Exploring relationships between two variables

Corinne Riddell (Instructor: Tomer Altman)

September 9, 2025

Administrivia

- Please *only* use the `142gsi@berkeley.edu` email address for contacting the instructors
- Lectures and R Studio
  - We will explain concepts and demonstrate them with code in R Studio
  - Please focus on the concepts, and don't try to use R Studio at the same time
  - Please use the Zoom Chat to ask conceptual questions, and not technical ones

Recap of Chapters 1 and 2

- Histograms and bar charts to plot the distribution of a variable
- Measures of central tendency (e.g., mean, median) and spread (e.g., standard deviation, IQR)
- Time plots to examine the *relationship* between a variable and time

Learning objectives for today

- Explore the relationship between two quantitative variables
  - Direction, form, strength, outliers
  - Association vs. causation
- Make scatter plots to visualize bivariate relationships
  - using `geom_point()`
- Calculate the **correlation coefficient** to quantify the strength of linear relationships
  - using the `cor()` function

Readings

- Chapter 3 of Baldi and Moore
- Visual Distribution of different correlation coefficients (See section 5.7.4)
- Interpreting Correlation Coefficients (See section 5.7.5)

Why visualize bivariate relations?

Why is it important to examine bivariate relations (i.e., the relationship between two variables)? We learned last week about visualizing the distribution of individual variables using bar plots and histograms. Isn't that enough?

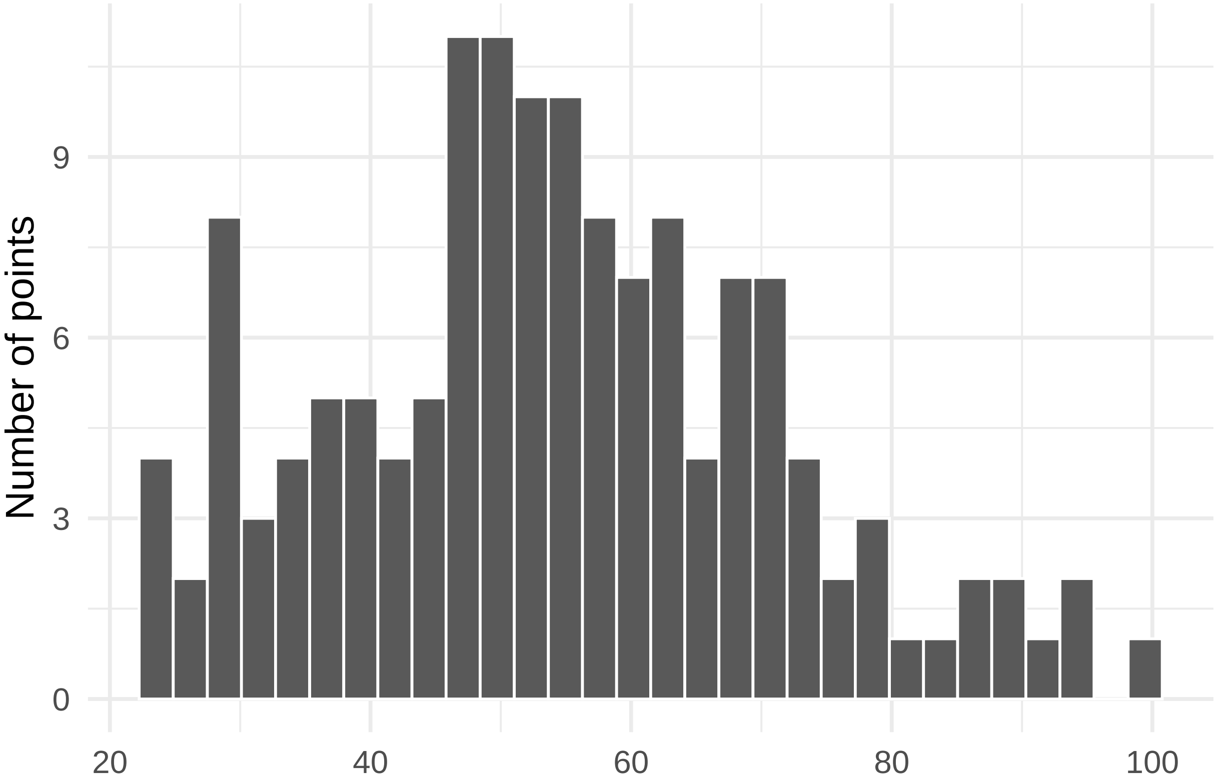
I will load a mystery data set that we want to examine.

Why visualize bivariate relations? X-variable

First, we look at the histogram of the `x` variable. What does the data tell you?

```r
ggplot(data = data_points, aes(x = x)) +
  geom_histogram(col = "white") +
  labs(x = "",
       y = "Number of points") +
  theme_minimal(base_size = 15)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

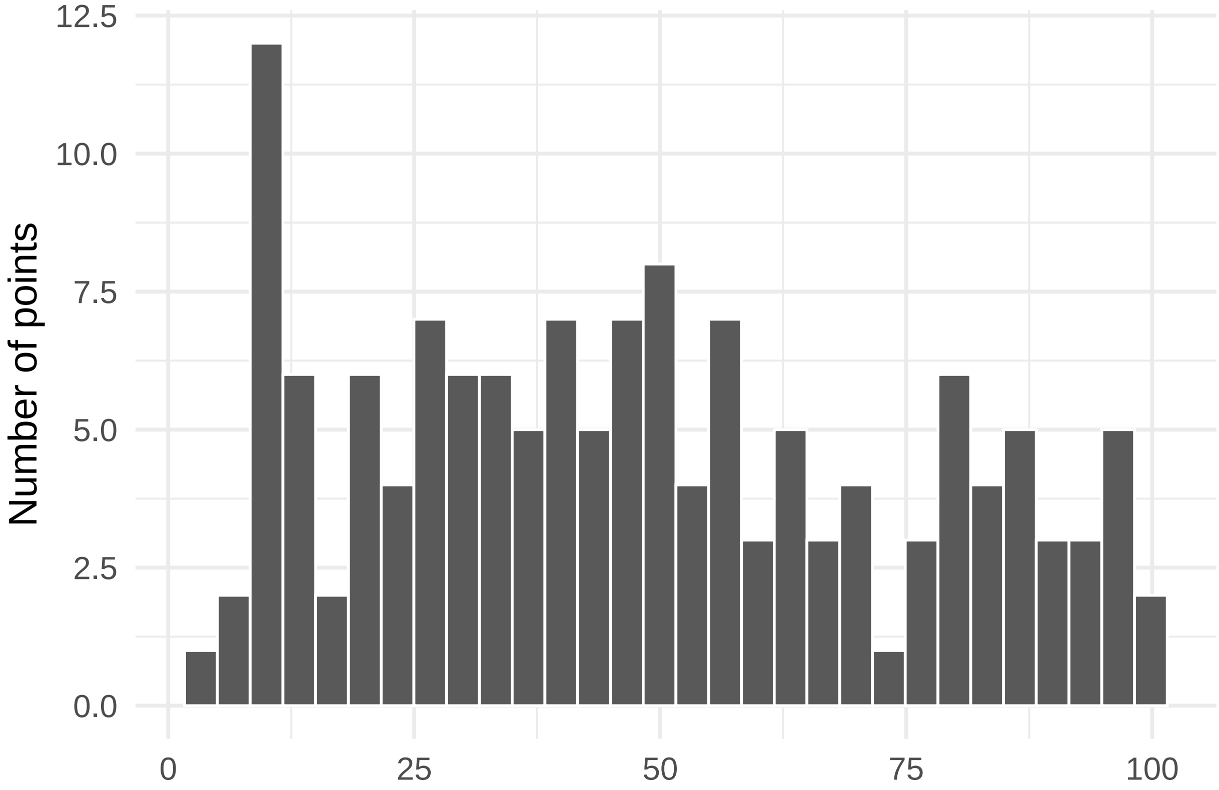


Why visualize bivariate relations? Y-variable

Next, we look at the histogram of the `y` variable. See anything interesting?

```
ggplot(data = data_points, aes(x = y)) +
  geom_histogram(col = "white") +
  labs(x = "",
       y = "Number of points") +
  theme_minimal(base_size = 15)
```

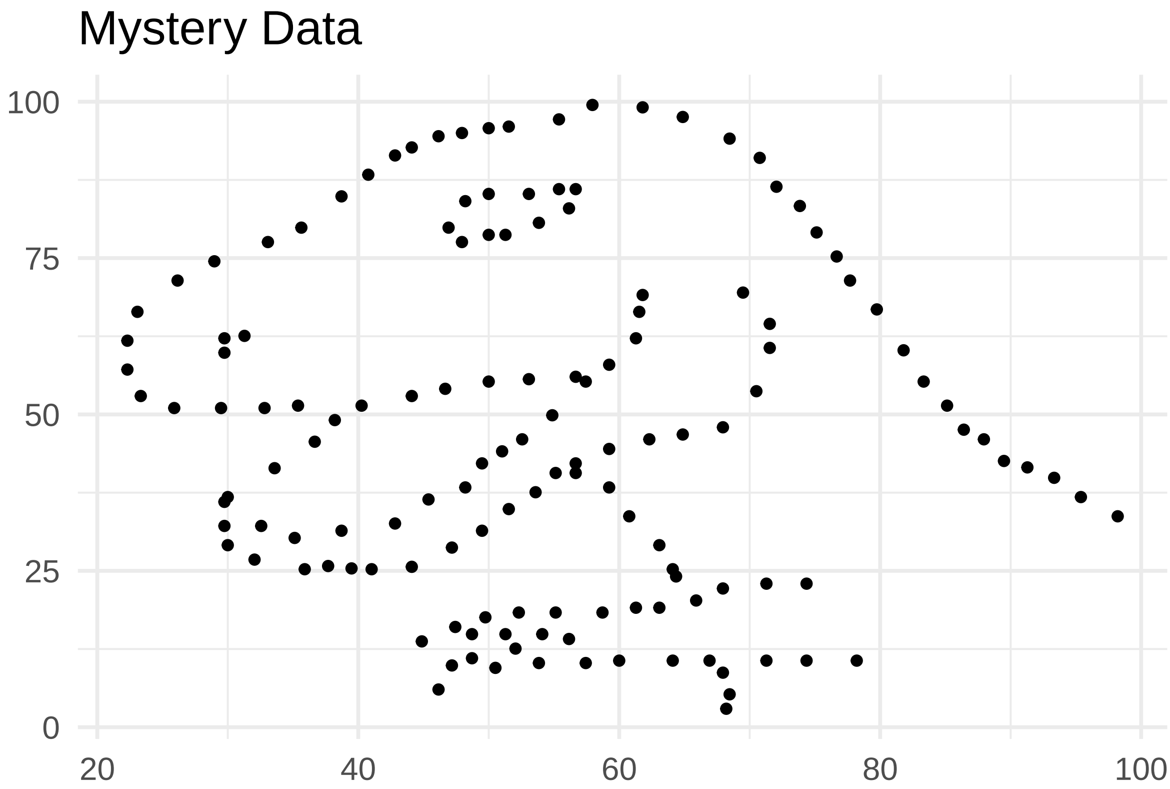
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Why visualize bivariate relations? Scatter-Plot

Now, let's plot the x and y variables together in a scatter-plot. See any pattern?

```
ggplot(data_points, aes(x = x, y = y)) +
  geom_point() +
  theme_minimal(base_size = 15) +
  labs(x = "",
       y = "",
       title = "Mystery Data")
```



Explanatory (X) and response (Y) variables

Bi-directional statements:

- "X predicts Y", or "Y predicts X"
- "X is associated with Y", or "Y is associated with X"
- These statements don't comment on causation. Only that two variables are related.

Unidirectional statements:

- "X causes Y"
- This statement is stronger. Not only are X and Y related, X is a cause of Y. That is, if you change X, then Y will also change. Researchers conduct studies to investigate causal claims.

Which variable is x and which is y?

- In **prediction** modeling, X denotes the variable used to predict the variable of interest (Y)
- In **causal** modeling, X denotes the explanatory (independent) variable and Y denotes the response (dependent) variable
- Graphically, the X variable is on the X (horizontal) axis and the Y variable is the Y (vertical) axis

Which variable is x and which is y?

1. Each hospital's rate of hospital-acquired infections, and whether the hospital has implemented a hand-washing intervention as part of a cluster randomized trial.
2. A person's leg length and arm length, in centimeters
3. Inches of rain in the growing season and the yield of corn in bushels per day

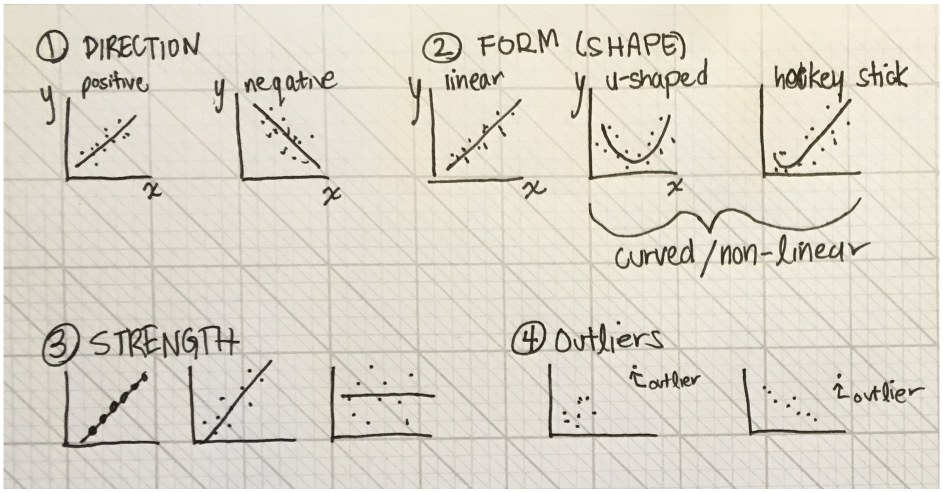
4. The number of steps a person takes each day and a person's mental health

How to investigate causation

- Experimentally: Using a randomized controlled trial (RCT) to randomize individuals to different levels
- Observationally: Conduct an observational study that is specifically designed to investigate causation and reduce the risk of bias
- To know more, take a class specifically about clinical trial design or take introduction to epidemiology to learn all about conducting observational studies
- In both settings, biostatistics is used to perform the calculations that are informed by the study design

Scatter plots

- Scatter plots are a preferred way to visualize a relationship between two variables
- They are used to evaluate:
  - **Direction**: Positive or negative?
  - **Form**: Linear or curved?
  - **Strength**: How close do the points lie to a line?
  - **Outliers**: Any individuals outside the general pattern?



Bi-directional relationships ex: systolic and diastolic BP

Read in NHANES dataset

```r
library(readr)
nhanes <- read_csv("./data/BPXI_I.csv")
```

```
## New names:
## Rows: 9544 Columns: 22
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (22): ...1, SEQN, PEASCCT1, BPXCHR, BPAARM, BPACSZ, BPXPLS, BPXPULS, BPX...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
head(nhanes)
```

```
## # A tibble: 6 x 22
##    ...1  SEQN PEASCCT1 BPXCHR BPAARM BPACSZ BPXPLS BPXPULS BPXPTY BPXML1 BPXSY1
##   <dbl> <dbl>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
## 1      1 83732       NA      NA       1       4      76        1       1     150     128
## 2      2 83733       NA      NA       1       4      72        1       1     170     146
## 3      3 83734       NA      NA       1       4      56        1       1     160     138
## 4      4 83735       NA      NA       1       5      78        1       1     150     132
## 5      5 83736       NA      NA       1       3      76        1       1     130     100
## 6      6 83737       NA      NA       1       4      64        1       1     140     116
## # i 11 more variables: BPXDI1 <dbl>, BPAEN1 <dbl>, BPXSY2 <dbl>, BPXDI2 <dbl>,
## #   BPAEN2 <dbl>, BPXSY3 <dbl>, BPXDI3 <dbl>, BPAEN3 <dbl>, BPXSY4 <dbl>,
## #   BPXDI4 <dbl>, BPAEN4 <dbl>
```

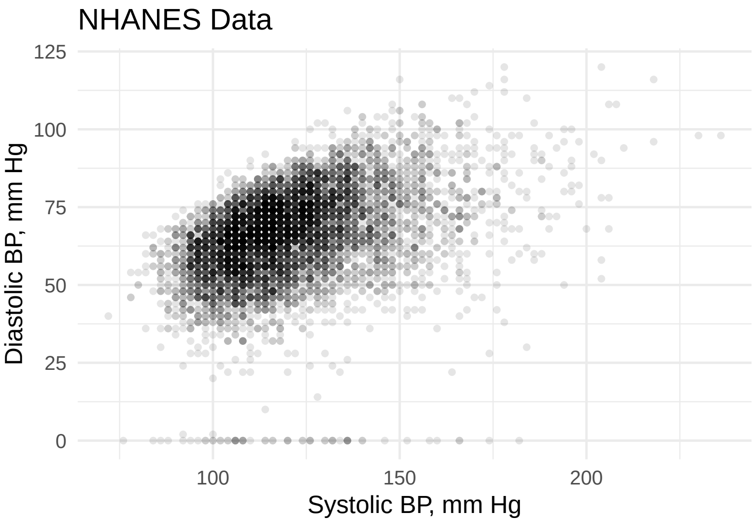
```
# View(nhanes) #Viewer provides data labels which are very useful for picking which variables to plot
```

Bi-directional relationships ex: systolic and diastolic BP

```
library(ggplot2)
bp_plot <- ggplot(nhanes, aes(x = BPXSY1, y = BPXDI1)) +
  geom_point(alpha = 0.1) +
  theme_minimal(base_size = 15) +
  labs(x = "Systolic BP, mm Hg",
       y = "Diastolic BP, mm Hg",
       title = "NHANES Data")
```

Bi-directional relationships ex: systolic and diastolic BP

```
## Warning: Removed 2399 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Bi-directional relationships ex: systolic and diastolic BP

What do we notice from the plot?

- **Direction**: Positive or negative?
- **Form**: Linear or curved?
- **Strength**: How close do the points lie to a line?
- **Outliers**: Any individuals outside the general pattern?

Association with a plausible direction: motor boats and manatees

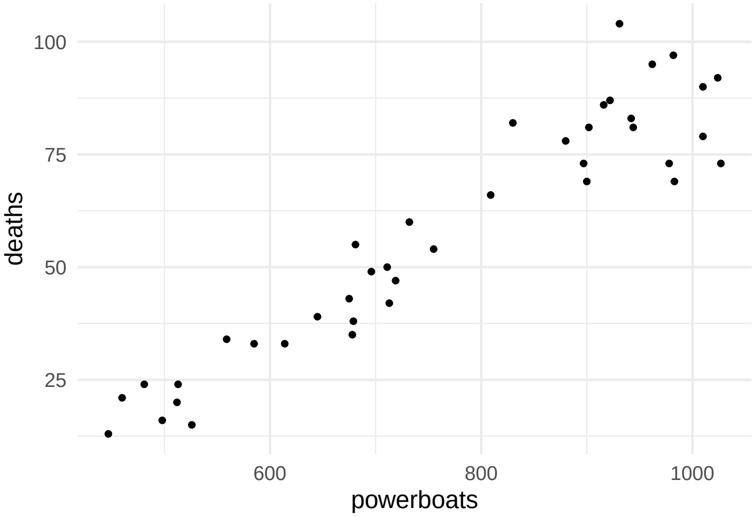
Read in the manatee data set (from the text book):

```r
library(readr)
mana_data <- read_csv("./data/Ch03_Manatee-deaths.csv")
```

```
## Rows: 40 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): year, powerboats, deaths
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Association with a plausible direction: motor boats and manatees

```r
mana_scatter <- ggplot(data = mana_data, aes(x = powerboats, y = deaths)) +
  geom_point() +
  theme_minimal(base_size = 15)

mana_scatter
```



Association with a plausible direction: motor boats and manatees

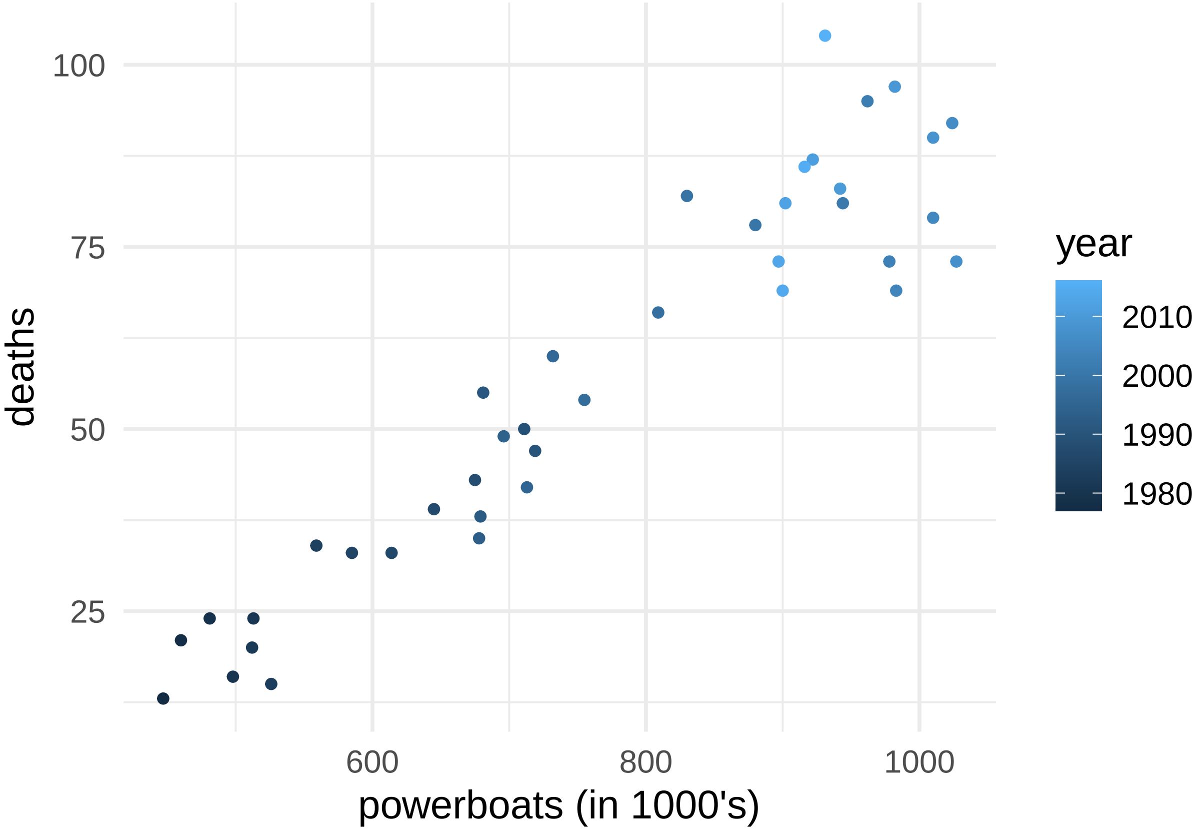
What do we notice from the plot?

- **Direction**: Positive or negative?
- **Form**: Linear or curved?
- **Strength**: How close do the points lie to a line?
- **Outliers**: Any individuals outside the general pattern?

Exercise: Power boats and Manatees

- Add (in thousands) to the x-axis title
- Change the point colour
- Is there a way to incorporate information on year into the graph?

```
ggplot(data = mana_data, aes(x = powerboats, y = deaths)) +
  geom_point(aes(col=year)) +
  theme_minimal(base_size = 15)+labs(x="powerboats (in 1000's)")
```



Example 3: Enzyme activity and temperature

- A study examined the activity rate (in micromoles per second) of a digestive enzyme at varying temperatures

```
# this dataset was provided in Baldi and Moore Ed#4 Apply your knowledge 3.4
enzyme_data <- read_csv("./data/Ch03_Enzyme-data.csv")
```

```
## Rows: 36 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): temperature, rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

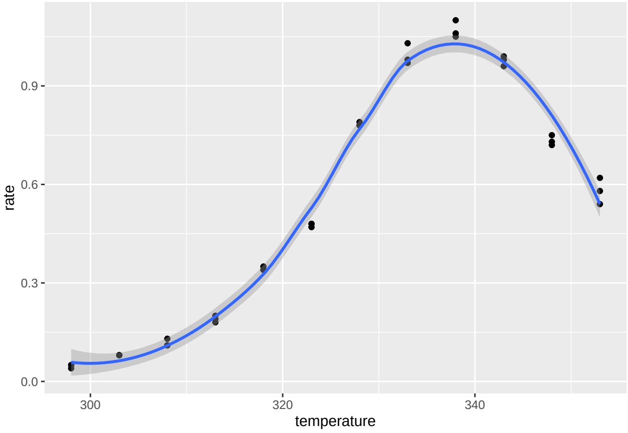
```
head(enzyme_data)
```

```
## # A tibble: 6 x 2
##   temperature  rate
##         <dbl> <dbl>
## 1         298  0.04
## 2         298  0.05
## 3         298  0.05
## 4         303  0.08
## 5          303  0.08
## 6          303  0.08
```

Scatter plot for enzyme data

```
ggplot(enzyme_data, aes(x = temperature, y = rate)) +
  geom_point() +
  geom_smooth()
```

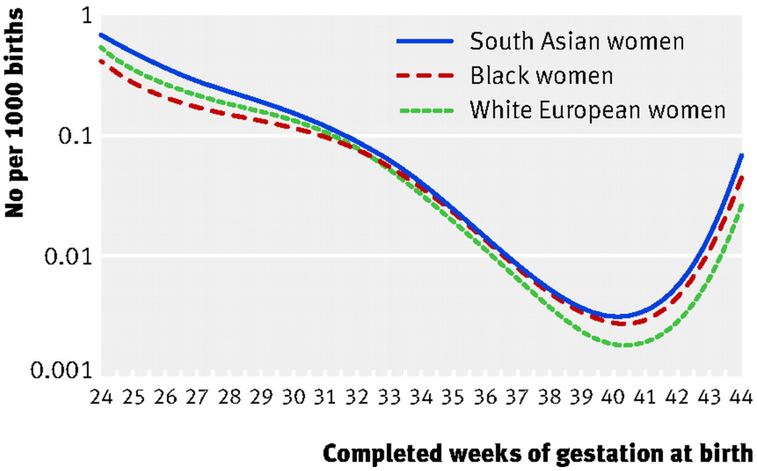
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Direction? Form? Strength? Outliers?

```
#Direction: `r #not strictly positive or negative.`

#Form: `r #non-linear`

#Strength: `r #strong`

#Outliers: `r #none`
```

Example 4: Gestational age and perinatal mortality



Source: Balchin et al. BMJ. 2007.

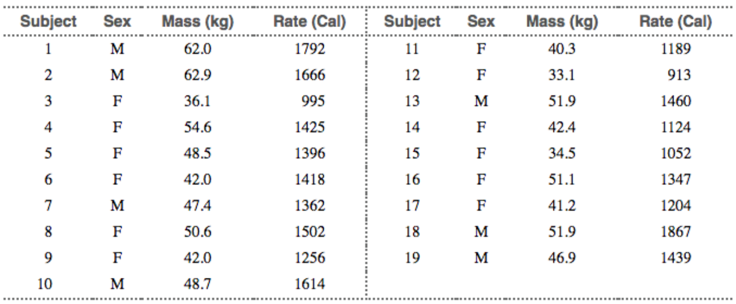
Example 5: Lean body mass and metabolic rate

Problem: Is lean body mass (person's weight after removing the fat) associated with metabolic rate (kilocalories burned in 24 hours)?

Plan: A diet study was conducted on 12 women and 7 men that measured lean body weight and metabolic rate for each individual.

Lean body mass and metabolic rate

Data:



Subject	Sex	Mass (kg)	Rate (Cal)
1	M	62.0	1792
2	M	62.9	1666
3	F	36.1	995
4	F	54.6	1425
5	F	48.5	1396
6	F	42.0	1418
7	M	47.4	1362
8	F	50.6	1502
9	F	42.0	1256
10	M	48.7	1614
11	F	40.3	1189
12	F	33.1	913
13	M	51.9	1460
14	F	42.4	1124
15	F	34.5	1052
16	F	51.1	1347
17	F	41.2	1204
18	M	51.9	1867
19	M	46.9	1439

- What would the corresponding data frame look like in R?
- How many variables does it have?
- How many rows?

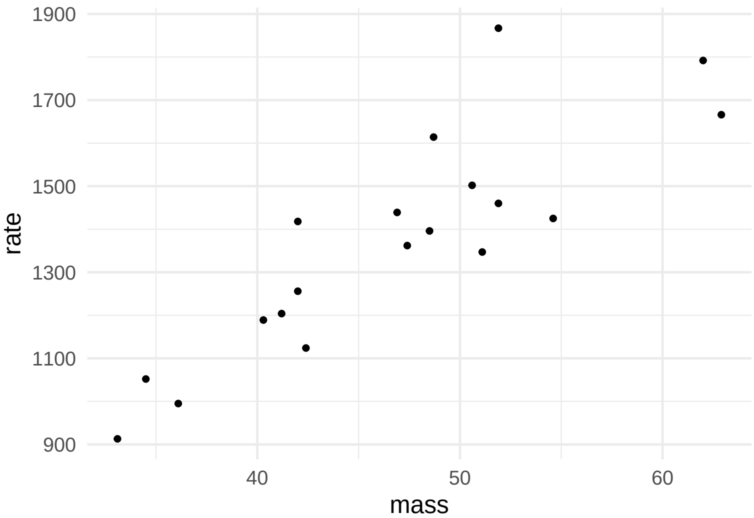
Lean body mass and metabolic rate

```r
# Note: you won't be tested on writing code using tibble::tribble()
# **Do** know how to look at this code and recognize that it is creating a data set

weight_data <- tibble::tribble(
  ~subject, ~gender, ~mass, ~rate,
  1, "M", 62.0, 1792,
  2, "M", 62.9, 1666,
  3, "F", 36.1, 995,
  4, "F", 54.6, 1425,
  5, "F", 48.5, 1396,
  6, "F", 42.0, 1418,
  7, "M", 47.4, 1362,
  8, "F", 50.6, 1502,
  9, "F", 42.0, 1256,
  10, "M", 48.7, 1614,
  11, "F", 40.3, 1189,
  12, "F", 33.1, 913,
  13, "M", 51.9, 1460,
  14, "F", 42.4, 1124,
  15, "F", 34.5, 1052,
  16, "F", 51.1, 1347,
  17, "F", 41.2, 1204,
  18, "M", 51.9, 1867,
  19, "M", 46.9, 1439
)
```

Analysis

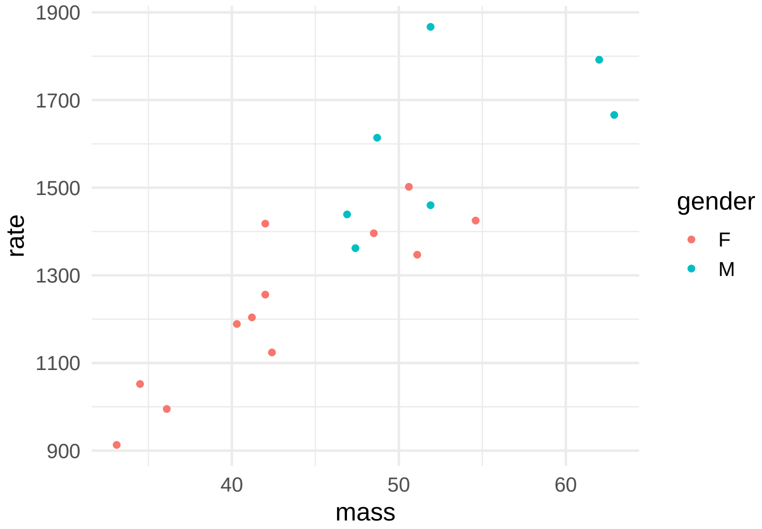
Exploratory data analysis using scatter plots

```
weight_scatter <- ggplot(weight_data, aes(x = mass, y = rate)) +
  geom_point() +
  theme_minimal(base_size = 15)

weight_scatter
```



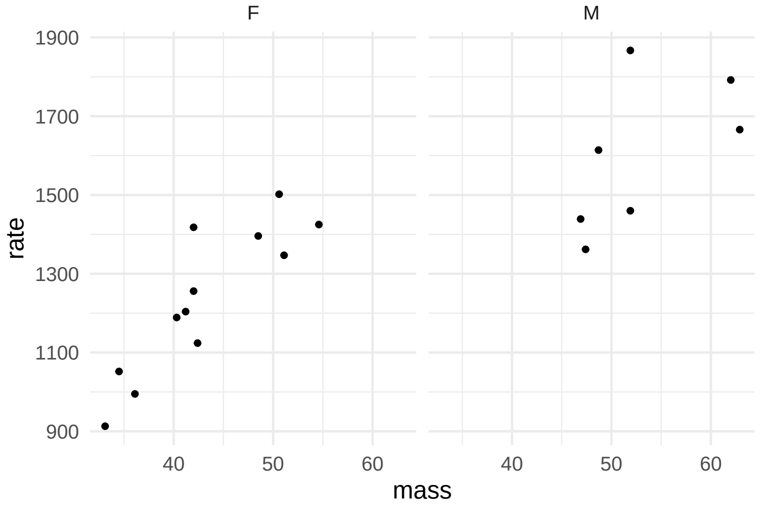
Analysis: Colour the points by gender

```
#Fill in during class
weight_scatter <- ggplot(weight_data, aes(x=mass, y=rate)) +
  geom_point(aes(col = gender)) +
  theme_minimal(base_size = 15)

weight_scatter
```



Analysis: Create separate plots for men and women

```
#Fill in during class
weight_scatter_by_gender <- ggplot(weight_data, aes(x=mass, y=rate)) +
  geom_point() +
  theme_minimal(base_size = 15) +
  facet_wrap(~gender)

weight_scatter_by_gender
```



Let's test our knowledge!

Direction:

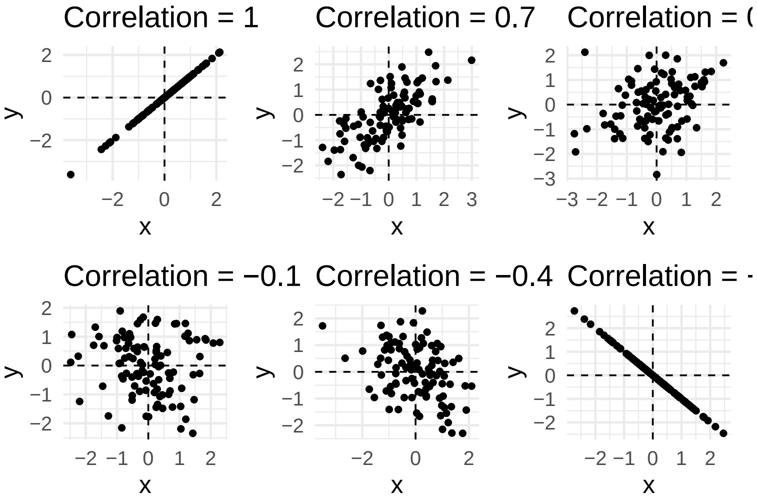
Form:

Strength:

Outliers:

Pearson's correlation

Using just our eyes, we can often say something about whether an association between two variables is weak or strong.



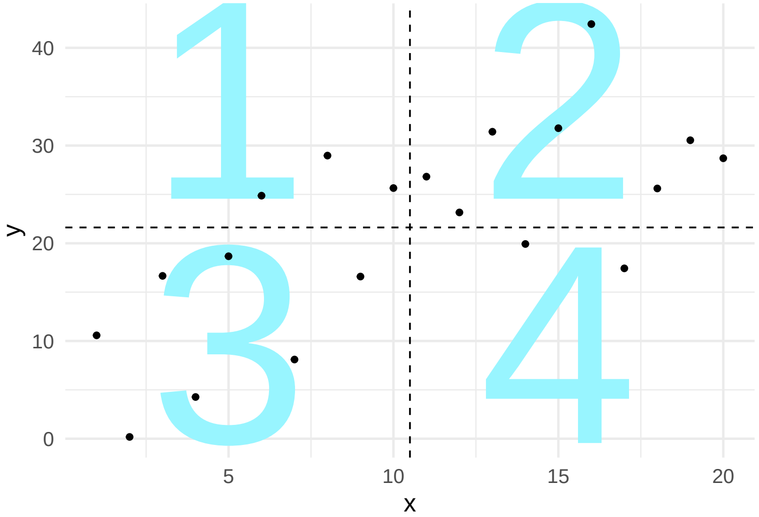
Pearson's correlation

- For **linear** associations, we can use **Pearson's correlation coefficient** (denoted by r) to **quantify the strength** of a linear relationship between two variables.
- The correlation between x and y is:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

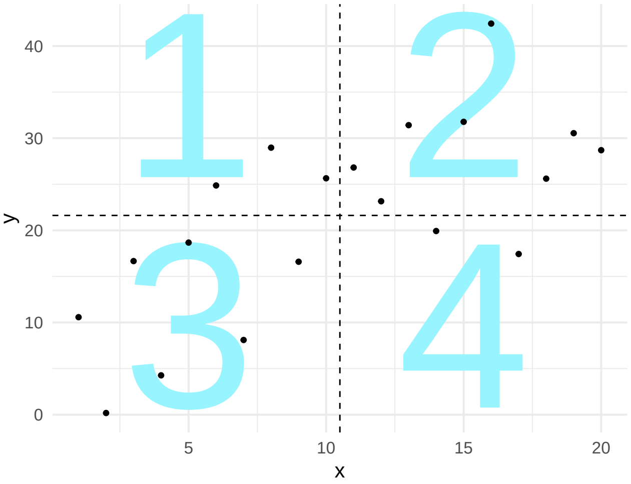
Intuition about Pearson's correlation

To understand this formula, first only consider the numerators of the fractions (i.e., $x_i - \bar{x}$ and $y_i - \bar{y}$). If you imagine a scatter plot of x and y, we can also add a dashed line at the mean x value of $\bar{x}$ and a dashed line line at the mean y value ($\bar{y}$):



Intuition about Pearson's correlation

$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$



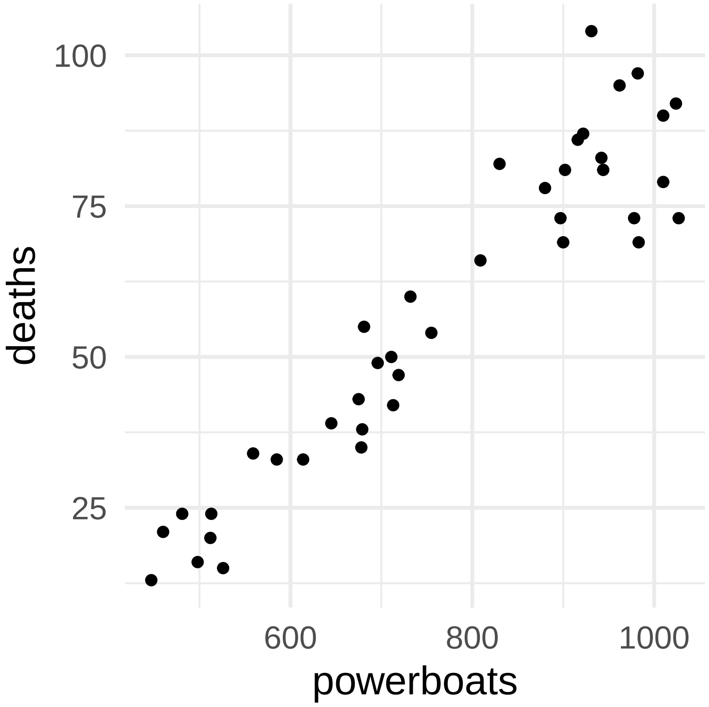
- Points in Q2 and Q3 contribute positive products to r
- Points in Q1 and Q4 contribute negative products to r
- The more there are points in Q2 and Q3 vs. Q1 and Q4, the more the value of the correlation coefficient will be higher and positive
- If you want even more of an explanation see the response to this stack overflow post or take an intermediate statistics class!

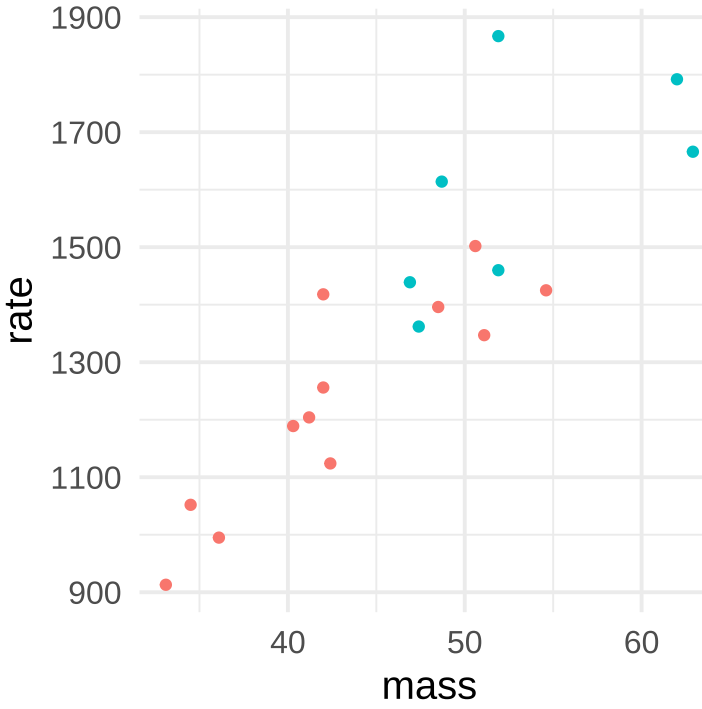
Syntax: Pearson's correlation using `cor()`

```
# Students, if you copy this code chunk, you need to set eval = T in the code chunk header for the code

correlation_coeff <- dataset %>%
  summarize(new_var = cor(x_variable, y_variable))
```

Syntax: Pearson's correlation using `cor()`

Remember the manatee plot and the weight plot:





Syntax: Pearson's correlation using `cor()`

Now, calculate the correlations between X and Y for manatees:

```
mana_cor <- mana_data %>%
  summarize(corr_mana = cor(powerboats, deaths))

mana_cor
```

```
## # A tibble: 1 x 1
##   corr_mana
##       <dbl>
## 1     0.945
```

Syntax: Pearson's correlation using `cor()`

And for the weight data:

```
weight_cor <- weight_data %>%
  summarize(corr_weight = cor(mass, rate))
weight_cor
```

```
## # A tibble: 1 x 1
##   corr_weight
##         <dbl>
## 1       0.865
```

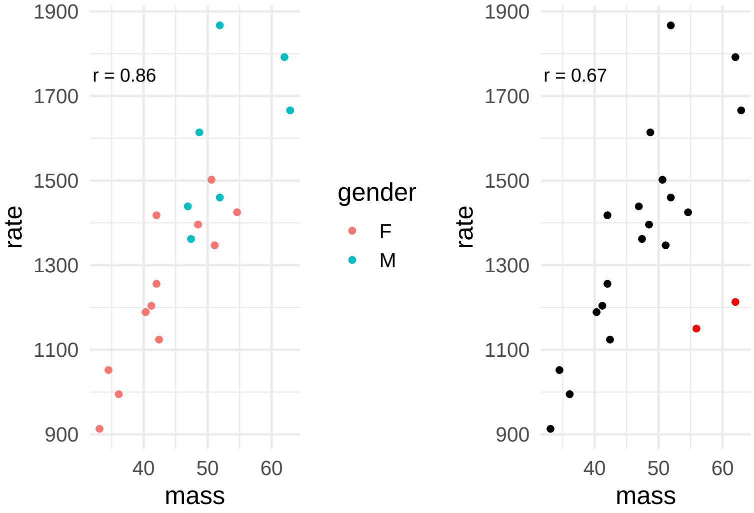
Properties of the correlation coefficient

- Always a number between -1 and 1
  - -1: A perfect, negative linear association
  - 1: A perfect, positive linear association
  - 0: No linear association

- Don’t confuse the correlation coefficient with the slope of the linear association!
- Measures association, *not* causation. Even a very strong association doesn’t mean that one variable causes the other.
- Is used to measure the association between two *quantitative* variables
- Only useful for *linear* associations!

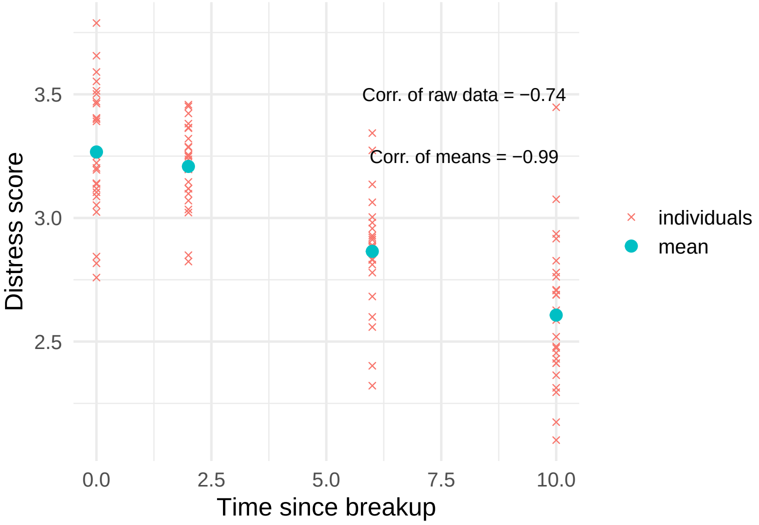
Properties of the correlation coefficient

- The correlation coefficient is not resistant to outliers
- E.g., I added two outliers (in red) to the `weight_data` and recalculated correlation. How much did the correlation change? (It is labeled on each plot.)



Properties of the correlation coefficient

- Correlations for average measures are typically stronger than correlations for individual data



Recap: What functions did we use?

- `geom_point()`, `aes(col = gender)` to color points by levels of `gender`
- `summarize()` to calculate correlation using `cor(var1, var2)`

Important concepts

- Determine which variable is explanatory and which is response, or when there is a bidirectional relationship (e.g., associated)
- Describe the relationship between two variables (e.g., form, direction, strength, and outliers)
- Formula for and properties of the correlation coefficient r