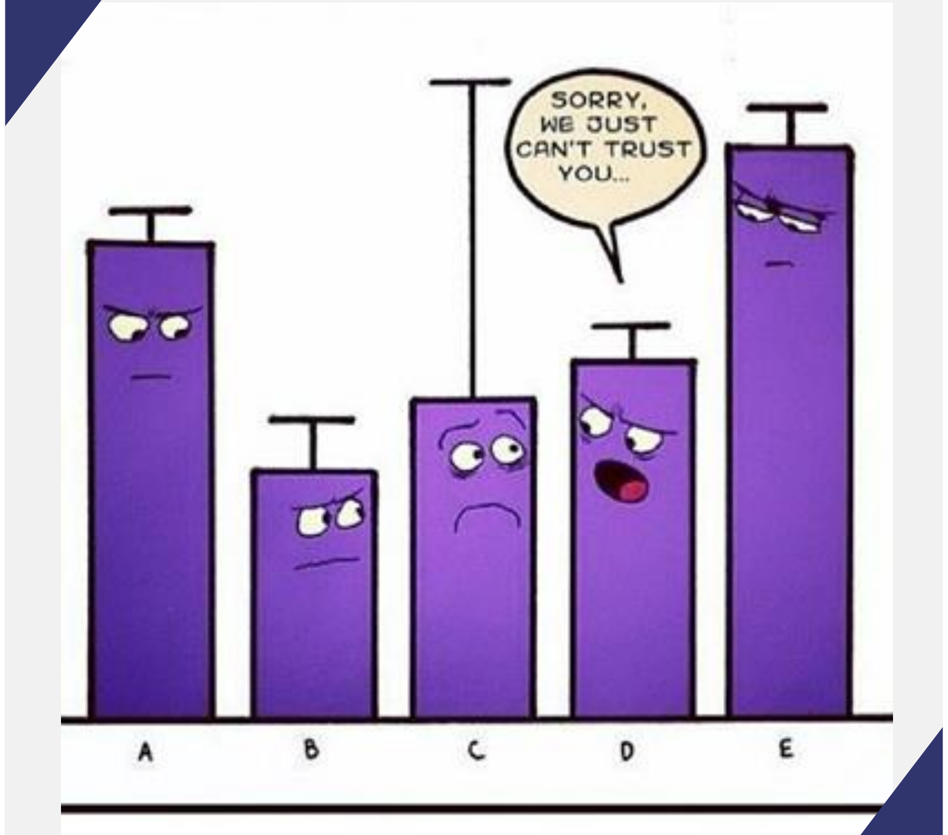
# Outlier, Noise and Missing Value

Penyusun Modul: Chairul Aulia

Editor: Rina Fitriyani, Silfa Rahma Aulia

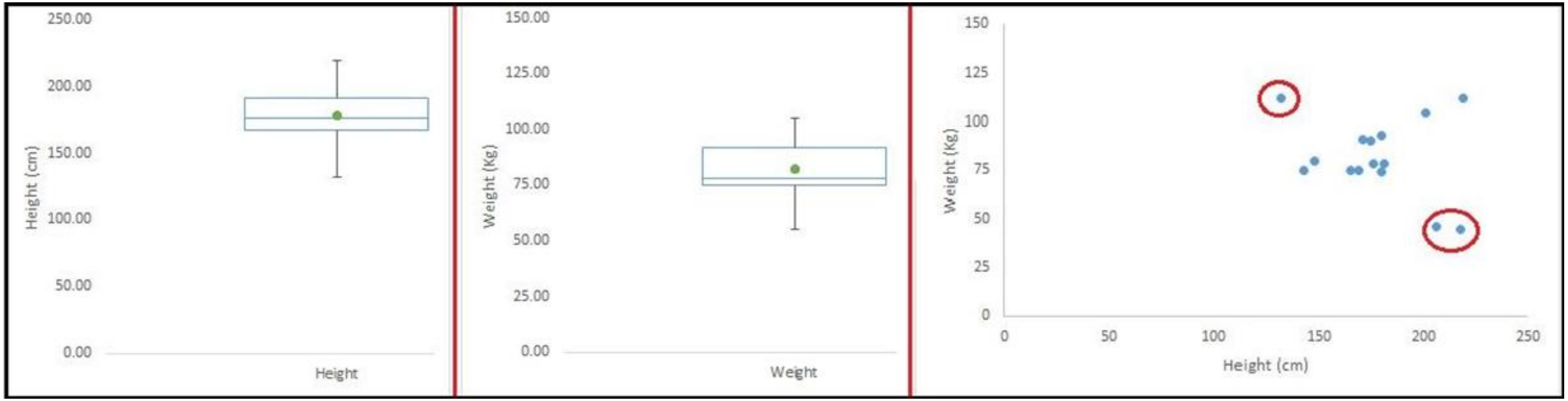
# Outlier

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.



Source : https://medium.com/

# Type of Outlier



Source : https://analyticsvidhya.com/

- **Univariate Outlier**: A univariate outlier is a data point that consists of an extreme value on one variable.
- **Multivariate Outlier**: A multivariate outlier is a combination of unusual scores on at least two variables/in an n-dimensional space

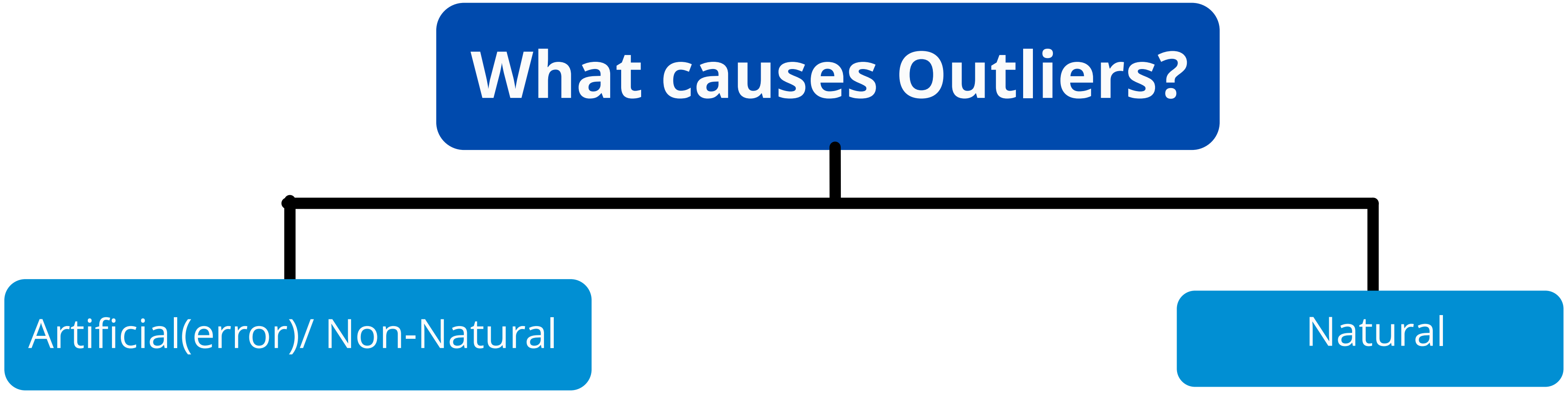
# What is the impact of Outliers on a dataset?

- It increases the error variance and reduces the power of statistical tests

- If the outliers are non-randomly distributed, they can decrease normality

- They can bias or influence estimates that may be of substantive interest.

- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.



What causes Outliers?

Artificial(error)/ Non-Natural

Natural

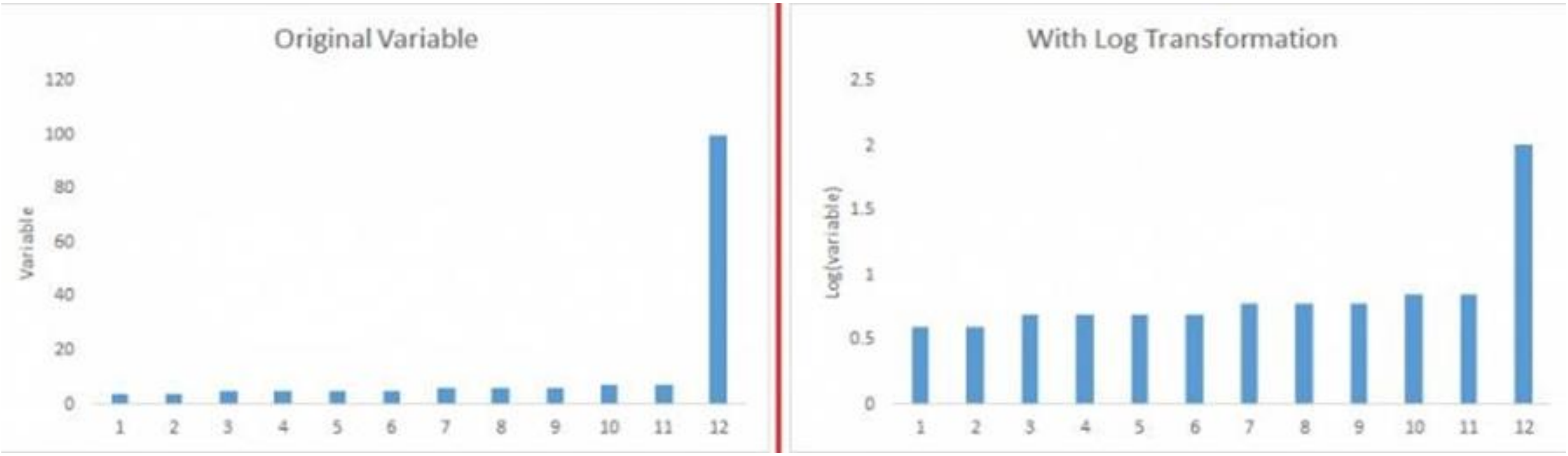
Most common causes of outliers on a data set:

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation errors)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

# How to remove the outlier?

The common techniques used to deal with outliers are:

1. Deleting observations
2. Transforming and binning values



Source : https://medium.com/

3. Imputing
4. Treat Outliers separately.

# As usual, let's try it out!

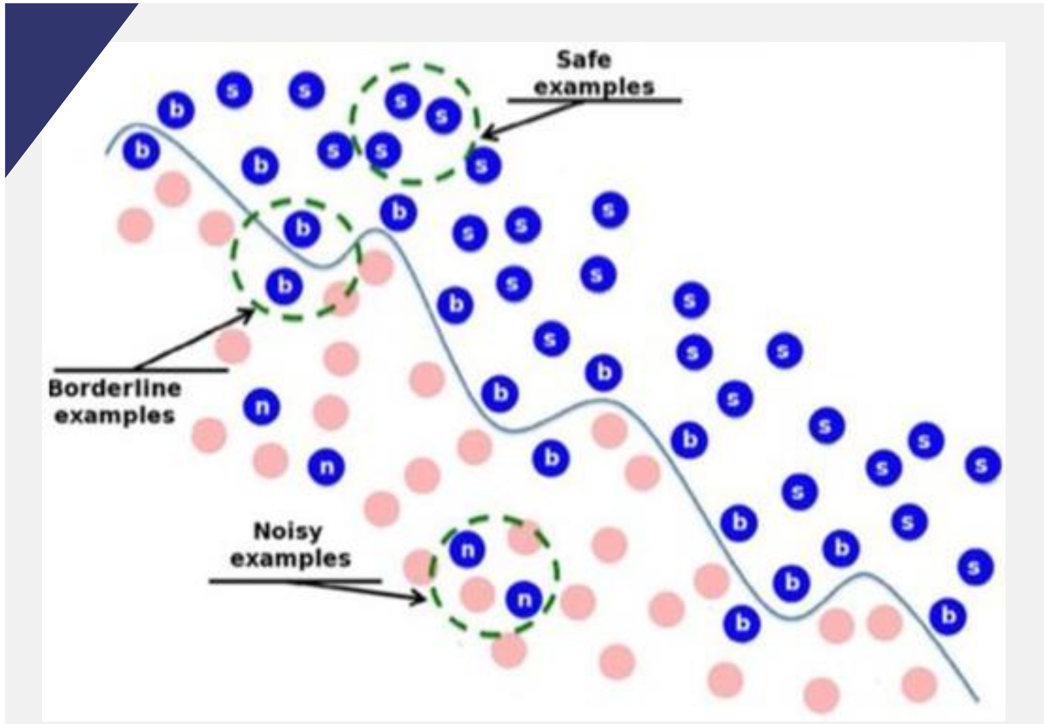
Let’s try to detect and remove outliers

Open the `Outliers` notebook file on JupyterLab

# Exercise

Now, try doing the same thing with the winequality-white dataset

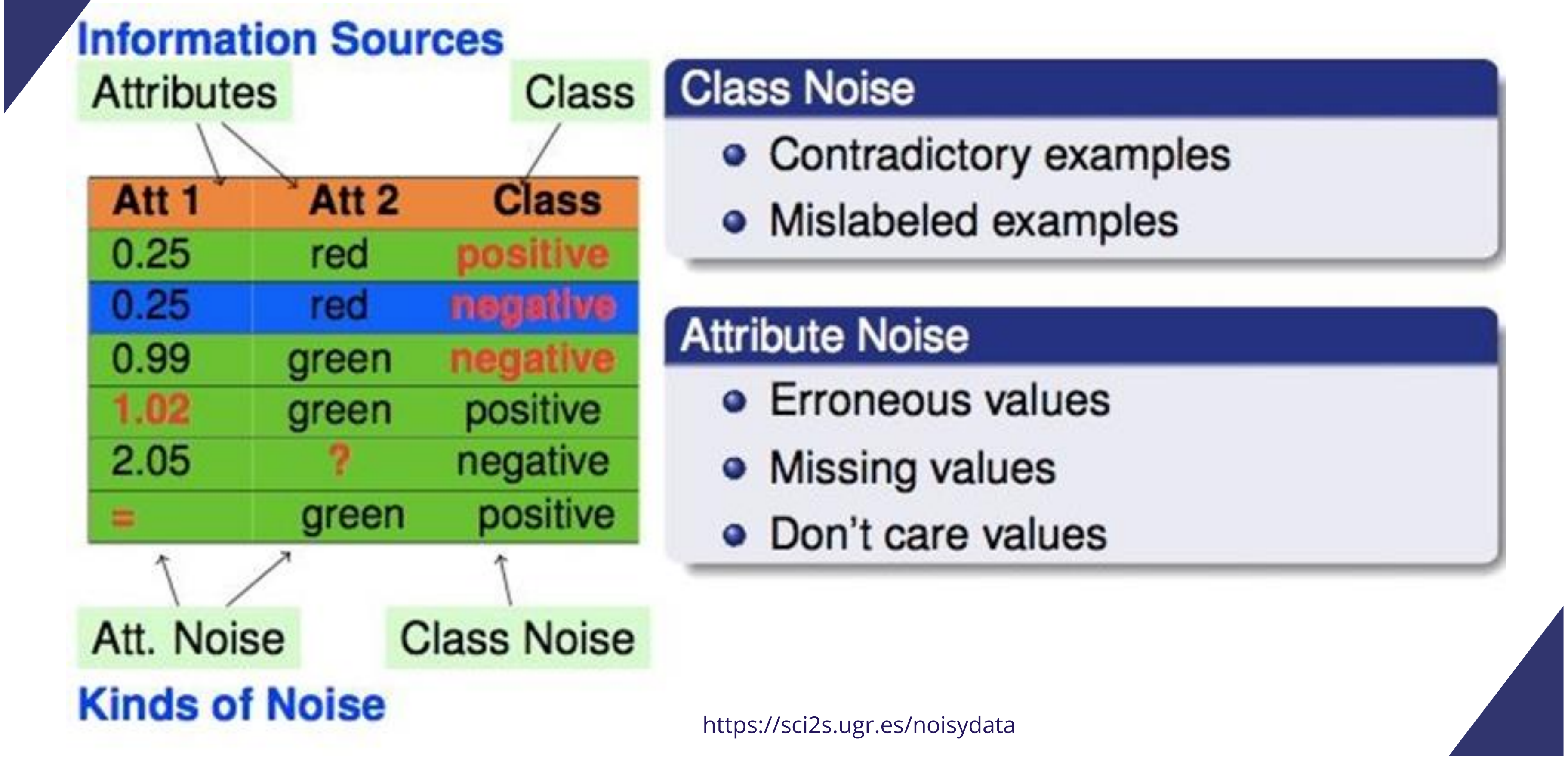
# Data Noise



https://sci2s.ugr.es/noisydata

Noisy data is data with a large amount of additional meaningless information in it called noise. This includes corrupted data. It also includes any data that a user system cannot understand and interpret correctly.

# Noise Types



**Information Sources**

| Att 1 | Att 2 | Class |
|---|---|---|
| 0.25 | red | positive |
| 0.25 | red | negative |
| 0.99 | green | negative |
| 1.02 | green | positive |
| 2.05 | ? | negative |
| = | green | positive |

**Kinds of Noise**

**Class Noise**

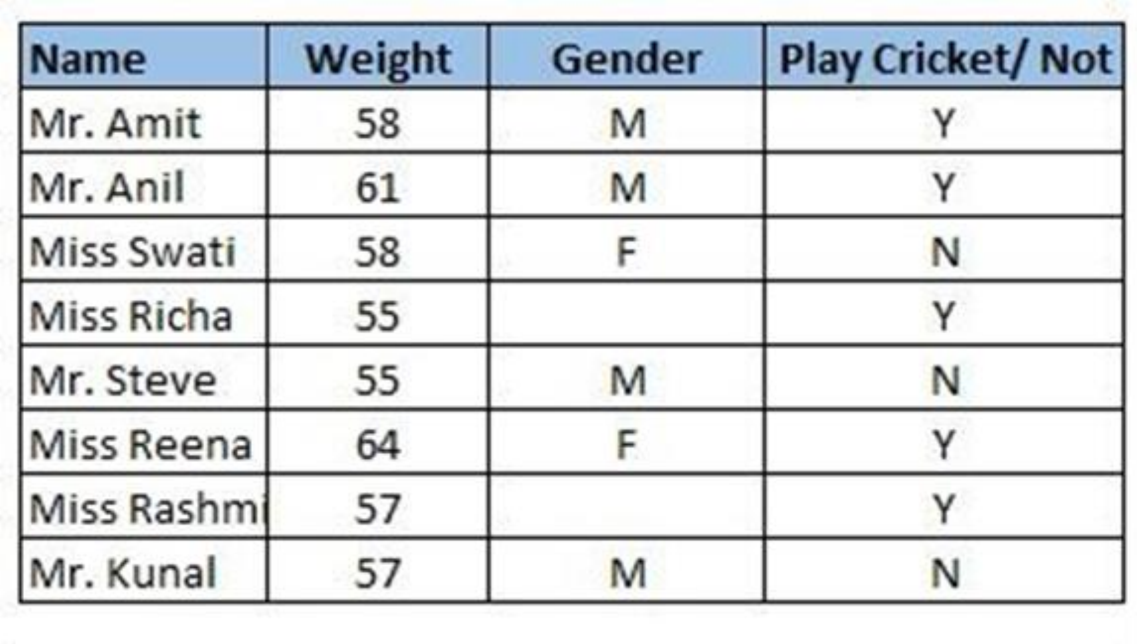
- Contradictory examples
- Mislabeled examples

**Attribute Noise**

- Erroneous values
- Missing values
- Don't care values

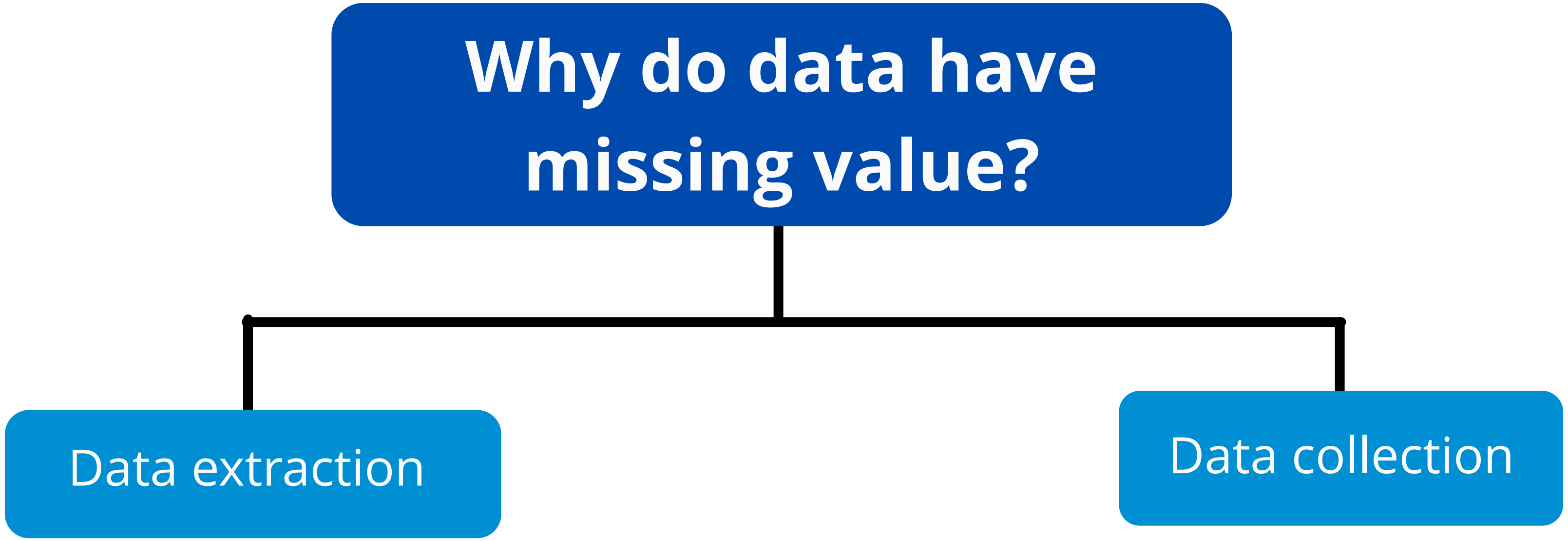
https://sci2s.ugr.es/noisydata

# Missing Value



| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 |  | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashmi | 57 |  | Y |
| Mr. Kunal | 57 | M | N |

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.



# Why do data have missing value?

Data extraction

Data collection

Missing completely at random

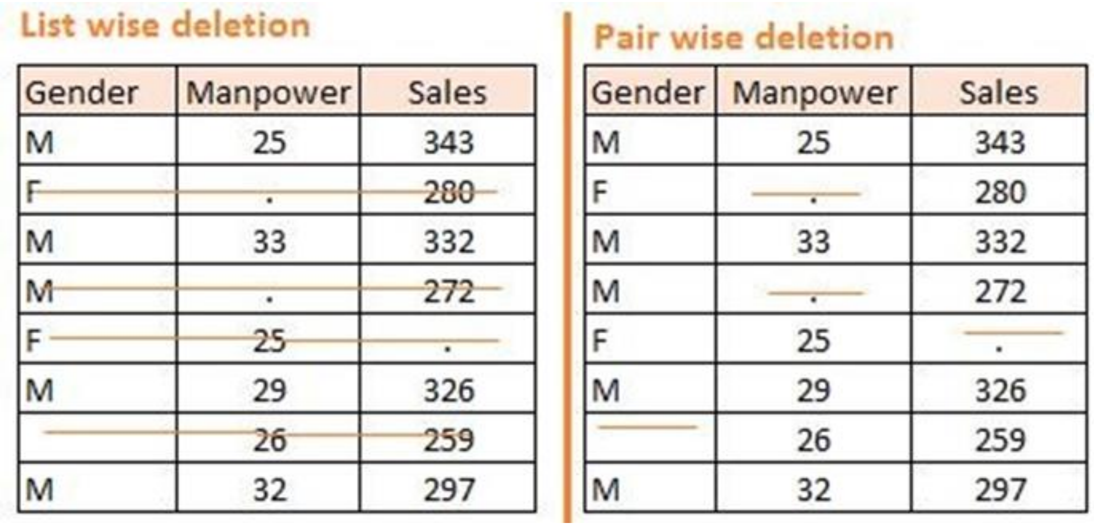
Missing at random

Missing that depends on unobserved predictors

Missing that depends on the missing value itself

# Which are the methods to treat missing values ?

1. Deletion



List wise deletion

| Gender | Manpower | Sales |
|---|---|---|
| M | 25 | 343 |
| ~~F~~ | ~~.~~ | ~~280~~ |
| M | 33 | 332 |
| ~~M~~ | ~~.~~ | ~~272~~ |
| ~~F~~ | ~~25~~ | ~~.~~ |
| M | 29 | 326 |
|  | ~~26~~ | ~~259~~ |
| M | 32 | 297 |

Pair wise deletion

| Gender | Manpower | Sales |
|---|---|---|
| M | 25 | 343 |
| F | ~~.~~ | 280 |
| M | 33 | 332 |
| M | ~~.~~ | 272 |
| F | 25 | ~~.~~ |
| M | 29 | 326 |
|  | 26 | 259 |
| M | 32 | 297 |

2. Mean/ Mode/ Median Imputation
    - Generalized Imputation
    - Similar case Imputation
3. Prediction Model
4. KNN Imputation

# As Usual

## Let's Try it Out!

Let’s try to detect and remove null values

Open the `Titanic_Statistics` notebook file on JupyterLab and head to the Working with null values section

# Let's EXERCISE!

Now, try doing the same thing with the `winequality-white` dataset

Please detect and remove outliers from a variable, you can choose one variable freely

Or if you feel that you can do all of the variable, then just do all of it

# Exploratory Data Analysis

Now that you've learn the fundamentals of Exploratory Data Analysis, how about we go take an example of how it is fully used as a whole?

Open and explore the `White_Wine_EDA` file on your JupyterLab