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.

Method

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

Result

We first take a look at the raw data set:

```
##   state  startdate    enddate
## 1  U.S. 2016-11-03 2016-11-06
## 2  U.S. 2016-11-01 2016-11-07
## 3  U.S. 2016-11-02 2016-11-06
## 4  U.S. 2016-11-04 2016-11-07
## 5  U.S. 2016-11-03 2016-11-06
## 6  U.S. 2016-11-03 2016-11-06
##                                                    pollster grade samplesize
## 1                                ABC News/Washington Post    A+       2220
## 2                                 Google Consumer Surveys     B      26574
## 3                                                   Ipsos    A-       2195
## 4                                                  YouGov     B       3677
## 5                                           Gravis Marketing    B-      16639
## 6 Fox News/Anderson Robbins Research/Shaw & Company Research     A       1295
##   population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1         lv           47.00         43.00            4.00               NA
## 2         lv           38.03         35.69            5.46               NA
## 3         lv           42.00         39.00            6.00               NA
## 4         lv           45.00         41.00            5.00               NA
## 5         rv           47.00         43.00            3.00               NA
## 6         lv           48.00         44.00            3.00               NA
##   adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1        45.20163      41.72430        4.626221               NA
## 2        43.34557      41.21439        5.175792               NA
```

```
## 3        42.02638      38.81620        6.844734              NA
## 4        45.65676      40.92004        6.069454              NA
## 5        46.84089      42.33184        3.726098              NA
## 6        49.02208      43.95631        3.057876              NA
```

As we can see, there are a few variables with missing values:

```
##     state             startdate           enddate            pollster
##  Length:4208        Length:4208        Length:4208        Length:4208
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     grade              samplesize       population        rawpoll_clinton
##  Length:4208        Min.   :   35.0   Length:4208        Min.   :11.04
##  Class :character   1st Qu.:  447.5   Class :character   1st Qu.:38.00
##  Mode  :character   Median :  772.0   Mode  :character   Median :43.00
##                     Mean   : 1148.2                      Mean   :41.99
##                     3rd Qu.: 1236.5                      3rd Qu.:46.20
##                     Max.   :84292.0                      Max.   :88.00
##                     NA's   :1
##  rawpoll_trump   rawpoll_johnson  rawpoll_mcmullin adjpoll_clinton
##  Min.   : 4.00   Min.   : 0.000   Min.   : 9.0     Min.   :17.06
##  1st Qu.:35.00   1st Qu.: 5.400   1st Qu.:22.5     1st Qu.:40.21
##  Median :40.00   Median : 7.000   Median :25.0     Median :44.15
##  Mean   :39.83   Mean   : 7.382   Mean   :24.0     Mean   :43.32
##  3rd Qu.:45.00   3rd Qu.: 9.000   3rd Qu.:27.9     3rd Qu.:46.92
##  Max.   :68.00   Max.   :25.000   Max.   :31.0     Max.   :86.77
##                  NA's   :1409     NA's   :4178
##  adjpoll_trump    adjpoll_johnson  adjpoll_mcmullin
##  Min.   : 4.373   Min.   :-3.668   Min.   :11.03
##  1st Qu.:38.429   1st Qu.: 3.145   1st Qu.:23.11
##  Median :42.765   Median : 4.384   Median :25.14
##  Mean   :42.674   Mean   : 4.660   Mean   :24.51
##  3rd Qu.:46.290   3rd Qu.: 5.756   3rd Qu.:27.98
##  Max.   :72.433   Max.   :20.367   Max.   :31.57
##                   NA's   :1409     NA's   :4178
```

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.

```
#Create frame contained state, poll number for each candidate
states <- ElectionPoll$state
prop_raw_clinton <- ElectionPoll$rawpoll_clinton
prop_raw_trump <- ElectionPoll$rawpoll_trump
prop_raw_johnson <- ElectionPoll$rawpoll_johnson
prop_raw_mcmullin <- ElectionPoll$rawpoll_mcmullin
size <- ElectionPoll$samplesize

poll_by_state <- data.frame(
  state = states,
  prop_clinton = prop_raw_clinton,
  prop_trump = prop_raw_trump,
  prop_johnson = prop_raw_johnson,
  prop_mcmullin = prop_raw_mcmullin,
  size = size,
  NumVote_clinton = size * (prop_raw_clinton/100),
  NumVote_trump = size * (prop_raw_trump/100),
  NumVote_johnson = size * (prop_raw_johnson/100),
  NumVote_mcmullin = size * (prop_raw_mcmullin/100)
)
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
```

```
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")
View(NumVote_State)
```

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.

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

We use the `pivot_longer` function to reshape the data and obtain long-format data `NumVoteState`, which is easier to analyze and visualize.

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.

- Clinton:

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

```
## # A tibble: 6 x 2
##   state      ClintonPolls
##   <chr>             <dbl>
## 1 Alabama           8711.
## 2 Alaska            4150.
## 3 Arizona          28816.
## 4 Arkansas          6240.
## 5 California       54446.
## 6 Colorado         30808.
```

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

- Trump:

```
# Trump total raw polls by state:
poll_trump <- filter(NumVoteState, candidate == "Trump")
Trump_state <- poll_trump %>%
  group_by(state) %>%
  summarize(TrumpPolls = sum(PollNumber))
head(Trump_state)
```

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

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

- Johnson:

```
# Johnson total raw polls by state:
poll_johnson <- filter(NumVoteState, candidate == "Johnson")
Johnson_state <- poll_johnson %>%
  group_by(state) %>%
  summarize(JohnsonPolls = sum(PollNumber))
head(Johnson_state)
```

```
## # A tibble: 6 x 2
##   state      JohnsonPolls
##   <chr>             <dbl>
## 1 Alabama            840.
## 2 Alaska            1424.
## 3 Arizona           4252.
## 4 Arkansas           870.
## 5 California        5260.
## 6 Colorado          5928.
```

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

- Mcmullin:

```
# Mcmullin total raw polls by state:
poll_mcmullin <- filter(NumVoteState, candidate == "Mcmullin")
Mcmullin_state <- poll_mcmullin %>%
  group_by(state) %>%
  summarize(McmullinPolls = sum(PollNumber))
head(Mcmullin_state)
```

```
## # A tibble: 6 x 2
##   state      McmullinPolls
##   <chr>              <dbl>
## 1 Alabama                0
## 2 Alaska                 0
## 3 Arizona                0
## 4 Arkansas               0
## 5 California             0
## 6 Colorado               0
```

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)
```

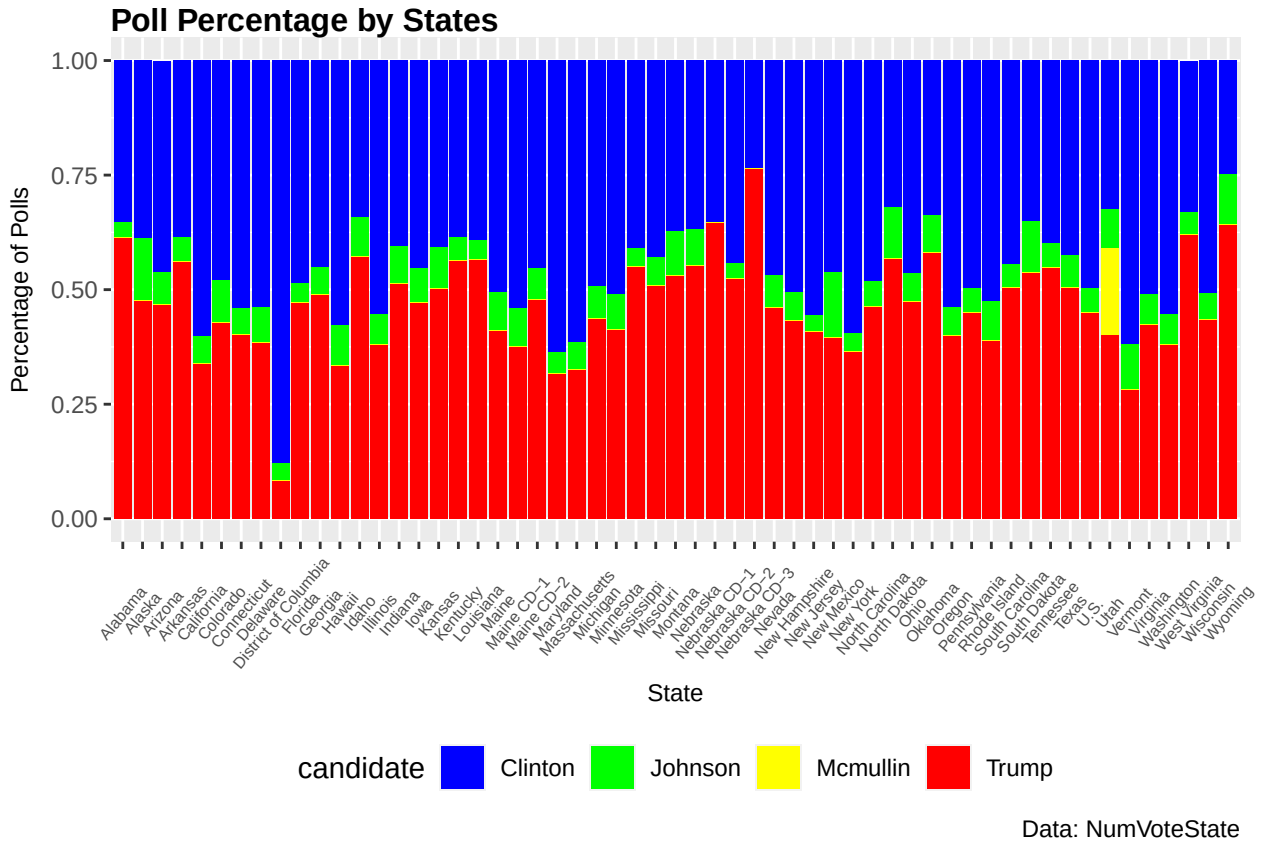

Figure 1: Poll Percentage by States

Figure @ref(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