Econ 106: Data Analysis for Economics

Lecture 9

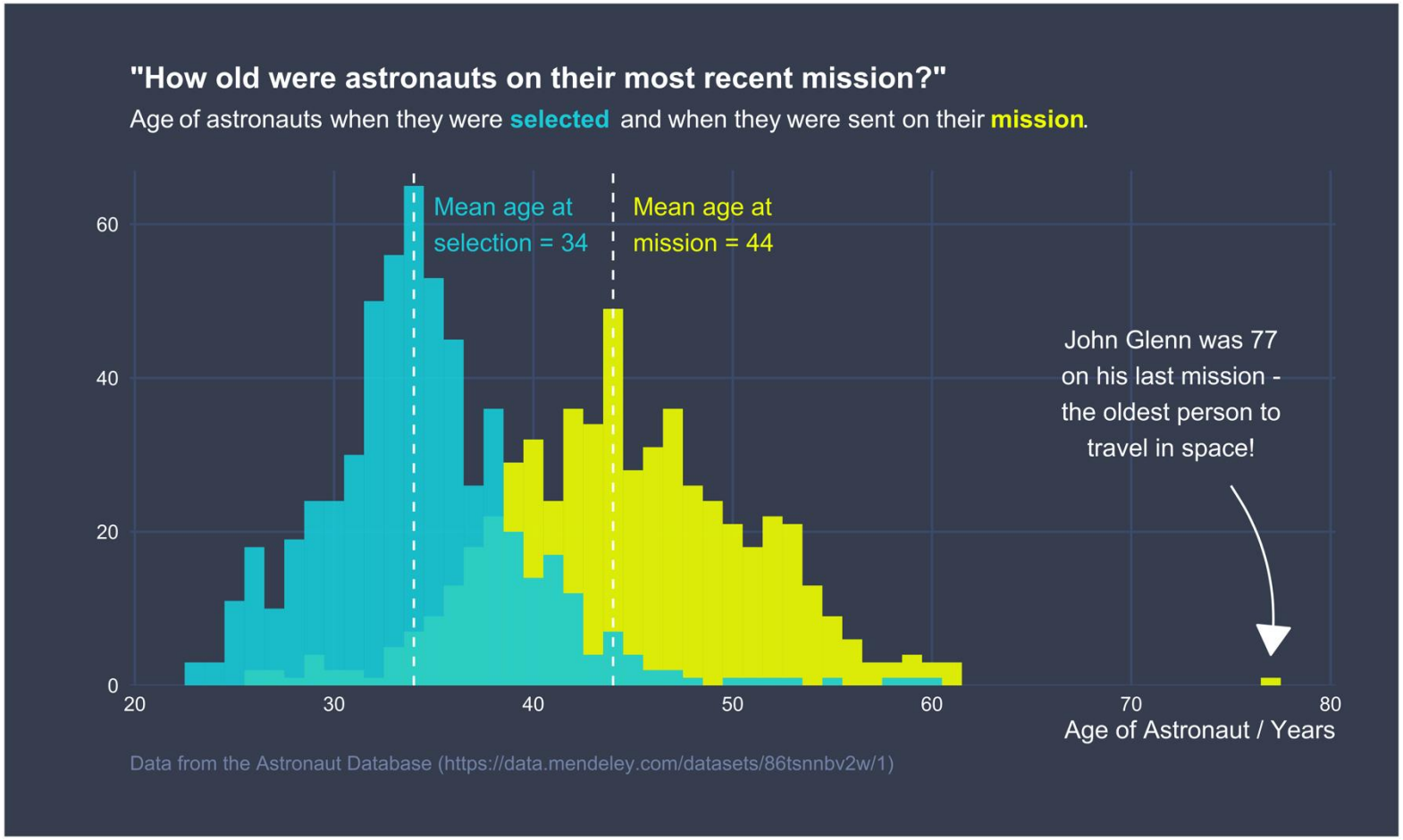
slides adapted from: https://jhudatascience.org/tidyversecourse/model.html#descriptive-and-exploratory-analysis

Reminder

- Research Milestone #1 is due Sunday 11:59pm
- Please reach out to me if you have any questions about your choice of dataset (I'm happy to give suggestions)
- You can also post your dataset here and I will give feedback

#tidytuesday

- code here



Outline for Today

- Descriptive vs. Exploratory Analysis
- Summary tables and figures for:
  - single quantitative variable
  - single categorical variable

Descriptive vs. Exploratory Analysis

- The goal of a descriptive analysis is to generate simple **summaries** to **describe** the data you're working with
- The goal of an exploratory analysis is to **explore** the data and find **relationships** that weren't previously known.
- Today: descriptive
- next lecture: exploratory

Descriptive Analysis Example: Single Variable

- In the US census, the government collects a series of measurements on all the country's residents.
- This table <u>describes</u> the age distribution of the population

Subject	Total Estimate
Total population	309,349,689
AGE	
Under 5 years	6.5%
5 to 9 years	6.6%
10 to 14 years	6.7%
15 to 19 years	7.1%
20 to 24 years	7.0%
25 to 29 years	6.8%
30 to 34 years	6.5%
35 to 39 years	6.5%
40 to 44 years	6.8%
45 to 49 years	7.3%
50 to 54 years	7.2%
55 to 59 years	6.4%
60 to 64 years	5.5%
65 to 69 years	4.0%
70 to 74 years	3.0%
75 to 79 years	2.3%
80 to 84 years	1.9%
85 years and over	1.8%

2010 US Census Data Summary Table (broken down by age)

https://factfinder.census.gov/

Exploratory Analysis Example: Two Variables

- We can <u>explore</u> whether there is a relationship between age and gender
- The idea: does the age distribution look different for men vs women?

Subject	United States		
	Total	Male	Female
	Estimate	Estimate	Estimate
Total population	309,349,689	152,089,450	157,260,239
AGE			
Under 5 years	6.5%	6.8%	6.3%
5 to 9 years	6.6%	6.8%	6.4%
10 to 14 years	6.7%	7.0%	6.4%
15 to 19 years	7.1%	7.5%	6.8%
20 to 24 years	7.0%	7.3%	6.7%
25 to 29 years	6.8%	6.9%	6.6%
30 to 34 years	6.5%	6.6%	6.4%
35 to 39 years	6.5%	6.6%	6.5%
40 to 44 years	6.8%	6.9%	6.7%
45 to 49 years	7.3%	7.3%	7.3%
50 to 54 years	7.2%	7.2%	7.2%
55 to 59 years	6.4%	6.3%	6.5%
60 to 64 years	5.5%	5.4%	5.6%
65 to 69 years	4.0%	3.9%	4.2%
70 to 74 years	3.0%	2.8%	3.2%
75 to 79 years	2.3%	2.1%	2.6%
80 to 84 years	1.9%	1.5%	2.2%
85 years and over	1.8%	1.2%	2.4%

... and stratified by sex

https://factfinder.census.gov/

Numeric vs. Factor Variables

- How we summarize/describe a variable will depend on whether it is quantitative (numeric) or categorical (factor):
  - Quantitative:
    - histograms
    - density plot
  - Categorical:
    - bar plot

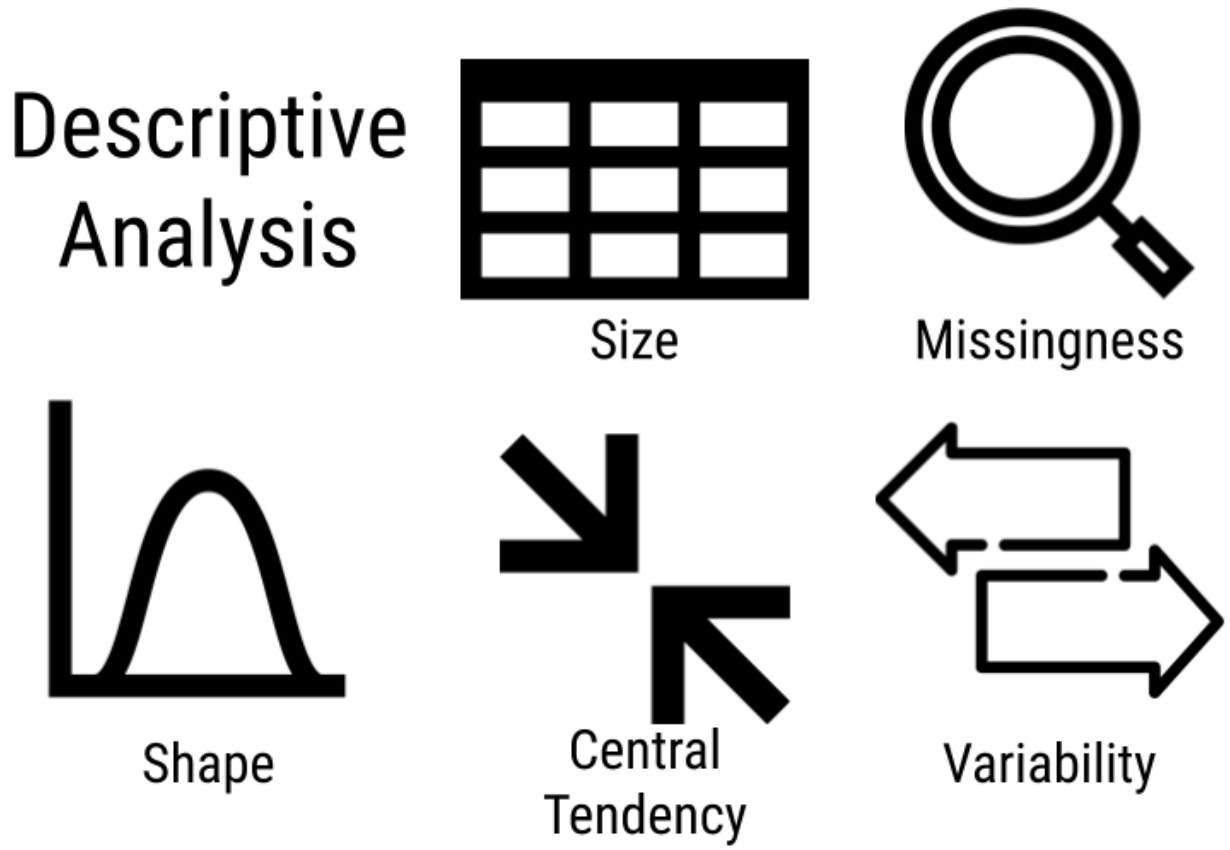
First, look at your data

- Remember to always look at your data
- GSS users: this is where you will see that variables you thought were quantitative are in fact categorical

	year	marital	age	race	rincome	partyid	relig	denom	tvhours
1	2000	Never married	26	White	$8000 to 9999	Ind,near rep	Protestant	Southern baptist	12
2	2000	Divorced	48	White	$8000 to 9999	Not str republican	Protestant	Baptist-dk which	NA
3	2000	Widowed	67	White	Not applicable	Independent	Protestant	No denomination	2
4	2000	Never married	39	White	Not applicable	Ind,near rep	Orthodox-christian	Not applicable	4
5	2000	Divorced	25	White	Not applicable	Not str democrat	None	Not applicable	1
6	2000	Married	25	White	$20000 - 24999	Strong democrat	Protestant	Southern baptist	NA
7	2000	Never married	36	White	$25000 or more	Not str republican	Christian	Not applicable	3
8	2000	Divorced	44	White	$7000 to 7999	Ind,near dem	Protestant	Lutheran-mo synod	NA
9	2000	Married	44	White	$25000 or more	Not str democrat	Protestant	Other	0
10	2000	Married	47	White	$25000 or more	Strong republican	Protestant	Southern baptist	3
11	2000	Married	53	White	$25000 or more	Not str democrat	Protestant	Other	2

Examining Quantitative Variables

- Missingness
- Shape
- Center
- Spread
- Unusual Values



summary()

- We can see some detailed information on our quantitative data (year, age, tvhours)

```
> summary(gss_cat)
      year               marital           age                     race                  rincome
 Min.   :2000   No answer    :   17   Min.   :18.00   Other         : 1959   $25000 or more:7363
 1st Qu.:2002   Never married: 5416   1st Qu.:33.00   Black         : 3129   Not applicable:7043
 Median :2006   Separated    :  743   Median :46.00   White         :16395   $20000 - 24999:1283
 Mean   :2007   Divorced     : 3383   Mean   :47.18   Not applicable:    0   $10000 - 14999:1168
 3rd Qu.:2010   Widowed      : 1807   3rd Qu.:59.00                           $15000 - 19999:1048
 Max.   :2014   Married      :10117   Max.   :89.00                           Refused       : 975
                                       NA's   :76                              (Other)       :2603
               partyid            relig                     denom          tvhours
 Independent       :4119   Protestant:10846   Not applicable  :10072   Min.   : 0.000
 Not str democrat  :3690   Catholic  : 5124   Other           : 2534   1st Qu.: 1.000
 Strong democrat   :3490   None      : 3523   No denomination : 1683   Median : 2.000
 Not str republican:3032   Christian :  689   Southern baptist: 1536   Mean   : 2.981
 Ind,near dem      :2499   Jewish    :  388   Baptist-dk which: 1457   3rd Qu.: 4.000
 Strong republican :2314   Other     :  224   United methodist: 1067   Max.   :24.000
 (Other)           :2339   (Other)   :  689   (Other)         : 3134   NA's   :10146
>
```

Better summaries with skimr

```
> skim(gss_cat)
── Data Summary ────────────────────────
                           Values
Name                       gss_cat
Number of rows             21483
Number of columns          9
_______________________
Column type frequency:
  factor                   6
  numeric                  3
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 marital               0             1 FALSE          6 Mar: 10117, Nev: 5416, Div: 3383, Wid: 1807
2 race                  0             1 FALSE          3 Whi: 16395, Bla: 3129, Oth: 1959, Not: 0
3 rincome               0             1 FALSE         16 $25: 7363, Not: 7043, $20: 1283, $10: 1168
4 partyid               0             1 FALSE         10 Ind: 4119, Not: 3690, Str: 3490, Not: 3032
5 relig                 0             1 FALSE         15 Pro: 10846, Cat: 5124, Non: 3523, Chr: 689
6 denom                 0             1 FALSE         30 Not: 10072, Oth: 2534, No : 1683, Sou: 1536

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean    sd   p0  p25  p50  p75 p100 hist
1 year                  0         1     2007.    4.45 2000 2002 2006 2010 2014 ▅▅▅▅
2 age                  76         0.996   47.2  17.3    18   33   46   59   89 ▅▅▅▃
3 tvhours           10146         0.528    2.98  2.59    0    1    2    4   24 ▅▂___
> |
```

Missingness

- Look for variables that have a lot of missing data:
  - tvhours
- Why is it missing? What should we do about it?

```
> skim(gss_cat)
── Data Summary ────────────────────────
                           Values
Name                       gss_cat
Number of rows             21483
Number of columns          9
_______________________
Column type frequency:
  factor                   6
  numeric                  3
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 marital               0             1 FALSE          6 Mar: 10117, Nev: 5416, Div: 3383, Wid: 1807
2 race                  0             1 FALSE          3 Whi: 16395, Bla: 3129, Oth: 1959, Not: 0
3 rincome               0             1 FALSE         16 $25: 7363, Not: 7043, $20: 1283, $10: 1168
4 partyid               0             1 FALSE         10 Ind: 4119, Not: 3690, Str: 3490, Not: 3032
5 relig                 0             1 FALSE         15 Pro: 10846, Cat: 5124, Non: 3523, Chr: 689
6 denom                 0             1 FALSE         30 Not: 10072, Oth: 2534, No : 1683, Sou: 1536

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean    sd   p0  p25  p50  p75 p100 hist
1 year                  0         1      2007.    4.45 2000 2002 2006 2010 2014 ▅▅▅▅
2 age                  76         0.996    47.2  17.3    18   33   46   59   89 ▅▅▅▃
3 tvhours           10146         0.528     2.98  2.59     0    1    2    4   24 ▅▂▁▁
>
```

Missingness

- Data for a variable can be missing for a variety of reasons:
  - the variable was included in some survey years, not others
  - only a subset of people answered the question based on some characteristic
  - missing at random
- In the first two cases, missingness is conveying information

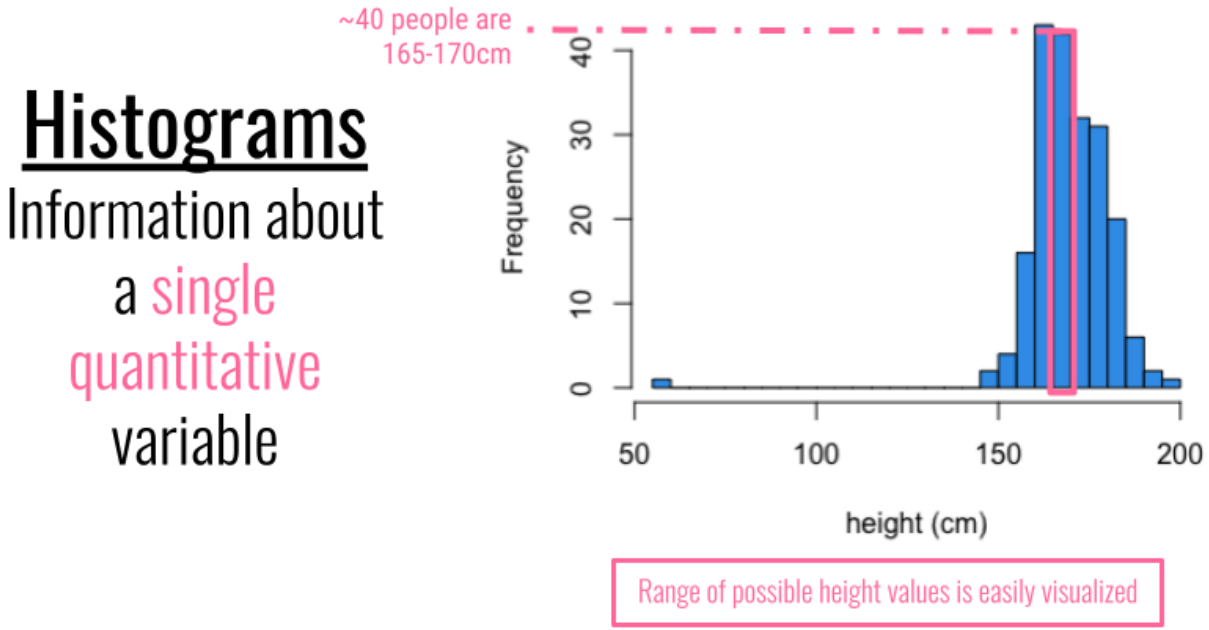
Example: Commute Times

- Examine the cases that have missing commute times and distance
- What is your student sample if you filter out cases where commute time is missing?

Live off campus?	commute time	distance from campus
TRUE	75	105
FALSE		
TRUE	45	54
FALSE		
TRUE	10	5
TRUE	5	65
TRUE	15	438
TRUE	60	6248

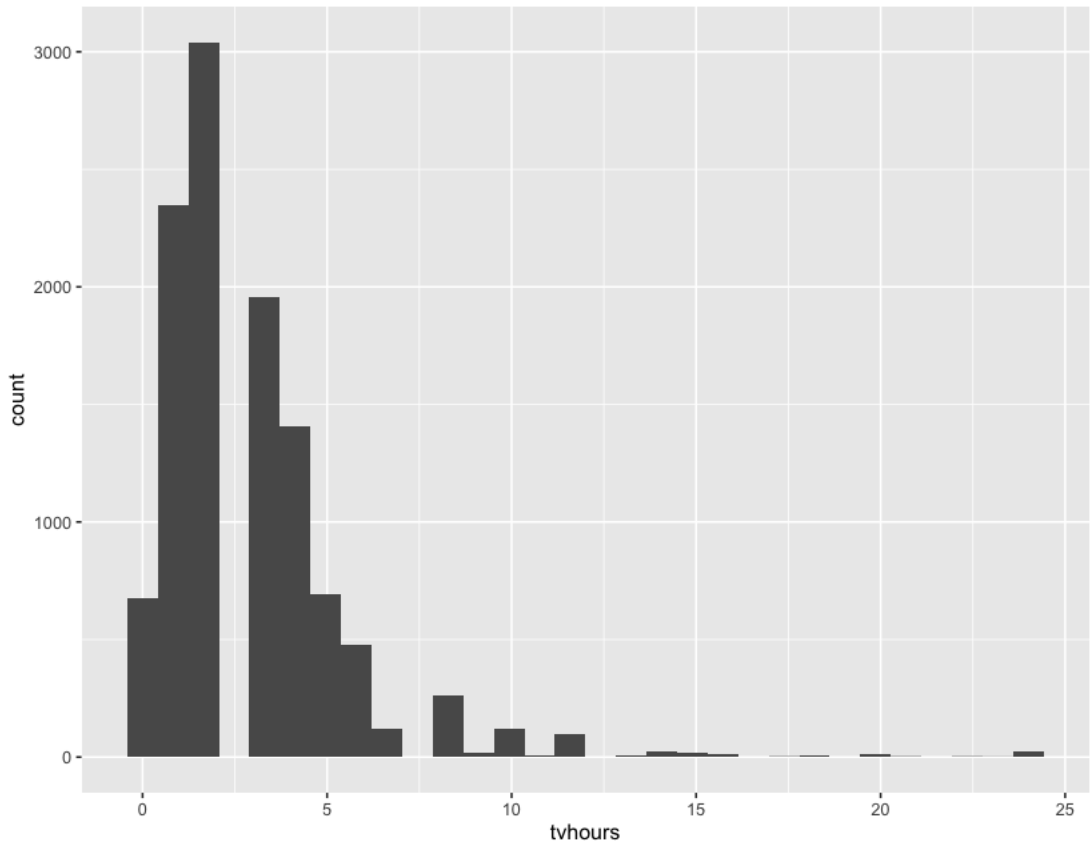
Summarizing a Quantitative Variable: Histograms

- Histograms are helpful when you want to understand what values you have in your dataset for a single variable.



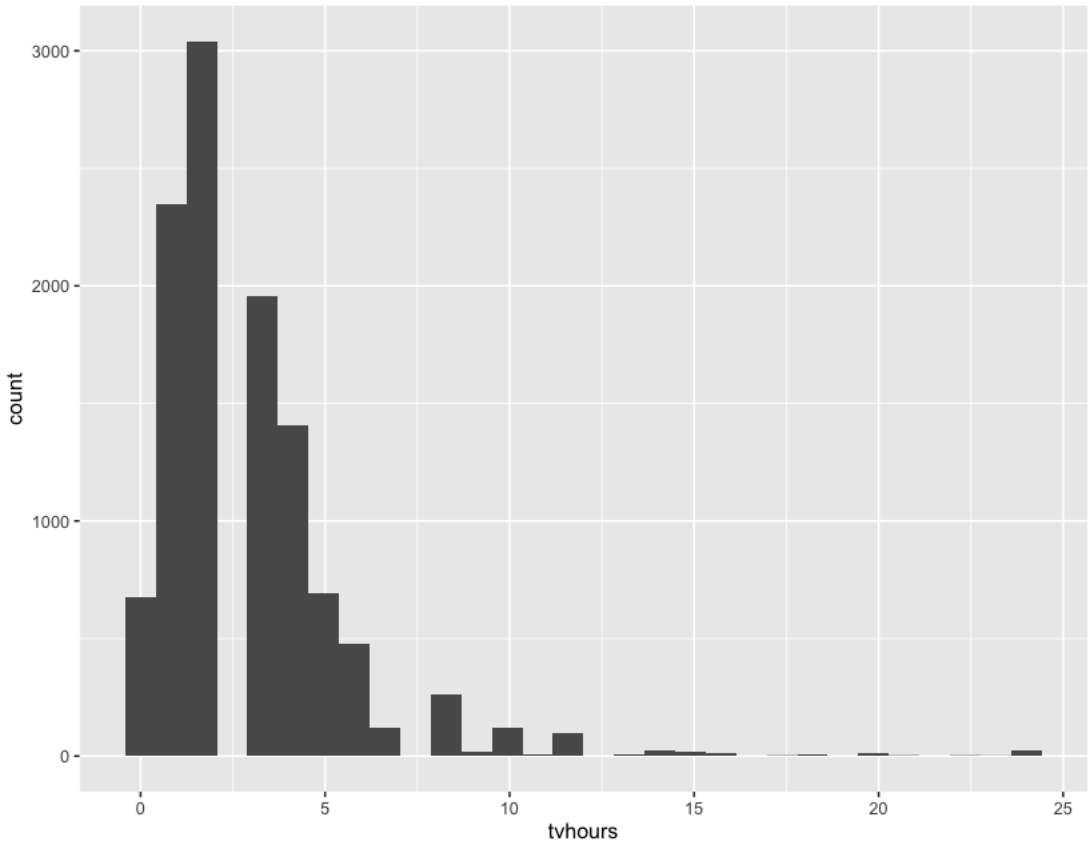
geom_histogram()

```
ggplot(data = gss_cat,
       mapping=aes(x=tvhours)) +
 geom_histogram()
```



Looks kind of funny

- Weird looking distribution
- Why did this happen?
  - tvhours are reported in increments of 1 (values range from 0-24)
  - ggplot default is to create 30 bins
- Result:
  - bins start at some negative value
  - bins are in increments of less than 1



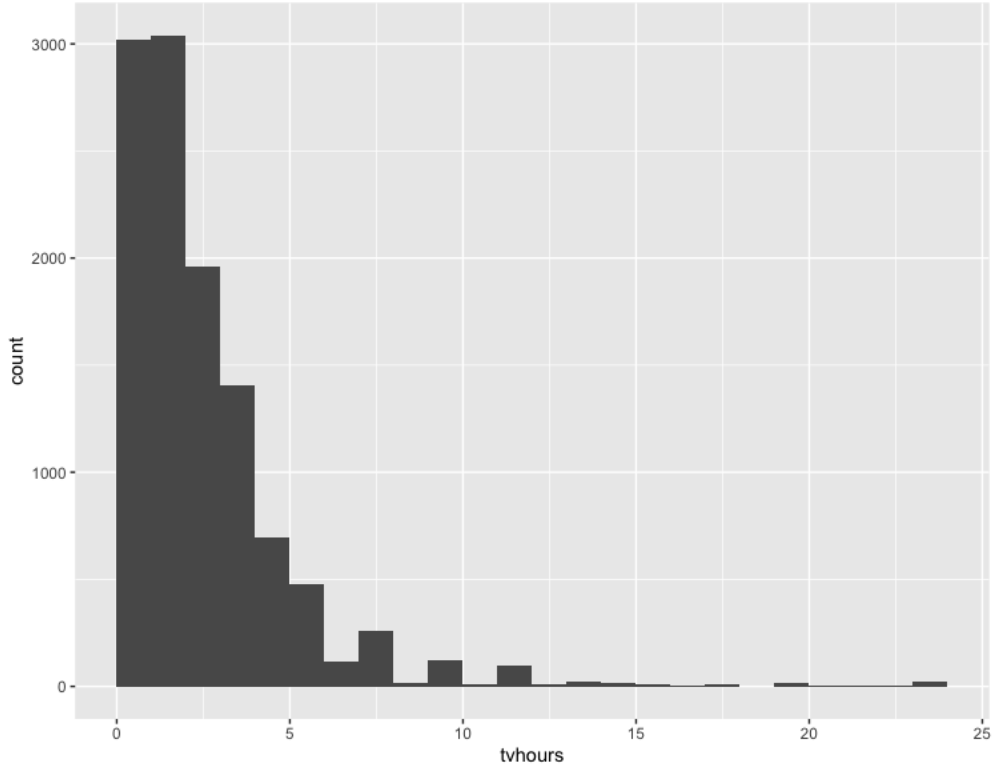
A Better Histogram

- Let's set the starting value (boundary)
- Let's also set the width of the bins (binwidth)

```
ggplot(data = gss_cat,
       mapping=aes(x=tvhours)) +
 geom_histogram(boundary=0, binwidth=1)
```

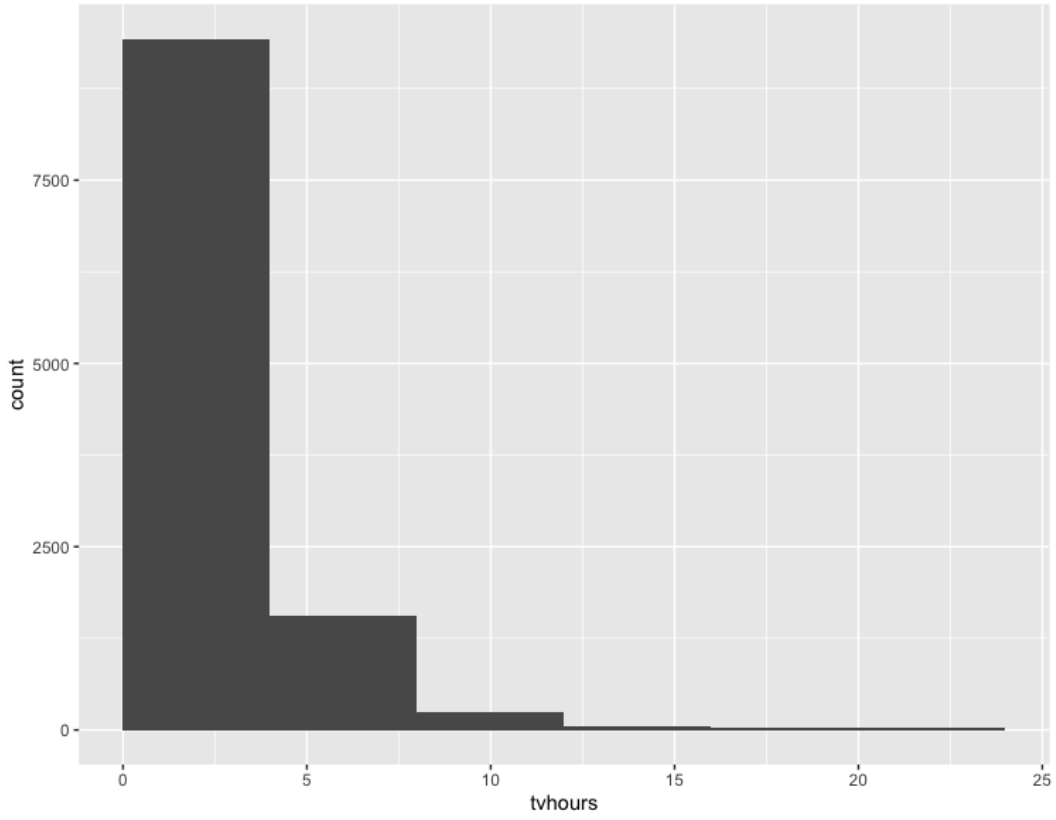
A Better Histogram

```
ggplot(data = gss_cat,
        mapping=aes(x=tvhours)) +
 geom_histogram(boundary=0, binwidth=1 )
```



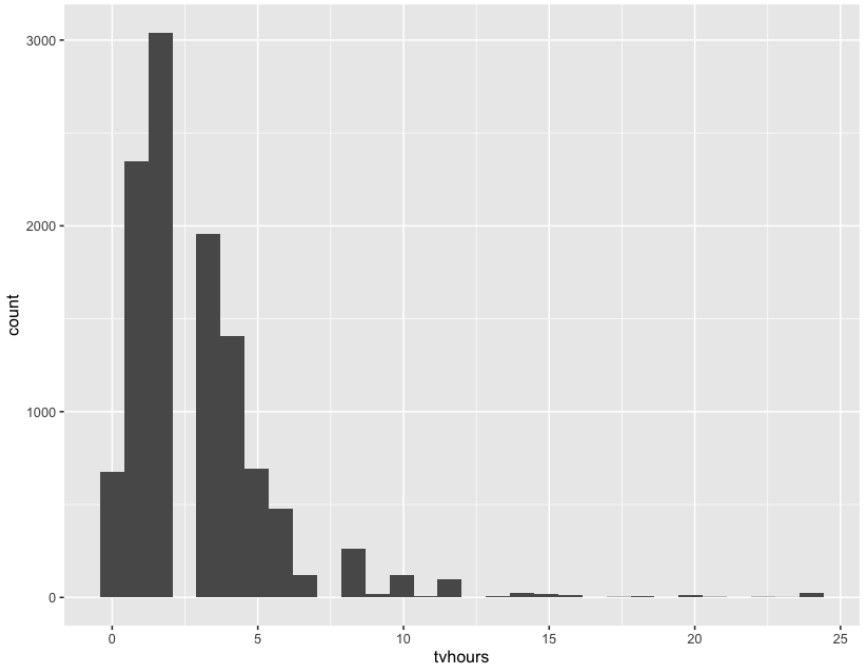
A Better Histogram (?)

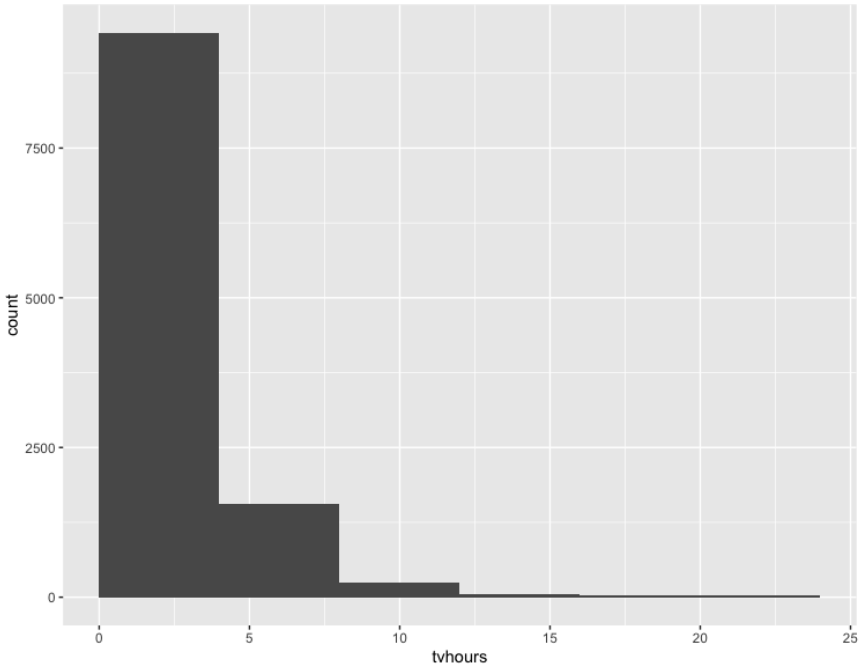
```
ggplot(data = gss_cat,
        mapping=aes(x=tvhours))+
 geom_histogram(boundary=0, binwidth=4 )
```



A Better Histogram (?)

- With histograms, it can be hard to decide what's the "correct" bin width
- when your bins get too narrow, it creates patterns that aren't really there
- when your bins get too wide, it erases the finer details of the distribution





Exercise

- Make a histogram for age, then adjust the bin width as needed

https://pollev.com/vsovero

Densityplot

- Densityplots are smoothed versions of histograms, visualizing the distribution of a continuous variable.
- Compared to histograms, they are less sensitive to the number of bins chosen for visualization.

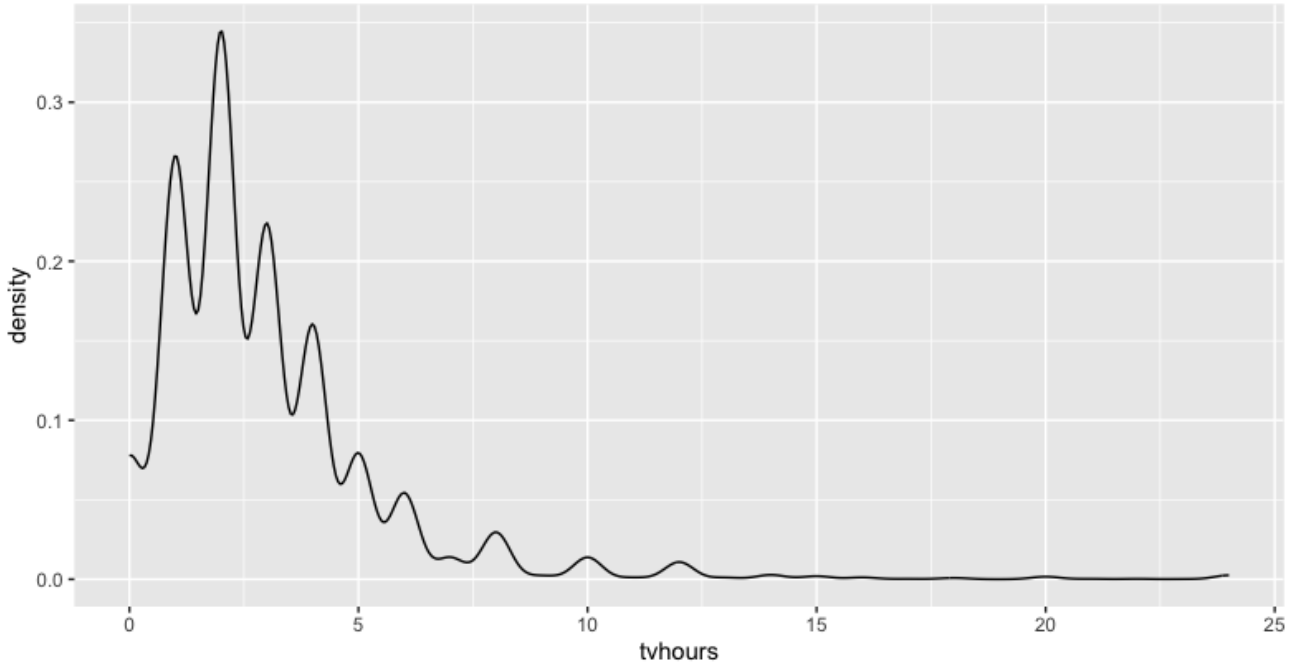
Densityplot

Information about a single quantitative variable

A smoothed version of a histogram - demonstrates the *distribution* of the data; helps to identify extreme values

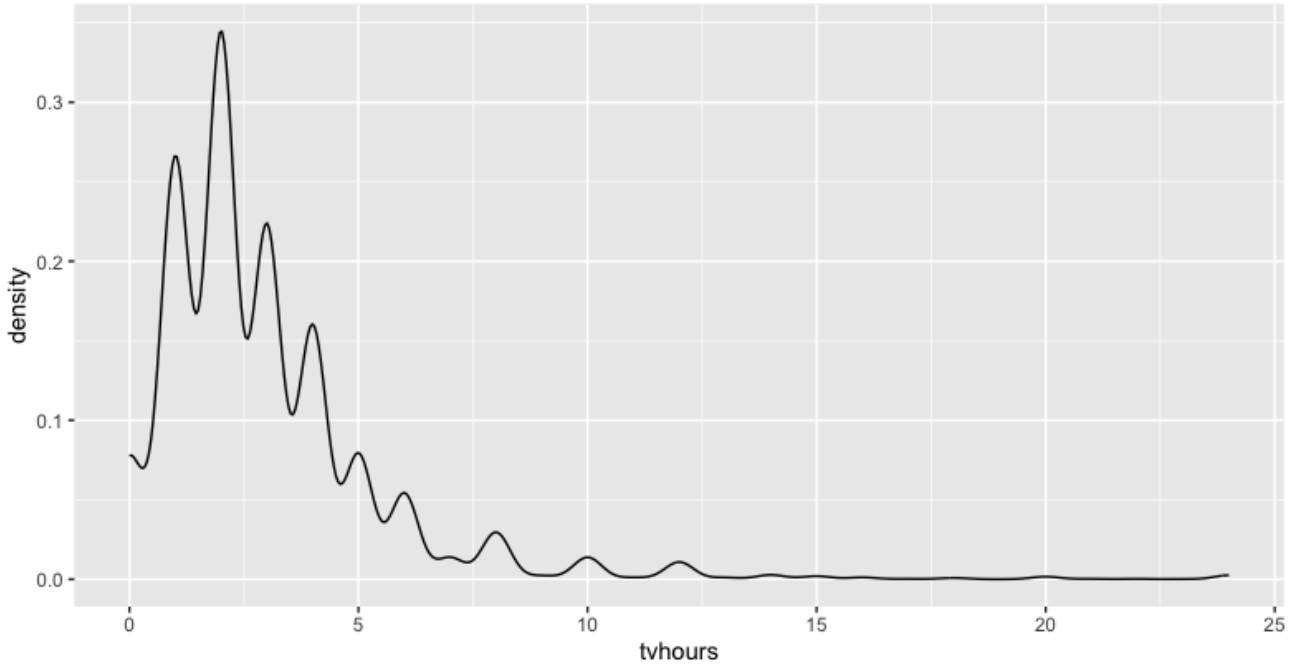
geom_density()

```
ggplot(data = gss_cat,
       mapping=aes(x=tvhours)) +
 geom_density()
```



Too Lumpy

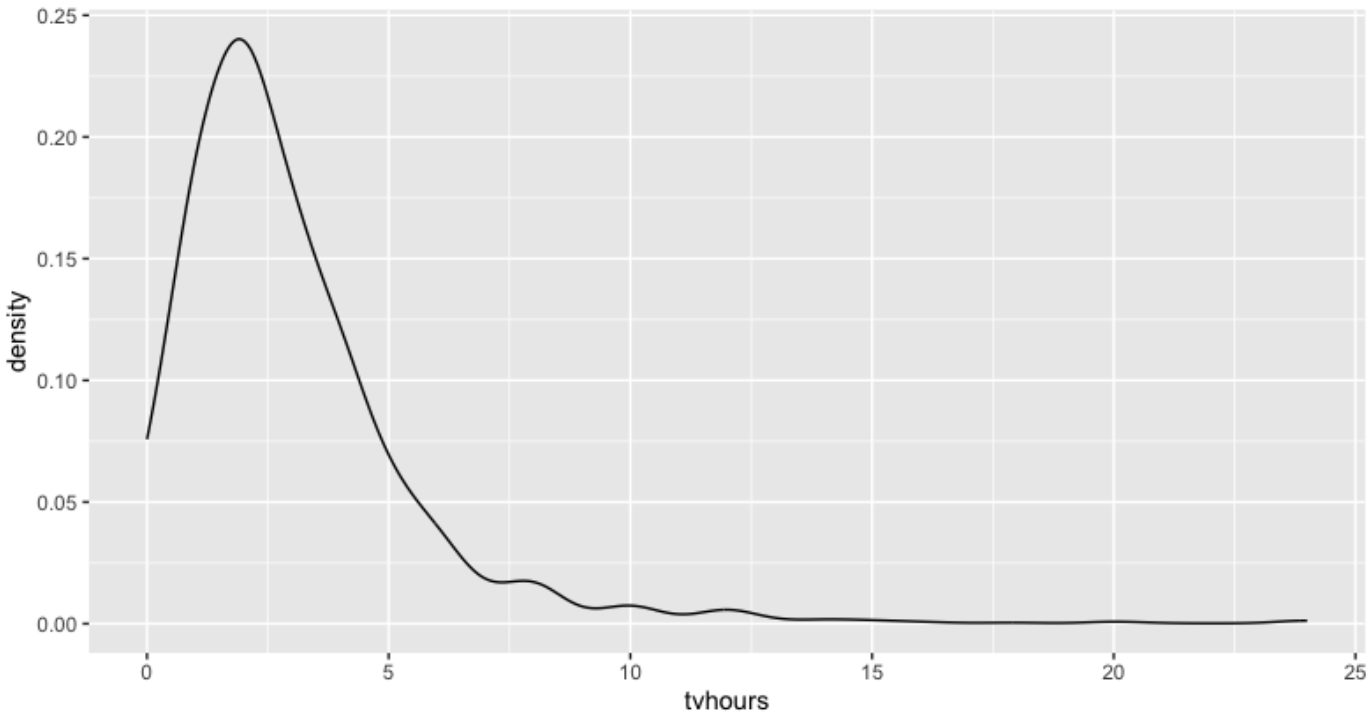
- The density plot looks “lumpy” because tvhours is an integer
- the lumps are at every integer (1,2,3,etc.)



Adjusting the bandwidth

Let's select a larger bandwidth (smooth out the bumps):

```
ggplot(data = gss_cat,
            mapping=aes(x=tvhours)) +
 geom_density(adjust=2)
```

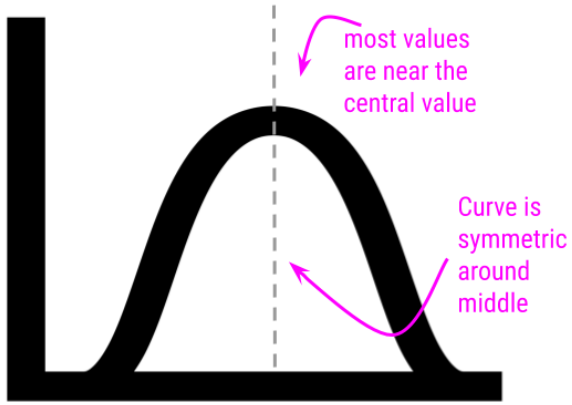


Exercise

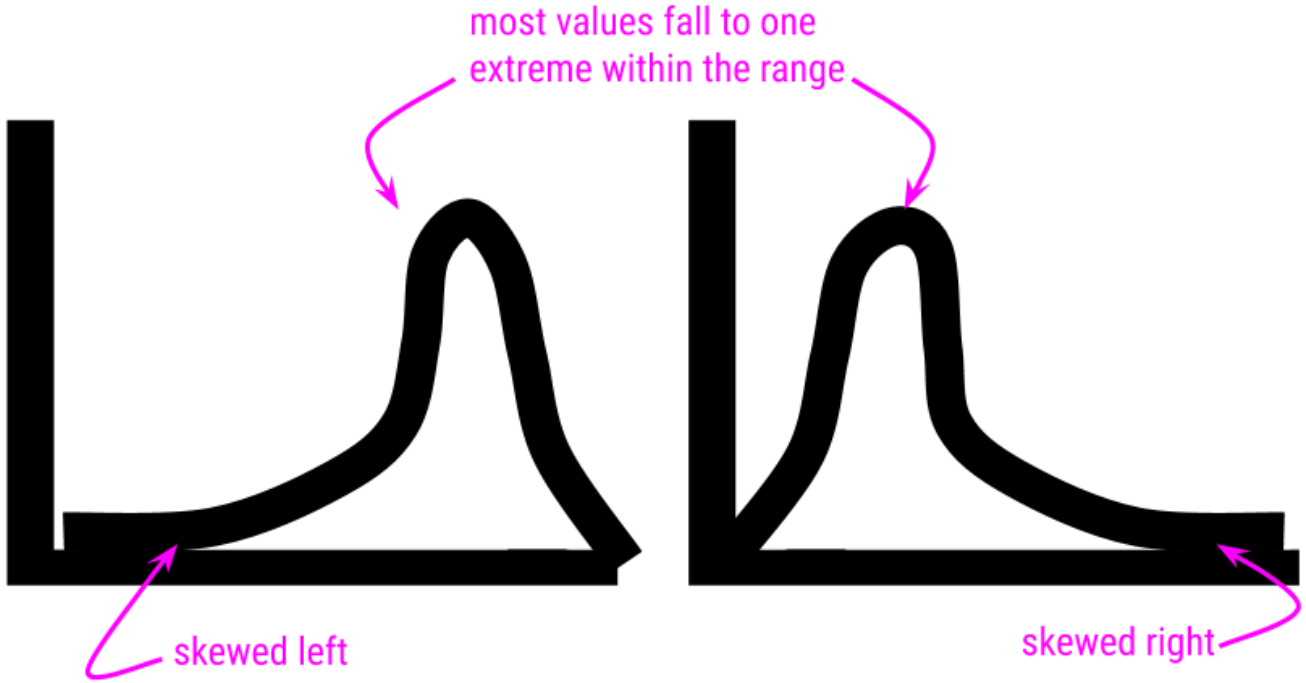
- Make a density plot for age, then adjust the bandwidth as needed

Describing the Shape of the Distribution

A Normal Distribution

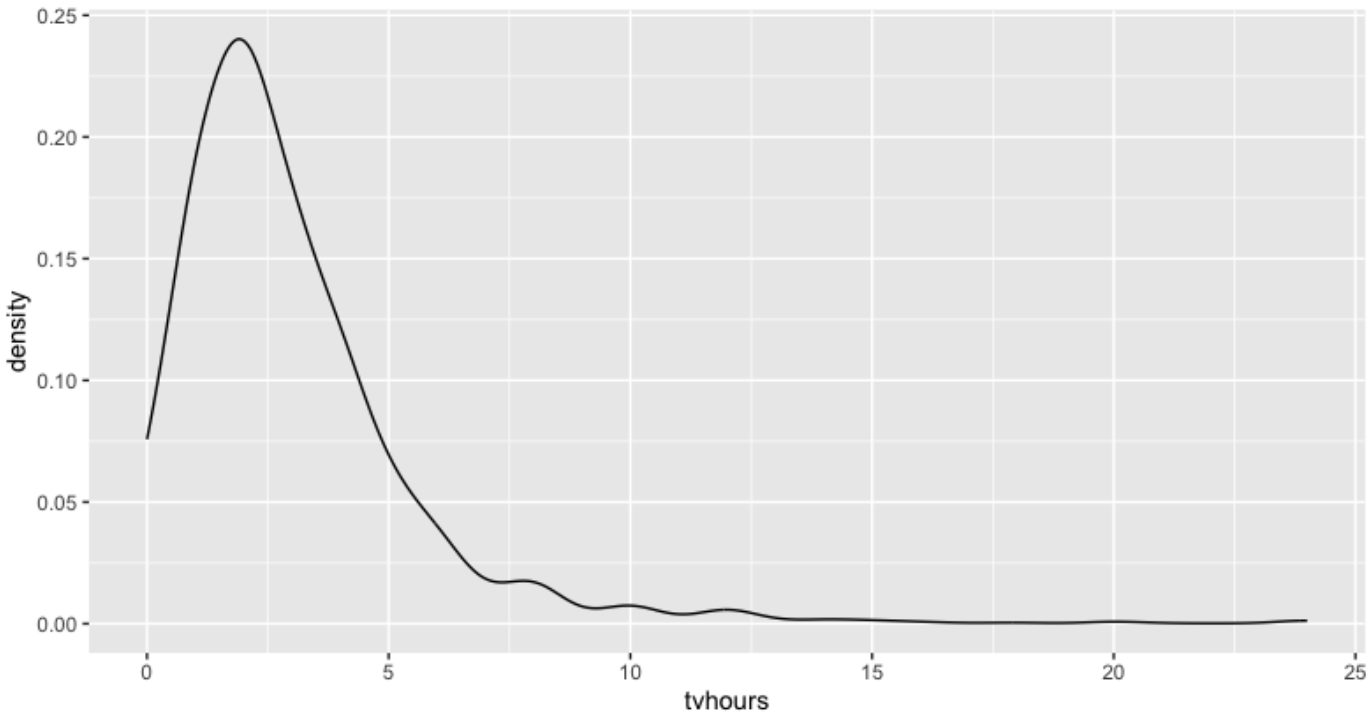


A Skewed Distribution



Tvhours Example

- Right-skewed distribution
- most people watch tv for between 0 and 5 hours per day
- very few watch 10 or more



Tvhours Example

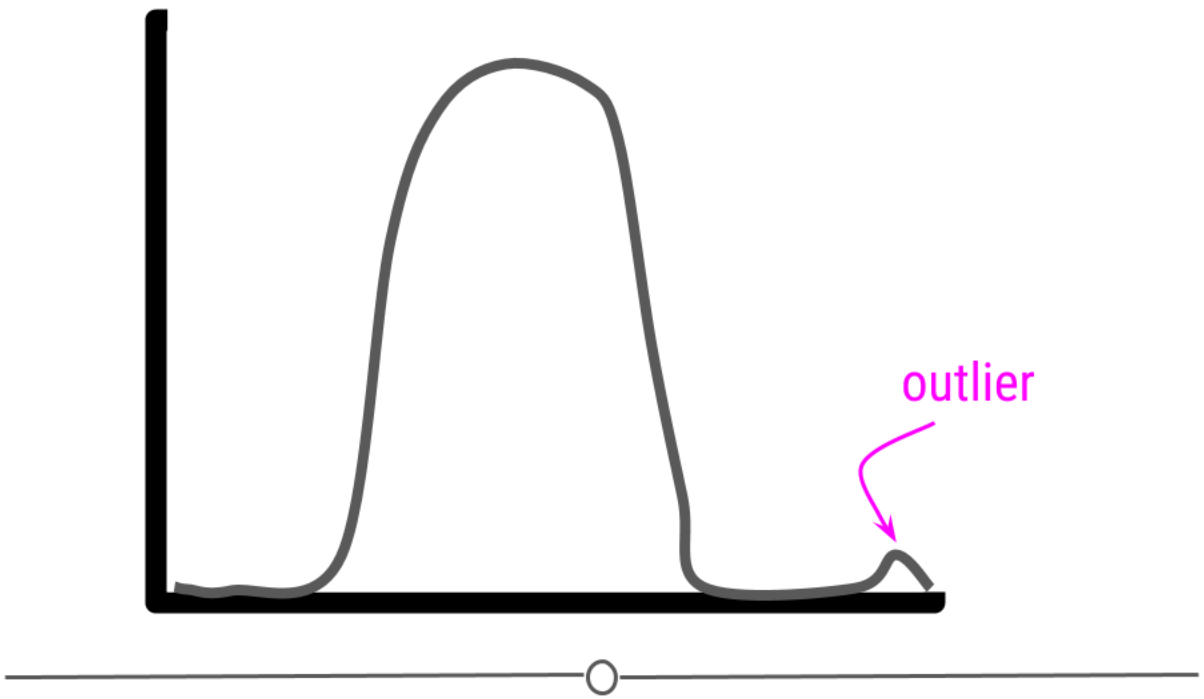
- Right-skewed distribution
- mean>median
- 75% of the sample watches less than 4 hours of tv per day

```
> skim(gss_cat)
── Data Summary ────────────────────────
                           Values
Name                       gss_cat
Number of rows             21483
Number of columns          9
_______________________
Column type frequency:
  factor                   6
  numeric                  3
_______________________
Group variables            None

── Variable type: factor ─────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 marital               0             1 FALSE          6 Mar: 10117, Nev: 5416, Div: 3383
2 race                  0             1 FALSE          3 Whi: 16395, Bla: 3129, Oth: 1959
3 rincome               0             1 FALSE         16 $25: 7363, Not: 7043, $20: 1283,
4 partyid               0             1 FALSE         10 Ind: 4119, Not: 3690, Str: 3490,
5 relig                 0             1 FALSE         15 Pro: 10846, Cat: 5124, Non: 3523
6 denom                 0             1 FALSE         30 Not: 10072, Oth: 2534, No : 1683

── Variable type: numeric ────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean    sd   p0  p25  p50  p75 p100 hist
1 year                  0         1     2007.    4.45 2000 2002 2006 2010 2014 ▅▃▅▃
2 age                  76         0.996   47.2  17.3    18   33   46   59   89 ▅▅▅▅▃▂▁
3 tvhours           10146         0.528    2.98  2.59    0    1    2    4   24 ▅▂▁▁▁▁▁
>
```

Outliers

- Plotting the distribution is also helpful for identifying any unusually large or small values (outliers)

- **Should we get rid of them? How do we decide?**



Extreme Values

- Values that are much larger or smaller than the rest of your distribution should be investigated until you are able to classify them into one of these categories:

1. They are erroneous (and should be excluded).
   - Example: missing values coded as 9999998
   - a value is measured in meters when it should be in cm
2. They are correct and produced by the same process as less extreme values (and should be retained).

What do we do about outliers? Mostly nothing.

🗣I don't know who needs to hear this but we ✨don't✨ get rid of outliers *because* they're extreme...

we get rid of them when their extremeness indicates they're not a part of the data generating process we want to study (like a typo that says your newborn is 1000 lbs)
— Chelsea Parlett-Pelleriti (@ChelseaParlett) February 1, 2021

https://pollev.com/vsovero

Categorical Variables: summary()

- Be careful: summary() doesn't show all the levels in a factor if you are summarizing the entire data frame

```
> summary(gss_cat)
      year                marital            age                      race                   rincome
 Min.   :2000   No answer    :   17   Min.   :18.00   Other         : 1959   $25000 or more:7363
 1st Qu.:2002   Never married: 5416   1st Qu.:33.00   Black         : 3129   Not applicable:7043
 Median :2006   Separated    :  743   Median :46.00   White         :16395   $20000 - 24999:1283
 Mean   :2007   Divorced     : 3383   Mean   :47.18   Not applicable:    0   $10000 - 14999:1168
 3rd Qu.:2010   Widowed      : 1807   3rd Qu.:59.00                           $15000 - 19999:1048
 Max.   :2014   Married      :10117   Max.   :89.00                           Refused       : 975
                                       NA's   :76                              (Other)       :2603
                partyid            relig                     denom          tvhours
 Independent       :4119   Protestant:10846   Not applicable  :10072   Min.   : 0.000
 Not str democrat  :3690   Catholic  : 5124   Other           : 2534   1st Qu.: 1.000
 Strong democrat   :3490   None      : 3523   No denomination : 1683   Median : 2.000
 Not str republican:3032   Christian :  689   Southern baptist: 1536   Mean   : 2.981
 Ind,near dem      :2499   Jewish    :  388   Baptist-dk which: 1457   3rd Qu.: 4.000
 Strong republican :2314   Other     :  224   United methodist: 1067   Max.   :24.000
 (Other)           :2339   (Other)   :  689   (Other)         : 3134   NA's   :10146
>
```

Categorical Variables: summary()

- summary() will show all the levels in a factor if you summarize a single variable

```
summary(gss_cat$relig)
          No answer            Don't know Inter-nondenominational        Native american
                 93                    15                     109                     23
          Christian     Orthodox-christian           Moslem/islam          Other eastern
                689                    95                     104                     32
           Hinduism               Buddhism                   Other                   None
                 71                   147                     224                   3523
             Jewish               Catholic              Protestant          Not applicable
                388                  5124                   10846                      0
```

Summaries with skimr

- With skim() you won't see all of the levels of a factor variable, but you will at least know how many levels there are

```
> skim(gss_cat)
── Data Summary ────────────────────────
                           Values
Name                       gss_cat
Number of rows             21483
Number of columns          9
_______________________
Column type frequency:
  factor                   6
  numeric                  3
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 marital               0             1 FALSE         6 Mar: 10117, Nev: 5416, Div: 3383, Wid: 1807
2 race                  0             1 FALSE         3 Whi: 16395, Bla: 3129, Oth: 1959, Not: 0
3 rincome               0             1 FALSE        16 $25: 7363, Not: 7043, $20: 1283, $10: 1168
4 partyid               0             1 FALSE        10 Ind: 4119, Not: 3690, Str: 3490, Not: 3032
5 relig                 0             1 FALSE        15 Pro: 10846, Cat: 5124, Non: 3523, Chr: 689
6 denom                 0             1 FALSE        30 Not: 10072, Oth: 2534, No : 1683, Sou: 1536

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean    sd   p0  p25  p50  p75 p100 hist
1 year                  0         1     2007.    4.45 2000 2002 2006 2010 2014 ▅▅▅▅
2 age                  76         0.996   47.2  17.3    18   33   46   59   89 ▅▅▅▅▂
3 tvhours           10146         0.528    2.98  2.59    0    1    2    4   24 ▅▂___
>
```

Best Option: dplyr

```
marital_count<-gss_cat%>%
                count(marital)%>%
                arrange(desc(n))
```

	marital	n
1	Married	10117
2	Never married	5416
3	Divorced	3383
4	Widowed	1807
5	Separated	743
6	No answer	17

```
marital_summarize<-gss_cat%>%
  group_by(marital)%>%
  summarize(freq=n())%>%
  arrange(desc(freq))
```

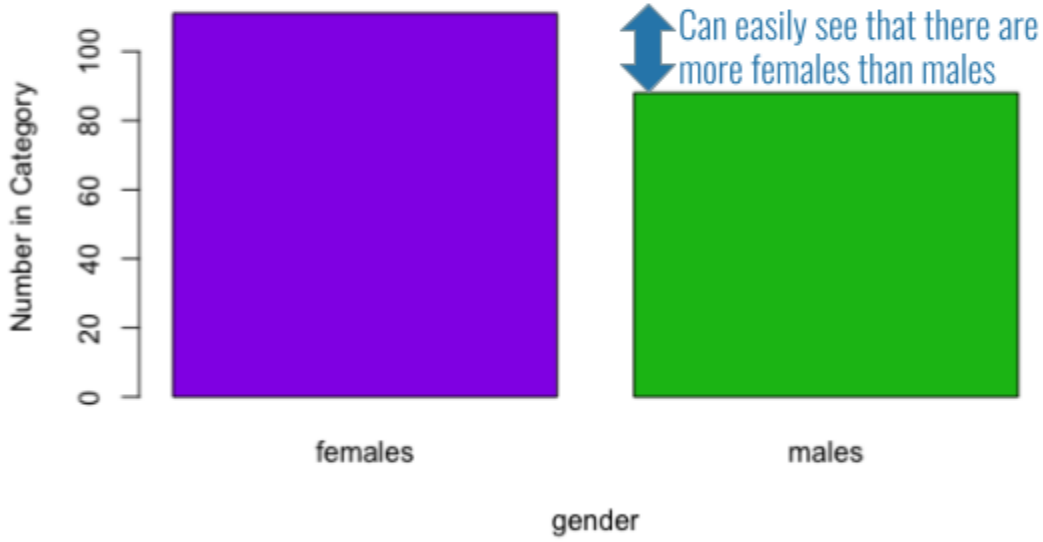
	marital	freq
1	Married	10117
2	Never married	5416
3	Divorced	3383
4	Widowed	1807
5	Separated	743
6	No answer	17

Exercise

- Create a frequency table for relig. Sort the levels from most to least frequent

Summarizing a Categorical Variable

- There are actually many options, but the most common is the bar plot

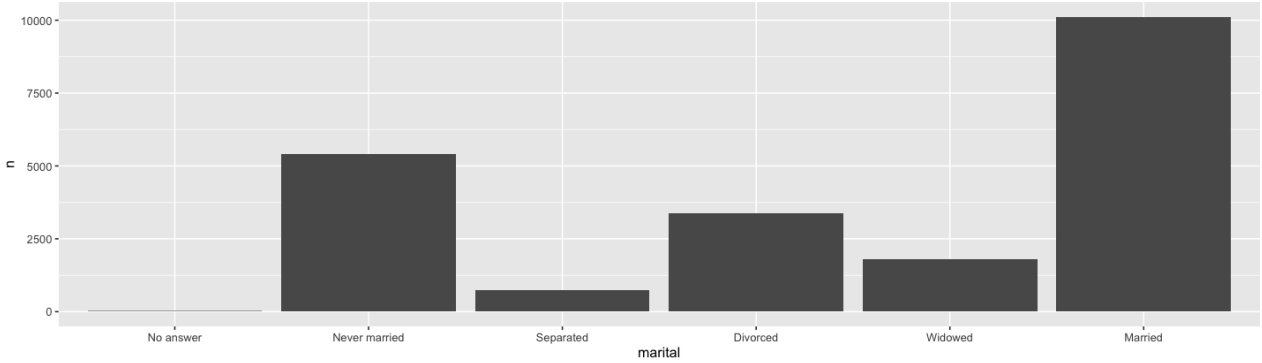


Bar plot from a Frequency Table

- Bar plot: reports frequencies of each level
- If you already created a frequency table (marital_count), use geom_col()

```
ggplot(data = marital_count,
        mapping=aes(x=marital, y=n )) +
 geom_col()
```

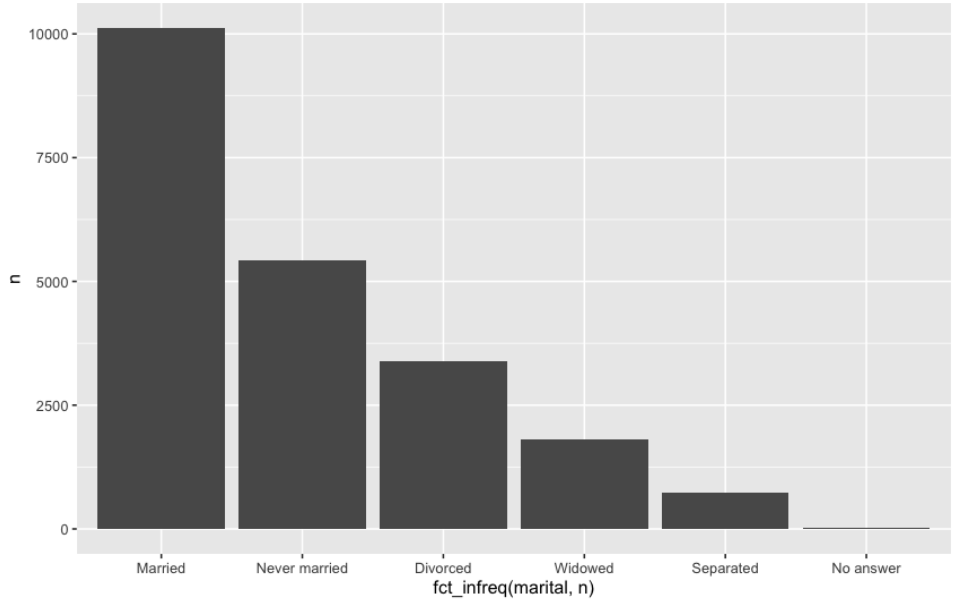
	marital	n
1	Married	10117
2	Never married	5416
3	Divorced	3383
4	Widowed	1807
5	Separated	743
6	No answer	17



Bar plot ordered by frequency

- If we want to sort the bar plot by frequency, we can use fct_infreq()
- Arguments:
  - the name of the factor variable
  - The variable that counts the frequencies (only when your data is a frequency table)

```
ggplot(data = marital_count,
        mapping=aes(x= fct_infreq(marital,n), y=n ))+
geom_col()
```



Exercise

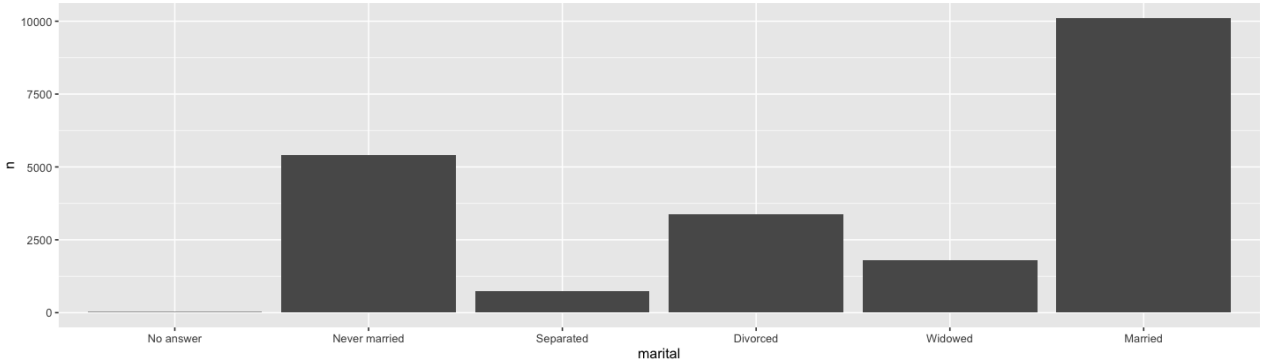
- Create a bar chart for relig using geom_col().
- Display the levels by frequency.

https://pollev.com/vsovero

Bar plot from the original data

- Bar plot: reports frequencies of each level
- If you didn't create a frequency table, use geom_bar() with the original data frame (gss_cat)

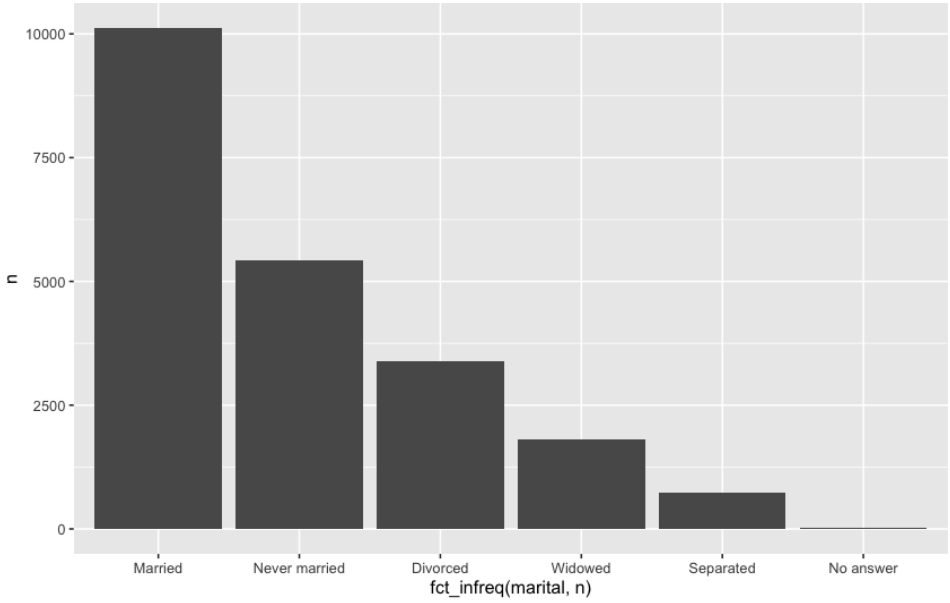
```
ggplot(data = gss_cat,
        mapping=aes(x=marital)) +
 geom_bar()
```



Bar plot ordered by frequency

- If we want to sort the bar plot by frequency, we can use fct_infreq()
- Arguments:
  - the name of the factor variable

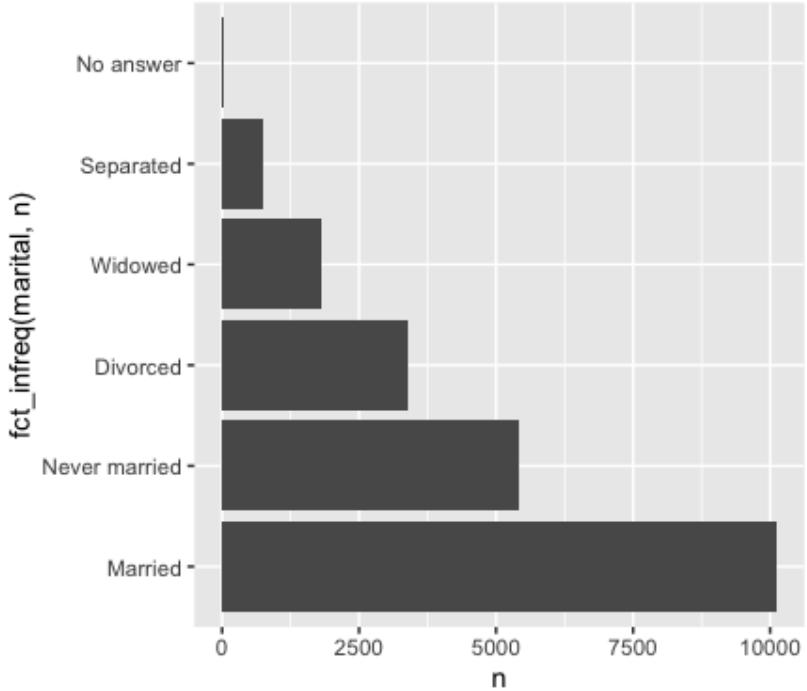
```
ggplot(data = gss_cat,
        mapping=aes(x= fct_infreq(marital) ))+
 geom_bar()
```



Horizontal Bar plot ordered by frequency

- When there are a lot of levels, it looks better to flip the direction of your bar plot
- we can use coord_flip()
- note that we add this to our ggplot using the + operator

```
ggplot(data = gss_cat,
        mapping=aes(x= fct_infreq(marital)))+
 geom_col() +
 coord_flip()
```



Exercise

- #Create a bar chart for relig using geom_bar().
- Display the levels by frequency.
- Rotate the bars so they're horizontal.