Election2016

MukilaR

10 October 2018

About the data

The data used in this project is a public polling data organized by FiveThirtyEight for the 2016 presidential election. The table includes results for national polls, as well as state polls, taken in the year before the election. It contains 12625 rows and 27 columns.

```
##
## 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
```

95% Confidence Interval

Here we will use all the national polls that ended within a few weeks before the election.

Assuming there are only two candidates, for each poll in the polling data set, we use the CLT to create a 95% confidence interval for the spread.

```
# Creating a table called `polls` that filters by  state, date, and reports
the spread
polls <- polls_us_election_2016 %>%
  filter(state != "U.S." & enddate >= "2016-10-31") %>%
  mutate(spread = rawpoll_clinton/100 - rawpoll_trump/100)

# An object called `cis` contains the columns for the lower and upper
confidence intervals.
cis <- polls %>% mutate(X_hat = (spread+1)/2, se = 2*sqrt(X_hat*(1-
X_hat)/samplesize),
                  lower = spread - qnorm(0.975)*se, upper = spread +
qnorm(0.975)*se) %>%
  select(state, startdate, enddate, pollster, grade, spread, lower, upper)
```

Now we add actual results calculated to the previous table and determine how often the 95% confidence interval includes the actual result.

```
# Adding the actual results to the `cis` data set
add <- results_us_election_2016 %>% mutate(actual_spread = clinton/100 -
trump/100) %>% select(state, actual_spread)
cis <- cis %>% mutate(state = as.character(state)) %>% left_join(add, by =
"state")

# `p_hits` summarizes the proportion of confidence intervals that contain the
actual value.
p_hits<-cis %>% mutate(hit = lower <= actual_spread & upper >= actual_spread)
%>%
summarize(proportion_hits=mean(hit))
p_hits
```

```
##   proportion_hits
## 1         0.66133
```

Proportion of hits for each pollster

Then we find the proportion of hits for each pollster showing only pollsters with more than 5 polls and ordering them from best to worst. Also we show the number of polls conducted by each pollster and the FiveThirtyEight grade of each pollster.

```
p_hits <- cis %>% mutate(hit = lower <= actual_spread & upper >=
actual_spread) %>%
  group_by(pollster) %>%
  filter(n() >=  5) %>%
  summarize(proportion_hits = mean(hit), n = n(), grade = grade[1]) %>%
  arrange(desc(proportion_hits))
p_hits
```

```
## # A tibble: 13 x 4
##    pollster                                proportion_hits     n grade
##    <fct>                                             <dbl> <int> <fct>
##  1 Quinnipiac University                             1         6 A-
##  2 Emerson College                                   0.909    11 B
##  3 Public Policy Polling                             0.889     9 B+
##  4 University of New Hampshire                       0.857     7 B+
##  5 Ipsos                                             0.807   119 A-
##  6 Mitchell Research & Communications                0.8       5 D
##  7 Gravis Marketing                                  0.783    23 B-
##  8 Trafalgar Group                                   0.778     9 C
##  9 Rasmussen Reports/Pulse Opinion Research          0.774    31 C+
## 10 Remington                                         0.667     9 <NA>
## 11 Google Consumer Surveys                           0.588   102 B
## 12 SurveyMonkey                                      0.577   357 C-
## 13 YouGov                                            0.544    57 B
```

Proportion of hits for each state

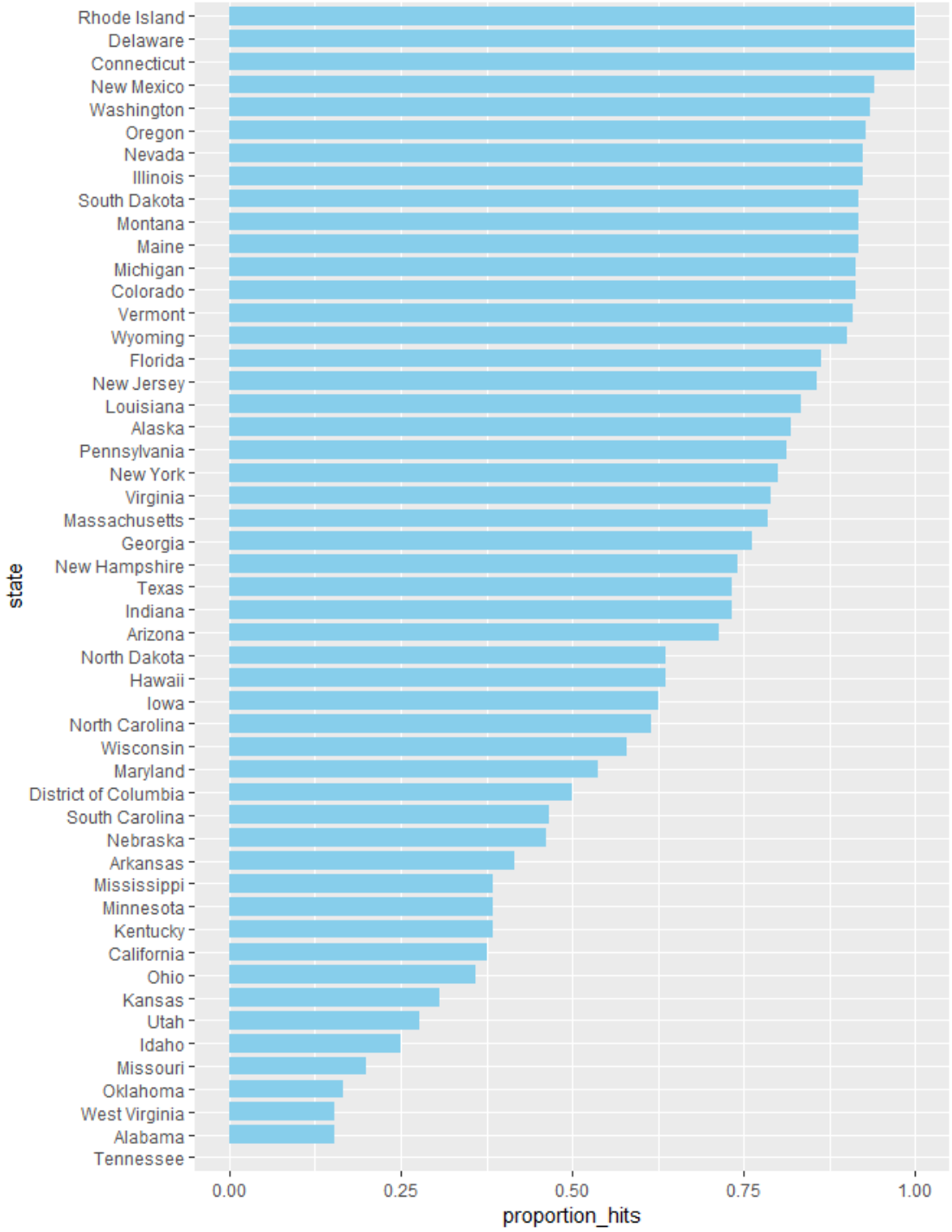
Repeating the previous code, but instead of pollster, we stratify by state. Here we cannot include grades.

```
p_hits <- cis %>% mutate(hit = lower <= actual_spread & upper >=
actual_spread) %>%
  group_by(state) %>%
  filter(n() >=  5) %>%
  summarize(proportion_hits = mean(hit), n = n()) %>%
  arrange(desc(proportion_hits))
p_hits
```

```
## # A tibble: 51 x 3
##    state        proportion_hits     n
##    <chr>                  <dbl> <int>
##  1 Connecticut            1        13
##  2 Delaware               1        12
##  3 Rhode Island           1        10
##  4 New Mexico             0.941    17
##  5 Washington             0.933    15
##  6 Oregon                 0.929    14
##  7 Illinois               0.923    13
##  8 Nevada                 0.923    26
##  9 Maine                  0.917    12
## 10 Montana                0.917    12
## # ... with 41 more rows
```

Visualisation

A barplot is created for this new state table.

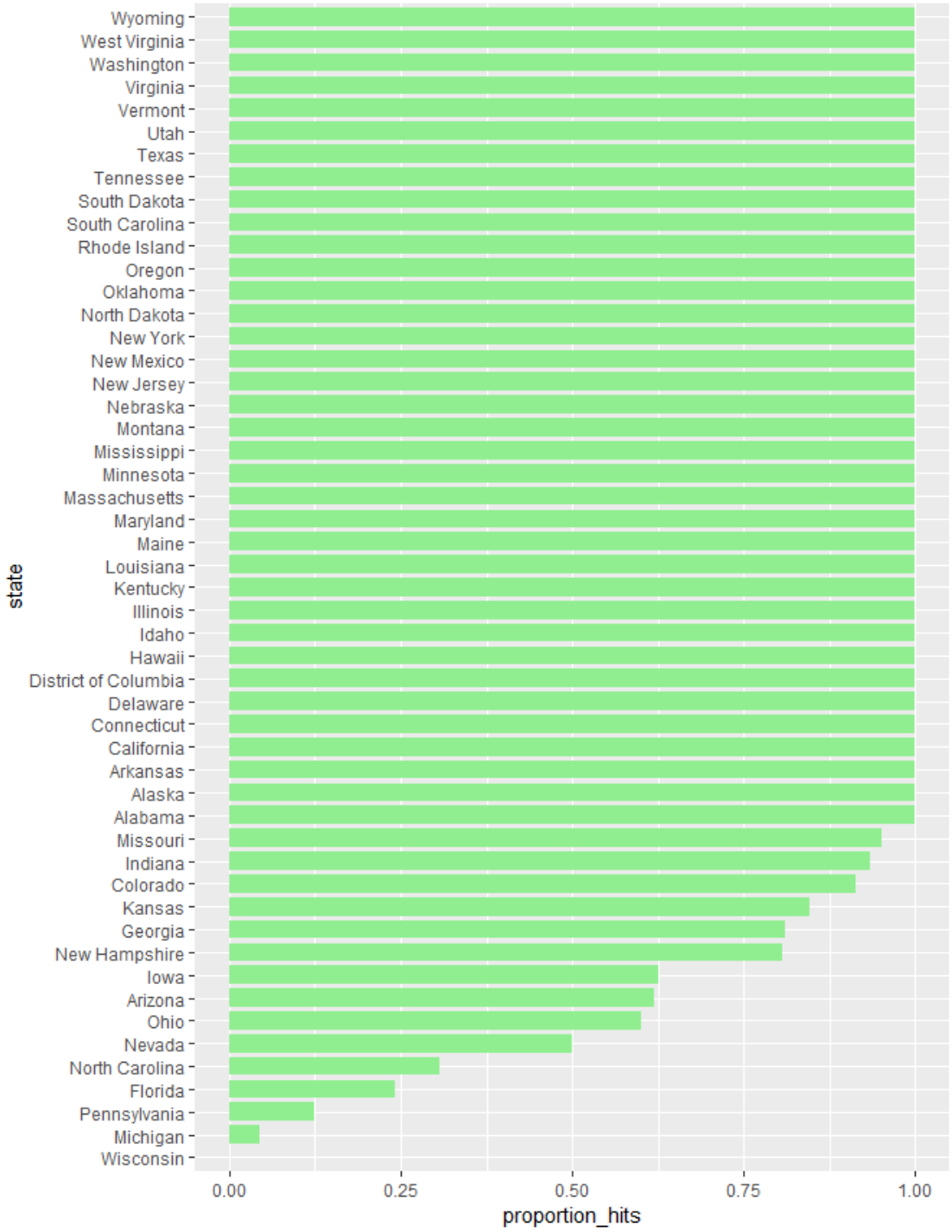


Difference between the predicted spread and the actual spread

Even if a forecaster's confidence interval is incorrect, the overall predictions will do better if they correctly called the right winner. So we add two columns to the cis table by computing, for each poll, the difference between the predicted spread and the actual spread, and define a column hit that is true if the signs are the same.

```
#An object called `errors` calculates the difference between the predicted
and actual spread and indicates if the correct winner was predicted
errors <- cis %>% mutate(error = spread - actual_spread, hit = sign(spread)
== sign(actual_spread))
```

Now we make a barplot based on the result from the previous code that shows the proportion of times the sign of the spread matched the actual result for the data in p_hits.



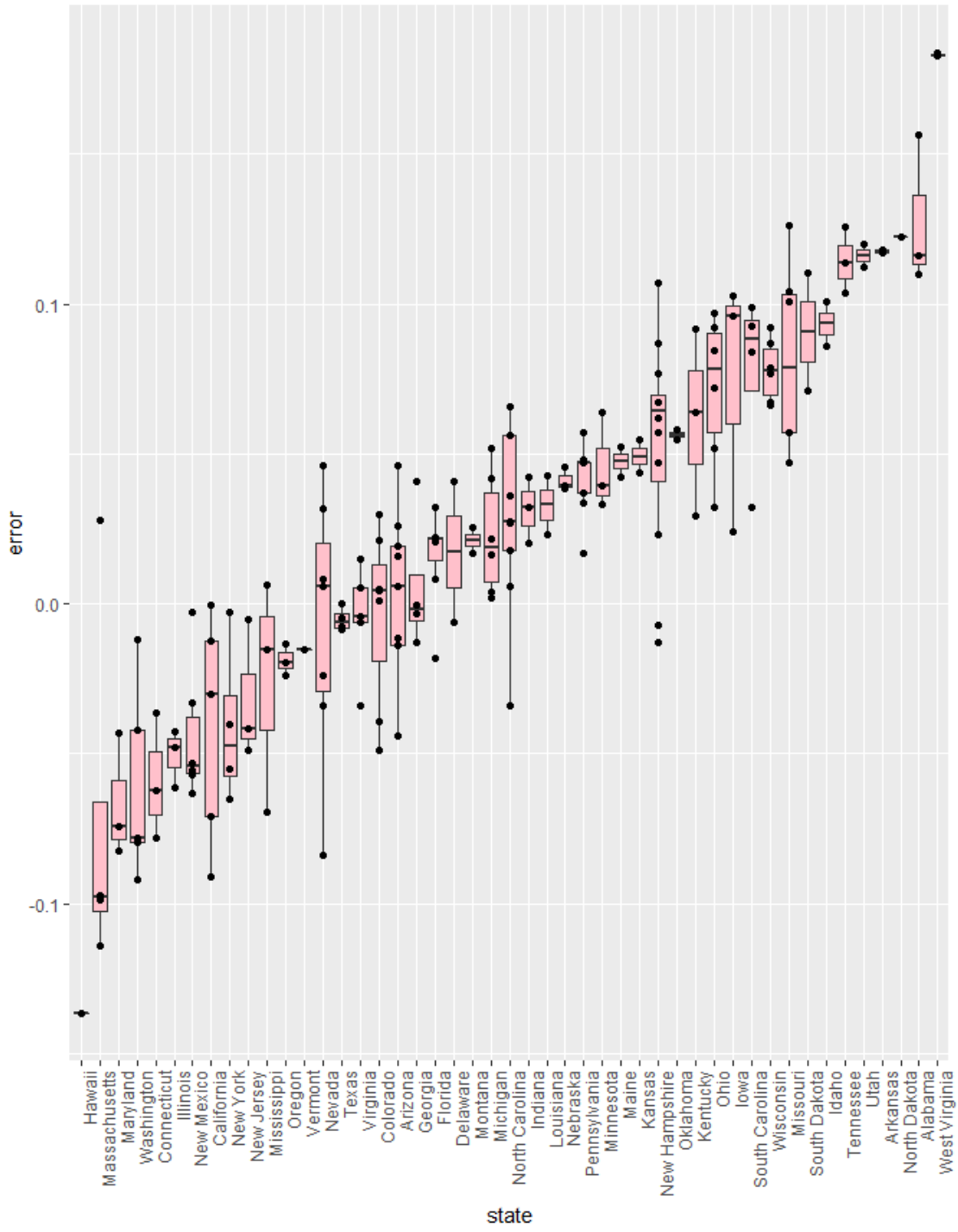
In the previous graph, we see that most states' polls predicted the correct winner 100% of the time. Only a few states polls' were incorrect more than 25% of the time. Wisconsin got every single poll wrong. In Pennsylvania and Michigan, more than 90% of the polls had the signs wrong.

Histogram

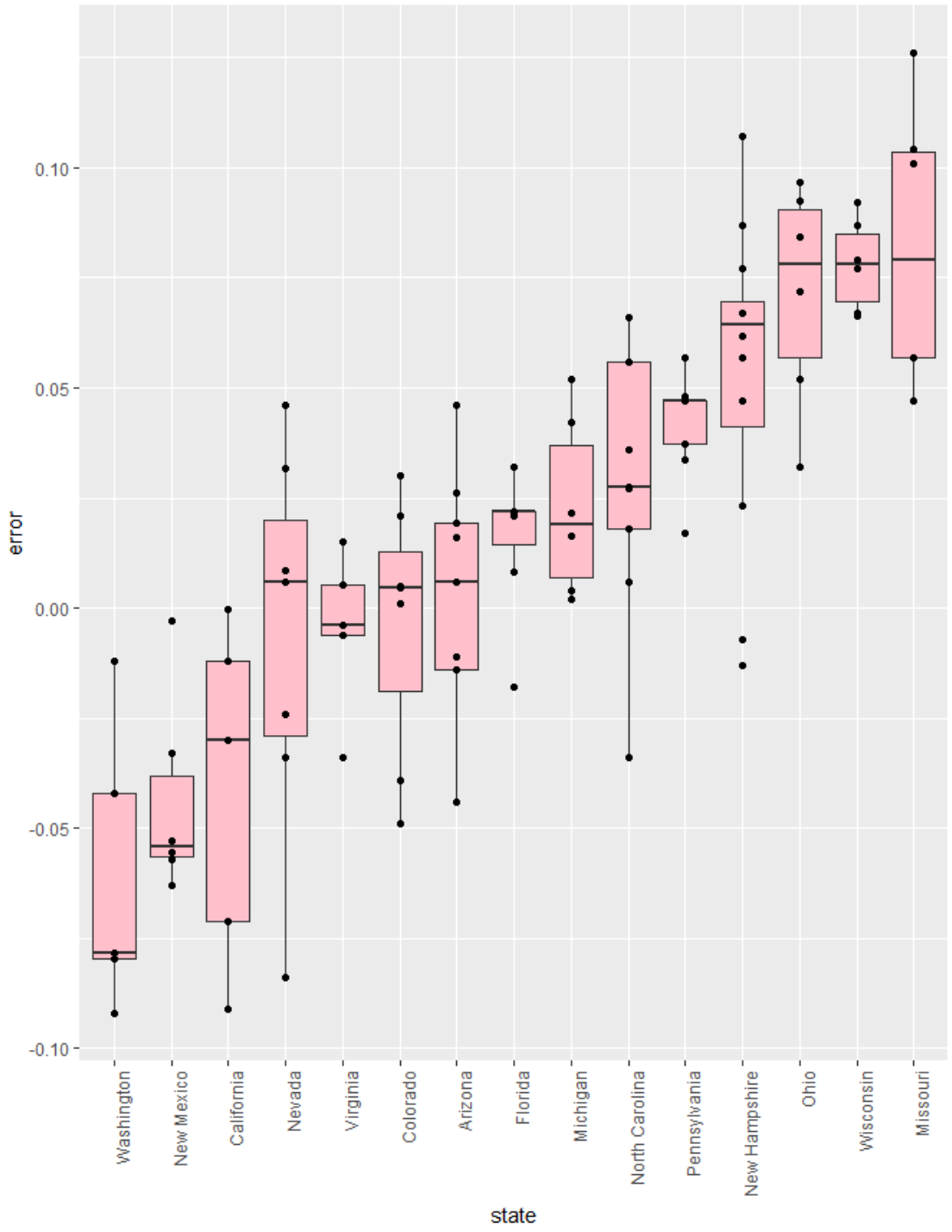
Lets make a histogram of the errors and find the median of these errors.


```
## [1] 0.037
```

We see that, at the state level, the median error was slightly in favor of Clinton. The distribution is not centered at 0, but at 0.037. This value represents the general bias. So we create a boxplot to examine if the bias was general to all states or if it affected some states differently. Also the data is filtered to include only pollsters with grades B+ or higher.



Some of these states only have a few polls. So lets repeat the previous process to plot the errors for each state, but only include states with five good polls or more.


0.10
0.05
0.00
-0.05
-0.10
error
Washington
New Mexico
California
Nevada
Virginia
Colorado
Arizona
Florida
Michigan
North Carolina
Pennsylvania
New Hampshire
Ohio
Wisconsin
Missouri
state