Final Project EDA and Data Evaluation

Your Name

November 2024

Question 3: Evaluating Data Quality

This question encourages you to begin the exploratory data analysis (EDA) for your final project. By addressing potential data quality issues early, you can identify and rectify problems promptly. For each important variable in your dataset, assess its quality by creating a table that includes the following:

- **Continuous variables**:
  - The number of non-missing observations.
  - The number of missing observations.
  - Measures of central tendency (e.g., mean, median).
  - Measures of variability (e.g., standard deviation, interquartile range [IQR]).
- **Categorical variables**:
  - The levels of the variable.
  - For each level:
    - The number of non-missing observations.
    - The number of missing observations.

Answer: Comparing Predictive Performance of Betting Markets and Polls

```
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Our project aims to compare the predictive accuracy of betting markets and polls in forecasting the outcome of the 2024 U.S. presidential election by state. We use the following datasets:

1. **Betting Market Data**
   - Sourced from a Kaggle dataset, this dataset scrapes Polymarket's 2024 election results by state and consolidates them into a CSV file (source).
2. **Poll Data**
   - Sourced from FiveThirtyEight, this dataset provides polling averages and raw data from the 2024 U.S. presidential election (source).
3. **Actual Election Results**
   - Sourced from CBS News, this dataset reports the official 2024 presidential election results (source).

Exploratory Data Analysis (EDA) Dataset 1: Betting Market Data

The dataset provides separate files for each state, containing the probability of a Republican win under the column "Donald Trump" and a Democratic win under "Kamala Harris." These probabilities are based on the amount bet on Polymarket for each candidate. Data was aggregated by month, resulting in a final dataset of state-level probabilities for Republican and Democratic wins from April 17, 2024, to November 4, 2024.

Columns Included "Date (UTC)","Timestamp (UTC)","Donald Trump","Kamala Harris", and "Other". All values of the table were continuous, enumerating the date, time, percentage probability for Trump, percentage probability for Kamala, and percentage probability for a different candidate. No cateogrical variables existed.

First non-missing and missing observations were quantified for each probability column. Mean probabilities and standard deviations for Republican and Democratic wins nationwide were calculated.

```
#Access csv files in directory
file_path <- "data/betting_data/polymarket/csv_month/"

file_list <- list.files(path = file_path, pattern = "*.csv", full.names = TRUE)

# Data frames to store results
final_data <- data.frame()
missing_data_summary <- data.frame()

for (file in file_list) {
  state_data <- read_csv(file, show_col_types = FALSE)

  # Extract the state abbreviation from the file name
  state_abbrev <- tools::file_path_sans_ext(basename(file)) %>%
    stringr::str_extract("^[A-Z]{2}")

  # Calculate averages for  Trump and Harris
  avg_trump <- mean(state_data$`Donald Trump`, na.rm = TRUE)
  avg_harris <- mean(state_data$`Kamala Harris`, na.rm = TRUE)

  # Count missing and non-missing observations for each candidate
  non_missing_trump <- sum(!is.na(state_data$`Donald Trump`))
  missing_trump <- sum(is.na(state_data$`Donald Trump`))
  non_missing_harris <- sum(!is.na(state_data$`Kamala Harris`))
  missing_harris <- sum(is.na(state_data$`Kamala Harris`))

  # Append missing data summary for this state
  missing_data_summary <- bind_rows(
    missing_data_summary,
    data.frame(
      state = state_abbrev,
      candidate = "Donald Trump",
      non_missing = non_missing_trump,
```

```
      missing = missing_trump
    ),
    data.frame(
      state = state_abbrev,
      candidate = "Kamala Harris",
      non_missing = non_missing_harris,
      missing = missing_harris
    )
  )

  # Create a new data structure for average percentages
  state_results <- data.frame(
    candidate = c("Donald Trump", "Kamala Harris"),
    percentage = c(avg_trump, avg_harris),
    state = c(state_abbrev, state_abbrev)
  )

  final_data <- bind_rows(final_data, state_results)
}

# Calculate nationwide statistics
nationwide_stats <- final_data %>%
  group_by(candidate) %>%
  summarise(
    nationwide_mean = mean(percentage, na.rm = TRUE),
    nationwide_sd = sd(percentage, na.rm = TRUE),
    total_non_missing = sum(!is.na(percentage)),
    total_missing = sum(is.na(percentage))
  )

print(missing_data_summary)
```

```
##     state     candidate non_missing missing
## 1      AK  Donald Trump           8       0
## 2      AK Kamala Harris           8       0
## 3      AL  Donald Trump           8       0
## 4      AL Kamala Harris           8       0
## 5      AR  Donald Trump           8       0
## 6      AR Kamala Harris           8       0
## 7      AZ  Donald Trump           9       0
## 8      AZ Kamala Harris           9       0
## 9      CA  Donald Trump           8       0
## 10     CA Kamala Harris           8       0
## 11     CO  Donald Trump           8       0
## 12     CO Kamala Harris           8       0
## 13     CT  Donald Trump           8       0
## 14     CT Kamala Harris           8       0
## 15     DE  Donald Trump           8       0
## 16     DE Kamala Harris           8       0
## 17     FL  Donald Trump           8       0
## 18     FL Kamala Harris           8       0
## 19     GA  Donald Trump           9       0
## 20     GA Kamala Harris           9       0
## 21     HI  Donald Trump           8       0
```

```
## 22     HI Kamala Harris          8       0
## 23     IA  Donald Trump          8       0
## 24     IA Kamala Harris          8       0
## 25     ID  Donald Trump          8       0
## 26     ID Kamala Harris          8       0
## 27     IL  Donald Trump          8       0
## 28     IL Kamala Harris          8       0
## 29     IN  Donald Trump          8       0
## 30     IN Kamala Harris          8       0
## 31     KS  Donald Trump          8       0
## 32     KS Kamala Harris          8       0
## 33     KY  Donald Trump          8       0
## 34     KY Kamala Harris          8       0
## 35     LA  Donald Trump          8       0
## 36     LA Kamala Harris          8       0
## 37     MA  Donald Trump          8       0
## 38     MA Kamala Harris          8       0
## 39     MD  Donald Trump          8       0
## 40     MD Kamala Harris          8       0
## 41     ME  Donald Trump          8       0
## 42     ME Kamala Harris          8       0
## 43     MI  Donald Trump          9       0
## 44     MI Kamala Harris          9       0
## 45     MN  Donald Trump          8       0
## 46     MN Kamala Harris          8       0
## 47     MO  Donald Trump          8       0
## 48     MO Kamala Harris          8       0
## 49     MS  Donald Trump          8       0
## 50     MS Kamala Harris          8       0
## 51     MT  Donald Trump          8       0
## 52     MT Kamala Harris          8       0
## 53     NC  Donald Trump          9       0
## 54     NC Kamala Harris          9       0
## 55     ND  Donald Trump          8       0
## 56     ND Kamala Harris          8       0
## 57     NE  Donald Trump          8       0
## 58     NE Kamala Harris          8       0
## 59     NH  Donald Trump          8       0
## 60     NH Kamala Harris          8       0
## 61     NJ  Donald Trump          8       0
## 62     NJ Kamala Harris          8       0
## 63     NM  Donald Trump          8       0
## 64     NM Kamala Harris          8       0
## 65     NV  Donald Trump          9       0
## 66     NV Kamala Harris          9       0
## 67     NY  Donald Trump          8       0
## 68     NY Kamala Harris          8       0
## 69     OH  Donald Trump          8       0
## 70     OH Kamala Harris          8       0
## 71     OK  Donald Trump          8       0
## 72     OK Kamala Harris          8       0
## 73     OR  Donald Trump          8       0
## 74     OR Kamala Harris          8       0
## 75     PA  Donald Trump          9       0
```

```
## 76     PA Kamala Harris           9       0
## 77     RI  Donald Trump           8       0
## 78     RI Kamala Harris           8       0
## 79     SC  Donald Trump           8       0
## 80     SC Kamala Harris           8       0
## 81     SD  Donald Trump           8       0
## 82     SD Kamala Harris           8       0
## 83     TN  Donald Trump           8       0
## 84     TN Kamala Harris           8       0
## 85     TX  Donald Trump           8       0
## 86     TX Kamala Harris           8       0
## 87     UT  Donald Trump           8       0
## 88     UT Kamala Harris           8       0
## 89     VA  Donald Trump           8       0
## 90     VA Kamala Harris           8       0
## 91     VT  Donald Trump           8       0
## 92     VT Kamala Harris           8       0
## 93     WA  Donald Trump           8       0
## 94     WA Kamala Harris           8       0
## 95     WI  Donald Trump           9       0
## 96     WI Kamala Harris           9       0
## 97     WV  Donald Trump           8       0
## 98     WV Kamala Harris           8       0
## 99     WY  Donald Trump           8       0
## 100    WY Kamala Harris           8       0
```

```
print(nationwide_stats)
```

```
## # A tibble: 2 x 5
##   candidate     nationwide_mean nationwide_sd total_non_missing total_missing
##   <chr>                   <dbl>         <dbl>             <int>         <int>
## 1 Donald Trump            0.558         0.404                50             0
## 2 Kamala Harris           0.431         0.406                50             0
```

While we will be aggregating over time, we nevertheless ran a missing information test on data and timestamp.

```
missing_data_summary <- data.frame()

for (file in file_list) {
  state_data <- read_csv(file, show_col_types = FALSE)

  # Extract the state abbreviation from the file name
  state_abbrev <- tools::file_path_sans_ext(basename(file)) %>%
    stringr::str_extract("^[A-Z]{2}")

  # Count missing and non-missing observations for each candidate
  non_missing_date <- sum(!is.na(state_data$`Date (UTC)`))
  missing_date <- sum(is.na(state_data$`Date (UTC)`))
  non_missing_timestamp <- sum(!is.na(state_data$`Timestamp (UTC)`))
  missing_timestamp <- sum(is.na(state_data$`Timestamp (UTC)`))

  # Append missing data summary for this state
  missing_data_summary <- bind_rows(
    missing_data_summary,
    data.frame(
      state = state_abbrev,
```

```
      variable = "Date",
      non_missing = non_missing_date,
      missing = missing_date
    ),
    data.frame(
      state = state_abbrev,
      variable = "Timestamp",
      non_missing = non_missing_timestamp,
      missing = missing_timestamp
    )
  )
}

print(missing_data_summary)
```

```
##     state  variable non_missing missing
## 1      AK      Date           8       0
## 2      AK Timestamp           8       0
## 3      AL      Date           8       0
## 4      AL Timestamp           8       0
## 5      AR      Date           8       0
## 6      AR Timestamp           8       0
## 7      AZ      Date           9       0
## 8      AZ Timestamp           9       0
## 9      CA      Date           8       0
## 10     CA Timestamp           8       0
## 11     CO      Date           8       0
## 12     CO Timestamp           8       0
## 13     CT      Date           8       0
## 14     CT Timestamp           8       0
## 15     DE      Date           8       0
## 16     DE Timestamp           8       0
## 17     FL      Date           8       0
## 18     FL Timestamp           8       0
## 19     GA      Date           9       0
## 20     GA Timestamp           9       0
## 21     HI      Date           8       0
## 22     HI Timestamp           8       0
## 23     IA      Date           8       0
## 24     IA Timestamp           8       0
## 25     ID      Date           8       0
## 26     ID Timestamp           8       0
## 27     IL      Date           8       0
## 28     IL Timestamp           8       0
## 29     IN      Date           8       0
## 30     IN Timestamp           8       0
## 31     KS      Date           8       0
## 32     KS Timestamp           8       0
## 33     KY      Date           8       0
## 34     KY Timestamp           8       0
## 35     LA      Date           8       0
## 36     LA Timestamp           8       0
## 37     MA      Date           8       0
## 38     MA Timestamp           8       0
```

```
## 39     MD      Date           8       0
## 40     MD Timestamp           8       0
## 41     ME      Date           8       0
## 42     ME Timestamp           8       0
## 43     MI      Date           9       0
## 44     MI Timestamp           9       0
## 45     MN      Date           8       0
## 46     MN Timestamp           8       0
## 47     MO      Date           8       0
## 48     MO Timestamp           8       0
## 49     MS      Date           8       0
## 50     MS Timestamp           8       0
## 51     MT      Date           8       0
## 52     MT Timestamp           8       0
## 53     NC      Date           9       0
## 54     NC Timestamp           9       0
## 55     ND      Date           8       0
## 56     ND Timestamp           8       0
## 57     NE      Date           8       0
## 58     NE Timestamp           8       0
## 59     NH      Date           8       0
## 60     NH Timestamp           8       0
## 61     NJ      Date           8       0
## 62     NJ Timestamp           8       0
## 63     NM      Date           8       0
## 64     NM Timestamp           8       0
## 65     NV      Date           9       0
## 66     NV Timestamp           9       0
## 67     NY      Date           8       0
## 68     NY Timestamp           8       0
## 69     OH      Date           8       0
## 70     OH Timestamp           8       0
## 71     OK      Date           8       0
## 72     OK Timestamp           8       0
## 73     OR      Date           8       0
## 74     OR Timestamp           8       0
## 75     PA      Date           9       0
## 76     PA Timestamp           9       0
## 77     RI      Date           8       0
## 78     RI Timestamp           8       0
## 79     SC      Date           8       0
## 80     SC Timestamp           8       0
## 81     SD      Date           8       0
## 82     SD Timestamp           8       0
## 83     TN      Date           8       0
## 84     TN Timestamp           8       0
## 85     TX      Date           8       0
## 86     TX Timestamp           8       0
## 87     UT      Date           8       0
## 88     UT Timestamp           8       0
## 89     VA      Date           8       0
## 90     VA Timestamp           8       0
## 91     VT      Date           8       0
## 92     VT Timestamp           8       0
```

```
## 93     WA      Date           8       0
## 94     WA Timestamp           8       0
## 95     WI      Date           9       0
## 96     WI Timestamp           9       0
## 97     WV      Date           8       0
## 98     WV Timestamp           8       0
## 99     WY      Date           8       0
## 100    WY Timestamp           8       0
```

Dataset 2: Poll Data

The dataset contains average polling data by candidate, date, and their adjusted percentage of being favored. We started out by exploring how many missing and non-missing data point there were for Harris and Trump. We then aggregated data over time, such that we are left with poll percentages for Harris/Trump per state. We take the mean to find the average poll percentage for Harris and Trump nationwide. We then take the standard deviation.

Dataset 3: Election Results

The dataset contains the number of votes for Harris and Trump, separated by state. We find the percentage who voted for Harris and Trump, the amount of missing and non-missing data, and the mean and variance.

```
actual_results <- read_csv("data/actual_results_data/state_results_2024.csv")
```

```
## Rows: 51 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): State
## dbl (2): Harris_Votes, Trump_Votes
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Add percentage columns for Harris and Trump
actual_results <- actual_results %>%
  mutate(
    percent_harris = (Harris_Votes / (Harris_Votes + Trump_Votes)) * 100,
    percent_trump = (Trump_Votes / (Harris_Votes + Trump_Votes)) * 100
  )

# Calculate the number of missing and non-missing observations for each column
missing_summary <- actual_results %>%
  summarise(
    missing_harris = sum(is.na(Harris_Votes)),
    non_missing_harris = sum(!is.na(Harris_Votes)),
    missing_trump = sum(is.na(Trump_Votes)),
    non_missing_trump = sum(!is.na(Trump_Votes)),
    missing_state = sum(is.na(State)),
    non_missing_trump = sum(!is.na(State))
  )

# Calculate the mean and variance for votes and percentages
stats_summary <- actual_results %>%
  summarise(
    mean_percent_harris = mean(percent_harris, na.rm = TRUE),
    var_percent_harris = var(percent_harris, na.rm = TRUE),
    mean_percent_trump = mean(percent_trump, na.rm = TRUE),
```

```
    var_percent_trump = var(percent_trump, na.rm = TRUE)
  )

print(missing_summary)
```

```
## # A tibble: 1 x 5
##   missing_harris non_missing_harris missing_trump non_missing_trump
##            <int>              <int>         <int>             <int>
## 1              0                 51             0                51
## # i 1 more variable: missing_state <int>
```

```
print(stats_summary)
```

```
## # A tibble: 1 x 4
##   mean_percent_harris var_percent_harris mean_percent_trump var_percent_trump
##                 <dbl>              <dbl>              <dbl>             <dbl>
## 1                47.4               145.               52.6              145.
```

Bibliography

- Kaggle. *Polymarket 2024 US Election State Data*. Retrieved from: https://www.kaggle.com/datasets/pbizil/polymarket-2024-us-election-state-data
- FiveThirtyEight. *2024 Presidential Polls*. Retrieved from: https://projects.fivethirtyeight.com/polls/president-general/2024/national/
- CBS News. *2024 Presidential Election Results*. Retrieved from: https://www.cbsnews.com/elections/2024/president/