

# Multivariate Analysis

## Part 1: Conditional Relationships

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/09/25

Slides Updated: 2023-09-23

# Agenda

1. Multivariate
2. What is "conditional"?
3. (Re-)Introducing the data
4. Visualization Principles

# Definition

- Multi + variate
  - Many + variables
  - Analysis of multiple variables
- When we analyze **multiple** variables, we are in the world of "conditional analysis"

# What is conditional?

- Put simply: "conditional" means "depending on"
  - I.e., How does a variable of interest vary *depending on* some other variable?
  - "Variable of interest": the **outcome** (or **dependent** variable  $Y$ )
  - "Some other variable": the **predictor** (or **independent** variable  $X$ )
  - "Vary depending on": the **relationship**
- Mapping concepts into data science
  - The relationship between the outcome and the predictor

# What is conditional?

- "Depending on" suggests a **causal** interpretation
  - High wages "depend on" education → education **causes** high wages
  - In **theory**, this is reasonable: students acquire skills in school which are valued by the labor market.
  - But the positive correlation between education and wages might also be "**spurious**"
  - Higher education *AND* higher wages are outcomes of some **true cause** (i.e., upbringing, SES, etc.)

**NOTE: The logic for why a relationship might be spurious is itself CAUSAL.**

# (Re-)Introducing the Data

- Using the Michigan exit poll data
- Download pre-wrangled data from [GitHub](#) and save to your `data` folder.
- `require(tidyverse)` and `readRDS()` the data to `mi_ep` object

```
require(tidyverse)
mi_ep <- read_rds('../data/MI2020_ExitPoll_small.Rds')
```

# Some Light Data Science

- The "gender gap" in Trump support
- **Theory**: Trump has expressed sexist views against women. Therefore, women should be less likely to support him.
  - **NOTE** the causal assumptions in this theory!
- **Analysis**: compare support for Trump among men and women
- But first, some quick data wrangling

```
MI_final_small <- mi_ep %>%  
  filter(preschoice=="Donald Trump, the Republican" |  
  preschoice=="Joe Biden, the Democrat") %>%  
  mutate(BidenVoter=ifelse(preschoice=="Joe Biden, the  
Democrat",1,0),  
         TrumpVoter=ifelse(BidenVoter==1,0,1),  
         AGE10=ifelse(AGE10==99,NA,AGE10))
```

# Conditional Means

```
MI_final_small %>%  
  count(preschoice,SEX) %>%  
  mutate(PctSupport = n/sum(n),  
         PctSupport = round(PctSupport, digits=2))
```

```
## # A tibble: 4 × 4  
##   preschoice      SEX      n PctSupport  
##   <chr>      <dbl> <int>      <dbl>  
## 1 Donald Trump, the Republican      1    247      0.21  
## 2 Donald Trump, the Republican      2    212      0.18  
## 3 Joe Biden, the Democrat           1    304      0.26  
## 4 Joe Biden, the Democrat           2    419      0.35
```

- **Results** are **consistent** with the **theory**
  - NB: **results** do not **prove** the **theory**



# Conditional Means

- However, note that these proportions are out of *all* voters.
- This isn't directly addressing the [theory](#)
  - We want to know the proportion **of women** who supported Trump

```
MI_final_small %>%  
  count(preschoice,SEX) %>%  
  group_by(SEX) %>%  
  mutate(totGender = sum(n)) %>%  
  mutate(pctSupport = n / totGender)
```

```
## # A tibble: 4 × 5  
## # Groups:   SEX [2]  
##   preschoice      SEX      n totGender pctSu...1  
##   <chr>      <dbl> <int>      <int>    <dbl>  
## 1 Donald Trump, the Republican      1    247      551    0.448  
## 2 Donald Trump, the Republican      2    212      631    0.336  
## 3 Joe Biden, the Democrat           1    304      551    0.552  
## 4 Joe Biden, the Democrat           2    419      631    0.664  
## # ... with abbreviated variable name 1pctSupport
```

# Additional Theorizing

- The strength of the theorized relationship might vary by age
  - Younger women might be more offended by Trump's casual sexism
  - Older women might be more inured to Trump's casual sexism
- **Theory**: the "gender gap" will be larger among younger voters
  - (But also recognize that younger Americans are generally more progressive...meaning that **both** younger men and women are more offended by Trump's casual sexism!)

# Two-Way Conditional Means

- We could just subset with `filter()`

```
MI_final_small %>%  
  filter(AGE10==1) %>%  
  group_by(SEX) %>%  
  count(preschoice) %>%  
  mutate(PctSupport = n/sum(n),  
         PctSupport = round(PctSupport, digits=2))
```

```
## # A tibble: 4 × 4  
## # Groups:   SEX [2]  
##   SEX preschoice          n PctSupport  
##   <dbl> <chr>          <int>      <dbl>  
## 1     1 Donald Trump, the Republican     7      0.44  
## 2     1 Joe Biden, the Democrat         9      0.56  
## 3     2 Donald Trump, the Republican     1      0.06  
## 4     2 Joe Biden, the Democrat        15      0.94
```

# Two-Way Conditional Means

- Or we could add `AGE10` to the `group_by`

```
MI_final_small %>%  
  group_by(SEX, AGE10) %>%  
  summarize(PctTrump = mean(TrumpVoter), .groups = 'drop') %>%  
  mutate(PctTrump = round(PctTrump, digits = 2))
```

```
## # A tibble: 22 × 3  
##       SEX AGE10 PctTrump  
##   <dbl> <dbl>   <dbl>  
## 1     1     1     0.44  
## 2     1     2     0.42  
## 3     1     3     0.42  
## 4     1     4     0.24  
## 5     1     5     0.42  
## 6     1     6     0.58  
## 7     1     7     0.54  
## 8     1     8     0.44  
## 9     1     9     0.39  
## 10    1    10     0.43  
## # ... with 12 more rows
```

# Two-Way Conditional Means

- A little hard to make comparisons

```
MI_final_small %>%  
  group_by(SEX, AGE10) %>%  
  summarize(PctTrump = mean(TrumpVoter), .groups = 'drop') %>%  
  spread(SEX, PctTrump) %>% rename(Male = `1`, Female = `2`)
```

```
## # A tibble: 11 × 3  
##   AGE10  Male Female  
##   <dbl> <dbl> <dbl>  
## 1      1  0.438 0.0625  
## 2      2  0.417 0.0714  
## 3      3  0.423 0.308  
## 4      4  0.241 0.294  
## 5      5  0.419 0.484  
## 6      6  0.583 0.4  
## 7      7  0.537 0.367  
## 8      8  0.443 0.263  
## 9      9  0.395 0.311  
## 10     10  0.425 0.387  
## 11     NA  0.667 0.571
```

# Introducing `spread()` & `gather()`

- Data in `R` is either "long" or "wide"
- **Long**: One column for a categorical label and multiple rows
  - I.e., For each age group, we have one **row** for men and one **row** for women
- **Wide**: Multiple columns for each categorical label and a single row
  - I.e., For each age group, we have one **column** for men and one **column** for women
- In `R`, we can switch between **wide** and **long** with two functions:
  1. `spread()`: converts from long to wide
  2. `gather()`: converts from wide to long

# spread([key], [value])

- **key**: categories that you want to make into columns
- **value**: values you want to put into these new columns

wide

id	x	y	z
1	a	c	e
2	b	d	f

# spread()

```
MI_final_small %>%  
  group_by(SEX, AGE10) %>%  
  summarize(PctTrump = mean(TrumpVoter), .groups = 'drop') %>%  
  spread(key = SEX, value = PctTrump, fill = NA) %>%  
  rename(Male = `1`, Female = `2`)
```

```
## # A tibble: 11 × 3  
##   AGE10  Male Female  
##   <dbl> <dbl> <dbl>  
## 1      1 0.438 0.0625  
## 2      2 0.417 0.0714  
## 3      3 0.423 0.308  
## 4      4 0.241 0.294  
## 5      5 0.419 0.484  
## 6      6 0.583 0.4  
## 7      7 0.537 0.367  
## 8      8 0.443 0.263  
## 9      9 0.395 0.311  
## 10     10 0.425 0.387  
## 11     NA 0.667 0.571
```



# gather()

```
MI_final_small %>%  
  group_by(SEX, AGE10) %>%  
  summarize(PctTrump = mean(TrumpVoter), .groups = 'drop') %>%  
  spread(key = SEX, value = PctTrump, fill = NA) %>%  
  rename(Male = `1`, Female = `2`) %>%  
  gather(SEX, PctTrump, -AGE10)
```

```
## # A tibble: 22 × 3  
##   AGE10 SEX    PctTrump  
##   <dbl> <chr>    <dbl>  
## 1     1  1 Male    0.438  
## 2     2  2 Male    0.417  
## 3     3  3 Male    0.423  
## 4     4  4 Male    0.241  
## 5     5  5 Male    0.419  
## 6     6  6 Male    0.583  
## 7     7  7 Male    0.537  
## 8     8  8 Male    0.443  
## 9     9  9 Male    0.395  
## 10    10 10 Male    0.425  
## # ... with 12 more rows
```

# Save Summary for Later Use

```
SexAge <- MI_final_small %>%  
  group_by(SEX, AGE10) %>%  
  summarize(PctTrump = mean(TrumpVoter), .groups = 'drop')  
  
SexAge %>% filter(SEX == 2)
```

```
## # A tibble: 11 × 3  
##       SEX AGE10 PctTrump  
##   <dbl> <dbl>   <dbl>  
## 1     2     1    0.0625  
## 2     2     2    0.0714  
## 3     2     3    0.308  
## 4     2     4    0.294  
## 5     2     5    0.484  
## 6     2     6    0.4  
## 7     2     7    0.367  
## 8     2     8    0.263  
## 9     2     9    0.311  
## 10    2    10    0.387  
## 11    2    NA    0.571
```

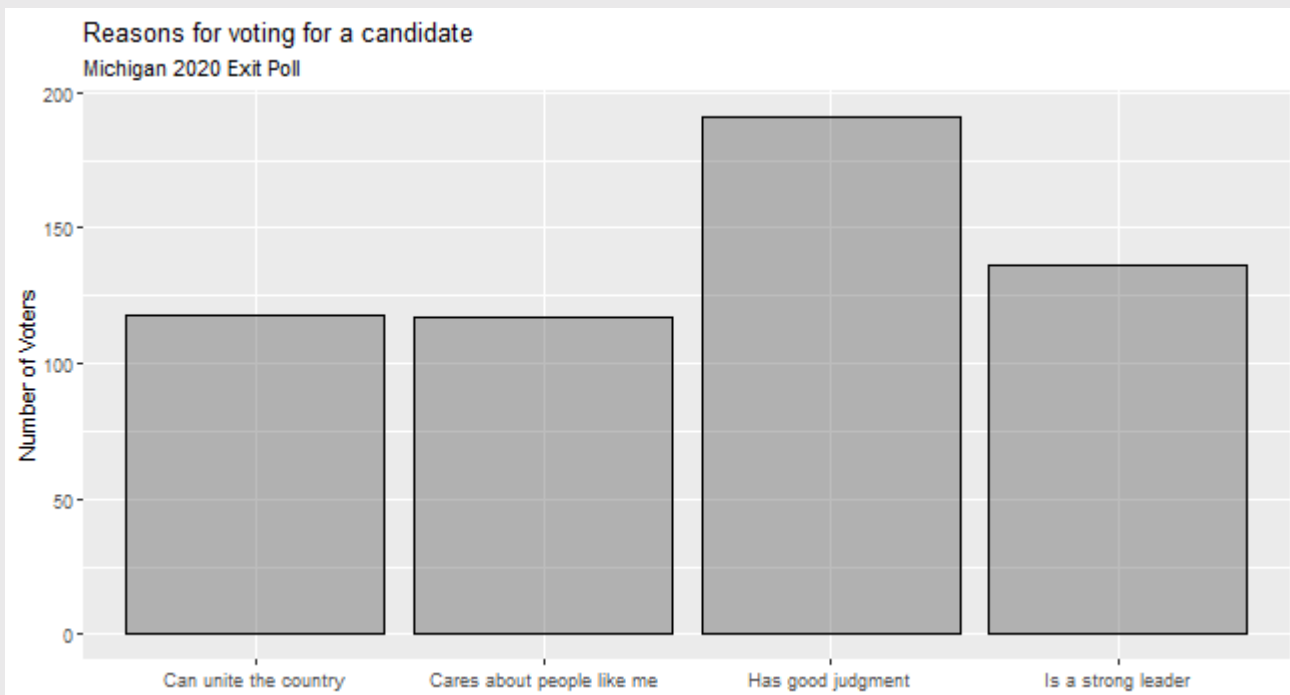
# Conditional Categorical Analysis

- Want to know **reason** for voting for candidate by **vote choice**
  - **Quality**: 4 category unordered
  - **preschoice**: 2 category unordered
- Some light data wrangling

```
toplot <- mi_ep %>%  
  select(Quality,preschoice,SEX) %>%  
  filter(grepl('Biden|Trump',preschoice)) %>%  
  drop_na() %>%  
  filter(Quality != "[DON'T READ] Don't know/refused")
```

# Conditional Categorical Analysis

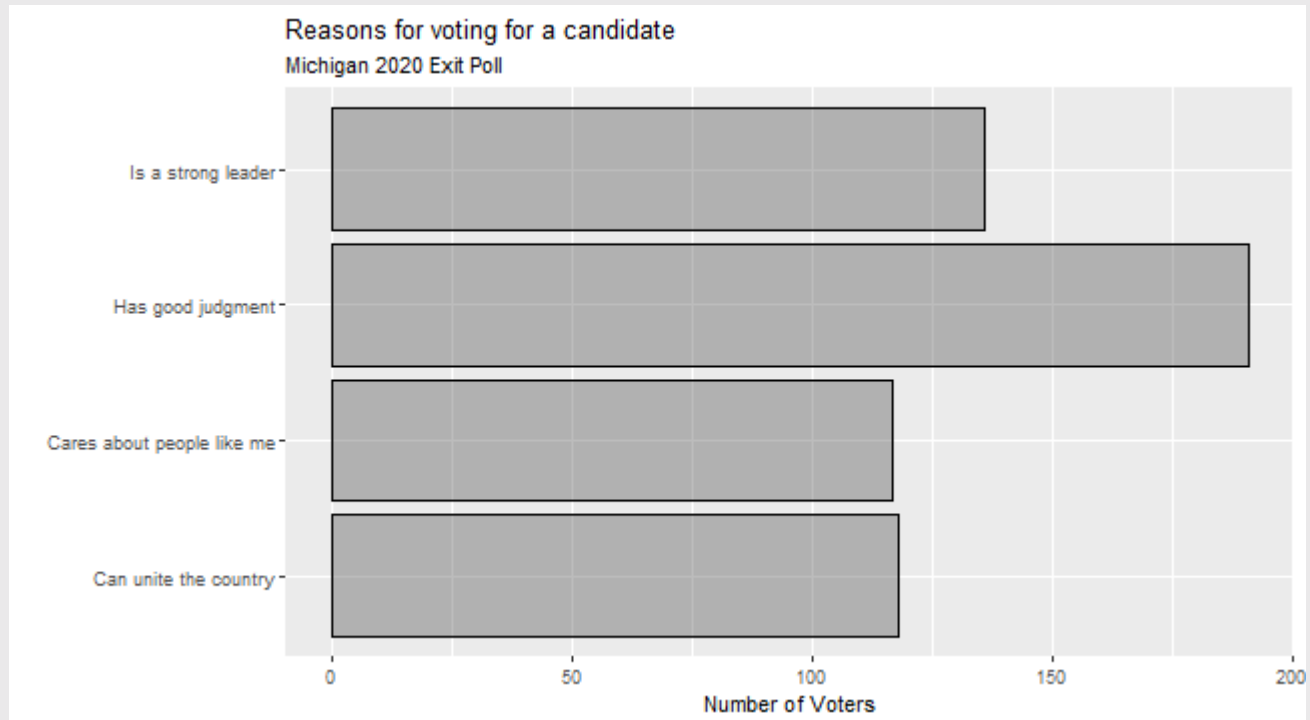
```
(pReasonOverall <- toplot %>%  
  ggplot(aes(x = Quality)) +  
  labs(y = "Number of Voters", x = "",  
        title = "Reasons for voting for a candidate",  
        subtitle = "Michigan 2020 Exit Poll") +  
  geom_bar(color="black", alpha = .4))
```



# Conditional Categorical Analysis

- Can swap axes with `coord_flip()`

```
pReasonOverall + coord_flip()
```



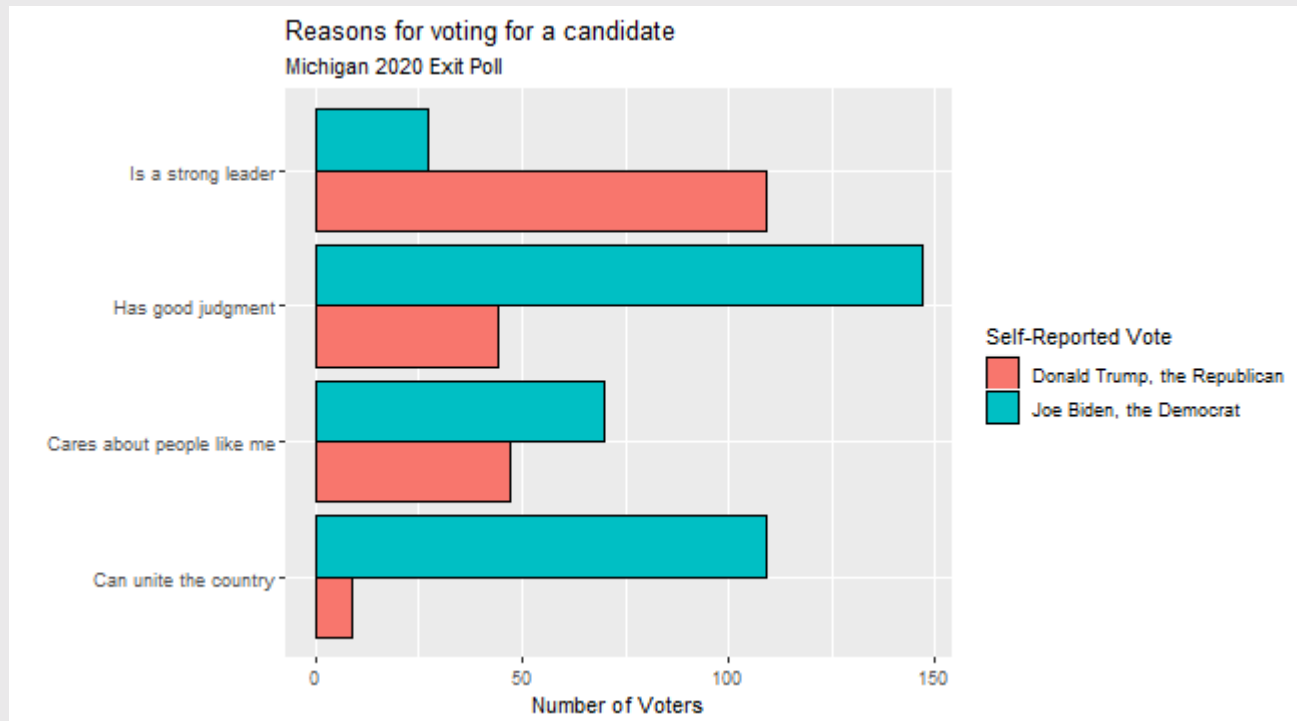
# Conditional Categorical Analysis

- `fill` and `position = "dodge"` for **conditional** analysis

```
pReasonChoice <- toplot %>%  
  ggplot(aes(x = Quality, fill = preschoice)) +  
  labs(y = "Number of Voters", x = "",  
        title = "Reasons for voting for a candidate",  
        subtitle = "Michigan 2020 Exit Poll",  
        fill = 'Self-Reported Vote') +  
  geom_bar(color="black", position = "dodge") +  
  coord_flip()
```

# Conditional Categorical Analysis

pReasonChoice



# Conditional Categorical Analysis

- What about if we do this by **SEX**?

```
toplot %>%  
  ggplot(aes(x= preschoice, fill = SEX)) +  
  labs(y = "Number of Respondents", x = "",  
       title = "Vote by Respondent Sex", fill = "Sex") +  
  geom_bar(position="dodge") + coord_flip()
```

```
## Warning: The following aesthetics were dropped during statistical  
## transformation: fill  
## i This can happen when ggplot fails to infer the correct  
##   grouping structure in the data.  
## i Did you forget to specify a `group` aesthetic or to  
##   convert a numerical variable into a factor?
```





# Be Attentive to `class()`

- How is `SEX` stored in the data?

```
class(mi_ep$SEX)
```

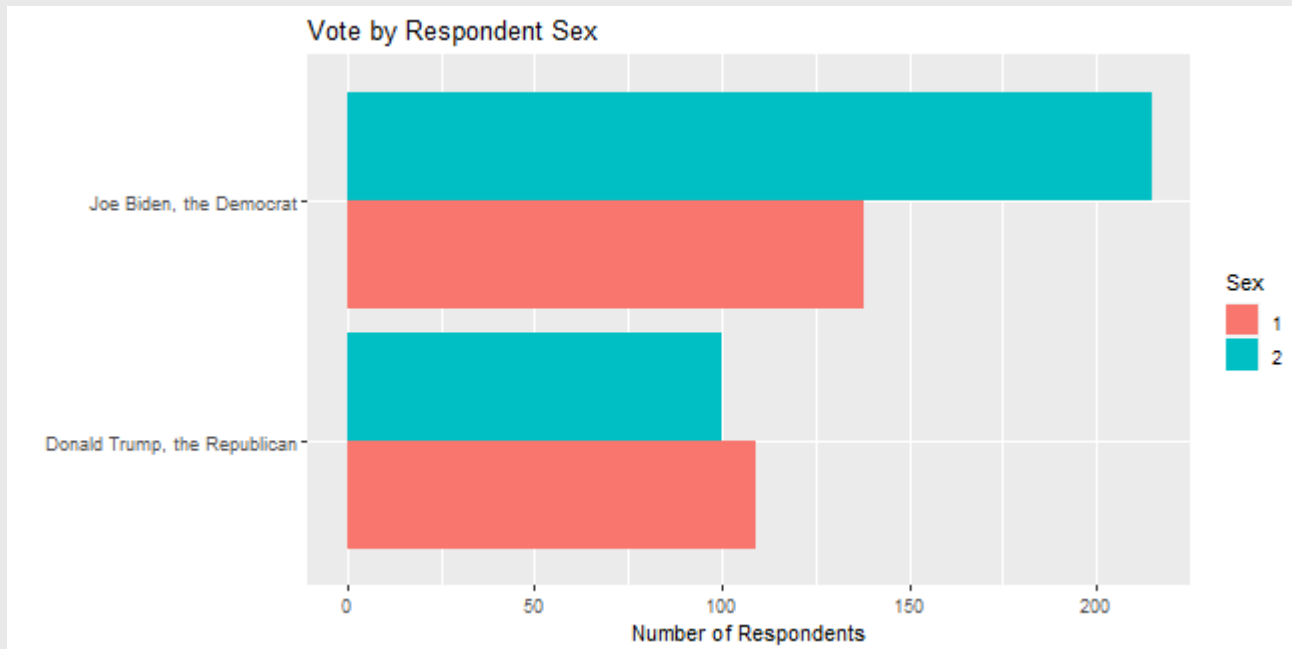
```
## [1] "numeric"
```

- Need to convert it to a `character` or `factor`

```
pVoteSex <- topplot %>%  
  ggplot(aes(x= preschoice, fill = factor(SEX))) +  
  labs(y = "Number of Respondents", x = "",  
       title = "Vote by Respondent Sex", fill = "Sex") +  
  geom_bar(position="dodge") + coord_flip()
```

# Be Attentive to `class()`

pVoteSex



- Why is this a bad visualization? **Poorly labeled legend!**

# Continuous by Categorical

- Let's introduce a different dataset!
  - Download and open [Pres2020\\_PV.Rds](#)
  - Wrangle to get the popular vote margin, expressed in decimals

```
poll <- read_rds('../data/Pres2020_PV.Rds')

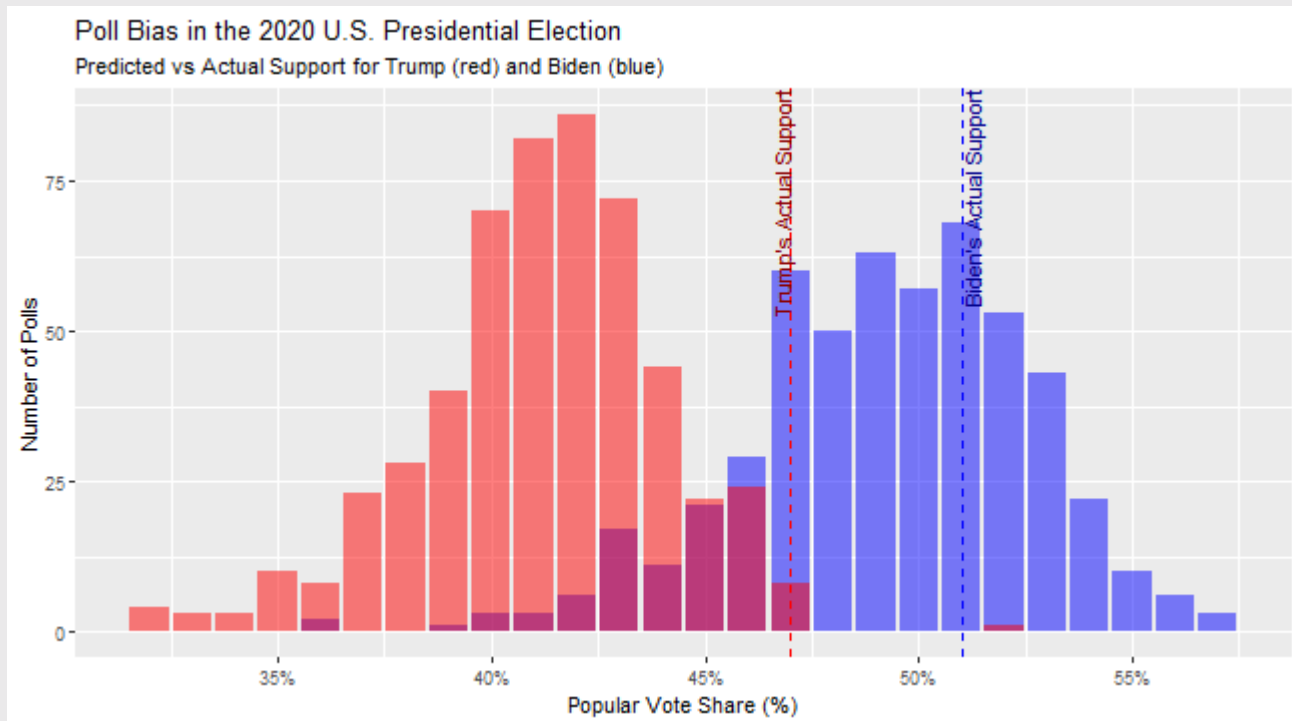
poll <- poll %>%
  mutate(Trump = Trump/100,
         Biden = Biden/100,
         margin = Biden - Trump)
```

# 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

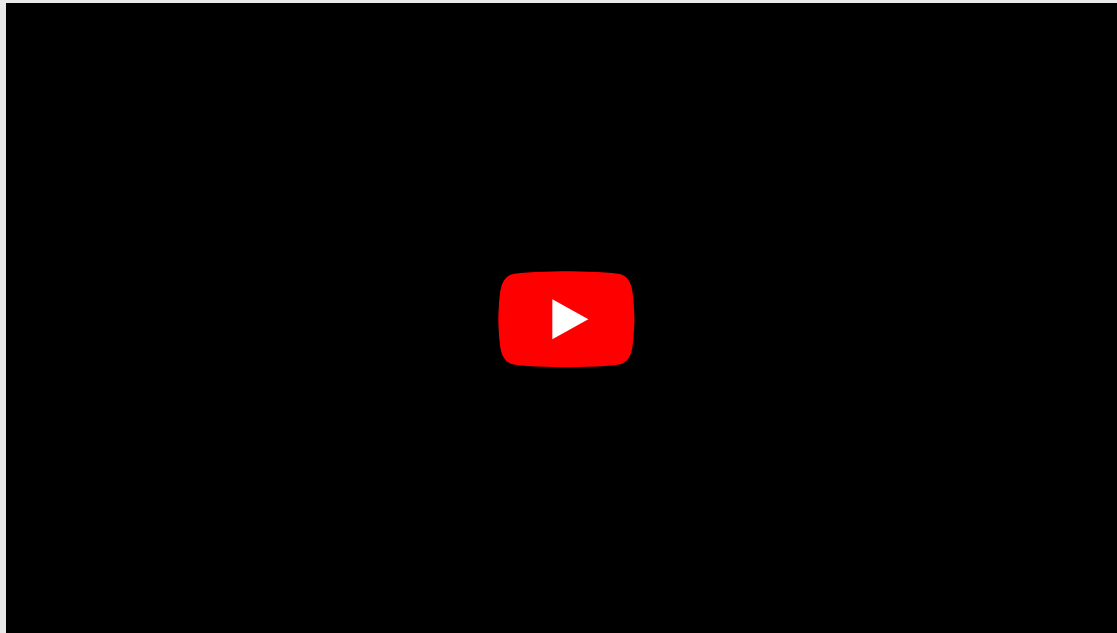
- A fifth explanation?
- Anti-Trump media!





# 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!

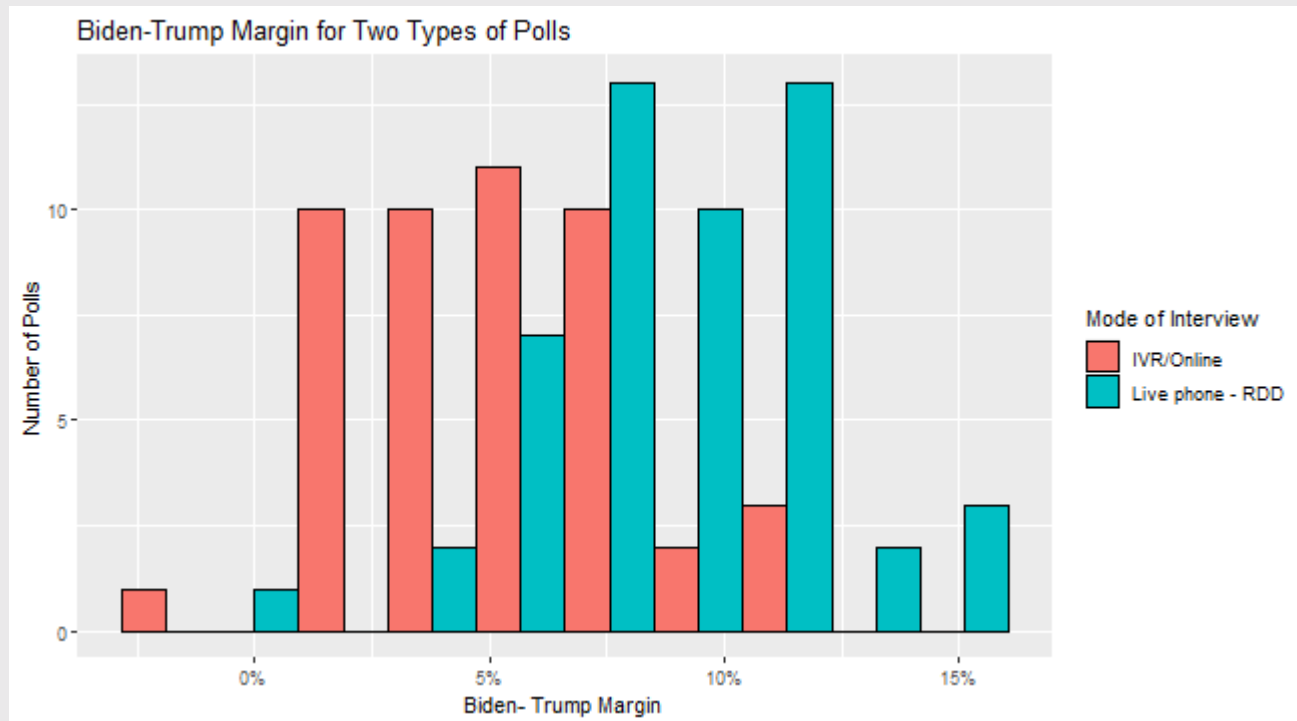
# Analyzing

- For now, just focus on **IVR/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")

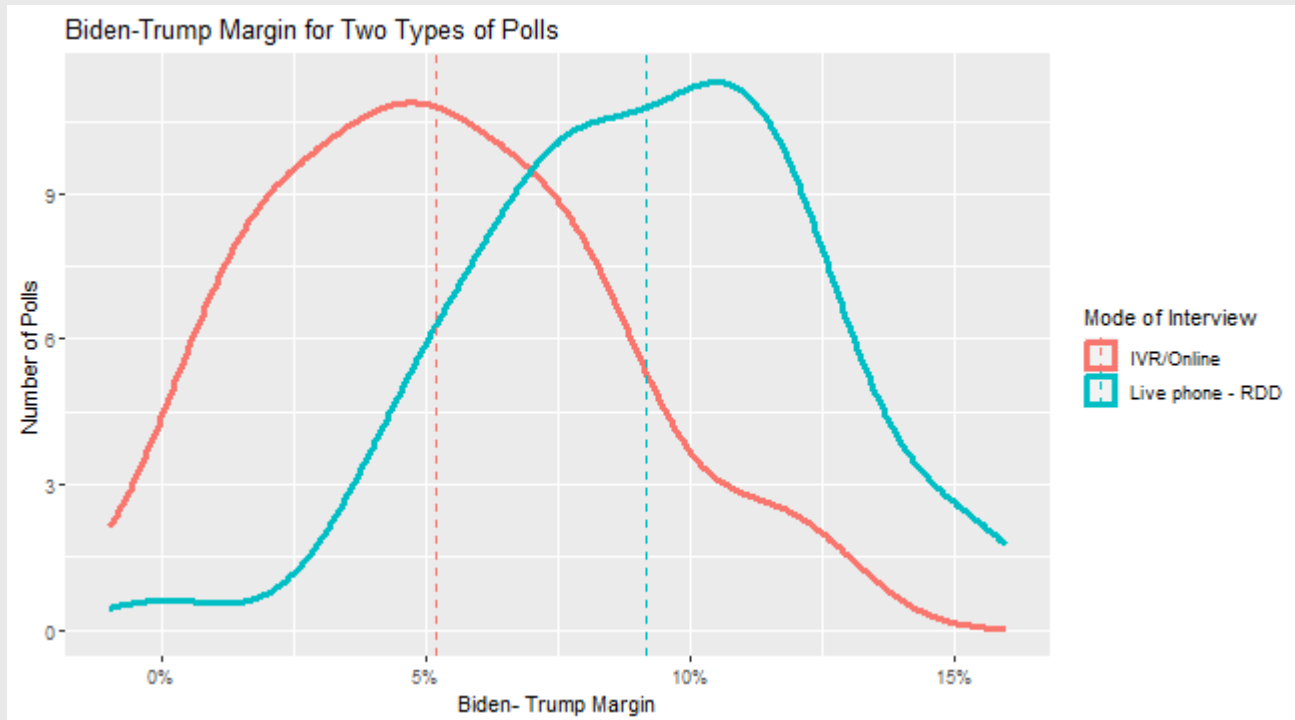
pModeDens <- toplot %>%
  ggplot(aes(x= margin, color = Mode)) +
  labs(y = "Number of Polls",
       x = "Biden- Trump Margin",
       title = "Biden-Trump Margin for Two Types of Polls",
       color = "Mode of Interview") +
  geom_density(lwd = 1.2) +
  scale_x_continuous(breaks=seq(-.1,.2,by=.05),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

```
## Warning: Using `size` aesthetic for lines was deprecated in
ggplot2
## 3.4.0.
```

# 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

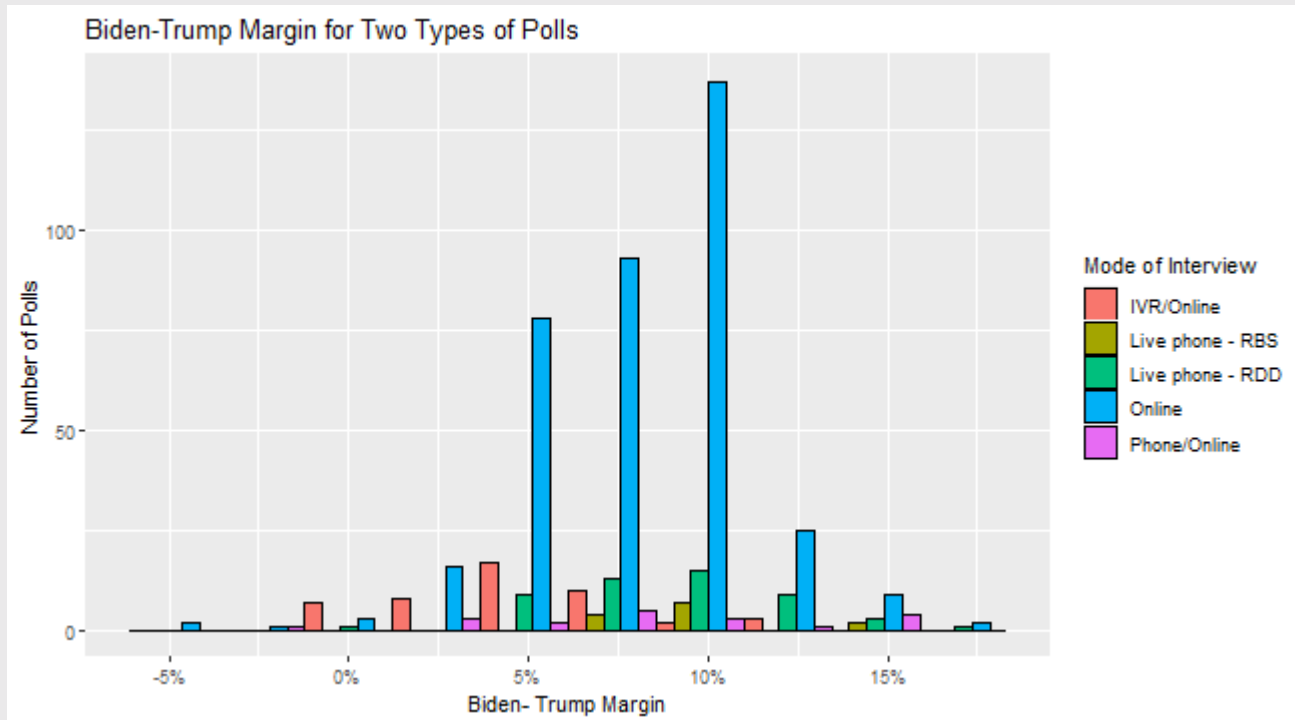
- How hard is `geom_histogram()` with more categories?

```
pModeHist <- topplot %>%  
  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

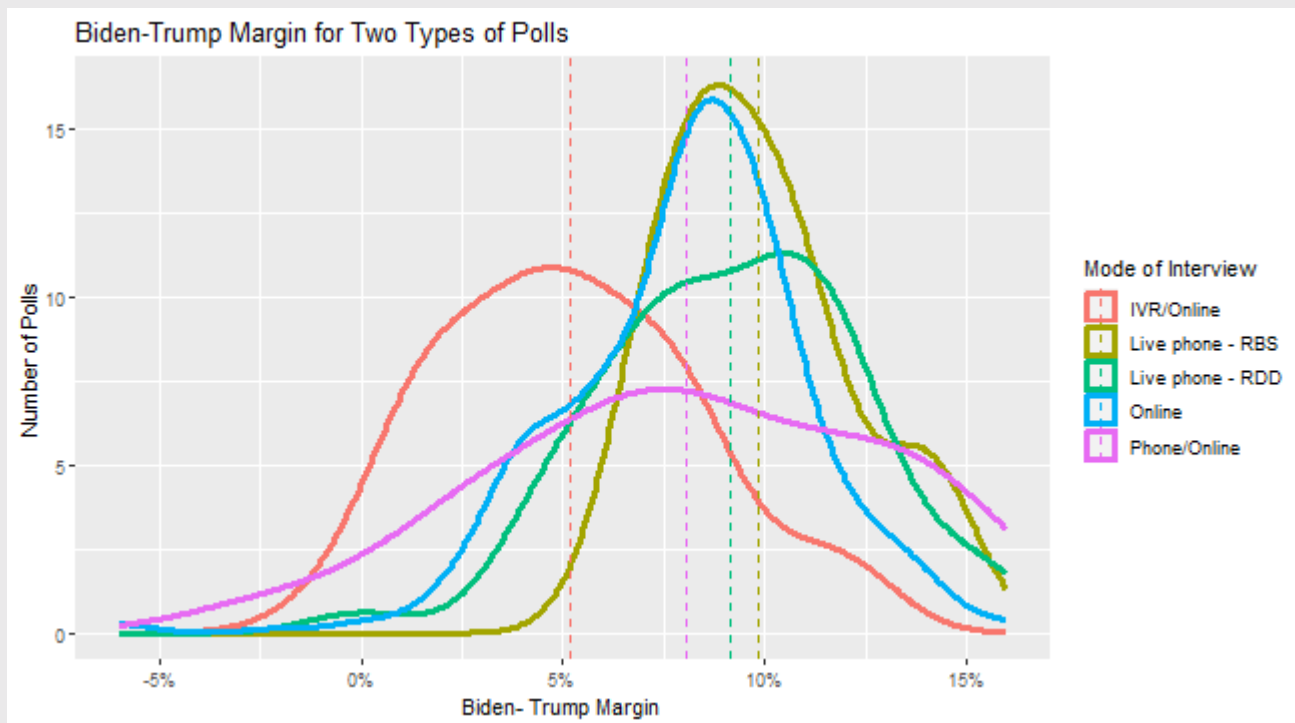
- How hard is `geom_density()` with more categories?

```
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()

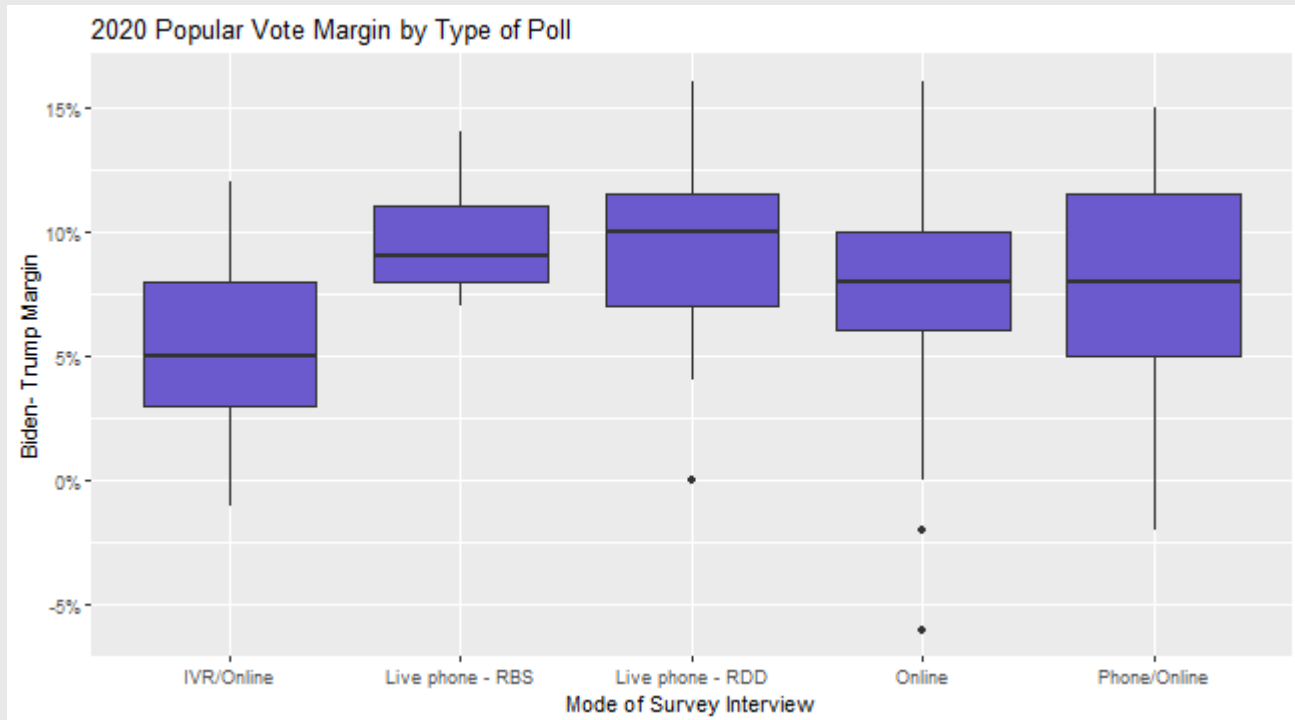
- More categories requires more compact ways of visualizing distributions

```
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

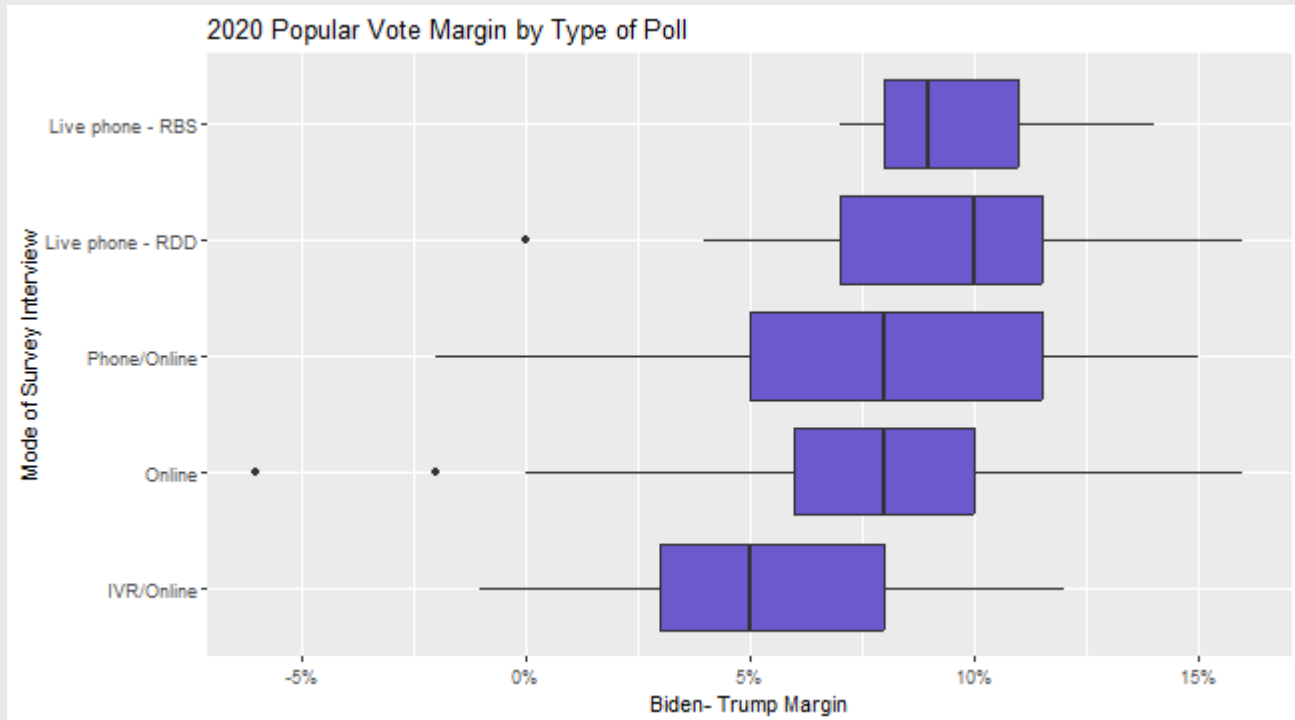
- We can use `reorder()` to arrange categories by the data

```
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()
```





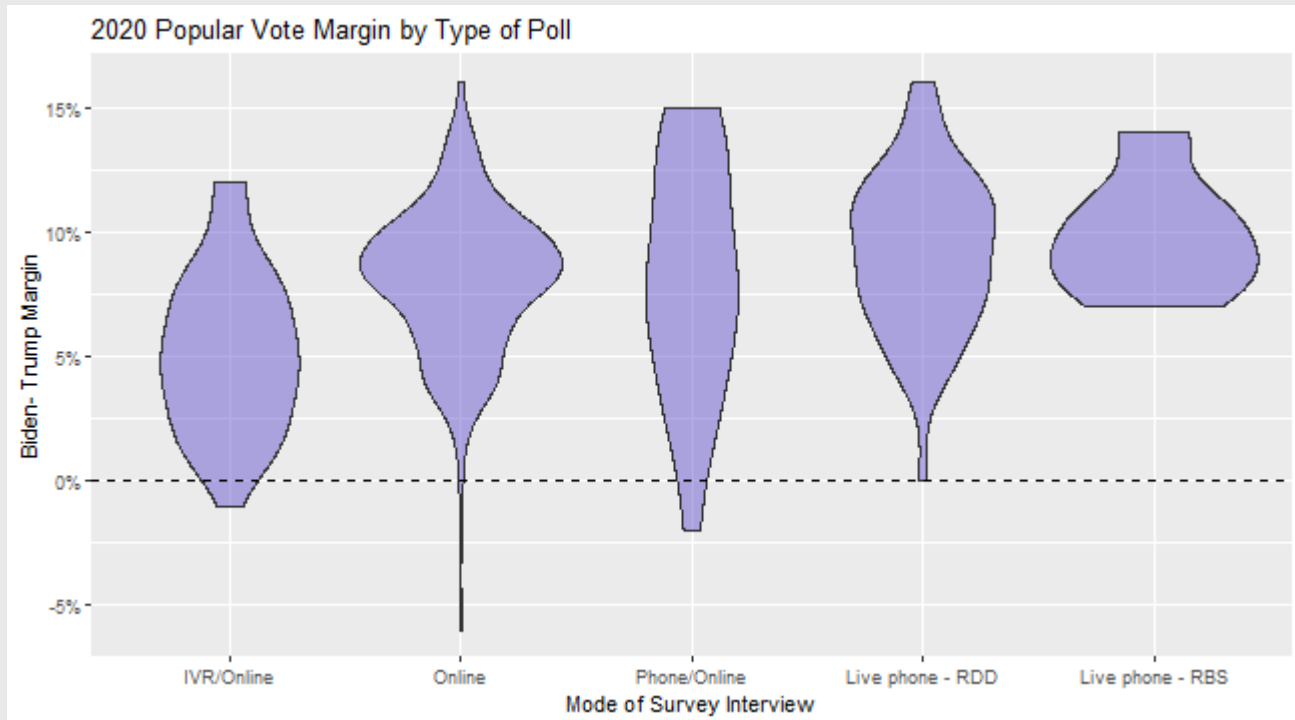
# geom\_violin()

- 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

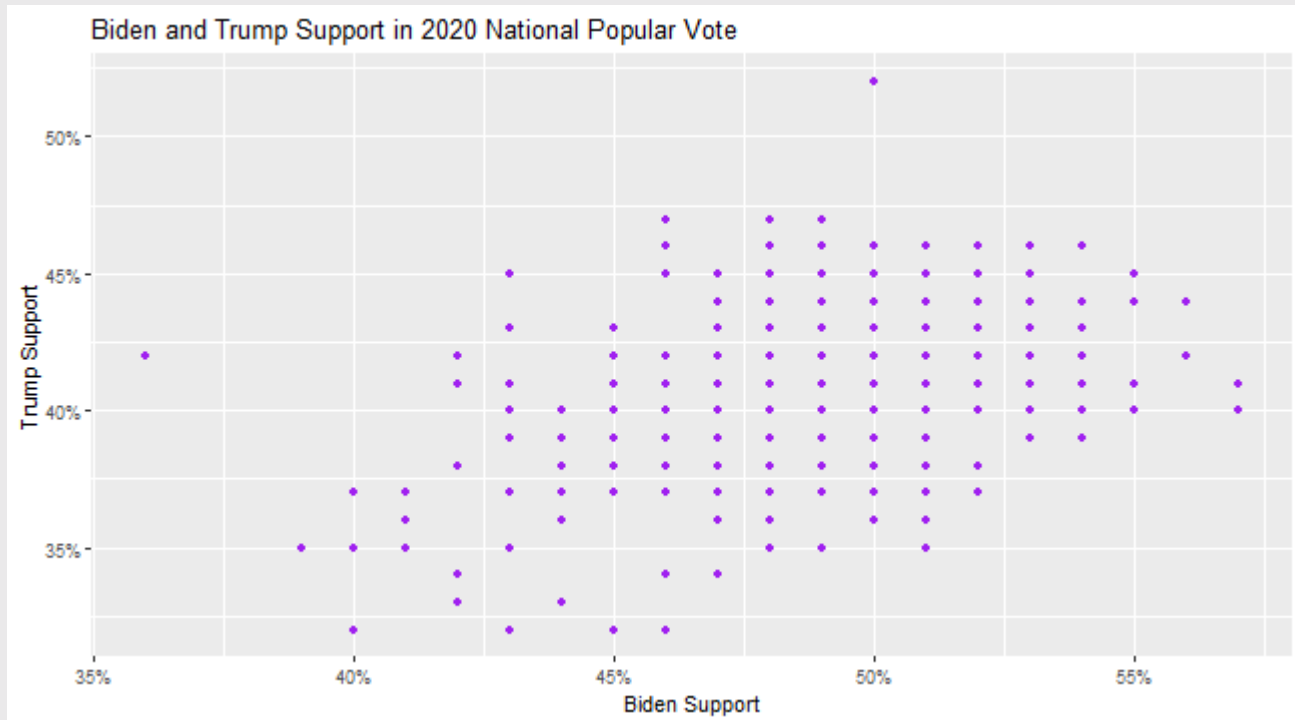
# Analysis

- 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?

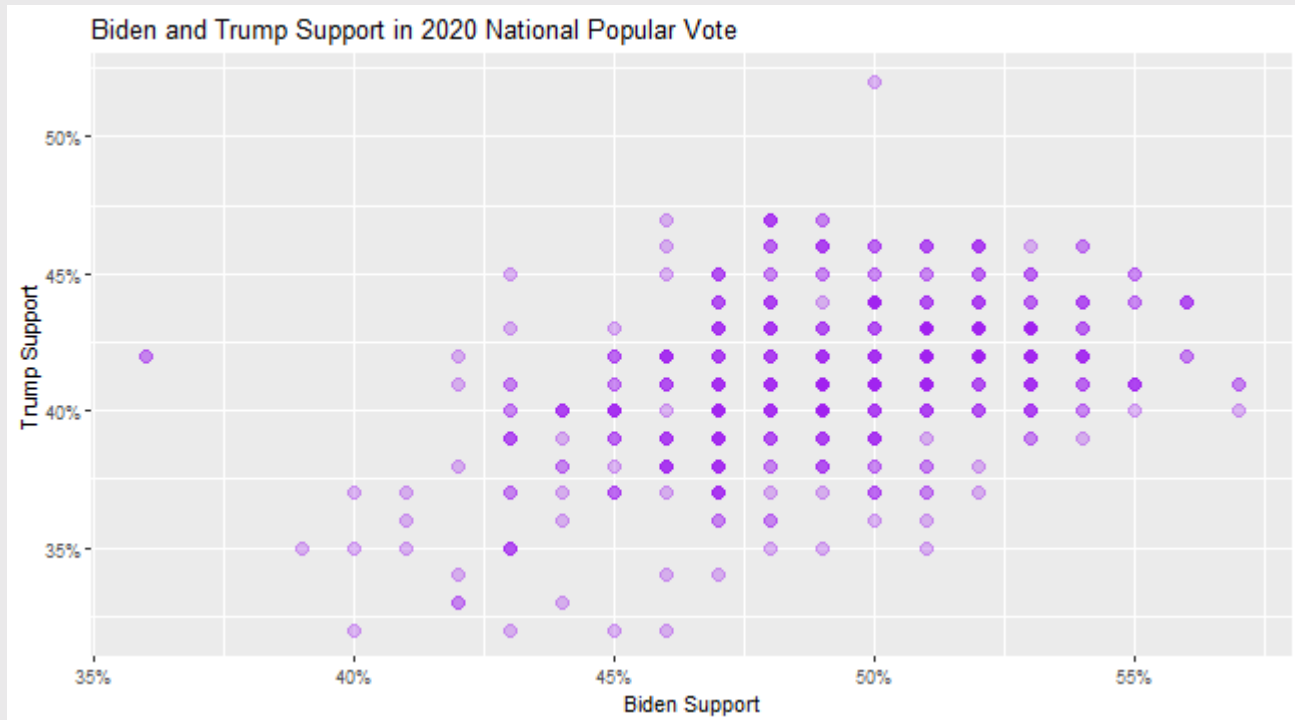
# Tweaking `alpha`

- 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 $\alpha$

pSupp



# geom\_jitter()

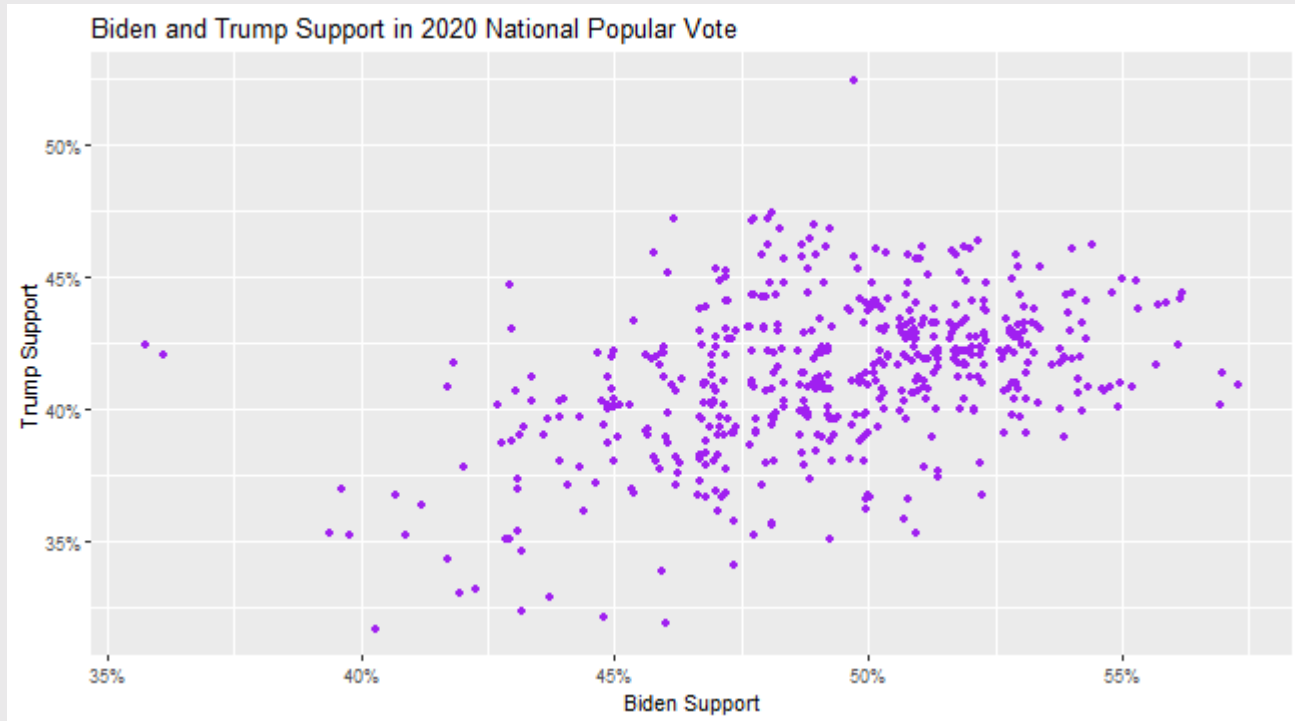
- Instead, we could "jitter" the points
  - This adds some random noise to each point to shake them off each other

```
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

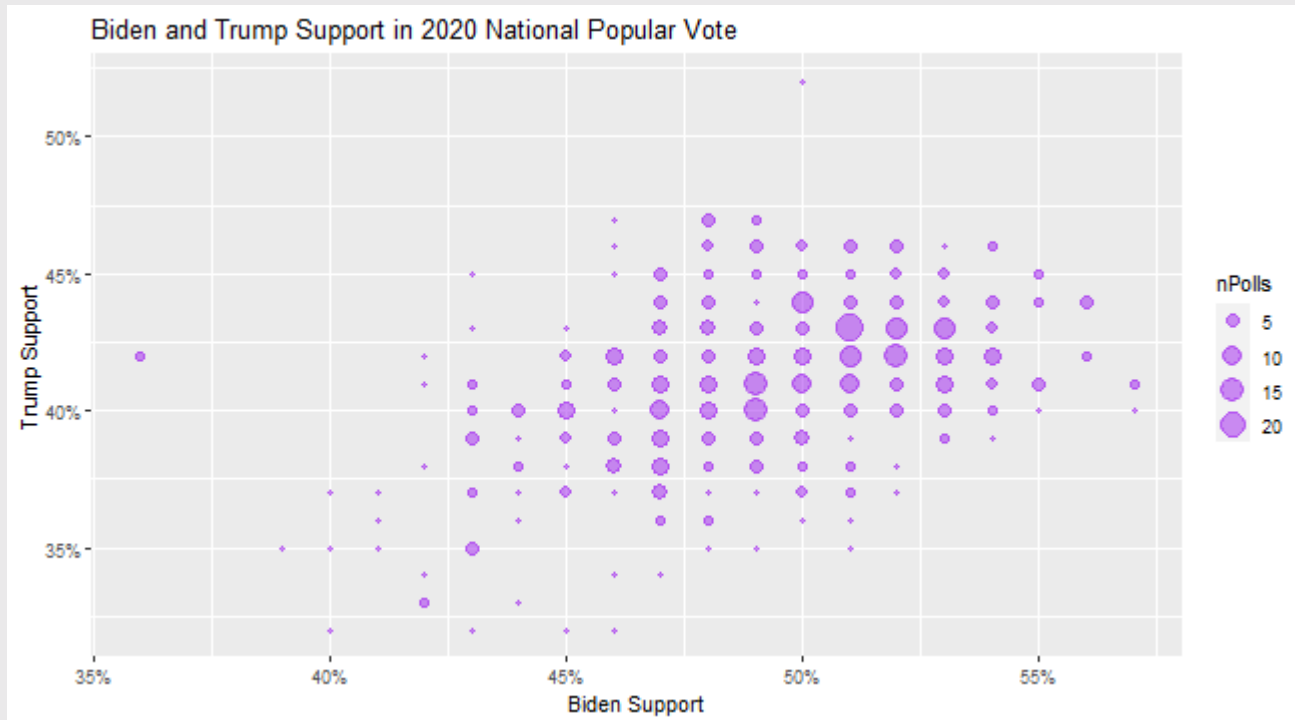
- Finally, we could simply count the number of polls at each x,y coordinate
  - Then size the points by the number of polls

```
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

# Quiz & Homework

- Go to Brightspace and take the **7th** quiz
  - The password to take the quiz is ####
- **Homework:**
  1. Work through Multivariate\_Analysis\_part1\_hw.Rmd
  2. Problem Set 3 (on Brightspace)