

# Intro to Data Wrangling and EDA

CS 459 Business Intelligence



# Data Wrangling



## Data Wrangling

also called Data Munging

- Data Wrangling is the process of gathering, collecting, and transforming **Raw data into another format for better understanding, decision-making, accessing, and analysis in less time.**
- *All the activity that you do on the raw data to make it “clean” enough to input to your analytical algorithm is called data wrangling or data munging. — Shubham Simar Tomar 2016*

# 1. Discovering

- Getting familiar with the data
- Identify multiple ways to use data for different purposes – check the ingredients before cooking a meal
- Data possibly collected from multiple sources; formatting is required to understand relationships.



## 2. Structuring

- Data structuring transforms raw data into a structured format for easier interpretation and analysis.
- Raw data doesn't help analysts because it's incomplete or incomprehensible.
- It needs to be parsed so that analysts can extract relevant information.



## 3. Cleaning

- People often use **data cleaning** and **data wrangling** interchangeably.
- However, data cleaning is one step in the data wrangling process.
- Clean and resolve issues with the data.



## 4. Enriching

- After transforming data into a usable format, find whether data from other datasets can make your analysis more effective.
- Helps improve quality of the data if it does not meet the requirement.
- Enrich with data example:
  - Combine 2 databases where one contains phone numbers and others don't
  - Create new columns

## 5. Validating

- Check for data accuracy and quality.
- Data validation ensures that data is fit for analysis.



## 6. Publishing

- Publish the data after validating.
- Shared as report, electronic document or deposited into a database which can be processed further to create larger and more complex structures such as data warehouses.
- **Once published, data is all set for analysis.**

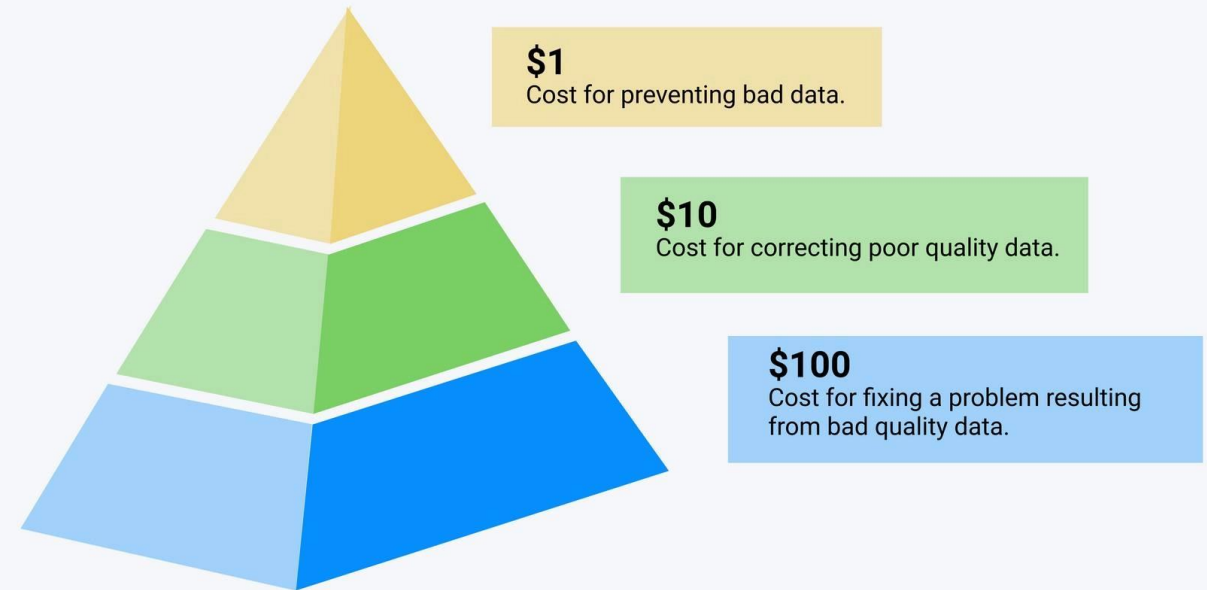
# Summarizing 6-steps of Data Wrangling



# Importance of Data Wrangling

- In data science & analysis, the amount of work that goes into data wrangling is embodied by the **80/20 rule**
- Data scientists typically spend **80% of their time 'wrangling' or preparing data** and **20% of their time actually analyzing the data.**

The Cost of Bad Quality Data Over Time



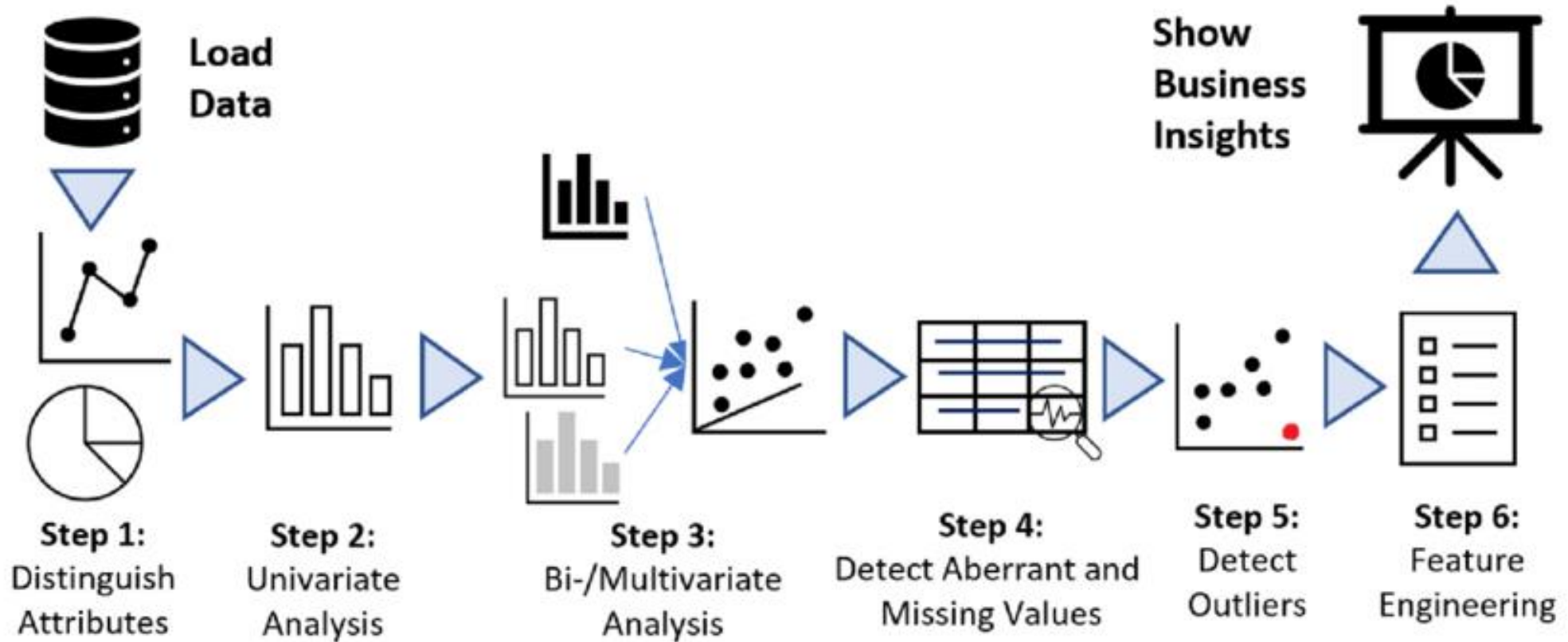
# Exploratory Data Analysis (EDA)



**E**xploratory **D**ata **A**nalysis involves:

- Examining the distribution of various variables in the dataset
- Identifying outliers
- Discover trends and patterns
- Analyze relationships between variables by using heat maps or correlation metrics.

# EDA



# Data Wrangling





# Types of dirty data



**Duplicate data**



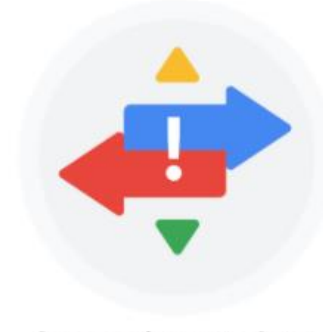
**Outdated data**



**Incomplete data**



**Incorrect/inaccurate data**



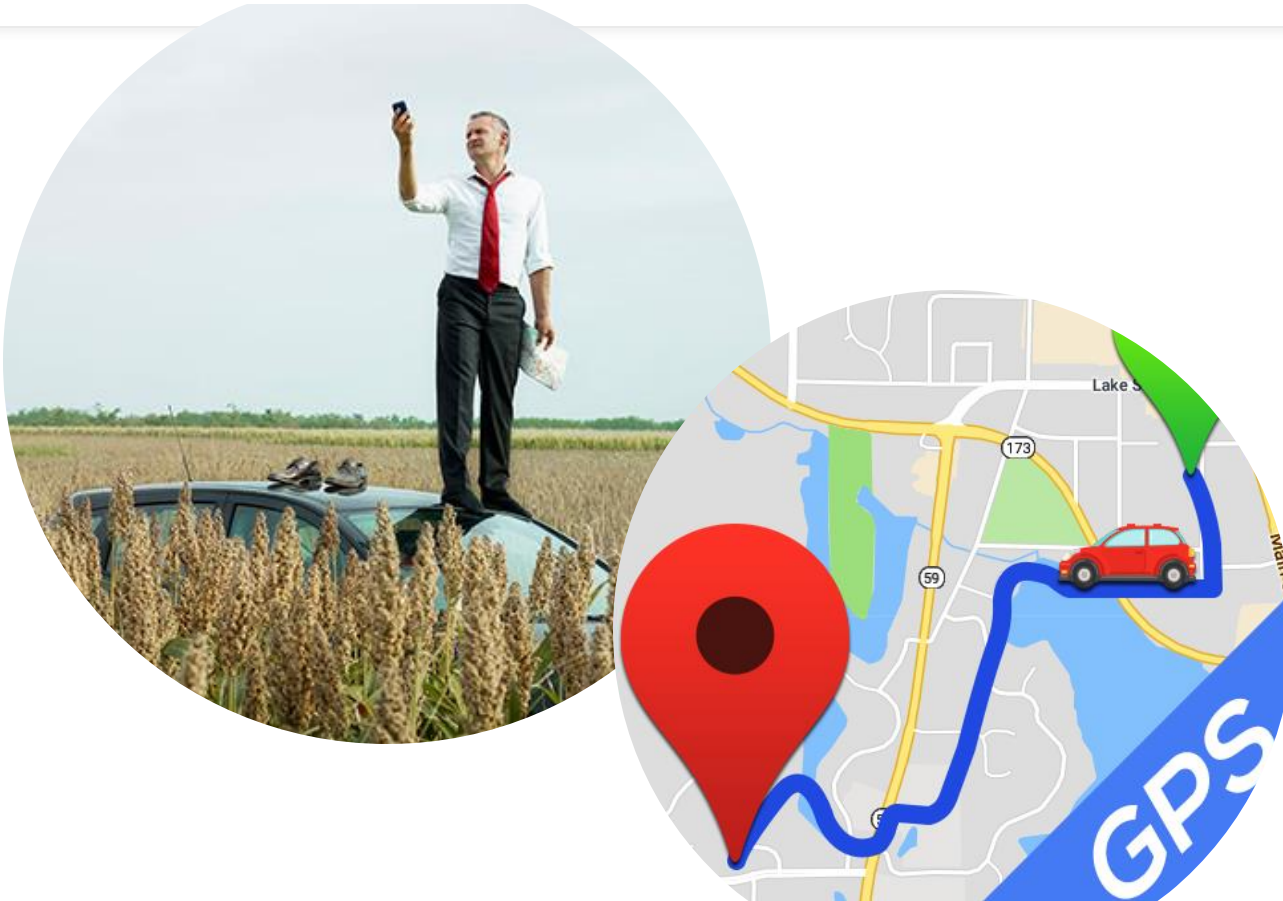
**Inconsistent data**

# Duplicate data

Repeated entries or records

S.N°	First Name	Last Name	Title	Company
1	Mary	Sue	Senior Marketing Manager	ABC Ltd.
2	Janet	Martin	Marketing Executive	ABC Ltd.
3	Bryan	Oscar	SEO Manager	ABC Ltd.
4	Jude	Taylor	Marketing Manager	ABC Ltd.
5	Mary S	Sue	Senior Marketing Manager	ABC Ltd.

# Outdated Data



- Data that no longer reflects the current state, trends, or conditions, often due to time lapse or changes in the system.

# Incomplete data or Missing Data

Customer ID	Last Purchase Date	Last Purchase Price	Age
Customer 1		\$130	37
Customer 2	24-Apr-20	\$310	40
	15-May-20	\$386	45
Customer 4	18-May-20		55
Customer 5	6-Mar-20	\$453	

Incomplete data



# Missing Values

# Missing Values

- Every value in every column has a certain probability of being missing (Rubin, 1976)
  - Generally, there is a probability distribution of any column in any data, i.e., which defines the shape of the probabilities of occurrence of that column (e.g., bell curve, exponential, logarithmic etc.)
- **Missing Completely at Random (MCAR)**
- **Missing at Random (MAR)**
- **Missing Not at Random (MNAR)**

# Missing Values

## Missing Completely at Random (MCAR)

- Every value in a column has the **same probability** of being missing.
- The cause of missingness is **unrelated** to the data itself.

# Missing Values

## Missing Completely at Random (MCAR)

### EXAMPLES:

- A weighing scale generates missing data because the **batteries die**.
- Sales data for an outlet is missing because the **store was closed for maintenance**.
- ATM transaction data is missing because the **machine was being refilled or a technical glitch** affected multiple locations.
- Data is missing **randomly and unpredictably**, with no systematic pattern.

# Missing Values

## Missing at Random (MAR)

- Different column values (e.g., different groups) can have **different probabilities** of being missing – **most common case**
- Causes of the missing data are **related** to the data

# Missing Values

## Missing at Random (MAR)

### EXAMPLES:

- A weighing scale produces **more missing values for heavier products**
- Sales data missing for **teenage customers** because no promotion targeted to them
- ATM data is missing on **weekends/holidays** due to lower transaction volumes. The missingness is related to the observed variable (day of the week) but not directly to the missing values.
- Missing values **follow a pattern**, making them **predictable based on other variables**.

# Missing Values

## Missing Not at Random (MNAR)

- The **probability of missingness depends on unobserved factors** or the **missing values themselves**.
- Neither MCAR nor MAR fully explains the missing data.

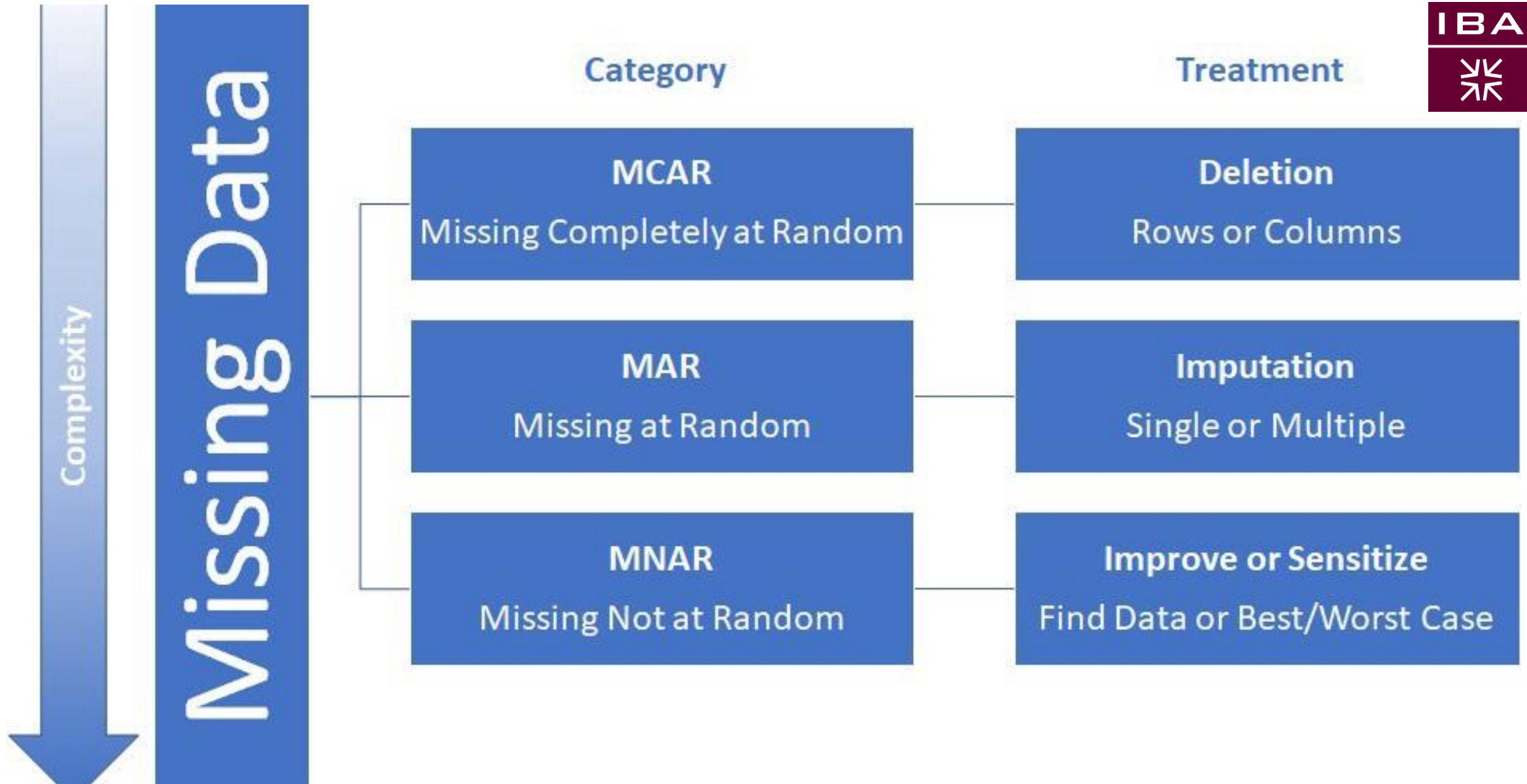
# Missing Values

## Missing Not at Random (MNAR)

### EXAMPLES:

- A weighing scale **wears out over time**, leading to **progressively more missing values**.
- Sales data is increasingly missing because **customers are relocating** (not recorded).
- ATM transactions decline because **people fear theft**, leading to **missing data correlated with safety concerns**.

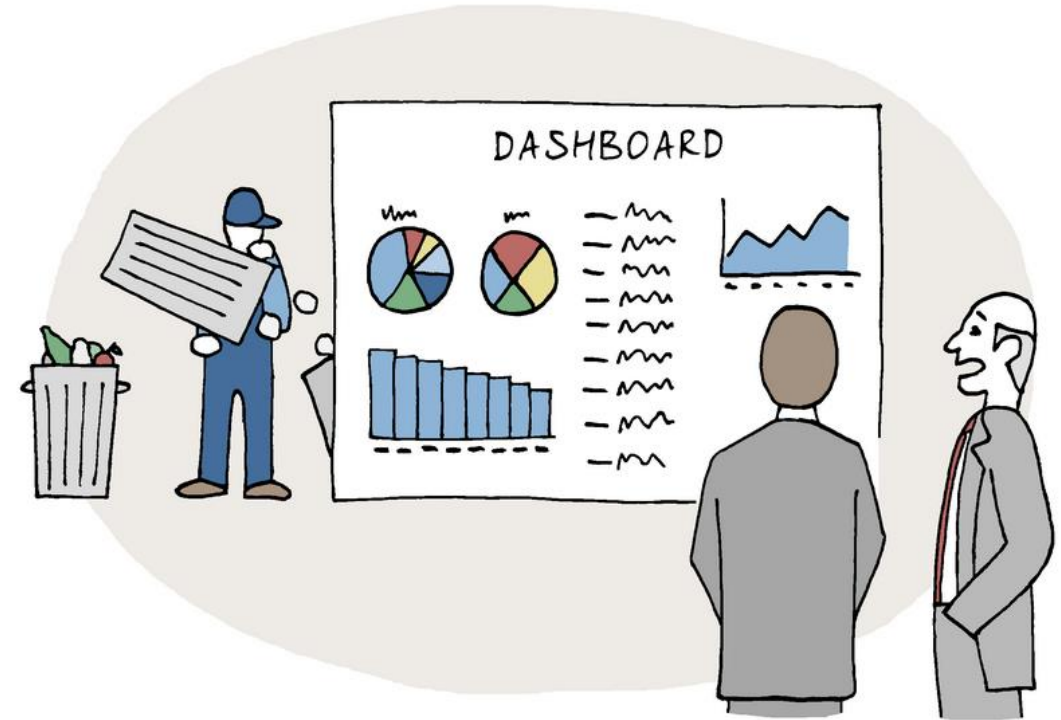
The **reason for missingness is unknown or unmeasured**, making it **difficult to handle without external data**.



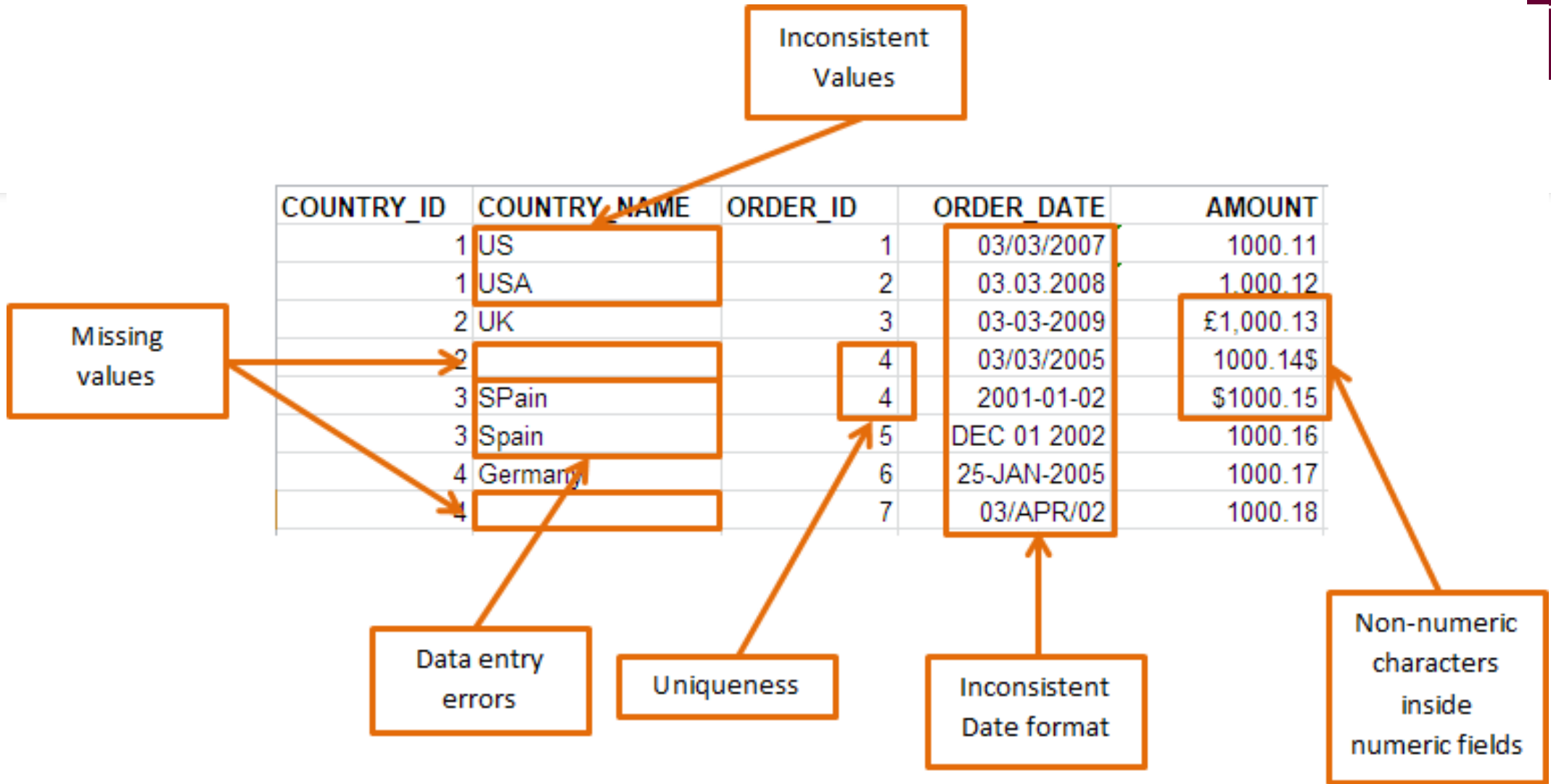
# Incorrect/Inaccurate Data

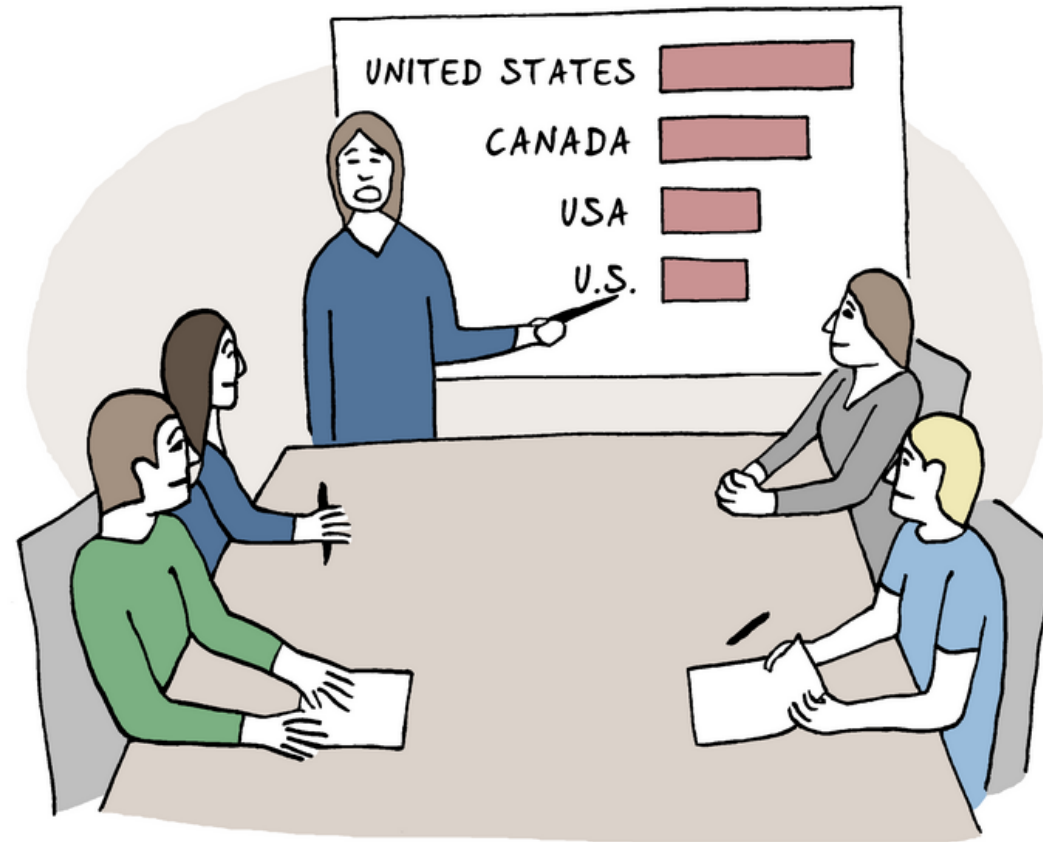
- If an online store records double the number of sales in a certain month, it could lead to an increased average customer spend value.
- While the data might make it seem like the store is performing well, this **false information** could lead to poor decision-making

- Incorrect data leads to incorrect insights
- Will the analysis be useful?
- A waste of time, energy and resources.



DO WE TRUST THIS DATA?





AS YOU CAN SEE, OUR TOP MARKETS ARE  
UNITED STATES, CANADA, USA AND THE U.S.

# Types of dirty data



**Duplicate data**



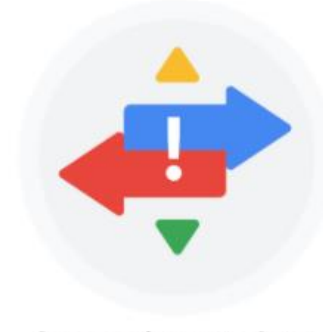
**Outdated data**



**Incomplete data**




**Incorrect/inaccurate data**





**Inconsistent data**

# DATA CLEANING CHECKLIST





**Up-to-date data**

Data should be up-to-date in order to obtain maximum value from the data analysis.



**Missing values**

Count missing values and analyze where in the data they are missing. Missing values can disrupt some analyses and skew the results.



**Duplicates**

Duplicate IDs indicate multiple records for one person, e.g. someone holds multiple functions at the same time.



**Numerical outliers**

Numerical outliers are fairly easy to detect and remove. Define minimum and maximum to spot outliers easily.


**Check IDs**

Check data labels of all the fields to see whether some categorical values are mislabeled.

**Define valid output**

Define valid data labels for categorical data. Define data ranges for numerical variables. Non-matching data is presumably wrong.



## Data Cleaning

# Problems with the Data

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values

Invalid values

Misfielded values

Misspellings

Uniqueness

Formats

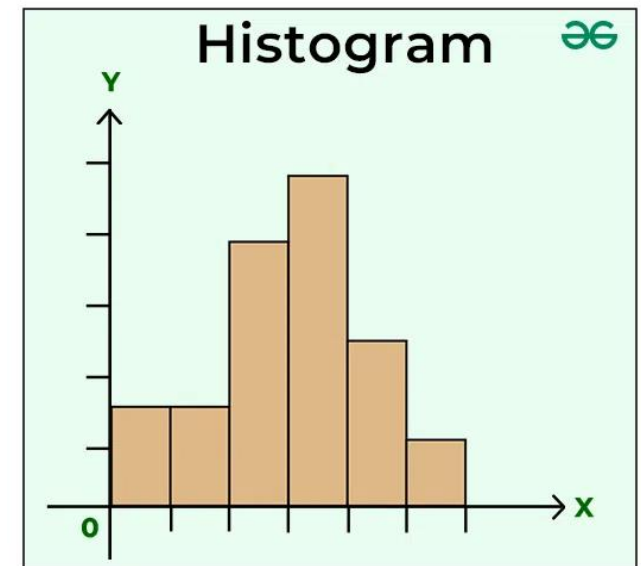
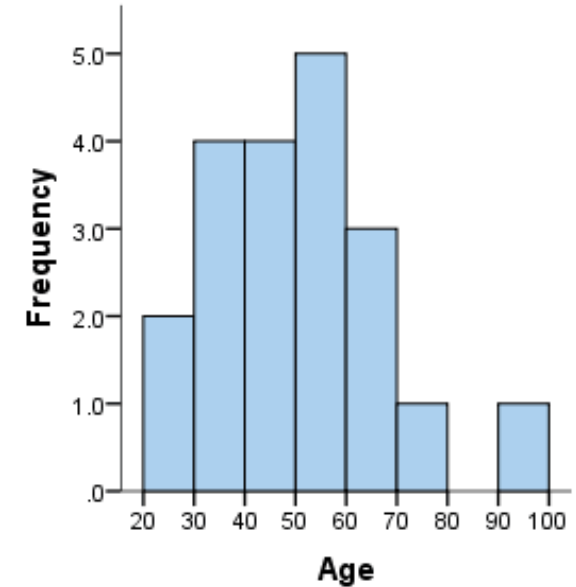
Attribute dependencies

# Interpreting Histograms and Box plots

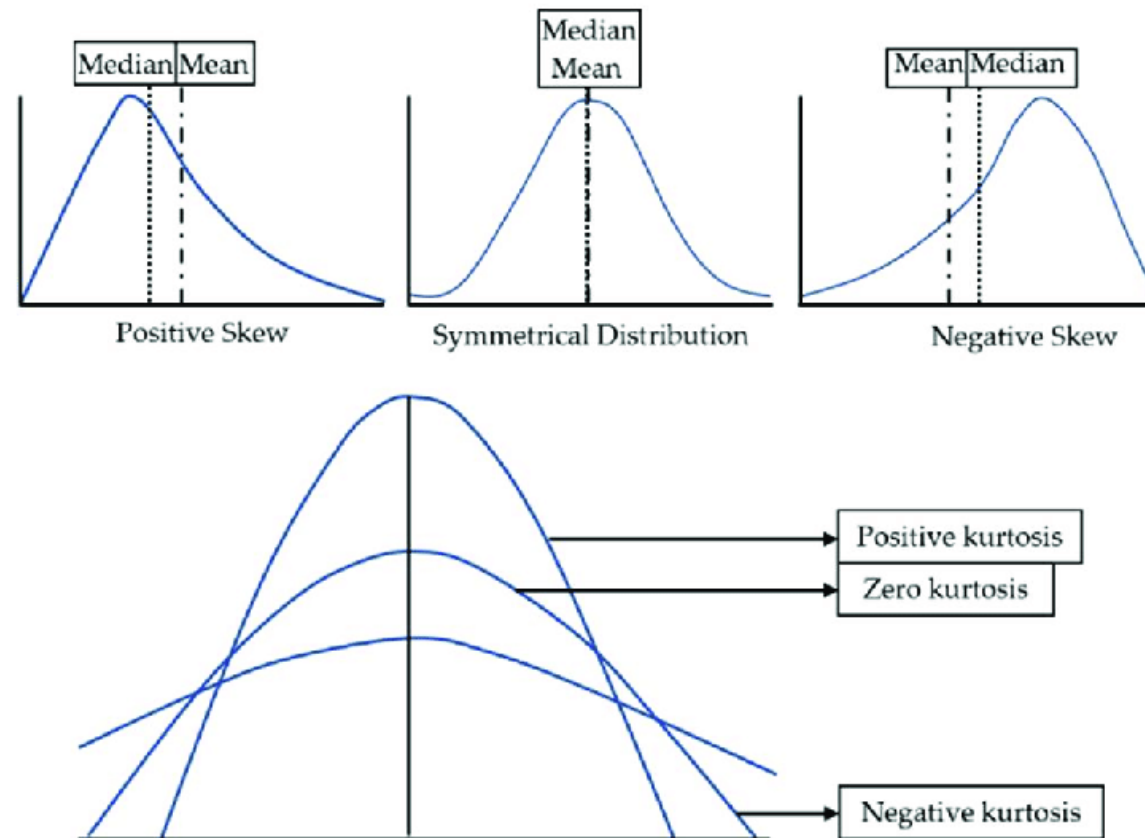


# What is a Histogram?

A **histogram** is a graphical representation of the frequency distribution of continuous series using rectangles. The x-axis of the graph represents the class interval, and the y-axis shows the various frequencies corresponding to different class intervals.



# Analyzing Histograms: Shape, Skew and Kurtosis



# Mean, Median, Mode

- **Mean:** The "average" number; found by adding all data points and dividing by the number of data points.

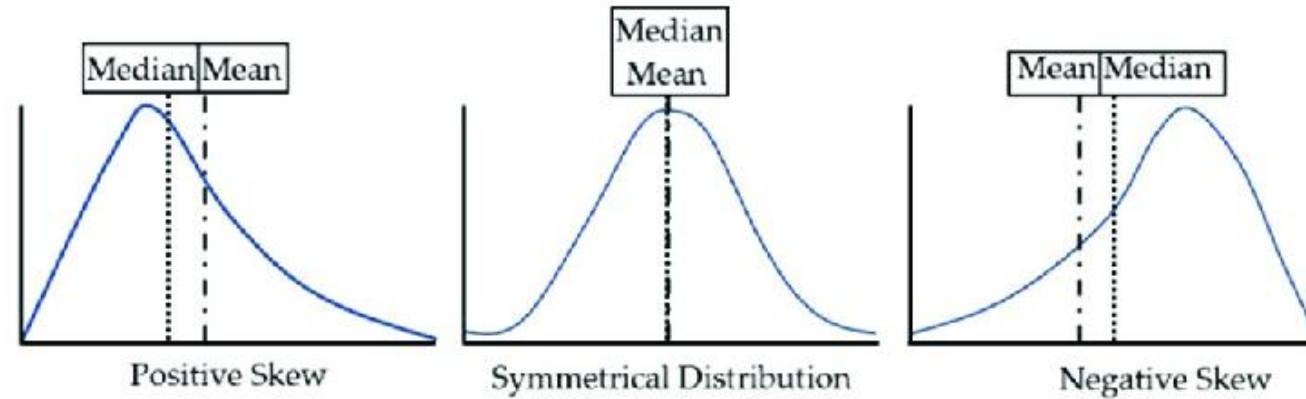
(impacted by outlier)

- **Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

(Not impacted by outlier)

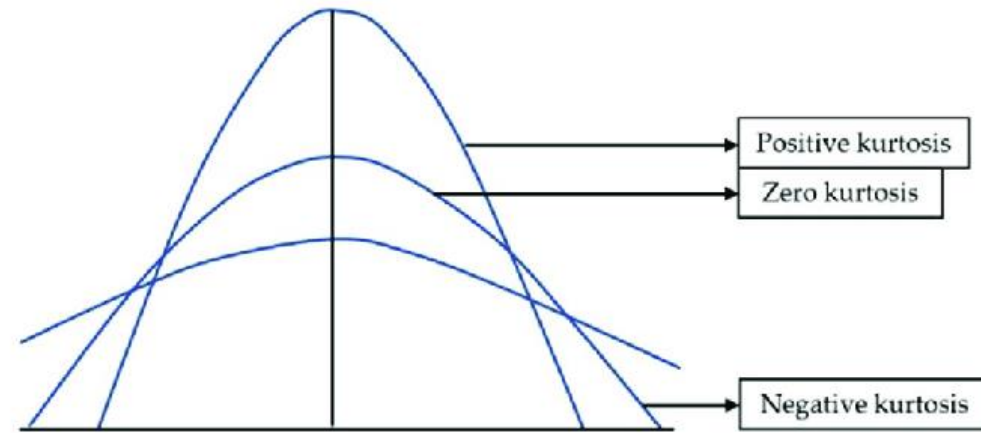
- **Mode:** The most frequent number—that is, the number that occurs the highest number of times.

# Skew



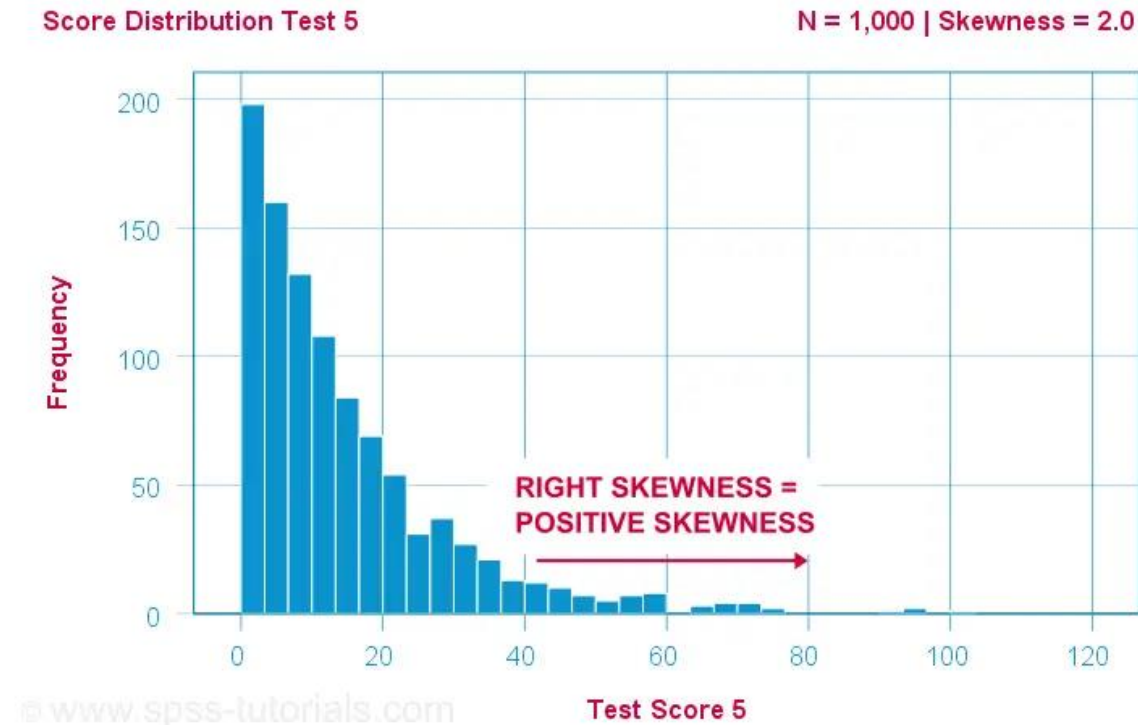
- Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side. Positive (long tail on right) and Negative (long tail on left)

# Kurtosis

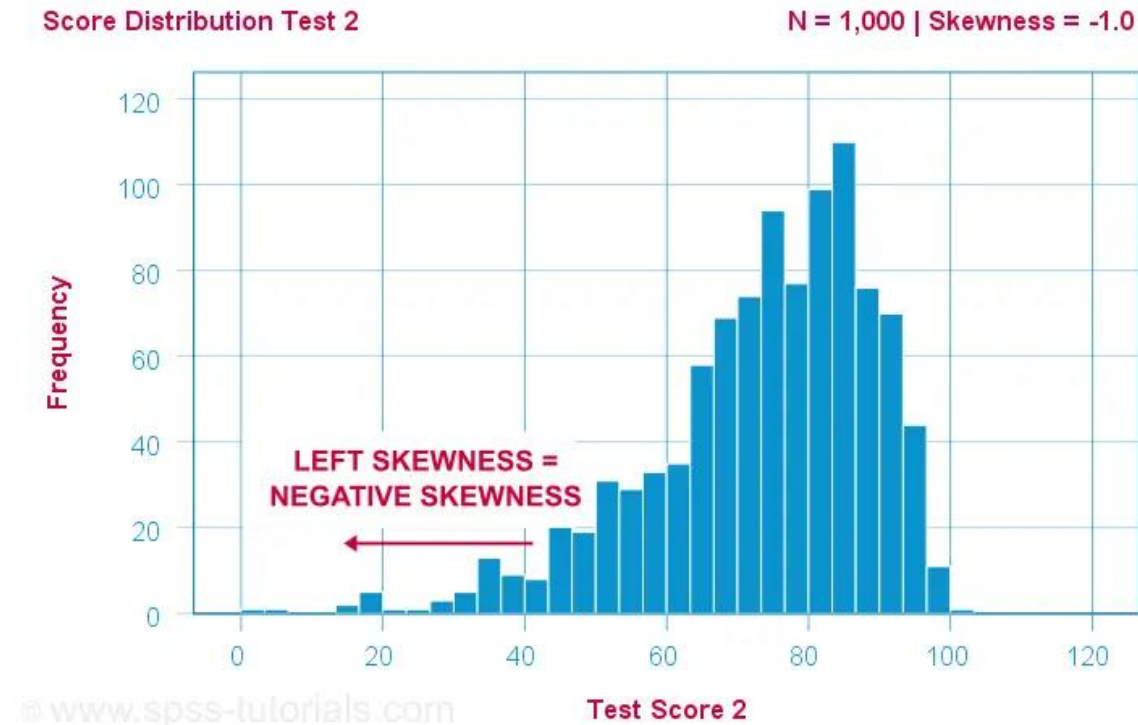


- Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.
- Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution.

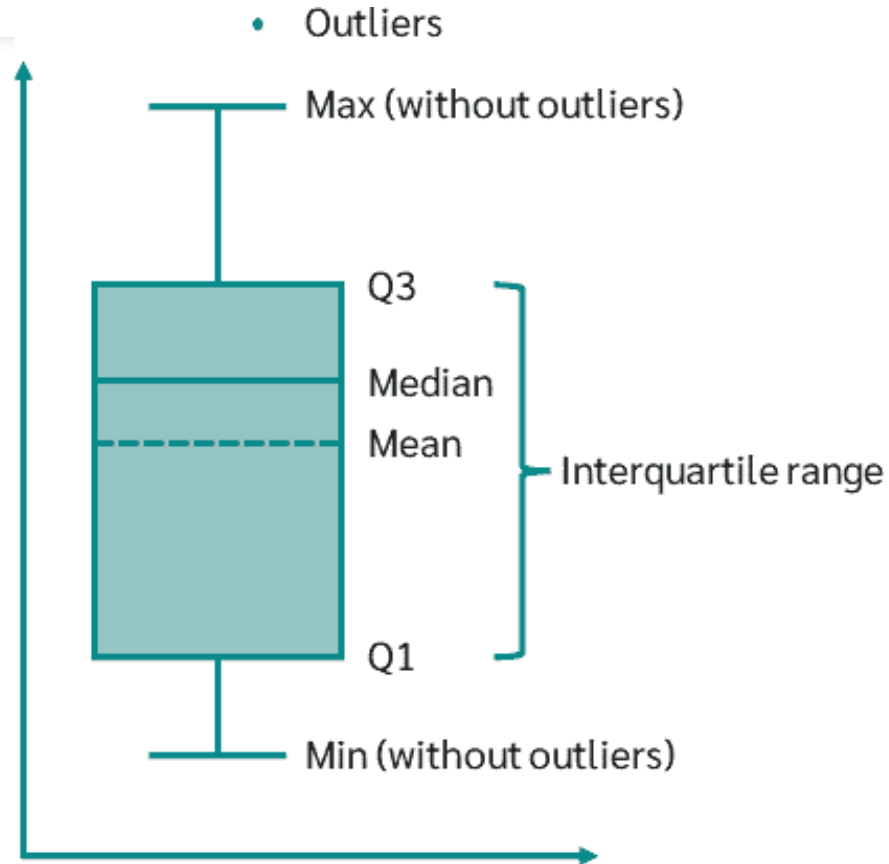
# Example: Scores on a Test



# Example: Scores on a Test



# Interpreting Box Plots



The box indicates the range in which the middle 50% of all data lies

Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile

Between Q1 and Q3, is the interquartile range

In the boxplot, the solid line indicates the median and the dashed line indicates the mean.

The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range.

Points that are further away are considered extreme values (outliers).

# Histograms and Box Plots

