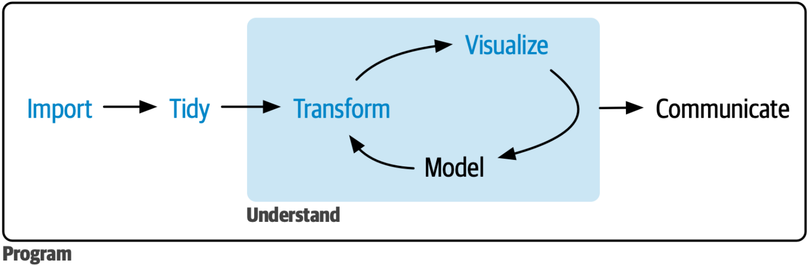
https://intro2r.com/ Chapter 5

CSCI 297b, Spring 2023

April 28, 2023

The Big Picture



"The simple graph has brought more information to the data analyst's mind than any other device." — John Tukey

The tidyverse

```
install.packages("tidyverse")
install.packages("palmerpenguins")
install.packages("ggthemes")

library("tidyverse")
library("palmerpenguins")
library("ggthemes")
```

- `tidyverse` and `ggthemes` already on RStudio Workbench
- Only have to install once. Not needed in scripts.
- Instead of `library(palmerpenguins)` can use

  `penguins <- palmerpenguins::penguins`

Questions

- Do penguins with longer flippers weigh more or less than penguins with shorter flippers?
- Try to make your answer precise.
- What does the relationship between flipper length and body mass look like?
- Is it positive? Negative? Linear? Nonlinear?
- Does the relationship vary by the species of the penguin?
- And how about by the island where the penguin lives?

Data frame vocabulary

variable A variable is a quantity, quality, or property that you can measure.

value A value is the state of a variable when you measure it. The value of a variable may change from measurement to measurement.

observation An observation is a set of measurements made under similar conditions. An observation will contain several values, each associated with a different variable. We'll sometimes refer to an observation as a data point.

Tabular data Tabular data is a set of values, each associated with a variable and an observation. Tabular data is tidy if each value is placed in its own "cell", each variable in its own column, and each observation in its own row.

penguins

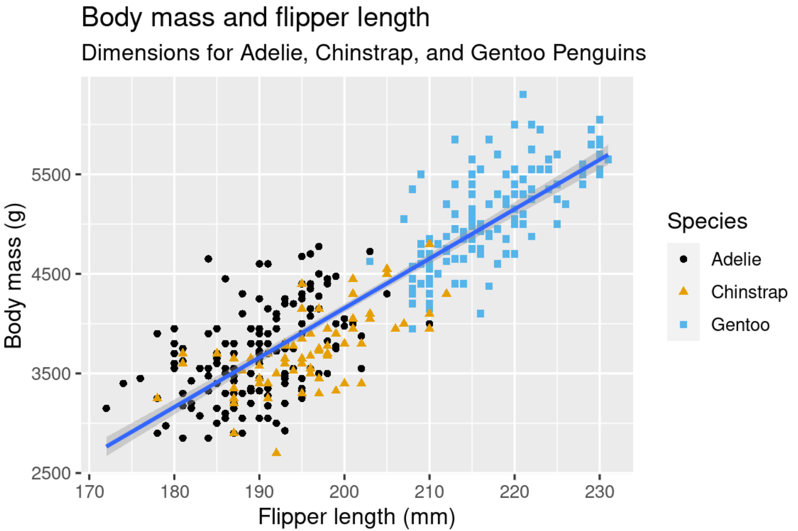
- penguins contains 344 observations collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER

```
> str(penguins)
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 42
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA N
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007
```

penguins

```
> glimpse(penguins)
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Ade...
$ island            <fct> Torgersen, Torgersen, Torgersen, To...
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 3...
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 1...
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 1...
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3...
$ sex               <fct> male, female, female, NA, female, m...
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007,...
```

Ultimate goal

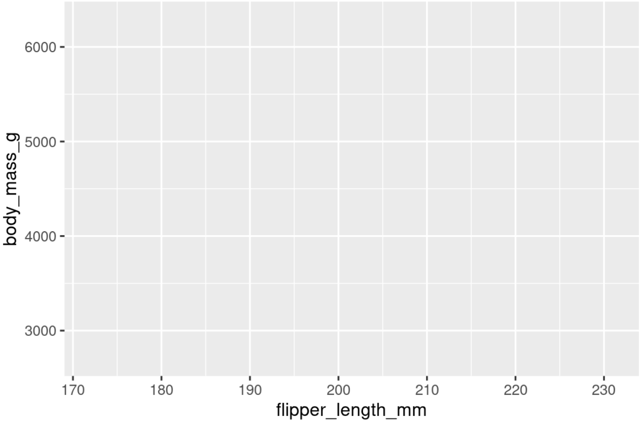


Creating a ggplot step by step

```
ggplot(data = penguins)
```

Add mappings for *x* and *y*

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
)
```

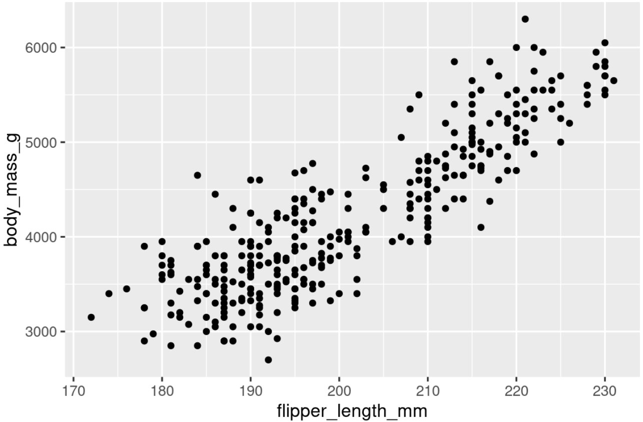


Specify a geom

- How do we represent the data on our plot?
- `geom_point()`
- `geom_bar()`
- `geom_line()`
- `geom_boxplot()`

geom_point()

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +  geom_point()
#> Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```

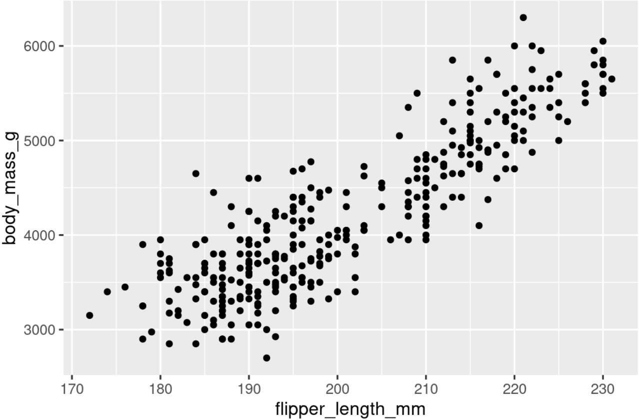


Missing values

```
Warning: Removed 2 rows containing missing values (‘geom\_point()‘).}
```

- Missing values in either the flipper length or body mass.
- Missing values are very common in real data.
- In remaining plots we won’t print this warning.

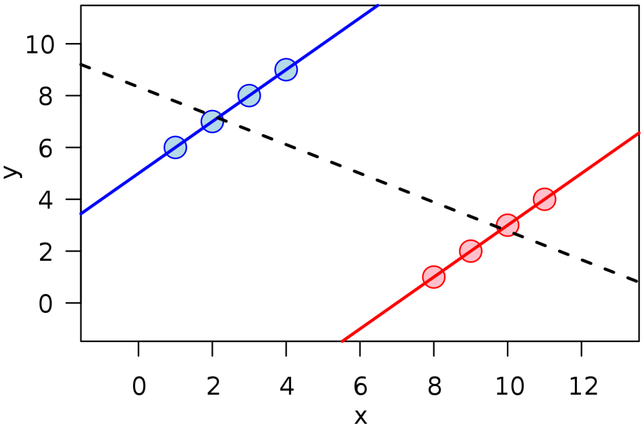
Checking hypotheses



- The relationship appears linear.
- The relationship appears strong.

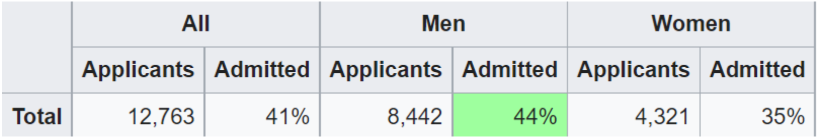
- Is it true within each species?
- Is it mainly an effect of having different species in the same dataset?

Simpson's paradox



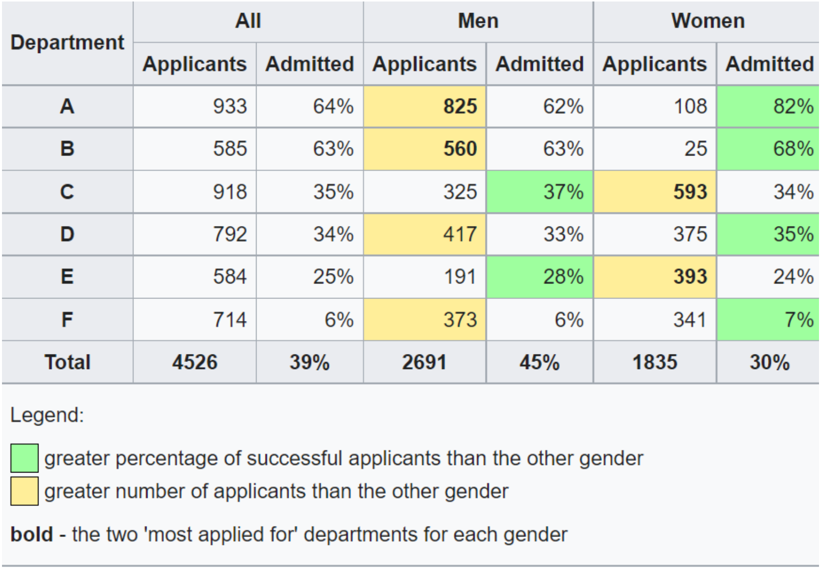
- What's good for each group is bad for the whole.

1973 UC Berkeley gender bias



	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

1973 UC Berkeley gender bias



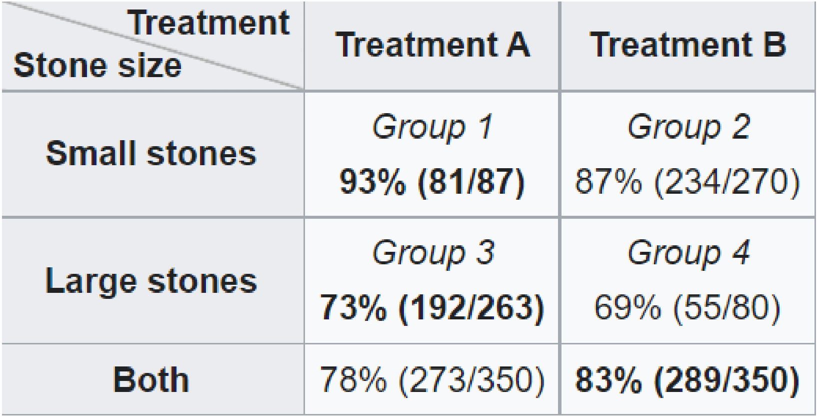
Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

🟩 greater percentage of successful applicants than the other gender
🟨 greater number of applicants than the other gender

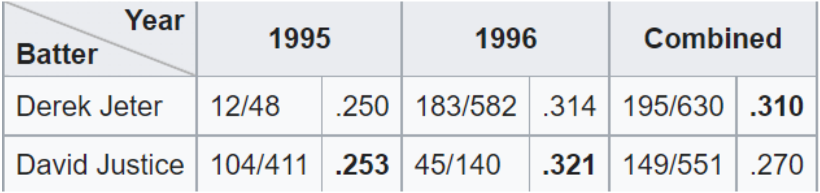
bold - the two 'most applied for' departments for each gender

Kidney Stones



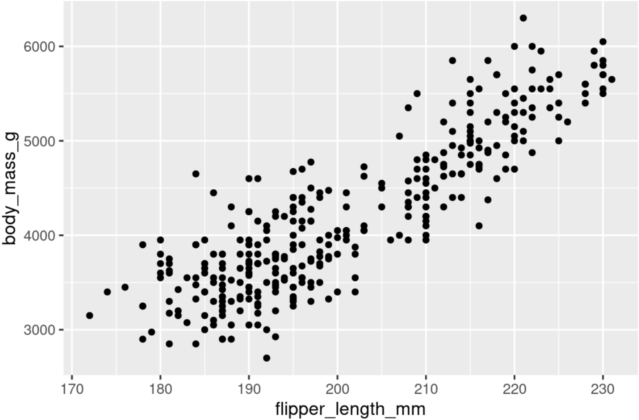
Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Batting averages



Year Batter	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	**.310**
David Justice	104/411	**.253**	45/140	**.321**	149/551	.270

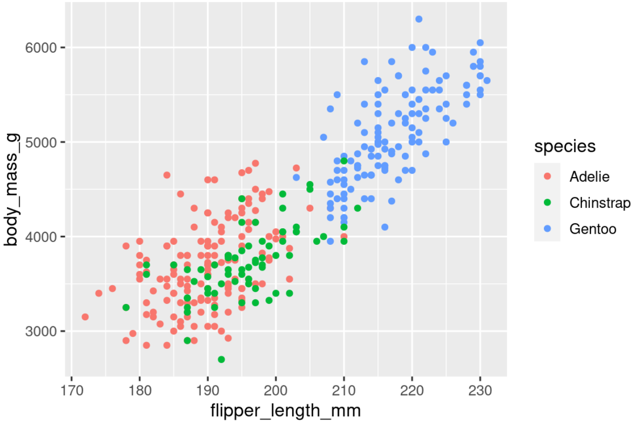
Simplson's paradox?



- We want to color by species.
- `geom` or `aes`?

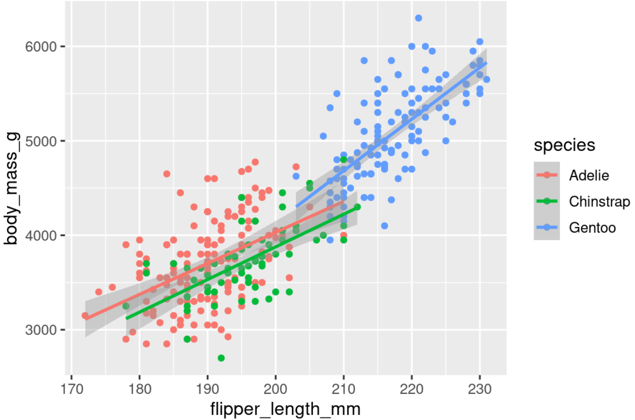
Creating a ggplot step by step

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)
) + geom_point()
```



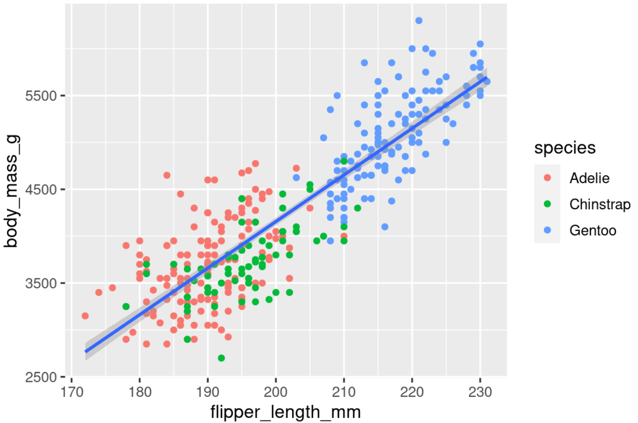
Add a smooth curve

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)
) +
  geom_point() +
  geom_smooth(method = "lm")
```



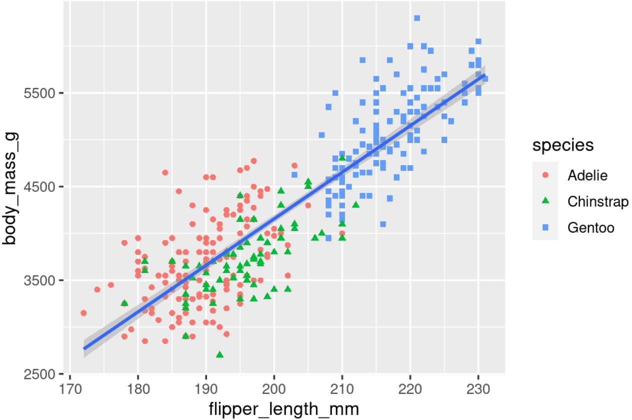
Aesthetics for everything, or just for some things

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(mapping = aes(color = species)) +
  geom_smooth(method = "lm")
```



Add shapes

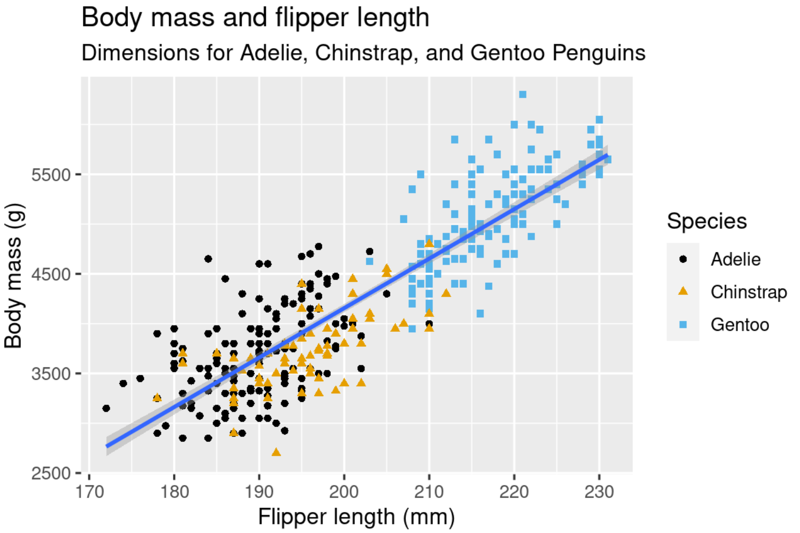
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(mapping = aes(color = species, shape = species)) +
  geom_smooth(method = "lm")
```



Add labels

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(aes(color = species, shape = species)) +
  geom_smooth(method = "lm") +
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",
    x = "Flipper length (mm)", y = "Body mass (g)",
    color = "Species", shape = "Species"
  ) +
  scale_color_colorblind()
```

Add labels


Body mass and flipper length
Dimensions for Adelie, Chinstrap, and Gentoo Penguins
Body mass (g)
5500
4500
3500
2500
170
180
190
200
210
220
230
Flipper length (mm)
Species
● Adelie
▲ Chinstrap
■ Gentoo

Do exercise 5