Chapter 3: Relationships Between Variables

DSCC 462

Computational Introduction to Statistics

Anson Kahng

Fall 2022

Plan for Today

- Visualize relationships between variables
- Determine whether variables are correlated

Summaries for Two Variables

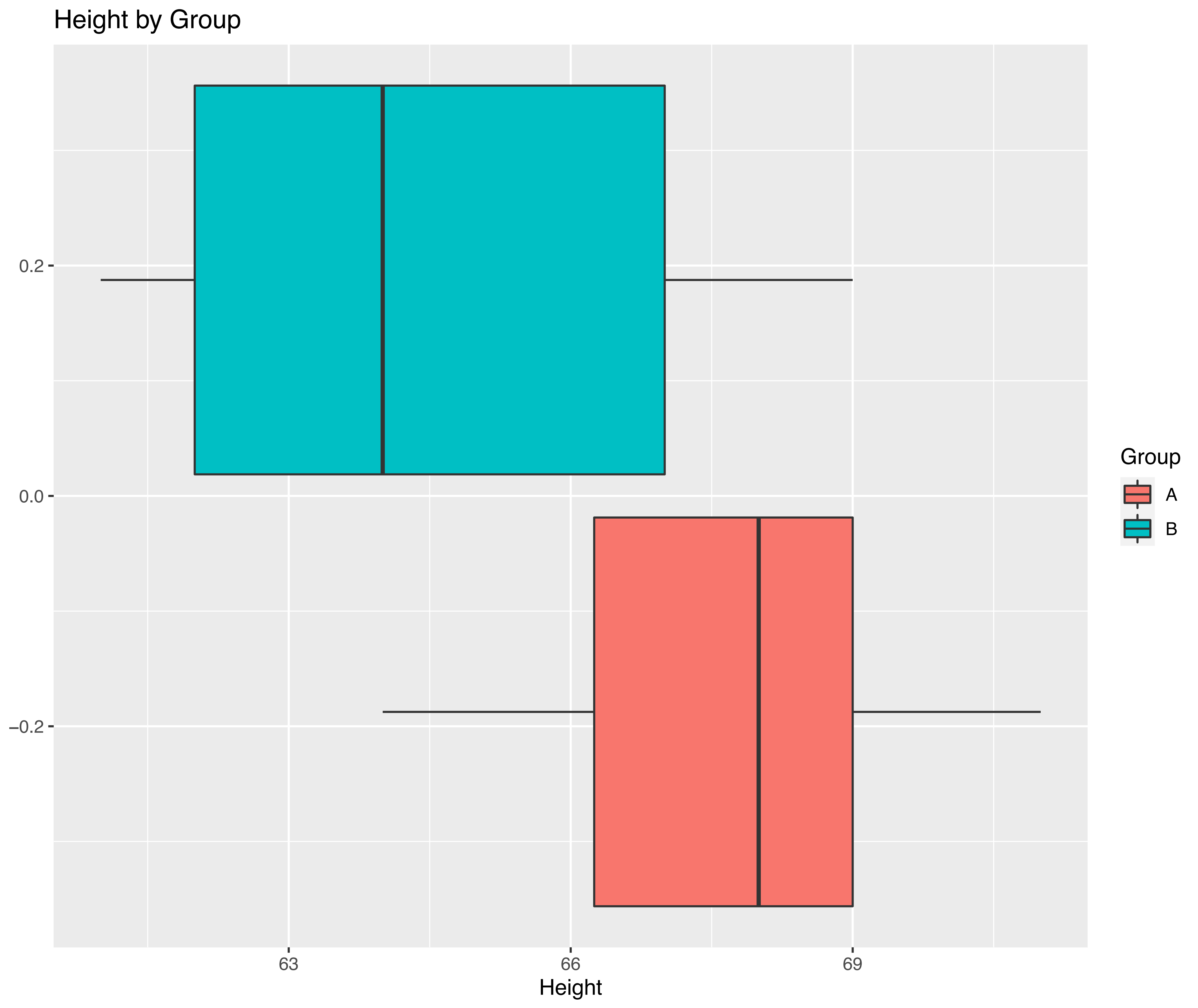
- Recall that we have discussed summaries of center and spread for one variable
- Suppose we wanted to summarize height by sex, or summarize the relationship between hip length and weight
- Much of what we did for one variable can be extended to two variables

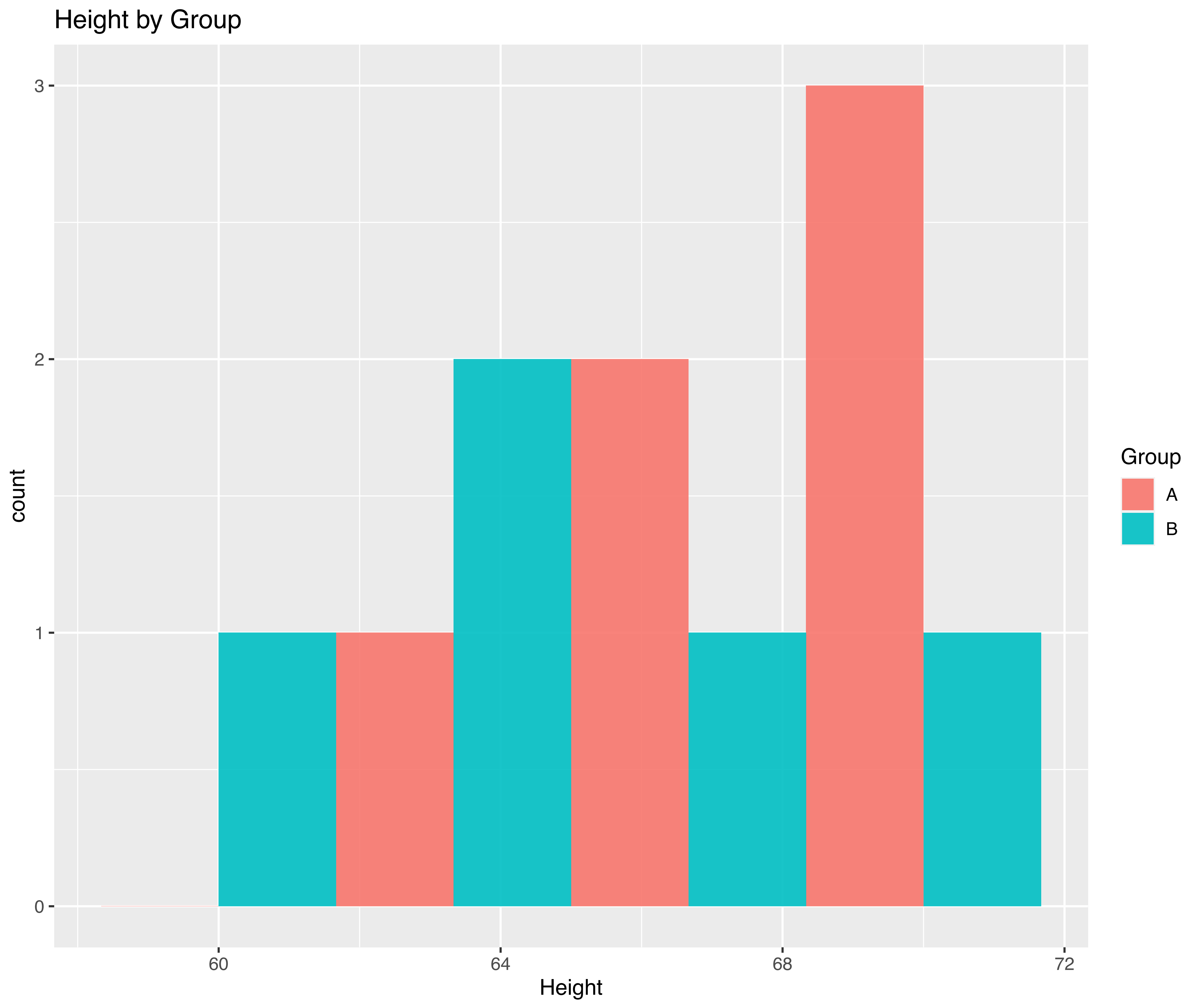
Case CQ: Categorical and Quantitative

- If we have a quantitative variable that we want to summarize over multiple categories/groups, we can simply calculate quantitative variable summary statistics (e.g., mean, median, SD, IQR, etc.) for each category/group
- Heights in Group A (in): 64, 66, 67, 69, 69, 71
- Heights in Group B (in): 61, 62, 64, 67, 69
- Mean for Group A: $\bar{x}_A = 67.7$
- Mean for Group B: $\bar{x}_B = 64.6$

Case CQ: Categorical and Quantitative

- We can make histograms for each group, or side-by-side boxplots






```
dat1<-data.frame(Height=c(64,66,67,69,69,71,61,62,64,67,69), Group=c(rep("A",6),rep("B",5)))
ggplot(dat1, aes(x=Height,fill=Group)) + geom_boxplot() + labs(title="Height by Group")
ggplot(dat1, aes(x=Height,fill=Group)) + geom_histogram(bins=4,alpha=0.9,position='dodge') + labs(title="Height by Group")
```

Case CC: Categorical and Categorical

- If we have two categorical variables, we want to make a *two-way table* to describe the results
  - Cross tabulation of two categorical variables
- Extend the frequency table we made for one categorical variable and extend it to two variables
- Consider the variables group (A/B) and smoking status (smoker/non-smoker)

	Smoker	Non-Smoker
Group A	15	22
Group B	26	18

Case CC: Categorical and Categorical

- From this table, we can determine the total number of people in Group A, people in Group B, smokers, and non-smokers
- These sub-totals are known as *marginal values* for each variable
- The marginal distributions for each variable can be summarized exactly as we did for the one variable case; we can make a bar plot for each and calculate marginal frequencies

	Smoker	Non-Smoker	**Total**
Group A	15	22	**37**
Group B	26	18	**44**
Total	**41**	**40**	**81**

Case CC: Conditional Distributions

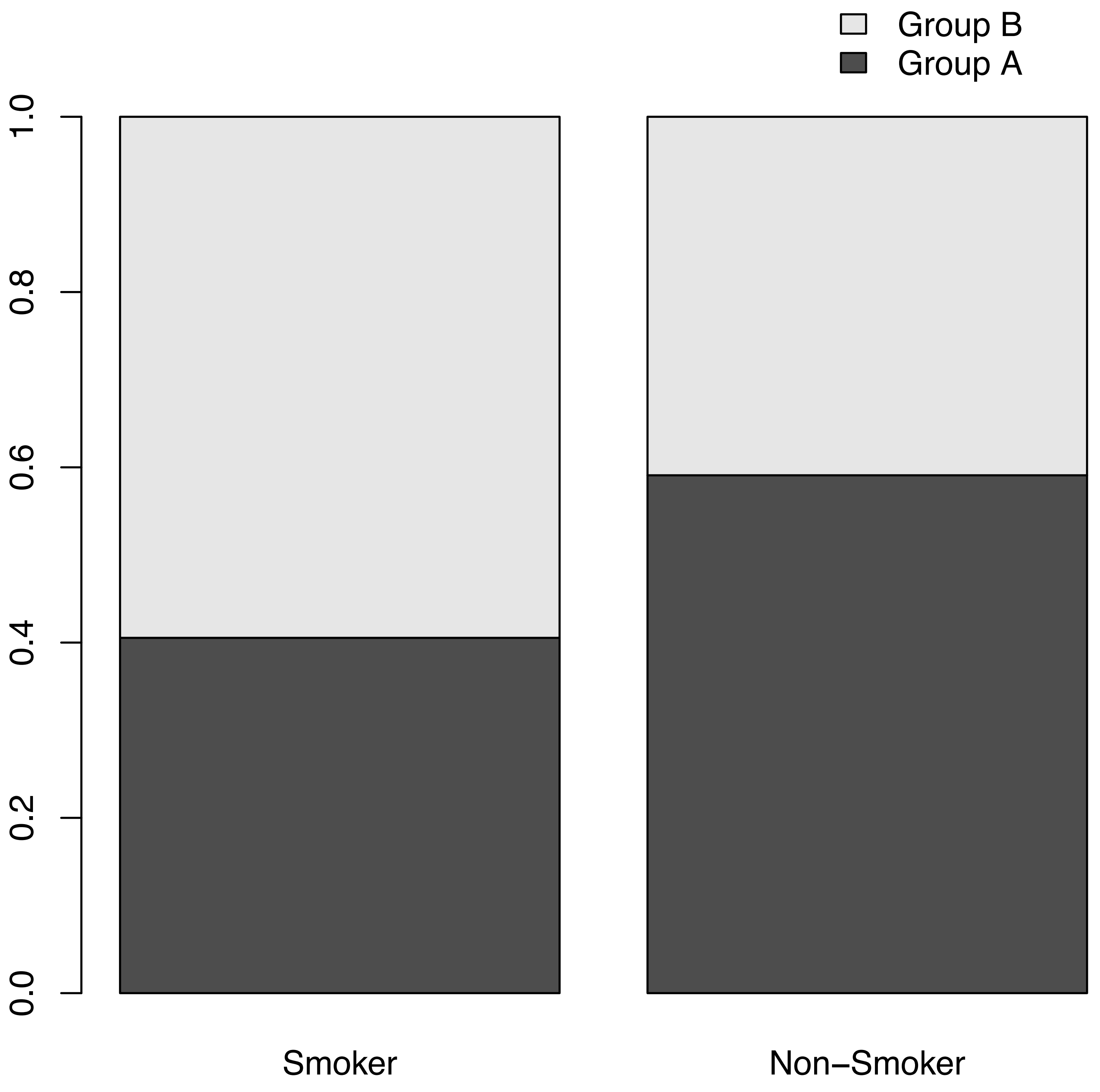
	Smoker	Non-Smoker	**Total**
Group A	15	22	**37**
Group B	26	18	**44**
Total	**41**	**40**	**81**

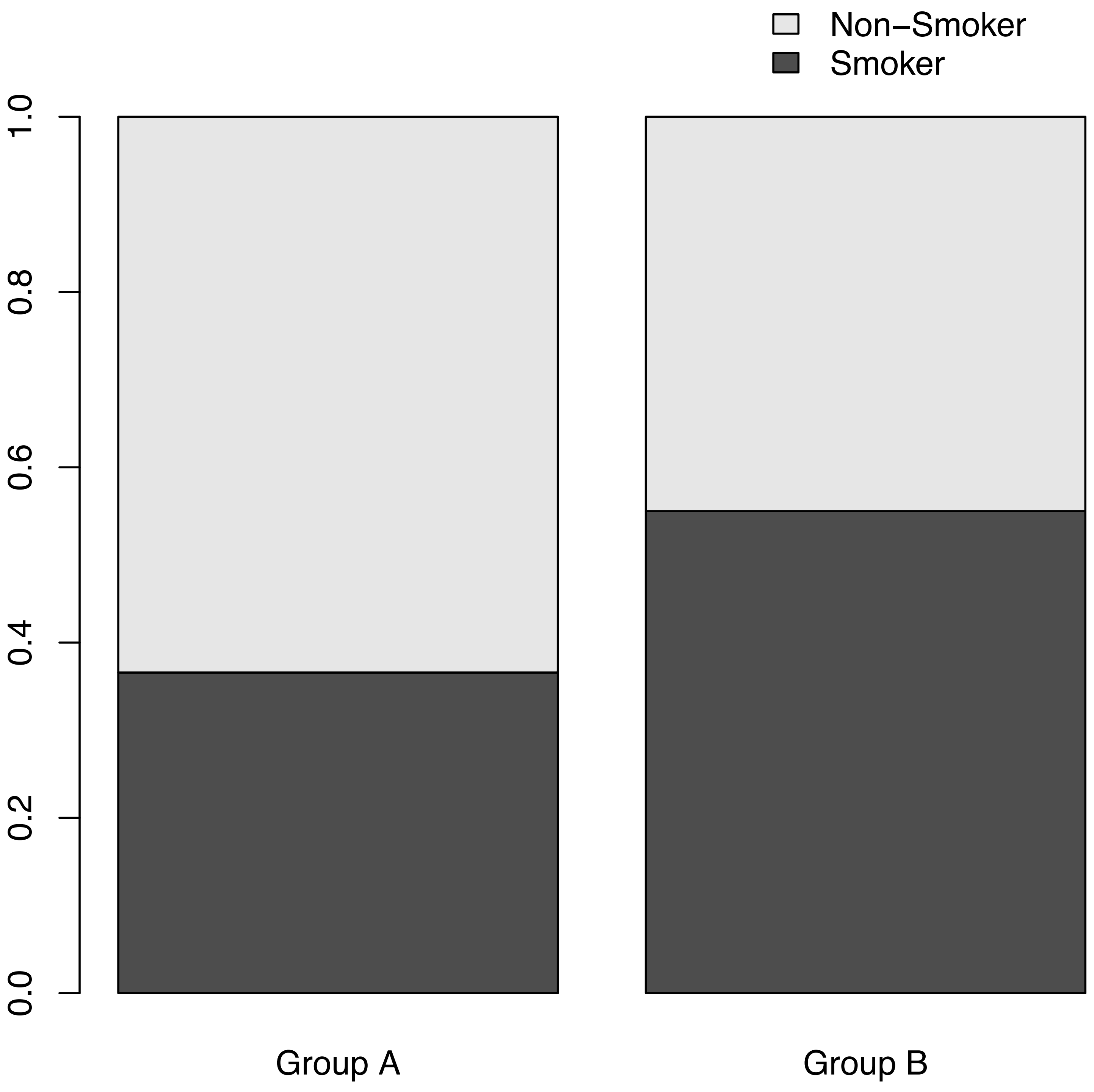
- What is the probability of smoking given that you are in Group B?

- What is the probability of being in Group A given that you smoke?

Case CC: Conditional Distributions

	Smoker	Non-Smoker	**Total**
Group A	15	22	**37**
Group B	26	18	**44**
Total	**41**	**40**	**81**

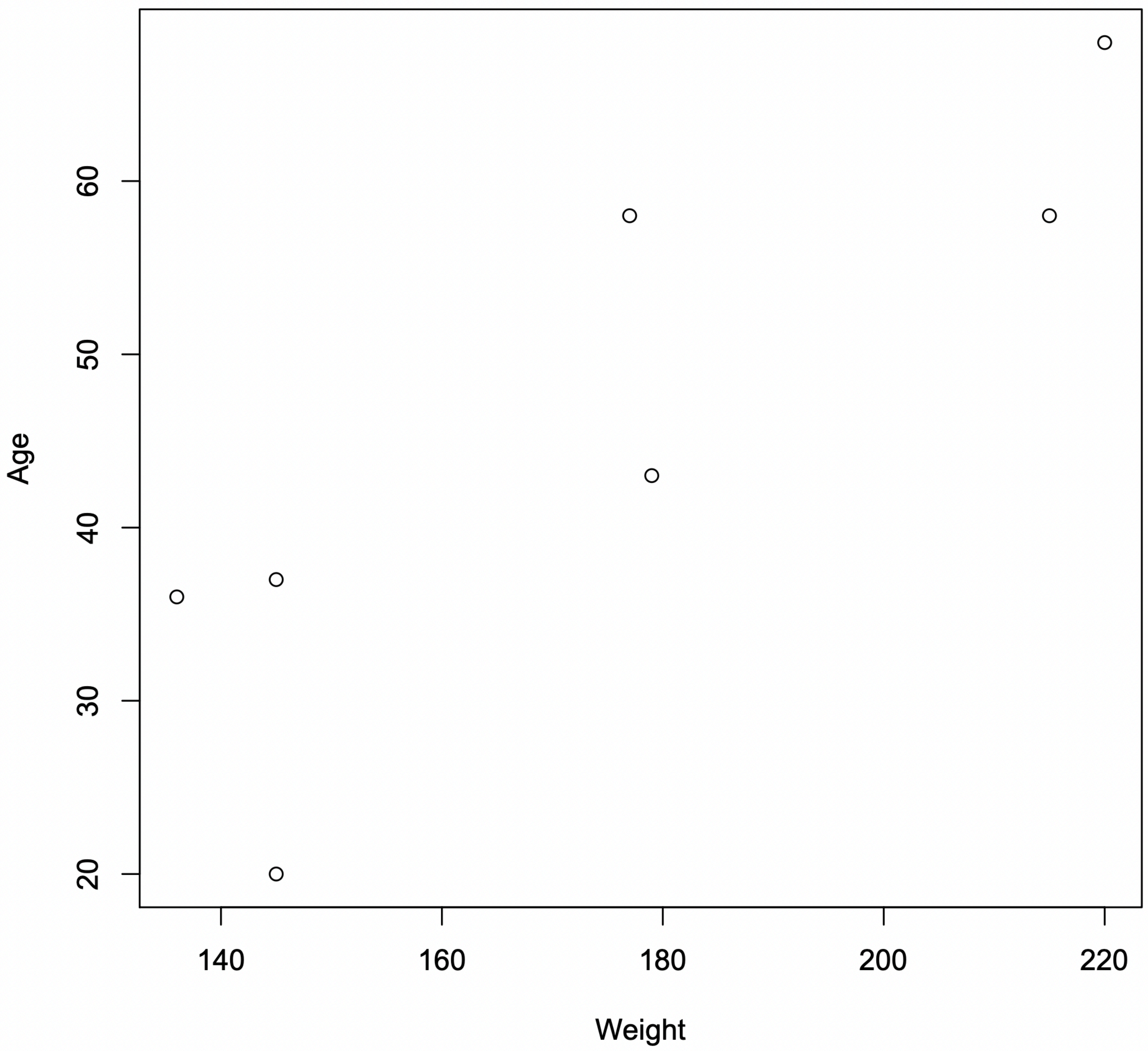




Case QQ: Quantitative and Quantitative

- Suppose we are interested in examining the relationship between diabetic patients' weights and ages
- We can graphically display this relationship with a *two-way scatterplot*
- When we make a scatterplot, we have our two variables as our two axes, and points are plotted based on their corresponding values for each variable
- R code: `plot(x=weight,y=age,xlab="Weight",ylab="Age")`

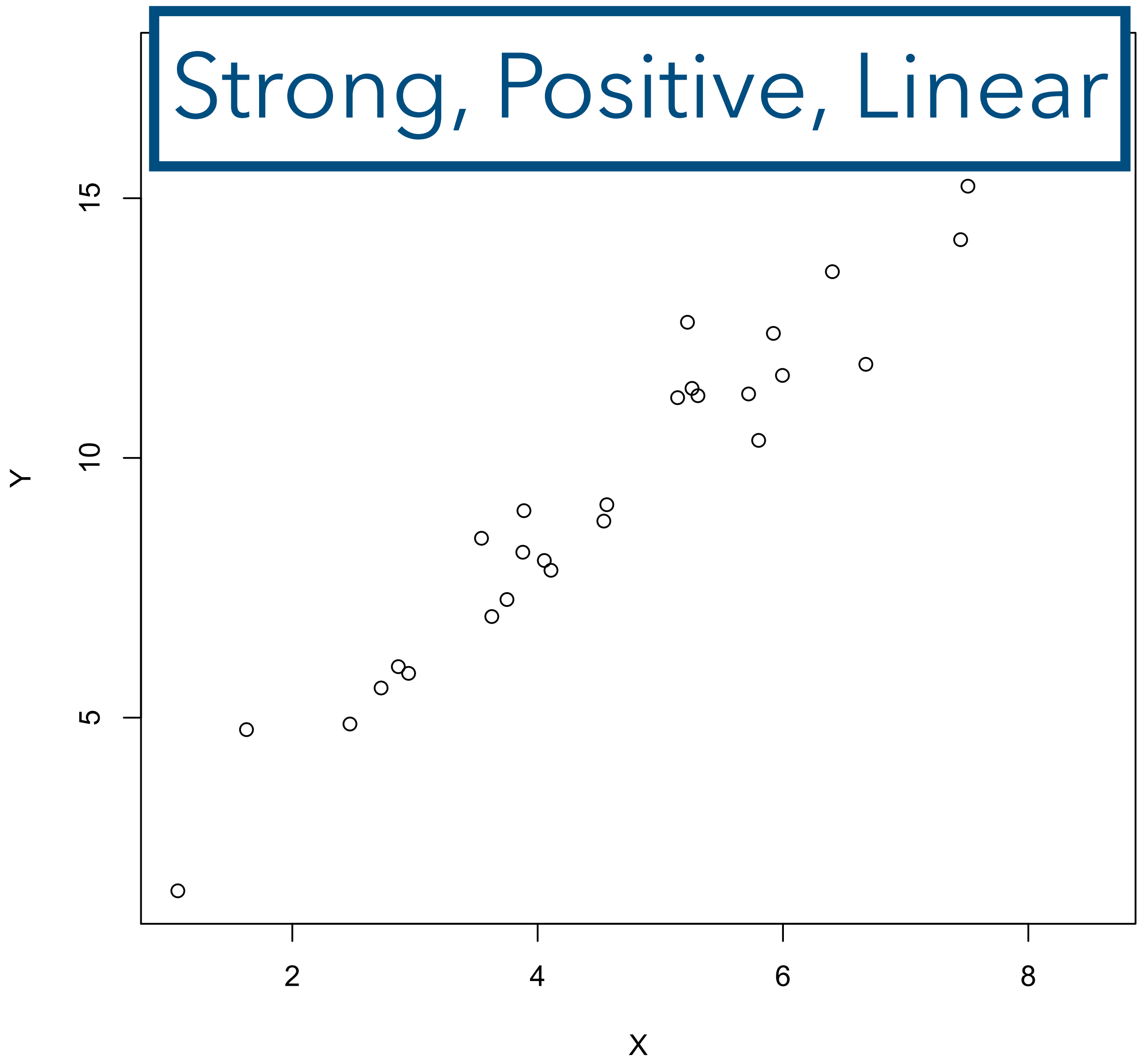
Scatterplot

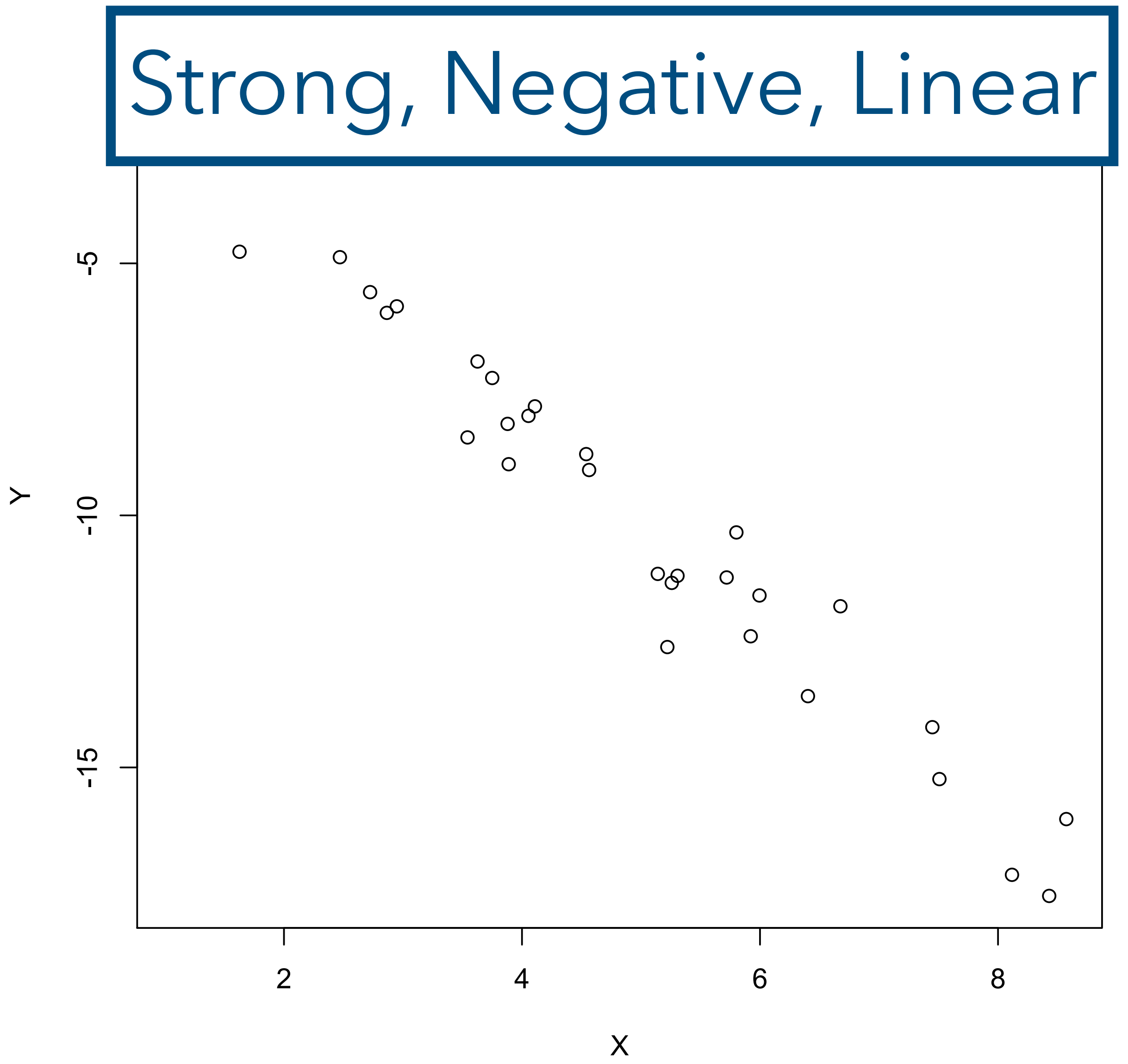


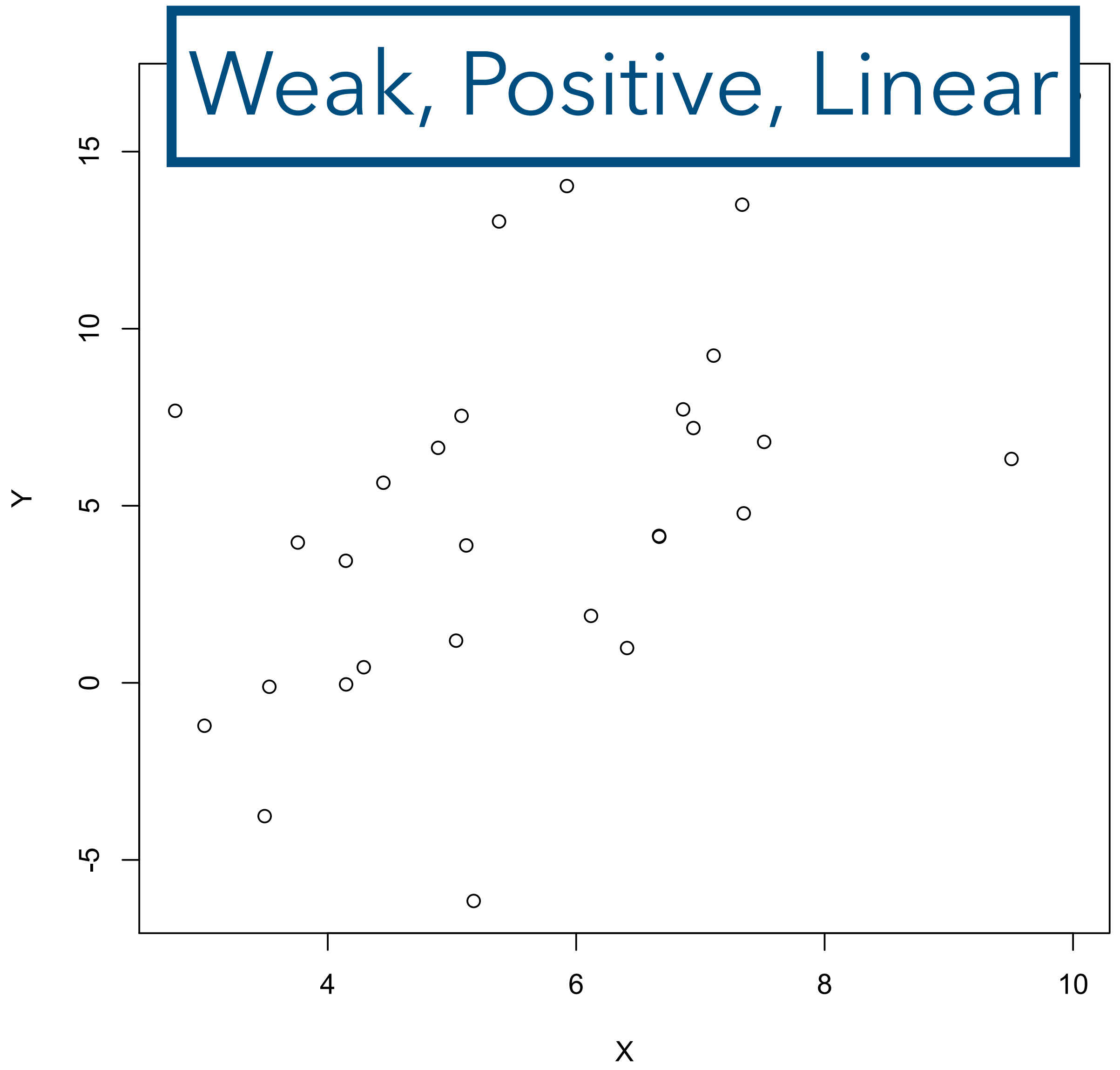
Scatterplot

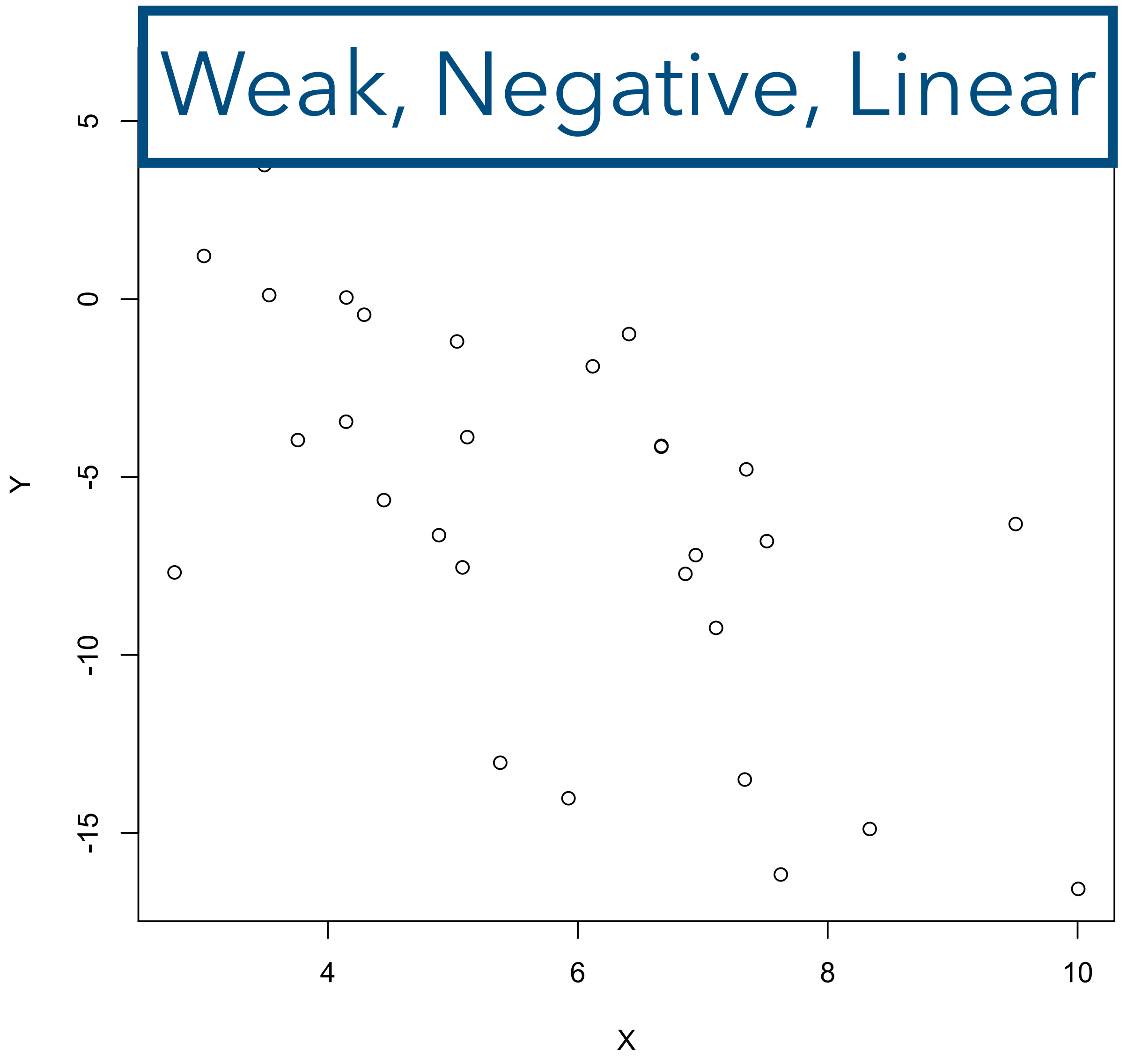
- Want to discuss the direction, form, and strength
  - ***Direction***: positive, negative, or neither
  - ***Form***: linear, non-linear, or no relationship
  - ***Strength***: strong, weak, or none

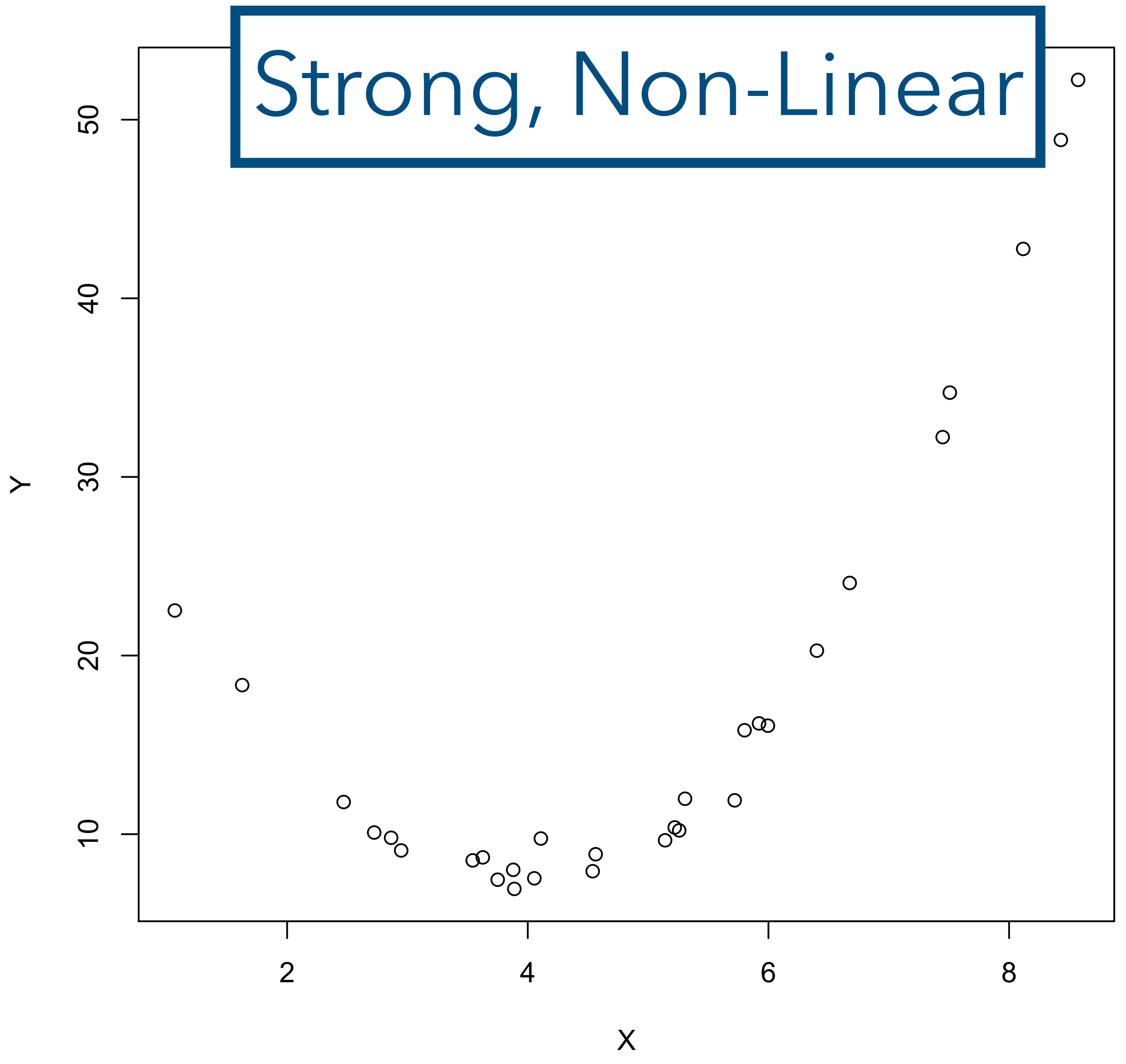
Examples: Strength, Direction, and Form

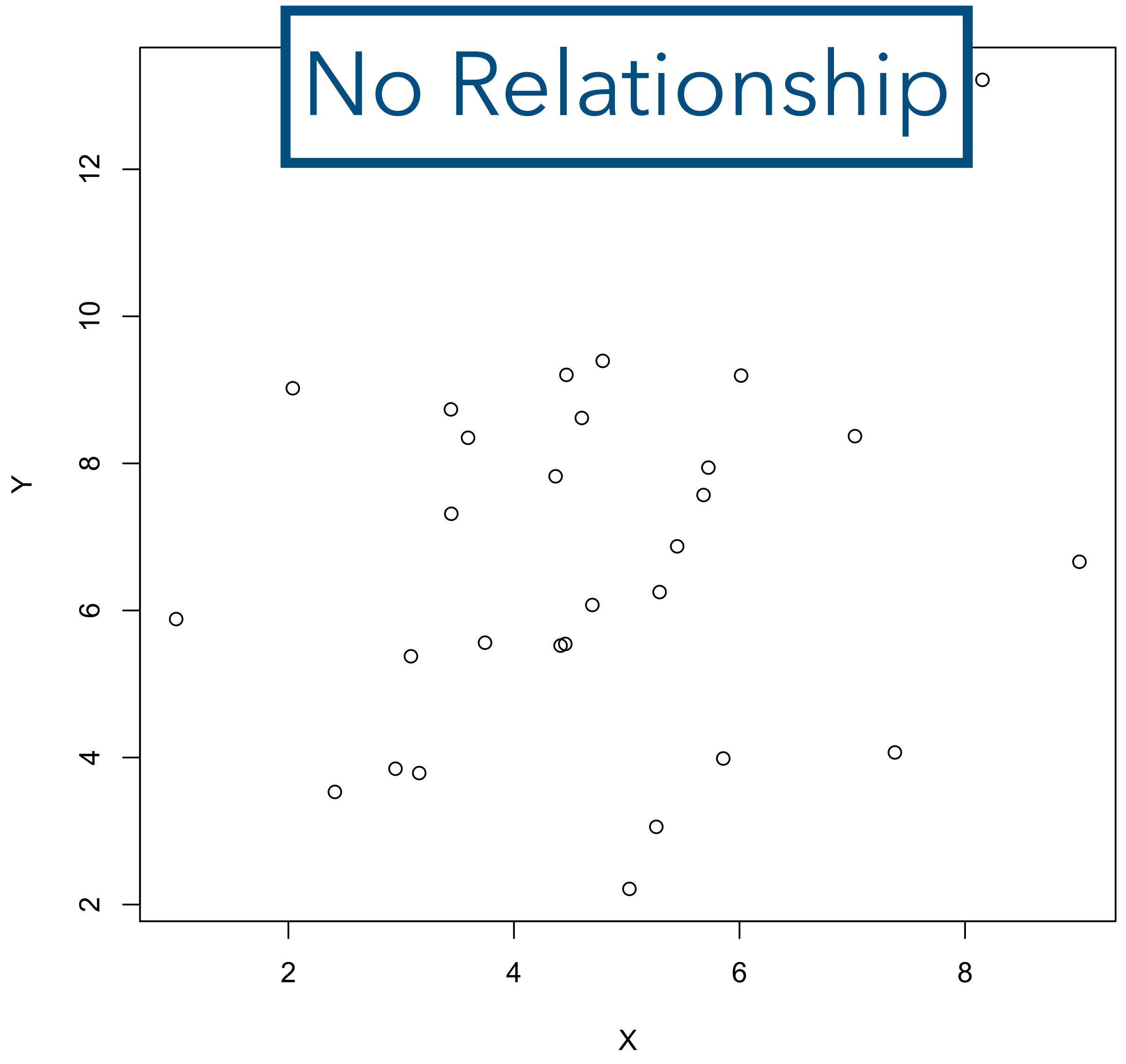








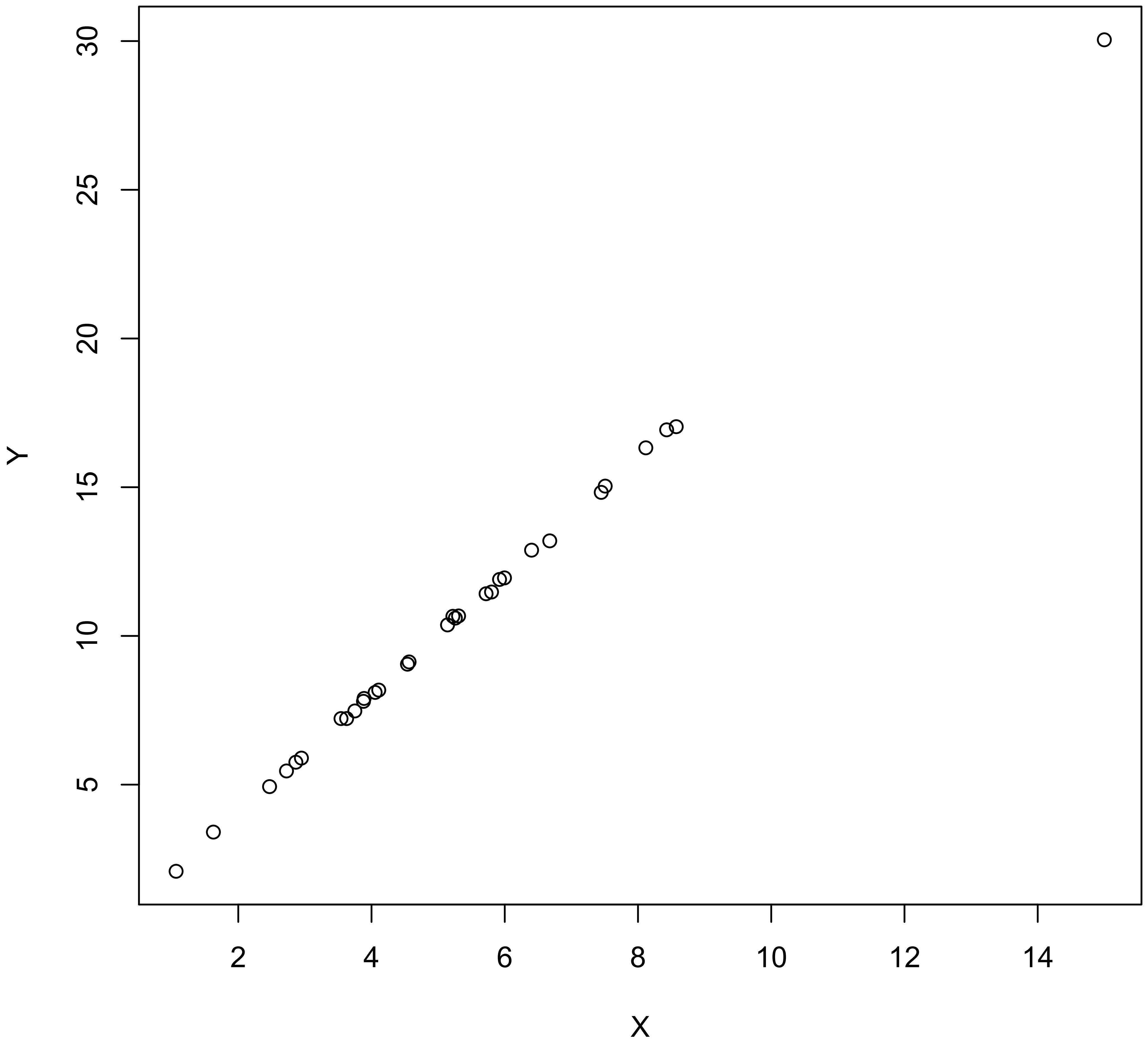


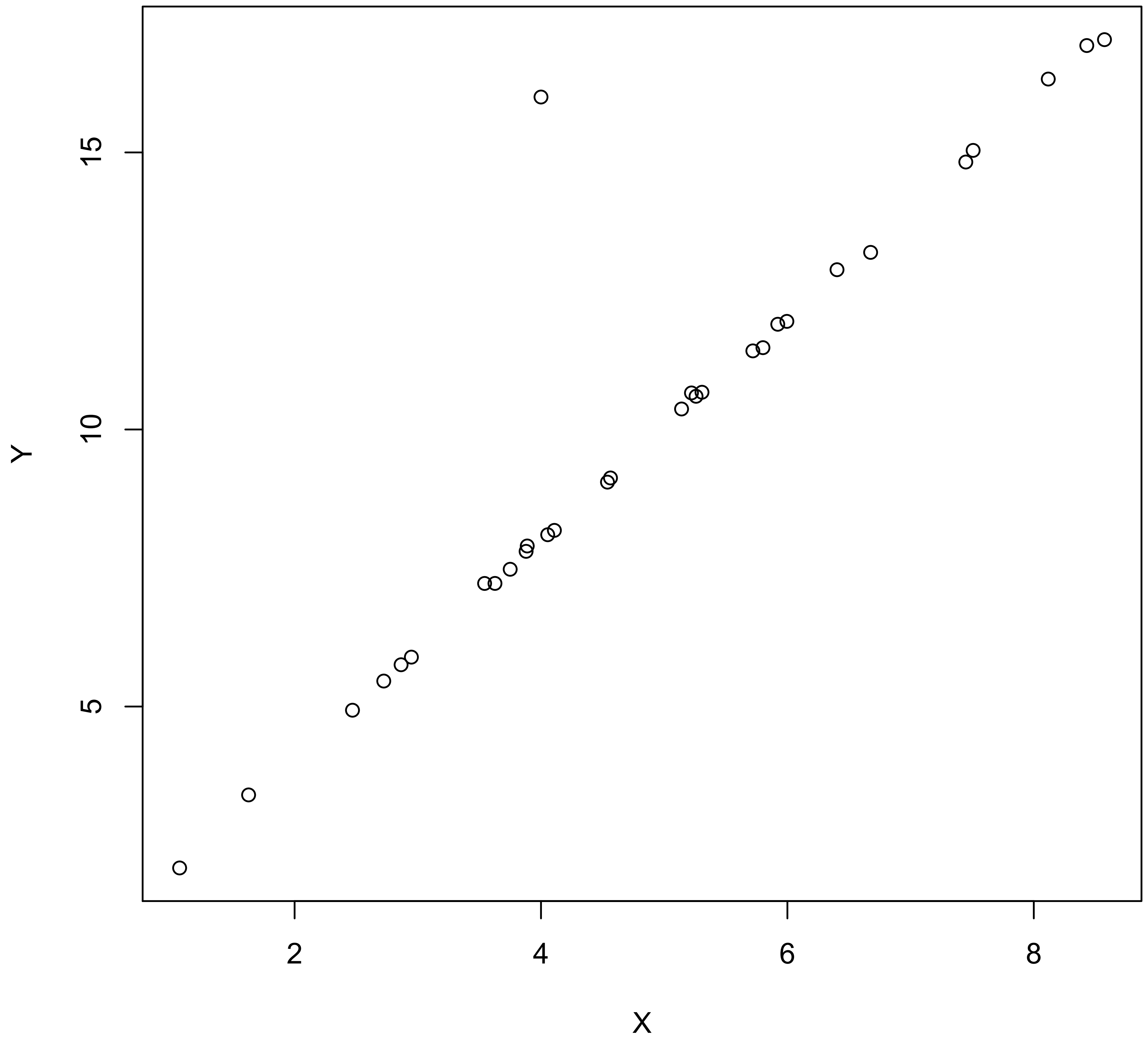


Scatterplot: Outliers

- From a scatterplot, we are able to visually identify unusual points and features of the data
- Examine the scatterplot to see if there are any points that do not seem to follow the trend of the data
    - These points are outliers

Outliers: Examples





Correlation

- From a scatterplot, we can see the relationship between two variables
- *Correlation* tells us the degree to which two random variables are (linearly) associated or related
- Setup: two quantitative variables, X and Y; X is on the horizontal axis of the scatterplot and Y is plotted on the vertical axis

Pearson's Correlation Coefficient (r)

- *Pearson's coefficient of correlation*, or *sample correlation coefficient*, r:

$$r = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}}$$

- A quantity related to the correlation is the *sample covariance:*

$$s_{xy} = \frac{1}{(n-1)}\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

Correlation vs. Covariance

- Both correlation and covariance measure the relationship between variables
- Positive values = positive (linear) relationship
- Negative values = negative (linear) relationship
- Covariance indicates direction
- Correlation indicates direction and strength
- Correlation values are standardized between -1 and 1
- Covariance values are not standardized

Pearson's Correlation Coefficient: Alternative Form

- We can define *Sums of Squares*:

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s_x^2$$

$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = (n-1)s_{xy}$$

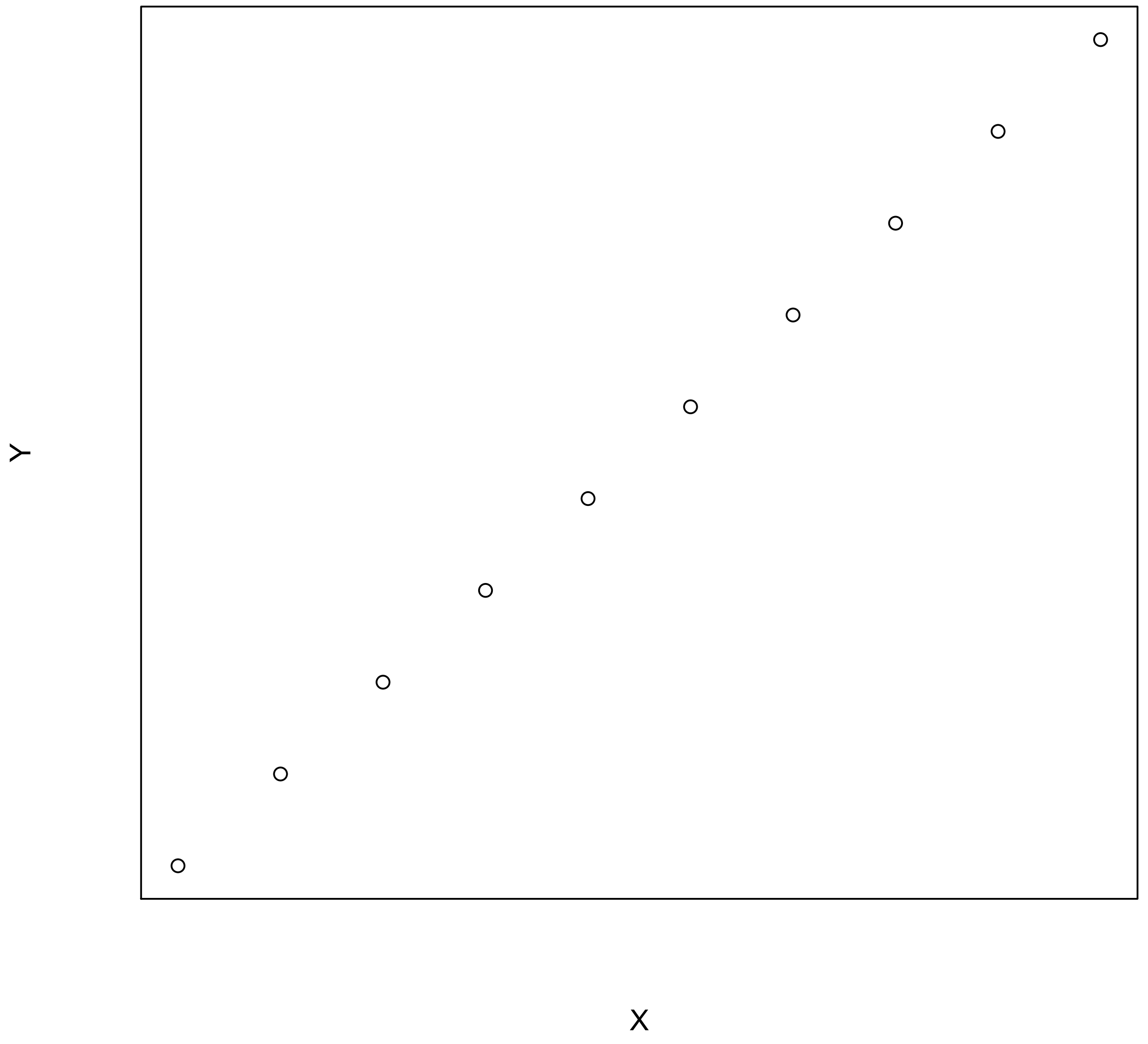
- Rewriting the sample correlation:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

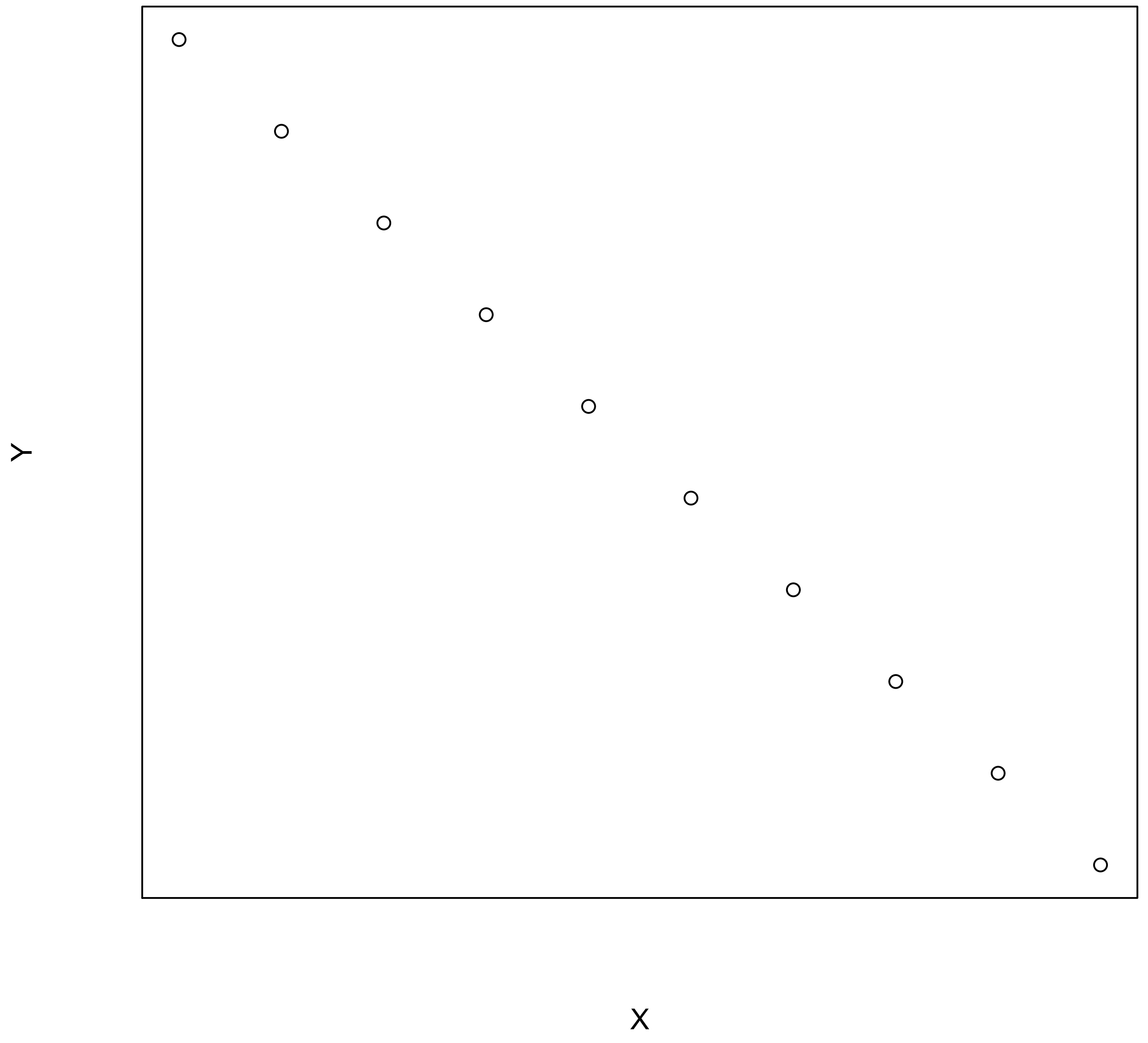
Pearson's Correlation Coefficient: Interpretation

- The correlation coefficient does not have units and is bounded: $-1 \leq r \leq 1$
- If $r = 1$ (resp. $r = -1$), then X and Y have a perfect linear relationship in the positive (resp. negative) direction, i.e., for each increase in X, we have a perfect increase (resp. decrease) in Y
  - In the cases of $r = \pm 1$, pairs of outcomes (x, y) lie on a straight line
- Any $r > 0$ indicates a positive relationship between X and Y ($x \uparrow \rightarrow y \uparrow$)
- Any $r < 0$ indicates a negative relationship between X and Y ($x \uparrow \rightarrow y \downarrow$)
- When $r = 0$, X and Y have no linear relationship at all (could be non-linear)

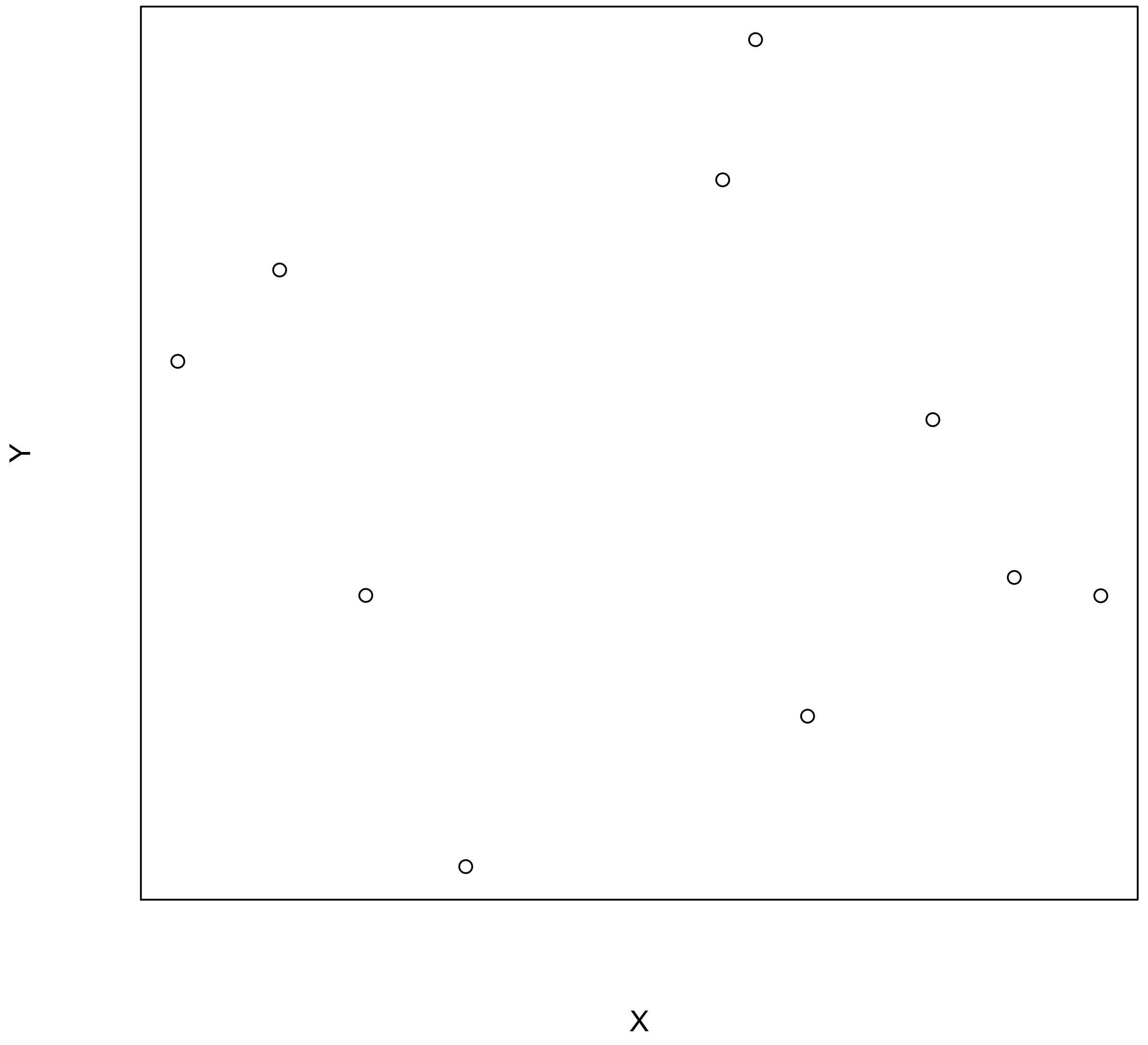
Pearson's Correlation Coefficient: Examples



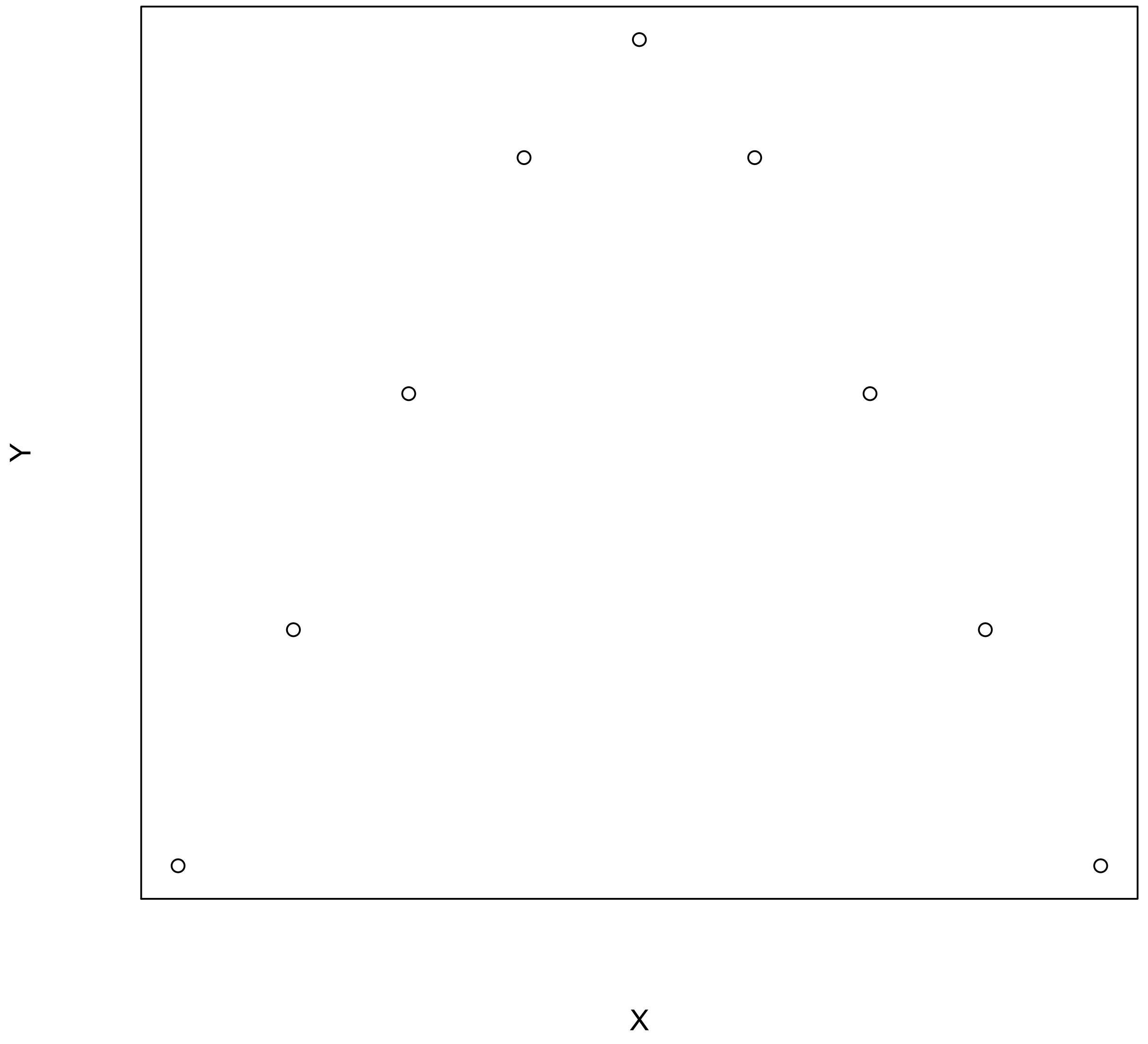
$r = 1$



$r = -1$



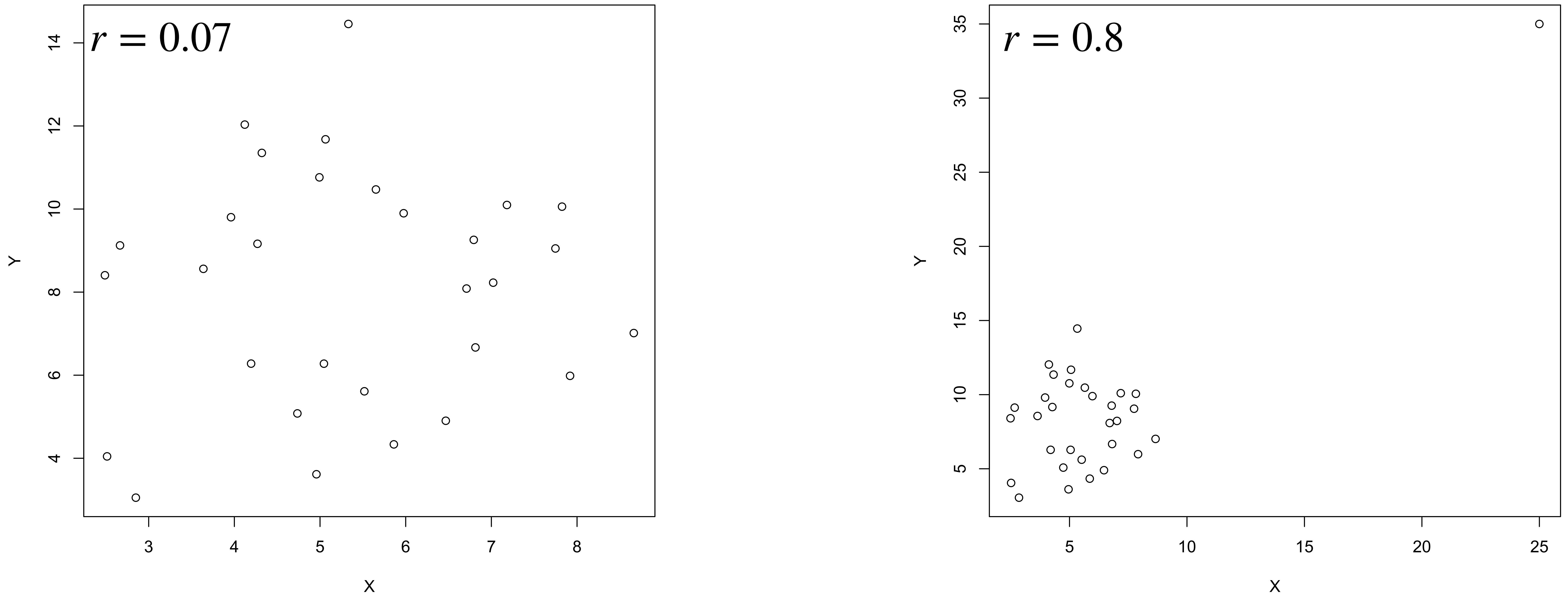
$r = 0$



$r = 0$

Correlation and Outliers

- Correlation can be sensitive to outliers
- A highly influential outlier can cause correlation to look strong when in fact not much of a relationship actually exists



Correlation: Example

- Find the correlation between weight (X) and age (Y) for the following data

Patient	Weight (lbs)	Age
1	220	68
2	215	58
3	179	43
4	145	37
5	145	20
6	177	58
7	136	36

Average weight: $\bar{x} = \dfrac{220 + 215 + 179 + 145 + 145 + 177 + 136}{7} = 173.86$

Average age: $\bar{y} = \dfrac{68 + 58 + 43 + 37 + 20 + 58 + 36}{7} = 45.71$

$$\sum_{i=1}^{7} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{7} (x_i - 173.86)(y_i - 45.71) = 2919.714$$

$$\sum_{i=1}^{7} (x_i - \bar{x})^2 = \sum_{i=1}^{7} (x_i - 173.86)^2 = 6956.857$$

$$\sum_{i=1}^{7} (y_i - \bar{y})^2 = \sum_{i=1}^{7} (y_i - 45.71)^2 = 1637.429$$

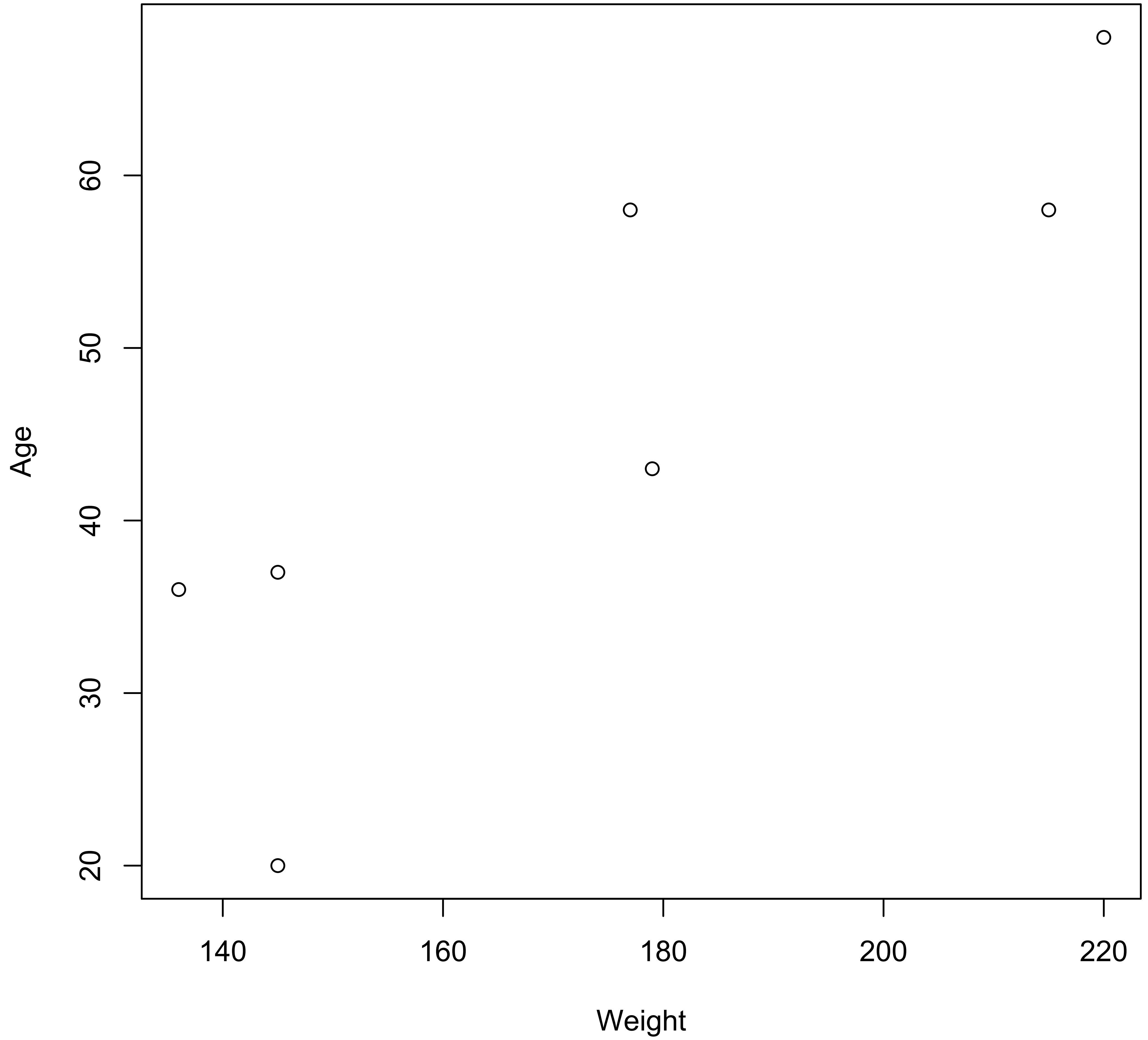
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]}} = \frac{2919.714}{\sqrt{6956.857 \times 1637.429}} = 0.865$$

Strong, positive, linear relationship between weight and age

Correlation: Example

- Find the correlation between weight (X) and age (Y) for the following data

Patient	Weight (lbs)	Age
1	220	68
2	215	58
3	179	43
4	145	37
5	145	20
6	177	58
7	136	36

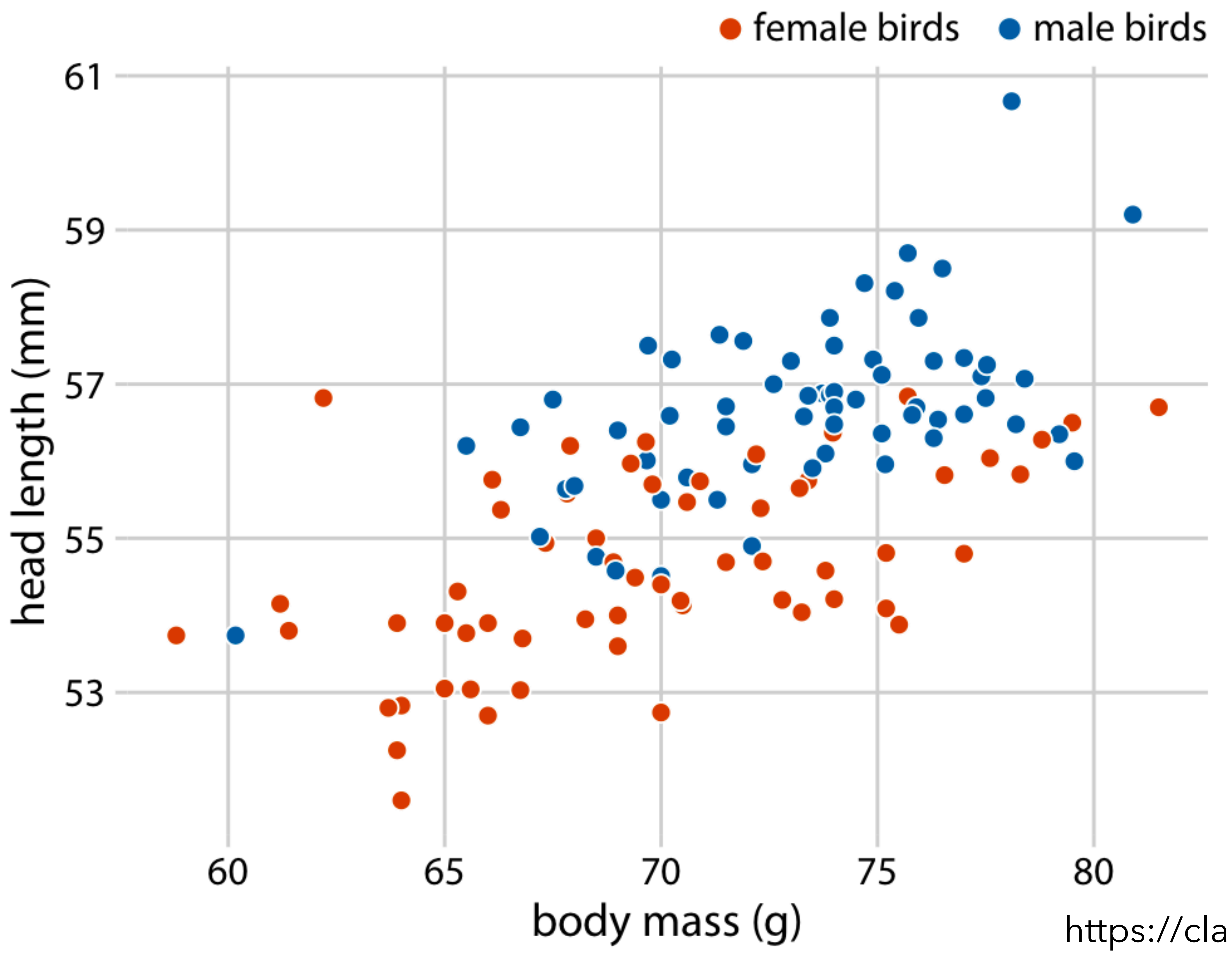


Correlation: Caveat

- Correlation does not imply causation (!!)
- We are only noting that a relationship exists; we are not specifying any cause-and-effect relationship

Three Variables: Add Color

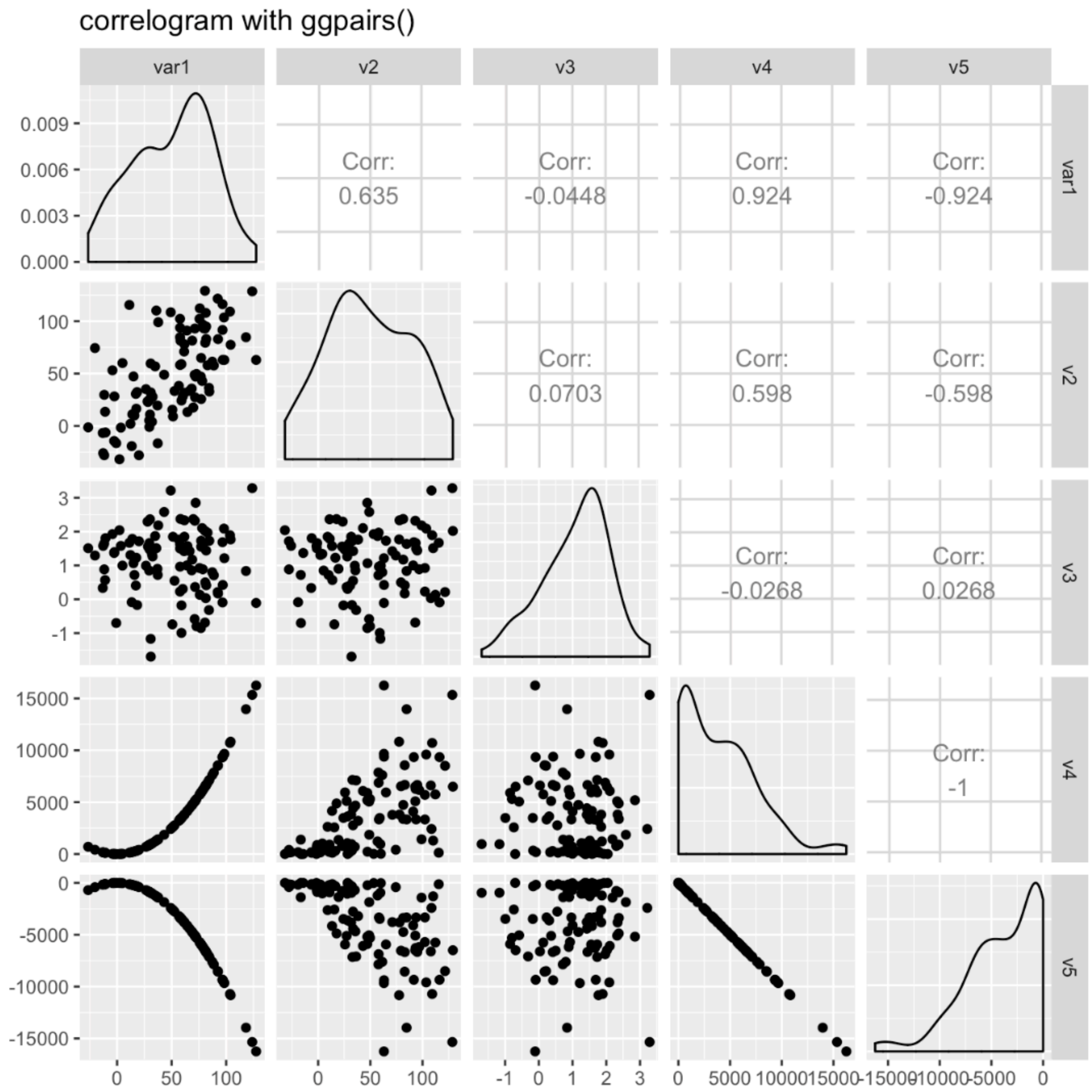
- Imagine we have two quantitative variables and one other variable (either Q or C)
- We can augment our QQ approach (scatterplots) with color corresponding to the third variable

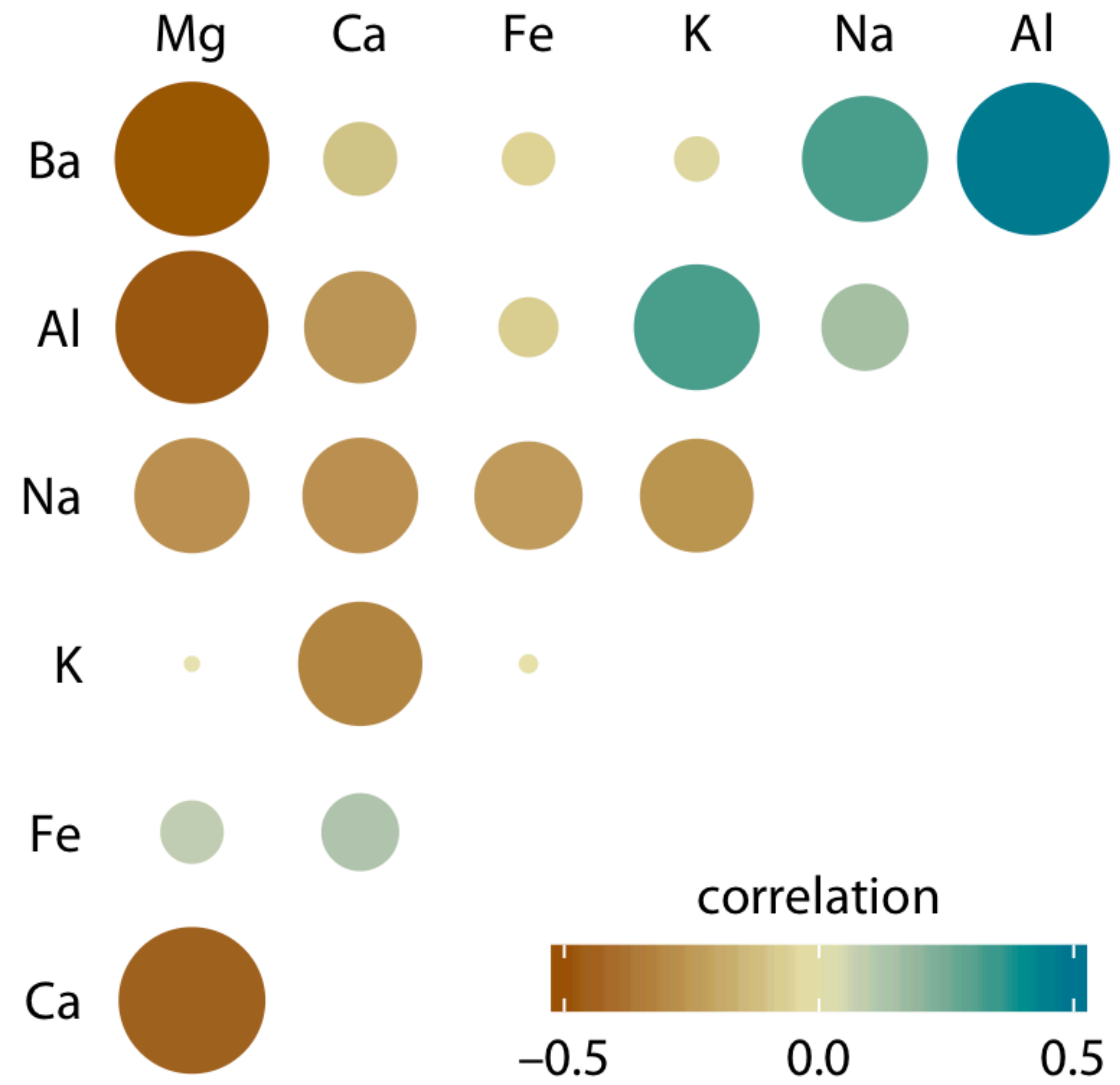


https://clauswilke.com/dataviz/visualizing-associations.html

Correlograms

- Correlograms: Visualize correlation coefficients between pairs of variables
- Very useful for looking at all pairwise relationships in large datasets



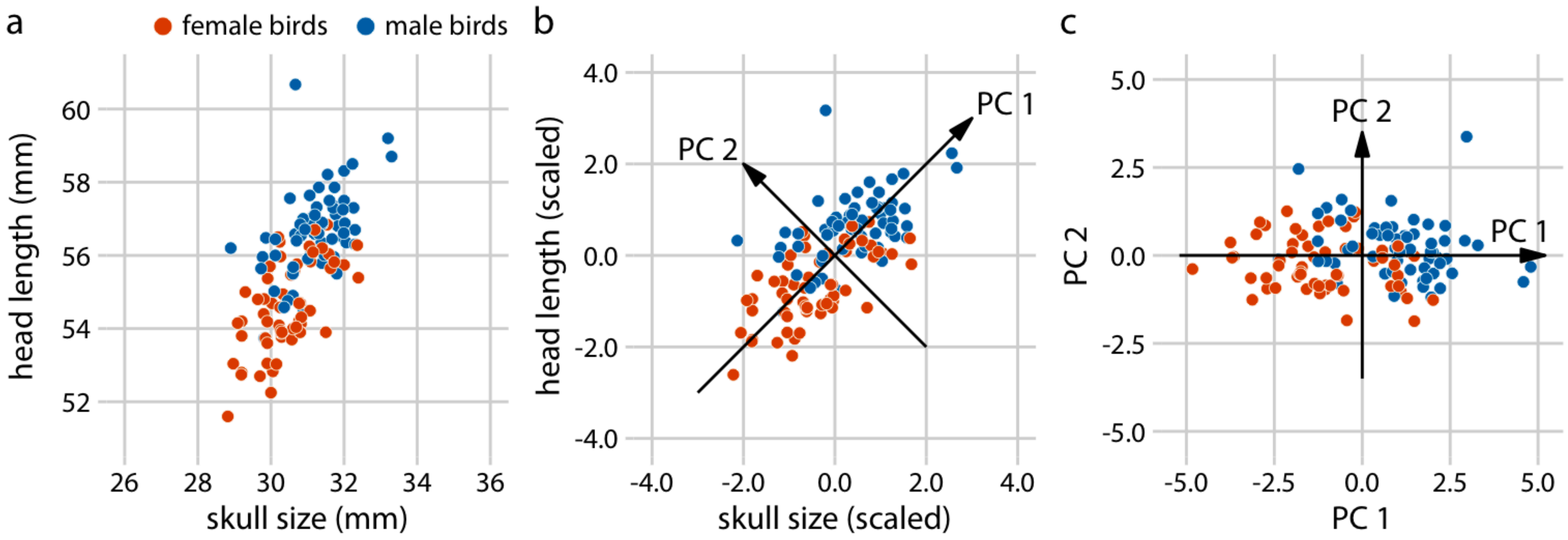


https://clauswilke.com/dataviz/visualizing-associations.html

Dimension Reduction: PCA

- Imagine we have a dataset with many variables (far too many to visualize)
- Intuitively, many of them may be correlated
- Dimension reduction: Reduction in the number of key dimension without losing much information
- **Principal Component Analysis (PCA)**:
    - Principal components: Linear combinations of original variables
    - All uncorrelated (i.e., orthogonal)
    - The n^{th} principal component explains the n^{th} largest amount of variation in the data

Dimension Reduction: PCA



https://clauswilke.com/dataviz/visualizing-associations.html