MSCI152: Introduction to Business Intelligence and Analytics

Lecture 8: Correlation

Dr Anna Sroginis

Lancaster University Management School

Agenda

Measures of Association Between Two Variables

1. Qualitative: Association between categorical variables
2. Quantitative: Association between numerical variables (correlation)

More details can be found in Camm et al., Section 2.8

Overview

Measures of Association Between Two Variables

1. Qualitative: Association between categorical variables
2. Quantitative: Association between numerical variables (correlation)

Survey example

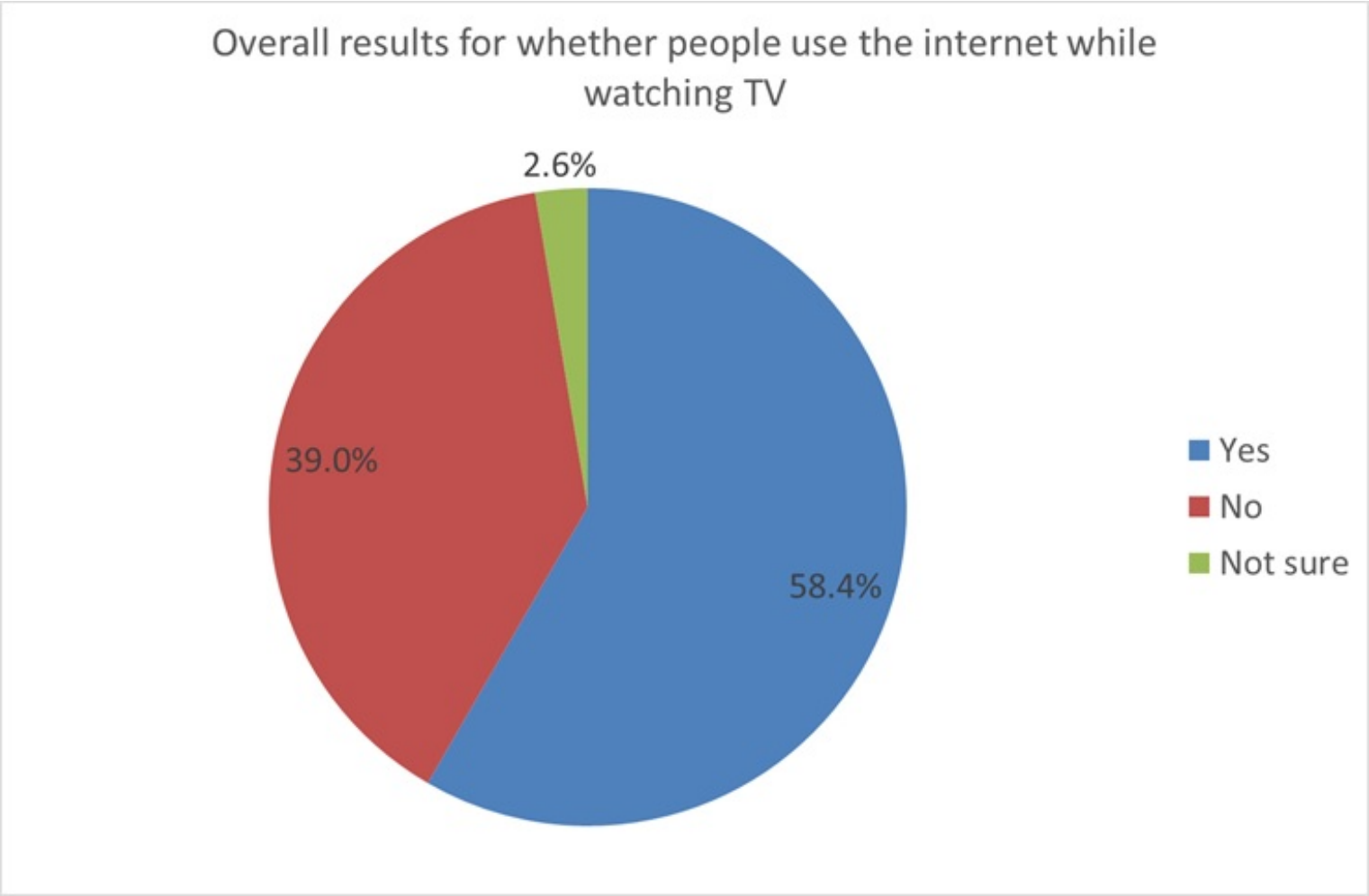
- Do you ever use the Internet at the same time as you are watching programmes on your TV?

Overall answers:

Yes	1169	58.4%
No	781	39.0%
Not sure	53	2.6%
Total	**2003**	**100.0%**

Qualitative Data: Survey Example

For such surveys, **percentages** are relevant:



Qualitative Data: Survey Example

We are often interested in splitting the respondents into different groups:

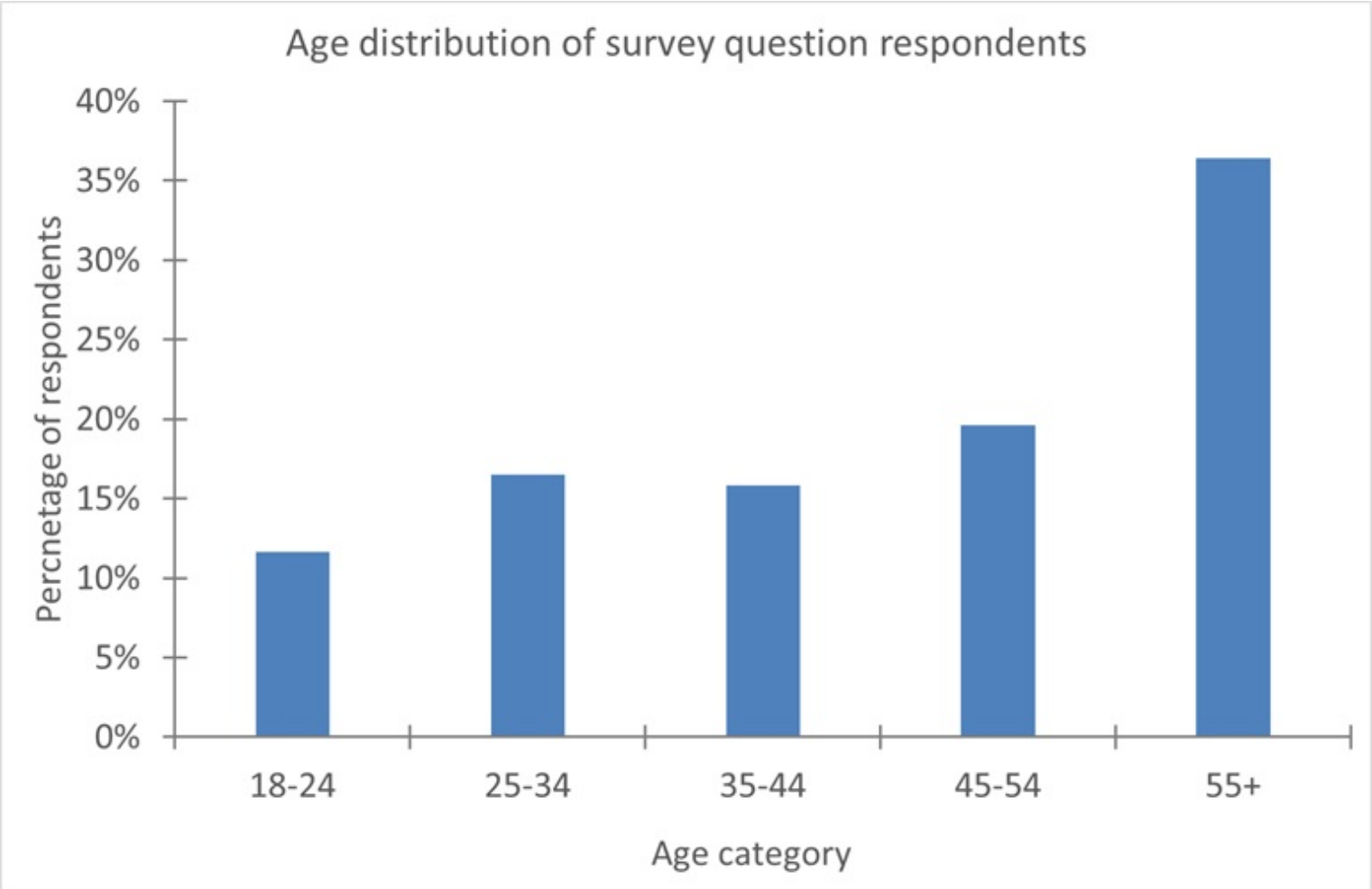
- e.g. age, gender, region, social group
- the survey data does all of these

The survey data is split into age groups:

Age	18-24	25-34	35-44	45-54	55+	Total
No.	233	331	317	393	729	2003
%	11.6%	16.5%	15.8%	19.6%	36.4%	100%

Qualitative Data: Survey Example

The categories are **ordered**



Survey Example: Contingency Table

Age	18-24	25-34	35-44	45-54	55+	Total
Yes	174	254	229	227	285	1169
No	51	64	83	158	425	781
Not sure	8	13	5	8	19	53
Total	233	331	317	393	729	2003

- Columns are one category: **Age**
- Rows are the other category: **Answer**
- Cells are **mutually exclusive**: each case appears in one cell
  - **Not** the case if a multiple response question

Survey Example: Contingency (Crosstabs) Table

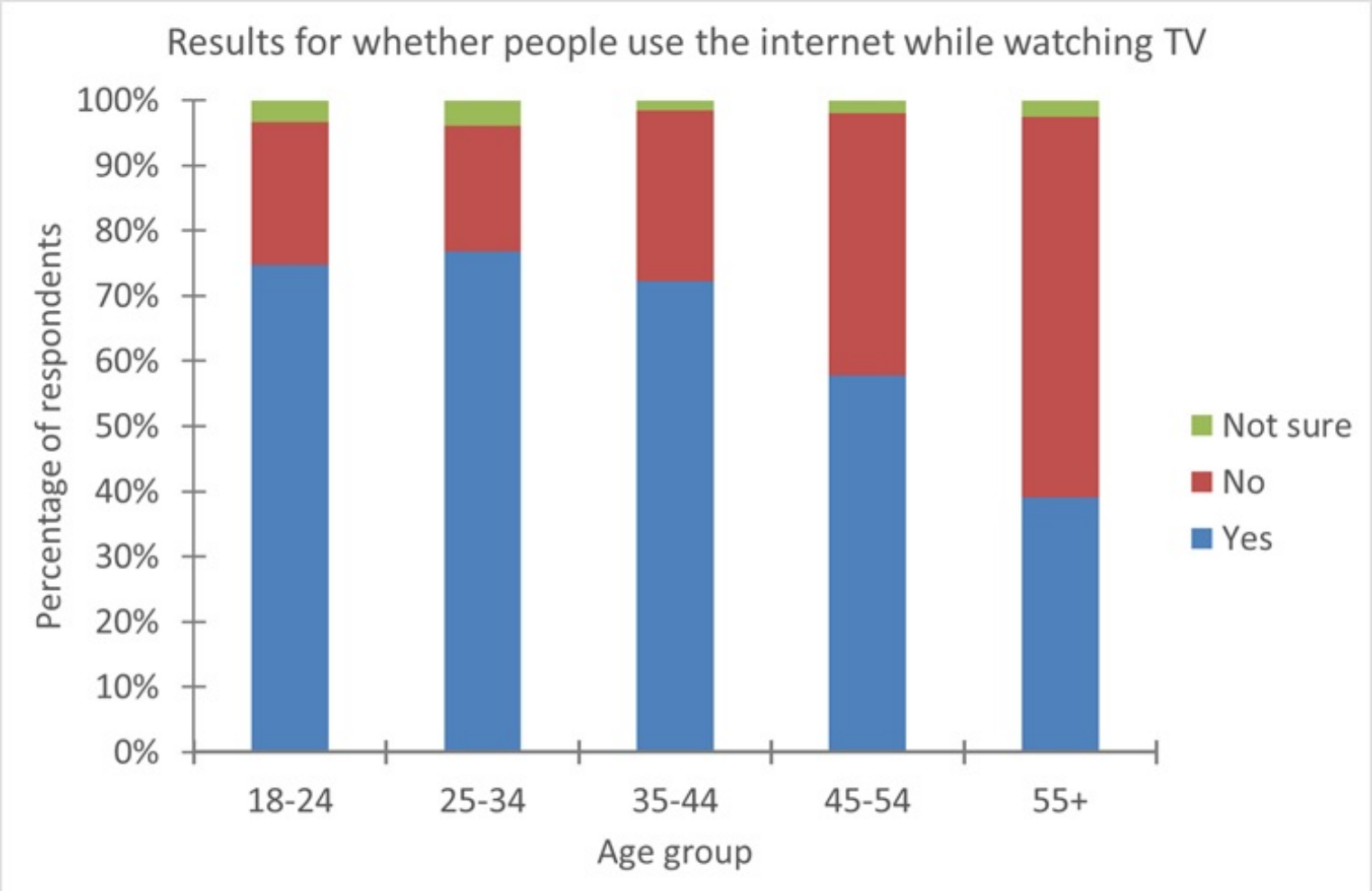
Age	18-24	25-34	35-44	45-54	55+	Total
Yes	74.7%	76.7%	72.2%	57.8%	39.1%	58.4%
No	21.9%	19.3%	26.2%	40.2%	58.3%	39.0%
Not sure	3.4%	3.9%	1.6%	2.0%	2.6%	2.6%
Total	100%	100%	100%	100%	100%	100%

Is there a **difference** in percentages between the age groups?

Look at the columns and see if there is a difference

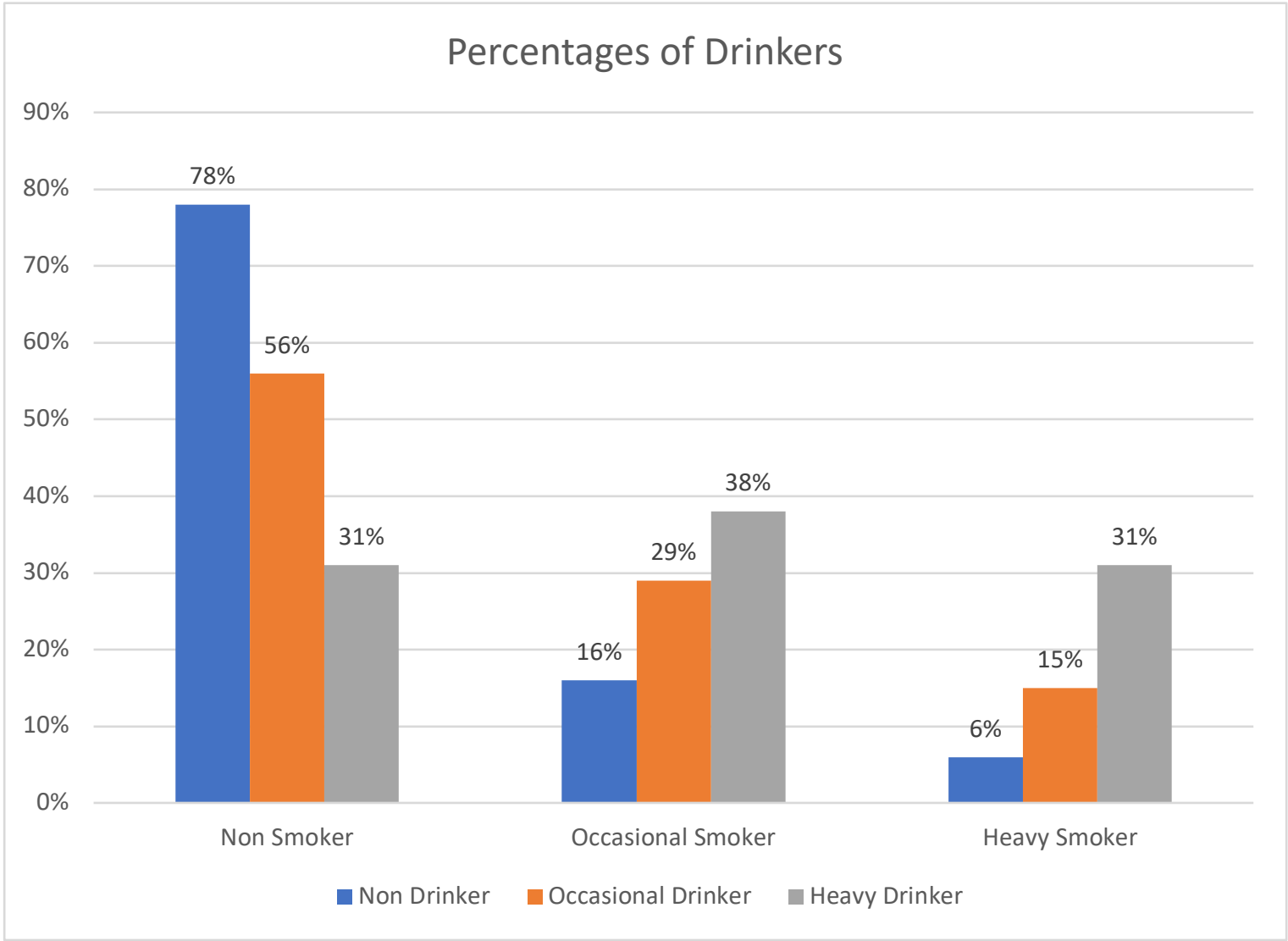
- 18–24, 25–34, 35–44 are fairly similar
- 45–54 has smaller % saying "Yes"
- 55+ has an even smaller % saying "Yes"

Survey Example: Stacked Column Chart



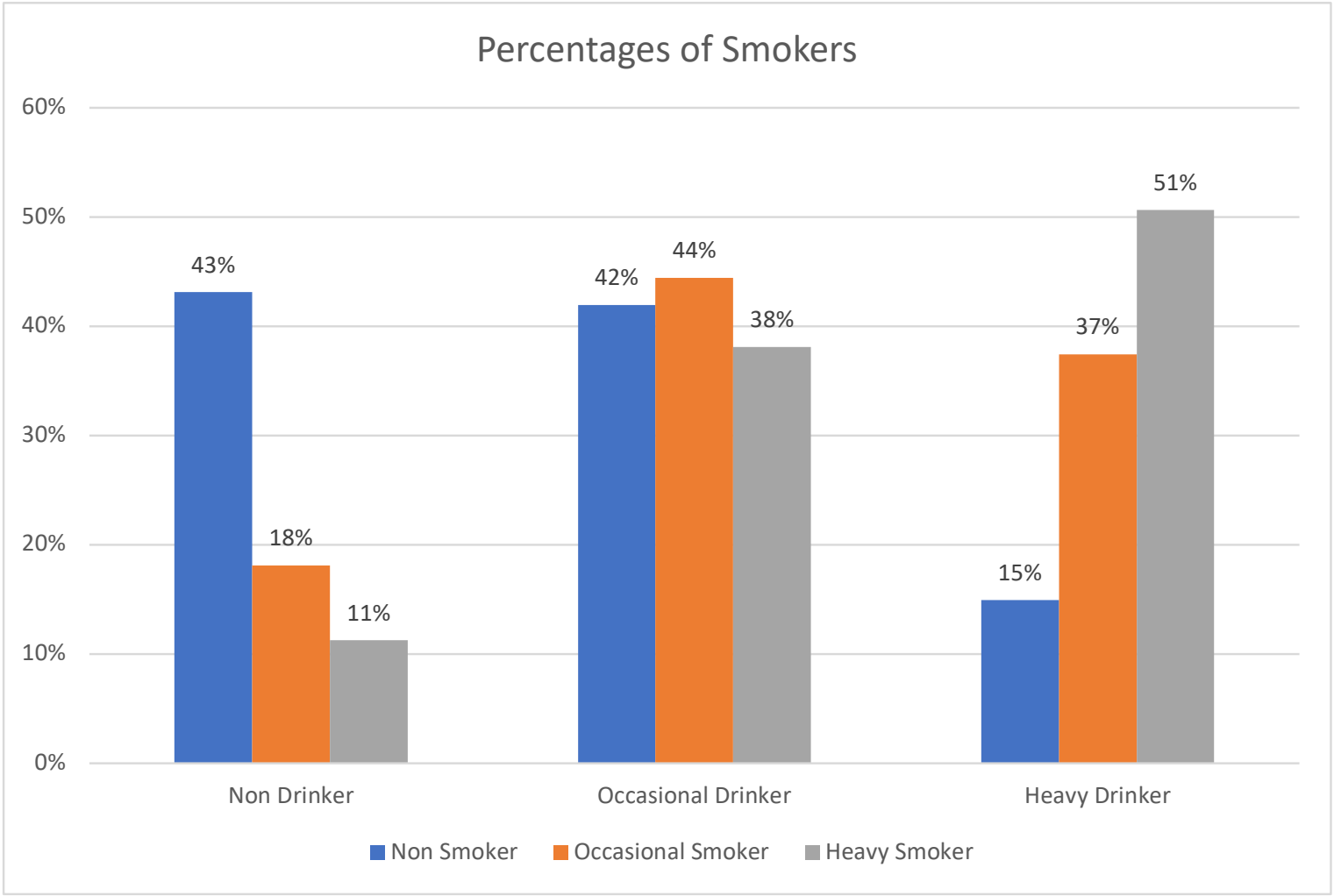
Discuss it in your groups

Can we say that smoking and drinking habits are related?



Discuss it in your groups

Can we say that smoking and drinking habits are related now?



Overview

Measures of Association Between Two Variables

1. Qualitative: Association between categorical variables

2. Quantitative: Association between numerical variables (correlation)

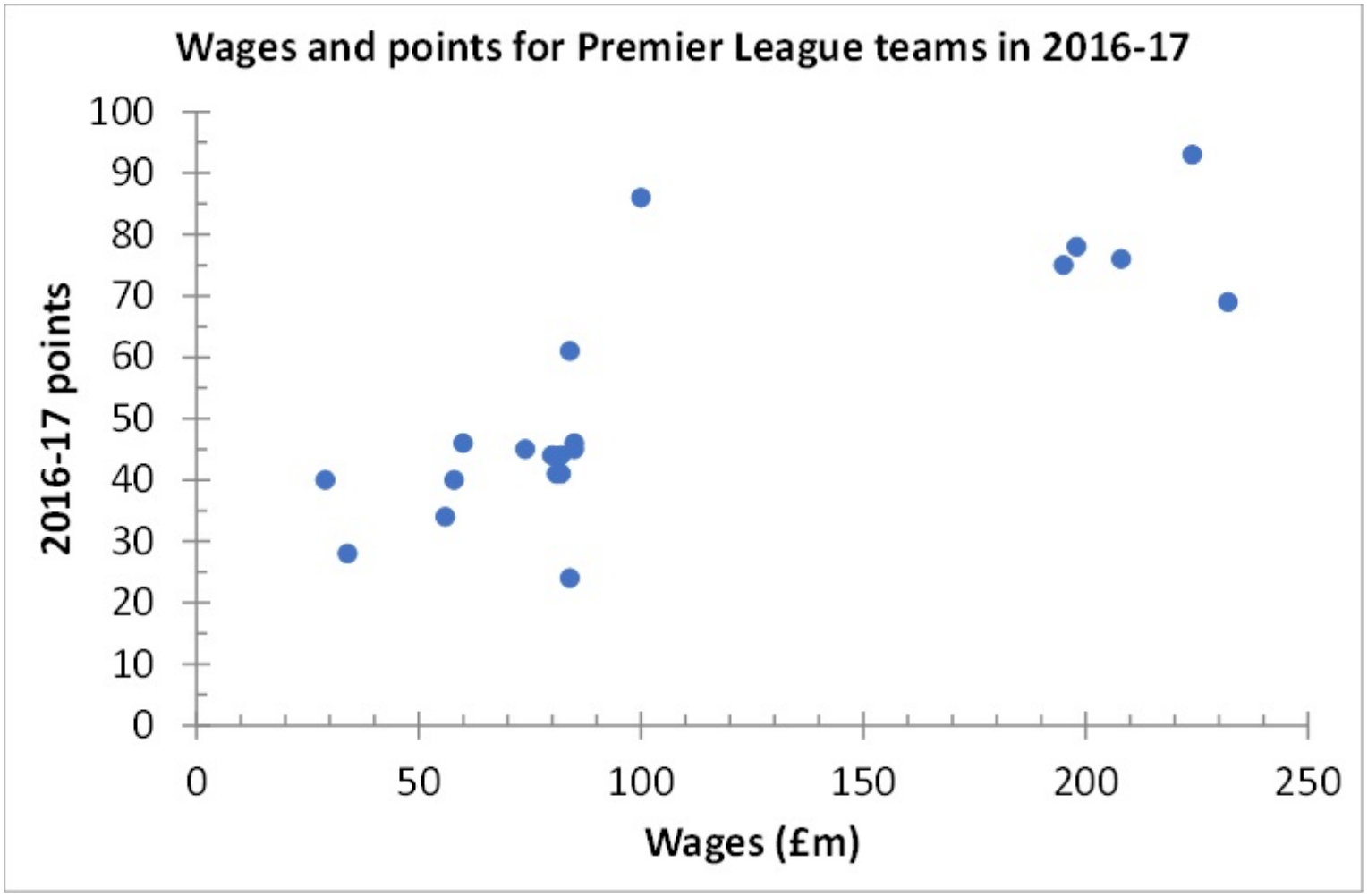
Example 1: Premier League

What variable might you use to measure **performance**?

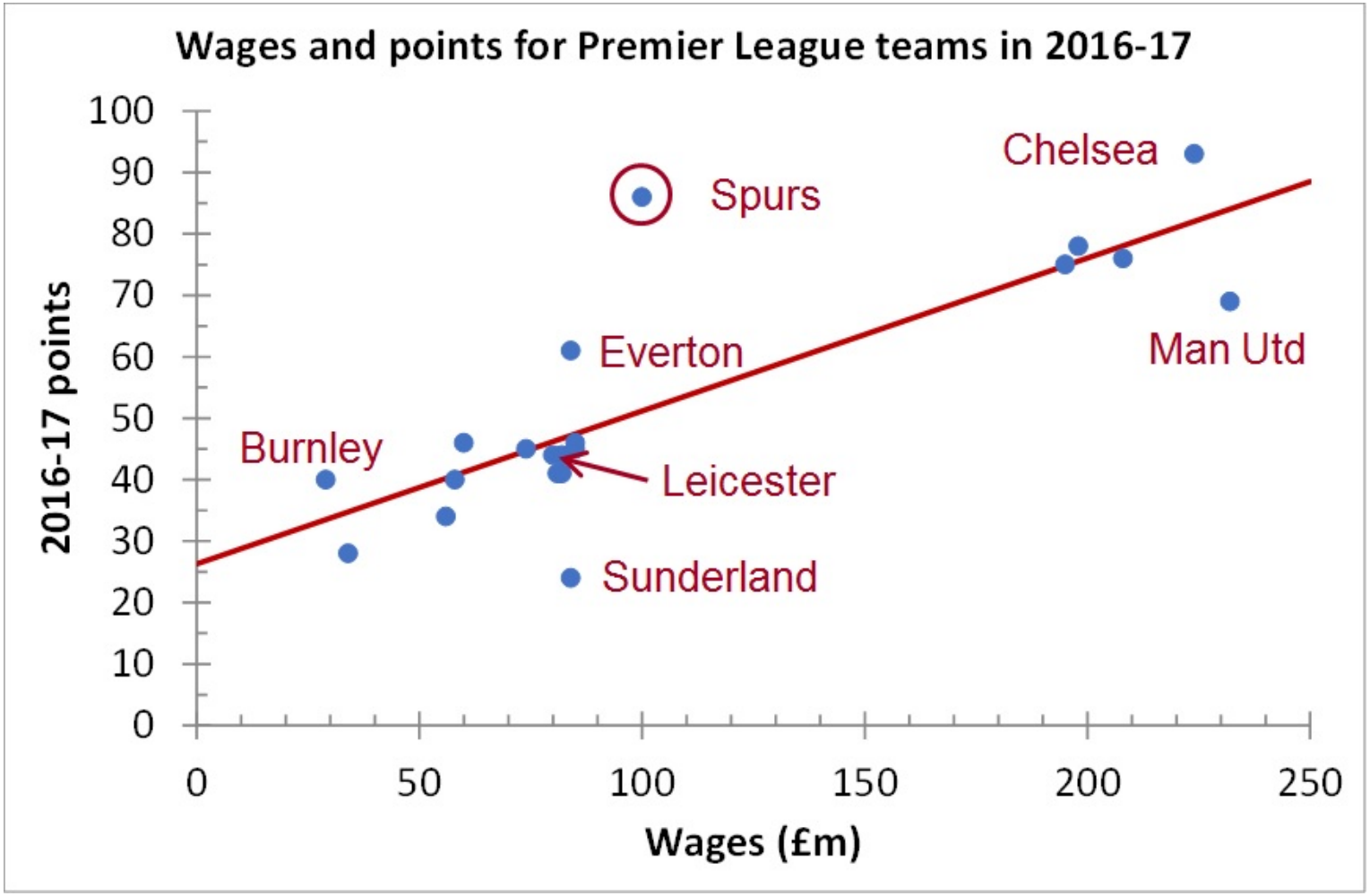
What **factors** might affect performance? How might we **measure** them?

Data Source: https://www.theguardian.com/football/2017/jun/01/premier-league-finances-club-by-club

Example 1: Wages and league points



Example 1: Wages and league points



Correlation

Examining relationships between two variables:

- For each element in our sample we record two values
- These values are our two variables

Looking for association

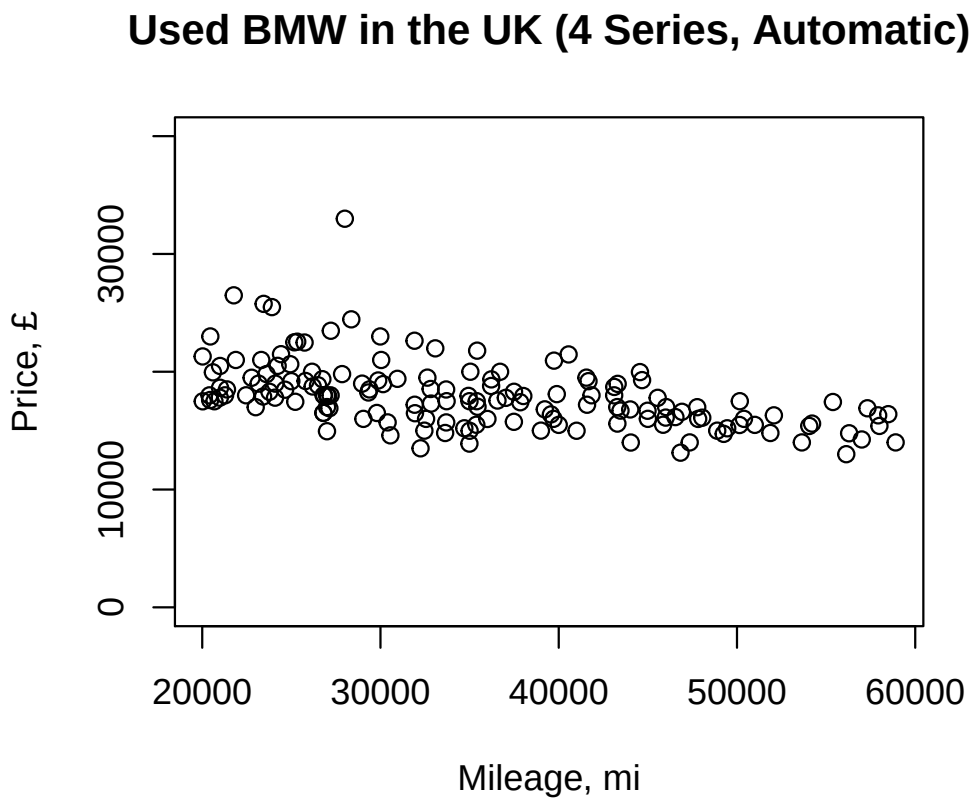
- as for categorical variables: the distribution of one variable depends on the value of the other

Leads on to regression

- model the relationship

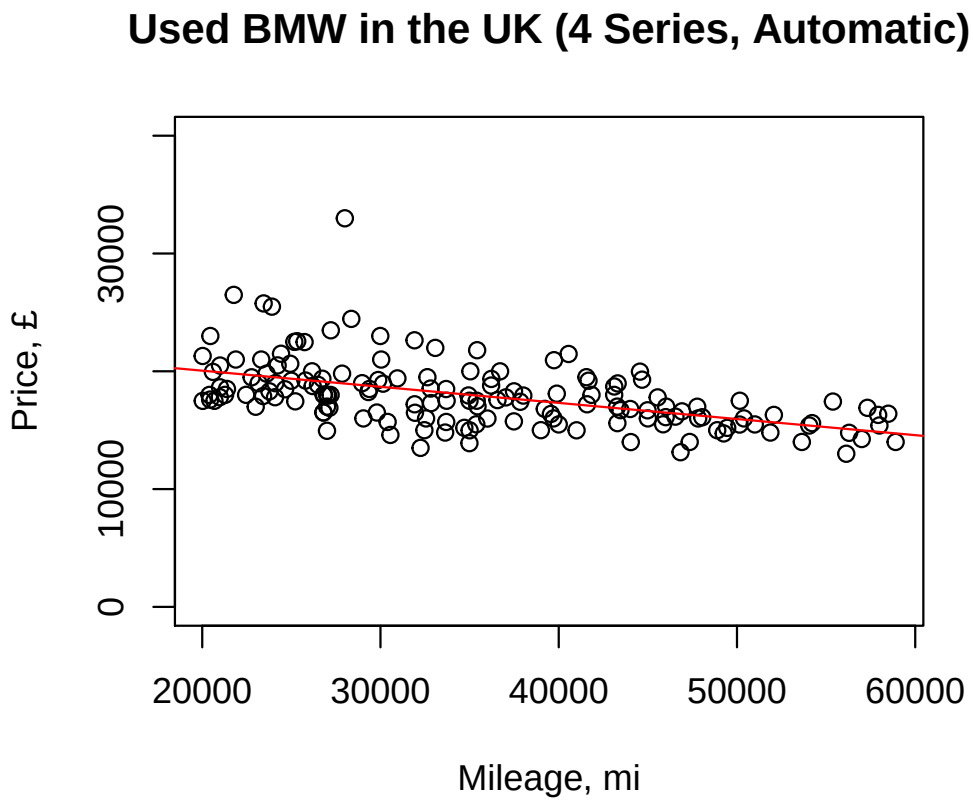
Example 2: BMW cars

Consider the following scatterplot of BMW cars, prices vs the mileage (mpg):



What can we conclude based on that?

Example 2: BMW cars



There is a connection between price and mileage:

- With the growth of mileage, the price tends to reduce,
- This is not a strict linear relation,
- But it is strong nonetheless.

Important notes about x and y

Definition

x is the **explanatory** (or independent) variable
y is the **response** (or dependent) variable

Important to get these the right way round

Ways of thinking about this:

- **y** is the variable we are trying to predict
- **x** is the variable we can control (sometimes)
- we think that x might cause y (but see later: the data can't prove this)

To quantify the association, use covariance and correlation.

Correlation coefficient

We can introduce a coefficient, which will measure the linear relation between the variables. It's called "Pearson's correlation coefficient":

Pearson's correlation coefficient

$$\text{Correlation } = r = corr(x, y) = \frac{cov(x, y)}{s_X s_Y}$$

where

- $cov(x, y)$ is the covariance between x and y:

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

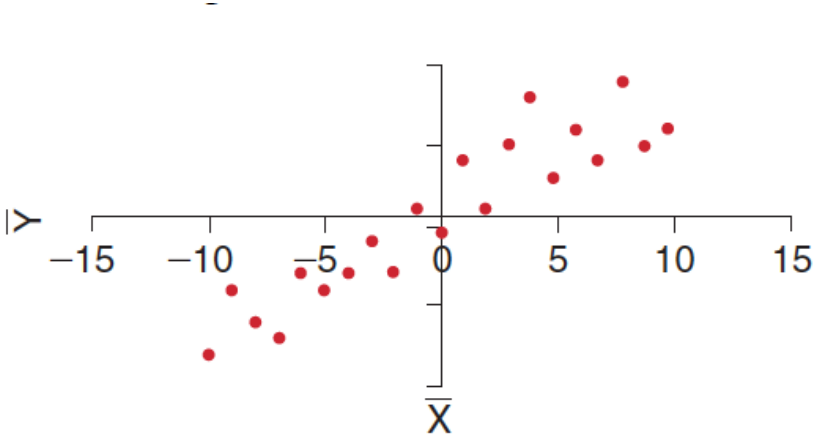
- s_X is the sample standard deviation for variable x
- s_Y is the sample standard deviation for variable y

Correlation coefficient

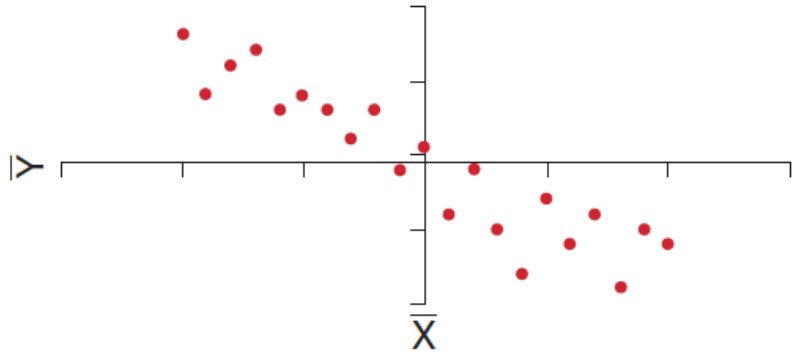
Properties of the coefficient:

1. $r_{x,y}$ **lies between -1 and 1**;
2. The values closer to -1 or 1 imply stronger linear relationship;
3. $r_{x,y} = 0$ implies no **linear** relationship;
4. $r_{x,y} = 1$ or $r_{x,y} = -1$ means that there is a perfect linear relationship (just a straight line of dots);
5. The sign of $r_{x,y}$ tells whether the relationship is positive or negative;
6. The value of $r_{x,y}$ tells you **nothing** about the steepness of the relationship.

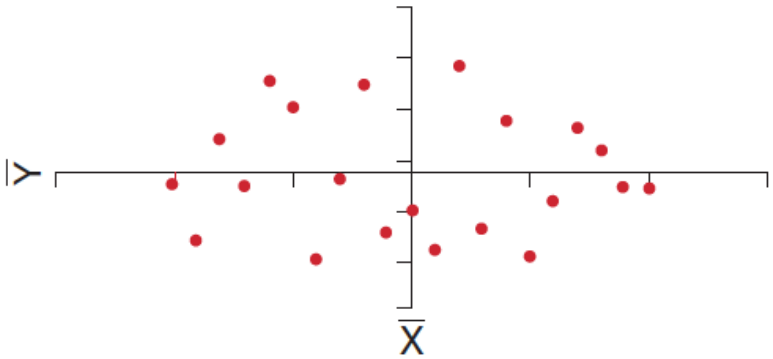
First, draw a scatter graph



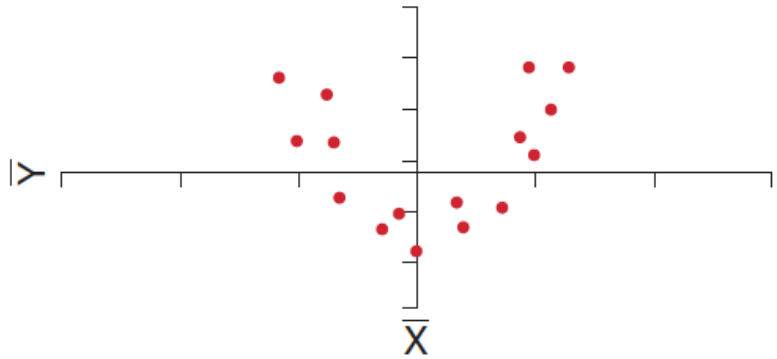
(a) Positive Correlation



(b) Negative Correlation

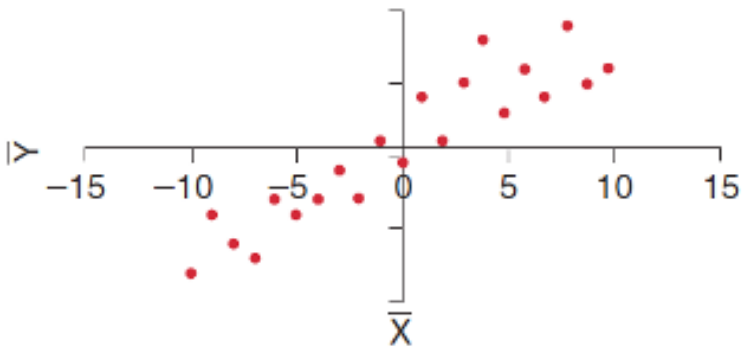


(c) No Correlation

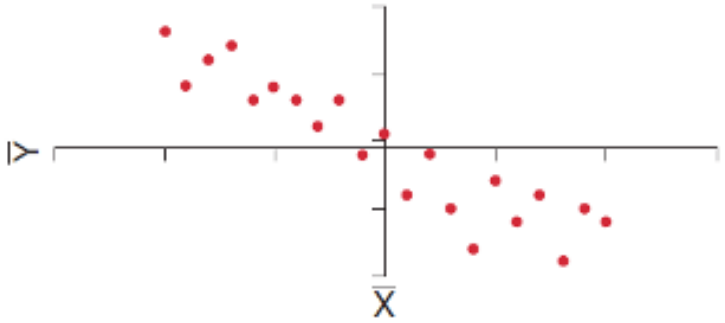


(d) A Nonlinear Relationship with No Linear Correlation

Then identify the sign



(a) Positive Correlation



(b) Negative Correlation

(a) **Perfect positive correlation**

- can draw an exact straight line through all points, with a positive slope.
- As x increases, so does y – a **positive** relationship

(b) **Perfect negative correlation**

- can draw an exact straight line through all points, with a negative slope.
- As x increases, y decreases – a **negative** relationship

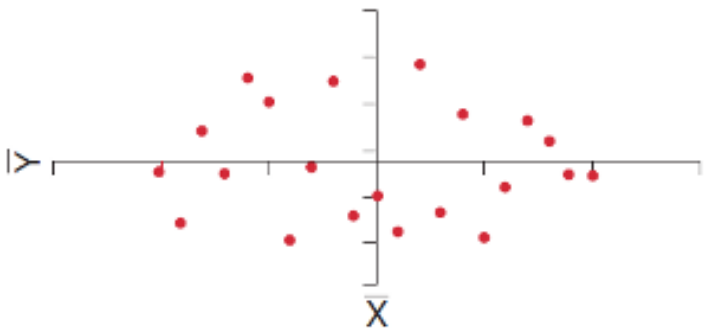
Then identify the sign

(c) **No correlation**

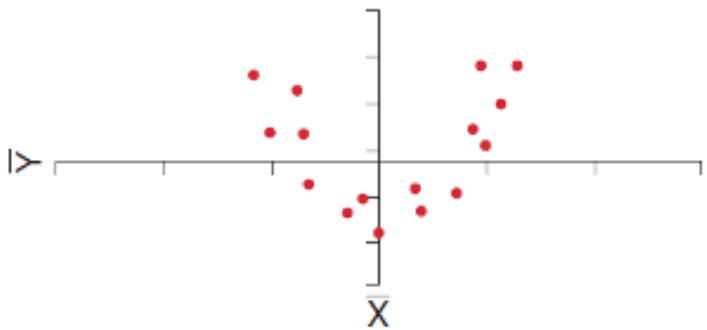
- An increase or decrease in x doesn't appear to affect the value of y.

(d) **Non-linear correlation**

- It is complicated.



(c) No Correlation



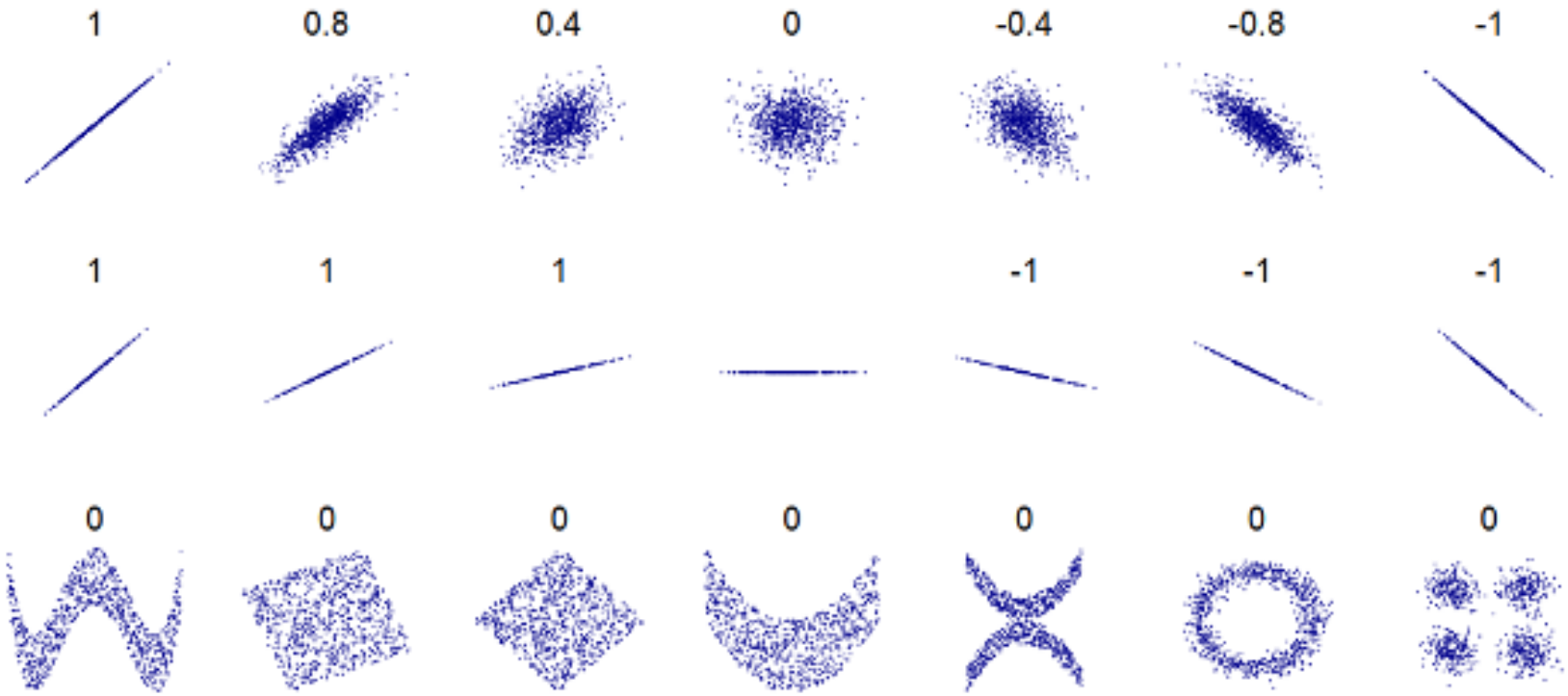
(d) A Nonlinear Relationship with No Linear Correlation

Describing scatter plots

1. **Direction:** does the trend go up or down?
2. **Curvature:** is the pattern linear or curved?
3. **Variation:** are the points tightly clustered around the trend?
4. **Outliers:** is there something unexpected?
5. **Amount of data:** lots of points or only a few?

Correlation coefficient

Examples of scatterplots vs correlation coefficient values (noise, slope effects and nonlinear patterns):



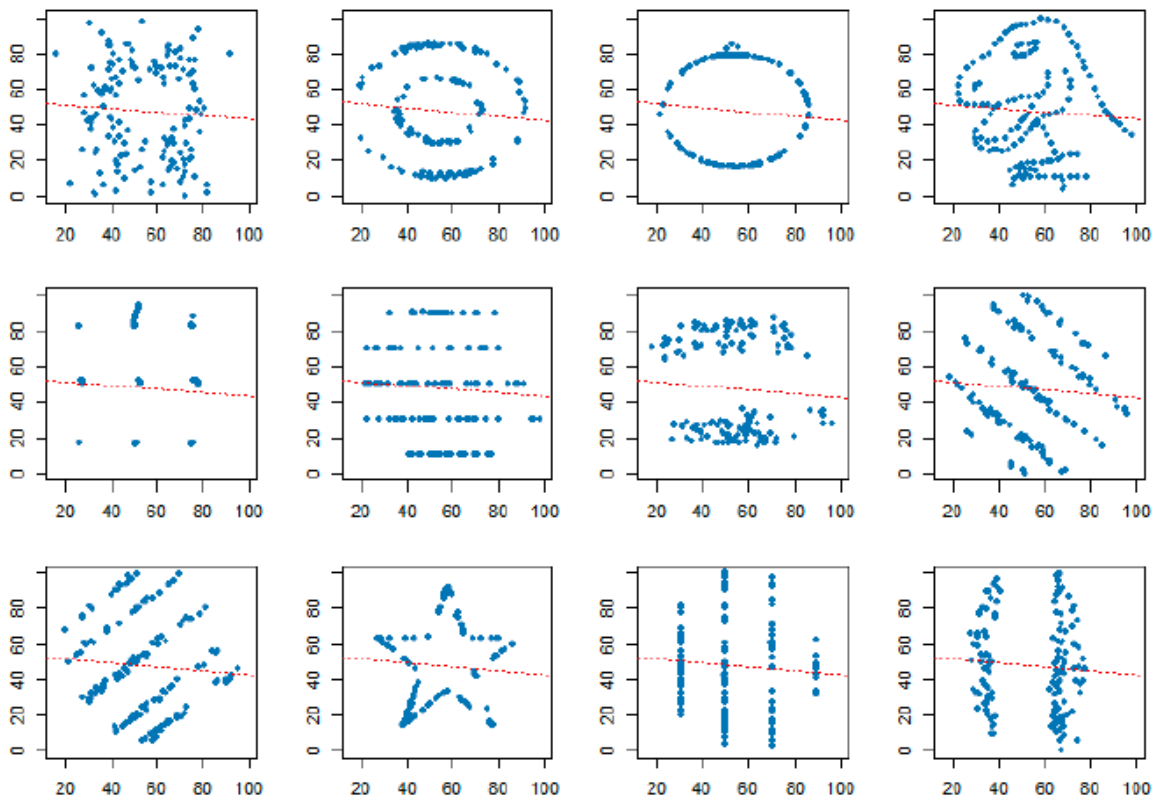
Source of the image: https://en.wikipedia.org/wiki/Correlation_and_dependence

So, **always plot your data**, don't just rely on $r_{x,y}$!

Always plot your data!

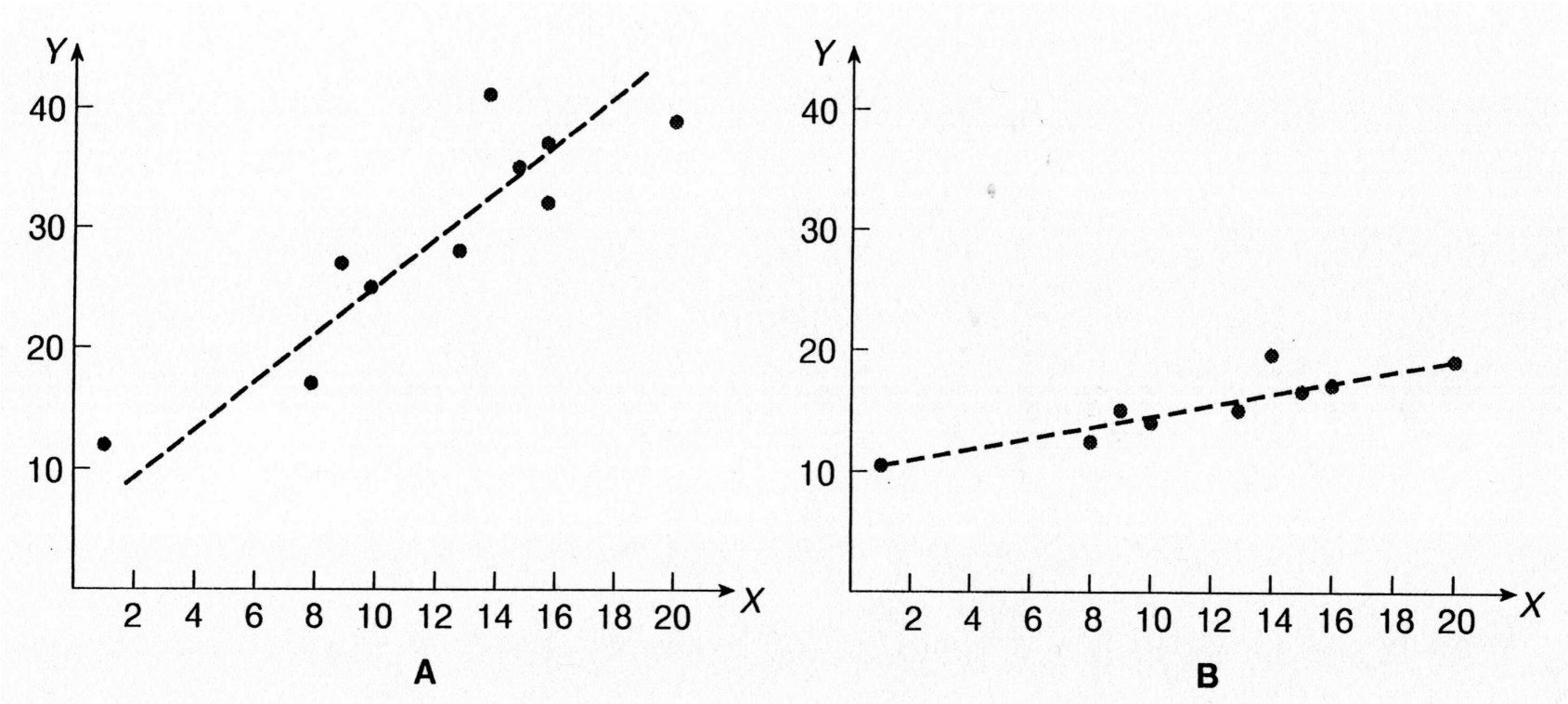
All 12 datasets have the same descriptive statistics with the same regression line: $y = 53.55 - 0.11x$.

X mean	54.26
Y mean	47.83
X st. dev	16.76
Y st. dev	26.93
Correlation	-0.06

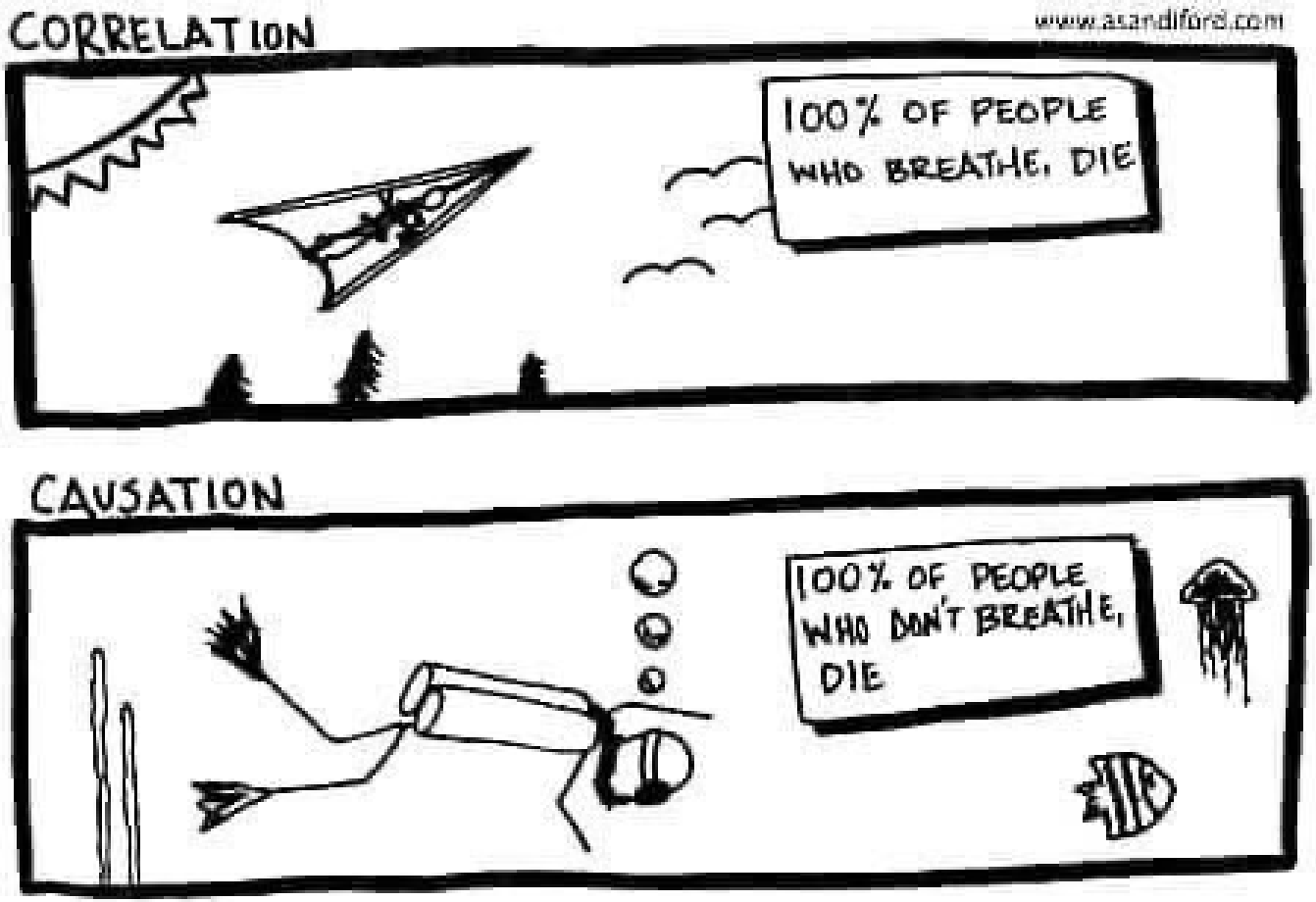


Correlation coefficient

Which correlation coefficient is bigger?



Correlation vs causation



- While causation and correlation can exist at the same time, correlation does not imply causation.
- Causation explicitly applies to cases where action A causes B.
- Correlation is simply a relationship. Action A relates to Action B — but one event doesn't necessarily cause the other event to happen.

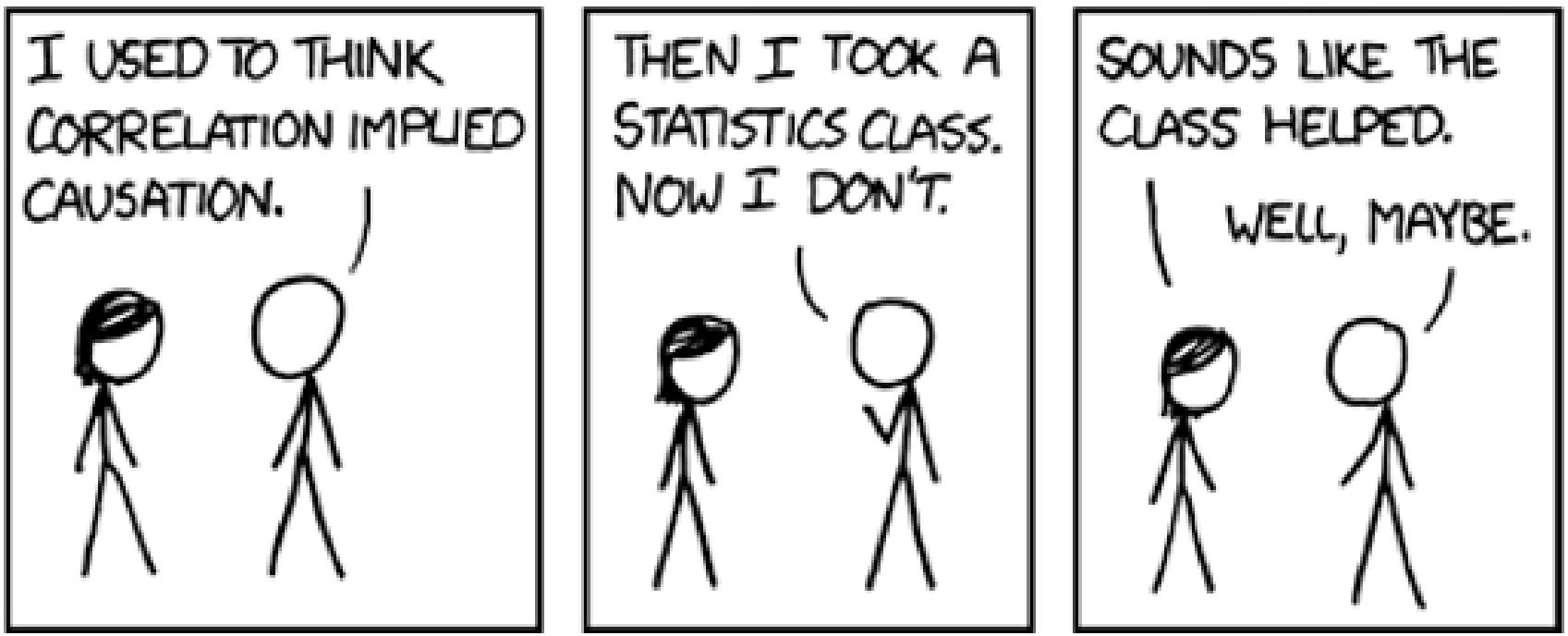
How do we interpret results?

A high correlation tells us there is a strong pattern.

We need to think about the most likely explanation, but the data cannot tell us. To some extent, the interpretation is subjective. However, we can use:

- Our knowledge of the situation
- Logic
- Imagination
- Further evidence and data

Correlation vs causation



Source: https://www.explainxkcd.com/wiki/index.php/552:_Correlation

Wrap up

Today we:

- Looked at the relationships between variables;
- Defined a correlation coefficient;
- Touched upon "correlation does not imply causation".

Next time:

- Introduction to the simple linear regression
  - Estimating the regression (least squares)
  - Confidence intervals
  - Measuring quality of a fit