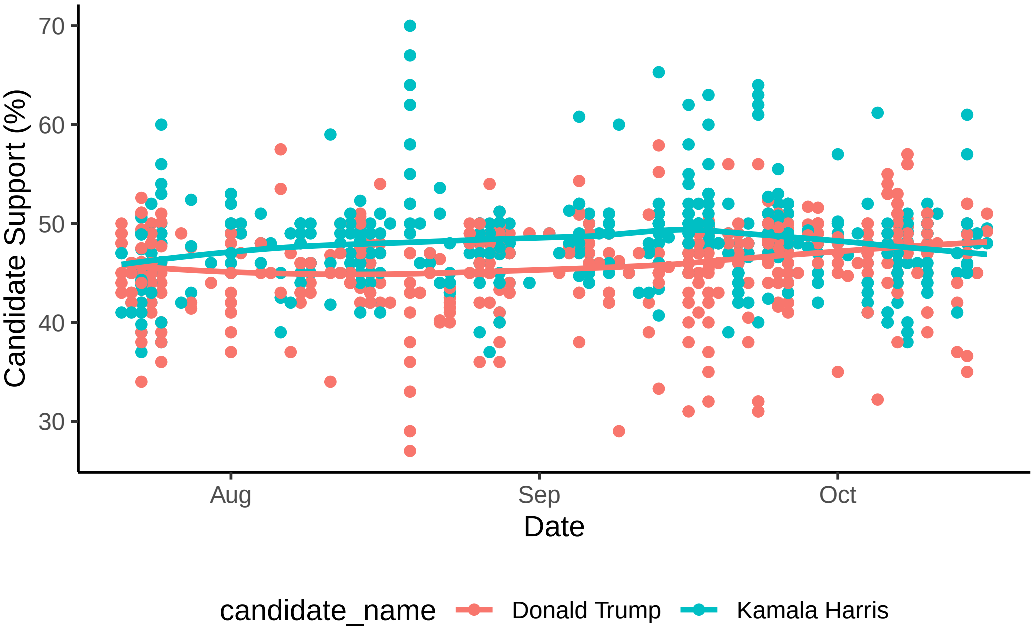
Forecasting the 2024 US Presidential Election*

A Poll-of-Polls Approach Using Generalized Linear Modeling

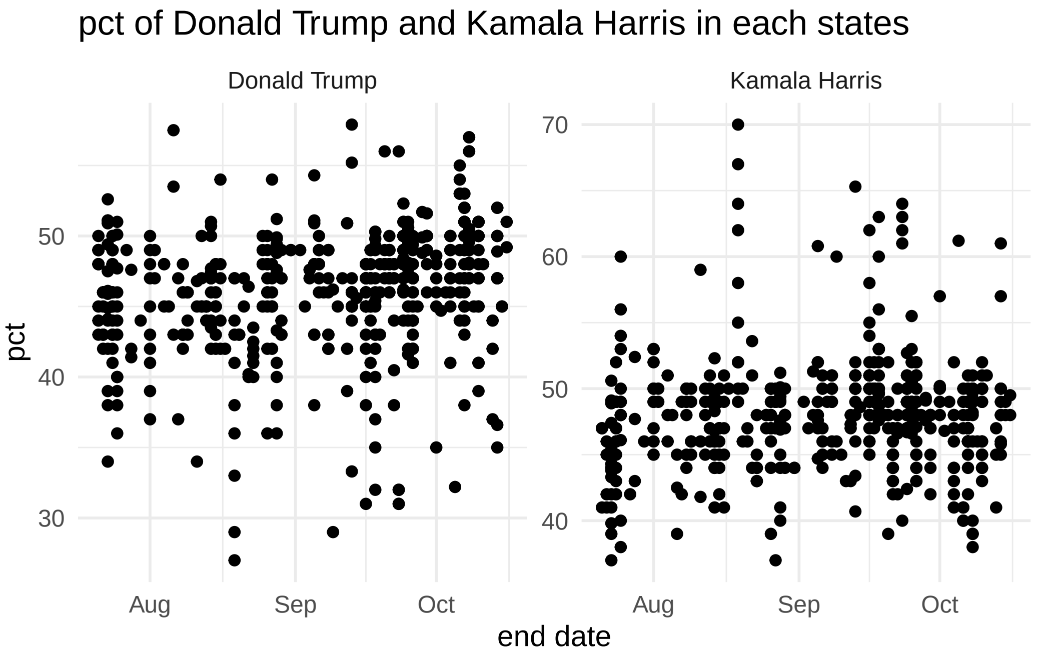
Xinqi Yue Yawen Tan Duanyi Su

November 3, 2024

This paper builds a predictive model to forecast the 2024 US presidential election using a "poll-of-polls" approach, analyzing high-quality polls for Kamala Harris and Donald Trump. The model incorporates data on polling methodology, state trends, and candidate support. Our findings indicate that xxxxxxxx. This study contributes to a deeper understanding of electoral forecasting, helping us interpret aggregated polling data to anticipate election results accurately.



*Code and data are available at: https://github.com/xinqiyue/2024_US_Election_Prediction.



1 Introduction

The 2024 U.S. presidential election is generating significant public interest, with recent polling reflecting shifts in candidate support across various demographic and regional groups. Polling data, while insightful, presents challenges due to inherent biases, variability in polling methodology, and regional influences on voter behavior. This study aims to address these challenges using a "poll-of-polls" approach, aggregating high-quality national and state polls to forecast voter support for Kamala Harris and Donald Trump. By applying Bayesian modeling, this analysis incorporates both the temporal dynamics of candidate support and state-specific trends.

In this paper, we explore the following questions: How is support for each candidate likely to evolve as the election approaches? How do regional differences influence overall voter sentiment? This paper contributes to electoral forecasting literature by synthesizing aggregated polling data to provide a nuanced view of candidate support trends.

2 Estimand

Our primary estimand is the percentage of voter support for Donald Trump and Kamala Harris on election day. We estimate this support percentage both at the national level and by state, incorporating time and state-specific effects.

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 3....

3 Data

3.1 Data Overview

The dataset encompasses polling data focused on two candidates, Donald Trump and Kamala Harris. Data has been filtered to include only results from pollsters with a numeric grade above 2.5 to maintain quality and consistency. This selection allows us to focus on higher-rated sources, which may enhance the reliability of predictions. Each record includes information on the pollster, their grading, polling methodology, transparency score, and specific polling results for each candidate.

The analysis utilizes FiveThirtyEight's dataset of national presidential general election polls (FiveThirtyEight 2024). Following the approach outlined in (Alexander 2023), we aim to predict the election outcome based on this polling data. The analyses were conducted in R R Core Team (2023), with support from several packages. The `tidyverse` packages (Wickham et al. 2019) were used in the process of data simulation, testing beforehand. After the original raw data was downloaded by using `tidyverse` package (Wickham et al. 2019), data cleaning process was done by using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), and `arrow` package (Richardson et al. 2024). Then, models were constructed using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), `rstanarm` (Goodrich et al. 2022) package, and `splines` package (R Core Team 2024). The model results are then presented by `modelsummary` (Arel-Bundock 2022) package, and graphs were made with `ggplot2` package (Wickham 2016).

3.2 Data Clean

We clean the raw data to produce the analysis dataset using the tools listed in the Section 3.1, providing high-quality data for subsequent election analysis. Firstly, we standardize column names to lowercase and remove special characters. Then, we retain only the fields relevant to the election analysis, such as `state`, `end_date`, `sample_size`, `candidate_name`, and `pct` (percentage support), simplifying the data structure by removing unnecessary columns. Next, we remove rows with missing values and replace empty values in the `state` column with "National," indicating these records pertain to nationwide polls. Additionally, we convert the `end_date` column to date format, ensuring accuracy in date-based filtering. Then, we filter the data to retain only records that meet specific criteria: a `numeric_grade` of 2.5 or higher, a `candidate_name` of either "Kamala Harris" or "Donald Trump" (focusing on these two candidates' support levels), an `end_date` on or after July 21, 2024 (focusing on recent polling when Biden withdrawal from the election), and a non-empty `end_date` (filtering out records with missing dates).

The cleaned data is displayed in Table 1, which contains a total of 5 variables:

Table 1: First 6 entries of Analysis Dataset

State	End date	Sample Size	Canadidate Name	PCT
National	2024-10-16	1000	Kamala Harris	49.5
National	2024-10-16	1000	Donald Trump	49.2
Arizona	2024-10-16	1435	Kamala Harris	48.0
Arizona	2024-10-16	1435	Donald Trump	51.0
National	2024-10-15	1457	Kamala Harris	48.0
National	2024-10-15	1457	Donald Trump	45.0

3.3 Predictors Explanation

- **state**: The U.S. state where the poll was conducted, which allows for state-by-state analysis and comparison of support levels.
- **end date**: The date the poll concluded, marking the end of data collection for that specific poll.
- **sample size**: The number of respondents in the poll, indicating the scope and potential statistical reliability of the results.
- **candidate_name**: The name of the candidate being polled, which in this data set focuses on Donald Trump and Kamala Harris.

3.4 Outcome Exaplanation

- **pct**: The support percentage each candidate received in the poll, which serves as the outcome variable for analysis.

3.5 Explanation of Other Variables Used

- **pollster**: The organization that conducted the poll, providing information on who gathered the data.
- **numeric grade**: A quality score assigned to the pollster, with higher values indicating greater reliability. Only pollsters with a score above 2.5 are included to enhance accuracy.

- **pollscore**: A specific rating of the individual poll, reflecting additional factors that impact the poll's quality and reliability.
- **methodology**: The method used by the pollster to conduct the poll, which may affect the reliability and interpretation of the results.
- **transparency score**: A measure of how openly the pollster reports their methodology and results. Higher scores suggest greater transparency and reliability.

3.6 Measurement

The goal of this measurement is to turn individual voter opinions into a reliable estimate of electoral college outcomes by analyzing polling data. This data, sourced from (**fivethirty?**)—a trusted platform known for high standards—includes only polls that meet rigorous quality criteria, ensuring broad representation of likely U.S. voters. Polls provide essential details, such as the pollster's identity, survey dates, sample size, and methodology, covering data collection mode, demographic weighting, and adjustments to create a representative sample. These quality standards, while robust, are based on historical practices, which may overlook recent methodological shifts or emerging biases.

3.7 Limitation

Polling has inherent limitations. It provides snapshots of voter sentiment at specific times rather than ongoing updates, potentially missing rapid shifts in public opinion, especially near Election Day. Adjustments are made for recency, but participation and response biases—like self-selection and social desirability bias—can still impact accuracy by creating gaps between expressed opinions and actual voting behavior. Additionally, regional polling disparities, with battleground states polled more frequently than "safe" states, can lead to imbalances in representation, highlighting the challenge of achieving uniformly accurate forecasts across the nation.

3.8 Data visualization

3.8.1 Exploring the Relationship Between State and PCT

3.8.2 Exploring the Relationship Between End Date and PCT

4 Model

In this analysis, we use a Bayesian linear regression model to estimate the percentage of support (`pct`) for each candidate in the upcoming US presidential election. The model includes key

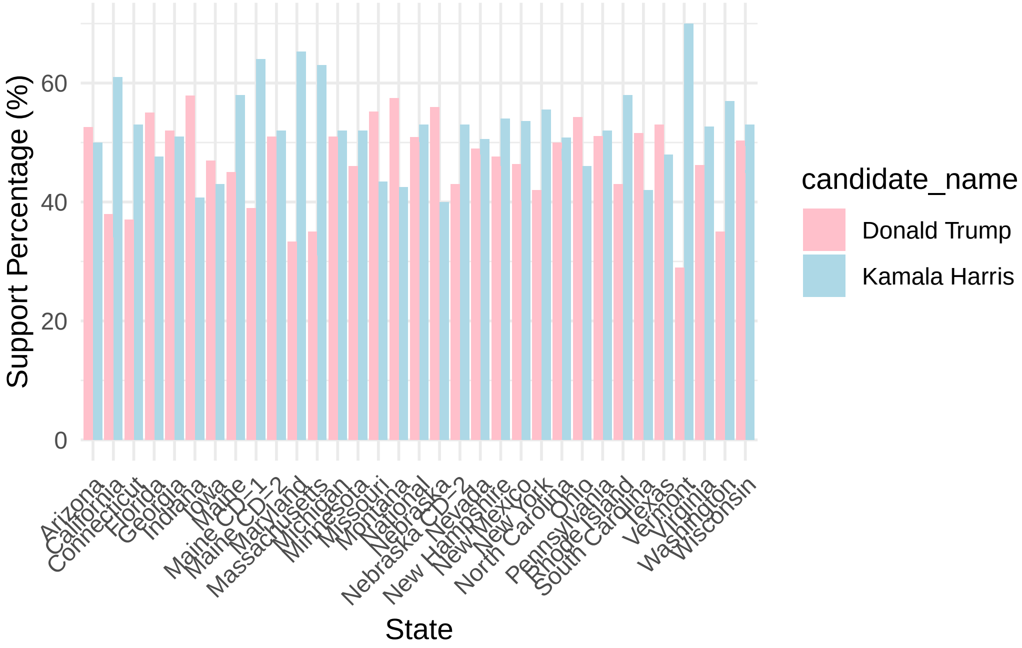


Figure 1: Support for Candidates by State

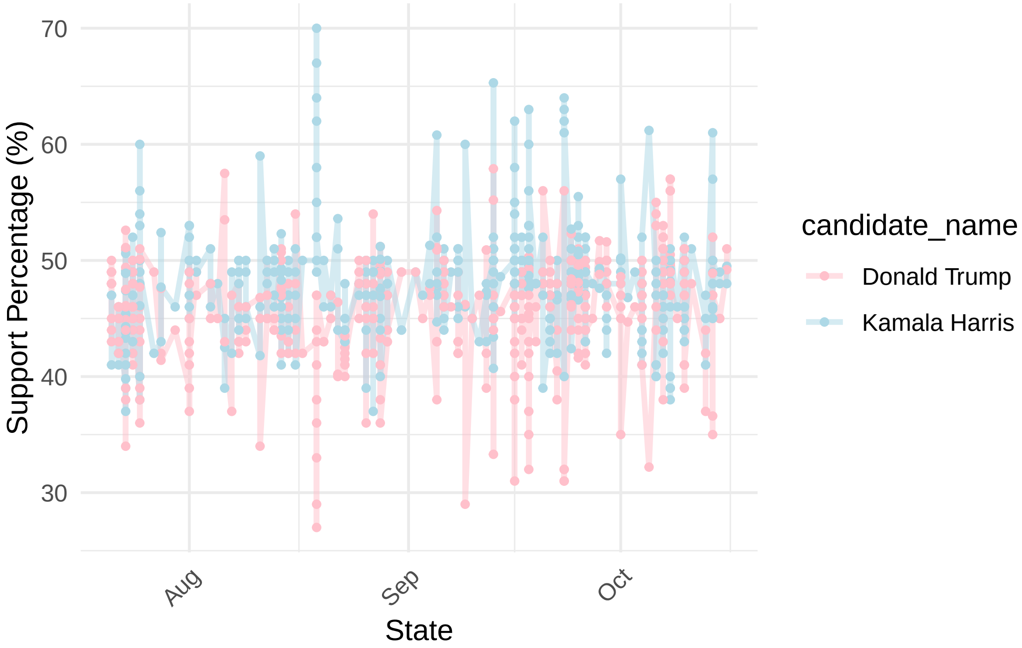


Figure 2: Trends in Candidate Support Over Time

predictors such as polling end date (`end_date`), candidate (`candidate_name`), sample size (`sample_size`), and state (`state`). Each predictor is selected based on its relevance to the variation in candidate support, as discussed in the data section.

4.1 Model set-up

In this Bayesian framework, we assume a normal distribution for poll results influenced by several predictors: end date, candidate name, sample size, and state. We denote y_i as the percentage of votes for a candidate in a given poll, where $beta_i$ represents the candidate effect, $gamma_i$ reflects the influence of sample size, and $delta_i$ corresponds to the state effect.

The model can be mathematically expressed as follows:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{candidate\_name}_i + \beta_3 \cdot \text{sample\_size}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{state}_j && (2)\\
\alpha &\sim \text{Normal}(50, 10) && (3)\\
\beta_1, \beta_2, \beta_3 &\sim \text{Normal}(0, 2.5) && (4)\\
\gamma_j &\sim \text{Normal}(0, 2.5) && (5)\\
\sigma &\sim \text{Exponential}(1) && (6)
\end{aligned}
$$

In this setup, the intercept α represents the baseline poll result, while each β coefficient captures the specific influence of its associated predictor on the vote percentage. The variable state_j accounts for effects attributed to different states, while end_date_i models temporal trends.

Priors are defined as follows: *alpha* follows a normal distribution with a mean of 50 and a standard deviation of 10. Each of the β and γ parameters has a normal prior centered at 0 with a standard deviation of 2.5, indicating a mild assumption about their potential effects without introducing bias. The noise parameter σ adheres to an exponential distribution, facilitating flexibility in the unexplained variation.

We run the model using the R Core Team (2023) package of Goodrich et al. (2024), focusing on how the vote percentage varies based on the predictors.

The complete model, summarizing the components, can be expressed as:

$$
y_i \sim \text{Normal}\left(\beta_0 + \beta_1 \cdot \text{end\_date}_i + \sum_{k=1}^{K} \gamma_k B_k(\text{state}_j), \sigma^2\right)
$$

In this expression, y_i represents the predicted vote percentage for observation i, modeled as a function of the predictors, where the mean μ_i captures the expected percentage, and σ reflects the variability.

To further clarify the model:

- **Formula (1)** shows that the response variable y_i is normally distributed, with μ_i as the expected value and σ representing standard deviation.
- **Formula (2)** illustrates that μ_i is derived from the intercept α and the predictors.
- **Formula (3)** details the basis functions applied to the state variable, allowing for a nuanced relationship between states and vote percentages. The prior distributions are:
- **Formula (4)** specifies that α has a normal distribution with mean 50 and variance 10^2, indicating our expectation of average vote percentage.
- **Formula (5)** outlines the priors for γ_k, which are normal distributions centered at 0 with variance 2.5^2, reflecting the belief in potentially small effects from state variables.
- **Formula (6)** assigns an exponential distribution to the standard deviation σ, suggesting that smaller variance is more likely while allowing for broader variations.

4.2 Bayesian Priors

For Bayesian inference, we apply the following prior distributions:

- **Intercept prior**: A `Normal(50, 10)` prior is used for the intercept, reflecting a moderately informative belief that baseline support percentage falls within the plausible range (centered around 50% with a moderate spread).
- **Coefficients prior**: A `Normal(0, 2.5)` prior is assigned to each regression coefficient, which is weakly informative and allows flexibility without over-restricting the influence of predictors.

These priors are selected to strike a balance between allowing the data to inform estimates while providing regularization to prevent extreme parameter values. The priors are justified based on general expectations of support ranges and aim to avoid overfitting.

4.3 Assumptions and Limitations

1. **Linearity**: The model assumes a linear relationship between independent variables (predictors) and the dependent variable (outcome, `pct`). The expected outcome is represented as a linear combination of the predictors. Nonlinear effects, if present, may not be fully captured, potentially affecting the model's fit.

2. **Normality of Errors**: It is assumed that the residuals (the differences between observed and predicted values) follow a normal distribution. This assumption is crucial for making valid inferences; violations can lead to incorrect conclusions and unreliable credible intervals.
3. **Homoscedasticity**: The variance of the residuals is expected to be constant across all combinations of the predictors. If residuals display heteroscedasticity (non-constant variance), it can result in inefficient estimates and biased interpretations of the modeled relationships.
4. **Independence of Errors**: The model assumes that errors are independent of one another, meaning the error term for any single observation does not affect the error term for another observation. However, polling data often includes repeated measures from the same pollster, which may introduce dependence among observations.
5. **Additivity**: Predictors are assumed to additively influence `pct`, meaning interactions or non-additive effects are not considered. Future versions of the model might benefit from testing interactions, such as between state and candidate.
6. **Prior Distributions**: Bayesian regression requires the specification of prior distributions for the model parameters (coefficients). The selection of these priors is significant, as they can impact the resulting posterior distributions, particularly when data is limited.
7. **Parameter Estimation**: In Bayesian linear regression, parameters are estimated using a posterior distribution derived from both the likelihood of the observed data and the prior distributions. This framework allows for uncertainty quantification and more flexible inference.
8. **Error Distribution**: The model presumes normally distributed residuals. Any deviation from this assumption may influence the validity of the inference. Diagnostic checks and posterior predictive checks are used to assess model fit and adherence to assumptions.

4.4 Software and Validation

The model is implemented using the `rstanarm` package in R, which allows for efficient Bayesian inference with automatic convergence diagnostics. Model fit and convergence were monitored through trace plots and effective sample sizes. To further ensure robustness, we performed posterior predictive checks using `pp_check(model_bayes)` to evaluate model fit and identify any systematic deviations.

Overall, this Bayesian model provides a balanced approach, capturing key effects while maintaining interpretability and flexibility, making it suitable for predicting candidate support percentages in the election context.

Table 2: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

4.4.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

5 Results

Our results are summarized in Table 2.

6 Discussion

6.1 First discussion point

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

6.2 Second discussion point

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

6.3 Third discussion point

6.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

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

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

B.2 Diagnostics

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

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

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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.

FiveThirtyEight. 2024. "Presidential General Election Polls (Current Cycle)." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

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

———. 2024. "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/.

———. 2024. *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://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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.