Predicting US Presidential Elections*

Analysing Electoral Polls to Model Forecast the Winner of the Upcoming US Presidential Election

Aamishi Avarsekar Gauravpreet Thind Divya Gupta

November 4, 2024

Table of contents

*Code and data are available at: https://github.com/aamishi/PredictingUSPresidentialElections

1 Abstract

This research aims to develop a generalized linear model to predict the outcome of the 2024 U.S. Presidential Election. The model aggregates polling data from various sources via the poll-of-polls method to provide a forecast of voter behavior regarding Kamala Harris, the Democratic Candidate, and Donald Trump, the Republican Candidate. This paper presents our exploratory data analysis (EDA) to clean and prepare the dataset followed by the predictive model and its results, which project a Kamala Harris win. The paper further presents an idolized survey to test these findings and discusses the methodology to run it using a USD $100K budget.

2 Introduction

The 2024 U.S. Presidential Election is set to take place on 5th November this year. This year, it is dominated by the competition between the Democratic candidate, Vice President Kamala Harris and her Vice President pick, Tim Walz, and the Republican candidate, former President Donald Trump, and his Vice President pick, JD Vance. Using regularly updated polling data allows one to forecast the outcome of the elections ahead of time and draw meaningful conclusions about voter behavior and the US electoral system. Forecasting the election also allows presidential candidates to recognise gaps in their campaigns and compels them to re-assess their platform to one that is better suited and favorable for the majority voting population.

Accurately forecasting the election outcomes is a dynamic and challenging process which requires one to assess voter behavior (which may vary county by county based on a plethora of factors such as age, sex, income levels and race), polling methodologies, etc. In this paper, we aim to develop a general linear predictive model that forecasts the outcome of the 2024 elections through a polls-of-polls approach. This approach aims to aggregate the results of multiple polls to develop a more reliable estimate of voter behavior across different sample sizes, counties, and demographic features. To develop this model, we use the publicly available

polling data retrieved from FiveThirtyEight. Our analysis focused on data collected between July and October 2024, and includes key voter demographic variables such as state of residence, pollster rating, and sample size.

This paper is structured such that Section 2 describes the dataset, including high-level data cleaning through our exploratory data analysis, variables used in the analysis, summary of statistics and necessary data visualizations. Section 3 presents the predictive model, its results, strengths and limitations. This allows the researchers to discuss next steps to improve the model and conclude our findings in Section 4. Appendix A presents an analysis of the New York Times/Siena College pollster methodology, and Appendix B presents a sample survey we could run to assess the efficacy of our model along with a potential allocation of a USD $100K budget to run this survey.

3 Data

3.1 Data Overview

The data that we are working with for this paper has been obtained from 385(R Core Team (2023)), using the Presidential General Election Polls (Current Cycle) data. The website also contains the data for the previous US Presidential Election from 2020 and Polling averages for predicting the next President. The website does not offer an API so the data must be downloaded manually and is available in .csv format. We have employed the dplyr package (Wickham et al. (2023)) to select the variables of our interest by omitting unused unnecessary variables. We used the janitor package (Firke (2023)) to clean all variable names for uniformity. For the creation of new variables or the mutation of existing variables, we used Grolemund and Wickham (2011) package's mdy() function. The csv data was converted to a dataframe in and subsequently converted back to a parquet file using Richardson et al. (2024) for ease of use.

The original data set provides 52 variables that reflect the data that each pollster has collected. These include each pollstor's sponsor, their respective identification numbers, the duration of the poll, the methodology used, the percentage of support received by each leading candidate, and numerics that evaluate the reliability and strength of the poll. 385 determines "major" candidates against a matrix of evaluation and aggregates the polls that focus on these candidates ("538's Polls Policy FAQs" 2024).

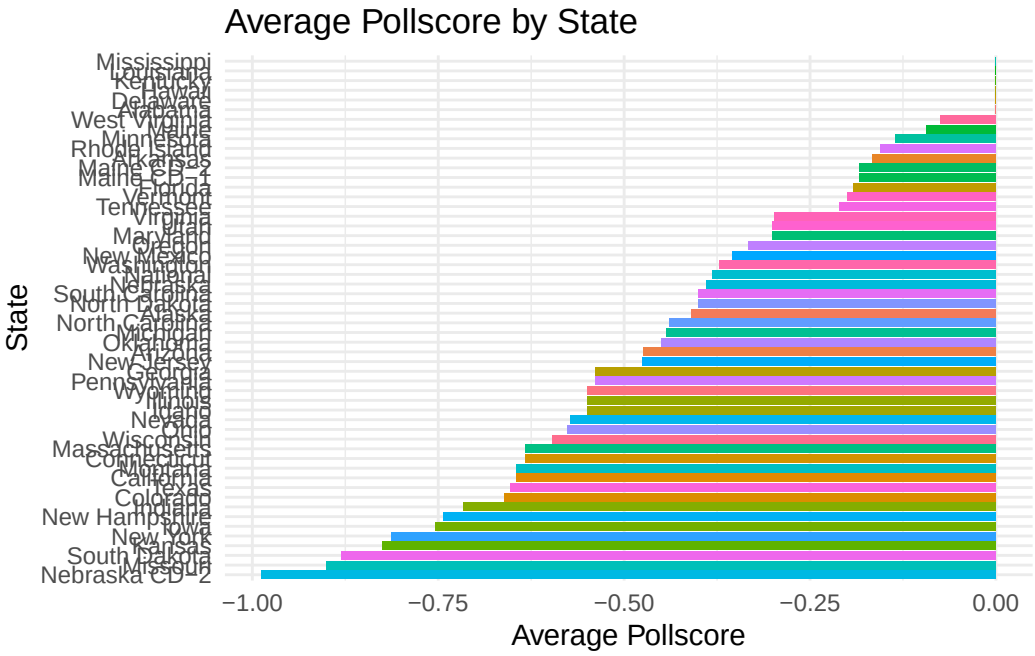
The website uses a poll-of-polls method to calculate the support for a major candidate. Using this method increase reliability by reducing the effect of individual bias for each pollster and noise across the data.

4 Measurement

The most crucial part of this analysis is to convert an individual's conflicting and convoluted opinions regarding their voter intent into a measurable metric. The pollster methodologies and surveys conducted by various pollsters in our data help situate and represent various aspects of a rapidly changing voter sentiment ahead of the 2024 Elections. Specifically, these surveys aim to sample registered voters and gain information regarding their demographics, party affiliation, voter intent, and guiding forces behind their vote (such as the primary issues that drive their vote). Weighting these individual and highly complex voter responses into stratified groups helps not only avoid overrepresentation of certain groups, but helps pollsters draw conclusions regarding extremely specific voter bases and how they stand on certain issues. Weighting the sample responses by state and demographics further ensures that traditionally underrepresented voter groups, which can often be instrumental in turning over election results (specifically in swing states), are accurately represented. A poll-of-polls approach of these accurate pollsters will later allow us, in our model, to aptly forecast the results for this election.

The dataset in use is downloaded from FiveThirtyEight and presents the aggregate data from various pollsters conducted in 2024 to indicate a preference for Vice President Kamala Harris - the Democratic party candidate, former President Donald Trump - the Republican party candidate, and third party candidates such as Jill Stein (Green Party), etc. All polls have an associated poll id along with information regarding the pollster, sample size, and methodology.

The main challenge here is to analyse voter intent data from the polls and aptly translate it into structured data useful for a forecasting model. This can be achieved by aggregating polling results from various pollster surveys, employing apt weighting and adjustment methods to ensure the survey results best reflect voter sentiment and demographic distributions, and report the results according to the time they were captured, so that we have real-time, accurate data while acknowledging the limits to our data's forecasting capacity due to changing voter sentiment over time.



4.1 Variables

4.1.1 Outcome Variable

The outcome variable that we decided to focus on is `pct`, which is the proportion of support that a major candidate is projected to receive through that poll. This proportion accounts for the candidate's support against all other candidates. We use this variable to estimate which candidate is likely to win based on popular vote.

4.1.2 Predictor Variables

Our predictor variables are as follows:

- **pollster**: The pollster conducting the poll. Each pollster also had an associated sponsor candidate, but this occurrence was rare, so we chose to omit it.
- **pollscore**: The score assigned to the reliability of the respective pollster. It includes error and bias that can arise within each pollster's methodology; hence, negative numbers are better. The original website describes it as "Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record, Empirically." `pollscore` is used to assign weight, as we will see later in the model section.

- **transparency_score**: A numerical value, graded on a scale of 1 to 10, that indicates how transparent a pollster is based on the amount of information they disclose and how recent their polls are.
- **end_days**: The number of days that each poll was conducted until November 3rd. This number was generated using `start_date` and `end_date`, and it is used as the spline term for our model with 3 degrees of freedom. This variable is used to assign greater weight, as we will see in the model section.
- **state**: The United States state where the poll was conducted. This variable has been modified to include a "national" category if the poll was not specific to any particular state.

5 Model

5.1 Model Overview

This model employs Bayesian analysis to fit the predictive model based on the current polling data. Spline functions have been used to fit the `end_days` variable, capturing the weight of more recent polls to reflect their potentially greater relevance in predicting support for Vice President Kamala Harris. The dependent variable, `pct`, represents the likelihood of Harris being elected as the next President of the USA, serving as the primary measure of voter support.

5.2 Model Equation

$$pct_i = \beta_0 + \beta_1 \cdot \text{pollscore}_i + \beta_2 \cdot \text{transparency\_score}_i + f(\text{end\_days}_i) + \beta_3 \cdot \text{state}_i + \epsilon_i$$

5.2.1 Where:

- pct_i = support for Kamala Harris in the (i^{th}) poll
- β_0 = intercept (baseline support when all predictors are zero)
- β_1 = coefficient for the pollster reliability score (`pollscore`)
- β_2 = coefficient for the pollster transparency score (`transparency_score`)
- end_days$_i$) = spline function capturing the non-linear effect of the recency of the poll (measured in days)
- β_3 = coefficients for the categorical variable representing state (each state would have its own coefficient)
- ϵ_i = error term (captures unexplained variance)

5.3 Predictor Variables

The predictor variables selected for this model include `pollscore`, `transparency_score`, and `state`, each chosen for their significance in understanding polling dynamics.

5.3.1 Pollscore

This variable reflects the reliability of the pollster, where lower scores indicate a more accurate historical performance. Including `pollscore` allows the model to account for biases and errors associated with different polling organizations, ensuring that more reputable polls are given greater weight in the analysis.

5.3.2 Transparency Score

The transparency score indicates how openly pollsters disclose their methodologies. This variable is critical because it helps assess the trustworthiness of the polls. A higher transparency score suggests a more reliable polling process, which can enhance the credibility of the results.

5.3.3 Spline for End Days

By incorporating a spline function for `end_days`, the model captures non-linear effects associated with the recency of the polls. This approach allows for a more nuanced understanding of how support for Harris changes over time, recognizing that more recent polls may reflect current voter sentiment more accurately.

5.3.4 State

Including state as a categorical variable helps account for regional differences in voter support. Each state may have unique demographic and political contexts that influence support for Harris, and capturing these variations is essential for improving the model's accuracy.

5.4 Model Priors

All variables in this model utilize default priors, with an additional weight applied to recent polls with higher `pollscore` values, emphasizing the importance of both reliability and recency in predicting electoral support for Kamala Harris.

6 Discussion

This model aimed to forecast the 2024 U.S. Presidential Election results using a poll-of-polls approach that combined the polling data from various pollsters into one. We employed Bayesian analysis and selected a set of predictor and outcome variables to accurately track voter sentiment and voter intent regarding the two main presidential candidates, Vice President Kamala Harris and former President Donald Trump. This section analyzes the strengths and limitations of our model with regards to being an accurate predictor of the election results.

6.1 Strengths

The model relies primarily on high quality polling data pulled from reputable pollsters such as the NYT/Siena college poll. These are further validated by pollscore and transcparency_score filters, ensuring that all polls included have a baseline level of reliability and transparency. This makes the data reliable and trustable, further enhancing the accuracy of the model's predictions. Secondly, by using stratified weighting for polls, we are able to understand nuanced voter sentiments as represented across different demographic characteristics such as age, race, gender, income levels, state and party affiliation. Further, by using polling recency and pollster reliability, we can increase the forecast's accuracy by ensuring it is based on changing voter sentiment over time. Using a Bayesian approach in developing the model allows for probabilistic inference which helps manage uncertainties in our data by allowing us to assign personalized weights to specific demographics (to manage under or overrepresentation), and by allowing us to understand change in voter sentiment over time. This provides a granular, data-driven look into a very broad and diverse voter base across the states.

6.2 Limitations

This model predicts the likelihood of Vice President Kamala Harris winning the upcoming USA Presidential Elections in 2024. This model is a simple representation of Kamala Harris's win or no-win. The model does not take into consideration that independent candidates could account for the lack of support for Harris but that does not accumulate into direct support for Trump. This model also only considers specific support for Harris. There is no exploration of the relationship between support for the Democrats and support for Harris. A state may have an overall Blue majority but not necessarily for Harris directly.

This model assumes a linear relation for the scores provided for understanding the reliability of each poll. pollscore and transparency_score are a general guideline to assess the reliability of a poll but they are based on historical evaluations of each pollstor. Thus, biases may still exists regardless. This model also has a simplified weight distribution for recent polls and polls with higher pollscores and transparency_scores. For examples, any poll within the 6 month period sense writing this paper is weighted at 1 and any older polls are weighted to be 0. This

does not capture the right sentiment associated with general people's opinions as the election day nears.

This model also does not take into consideration swing states and only uses them as categorical variables. This was done by design at this risk of overfitting. Further, the shy voter phenomenon may stop people from accurately reflecting their voter sentiment in the polls that our data is extracted from. For example, people voting for Donald Trump may feel that they will be ostracized for their choices and so, they may lie about their voting preferences by claiming to vote for Harris or say they're undecided. Similarly, voters who do not wish to vote for Kamala Harris due to her policies but do not wish to vote for Trump either, specifically in swing states, are increasingly engaging in vote swapping mechanisms where they make deals with people in guaranteed blue states such that a blue state voter votes for an independent party while a swing state voter votes for Harris to ensure a Democratic win over Trump. This is an uninformed and non regulated practice which impedes the data we have and leads to inaccurate predictions.

6.3 Next steps

Add a more educated method to assign weights to more reliable and more recent polls. For example, as Election Day nears, Trump has been focusing more on swing states like North Carolina ("US Election Live: Polls Show Close Race as Trump, Harris Hit North Carolina" 2024) for his last set of campaign speeches. This could potentially weigh more than polls taken slightly earlier but still within the 6 month period.

7 Appendix A - Pollster Deep Dive for Kamala Harris

In this appendix, we conduct an analysis of the polling methodology used by the New York Times/Siena College (NYT/Siena) partnership by examining the poll's population, sampling frame, recruitment methods and survey strategies especially as they relate to presidential candidate Kamala Harris. The NYT/Siena poll aims to capture voter sentiment on both the national and state levels by employing stratified sampling and advanced weighting techniques to capture a representative sample of the voting population. Kamala Harris has a strong positioning within the Democratic voter base, and this analysis deep dives into specific populations and methods she employs within her campaign to reach her diverse voter base.

Population, Frame, and Sample Population: The NYT/Sienna poll targets various key demographic groups, with specific emphasis on battleground or swing states such as Pennsylvania, Georgia and Michigan. Harris particularly aims to reach Democratic-leaning voters that belong to groups such as people of color, LGBTQ+ people, women, students and younger populations (ages 18-35), and Latinx voters (such as the Puerto Rican population).

Frame: The sampling frame is developed using voter registration records by state and other voter registration databases such as the L2 voter file. This demographic information allows accurate population arrangement according to factors such as age, gender, race, geographic region (urban/rural) and education level. This also ensures data accuracy by representing traditionally underrepresented voter groups and capturing voter intent in swing states.

Sample: The NYT/Siena poll typically samples 800 to 1500 respondents per poll (per state - with the late October 2024 poll with focus on swing states consisting of 7,575 voters (New York Times, 2024)), however, it may be bigger to ensure representativeness on a national level. Oversampling is conducted, specifically in swing states, to capture an accurate portrayal of underrepresented groups so that the accuracy of subgroup analysis can be increased.

Respondent Recruitment The NYT/Siena poll uses various methods such as random digit dialing, interviewing, and online polling to reach a wide sample of participants from a list of registered voters. Random Digit Dialing consists of calling landlines and mobile phones in order to reach a wide age range of voters, including non tech-savvy, potentially older voters, and those living in rural areas. Key states like Florida and Arizona also consist of a high Latinx voter population, which are key focuses in Kamala Harris's presidential campaign. To gain their voter perspective, the NYT/Siena poll also conducts online surveys and phone calls in both English and Spanish to enhance inclusivity, reduce language-based barriers, and accurately capture the diversity in the sample.

The NYT/Siena poll employs stratified sampling, defined as a probability-based method to divide a population into distinct strata based on characteristics (Alexander 2024) such as state, district, party, age, race, gender, and income level (signaled by home ownership) (New York Times, 2024). This is done to ensure accurate representativeness of various groups and reduce sampling bias. This allows traditionally underrepresented groups, such as people of color or low-income voters, to be aptly represented.

Post data collection, post-stratification weighting is applied to ensure accurate sampling representation with respect to the voter populations of each state, specifically the swing states that are crucial in determining the election's outcome. For example, older voters are more likely to respond than younger voters to techniques such as random digit dialing. This non-response bias needs to be adjusted to accurately reflect voter representativeness. The poll also adjusts state-by-state to outline voter behavior based on past polling history.

Strengths and Limitations The NYT/Siena poll lays great emphasis, through its probability-based stratified sampling approach, on a diverse sampling recruitment. This helps accurately represent election outcomes by ensuring underrepresented groups are aptly represented. Further, the poll focuses on battleground (or swing) states such as Pennsylvania, Wisconsin, etc. which provides deeper insights into the convoluted voter sentiment in such states. Lastly, post-stratification weighting reduces the non-response bias and reduces the risk of overrepresentation of certain voter strata.

On the other hand, the poll has certain weaknesses. The random digit dialing methodology of reaching out to voters has a tradition of low response rates, often falling lower than 2%. This

limits the sample size greatly and can potentially lead to an underrepresentation of certain, unreachable voter groups despite the post-stratification weighting due to an initial lack of data. Further, due to higher non-response rates in certain groups, it is harder to accurately divide the subgroups and represent them in the sample. For example, the poll may not be able to capture rural residents due to unreachability, which will bias the polling results. Further, the polling results may differ greatly depending on when the poll was conducted due to rapid changes in voter sentiment, increasing success of Kamala Harris' presidential campaign, and the social desirability factor to fit in with the majority group of voters.

8 Appendix B: Idealized Polling Methodology

In this appendix, we create an idealized polling methodology for forecasting the 2024 U.S. Presidential election with a budget of $100,000. This helps maximize the accuracy and representativeness of poll results, and gain specific information regarding voter demographics, intentions and motivations. To generate accurate and a wide array of results, the polling methodology includes stratified random sampling with multiple recruitment channels such as online recruitment, random digit dialing, and participation incentives. A detailed survey (https://docs.google.com/forms/d/e/1FAIpQLSdTjIs3mYx5rQCoexps3vc3bzcg15GptrRx1jVzVygUTY6zg/viewform), implemented using Google Forms, will be used to capture this data.

The object for finding the ideal methodology for forecasting the 2024 US Presidential Election starts off by gathering data on voter preferences, key issues, and demographic characteristics. The second objective would be to provide an accurate forecast of the election outcome based on the collected data.

Forecasting the 2024 US Presidential Election, starts with a sampling approach that will start by identifying the demographics of eligible voters. Starting by gathering the population over the age of 18 by election day, hold a valid US citizenship, and meet their respective state's residency requirements. Our sampling method would use the stratified random sampling approach to ensure representation across demographics. The Strata would go by the US state, age group, gender, ethnicity, and education level. With a sample size of approximately 10,000 respondents for a 95% confidence level and +3% margin of error.

The sample population will be divided into strata in the following way: age, gender, state, race or ethnicity, and education level. This stratification ensures that the survey represents traditionally underrepresented minority groups such as Latinx or African American voters. Further information that helps us understand the strata by their voting behavior will be collected, particularly voter's past voting behavior, voter intention for this election, and their primary areas of issue that motivate their voting intentions.

A large sample size helps provide sufficient data and statistical power to understand and draw conclusions about the data. Specifically, we aim to reach the sample size of approximately

10,000 respondents. This will ensure enough data is gathered from underrepresented groups as well as swing states to draw conclusions regarding their voting behaviors. Recruitment Strategy

To reach a wide voter base, we aim to employ the following participant recruitment strategy:

Online recruitment: Social media platforms such as Facebook, Twitter, Instagram, Reddit and TikTok are a primary source of gathering voter opinions as they serve as community platforms for a large part of our sampling populations. A portion of the budget will be allocated towards running targeted ads on such platforms and other online news and political discussion forums to reach our audience. We allocate $20,000 to this portion of the campaign.

Random-Digit Dialing: While online recruitment may be successful with tech-savvy voters, random digit dialing will help us reach participants that may not use social media (such as those in rural areas or older voters). This should occupy $30,000 of the total budget, along with an additional $10,000 allocated to costs such as wages for bilingual call center agents to reach a diverse voter base.

Participation Incentives: To motivate voters to actually and accurately reflect their voter intentions in our survey, they will be ensured of data confidentiality within the survey and be informed about our research intentions with the data beforehand. Further, they will be offered small monetary rewards (for eg. gift cards) to motivate them to complete the survey. This can also help capture underrepresented groups such as low-income voters, students and minorities. We allocate $20,000 to this portion of the campaign.

The survey was developed with the help of Google Forms. To ensure accuracy in data capturing along with ensuring participation confidentiality, we allocate $5,000 from the budget to enhanced technical support. After the data is captured, we allocate the last $15,000 from the budget to data cleaning, validation and post-stratification weighting.

Appendix B Survey Questions:

Demographics

What is your age?

What is your gender? Male Female Non-binary Prefer not to say

What is your ethnicity? (select all that apply) White Black or African American Hispanic or Latino Asian
Native American Other Prefer not to say

What is the highest level of education you have completed? Less than high school High school Graduate Some college Bachelor's degree Graduate or professional degree

In which state do you reside?

Voting Intentions

How likely are you to vote in the upcoming election? Very Likely 1 2 3 4 5 Definitely not

Which candidate do you intend to vote for in the upcoming presidential election?
Kamala D. Harris (Democratic) Donald J. Trump (Republican) Robert F. Kennedy Jr. (American Independent) Chase Oliver (Libertarian) Jill Stein (Green) Claudia De La Cruz (Peace and Freedom) Other:

Key Issues

Which issues are most important to you when making your voting decision? (Select all that apply) Economy Healthcare Education Climate Change Immigration Social Justice Foreign Policy Gun Control Other:

How satisfied are you with the current direction of the country? Very satisfied 1 2 3 4 5 Very dissatisfied

How much influence do you believe your vote has on the outcome of the election? A great deal 1 2 3 4 5 None at all

Past Voting Behavior

Have you ever voted in previous elections? Yes No Prefer not to say

If yes, which elections did you participate in? (Select all that apply) Presidential elections Midterm elections Local elections

How do you typically vote? (select all that apply) In-person on election day Early voting Absentee ballot Other

Additional Comments Is there anything else you would like to share about your voting preferences or issues that matter to you?

The data validations from the survey starts by offering diverse demographic questions including age, gender, ethnicity, and education level that allows for a comprehensive understanding of the respondent base and how different demographics may influence voting behavior. For voting Intentions, the survey questions about likelihood to vote and candidate preference directly gauge respondents' engagement in the electoral process. The survey offers key Issues of identification, by allowing respondents to select multiple key issues, the survey captures a nuanced view of what influences voter decision-making. Inquiring about past voting experiences provides context for current intentions, helping to identify trends in voting behavior and closing with the additional comments section allows respondents to express nuanced views that may not be captured in structured questions, providing qualitative insights.

For strength limitations of the survey, the options for gender and ethnicity may not encompass all identities, leading to possible exclusion of some respondents' experiences. Numeric scales (e.g., for satisfaction or influence) may be interpreted differently by respondents, making it challenging to quantify responses consistently. The open comments section at the end of the survey can yield varied qualitative data that may be difficult to analyze systematically. Also, the number of questions may lead to fatigue, affecting the quality of responses, especially if participants feel overwhelmed. Lastly, for geographic limitations, asking about the state of

residence is useful but may not capture regional issues that could influence voting behavior; more localized questions could provide deeper insights.

9 References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. Palmerpenguins:

Palmer Archipelago (Antarctica) Penguin Data. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. Vienna,

Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.

"538's Polls Policy FAQs." 2024. https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

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/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to Apache Arrow*. https://github.com/apache/arrow/.

"US Election Live: Polls Show Close Race as Trump, Harris Hit North Carolina." 2024. https://www.aljazeera.com/news/liveblog/2024/11/2/us-election-live-polls-show-close-race-as-trump-harris-hit-north-carolina.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.