Multivariate Analysis

Part 1: Conditional Relationships

Prof. Bisbee

Vanderbilt University

Slides Updated: 2024-08-10

Agenda

1. Mutivariate
2. What is "conditional"?
3. Understanding Trump support

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('https://github.com/jbisbee1/DS1000_F2024/raw/main/data/MI2020
```

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 pctSupport
##   <chr>                       <dbl> <int>     <int>      <dbl>
## 1 Donald Trump, the Republ…     1   247       551      0.448
## 2 Donald Trump, the Republ…     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
```

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
## # ℹ 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`)
```

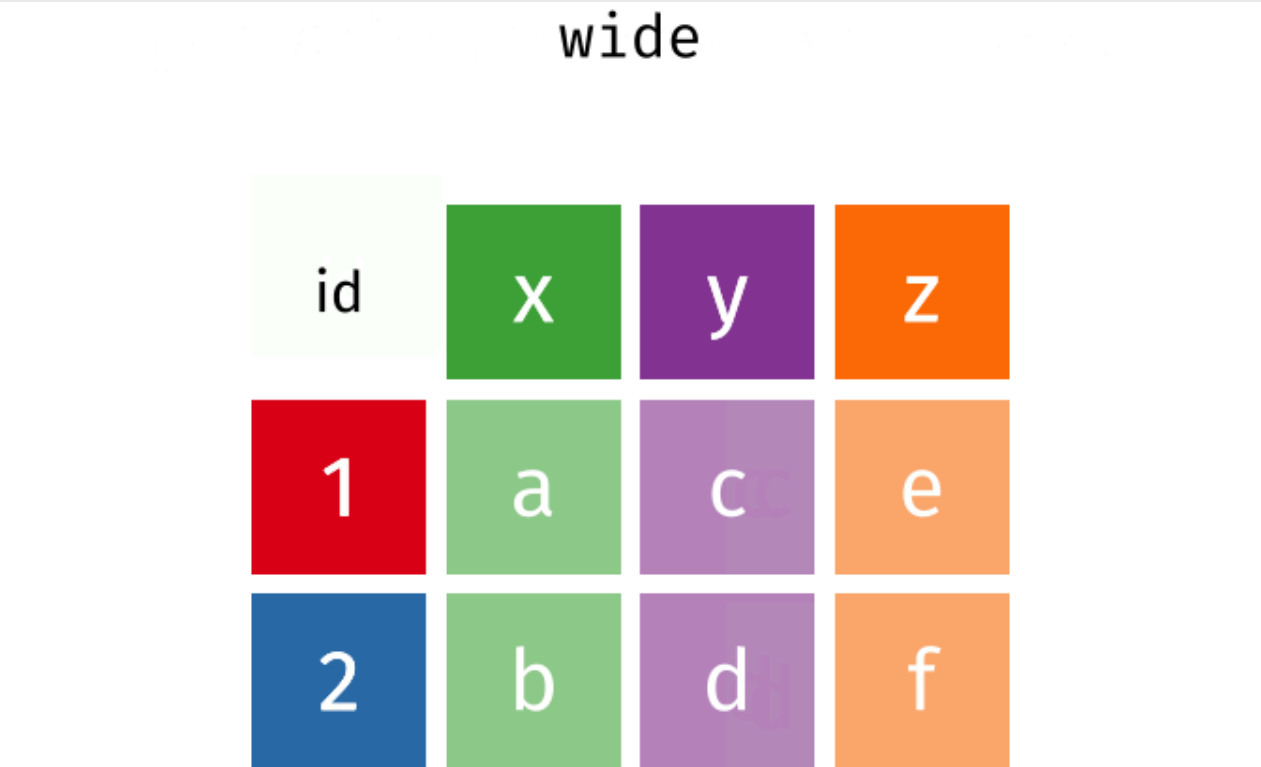
```
## # A tibble: 11 × 3
##    AGE10  Male Female
##    <dbl> <dbl>  <dbl>
##  1     1 0.438 0.0625
##  2     2 0.417 0.0714
##  3     3 0.423 0.308
##  4     4 0.241 0.294
##  5     5 0.419 0.484
##  6     6 0.583 0.4
##  7     7 0.537 0.367
##  8     8 0.443 0.263
##  9     9 0.395 0.311
## 10    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()` (or `pivot_wider()`): converts from long to wide
    2. `gather()` (or `pivot_longer()`): converts from wide to long

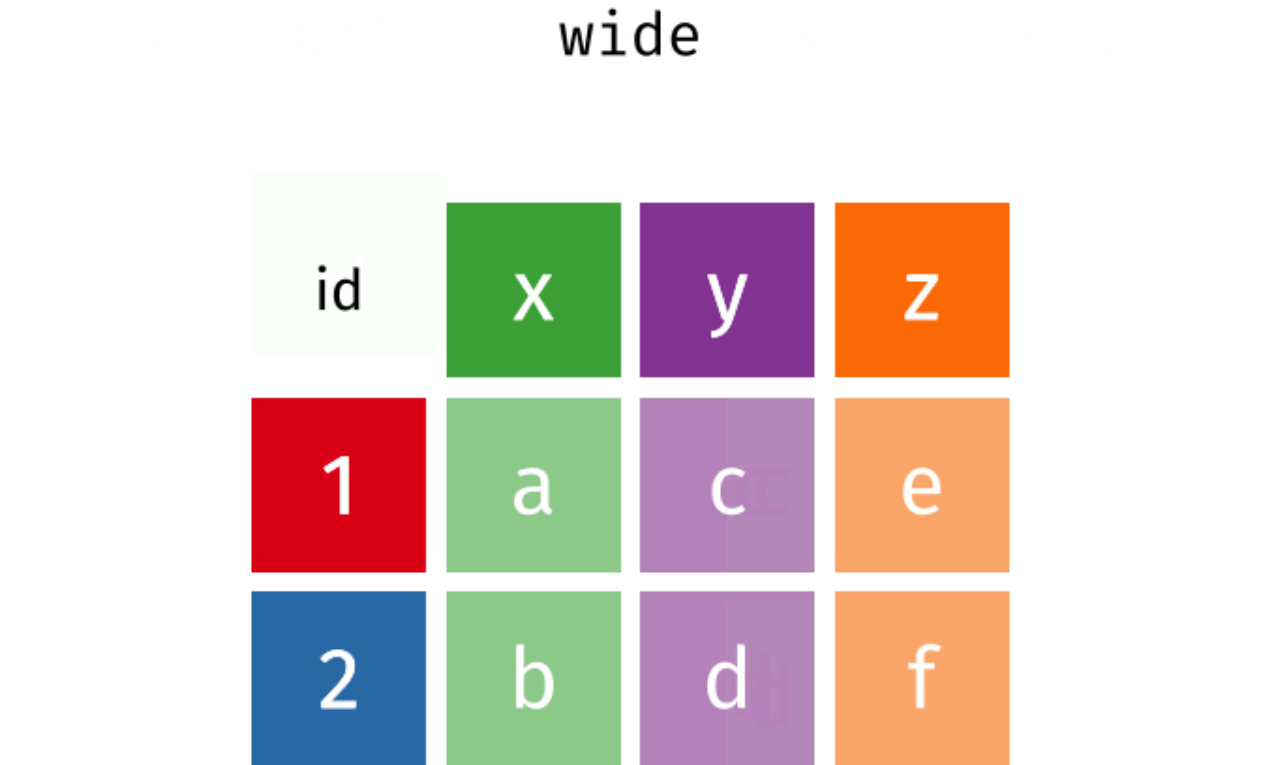
`spread()` and `gather()`

- `spread([key],[value])`
  - `key`: variable containing categories to make into columns labels
  - `value`: variable containing values put into these new columns



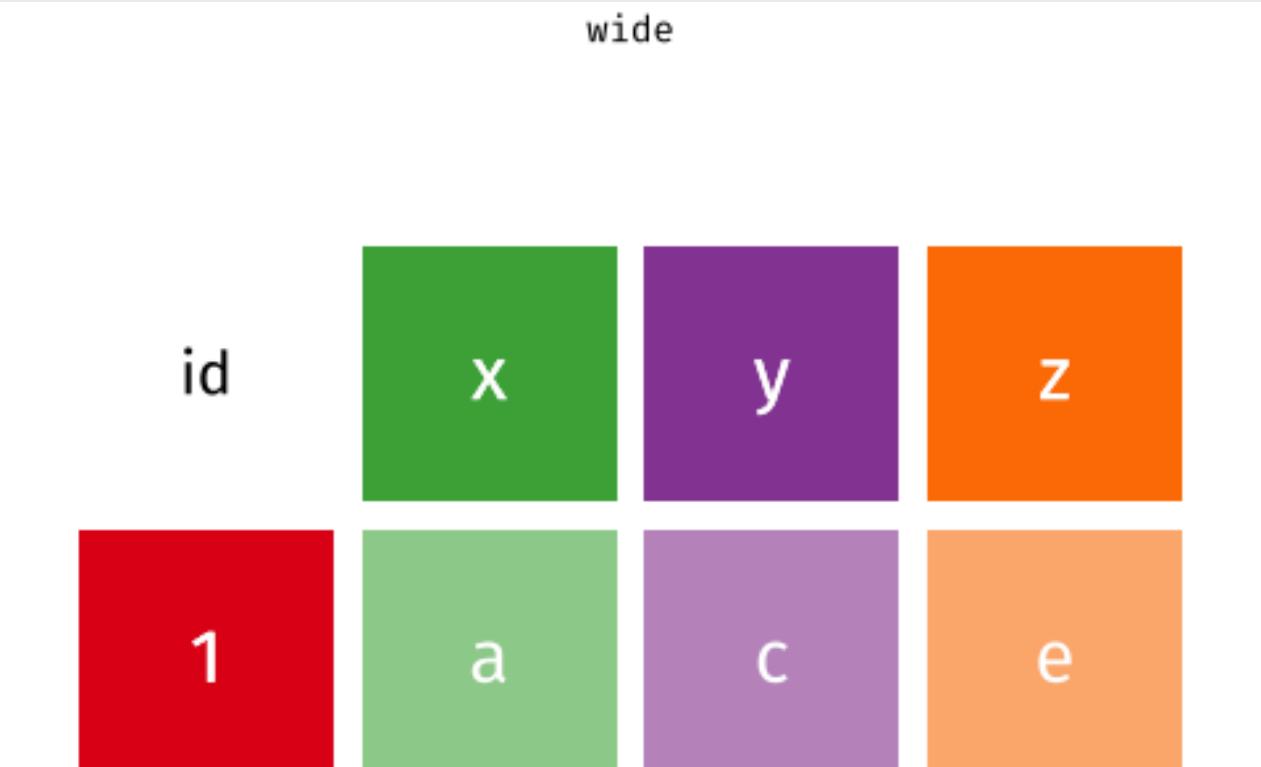
`spread()` and `gather()`

- `gather([key],[value],[columns])`
  - `key`: name of **new column** that contains categories
  - `value`: values you want to put into this new column



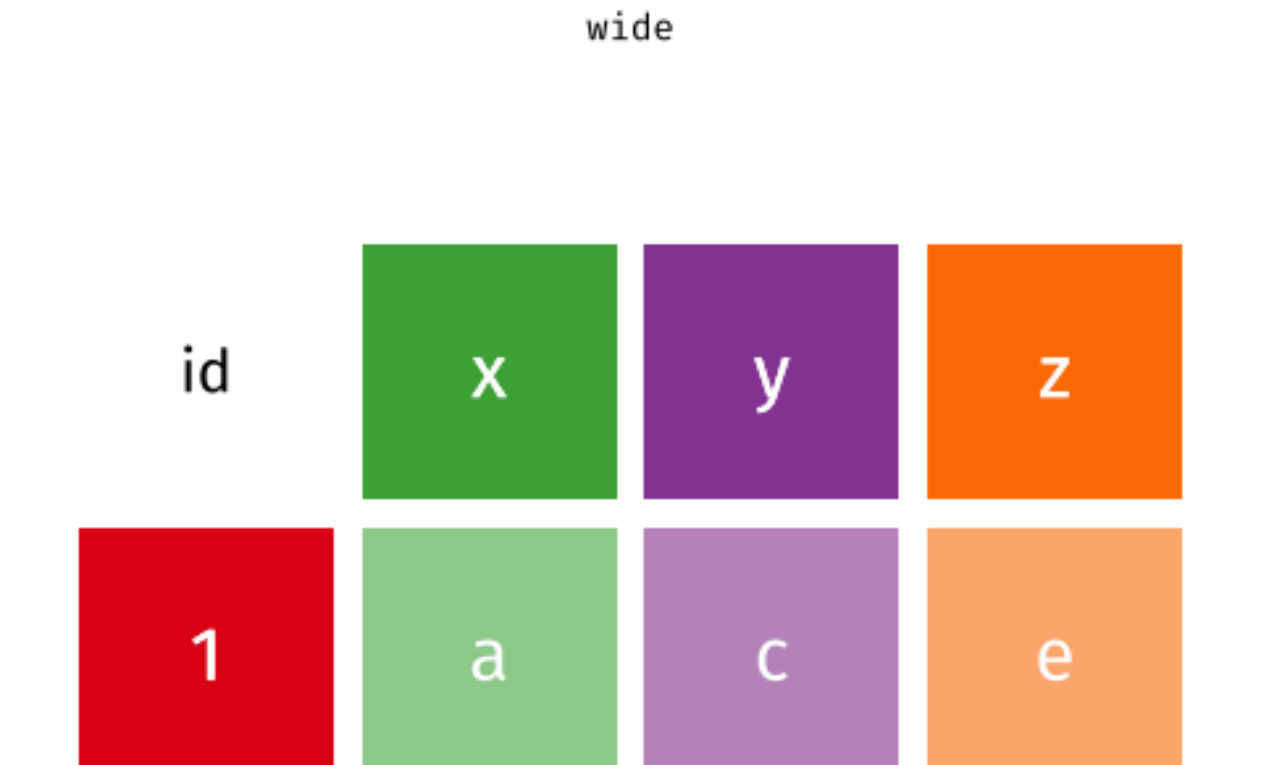
pivot_wider()

- `pivot_wider([names_from],[values_from])`
  - `names_from`: variable containing categories to make into column labels
  - `values_from`: variable containing values put into these new columns



OR `pivot_longer()`

- `pivot_longer([names_from],[values_from])`
  - `names_from`: variable containing categories to make into column labels
  - `values_from`: variable containing values put into these new columns



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 Male     0.438
##  2     2 Male     0.417
##  3     3 Male     0.423
##  4     4 Male     0.241
##  5     5 Male     0.419
##  6     6 Male     0.583
##  7     7 Male     0.537
##  8     8 Male     0.443
##  9     9 Male     0.395
## 10    10 Male     0.425
## # ℹ 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
```

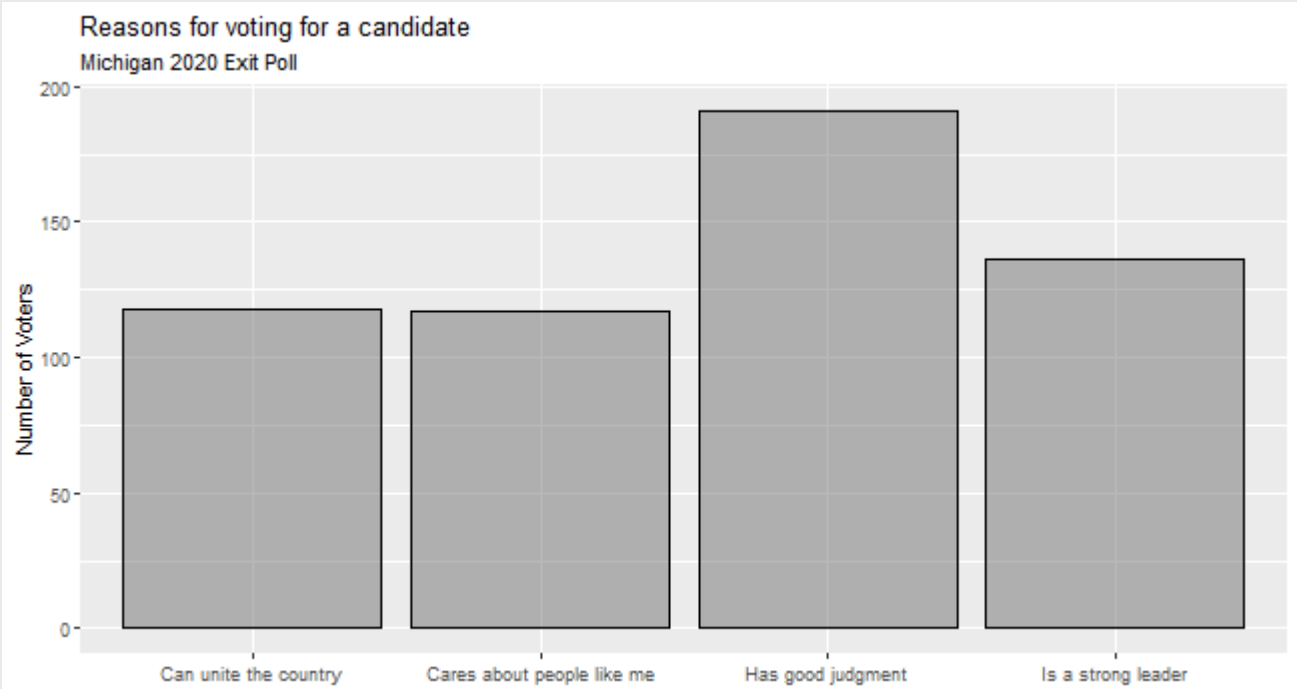
Conditional Categorical Analysis

- 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

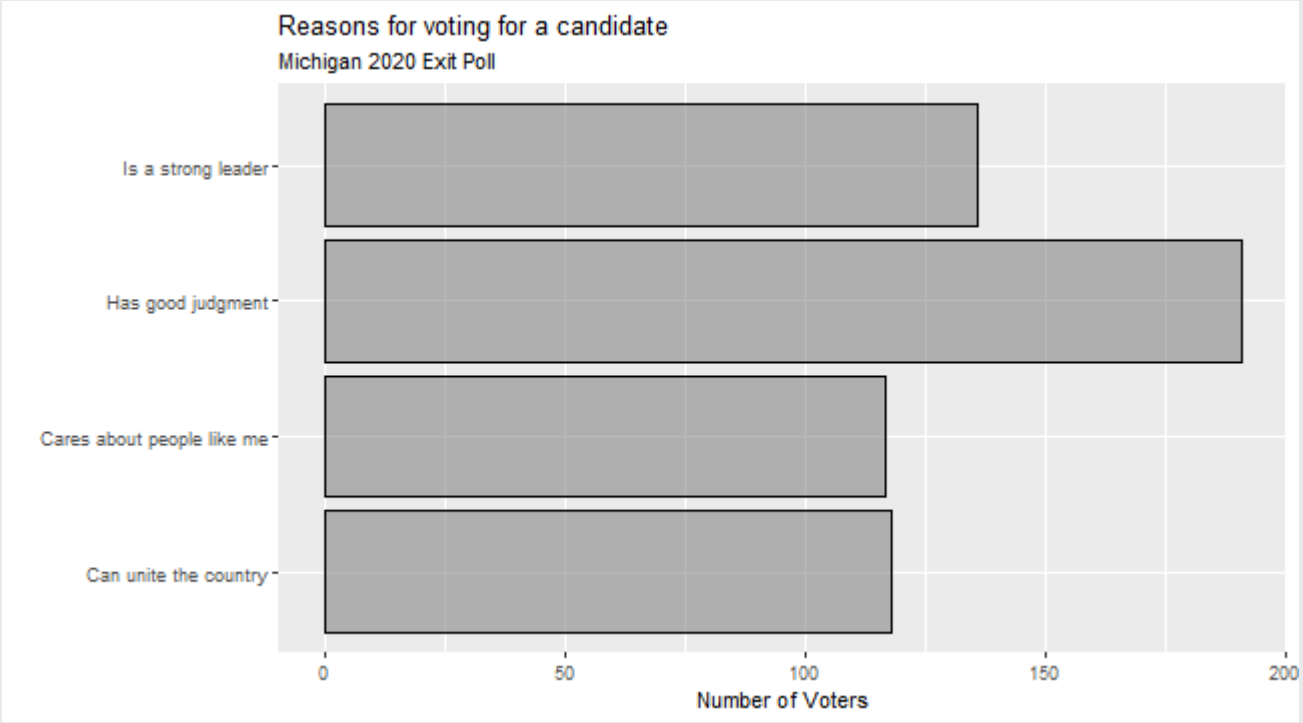
```
(pReasonOverall <- toplot %>%
  ggplot(aes(x = Quality)) +
  labs(y = "Number of Voters",x = "",
         title = "Reasons for voting for a candidate",
       subtitle = "Michigan 2020 Exit Poll") +
    geom_bar(color="black",alpha = .4))
```



Conditional Categorical Analysis

- Can swap axes with `coord_flip()`

```
pReasonOverall + coord_flip()
```



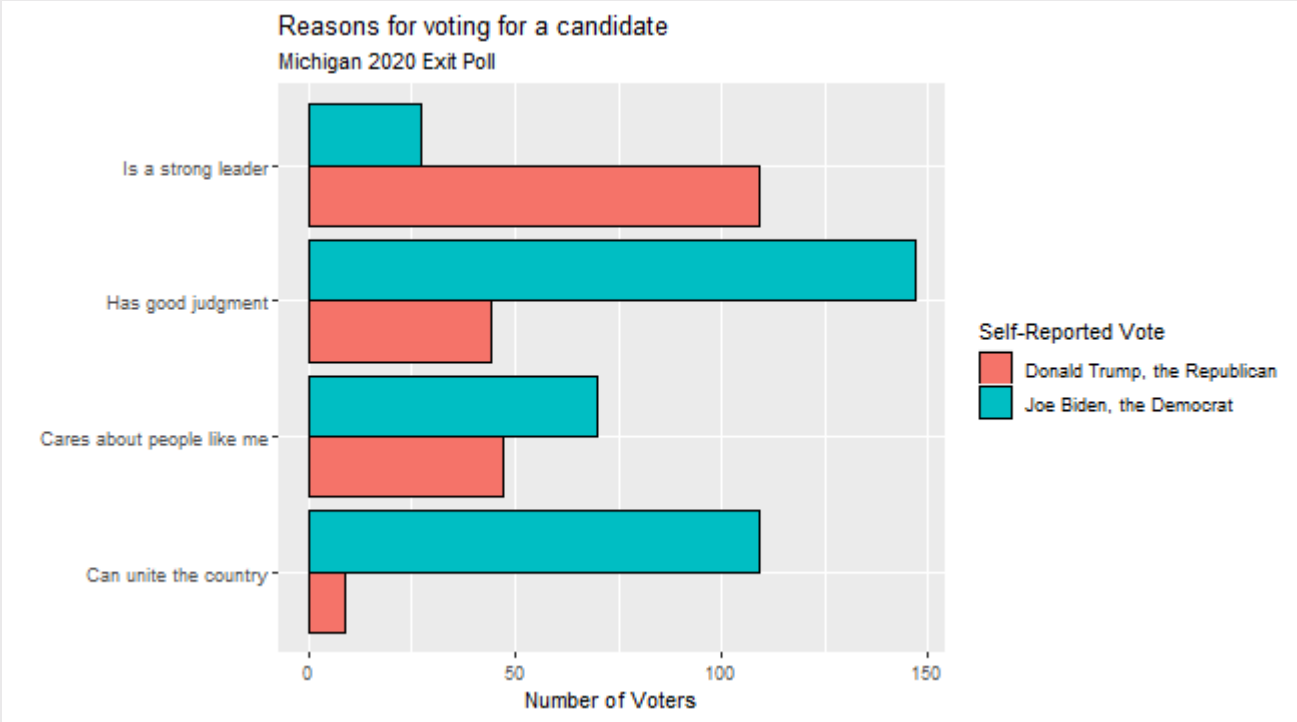
Conditional Categorical Analysis

- `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

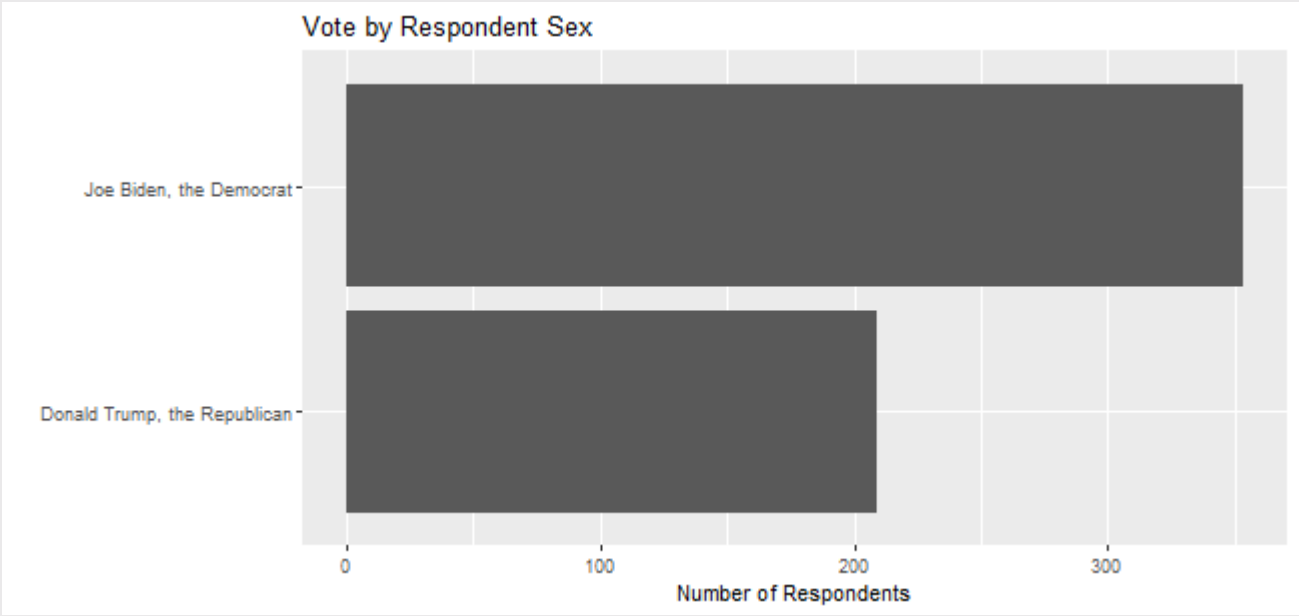
```
pReasonChoice
```



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()
```



Be Attentive to `class()`

- How is `SEX` stored in the data?

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

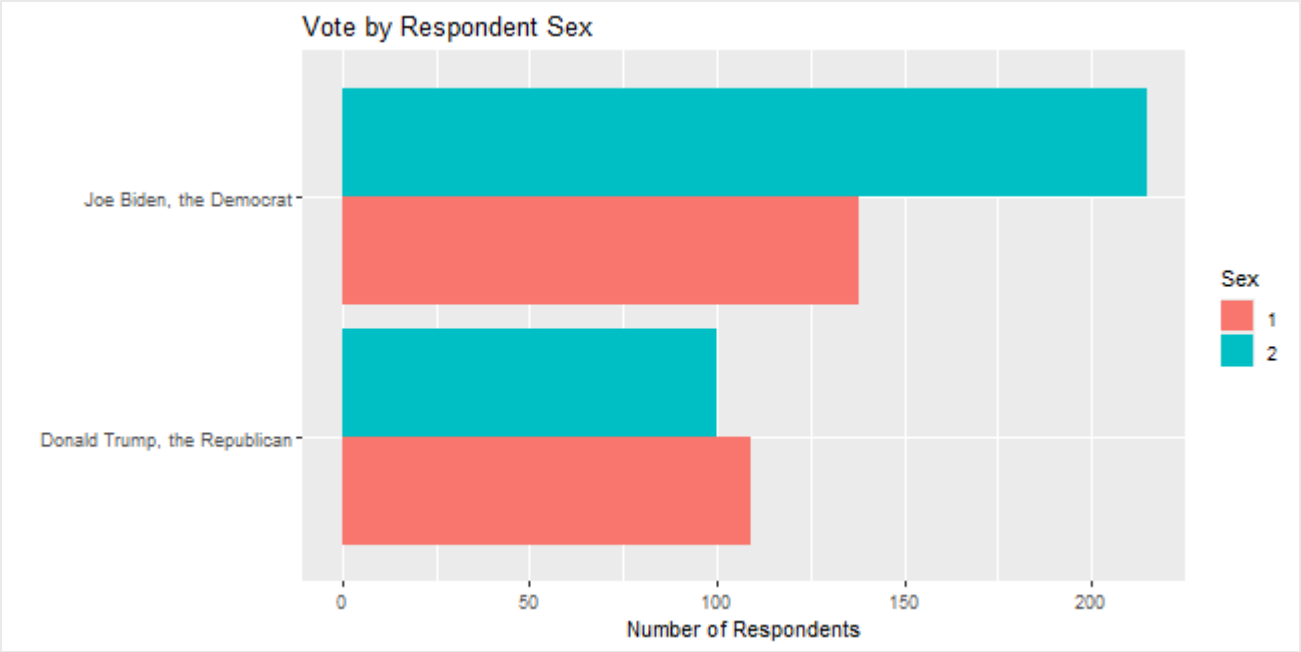
```
## [1] "numeric"
```

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

```
pVoteSex <- toplot %>%
    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!**

Quiz & Homework

- Go to Brightspace and take the **6th** quiz
  - The password to take the quiz is ####
- **Homework:**
  1. Work through ds1000_hw_7.Rmd
  2. Problem Set 4 (on Brightspace)