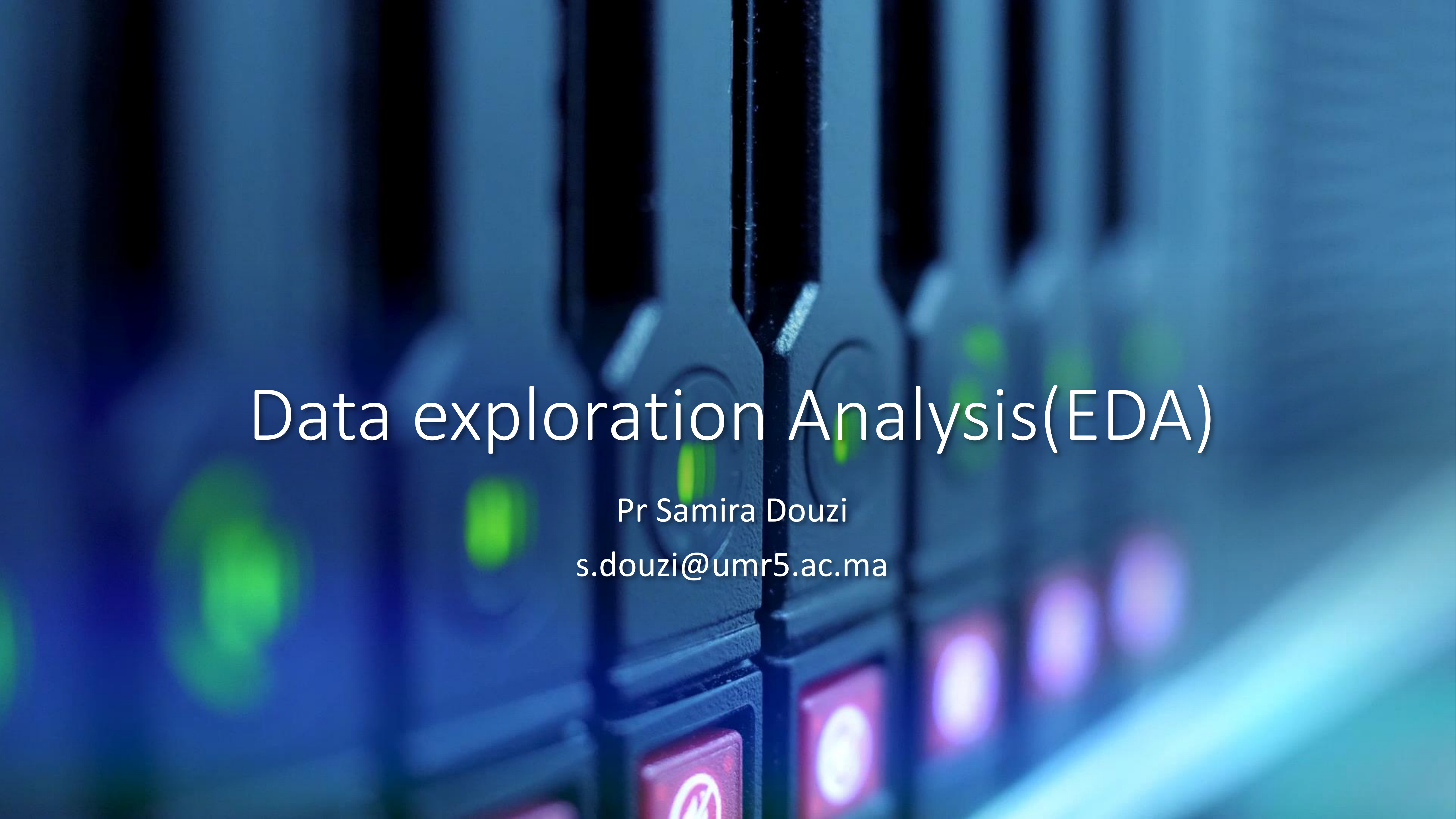


# Data exploration Analysis(EDA)

Pr Samira Douzi

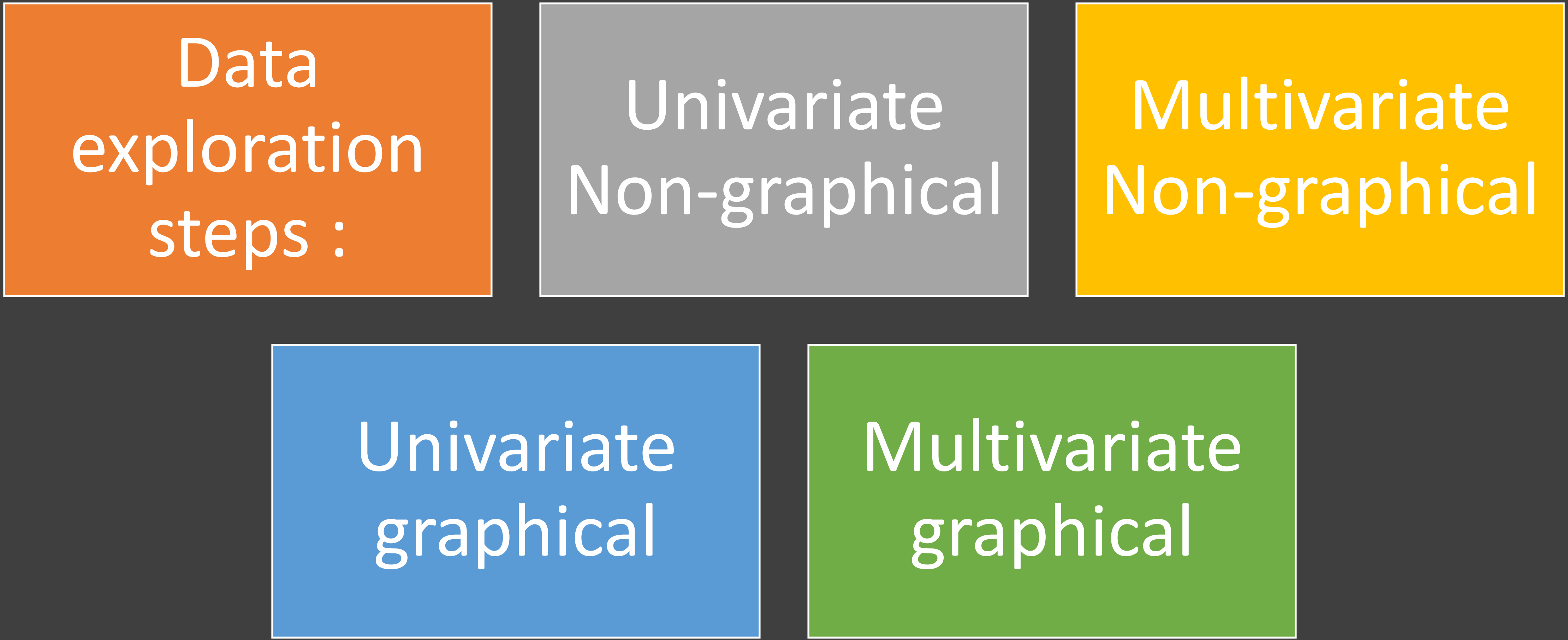
s.douzi@umr5.ac.ma

## What is Data Exploration Analysis?

**Data exploration** is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more.

What is Data Exploration Analysis tools ?

Data Exploration Analysis uses data visualization and statistical techniques to describe dataset characterizations in order to better understand the nature of the data.



Data exploration steps :

Univariate Non-graphical

Multivariate Non-graphical

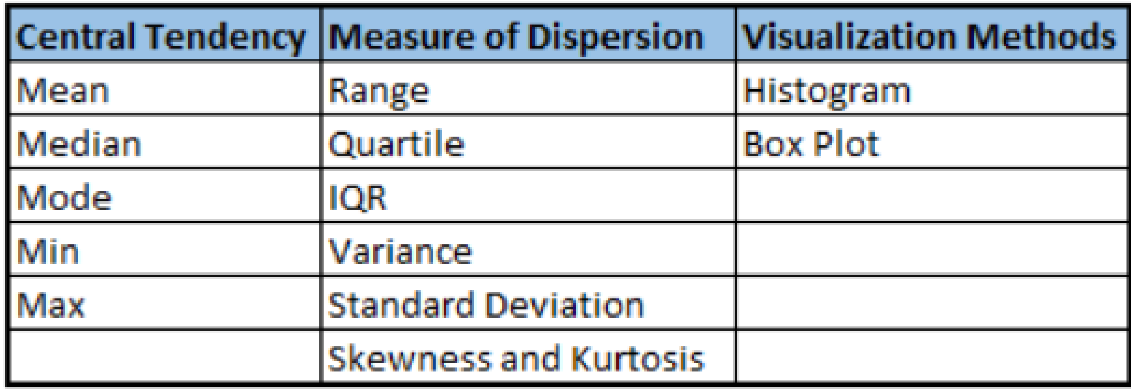
Univariate graphical

Multivariate graphical

# Univariate Analysis Non-graphical

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous.

- **Continuous Variables:-** In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown :



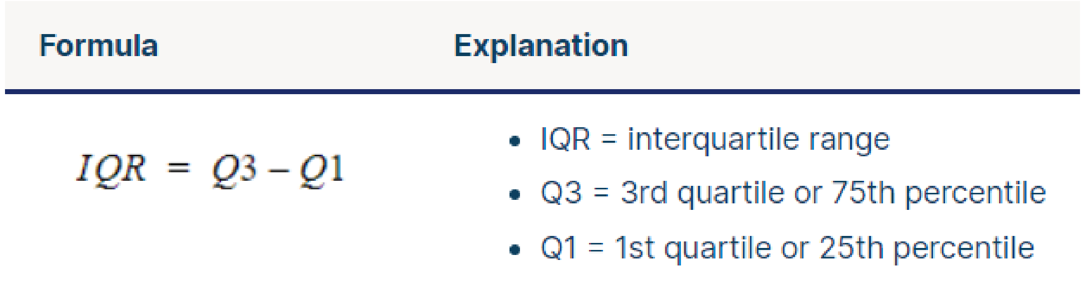
| Central Tendency | Measure of Dispersion | Visualization Methods |
|---|---|---|
| Mean | Range | Histogram |
| Median | Quartile | Box Plot |
| Mode | IQR | |
| Min | Variance | |
| Max | Standard Deviation | |
| | Skewness and Kurtosis | |

# Univariate Analysis Non-graphical

- There are three quartile values—a lower quartile, median, and upper quartile—to divide the data set into four ranges, each containing 25% of the data points.
  - **First interval**: The set of data points between the minimum value and the first quartile.
  - **Second interval**: The set of data points between the lower quartile and the median.
  - **Third interval**: The set of data between the median and the upper quartile.
  - **Fourth interval**: The set of data points between the upper quartile and the maximum value of the data set.

# Univariate Analysis Non-graphical

The interquartile range is found by subtracting the Q1 value from the Q3 value:



| Formula | Explanation |
|---|---|
| $IQR = Q3 - Q1$ | • IQR = interquartile range<br>• Q3 = 3rd quartile or 75th percentile<br>• Q1 = 1st quartile or 25th percentile |

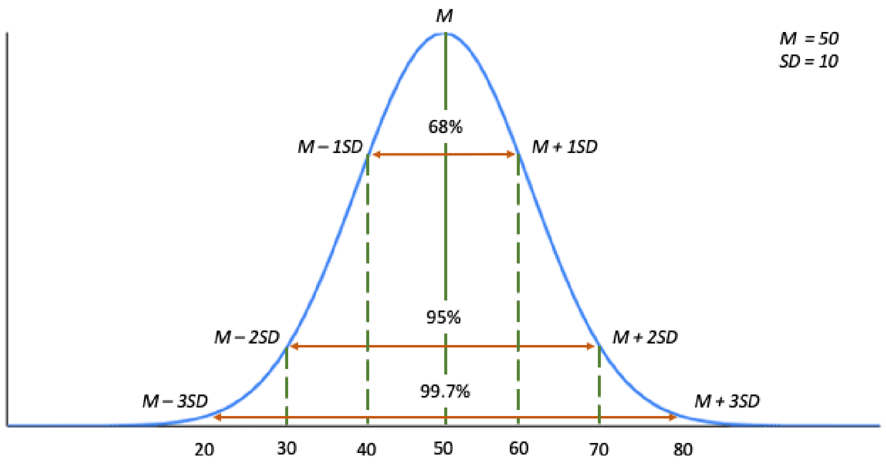
A smaller the interquartile range means you have less dispersion, while a larger the interquartile range means you have more dispersion.

# Univariate Analysis Non-graphical

Standard deviation is a useful measure of spread for **normal distributions**. In normal distributions, most values cluster around a central region. The standard deviation tells us how spread out from the center of the distribution your data is on average.

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

Standard deviations in a normal distribution

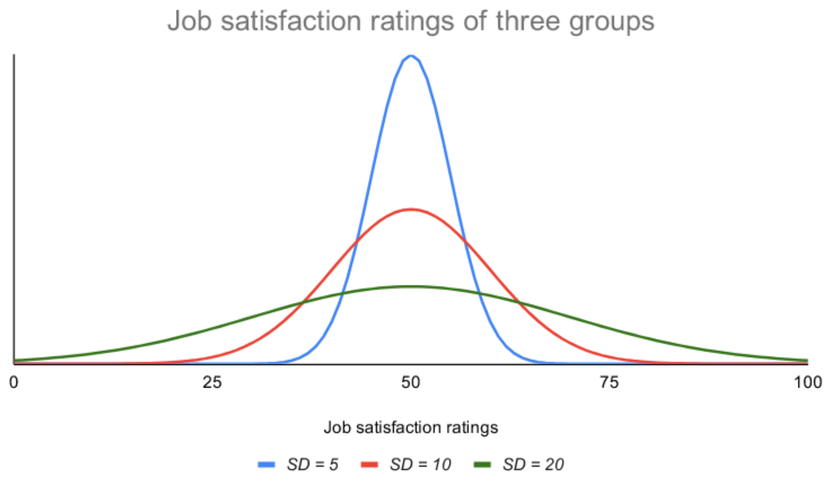


# Univariate Analysis Non-graphical

The **standard deviation** is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.

A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

$$\sigma = \sqrt{\frac{\Sigma f_i(x_i - \mu)^2}{N}}$$



The standard deviation reflects the dispersion of the distribution. The curve with the lowest standard deviation has a high peak and a small spread, while the curve with the highest standard deviation is flatter and more widespread.

# Univariate Analysis Non-graphical

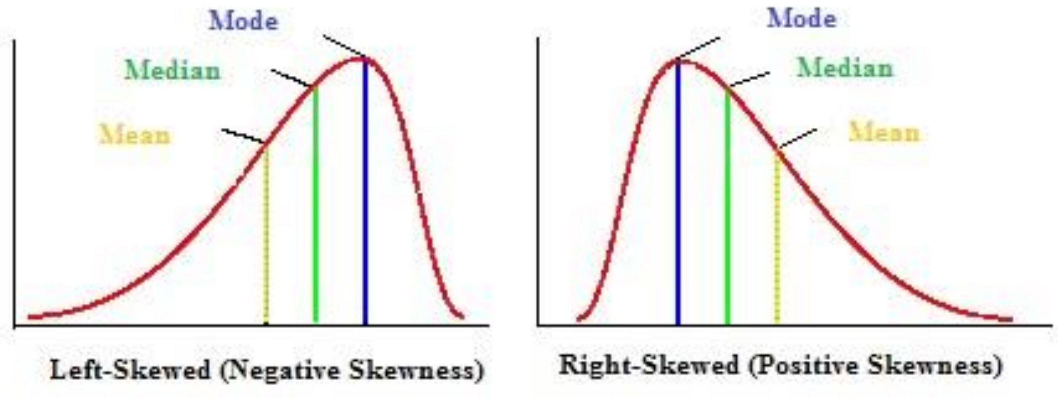
The term ‘skewness’ is used to mean the absence of symmetry from the mean of the dataset. It is characteristic of the deviation from the mean, to be greater on one side than the other, i.e., attribute of the distribution having one tail heavier than the other. Skewness is used to indicate the shape of the distribution of data.

$$\textbf{Skewness} = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3}$$

$$\mathrm{SK} = \frac{(\text{mean} - \text{median})}{SD} = \frac{(\tilde{X} - \text{median})}{s}$$

# Univariate Analysis Non-graphical

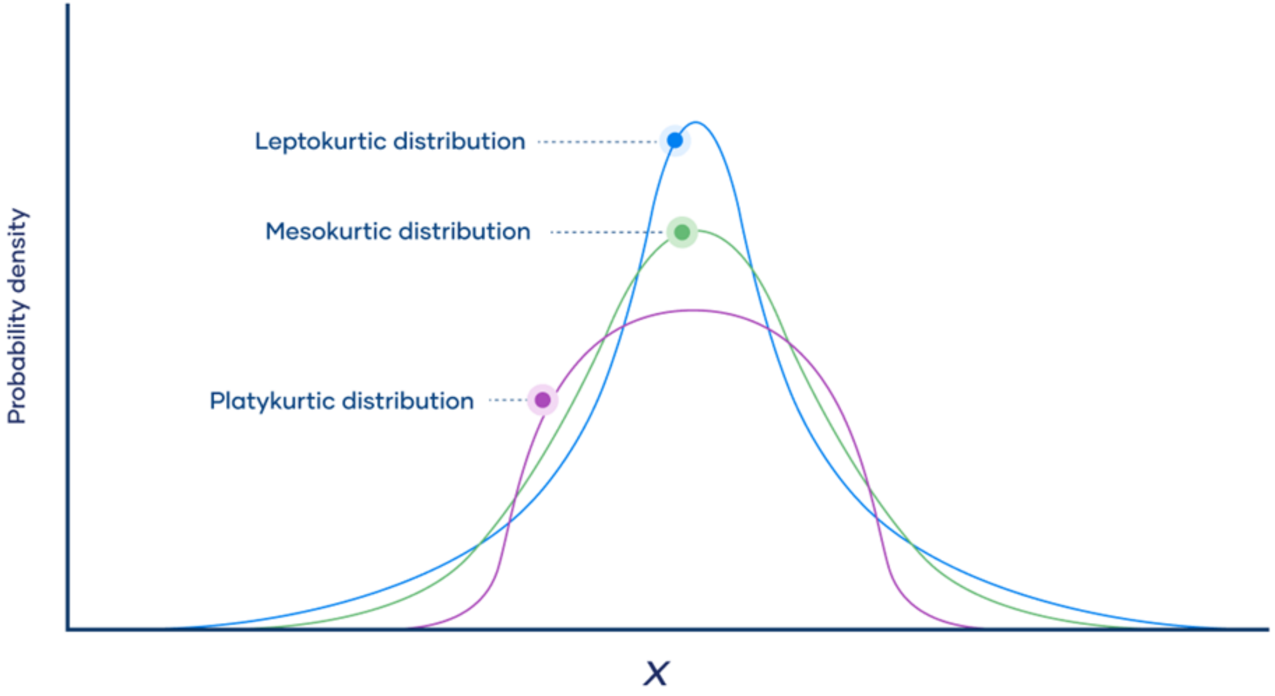
In a skewed distribution, the curve is extended to either left or right side. So, when the plot is extended towards the right side more, it denotes positive skewness, wherein mode < median < mean. On the other hand, when the plot is stretched more towards the left direction, then it is called as negative skewness and so, mean < median < mode.


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

# Univariate Analysis Non-graphical

Kurtosis is a measure of the tailedness of a distribution. Tailedness is how often outliers occur.

- Distributions with medium kurtosis (medium tails) are mesokurtic.
- Distributions with low kurtosis (thin tails) are platykurtic.
- Distributions with high kurtosis (fat tails) are leptokurtic.



# Univariate Analysis Non-graphical

Kurtosis is measured in comparison to normal distributions. Normal distributions have a kurtosis of 3, so any distribution with a kurtosis of approximately 3 is mesokurtic.

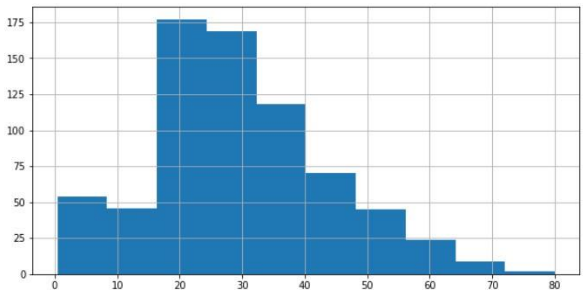
For a data analyst or statistician, the concept of kurtosis is very important as it indicates how are the outliers distributed across the distribution in comparison to a normal distribution. excess kurtosis wherein it is calculated by deducting 3 from the kurtosis, i.e. (kurtosis – 3)

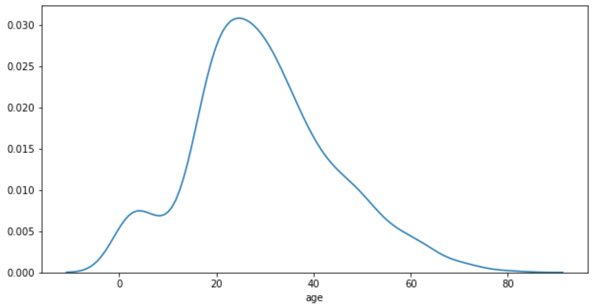
•If Zero, Then It Is a Mesokurtic Distribution
•If Negative, Then It Is a Platykurtic Distribution
•If Positive, Then It Is a Leptokurtic Distribution

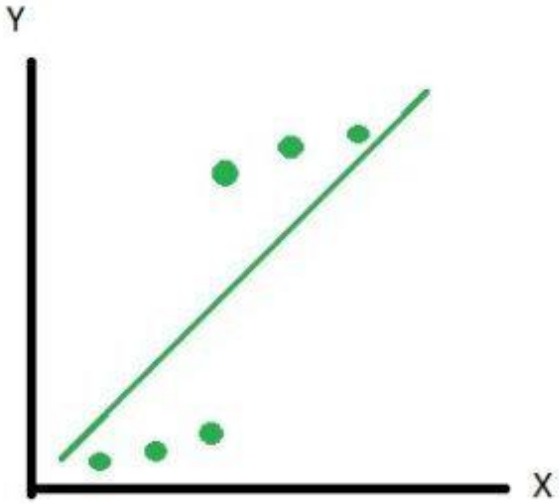
$$\textbf{Kurtosis} = n * \frac{\sum_i^n (X_i - \bar{X})^4}{\sum_i^n (X_i - \bar{X}\,)^2)^2}$$

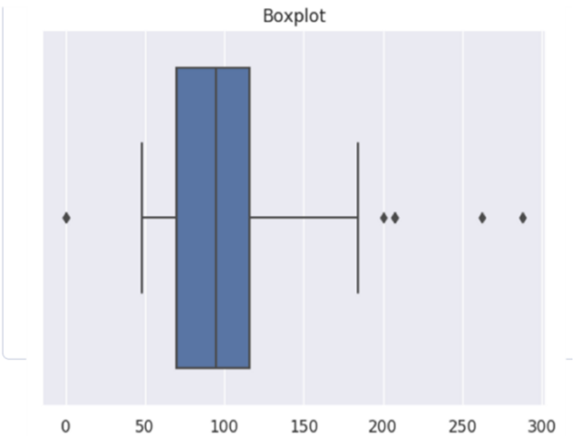
note that an investor is more comfortable with a platykurtic distribution of return as it indicates stable returns and lower risk of sudden shock of outliers, while leptokurtic distribution means chances of higher return but with higher risk.

# Univariate graphical









# Multivariate Non-graphical Analysis

Multi-variate Analysis finds out the relationship between two or multi variables. We look for association and disassociation between variables at a pre-defined significance level.

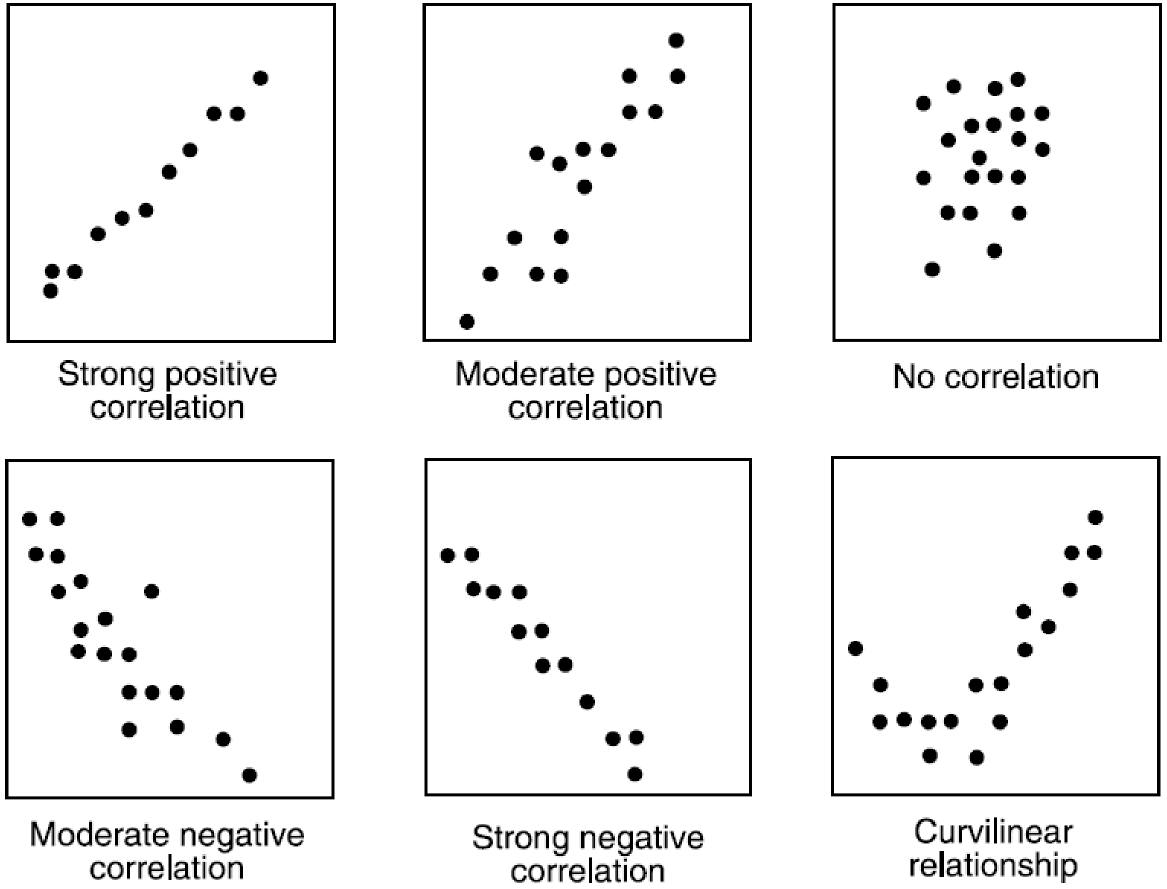
## Bi-Variate analysis

We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

# Multivariate analysis

**Continuous & Continuous :** While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.

Correlation = Covariance(X,Y) / SQRT( Var(X)* Var(Y))



Strong positive correlation

Moderate positive correlation

No correlation

Moderate negative correlation

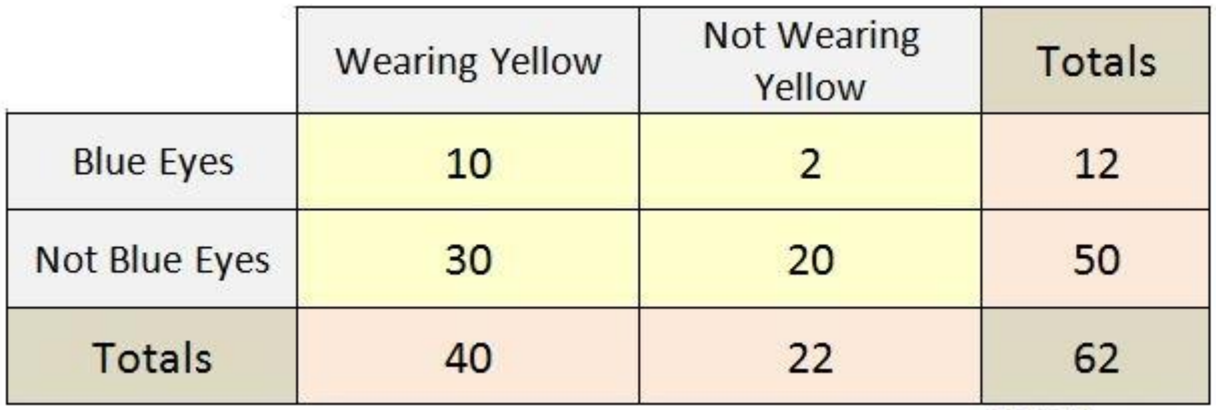
Strong negative correlation

Curvilinear relationship

# Multivariate analysis

**Categorical & Categorical** : To find the relationship between two categorical variables, we can use following methods:

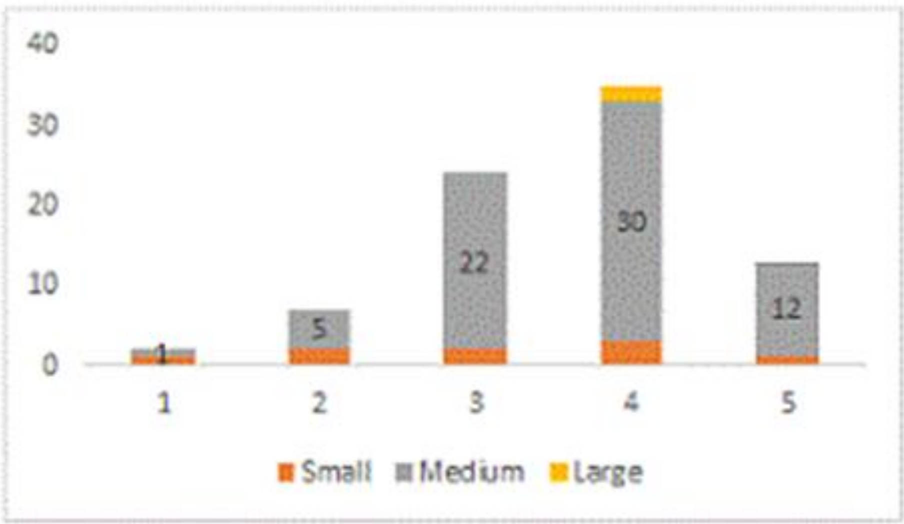
- Two-way table



| | Wearing Yellow | Not Wearing Yellow | Totals |
|---|---|---|---|
| Blue Eyes | 10 | 2 | 12 |
| Not Blue Eyes | 30 | 20 | 50 |
| Totals | 40 | 22 | 62 |

MathBits.com

**Stacked Column Chart:** This method is more of a visual form of Two-way table.



# Multivariate analysis

**Chi-Square Test:** This test is used to derive the statistical significance of relationship between the variables. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

$$X^2 = \sum (O - E)^2 / E$$

E is the expected frequency and O represents the observed.

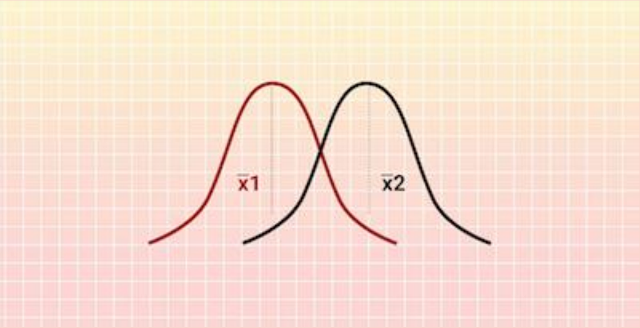
$$E = \frac{row\ total \times column\ total}{sample\ size}$$

Probability of 0 : It indicates that both categorical variable are dependent

Probability of 1 : It shows that both variables are independent.

Probability less than 0.05 : It indicates that the relationship between the variables is significant at 95% confidence.

# Bi-Variate analysis



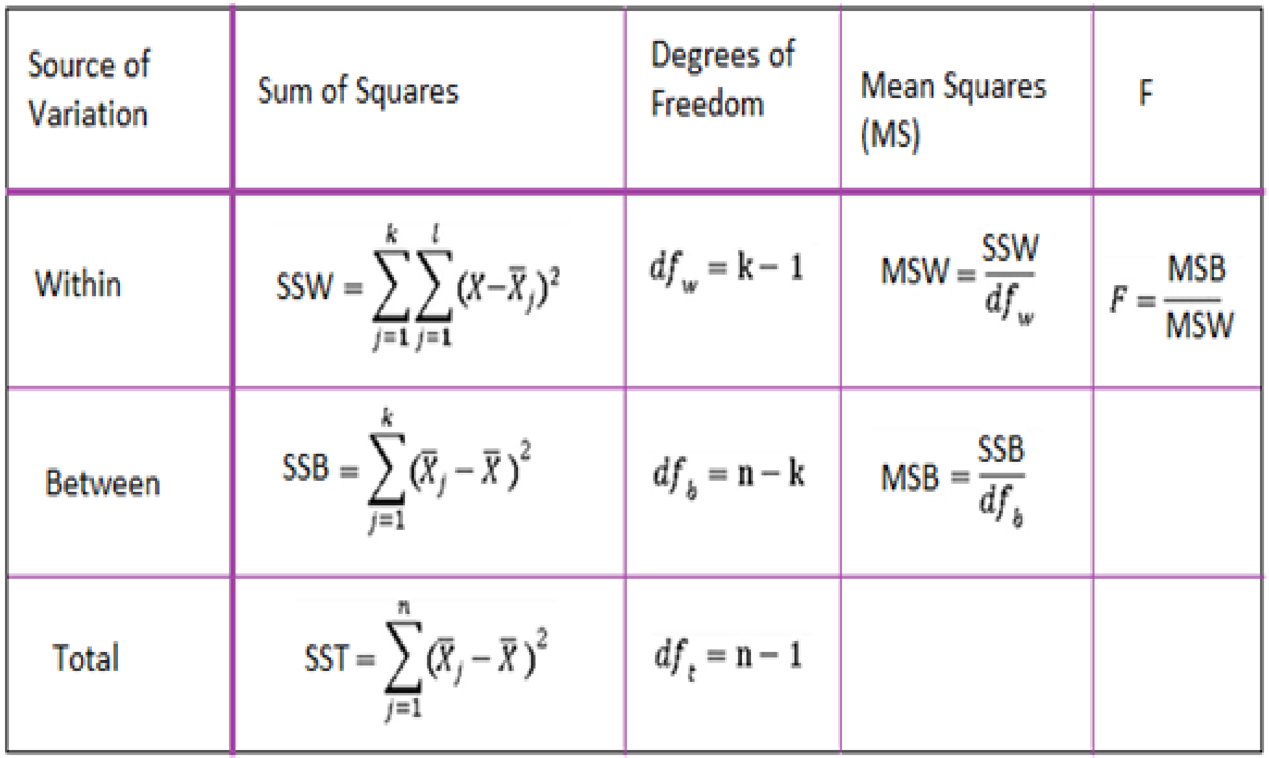
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

- **t-test** is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment influences the population of interest , or whether two groups are different from one another.

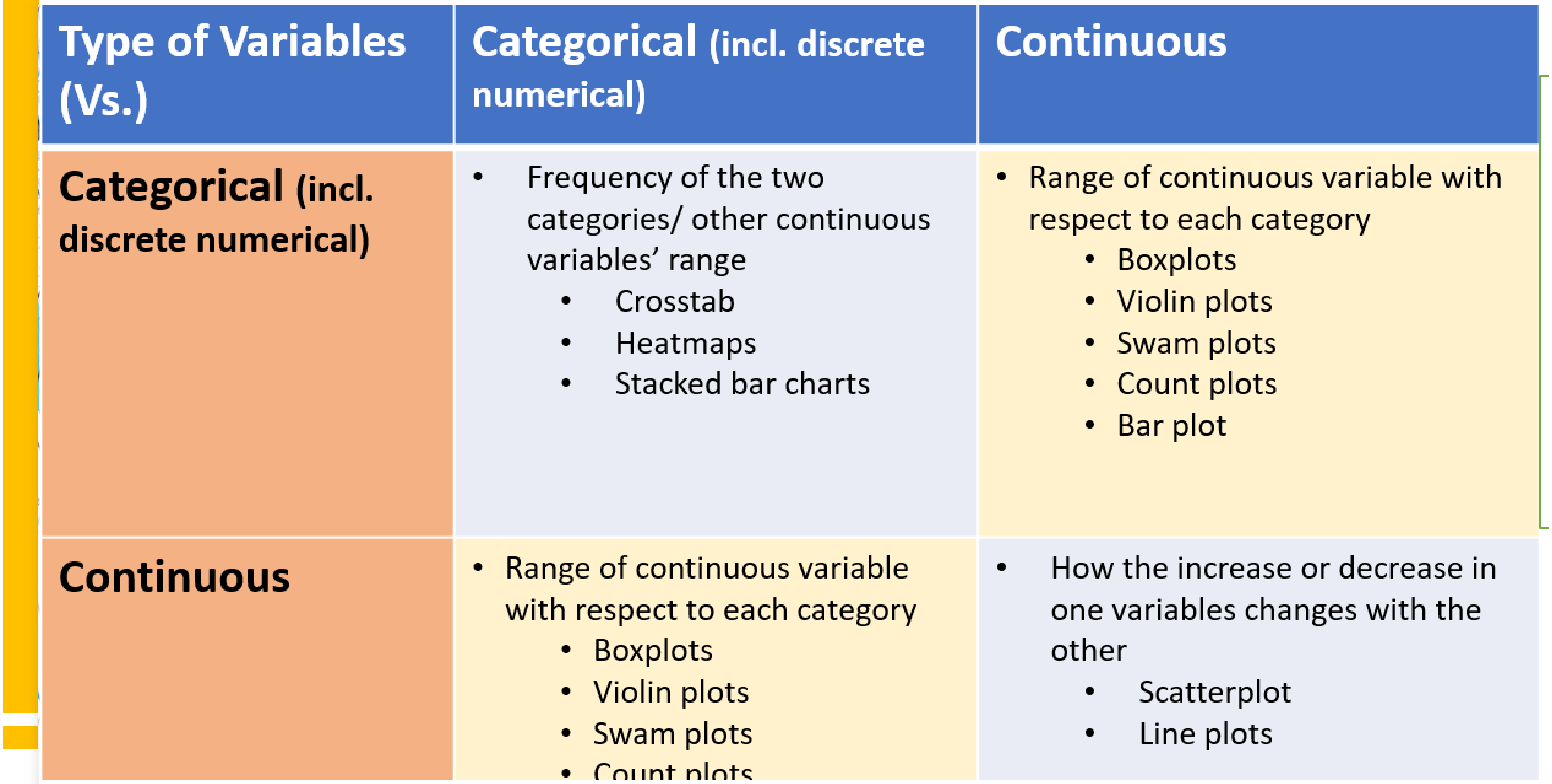
Example : A doctor may want to know if some new drug leads to a significant reduction in blood pressure compared to the current standard drug used.

# Multi-Variate analysis

The ANOVA test is used to determine whether there is a significant difference among the averages of more than two groups that are statistically different from each other.



| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares (MS) | F |
|---|---|---|---|---|
| Within | $SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$ | $df_w = k - 1$ | $MSW = \frac{SSW}{df_w}$ | $F = \frac{MSB}{MSW}$ |
| Between | $SSB = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$ | $df_b = n - k$ | $MSB = \frac{SSB}{df_b}$ | |
| Total | $SST = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$ | $df_t = n - 1$ | | |



| **Type of Variables (Vs.)** | **Categorical (incl. discrete numerical)** | **Continuous** |
|---|---|---|
| **Categorical (incl. discrete numerical)** | • Frequency of the two categories/ other continuous variables' range<br>• Crosstab<br>• Heatmaps<br>• Stacked bar charts | • Range of continuous variable with respect to each category<br>• Boxplots<br>• Violin plots<br>• Swam plots<br>• Count plots<br>• Bar plot |
| **Continuous** | • Range of continuous variable with respect to each category<br>• Boxplots<br>• Violin plots<br>• Swam plots<br>• Count plots | • How the increase or decrease in one variables changes with the other<br>• Scatterplot<br>• Line plots |

# Missing Value Treatment

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

# Missing Value Types

**Missing Completely at Random (MCAR)**
When we say data are missing completely at random, we mean that the missingness has nothing to do with the observation being studied

**Missing at Random (MAR)**
When we say data are missing at random, we mean that missing data on a partly missing variable (Y) is related to some other completely observed variables(X) in the analysis model but not to the values of Y itself.

**Missing not at Random (MNAR)**
When data are missing, not at random, the missingness is specifically related to what is missing, e.g. a person does not attend a drug test because the person took drugs the night before.

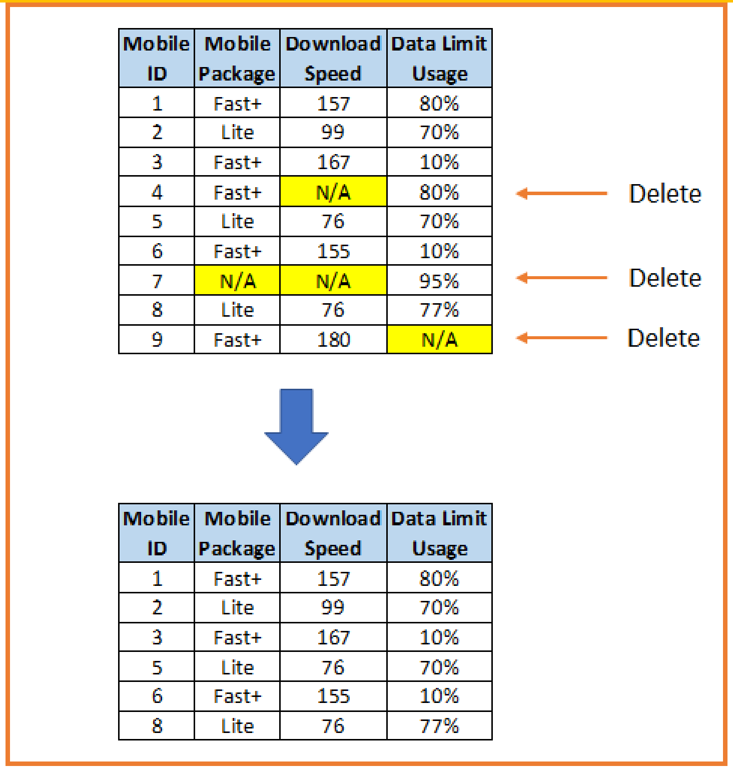
# Handling Missing Values : Discard Data

**1) list-wise (Complete-case analysis — CCA) deletion**

The most common approach to the missing data is to omit those cases with the missing data and analyze the remaining data.

If there is a large enough sample, where power is not an issue, and the assumption of MCAR is satisfied, the listwise deletion may be a reasonable strategy.

However, when there is not a large sample or the assumption of MCAR is not satisfied, then listwise deletion is not the optimal strategy.



| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | N/A | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 | N/A | N/A | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | N/A |

Delete
Delete
Delete

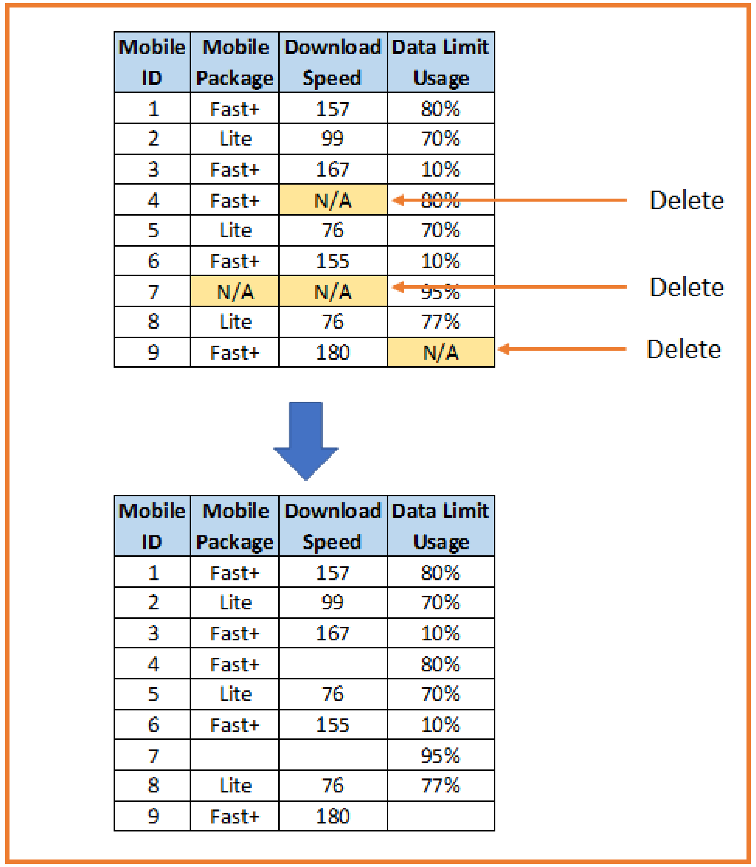
| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 8 | Lite | 76 | 77% |

# Handling Missing Values : Discard Data

**2) Pairwise (available case analysis — ACA) Deletion**

In this case, only the missing observations are ignored, and analysis is done on the variables present. If there is missing data elsewhere in the data set, the existing values are used. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion.

Pairwise deletion is known to be less biased for the MCAR or MAR data. However, if there are many missing observations, the analysis will be deficient. The problem with pairwise deletion is that even though it takes the available cases, one can’t compare analyses because they are different every time.



| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | N/A | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 | N/A | N/A | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | N/A |

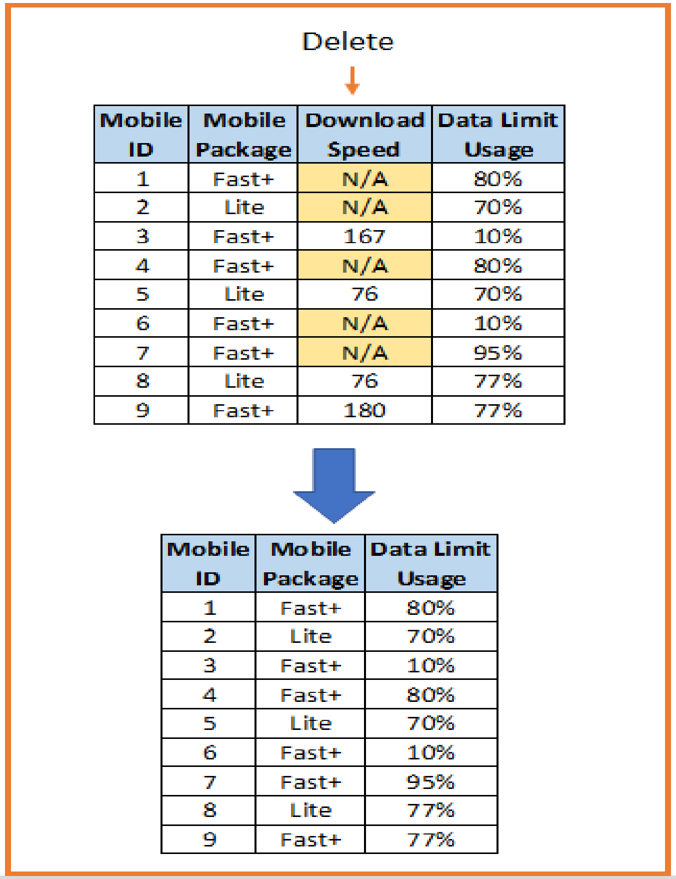
Delete
Delete
Delete

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ |  | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 |  |  | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 |  |

# Handling Missing Values : Discard Data

**3) Dropping Variables**

If there is too much data missing for a variable, it may be an option to delete the variable or the column from the dataset. There is no rule of thumbs for this, but it depends on the situation, and a proper analysis of data is needed before the variable is dropped altogether. This should be the last option, and we need to check if model performance improves after the deletion of a variable.



Delete

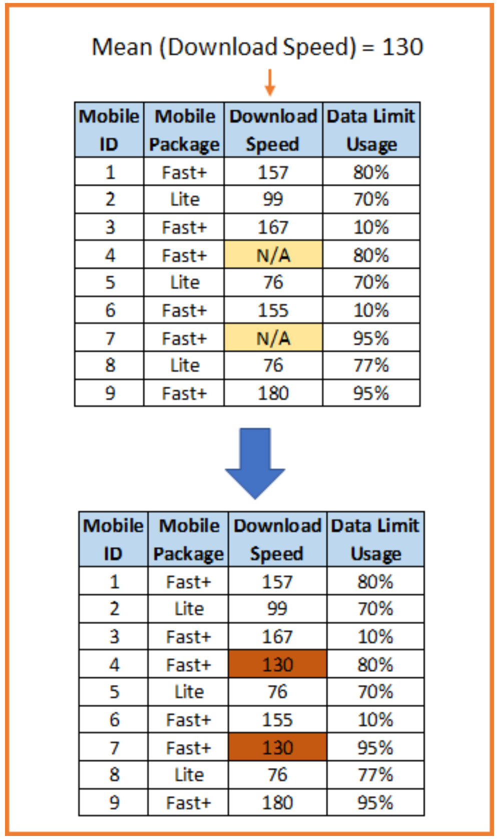
| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | N/A | 80% |
| 2 | Lite | N/A | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | N/A | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | N/A | 10% |
| 7 | Fast+ | N/A | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | 77% |

| Mobile ID | Mobile Package | Data Limit Usage |
|---|---|---|
| 1 | Fast+ | 80% |
| 2 | Lite | 70% |
| 3 | Fast+ | 10% |
| 4 | Fast+ | 80% |
| 5 | Lite | 70% |
| 6 | Fast+ | 10% |
| 7 | Fast+ | 95% |
| 8 | Lite | 77% |
| 9 | Fast+ | 77% |

# Handling Missing Values : imputation technique

## 1) Mean, Median and Mode

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable . However, with missing values that are not strictly random, especially in the presence of great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. Distortion of original variance and Distortion of co-variance with remaining variables within the dataset are two major drawbacks of this method.



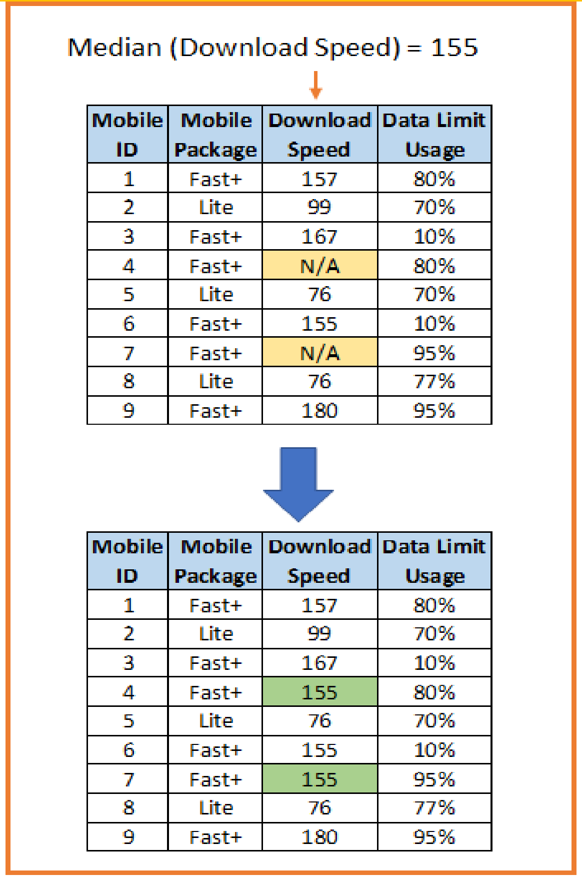
Mean (Download Speed) = 130

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | N/A | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 | Fast+ | N/A | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | 95% |

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | 130 | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 | Fast+ | 130 | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | 95% |

# Handling Missing Values : imputation technique

Median can be used when the variable has a skewed distribution.

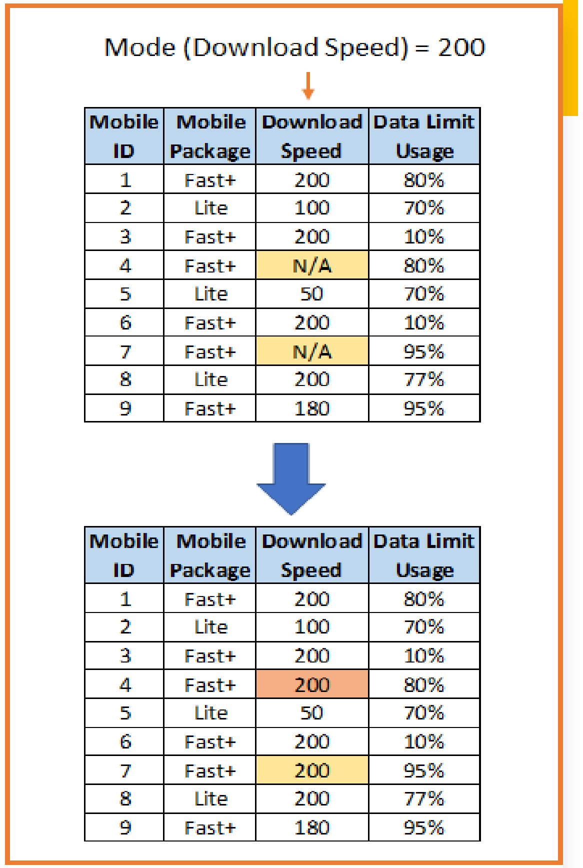


Median (Download Speed) = 155

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | N/A | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 | Fast+ | N/A | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | 95% |

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 157 | 80% |
| 2 | Lite | 99 | 70% |
| 3 | Fast+ | 167 | 10% |
| 4 | Fast+ | 155 | 80% |
| 5 | Lite | 76 | 70% |
| 6 | Fast+ | 155 | 10% |
| 7 | Fast+ | 155 | 95% |
| 8 | Lite | 76 | 77% |
| 9 | Fast+ | 180 | 95% |

The rationale for Mode is to replace the population of missing values with the most frequent value since this is the most likely occurrence.



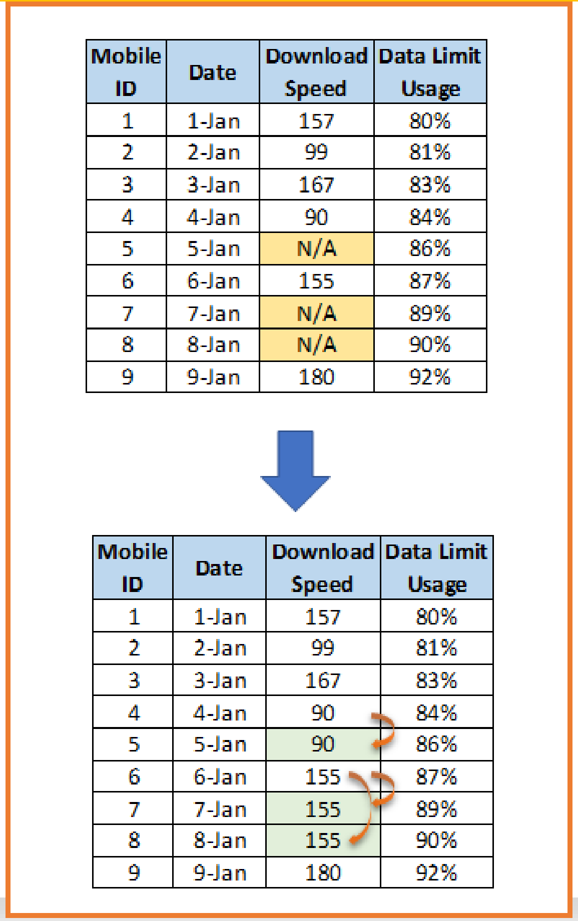
Mode (Download Speed) = 200

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 200 | 80% |
| 2 | Lite | 100 | 70% |
| 3 | Fast+ | 200 | 10% |
| 4 | Fast+ | N/A | 80% |
| 5 | Lite | 50 | 70% |
| 6 | Fast+ | 200 | 10% |
| 7 | Fast+ | N/A | 95% |
| 8 | Lite | 200 | 77% |
| 9 | Fast+ | 180 | 95% |

| Mobile ID | Mobile Package | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | Fast+ | 200 | 80% |
| 2 | Lite | 100 | 70% |
| 3 | Fast+ | 200 | 10% |
| 4 | Fast+ | 200 | 80% |
| 5 | Lite | 50 | 70% |
| 6 | Fast+ | 200 | 10% |
| 7 | Fast+ | 200 | 95% |
| 8 | Lite | 200 | 77% |
| 9 | Fast+ | 180 | 95% |

# Handling Missing Values : imputation technique

**2) Last Observation Carried Forward (LOCF)**
If data is time-series data, one of the most widely used imputation methods is the last observation carried forward (LOCF). Whenever a value is missing, it is replaced with the last observed value.

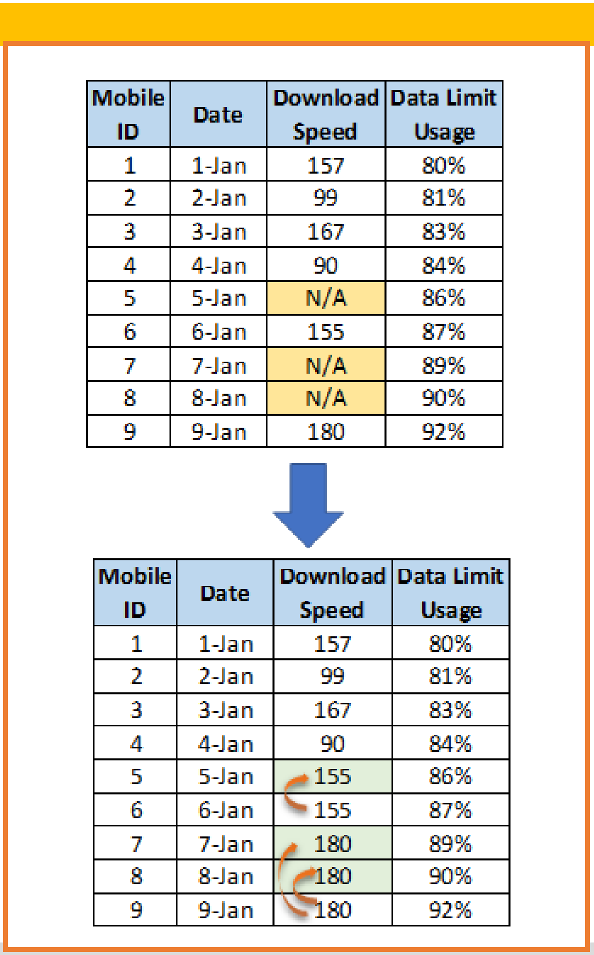


| Mobile ID | Date | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | 1-Jan | 157 | 80% |
| 2 | 2-Jan | 99 | 81% |
| 3 | 3-Jan | 167 | 83% |
| 4 | 4-Jan | 90 | 84% |
| 5 | 5-Jan | N/A | 86% |
| 6 | 6-Jan | 155 | 87% |
| 7 | 7-Jan | N/A | 89% |
| 8 | 8-Jan | N/A | 90% |
| 9 | 9-Jan | 180 | 92% |

| Mobile ID | Date | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | 1-Jan | 157 | 80% |
| 2 | 2-Jan | 99 | 81% |
| 3 | 3-Jan | 167 | 83% |
| 4 | 4-Jan | 90 | 84% |
| 5 | 5-Jan | 90 | 86% |
| 6 | 6-Jan | 155 | 87% |
| 7 | 7-Jan | 155 | 89% |
| 8 | 8-Jan | 155 | 90% |
| 9 | 9-Jan | 180 | 92% |

# Handling Missing Values : imputation technique

**3) Next Observation Carried Backward (NOCB)**
A similar approach like LOCF works oppositely by taking the first observation after the missing value and carrying it backward.



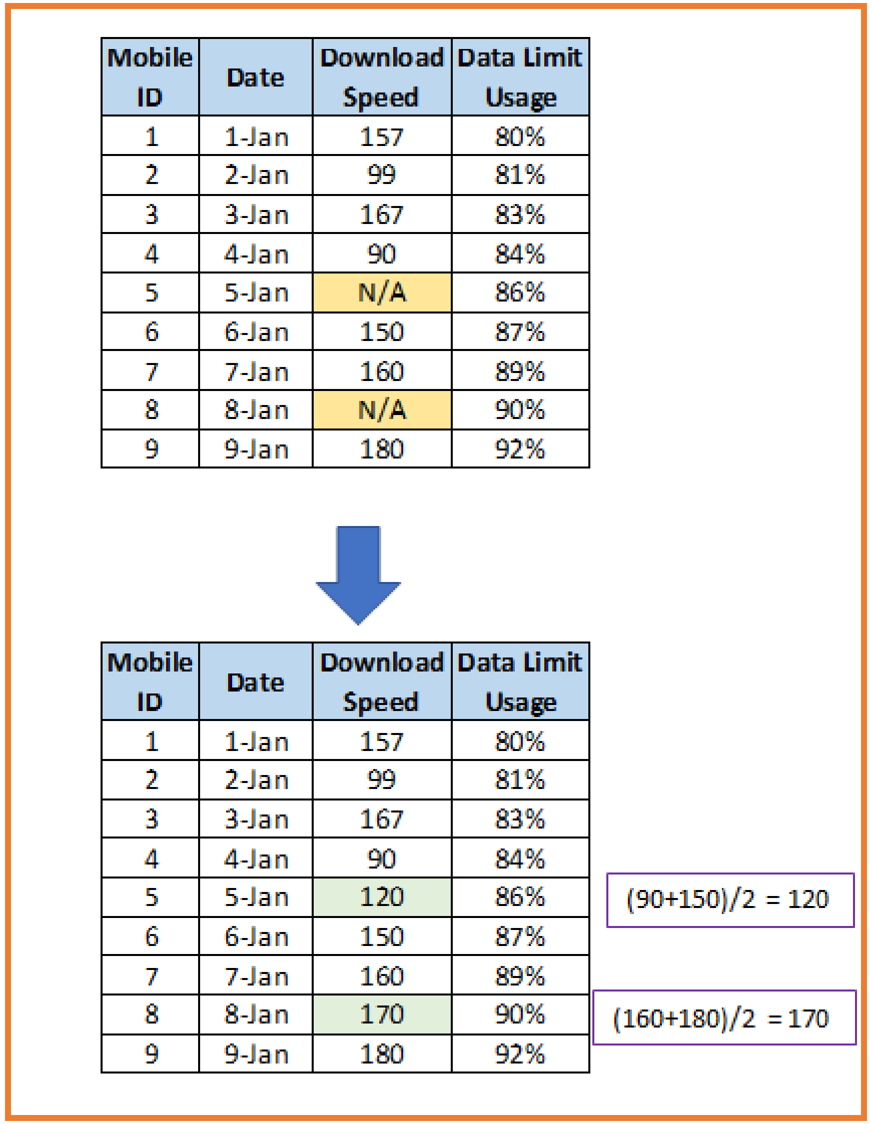
| Mobile ID | Date | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | 1-Jan | 157 | 80% |
| 2 | 2-Jan | 99 | 81% |
| 3 | 3-Jan | 167 | 83% |
| 4 | 4-Jan | 90 | 84% |
| 5 | 5-Jan | N/A | 86% |
| 6 | 6-Jan | 155 | 87% |
| 7 | 7-Jan | N/A | 89% |
| 8 | 8-Jan | N/A | 90% |
| 9 | 9-Jan | 180 | 92% |

| Mobile ID | Date | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | 1-Jan | 157 | 80% |
| 2 | 2-Jan | 99 | 81% |
| 3 | 3-Jan | 167 | 83% |
| 4 | 4-Jan | 90 | 84% |
| 5 | 5-Jan | 155 | 86% |
| 6 | 6-Jan | 155 | 87% |
| 7 | 7-Jan | 180 | 89% |
| 8 | 8-Jan | 180 | 90% |
| 9 | 9-Jan | 180 | 92% |

# Handling Missing Values imputation technique

## 4) Linear Interpolation

Interpolation is a mathematical method that adjusts a function to data and uses this function to extrapolate the missing data. The simplest type of interpolation is linear interpolation, which means between the values before the missing data and the value. Of course, we could have complex pattern in data, and linear interpolation could not be enough. There are several different types of interpolation. Just in Pandas, we have the following options like: ‘linear’, ‘quadratic’, ‘cubic’, ‘polynomial’, ‘spline’, and many more.



| Mobile ID | Date | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | 1-Jan | 157 | 80% |
| 2 | 2-Jan | 99 | 81% |
| 3 | 3-Jan | 167 | 83% |
| 4 | 4-Jan | 90 | 84% |
| 5 | 5-Jan | N/A | 86% |
| 6 | 6-Jan | 150 | 87% |
| 7 | 7-Jan | 160 | 89% |
| 8 | 8-Jan | N/A | 90% |
| 9 | 9-Jan | 180 | 92% |

| Mobile ID | Date | Download Speed | Data Limit Usage |
|---|---|---|---|
| 1 | 1-Jan | 157 | 80% |
| 2 | 2-Jan | 99 | 81% |
| 3 | 3-Jan | 167 | 83% |
| 4 | 4-Jan | 90 | 84% |
| 5 | 5-Jan | 120 | 86% |
| 6 | 6-Jan | 150 | 87% |
| 7 | 7-Jan | 160 | 89% |
| 8 | 8-Jan | 170 | 90% |
| 9 | 9-Jan | 180 | 92% |

(90+150)/2 = 120

(160+180)/2 = 170

# Handling Missing Values imputation technique

**Linear Regression**

In regression imputation, the existing variables are used to predict, and then the predicted value is substituted as if an actually obtained value. This approach has several advantages because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution.

**k-NN (k Nearest Neighbors)**

k-NN imputes the missing attribute values based on the nearest K neighbor. Neighbors are determined based on a distance measure. Once K neighbors are determined, the missing value is imputed by taking mean/median or mode of known attribute values of the missing attribute.

# Outlier Detection and Treatment

**What is an Outlier?**

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

## What causes Outliers?

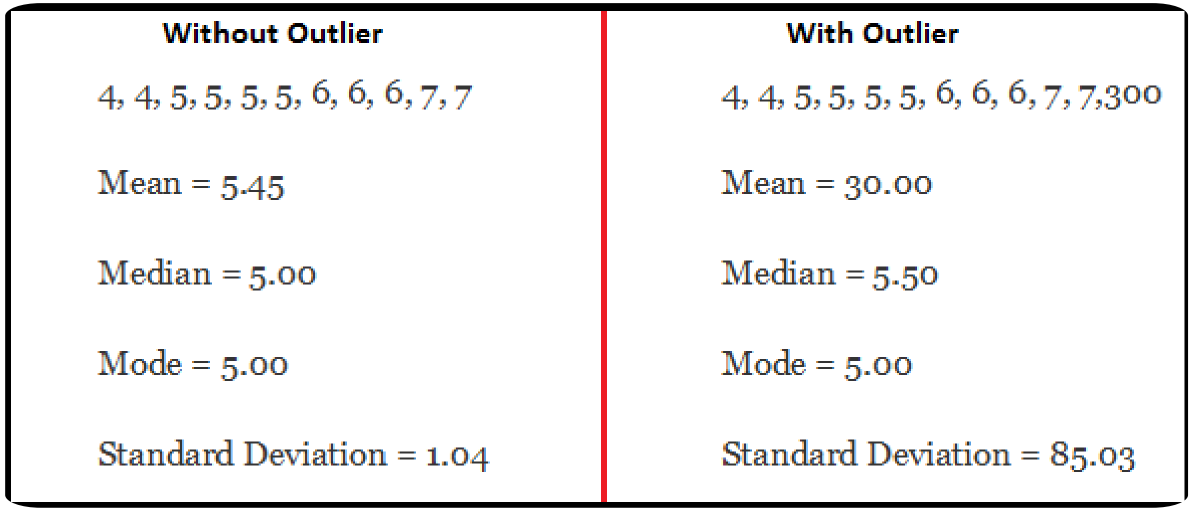
- Data Entry Errors
- Measurement Error
- Experimental Error
- Intentional Outlier
- Data Processing Error
- Natural Outlier

# Outlier Detection and Treatment

**What is the impact of Outliers on a dataset?**

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:

- 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.



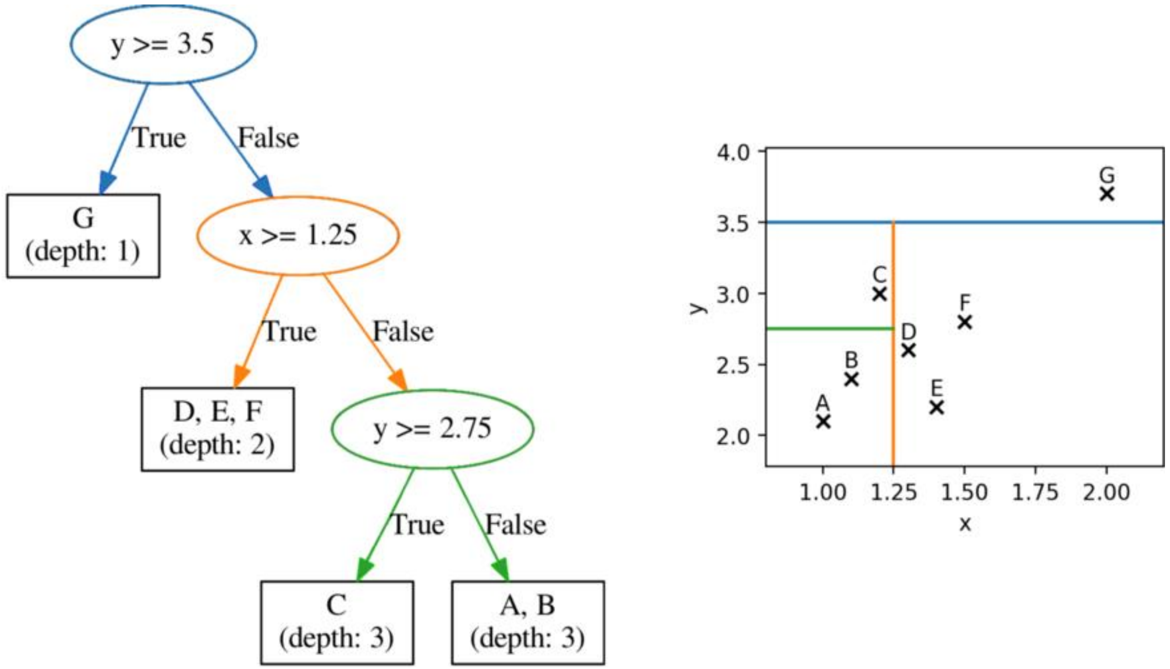
| Without Outlier | With Outlier |
|---|---|
| 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7 | 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300 |
| Mean = 5.45 | Mean = 30.00 |
| Median = 5.00 | Median = 5.50 |
| Mode = 5.00 | Mode = 5.00 |
| Standard Deviation = 1.04 | Standard Deviation = 85.03 |

# Outlier Detection and Treatment

**How to detect Outliers?**

- Most commonly used method to detect outliers is visualization. We use various visualization methods, like **Box-plot**, **Histogram**, **Scatter Plot**. Some analysts also various thumb rules to detect outliers. Some of them are:
- Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR
- Data points, three or more standard deviation away from mean are considered outlier
- Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding
- Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices such as Mahalanobis' distance and Cook's *D* are frequently used to detect outliers.

# Isolation Forest



y >= 3.5
True
False
G (depth: 1)
x >= 1.25
True
False
D, E, F (depth: 2)
y >= 2.75
True
False
C (depth: 3)
A, B (depth: 3)
4.0
3.5
3.0
2.5
2.0
y
1.00
1.25
1.50
1.75
2.00
x
G
C
F
D
B
E
A