Lecture 9 Notes

2024-02-13

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

```
poll <- read_rds("https://github.com/jbisbee1/DS1000_S2024/raw/main/data/Pres2020_PV.Rd
s")

# Initial Wrangling
poll <- poll %>%
  mutate(Trump = Trump/100,
         Biden = Biden/100,
         margin = Biden - Trump)
```

Introducint `as.Date()`

```
as.Date('02/13/2024','%m/%d/%Y') - as.Date('02/01/2021','%m/%d/%Y')
```

```
## Time difference of 1107 days
```

```
as.Date('13/02/2024','%d/%m/%Y')
```

```
## [1] "2024-02-13"
```

```
as.Date('2024-02-13','%Y-%m-%d')
```

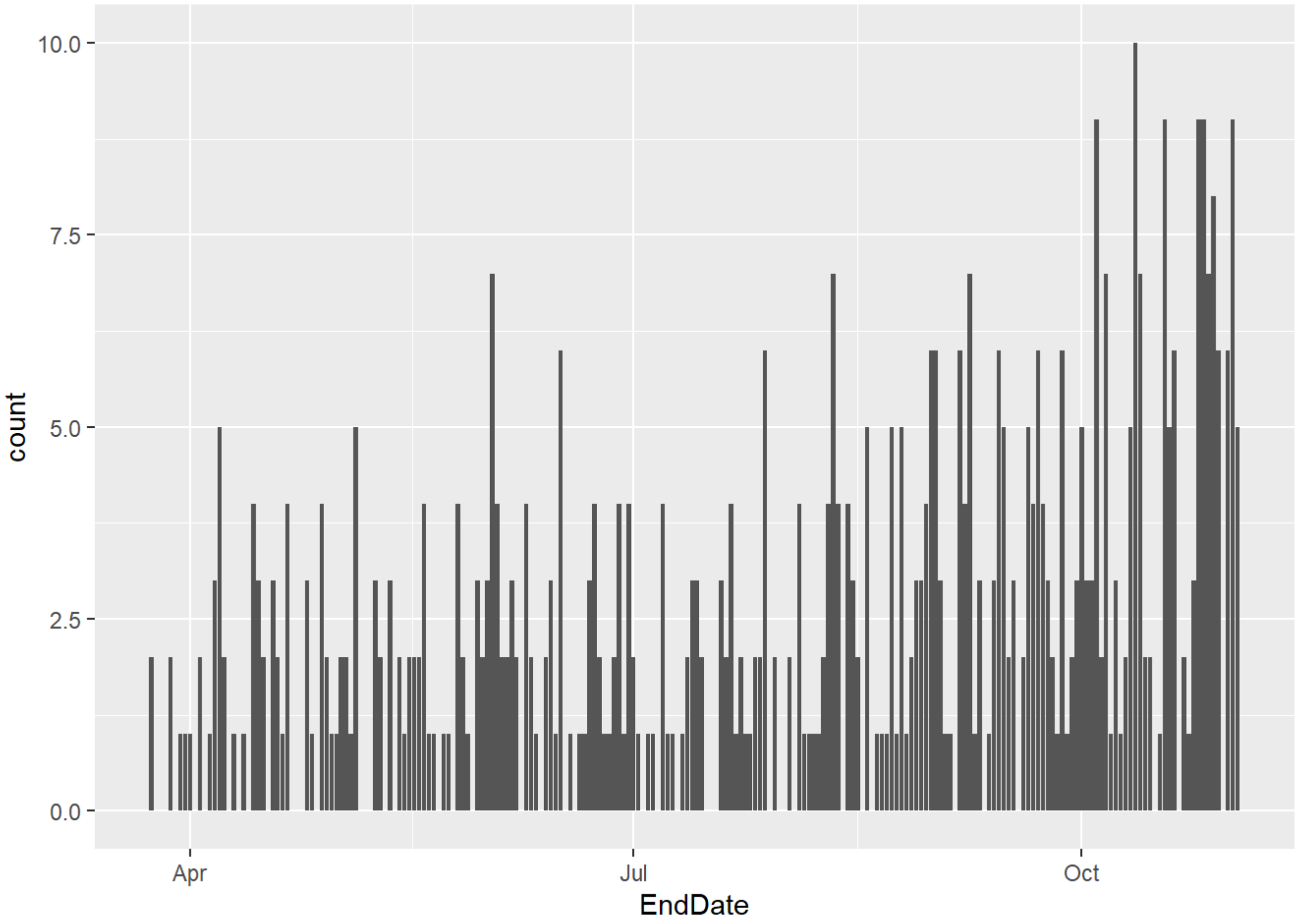
```
## [1] "2024-02-13"
```

```
election.day <- as.Date('11/3/2020','%m/%d/%Y')

poll <- poll %>%
  mutate(StartDate = as.Date(StartDate,'%m/%d/%Y'),
         EndDate = as.Date(EndDate,'%m/%d/%Y'),
         DaysToElection = as.numeric(election.day - EndDate))
```

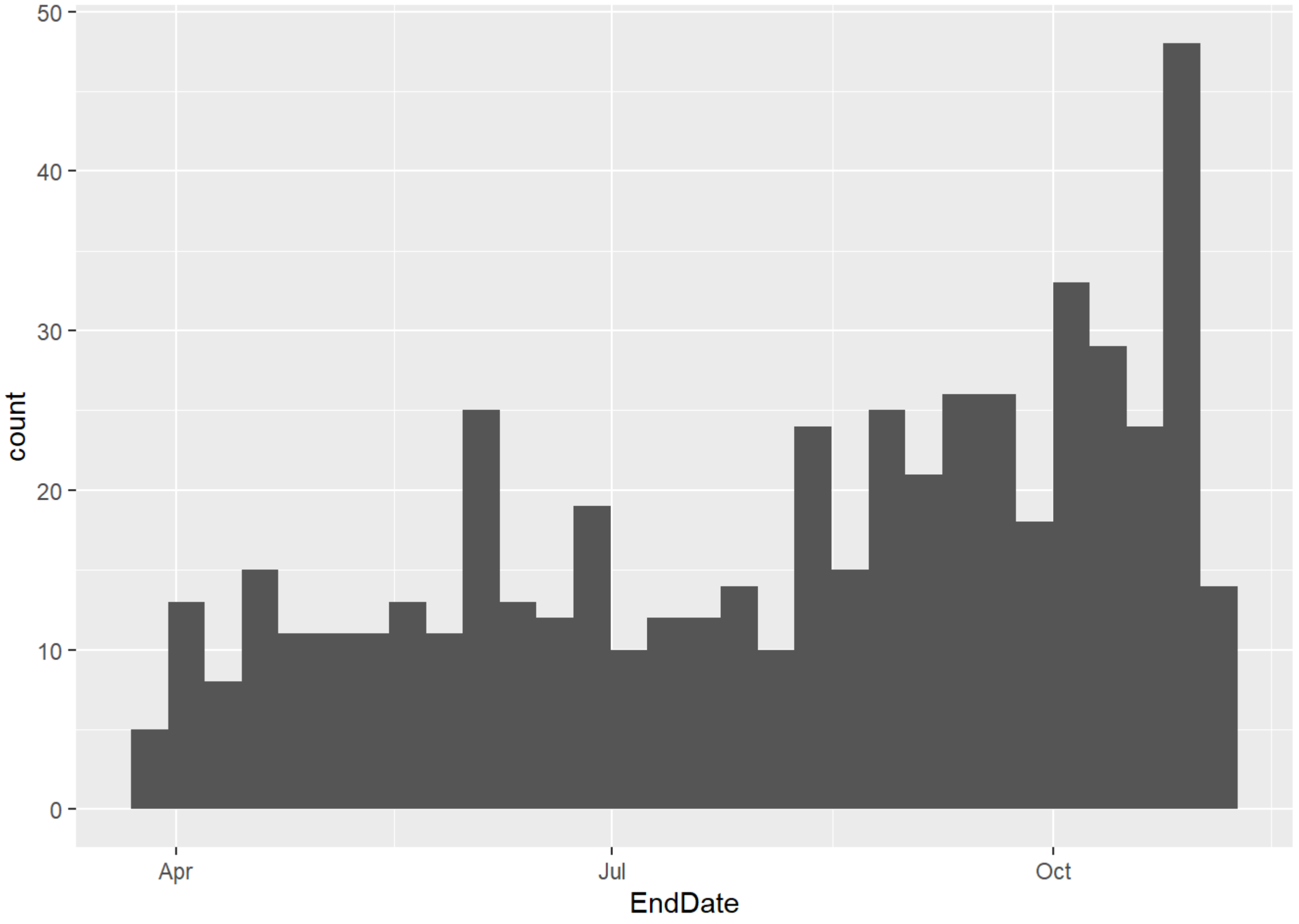
Multivariate VIsualization with Dates

```
poll %>%
  ggplot(aes(x = EndDate)) +
  geom_bar()
```



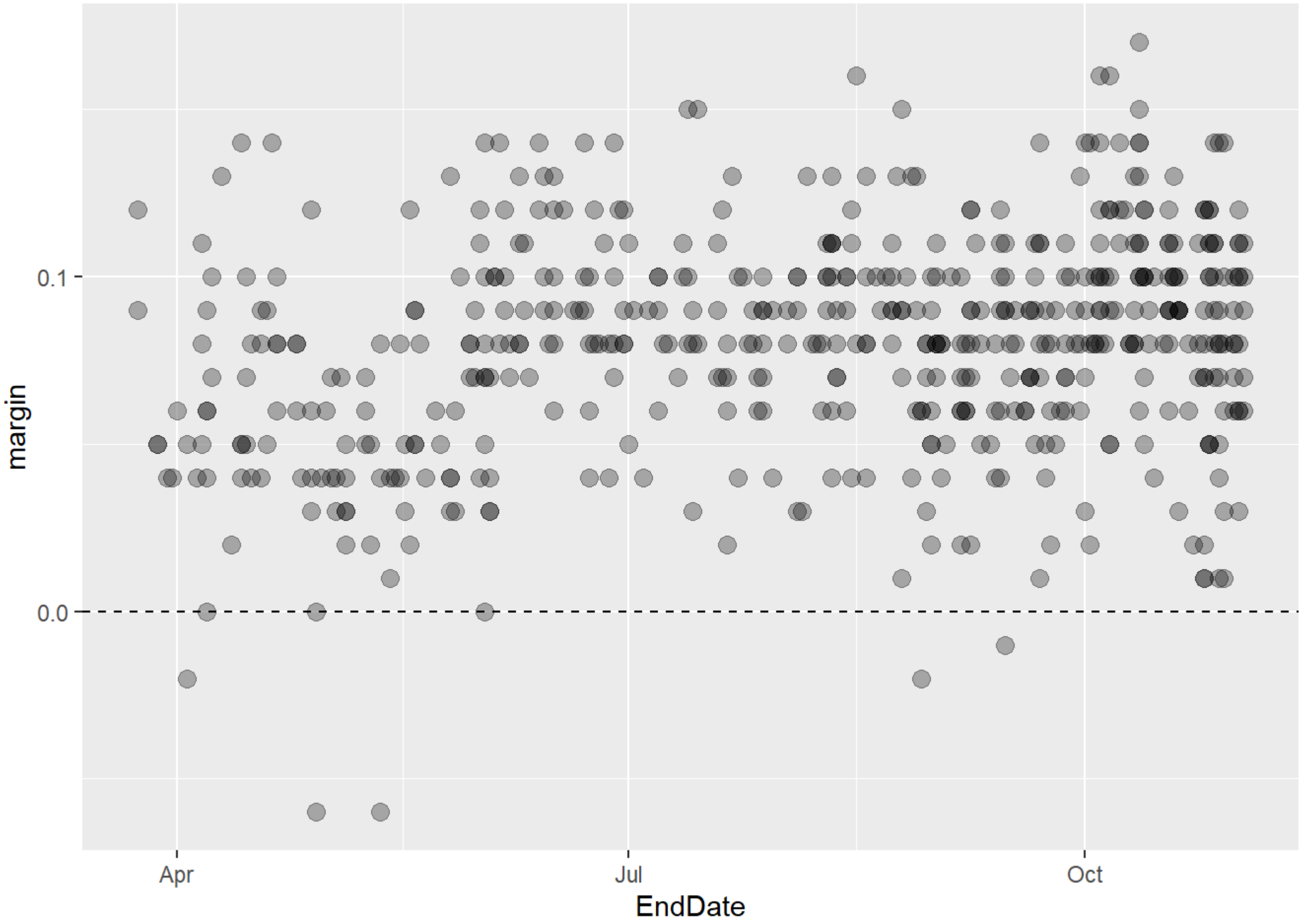
```
poll %>%
  ggplot(aes(x = EndDate)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Look at margin first

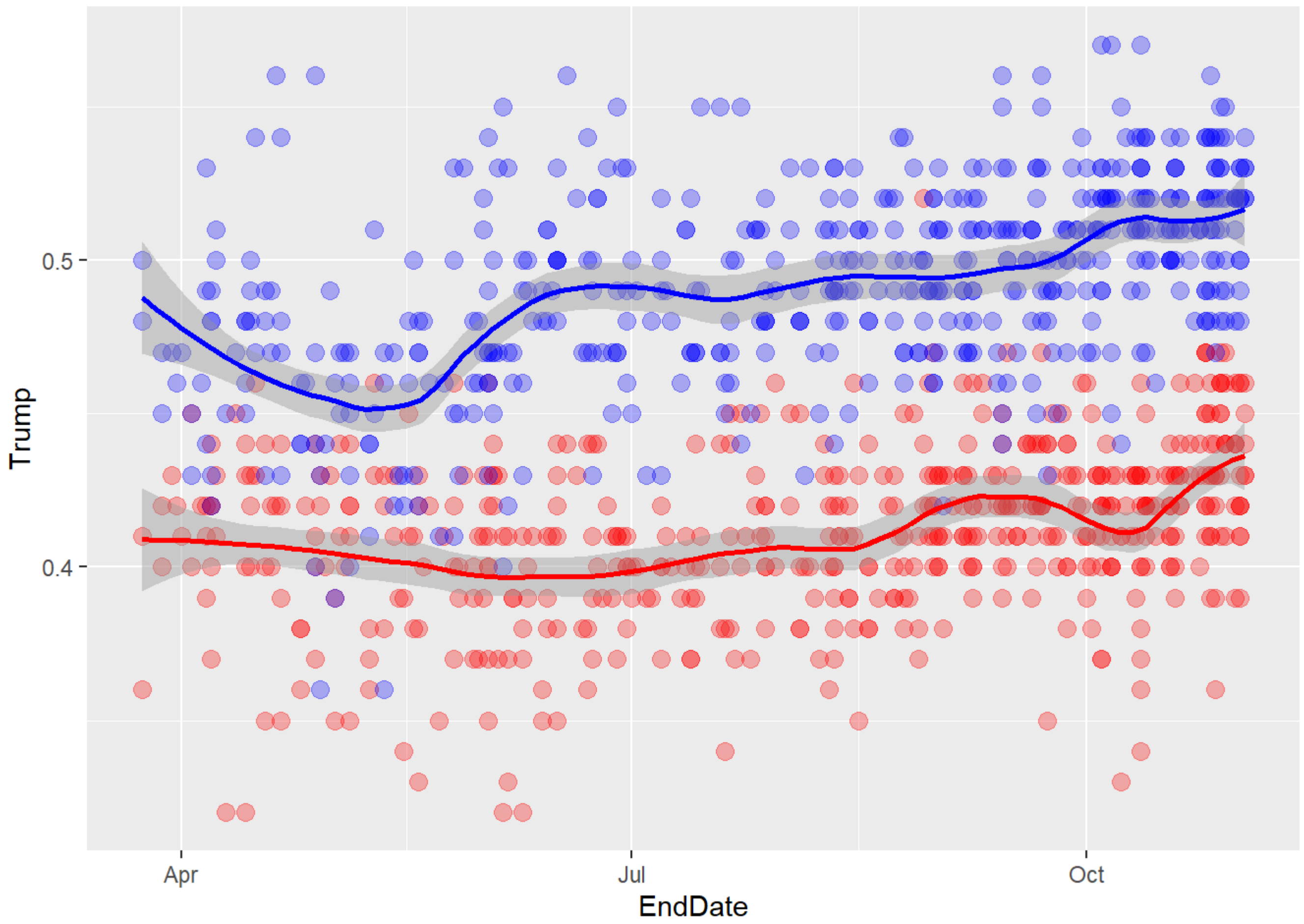
```
poll %>%
  ggplot(aes(x = EndDate,y = margin)) +
  geom_point(size = 3,alpha = .3) +
  geom_hline(yintercept = 0,linetype = 'dashed')
```



Look at Trump and Biden separately

```
poll %>%
  ggplot(aes(x = EndDate)) +
  geom_point(aes(y = Trump),color = 'red',size = 3,alpha = .3) +
  geom_point(aes(y = Biden),color = 'blue',size = 3,alpha = .3) +
  geom_smooth(aes(y = Biden),color = 'blue',span = .3) +
  geom_smooth(aes(y = Trump),color = 'red',span = .3)
```

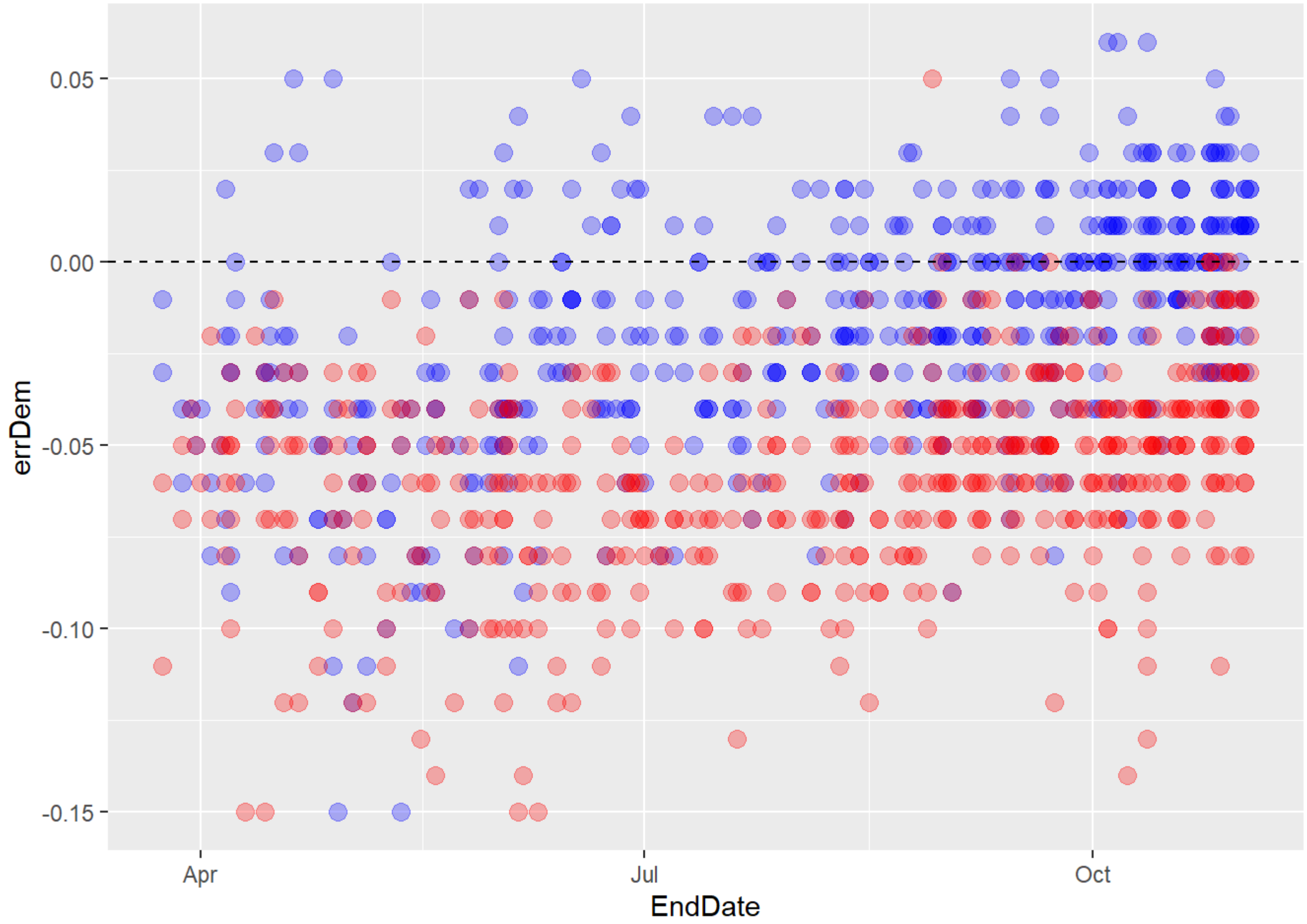
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Look at prediction error

```
poll <- poll %>%
  mutate(errDem = Biden - DemCertVote/100,
         errRep = Trump - RepCertVote/100)

poll %>%
  ggplot(aes(x = EndDate)) +
  geom_point(aes(y = errDem),color = 'blue',size = 3,alpha = .3) +
  geom_point(aes(y = errRep),color = 'red',size = 3,alpha = .3) +
  geom_hline(yintercept = 0,linetype = 'dashed')
```



State-Level Polling

```
statePoll <- read_rds("https://github.com/jbisbee1/DS1000_S2024/raw/main/data/Pres2020_S
tatePolls.Rds")

statePoll
```

```
## # A tibble: 1,545 × 19
##    StartDate  EndDate    DaysinField   MoE Mode    SampleSize Biden Trump Winner
##    <date>     <date>           <dbl> <dbl> <chr>        <dbl> <dbl> <dbl> <chr> 
##  1 2020-03-21 2020-03-30          10   2.8 Phone/…       1331    41    46 Rep   
##  2 2020-03-24 2020-04-03          11   3   Phone/…       1000    47    34 Dem   
##  3 2020-03-24 2020-03-29           6   4.2 Live p…        813    48    45 Dem   
##  4 2020-03-28 2020-03-29           2  NA   Live p…        962    67    29 Dem   
##  5 2020-03-30 2020-04-01           3   4   IVR            602    46    46 Dem   
##  6 2020-03-31 2020-04-04           5   1.7 Online        3244    46    40 Rep   
##  7 2020-03-31 2020-04-01           2   3   Phone …       1035    46    48 Dem   
##  8 2020-03-31 2020-04-01           2   3.1 Live p…       1019    48    45 Dem   
##  9 2020-03-31 2020-04-06           7   4.1 Online         583    52    39 Dem   
## 10 2020-04-05 2020-04-07           3   4.4 Live p…        500    42    49 Rep   
## # ℹ 1,535 more rows
## # ℹ 10 more variables: poll.predicted <dbl>, Funded <chr>, Conducted <chr>,
## #   margin <dbl>, DaysToED <drtn>, StateName <chr>, EV <int>, State <chr>,
## #   BidenCertVote <dbl>, TrumpCertVote <dbl>
```

Wrangling to analyze the data

```
statePoll <- statePoll %>%
  mutate(Biden2W = Biden / (Biden + Trump),
         Biden = Biden / 100,
         Trump = Trump / 100)

view(statePoll)

stateProbs <- statePoll %>%
  group_by(State,EV) %>%
  summarise(BidenProb1 = mean(Biden > Trump),
            BidenProb2 = mean(Biden),
            BidenProb3 = mean(Biden2W))
```

```
## `summarise()` has grouped output by 'State'. You can override using the
## `.groups` argument.
```

Cool visualizations

```
require(plotly)
```

```
## Loading required package: plotly
```

```
## Warning: package 'plotly' was built under R version 4.3.2
```

```
##
## Attaching package: 'plotly'
```

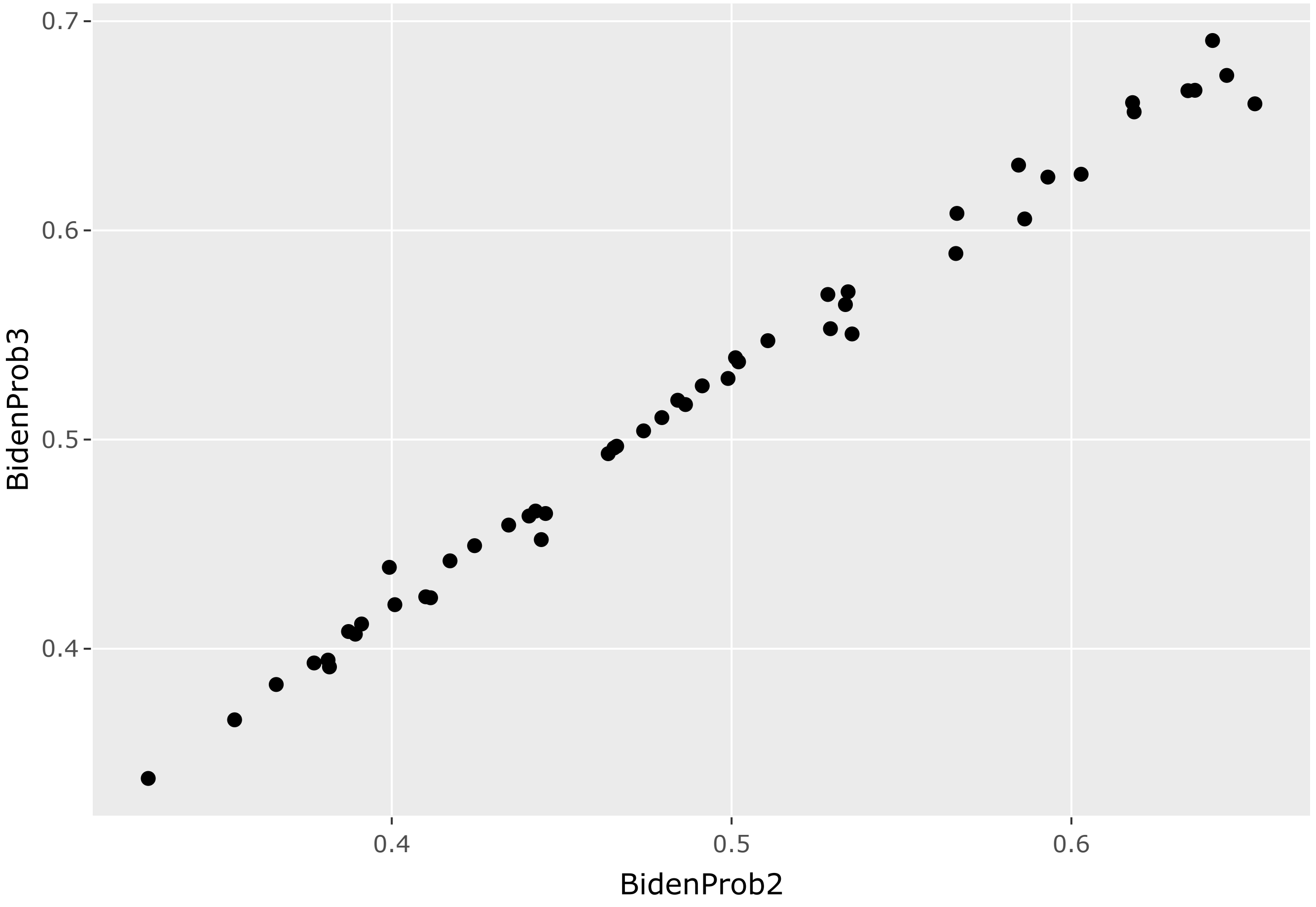
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
p <- stateProbs %>%
  ggplot(aes(x = BidenProb2,y = BidenProb3,text = State)) +
  geom_point()

ggplotly(p,tooltip = 'text')
```



```
# Different visualization
p <- stateProbs %>%
  ggplot(aes(x = BidenProb2,y = BidenProb1,text = State,
             color = BidenProb1)) +
  geom_point()
```