

# Week 4: *Correlation*

☰ EMSE 4575: Exploratory Data Analysis

👤 John Paul Helveston

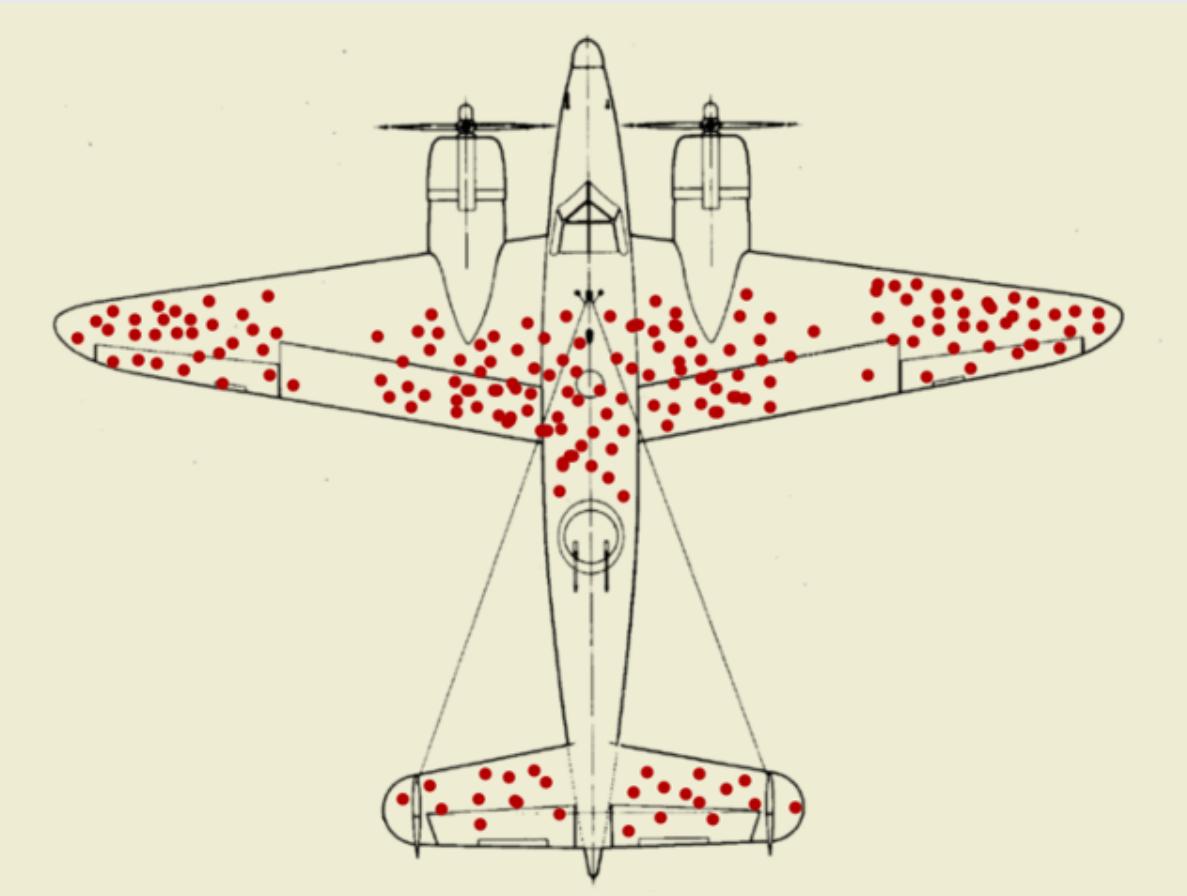
📅 February 03, 2021

## *Tip of the week*

All data are biased



Abraham Wald



# Today's data

```
msleep <- read_csv(here::here('data', 'msleep.csv'))
```

## New packages:

```
install.packages('HistData')
install.packages('palmerpenguins')
install.packages('GGally')
```

# Week 4: *Correlation*

1. What is correlation?
  2. Visualizing correlation
- BREAK
3. Linear models
  4. Visualizing linear models

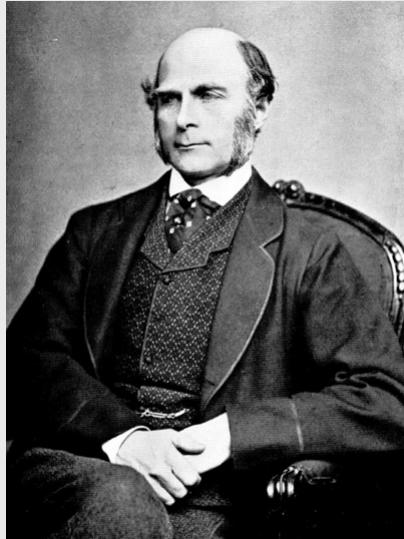
# Week 4: *Correlation*

1. What is correlation?
2. Visualizing correlation
- BREAK
3. Linear models
4. Visualizing linear models

# Some pretty racist origins in **eugenics** ("well born")

## Sir Francis Galton (1822 - 1911)

- Charles Darwin's cousin.
- "Father" of **eugenics**.
- Interested in heredity.



## Karl Pearson (1857 - 1936)

- Galton's (**hero-worshiping**) protégé.
- Defined correlation equation.
- "Father" of mathematical statistics.



# Galton's family data

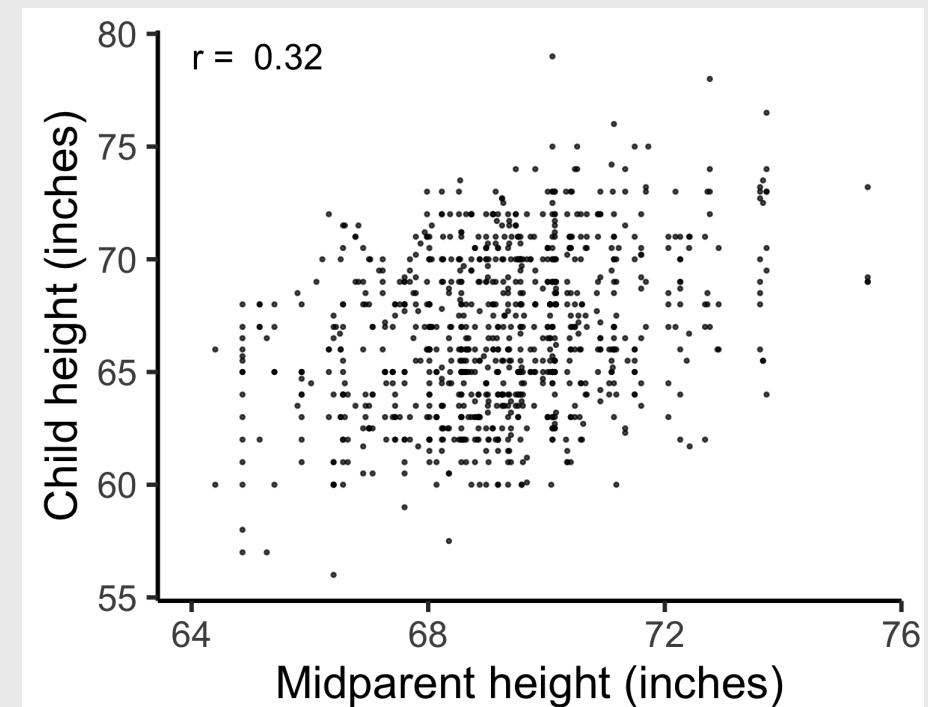
Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246-263.

**Galton's question:** Does marriage selection indicate a relationship between the heights of husbands and wives?  
(He called this "assortative mating")

"midparent height" is just a scaled mean:

```
midparentHeight = (father + 1.08*mother)/2
```

```
library(HistData)  
  
galtonScatterplot <- ggplot(GaltonFamilies) +  
  geom_point(aes(x = midparentHeight,  
                 y = childHeight),  
             size = 0.5, alpha = 0.7) +  
  theme_classic() +  
  labs(x = 'Midparent height (inches)',  
       y = 'Child height (inches)')
```



# How do you measure correlation?

Pearson came up with this:

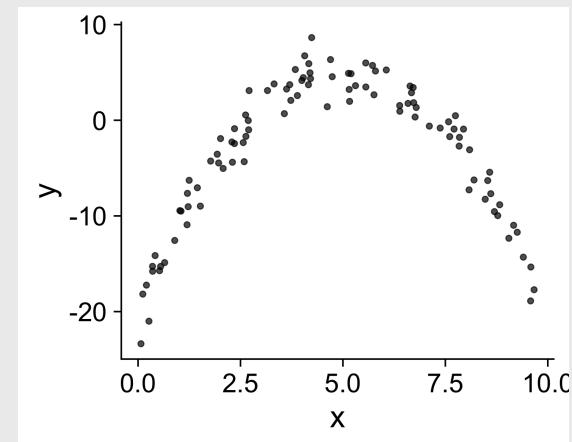
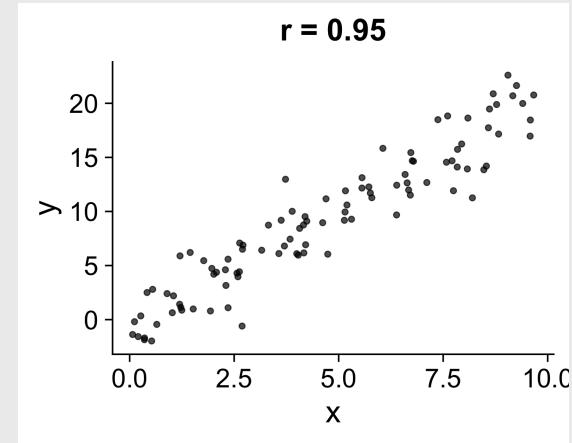
$$r = \frac{\text{Cov}(x,y)}{\text{sd}(x)*\text{sd}(y)}$$

# How do you measure correlation?

$$r = \frac{\text{Cov}(x,y)}{\text{sd}(x)*\text{sd}(y)}$$

Assumptions:

1. Variables must be interval or ratio
2. Linear relationship



# How do you *interpret* $r$ ?

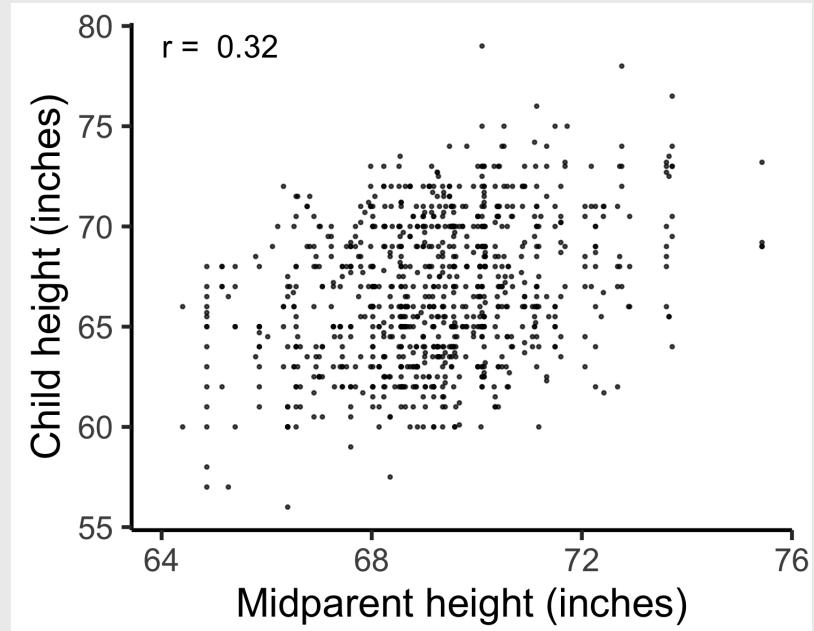
$$r = \frac{\text{Cov}(x,y)}{\text{sd}(x)*\text{sd}(y)}$$

```
cor(x = GaltonFamilies$midparentHeight,  
     y = GaltonFamilies$childHeight,  
     method = 'pearson')
```

```
#> [1] 0.3209499
```

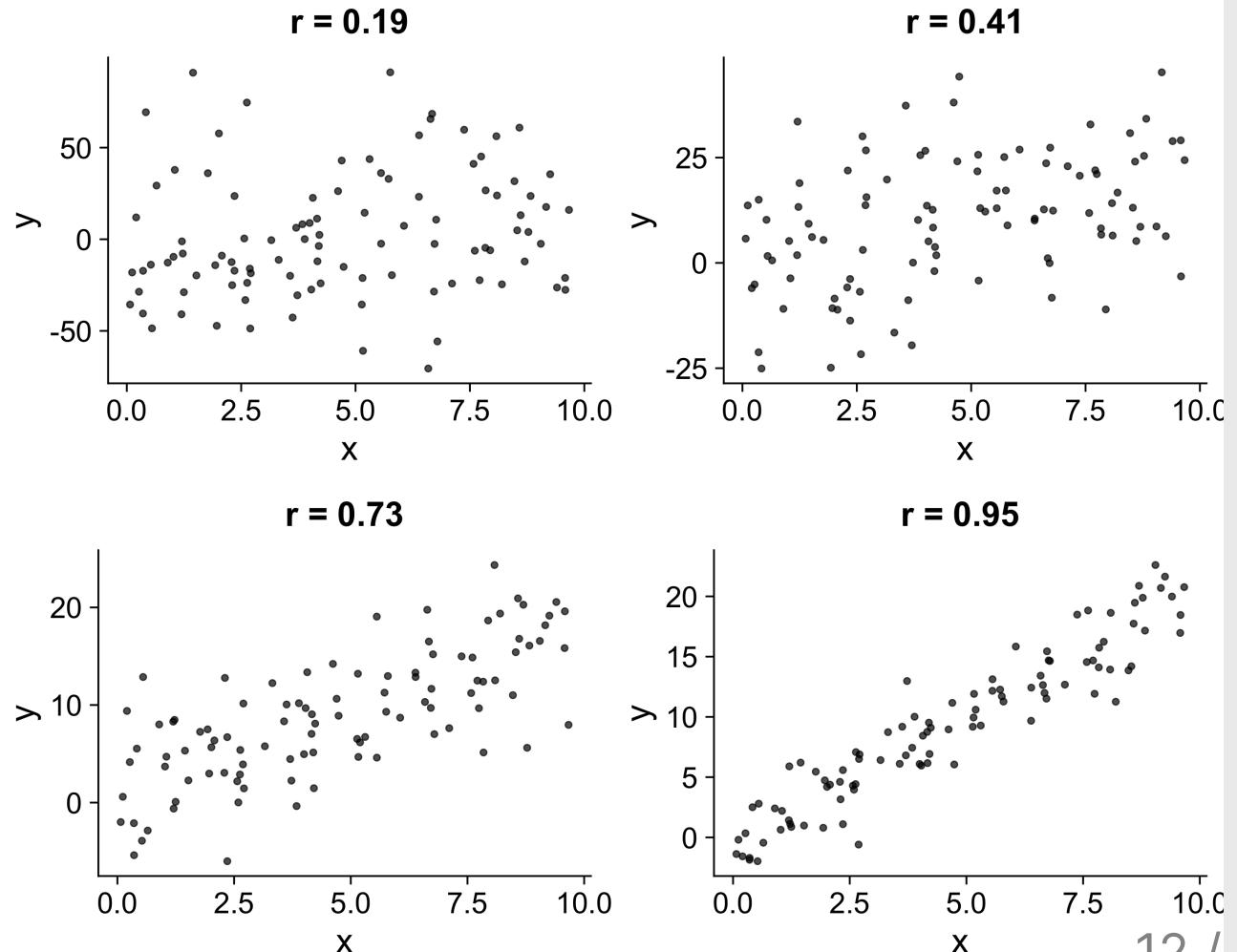
Interpretation:

- $-1 \leq r \leq 1$
- Closer to 1 is stronger correlation
- Closer to 0 is weaker correlation



# What does $r$ mean?

- $\pm 0.1 - 0.3$ : Weak
- $\pm 0.3 - 0.5$ : Moderate
- $\pm 0.5 - 0.8$ : Strong
- $\pm 0.8 - 1.0$ : Very strong



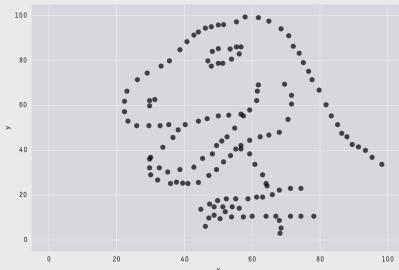
Visualizing correlation is...um...easy, right?

[guessthecorrelation.com](http://guessthecorrelation.com)

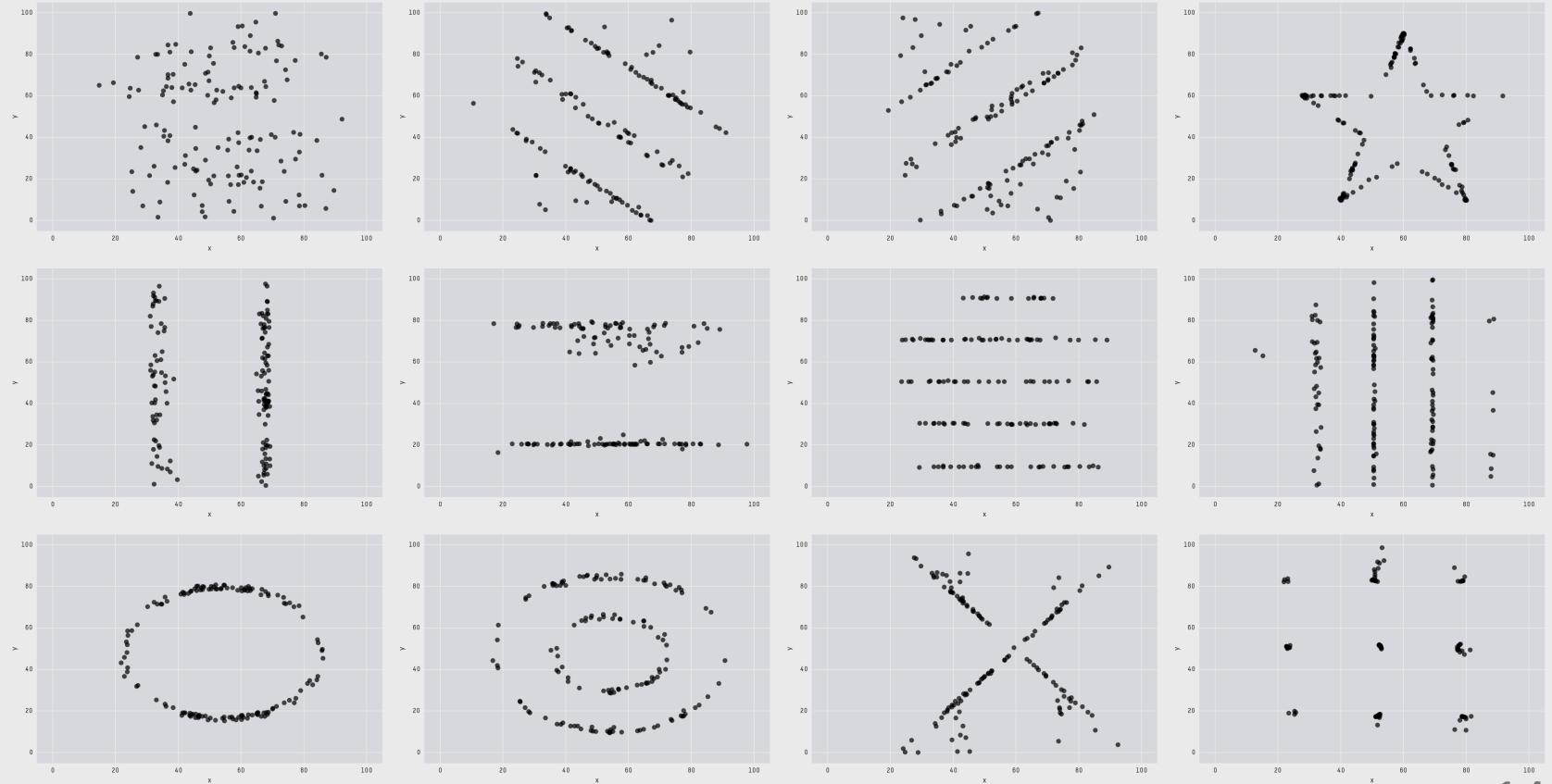
Click [here](#) to vote!

# The datasaurus

(More [here](#))



X Mean: 54.26  
Y Mean: 47.83  
X SD : 16.76  
Y SD : 26.93  
Corr. : -0.06



# Coefficient of determination: $r^2$

Percent of variance in one variable that is explained by the other variable

$r$	$r^2$
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25
0.6	0.36
0.7	0.49
0.8	0.64
0.9	0.81
1.0	1.00

You should report both  $r$  and  $r^2$

Correlation between parent and child height is 0.32, therefore 10% of the variance in the child height is explained by the parent height.

# Correlation != Causation

X causes Y

- Training causes improved performance

Y causes X

- (Good / bad) performance causes people to train harder.

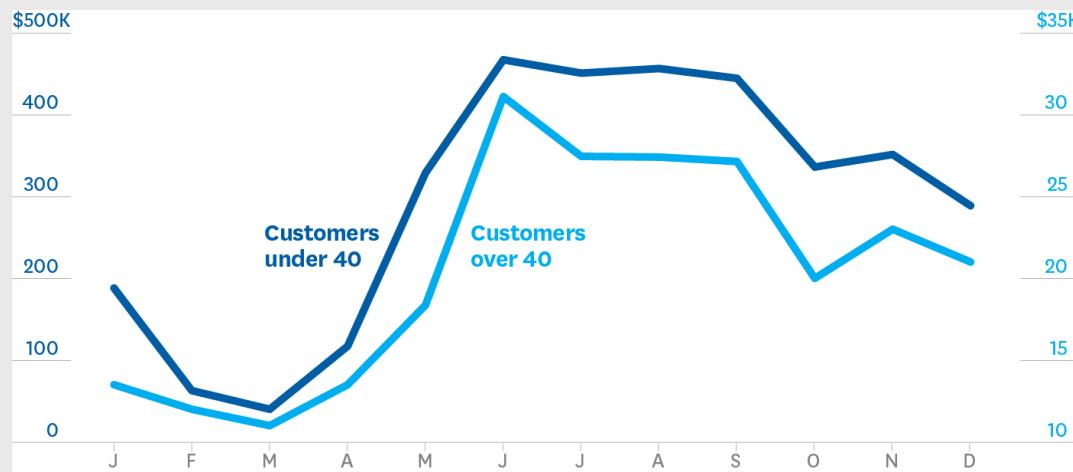
Z causes both X & Y

- Commitment and motivation cause increased training and better performance.

# Be weary of dual axes!

(They can cause spurious correlations)

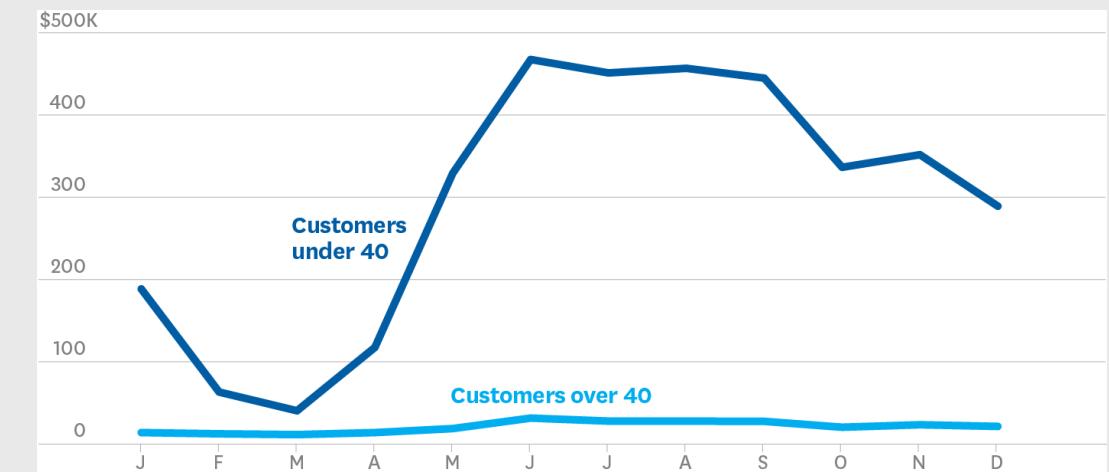
Dual axes



© HBR.ORG

FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

Single axis

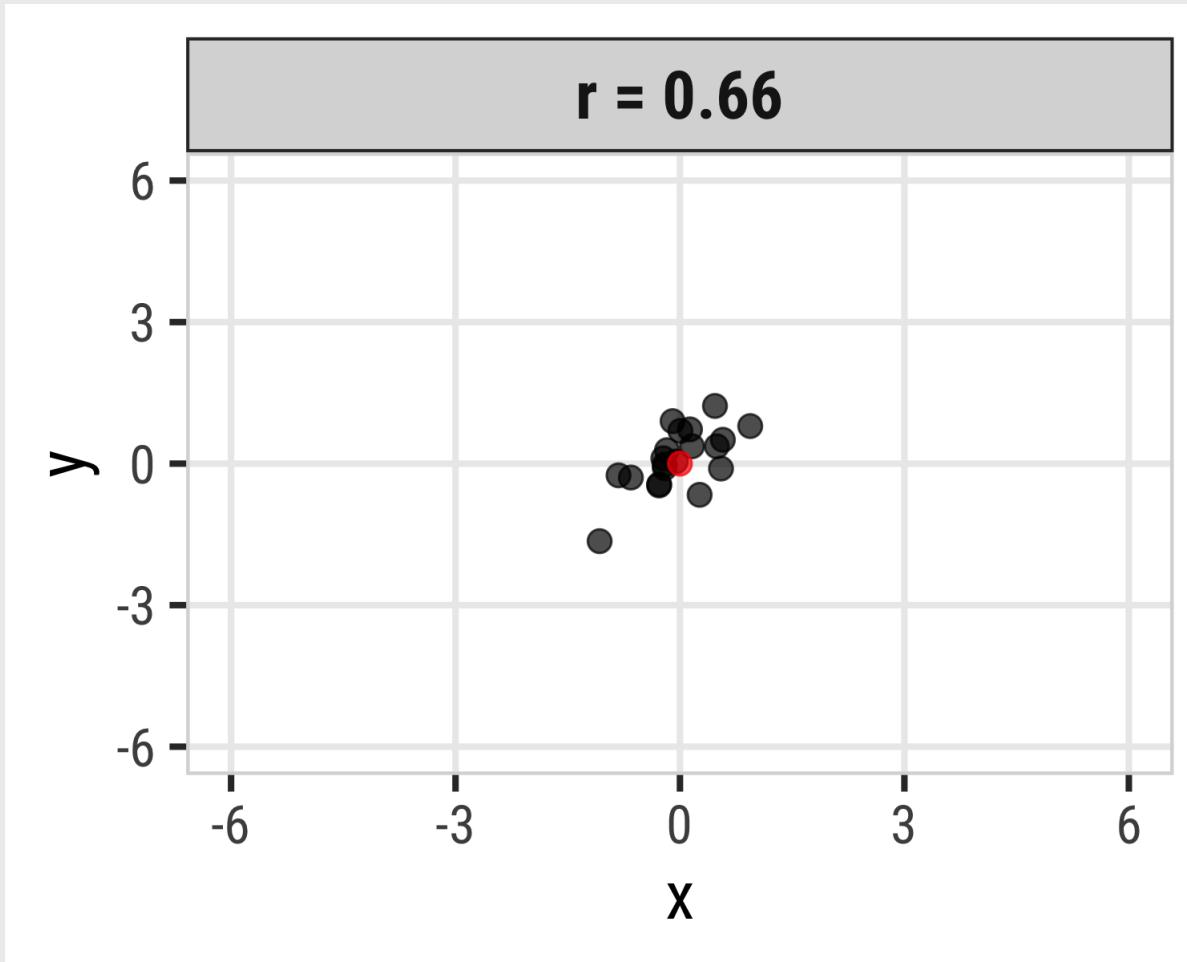


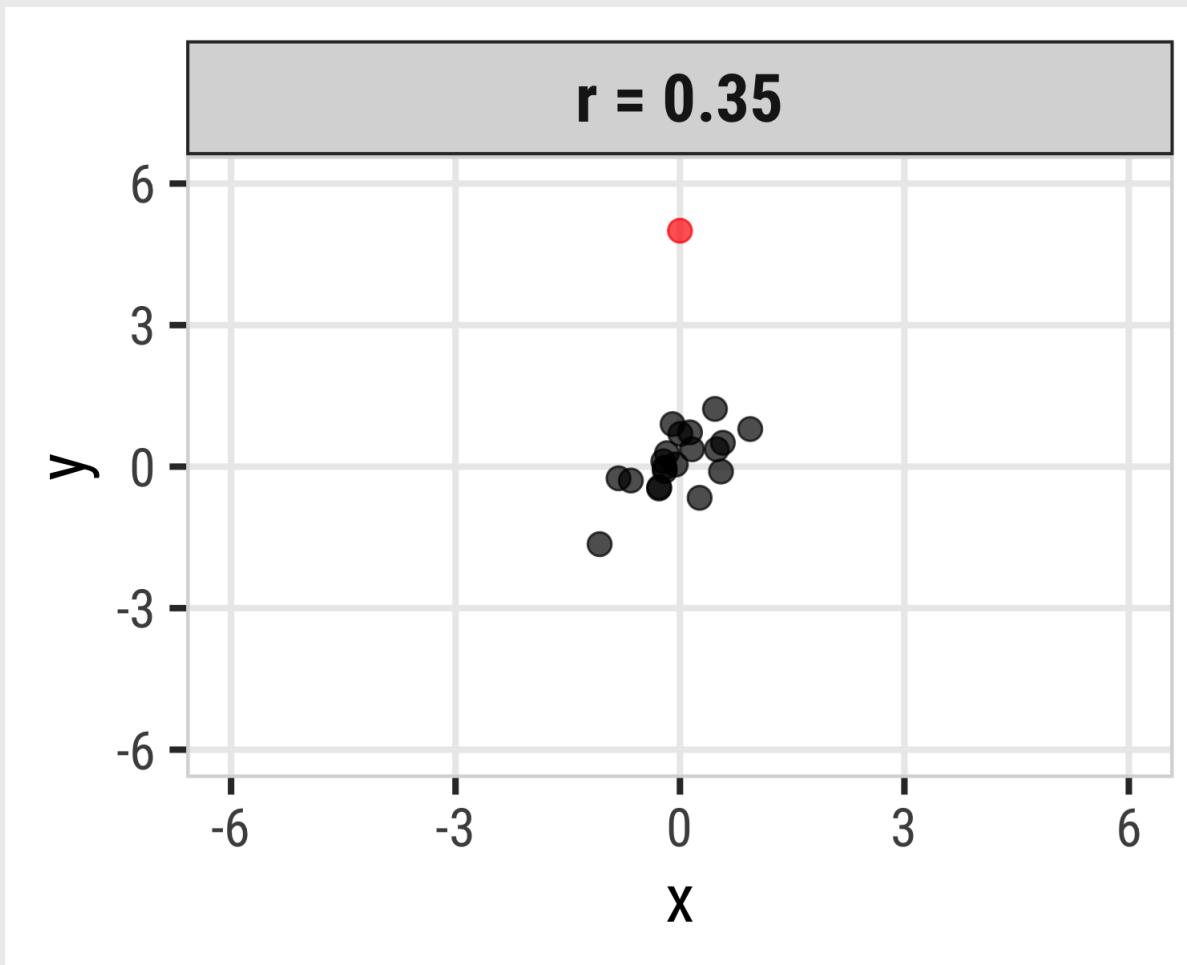
© HBR.ORG

FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

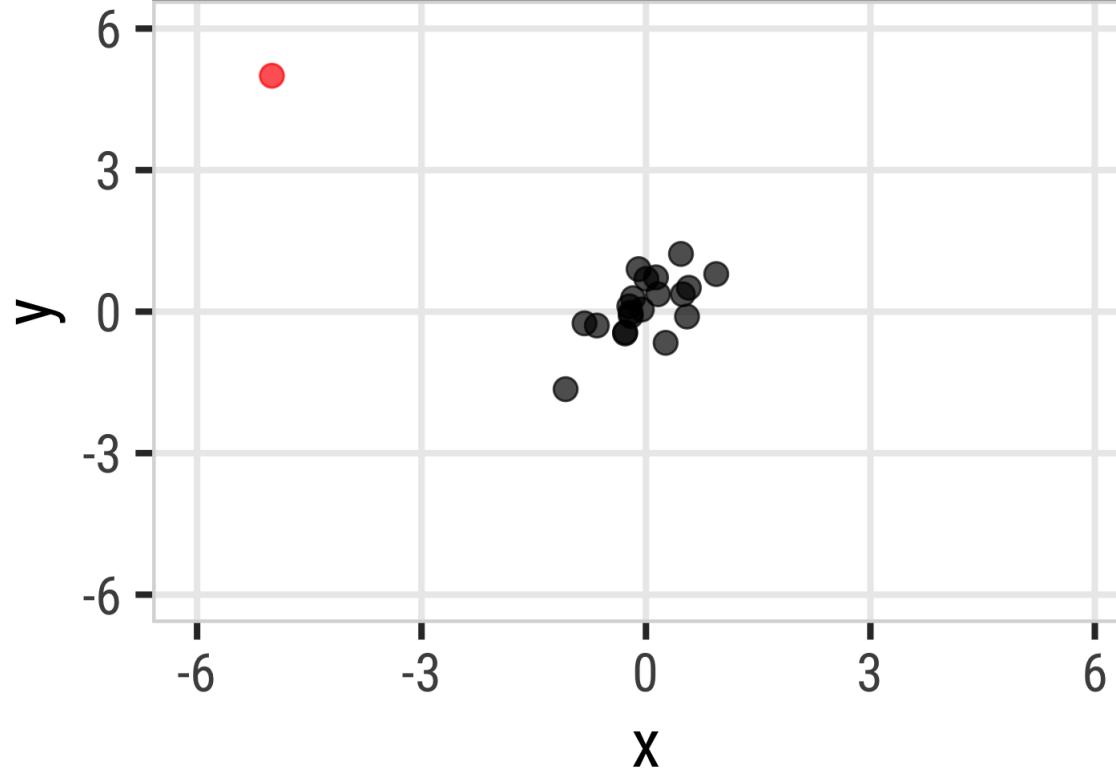
# Outliers



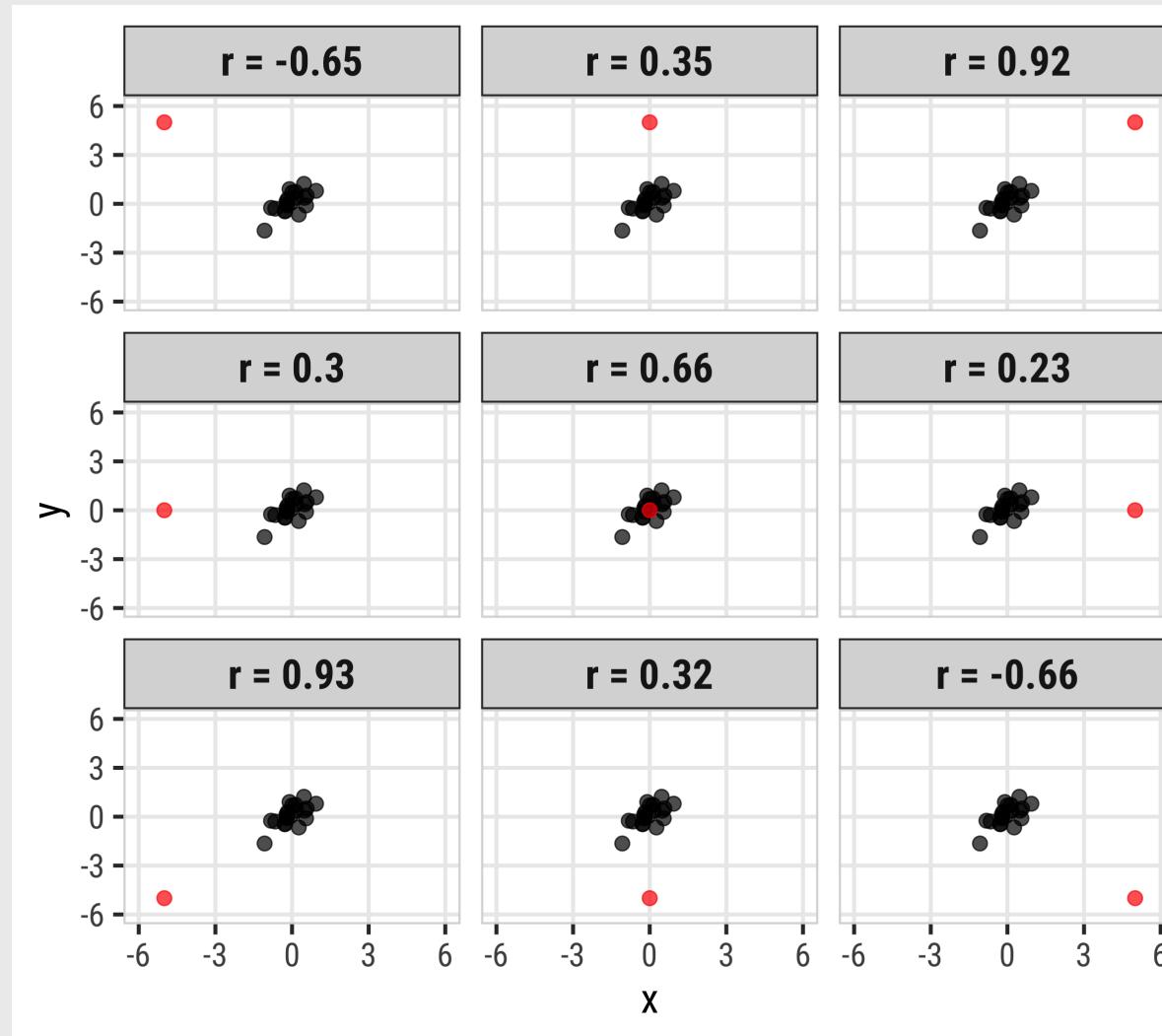




**$r = -0.65$**



# Pearson correlation is highly sensitive to outliers



# Spearman's rank-order correlation

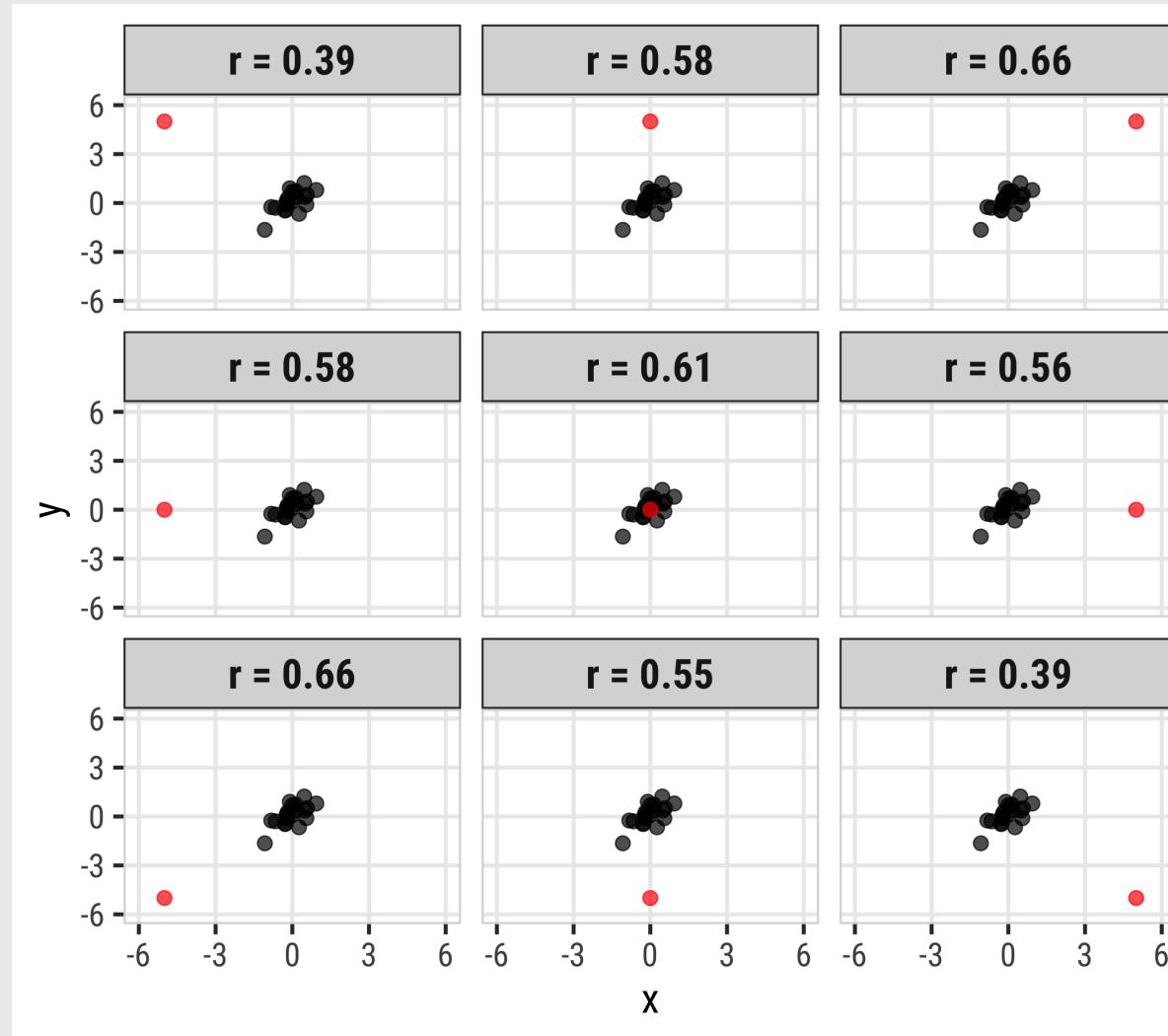
$$r = \frac{\text{Cov}(x,y)}{\text{sd}(x)*\text{sd}(y)}$$

- Separately rank the values of X & Y.
- Use Pearson's correlation on the *ranks* instead of the  $x$  &  $y$  values.

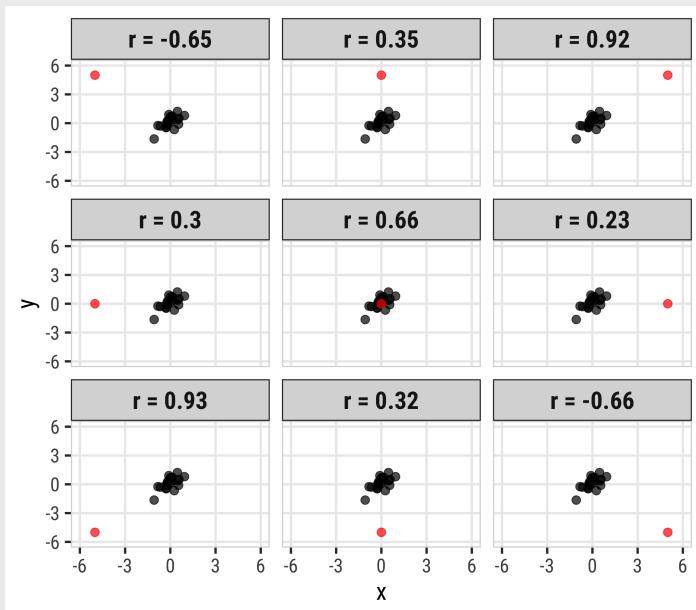
Assumptions:

- Variables can be ordinal, interval or ratio
- Relationship must be monotonic (i.e. does not require linearity)

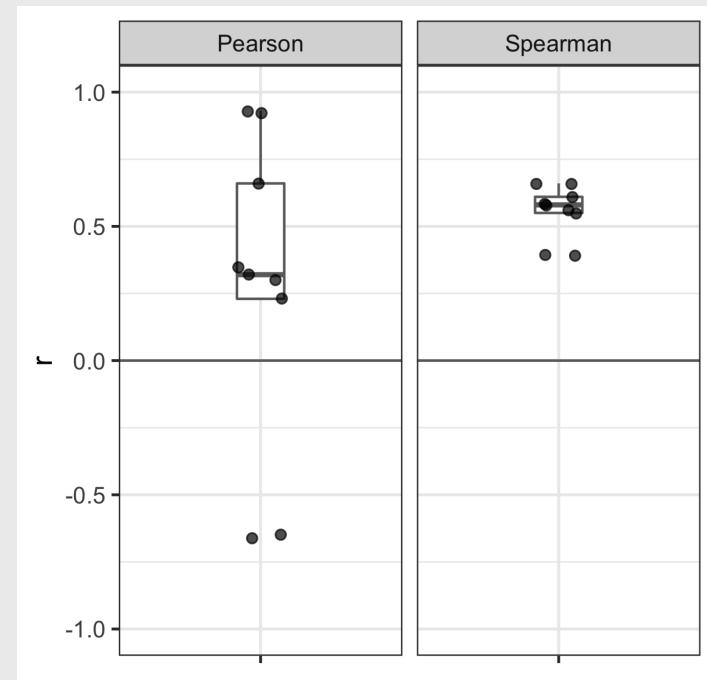
# Spearman correlation more robust to outliers



# Spearman correlation more robust to outliers



	Pearson	Spearman
$r = -0.65$	-0.56	0.53
$r = 0.35$	0.39	0.69
$r = 0.92$	0.94	0.81
$\rho = 0.3$	0.38	0.76
$\rho = 0.66$	0.81	0.79
$\rho = 0.23$	0.31	0.70
$\rho = 0.93$	0.95	0.81
$\rho = 0.32$	0.51	0.75
$\rho = -0.66$	-0.56	0.53



# Summary of correlation

- **Pearon's correlation:** Described the strength of a **linear** relationship between two variables that are interval or ratio in nature.
- **Spearman's rank-order correlation:** Describes the strength of a **monotonic** relationship between two variables that are ordinal, interval, or ratio. **It is more robust to outliers.**
- The **coefficient of determination** ( $r^2$ ) describes the amount of variance in one variable that is explained by the other variable.
- **Correlation != Causation**

R command (hint: add `use = "complete.obs"` to drop NA values)

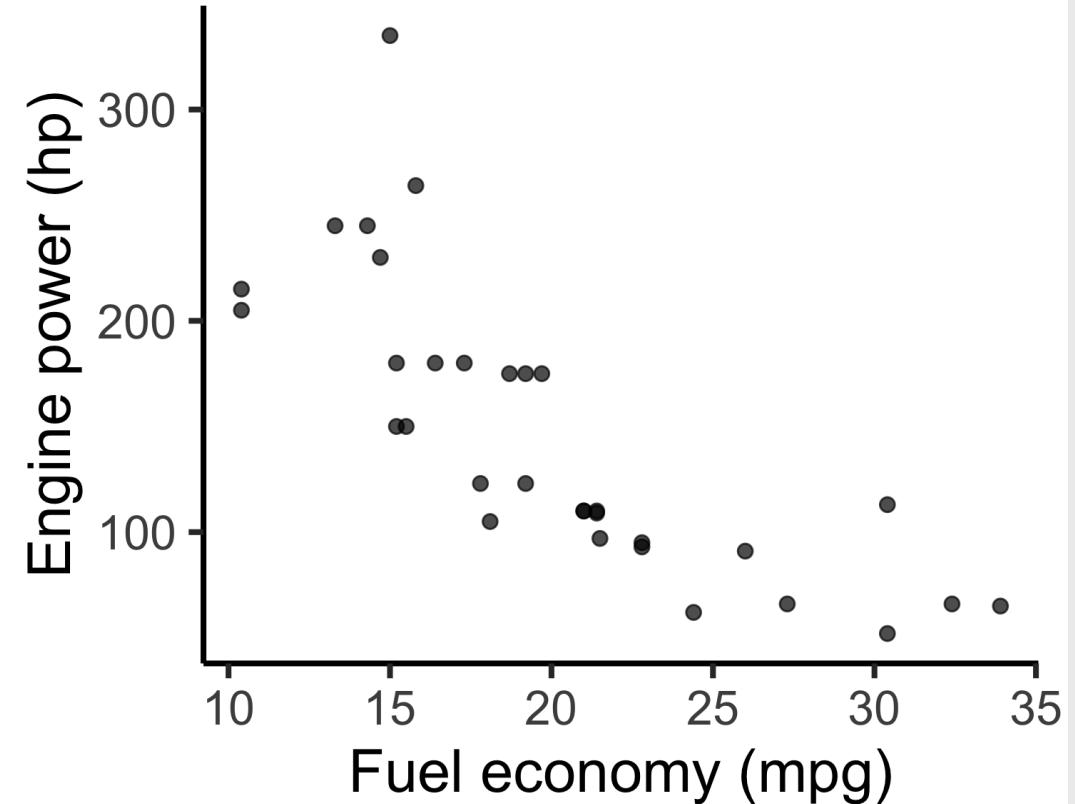
```
pearson <- cor(x, y, method = "pearson", use = "complete.obs")
spearman <- cor(x, y, method = "spearman", use = "complete.obs")
```

# Week 4: *Correlation*

1. What is correlation?
  2. Visualizing correlation
- BREAK
3. Linear models
  4. Visualizing linear models

# Scatterplots: The correlation workhorse

```
scatterplot <- ggplot(mtcars) +  
  geom_point(aes(x = mpg, y = hp),  
             size = 2, alpha = 0.7) +  
  theme_classic(base_size = 20) +  
  labs(x = 'Fuel economy (mpg)',  
       y = 'Engine power (hp)')  
  
scatterplot
```



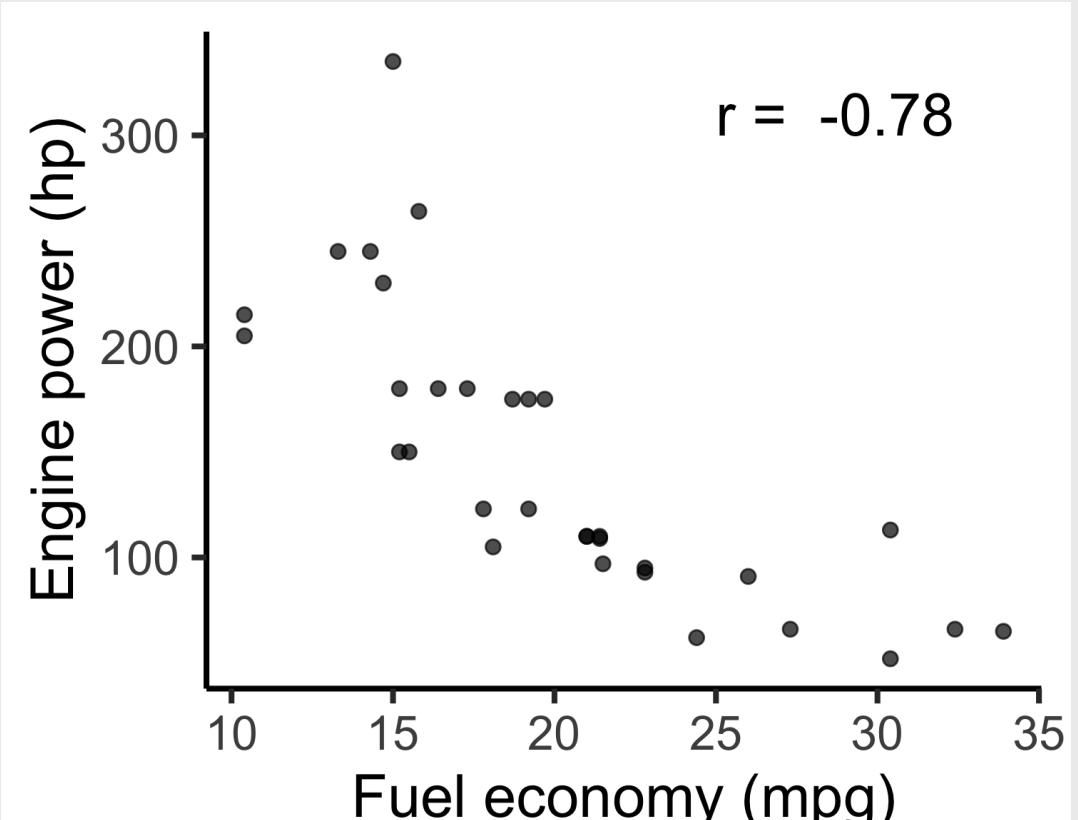
# Adding a correlation label to a chart

Make the correlation label

```
corr <- cor(  
  mtcars$mpg, mtcars$hp,  
  method = 'pearson')  
  
corrLabel <- paste('r = ', round(corr, 2))
```

Add label to the chart with `annotate()`

```
scatterplot +  
  annotate(geom = 'text',  
    x = 25, y = 310,  
    label = corrLabel,  
    hjust = 0, size = 7)
```



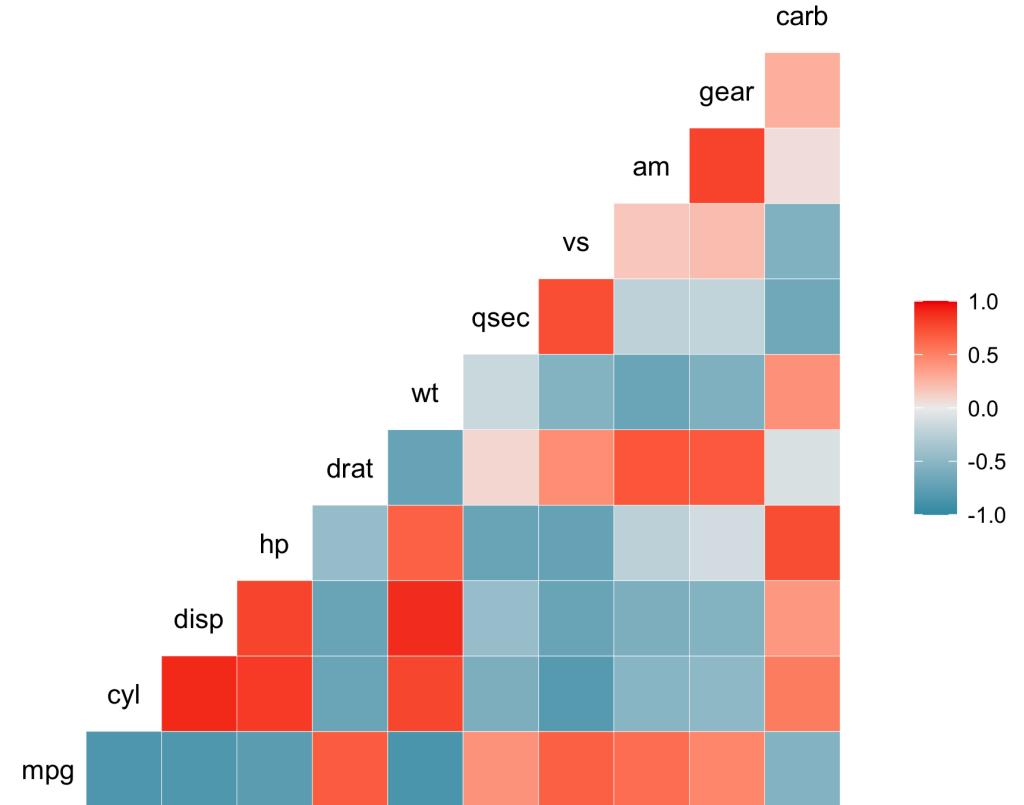
**Visualize all the correlations**



# Visualize all the correlations: `ggcorr()`

```
library('GGally')
```

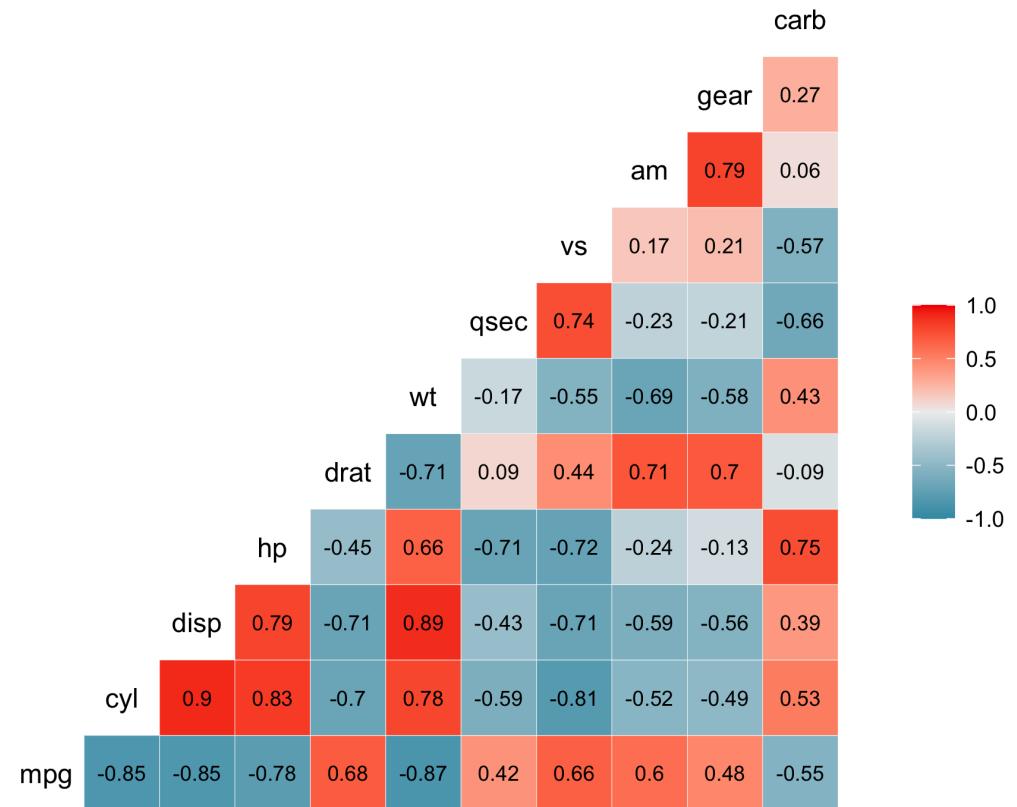
```
mtcars %>%  
  ggcorr()
```



# Visualizing correlations: ggcorr()

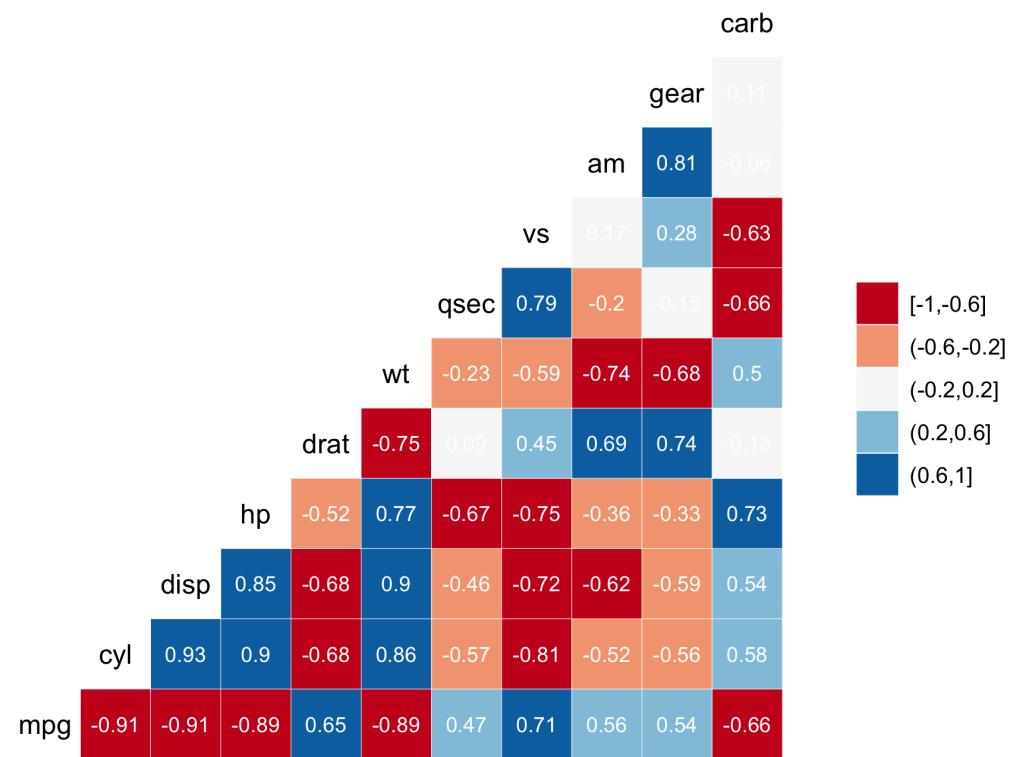
```
library('GGally')
```

```
mtcars %>%
  ggcorr(label = TRUE,
         label_size = 3,
         label_round = 2)
```



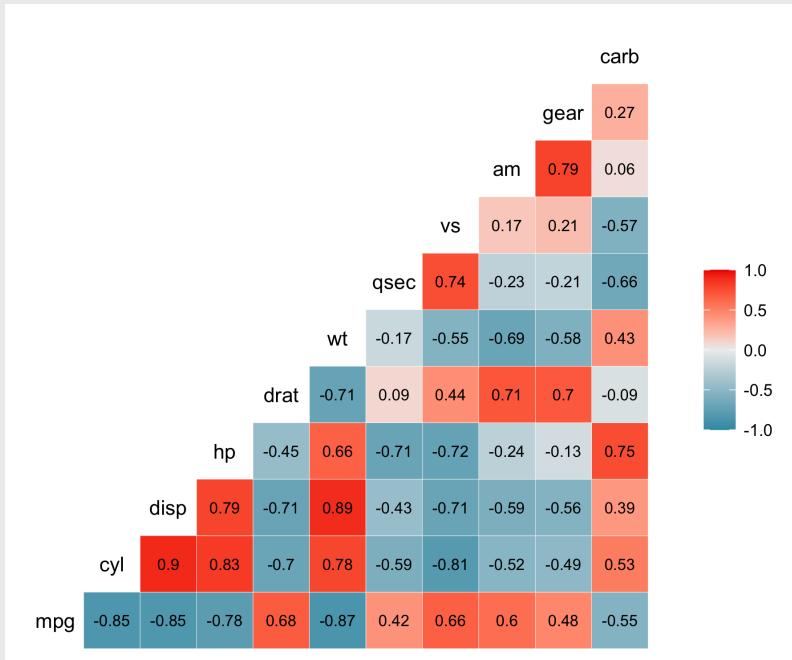
# Visualizing correlations: ggcorr()

```
ggcor_mtcars_final <- mtcars %>%
  ggcorr(label = TRUE,
         label_size = 3,
         label_round = 2,
         label_color = 'white',
         nbreaks = 5,
         palette = "RdBu")
```



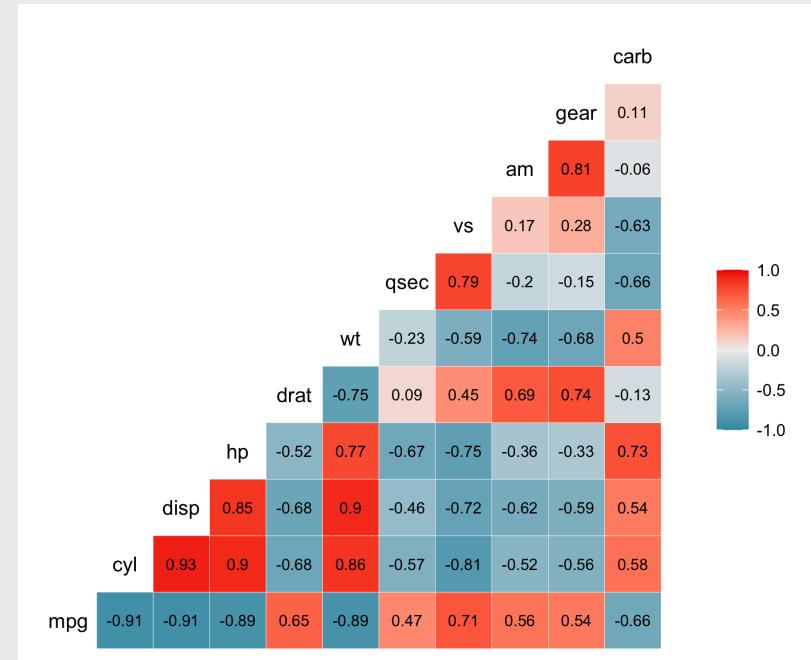
# Pearson

```
mtcars %>%  
  ggcorr(label = TRUE,  
         label_size = 3,  
         label_round = 2,  
         method = c("pairwise", "pearson"))
```



# Spearman

```
mtcars %>%  
  ggcorr(label = TRUE,  
         label_size = 3,  
         label_round = 2,  
         method = c("pairwise", "spearman"))
```

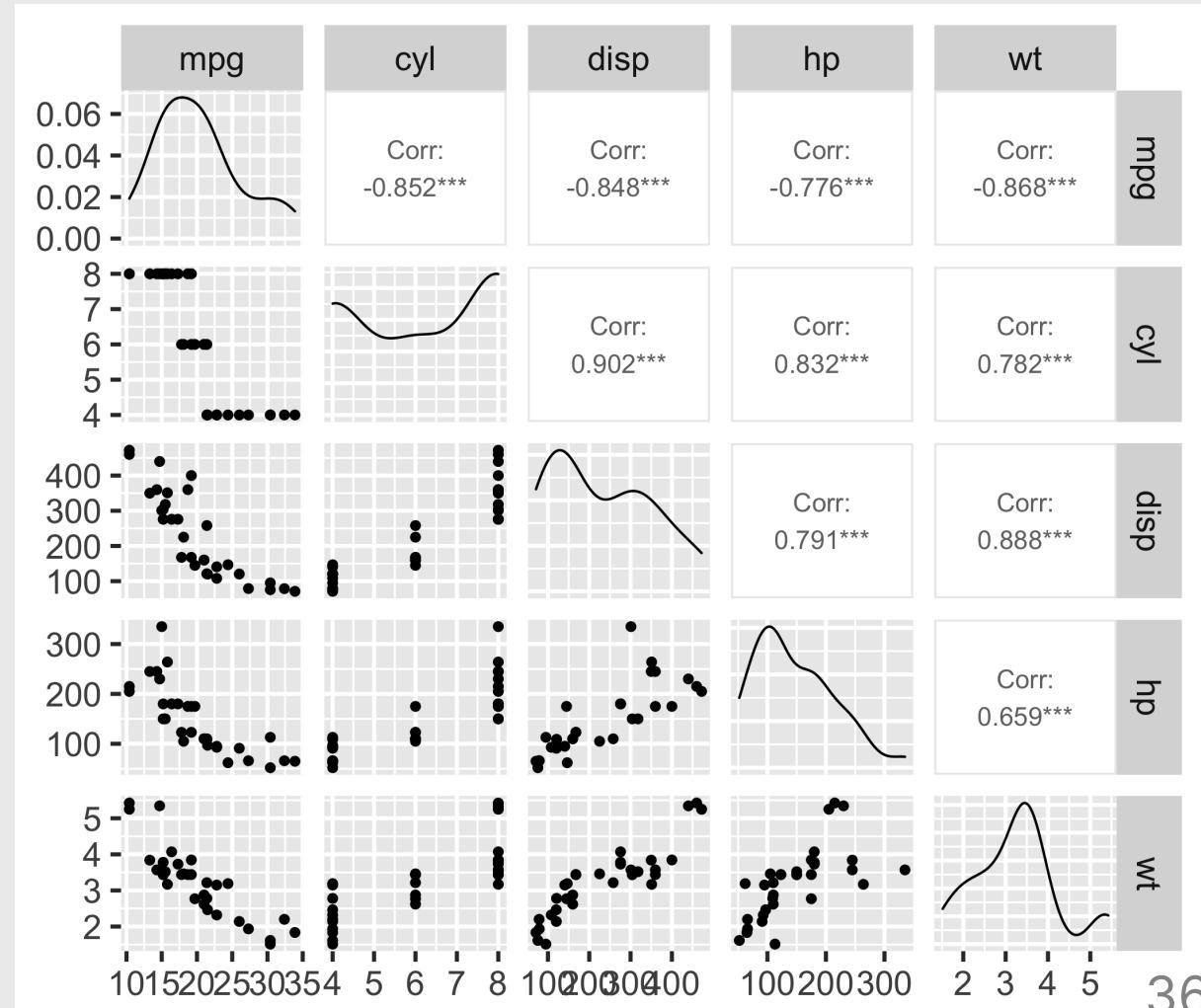


# Correlograms: `ggpairs()`

```
library('GGally')
```

```
mtcars %>%
  select(mpg, cyl, disp, hp, wt)
ggpairs()
```

- Look for linear relationships
- View distribution of each variable

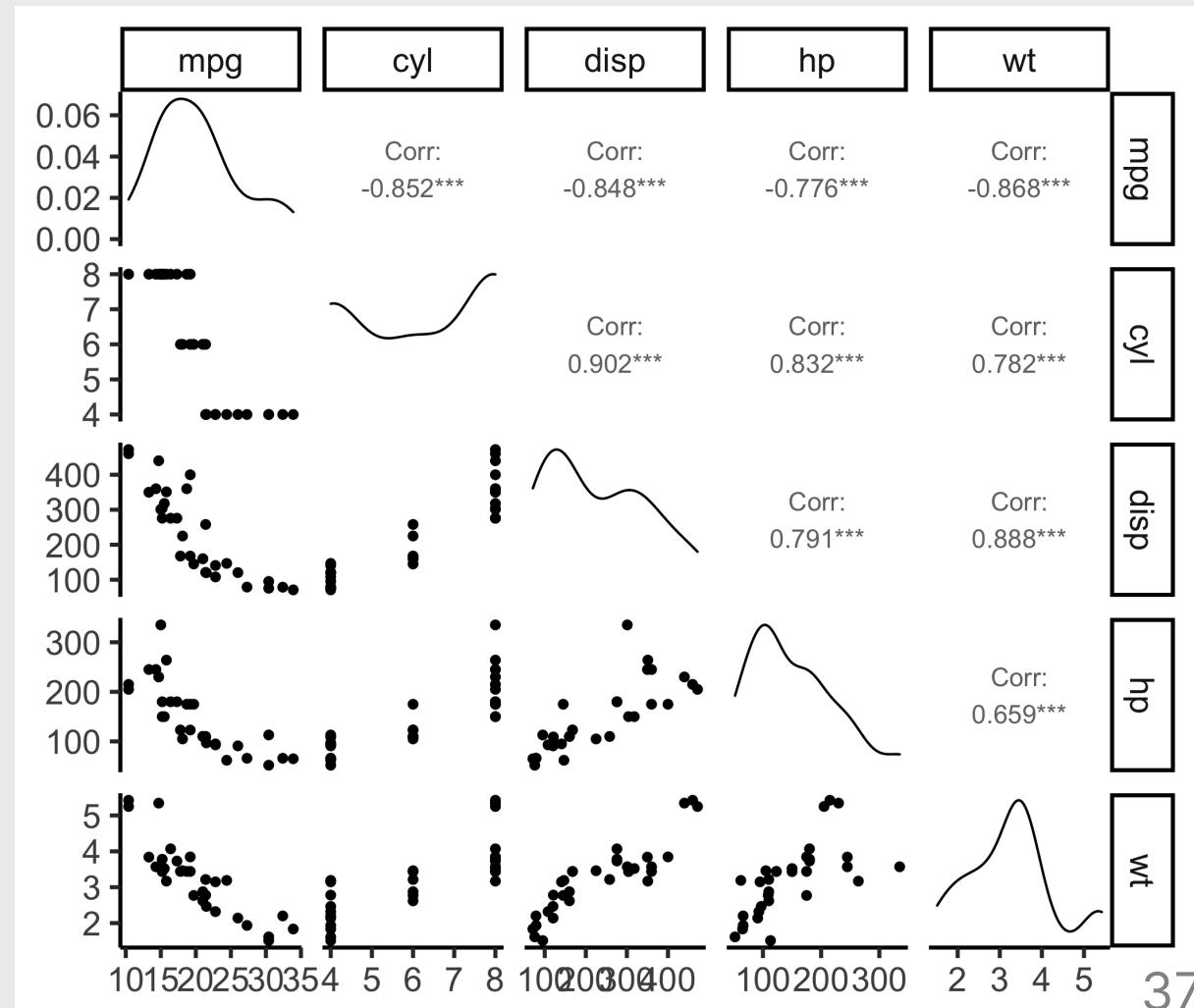


# Correlograms: `ggpairs()`

```
library('GGally')
```

```
mtcars %>%
  select(mpg, cyl, disp, hp, wt)
ggpairs() +
  theme_classic()
```

- Look for linear relationships
- View distribution of each variable



# Your turn

```
<div class="countdown blink-colon noudate-15" id="timer_6016f3eb" style="top:0;right:0;fo  
<code class="countdown-time"><span class="countdown-digits minutes">15</span><span class='</div>
```

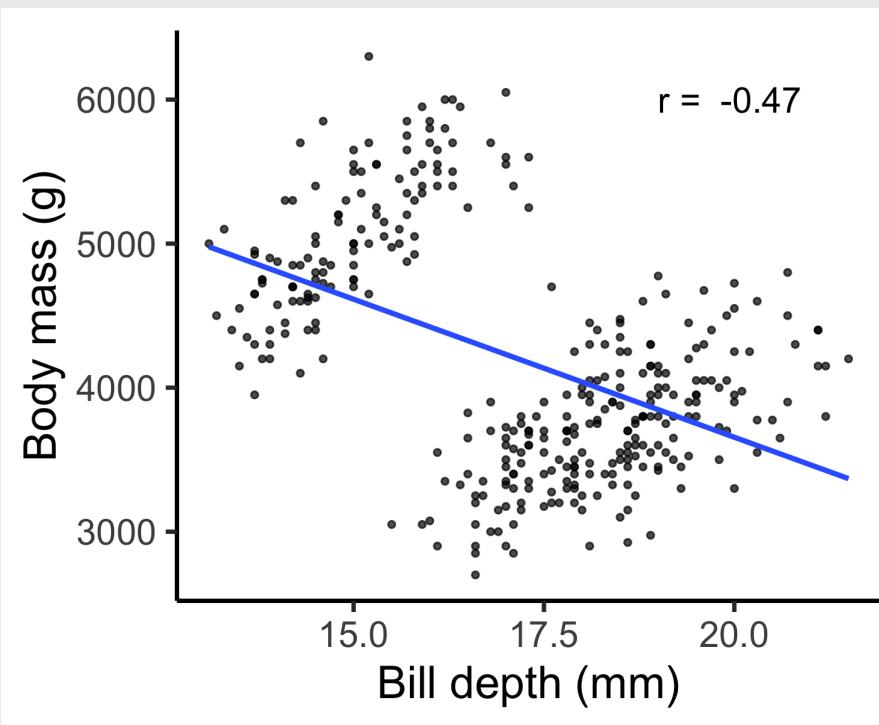
Using the **penguins** data frame:

[palmerpenguins library](#)

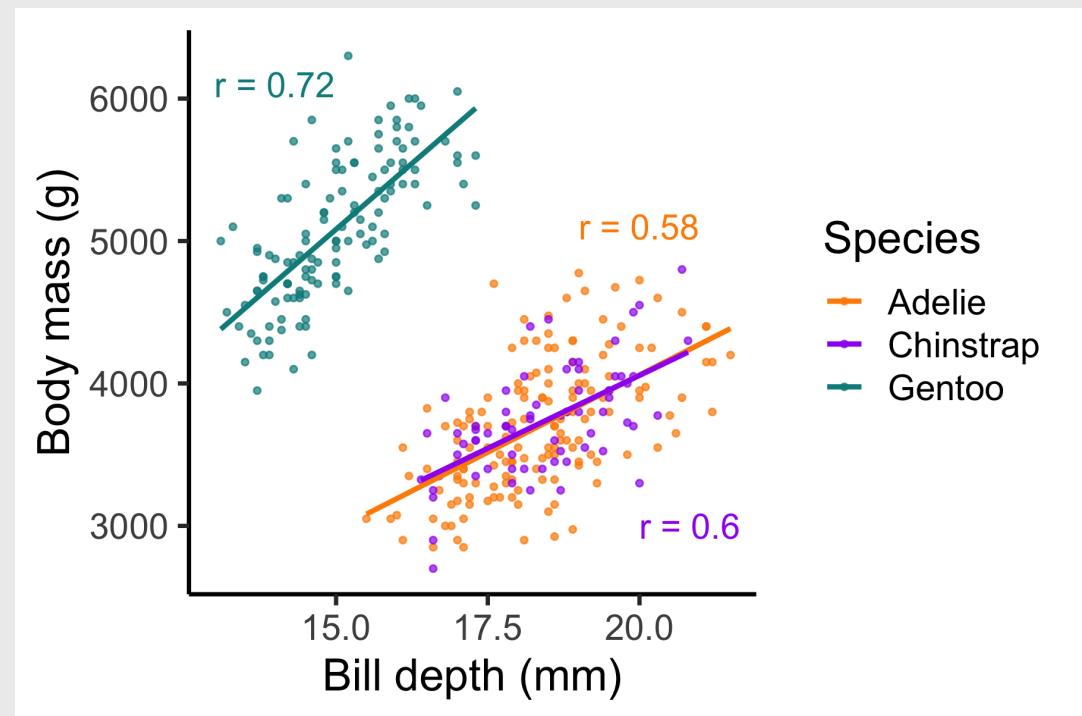
1. Find the two variables with the largest correlation in absolute value (i.e. closest to -1 or 1).
2. Create a scatter plot of those two variables.
3. Add an annotation for the Pearson correlation coefficient.

# Simpson's Paradox: when correlation betrays you

**Body mass vs. Bill depth**



**Body mass vs. Bill depth**



```
<div class="countdown" id="timer_6016f53c" style="top:1;right:0;bottom:0;left:0;margin:5%;<br><code class="countdown-time"><span class="countdown-digits minutes">05</span><span class="countdown-digits seconds">00</span></code></div>
```

# Break!

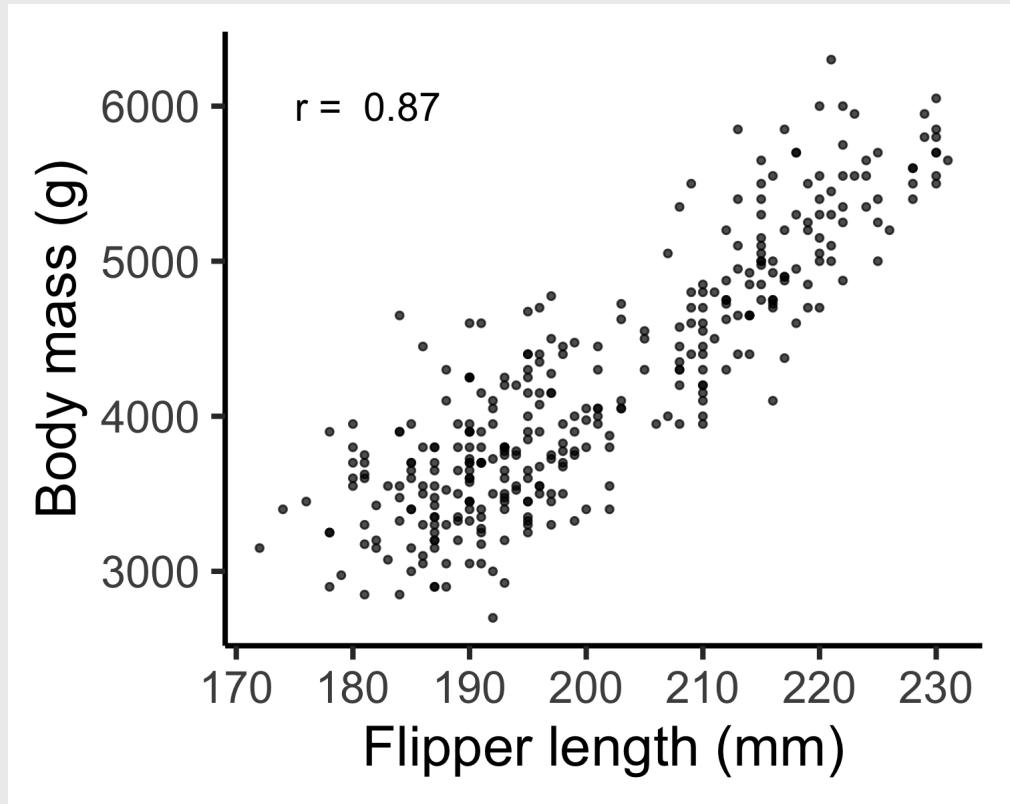
Stand up, Move around, Stretch!

# Week 4: *Correlation*

1. What is correlation?
  2. Visualizing correlation
- BREAK
3. Linear models
  4. Visualizing linear models

# Palmer Penguins

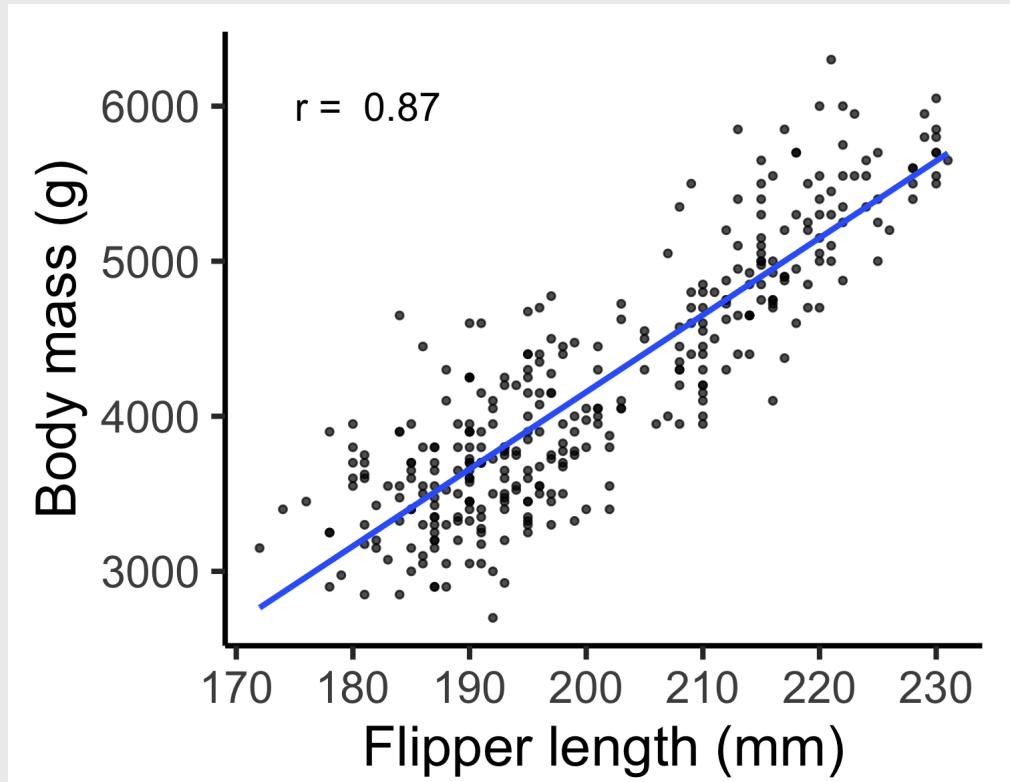
The correlation of 0.87 means that the body mass (g) explains about 75% of the variation in the flipper length (mm).



# Palmer Penguins

The correlation of 0.87 means that the body mass (g) explains about 75% of the variation in the flipper length (mm).

**Now let's fit a model to these points!**

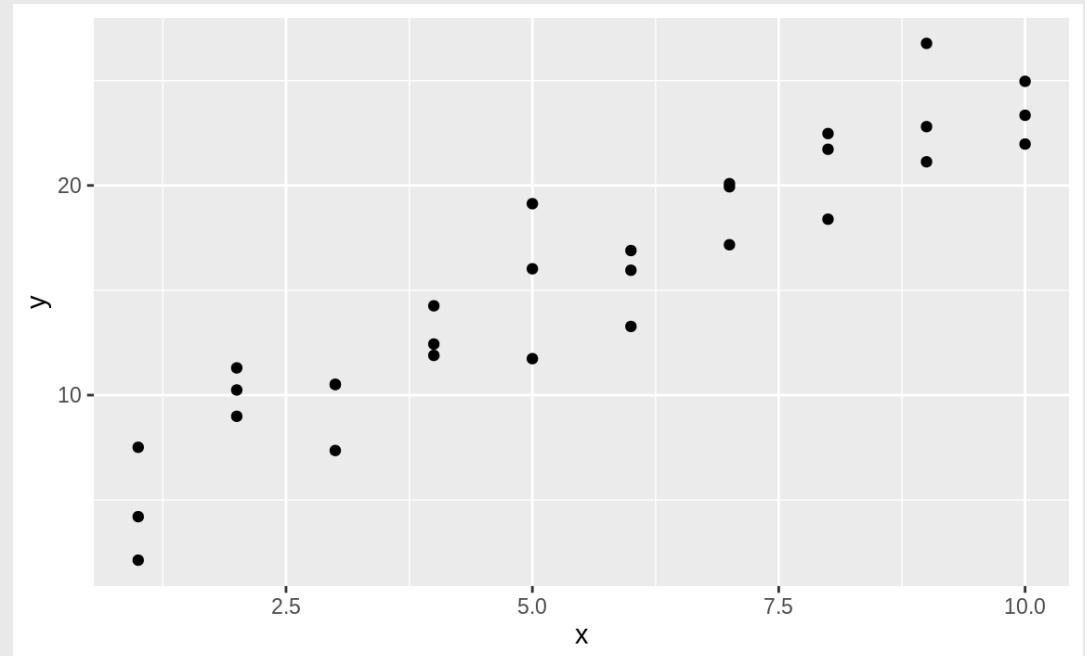


# Modeling basics

Two parts to a model:

1. **Model family:** e.g.,  $y = ax + b$
2. **Fitted model:** e.g.,  $y = 3x + 7$

Here is some simulated data



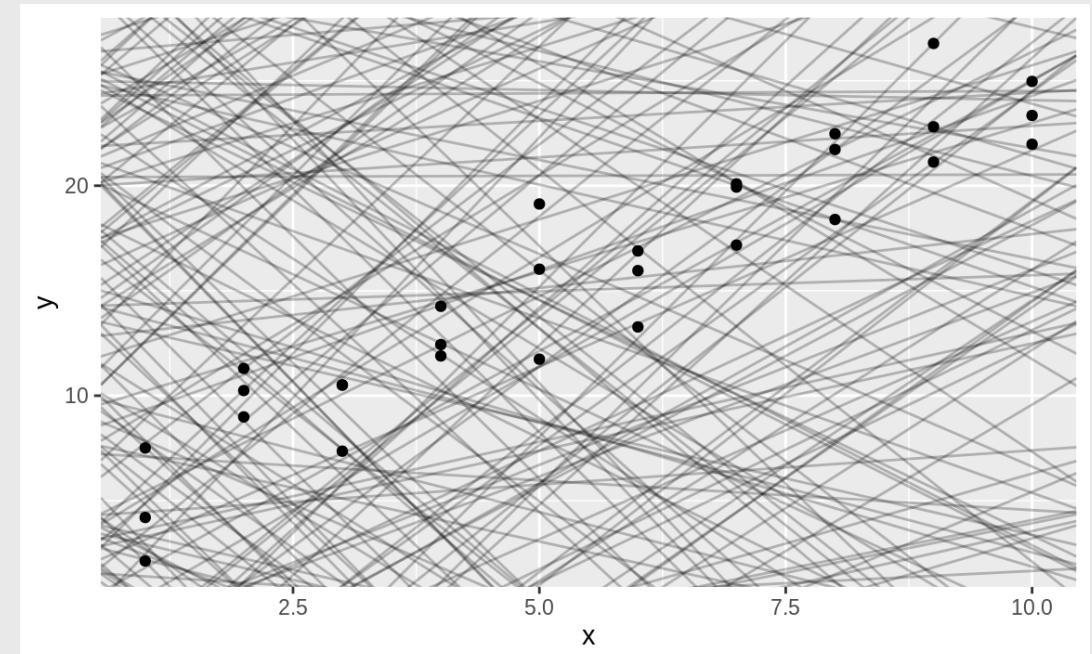
# Modeling basics

Two parts to a model:

1. **Model family:** linear model:

$$y = ax + b$$

There are an infinite number of possible models



# Modeling basics

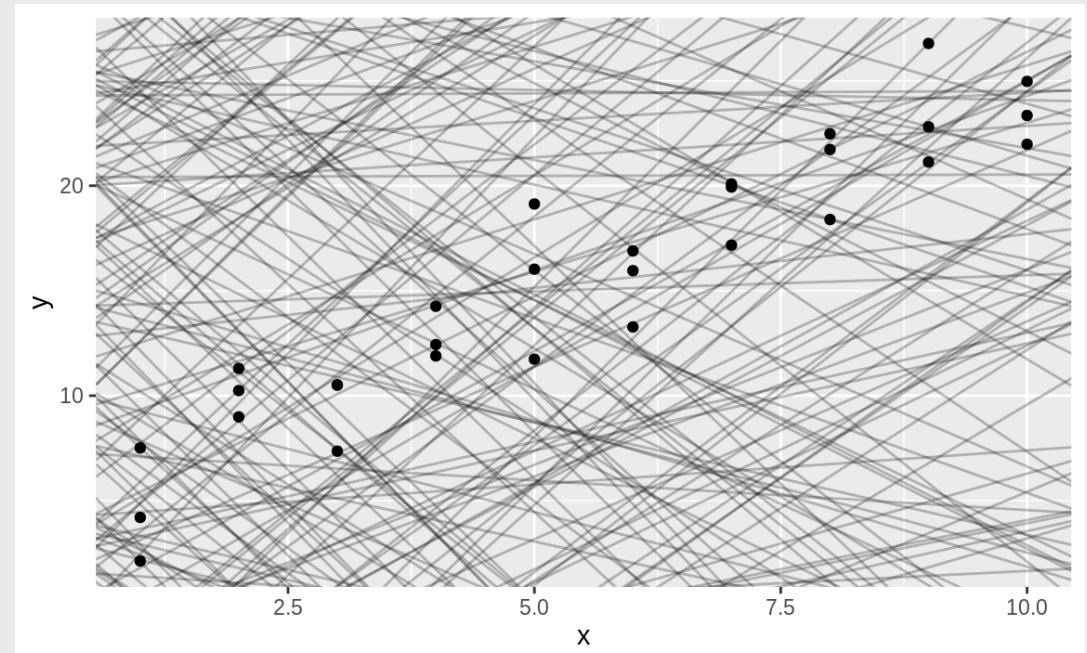
Two parts to a model:

1. **Model family:** linear model:

$$y = ax + b$$

2. **Fitted model:** How to choose the "best"  $a$  and  $b$ ?

There are an infinite number of possible models



# Modeling basics

Two parts to a model:

1. **Model family:** linear model:

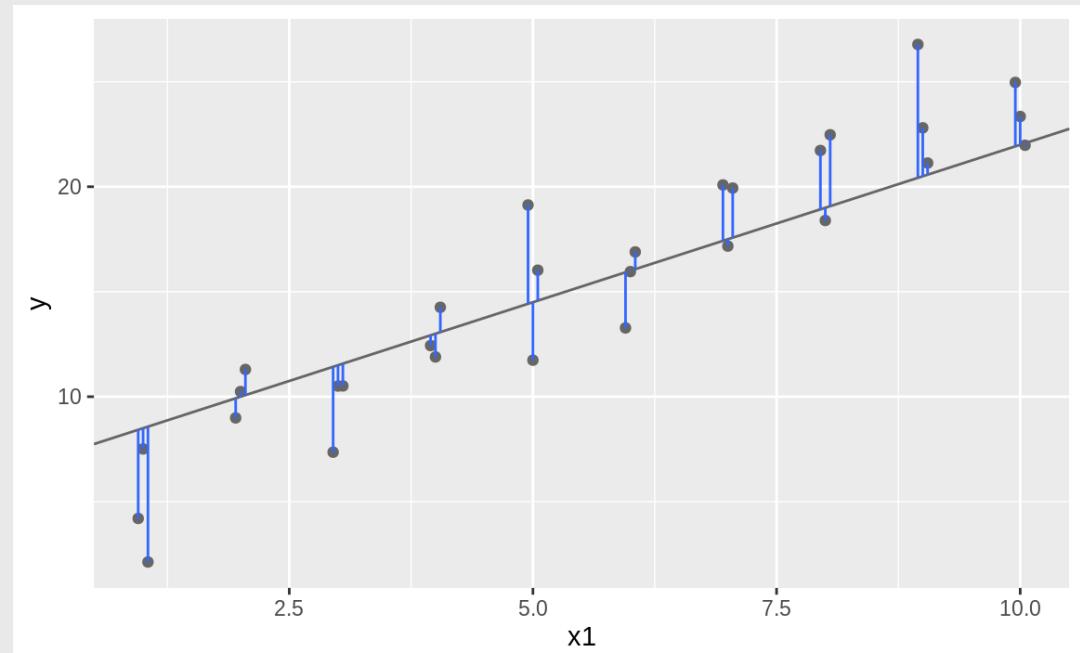
$$y = ax + b$$

2. **Fitted model:** How to choose the "best"  $a$  and  $b$ ?

*We need to come up with some measure of "distance" from the model to the data*

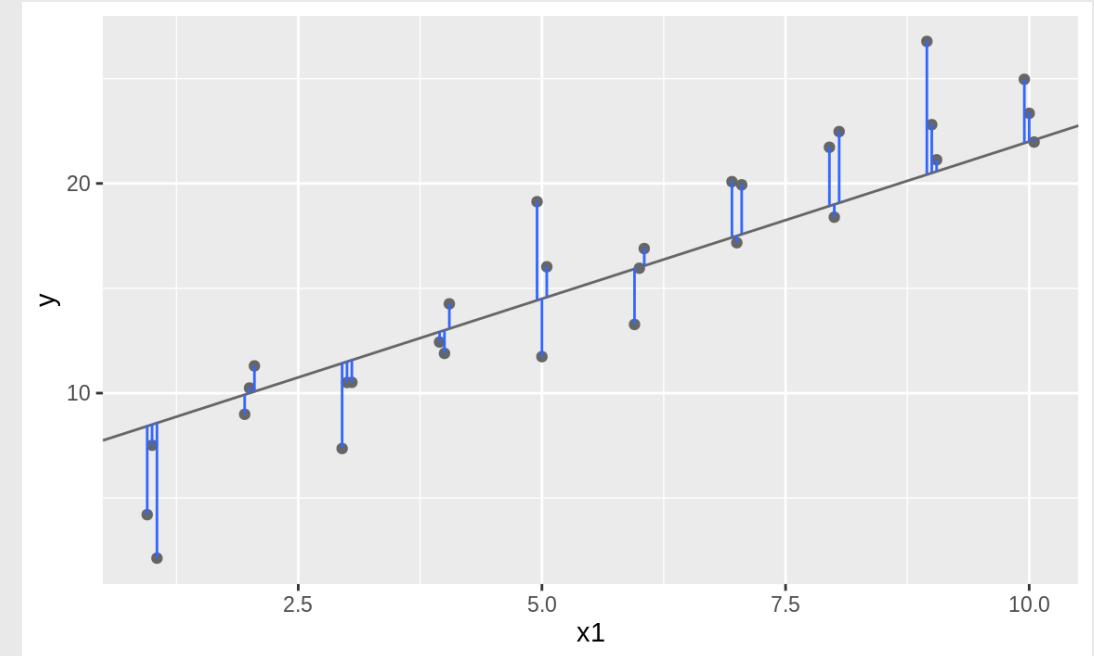
Compute the "**residuals**:

The distance between the model line and the data



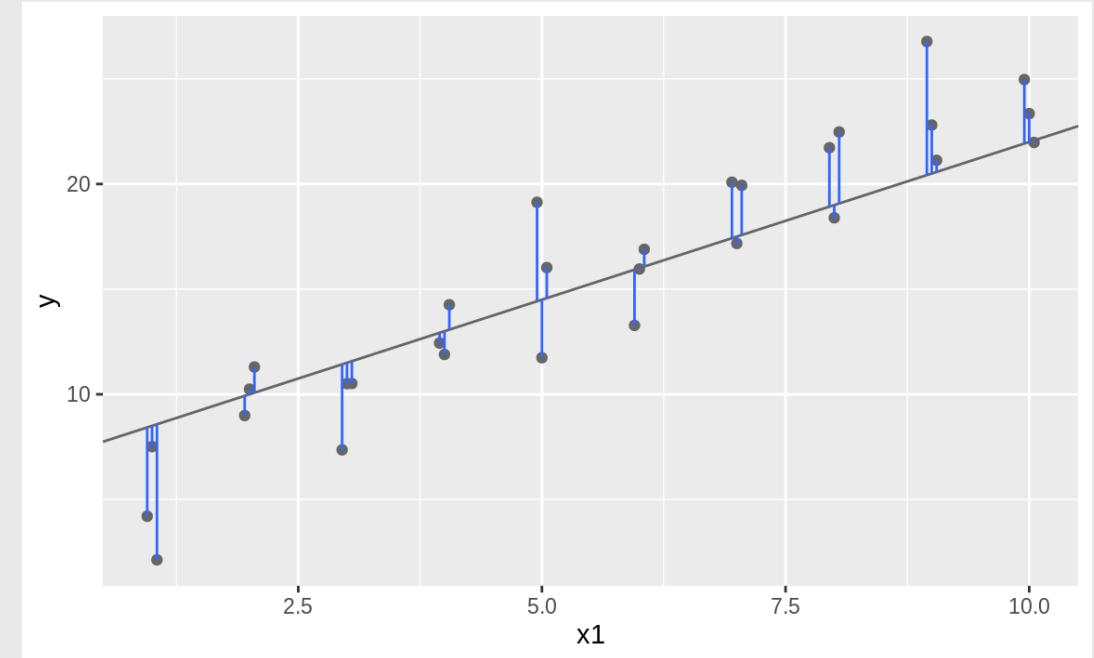
Residual:  $y_i - \hat{y}'_i$

**Residual:** The distance between the model line and the data



$$\text{Sum of squared residuals: } \text{SSR} = \sum_{i=1}^n (y_i - y'_i)^2$$

**Residual:** The distance between the model line and the data



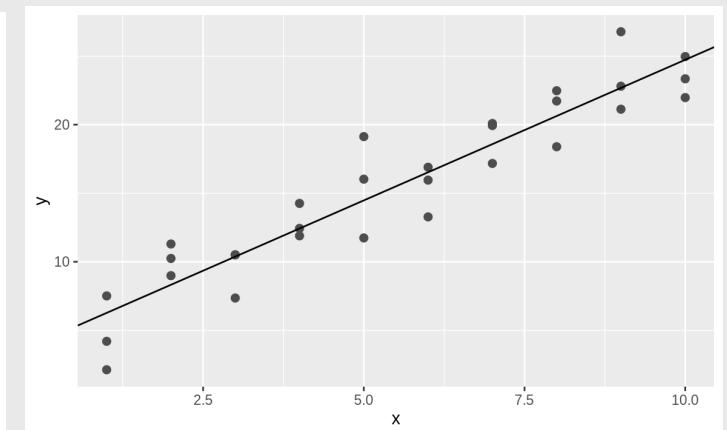
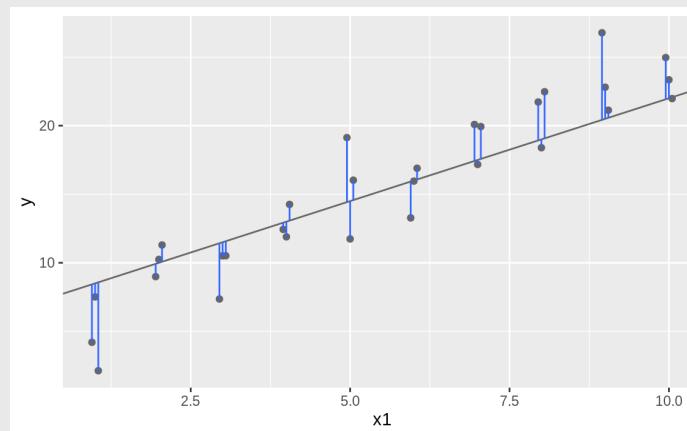
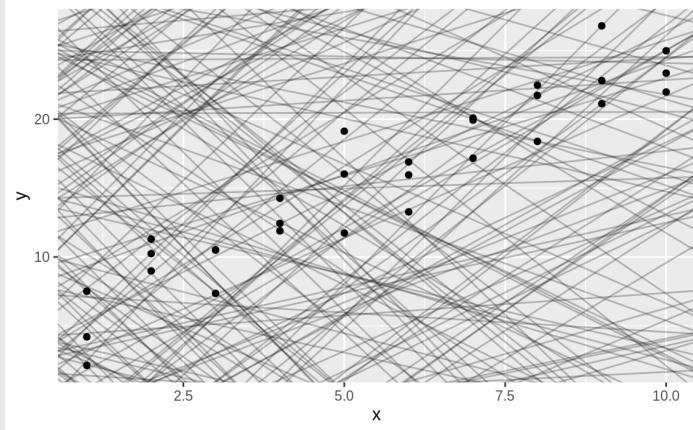
# Search algorithm

1): Pick a model (  $a$  and  $b$  ): 2): Compute the SSR:

$$y = ax + b$$

$$\text{SSR} = \sum_{i=1}^n (y_i - y'_i)^2$$

3): Repeat steps 1 & 2 until the smallest SSR is found



# Fitting a linear model in R

```
model <- lm(formula = y ~ x, data = data)
```

Example: Penguin data

```
model <- lm(  
  formula = body_mass_g ~ flipper_length_mm,  
  data     = penguins)
```

Get coefficients (  $a$  &  $b$  in  $y = ax + b$  )

```
coef(model)
```

```
#>      (Intercept) flipper_length_mm  
#> -5780.83136          49.68557
```

# Fitting a linear model in R

```
model <- lm(formula = y ~ x,  
            data = data)
```

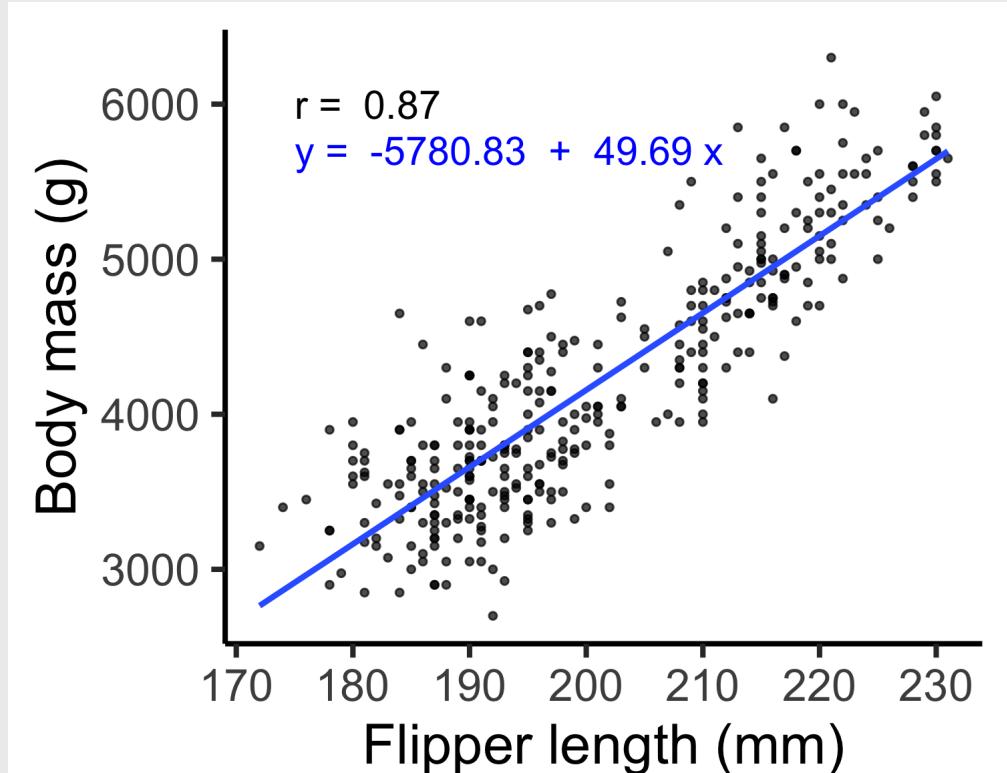
Example: Penguin data

```
model <- lm(  
  formula = body_mass_g ~ flipper_length_mm  
  data    = penguins)
```

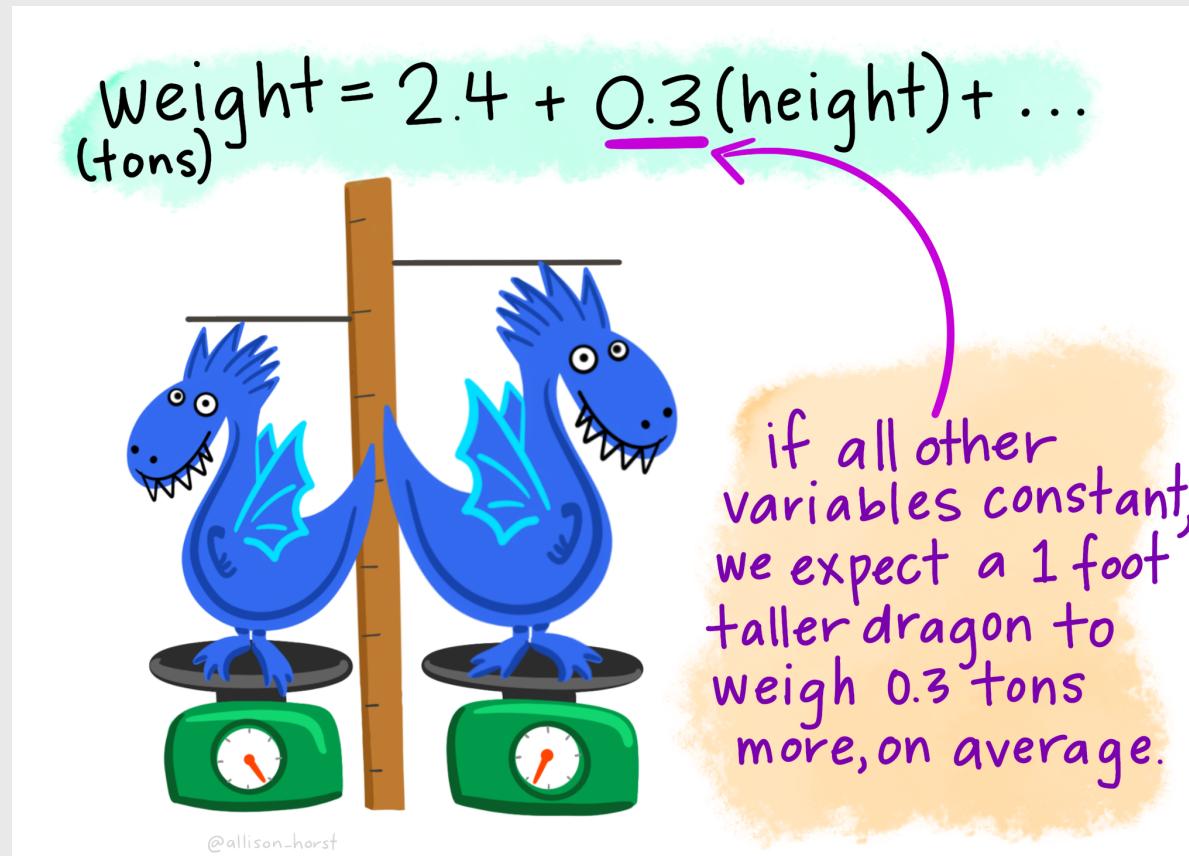
Get coefficients

```
coef(model)
```

```
#> (Intercept) flipper_length_mm  
#> -5780.83136      49.68557
```



# Interpreting results

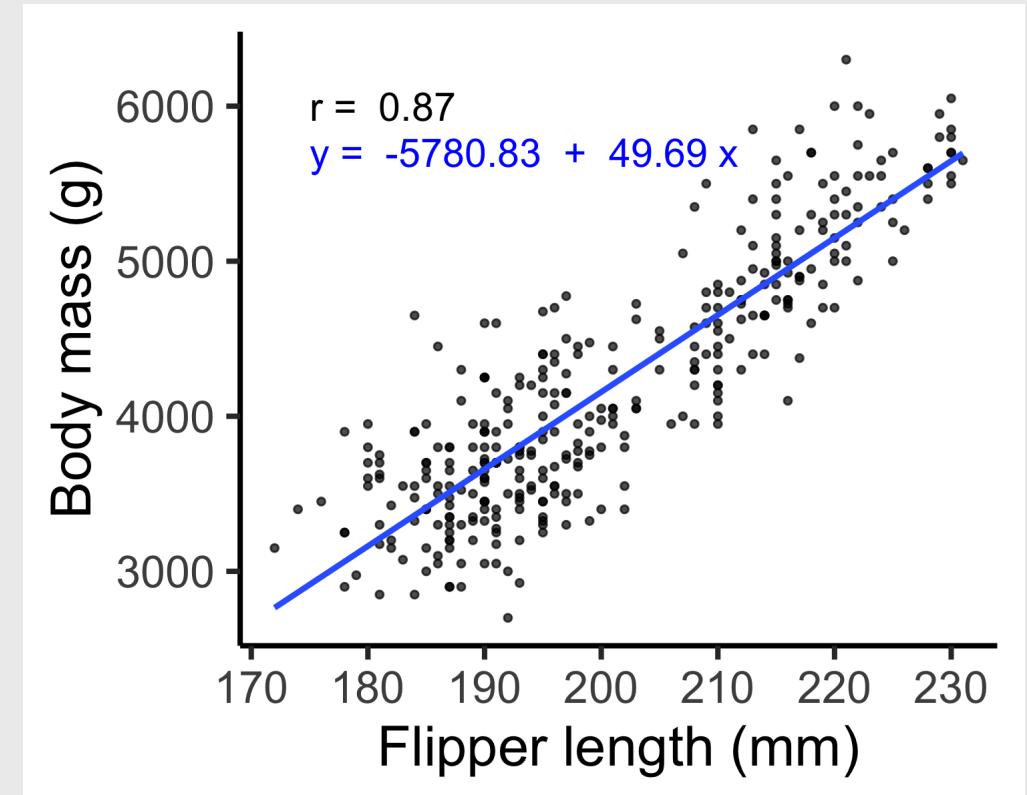


Artwork by [@allison\\_horst](#)

# Example write up for Penguin data

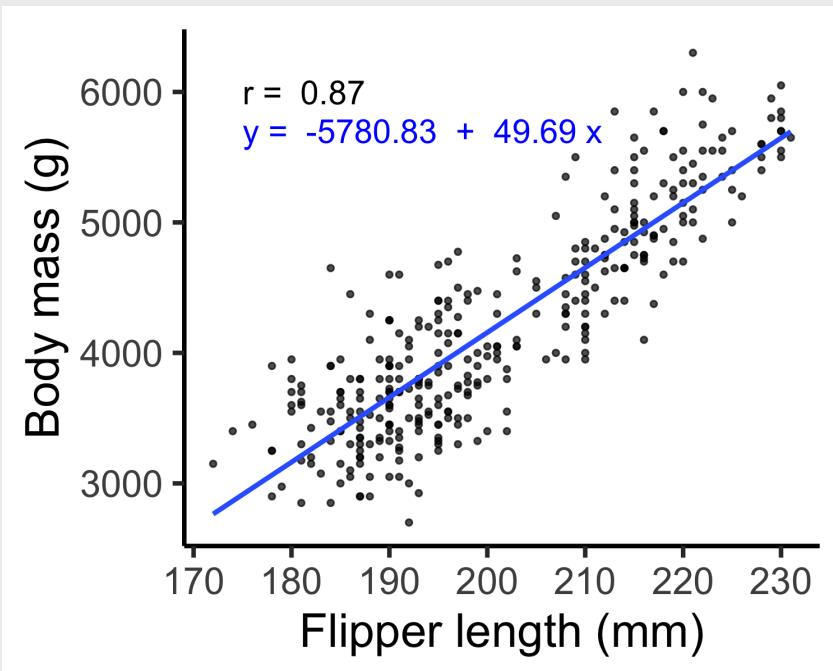
The correlation between flipper length (mm) and body mass (g) is **0.87**. Therefore, **~75%** of the variance in body mass is explained by flipper length.

The slope of the best fitting regression line indicates that body mass increased by **49.7 g** as flipper length increased by one mm.

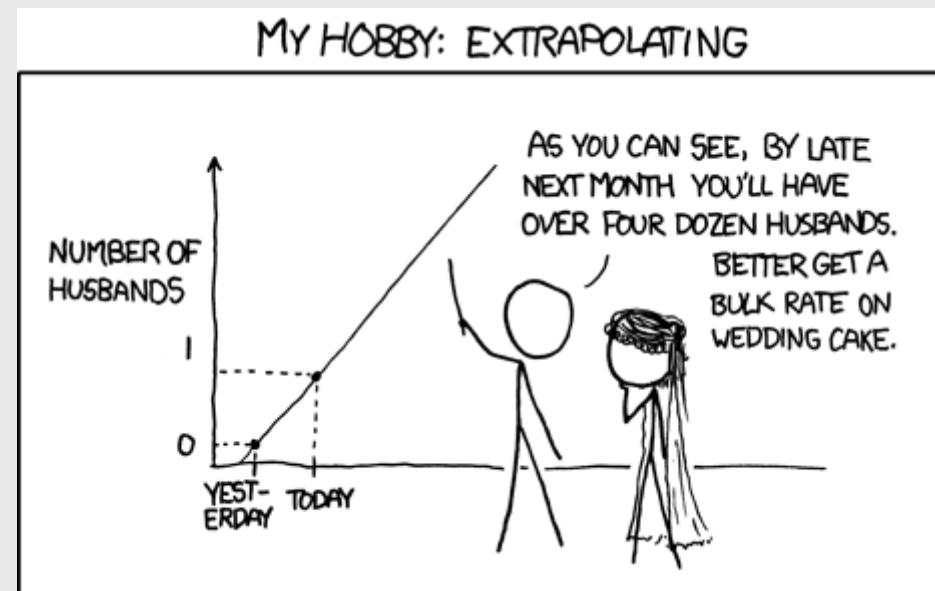


# Making predictions

**Interpolation is OK:** You may predict values of  $y$  for values of  $x$  that were not observed but are within the range of the observed values of  $x$ .



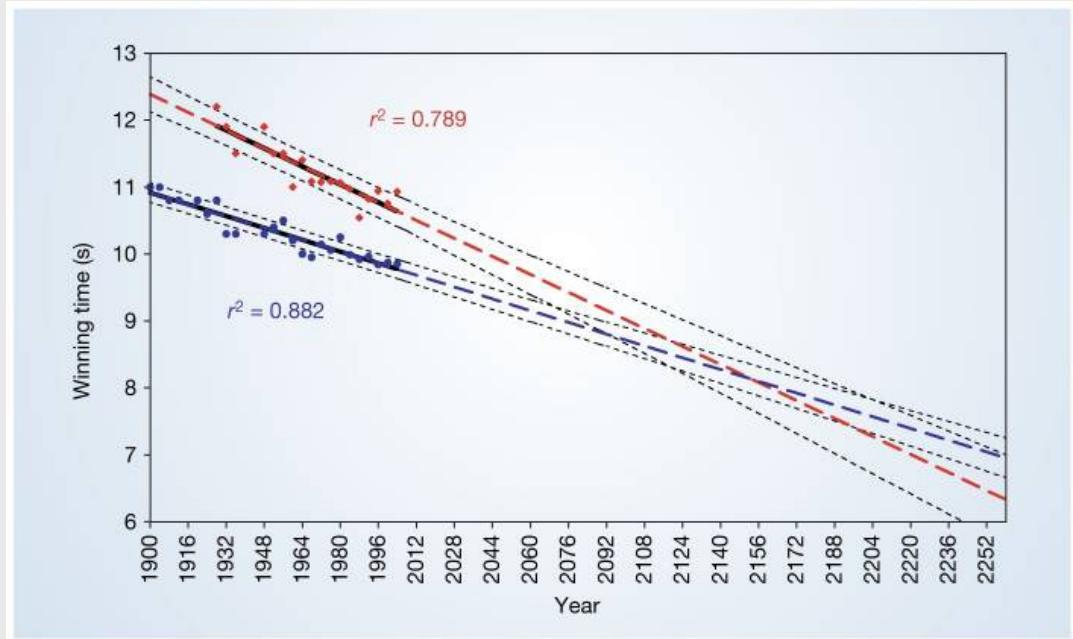
**Extrapolation is BAD:** You should NOT predict values of  $y$  using values of  $x$  that are outside the observed range of  $x$ .



xkcd

# Repeat: Extrapolation is **BAD**

"Extrapolation of these trends to the 2008 Olympiad indicates that the women's 100-metre race could be won in a time of  $10.57 \pm 0.232$  seconds and the men's event in  $9.73 \pm 0.144$  seconds. **Should these trends continue, the projections will intersect at the 2156 Olympics, when — for the first time ever — the winning women's 100-metre sprint time of 8.079 seconds will be lower than that of the men's winning time of 8.098 seconds (Fig. 1).**"



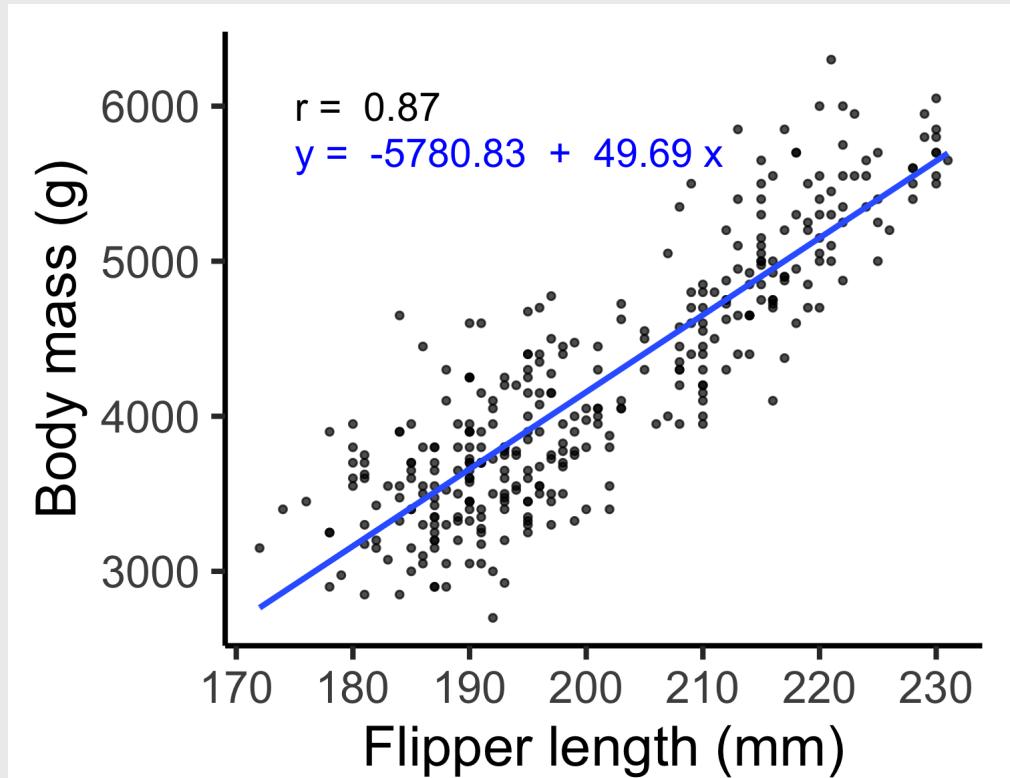
Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Momentous sprint at the 2156 Olympics? *Nature*, 431(7008), 525-525. [View online](#)

# Symantics

These all mean the same thing:

- "Use X to predict Y"
- "Regress Y on X"
- "Regression of Y on X"

```
model <- lm(formula = y ~ x,  
            data = data)
```



# Symantics

```
model <- lm(formula = y ~ x,  
            data = data)
```

## **Y: Dependent variable**

- Outcome variable
- Response variable
- Regressand
- Left-hand variable

## **X: Independent variable**

- Predictor variable
- Explanatory variable
- Regressor
- Right-hand variable

# Week 4: *Correlation*

1. What is correlation?
2. Visualizing correlation
3. Linear models
4. Visualizing linear models

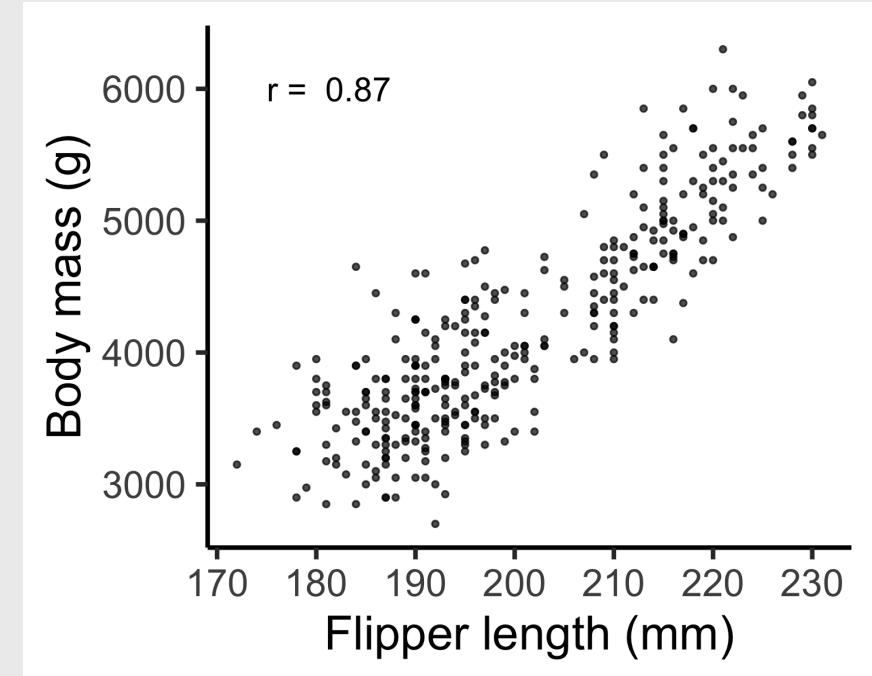
BREAK

# Adding the correlation annotation

```
# Make the correlation label
corr <- cor(
  penguins$body_mass_g, penguins$flipper_length_mm,
  method = 'pearson', use = "complete.obs")

corrLabel <- paste("r = ", round(corr, 2))

# Make the chart!
ggplot(penguins, aes(x = flipper_length_mm, y = body_
  geom_point(size = 1, alpha = 0.7) +
  annotate(geom = 'text', x = 175, y = 6000,
    label = corrLabel,
    hjust = 0, size = 5) +
  theme_classic(base_size = 20) +
  labs(x = "Flipper length (mm)",
    y = "Body mass (g)")
```



```

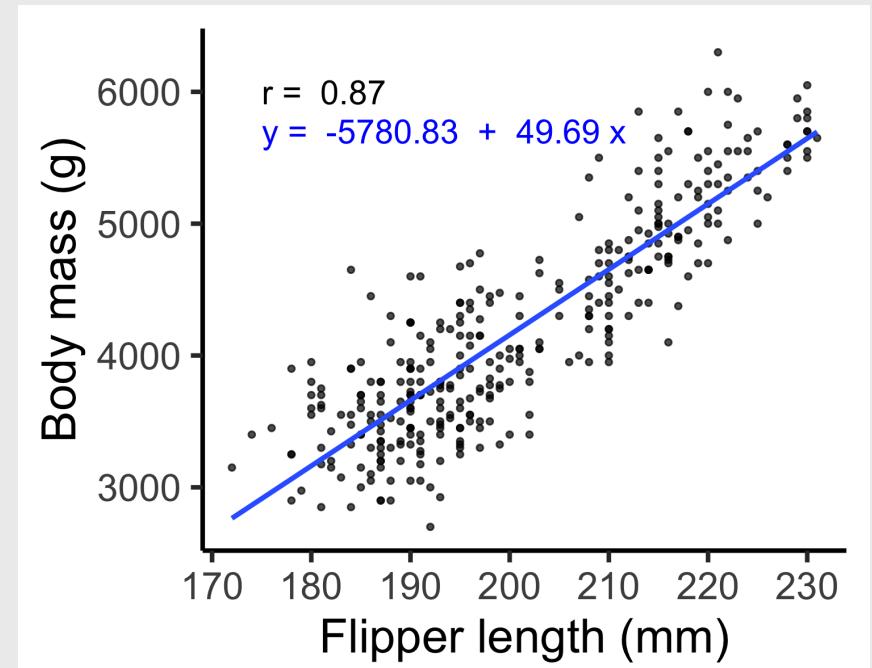
# Make correlation label
corrLabel <- paste("r = ", round(cor(
  penguins$body_mass_g, penguins$flipper_length_mm,
  method = 'pearson', use = "complete.obs"), 2))

# Make model label
model <- lm(
  formula = body_mass_g ~ flipper_length_mm,
  data    = penguins)
coefs <- round(coef(model), 2)
modelLabel <- paste('y = ', coefs[1], ' + ', coefs[2])

# Make the chart!
ggplot(penguins, aes(x = flipper_length_mm, y = body_
  geom_point(size = 1, alpha = 0.7) +
  geom_smooth(method = 'lm', se = FALSE) +
  annotate(geom = 'text', x = 175, y = 6000,
    label = corrLabel,
    hjust = 0, size = 5) +
  annotate(geom = 'text', x = 175, y = 5700,
    label = modelLabel, color = "blue",
    hjust = 0, size = 5)
  theme_classic(base_size = 20) +
  labs(x = "Flipper length (mm)",
    y = "Body mass (g)")

```

## Add correlation + model



# Your turn

```
<div class="countdown blink-colon noudate-15" id="timer_6016f19b" style="top:0;right:0;fo<code class="countdown-time"><span class="countdown-digits minutes">15</span><span class='</div>
```

Using the `msleep` data frame:

1. Create a scatter plot of `brainwt` versus `bodywt`.
2. Include an annotation for the Pearson correlation coefficient.
3. Include an annotation for the best fit line.

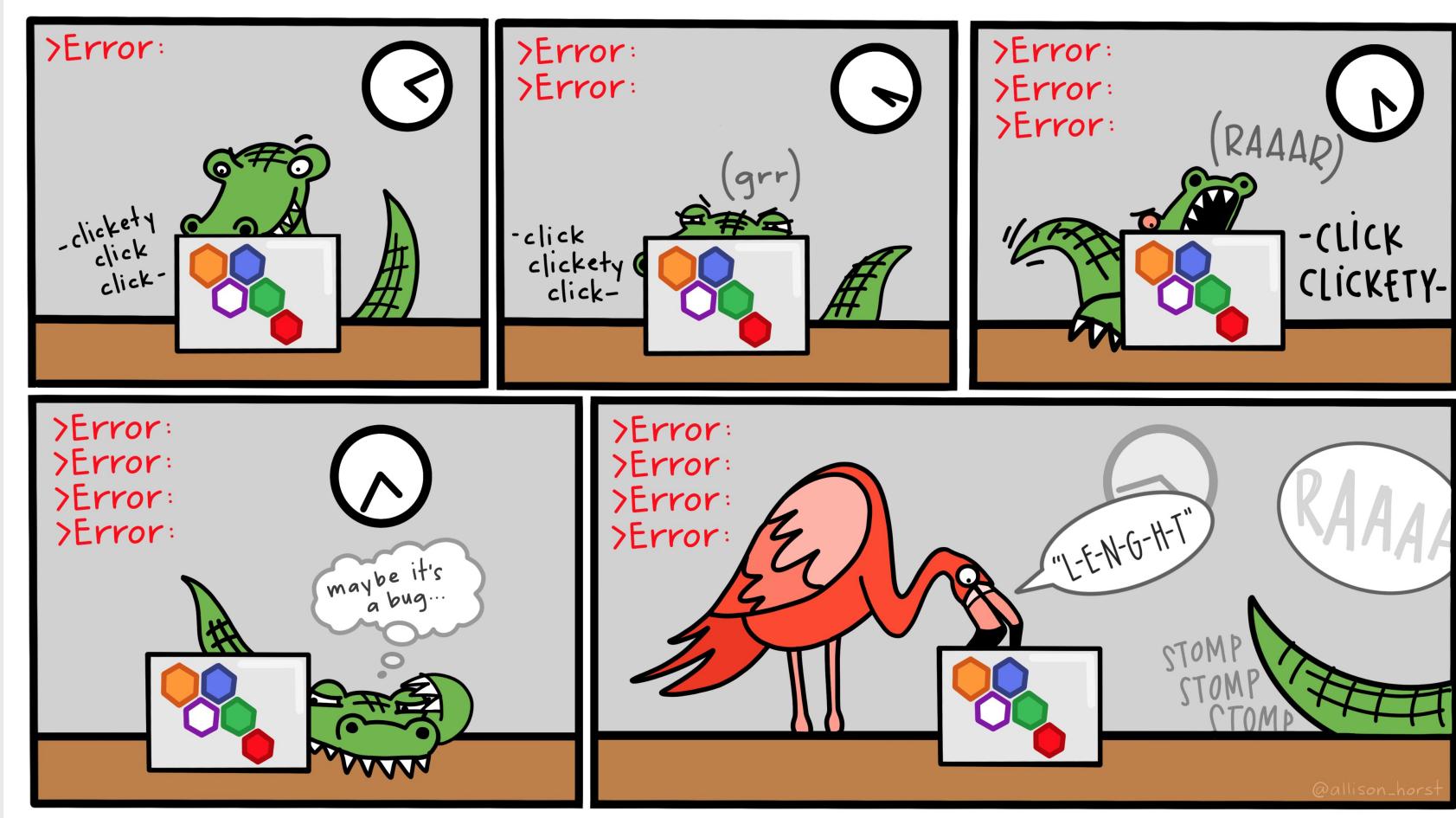
Bonus: Compare your results to a log-linear relationship by converting the x and y variables to the log of x and y, like this:

```
model <- lm(log(x) ~ log(y), data = msleep)
```

You can also convert your plot to log axes by adding these layers:

# Projects

# Take your time and take breaks





Artwork by @allison\_horst

Start thinking about research questions

# Writing a research question

Follow [these guidelines](#) - your question should be:

- **Clear:** your audience can easily understand its purpose without additional explanation.
- **Focused:** it is narrow enough that it can be addressed thoroughly with the data available and within the limits of the final project report.
- **Concise:** it is expressed in the fewest possible words.
- **Complex:** it is not answerable with a simple "yes" or "no," but rather requires synthesis and analysis of data.
- **Arguable:** its potential answers are open to debate rather than accepted facts (do others care about it?)

# Writing a research question

## **Bad question: Why are social networking sites harmful?**

- Unclear: it does not specify *which* social networking sites or state what harm is being caused; assumes that "harm" exists.

## **Improved question: How are online users experiencing or addressing privacy issues on such social networking sites as Facebook and Twitter?**

- Specifies the sites (Facebook and Twitter), type of harm (privacy issues), and who is harmed (online users).

**Other good examples:** See the [Example Projects Page](#) page