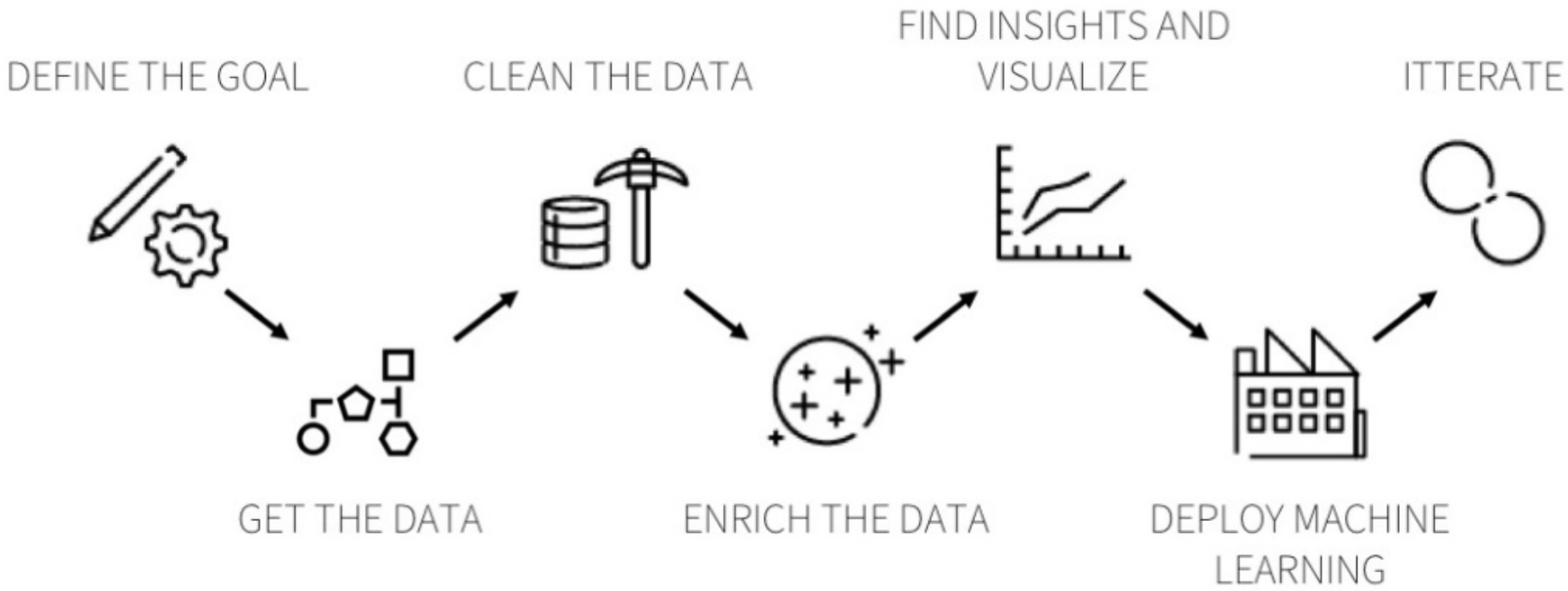
Exploring Data with `Statistics`

The Data Science Process



Agenda

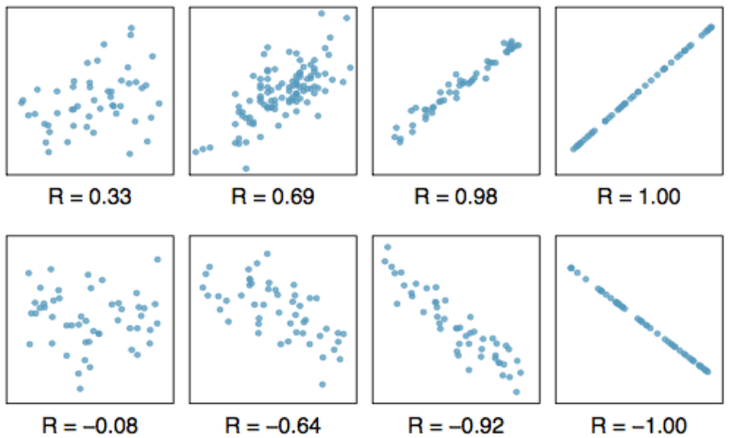
- Correlation
- Pearson's Correlation Coefficient and its limitations

Correlation

- Two variables are correlated **if knowing one gives you information about the other**.
  - For example, height and weight are related; people who are taller tend to be heavier.
- A **correlation** is a statistic intended to quantify the **strength** and **direction** of the relationship between two variables.
  - The mathematical formula for calculing the correlation coefficient is:

$$\rho = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

- First, we draw scatter plot of data. Place the independent variable along the x-axis, and the dependent variable along the y-axis.
- Note whether the data looks as if it lies approximately along a straight line.
  - If it has some other pattern, for example, a curve, then the linear correlation coefficient should not be used.



Interpreting ρ (Pearson's Correlation)

Calculated easily in Pandas using the `pd.DataFrame.corr()` method, which produces a **Correlation Matrix.**

- Pearson's correlation is always between -1 and +1.
- The **sign** of ρ indicates the **direction of the correlation**. If ρ is positive, we say that the correlation is positive, which means that the variables are directly proportional. If ρ is negative, the correlation is negative, which implies that the variables are inversely proportional.
- The **magnitude** of ρ indicates the **strength of the correlation**. If ρ is 1 or -1, the variables are perfectly correlated, which means that if you know one, you can make a perfect prediction about the other.

NOTE

> Pearson's correlation only measures **linear relationships**. If there's a nonlinear relationship, ρ understates its strength. If the ρ between two variables is near 0, we cannot conclude that there is no relationship (we may claim that there is no *linear* relationship.)

Scatterplots and Correlations

- The top row shows linear relationships with a range of correlations;
  - you can use this row to get a sense of what different values of ρ look like.
- The second row shows perfect correlations with a range of slopes, which demonstrates that **correlation is unrelated to slope**
- The third row shows variables that are **clearly related, but because the relationship is non-linear, the correlation coefficient is 0.**

> The moral of this story is that you should **always look at a scatter plot of your data before blindly computing a correlation coefficient**.

Summary

Pearson's correlation works well if

- the relationship between variables is linear and
- if the variables are roughly normal.

> But it is not robust if the relationship isn't linear, or in the presence of outliers.

Correlation Does Not Imply Causation

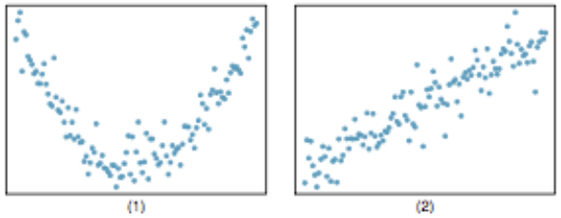
> Something other than X (but related to X) could be causing the changes in Y. So we can use X to predict Y, but a change in X does not necessarily cause a change in Y.

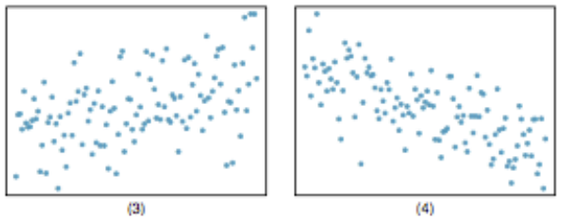
A classic example is that, before the polio vaccine, there was a positive correlation between sales of soda and the outbreak of polio. This doesn't mean that soda caused polio. So what was going on? It turned out that polio spread more easily in the summer when many people played together at pools and on the beach. That was also the time of year when soda sales went up.

Quiz

7.7 Match the correlation, Part I. Match the calculated correlations to the corresponding scatterplot.

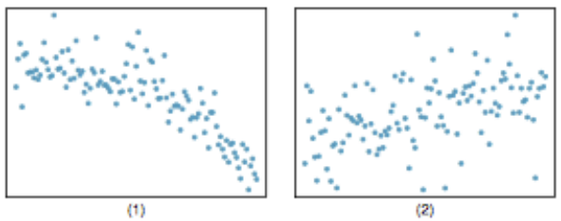
(a) $r = -0.7$
(b) $r = 0.45$
(c) $r = 0.06$
(d) $r = 0.92$

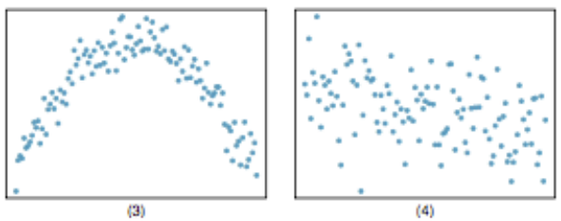




7.8 Match the correlation, Part II. Match the calculated correlations to the corresponding scatterplot.

(a) $r = 0.49$
(b) $r = -0.48$
(c) $r = -0.03$
(d) $r = -0.85$





- Determine if the following statements are true or false. If false, explain why.
  - A correlation coeffcient of -0.90 indicates a stronger linear relationship than a correlation coeffcient of 0.5.
  - Correlation is a measure of the association between any two variables.

Assignment

7.17 Correlation, Part I. What would be the correlation between the ages of husbands and wives if men always married woman who were

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

7.18 Correlation, Part II. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

(a) \$5,000 more than women?

(b) 25% more than women?

(c) 15% less than women?