Presidential Election Predictions 2016

Jo Hardin

August 16, 2016

Getting the Data

Thanks to the Internet, there is a lot of polling data which is publicly accessible. For the competition, you are welcome to get your data from anywhere. However, I'm going to take mine from 538. http://projects.fivethirtyeight.com/2016-election-forecast/national-polls/ (Other good sources of data are http://www.realclearpolitics.com/epolls/latest_polls/ and http://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton and http://www.gallup.com/products/170987/gallup-analytics.aspx)

Note the date indicated above as to when this R Markdown file was written. That's the day the data were scraped from 538. If you run the Markdown file on a different day, you are likely to get different results as the polls are constantly being updated.

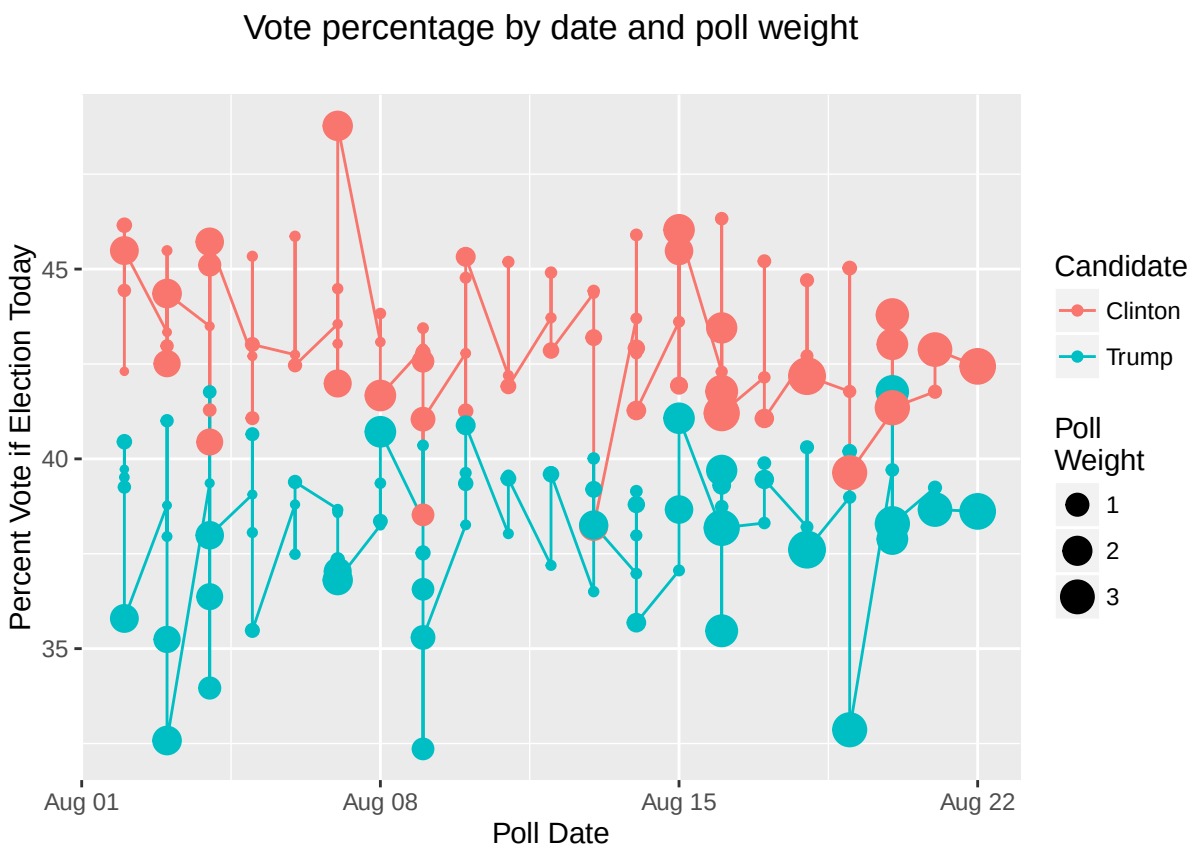
Because the original data were scraped as a JSON file, it gets pulled into R as a list of lists. The data wrangling used to convert it into a tidy format is available from the source code at .

```
url = "http://projects.fivethirtyeight.com/2016-election-forecast/national-polls/"
doc <- htmlParse(url, useInternalNodes = TRUE)

sc = xpathSApply(doc, "//script[contains(., 'race.model')]",
                 function(x) c(xmlValue(x), xmlAttrs(x)[["href"]]))

jsobj = gsub(".*race.stateData = (.*);race.pathPrefix.*", "\\1", sc)

data = fromJSON(jsobj)
allpolls <- data$polls

#unlisting the whole thing
indx <- sapply(allpolls, length)
pollsdf <- as.data.frame(do.call(rbind, lapply(allpolls, `length<-`, max(indx))))
```

A Quick Visualization

Before I try to come up with a prediction for the winner of the 2016 US Presidential Race, I think it is worth trying to look at the data. I read the data into a tidy form, so ggplot2 will be the right tool for visualizing the data.

```
ggplot(subset(allpolldata, ((polltypeA == "now") & (endDate > ymd("2016-08-01")))),
                        aes(y=adj_pct, x=endDate, color=choice)) +
  geom_line() + geom_point(aes(size=wtnow)) +
  labs(title = "Vote percentage by date and poll weight\n",
     y = "Percent Vote if Election Today", x = "Poll Date",
     color = "Candidate", size="Poll\nWeight")
```

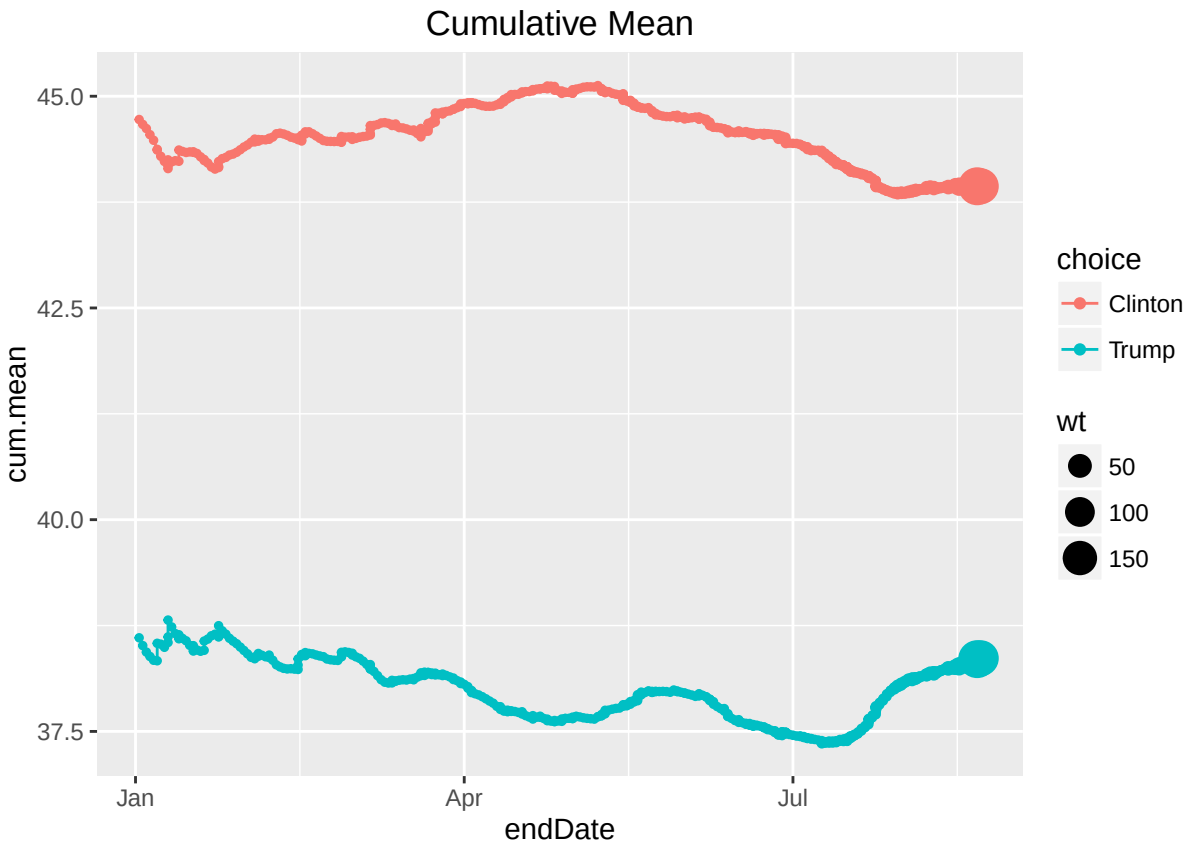


A Quick Analysis

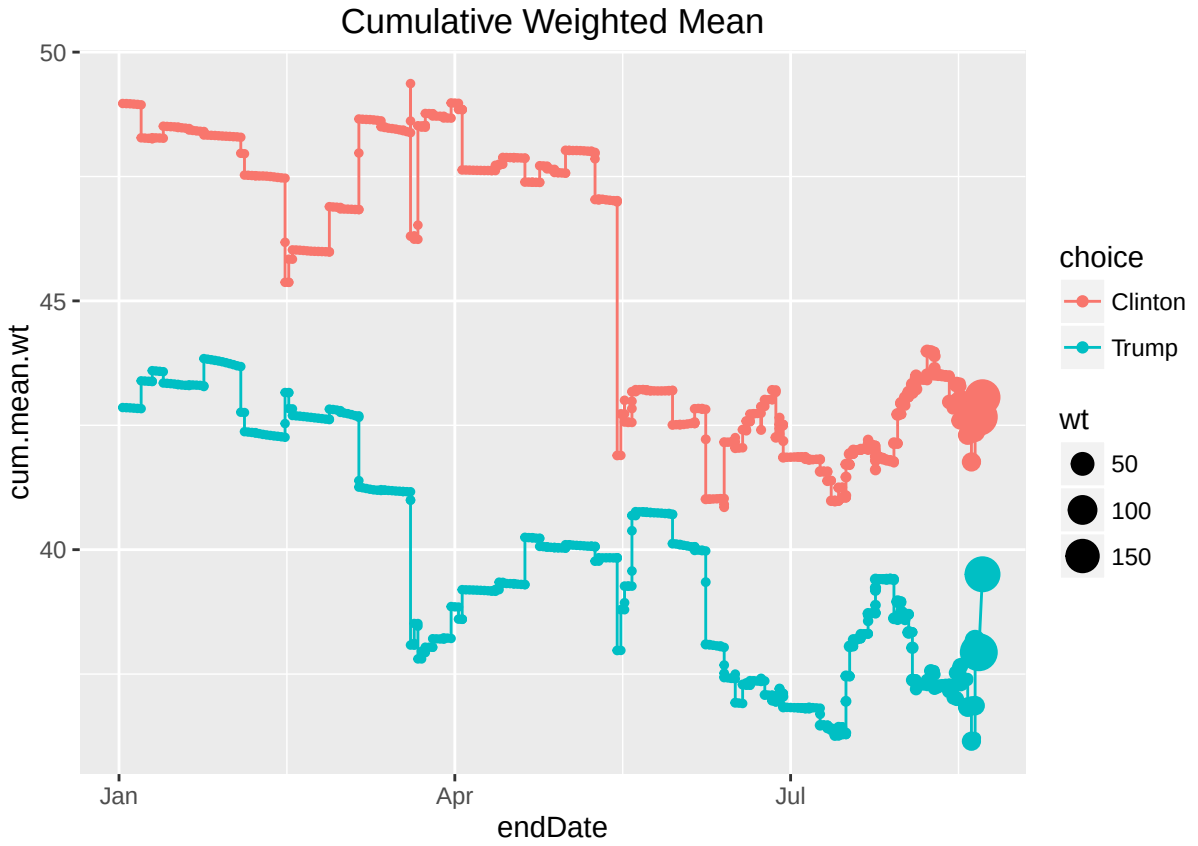
Let's try to think about the percentage of votes that each candidate will get based on the current polls. We'd like to weight the votes based on what 538 thinks (hey, they've been doing this longer than I have!), the sample size, and the number of days since the poll closed. Using the weight, I'll calculate a weighted average and a weighted SE for the predicted percent of votes.

$$w = \frac{w_{538}}{dayssincepoll}\sqrt{samplesize}$$

```
allpolldata2 <- allpolldata %>%
  filter(wtnow > 0) %>%
  filter(polltypeA == "now") %>%
  mutate(dayssince = as.numeric(today() - endDate)) %>%
  mutate(wt = wtnow * sqrt(sampleSize) / dayssince) %>%
  mutate(votewt = wt*pct) %>%
  group_by(choice) %>%
  arrange(choice, -dayssince) %>%
  mutate(cum.mean.wt = cumsum(votewt) / cumsum(wt)) %>%
  mutate(cum.mean = cummean(pct))

ggplot(subset(allpolldata2, ( endDate > ymd("2016-01-01"))),
                          aes(y=cum.mean, x=endDate, color=choice)) +
  geom_line() + geom_point(aes(size=wt)) + ggtitle("Cumulative Mean")
```




```
ggplot(subset(allpolldata2, (endDate > ymd("2016-01-01"))),
                          aes(y=cum.mean.wt, x=endDate, color=choice)) +
  geom_line() + geom_point(aes(size=wt)) + ggtitle("Cumulative Weighted Mean")
```




```
pollsummary <- allpolldata2 %>%
  select(choice, pct, wt, votewt, sampleSize, dayssince) %>%
  group_by(choice) %>%
  summarise(mean.vote = weighted.mean(pct, wt, na.rm=TRUE),
            std.vote = sqrt(weighted.var.se(pct, wt, na.rm=TRUE)))

pollsummary
```

```
## # A tibble: 2 x 3
##    choice mean.vote  std.vote
##     <chr>     <dbl>     <dbl>
## 1 Clinton  43.06351 0.5543392
## 2   Trump  39.50546 1.3914807
```

Other people's advice

Prediction is very difficult, especially about the future. - Niels Bohr

Andrew Gelman: http://andrewgelman.com/2016/08/17/29654/