depth.

2 Data

2.1 Overview

As our paper is about the 2024 United States federal election and more specifically we are looking at the polling Data and the polling data comparing the two candidates, former President Donald Trump and Vice President Kamala Harris, for the upcoming election. Our original Data set had over 16,000 unique entries from different pollsters, the business or Organization that conducts the poll. Each poll has two entries, one giving the Data for the polling opinions of Donald Trump and the second for Kamala Harris. We have acquired our polling data from FiveThirtyEight (2024)

As there are over 50 variables many of which are redundant to our paper we will only discuss those that we have kept in our clean data as they are the only ones we use in our analysis.

- `pollster`: Shows name of the Pollster that conducted the poll.
- `sample_size`: The number of people that participated in the specific poll.
- `state`: This variable tells us in which States the poll was conducted.
- `candidate_name`: This is the full name of the selected candidate.
- `pct`: This tells us the percentage of participants that intend to vote for the selected candidate.
- `end_date`: The date the pollster finished conducting the poll.

The variables that we have create are:

- `num_harris`: The number of participants that intend to vote for Kamala Harris.
- `end_date_num`: The number of days since the first poll since Harris announced her candidacy, calculated using `end_date`.

Each pollster has a numeric grade from 1.0 to 3.0, which indicates the quality/ reliability of the respective pollster. Additionally, each pollster is also given a transparency score from 1.0 to 10.0 reflecting how 'transparent' the pollster is, or how much information is disclosed about its polls and methodology. It is important for pollsters to maintain high numeric grades and transparency scores because these metrics directly reflect the quality and reliability of their data. To ensure the highest level of accuracy in our predictions, we only include polls with a numeric grade of 1.5 or above and a transparency score of 6.0 or above.

¿ does this go into data cleaning ?

We use the statistical programming language R (R Core Team 2023). Our data comes from (FiveThirtyEight 2024) and was cleaned, modeled and graphed, using Robinson, Hayes, and Couch (2023), Wickham et al. (2019), Goodrich et al. (2022), Firke (2023), Grolemund and Wickham (2011), Arel-Bundock (2022).

2.2 Methodology and Measurement

This paper employs a "polls-of-polls" approach to analyze and predict our outcome, combining data from multiple pollsters. By integrating results from various sources, this method captures a broader range of perspectives, reducing the influence of individual biases and offering a more balanced view. Unlike relying on a single poll, which may be shaped by its unique methodological biases, the "polls-of-polls" approach enhances reliability and increases the overall validity of our predictions.

Each pollster uses methodologies they believe are most effective for capturing an accurate snapshot of public opinion. Pollsters begin by selecting a sample of voters they consider representative of the broader population, factoring in various demographics such as age, region, and socio-economic background. However, polling methods differ significantly; some pollsters rely on phone interviews, while others use online surveys, text messaging, or a combination of methods (mixed-mode). After data collection, pollsters apply weighting adjustments to account for demographic imbalances, ensuring that underrepresented groups are adequately reflected in the results.

The diversity in methodologies helps reduce systematic bias compared to using a single poll source. Different methods may also be more effective in reaching certain demographics—such as older voters through phone surveys and younger voters through online platforms—broadening the inclusivity and representativeness of the sample.

n predictive modeling, accounting for pollster variability and methodology is essential to avoid overfitting to any single source. By including pollster as a predictor in the model, we allow it to recognize that poll results may vary not only due to real changes in public opinion but also because of differences in survey methodologies. This consideration enables the model to generate more accurate, generalized predictions, offering a reliable forecast that reflects a well-rounded view of public opinion

2.3 Data Visualization

In this analysis, we examine polling trends for Kamala Harris across key swing states, focusing on recent polling data from Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin [CITE NYT PAPER]. As shown in Figure 1, which displays the support percentages for both Harris and Donald Trump side-by-side by state, the data reveals a competitive race, with relatively close support levels for both candidates in most states. While Trump has a slight edge in Arizona, Georgia, and North Carolina, Harris maintains a modest lead in Michigan, Nevada, and Pennsylvania.

Moving to Figure 2, which isolates Harris's support percentages across states, we gain insight into the consistency and variability of her support. The box plot highlights states like Wisconsin and Pennsylvania, where the smaller interquartile ranges indicate less variability, suggesting that Harris's support is more stable in these regions. Conversely, states such as Nevada and Georgia show greater variability, as reflected in the larger spread of the boxes and whiskers, indicating that Harris's support fluctuates more in these areas.

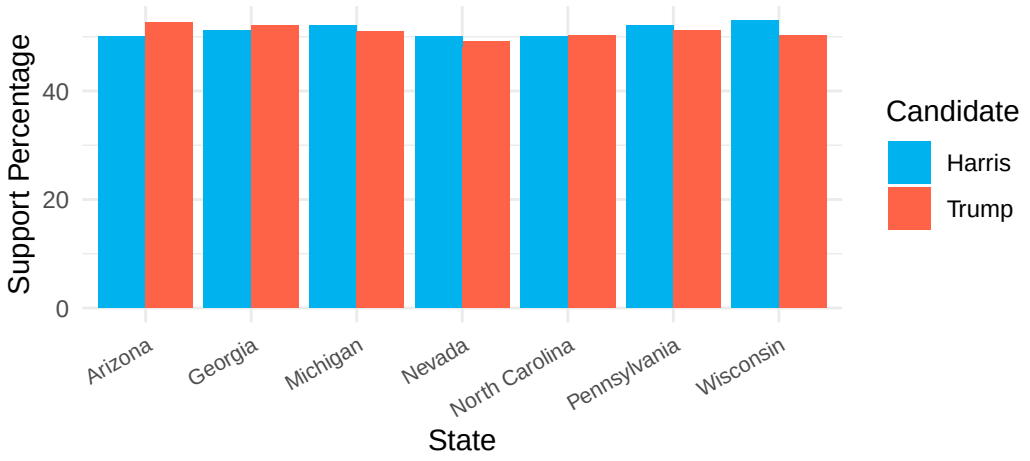

Figure 1: Distribution of Support Percentage by State

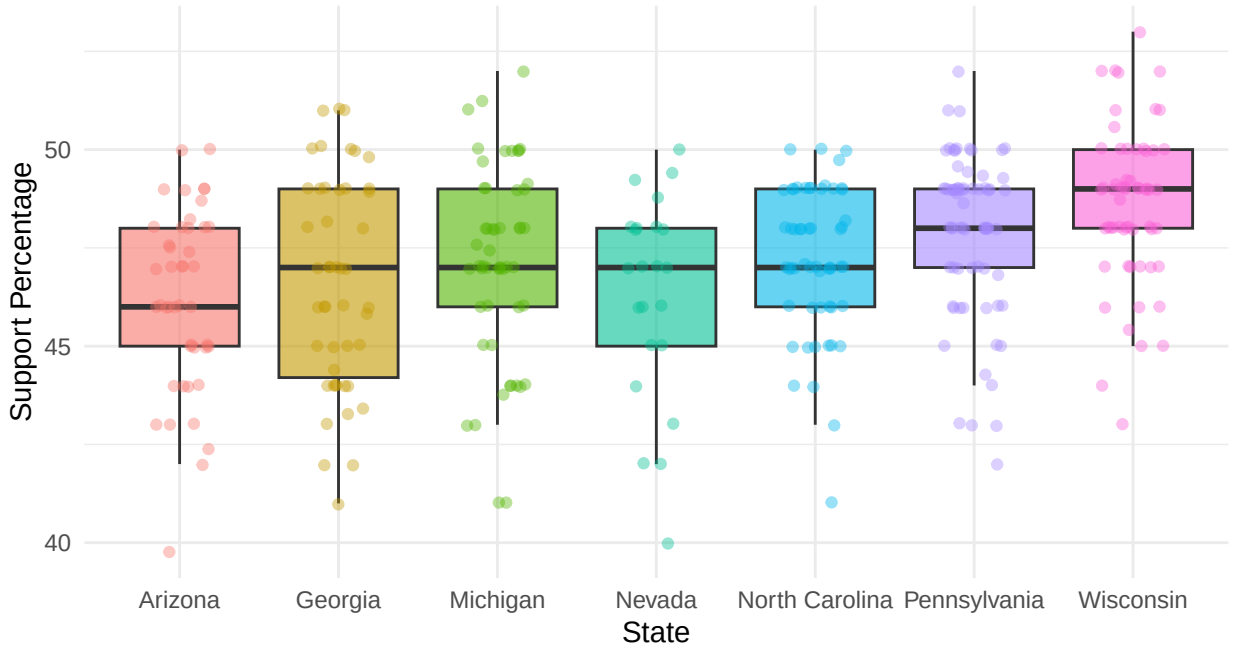

Figure 2: Support Percentage for Harris by State

To explore temporal trends, we introduce the variable `end_date_num`, representing the number of days since Harris declared her candidacy. The next figures examine support trends over time in individual states.

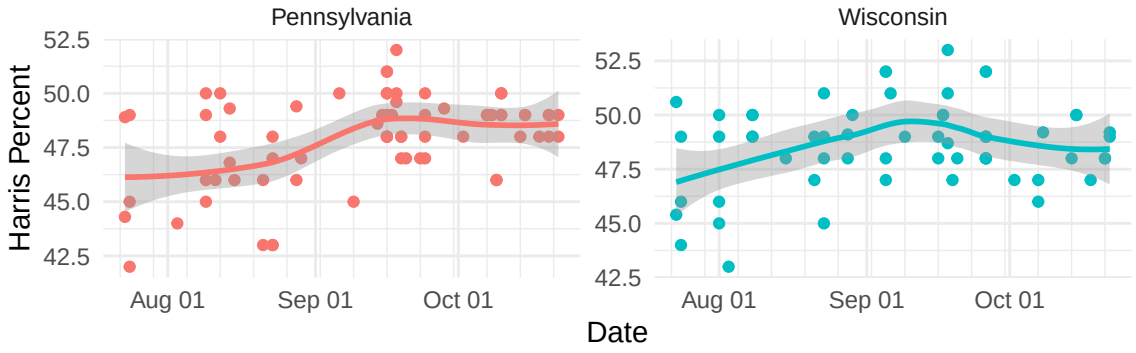

Figure 3: Support Percentage for Harris by State since Harris declared as Candidate for Wisconsin and Pennsylvania

In Figure 3, Pennsylvania and Wisconsin display distinct trends. In Pennsylvania, support for Harris has been relatively consistent since early September, stabilizing around 48.5%. In contrast, Wisconsin shows a peak in support in mid-September, followed by a gradual decline that stabilizes in October. These observations suggest that while Pennsylvania remains a consistent stronghold for Harris, Wisconsin's voter sentiment may be more susceptible to fluctuations.

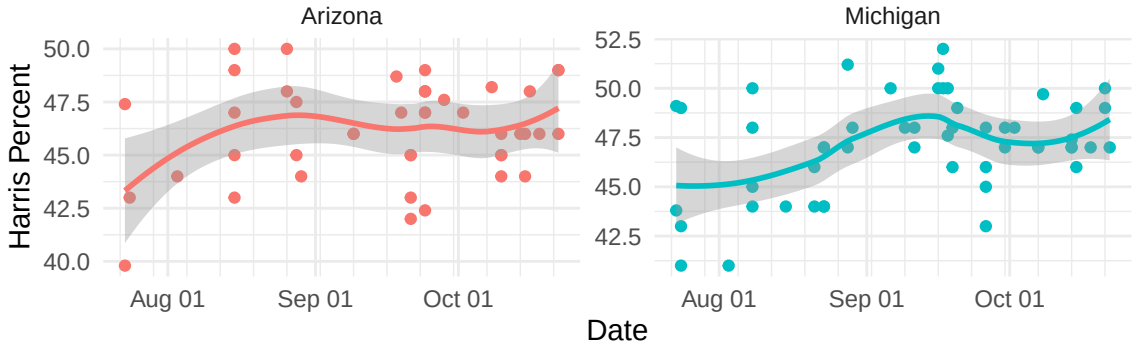

Figure 4: Support Percentage for Harris by State since Harris declared as Candidate

Figure 4 further examines Harris's support trends in Arizona and Michigan. Both states exhibit an upward trend in support starting in early October, which may suggest increasing favorability or strategic campaign efforts in these regions. This upward trajectory could indicate an opportunity for Harris to solidify her support in these competitive areas.

In Figure 5, we analyze Georgia, Nevada, and North Carolina. Nevada demonstrates the highest variability in Harris's support, which could reflect the state's dynamic political response to recent campaign efforts. Georgia shows a steady upward trend, indicating growing support for Harris, while North Carolina reveals a slight decline since October, suggesting a possible shift in voter support.

In Figure 6, we examine the impact of pollster variability on Harris's support percentage over time. This visualization introduces the variable pollster, highlighting the differences in polling results that arise from varying methodologies. The chart reveals considerable fluctuations and occasional outliers in Harris' support, likely due to the distinct approaches each pollster employs in sampling and weighting responses.

Overall, these visualizations illustrate that Harris's voter support is not only state-dependent but also fluctuates significantly over time. In particular, many states exhibit notable shifts around mid-September, pointing

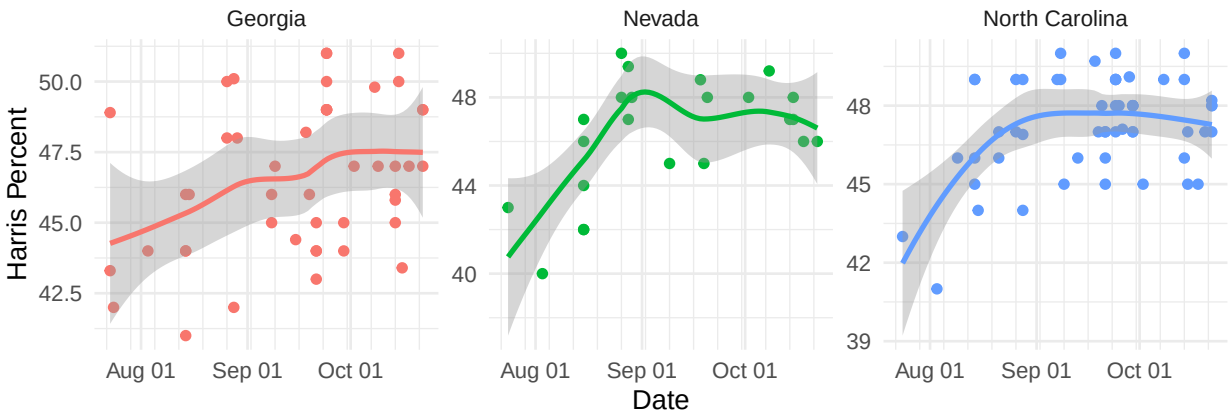

Figure 5: Support Percentage for Harris by State since Harris declared as Candidate

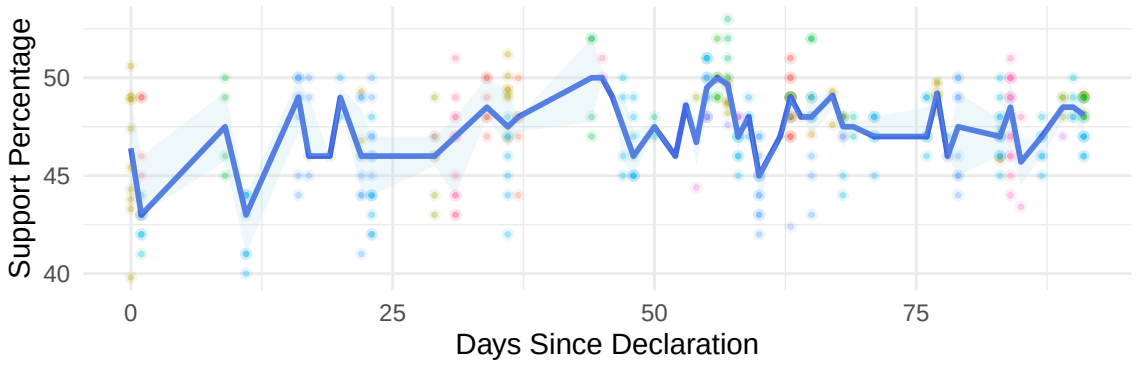

Figure 6: Variation in Harris's Support Percentage Across Pollsters Over Time

to potential influences from external events or campaign dynamics during this period. Furthermore, the observed variability across pollsters emphasizes the necessity of a robust model that can account for differences in pollster methodologies. By incorporating temporal patterns, state-specific characteristics, and pollster variability, our predictive model will aim to provide a more comprehensive and nuanced forecast of support for Harris in key regions, helping to isolate genuine shifts in public opinion from noise introduced by polling differences.

2.4 Outcome and Predictor variables

In this analysis, our primary outcome variable is the support percentage for Kamala Harris. This variable represents the proportion of respondents in each poll who indicated support for Harris. Given the competitive nature of the election, understanding how this support varies across states and over time is crucial for identifying trends and forecasting her overall performance.

In this analysis, our primary outcome variable is the support percentage for Kamala Harris. This variable represents the proportion of respondents in each poll who indicated support for Harris. Given the competitive nature of the election, understanding how this support varies across states and over time is crucial for identifying trends and forecasting her overall performance.

To explain the outcome, we consider several predictor variables:

1. **State**: Since support for candidates can vary significantly across different states due to regional demographics, political history, and local issues, the state variable allows us to capture these geographical differences. As shown in the earlier visualizations, some states demonstrate consistently higher or lower support for Harris, emphasizing the importance of including state-level distinctions in our analysis.
2. **Days Since Declaration**: Representing the number of days since Harris declared her candidacy, this temporal variable helps us capture shifts in voter sentiment over time. Our analysis reveals several critical points where support trends noticeably change, such as the upward or downward shifts in mid-September. This predictor allows the model to account for these temporal patterns and track how Harris's support evolves as the campaign progresses.
3. **Pollster**: Given that different pollsters employ varied methodologies, sampling techniques, and weighting schemes, the pollster variable captures the potential variability introduced by each polling organization. As illustrated in Figure 6, the results can vary considerably depending on the pollster, with some reporting notably higher or lower support percentages for Harris. By including pollster as a predictor, we aim to control for this source of variability, helping the model focus on underlying trends rather than discrepancies introduced by polling differences.

Each of these predictors provides a distinct layer of insight into Harris's support trends. Together, they form the foundation of our predictive model, which will use these variables to more accurately forecast shifts in voter support and clarify the impact of state, time, and pollster methodology on public opinion.

3 Model

3.1 Model Step- Up

We used the `stan_glmer` function from the `rstanarm` package to create a Bayesian regression model with a binomial family and logit link. This model is designed to predict the proportion of respondents who support Kamala Harris based on several key factors.

The model specification is as follows:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \beta_0 + \beta_1 \times \text{State}_i + \beta_2 \times \text{Pollster}_i + \beta_3 \times \text{Days since first poll}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$
$$\sigma \sim \text{Exponential}(1) \tag{7}$$

The outcome variable in this model is the proportion of respondents who support Harris in each poll, represented as y_i. This binomial outcome (successes out of the total sample size) is modeled as a function of several predictor variables that account for differences by state, pollster, and time.

- y_i is the dependent variable, representing the proportion of respondents who support Harris.
- β_0: The intercept term, representing the expected proportion of y_i when all other predictors are set to zero.
- **State** (β_1): Categorical variable that accounts for differences in support across seven key swing states.
- **Pollster** (β_2): Categorical variable capturing variability introduced by the polling organization.
- **Days since first poll** (β_3): Numeric variable representing the number of days since Harris declared her candidacy.

Each coefficient β follows the normal distribution with a mean of 0 and a standard deviation of 2.5, providing some regularization while allowing for a range of possible effects. The standard deviation σ is given an Exponential(1) prior, which constrains it to positive values and provides a realistic expectation of variability around the mean Each predictor captures a unique dimension of voter sentiment:

- **State** accounts for regional differences in support due to factors like demographics and political history.
- **Pollster** adjusts for methodological variations in polling, which we observed to influence support percentages.
- **Days Since Declaration** captures temporal trends, allowing the model to incorporate fluctuations in voter sentiment over the campaign period.

By combining these predictors, the model aims to forecast Harris's support by isolating genuine trends from random variability.

3.1.1 Model justification

Based on the data overview in Section 2.3, we observe that Harris's support varies across key swing states, and we expect these regional differences to influence our model's predictions. The state variable is anticipated to capture these geographic variations, with states like Arizona, Michigan, Georgia, and Pennsylvania showing relatively consistent high levels of support for Harris. In contrast, Nevada, North Carolina, and Wisconsin exhibit a declining trend in support. We expect the intercept to be around 50%, as this reflects the average level of support observed across all states and pollsters.

A model validation is included in `scripts/05-exploratory_data_analysis.R`, and a Posterior Predictive Check and summary statistics for the model are included in Section C.

3.2 Model Results

The model predicts an average support percentage of approximately 47.33% for Kamala Harris across the selected states. This percentage reflects the central tendency of predicted support within the dataset, factoring in the influence of state, pollster, and the number of days since the first poll.

The results indicate variations in predicted support across states. For instance, Pennsylvania and Wisconsin show the highest average predicted support percentages (48.02% and 48.38%, respectively), while states like Nevada and Arizona display slightly lower support.

In states like Arizona, Georgia, Michigan, and North Carolina, the model predicts a steady increase in support for Harris over time, as shown in Figure 7. In states such as Nevada, Pennsylvania, and Wisconsin, the model predicts a decline in support over time, as seen in Figure 8.

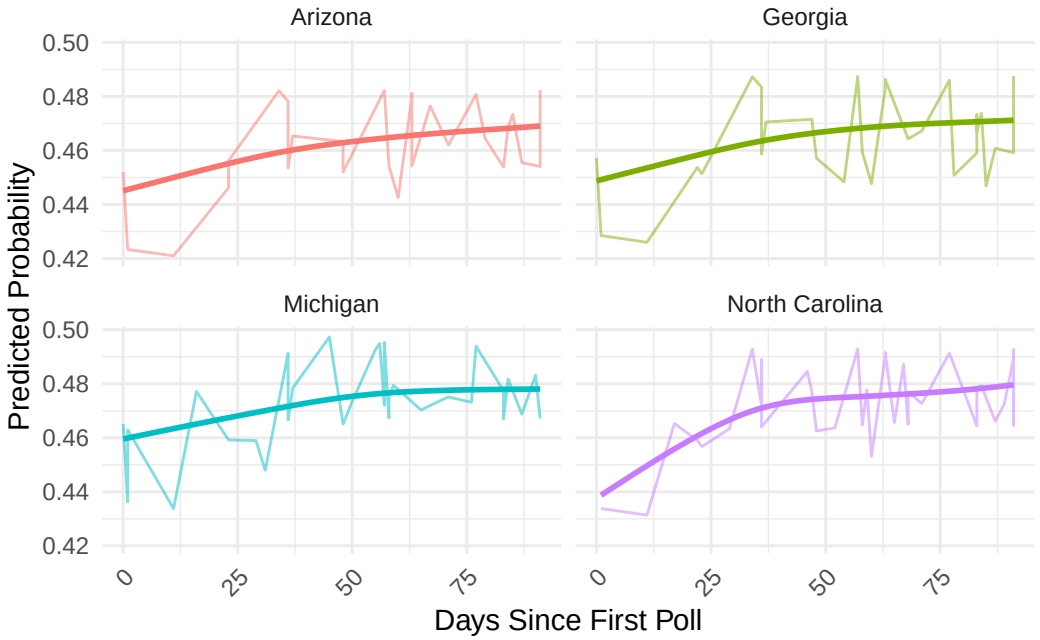

Figure 7: Poll predictions of Arizona, Geogria, Michigan, North Carolina

In Figure 8, we see states of Nevada, Pennsylvania and Wisconsin have a singular declining trend.

CONFIDENCE INTERVAL TABLE ??

State	Avg. Predicted Support for Harris
Arizona	46.26
Nevada	46.43
Georgia	46.59
North Carolina	47.26
Michigan	47.30
Pennsylvania	48.02
Wisconsin	48.38

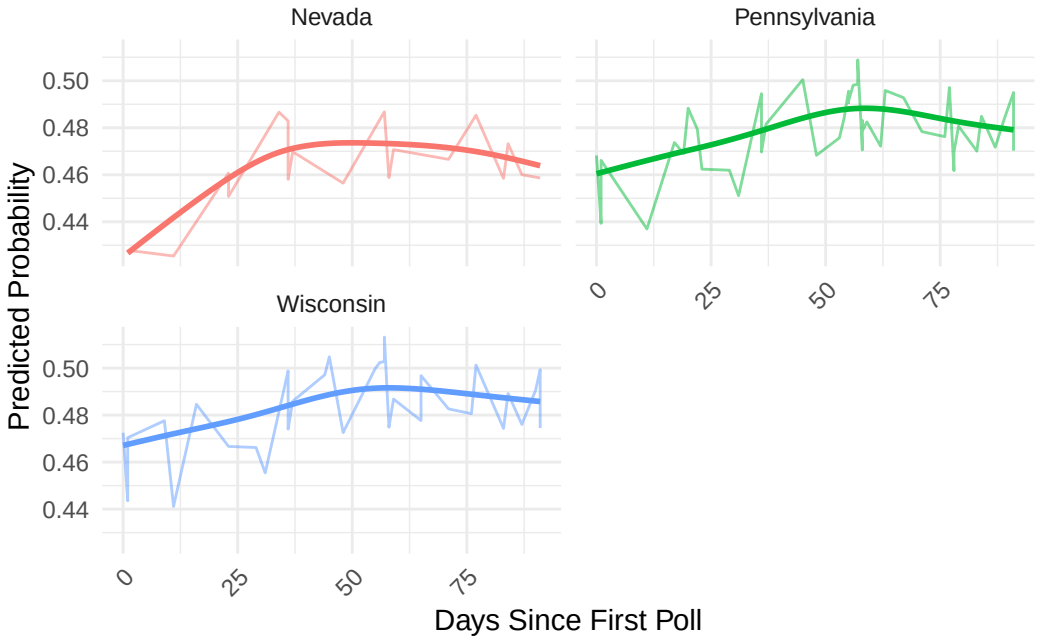

Figure 8: Poll predictions of Nevada, Pennsylvania, Wisconsin

4 Discussion

4.1 Arizona

Arizona is a very important swing state especially as the state has switches which party they voted for in in the last two elections with Trump winning 49% of the vote and Clinton with 45.5%, which considering Arizona is a swing state is a rather significant6 margin. While in the 2020 election Biden ended up winning Arizona with 49.4% of the vote and Trump 49.1%, which is a much smaller margin of victory and what makes Arizona so questionable in the 2024 election. We can see in Figure 4 that while the polling numbers for Harris started somewhat low around 42%, the have gone up significantly and have continued to trend upwards in the most recent polls. Although, in the Figure 1 trump still has larger support in the state, and the largest polling difference in Trump's favour out of all the swing states polled.

We believe Arizona will vote: **Donald Trump**

4.2 Georgia

Considering the last two elections Georgia is very similar to Arizona, as in 2016 Trump won the state with 50.8% to Clinton's 45.6%. While in the 2020 election Biden won with a slim victory of 49.5% to Trump's 49.3%, which again is very similar to Arizona's last two election results with significant republican victory in 2016 and a rather marginal win in 2020, which was the first democratic victory for the state since 1992. Georgia is in an interesting situation as lot of the democratic vote in the state comes from the black community, although Harris a has less support from male black population, as the last victory for Biden was as a small margin, even a small percent of the population switching their vote could heavily impact the vote in 2024.

In Figure 5 at the beginning of the polling data the support for Harris was relatively low, while it did start to grow the polling percentage somewhat plateaued at around 47.5% and does not show any signs of growing.

Furthermore, in Figure 1 we can see that Trump has the edge over Harris considering all of these factors there is a very real possibility that Trump does win Georgia at the end of the day.

We believe Georgia will vote: **Donald Trump**

4.3 Nevada

Nevada is a state who's majority is dessert and wilderness, with two large cities being Reno and Las Vegas, as the case is with many states a Majority of democratic votes come from large cities and the majority of republican votes coming from rural areas. Nevada has been a democratic state for the last 4 elections with at least a 2% difference in favor of the democrats in the last two elections.

Interestingly enough, if we look at Figure 5 we can see Nevada, similar to several other states had a spike in support of Harris in early September peaking north of 48% in the polls and their recent history of voting in favor of the democratic candidates. Nevertheless, Harris is seen to have had a gradual decline in the state stabilizing at around 46%, which which seems low in comparison to other states that are looking more favorably at Harris than Trump. Still, when we look at (**supportstate?**) Harris is still has a marginal to Trump, considering all of these factors it is likely Harris will in fact take Nevada for the democrats.

We believe Nevada will vote: **Kamala Harris**

4.4 North Carolina

4.5 Michigan

Harris is currently leading the polls in Michigan with 49% of the vote and Trump with 47%, this is a rather slim margin, but the first in the discussed states where Harris is leading in the polls. In Figure 4 we can see that there was a significant jump for Harris and a small dip in October followed by a gradual increase leading up to the elections. When we look at the voting history of the state in the last two elections, the have voted for the same candidates as Arizona and Georgia, where they voted for Biden in 2020 and for Trump in 2016, with the voting percentages at 50.6% to 47.8% in 2020 and 47.5% to 47.3 in 2016 with respect to the winner. Although unlike the previously mentioned states 2016 was the first time the state voted for the republican candidate since 1992. Considering Michigan is home to the largest black majority city in the United states, Detroit, and a large majority of union workers who often vote democratic as it is seen as the more union-friendly party between the two it is rather like that Harris will win the state.

We believe Michigan will vote for: **Kamala Harris**

4.6 Pennsylvania

Pennsylvania is an interesting for several reasons, the first is that before Tim Walz was announced as Kamala Harris's running mate, one of the speculated candidates was Josh Shapiro, the current governor of Pennsylvania, which presidential candidates often consider from where their potential running mates are from as they tend to help with swing states nevertheless, she did not choose Shapiro as her running mate and this is no longer a factor.

Second, Similarly to Nevada and Michigan Pennsylvania Voted for Trump in 2016 with a 48.2% to a 47.5% victory, and for Biden in 2020 with a slightly more significant but not overwhelming victory at 50% to 48.8%. 2016 was the first time Pennsylvania voted for a republican candidate since 1988.

Lastly, if we consider Figure 3 we see that harris has been polling at a consistent 48% since mid september which is somewhat similar to Georgia, while Pennsylvania seems to be leaning more democratic than republican than Georgia.@fig-supportGNN, they have both similarly increased their interest in Harris since the polling began until mid september which is when they both becan to plateau at their current polling numbers.

We believe Pennsylvania will vote: **Kamala Harris**

4.7 Wisconsin

4.8 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix 1 - Pollster Methodology

The New York Times/Siena College Poll is a respected polling partnership that provides data-driven insights into U.S. public opinion, particularly on political, social, and economic issues. Known for its rigorous methodology, this poll focuses on capturing the views of registered and likely voters across the country, with particular emphasis on key battleground states during election cycles. The poll is frequently cited for its accuracy and transparency, and is widely regarded as reliable by the public, political analysts, and the media.

The New York Times/Siena College Poll is conducts live telephone interviews with a national sample of registered voters, using both landlines and cellphones, with over 90% of contacts made via cellphones. The sample frame is drawn from a comprehensive voter file covering 200 million registered voters. This file provides demographic and geographic information, enabling representative sampling across age, race, and political affiliation. The poll targets registered U.S. voters, with special focus on battleground states like Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin, determined as swing states. Respondents are randomly selected from the voter file, ensuring a diverse sample by region, party, and demographics. To mitigate non-response bias, the poll applies demographic weighting, increasing the weight of underrepresented groups, such as young adults or non-college graduates. Low response rates (around 2%) present challenges, to mitigate this, weighing adjustments are applied to correct for demographic imbalances and potential non-response bias. Questions are structured to be clear, neutral, and representative across political views, with each question carefully reviewed for balance and clarity. This process reduces the risk of bias in responses, though the brief format of phone surveys limits depth.

B Appendix 2 - Survey

Here we will be talking about how we would conduct a $100,000 surevy to gather data about this upcoming election to ensure there is minimal error in the data we are collecting and, our data represents the facts we want them to represent. Out of the $100,000 we will allocate $15000 for the necessary development and administration of the data. This means we will be spending this money to create the survey, and have a infrastructure in place to hold the large amount of data as well as cover any security fees to keep this data private and safe. Then we allocate $50000 to advertise the survey. We will need a large sample so we will be spending most of our budget for this. We will have a different urls for each advertisement we have so that we can incorporate this in our data to see which demographic is accessing the survey through which platform. The platforms to advertise this is through spotify, facebook, instagram, and some news networks. If possible we will try to use this money to get it endorsed by branches of government to show our reliability. Reliability is also what citizens respond to so this will get a lot of respondents. We will spend $25,000 for modeling our data as they tend to be expensive with data this size. And the rest of the $15,000 will be leftover cost for anything that we don't foresee. If there is any leftover we can add a survey participation price such as having a free 3 month trial for the platforms we mentioned above (that is if they have a subscription based membership).

We will not be using telephone surveys because according to research (Survey 2024) we find the most people, don't pick up calls from unknown numbers and even the people that pick it up they are less inclined to answer the questions of the survey.

Next to look at the actual contents of the survey. They can be accessed through this link: Sample survey question. There are 3 things that we were careful of when we were creating this survey. First thing we considered was transperency. People need the reassurance that the data that is being collected will not be used against them, and so we feel them what data we do collect and how the data we collect cannot be used to identify a person. The second thing we focus on is readability of the questions. We tried to make them as simple as possible using accessibile language, and tried to keep it short as well (Tourangeau, Rips, and Rasinski (2000)). We also prioritised the size of the survey. We kept it to a short 11 questions that will tell us their political standing in the past and present. We know what current issues are important to them as well as their age group and the state they are from. This will help us gather data without inconveniencing

the person. There are disadvanatages to the way we structured the survey. For example, we were keeping time and security in mind, so we restricted the survey to 11 questions, however, this limits the amount of answers we get from the reader. We may be missing some crucial information that helps us figure out which group of people may vote for whom, such as ethnicity and wages (Stantcheva (2023)).

Finally we take a look at how we sample from the gathered data. We thought about random sampling but we were worried about having unequal proportions of the demographic we would advertise to. So if more people that go to facebook responded, that doesn't necessarily mean that people use facebook more than the other platforms. So we we decided to combine stratified and cluster sampling(Stantcheva (2023)). The idea is that we will choose based on the different platforms we advertise to first to have a cluster sampling method. And then we will sample the data by dividing the people by the certain aspects of the surevy, like if they are in different age groups or if they are registered voters. This is called Stratified Sampling (Stantcheva (2023)). Then we take a union of all the samples and we gather a group that we are able to build models based off of.

There are obvious disadvantages as well. For example, we have fewer questions meaning when we sample based on the question we might have cases where, some groups are over or under represented. This can happen for example when most of the people who answered might be from age 18-50 because those are the age groups with access to a phone and are frequently visiting the sites we advertise our survey on. Also the different apps may skew the data as Facebook has a older audience as opposed to Instagram and Tiktok. Another disadvantage is the fact that we have 2 different sampling methods that we use and we combine them both, meaning that this can be an expensive process. Doing so may exceed the allocated budget however, it will be worth it because we minimize the sampling bias we talk about earlier where we may have a group that has more access to the survey. The last disadvantage we also foresee is the fact that younger audience may not be as interested in filling out a survey about politics as they feel, it might not directly affect them. We try to minimize this by offering compensation for their time when they fill out the survey, i.e. free membership services for 3 months on Spotify.

Although there are both advatages and disadvantages, we feel that this survey is well though out and doesn't allow for ambiguity in the answers of the participants.

C Appendix 3- Posterior Predictive Check

In Figure 9, we check the Posterior predictive check of model of our model. Since the raw data aligns closely with our prdictions, our model is a good predicton of the data.

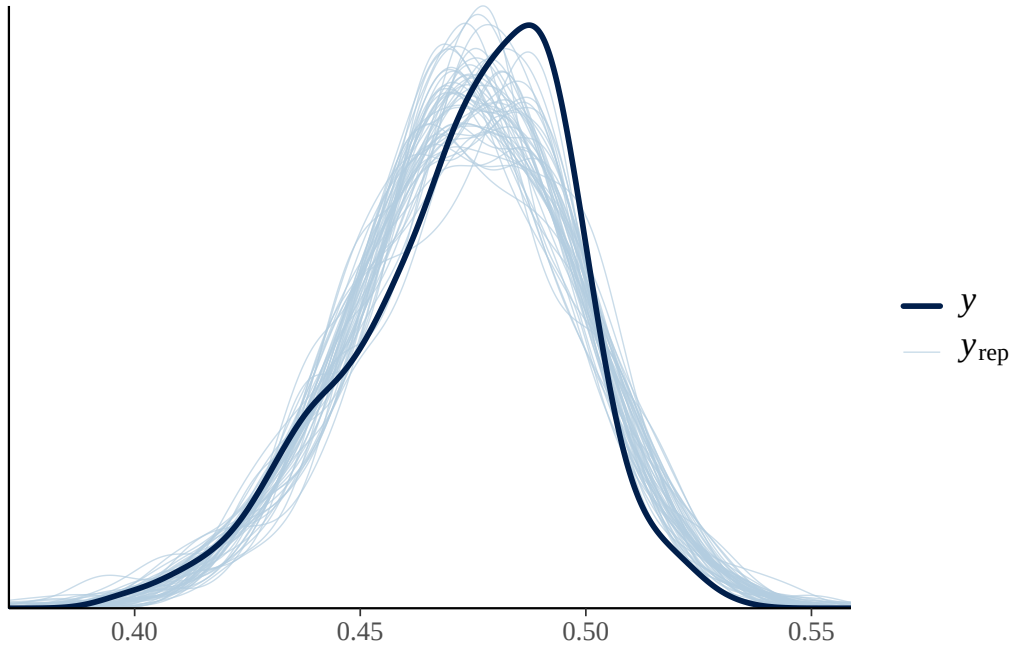

Figure 9: Posterior predictive check of model

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. "2024 National Presidential Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." Journal Article. *Annual Review of Economics* 15 (Volume 15, 2023): 205–34. https://doi.org/https://doi.org/10.1146/annurev-economics-091622-010157.

Survey, ID. 2024. "Phone Survey." https://www.idsurvey.com/en/phone-survey/.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. "Respondents' Understanding of Survey Questions." In *The Psychology of Survey Response*, 23–61. Cambridge University Press.

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.