Donald Trump projected to win Popular Vote in 2024 US Election with 47.75% of Vote*

Yuxin Sun

November 4, 2024

The 2024 U.S. Presidential Election has intensified, with Vice President Kamala Harris and former President Donald Trump locked in a close race marked by a narrow and fluctuating gap. This paper utilizes a poll-of-polls approach combined with a Bayesian generalized linear model to project the popular vote share for these two leading candidates. By applying a time-decayed weighting system within the last 30 days and focusing on high-quality pollsters, while also accounting for the dynamics of swing states, the analysis predicts that Trump holds a slight edge over Harris, with an estimated 47.75% of nationwide support. These projections are crucial for understanding state-specific priorities and offer strategic insights that could shape campaign efforts and voter outreach initiatives.

1 Introduction

The 2024 United States presidential election is poised to be one of the most consequential and unprecedented in the nation's history. A series of dramatic events have already reshaped the political landscape: former President Donald Trump survived an assassination attempt during a public speech, forcing heightened security measures, while the Democratic candidate unexpectedly shifted from President Joe Biden to Vice President Kamala Harris. Additionally, natural disasters such as a devastating hurricane have injected further uncertainty, complicating campaign efforts across key states. A groundbreaking TV candidate forum has also set a new precedent, steering public discourse and influencing voter sentiment in unforeseen ways. This election is the first major national vote following the global COVID-19 pandemic, and the rapid spread of ideas through social media has made the political conversation more volatile and polarized than ever. In this historic race, Vice President Kamala Harris represents the Democratic Party, leveraging her tenure as Biden's second-in-command and a track record of pushing progressive policies. On the other side, Donald Trump stands as the Republican

*Code and data are available at: https://github.com/Yuxin-Sun-Caroline/US_Presidential_Forecast_2024 contender, having maintained a dedicated and fervent base since his first presidency, characterized by his signature populist and nationalist rhetoric. Both candidates bring contrasting visions for the future, fueling a deeply polarized electorate.Understanding these dynamics is crucial to grasping the complexities of this election and anticipating the possible trajectories that could define the future of the United States. In this paper, the estimand under investigation is the popular vote for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election. To forecast the outcomes, I employ a Bayesian generalized linear model, incorporating the following predictors: the age of days before the election, a combined weight that integrates time-decayed influence and the numeric grade reflecting the quality of the pollster, and a final state weight based on electoral votes, referencing the state-wise population and the strategic importance of swing states. This multi-faceted approach aims to capture the nuances of voter sentiment across the nation. Using the poll-of-polls method, which aggregates various state-level polls to enhance the accuracy of national predictions, the model forecasts a razor-thin popular vote margin, with Trump at 47.75% and Harris at 47.73%. This result underscores the intense competitiveness of the race, with only a 0.02% difference in national support between the two main candidates. State-level analysis reveals crucial insights into the dynamics of swing states. Arizona, Georgia, and North Carolina are currently leaning toward Trump, suggesting a potential advantage for the Republican nominee. Conversely, Harris shows stronger support in Michigan and Wisconsin. Meanwhile, Pennsylvania and Nevada emerge as critical battlegrounds with exceptionally close competition, making them pivotal to the overall outcome of the election. Understanding state preferences and their impact on the national support ratio is essential for predicting electoral results. This analysis highlights how regional trends could influence strategic decisions by both campaigns, shaping their voter outreach and resource allocation in a highly contested election year.

2 Data

2.1 Overview

The data utilized in this analysis is sourced from the 'Presidential General Election Polls (Current Cycle)' by FiveThirtyEight, with poll results updated through November 1st. The original dataset comprises a comprehensive collection of polls, capturing the preferences of voters across various states and conducted by different polling organizations.

Here are the key variables employed in this analysis, described in simplified terms:

pollster: The name of the polling organization that conducted each poll, which helps identify the source and potential reliability. numeric_grade: A numeric rating that reflects the reliability of each pollster, allowing for adjustments in analysis based on poll quality. state: The specific U.S. state where the poll was conducted, highlighting the geographical focus of the data and relevance to state-level electoral analysis. start_date and end_date: The dates marking when each poll began and ended, used to track the timing and freshness of the data. candidate_name: The name of the candidate receiving support in the poll, essential for categorizing voter preferences. pct: The percentage of respondents who support each candidate in the poll, a critical metric for determining popular support. New Variables Created for the Analysis: final_state_weight: A calculated weight that combines state population size and the strategic importance of swing states. This variable emphasizes the electoral impact of key battleground states, enhancing the model's sensitivity to regional significance. combined_weight: A comprehensive weight incorporating a 30-day half-life decay function (to prioritize recent polls) and a quality decay factor (to down-weight lower-quality pollsters and give higher significance to reliable sources). This ensures that the analysis prioritizes up-to-date and high-quality data. mid_date: The midpoint date between the start and end dates of each poll, used to represent the exact time the poll data reflects. age_in_days: The number of days between the mid_date and November 5, capturing the recency of the poll and informing the time decay weighting. The model places higher significance on polls that are conducted closer to the election date, are of better quality, and focus on swing states, as these factors are critical for achieving a more accurate forecast. By carefully weighting these variables, the model accounts for the intense and unpredictable nature of the 2024 election landscape, aiming for a robust and precise prediction.

Pollsters utilize a range of survey methods to translate the complexities of voter behavior into actionable numerical data. A reliable pollster is one who conducts surveys using well-designed and unbiased questions, selects a representative sample of the population, and employs transparent and scientifically sound methodologies. The sampling methods are critical: national polls aim to capture broad trends across the United States, while state polls often focus on regional dynamics and voter preferences in specific states. High-quality pollsters also take care to ask demographic questions, such as age, gender, and education, to ensure that the sample accurately reflects the broader electorate. ## Methodology and Measurement The forecasting method used in this analysis is the "poll-of-polls" approach. This method aggregates both national and state-level polls to create a more robust and accurate prediction model. The dataset combines national polls, which generally have higher quality but tend to be slightly older, and state polls, which are more recent but often vary in quality and reflect specific state preferences. To manage these differences, we apply a reweighting process. First, we calculate the mid_date of each poll and determine the age_in_days, which measures how long it has been since the poll was conducted, relative to November 5. We then filter out polls that are more than 30 days old and apply a half-life decay weight, giving more significance to more recent polling data.

Once the dataset is filtered for recency, we separate the data into state and national polls. For standardization, we simplify district names (e.g., "Maine Cd-1" and "Nebraska Cd-1") to their respective state names. To create population weights, we match the voting data for each state with its corresponding name and exclude states for which we have insufficient data. We then calculate a proportion weight for each state, reflecting both its population size and electoral significance. This cleaning and weighting process ensures that our data is reliable, up-to-date, and capable of accurately capturing the nuanced dynamics of the 2024 election landscape.

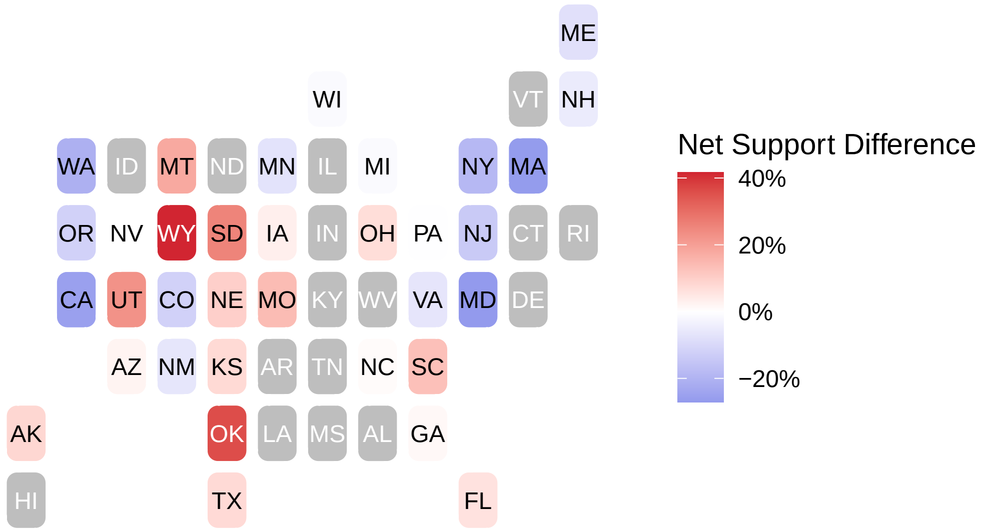
This structured approach helps balance the value of fresh, state-specific data with the reliability of national polling, providing a comprehensive and well-weighted dataset for making informed electoral predictions.

2.2 Data Visualization

The graph below employs a sophisticated weighted-averaging approach to analyze the state-level electoral landscape between Donald Trump and Kamala Harris. The methodology calculates state-specific support levels by applying a combined weight that accounts for both poll quality and temporal relevance. The analysis creates a net support differential metric, computed as the difference between Trump's and Harris's weighted average support in each state, which is then visualized using a statebins representation of the U.S. map. The color gradient scheme employs royalblue for Harris-leaning states and red (#d12531) for Trump-leaning states, with white indicating closely contested regions. States without polling data are represented in grey, acknowledging data limitations. This visualization methodology is particularly effective as it avoids the common distortion of traditional geographic maps where larger states appear more significant, instead using equally-sized bins to represent each state's electoral status, thereby providing a more balanced representation of the electoral landscape. he graph clearly illustrates the distinct state-level preferences for Trump and Harris. States like Wyoming, South Dakota, and Oklahoma show a strong and consistent preference for Trump, whereas California, Massachusetts, Maryland, and Washington are firmly aligned with Harris. However, the chart also highlights the less definitive, contested areas: Arizona, Georgia, North Carolina, Michigan, Wisconsin, Pennsylvania, and Nevada emerge as key swing states where voter preference is less clear and color-coded representations are more ambiguous. These swing states are pivotal in the overall electoral outcome, warranting a focused and deeper analysis to understand the nuanced dynamics and factors influencing voter behavior in these regions.

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(tidyr)  # For pivot_wider
library(statebins)

# Calculate state-wise weighted average support for Trump and Harris
state_support <- state_polls %>%
  filter(candidate_name %in% c("Donald Trump", "Kamala Harris")) %>%
  group_by(state, candidate_name) %>%
  summarise(
    average_support_ratio = sum(pct * combined_weight, na.rm = TRUE) / sum(combined_weight, n
    .groups = 'drop'
  ) %>%
  pivot_wider(names_from = candidate_name, values_from = average_support_ratio)  # Use pivot_
```

```
# Create a dataframe with all states
all_states <- data.frame(
  state = c(
    "Alabama", "Alaska", "Arizona", "Arkansas", "California",
    "Colorado", "Connecticut", "Delaware", "Florida", "Georgia",
    "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa",
    "Kansas", "Kentucky", "Louisiana", "Maine", "Maryland",
    "Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri",
    "Montana", "Nebraska", "Nevada", "New Hampshire", "New Jersey",
    "New Mexico", "New York", "North Carolina", "North Dakota",
    "Ohio", "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island",
    "South Carolina", "South Dakota", "Tennessee", "Texas",
    "Utah", "Vermont", "Virginia", "Washington", "West Virginia",
    "Wisconsin", "Wyoming"
  )
)

# Merge with calculated support data
map_data <- all_states %>%
  left_join(state_support, by = "state") %>%
  mutate(
    # Calculate the net support difference: positive for Trump, negative for Harris
    net_support = `Donald Trump` - `Kamala Harris`
  )

# Plotting the state-level support using a gradient color scale
ggplot(map_data, aes(fill = net_support, state = state)) +
  geom_statebins() +
  scale_fill_gradient2(
    low = "royalblue", high = "#d12531", mid = "white",
    midpoint = 0, na.value = "grey",
    name = "Net Support Difference",
    labels = scales::percent_format(scale = 1)
  ) +
  theme_classic() +
  theme(
    axis.line = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank()
  ) +
  labs(
    fill = "Net Support",
    title = "U.S. Map of Weighted Support for Trump vs. Harris",

  )
```


U.S. Map of Weighted Support for Trump vs. Harris

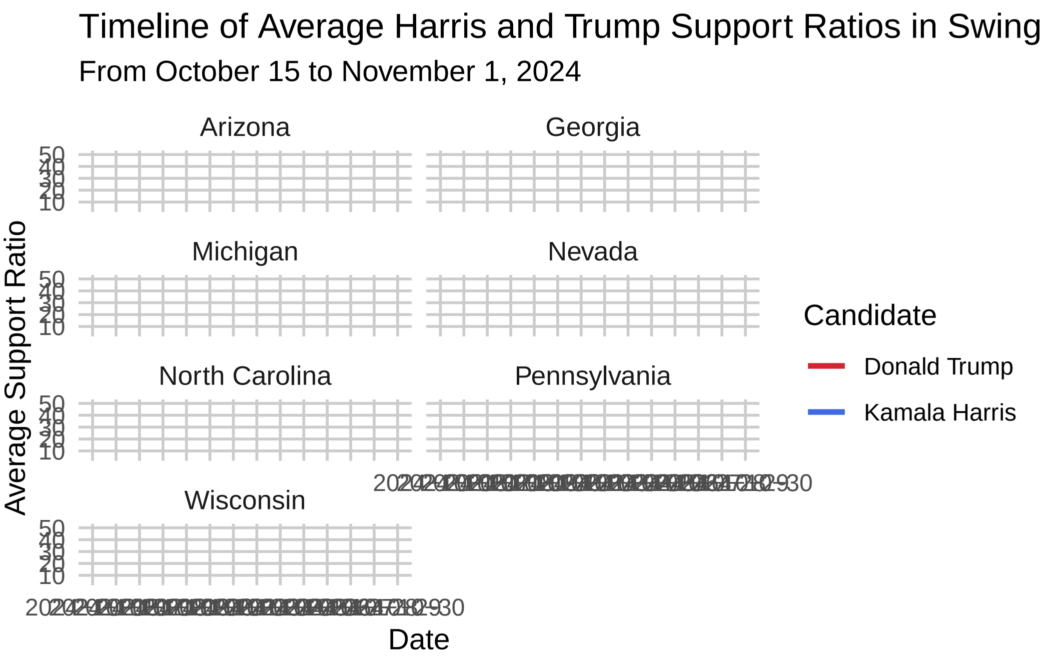
The graph below reveals a time-based trend for swing states between October 15 and November 1. In Arizona, Georgia, and North Carolina, there is a noticeable lean towards Trump, suggesting a higher likelihood of these states favoring the Republican candidate. On the other hand, Michigan and Wisconsin display a consistent preference for Harris, indicating stronger support for the Democrats in these areas. However, Nevada and Pennsylvania present a more ambiguous picture, with no clear trend emerging for either candidate. To enhance the accuracy of our predictions, we will increase the significance of these critical swing states in our analysis, focusing on the uncertainty and potential impact they may have on the overall election outcome.

```
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Define the swing states
swing_states <- c("Nevada", "Wisconsin", "Michigan", "Pennsylvania",
                  "North Carolina", "Arizona", "Georgia")

# Filter the state_polls for swing states, both candidates, and the date range
swing_state_polls <- state_polls %>%
  filter(
    state %in% swing_states &
    candidate_name %in% c("Donald Trump", "Kamala Harris") &
    mid_date >= as.Date("2024-10-15") & mid_date <= as.Date("2024-11-01")
  ) %>%
  group_by(mid_date, state, candidate_name) %>%
  summarise(
    average_support_ratio = mean(pct * combined_weight, na.rm = TRUE) / sum(combined_weight,
    .groups = 'drop'
  )

# Plotting the timeline of average support ratios from October 15 to November 1
ggplot(swing_state_polls, aes(x = mid_date, y = average_support_ratio, color = candidate_name
  geom_line(size = 1) +
  facet_wrap(~ state, ncol = 2) +  # Create separate panels for each swing state
  scale_color_manual(values = c("Donald Trump" = "#d12531", "Kamala Harris" = "royalblue")) +
  labs(
    title = "Timeline of Average Harris and Trump Support Ratios in Swing States",
    subtitle = "From October 15 to November 1, 2024",
    x = "Date",
    y = "Average Support Ratio",
    color = "Candidate"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_line(color = "grey80"),
    panel.grid.minor = element_blank(),
    strip.text = element_text(size = 10)
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```


Timeline of Average Harris and Trump Support Ratios in Swing

From October 15 to November 1, 2024

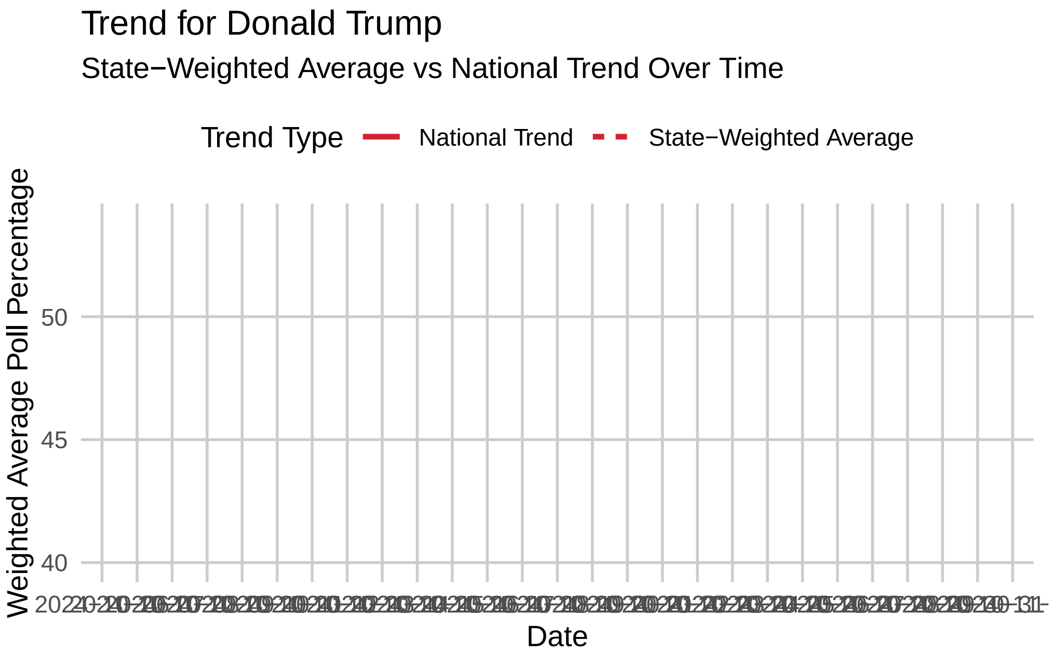
```r
# Load necessary libraries
library(dplyr)

# Step 1: Define Electoral Votes for Each State
electoral_votes <- data.frame(
  state = c("Alabama", "Kentucky", "North Dakota", "Alaska", "Louisiana",
            "Ohio", "Arizona", "Maine", "Oklahoma", "Arkansas",
            "Maryland", "Oregon", "California", "Massachusetts", "Pennsylvania",
            "Colorado", "Michigan", "Rhode Island", "Connecticut", "Minnesota",
            "South Carolina", "Delaware", "Mississippi", "South Dakota",
            "District of Columbia", "Missouri", "Tennessee", "Florida",
            "Montana", "Texas", "Georgia", "Nebraska", "Utah",
            "Hawaii", "Nevada", "Vermont", "Idaho", "New Hampshire",
            "Virginia", "Illinois", "New Jersey", "Washington", "Indiana",
            "New Mexico", "West Virginia", "Iowa", "New York", "Wisconsin",
            "Kansas", "North Carolina", "Wyoming"),
  electoral_votes = c(9, 8, 3, 3, 8, 17, 11, 4, 7, 6, 10, 8, 54, 11, 19, 10, 15, 4, 7, 10,
                      9, 3, 6, 3, 3, 10, 11, 30, 4, 40, 16, 5, 6, 4, 6, 3, 4, 4, 13, 19, 14,
                      12, 11, 5, 4, 6, 28, 10, 6, 16, 3)
)
electoral_votes <- electoral_votes %>%
  filter(state %in% state_polls$state)

# Calculate the total number of electoral votes
total_electoral_votes <- sum(electoral_votes$electoral_votes)

# Calculate normalized electoral weight
electoral_votes <- electoral_votes %>%
  mutate(electoral_weight = electoral_votes / total_electoral_votes)

# Step 2: Apply Swing State Adjustment
# Define the swing states and their mild adjustment factor
swing_states <- c("Nevada", "Wisconsin", "Michigan", "Pennsylvania",
                  "North Carolina", "Arizona", "Georgia")
swing_state_adjustment <- 1.2  # Mild adjustment factor

# Add the swing state weight adjustment
electoral_votes <- electoral_votes %>%
  mutate(
    swing_weight = ifelse(state %in% swing_states, swing_state_adjustment, 1),
    state_weight = electoral_weight * swing_weight  # Calculate the adjusted state weight
  )

# Normalize the state weights so they sum to 1
total_state_weight <- sum(electoral_votes$state_weight)
electoral_votes <- electoral_votes %>%
  mutate(final_state_weight = state_weight / total_state_weight)  # Renamed to "final_state_

# Display the final state weights
#print(electoral_votes)
state_polls <- state_polls %>%
  left_join(electoral_votes %>% select(state, final_state_weight), by = "state")
```

```
# Load required libraries
library(tidyverse)
library(lubridate)

# 1. First ensure electoral_votes is properly set up
```

```
electoral_votes <- data.frame(
  state = c("Alabama", "Kentucky", "North Dakota", "Alaska", "Louisiana",
            "Ohio", "Arizona", "Maine", "Oklahoma", "Arkansas",
            "Maryland", "Oregon", "California", "Massachusetts", "Pennsylvania",
            "Colorado", "Michigan", "Rhode Island", "Connecticut", "Minnesota",
            "South Carolina", "Delaware", "Mississippi", "South Dakota",
            "District of Columbia", "Missouri", "Tennessee", "Florida",
            "Montana", "Texas", "Georgia", "Nebraska", "Utah",
            "Hawaii", "Nevada", "Vermont", "Idaho", "New Hampshire",
            "Virginia", "Illinois", "New Jersey", "Washington", "Indiana",
            "New Mexico", "West Virginia", "Iowa", "New York", "Wisconsin",
            "Kansas", "North Carolina", "Wyoming"),
  electoral_votes = c(9, 8, 3, 3, 8, 17, 11, 4, 7, 6, 10, 8, 54, 11, 19, 10, 15, 4, 7, 10,
                      9, 3, 6, 3, 3, 10, 11, 30, 4, 40, 16, 5, 6, 4, 6, 3, 4, 4, 13, 19, 14,
                      12, 11, 5, 4, 6, 28, 10, 6, 16, 3)
)

# 2. Calculate state weights
electoral_votes <- electoral_votes %>%
  mutate(
    electoral_weight = electoral_votes / sum(electoral_votes),
    swing_weight = ifelse(state %in% c("Nevada", "Wisconsin", "Michigan", "Pennsylvania",
                                      "North Carolina", "Arizona", "Georgia"), 1.2, 1),
    state_weight = electoral_weight * swing_weight,
    final_state_weight = state_weight / sum(state_weight)
  )

# 3. Standardize state names
electoral_votes <- electoral_votes %>%
  mutate(state = str_to_title(str_trim(state)))

# 4. Process state polls
state_trend <- state_polls %>%
  filter(candidate_name %in% c("Donald Trump", "Kamala Harris")) %>%
  # Group by all necessary variables including state
  group_by(mid_date, candidate_name, state) %>%
  # First get average by state and date
  summarise(
    avg_pct = mean(pct, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  # Join with electoral votes to get weights
```

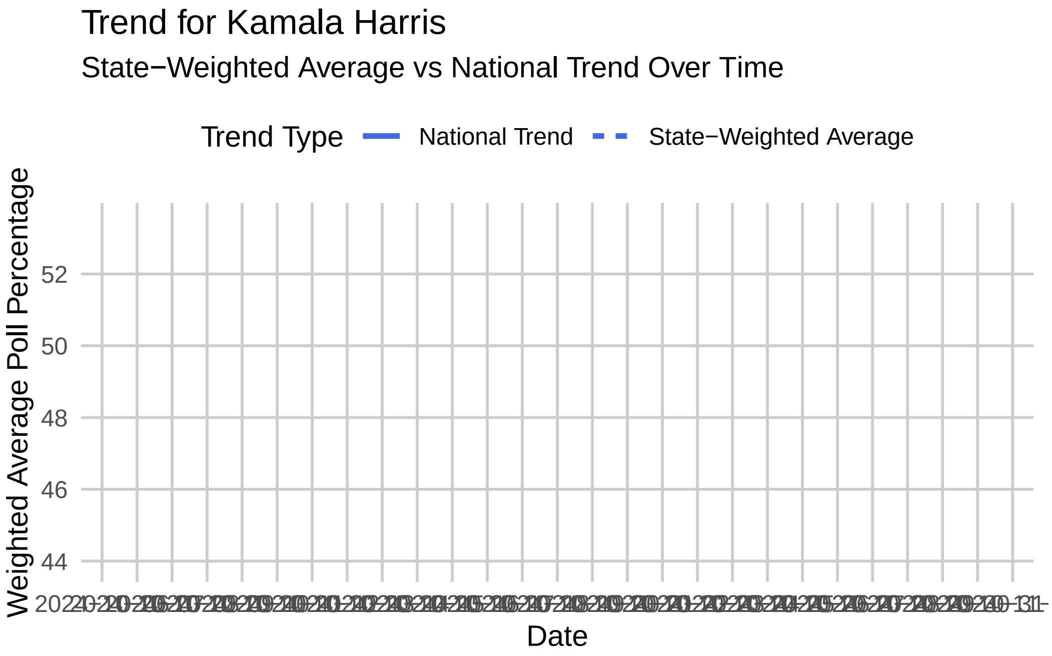
```r
  left_join(electoral_votes %>% select(state, final_state_weight), by = "state") %>%
  # Now group by date and candidate to get weighted averages
  group_by(mid_date, candidate_name) %>%
  summarise(
    weighted_avg_pct = weighted.mean(avg_pct, final_state_weight, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  mutate(type = "State-Weighted Average")

# 5. Process national polls
national_trend <- national_polls %>%
  filter(candidate_name %in% c("Donald Trump", "Kamala Harris")) %>%
  group_by(mid_date, candidate_name) %>%
  summarise(
    weighted_avg_pct = weighted.mean(pct, combined_weight, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  mutate(type = "National Trend")

# 6. Combine trends
combined_trend <- bind_rows(state_trend, national_trend)

# 7. Create separate trends for each candidate
trump_trend <- combined_trend %>%
  filter(candidate_name == "Donald Trump")

harris_trend <- combined_trend %>%
  filter(candidate_name == "Kamala Harris")

# 8. Create plots
# Trump plot
trump_plot <- ggplot(trump_trend, aes(x = mid_date, y = weighted_avg_pct, linetype = type)) +
  geom_line(color = "#d12531", size = 1) +
  labs(
    title = "Trend for Donald Trump",
    subtitle = "State-Weighted Average vs National Trend Over Time",
    x = "Date",
    y = "Weighted Average Poll Percentage",
    linetype = "Trend Type"
  ) +
  theme_minimal() +
  theme(
    legend.position = "top",
    panel.grid.major = element_line(color = "grey80"),
    panel.grid.minor = element_blank()
  )

# Harris plot
harris_plot <- ggplot(harris_trend, aes(x = mid_date, y = weighted_avg_pct, linetype = type)
  geom_line(color = "royalblue", size = 1) +
  labs(
    title = "Trend for Kamala Harris",
    subtitle = "State-Weighted Average vs National Trend Over Time",
    x = "Date",
    y = "Weighted Average Poll Percentage",
    linetype = "Trend Type"
  ) +
  theme_minimal() +
  theme(
    legend.position = "top",
    panel.grid.major = element_line(color = "grey80"),
    panel.grid.minor = element_blank()
  )

# 9. Display plots
print(trump_plot)
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```



```
print(harris_plot)
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```



```
# 10. Print summary statistics
cat("\nFinal Weighted Averages:\n")
```

```
Final Weighted Averages:
```

```
combined_trend %>%
  group_by(candidate_name, type) %>%
  summarise(
    avg_support = mean(weighted_avg_pct, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  arrange(type, desc(avg_support)) %>%
  print()
```

```
# A tibble: 4 x 3
  candidate_name type                    avg_support
  <chr>          <chr>                         <dbl>
1 Kamala Harris  National Trend                 48.5
2 Donald Trump   National Trend                 47.0
3 Kamala Harris  State-Weighted Average         48.4
4 Donald Trump   State-Weighted Average         46.7
```

2.3 Outcome variables

3 Model

3.1 Model Results

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.2 Model Setup

$y_i|\mu_i \sim \text{Normal}(\mu_i, \sigma)$ $\mu_i \quad = \beta_0 + \beta_1 \times \text{IsTrump}_i + \beta_2 \times \text{CombinedWeight}_i + \beta_3 \times \text{FinalStateWeight}_i + \beta_4 \times \text{Nu}$

where: y_i is the polling percentage for observation i IsTrump_i is a binary indicator (1 for Trump, 0 for Harris) CombinedWeight_i is the combined poll weight (quality × temporal) $\text{FinalStateWeight}_i$ is the electoral importance weight NumericGrade_i is the poll quality grade AgeDays_i is the age of the poll in days $\alpha_{\text{state}[i]}$ represents state-level random effects τ is the standard deviation of the state-level random effects

We implemented this model in R using the `rstanarm` package [**?**]. The priors were chosen to be weakly informative to allow for flexibility without overly constraining the parameter space.

3.2.1 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.

4 Results

5 Discussion

5.1 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.

5.2 Second discussion point

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

5.3 Third discussion point

5.4 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

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...

‘

C References