

Multivariate Analysis

Part 2: Visualizations

Prof. Bisbee

Vanderbilt University

Slides Updated: 2024-01-08

Agenda

1. Why did the polls underestimate Trump support?
2. Rules of visualization

2020 polling

- Let's introduce a different dataset!
 - Download and open [Pres2020_PV.Rds](#)
 - Wrangle to get the popular vote margin, expressed in decimals

```
require(tidyverse)
poll <-
read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/Pres2020_PV.Rds')

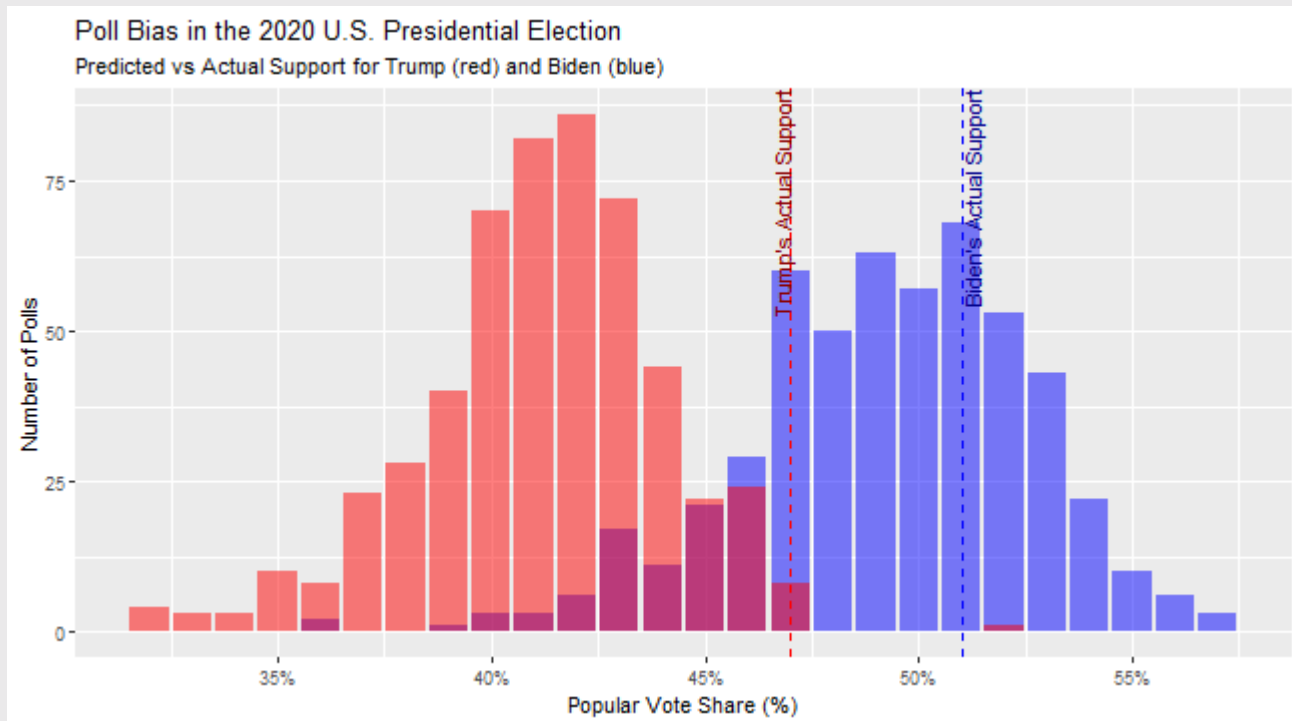
poll <- poll %>%
  mutate(Trump = Trump/100,
         Biden = Biden/100,
         margin = Biden - Trump)
```

The Research Question

```
pRQ <- poll %>%
  ggplot() +
  geom_bar(aes(x = Biden*100),fill = 'blue',alpha = .5) +
  geom_bar(aes(x = Trump*100),fill = 'red',alpha = .5) +
  geom_vline(xintercept = 47,linetype = 'dashed',color= 'red') +
  geom_vline(xintercept = 51,linetype = 'dashed',color= 'blue')+
  annotate(geom = 'text',x = c(47),y = Inf,angle = 90,hjust = 1,vjust
= 0,label = c("Trump's Actual Support"),color = 'darkred') +
  annotate(geom = 'text',x = c(51),y = Inf,angle = 90,hjust = 1,vjust
= 1,label = c("Biden's Actual Support"),color = 'darkblue') +
  labs(title = 'Poll Bias in the 2020 U.S. Presidential Election',
        subtitle = 'Predicted vs Actual Support for Trump (red) and
Biden (blue)',
        x = 'Popular Vote Share (%)',
        y = 'Number of Polls') +
  scale_x_continuous(breaks = seq(30,60,by = 5),labels = function(x)
paste0(x,'%'))
```

The Research Question

pRQ



The Research Question

```
poll %>% # Proportion that under-predict
  summarise(propBidenUP = mean(Biden < .51),
            propTrumpUP = mean(Trump < .47))
```

```
## # A tibble: 1 × 2
##   propBidenUP propTrumpUP
##   <dbl>      <dbl>
## 1      0.612      0.983
```

```
poll %>% # Average under-prediction
  summarise(avgBidenErr = mean(.51 - Biden),
            avgTrumpErr = mean(.47 - Trump))
```

```
## # A tibble: 1 × 2
##   avgBidenErr avgTrumpErr
##   <dbl>      <dbl>
## 1      0.0175      0.0577
```

Theorizing

- **Research Question:** Why do polls under-predict Trump more than Biden?
 1. Unrepresentative samples (how were respondents contacted?)
 2. Small samples (how many respondents?)
 3. Shy Trump Voters / trolls (lying respondents)
 4. Timing (closer to the election → less biased)

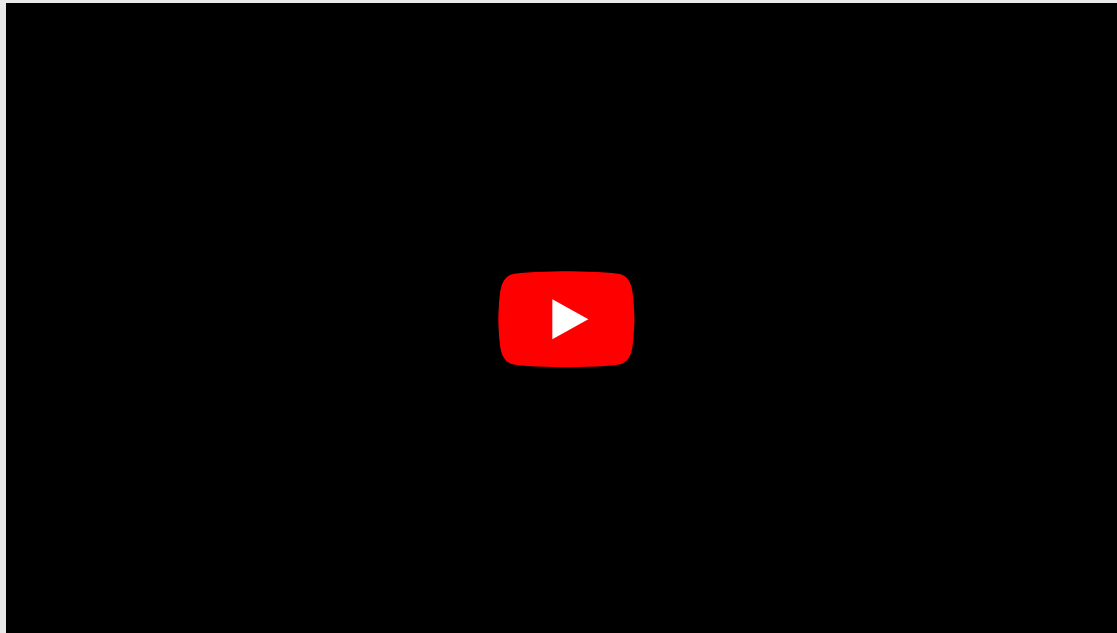
Theorizing

- A fifth explanation?
- Anti-Trump media!



Theorizing

- However...



Theorizing

- Theory #1: Does the "mode" of the survey matter?
 - I.e., if you only call people on landlines, who do you reach?
 - And how might they differ from the general population?
- **Assumption 1**: Younger people do not use landlines, meaning that surveys which rely on **random digit dialing** (RDD) will get disproportionately older respondents.
- **Assumption 2**: Younger voters are more progressive, making them less likely to support Trump.
- **Theory**: Surveys that use RDD will find more support for Trump than Biden.

Analyzing

- Plot the Biden-Trump vote margin by mode type

```
poll %>%  
  count(Mode)
```

```
## # A tibble: 9 × 2  
##   Mode          n  
##   <chr>      <int>  
## 1 IVR          1  
## 2 IVR/Online   47  
## 3 Live phone - RBS 13  
## 4 Live phone - RDD 51  
## 5 Online      366  
## 6 Online/Text    1  
## 7 Phone - unknown 1  
## 8 Phone/Online   19  
## 9 <NA>          29
```

- So many modes of interviewing people!

(Soft) Rules of Visualization

- Variable `type` informs visualization

1. Univariate

- Categorical data: `geom_bar()`
- Continuous data: `geom_histogram()` or `geom_density()`

2. Bivariate

- Categorical X Categorical: `geom_bar()`
- Binary X Continuous: `geom_histogram()` or `geom_density()`
- Categorical X Continuous: `geom_boxplot()` or `geom_violin()`
- Continuous X Continuous: `geom_point()`

Beyond Bivariate

1. Trivariate

- Categorical X Categorical X Continuous: `geom_tile()`
- Continuous X Continuous X Categorical: `geom_point()` + `color`
- Continuous X Continuous X Continuous: `geom_point()` + `color/size`
- Latitude X Longitude X Categorical / Continuous: Maps!
- Var X Var X Time: Animated!
- (Beyond the scope of this course, but get creative!)

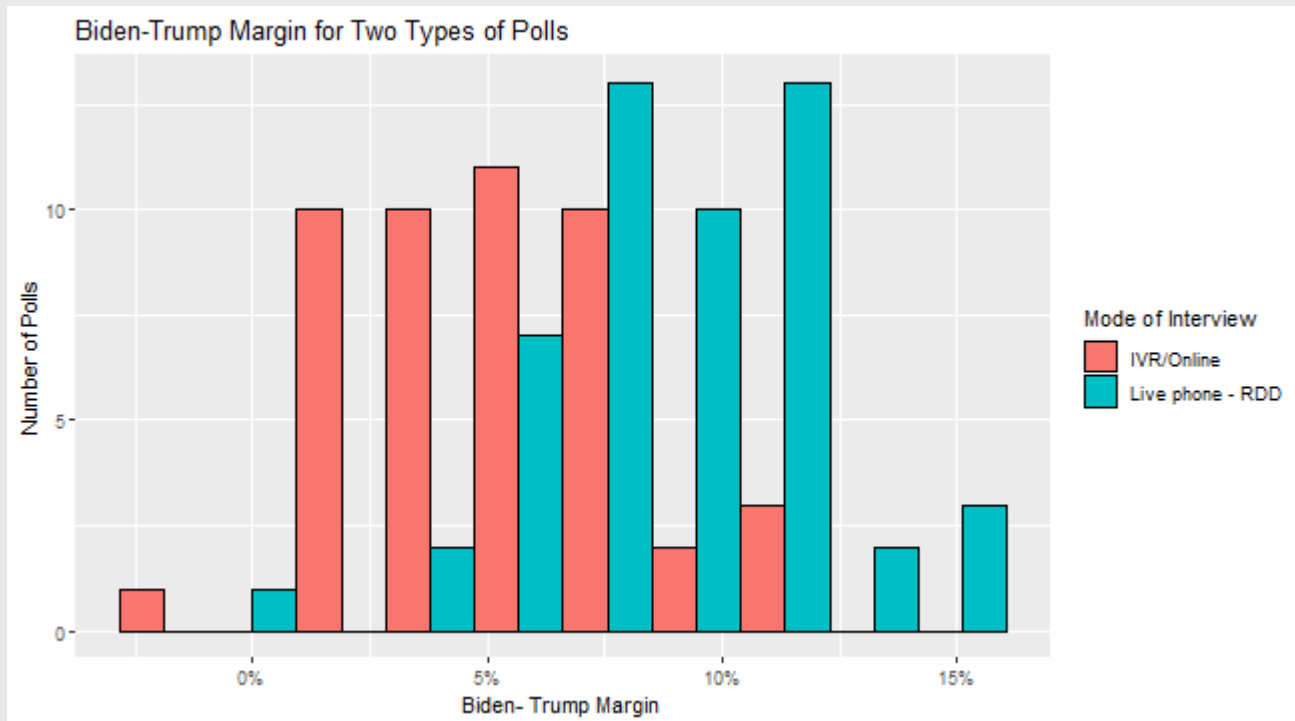
Analyzing

- For now, just focus on **IRV/Online** versus **Live phone - RDD**
- Since **margin** is a continuous variable, use **geom_histogram**

```
pMode <- poll %>%  
  filter(Mode == "IVR/Online" | Mode == "Live phone - RDD") %>%  
  ggplot(aes(x= margin, fill = Mode)) +  
  labs(y = "Number of Polls",  
        x = "Biden- Trump Margin",  
        title = "Biden-Trump Margin for Two Types of Polls",  
        fill = "Mode of Interview") +  
  geom_histogram(bins=10, color="black", position="dodge") +  
  scale_x_continuous(breaks=seq(-.1,.2,by=.05),  
                    labels= scales::percent_format(accuracy = 1))
```

Mode Matters!

pMode



- But results are **inconsistent** with our [theory](#)!

Visualization

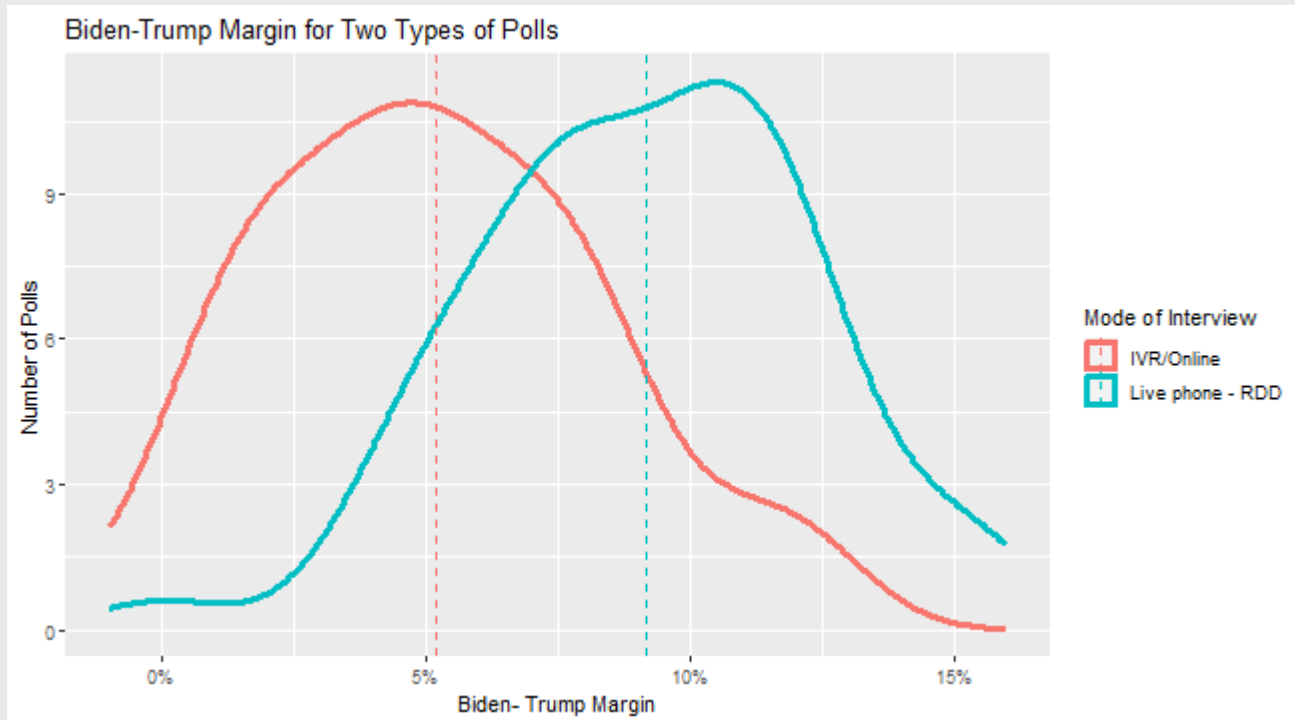
- How can we improve this? Perhaps `geom_density()` and `geom_vline()`?

```
toplot <- poll %>%  
  filter(Mode == "IVR/Online" | Mode == "Live phone - RDD")  
  
pModeDens <- toplot %>%  
  ggplot(aes(x= margin, color = Mode)) +  
  labs(y = "Number of Polls",  
       x = "Biden- Trump Margin",  
       title = "Biden-Trump Margin for Two Types of Polls",  
       color = "Mode of Interview") +  
  geom_density(lwd = 1.2) +  
  scale_x_continuous(breaks=seq(-.1,.2,by=.05),  
                    labels= scales::percent_format(accuracy = 1)) +  
  geom_vline(data = toplot %>%  
            group_by(Mode) %>%  
            summarise(margin = mean(margin)),aes(xintercept =  
margin,color = Mode),linetype = 'dashed')
```


Visualization

- How can we improve this? Perhaps `geom_density()` and `geom_vline()`?

pModeDens



More Modes

- `geom_histogram()` and `geom_density()` less useful for more comparisons
- First, let's drop modes that were hardly used

```
(toKeep <- poll %>%  
  count(Mode) %>%  
  filter(n > 5,  
    !is.na(Mode)))
```

```
## # A tibble: 5 × 2  
##   Mode          n  
##   <chr>      <int>  
## 1 IVR/Online      47  
## 2 Live phone - RBS  13  
## 3 Live phone - RDD  51  
## 4 Online        366  
## 5 Phone/Online    19
```

```
toplot <- poll %>% filter(Mode %in% toKeep$Mode)
```

More Modes

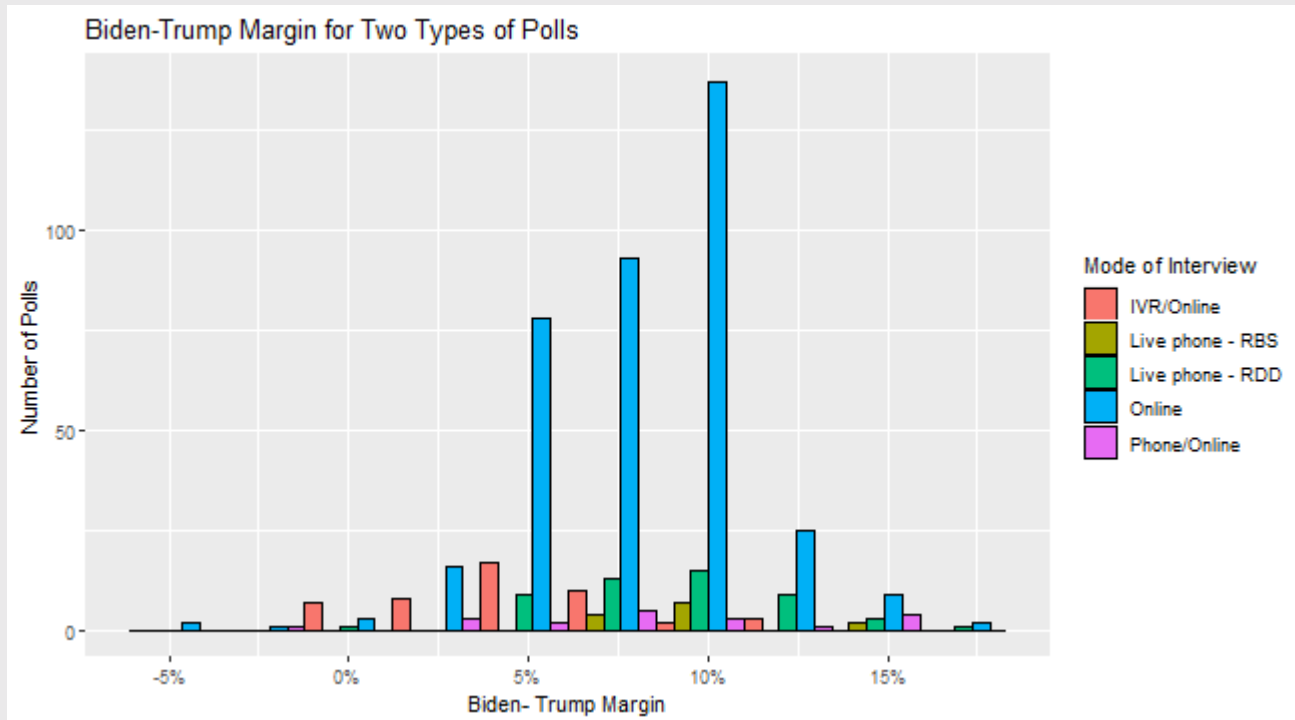
- How hard is `geom_histogram()` with more categories?

```
pModeHist <- toplot %>%  
  ggplot(aes(x= margin, fill = Mode)) +  
  labs(y = "Number of Polls",  
       x = "Biden- Trump Margin",  
       title = "Biden-Trump Margin for Two Types of Polls",  
       fill = "Mode of Interview") +  
  geom_histogram(color = 'black', position = 'dodge', bins = 10) +  
  scale_x_continuous(breaks=seq(-.1,.2,by=.05),  
                    labels= scales::percent_format(accuracy = 1))
```

More Modes

- How hard is `geom_histogram()` with more categories?

pModeHist



More Modes

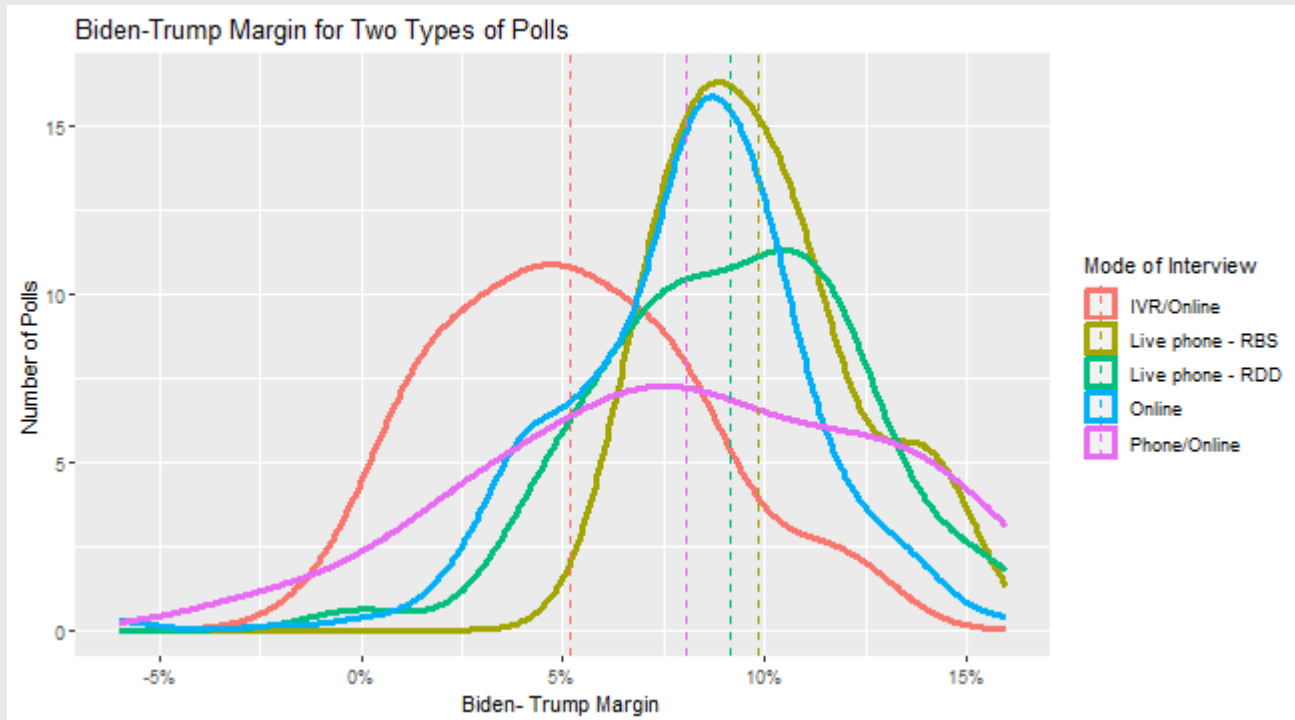
- How hard is `geom_density()` with more categories?

```
pModeDens <- toplot %>%  
  ggplot(aes(x= margin, color = Mode)) +  
  labs(y = "Number of Polls",  
       x = "Biden- Trump Margin",  
       title = "Biden-Trump Margin for Two Types of Polls",  
       color = "Mode of Interview") +  
  geom_density(lwd = 1.2) +  
  scale_x_continuous(breaks=seq(-.1,.2,by=.05),  
                    labels= scales::percent_format(accuracy = 1)) +  
  geom_vline(data = toplot %>%  
            group_by(Mode) %>%  
            summarise(margin = mean(margin)),aes(xintercept =  
margin,color = Mode),linetype = 'dashed')
```

More Modes

- How hard is `geom_density()` with more categories?

pModeDens



geom_boxplot()

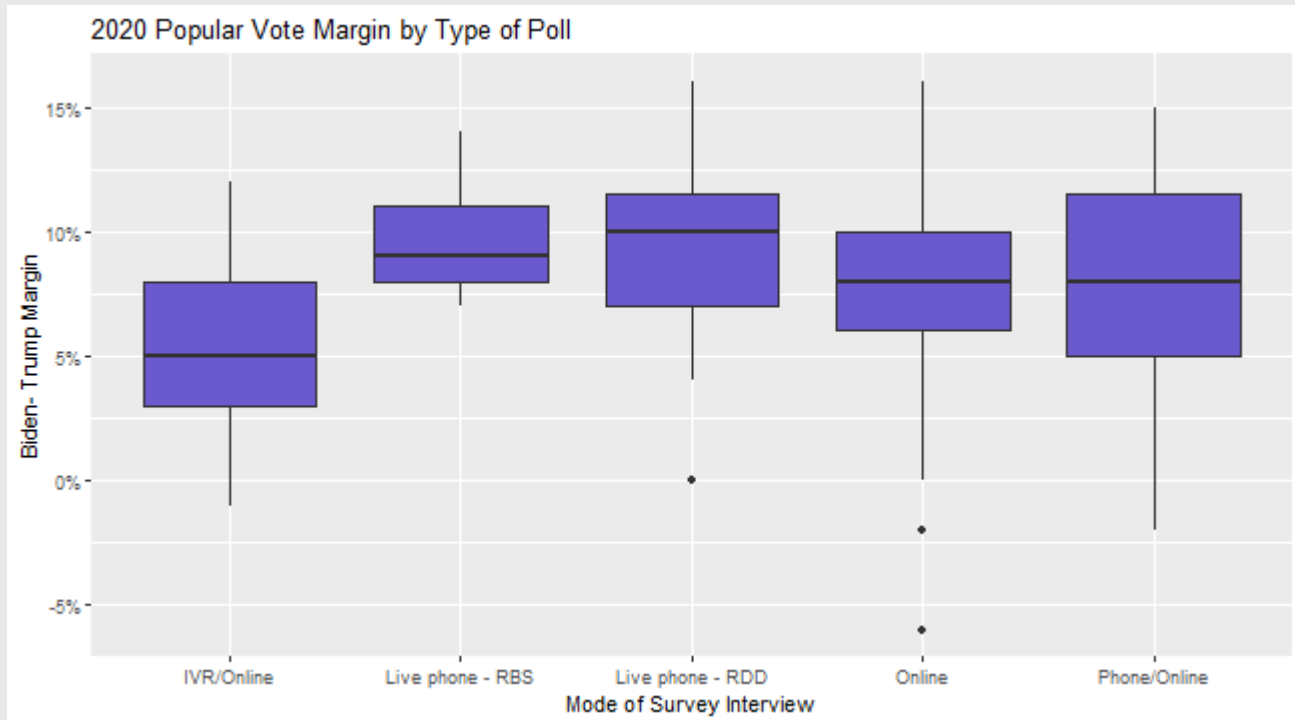
- More categories requires more compact ways of visualizing distributions

```
pModeBox <- topplot %>%  
  ggplot(aes(x = Mode, y = margin)) +  
    labs(x = "Mode of Survey Interview",  
         y = "Biden- Trump Margin",  
         title = "2020 Popular Vote Margin by Type of Poll") +  
    geom_boxplot(fill = "slateblue") +  
    scale_y_continuous(breaks=seq(-.1,.2,by=.05),  
                      labels= scales::percent_format(accuracy = 1))
```

geom_boxplot()

- More categories requires more compact ways of visualizing distributions

pModeBox



Ordering Unordered Categories

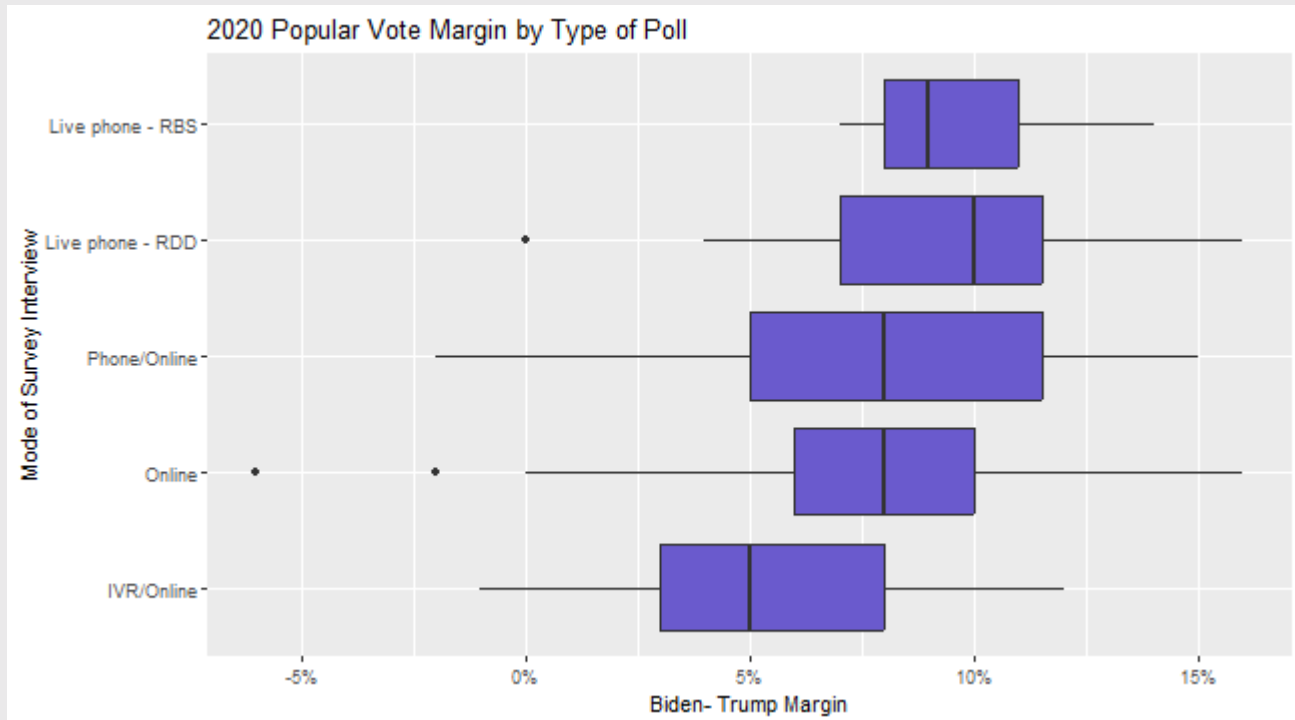
- We can use `reorder()` to arrange categories by the data

```
pModeBox <- topplot %>%  
  ggplot(aes(x = reorder(Mode,margin), y = margin)) +  
    labs(x = "Mode of Survey Interview",  
         y = "Biden- Trump Margin",  
         title = "2020 Popular Vote Margin by Type of Poll") +  
    geom_boxplot(fill = "slateblue") +  
    scale_y_continuous(breaks=seq(-.1,.2,by=.05),  
                      labels= scales::percent_format(accuracy = 1))
```

Ordering Unordered Categories

- We can use `reorder()` to arrange categories by the data

```
pModeBox + coord_flip()
```



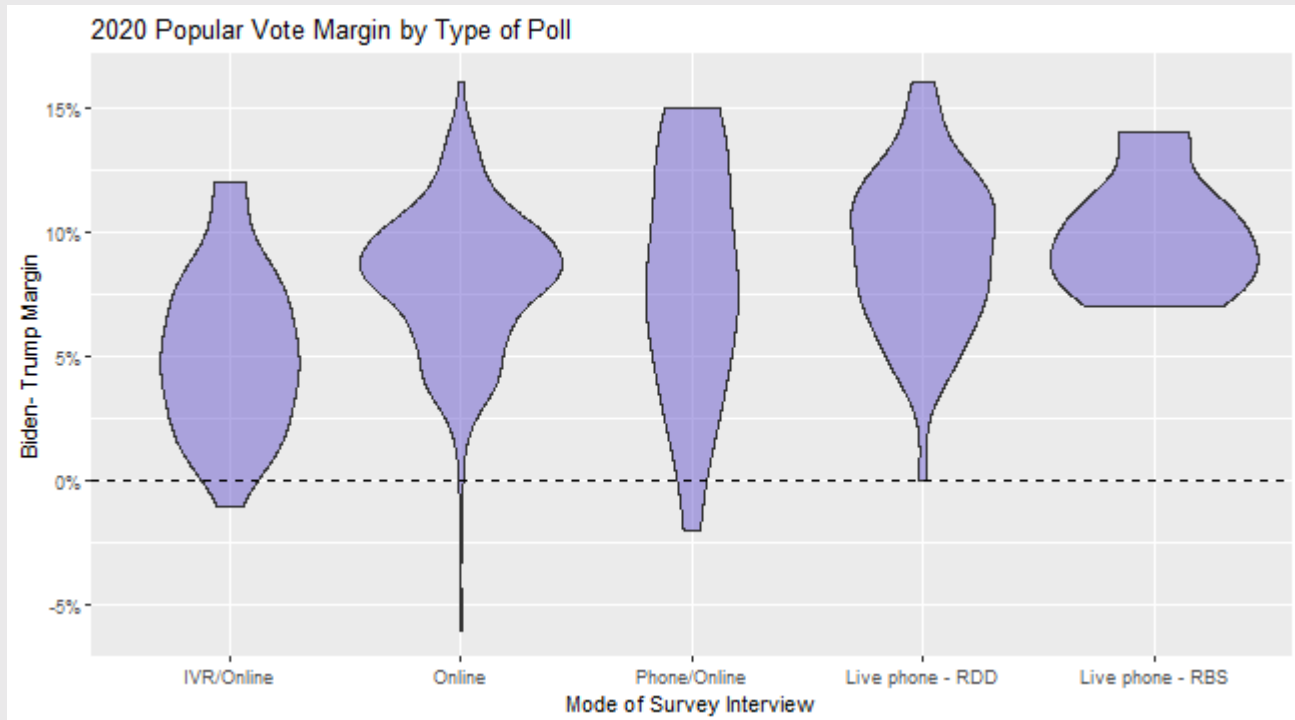
geom_violin()

- Boxplots are cleaner than densities and histograms for multiple categories
- But we lose ability to see distributions within the 80% box

```
pModeViol <- topplot %>%  
  ggplot(aes(x = reorder(Mode,margin), y = margin)) +  
    labs(x = "Mode of Survey Interview",  
         y = "Biden- Trump Margin",  
         title = "2020 Popular Vote Margin by Type of Poll") +  
    geom_violin(fill = "slateblue",alpha = .5) +  
    scale_y_continuous(breaks=seq(-.1,.2,by=.05),  
                      labels= scales::percent_format(accuracy = 1))
```

geom_violin()

```
pModeViol + geom_hline(yintercept = 0, linetype = 'dashed')
```



Continuous by Continuous

- For **conditional relationships** between two **continuous variables**, use `geom_point()`
- **Theory**: Are polls politically biased?
 - I.e., a Biden-friendly poll might **under**predict Trump support and **over**predict Biden support
- **Data**: Trump support conditional on Biden support

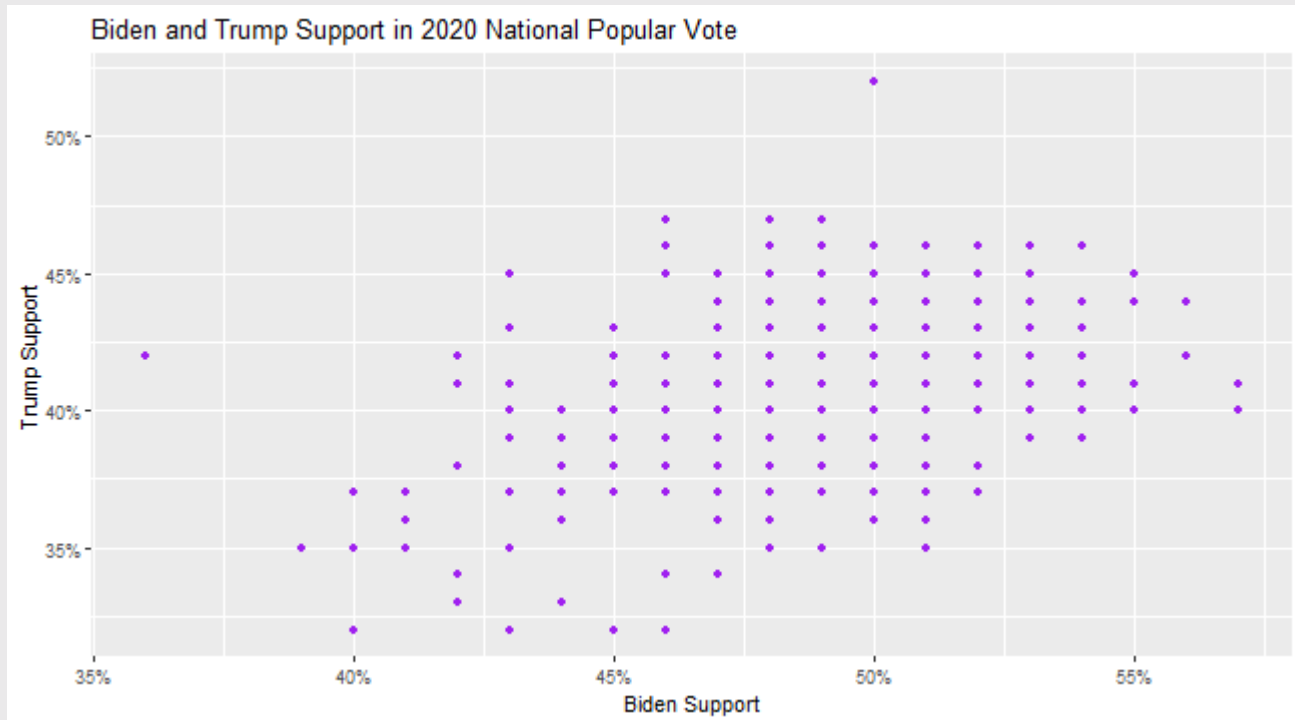
Analysis

- Plot Trump support versus Biden support

```
pSupp <- poll %>%  
  ggplot(aes(x = Biden, y = Trump)) +  
  labs(title="Biden and Trump Support in 2020 National Popular Vote",  
        y = "Trump Support",  
        x = "Biden Support") +  
  geom_point(color="purple") +  
    scale_y_continuous(breaks=seq(0,1,by=.05),  
                      labels= scales::percent_format(accuracy = 1)) +  
  scale_x_continuous(breaks=seq(0,1,by=.05),  
                    labels= scales::percent_format(accuracy = 1))
```

geom_scatter()

pSupp



- How many observations are at each point?

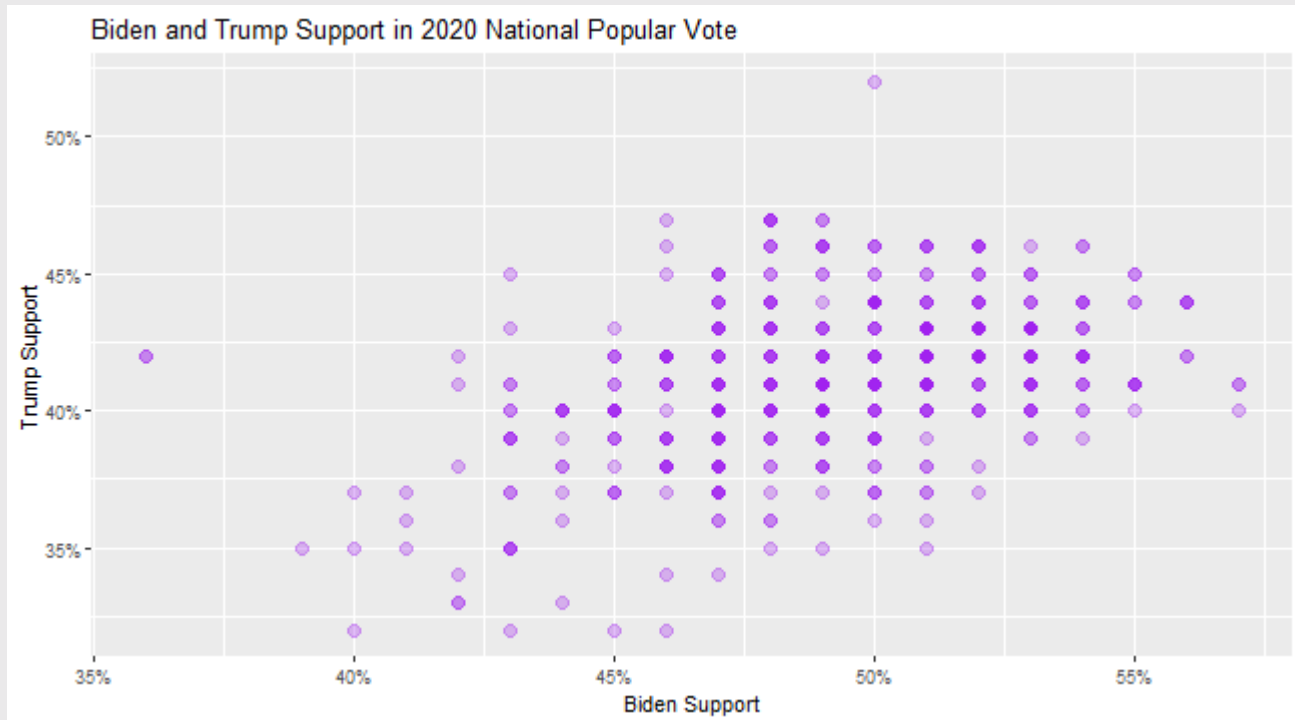
Tweaking `alpha`

- We can set the transparency of each point such that multiple points will show up darker
 - I.e., `alpha=.3` means that a single point will be 70% transparent, but 3 points on top of each other will be 10% transparent

```
pSupp <- poll %>%
  ggplot(aes(x = Biden, y = Trump)) +
  labs(title="Biden and Trump Support in 2020 National Popular Vote",
        y = "Trump Support",
        x = "Biden Support") +
  geom_point(color="purple", alpha = .3, size = 3) +
  scale_y_continuous(breaks=seq(0,1,by=.05),
                    labels= scales::percent_format(accuracy = 1)) +
  scale_x_continuous(breaks=seq(0,1,by=.05),
                    labels= scales::percent_format(accuracy = 1))
```


Tweaking α

pSupp



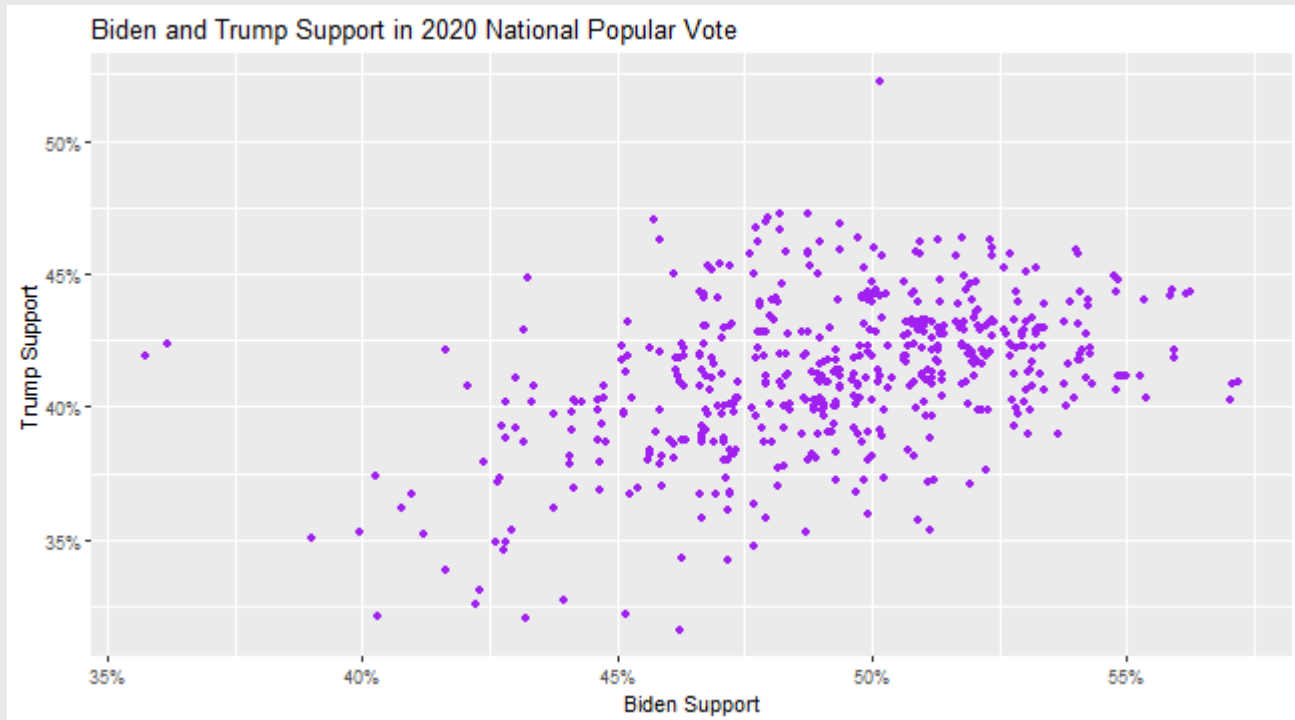
geom_jitter()

- Instead, we could "jitter" the points
 - This adds some random noise to each point to shake them off each other

```
pSupp <- poll %>%  
  ggplot(aes(x = Biden, y = Trump)) +  
  labs(title="Biden and Trump Support in 2020 National Popular Vote",  
        y = "Trump Support",  
        x = "Biden Support") +  
  geom_jitter(color="purple") +  
    scale_y_continuous(breaks=seq(0,1,by=.05),  
                      labels= scales::percent_format(accuracy = 1)) +  
  scale_x_continuous(breaks=seq(0,1,by=.05),  
                    labels= scales::percent_format(accuracy = 1))
```

geom_jitter()

pSupp



size

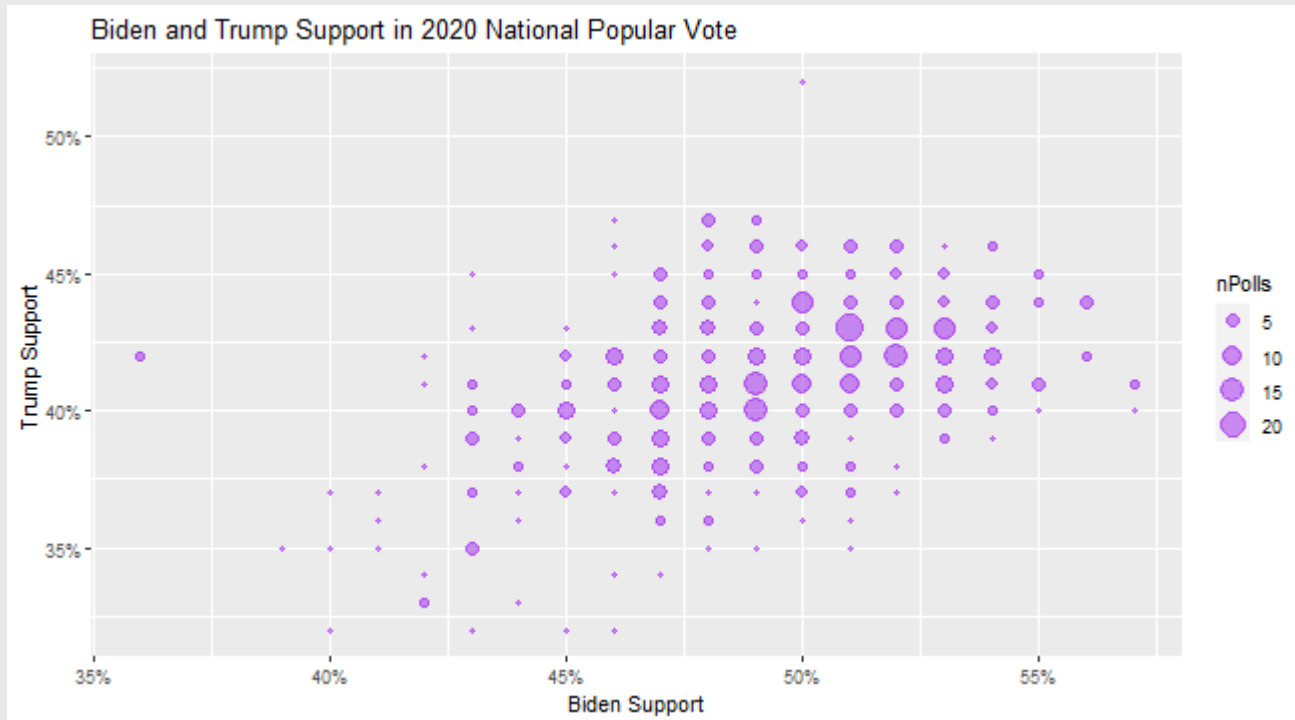
- Finally, we could simply count the number of polls at each x,y coordinate
 - Then size the points by the number of polls

```
pSupp <- poll %>%  
  group_by(Biden,Trump) %>%  
  summarise(nPolls = n()) %>%  
  ggplot(aes(x = Biden, y = Trump,size = nPolls)) +  
  labs(title="Biden and Trump Support in 2020 National Popular Vote",  
        y = "Trump Support",  
        x = "Biden Support") +  
  geom_point(color="purple",alpha = .5) +  
    scale_y_continuous(breaks=seq(0,1,by=.05),  
                      labels= scales::percent_format(accuracy = 1)) +  
  scale_x_continuous(breaks=seq(0,1,by=.05),  
                    labels= scales::percent_format(accuracy = 1))
```

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

size

pSupp



Theory

- These results indicate that polls which predict greater support for Biden **also** predict greater support for Trump
 - Is this consistent with the theory?
 - Recall that **Biden-biased** polls should underpredict Trump support and overpredict Biden support
 - In the **data**, this would suggest a **negative** relationship
 - But we find a **positive** relationship
- **Inconsistent** with the theory, but raises another puzzle
- Why do polls that underpredict support for Biden also underpredict support for Trump?
 - **Third party bias?** Polls bias against 3rd party candidates
 - **Timing of poll?** Fewer uncertain responses closer to election
 - More next time!

Quiz & Homework

- Go to Brightspace and take the **7th** quiz
 - The password to take the quiz is ####
- **Homework:**
 1. Work through ds1000_hw_8.Rmd
 2. Problem Set 4 (Brightspace)