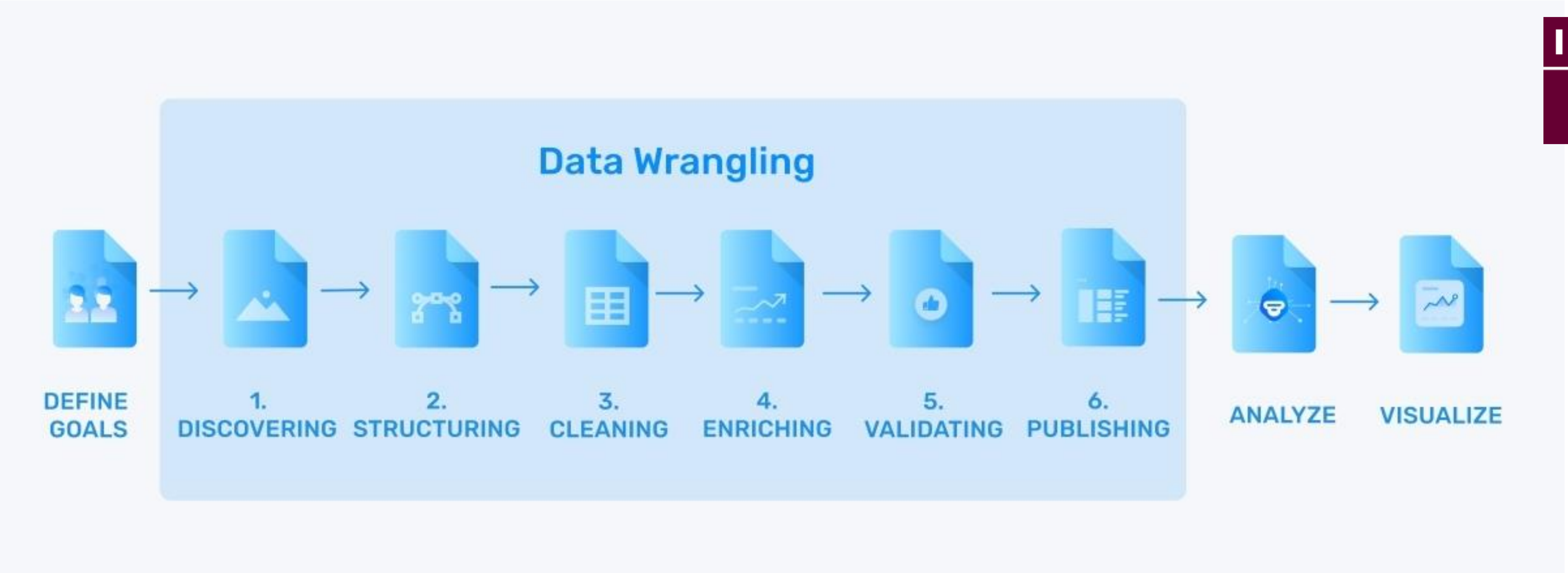
Intro to Data Wrangling and EDA

CS 459 Business Intelligence

Data Wrangling

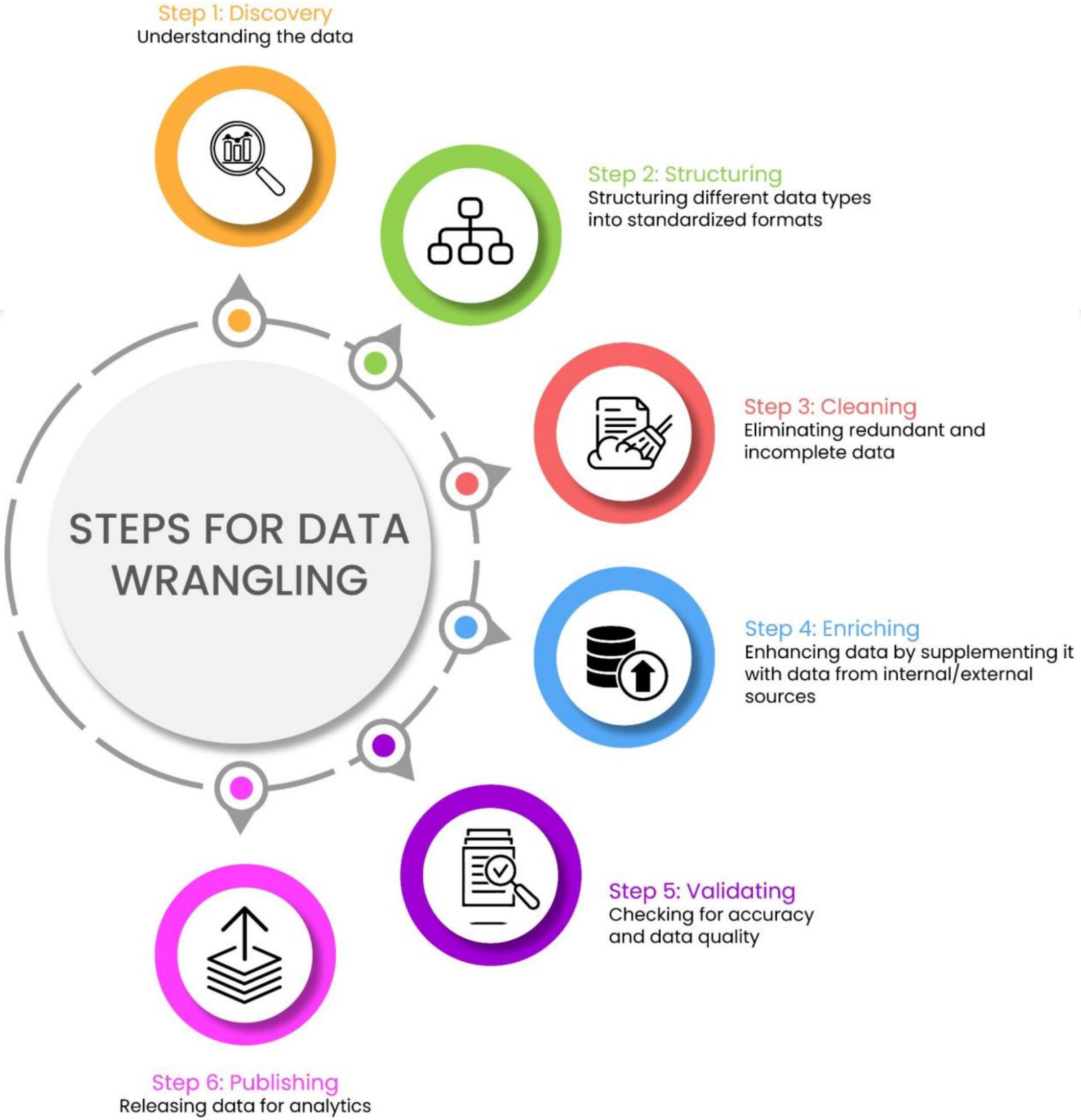


Data Wrangling

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*

Summarizing 6-steps of Data Wrangling



Step 1: Discovery
Understanding the data

Step 2: Structuring
Structuring different data types into standardized formats

Step 3: Cleaning
Eliminating redundant and incomplete data

Step 4: Enriching
Enhancing data by supplementing it with data from internal/external sources

Step 5: Validating
Checking for accuracy and data quality

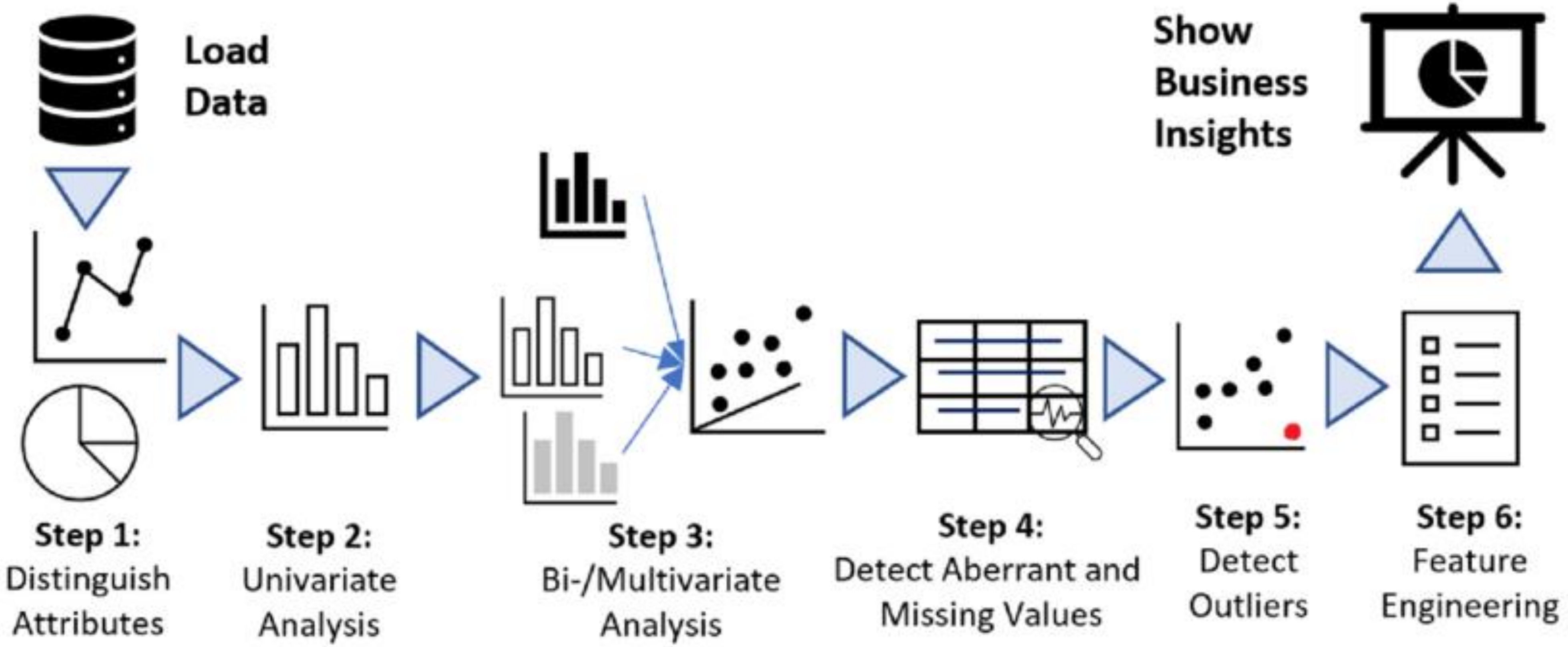
Step 6: Publishing
Releasing data for analytics

Exploratory Data Analysis (EDA)

Exploratory **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.

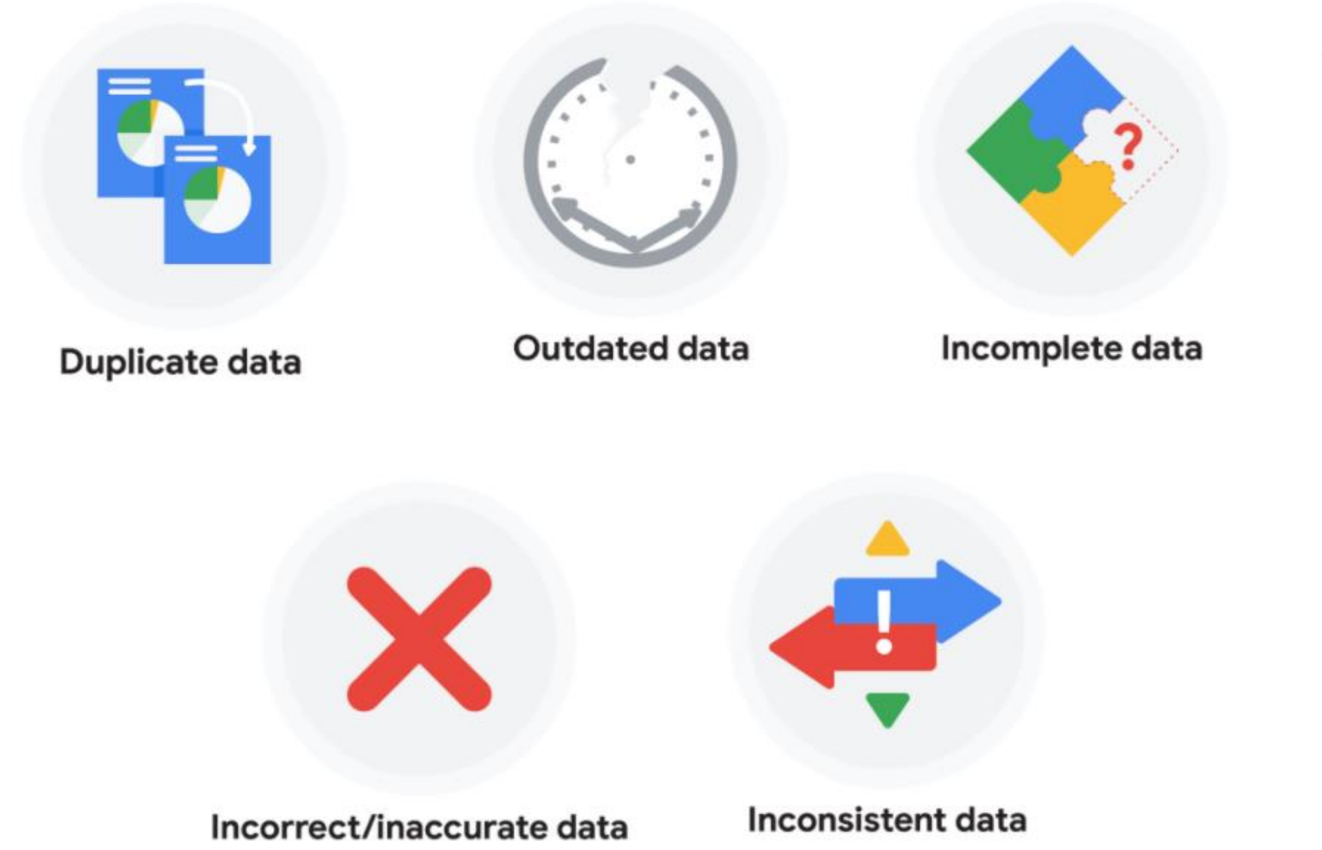
EDA


Load
Data
Show
Business
Insights
Step 1:
Distinguish
Attributes
Step 2:
Univariate
Analysis
Step 3:
Bi-/Multivariate
Analysis
Step 4:
Detect Aberrant and
Missing Values
Step 5:
Detect
Outliers
Step 6:
Feature
Engineering

Data Wrangling

Data Cleaning

Types of dirty data



Duplicate data

Outdated data

Incomplete data

Incorrect/inaccurate data

Inconsistent 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.

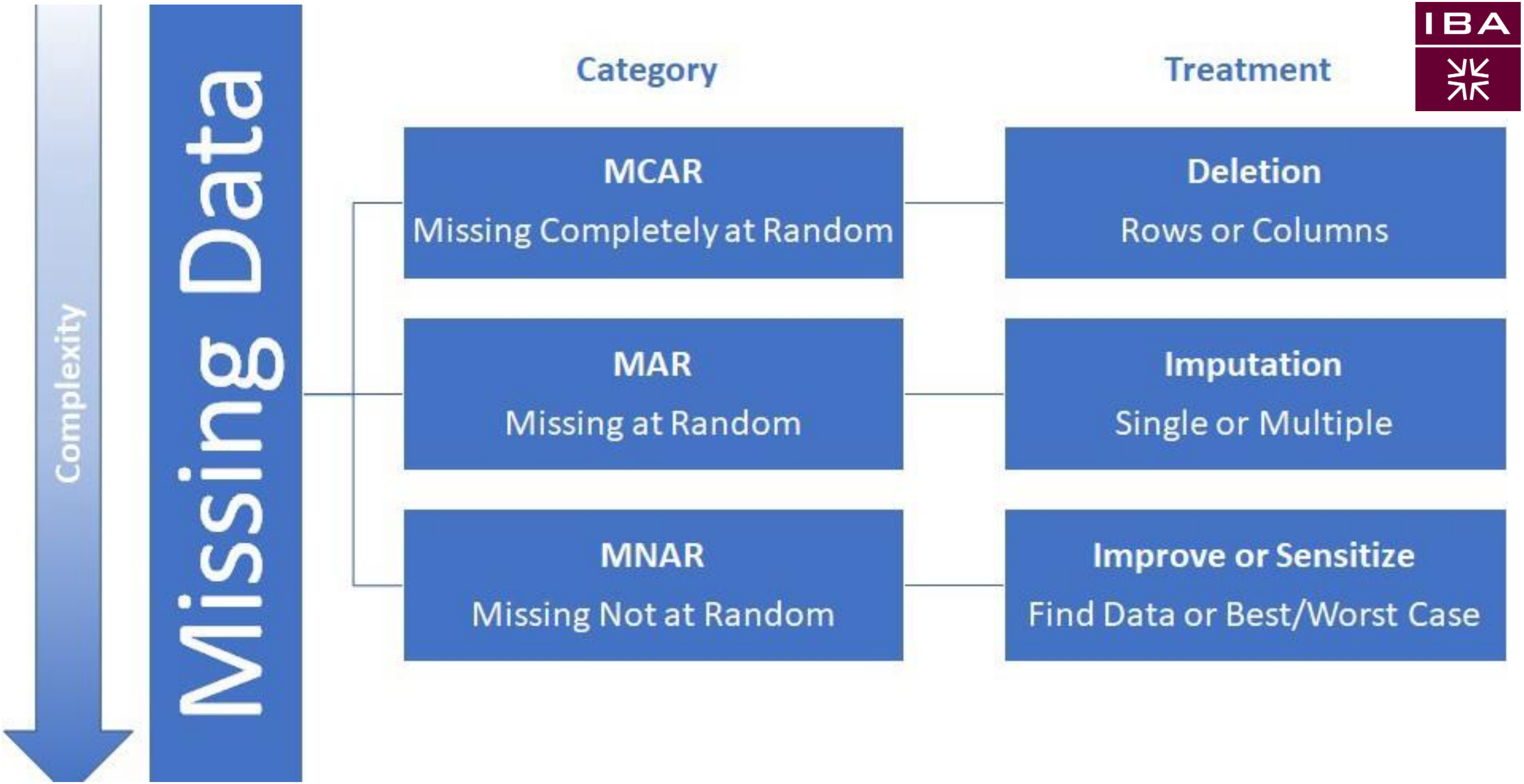
Missing Values
Missing at Random (MAR)

- 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 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.

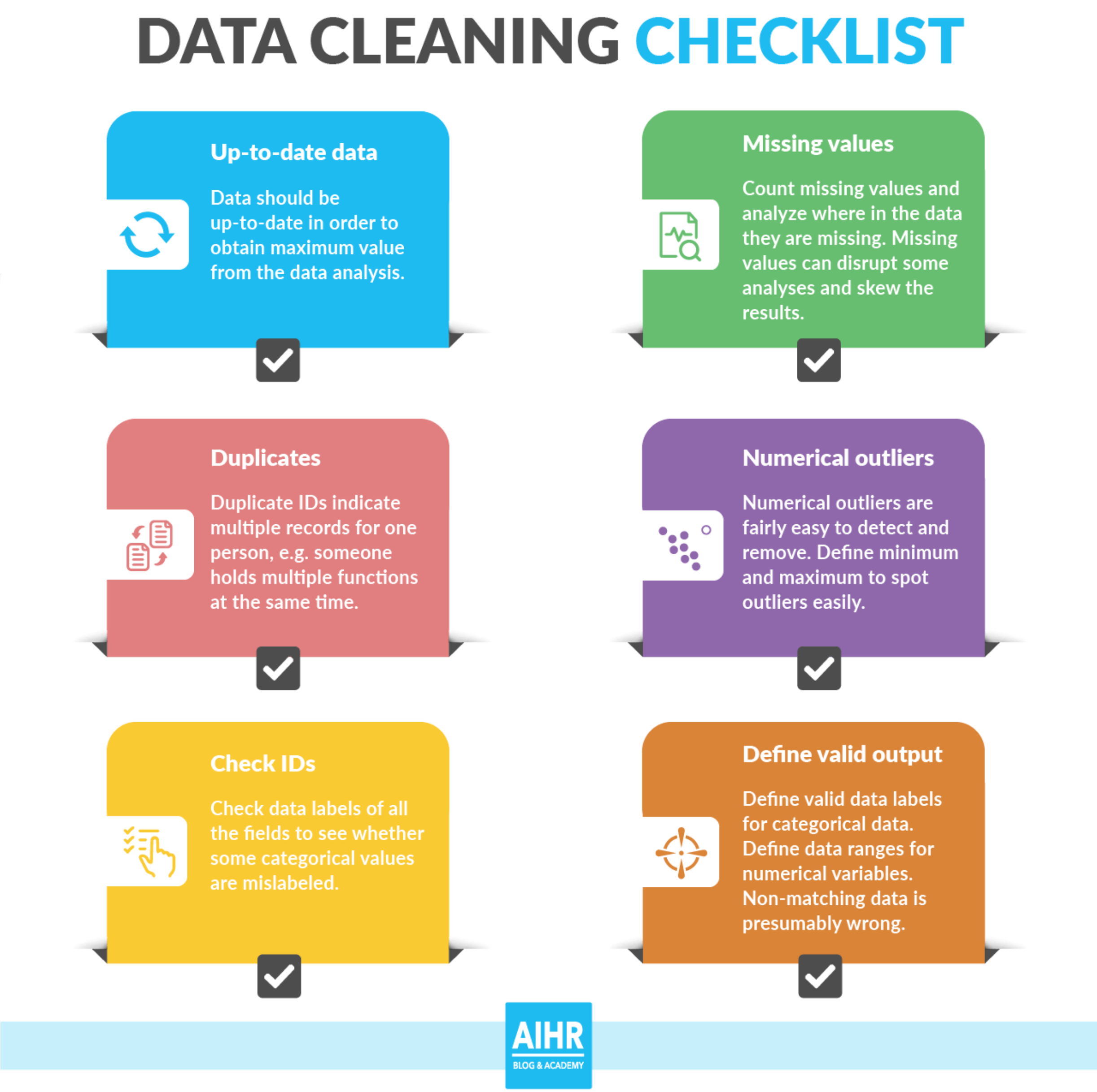


Complexity

Missing Data

Category	Treatment
MCAR Missing Completely at Random	**Deletion** Rows or Columns
MAR Missing at Random	**Imputation** Single or Multiple
MNAR Missing Not at Random	**Improve or Sensitize** Find Data or Best/Worst Case

Data Cleaning



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.

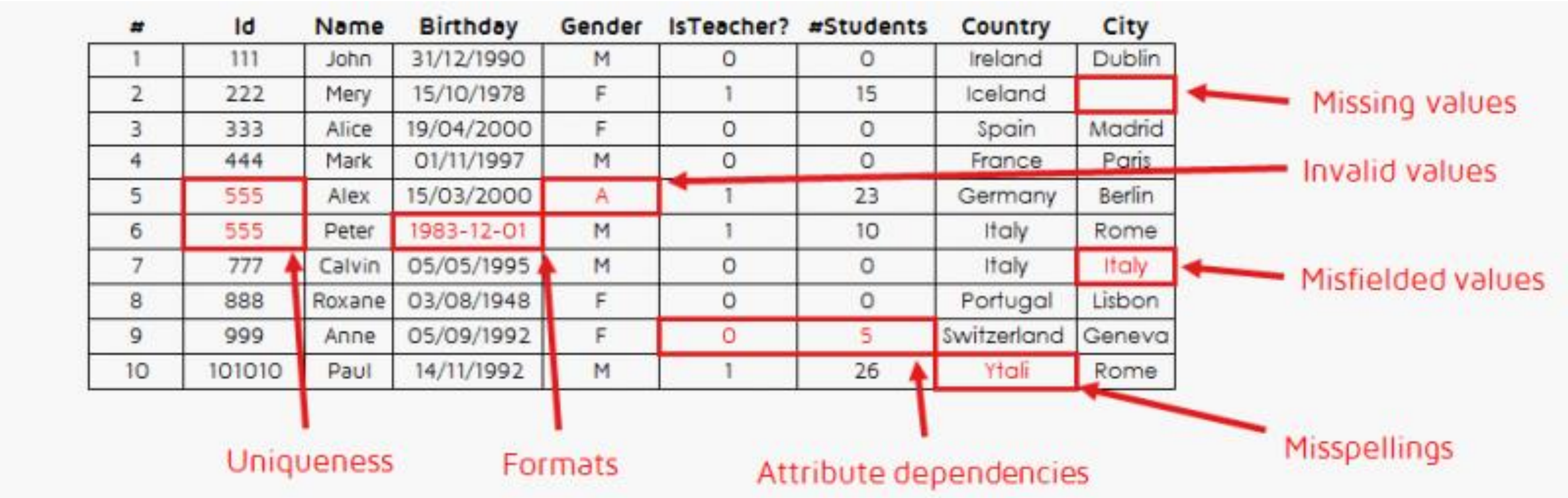
Check IDs

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.

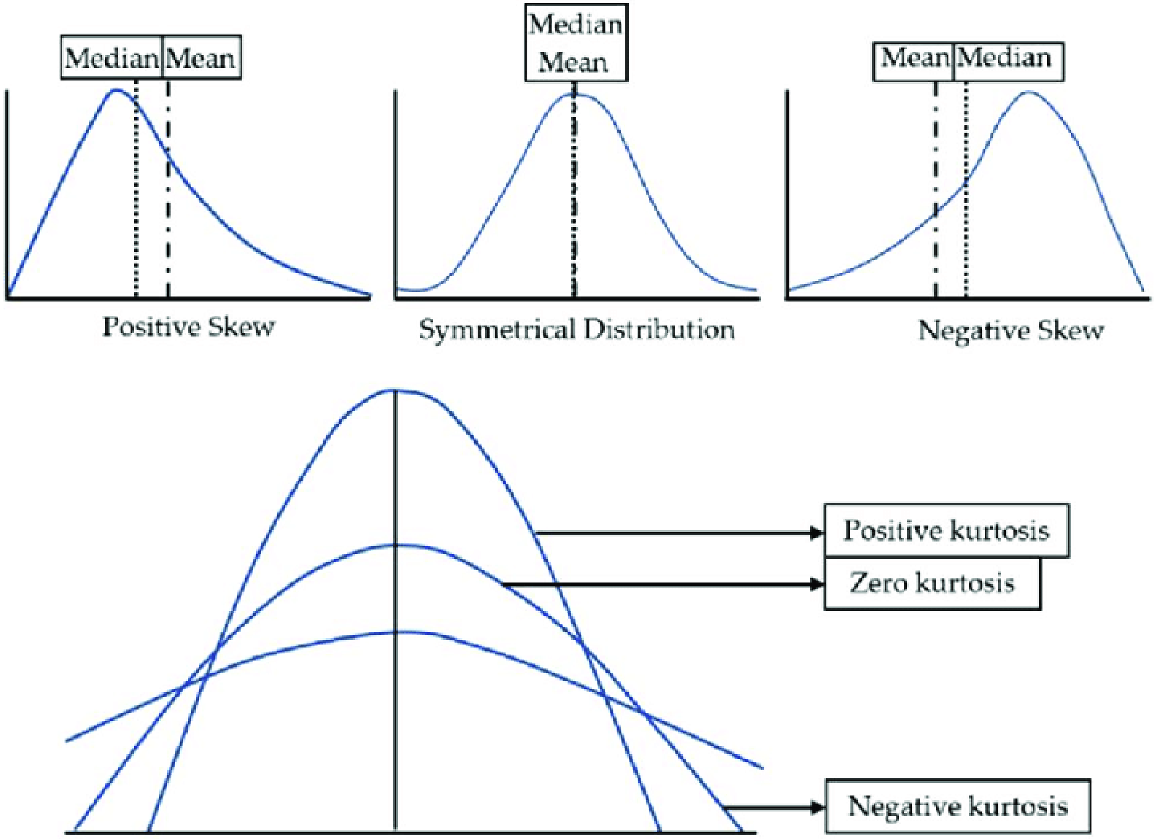
Problems with the Data



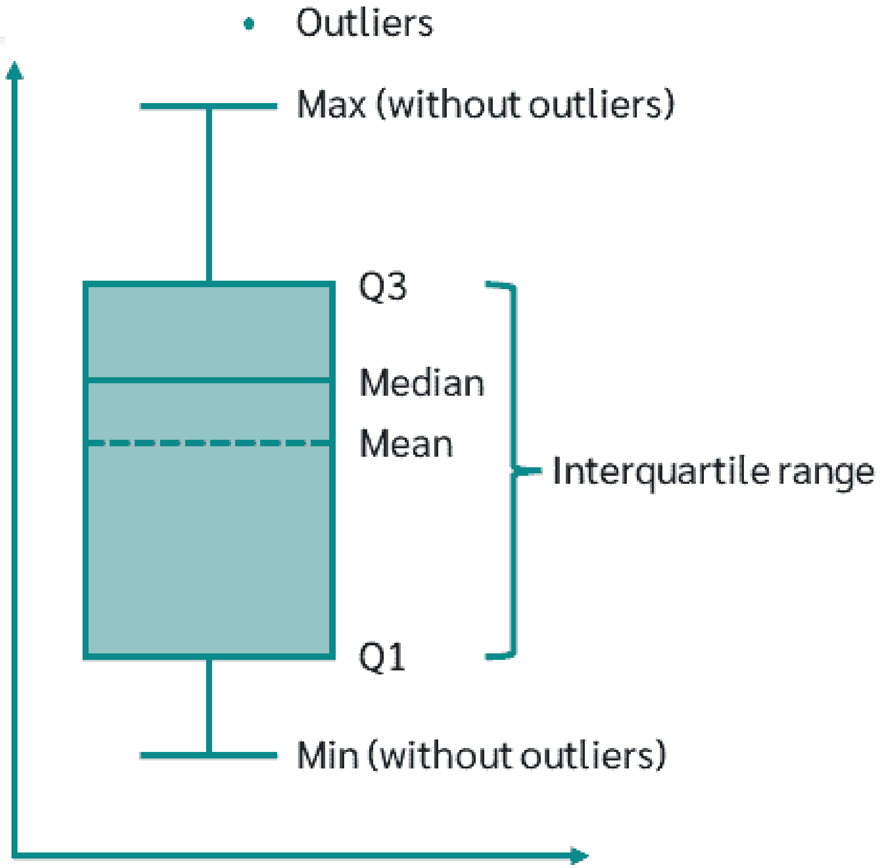
#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	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

Interpreting Histograms and Box plots

Analyzing Histograms: Shape, Skew and Kurtosis



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

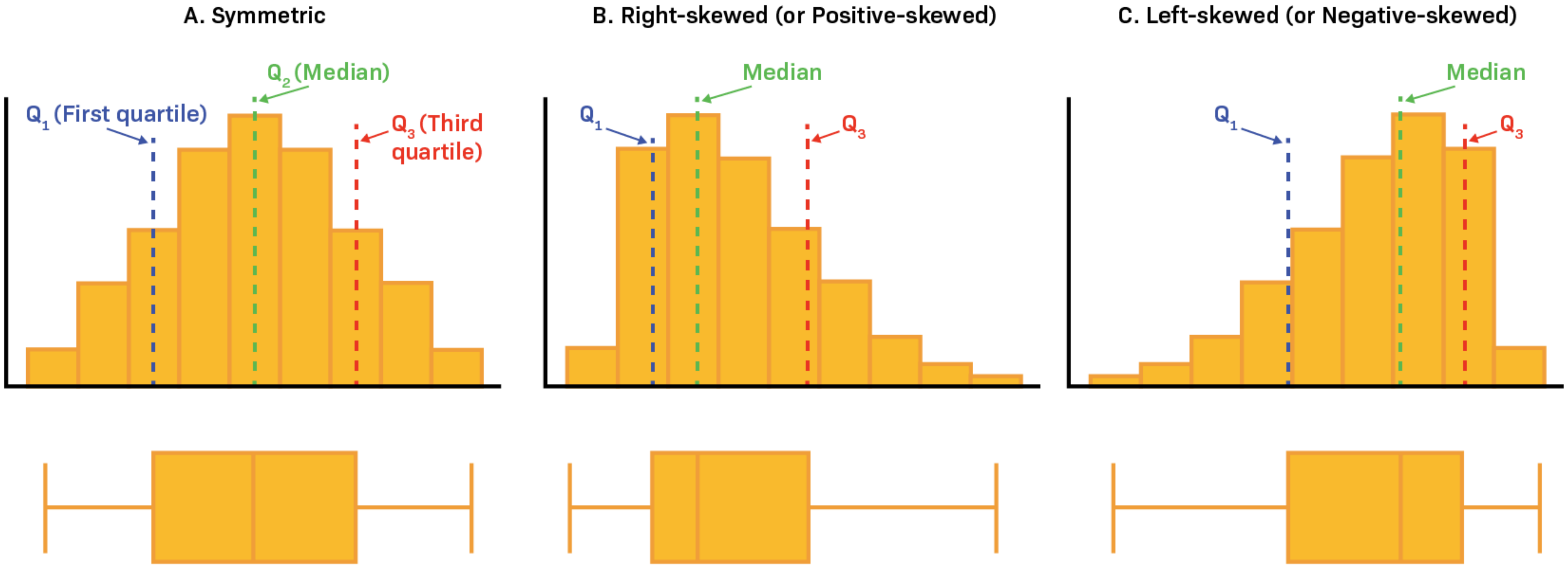
Between Q1 and Q3, is the interquartile range

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).

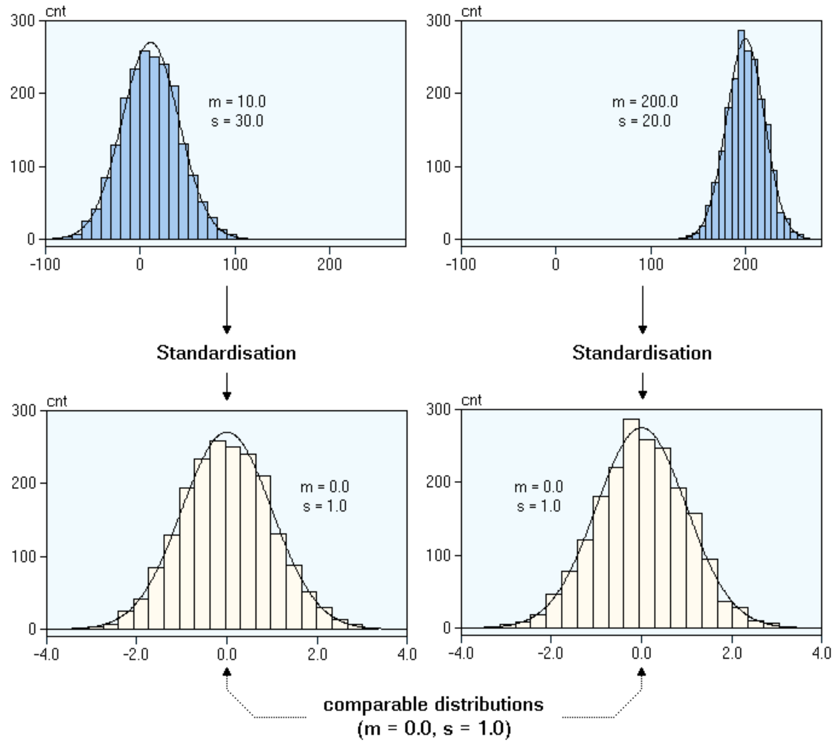
Histograms and Box Plots

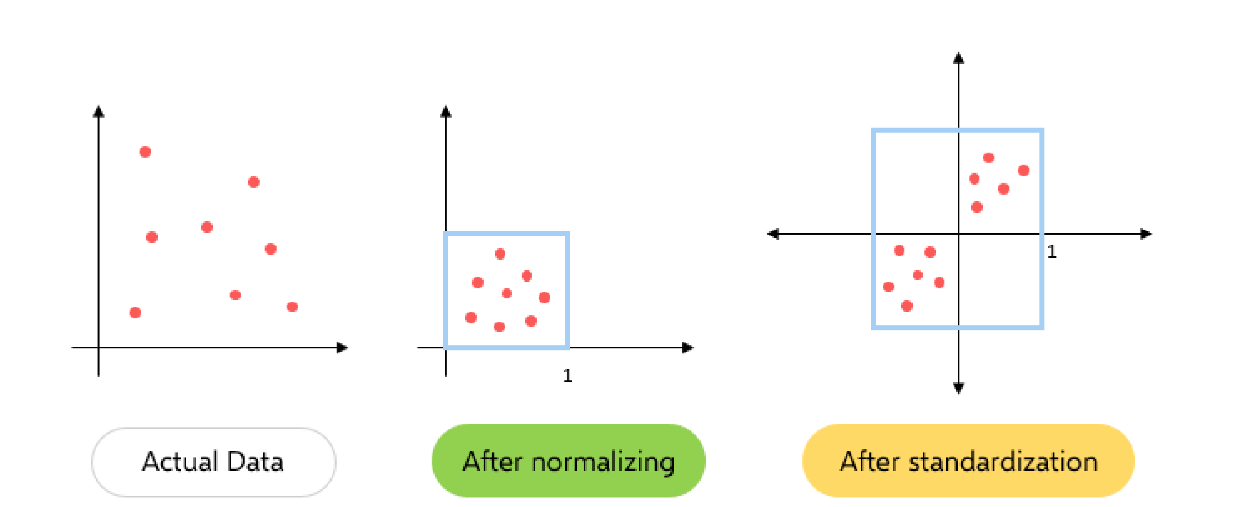


Wrangling Techniques

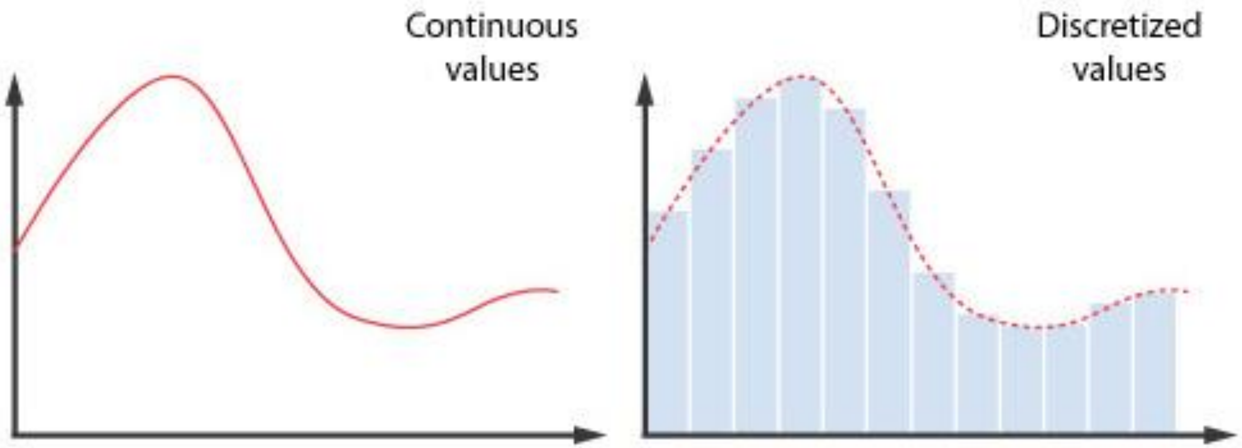
Standardization vs Normalization

- Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
- Normalization typically means rescales the values into a range of [0,1] or [-1,1].





Discretization



- Discretization is the process through which we can transform continuous variables, models or functions into a discrete form.
- For categorical variables to reduce the number of possible groups.

Example – Price of commonly sold products

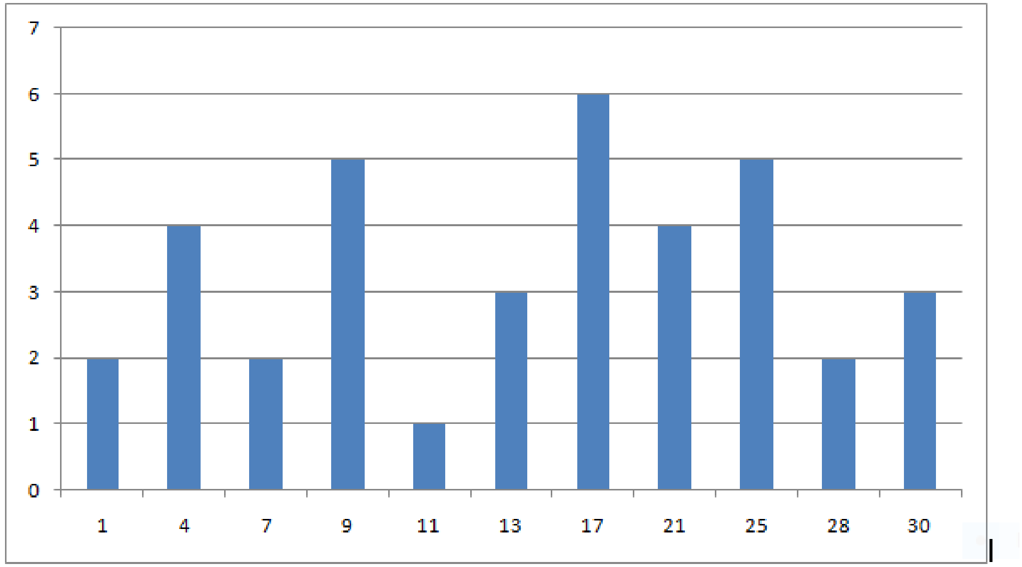


Figure 1 Histogram using price where one bucket represents one value

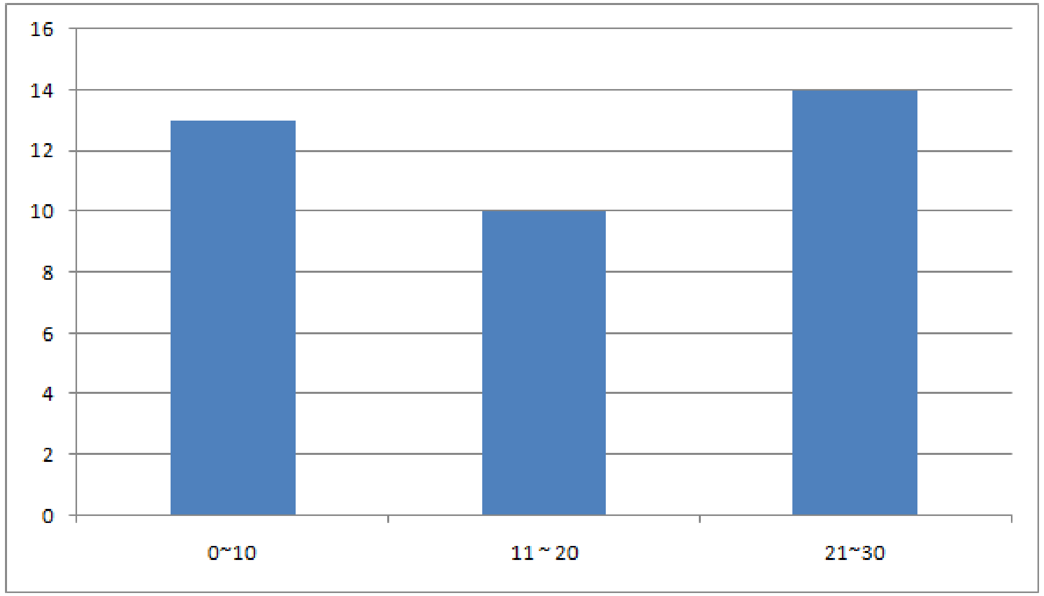
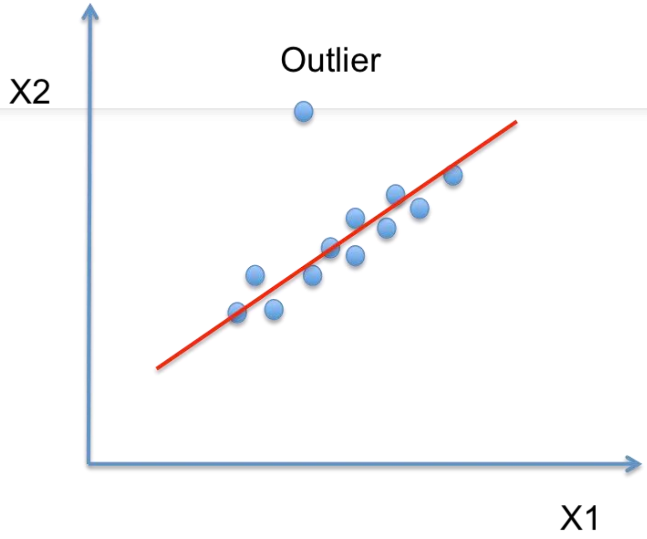
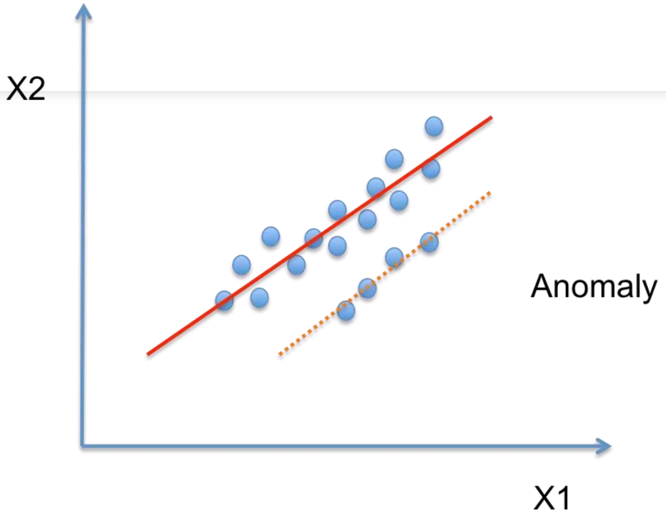


Figure 2 : Equal width Histogram

Outlier Analysis

Outliers Vs Anomalies

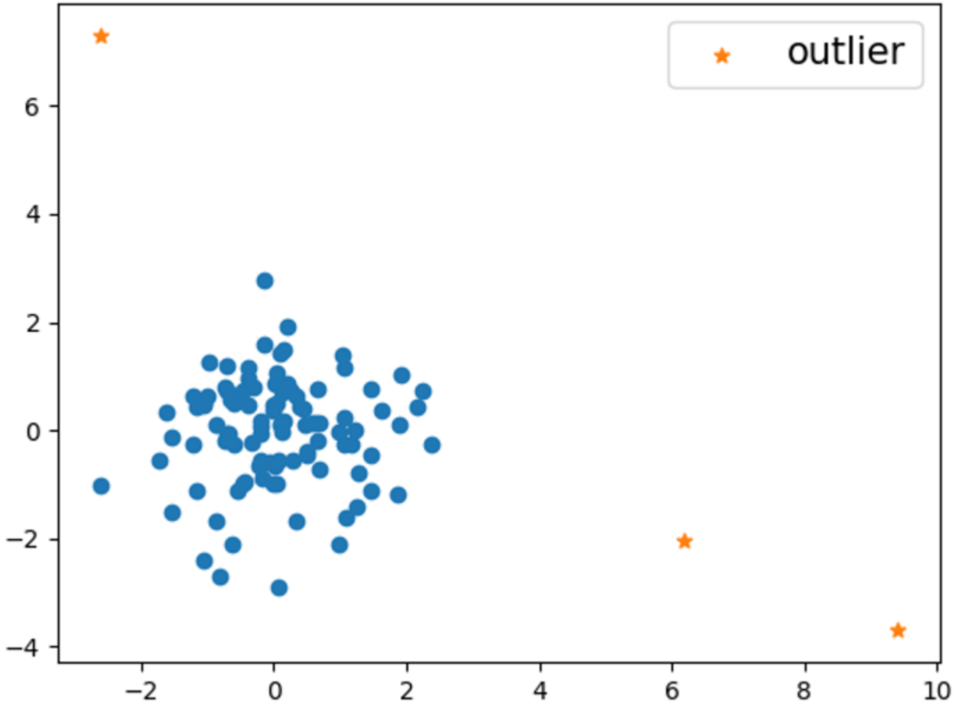




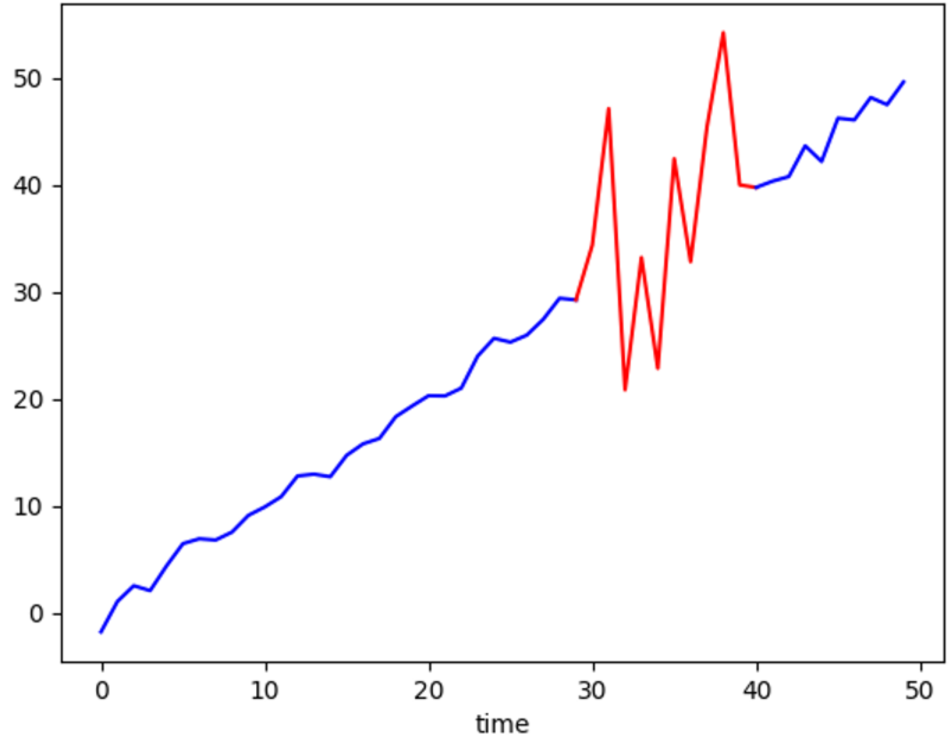
Outlier is usually a single observation, which is extreme from "Median" and can fall on either side of it.

Anomalies are observations (usually more than one) where they *don't confirm to pattern* exhibited by certain variable.

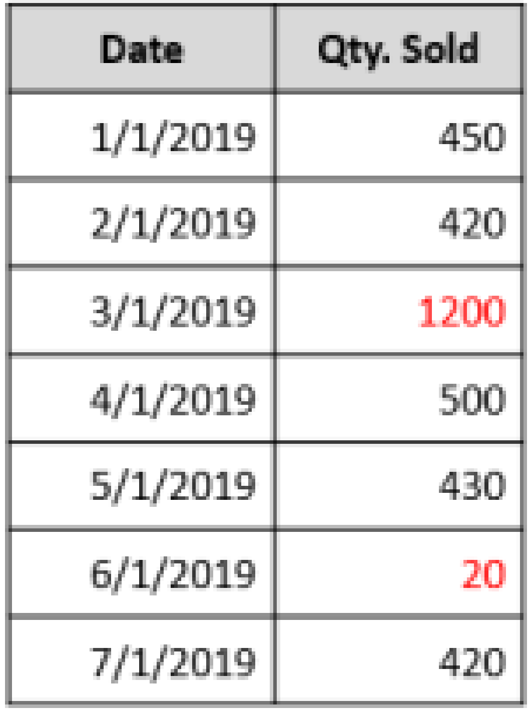
Outlier



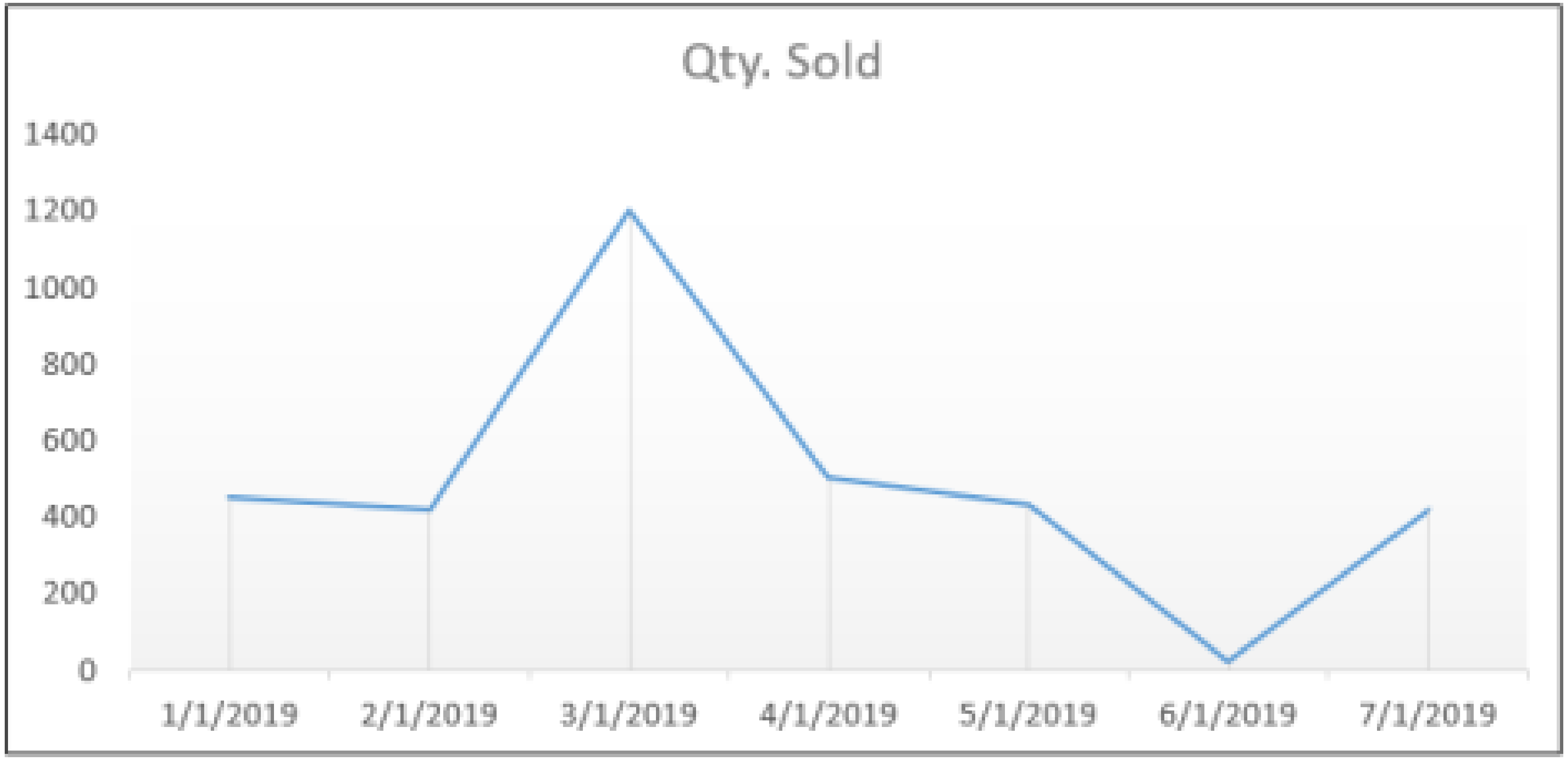
Anomaly



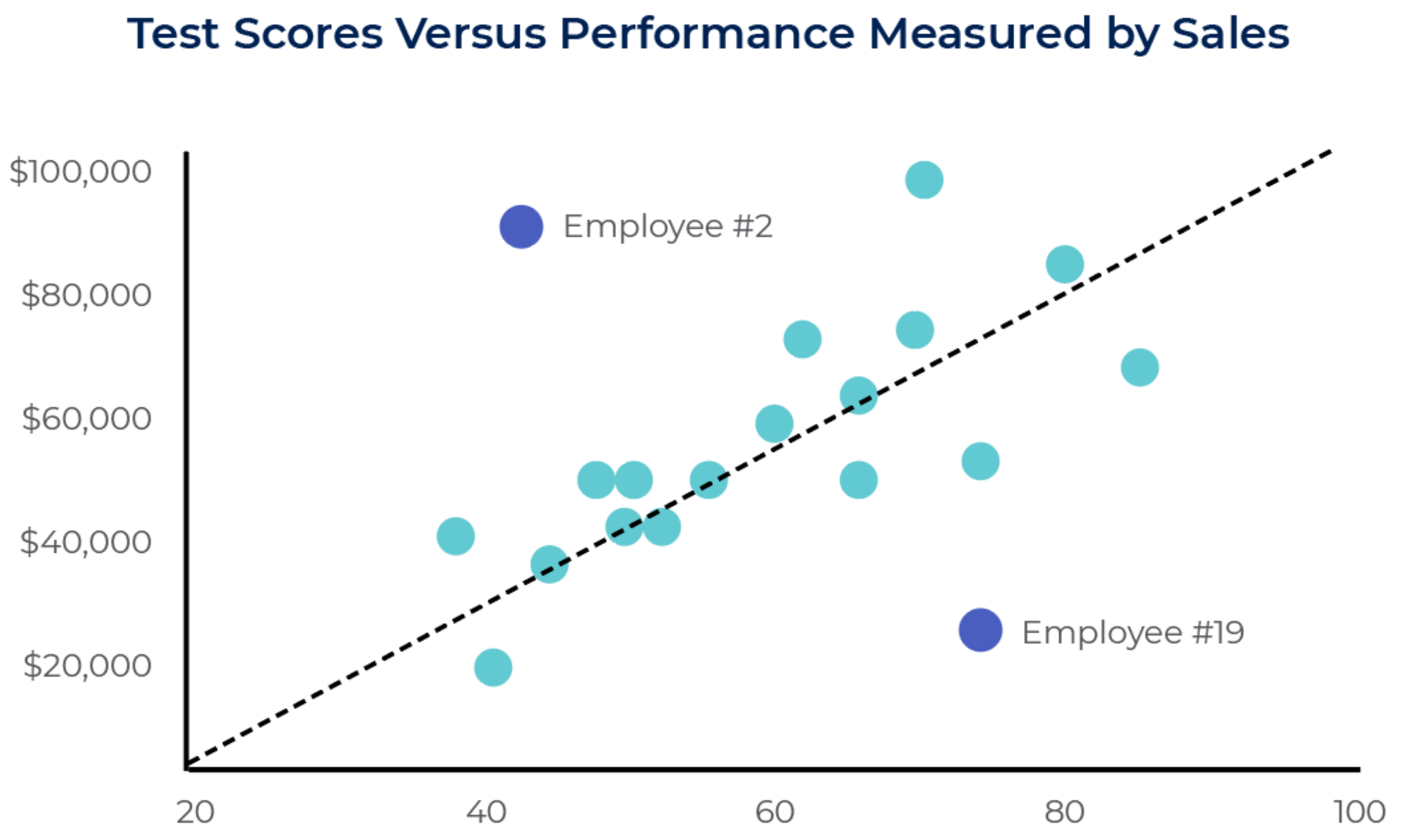
Outliers- Example



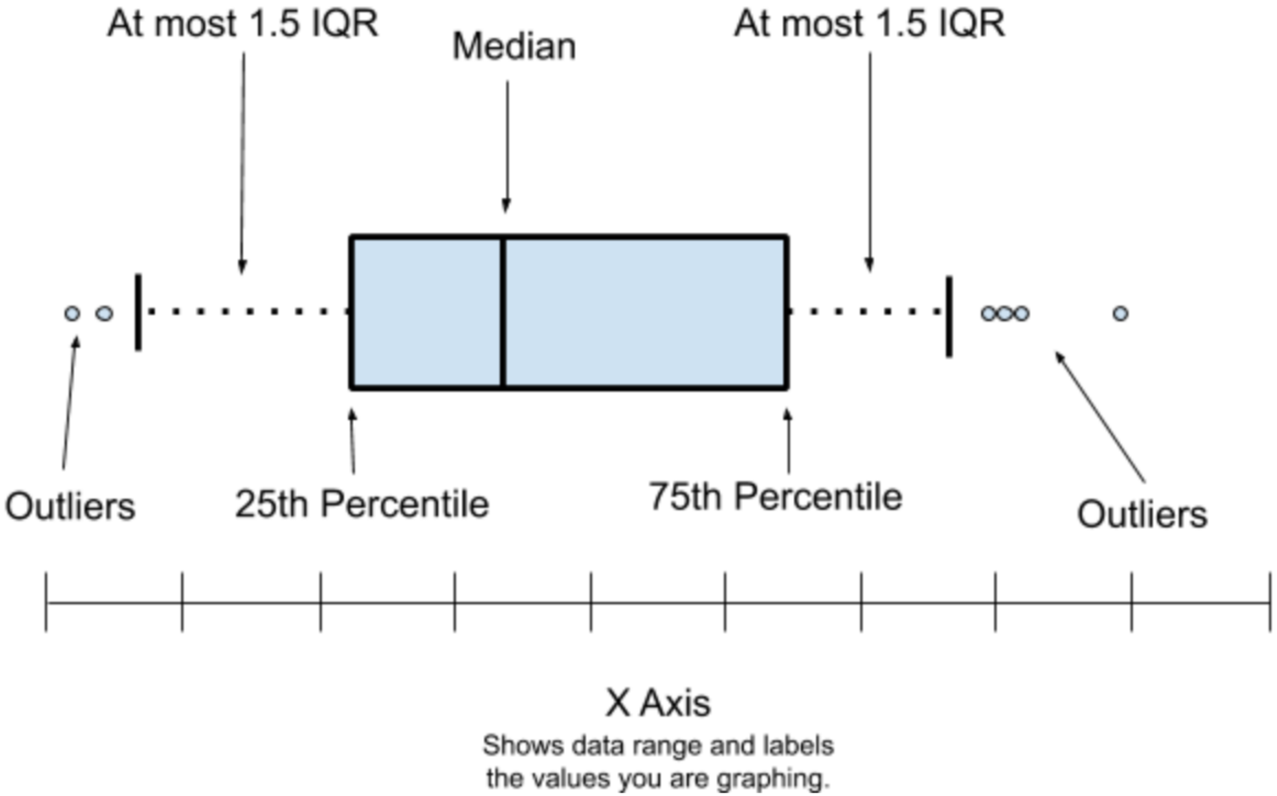
Date	Qty. Sold
1/1/2019	450
2/1/2019	420
3/1/2019	1200
4/1/2019	500
5/1/2019	430
6/1/2019	20
7/1/2019	420



Outliers- Example



Outliers with Box Plots



Outliers

Outliers in data may contain valuable information.

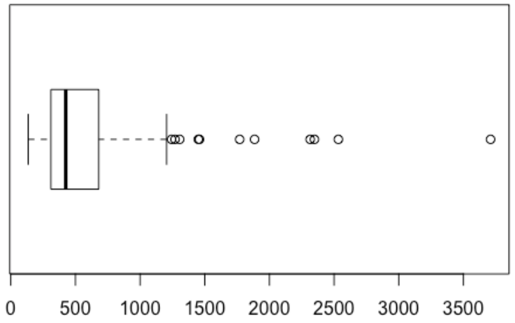
Or be meaningless aberrations caused by measurement and recording/data entry errors. E,g , not converting weight, a typo in sales value with an extra zero.

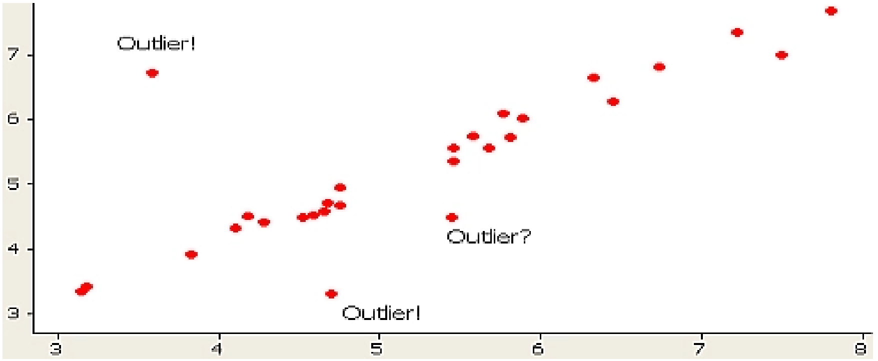
Investigate why are they occurring? Where—and what—might the meaning be?

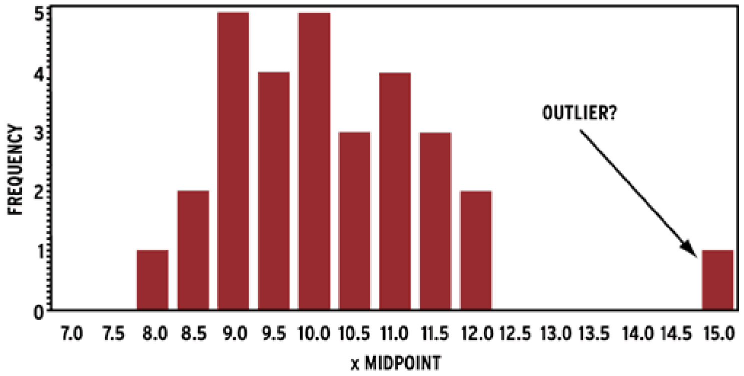
The answer could differ from business to business, but it’s important to have the conversation rather than ignoring the data.

Outliers Testing and Visualization

- Visualization : Boxplot and the scatterplot







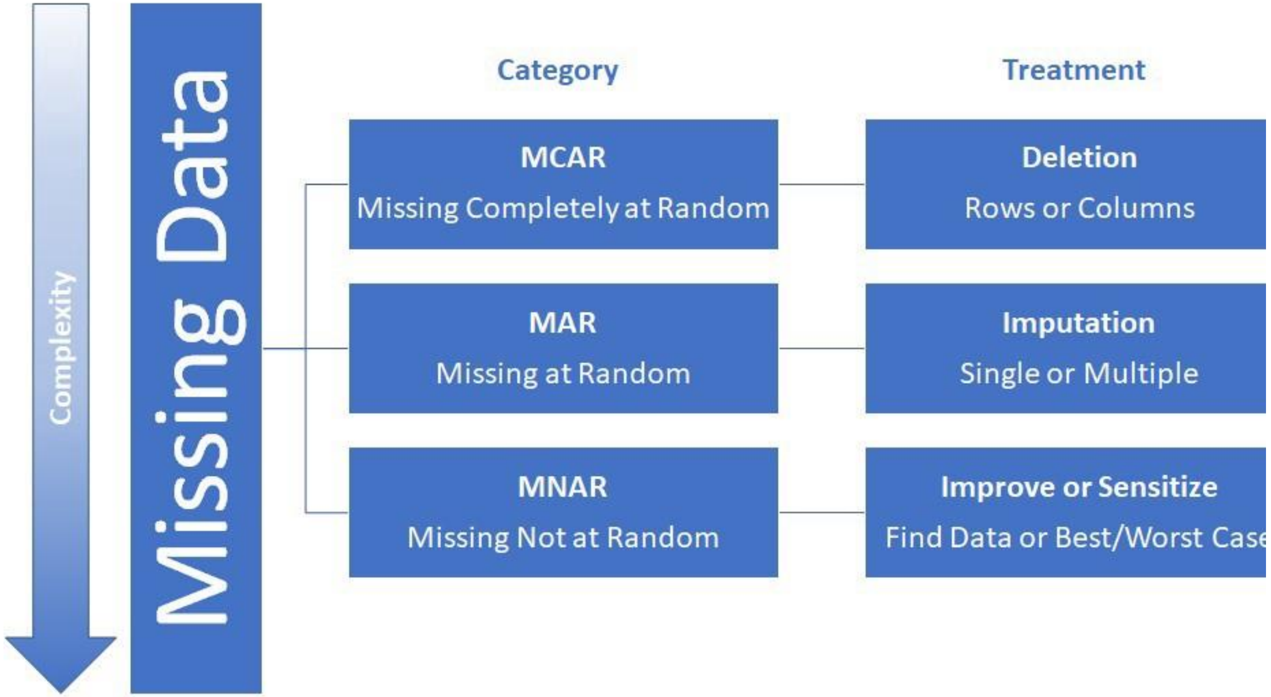
- The **Tietjen-Moore** test is useful for determining multiple outliers in a data set with the null hypothesis for this test is – there are no outliers in the data.

What should I do with outliers?

- Much dependent on the business needs.
- A good BI dashboard should be able to detect outliers for the right decision making at the right time.
- Outlier Detection is important, treatment is dependent on the requirements of analysis.
- Removal/Imputation may become important when it is essential to have a normal distribution for some statistical testing or machine learning algorithms.

Missing Value Analysis (MVA)

Missing Values



- Missing values are usually represented in the form of **NaN** or **Null** or **None** in the dataset.

Dropping Rows and Columns

- Data not in use →Not useful for your analysis
- Contains the same value (with missing values or not)
- Very few rows with missing values in comparison to the full size of the dataset and information in multiple columns is missing.
- Use this method in extreme cases when there are too many null values in the column or row.
- *Tradeoff: Loss of information.*

Imputation

NUMERICAL

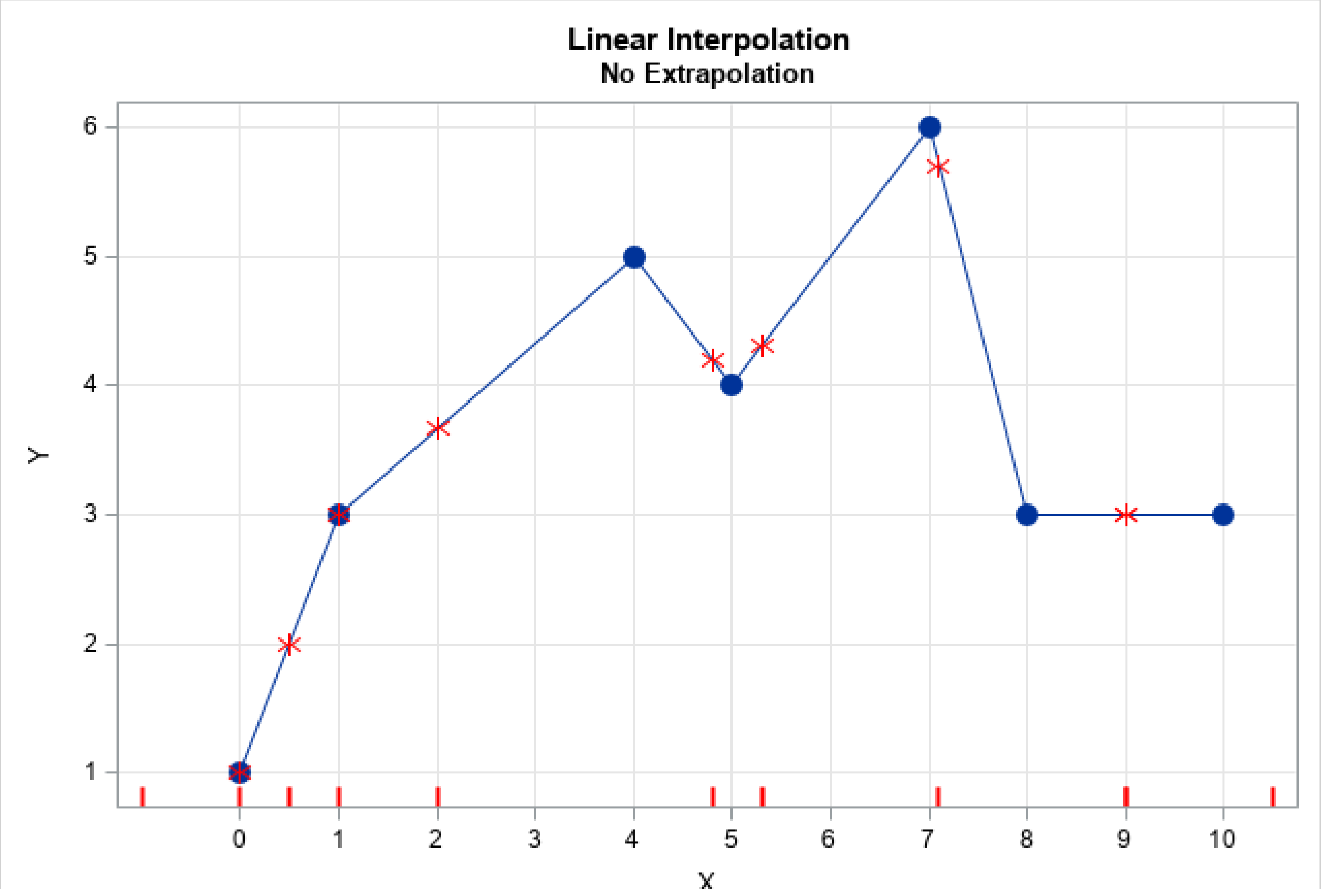
1. Filling the missing data with the **mean**
2. Filling the missing data with the **median**.

CATEGORICAL

1. Filling the missing data with **mode**
2. Filling with a **new type** for the missing values.

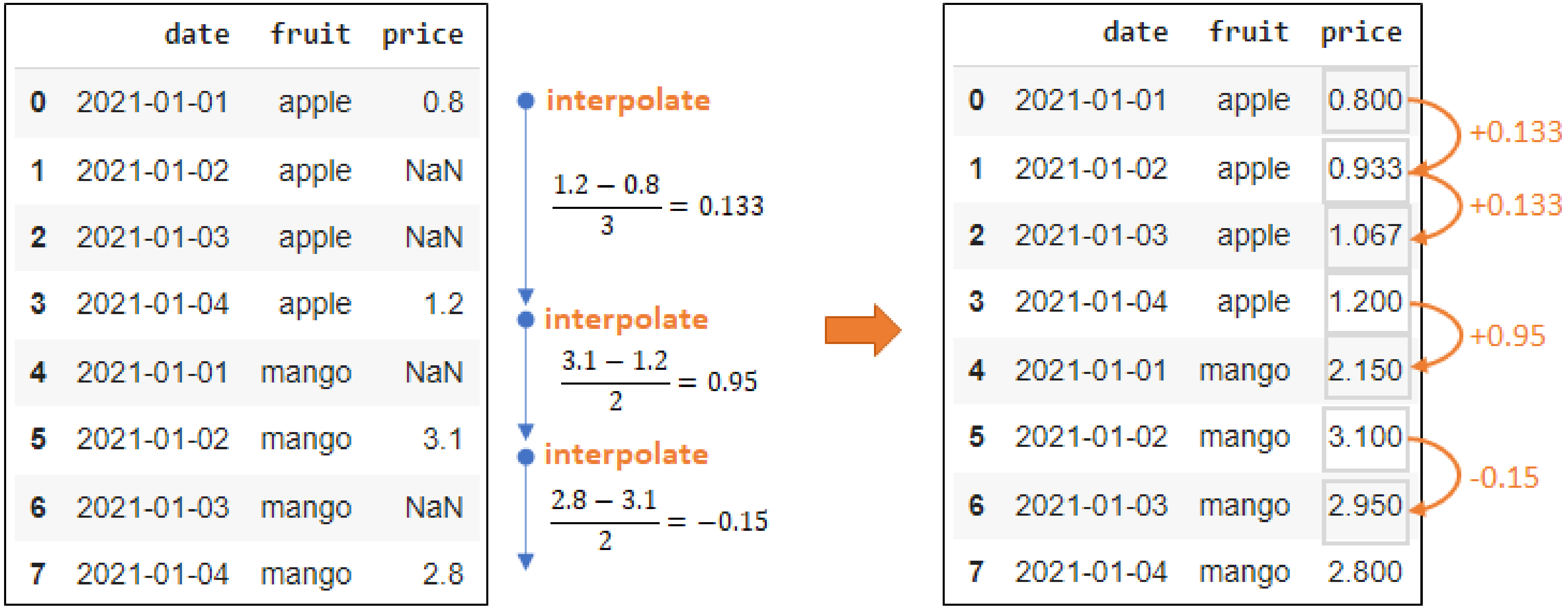
Last observation carried forward (LOCF)

Interpolation – Linear



- It's the method of approximating a missing value by joining dots in increasing order along a straight line.
- In a nutshell, it calculates the unknown value in the same ascending order as the values that came before it

Forward Interpolation



	date	fruit	price
0	2021-01-01	apple	0.8
1	2021-01-02	apple	NaN
2	2021-01-03	apple	NaN
3	2021-01-04	apple	1.2
4	2021-01-01	mango	NaN
5	2021-01-02	mango	3.1
6	2021-01-03	mango	NaN
7	2021-01-04	mango	2.8

$$\frac{1.2-0.8}{3}=0.133$$

$$\frac{3.1-1.2}{2}=0.95$$

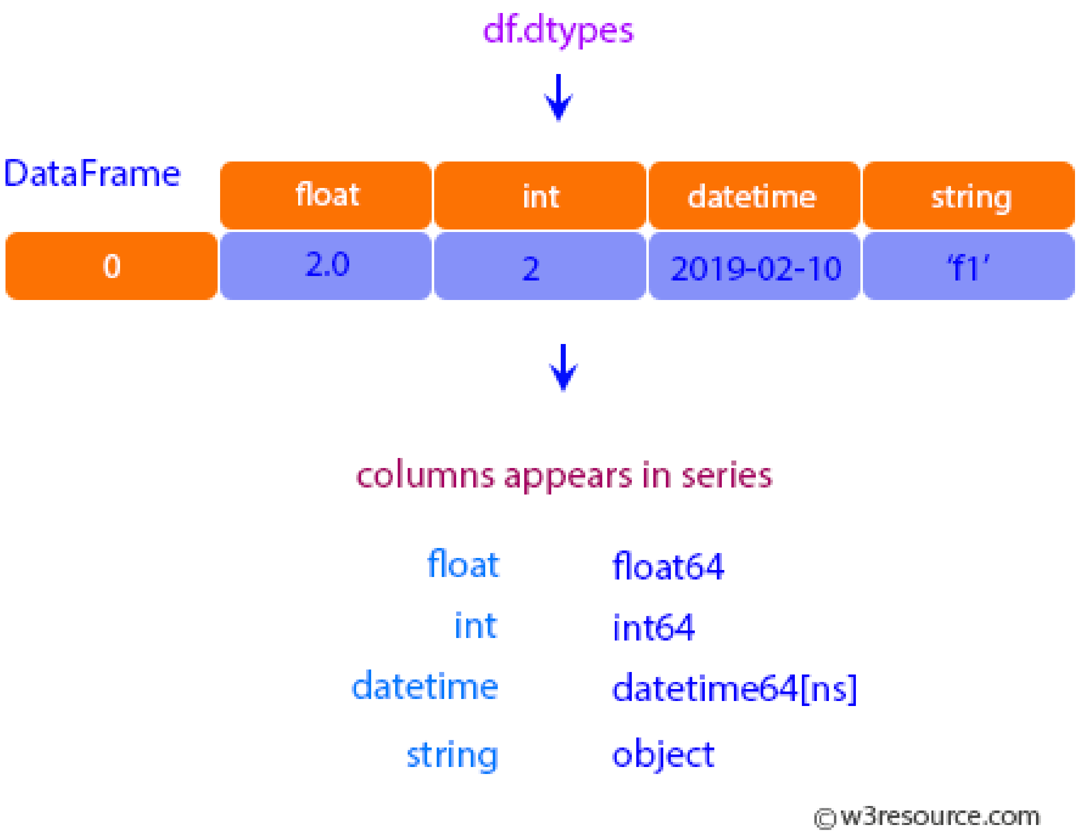
$$\frac{2.8-3.1}{2}=-0.15$$

	date	fruit	price
0	2021-01-01	apple	0.800
1	2021-01-02	apple	0.933
2	2021-01-03	apple	1.067
3	2021-01-04	apple	1.200
4	2021-01-01	mango	2.150
5	2021-01-02	mango	3.100
6	2021-01-03	mango	2.950
7	2021-01-04	mango	2.800

Python Notebook

DataWrangling.ipynb

```
#importing the basic libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plot

import missingno as mano

%matplotlib inline
```



df.dtypes

DataFrame	float	int	datetime	string
0	2.0	2	2019-02-10	‘f1’

columns appears in series

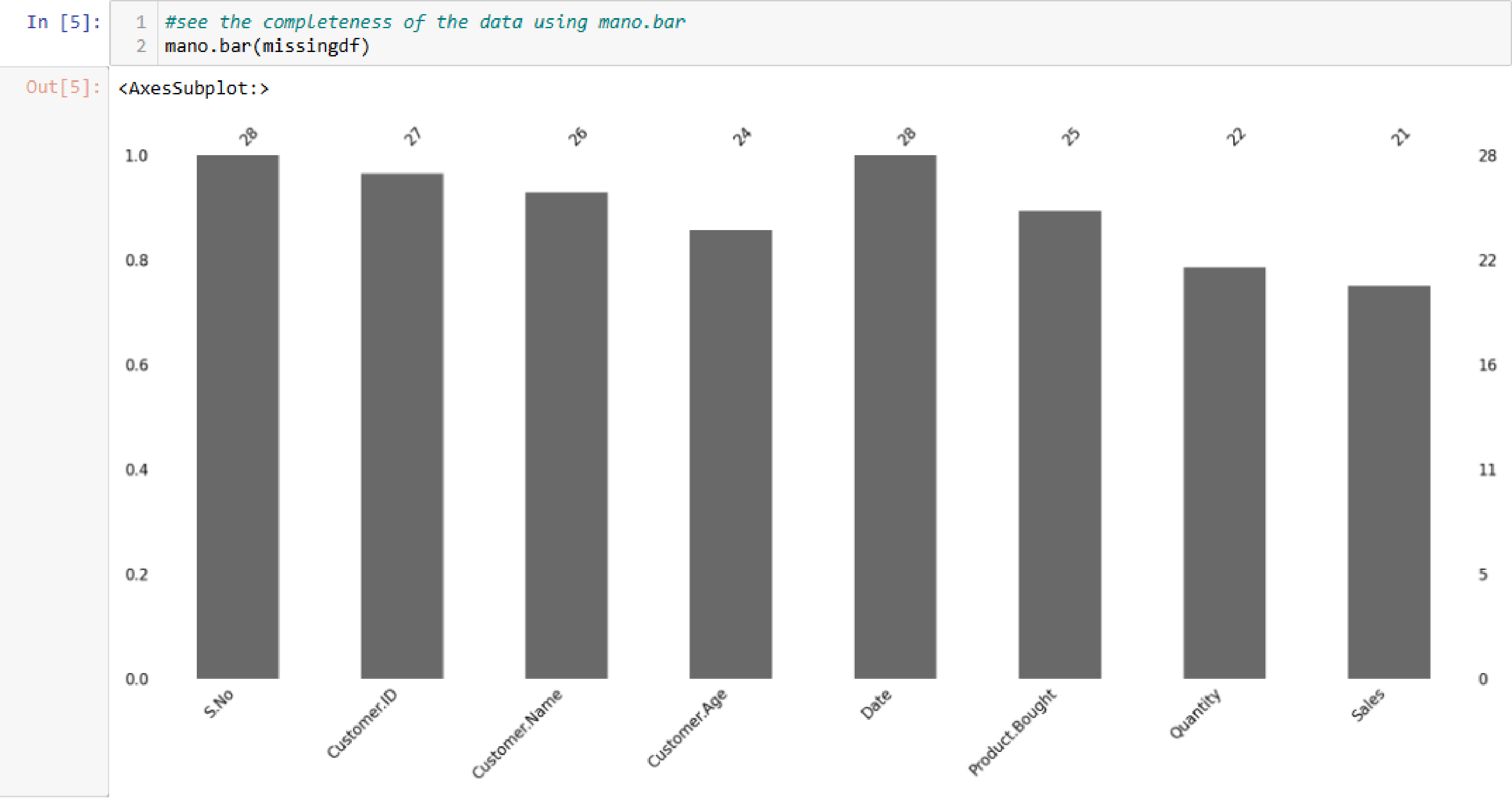
float	float64
int	int64
datetime	datetime64[ns]
string	object

©w3resource.com

Detecting MV Type before Treating it

DataWrangling_MVA.ipynb

MissingNo Library – Missingness Bar



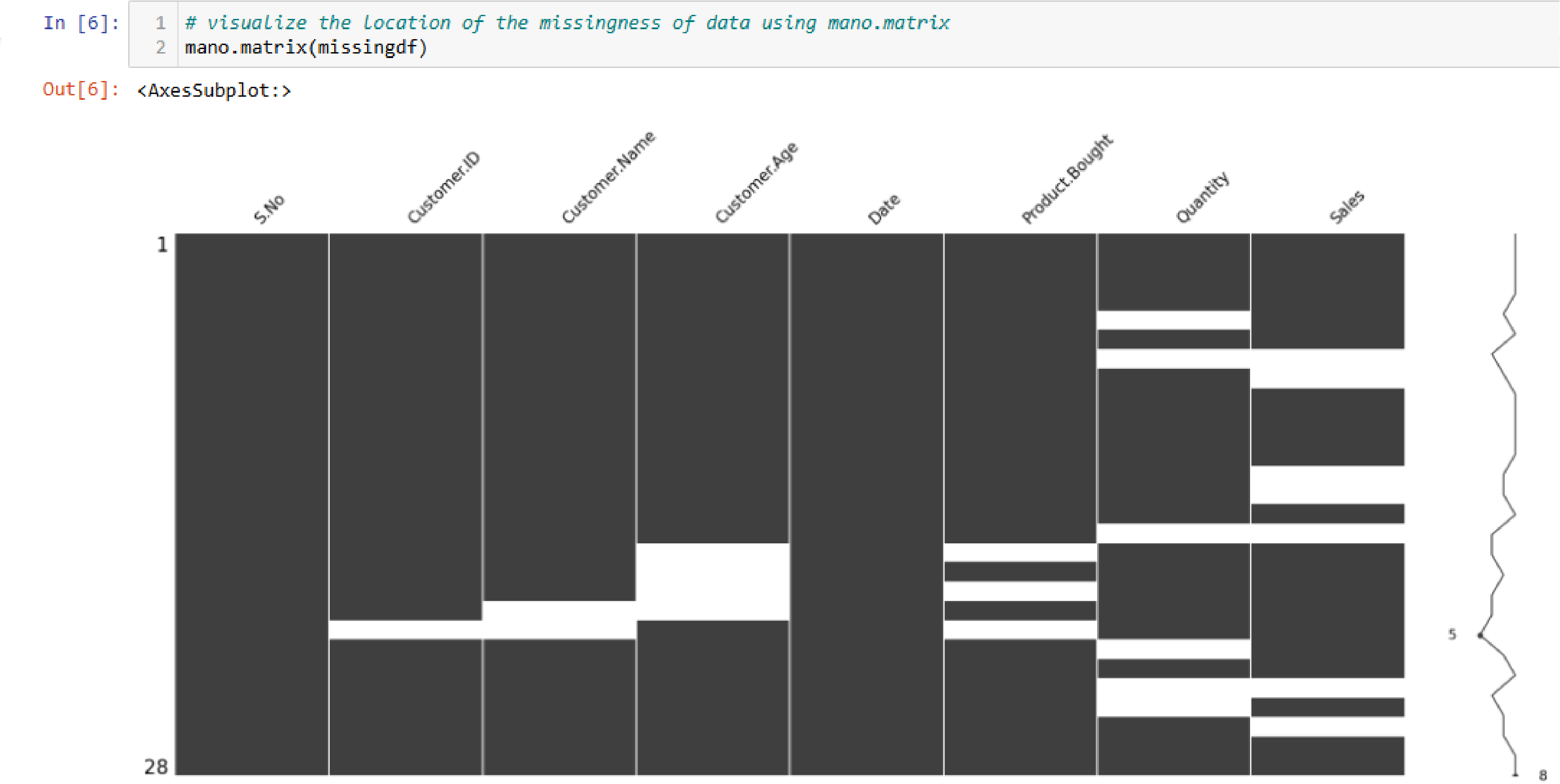
In [5]:

```
#see the completeness of the data using mano.bar
mano.bar(missingdf)
```

Out[5]:

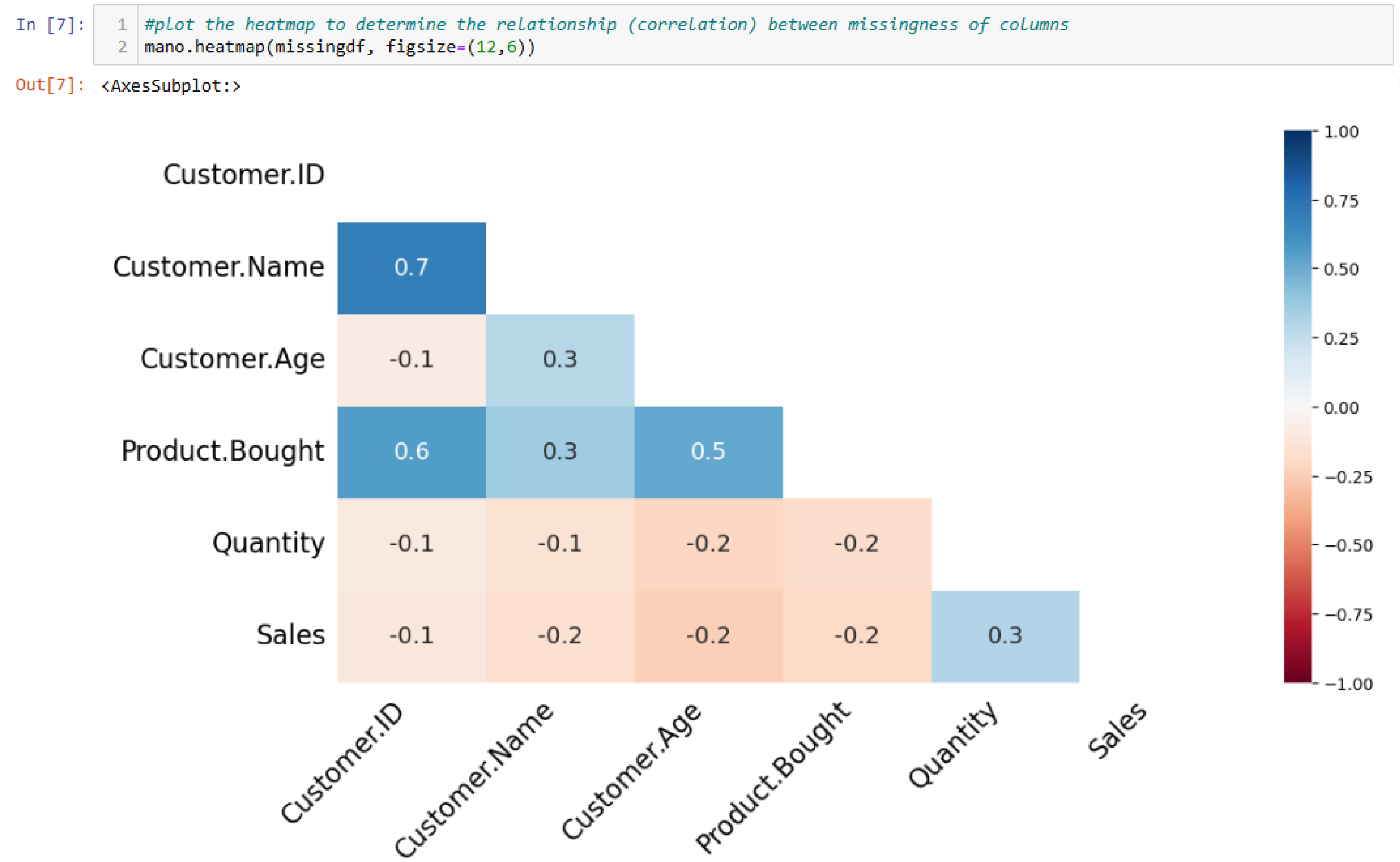
```
<AxesSubplot:>
```

MissingNo Library – Missingness Matrix



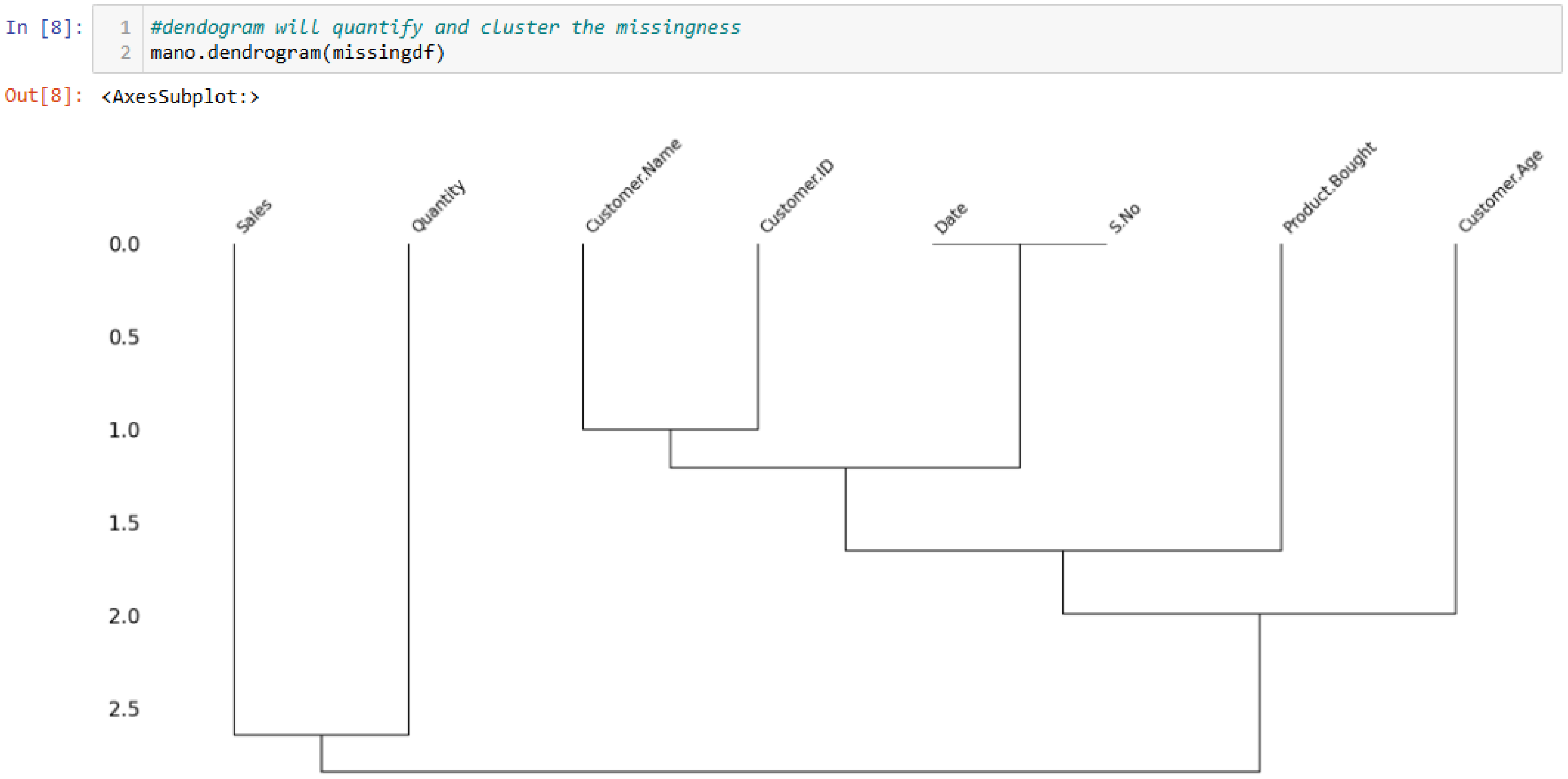
```
In [6]:  # visualize the location of the missingness of data using mano.matrix
         mano.matrix(missingdf)

Out[6]:  <AxesSubplot:>
```

MissingNo Library – Heatmap of missingness



```
In [7]:  #plot the heatmap to determine the relationship (correlation) between missingness of columns
         mano.heatmap(missingdf, figsize=(12,6))
Out[7]:  <AxesSubplot:>
```

MissingNo Library – Dendrogram



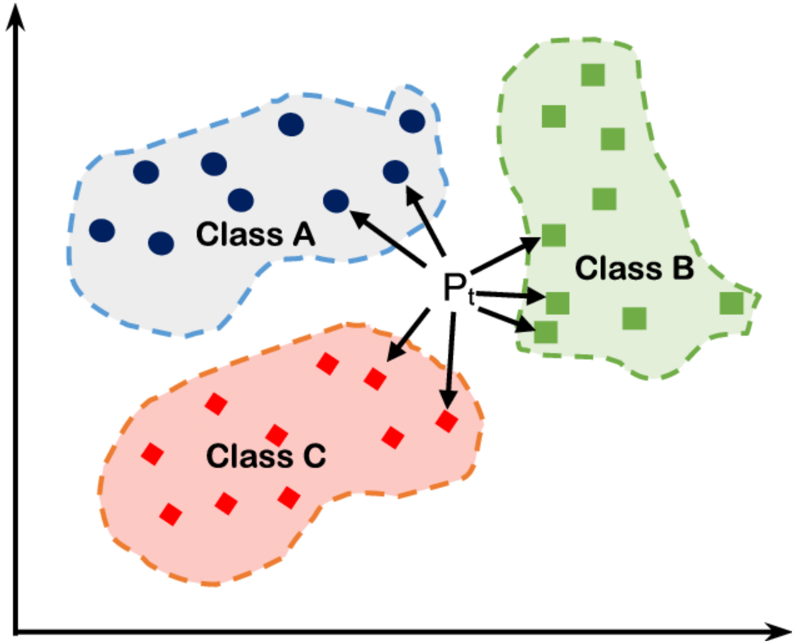
```
In [8]:  #dendogram will quantify and cluster the missingness
          mano.dendrogram(missingdf)
```

```
Out[8]: <AxesSubplot:>
```

Few more Imputation Strategies

Imputation by KNN

- A fundamental classification approach is the k-nearest-neighbors (kNN) algorithm.
- Class membership is the outcome of k-NN categorization
- If k = 1, the item is simply assigned to the class of the item's closest neighbor.
- Finding the k's closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighborhood might help generate predictions about the missing values.



MICE - Multiple Imputation by Chained Equation

- Multiple Imputation by Chained Equation assumes that data is MAR, i.e. missing at random.
- Sometimes data missing in a dataset and is related to the other features and can be predicted using other feature values.
- It cannot be imputed with general ways of using mean, mode, or median.

IterativeImputer class

- Models each feature with missing values as a function of other features and uses that estimate for imputation.
- It does so in an iterated round-robin fashion: at each step, a feature column is designated as output y and the other feature columns are treated as inputs X.
- A regressor is fit on (X, y) for known y. Then, the regressor is used to predict the missing values of y. This is done for each feature in an iterative fashion, and then is repeated for max_iter imputation rounds. The results of the final imputation round are returned.