

Topic 3. Data Wrangling

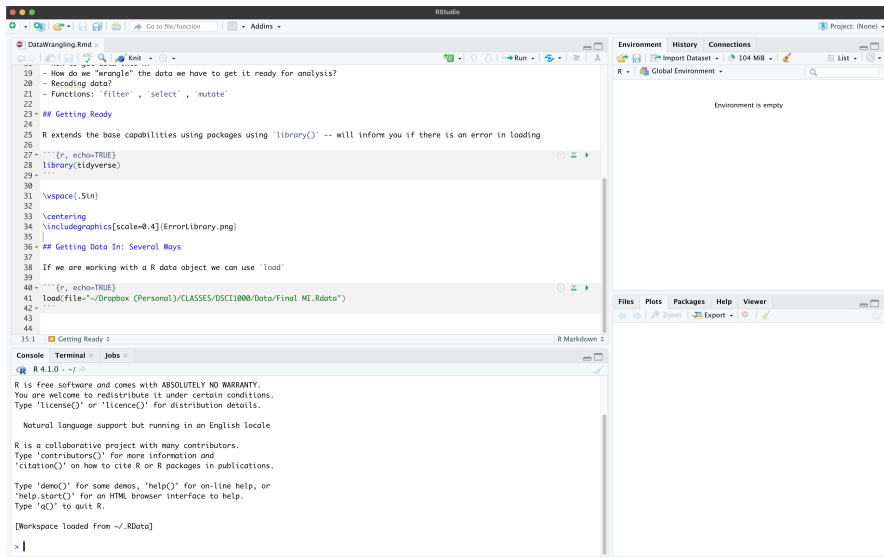
Josh Clinton

Objectives for this Topic

- ▶ Review: how to get data into R?
- ▶ How do we “wrangle” the data we have to get it ready for analysis?
- ▶ `filter` , `select` , `mutate` , `summarize`
- ▶ Applications: Johns Hopkins' Covid Data, 2020 MI Exit Poll #

Data Wrangling: The process of getting the data ready for analysis, including: accessing data, reformatting in appropriate ways (format, orientation), creating variables, recoding values, and selecting variables and/or observations of interest.

Starting Out:



The screenshot displays the RStudio environment. The main script editor shows a file named 'DataWrangling.Rmd' with the following content:

```
19 - How do we "arrange" the data we have to get it ready for analysis?
20 - Recoding data?
21 - Functions: 'filter', 'select', 'mutate'
22
23 ## Getting Ready
24
25 R extends the base capabilities using packages using 'library()' -- will inform you if there is an error in loading
26
27 ```{r, echo=TRUE}
28 library(tidyverse)
29 ...
30
31 \vspace{.5in}
32
33 \centering
34 \includegraphics[scale=0.4]{ErrorLibrary.png}
35
36 ## Getting Data In: Several Ways
37
38 If we are working with a R data object we can use 'load'
39
40 ```{r, echo=TRUE}
41 load(file=~/.Dropbox (Personal)/CLASSES/DSCI1000/Data/Final MI.Rdata")
42 ...
43
44
45 35.1 Getting Ready :
```

The console at the bottom shows the R startup message:

```
R 4.1.0 . ~/R
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.Rdata]
> |
```

The environment pane on the right shows 'Global Environment' and states 'Environment is empty'.

NOTE: Global Environment is empty!

Getting Data In: Basic Principles

- ▶ Goal is always replicability! Art depends on *personality*, science depends on *replication*.
- ▶ How you process and recode data is critical!
- ▶ You cannot (easily) fix in your analysis what you screw up in your data! (Especially if you are unaware!)

Reminder: Always know where you are!

Starting out it can be easy to forget that R needs to be told where to look to find things!

1. For every assignment/lecture create a new folder on your computer. Save your code and your data to this folder.
2. Figure out where R thinks it is using `getwd()`

```
getwd()
```

```
[1] "/Users/clintojd/GitHub/vandy_ds_1000_materials/Lectures/Lec
```

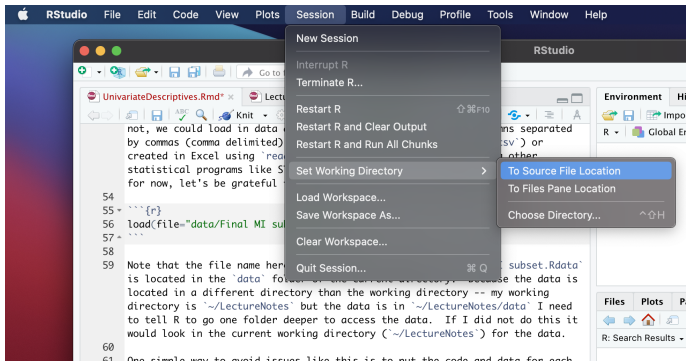
If this is different than where your code and data is there will be trouble!

3. Set the working directory to be the directory where your code and data is located!

```
setwd("/Users/clintojd/GitHub/vandy_ds_1000_materials/Lecture4")
```

Using RStudio to setwd()

1. Open your code in R-Studio



2.

3. Copy and paste resulting code snippet to your code!

Or just access the data from the web using url

To read in the data from before

```
```{r}
df <- readRDS(url("https://github.com/wdoyle42/vandy_ds_1000/raw/main/Lectures/Lecture2>HelloWorld/sc_debt.Rds"))
```
```

- ▶ `url()` tells R to look on the web for the location

Danger Will Robinson!

There is a difference in what R sees when working in the Console window (bad!) or using an R Script and what it sees when using RMarkdown!

- ▶ Objects in the Global Environment are not “seen” by RMarkdown.
- ▶ Nor does Knitting RMarkdown create objects in the Global Environment!

It can a challenge to develop code in RMarkdown if you forget what objects you have to work with!

```
```${r, echo=TRUE}  
load(file="data/Final MI subset.Rdata")
```
```



To recap....

So getting started we *always* start our code by:

1. Loading the tidyverse library

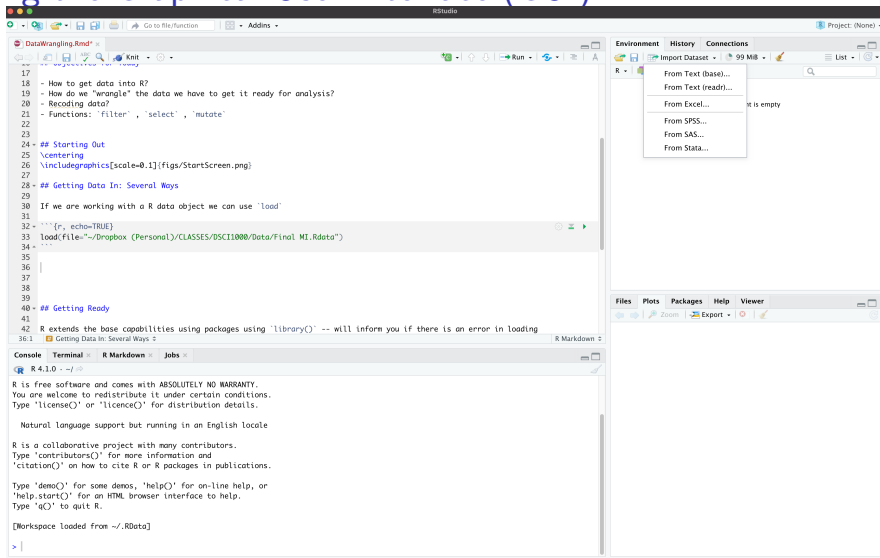
```
library(tidyverse)
```

2. Checking to sure that R is looking in right spot to find your data and code. It typically makes sense to `setwd` to tell R where in your computer it should be looking for additional data (and code).

```
getwd()  
setwd()
```

3. Now get that data!

Using the Graphical User Interface (GUI)



- ▶ If you insist on loading it this way, copy and paste resulting code into a script!

Importing data with functions

R gives us a few ways to access data using functions.

If we are working with a R data object we can use `load`

```
load(file="data/Final MI subset.Rdata")
```

NOTE: R will look for this file relative to where it is (as is given by `getwd`). Here I have created a data folder located within the current directory and telling R to look for my data in that folder!

NOTE: `load` will read in the R object stored under that file name. The name of the object being read in `load` may be different from the file name!

Can read lots of files!

Reading in non-R objects/data requires reading in the data and also assigning the data to a new R object.

- ▶ CSV (comma delimited) - generic "flat" file (i.e., just "rows and columns")

```
read.csv(file="data/JohnsHopkinsStateCasesTS.csv")
```

Remember to assign (<-) the data being read in to an object!

```
jh.covid <- read.csv(file="data/JohnsHopkinsStateCasesTS.csv")
```

And now we have two objects to work with in our Global Environment!

```
objects()
```

```
[1] "jh.covid" "MI_final"
```

NOTE: can also read in tab-delimited files (read.delim), Excel-created files (read.xls, read.xlsx), and files from other statistical languages (e.g., read.dta, read.sav).

Load and look...

Always see what you have loaded in! Look at data to see what and how much.

With tidyverse we can glimpse...

```
glimpse(jh.covid)
```

```
Rows: 33,234
```

```
Columns: 4
```

```
$ X      <int> 1, 2, 3, 4, 5, 6, 7, 8, ~
```

```
$ date   <chr> "2020-01-22", "2020-01-2~
```

```
$ state  <chr> "Alabama", "Alabama", "A~
```

```
$ cases  <int> 0, 0, 0, 0, 0, 0, 0, 0, ~
```

NOTE: define the type of variable!

Trouble with tibbles?

- ▶ Data objects in `tidyverse` are called `tibbles`.
- ▶ Data objects in base R are called `dataframes`.

Dimensions of data

- ▶ tibbles (and dataframes) can be thought of as a matrix with rows and columns:

```
dim(MI_final)
```

```
[1] 1231    13
```

```
dim(jh.covid)
```

```
[1] 33234     4
```

Using the Matrix...

We can examine/extract rows and columns of tibbles/dataframes.

Selecting rows/columns in a sequence:

```
jh.covid[1:3,]
```

| | X | date | state | cases |
|---|---|------------|---------|-------|
| 1 | 1 | 2020-01-22 | Alabama | 0 |
| 2 | 2 | 2020-01-23 | Alabama | 0 |
| 3 | 3 | 2020-01-24 | Alabama | 0 |

```
jh.covid[,1:3]
```

```
jh.covid[1:4,1:2]
```

| | X | date |
|---|---|------------|
| 1 | 1 | 2020-01-22 |
| 2 | 2 | 2020-01-23 |
| 3 | 3 | 2020-01-24 |
| 4 | 4 | 2020-01-25 |

Using the Matrix...

Selecting a collection/combination of rows

```
jh.covid[c(1:3,45:49,100),]
```

| | X | date | state | cases |
|-----|-----|------------|---------|-------|
| 1 | 1 | 2020-01-22 | Alabama | 0 |
| 2 | 2 | 2020-01-23 | Alabama | 0 |
| 3 | 3 | 2020-01-24 | Alabama | 0 |
| 45 | 45 | 2020-03-06 | Alabama | 0 |
| 46 | 46 | 2020-03-07 | Alabama | 0 |
| 47 | 47 | 2020-03-08 | Alabama | 0 |
| 48 | 48 | 2020-03-09 | Alabama | 0 |
| 49 | 49 | 2020-03-10 | Alabama | 0 |
| 100 | 100 | 2020-04-30 | Alabama | 7187 |

Creating New Objects:

Nothing we just did created a new object (tibble).

We need to assign (<-) what we do to a new object!

```
jh.covid.small <- jh.covid[1:3,]
```

```
jh.covid.small
```

| | X | date | state | cases |
|---|---|------------|---------|-------|
| 1 | 1 | 2020-01-22 | Alabama | 0 |
| 2 | 2 | 2020-01-23 | Alabama | 0 |
| 3 | 3 | 2020-01-24 | Alabama | 0 |

Referencing objects

We can also use the column (variable) names to extract columns.

```
jh.covid$state[1:10]
```

```
[1] "Alabama" "Alabama" "Alabama"  
[4] "Alabama" "Alabama" "Alabama"  
[7] "Alabama" "Alabama" "Alabama"  
[10] "Alabama"
```

NOTE: \$ tells us to look in the dataframe object `jh.covid` to find the state we are looking for.

2020 MI Exit Poll

To do our data wrangling we are going to wrangle the 2020 National Exit Poll from the National Election Pool in the state of Michigan.

- ▶ We are going to use the actual data we got on Election Night!
- ▶ To answer some of the same questions we were answering.
- ▶ But not much has been cleaned up since then so lots of work to do! (Ugh. . .)

2020 MI Exit Poll



**YOUR ANSWERS ARE
CONFIDENTIAL**
Please check only ONE
response for each
question.
Version 1

[A] Are you:

- 1 ☐ Male
2 ☐ Female

[B] Are you:

- 1 ☐ White
2 ☐ Black
3 ☐ Hispanic/Latino
4 ☐ Asian
5 ☐ American Indian
6 ☐ Other

[C] In today's election for president, did you just vote for:

- 1 ☐ Joe Biden (Dem)
2 ☐ Donald Trump (Rep)
9 ☐ Other: Who? _____
0 ☐ Did not vote

[D] When did you finally decide for whom to vote in the presidential election?

- 1 ☐ In the last few days
2 ☐ In the last week
3 ☐ In October
4 ☐ In September
5 ☐ Before that

[E] Are you of Hispanic or Latino descent?

- 1 ☐ Yes 2 ☐ No

[F] In which age group are you?

- 1 ☐ 18-24 6 ☐ 45-49
2 ☐ 25-29 7 ☐ 50-59
3 ☐ 30-34 8 ☐ 60-64
4 ☐ 35-39 9 ☐ 65-74
5 ☐ 40-44 10 ☐ 75 or over

[G] In today's election for U.S. Senate, did you just vote for:

- 1 ☐ Gary Peters (Dem)
2 ☐ John James (Rep)
9 ☐ Other: Who? _____
0 ☐ Did not vote

[H] Which best describes your education? You have:

- 1 ☐ Never attended college
2 ☐ Attended college but received no degree
3 ☐ Associate's degree (AA or AS)
4 ☐ Bachelor's degree (BA or BS)
5 ☐ An advanced degree after a bachelor's degree (such as JD, MA, MBA, MD, PhD)

[I] Compared to four years ago, is your family's financial situation:

- 1 ☐ Better today
2 ☐ Worse today
3 ☐ About the same

[J] Which ONE of these five issues mattered most in deciding how you voted for president?

(CHECK ONLY ONE)

- 1 ☐ Racial inequality
2 ☐ The coronavirus pandemic
3 ☐ The economy
4 ☐ Crime and safety
5 ☐ Health care policy

[K] Which ONE of these four candidate qualities mattered most in deciding how you voted for president?

(CHECK ONLY ONE)

- 1 ☐ Can unite the country
2 ☐ Is a strong leader
3 ☐ Cares about people like me
4 ☐ Has good judgment

[L] Does anyone in your household belong to a labor union?

- 1 ☐ Yes 2 ☐ No

[M] Do you think Joe Biden has the temperament to serve effectively as president?

- 1 ☐ Yes 2 ☐ No

[N] Do you think Donald Trump has the temperament to serve effectively as president?

- 1 ☐ Yes 2 ☐ No

[O] Would you rather see the U.S. Senate controlled by:

- 1 ☐ The Democratic Party
2 ☐ The Republican Party

[P] Which is more important?

- 1 ☐ Containing the coronavirus now, even if it hurts the economy
2 ☐ Rebuilding the economy now, even if it hurts efforts to contain the coronavirus

[Q] Is your opinion of Joe Biden:

- 1 ☐ Favorable
2 ☐ Unfavorable

[R] Is your opinion of Donald Trump:

- 1 ☐ Favorable
2 ☐ Unfavorable

[S] No matter how you voted today, do you usually think of yourself as a:

- 1 ☐ Democrat
2 ☐ Republican
3 ☐ Independent
4 ☐ Something else

[T] On most political matters, do you consider yourself:

- 1 ☐ Liberal
2 ☐ Moderate
3 ☐ Conservative

[U] 2019 total family income:

- 1 ☐ Under \$30,000
2 ☐ \$30,000 - \$49,999
3 ☐ \$50,000 - \$99,999
4 ☐ \$100,000 - \$199,999
5 ☐ \$200,000 or more

PLEASE TURN THE QUESTIONNAIRE OVER

Please fold questionnaire and put it in the box. Thank you.

Lots of interesting questions!

Predictive: Use the data to *predict* an outcome of interest.

- ▶ How many voters report voting for Biden vs. Trump?
- ▶ What predicts who supports Trump? And Biden?

Descriptive: Use the data to *describe* an event.

- ▶ How did the support for Trump and Biden vary by: gender? race? age? education?
- ▶ When did they make up their minds?
- ▶ *Why* did voters choose to vote for Trump? Or Biden?
- ▶ How do Trump and Biden voters vary in their opinions toward: COVID? Race relations?

(Some of) the data:

```
glimpse(MI_final)
```

```
Rows: 1,231
```

```
Columns: 13
```

```
$ SEX      <dbl> 2, 2, 2, 1, 2, 2, 1~  
$ AGE10    <dbl> 2, 10, 7, 9, 8, 7, ~  
$ PRSMI20  <dbl> 1, 1, 1, 1, 1, 1, 1~  
$ PARTYID  <dbl> 3, 1, 1, 3, 3, 3, 1~  
$ WEIGHT   <dbl> 0.4045421, 1.805261~  
$ QRACEAI  <dbl> 1, 2, 1, 1, 1, 1, 1~  
$ EDUC18   <dbl> 4, 1, 5, 4, 5, 3, 3~  
$ LGBT     <dbl> NA, 2, 2, NA, NA, 2~  
$ BRNAGAIN <dbl> NA, 1, 2, NA, NA, 2~  
$ LATINOS  <dbl> 2, 2, 2, 2, 2, 2, 2~  
$ RACISM20 <dbl> NA, 2, 2, NA, NA, 2~  
$ QLT20    <fct> Has good judgment, ~  
$ preschoice <chr> "Joe Biden, the Dem~
```

Data can differ!

There are several types of data:

1. `<dbl>` Double. “Numbers as a number.” Numbers stored to a high level of scientific precision. Mathematical operations are defined. (At least in theory!) e.g., `SEX`
2. `<int>` Integer. “Numbers as a number.” Mathematical operations are defined. (At least in theory!) R treats `<dbl>` and `<int>` as largely interchangeable.
3. `<chr>` Character. A variable with letter and/or number values. Mathematical operations are *not* defined, but other functions exist (e.g., extract the first and last characters, etc.) e.g., `preschoice`
4. `<fct>` Factor. A variable defining group membership. Mathematical operations are *not* defined, but they can be used in special ways in R. e.g. `QLT20`

NOTE: There are also `list` objects, but we will cover them when needed.

Setting the table?

Always start by exploring your data.

There are several functions depending on the type of data.

- ▶ All-purpose, but outdated: `table`

```
table(MI_final$preschoice)
```

```
          Another candidate
                25
Donald Trump, the Republican
                459
    Joe Biden, the Democrat
                723
                Refused
                14
    Undecided/Don't know
                4
Will/Did not vote for president
                6
```

Can I count on you?

- ▶ All-purpose (modern) tidyverse version of table: count:

```
count(MI_final,preschoice)
```

```
# A tibble: 6 x 2
```

| preschoice | n |
|-----------------------------------|-------|
| <chr> | <int> |
| 1 Another candidate | 25 |
| 2 Donald Trump, the Republican | 459 |
| 3 Joe Biden, the Democrat | 723 |
| 4 Refused | 14 |
| 5 Undecided/Don't know | 4 |
| 6 Will/Did not vote for president | 6 |

Can I count on you?

But not so useful for many-valued variables...

```
count(MI_final,WEIGHT)
```

```
# A tibble: 411 x 2
```

| | WEIGHT | n |
|--|--------|-------|
| | <dbl> | <int> |

| | | |
|---|-------|---|
| 1 | 0.100 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 2 | 0.113 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 3 | 0.119 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 4 | 0.133 | 2 |
|---|-------|---|

| | | |
|---|-------|---|
| 5 | 0.141 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 6 | 0.142 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 7 | 0.144 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 8 | 0.146 | 1 |
|---|-------|---|

| | | |
|---|-------|---|
| 9 | 0.147 | 1 |
|---|-------|---|

| | | |
|----|-------|---|
| 10 | 0.149 | 5 |
|----|-------|---|

```
# ... with 401 more rows
```

summary-izing numerics

If we have a numeric variable we can use the summary variable:

```
summary(MI_final$WEIGHT)
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. |
|--------|---------|--------|--------|---------|
| 0.1003 | 0.3775 | 0.8020 | 1.0000 | 1.4498 |
| Max. | | | | |
| 5.0853 | | | | |

Because everything is an object we can do cool things.

```
SummaryWeight <- summary(MI_final$WEIGHT)  
SummaryWeight[c(1,6)]
```

What does this do?

| Min. | Max. |
|-----------|-----------|
| 0.1002913 | 5.0852765 |

Summarizing factors and characters

```
summary(MI_final$QLT20)
```

```
[DON'T READ] Don't know/refused
                26
      Can unite the country
                125
    Cares about people like me
                121
      Has good judgment
                205
    Is a strong leader
                138
                NA's
                616
```

```
summary(MI_final$preschoice)
```

```
Length      Class      Mode
 1231 character character
```

You can even summary the whole tibble!

```
summary(MI_final)
```

| SEX | | AGE10 | |
|----------|-------|----------|---------|
| Min. | :1.00 | Min. | : 1.000 |
| 1st Qu.: | 1.00 | 1st Qu.: | 6.000 |
| Median | :2.00 | Median | : 8.000 |
| Mean | :1.53 | Mean | : 8.476 |
| 3rd Qu.: | 2.00 | 3rd Qu.: | 9.000 |
| Max. | :2.00 | Max. | :99.000 |

| PRSMI20 | | PARTYID | |
|----------|-------|----------|--------|
| Min. | :0.00 | Min. | :1.000 |
| 1st Qu.: | 1.00 | 1st Qu.: | 1.000 |
| Median | :1.00 | Median | :2.000 |
| Mean | :1.63 | Mean | :2.236 |
| 3rd Qu.: | 2.00 | 3rd Qu.: | 3.000 |
| Max. | :9.00 | Max. | :9.000 |

| WEIGHT | | QRACEAI | |
|----------|---------|----------|--------|
| Min. | :0.1003 | Min. | :1.000 |
| 1st Qu.: | 0.3775 | 1st Qu.: | 1.000 |

Enough playing around, let's get going

Now we will use the tools to help explore patterns of voting in Michigan in the 2020 presidential election.

- ▶ Which voters tended to support which candidate?
- ▶ How large are the differences in opinion between groups of voters thought to be politically relevant?

NOTE: This is exactly the kind of analysis that analysts do on Election Night (and thereafter).

Our toybox!



**YOUR ANSWERS ARE
CONFIDENTIAL**
Please check only ONE
response for each
question.
Version 1

[A] Are you:

- 1 ☐ Male
2 ☐ Female

[B] Are you:

- 1 ☐ White
2 ☐ Black
3 ☐ Hispanic/Latino
4 ☐ Asian
5 ☐ American Indian
6 ☐ Other

[C] In today's election for president, did you just vote for:

- 1 ☐ Joe Biden (Dem)
2 ☐ Donald Trump (Rep)
9 ☐ Other: Who? _____
0 ☐ Did not vote

[D] When did you finally decide for whom to vote in the presidential election?

- 1 ☐ In the last few days
2 ☐ In the last week
3 ☐ In October
4 ☐ In September
5 ☐ Before that

[E] Are you of Hispanic or Latino descent?

- 1 ☐ Yes 2 ☐ No

[F] In which age group are you?

- 1 ☐ 18-24 6 ☐ 45-49
2 ☐ 25-29 7 ☐ 50-59
3 ☐ 30-34 8 ☐ 60-64
4 ☐ 35-39 9 ☐ 65-74
5 ☐ 40-44 10 ☐ 75 or over

[G] In today's election for U.S. Senate, did you just vote for:

- 1 ☐ Gary Peters (Dem)
2 ☐ John James (Rep)
9 ☐ Other: Who? _____
0 ☐ Did not vote

[H] Which best describes your education? You have:

- 1 ☐ Never attended college
2 ☐ Attended college but received no degree
3 ☐ Associate's degree (AA or AS)
4 ☐ Bachelor's degree (BA or BS)
5 ☐ An advanced degree after a bachelor's degree (such as JD, MA, MBA, MD, PhD)

[I] Compared to four years ago, is your family's financial situation:

- 1 ☐ Better today
2 ☐ Worse today
3 ☐ About the same

[J] Which ONE of these five issues mattered most in deciding how you voted for president?

(CHECK ONLY ONE)

- 1 ☐ Racial inequality
2 ☐ The coronavirus pandemic
3 ☐ The economy
4 ☐ Crime and safety
5 ☐ Health care policy

[K] Which ONE of these four candidate qualities mattered most in deciding how you voted for president?

(CHECK ONLY ONE)

- 1 ☐ Can unite the country
2 ☐ Is a strong leader
3 ☐ Cares about people like me
4 ☐ Has good judgment

[L] Does anyone in your household belong to a labor union?

- 1 ☐ Yes 2 ☐ No

[M] Do you think Joe Biden has the temperament to serve effectively as president?

- 1 ☐ Yes 2 ☐ No

[N] Do you think Donald Trump has the temperament to serve effectively as president?

- 1 ☐ Yes 2 ☐ No

[O] Would you rather see the U.S. Senate controlled by:

- 1 ☐ The Democratic Party
2 ☐ The Republican Party

[P] Which is more important?

- 1 ☐ Containing the coronavirus now, even if it hurts the economy
2 ☐ Rebuilding the economy now, even if it hurts efforts to contain the coronavirus

[Q] Is your opinion of Joe Biden:

- 1 ☐ Favorable
2 ☐ Unfavorable

[R] Is your opinion of Donald Trump:

- 1 ☐ Favorable
2 ☐ Unfavorable

[S] No matter how you voted today, do you usually think of yourself as a:

- 1 ☐ Democrat
2 ☐ Republican
3 ☐ Independent
4 ☐ Something else

[T] On most political matters, do you consider yourself:

- 1 ☐ Liberal
2 ☐ Moderate
3 ☐ Conservative

[U] 2019 total family income:

- 1 ☐ Under \$30,000
2 ☐ \$30,000 - \$49,999
3 ☐ \$50,000 - \$99,999
4 ☐ \$100,000 - \$199,999
5 ☐ \$200,000 or more

PLEASE TURN THE QUESTIONNAIRE OVER

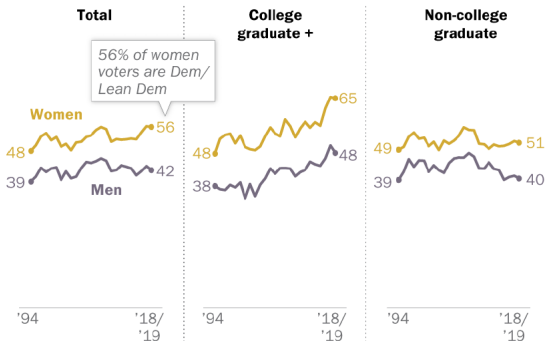
Please fold questionnaire and put it in the box. Thank you.

Question: “Gender” Gap Among Younger Voters

Of historical interest: is there a difference in how men and women vote?

Wide gender gap in leaned partisanship, especially among college graduates

% of ___ registered voters who identify as Democrats or lean toward the Democratic Party



Notes: Based on registered voters. Due to smaller sample sizes in 2018 and 2019, the data from those years has been combined.

Source: Annual totals of Pew Research Center survey data (U.S. adults).

PEW RESEARCH CENTER

Is this as easy as we think? Of course not. . .

So all we need to do is to compare males vs. females?!?

Tasks Ahead

Questions of Interest:

- ▶ How does the support for Biden and Trump vary for younger males and females?
- ▶ How does the “gender gap” vary by: age, race, and education?

To do this:

1. Create the variables needed for analysis (`mutate`)
2. Select the relevant data for analysis (`select`, `filter`)
3. Summarize/quantify the difference (using `mean`)

We are going to doing some abstract work to learn some tools and then apply what we need to answer the question.

Principles of Data Wrangling

- ▶ *Replication*: Can others do what you did?
- ▶ *Understanding*: Can others understand what you did and why? Can you follow what you did if you come back to the code in a year?
- ▶ *Robustness*: Does your code “break” easily?

Selecting variables using select

Many times we want to work with only the portion of the tibble that we need – especially if the tibble is large! So we often begin by selecting variables (columns) and filtering (rows) to the relevant data.

This requires you to know what question you are asking of the data!

Tools are in the dplyr library so we need to install and load that library first.

```
#install.packages(dplyr)      # Already installed for me  
library(dplyr)
```

Selecting variables using select

What we want to do is to create a new tibble with 4 variables from MI_final.

```
MI_small <- select(MI_final, c(SEX,AGE10,PRSMI20,PARTYID))  
glimpse(MI_small)
```

Rows: 1,231

Columns: 4

\$ SEX <dbl> 2, 2, 2, 1, 2, 2, 1, 1~

\$ AGE10 <dbl> 2, 10, 7, 9, 8, 7, 9, ~

\$ PRSMI20 <dbl> 1, 1, 1, 1, 1, 1, 1, 1~

\$ PARTYID <dbl> 3, 1, 1, 3, 3, 3, 1, 1~

NOTE: Make sure to define the selection to a new object!

Dropping variables using select

We can also drop variables by negatively selecting them. To drop AGE10:

```
MI_small_1 <- select(MI_small, -AGE10)
glimpse(MI_small_1)
```

Rows: 1,231

Columns: 3

\$ SEX <dbl> 2, 2, 2, 1, 2, 2, 1, 1~

\$ PRSMI20 <dbl> 1, 1, 1, 1, 1, 1, 1, 1~

\$ PARTYID <dbl> 3, 1, 1, 3, 3, 3, 1, 1~

Selecting based on variable names

Can also select based on variable names:

```
MI_small_2 <- select(MI_small, starts_with("P"))  
glimpse(MI_small_2)
```

Rows: 1,231

Columns: 2

\$ PRSMI20 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1~

\$ PARTYID <dbl> 3, 1, 1, 3, 3, 3, 1, 1, 1~

NOTE: Useful if variables are sensibly named (e.g., "d_age")

Selecting based on variable names

```
MI_small_3 <- select(MI_small, !ends_with("0"))  
glimpse(MI_small_3)
```

Rows: 1,231

Columns: 2

\$ SEX <dbl> 2, 2, 2, 1, 2, 2, 1, 1~

\$ PARTYID <dbl> 3, 1, 1, 3, 3, 3, 1, 1~

Selecting a range of variables

```
MI_small_4 <- select(MI_small,  
                     SEX:PRSMI20)  
glimpse(MI_small_4)
```

Rows: 1,231

Columns: 3

\$ SEX <dbl> 2, 2, 2, 1, 2, 2, 1, 1~

\$ AGE10 <dbl> 2, 10, 7, 9, 8, 7, 9, ~

\$ PRSMI20 <dbl> 1, 1, 1, 1, 1, 1, 1, 1~

Why is this not good practice?

Conditionals: “AND”

- ▶ Select variables “if and only if” multiple conditions are true: & (AND)

```
MI_small <- select(MI_final, SEX & starts_with("P"))  
glimpse(MI_small)
```

Rows: 1,231

Columns: 0

Conditionals: “OR”

- Select variables “if and only if” multiple conditions are true: | (OR)

```
MI_small <- select(MI_final, SEX | starts_with("P"))  
glimpse(MI_small)
```

Rows: 1,231

Columns: 4

```
$ SEX      <dbl> 2, 2, 2, 1, 2, 2, 1~  
$ PRSMI20  <dbl> 1, 1, 1, 1, 1, 1, 1~  
$ PARTYID  <dbl> 3, 1, 1, 3, 3, 3, 1~  
$ preschoice <chr> "Joe Biden, the Dem~
```

Working with a subset of rows: filter

filter selects all rows satisfying the specified condition(s)

```
filter(MI_final,SEX==2)
```

```
# A tibble: 652 x 13
```

| | SEX | AGE10 | PRSMI20 | PARTYID | WEIGHT |
|----|-------|-------|---------|---------|--------|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 2 | 2 | 1 | 3 | 0.405 |
| 2 | 2 | 10 | 1 | 1 | 1.81 |
| 3 | 2 | 7 | 1 | 1 | 0.860 |
| 4 | 2 | 8 | 1 | 3 | 0.177 |
| 5 | 2 | 7 | 1 | 3 | 0.492 |
| 6 | 2 | 6 | 2 | 2 | 1.50 |
| 7 | 2 | 1 | 1 | 4 | 0.593 |
| 8 | 2 | 10 | 1 | 1 | 1.59 |
| 9 | 2 | 8 | 1 | 1 | 1.44 |
| 10 | 2 | 9 | 2 | 3 | 0.438 |

```
# ... with 642 more rows, and 8 more
```

```
# variables: QRACEAI <dbl>,  
# EDUC18 <dbl>, LGBT <dbl>,  
# BRNAGAIN <dbl>, LATINOS <dbl>,  
# RACISM20 <dbl>, CLT20 <dbl>
```

Define a new object?

If you are going to use the subset again, define a new object!

```
dim(MI_final)
```

```
[1] 1231  13
```

```
MI_female <- filter(MI_final,SEX==2)
```

```
dim(MI_female)
```

```
[1] 652  13
```

NOTE: Sometimes we will just pass the filter condition e.g. ,
`filter(MI_final,SEX==2)` to another function rather than defining a
new object if we only need to filter once.

filter if...

Lke select, we can also filter using conditions.

- ▶ & (AND) selects if *both* conditions are true
- ▶ | (OR) selects if *either* condition is true

Applying filters

Select observations of females under the age of 24?

```
MI_female_under24 <- filter(MI_final,  
                             SEX == 2 & AGE10 == 1)
```

```
dim(MI_female_under24)
```

```
[1] 17 13
```

Sequential filters are the same as a joint filter!

```
MI_female_under24 <- filter(MI_female,  
                             AGE10 == 1)
```


Applying filters

Select observations of females under the age of 24 *or* over the age of 75?

```
MI_female_LT24orGT75 <- filter(MI_final,  
                                SEX == 2 & (AGE10 == 1 | AGE10 == 10))
```

```
dim(MI_female_LT24orGT75)
```

```
[1] 128 13
```

Applying filters

Select observations of females under the age of 24 *and* over the age of 75?

```
MI_female_LT24andGT75 <- filter(MI_final,  
                                SEX == 2 & (AGE10 == 1 & AGE10 == 10))
```

```
dim(MI_female_LT24andGT75)
```

```
[1] 0 13
```

Creating new variables

Do our variables make sense for what we want to do?

And are they sensibly named and coded?

```
mean(MI_final$SEX)
```

```
[1] 1.529651
```

```
count(MI_final,SEX)
```

```
# A tibble: 2 x 2
```

| | SEX | n |
|---|-------|-------|
| | <dbl> | <int> |
| 1 | 1 | 579 |
| 2 | 2 | 652 |

What does that even mean? Make sure your variables are meaningfully defined! If not, may need to create some new ones.

Variable creation via assignment

We can always use the variable assignment to create a new variable.

How do these differ?

```
sex.recode <- MI_final$SEX
```

```
MI_final$sex.recode <- MI_final$SEX
```

What was the point of that?

```
MI_final$sex.recode[MI_final$sex.recode==1] <- "Male"
```

```
MI_final$sex.recode[MI_final$sex.recode==2] <- "Female"
```

Does this make us happy?

```
count(MI_final,sex.recode)
```

```
# A tibble: 2 x 2  
  sex.recode      n  
  <chr>      <int>  
1 Female      652  
2 Male       579
```

Better value labels, but can we do better still?

What about?

```
MI_final$FEMALE <- MI_final$SEX -1
```

ASIDE: indicator variables

When we have variables that describe group membership we can use that to create a variable that indicates if the observation belongs to the group (1) or not (0).

```
count(MI_final,FEMALE)
```

```
# A tibble: 2 x 2
```

```
  FEMALE      n
```

```
  <dbl> <int>
```

```
1      0   579
```

```
2      1   652
```

But now some things make sense mathematically!

```
mean(MI_final$FEMALE)
```

```
[1] 0.5296507
```

And also note the importance of the variable name for interpretation! The name is the meaning of a value of 1.

A better way? mutate

While we can create new variables one at a time and assign them to a tibble this can be a pain.

`mutate` lets us create new variables and also the potential to create a new tibble by creating a tibble consisting of all existing variables plus the new additions.

```
load(file="data/Final MI subset.Rdata")
MI_final_new <- mutate(MI_final,
                       FEMALE = SEX - 1)
```

```
dim(MI_final)
```

```
[1] 1231  13
```

```
dim(MI_final_new)
```

```
[1] 1231  14
```

Let's focus on vote choice...

[C] In today's election for president, did you just vote for:

1 ☐ Joe Biden (Dem)

2 ☐ Donald Trump (Rep)

9 ☐ Other: Who? _____

0 ☐ Did not vote

```
count(MI_final, preschoice)
```

```
# A tibble: 6 x 2
```

| preschoice | n |
|-----------------------------------|-------|
| <chr> | <int> |
| 1 Another candidate | 25 |
| 2 Donald Trump, the Republican | 459 |
| 3 Joe Biden, the Democrat | 723 |
| 4 Refused | 14 |
| 5 Undecided/Don't know | 4 |
| 6 Will/Did not vote for president | 6 |

Can use mutate to create multiple new variables

Let's create some more indicators...

```
MI_final <- mutate(MI_final,  
  FEMALE = SEX - 1,  
  WGT100 = WEIGHT*100,  
  BidenVoter = ifelse(PRSMI20==1, 1 , 0),  
  TrumpVoter = ifelse(PRSMI20==2, 1 , 0))
```

NOTE: we replaced MI_final with MI_final instead of creating a new tibble.

Should the percentage of Biden votes and Trump voters sum to 1?

```
mean(MI_final$BidenVoter)
```

```
[1] 0.5873274
```

```
mean(MI_final$TrumpVoter)
```

```
[1] 0.3728676
```

Enough screwing around, let's put this to work...

- ▶ What is the Gender gap in Michigan among the oldest and youngest voters?
- ▶ Which is larger – gender differences or age differences?

```
load(file="data/Final MI subset.Rdata")
```

```
MI_final <- mutate(MI_final,  
                   FEMALE = SEX - 1,  
                   BidenVoter = ifelse(PRSMI20==1, 1 , 0),  
                   TrumpVoter = ifelse(PRSMI20==2, 1 , 0))
```

```
MI_femaleunder24 <- filter(MI_final, FEMALE==1 & AGE10 == 1)
```

Creating new tibbles with summarize

- Summarize lets us create a new tibble that is a function of an existing tibble!

```
summarize(MI_femaleunder24,  
          N_Obs = length(BidenVoter),  
          BidenPct = mean(BidenVoter),  
          TrumpPct = mean(TrumpVoter),  
          WeightSD = sd(WEIGHT),  
          medianED = median(EDUC18),  
          ImpRacism = mean(RACISM20, na.rm=TRUE))
```

```
# A tibble: 1 x 6  
  N_Obs BidenPct TrumpPct WeightSD  
  <int>   <dbl>   <dbl>   <dbl>  
1    17    0.882    0.0588    1.22  
# ... with 2 more variables:  
#   medianED <dbl>, ImpRacism <dbl>
```

- So why doesn't BidenPct and TrumpPct sum to 1? Should it?

Let's focus on Biden and Trump voters...

```
MI_femaleunder24 <- filter(MI_final,  
                           (BidenVoter==1 | TrumpVoter==1) &  
                           FEMALE==1 & AGE10 == 1)
```

Now summarize...

```
summarize(MI_femaleunder24,  
          BidenPct = mean(BidenVoter),  
          TrumpPct = mean(TrumpVoter))
```

```
# A tibble: 1 x 2  
  BidenPct TrumpPct  
    <dbl>    <dbl>  
1    0.938    0.0625
```

- ▶ Now we don't need both because they add to 1!
- ▶ $\text{TrumpPct} = 1 - \text{BidenPct}$

Do again, and again, and again

```
MI_maleunder24 <- filter(MI_final,  
                          (BidenVoter==1 | TrumpVoter==1) &  
                          FEMALE==0 & AGE10 == 1)  
  
MI_maleover64 <- filter(MI_final,  
                        (BidenVoter==1 | TrumpVoter==1) &  
                        FEMALE==1 & AGE10 >= 9)  
  
MI_femaleover64 <- filter(MI_final,  
                          (BidenVoter==1 | TrumpVoter==1) &  
                          FEMALE==0 & AGE10 >= 9)
```

THIS IS VERY, VERY, VERY BAD CODING PRACTICE!

- ▶ Inefficient (too many tibbles!)
- ▶ Prone to error (“copy and paste” is your enemy)

Now Summarize!

```
male24Biden <- summarize(MI_maleunder24,  
  BidenPct = mean(BidenVoter))  
  
female24Biden <- summarize(MI_femaleunder24,  
  BidenPct = mean(BidenVoter))  
  
male64Biden <- summarize(MI_maleover64,  
  BidenPct = mean(BidenVoter))  
  
female64Biden <- summarize(MI_femaleover64,  
  BidenPct = mean(BidenVoter))
```

There has to be a better way... (and there is!)

Now infer!

Gender gap among young?

```
male24Biden - female24Biden
```

```
BidenPct
```

```
1    -0.375
```

Which Differences are largest?

Gender gap among old?

```
male64Biden - female64Biden
```

```
BidenPct
```

```
1 0.06888854
```

Age gap among men?

```
male64Biden - male24Biden
```

```
BidenPct
```

```
1 0.0909296
```

Age gap among female?

```
female64Biden - female24Biden
```

```
BidenPct
```

```
1 -0.3529589
```


Other Comparisons?

- ▶ Education gap? (EDUC18)
- ▶ Racial gap? (QRACEAI)
- ▶ Religion gap? (BRNAGAIN)
- ▶ NOTE: Value labels given by numeric codes on Questionnaire

Causality?

- ▶ What does it mean to look at opinions by gender/age/race?
- ▶ How do we interpret that relationship?
- ▶ Are we (explicitly/implicitly) suggesting that they are the *cause*?
- ▶ But more likely they are related to other experiences/aspects?
- ▶ If so, is it misleading to focus on demographics rather than underlying events?

Stepping Back

- ▶ How confident should we be based on *amount of data*?
- ▶ How confident should we be based on *how collected?* (i.e., who is included and excluded?)

YOUR CONCLUSIONS ARE ONLY AS GOOD AS THE DATA YOU HAVE!

Pipe dreams?

- ▶ What we just did was a coding nightmare.
- ▶ Lots of replicated code and copying and pasting.
- ▶ Lots of potential for user-error.
- ▶ Surely we can do it in a cleaner way?

Piping code through a tibble

Think of the pipe command %>% as “then”:

```
MI_final %>%  
  count(preschoice)
```

```
# A tibble: 6 x 2
```

| preschoice | n |
|-----------------------------------|-------|
| <chr> | <int> |
| 1 Another candidate | 25 |
| 2 Donald Trump, the Republican | 459 |
| 3 Joe Biden, the Democrat | 723 |
| 4 Refused | 14 |
| 5 Undecided/Don't know | 4 |
| 6 Will/Did not vote for president | 6 |

“Use MI_final *then* count the variable preschoice”

Stacking pipes!

Very useful for efficiently doing steps in a sequence:

```
MI_final %>%  
  filter(AGE10==1 & FEMALE==0) %>%  
  summarize(  
    BidenPct = mean(preschoice=="Joe Biden, the Democrat"),  
    TrumpPct = mean(preschoice=="Donald Trump, the Republican"))
```



```
# A tibble: 1 x 2  
  BidenPct TrumpPct  
    <dbl>    <dbl>  
1    0.562    0.438
```

Grouping!

```
MI_final %>%  
  filter(AGE10==1) %>%  
  group_by(FEMALE) %>%  
  summarize(  
    BidenPct = mean(preschoice=="Joe Biden, the Democrat"),  
    TrumpPct = mean(preschoice=="Donald Trump, the Republican"))
```

```
# A tibble: 2 x 3  
  FEMALE BidenPct TrumpPct  
  <dbl>   <dbl>   <dbl>  
1      0     0.562     0.438  
2      1     0.882     0.0588
```

Now we have the power!

```
AgeGenderPct <- MI_final %>%  
  group_by(AGE10, FEMALE) %>%  
  summarize(  
    BidenPct = mean(preschoice=="Joe Biden, the Democrat"),  
    TrumpPct = mean(preschoice=="Donald Trump, the Republican"))
```


Now we have the power!

```
AgeGenderPct
```

```
# A tibble: 22 x 4
# Groups:   AGE10 [11]
  AGE10 FEMALE BidenPct TrumpPct
  <dbl> <dbl>    <dbl>    <dbl>
1     1     0    0.562    0.438
2     1     1    0.882    0.0588
3     2     0     0.5     0.357
4     2     1    0.929    0.0714
5     3     0    0.556    0.407
6     3     1     0.6     0.267
7     4     0    0.759    0.241
8     4     1    0.706    0.294
9     5     0    0.543    0.391
10    5     1     0.5     0.469
# ... with 12 more rows
```

We can now work with this tibble!

```
filter(AgeGenderPct, FEMALE==0)
```

```
# A tibble: 11 x 4
```

```
# Groups:   AGE10 [11]
```

| | AGE10 | FEMALE | BidenPct | TrumpPct |
|----|-------|--------|----------|----------|
| | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 1 | 0 | 0.562 | 0.438 |
| 2 | 2 | 0 | 0.5 | 0.357 |
| 3 | 3 | 0 | 0.556 | 0.407 |
| 4 | 4 | 0 | 0.759 | 0.241 |
| 5 | 5 | 0 | 0.543 | 0.391 |
| 6 | 6 | 0 | 0.375 | 0.525 |
| 7 | 7 | 0 | 0.452 | 0.524 |
| 8 | 8 | 0 | 0.523 | 0.415 |
| 9 | 9 | 0 | 0.566 | 0.369 |
| 10 | 10 | 0 | 0.568 | 0.420 |
| 11 | 99 | 0 | 0.25 | 0.5 |

We can now work with this tibble!

```
filter(AgeGenderPct, AGE10==1)
```

```
# A tibble: 2 x 4
```

```
# Groups:   AGE10 [1]
```

| | AGE10 | FEMALE | BidenPct | TrumpPct |
|---|-------|--------|----------|----------|
| | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 1 | 0 | 0.562 | 0.438 |
| 2 | 1 | 1 | 0.882 | 0.0588 |

But how much data are we talking about?

```
MI_final %>%  
  group_by(AGE10, FEMALE) %>%  
  count(preschoice)
```

```
# A tibble: 75 x 4
```

```
# Groups:   AGE10, FEMALE [22]
```

| | AGE10 | FEMALE | preschoice | n |
|----|-------|--------|--------------------|-------|
| | <dbl> | <dbl> | <chr> | <int> |
| 1 | 1 | 0 | Donald Trump, the~ | 7 |
| 2 | 1 | 0 | Joe Biden, the De~ | 9 |
| 3 | 1 | 1 | Another candidate | 1 |
| 4 | 1 | 1 | Donald Trump, the~ | 1 |
| 5 | 1 | 1 | Joe Biden, the De~ | 15 |
| 6 | 2 | 0 | Another candidate | 1 |
| 7 | 2 | 0 | Donald Trump, the~ | 5 |
| 8 | 2 | 0 | Joe Biden, the De~ | 7 |
| 9 | 2 | 0 | Refused | 1 |
| 10 | 2 | 1 | Donald Trump, the~ | 1 |

```
# ... with 65 more rows
```

Can get both counts and proportions? Of course...

```
MI_final %>%  
  group_by(AGE10, FEMALE) %>%  
  count(preschoice) %>%  
  mutate(Prop = n / sum(n))
```

```
# A tibble: 75 x 5
```

```
# Groups:   AGE10, FEMALE [22]
```

| | AGE10 | FEMALE | preschoice | n | Prop |
|----|-------|--------|-------------|-------|--------|
| | <dbl> | <dbl> | <chr> | <int> | <dbl> |
| 1 | 1 | 0 | Donald Tru~ | 7 | 0.438 |
| 2 | 1 | 0 | Joe Biden,~ | 9 | 0.562 |
| 3 | 1 | 1 | Another ca~ | 1 | 0.0588 |
| 4 | 1 | 1 | Donald Tru~ | 1 | 0.0588 |
| 5 | 1 | 1 | Joe Biden,~ | 15 | 0.882 |
| 6 | 2 | 0 | Another ca~ | 1 | 0.0714 |
| 7 | 2 | 0 | Donald Tru~ | 5 | 0.357 |
| 8 | 2 | 0 | Joe Biden,~ | 7 | 0.5 |
| 9 | 2 | 0 | Refused | 1 | 0.0714 |
| 10 | 2 | 1 | Donald Tru~ | 1 | 0.0714 |

```
# ... with 65 more rows
```

We can also ungroup! How will these differ?

```
grouped <- MI_final %>%  
  filter(AGE10==1) %>%  
  group_by(FEMALE) %>%  
  count(preschoice) %>%  
  mutate(Prop = n / sum(n))
```

```
ungrouped <- MI_final %>%  
  filter(AGE10==1) %>%  
  group_by(FEMALE) %>%  
  count(preschoice) %>%  
  ungroup() %>%  
  mutate(Prop = n / sum(n))
```

grouped

```
# A tibble: 5 x 4
```

```
# Groups:   FEMALE [2]
```

| | FEMALE | preschoice | n | Prop |
|---|--------|--------------------|-------|--------|
| | <dbl> | <chr> | <int> | <dbl> |
| 1 | 0 | Donald Trump, the~ | 7 | 0.438 |
| 2 | 0 | Joe Biden, the De~ | 9 | 0.562 |
| 3 | 1 | Another candidate | 1 | 0.0588 |
| 4 | 1 | Donald Trump, the~ | 1 | 0.0588 |
| 5 | 1 | Joe Biden, the De~ | 15 | 0.882 |

ungrouped

```
# A tibble: 5 x 4
```

| | FEMALE | preschoice | n | Prop |
|---|--------|--------------------|-------|--------|
| | <dbl> | <chr> | <int> | <dbl> |
| 1 | 0 | Donald Trump, the~ | 7 | 0.212 |
| 2 | 0 | Joe Biden, the De~ | 9 | 0.273 |
| 3 | 1 | Another candidate | 1 | 0.0303 |
| 4 | 1 | Donald Trump, the~ | 1 | 0.0303 |
| 5 | 1 | Joe Biden, the De~ | 15 | 0.455 |