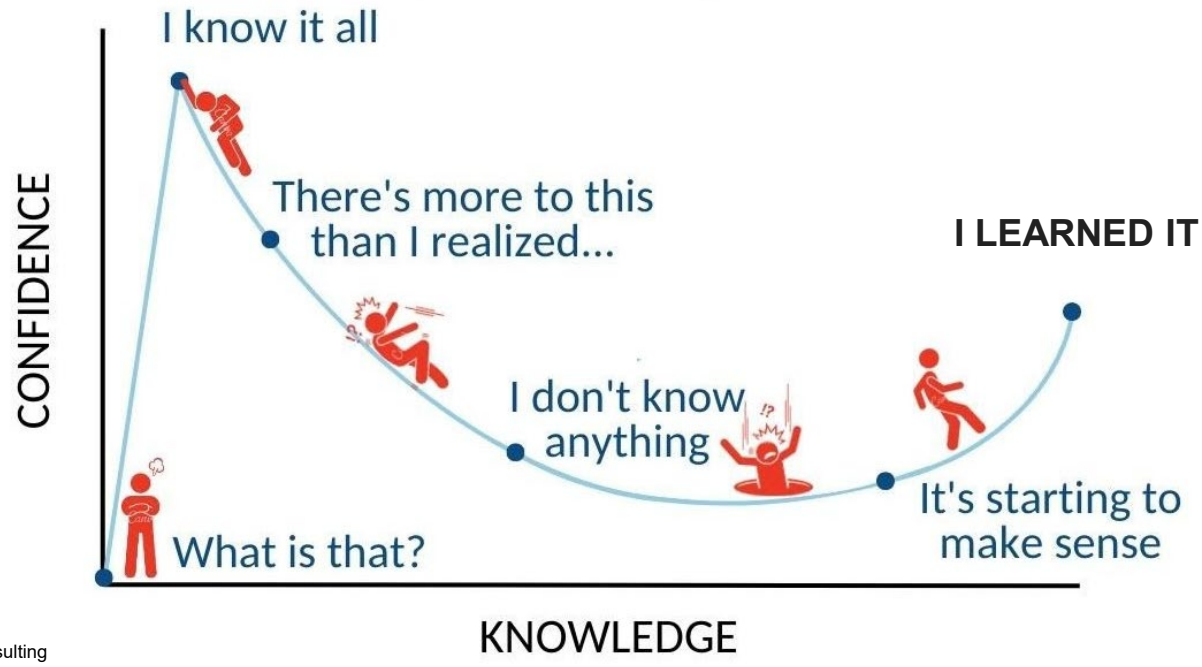
correlation.·one

TECH FOR JOBS

Dunning-Kruger Effect



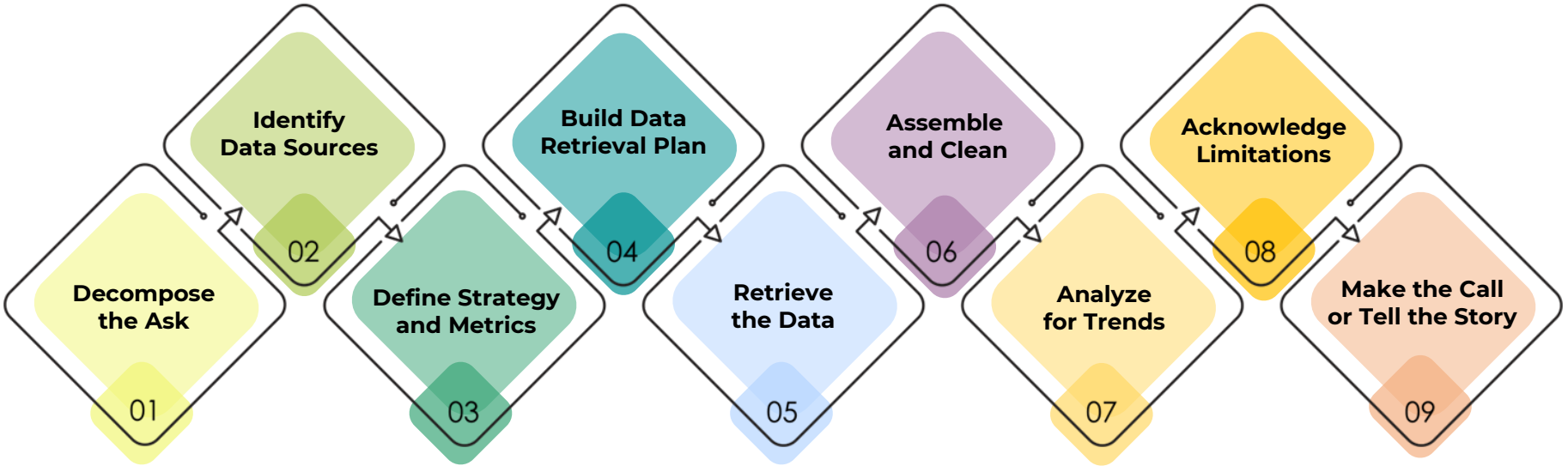
Credit: BVA Nudge Consulting

Data analytics is about what two things?

Fundamentally, data analytics
is about storytelling and truth-telling.

Analytics Paradigm

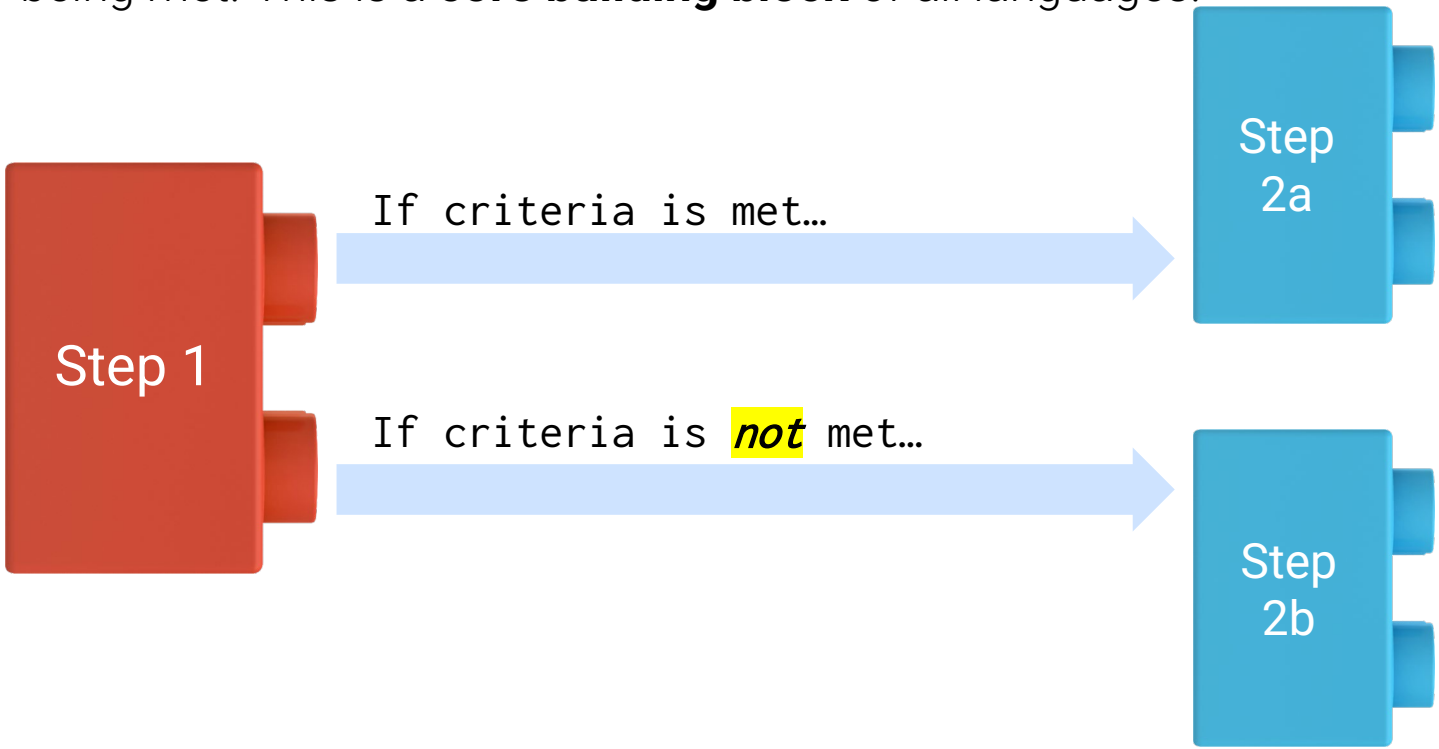
Regardless of type or industry, this paradigm provides a repeatable pathway for effective data problem solving.



Data Wrangling

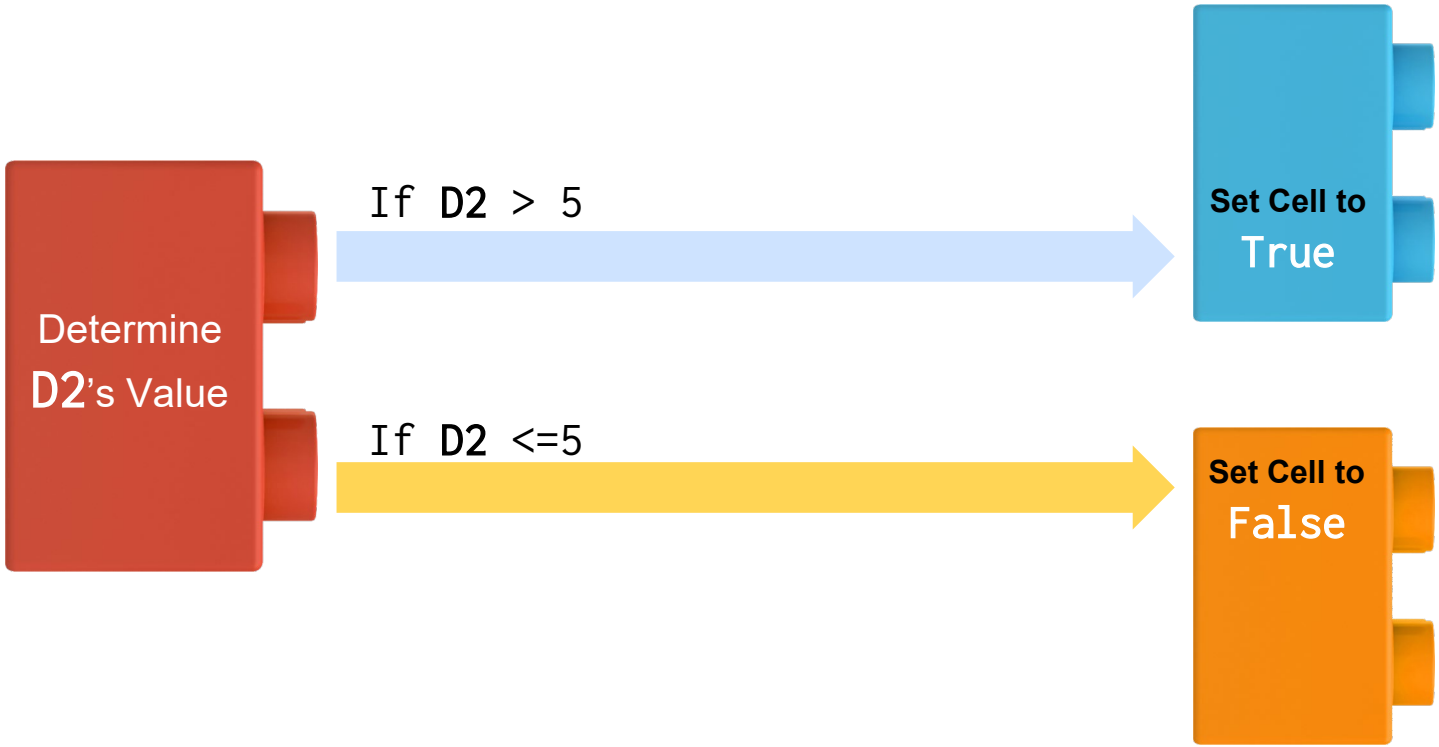
Conditionals: If This, Then That

Conditionals present a way to control the flow of logic based on certain criteria being met. This is a **core building block** of all languages.



Conditionals: If This, Then That

```
=IF(D2>5,TRUE,FALSE)
```



But what if we want to combine conditions?

AND, NOT, OR

Ooh…Coding! (Sort Of)

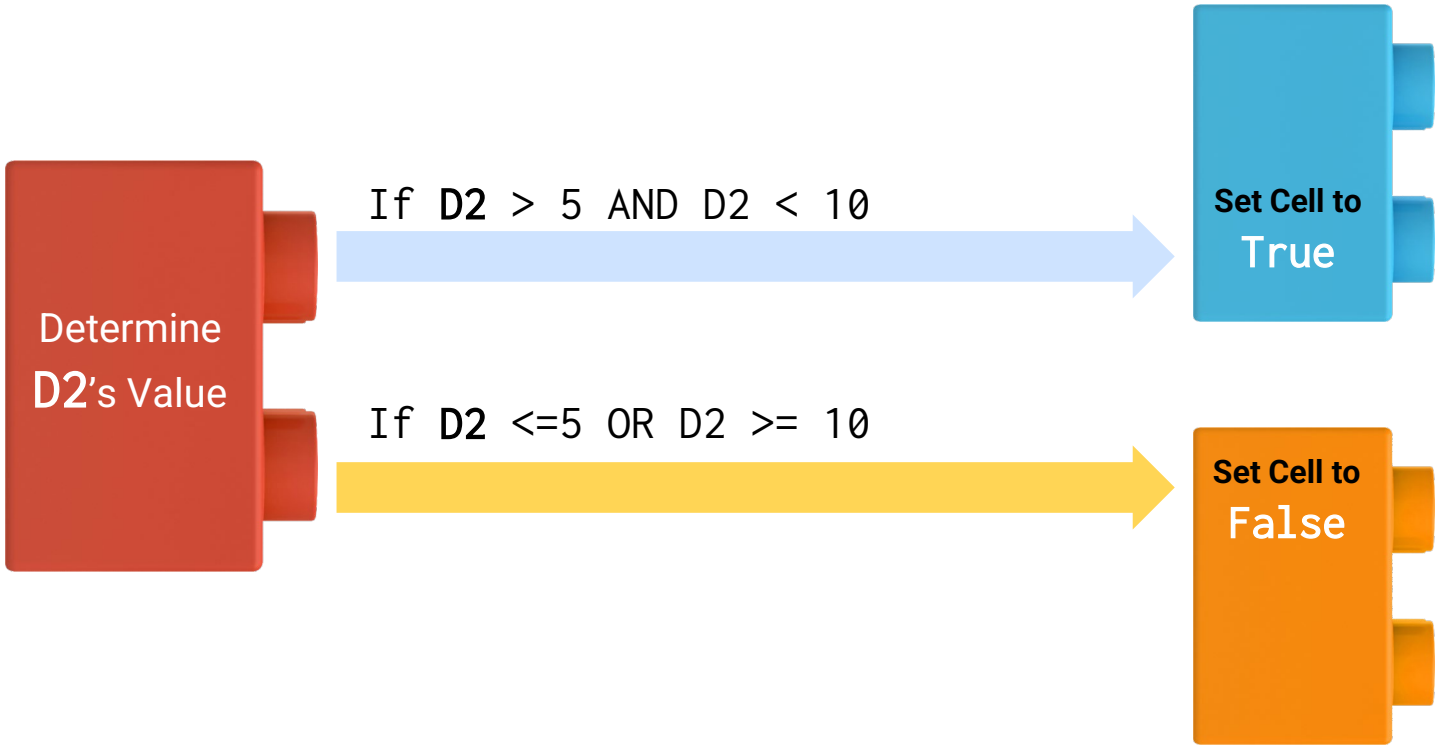
But what if we want
to **combine** conditions?

AND, NOT, OR

```
=IF(AND(D2>5, D2<10),TRUE,FALSE)
```

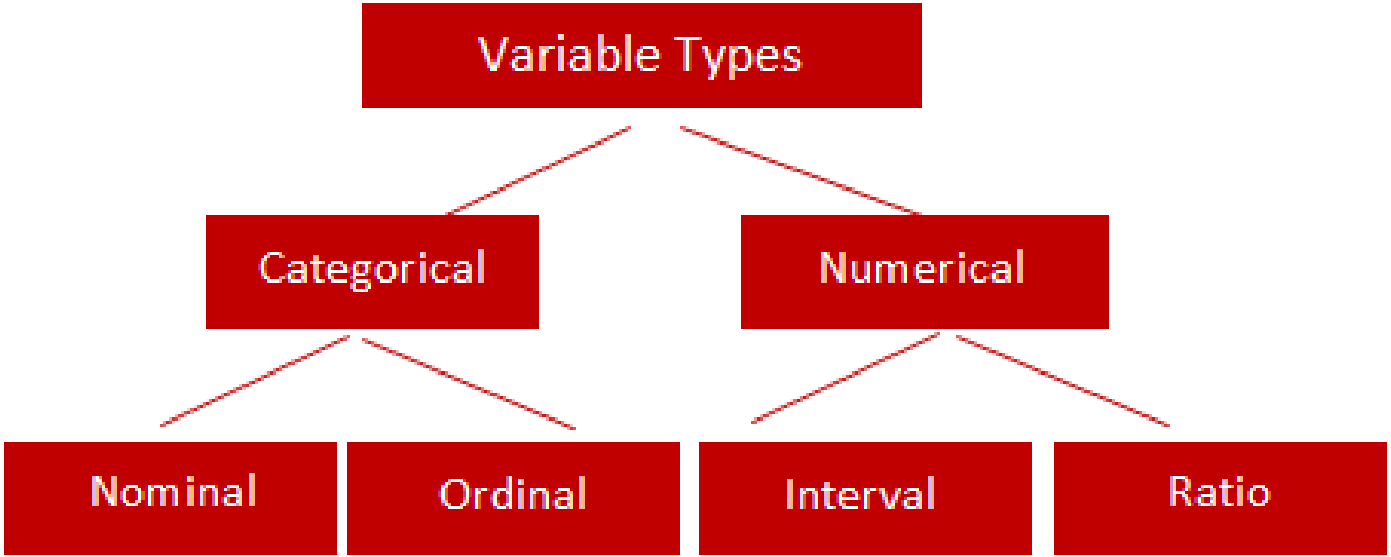
Conditionals: If This, Then That

Nesting conditionals are powerful, but can become convoluted very quickly!



Types of Data

Types of Data



What are “measures of central tendency”?

Values used to describe the center of a data set.

Central Tendency

Three most common measures of central tendency:

Mean

The “arithmetic” average

To calculate: The sum of all values, divided by the number of values

Median

The middle value of a data set

To calculate: Sort the data set and find the center

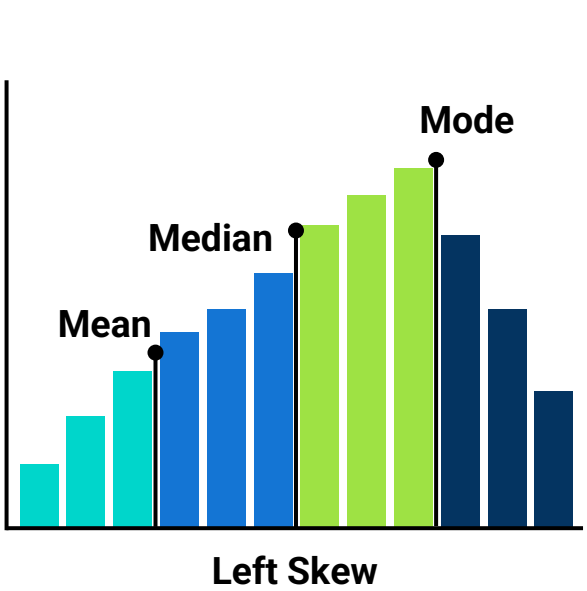
Mode

The most frequent value of a data set

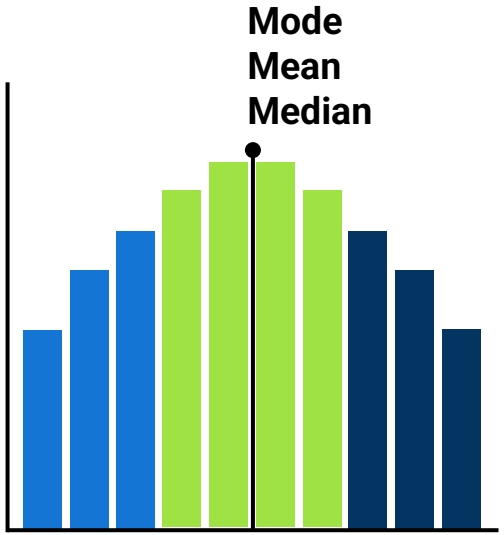
To calculate: Count the frequency of each value in a data set, determine the most frequent value

The mean, median and mode.

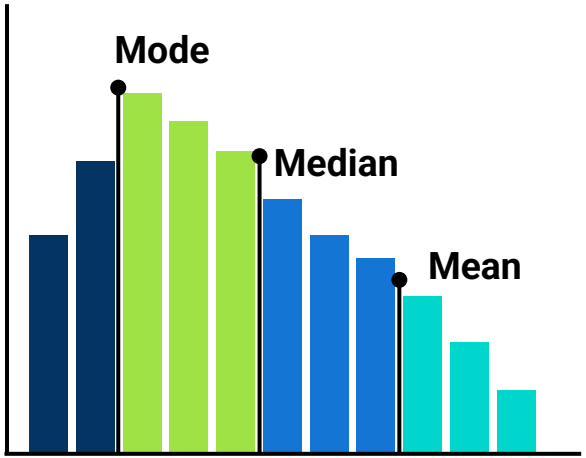
The mean, median and mode.



Left Skew



Normal Distribution



Right Skew

How do you describe the variability of a data set?

Variability of a Data Set

Three summary statistics metrics for describing variability:

01 Variance

02 Standard Deviation

03 Z-Score

Variance

- Used to describe how far values in the data set are from the mean
- Describes how much variation exists in the data
- Variance considers the distance of each value in the data set from the center of the data

The value of the one observation x_i

The mean value of all observations $\bar{x}$

Sample variance $S^2 = \dfrac{\sum (x_i - \bar{x})^2}{n - 1}$

The number of observations n

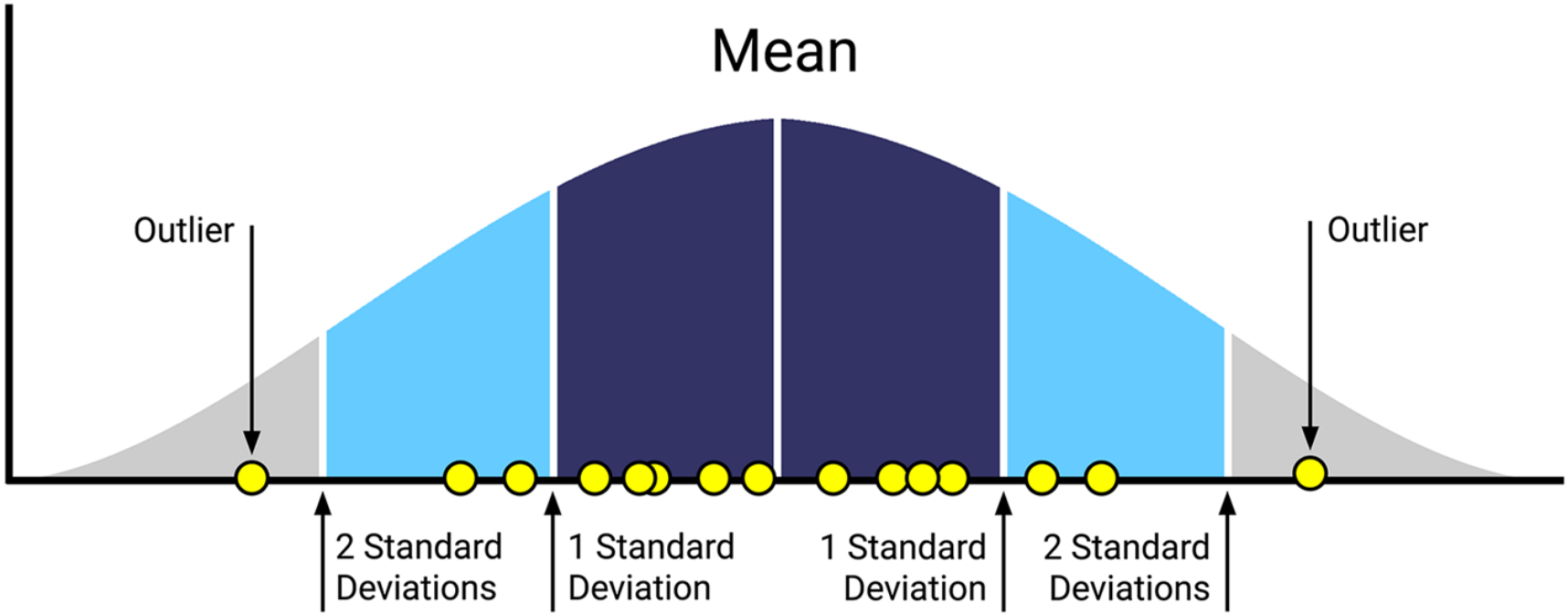
Standard Deviation

- Describes how spread out the data is from the mean
- Calculated from the square root of the variance
- In the same units of measurement as the mean

Standard deviation $\sigma = \sqrt{S^2}$ **The variance**

Standard Deviation

Square root of the variance; a measure used to quantify the dispersion of a set of observations.



Z-Score

Z-Score describes a single value's distance from the mean of the data set
The distance is in terms of standard deviations. Can be positive or negative:

If negative	the value is less than the mean
If positive	the value is greater than the mean.

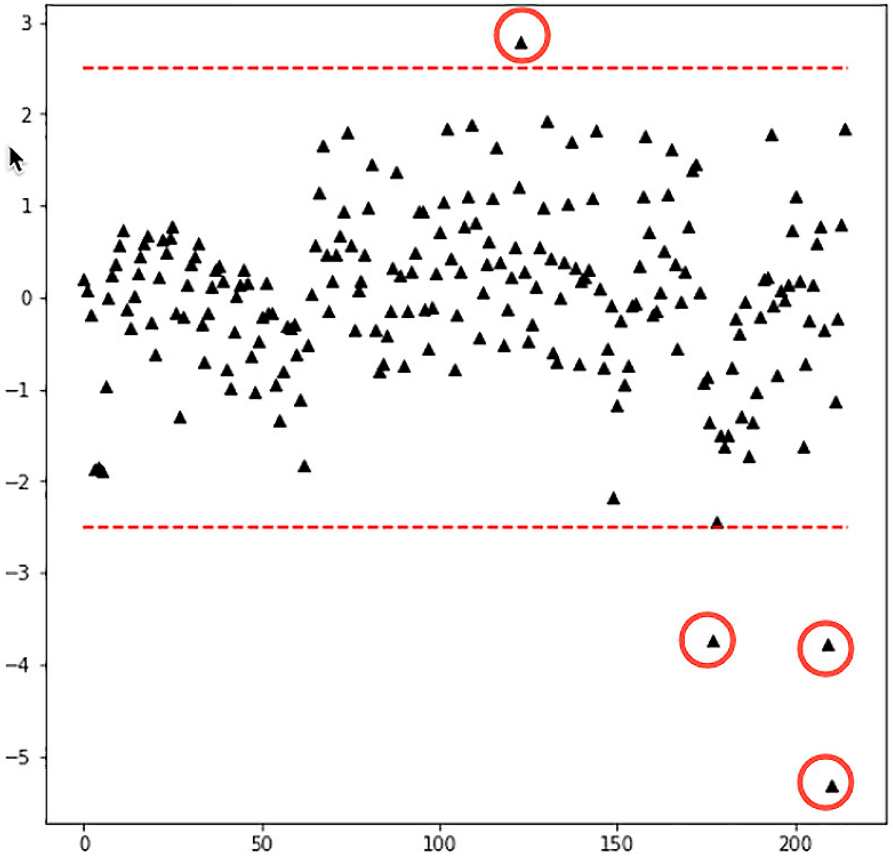
The smaller the z-score, the closer the value is to the mean

A single value X — μ **The mean of the dataset**

$$z = \frac{X - \mu}{\sigma}$$

σ **The standard deviation of the dataset**

Z-Score



But how can we summarize real-world data?

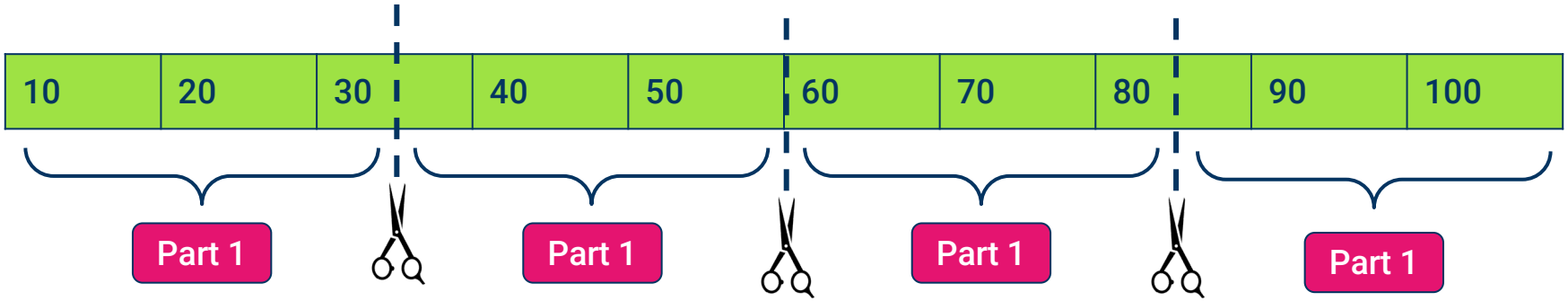
Quantiles: Used to Describe Segments of a Dataset

Quantiles separate a sorted dataset into equally sized fragments.

The two most popular types of quantiles are quartiles and percentiles.

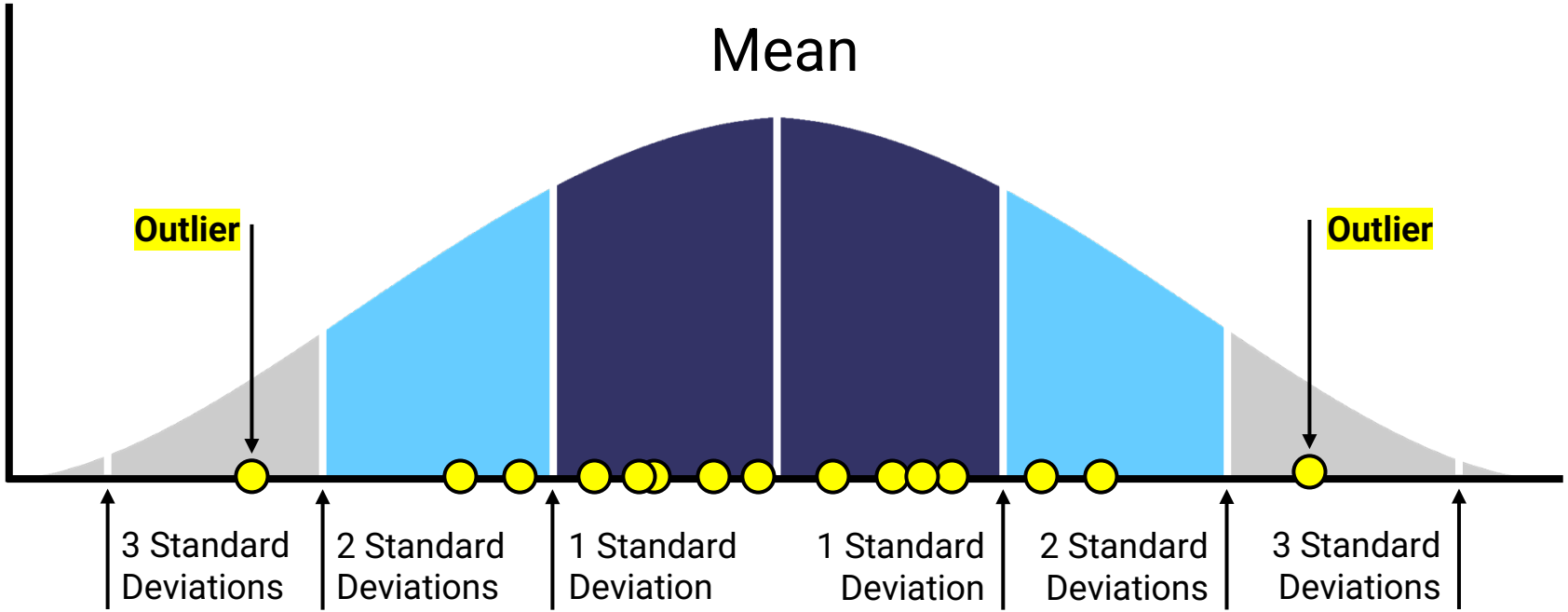
01 **Quartiles** divide the dataset into four equally sized parts.

02 **Percentiles** divide the dataset into 100 equally sized parts.



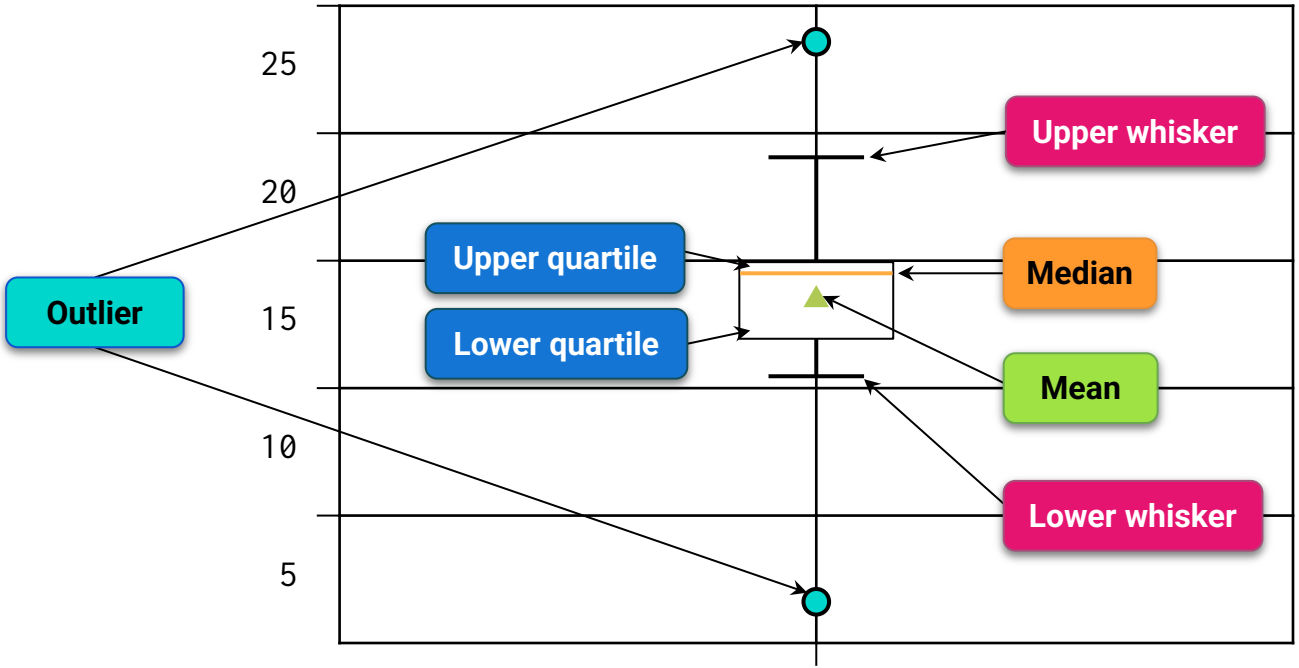
Outliers

Suspicious values are called potential outliers. An outlier is a data point that differs from the rest of a data set. Outliers can inaccurately skew a data set.



Qualitatively

Use **box-and-whisker plots** to visually identify potential outliers.



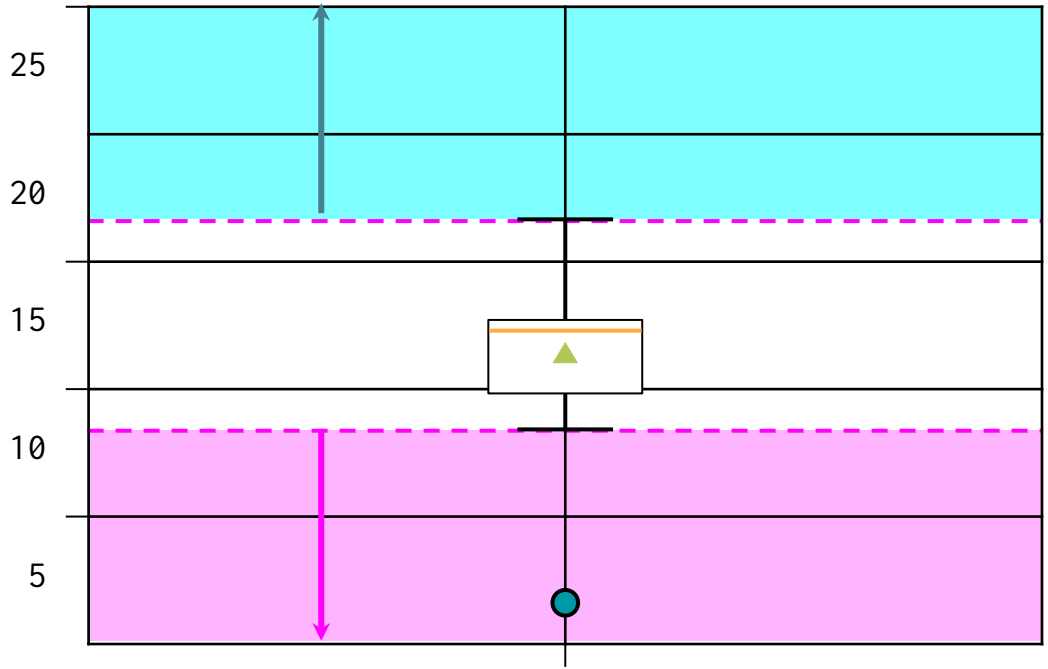
Quantitatively

Determine the outlier boundaries in a dataset by using the **1.5 × IQR rule**.

The IQR is the range between the first and the third quartile.

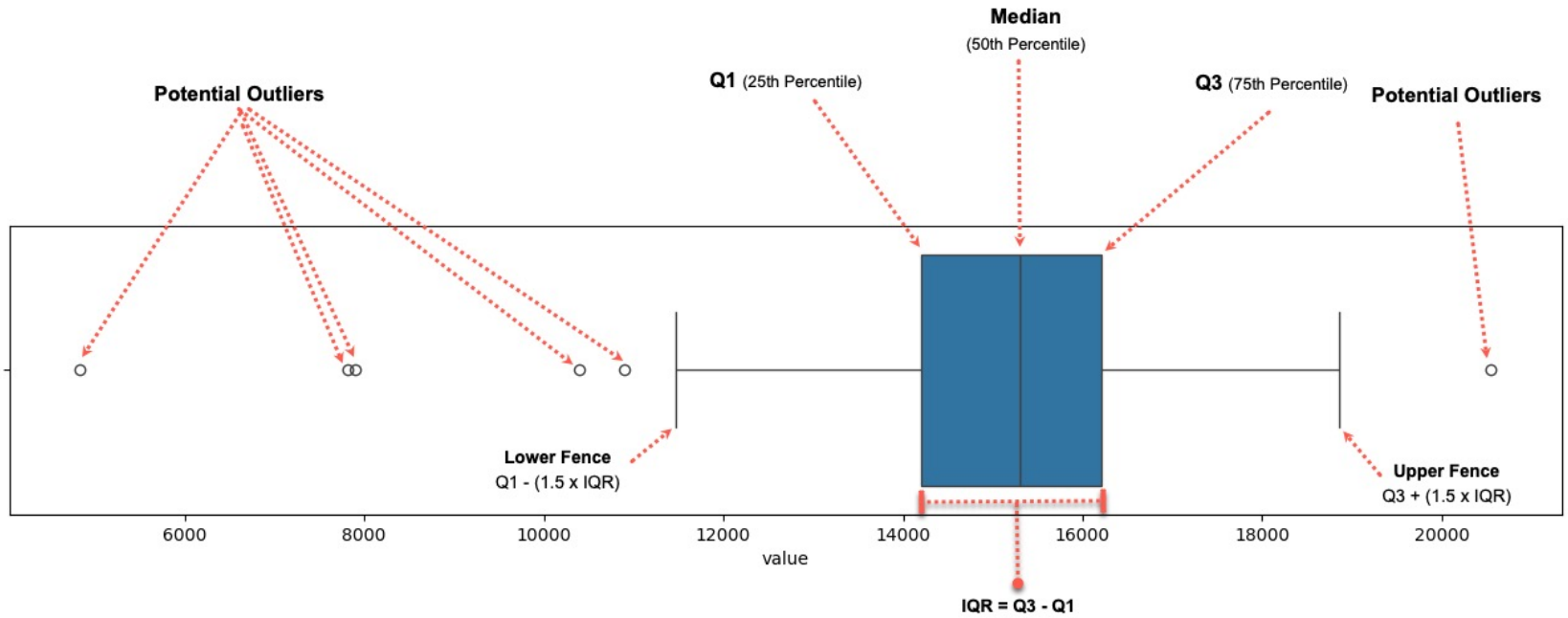
Anything less than, or below, Quartile 1 − (1.5 × IQR) might be an outlier.

Anything greater than, or above, Quartile 3 + (1.5 × IQR) might be an outlier.



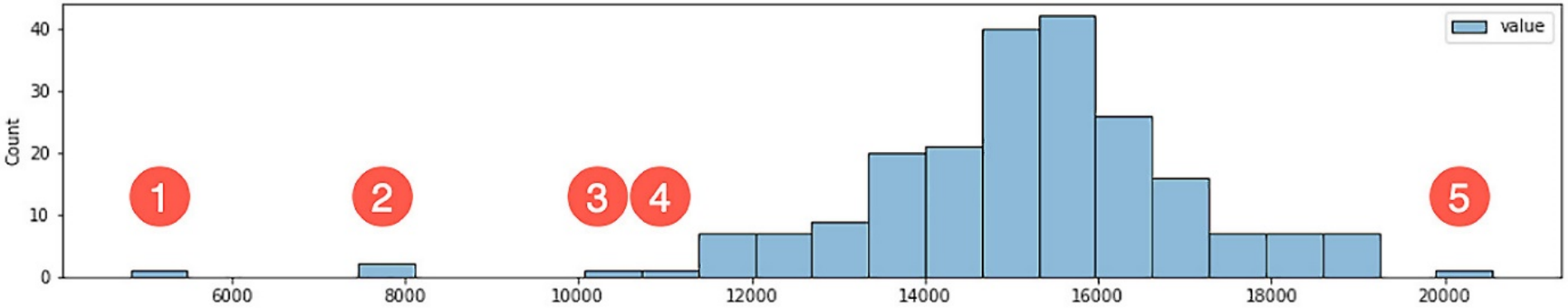
Qualitatively

Use **box-and-whisker plots** to visually identify potential outliers.



Other Visuals

Use **Histograms** to visually identify potential outliers.



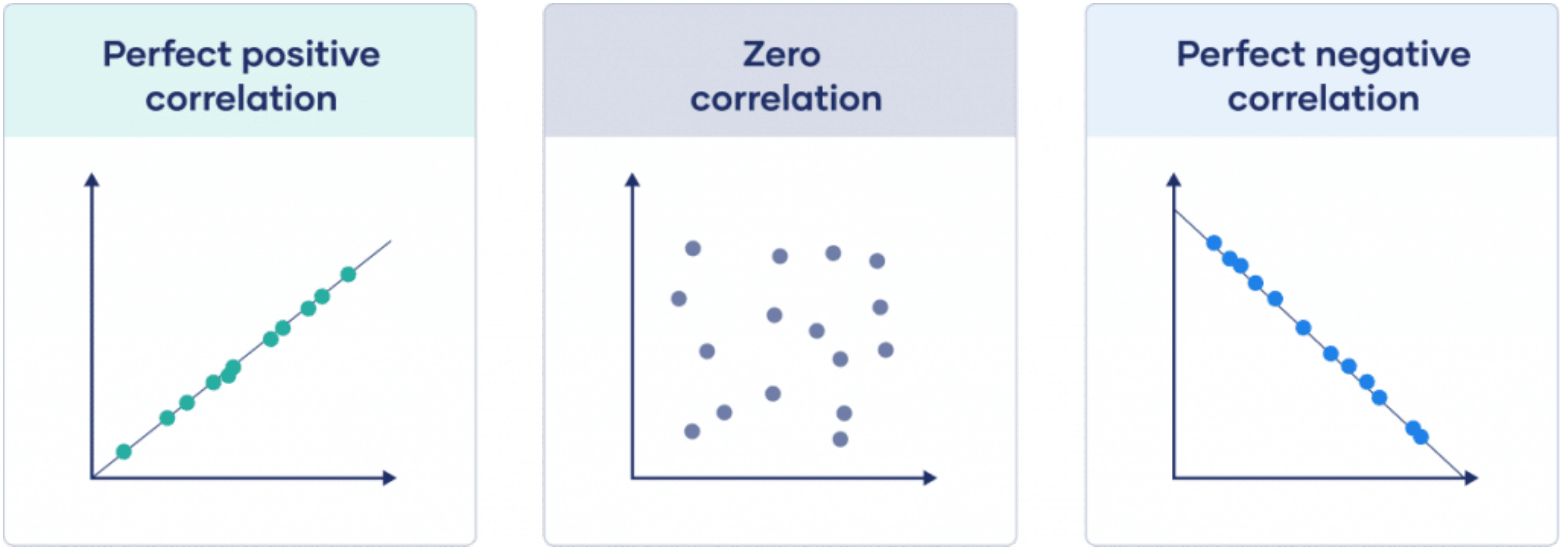
Scatter plots are a powerful visualization tool!

Visualizes the comparison between two variables:

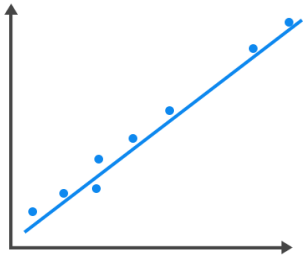
One variable	is located on the x-axis
Another variable	is plotted on the y-axis

- Each data point represents a pair of measurements
- Measurements on a scatter plot are independent
- Scatter plots can help to identify positive or negative relationships between two variables
- Adding a trend line to a scatterplot can visualize this relationship even easier!

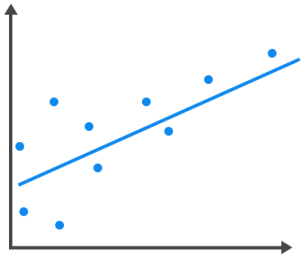
Correlation



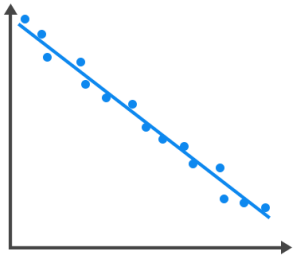
Correlation



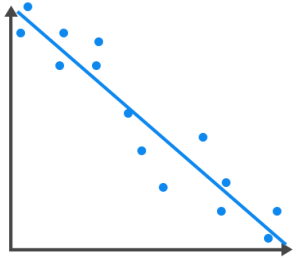
Strong positive correlation



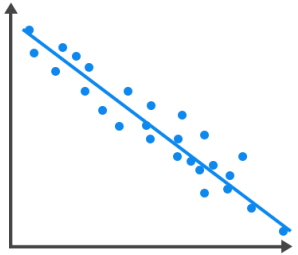
Weak positive correlation



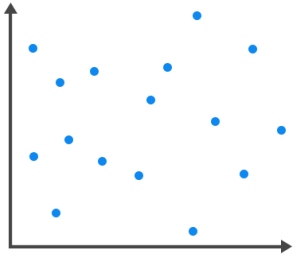
Strong negative correlation



Weak negative correlation



Moderate negative correlation



No correlation