What is it used for?

- Dotplot
- Histogram
- Density plot
- Bar chart
- Pie chart
- Boxplot

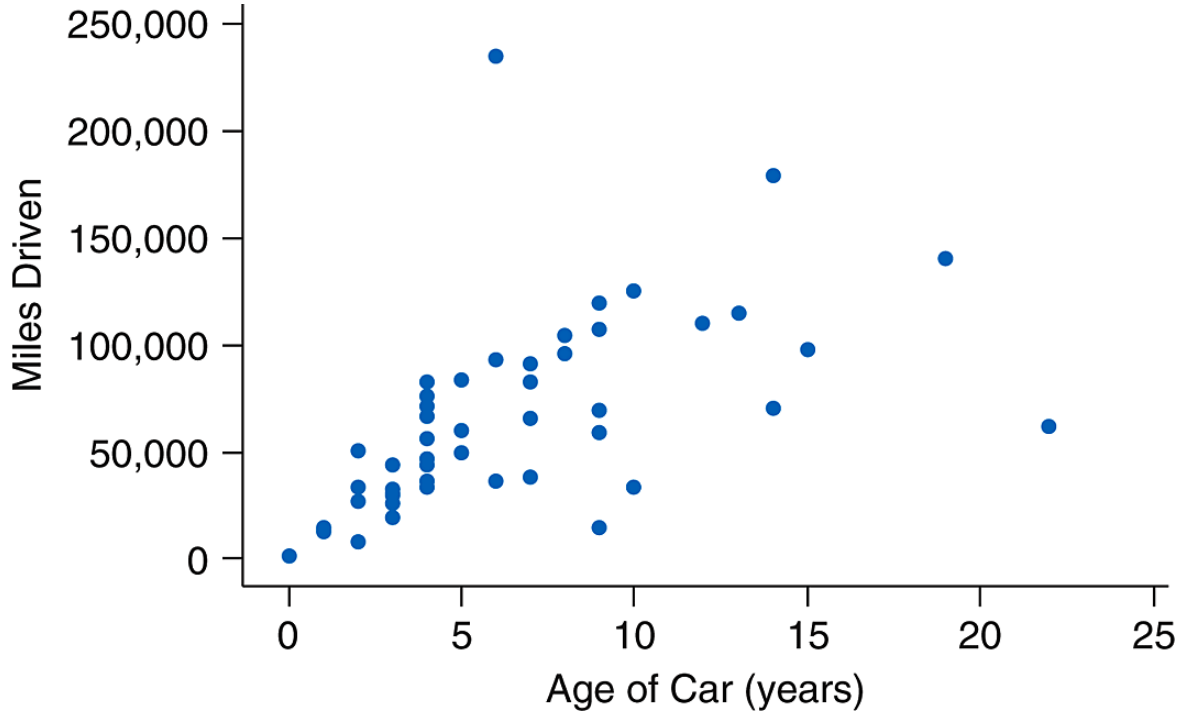
Section 4.1

Visualize correlations with scatterplot

Scatterplot

- The primary tool for examining relationships between two numerical data sets (variables).
- Each point in the scatterplot represents one observation.
- Usually created using technology such as a computer software program or a graphing calculator.

Age of Used Cars and Miles Driven



Each point in the scatterplot represents one car. Each data point has the form:

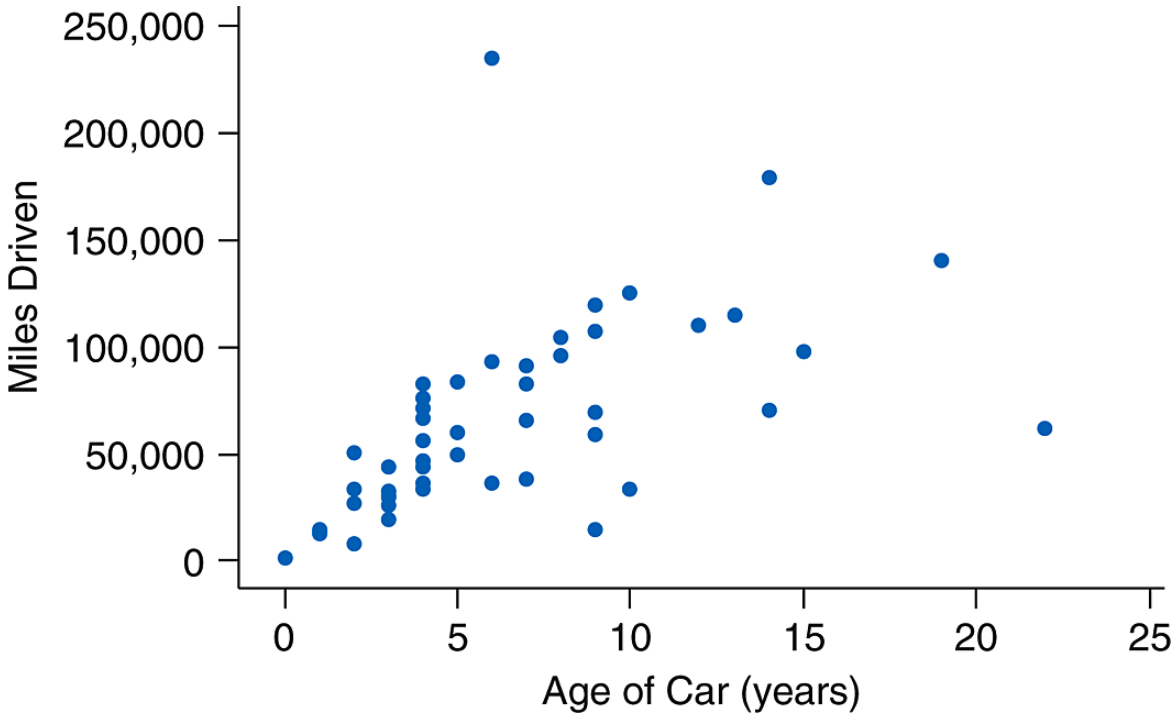
(age of car, miles driven).

Examining Scatterplots

Note three features:

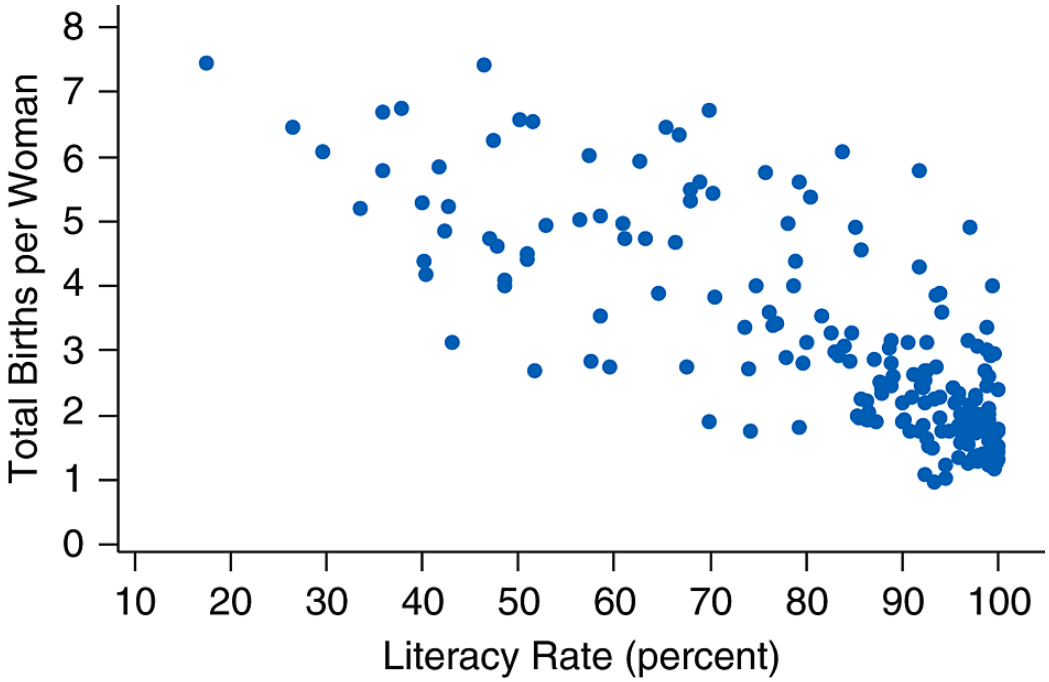
1. Trend: positive, negative, neither
2. Strength: weak, strong
3. Shape: linear, nonlinear

Trend: positive (uphill) correlation



Two variables tend to increase together. As the age of the car increases, the mileage also tends to increase.

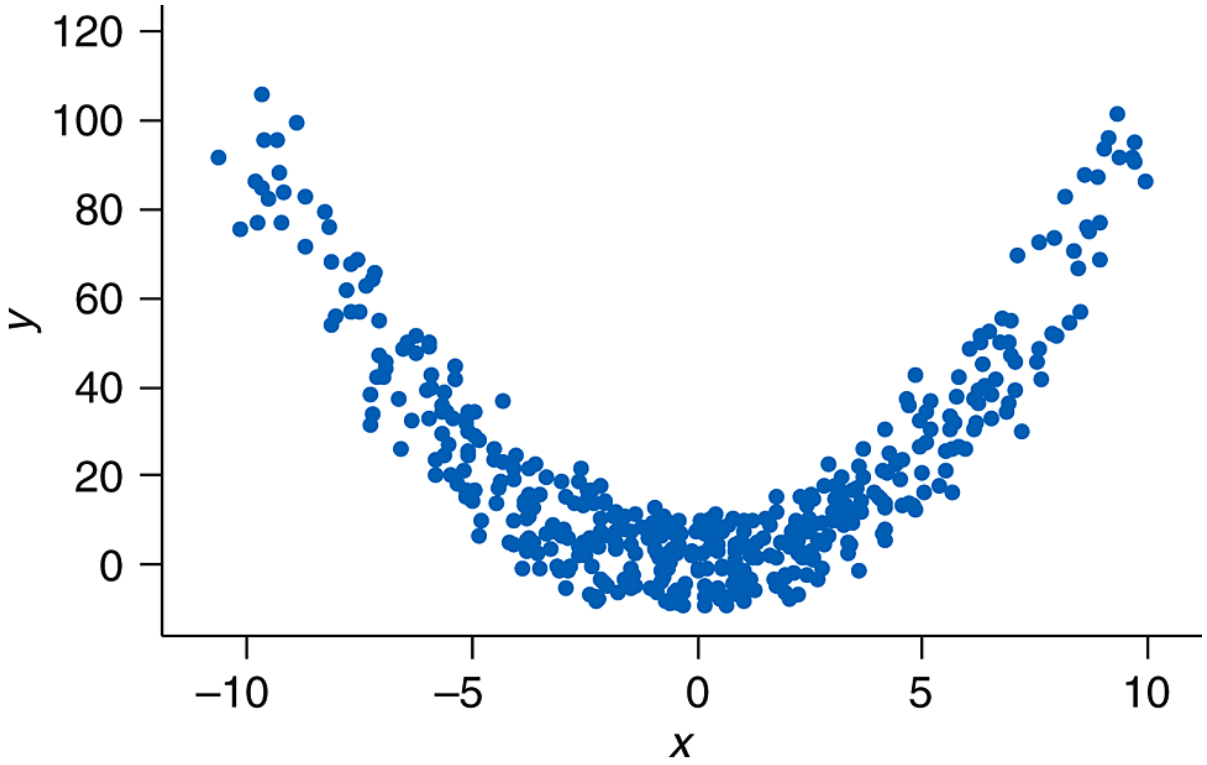
Trend: negative (downhill) correlation



Each dot represents a country

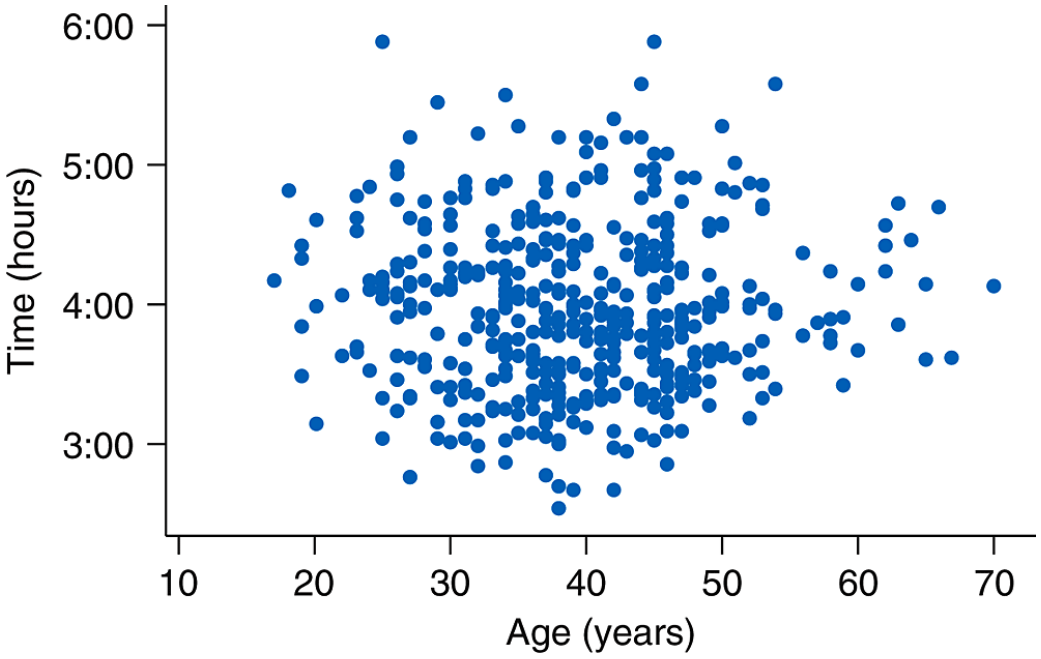
As one variable increases, the other tends to decrease. As literacy rate increases, total births per woman tends to decrease.

Trend: neither positive nor negative



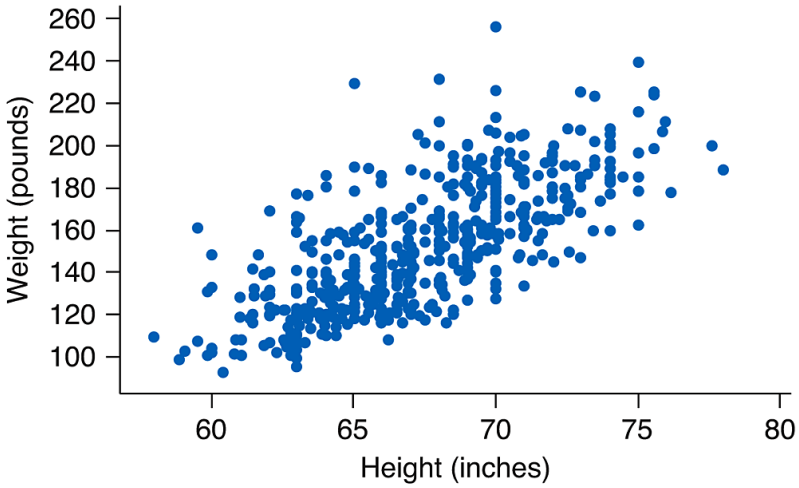
This data set shows an association between two variables, but it cannot be characterized as positive nor negative.

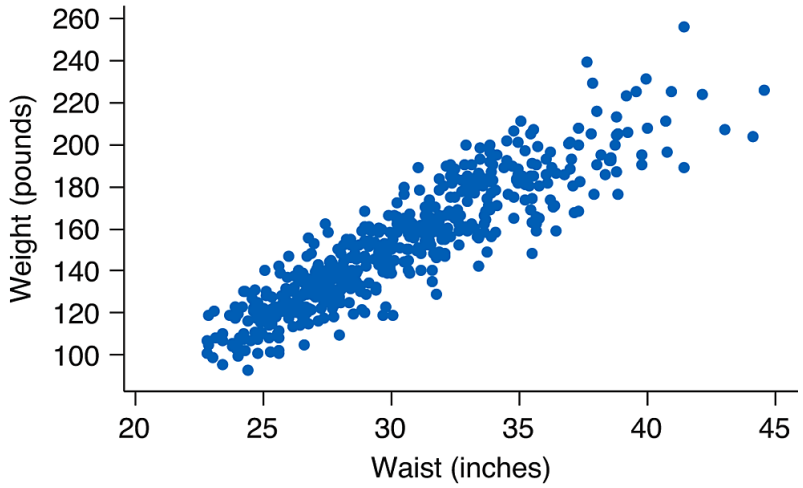
Trend: no correlation



No predictable pattern. Marathon running speed does not seem to be related to age of runner. For every age group we can find relatively fast and relative slow runners.

Strength of a correlation



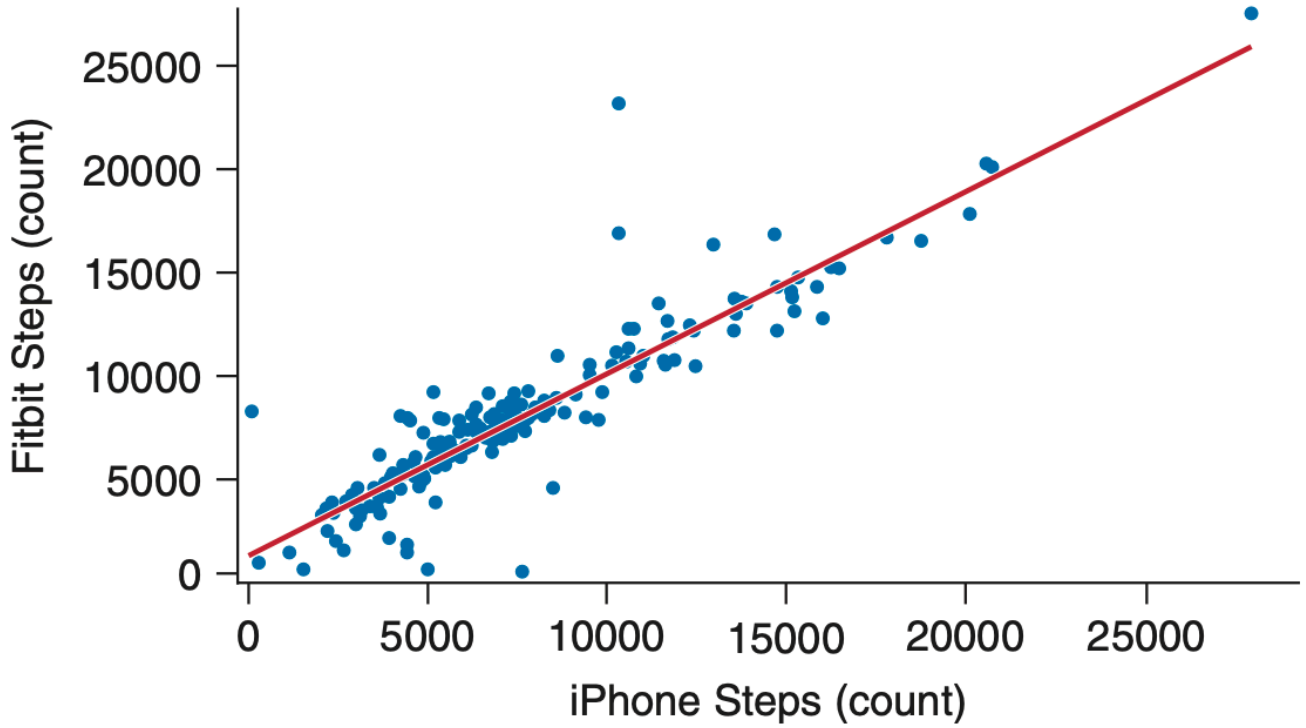


Scatterplots with large amounts of vertical variation indicate a **weak** association.

Scatterplots with small amounts of vertical variation indicate a **strong** association.

Shape: linear

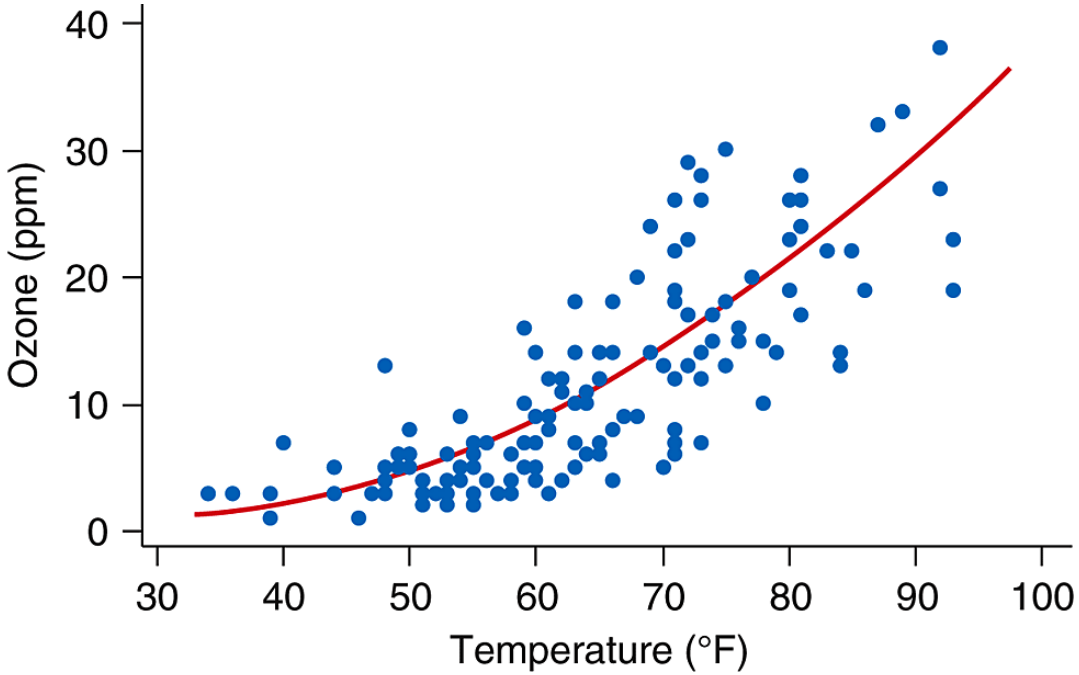
Data points tend to fit a straight line.



Linear association between steps counted by a Fitbit versus steps counted by an iPhone.

Shape: nonlinear

Data points tend to fit a curve rather than a line.



Non-linear trend between temperature and pollutant ozone levels.

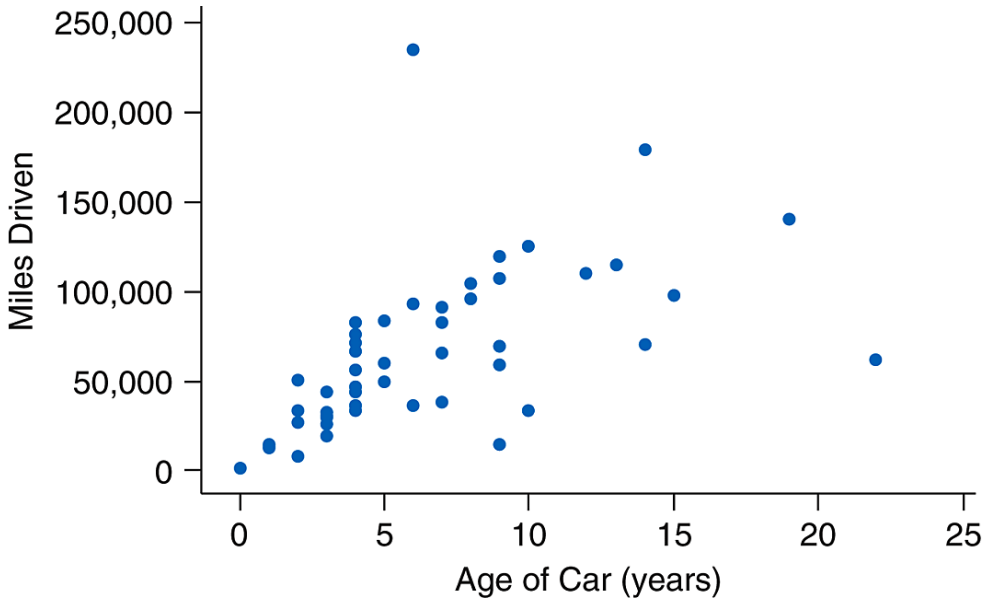
Describe a correlation / association

1. Trend: positive, negative, neither
2. Strength: weak, strong
3. Shape: linear, nonlinear

- Explain what all of these mean in the context of the data.
- Always use a phrase like “tends to” when describing an association because the association you are describing may not be true for all individuals.
- Always point out any data points that appear to be unusual or not part of the general pattern.

Example: Describing Associations

The association between the age and mileage of used cars is **positive and linear**. This means that older cars tend to have greater mileage.

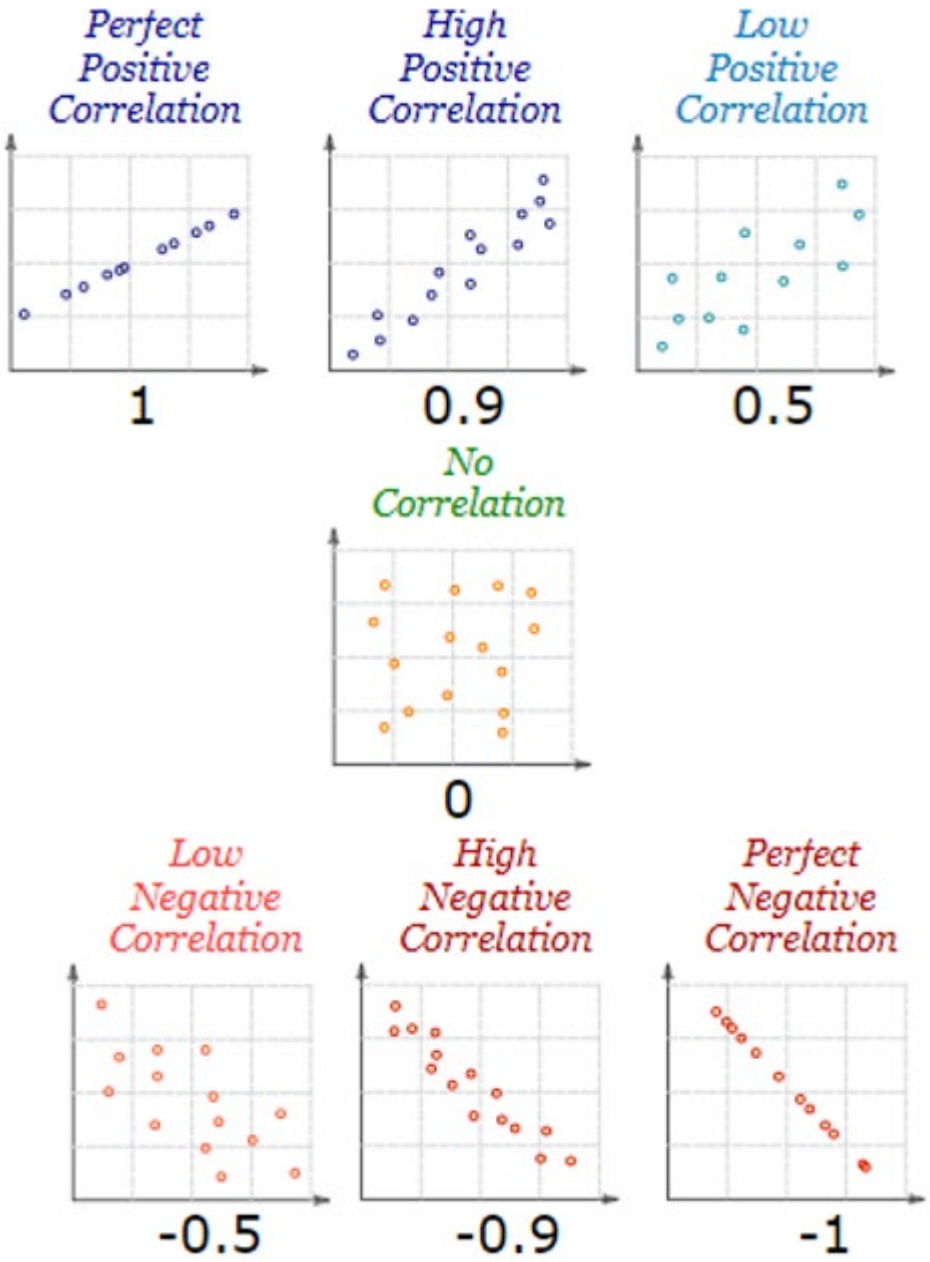


The association **is moderately strong**. There is one exceptional point: One car is only about 6 years old but has been driven many miles.

Section 4.2

Measure strength of linear correlation

- Correlation coefficient *r* measures the strength of a **linear** correlation.
- *r* is always between −1 and 1
- $r \approx 1$ indicates a strong linear positive correlation
- $r \approx -1$ indicates a strong linear negative correlation
- $r \approx 0$ indicates a weak or no linear correlation



General rule of thumb for correlation:

Strong: $0.8 \leq r \leq 1$
Medium: $0.5 \leq r \leq 0.7$
Weak: $0.1 \leq r \leq 0.4$
None: $r = 0$
(same for negative values)

Find the correlation coefficient

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left(n(\sum X^2) - (\sum X)^2\right)\left(n(\sum Y^2) - (\sum Y)^2\right)}}$$

Find *r* for the data points $(1, 3), (2, 4), (3, 2)$.

The number of data points is $n = 3$.

$$r = \frac{\mathbf{3(3 + 8 + 6) - (1 + 2 + 3)(3 + 4 + 2)}}{\sqrt{(\mathbf{3(1^2 + 2^2 + 3^2) - (1 + 2 + 3)^2)(3(3^2 + 4^2 + 2^2) - (3 + 4 + 2)^2)}}} = \mathbf{-0.5}$$

Find the correlation coefficient

Using StatCrunch:

- Stat → Regression → Simple Linear
- Select the columns for *x*-variable, *y*-variable
- Select **Compute**.

The table below shows the heights and weights for 6 people. Compute and interpret *r*.

Height	61	62	63	64	66	68
Weight	104	110	141	125	170	160

Notes about *r*

- Changing the order of the variables *(i.e., switching x and y)* does not change *r*.
- *r* has no units
- *r* is only useful to measure a linear trend. Try to graph your data first to make sure the association is linear before computing *r*.

End of 4.2