MAT5314 Project 1: Data Visualization

Teng Li(7373086)
Shiya Gao(300381032)
Chuhan Yue(300376046)
Yang Lyu(8701121)

Introduction

A data set of the 2016 US election polls was given. In this project we aim to understand the data structure by creating various visualizations.

The data set was published by FiveThirtyEight to illustrate the reliability and quality of each pollster to which a letter grade ranging from A+ to D- was given.

Method

We use various R packages to present the data set and to plot the graphs.

Result

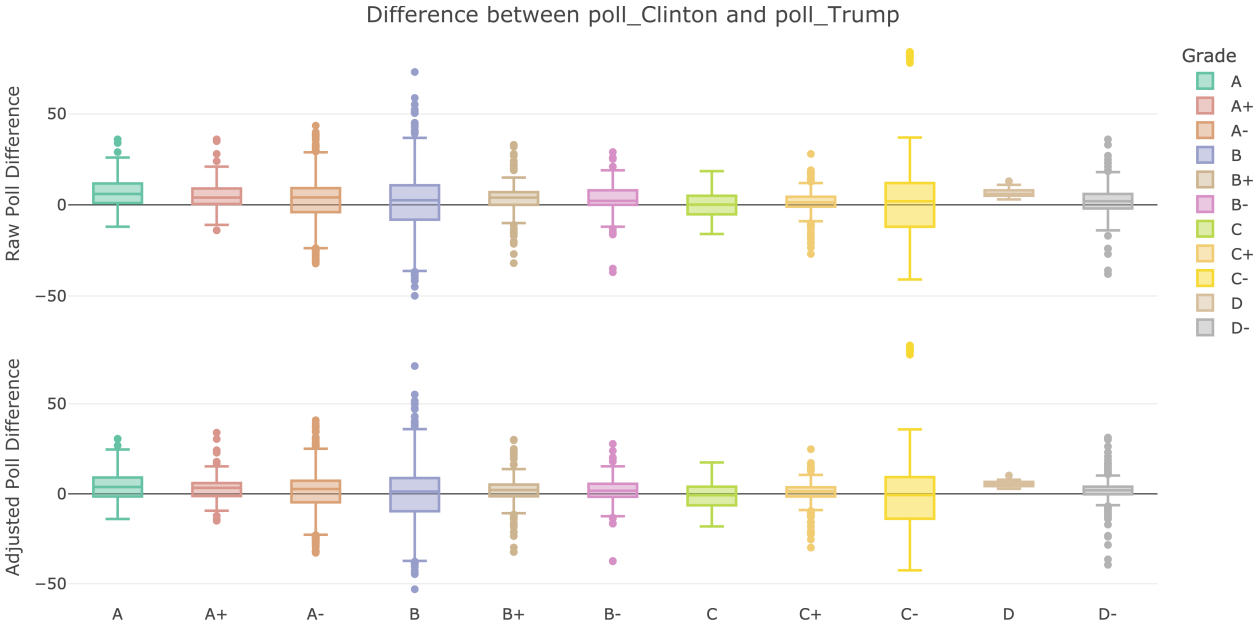
We first created a data variable definition table to give an initial understanding of the data. As one can see, there were a few variables with missing values:

Note that the poll results for Johnson and McMullin had lots of missing values. In particular, Johnson had 33.48% raw poll result and 33.48% adjusted poll result missing, and McMullin had 99.29% and 99.29% missing. Due to the fact that these two candidate didn't make to the final election, we chose to ignore their data in some of the analysis.

Since the final two candidates in Election 2016 are Clinton and Trump, we plotted box plots of the difference of their poll results, one for the raw poll and one for the adjusted poll. We saw that there's little difference between the distribution of the raw and the adjusted data. However, the mean of each grade of the adjusted poll result was a little closer to zero than that of the raw poll result. This indicated that the adjustment that FiveThirtyEight made was an improvement because the raw poll difference of each grade was mostly above zero, which clearly showed that the poll result was more in favour of Clinton yet Trump was the final winner of the election.

Table 1: Data Variable Definition

Variables	Size	Type	Example	Number.Unique	Number.Missing	Comment
state	4208	character	U.S., New Mexico, Virginia	57	0	The name of the state (or national) where the election is held
startdate	4208	character	2016-11-03, 2016-11-01, 2016-11-02	352	0	Start date of poll
enddate	4208	character	2016-11-06, 2016-11-07, 2016-11-05	345	0	End date of poll
pollster	4208	character	ABC News/Washington Post, Google Consumer Surveys, Ipsos	196	0	Organization name that conducts or analyzes opinion polls
grade	4208	character	A+, B, A-	11	429	Grade assigned by Fivethirtyeight to pollster
samplesize	4208	integer	2220, 26574, 2195	1767	1	Sample size of polls for each pollster
population	4208	character	lv, rv, a	4	0	Type of population being polled
rawpoll_clinton	4208	numeric	47, 38.03, 42	1312	0	Poll Percentage for Hillary Clinton
rawpoll_trump	4208	numeric	43, 35.69, 39	1385	0	Poll Percentage for Donald Trump
rawpoll_johnson	4208	numeric	4, 5.46, 6	585	1409	Poll Percentage for Gary Johnson
rawpoll_mcmullin	4208	numeric	NA, 24, 27.6	17	4178	Poll Percentage for Evan Mcmullin
adjpoll_clinton	4208	numeric	45.20163, 43.34557, 42.02638	4200	0	Adjusted percentage for Hillary Clinton
adjpoll_trump	4208	numeric	41.7243, 41.21439, 38.8162	4204	0	Adjusted percentage for Donald Trump
adjpoll_johnson	4208	numeric	4.626221, 5.175792, 6.844734	2211	1409	Adjusted percentage for Gary Johnson
adjpoll_mcmullin	4208	numeric	NA, 24, 27.70142	31	4178	Adjusted percentage for Evan Mcmullin

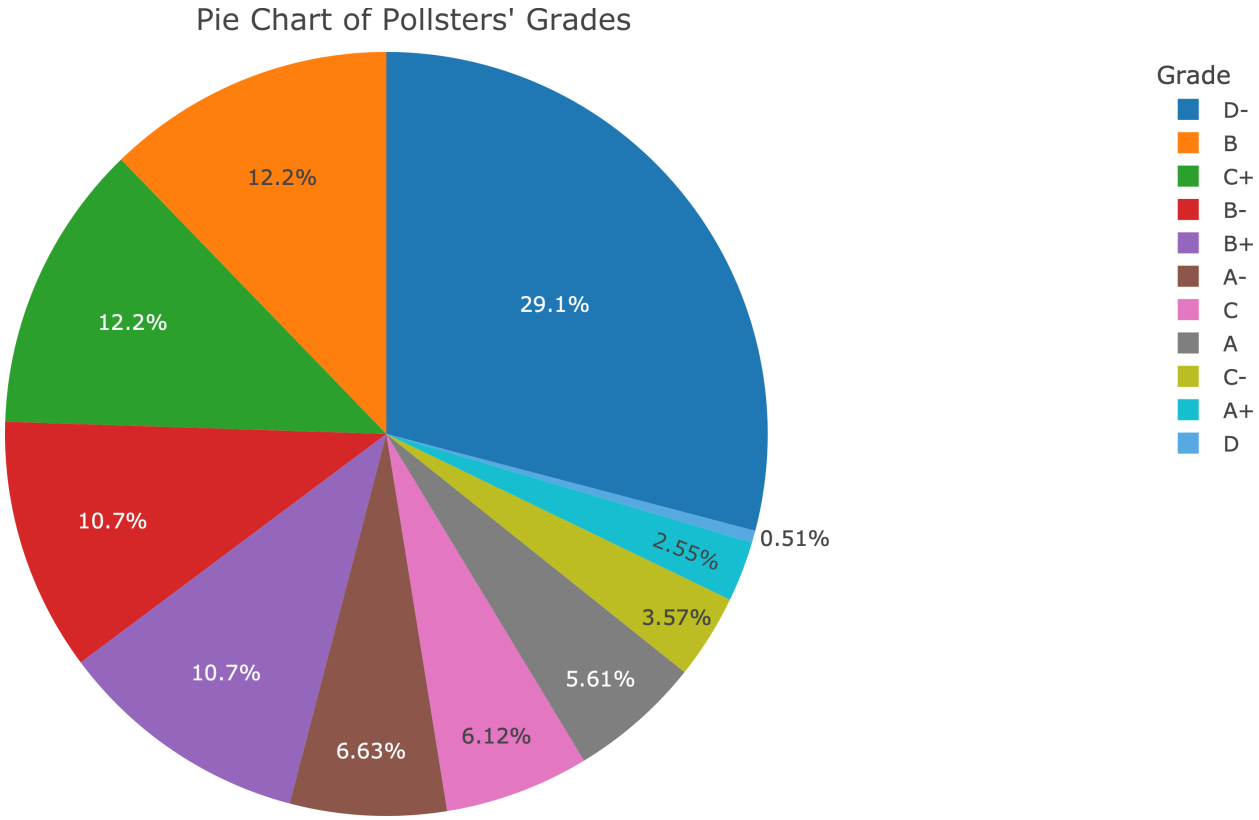


In the second step, we want to look at the percentage of different grades of pollsters. Therefore, we extract the "pollster" and "grade" columns from the original data frame, merge the duplicate rows and re-sort the data according to the grade.

```
## # A tibble: 196 x 2
##    grade pollster
##    <chr> <chr>
##  1 A     Behavior Research Center (Rocky Mountain)
##  2 A     Fairleigh Dickinson University (PublicMind)
##  3 A     Fox News/Anderson Robbins Research/Shaw & Company Research
```

```
##  4 A     Marist College
##  5 A     Marquette University
##  6 A     Muhlenberg College
##  7 A     National Journal
##  8 A     Public Policy Institute of California
##  9 A     Research & Polling, Inc.
## 10 A     Siena College
## # i 186 more rows
```

From the "Pie Chart of Grades" below, we could see that the pollsters with B grades account for nearly 50% of the total, C and A grades account for about 30% and 25% respectively. Pollsters rated D only account for less than 1%. Furthermore, grade B and C+ have the largest proportions, at about 17.3%, followed by B+ and B-, both at about 15.1%.



Comparison of candidates' polls in each state

First, we would like to give a brief introduction to the U.S. election system, because it is crucial to understand the background of the data. Voters in each state vote to choose the President of the United States. The candidate who wins the majority of the votes will receive all the electoral votes in that state. Then the sum of the electoral votes in each state is calculated. The total number of electoral votes is 538. The candidate who wins half of the votes plus 1 will win and become the new President of the United States.

Secondly, we want to analyze the key factors for the candidate's victory. Since each state has a different number of electoral votes, it is crucial for electors to win in several key states. The reason is that if a candidate wins a certain state, he will win all the electoral votes in that state. So there will be tight competition in states with more votes.

We want to process the metadata by counting the polls received by each of the four candidates in each state. This result is easier to obtain by multiplying the given size and the proportion. Regardless of the various

pollsters, we combine the number of polls for each candidate received in each state although the polls may come from different pollsters. In this case, we treat NA as 0 in the data frame `poll_by state`.

```
poll_by_state[is.na(poll_by_state)] <- 0
head(poll_by_state)
```

```
##   state prop_clinton prop_trump prop_johnson prop_mcmullin  size
## 1  U.S.        47.00      43.00         4.00             0  2220
## 2  U.S.        38.03      35.69         5.46             0 26574
## 3  U.S.        42.00      39.00         6.00             0  2195
## 4  U.S.        45.00      41.00         5.00             0  3677
## 5  U.S.        47.00      43.00         3.00             0 16639
## 6  U.S.        48.00      44.00         3.00             0  1295
##   NumVote_clinton NumVote_trump NumVote_johnson NumVote_mcmullin
## 1         1043.40       954.600           88.80                0
## 2        10106.09      9484.261         1450.94                0
## 3          921.90       856.050          131.70                0
## 4         1654.65      1507.570          183.85                0
## 5         7820.33      7154.770          499.17                0
## 6          621.60       569.800           38.85                0
```

We extracted the variables we were going to use and formed a new data structure `NumVote_State` with state and four candidates as variables. There are multiple identical values in the State column for a particular state because there are multiple pollsters for each state. In this component, we count the support of each candidate in each state based on the state as the standard, so we ignore the differences in different pollsters in the same state. The combination will take place later.

```
NumVote_State <- cbind(poll_by_state$state, poll_by_state$NumVote_clinton,
      poll_by_state$NumVote_trump, poll_by_state$NumVote_johnson,
      poll_by_state$NumVote_mcmullin)
colnames(NumVote_State) <- c("state","Clinton", "Trump", "Johnson", "Mcmullin")
head(NumVote_State)
```

```
##      state  Clinton      Trump       Johnson     Mcmullin
## [1,] "U.S." "1043.4"     "954.6"     "88.8"      "0"
## [2,] "U.S." "10106.0922" "9484.2606" "1450.9404" "0"
## [3,] "U.S." "921.9"      "856.05"    "131.7"     "0"
## [4,] "U.S." "1654.65"    "1507.57"   "183.85"    "0"
## [5,] "U.S." "7820.33"    "7154.77"   "499.17"    "0"
## [6,] "U.S." "621.6"      "569.8"     "38.85"     "0"
```

We use the `pivot_longer` function to reshape the data and obtain long-format data `NumVoteState`, which is easier to analyze and visualize. The type of data in `NumVore_State` is character. Converting data types is necessary.

```
NumVoteState <- as.data.frame(NumVote_State) %>%
  pivot_longer(cols = -state,
               names_to = "candidate",
               values_to = "PollNumber")
NumVoteState$PollNumber <- as.numeric(NumVoteState$PollNumber)
head(NumVoteState)
```

```
## # A tibble: 6 x 3
##   state candidate PollNumber
##   <chr> <chr>          <dbl>
## 1 U.S.  Clinton       1043.
## 2 U.S.  Trump          955.
## 3 U.S.  Johnson         88.8
## 4 U.S.  Mcmullin         0
## 5 U.S.  Clinton      10106.
## 6 U.S.  Trump         9484.
```

We calculate the total polls received by the four candidates in each state respectively. That is, we combine distinct pollsters if they are in the same state for each candidate. The following shows the support of the four candidates in each state. Code is provided only for Clinton due to resemblance.

- Clinton:

```
# Clinton total raw polls by states:
poll_clinton <- filter(NumVoteState, candidate == "Clinton")
Clinton_state <- poll_clinton %>%
  group_by(state) %>%
  summarize(ClintonPolls = sum(PollNumber))
Clinton_state
```

```
## # A tibble: 57 x 2
##    state                ClintonPolls
##    <chr>                       <dbl>
##  1 Alabama                     8711.
##  2 Alaska                      4150.
##  3 Arizona                    28816.
##  4 Arkansas                    6240.
##  5 California                 54446.
##  6 Colorado                   30808.
##  7 Connecticut                11579.
##  8 Delaware                    4731.
##  9 District of Columbia        4226.
## 10 Florida                    73912.
## # i 47 more rows
```

Clinton's total raw polls by state are presented in `Clinton_state`.

- Trump:

```
## # A tibble: 57 x 2
##    state                TrumpPolls
##    <chr>                     <dbl>
##  1 Alabama                  15130.
##  2 Alaska                    5092.
##  3 Arizona                  29132.
##  4 Arkansas                  9085.
##  5 California               30586.
##  6 Colorado                 27525.
##  7 Connecticut               8583.
```

```
##  8 Delaware                  3388.
##  9 District of Columbia       399.
## 10 Florida                  71730.
## # i 47 more rows
```

Trump's total raw polls by state are presented in `Trump_state`.

- Johnson:

```
## # A tibble: 57 x 2
##    state                JohnsonPolls
##    <chr>                       <dbl>
##  1 Alabama                      840.
##  2 Alaska                      1424.
##  3 Arizona                     4252.
##  4 Arkansas                     870.
##  5 California                  5260.
##  6 Colorado                    5928.
##  7 Connecticut                 1202.
##  8 Delaware                     670.
##  9 District of Columbia         181.
## 10 Florida                     6315.
## # i 47 more rows
```

Johnson's total raw polls by state are presented in `Johnson_state`.

- Mcmullin:

```
## # A tibble: 57 x 2
##    state                McmullinPolls
##    <chr>                        <dbl>
##  1 Alabama                          0
##  2 Alaska                           0
##  3 Arizona                          0
##  4 Arkansas                         0
##  5 California                       0
##  6 Colorado                         0
##  7 Connecticut                      0
##  8 Delaware                         0
##  9 District of Columbia             0
## 10 Florida                          0
## # i 47 more rows
```

Mcmullin's total raw polls by state are presented in `Mcmullin_state`.

Visualization of the poll proportion of the four candidates in each state:

```
g_11 <- ggplot(data = NumVoteState, mapping = aes( x = state, fill = candidate)) +
  geom_col(aes(y = PollNumber),position='fill') +
  labs(x = "State", y = "Percentage of Polls",
```

```
       title = "Poll Percentage by States",
       caption = "Data: NumVoteState") +
  scale_fill_manual(values=c("Clinton" = "blue", "Trump" = "red",
                             "Johnson" = "green", "Mcmullin" = "yellow")) +
  theme(axis.text.x = element_text(angle = 50, size = 6, vjust = 0.5)) +
  theme(axis.title.x = element_text(size = 9, vjust = 0),
        axis.title.y = element_text(size = 9, vjust = 3)) +
  theme(plot.title = element_text(size = 12, face = "bold",
                                  margin = margin(0, 0, 0, 0))) +
  theme(legend.position = "bottom")
print(g_11)
```

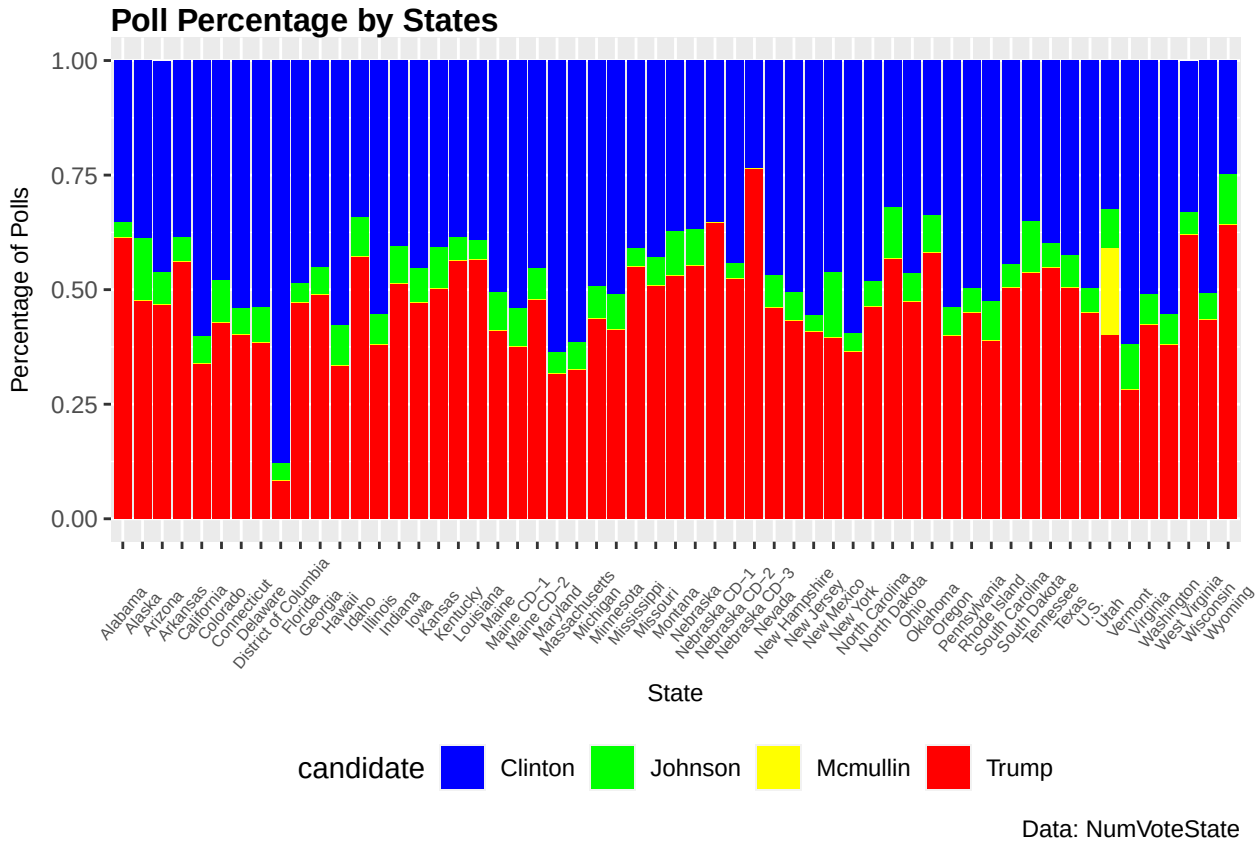


Figure 1: Poll Percentage by States.

Figure @ref(fig:g_11) shows the poll proportions of the four candidates in each state distinguished by colors. We can clearly observe which candidate is likely to win all the electoral votes in each state, which is helpful in estimating the outcome of the presidential election.

Discussion

Conclusion