Universidad
Francisco de Vitoria
UFV Madrid

Data visualisation

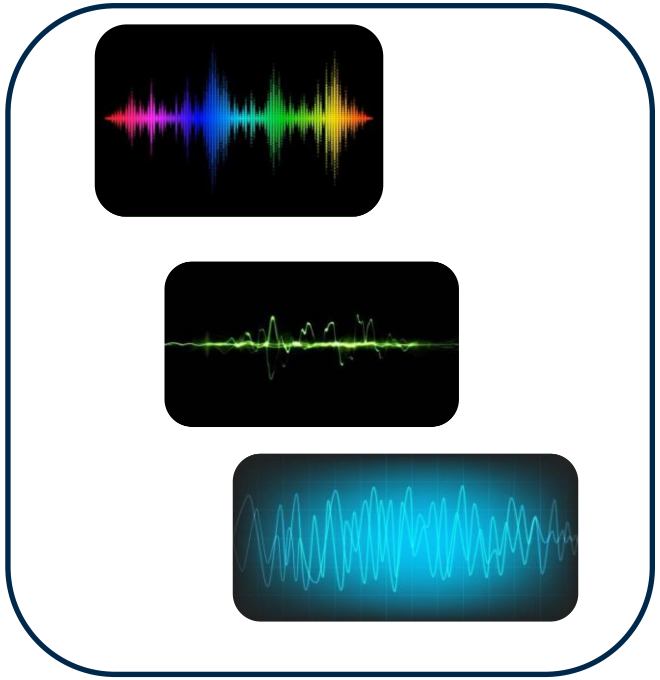
Luigi Borzì

DAUIN

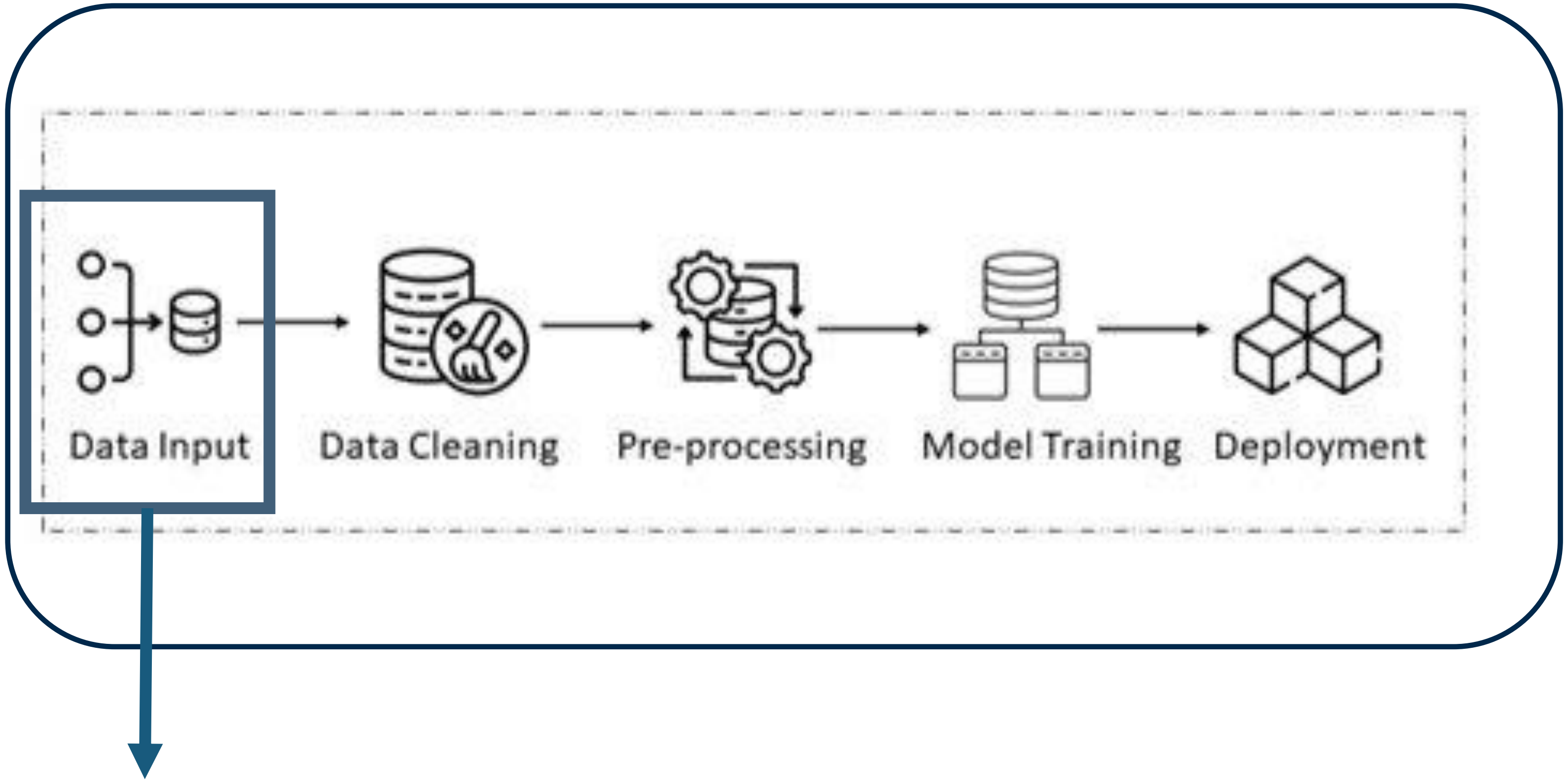
Politecnico
di Torino
1859

Overview

Data



Data processing



Data visualization and exploration

Data visualization and exploration

Data visualization and exploration

- The importance of data visualization
- Line plot
- Bar chart
- Pie chart
- Histogram
- Scatter plot
- Heat map
- Spider plot
- Correlation plot

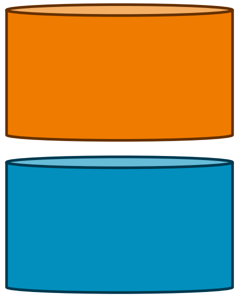
Data visualization: why?

By transforming big data into visual formats, data visualization tools allow us to comprehend complex data and make data-driven decisions effectively.

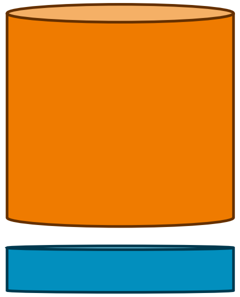
Data visualization allows for describing the dataset, understand the key components, potential and limitations. This is essential for the subsequent data processing steps.

Data visualization: why?

When performing classification, one of the most important points to consider is dataset unbalance.



If you want to classify subjects with or without a disease, you would expect a dataset like this, where 50% of the dataset is made of healthy and 50% of patients.



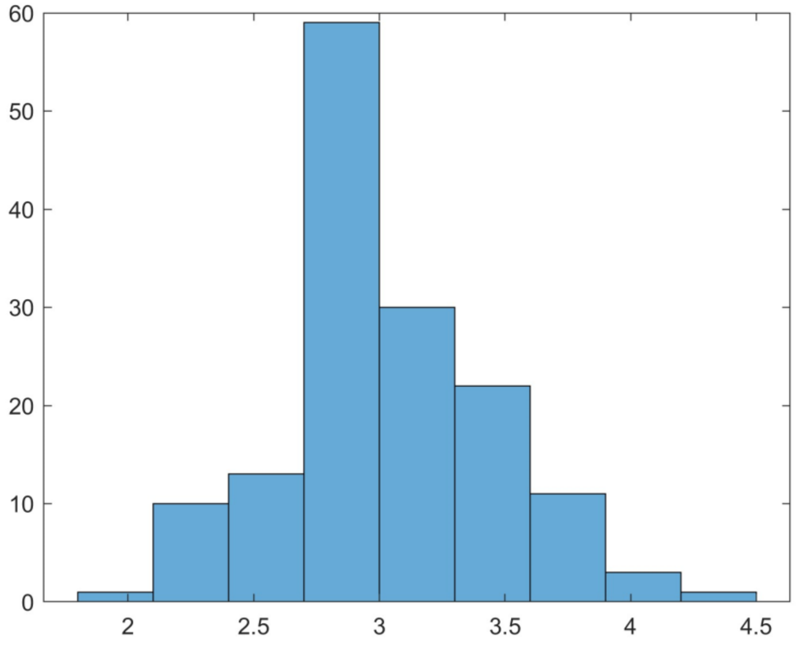
However, most of the times the dataset will be unbalanced.
You will have many more healthy subjects!
This can be problematic when using machine learning later on

BUT: there are methods to handle this problem
The most important thing is to UNDERSTAND the dataset and then decide how to process data.

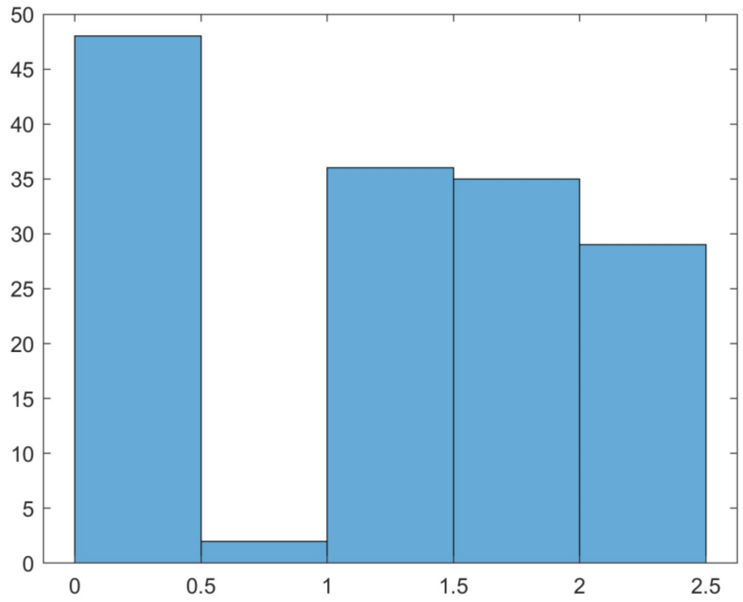
Data visualization: why?

Often it is important to assess the distribution of values in each variable. This is for several reason:

a) Specific statistical test are selected based on data distribution
b) Some models make assumptions on the distribution of data
c) Evaluate the scale/range of each variable is important



This is a normal distribution.
We will see it later

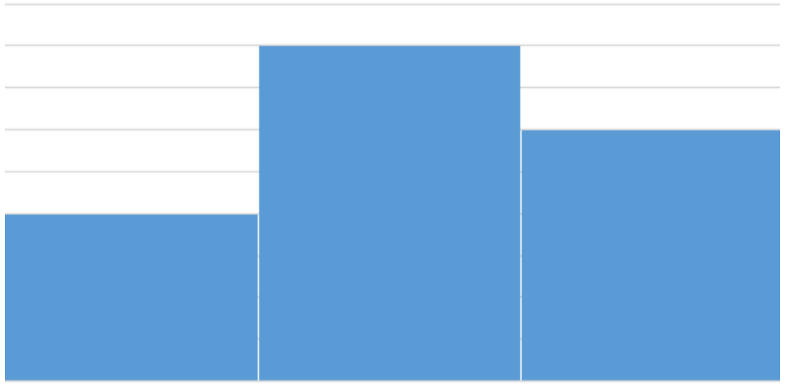


This is NOT a normal distribution.

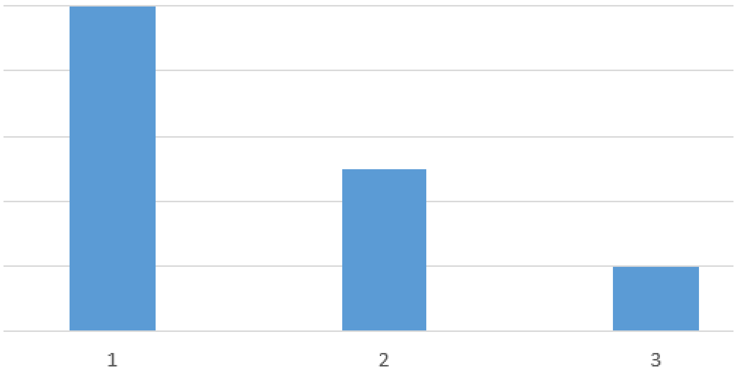
Data visualization: tools

PIE CHART

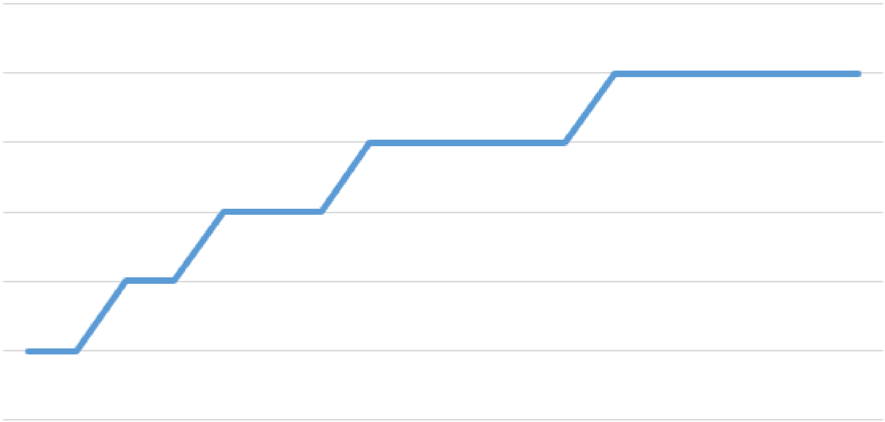
HISTOGRAM



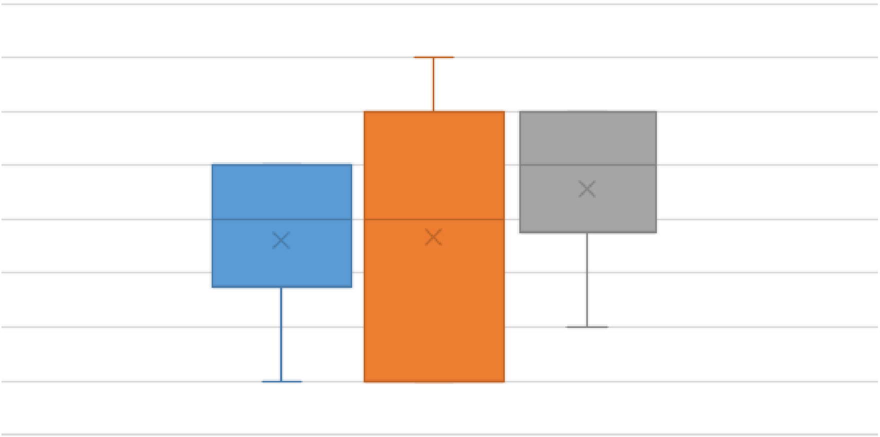
BAR CHART



LINE PLOT

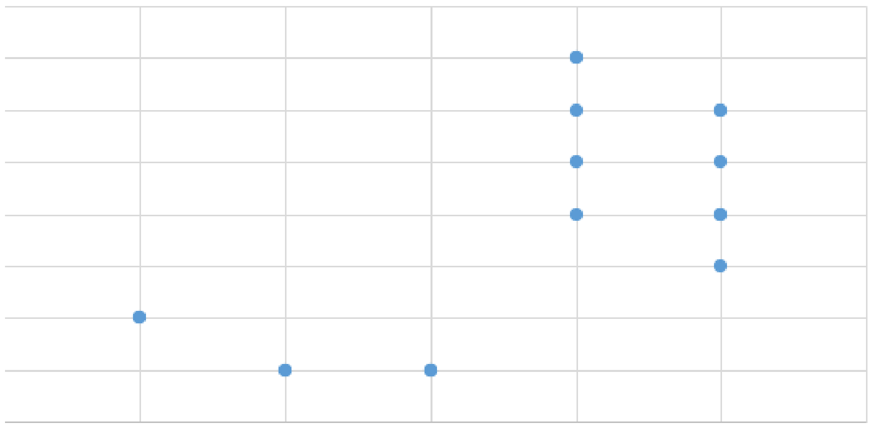


BOXPLOT

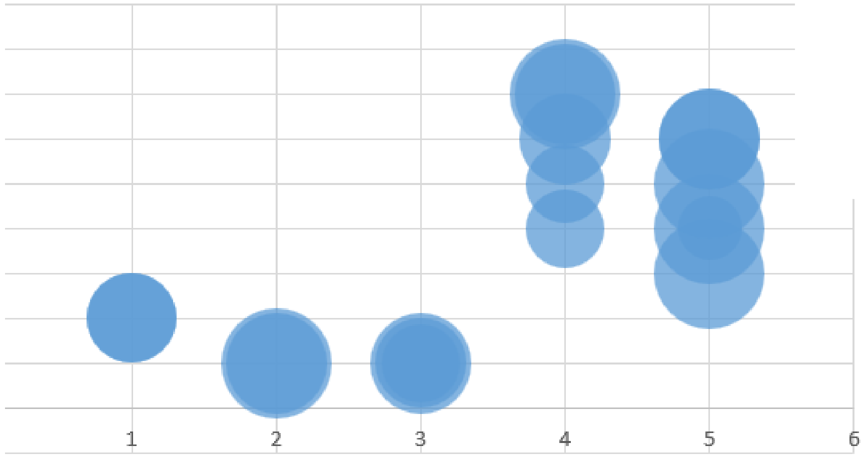


Data visualization: tools

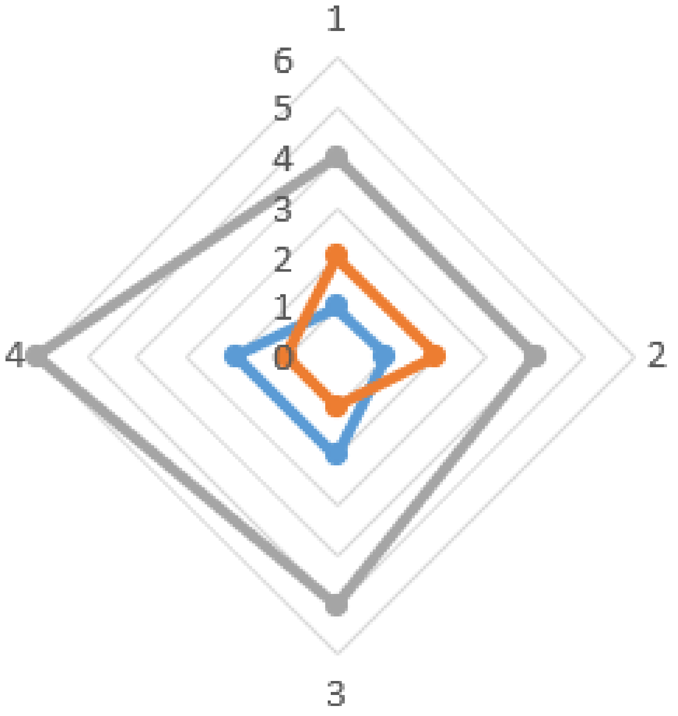
SCATTER



BUBBLE PLOT



SPIDER CHART



Data visualization: tools

Choosing the correct visualization tool is essential for provide the correct information in the most direct and comprehensible way.

For a given case, we can have more than one possible visualization tool.

In that case, choose the best one.

If they are equivalent → choose one

Data visualization: use cases

We will now move to live visualization scenarios.

There are plenty of software or online tools for data visualization.

For simplicity, we will use Excel.

For the sake of completeness, we will also use Matlab.

Data visualization: use cases

Let’s start with easy and immediate examples.

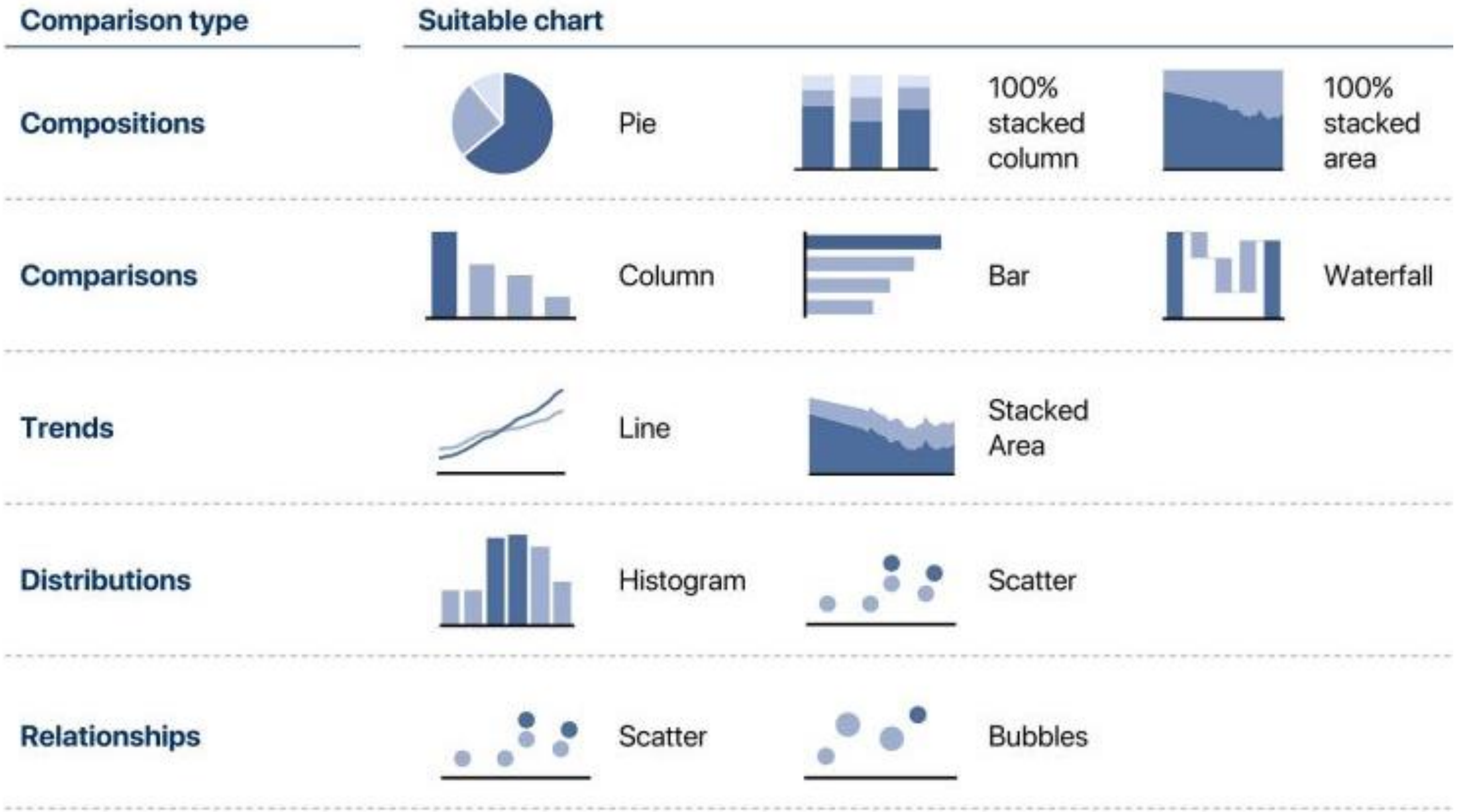
We will visualize data and information of this class.

Excel (file Plots.xslx)

Data visualization: how to choose?

It depends on:

- Objective
- Data structure

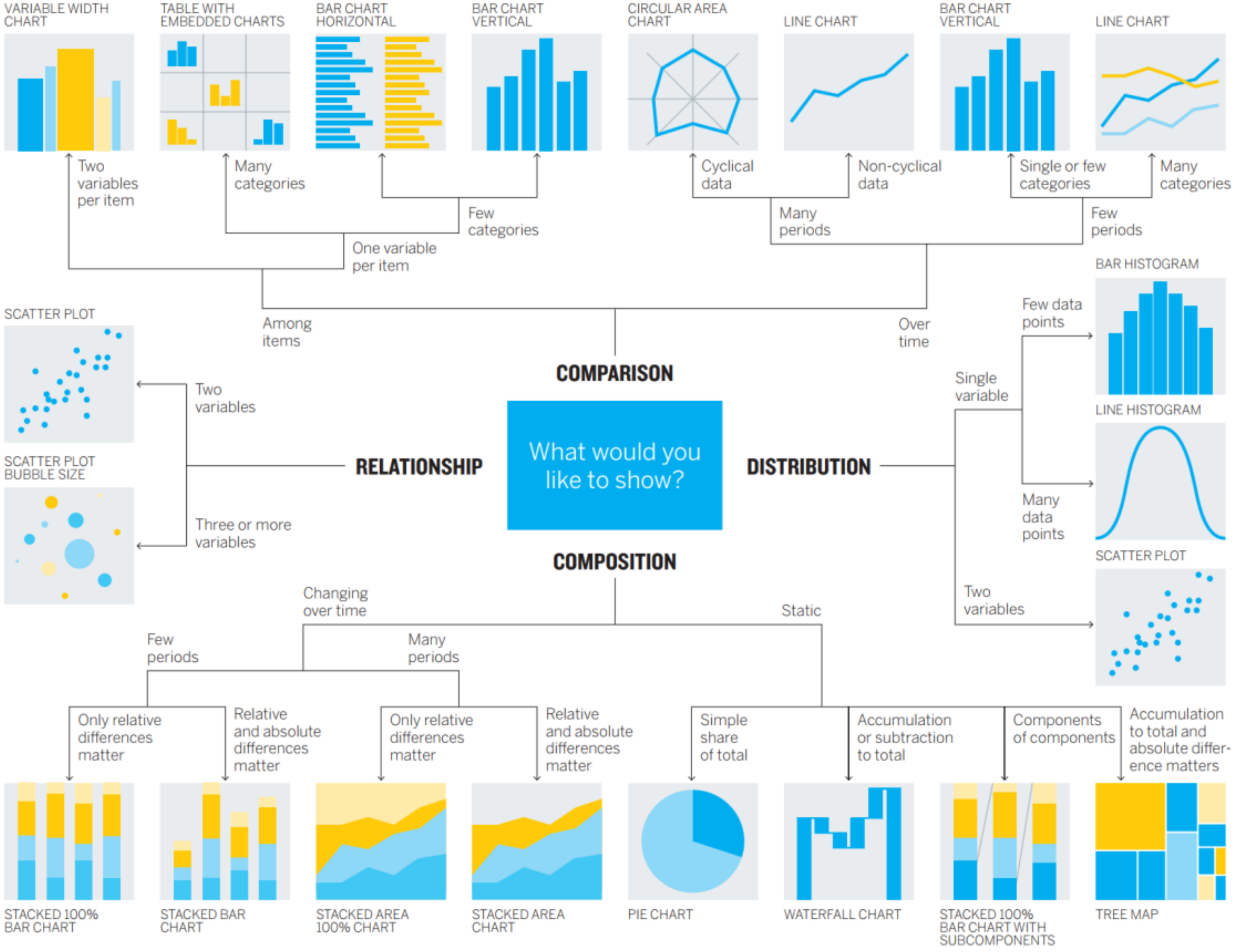


Comparison type	Suitable chart		
Compositions	Pie	100% stacked column	100% stacked area
Comparisons	Column	Bar	Waterfall
Trends	Line	Stacked Area	
Distributions	Histogram	Scatter	
Relationships	Scatter	Bubbles	

Data visualization: how to choose?

It depends on:

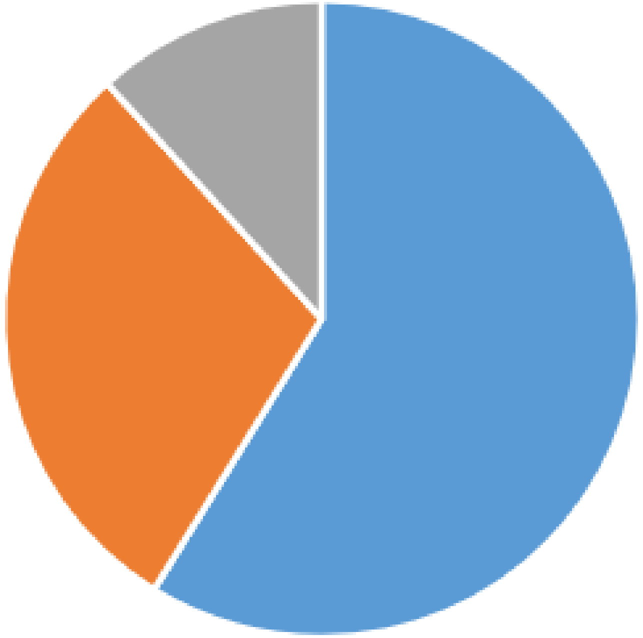
- Objective
- Data structure



SOURCE: ANDREW V. ABELA

Data visualization: pie chart

- Number of variables: more than 1 (e.g., number of males and females in a class)
- Number of items per variable: 1 (e.g., 1,10)
- Objective: describe the composition
- Sometimes equivalent to bar plot, but more immediate!



Data visualization: bar plot

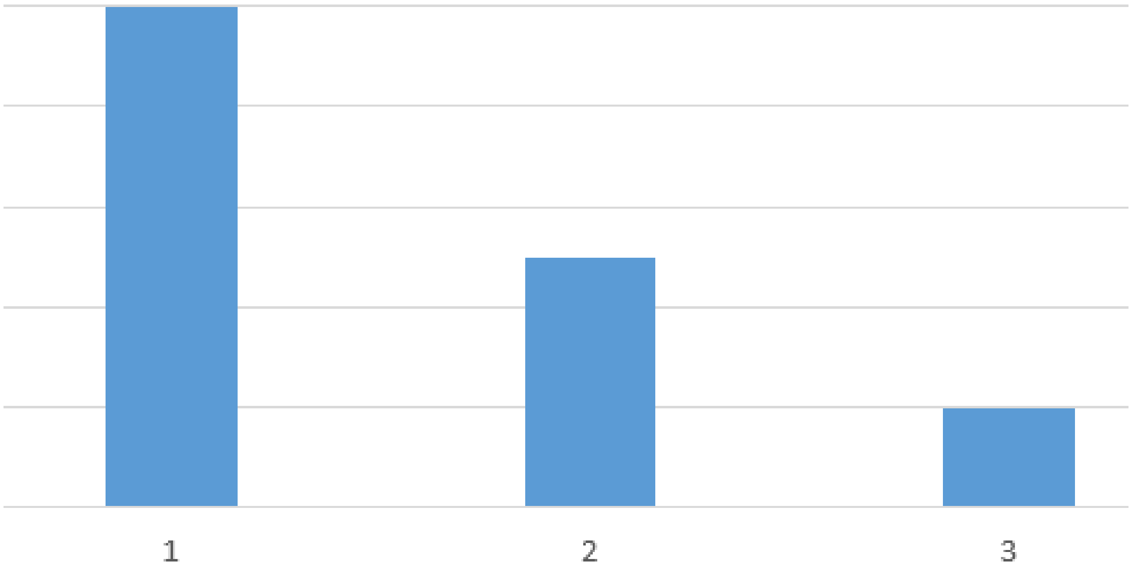
- Number of variables: more than 1 (e.g., number of males and females in a class)

- Number of items per variable: few (e.g., 1-3)

- Objective: describe the composition

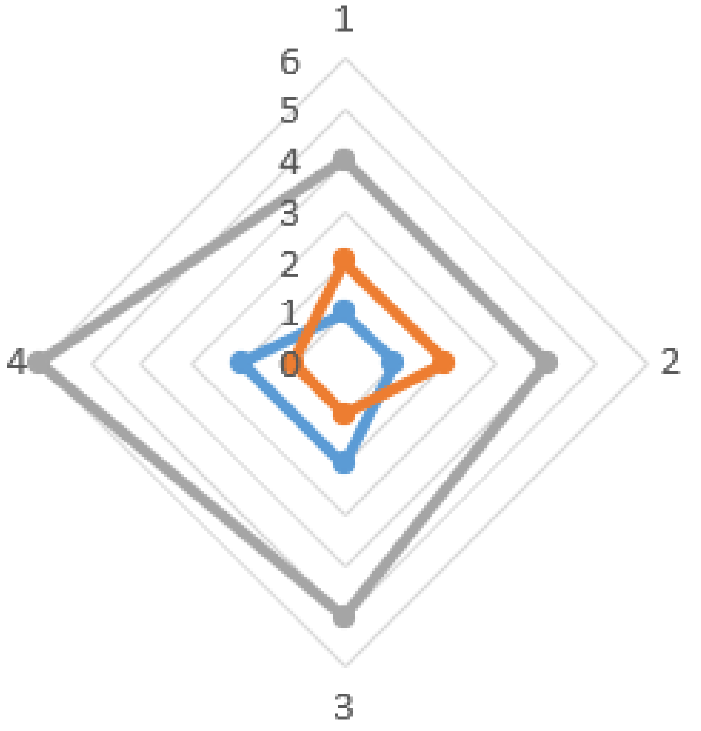
- Sometimes equivalent to pie chart.

- Better than pie chart in cases with different associated variables



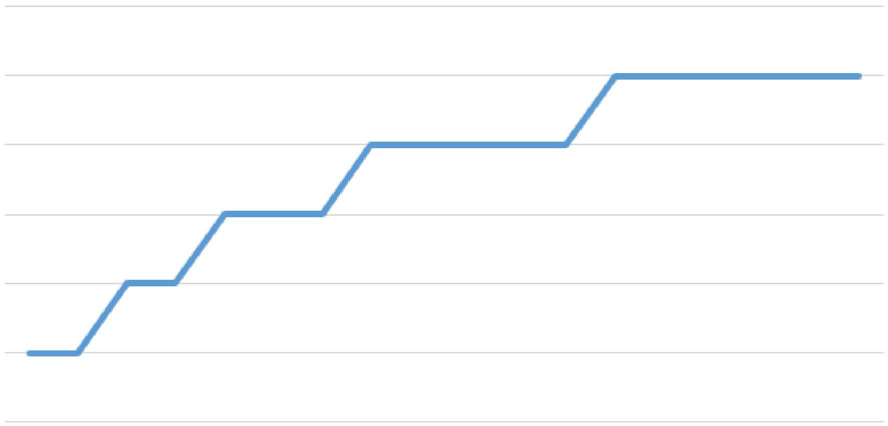
Data visualization: spider plot

- Number of variables: more than 1 (e.g., age, weight, height)
- Number of items per variable: few (e.g., 1-5)
- Objective: compare subjects/items
- Sometimes equivalent to bar chart.
- Better than bar chart because more immediate



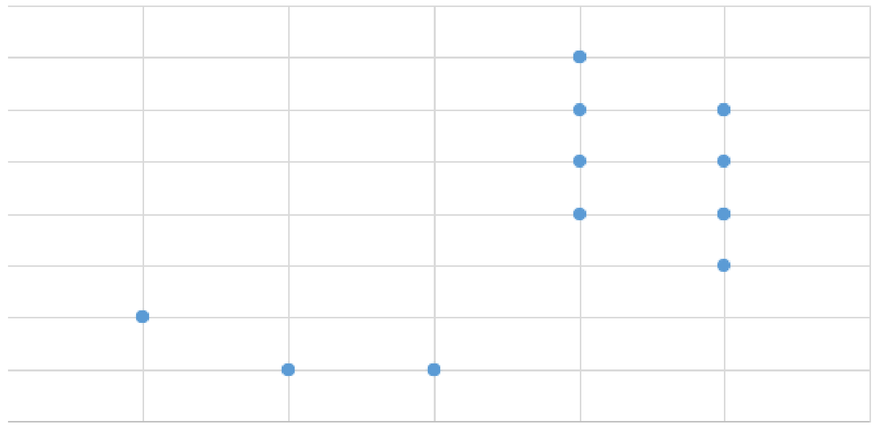
Data visualization: line plot

- Number of variables: few (e.g., subjects)
- Number of items per variable: many
- Objective: describe the evolution over time, compare trends (temperature indoor vs outdoor)
- Useful when visualizing signals or data points and different time stamps
- Can be used to plot two signals on the same temporal scale



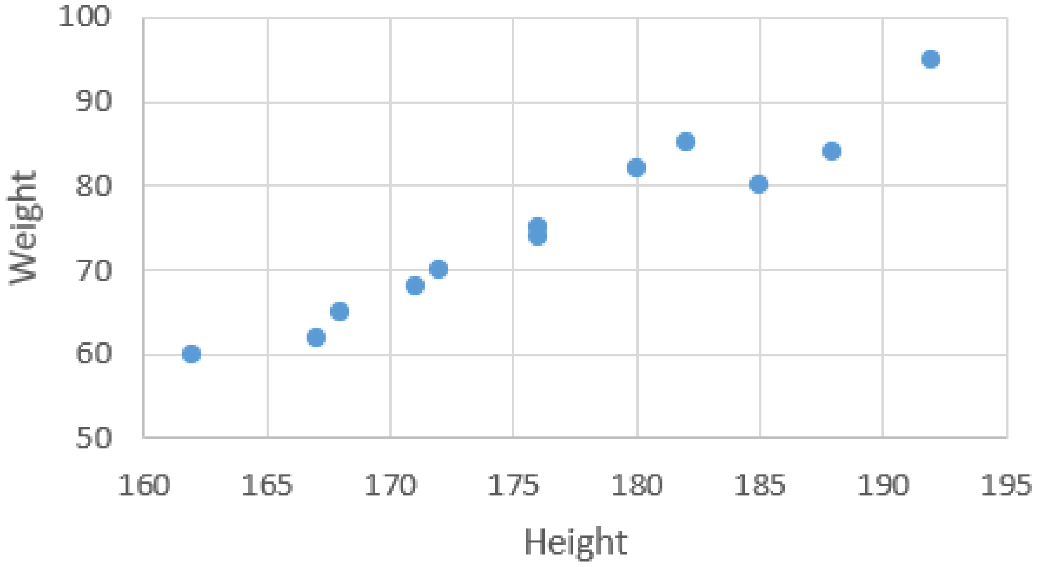
Data visualization: scatter plot

- Number of variables: 1-2 (3 if 3-dimensional)
- Number of items per variable: many
- Objective: describe the distribution of data, check the relationship between variables



Data visualization: scatter plot - correlation

- Number of variables: 2 (3 if 3-dimensional)
- Number of items per variable: many
- Objective: evaluate the relationship between variables

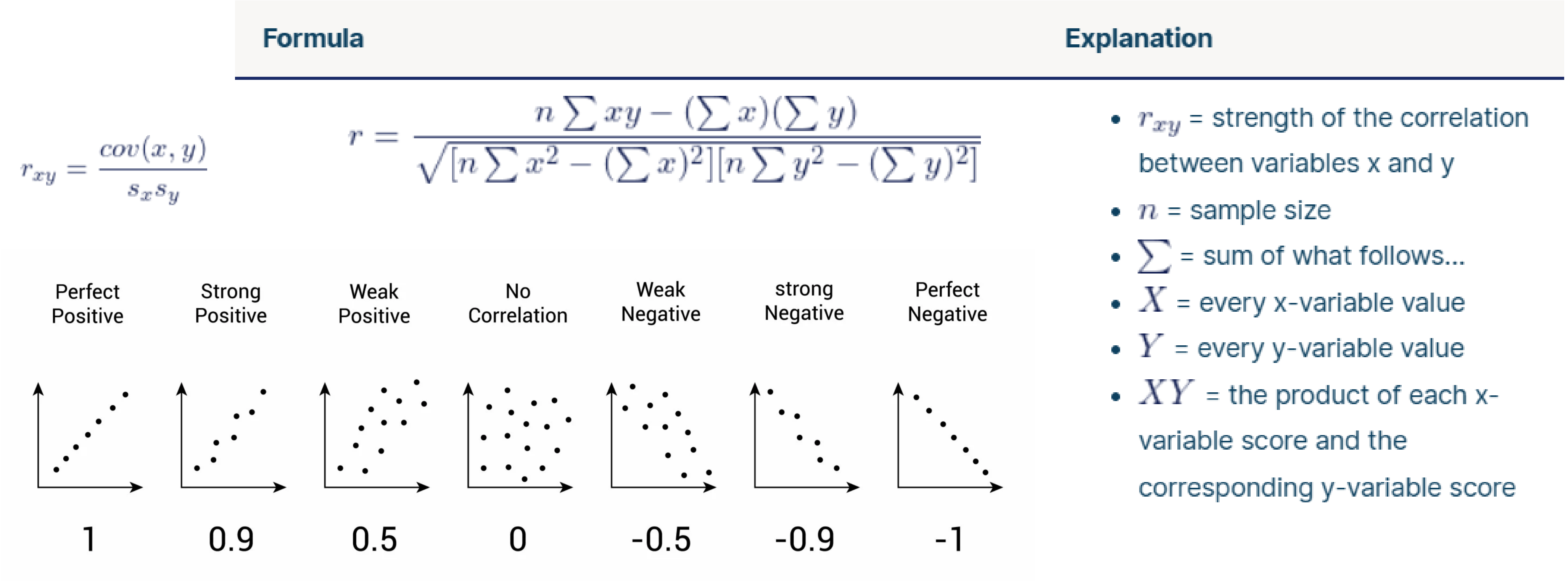


Data visualization: scatter plot - correlation

The strength and direction of the relationship between variables can be evaluated using:

- **Pearson correlation coefficient**: if linear relationship or
- Spearman correlation coefficient: if non-linear relationship

The correlation coefficient will be in the range from -1 to +1

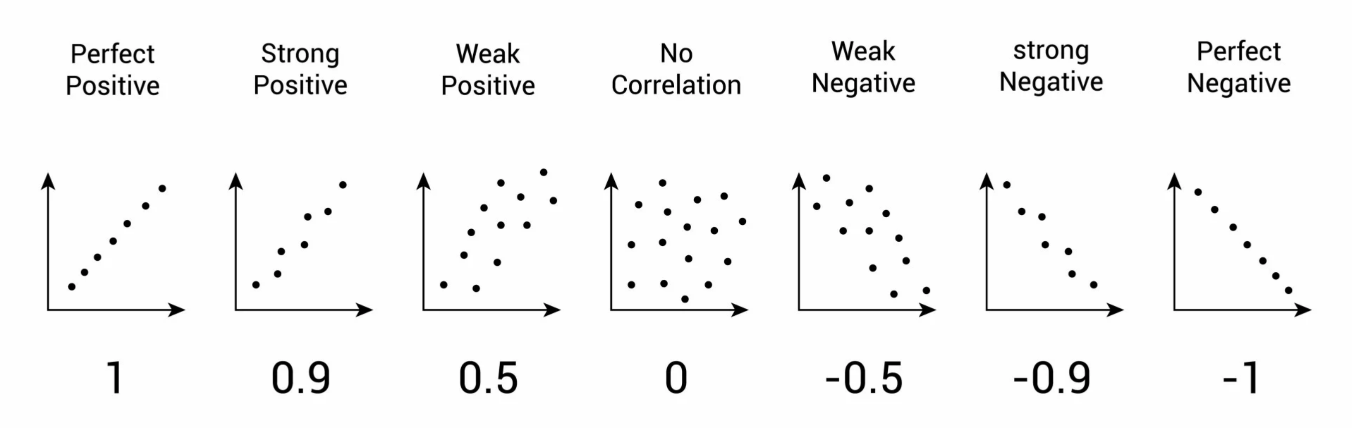


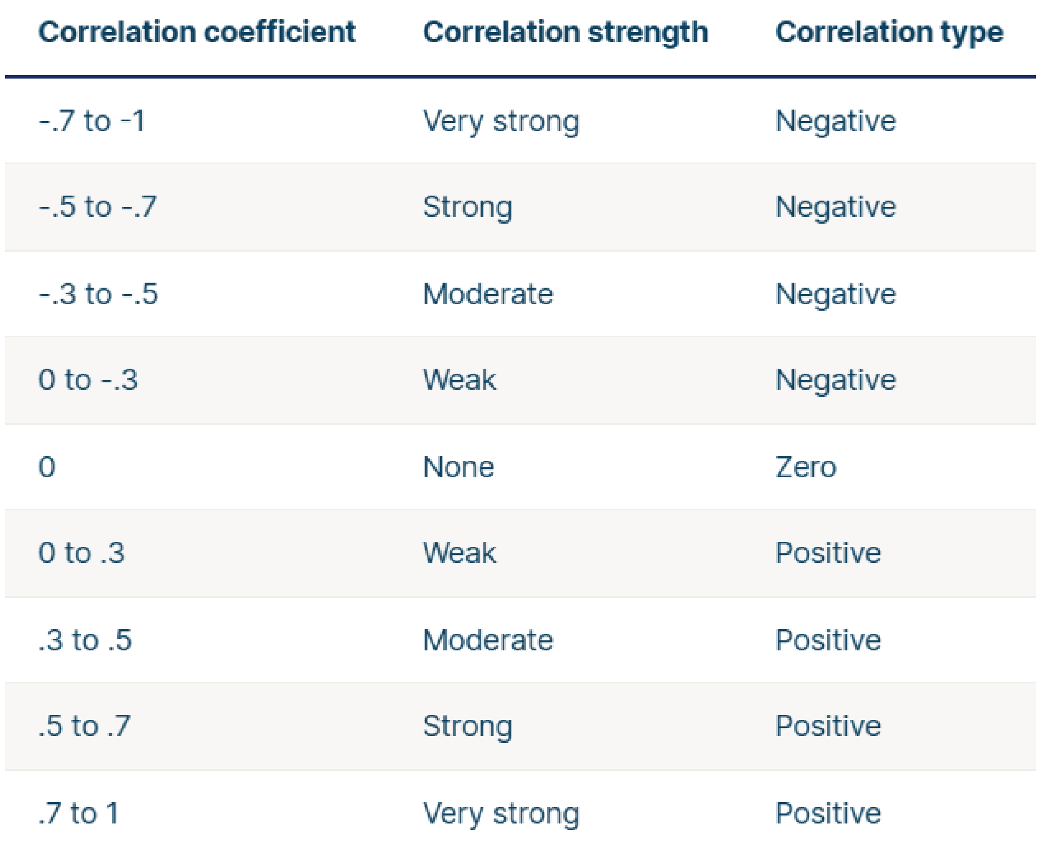
$$r_{xy} = \frac{cov(x,y)}{s_x s_y}$$

Formula	Explanation
$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$	• r_{xy} = strength of the correlation between variables x and y • n = sample size • $\sum$ = sum of what follows... • X = every x-variable value • Y = every y-variable value • XY = the product of each x-variable score and the corresponding y-variable score

Data visualization: scatter plot - correlation

The strength and direction of the relationship between variables can be evaluated using correlation coefficients





Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Data visualization: scatter plot - correlation

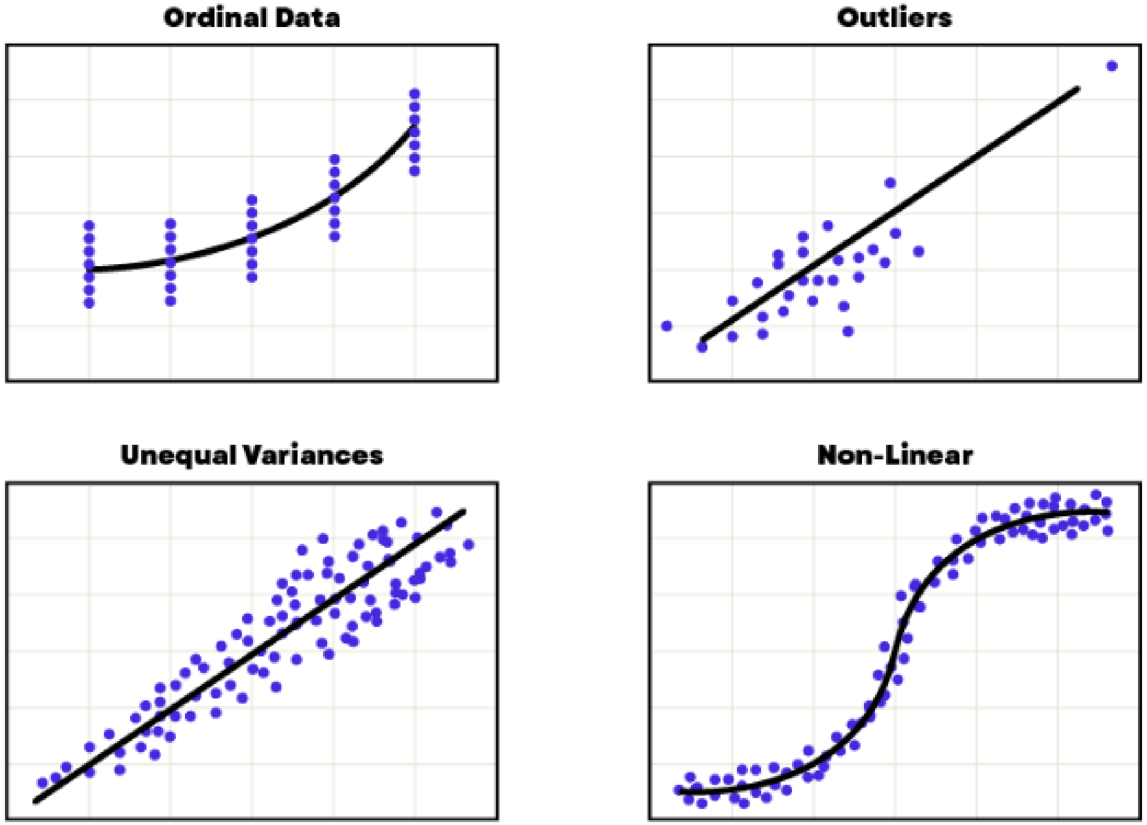
The strength and direction of the relationship between variables can be evaluated using:

- Pearson correlation coefficient: if linear relationship or
- **Spearman correlation coefficient**: if non-linear relationship

The correlation coefficient will be in the range from -1 to +1

There are some cases in which the relationship is not linear. In these cases, the Pearson correlation will fail to catch the relationship.

Spearman correlation evaluates the strength and direction of monotonical relationships

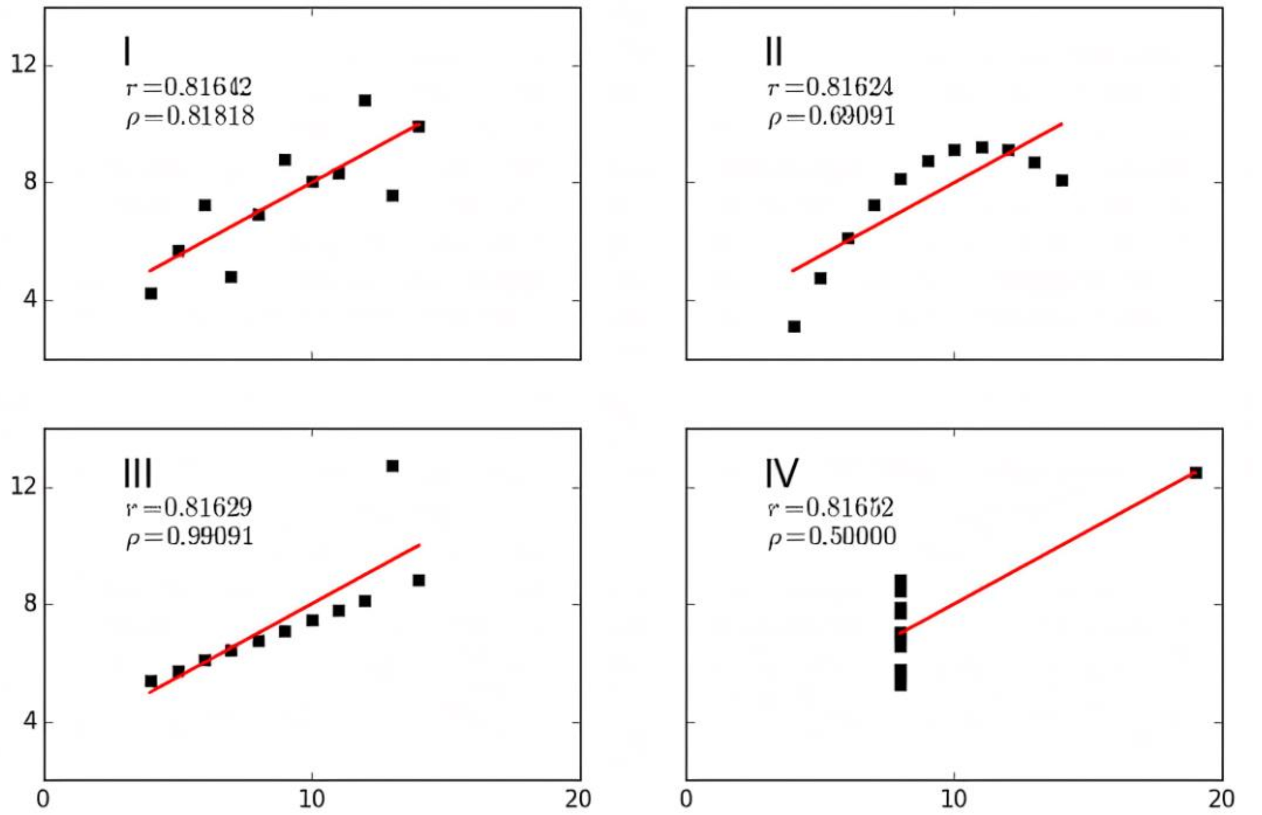


Data visualization: scatter plot - correlation

A method for understanding the relationship and distribution of data can be:

- Compute Pearson correlation coefficient
- Compute Spearman correlation coefficient
- Compare the two measures:
  - If they are similar → Linear relationship
  - If they are very different → Non-linear relationship

- Visualize the relationship to discover more
- Use Spearman correlation



Data visualization: heatmap- correlation

We can use a heatmap to visualize 3-dimensional data.

For example, we can visualize the values in a matrix, highlighting different values with different colors (e.g., red for high values, blue for low values).

This is effective in different applications.

One of this is the visualization of correlation between variables.

In a single plot, we want to easily identify relationships between all variables!

Sample	Entropy	Mean	std	range	energy
1	-15	0.5	0.2	0.8	172
2	-150	0.6	0.15	1.2	165
3	-75	0.7	0.5	0.7	180
4	-28	0.2	0.4	0.75	168
5	-50	0.1	0.6	0.90	185
6	-32	0.25	0.8	0.78	170
7	-55	0.27	0.5	0.95	185
8	-40	0.3	0.4	0.72	160
9	-48	0.2	0.6	0.85	175
10	-38	0.8	0.8	0.80	170
…					
…					
N	-40	1	0.6	0.72	160

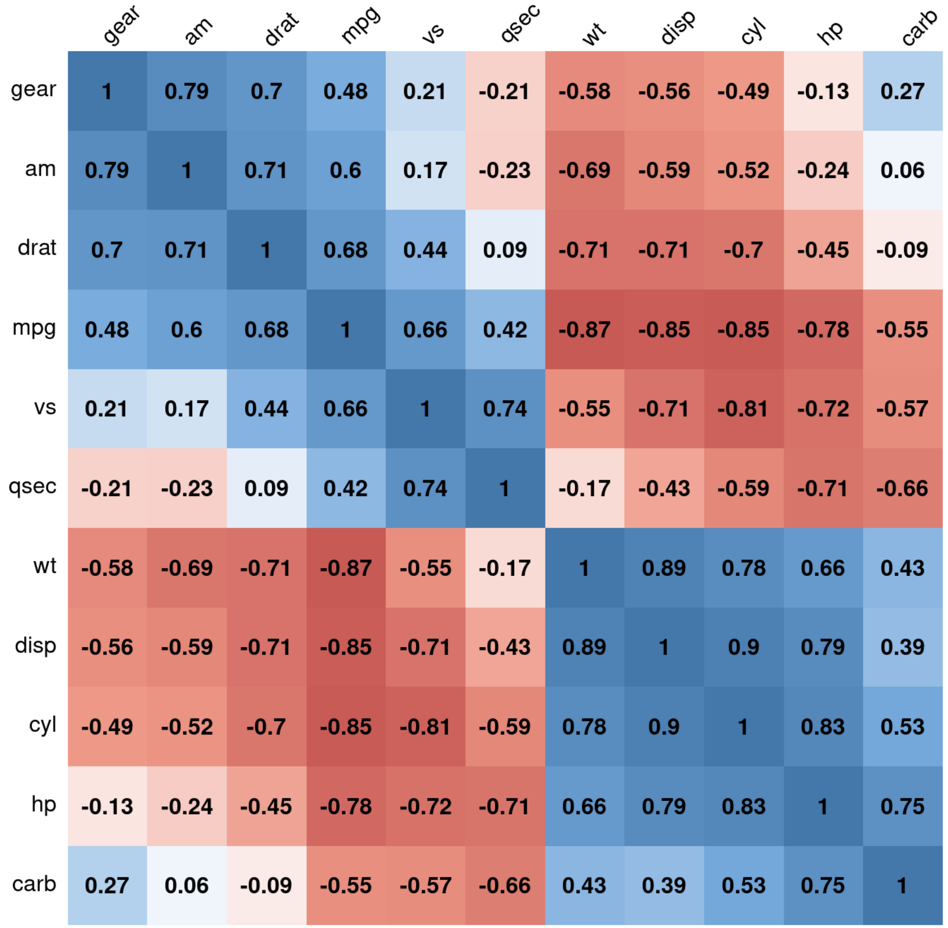
Data visualization: heatmap- correlation

We can use a heatmap to visualize 3-dimensional data.

For example, we can visualize the values in a matrix, highlighting different values with different colors (e.g., red for high values, blue for low values).

This is effective in different applications.

One of this is the visualization of correlation between variables.



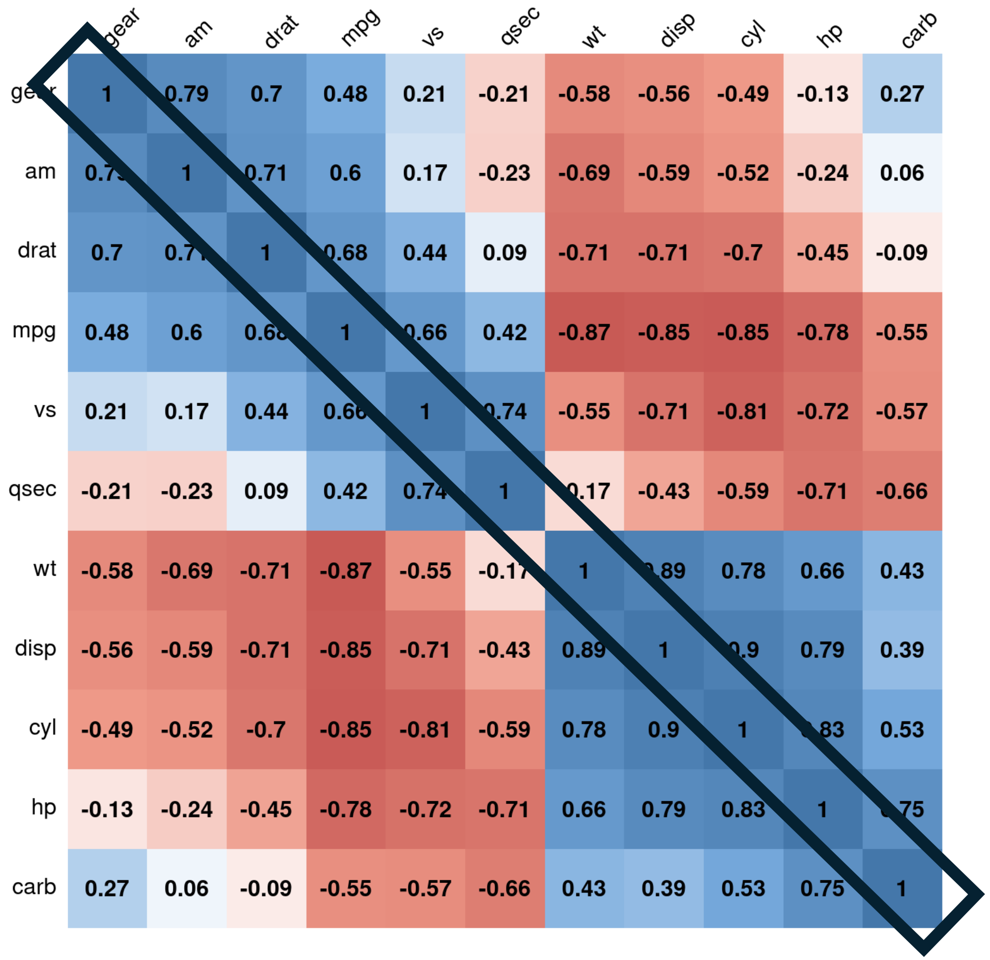
	gear	am	drat	mpg	vs	qsec	wt	disp	cyl	hp	carb
gear	1	0.79	0.7	0.48	0.21	-0.21	-0.58	-0.56	-0.49	-0.13	0.27
am	0.79	1	0.71	0.6	0.17	-0.23	-0.69	-0.59	-0.52	-0.24	0.06
drat	0.7	0.71	1	0.68	0.44	0.09	-0.71	-0.71	-0.7	-0.45	-0.09
mpg	0.48	0.6	0.68	1	0.66	0.42	-0.87	-0.85	-0.85	-0.78	-0.55
vs	0.21	0.17	0.44	0.66	1	0.74	-0.55	-0.71	-0.81	-0.72	-0.57
qsec	-0.21	-0.23	0.09	0.42	0.74	1	-0.17	-0.43	-0.59	-0.71	-0.66
wt	-0.58	-0.69	-0.71	-0.87	-0.55	-0.17	1	0.89	0.78	0.66	0.43
disp	-0.56	-0.59	-0.71	-0.85	-0.71	-0.43	0.89	1	0.9	0.79	0.39
cyl	-0.49	-0.52	-0.7	-0.85	-0.81	-0.59	0.78	0.9	1	0.83	0.53
hp	-0.13	-0.24	-0.45	-0.78	-0.72	-0.71	0.66	0.79	0.83	1	0.75
carb	0.27	0.06	-0.09	-0.55	-0.57	-0.66	0.43	0.39	0.53	0.75	1

Data visualization: heatmap- correlation

Heatmap for correlation

The diagonal values are all ones!

This is because the correlation between a variable and itself is 1, of course.



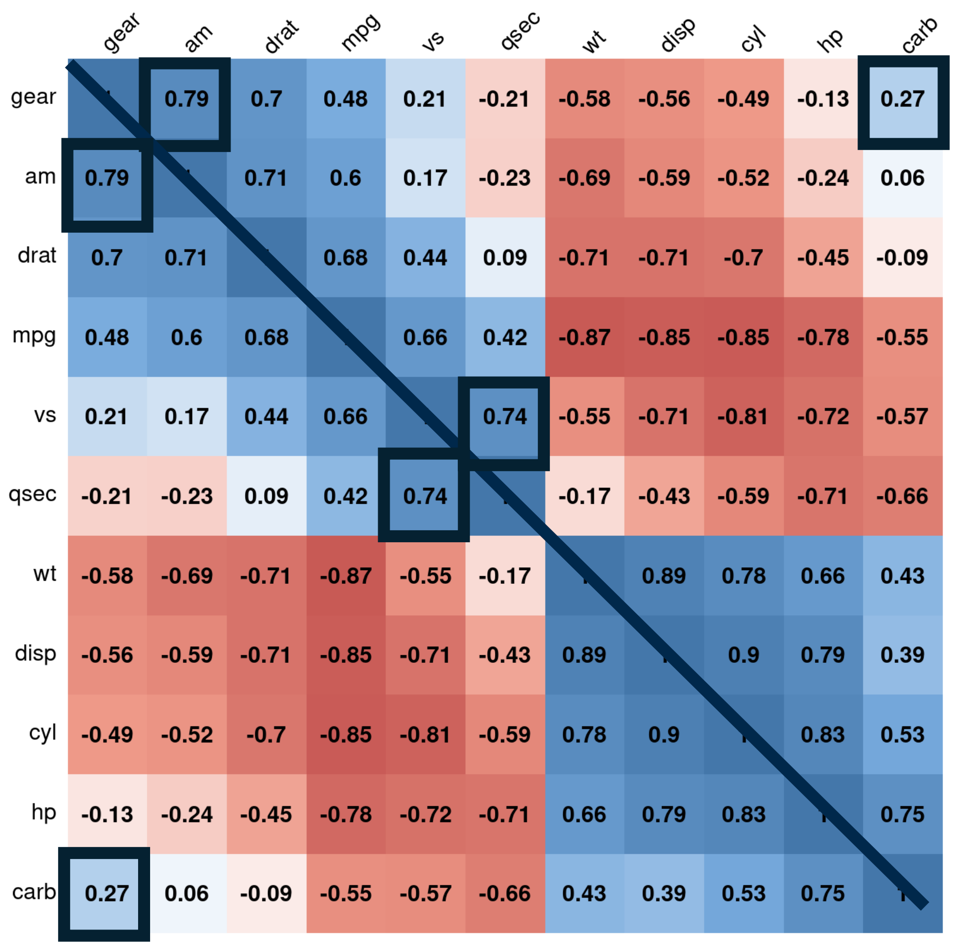
	gear	am	drat	mpg	vs	qsec	wt	disp	cyl	hp	carb
gear	1	0.79	0.7	0.48	0.21	-0.21	-0.58	-0.56	-0.49	-0.13	0.27
am	0.79	1	0.71	0.6	0.17	-0.23	-0.69	-0.59	-0.52	-0.24	0.06
drat	0.7	0.71	1	0.68	0.44	0.09	-0.71	-0.71	-0.7	-0.45	-0.09
mpg	0.48	0.6	0.68	1	0.66	0.42	-0.87	-0.85	-0.85	-0.78	-0.55
vs	0.21	0.17	0.44	0.66	1	0.74	-0.55	-0.71	-0.81	-0.72	-0.57
qsec	-0.21	-0.23	0.09	0.42	0.74	1	-0.17	-0.43	-0.59	-0.71	-0.66
wt	-0.58	-0.69	-0.71	-0.87	-0.55	-0.17	1	0.89	0.78	0.66	0.43
disp	-0.56	-0.59	-0.71	-0.85	-0.71	-0.43	0.89	1	0.9	0.79	0.39
cyl	-0.49	-0.52	-0.7	-0.85	-0.81	-0.59	0.78	0.9	1	0.83	0.53
hp	-0.13	-0.24	-0.45	-0.78	-0.72	-0.71	0.66	0.79	0.83	1	0.75
carb	0.27	0.06	-0.09	-0.55	-0.57	-0.66	0.43	0.39	0.53	0.75	1

Data visualization: heatmap- correlation

Heatmap for correlation

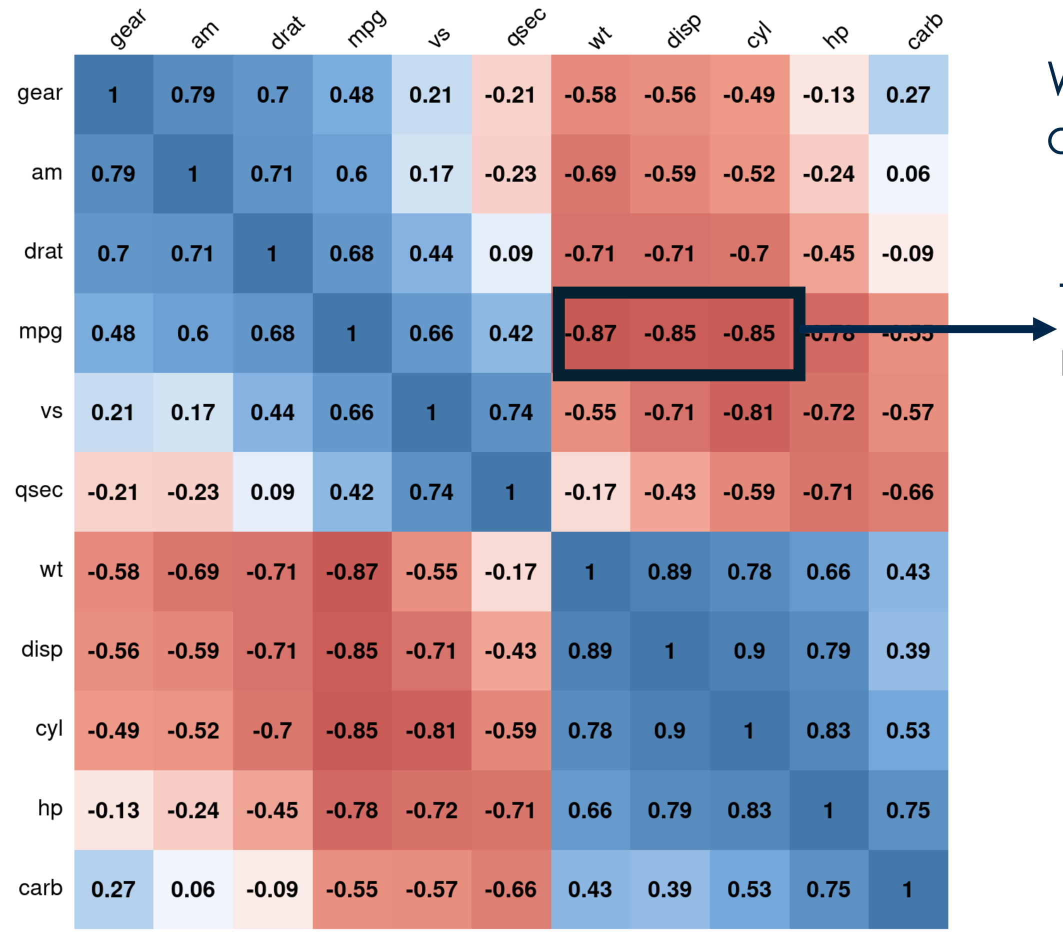
The diagonal values are all ones!

The matrix is symmetric → The triangle above and under the diagonal are specular



	gear	am	drat	mpg	vs	qsec	wt	disp	cyl	hp	carb
gear	1	0.79	0.7	0.48	0.21	-0.21	-0.58	-0.56	-0.49	-0.13	0.27
am	0.79	1	0.71	0.6	0.17	-0.23	-0.69	-0.59	-0.52	-0.24	0.06
drat	0.7	0.71	1	0.68	0.44	0.09	-0.71	-0.71	-0.7	-0.45	-0.09
mpg	0.48	0.6	0.68	1	0.66	0.42	-0.87	-0.85	-0.85	-0.78	-0.55
vs	0.21	0.17	0.44	0.66	1	0.74	-0.55	-0.71	-0.81	-0.72	-0.57
qsec	-0.21	-0.23	0.09	0.42	0.74	1	-0.17	-0.43	-0.59	-0.71	-0.66
wt	-0.58	-0.69	-0.71	-0.87	-0.55	-0.17	1	0.89	0.78	0.66	0.43
disp	-0.56	-0.59	-0.71	-0.85	-0.71	-0.43	0.89	1	0.9	0.79	0.39
cyl	-0.49	-0.52	-0.7	-0.85	-0.81	-0.59	0.78	0.9	1	0.83	0.53
hp	-0.13	-0.24	-0.45	-0.78	-0.72	-0.71	0.66	0.79	0.83	1	0.75
carb	0.27	0.06	-0.09	-0.55	-0.57	-0.66	0.43	0.39	0.53	0.75	1

Data visualization: heatmap- correlation



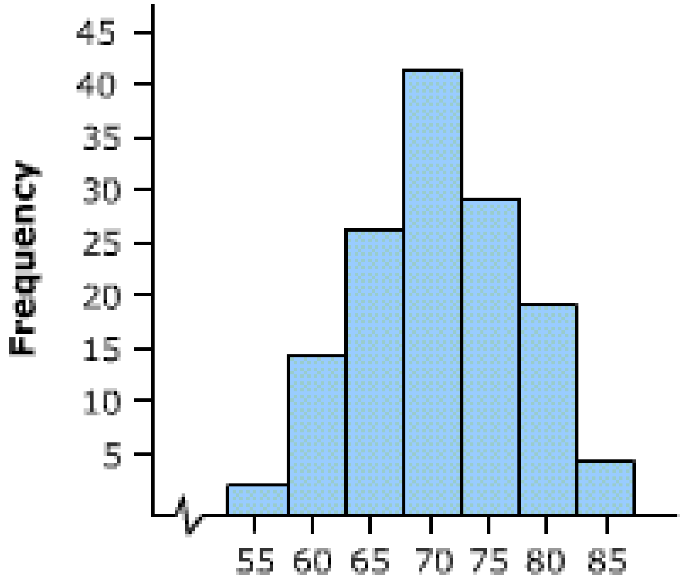
	gear	am	drat	mpg	vs	qsec	wt	disp	cyl	hp	carb
gear	1	0.79	0.7	0.48	0.21	-0.21	-0.58	-0.56	-0.49	-0.13	0.27
am	0.79	1	0.71	0.6	0.17	-0.23	-0.69	-0.59	-0.52	-0.24	0.06
drat	0.7	0.71	1	0.68	0.44	0.09	-0.71	-0.71	-0.7	-0.45	-0.09
mpg	0.48	0.6	0.68	1	0.66	0.42	-0.87	-0.85	-0.85	-0.78	-0.55
vs	0.21	0.17	0.44	0.66	1	0.74	-0.55	-0.71	-0.81	-0.72	-0.57
qsec	-0.21	-0.23	0.09	0.42	0.74	1	-0.17	-0.43	-0.59	-0.71	-0.66
wt	-0.58	-0.69	-0.71	-0.87	-0.55	-0.17	1	0.89	0.78	0.66	0.43
disp	-0.56	-0.59	-0.71	-0.85	-0.71	-0.43	0.89	1	0.9	0.79	0.39
cyl	-0.49	-0.52	-0.7	-0.85	-0.81	-0.59	0.78	0.9	1	0.83	0.53
hp	-0.13	-0.24	-0.45	-0.78	-0.72	-0.71	0.66	0.79	0.83	1	0.75
carb	0.27	0.06	-0.09	-0.55	-0.57	-0.66	0.43	0.39	0.53	0.75	1

We are interested in dark red (strong negative relationship) and dark blue (strong positive relationship) coordinates.

These values indicate that the variable **mpg** is strongly negatively correlated with the variables **wt, disp, and cyl**

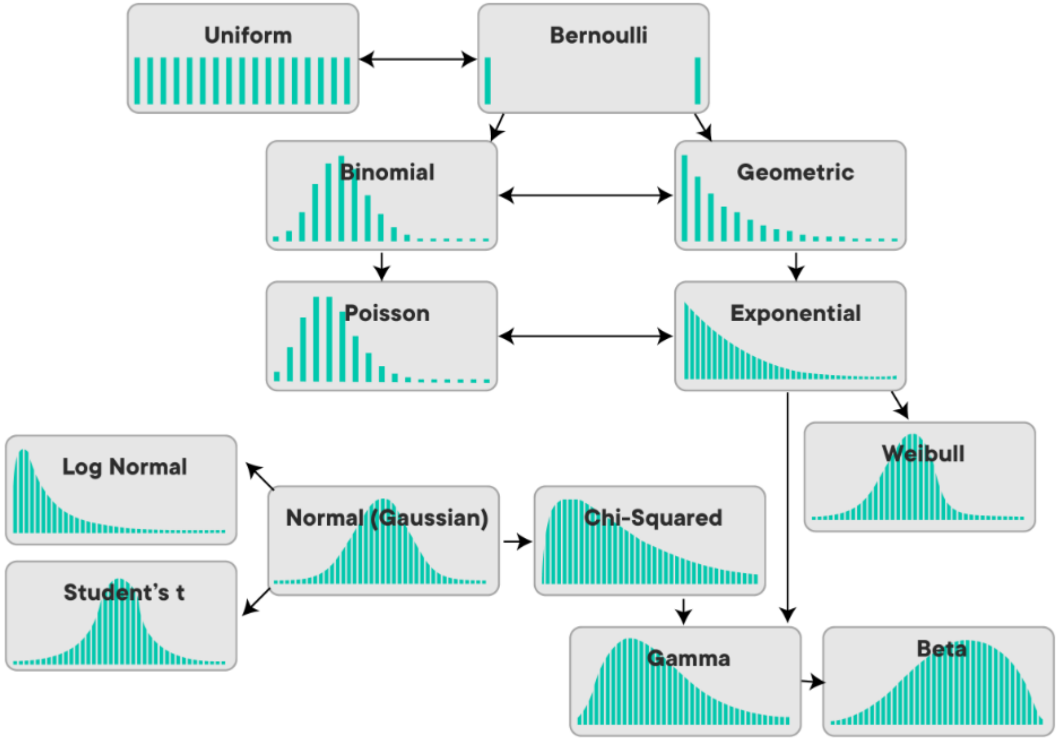
Data visualization: histogram

- Number of variables: 1 (or a few if for comparison)
- Number of items per variable: many
- Objective: evaluate the distribution of variables



Data visualization: histogram

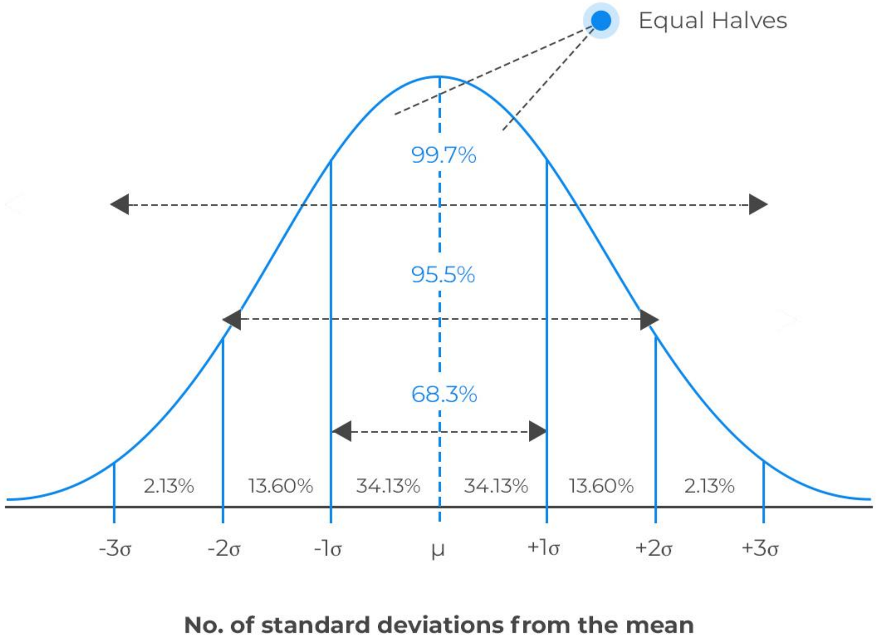
- We can have a wide range of distributions
- We will focus on the difference between normal (gaussian) and non-normal distributions
- We will also see the variants of normal distributions



Data visualization: normal distribution

Gaussian or Normal: continuous probability distribution that is symmetric around its mean (bell-shaped).

Normal distribution

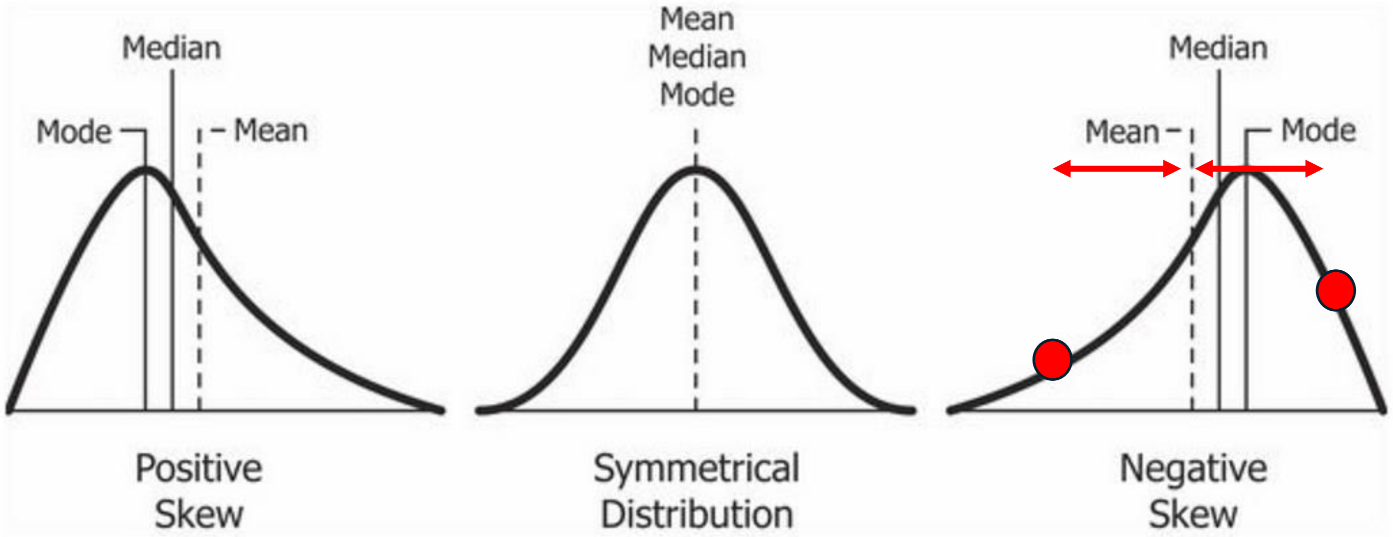


$$f\,(x|\mu,\sigma) = \frac{1}{\sigma \cdot 2\pi} \cdot \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{mean} = \frac{\sum_{\mathrm{i}=1}^{N} x_{\mathrm{i}}}{N} \qquad \text{std} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2}$$

Data visualization: normal distribution

Always consider skewness



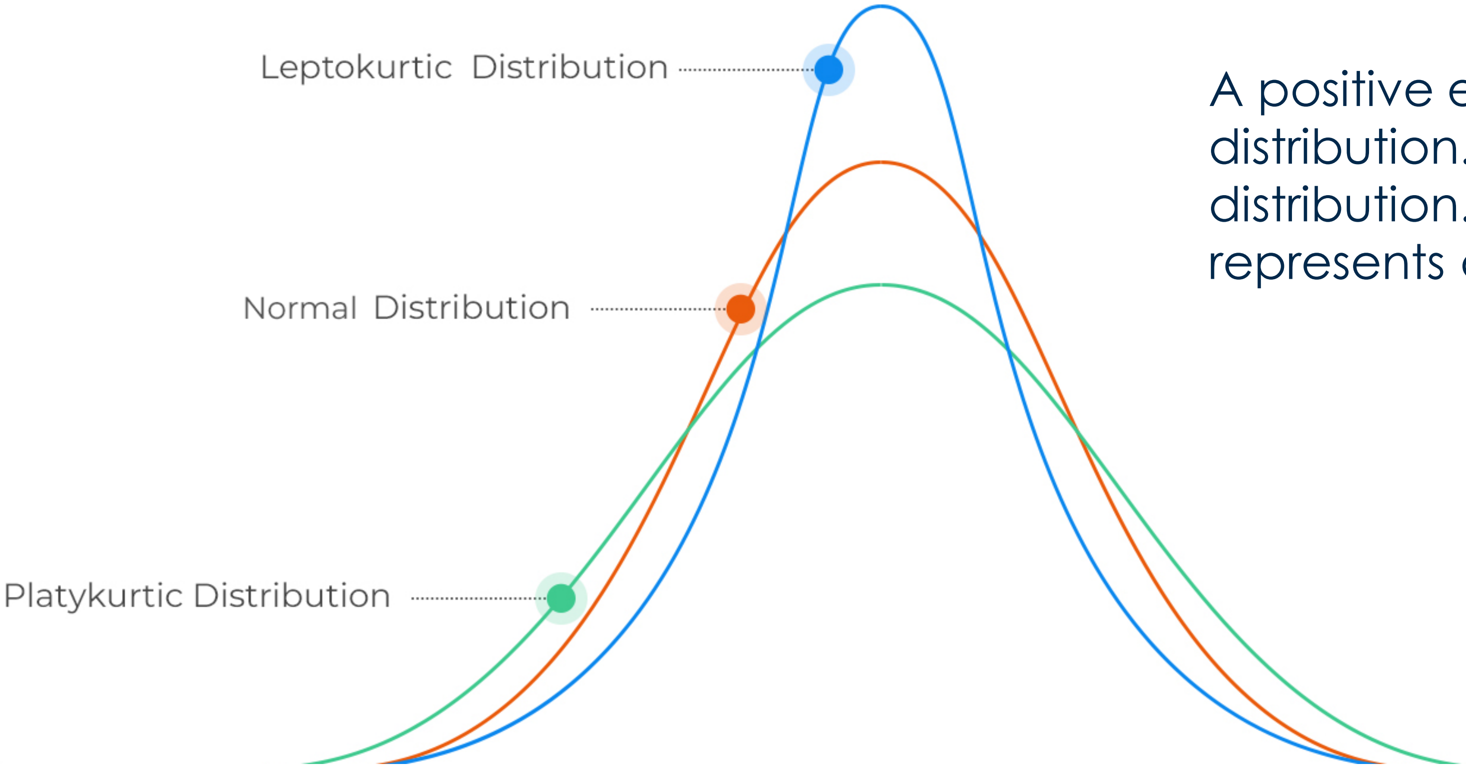
$$S_k = \frac{1}{n} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$$

A positive value indicates positive skewness. A ‘zero’ value indicates the data is not skewed. Lastly, a negative value indicates negative skewness or rather a negatively skewed distribution.

Data visualization: normal distribution

And kurtosis

$$S_{kr} = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^4}{s^4}$$



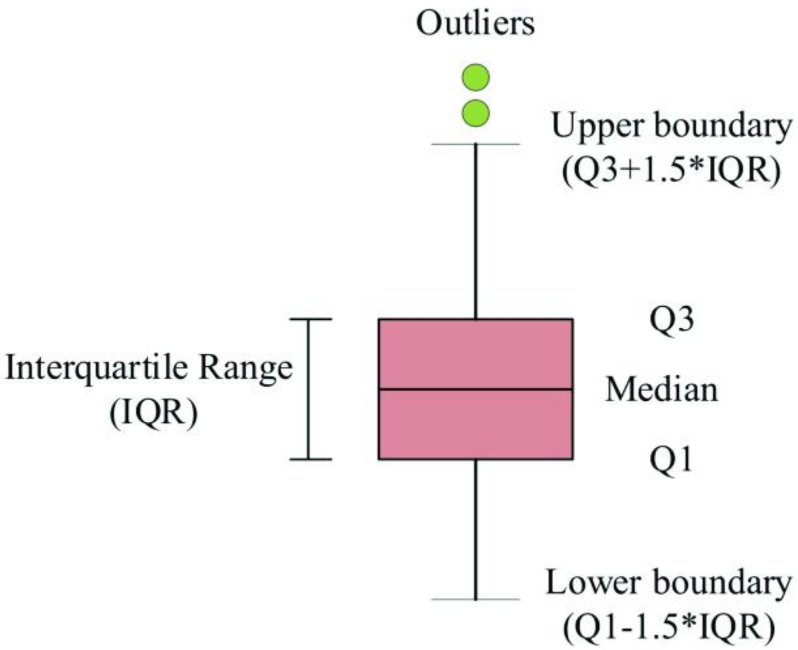
A positive excess kurtosis indicates a leptokurtic distribution. A zero value indicates a mesokurtic distribution. Lastly, a negative excess kurtosis represents a platykurtic distribution.

Data visualization: boxplot

- Number of variables: 1 (or a few if for comparison)
- Number of items per variable: many
- Objective: evaluate the distribution of variables, compare distributions

Data visualization: boxplot

Boxplots provide a comprehensive quantitative analysis of the distribution of data.
In a single plot, we can visualize: the average value (median actually), the variability (inter-quartile range), minimum, maximum, outliers



First Quartile (Q1): value below which 25% of the data fall. Q1 is also known as the 25th percentile.

Third Quartile (Q3): value below which 75% of the data fall. Q3 is also known as the 75th percentile.

Second Quartile: value below which 50% of the data fall. Q2 is also known as the 50th percentile or median.

Data visualization: suggestions

To recap:

Data visualisation provides essential information and insigths into the dataset.

This is helpful for showing others the structure of the dataset.

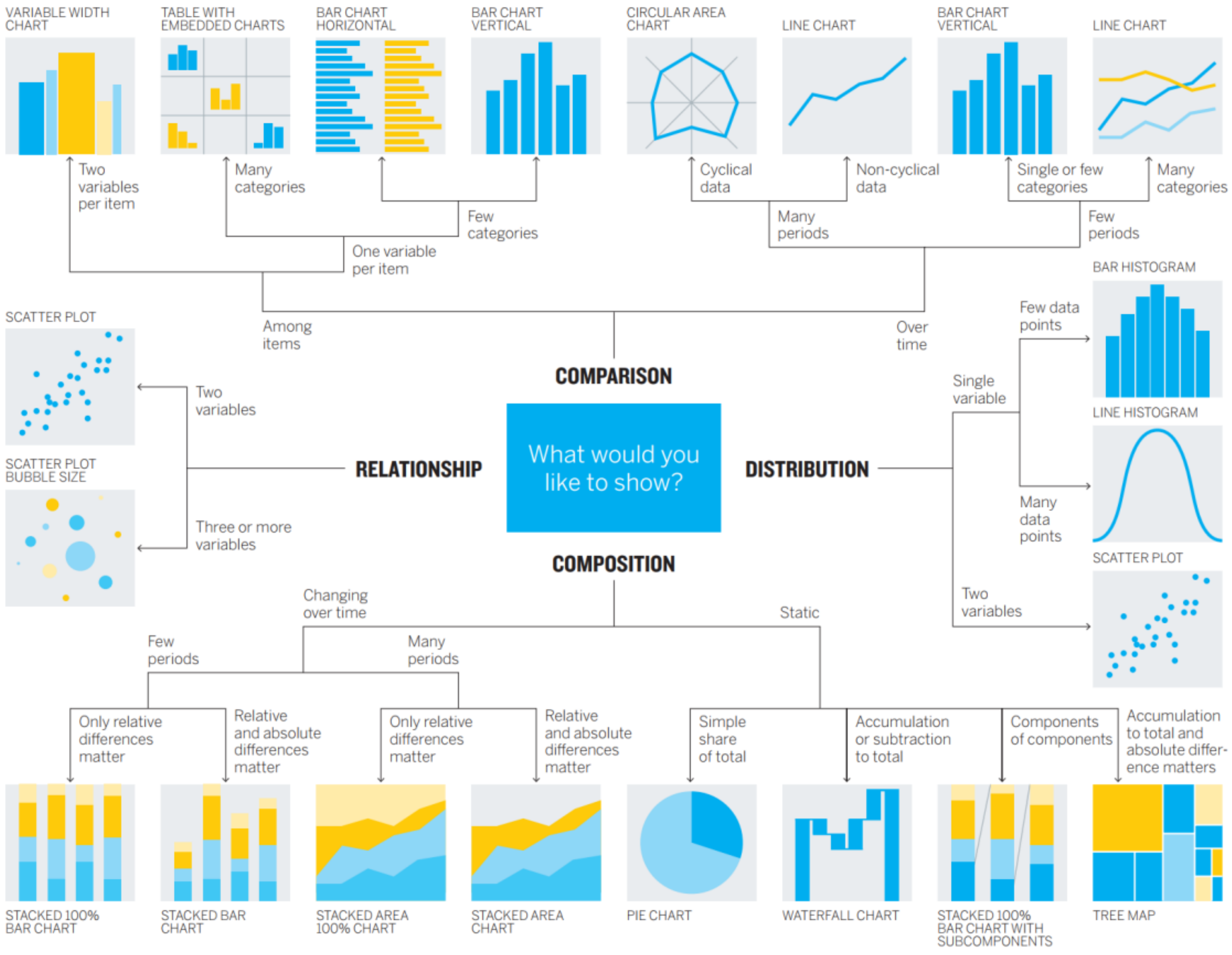
But first of all, it is useful for you. It allows for comprehensively understanding the data and plan how to process.

Data visualization: suggestions

To recap:

When providing visual presentations, the questions are:

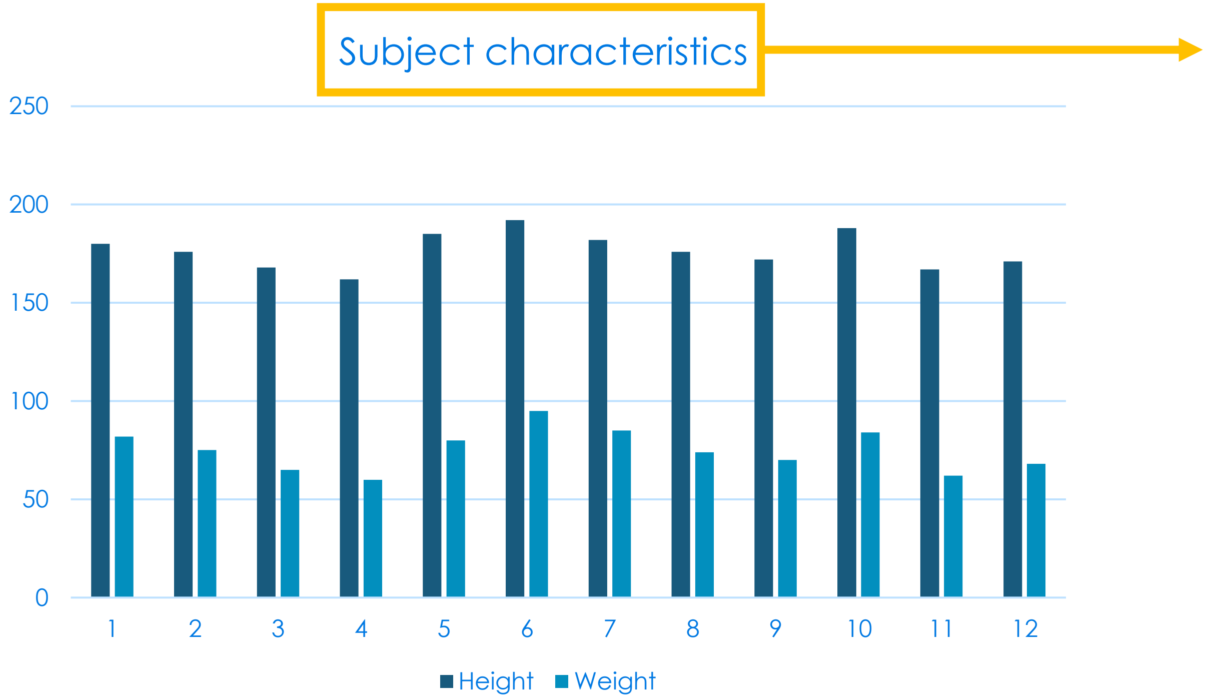
What I want to show? Based on this, select the appropriate visualization method



SOURCE: ANDREW V. ABELA

Data visualization: suggestions

In addition to the selection of the correct(s) method(s), some aspects should be considered.



The **title** summarise the content of the figure

Figure 1. Subject characteristics

Data visualization: suggestions

In addition to the selection of the correct(s) method(s), some aspects should be considered.

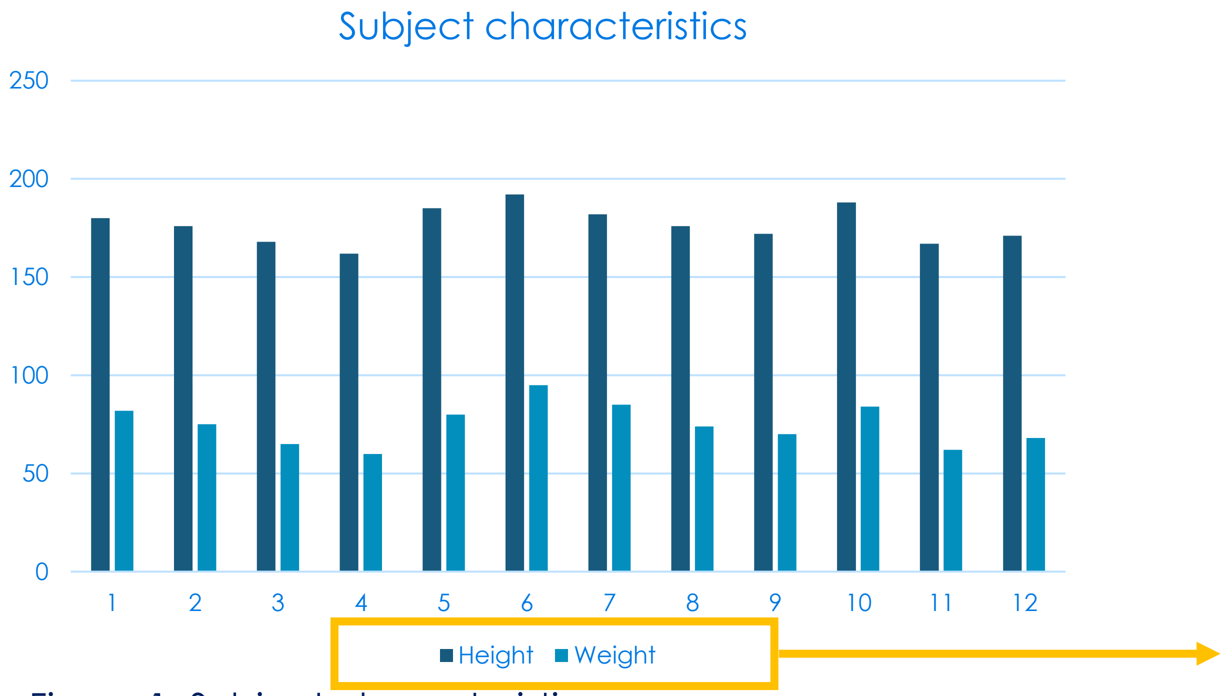


Figure 1. Subject characteristics

The **legend** describes the represented elements/variables

Data visualization: suggestions

In addition to the selection of the correct(s) method(s), some aspects should be considered.

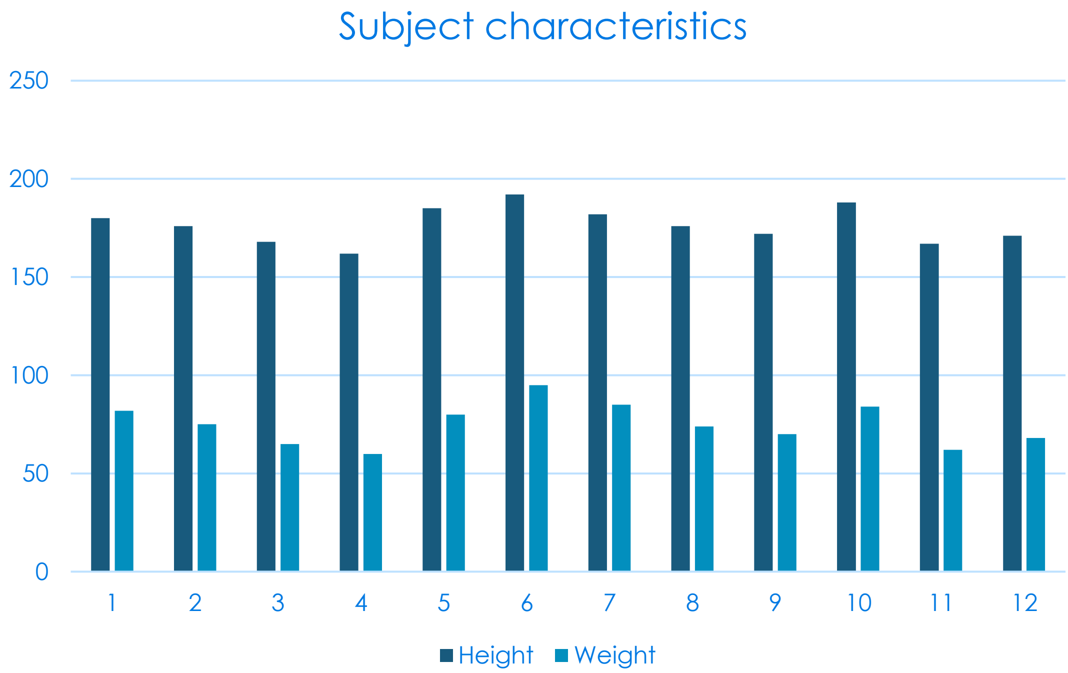


Figure 1. Subject characteristics

The **caption** describes the figure

Data visualization: suggestions

In addition to the selection of the correct(s) method(s), some aspects should be considered.

The **axes** allow for quantitative analysis

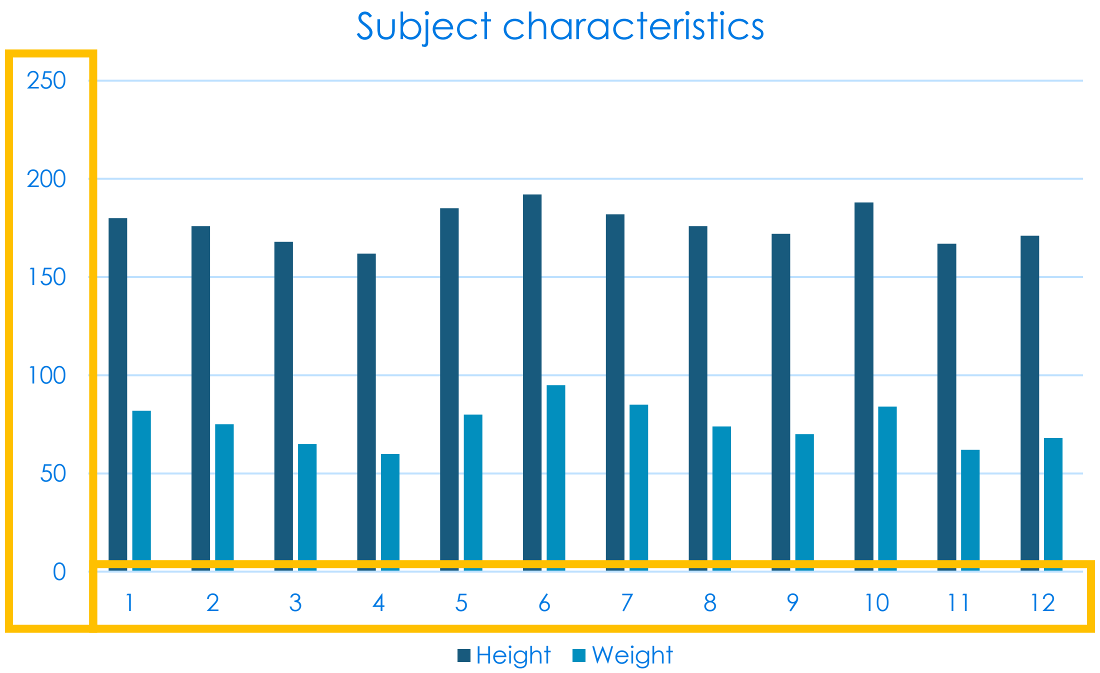


Figure 1. Subject characteristics

Data visualization: suggestions

In addition to the selection of the correct(s) method(s), some aspects should be considered.

The **axes title** indicate what each axis is showing

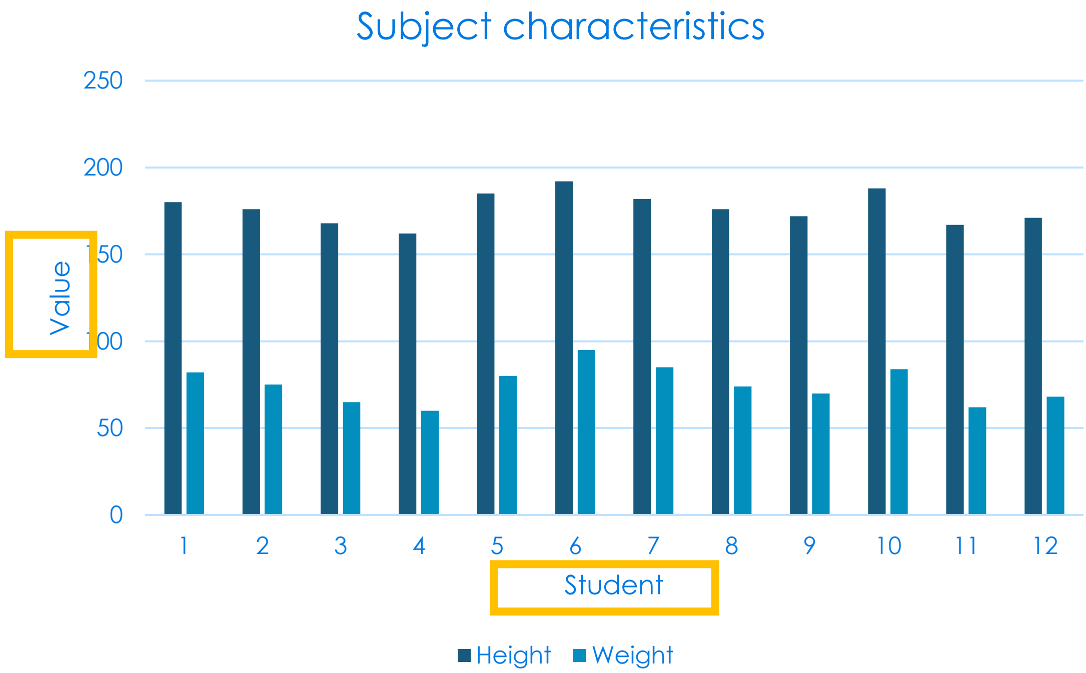
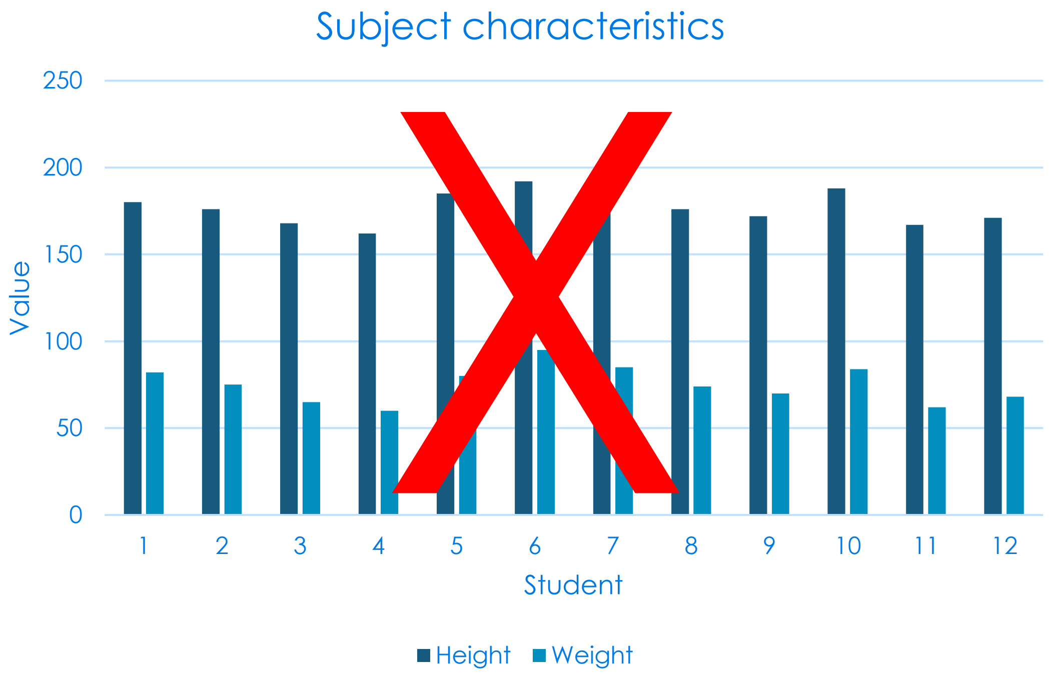
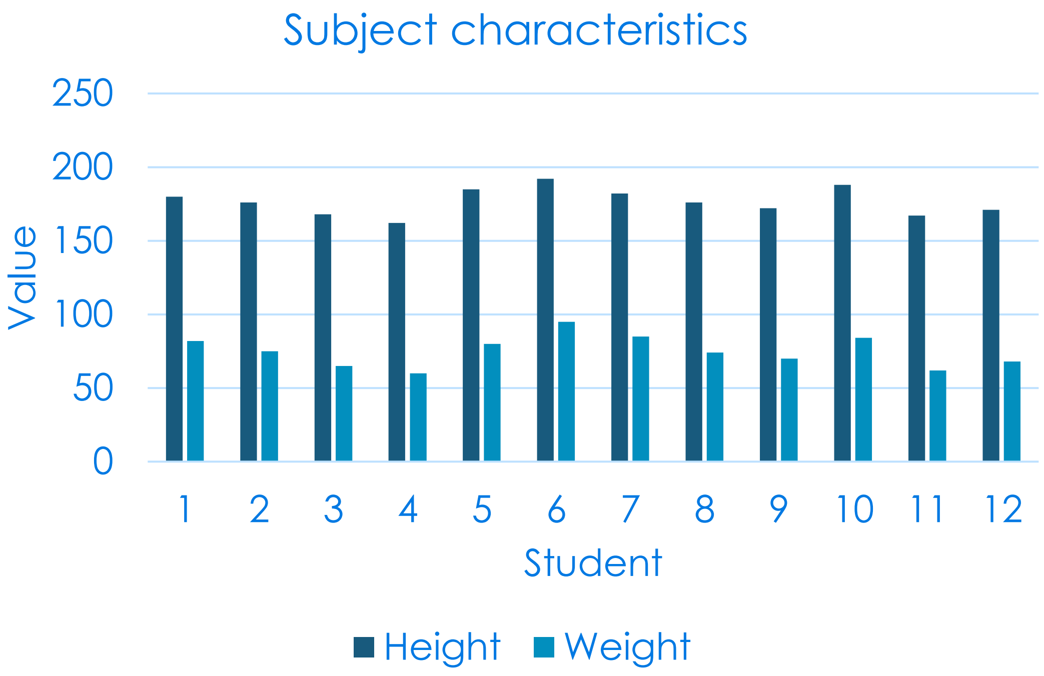


Figure 1. Subject characteristics

Data visualization: suggestions – font size

In addition to the selection of the correct(s) method(s), some aspects should be considered.

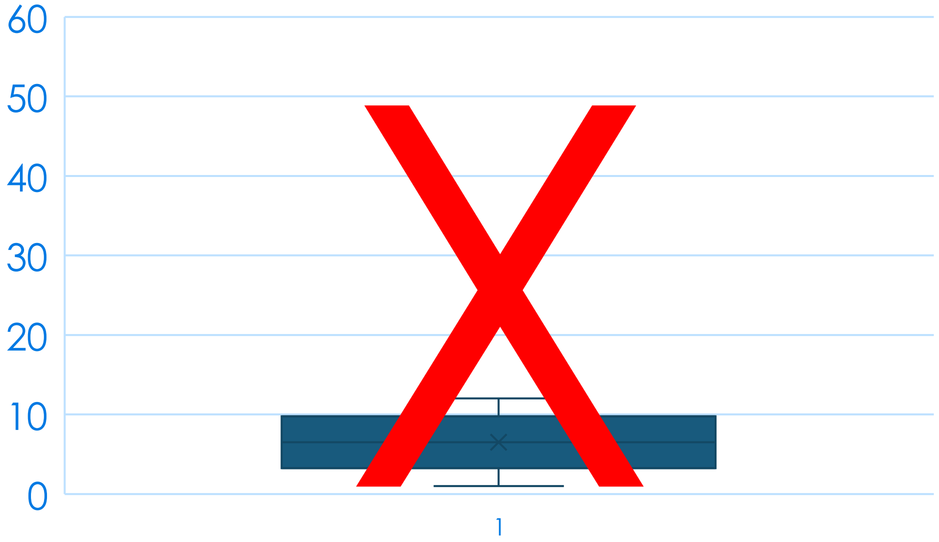


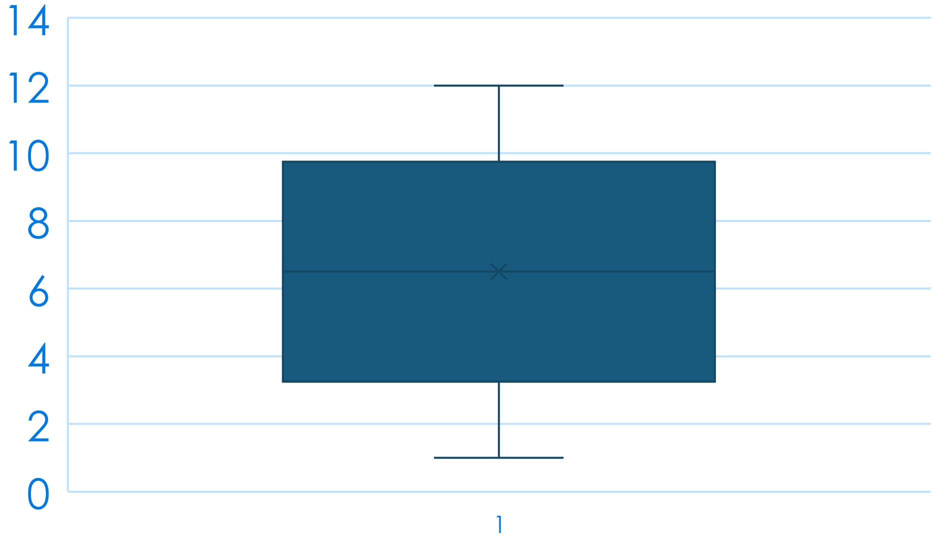


Always make sure the writings (axes, axes title, legend, title) are clearly visible!

Data visualization: suggestions – scale

In addition to the selection of the correct(s) method(s), some aspects should be considered.

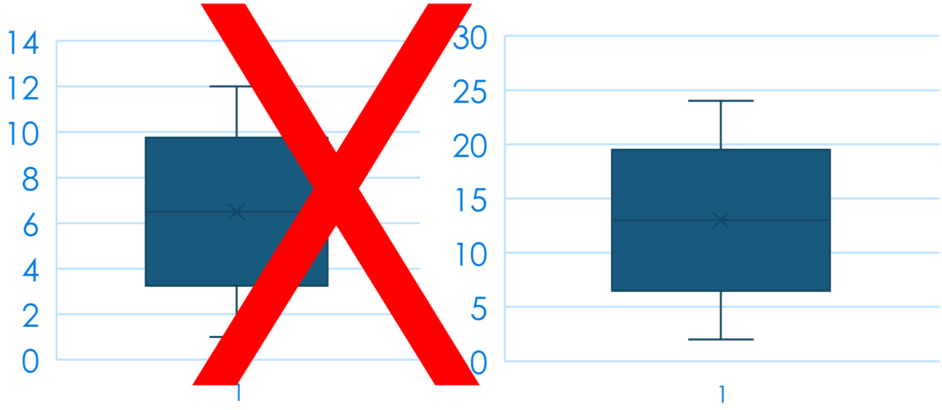


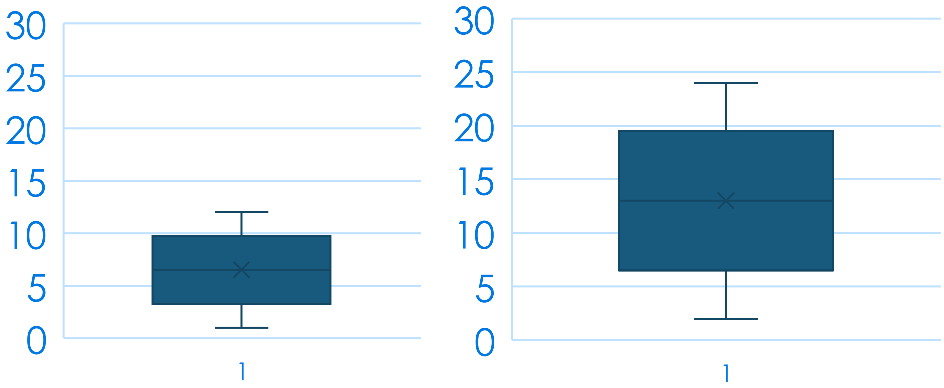


Always choose the correct scale. It is the scale (axis range) that allows for best representing data

Data visualization: suggestions – comparison

In addition to the selection of the correct(s) method(s), some aspects should be considered.





Always use the same scale for comparing plots