Econ 106: Data Analysis for Economics

Lecture 6
Fall 2024

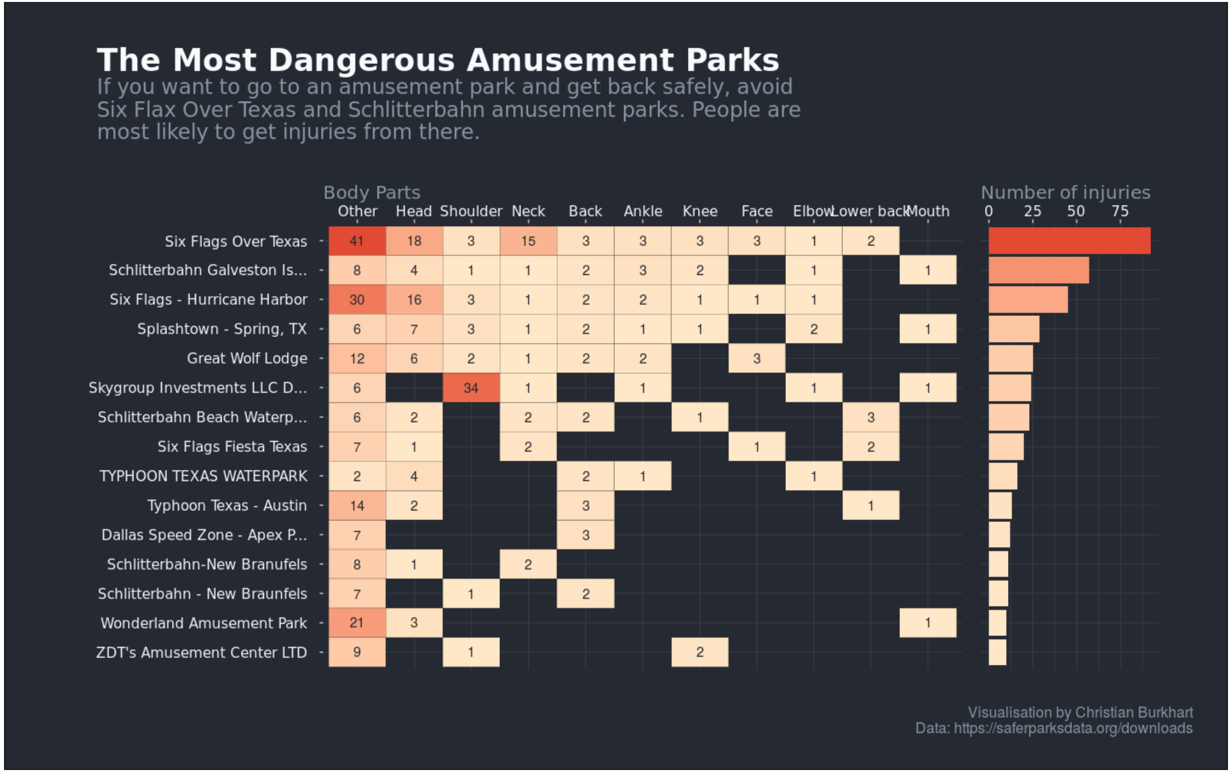
slides adapted from https://stats.oarc.ucla.edu/r/seminars/ggplot2_intro/

Reminders

- Lab 1 is due Sunday, 11:59pm (Q3 was updated)

https://pollev.com/vsovero

#tidytuesday



The Most Dangerous Amusement Parks

If you want to go to an amusement park and get back safely, avoid Six Flax Over Texas and Schlitterbahn amusement parks. People are most likely to get injuries from there.

Body Parts

	Other	Head	Shoulder	Neck	Back	Ankle	Knee	Face	Elbow	Lower back	Mouth
Six Flags Over Texas	41	18	3	15	3	3	3	3	1	2	
Schlitterbahn Galveston Is...	8	4	1	1	2	3	2		1		1
Six Flags - Hurricane Harbor	30	16	3	1	2	2	1	1	1		
Splashtown - Spring, TX	6	7	3	1	2	1	1		2		1
Great Wolf Lodge	12	6	2	1	2	2		3			
Skygroup Investments LLC D...	6		34	1		1			1		1
Schlitterbahn Beach Waterp...	6	2		2	2		1			3	
Six Flags Fiesta Texas	7	1		2				1		2	
TYPHOON TEXAS WATERPARK	2	4			2	1			1		
Typhoon Texas - Austin	14	2			3					1	
Dallas Speed Zone - Apex P...	7				3						
Schlitterbahn-New Branufels	8	1		2							
Schlitterbahn - New Braunfels	7		1		2						
Wonderland Amusement Park	21	3									1
ZDT's Amusement Center LTD	9		1				2				

Visualisation by Christian Burkhart
Data: https://saferparksdata.org/downloads

https://x.com/ChBurkhart/status/1173646158339727361?s=20

Outline

- Color setting vs mapping
- Bar graphs

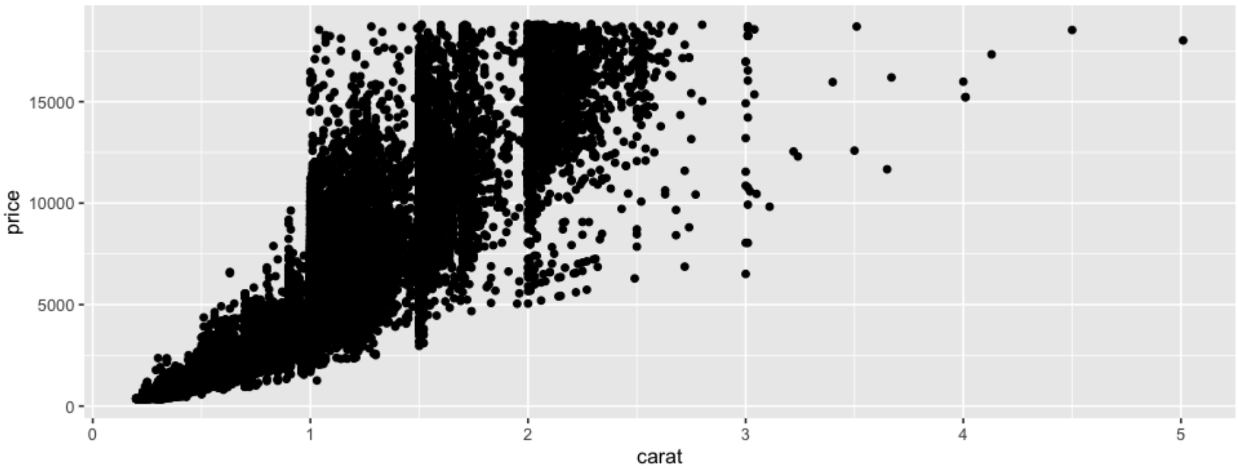
Basic Elements of ggplot

1. **Data**: the data you want to plot

2. **Layout**: mapping variables on the plot

3. **Data display**: how you want the data to be visualized (points, lines, bars, etc.)

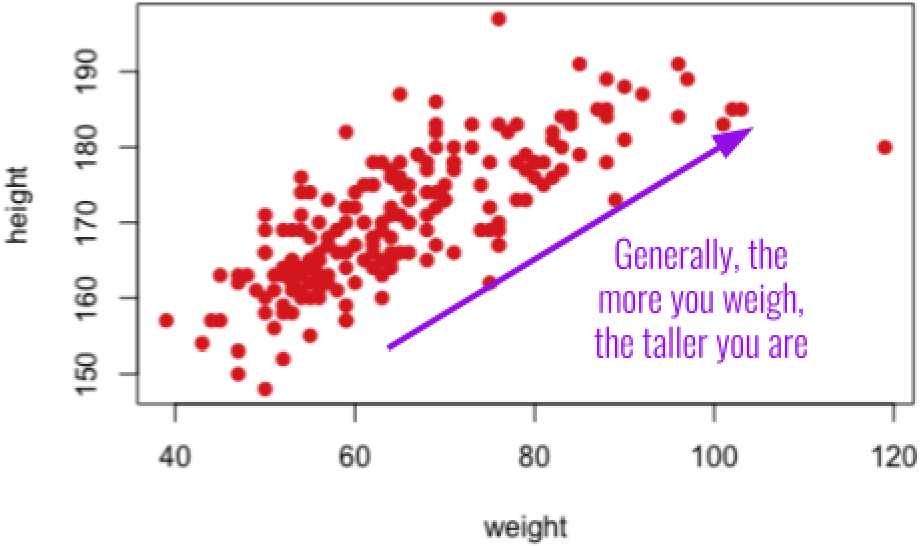
```
ggplot(data = diamonds,
          mapping = aes(x = carat,  y = price)) +
geom_point()
```



Scatter plots

- Which geom?
  geom_point()

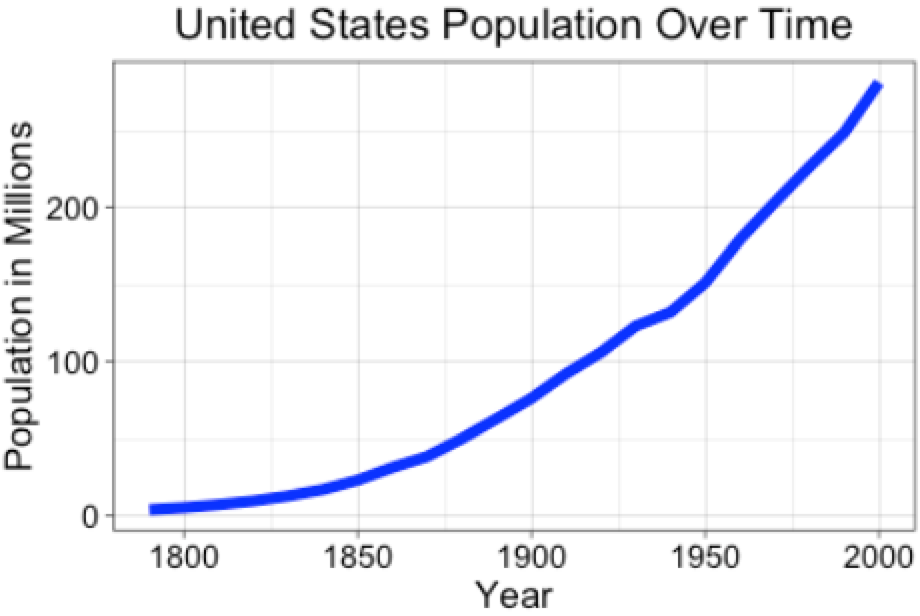
- What type of data?
  Two numeric variables



Line graphs

- Which geom? geom_line()

- What type of data? a date/time variable and a numeric variable



https://pollev.com/vsovero

Adjusting Plot Settings

- **color**: color of 1-d objects
- **fill**: fill color of 2-d objects
- **linetype**: how lines should be drawn (solid, dashed, dotted, etc.)
- **shape**: shape of markers in scatter plots
- **size**: how large objects appear
- **alpha**: transparency of objects (value between 0 and 1)

Color Mapping

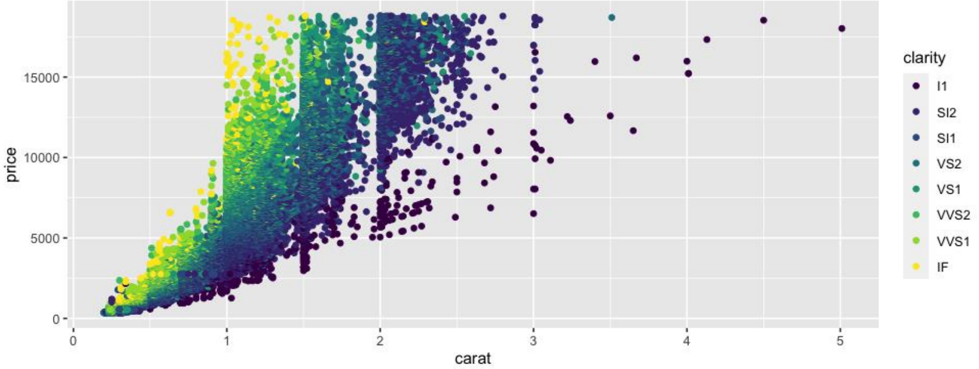
- Color points by clarity

- **Input**: color = categorical variable

- Remember, anything that references variables in the dataset must be inside aes():

- **Output**: a colored plot by clarity

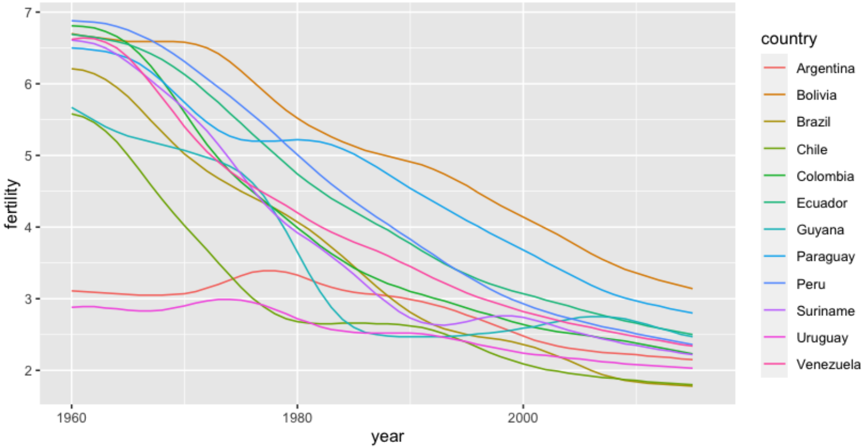
```
ggplot(data = diamonds,
          mapping= aes(x = carat,  y = price)) +
geom_point(aes(color = clarity))
```



Color Mapping (line graph)

- when we use the color argument for a line graph, it will:
- create a separate line for each country
- assign each country a unique color

```
gapminder %>%
    filter(region=="South America") %>%
ggplot(mapping=aes(x=year, y=fertility)) +
 geom_line(aes(color=country))
```



Color: Setting vs. Mapping

- **Color Setting**: color is a fixed value
- **Set** aesthetics to a constant *outside* the **aes**() function.

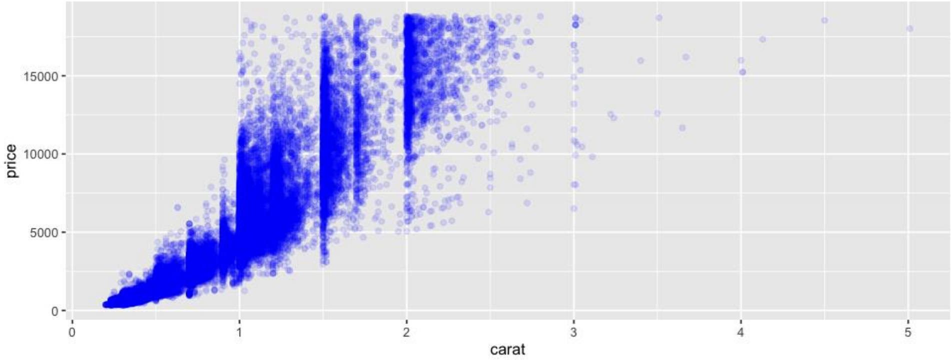
- **Color Mapping:** color will vary based on the value of a variable
- **Map** aesthetics to variables *inside* the **aes**() function

Setting vs. Mapping

https://pollev.com/vsovero

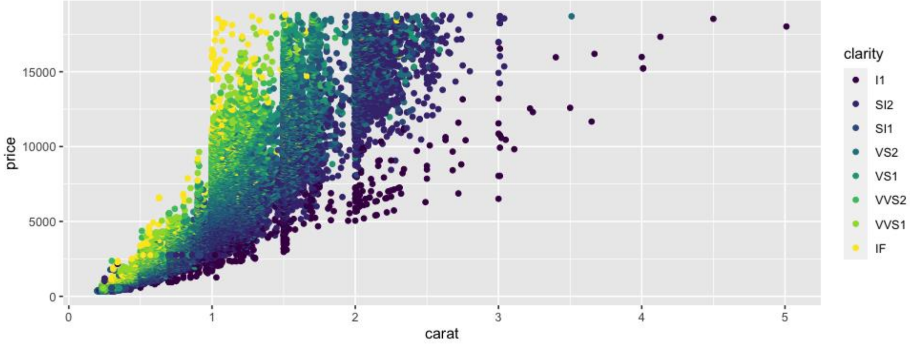
The color is set to blue (color setting):

```
ggplot(data = diamonds,
                mapping= aes(x = carat, y = price))+
geom_point(color = "blue")
```



Color is mapped to the clarity variable (color mapping):

```
ggplot(data = diamonds,
                mapping= aes(x = carat, y = price))+
geom_point(color = aes(clarity))
```



Exercise

- use the jobs_gender data frame to create a scatter plot of total_earnings on the x-axis and wage_percent_of_male on the y-axis, color mapping by major_category

Aesthetic Mappings

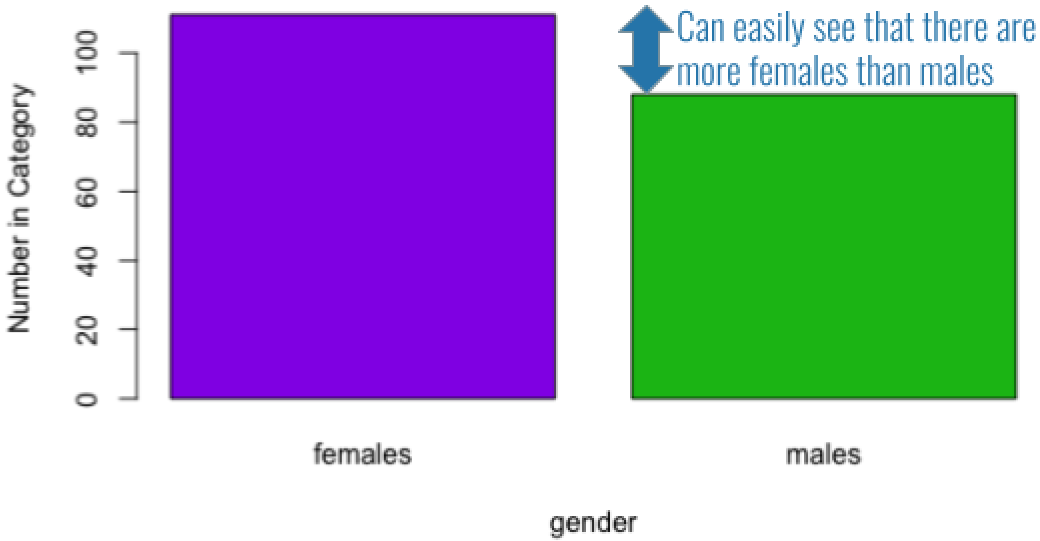
- Quantitative and Categorical variables work for:
  - color and fill: color gradient scales or evenly-spaced hue scales
- Only categorical variables work for:
  - shape
  - linetype
- Your code will run, but you really should only use quantitative variables for:
  - size
  - alpha

Exercise

- make a scatter plot of:
  - total_earnings on the x-axis and wage_percent_of_male on the y-axis
  - map total_employees to size
  - alpha of .2

Bar plot

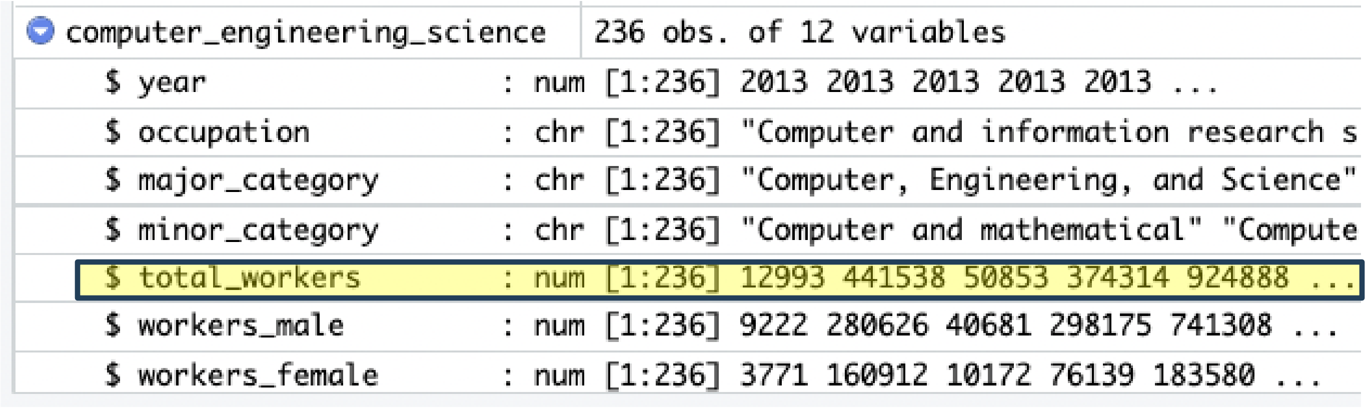
- Counting frequencies of a single categorical variable



Can easily see that there are more females than males

Bar plot (frequency counts are in the data)

```
computer_engineering_science <- jobs_gender %>%
   filter(major_category=="Computer, Engineering, and
Science")
```

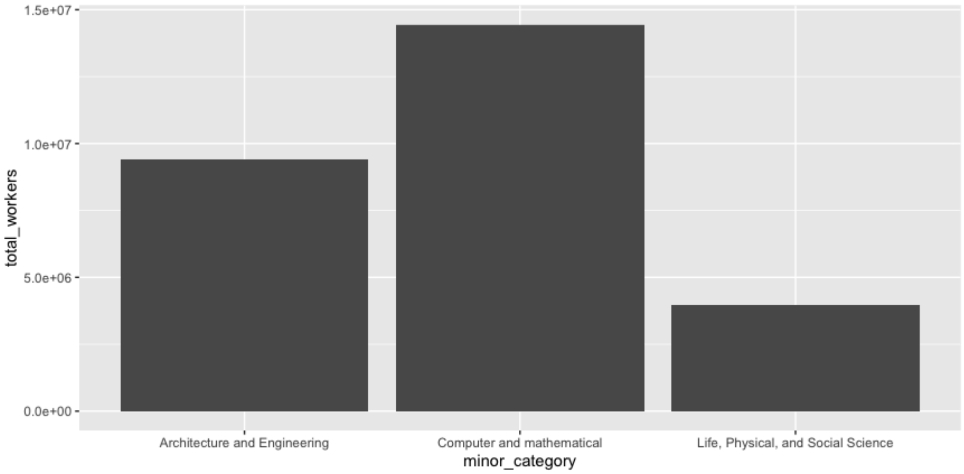


```
computer_engineering_science    236 obs. of 12 variables
    $ year              : num [1:236] 2013 2013 2013 2013 2013 ...
    $ occupation        : chr [1:236] "Computer and information research s
    $ major_category    : chr [1:236] "Computer, Engineering, and Science"
    $ minor_category    : chr [1:236] "Computer and mathematical" "Compute
    $ total_workers     : num [1:236] 12993 441538 50853 374314 924888 ...
    $ workers_male      : num [1:236] 9222 280626 40681 298175 741308 ...
    $ workers_female    : num [1:236] 3771 160912 10172 76139 183580 ...
```

Bar plot (frequency counts are in the data)

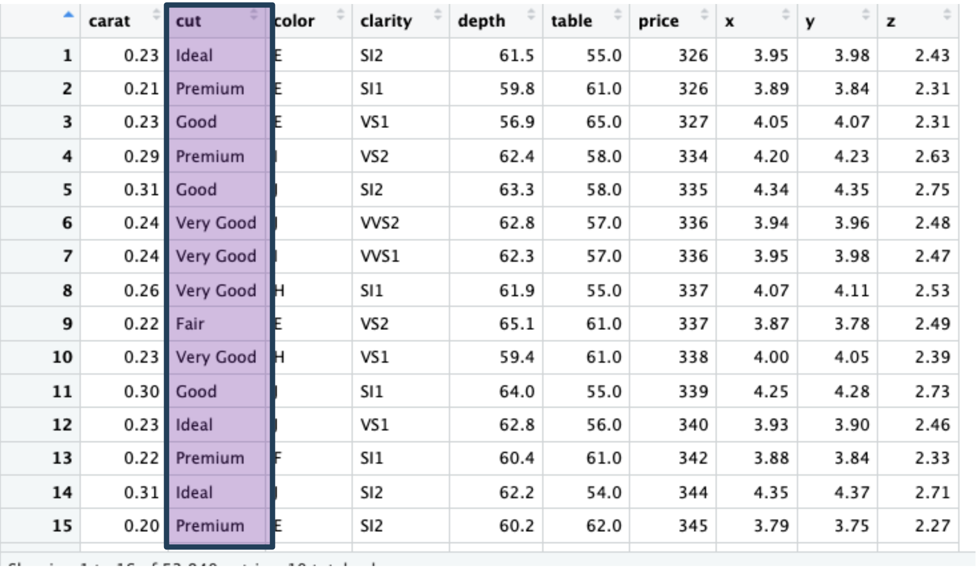
- **geom_col**() is adding up the total workers for each value of minor_category

```
ggplot(data = computer_engineering_science,
                 mapping=aes(x=minor_category, y=total_workers))+
 geom_col()
```



Bar plot (counts are not in the data)

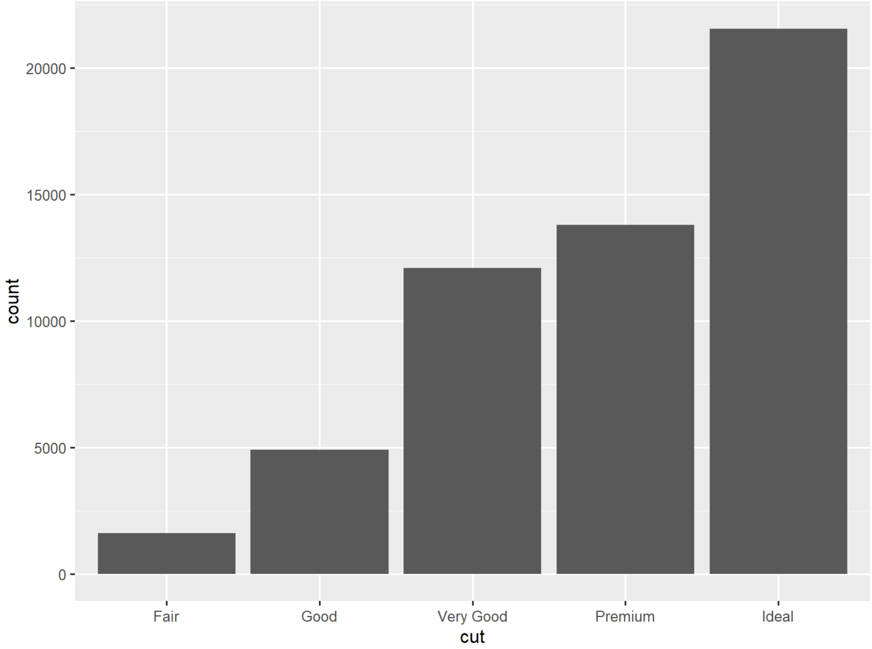
- in this dataset, there is no variable that counts the frequency of each cut
- we have to ask ggplot to count for us



	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27

Bar plot (ggplot counts the frequencies)

```
ggplot(data = diamonds,
          mapping=aes(x=cut))+
 geom_bar()
```



https://pollev.com/vsovero

Data Example

https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-09-10

- we will use the safer_parks data from this tidy Tuesday challenge (code to load the data is in the lecture script)

Exercise

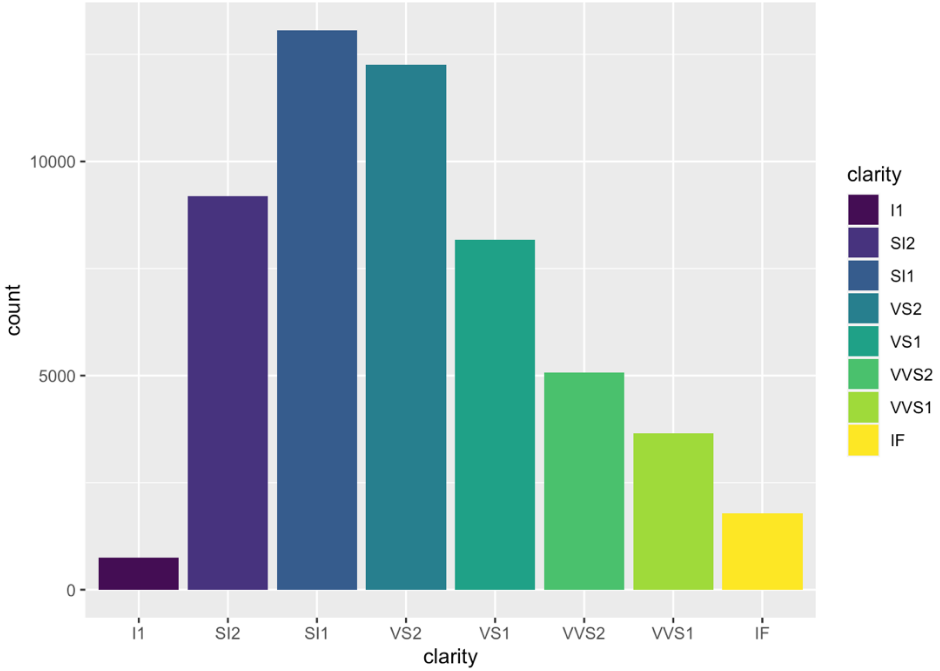
- create a bar graph that shows the frequencies of each value of industry sector

Color Mapping with **geom_bar**()

- How do we color map a bar plot?
- Same idea as before- has to go within the **aes**() function
- use **fill** option to color the entire bar

Color mapping with **fill**

```
ggplot(data=diamonds,
            mapping= aes(x=clarity)) +
 geom_bar(aes(fill=clarity))
```



Exercise

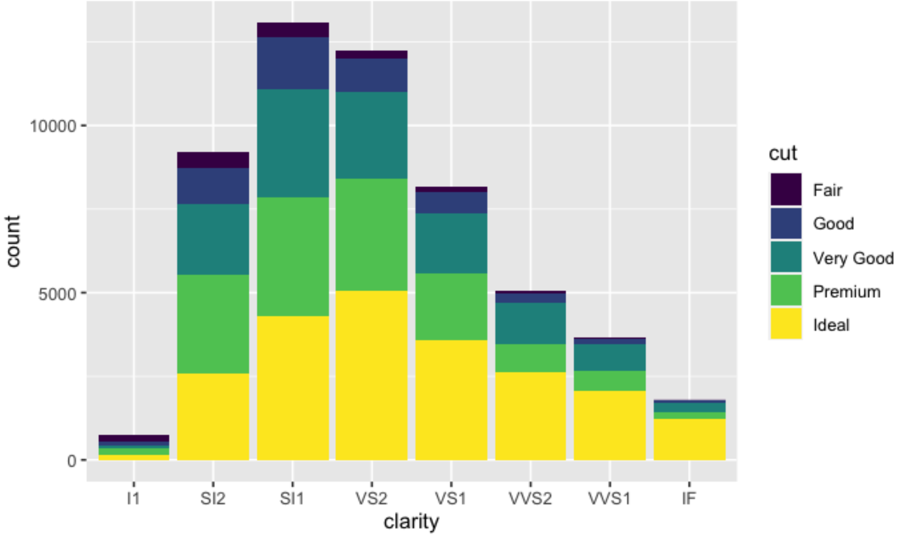
- create a bar graph that shows the frequencies of each value of industry sector
- color the bars by industry sector

https://pollev.com/vsovero

Stacked barplot (counts)

- The counts of clarity are broken down further by cut

```
ggplot(data=diamonds,
                    mapping= aes(x=clarity)) +
 geom_bar(aes(fill=cut))
```



Exercise

- filter for amusement rides
- only keep injury reports where the gender reported is male or female
- create a bar graph that shows the frequencies of each value of device category
- within each device category, show the gender counts

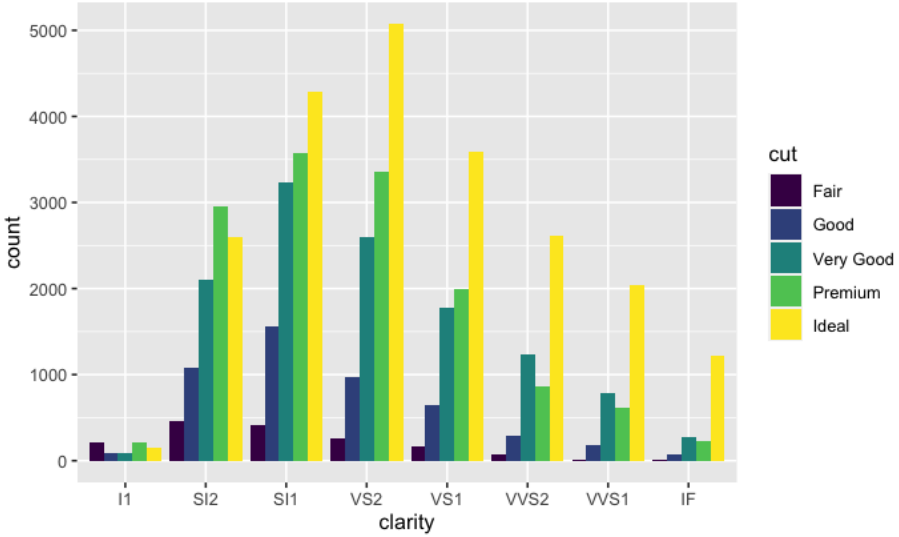
Exercise

- filter for amusement rides
- only keep injury reports where the gender reported is male or female
- create a bar graph that shows the frequencies of each value gender
- within each gender, show the device category counts

Grouped Bar plot (counts)

- the we use the dodge position to creates a separate bar for every combination of cut and clarity

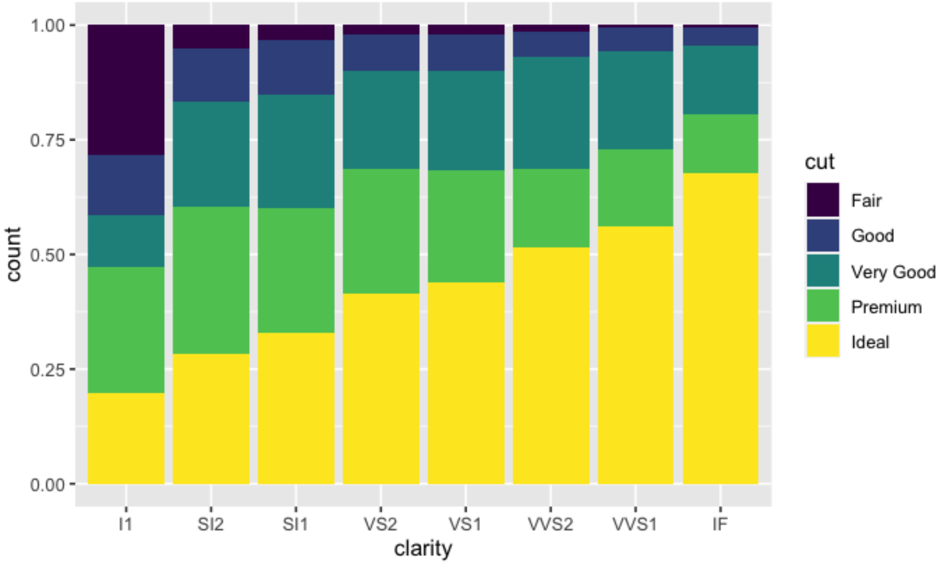
```
ggplot(data=diamonds,
              mapping= aes(x=clarity)) +
 geom_bar(aes(fill=cut), position="dodge")
```



Stacked Bar plot (proportions)

- we use the fill position to show proportions within each value of clarity
- we can see that the proportion of ideal cut diamonds is greater for diamonds with higher clarity

```
ggplot(data=diamonds,
          mapping= aes(x=clarity)) +
 geom_bar(aes(fill=cut), position="fill")
```



Exercise

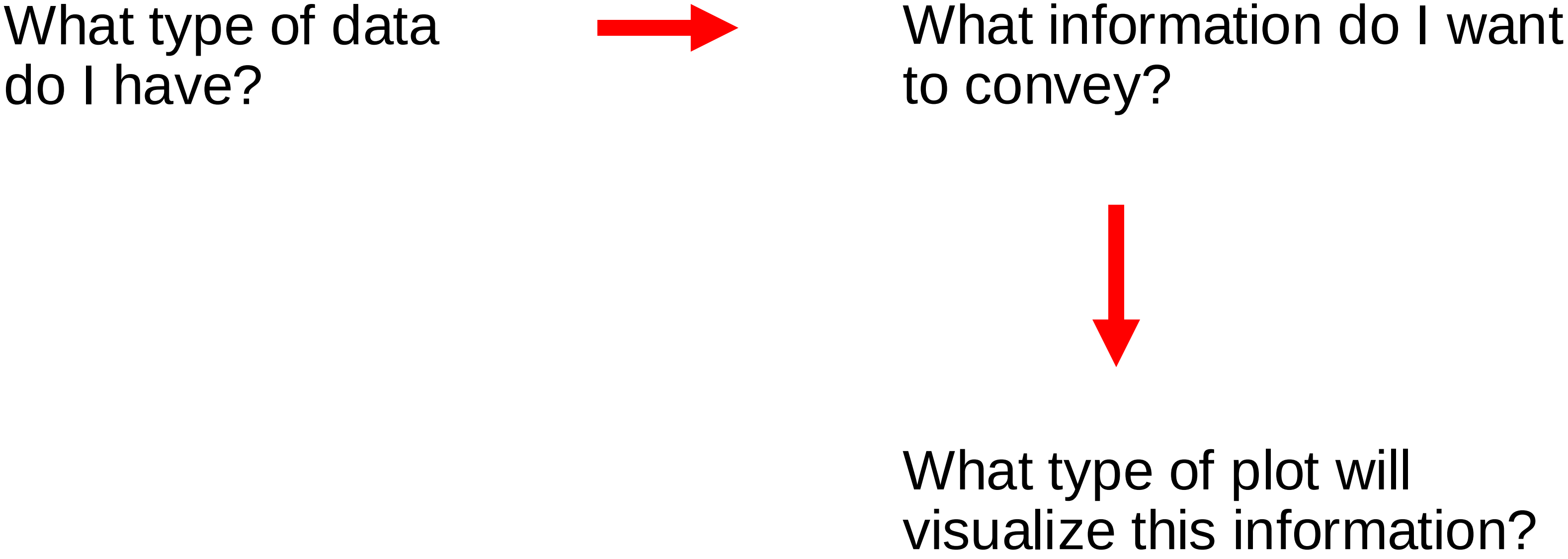
- filter for amusement rides
- only keep injury reports where the gender reported is male or female
- create a bar graph that shows the frequencies of each value of device category
- within each device category, show the gender proportions

Some parting words of wisdom

Things to be mindful of:

- Know if your variables are quantitative or categorical
- Know how your data is currently structured vs. how it needs to be structured for visualization (wrangle your data as needed)

Plan ahead



What type of data do I have?

What information do I want to convey?

What type of plot will visualize this information?

What plot do I need?

Data	**Information**	**Plot**
Two quantitative variables	Relationship between two variables	scatterplot
Quantitative variable and time	Trend over time	Line plot
Categorical variable	Frequencies within a single variable	barplot
Two categorical variables	Frequencies across variables	Grouped barplot
Two categorical variables	Relative frequencies across variables	Stacked barplot

What geom() do I need?

Plot	**geom**
scatterplot	**geom_point**()
Line plot	**geom_line**()
barplot	**geom_bar**() or **geom_col**()
Grouped barplot	**geom_bar**()
Stacked barplot	**geom_bar**()