Topic 6. Conditional (Bivariate) Visualization

Josh Clinton

9/22/2021

Today’s Agenda

- Becoming political pundits: summarizing the 2020 Popular Vote Over Time
- Conditional data: when a variable varies with respect to some other variable
- Visualizing conditional data
- “Smoothing” data
- (Intro) Looping

What is our question?

How did the support for Biden and Trump vary across the course of the 2020 Election?

- What should we measure?
- How do we summarize, visualize, and communicate?

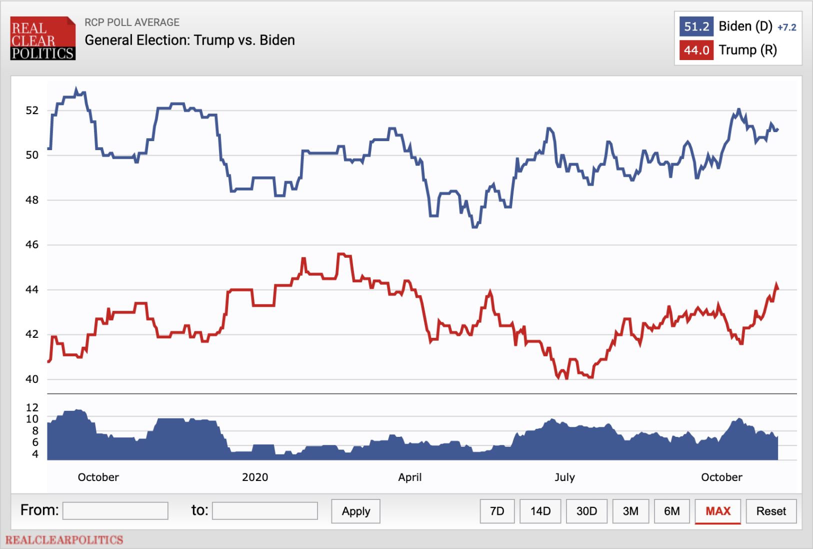
What is our question?

How did the support for Biden and Trump vary across the course of the 2020 Election?

- What should we measure?
- How do we summarize, visualize, and communicate?

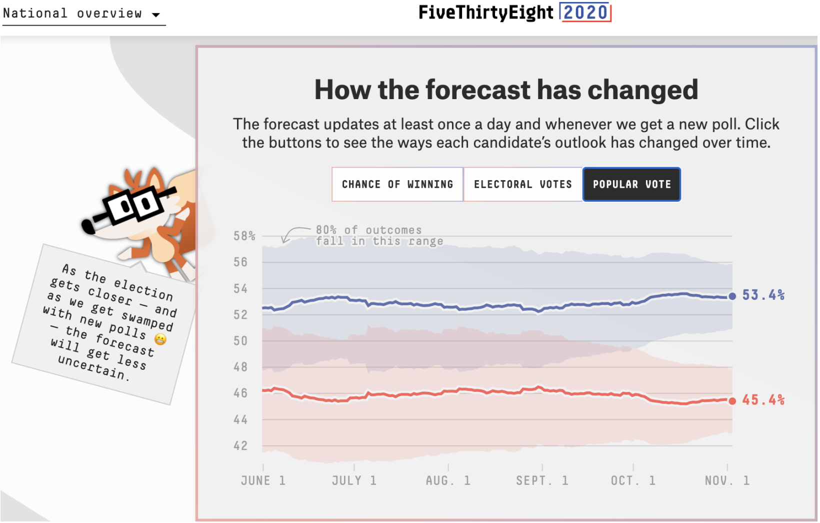
Give you some tools to do some *amazing* things!

End goal?



REAL CLEAR POLITICS
RCP POLL AVERAGE
General Election: Trump vs. Biden
51.2 Biden (D) +7.2
44.0 Trump (R)
52
50
48
46
44
42
40
12
10
8
6
4
October
2020
April
July
October
From:
to:
Apply
7D
14D
30D
3M
6M
MAX
Reset
REALCLEARPOLITICS

End goal?



National overview ▾

FiveThirtyEight 2020

How the forecast has changed

The forecast updates at least once a day and whenever we get a new poll. Click the buttons to see the ways each candidate's outlook has changed over time.

As the election gets closer — and as we get swamped with new polls 😬 — the forecast will get less uncertain.

Telling Time

- Time is often a critical *descriptive* variable. (Not causal!)
- Also useful for *prediction* ?

Telling Time

- Time is often a critical *descriptive* variable. (Not causal!)
- Also useful for *prediction* ?
- We want to evaluate the properties of presidential polling as Election Day 2020 approached.
- Necessary for prediction – we want most recent data to account for last-minute shift.
- Necessary for identifying when changes occurred (and why?)

Dates in R

- Dates are a special format in R (character with quasi-numeric properties)

```r
load(file="data/Pres2020.PV.Rdata")
election.day <- as.Date("11/3/2020", "%m/%d/%Y")
election.day16 <- as.Date("11/8/2016", "%m/%d/%Y")
```

Dates in R

- Dates are a special format in R (character with quasi-numeric properties)

```
load(file="data/Pres2020.PV.Rdata")
election.day <- as.Date("11/3/2020", "%m/%d/%Y")
election.day16 <- as.Date("11/8/2016", "%m/%d/%Y")
```

- Difference in “dates” versus difference in integers?

```
election.day - election.day16
```

```
## Time difference of 1456 days
```

```
as.numeric(election.day - election.day16)
```

```
## [1] 1456
```

Initial Questions

- How many polls were publicly done and reported in the media about the national popular vote?
- When did the polling occur? Did most of the polls occur close to Election Day?

Alternative Questions using similar code but different data!

- How does the pattern in 2020 compare to past patterns?
- What does the pattern look like in upcoming elections? (NJ/VA/2022)
- How has the (per capita?) number of COVID cases/deaths/hospitalizations (in a county/state/country?) changed over time?
- How does the performance of an NBA Team (or player) vary over the course of a season in terms of Y?

Conditional Relationships

- How does the value of the outcome of interest vary *depending* on the value of another variable of interest?
- Outcome of interest (dependent variable, Y)
- Other variables possibly related to the outcome (independent variable, X)

Conditional Relationships

- How does the value of the outcome of interest vary *depending* on the value of another variable of interest?
- Outcome of interest (dependent variable, Y)
- Other variables possibly related to the outcome (independent variable, X)

Y: Number of Polls being reported on X: Proximity to Election Day

Conditional Relationships

- How does the value of the outcome of interest vary *depending* on the value of another variable of interest?
- Outcome of interest (dependent variable, Y)
- Other variables possibly related to the outcome (independent variable, X)

Y: Number of Polls being reported on X: Proximity to Election Day

So, for every day, how many polls were reported by the media?

Let's Wrangle. . .

```
Pres2020.PV <- Pres2020.PV %>%
                mutate(EndDate = as.Date(Pres2020.PV$EndDate, "%
                      StartDate = as.Date(Pres2020.PV$StartDate,
                      DaysToED = as.numeric(election.day - EndDa
                      margin = Biden - Trump)
```

What are we plotting?

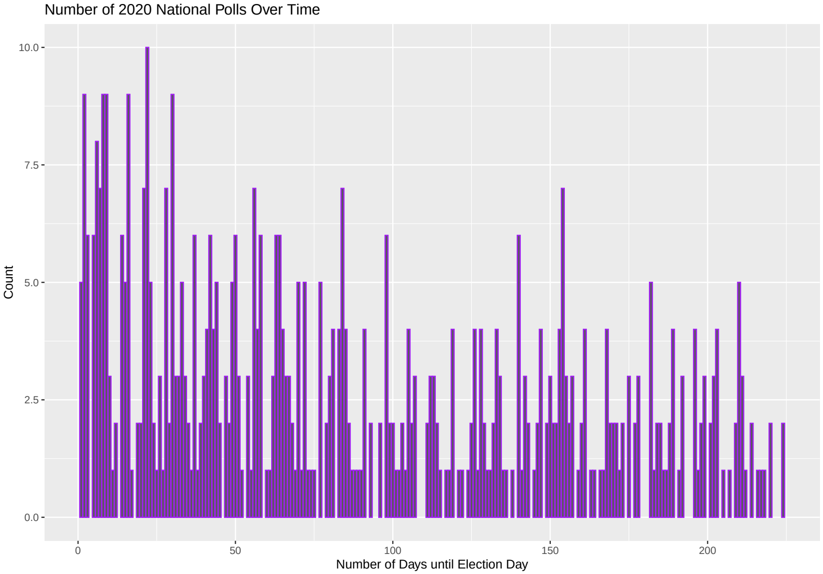
- Media Question: how does the number of polls change over time?
- Data Scientist Question: What do we need to plot? `margin` or `DaysToED`?
- What will each produce?
- Are they *categorical* (barplot) or *continuous* (histogram)?

Barplot

```r
p <- ggplot(data = Pres2020.PV, aes(x = DaysToED)) +
  labs(title = "Number of 2020 National Polls Over Time") +
  labs(x = "Number of Days until Election Day") +
  labs(y = "Count") +
  geom_bar(color="PURPLE")
```

When did polls occur?

```
p
```

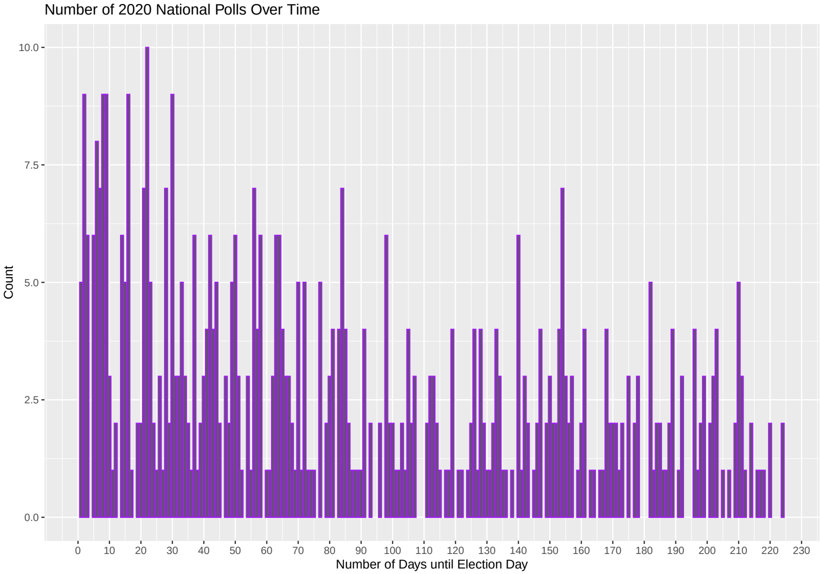


Hmm. . . Better axis?

```r
p <- ggplot(data = Pres2020.PV, aes(x = DaysToED)) +
  labs(title = "Number of 2020 National Polls Over Time") +
  labs(x = "Number of Days until Election Day") +
  labs(y = "Count") +
  geom_bar(color="PURPLE")  +
  scale_x_continuous(breaks=seq(0,230,by=10))
```

Hmm. . . Better axis?

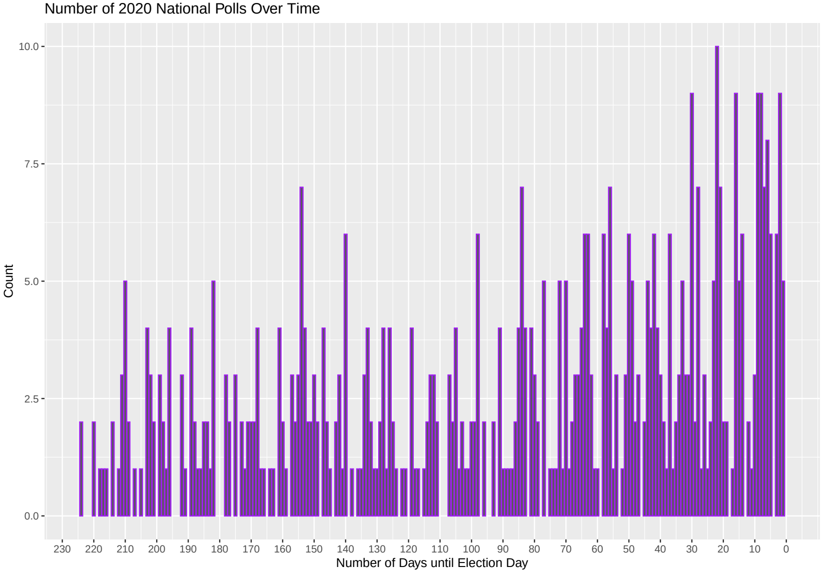
```
p
```



Flipping the scale: November > January

```r
p <- ggplot(data = Pres2020.PV, aes(x = DaysToED)) +
  labs(title = "Number of 2020 National Polls Over Time") +
  labs(x = "Number of Days until Election Day") +
  labs(y = "Count") +
  geom_bar(color="PURPLE")  +
  scale_x_reverse(breaks=seq(0,230,by=10))
```

When did polls occur? November > January



Histogram

```r
p <- ggplot(data = Pres2020.PV, aes(x = DaysToED)) +
  labs(title = "Number of 2020 National Polls Over Time") +
  labs(x = "Number of Days until Election Day") +
  labs(y = "Count") +
  geom_histogram(color="PURPLE",bins = 30) +
  scale_x_reverse()
```

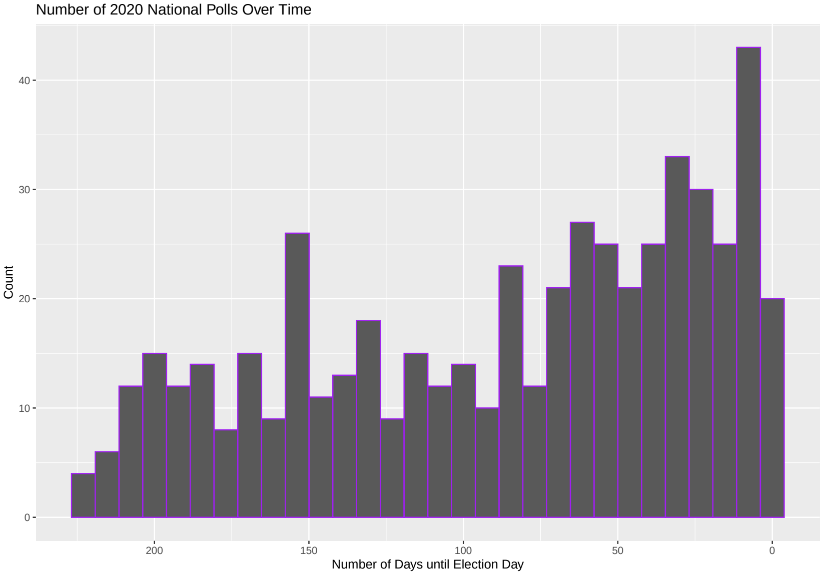
Histogram

```r
p <- ggplot(data = Pres2020.PV, aes(x = DaysToED)) +
  labs(title = "Number of 2020 National Polls Over Time") +
  labs(x = "Number of Days until Election Day") +
  labs(y = "Count") +
  geom_histogram(color="PURPLE",bins = 30) +
  scale_x_reverse()
```

- What does a bin mean in this context?

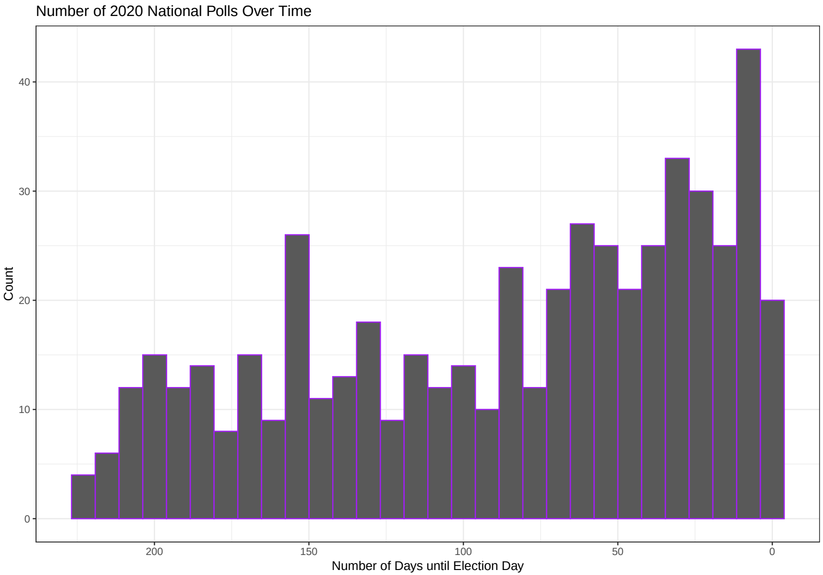
Histogram because nearly continuous?

```
p
```



Get rid of background?

```
p + theme_bw()
```



Bivariate/Multivariate relationships

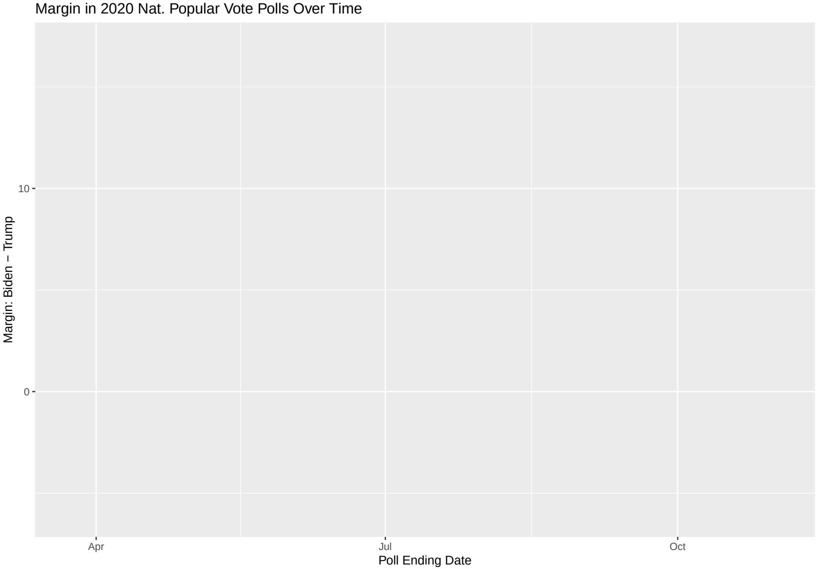
- Most of what we do is a relationship between (at least) 2 variables.
- Here we are interested in how the margin varies as Election Day approaches: `margin` by `DaysToED`.
- Want to plot X (variable that "explains") vs. Y (variable being "explained"):

Scatterplot: Relationship between Continuous Variables

```
margin.plot <- ggplot(Pres2020.PV,
                      aes(x = EndDate, y = margin)) +
  labs(title="Margin in 2020 Nat. Popular Vote Polls Over Time")
  labs(y = "Margin: Biden - Trump") +
  labs(x = "Poll Ending Date")
```

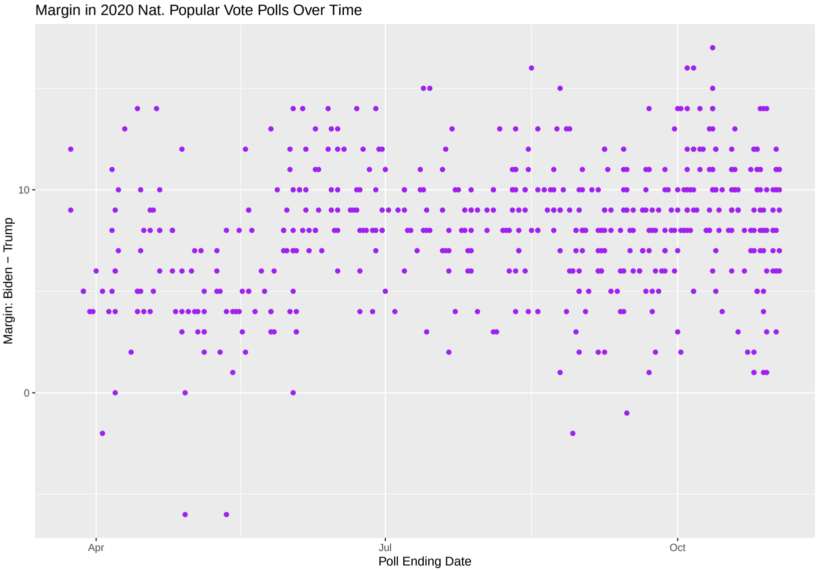
Scatterplot

```
margin.plot
```



Scatterplot: Add Points!

```
margin.plot  + geom_point(color = "PURPLE")
```



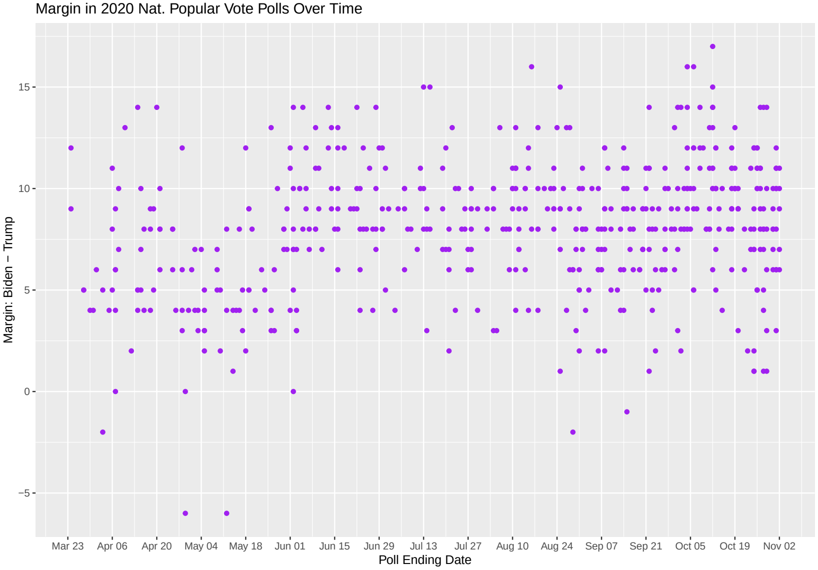
Things that make me sad...

1. Axis looks weird - lots of interpolation required by the consumer.
2. Data looks “chunky”? How many data points are at each point?

Fix Axis Scale!

```r
margin.plot  +
    geom_point(color = "PURPLE")  +
    scale_y_continuous(breaks=seq(-10,20,by=5)) +
    scale_x_date(date_breaks = "2 week", date_labels = "%b %d")
```

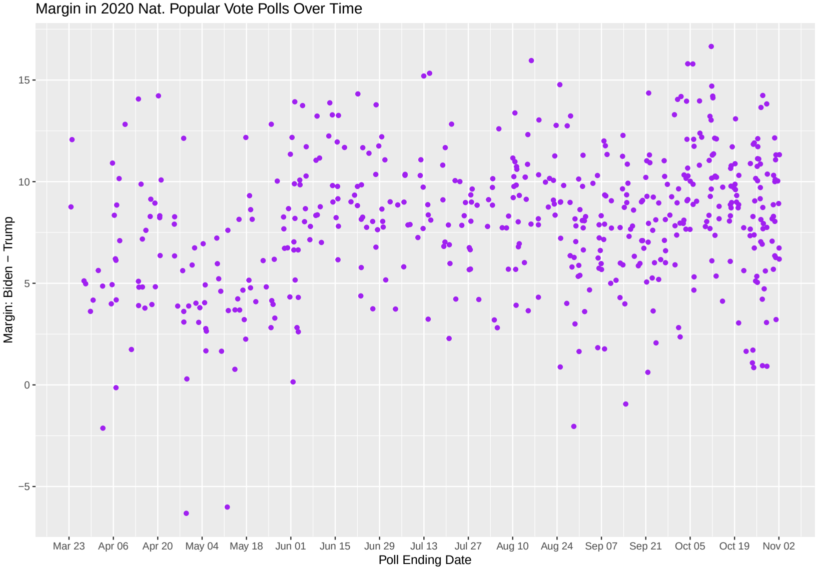
Fix Axis Scale!



Chunky Data? jitter points

```
margin.plot  +
    geom_point(color = "PURPLE", position="jitter") +
    scale_y_continuous(breaks=seq(-10,20,by=5)) +
    scale_x_date(date_breaks = "2 week", date_labels = "%b %d")
```

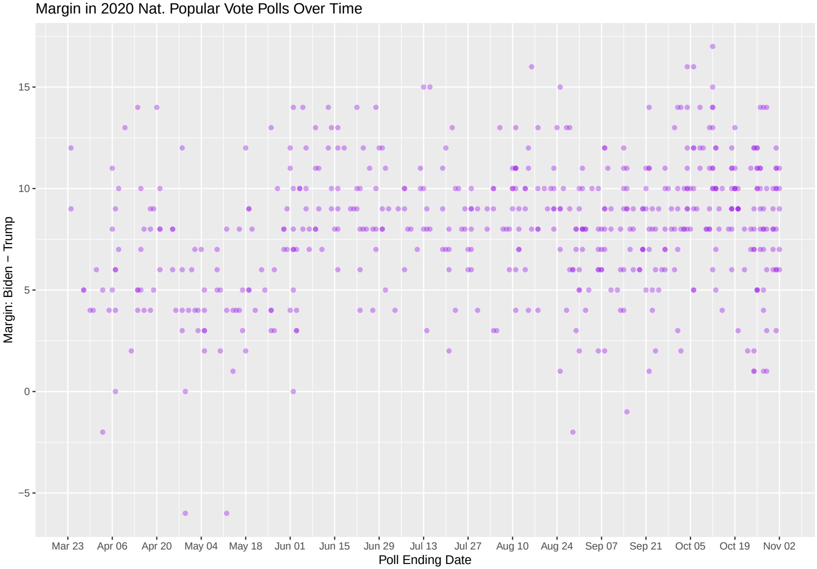
Chunky Data? `jitter points`



Chunky Data? `alpha` points

```
margin.plot  +
    geom_point(color = "PURPLE", alpha = .4) +
    scale_y_continuous(breaks=seq(-10,20,by=5)) +
    scale_x_date(date_breaks = "2 week", date_labels = "%b %d")
```

Chunky Data? `alpha` points



Create Object for later

```r
margin.plot <- margin.plot  +
  geom_point(color = "PURPLE", alpha = .4) +
    scale_y_continuous(breaks=seq(-10,20,by=5)) +
    scale_x_date(date_breaks = "2 week", date_labels = "%b %d")
```

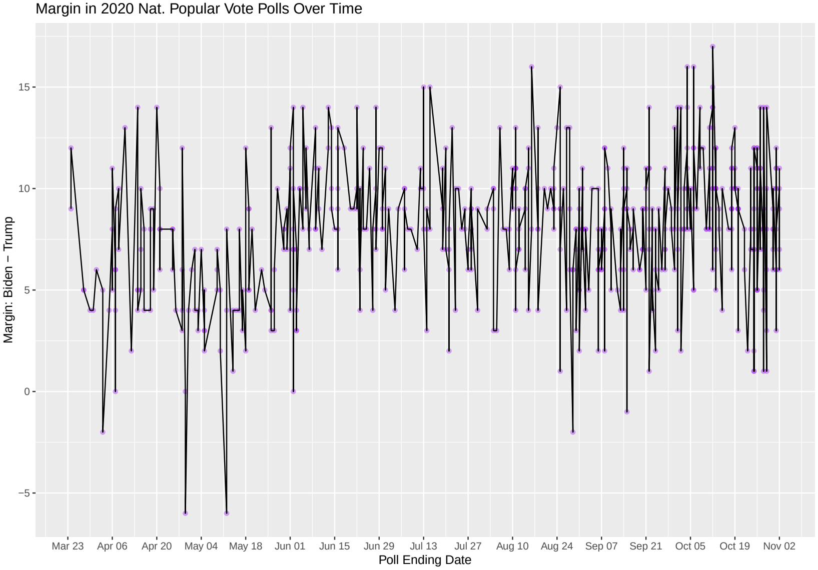
Scatterplot: Add Lines?

```
margin.plot + geom_line()
```

RECAP:

- `geom_point` adds a point at the `(x,y)` point defined in `ggplot` (unless defined in `geom_point`)
- `geom_line` adds a line connecting the `(x,y)` points

Scatterplot: Add Lines?



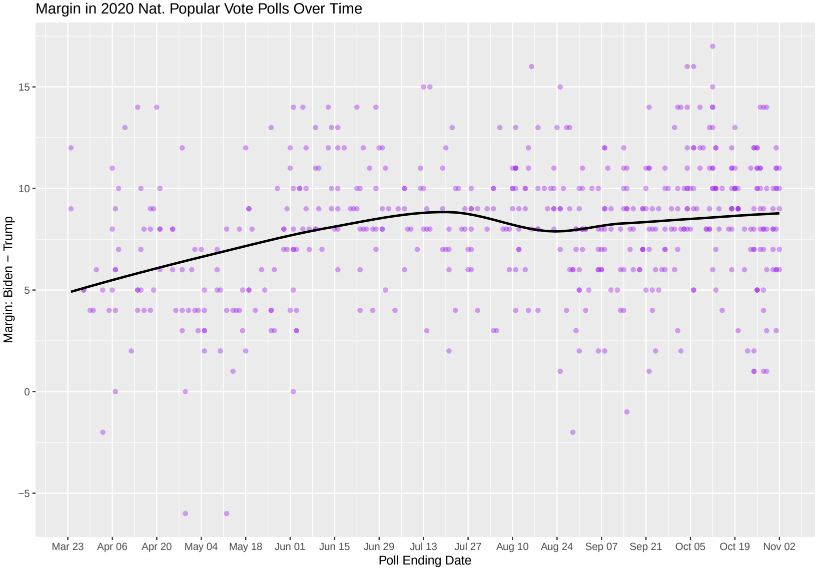
Scatterplot: Add “Smoother”?

```
margin.plot + geom_smooth(color = "BLACK", se=F)
```

- `geom_smooth` adds in a weighted (“smoothed”) average
- BUT: Don’t use what you don’t understand!
- More on smoothing soon!

Scatterplot: Add Smoother?

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



Plotting Multiple Variables Over Time (Time-Series)

- `margin` OK, but limited in what it shows

Plotting Multiple Variables Over Time (Time-Series)

- `margin` OK, but limited in what it shows
- Can we plot support for Biden and support for Trump separately over time (on the same plot)?

“Stretch” Extensions

- Comparing the change in `margin` over time for multiple election years?
- Comparing the support for candidates (`Biden` and `Trump`) in multiple states?
- Comparing the support for candidates according to different types of polls?
- Comparing the support for presidential candidates relative to senatorial and gubernatorial candidates in the same state?

“Stretch” Extensions

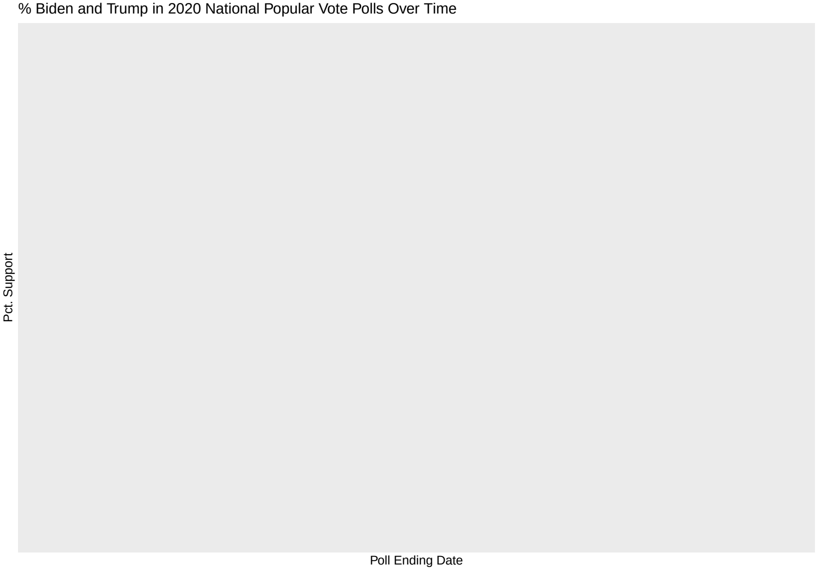
- Comparing the change in `margin` over time for multiple election years?
- Comparing the support for candidates (`Biden` and `Trump`) in multiple states?
- Comparing the support for candidates according to different types of polls?
- Comparing the support for presidential candidates relative to senatorial and gubernatorial candidates in the same state?
- Comparing the deaths/cases per capita over time (and also by county/state)?
- Comparing the performance of an NBA team/player in several dimensions over time?

First, define the canvas!

```
BidenTrumpplot <- ggplot(Pres2020.PV)  +
  labs(title="% Biden and Trump in 2020 National Popular Vote Po
  labs(y = "Pct. Support") +
  labs(x = "Poll Ending Date")
```

Blank scale!

```
BidenTrumpplot
```



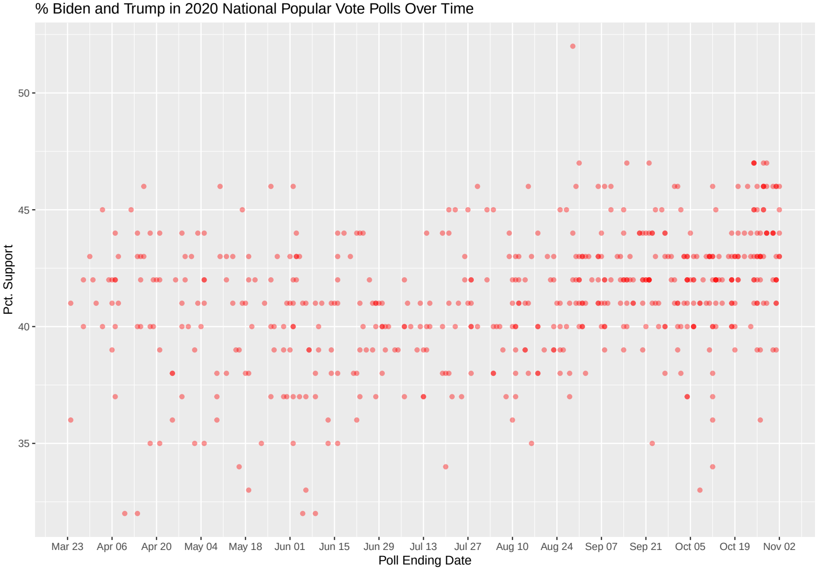
Now, add the points for Trump

```
BidenTrumpplot <- BidenTrumpplot +
  geom_point(aes(x = EndDate, y = Trump),
             color = "red", alpha=.4)  +
  scale_x_date(date_breaks = "2 week", date_labels = "%b %d")
```

- Note the use of aes in geom_point()!

What do you have?

```
BidenTrumpplot
```



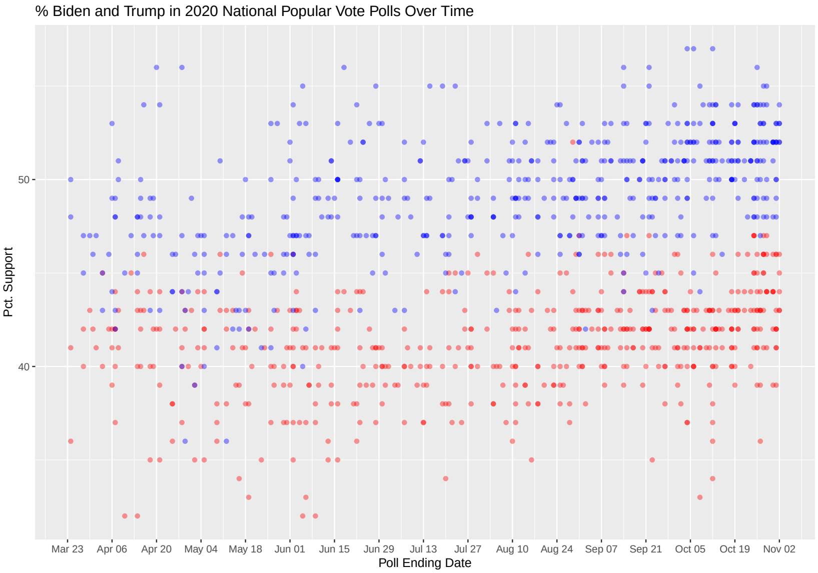
Now, add the points for Biden

```r
BidenTrumpplot <- BidenTrumpplot +
  geom_point(aes(x = EndDate, y = Biden),
             color = "blue", alpha=.4)
```

- ggplot will now rescale y-axis to fit both `Trump` and `Biden`

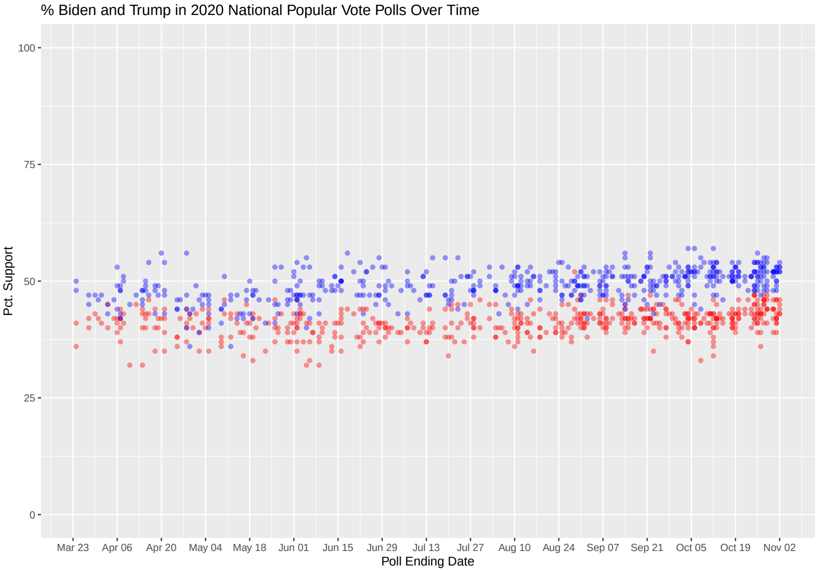
Adding Biden

```
BidenTrumpplot
```



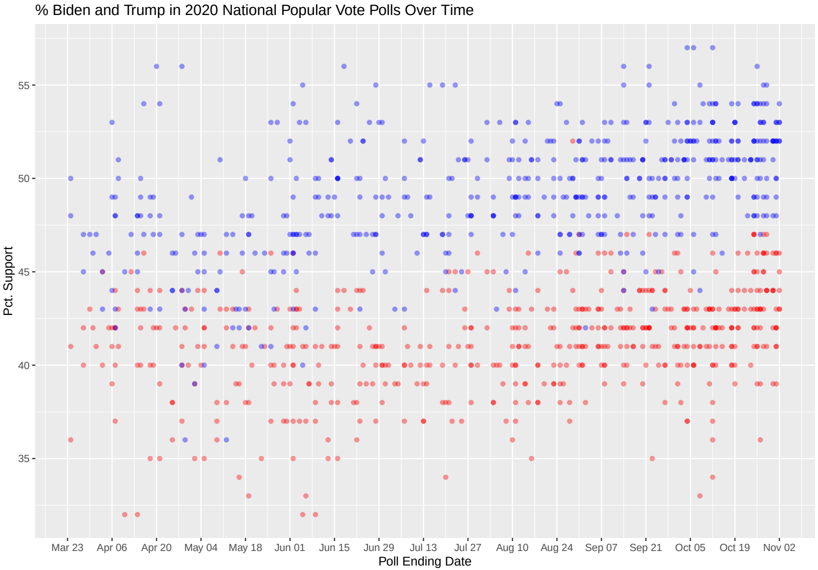
Set the Axis?

```
BidenTrumpplot + ylim(0,100)
```



For reals

```
BidenTrumpplot + scale_y_continuous(breaks=seq(30,70,by=5))
```



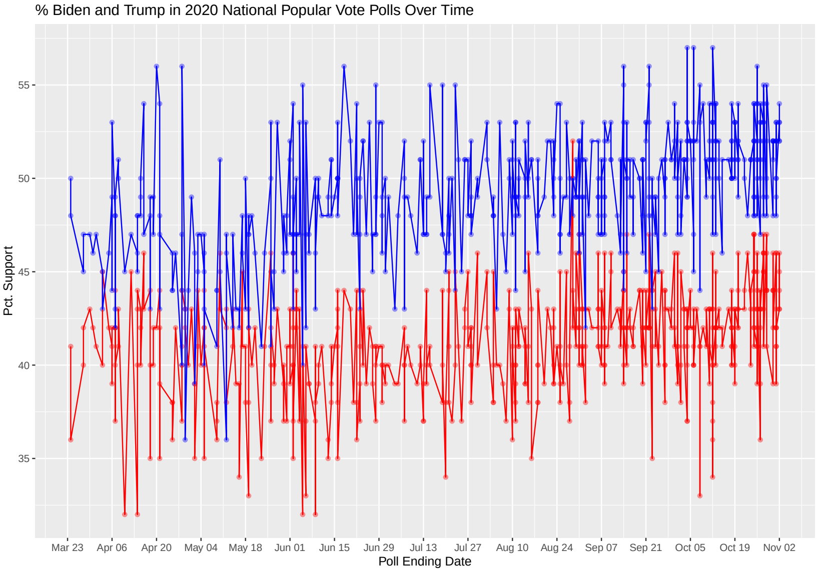
Adding some lines?

```
BTwithlines <- BidenTrumpplot +
  scale_y_continuous(breaks=seq(30,70,by=5)) +
  geom_line(aes(x = EndDate, y = Trump), color = "red") +
  geom_line(aes(x = EndDate, y = Biden), color = "blue")
```

- We add lines the same way we added points!

But we shouldn't...

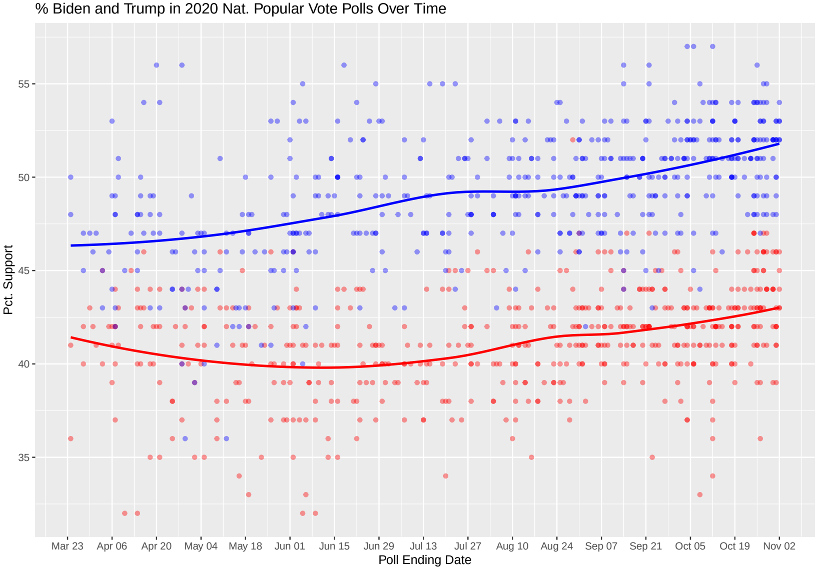
```
BTwithlines
```



Putting it all together

```
BTNational <- ggplot(Pres2020.PV) +
  geom_point(aes(x = EndDate, y = Trump),
             color = "red", alpha = .4) +
  geom_point(aes(x = EndDate, y = Biden),
             color = "blue", , alpha = .4)  +
  geom_smooth(aes(x = EndDate, y = Trump),
              color = "red",se=F) +
  geom_smooth(aes(x = EndDate, y = Biden),
              color = "blue",se=F) +
  labs(title="% Biden and Trump in 2020 Nat. Popular Vote Polls
  labs(y = "Pct. Support") +
  labs(x = "Poll Ending Date") +
  scale_x_date(date_breaks = "2 week", date_labels = "%b %d") +
  scale_y_continuous(breaks=seq(30,70,by=5))
```

BTNational


% Biden and Trump in 2020 Nat. Popular Vote Polls Over Time
Pct. Support
55
50
45
40
35
Mar 23
Apr 06
Apr 20
May 04
May 18
Jun 01
Jun 15
Jun 29
Jul 13
Jul 27
Aug 10
Aug 24
Sep 07
Sep 21
Oct 05
Oct 19
Nov 02
Poll Ending Date