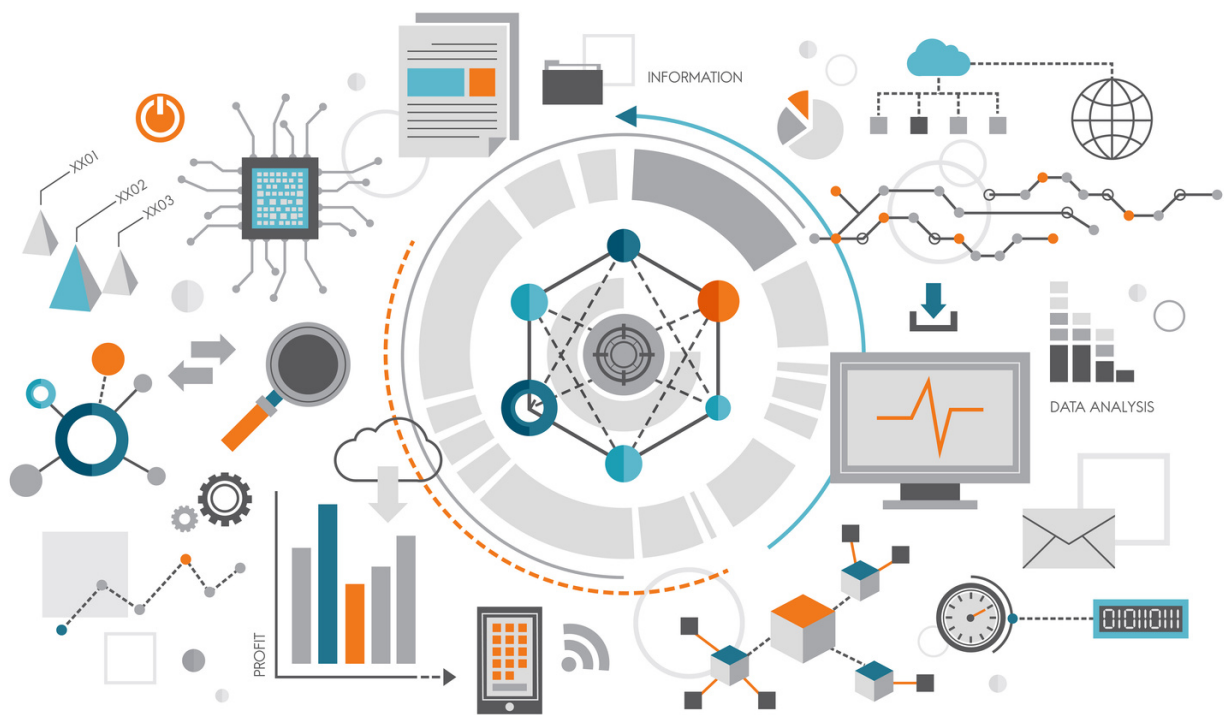


2023

# DATA SCIENCE NOTEBOOK



# INTRODUCTION

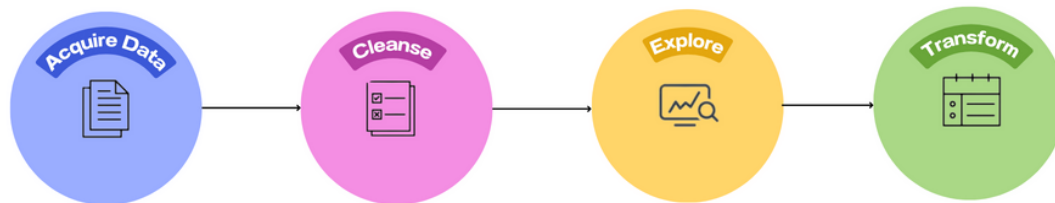
As the primary purpose of a data science project is to extract meaningful insights from large amounts of data, It's mandatory to focus on the step of data preparation to first understand your data.

In this notebook, I will reveal the steps of data preparation and the basics of each step, we will answer the following questions: how to detect and handle missing data and outliers, how to conduct an exploratory data analysis with comprehensive.

Let's start now:



# Data preparation



Here, we will focus on the data preparation workflow :

**Acquire Data:** In this first step, the analyst or the data scientist acquires data that will be used for their analysis. Usually, this data is delivered in structured data.

In our example, we choose to work on cars data: used cars data.csv

**Cleanse data:** This phase includes different tasks in order to improve the quality of data:

- Dealing with Outliers: outliers can be removed, transformed, or treated as special cases during analysis.
- Standardizing Variables: Standardize numerical variables by scaling them to a common range or transforming them into z-scores.
- Handling missing values: Options include removing rows with missing values, imputing values using statistical techniques, or using advanced imputation methods.
- Encoding Categorical Variables: Convert categorical variables into numerical representations suitable for analysis. Common encoding methods include one-hot encoding, label encoding, or target encoding.

the first step when it comes to dealing with data is knowing its type and properties. Let's explore now the types of data and their scale of measurement. This step is mandatory to choose the right statistical analysis.

As shown in the figure below, the collected data can be categorized as quantitative or qualitative.

## Quantitative = Quantity

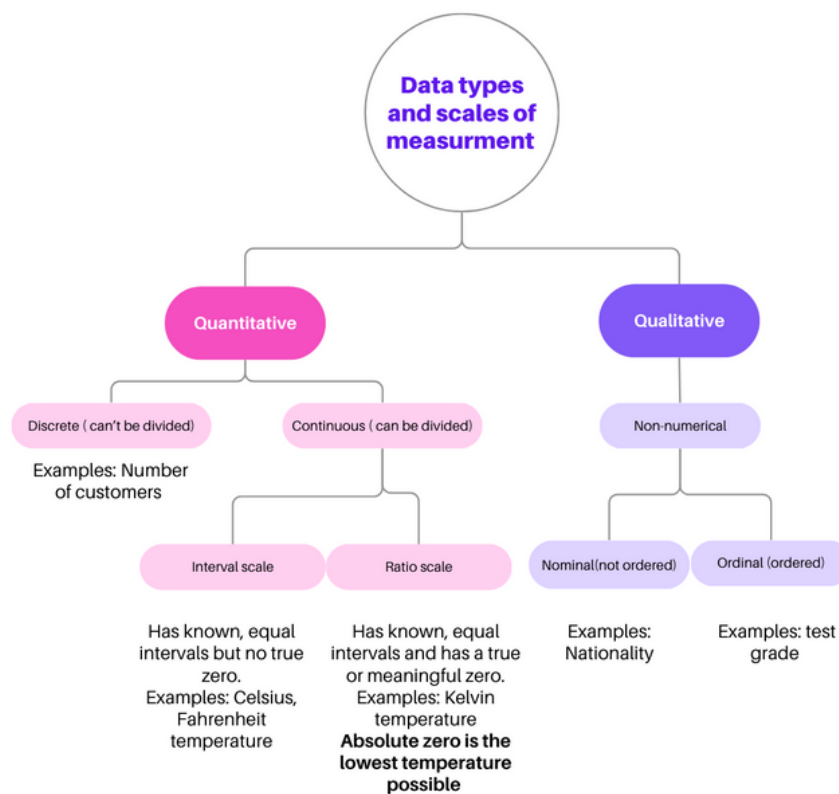
Quantitative data are :

- measures of values or counts and are expressed as numbers.
- data about numeric variables (e.g. how many, how much,s, or how often).

## Qualitative = Quality

Qualitative data are

- measures of 'types' and may be represented by a name, symbol, or a number code.
- Qualitative data are data about categorical variables (e.g. what type)



### Quantitative data can be:

- Discrete: can be counted and has a limited number of values.
- Continuous: can take any value like height, weight, and temperature.

### Continuous data can be classified into two categories :

- **Interval data:** an interval scale is one where there is order and the difference between two values is meaningful. Example: temperature (Fahrenheit), temperature (Celsius), pH.
- **Ratio data:** a ratio variable, has all the properties of an interval variable and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable. Example: temperature in Kelvin (0.0 Kelvin really does mean "no heat").

### Qualitative data can be :

- Ordinal data: an ordinal scale is one where the order matters but not the difference between values.

**Example:** Customer level of satisfaction (satisfied, neutral, dissatisfied)

- Nominal data: a nominal scale describes a variable with categories that do not have a natural order or ranking.

**Example:** Eye color (Black, Brown, Blue, Green)

**As machine learning or deep learning algorithms work only with numbers, we have to encode categorical data. Data encoding is an important pre-processing step in machine learning.**

### We have multiple Data encoding techniques:

- **Label encoding (Integer encoding):** assign an integer for each category. for example, we have a column for customer level satisfaction level having elements as dissatisfied, neutral, or satisfied in this case, we can replace these elements with 1,2,3. where 1 represents 'dissatisfied' 2 'neutral' and 3 'satisfied'.

→ Through this type of encoding, we try to preserve the meaning of the element where higher weights are assigned to the elements having higher priority.

**(+) advantages:**

- simple
- suited for ordinal data: this type of encoding preserves the order of ordinal data.

**(-) drawbacks:**

- Numeric values introduce arbitrary distances between categorical data that may not necessarily be realistic. For example, the three levels of satisfaction could also have been coded as 200, 101, 20.”
- not suitable for nominal data because we will introduce a non-existent order.

level of satisfaction	code
Satisfied	3
neutral	2
dissatisfied	1

**One-hot encoding:** let's consider a qualitative variable with n number of categories. We encode each category by mapping it to a vector in which the element corresponding to the category's dimension is 1, and the rest are 0, hence the name.

For example, let's suppose the categorical variable color which has three categories: red, green, and blue. Then, the encoding can be red = [1,0,0], green = [0,1,0], blue = [0,0,1] as shown in the figure below:

Island		Biscoe	Dream	Torgensen
Biscoe	→	1	0	0
Torgensen		0	0	1
Dream		0	1	0

**(+) advantages:** the transition from one category to another requires only one digit shift from 1 to 0 and another shift from 0 to 1.

**Missing data:** missing data is presented by Null, NaN(Not a Number), and NA (Not Available). Let us familiarize ourselves with the most common types of missing data:

-**Missing Completely at Random (MCAR):** the probability of a missing data value is independent of any observation in the data set.

-**Missing At Random (MAR):** The probability of a missing data value can depend on the values of other attributes.

-**Not Missing At Random (NMAR):** The probability of a missing data value depends on the value itself.

## Ways to handle missing data:

1. **Removal**: deleting Rows with missing values for example, missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped.

**Cons**: (-) loss of information

(-) not effective if the percentage of missing values is excessive in comparison with the complete dataset.

2. **Imputation**: instead of dropping the missing values, we can replace these values with imputation techniques.

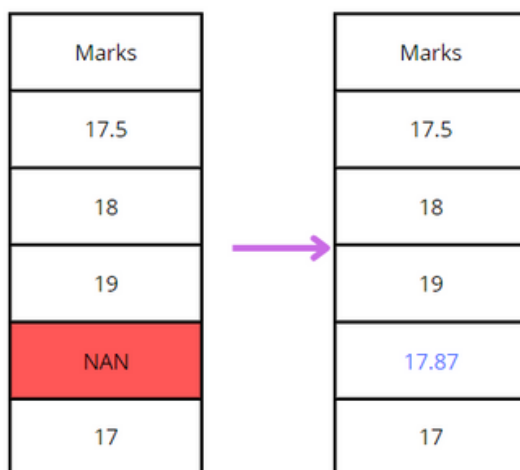
Choosing the best-suited imputation techniques depends on the type of missing value (MAR, NMAR, MCAR).

for example, in case we have the attribute1: 'Age', attribute2: 'minor/major' and the attribute2 is missing, we handle the value using the age.

technique 1: Mean/ Median /Mode imputation :

**Mean**: this technique is used in case the missing value is quantitative data, it is the average value.

$$\text{Mean} = (17.5 + 18 + 19 + 17) / 4 = 17.87$$



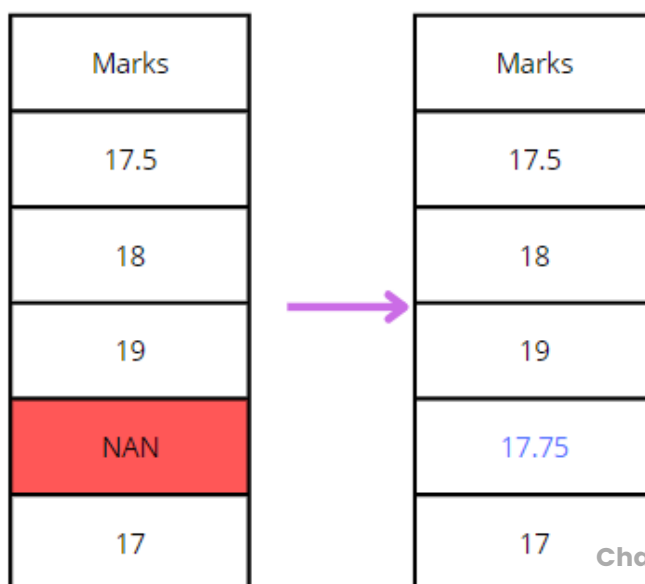
The diagram illustrates the mean imputation technique. It shows two vertical tables representing a column of 'Marks'. The first table has values 17.5, 18, 19, a red box containing 'NAN', and 17. A purple arrow points to the second table, which has the same values except the 'NAN' is replaced by '17.87'.

Marks
17.5
18
19
NAN
17

Marks
17.5
18
19
17.87
17

**Median**: this technique is also used in case the missing value is quantitative data, it is the midpoint value.

$$\text{Median} = \text{mean}(17.5, 18) = (17.5 + 18) / 2 = 17.75$$



The diagram illustrates the median imputation technique. It shows two vertical tables representing a column of 'Marks'. The first table has values 17.5, 18, 19, a red box containing 'NAN', and 17. A purple arrow points to the second table, which has the same values except the 'NAN' is replaced by '17.75'.

Marks
17.5
18
19
NAN
17

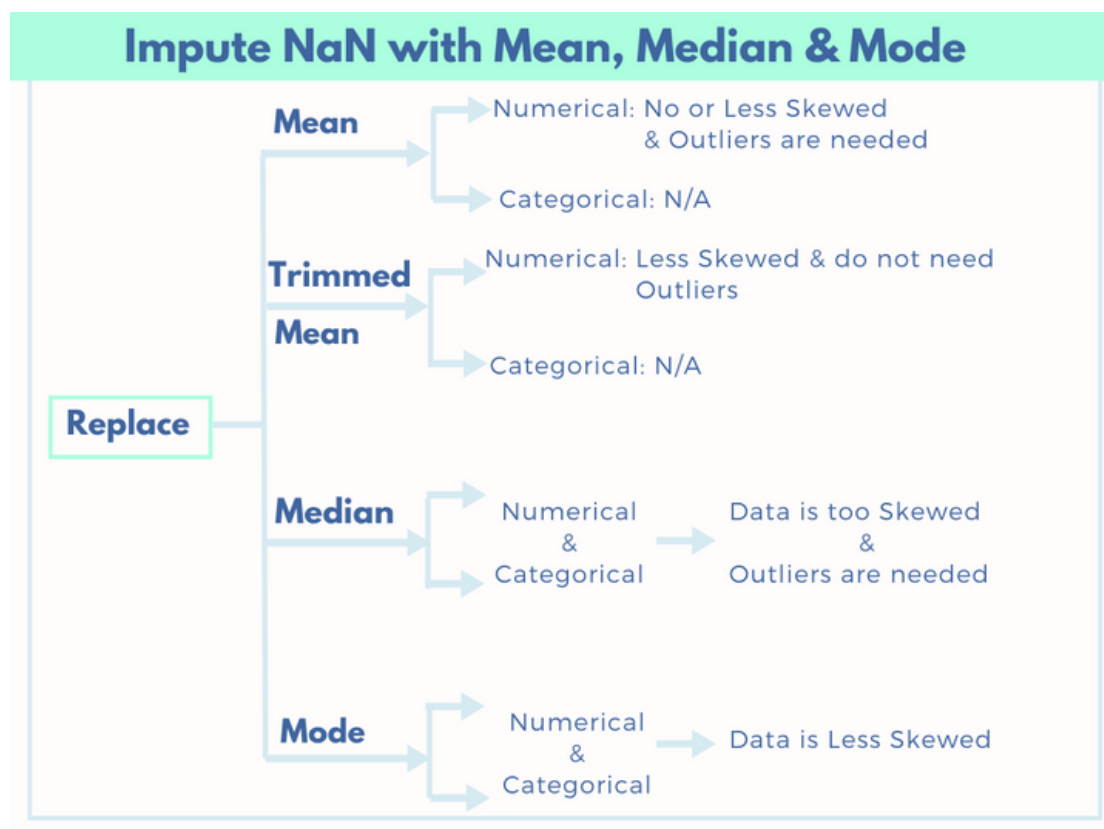
Marks
17.5
18
19
17.75
17

**Mode: It is the most common value.**

Mode

Marks		Marks
17		17.5
18		18
19	→	19
NAN		17
17		17

To use these techniques , certain conditions need to be validated as presented in the figure below.

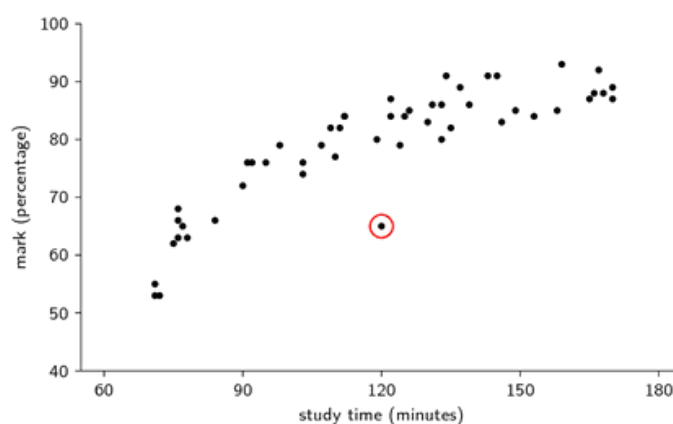


**Outliers:** an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. That's why it's mandatory to understand the cause in order to handle efficiently these outliers.

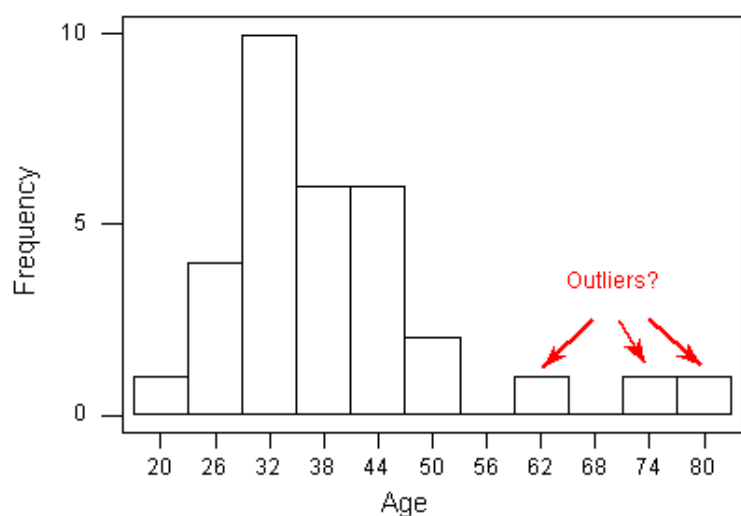
**Ways to identify the outliers:** we can identify them through visualization or statistical methods.

**graph your data:** use graphs like scatter plots or histograms. Graphs present your data visually, making it easy to see when a piece of data differs from the rest of the data set.

**A scatter plot** displays your data points as dots on a graph based on two variables on the x-axis and y-axis. Scatter plots are useful for visualizing outliers because you can see when one dot is far away from the other dots, which are usually clustered together. A far-away data point is likely the outlier.



**A histogram** displays data in groups called "bins." Histograms usually group data in ranges, differentiating histograms from bar graphs. Your range of data is usually the x-axis, and your other variable is typically the y-axis, which can help you identify unusual data points. For example, if most of your data points are on the right side of the graph and one bin of data is on the left side, you can deduce that the far left bin is an outlier.





**Boxplots:** are a popular and easy method for identifying outliers. this graph uses the interquartile range (IQR).

The formula of this distance is  $IQR = Q3 - Q1$ . Generally, a data point is an outlier if it is over 1.5 times the IQR below the first quartile or 1.5 times the IQR above the third quartile.

where:

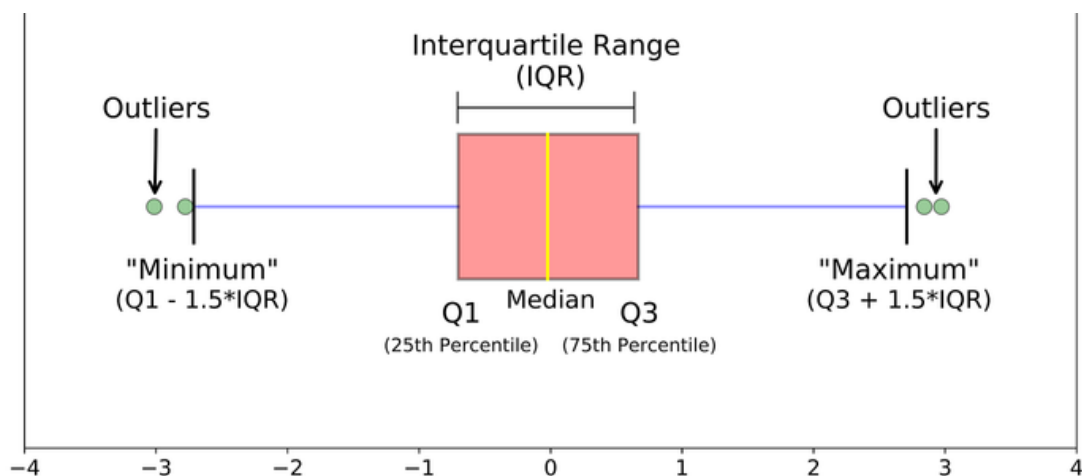
$Q3$  = the third quartile = the median of the upper half of the data set

$Q1$  = the first quartile = the median of the lower half of the data set

Then, you can use the IQR to find any outliers in your data set. The equations to calculate low or high outliers via the IQR range are:

High outlier  $\geq Q3 + (1.5 \times IQR)$

Low outlier  $\leq Q1 - (1.5 \times IQR)$



**Explore data:** during this phase, the analyst or the data scientist evaluates the quality of the dataset:

To accomplish this task, we perform an **Exploratory Data Analysis (EDA)**

#### Importance of EDA in a data science project :

This step is mandatory to understand and extract meaningful insights from the data set. Here are the following goals of conducting an EDA.

- Obtain a global overview of the data:** to get a general understanding of the data, including the distribution of the variables, the presence of outliers, and any potential trends or relationships between variables.

- Discover underlying patterns and structures:** to identify any patterns or structures in the data that are not immediately obvious. This can be done by using statistical techniques such as clustering, dimensionality reduction, and hypothesis testing.

- Extract important variables:** to identify the variables that are most important for understanding the data. This can be done by using statistical techniques such as feature selection and variable importance analysis.

- Formulate interesting questions or hypotheses:** use the insights gained from exploratory data analysis to generate new questions or hypotheses about the data.

**Types of Exploratory Data Analysis:** there are three main types of EDA:

- Univariate
- Bivariate
- Multivariate

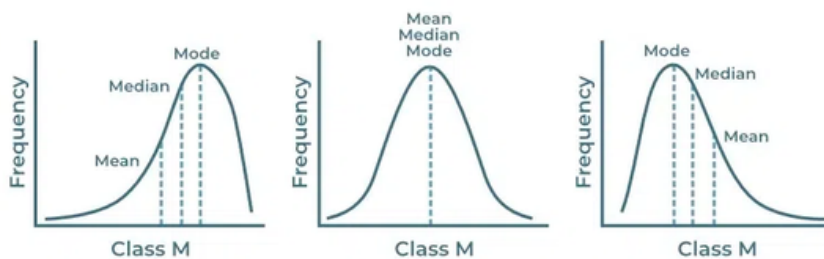
**Let's start with the type 'univariate': study each variable separately**

It is the simplest of all types of data analysis used in practice. Uni means only one variable is considered whose data (referred to as population) is compiled and studied.

- Univariate non-graphical EDA:

The main aim of univariate non-graphical EDA is to find out the details about the distribution of the population data and to know some specific parameters of statistics.

**Central Tendency:** define the values located at the data's central position or middle zone. The three generally estimated parameters of central tendency are mean, median, and mode.



**Range:** this information can be calculated by simply the difference between the maximum and minimum value in the data but this formula is not robust for outliers.

--> **Interquartile Range= Q3-Q1**

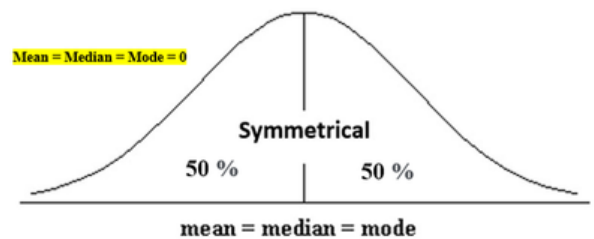
**Skewness Coefficient :**

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.

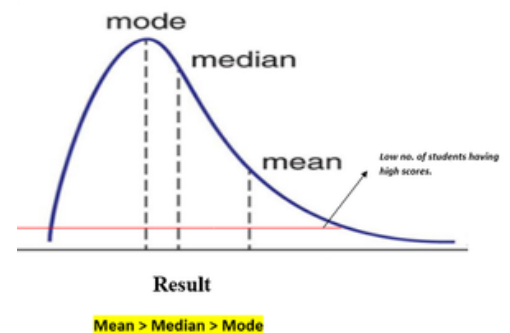
**skewness** is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.

Skewness: the third empirical moment=  $1/(n \cdot S^3) (\sum_{i=1}^n (x_i - \bar{x})^3)$

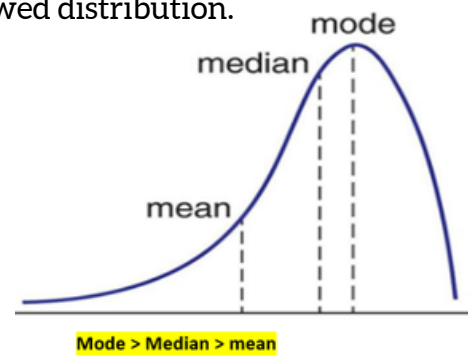
- if Skewness=0: data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations.



- if Skewness>0: data is a positively skewed or right-skewed distribution which has a long right tail.



- if  $\text{Skewness} < 0$ : data is a negatively skewed or left-skewed distribution has a long left tail; it is the complete opposite of a positively skewed distribution.



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.

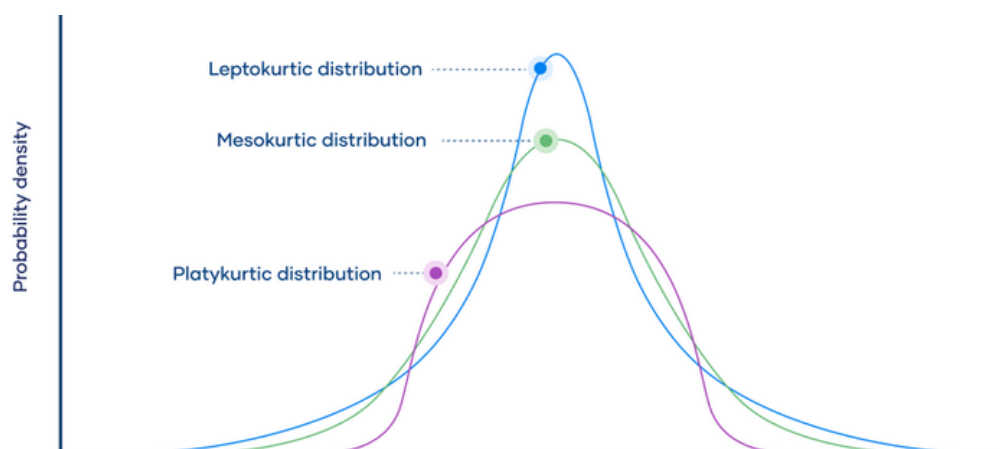
**kurtosis** : the fourth empirical moment =  $\frac{1}{(n \cdot S^4)} (\sum_{i=1}^n (x_i - \bar{x})^4)$

In the case of a normal distribution, kurtosis=3, we choose the normal distribution as a reference.

**Excess Kurtosis = kurtosis - 3**

the excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution).

- **Leptokurtic** or heavy-tailed distribution (kurtosis more than normal distribution): Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.
- **Platykurtic** having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.
- **Mesokurtic** is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.



- Univariate graphical EDA:

**Histograms (Bar Charts):** These plots are used to display both grouped or ungrouped data. On the x-axis, values of variables are plotted, while on the y-axis are the number of observations or frequencies. Histograms are very simple to quickly understand your data, which tell about values of data like central tendency, dispersion, outliers, etc.

There are many types of histograms:

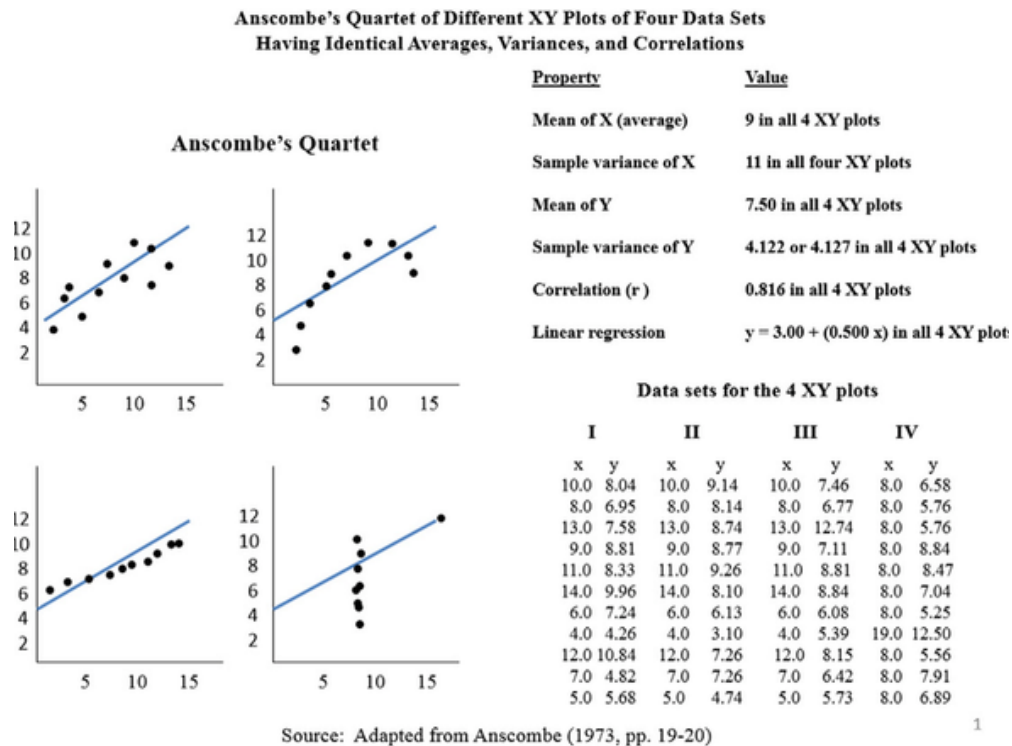
1. **Simple Bar Charts:** These are used to represent categorical variables with rectangular bars, where the different lengths correspond to the values of the variables.
2. **Percentage Bar Charts:** These are bar graphs that depict the data in the form of percentages for each observation.
3. **Box Plots:** These are used to display the distribution of quantitative value in the data. If the data set consists of categorical variables, the plots can show the comparison between them. Further, if outliers are present in the data, they can be easily identified.

Generally, it seems statistical graphics emphasize specific goals and answers being sought, whereas data visualization is more interested in creating the best way for data to be explored and questions to be discovered.

Anscombe's Quartet dataset showed the importance of data visualization compared to statistical graphics. This dataset is composed of 4 datasets, each one consisting of 11 pairs of x and y coordinates as shown in the table below:

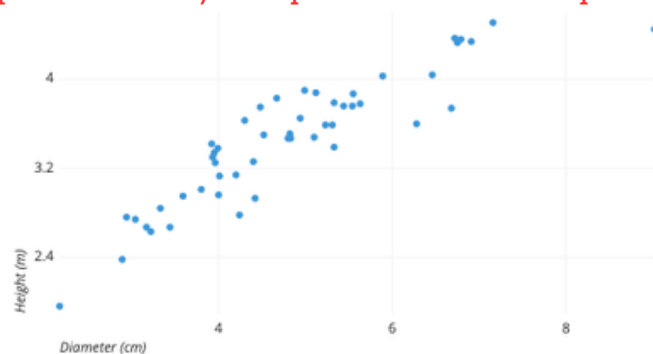
	I		II		III		IV	
	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

We may well have seen the identical summary statistics and assumed that all four datasets had the same distribution but as you can see, they don't have the same distribution.

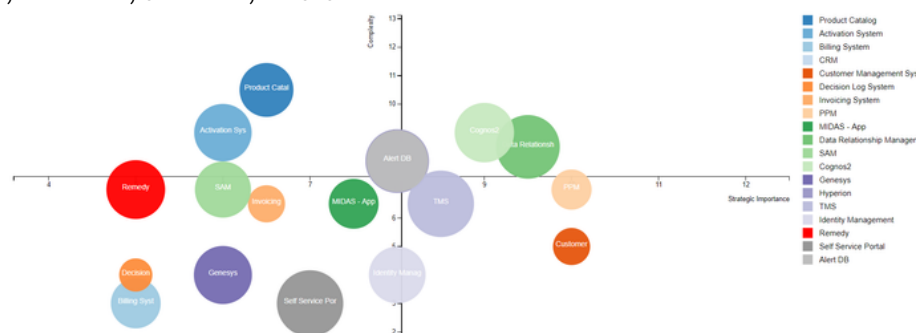


## Multivariate Graphical EDA

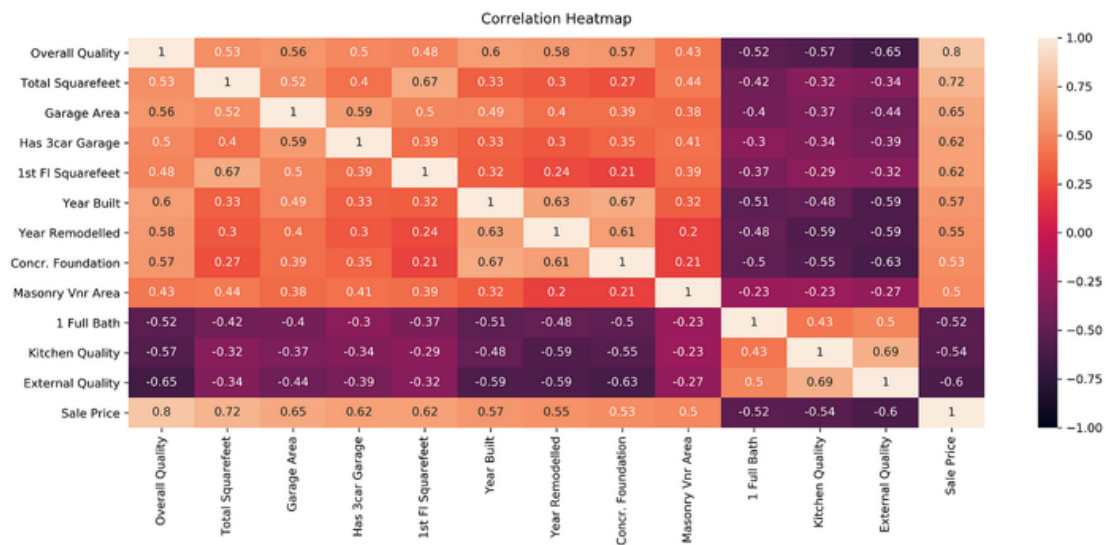
- **Scatter Plot:** the essential graphical EDA technique for two quantitative variables is the scatter plot, so one variable appears on the x-axis and the other on the y-axis and, therefore, the point for every case in your dataset. This can be used for bivariate analysis. -> **This graphical technique is used only to explore the relationships between two numerical variables.**



- **Bubble Chart:** bubble charts scatter plots that display multiple circles (bubbles) in a two-dimensional plot. These are used to assess the relationships between three or more numeric variables. In a bubble chart, every single dot corresponds to one data point, and the values of the variables for each point are indicated by different positions such as horizontal, vertical, dot size, and dot colors.



- **Heat map:** A heat map is a colored graphical representation of multivariate data structured as a matrix of columns and rows. The heat map transforms the correlation matrix into color coding and represents these coefficients to visualize the strength of correlation among variables. It assists in finding the best features suitable for building accurate Machine Learning models.



It's essential to acknowledge that data preparation is a fundamental requirement in every data science project, and we must thoroughly understand the basic techniques and remarks presented.

**Thank you for your attention**