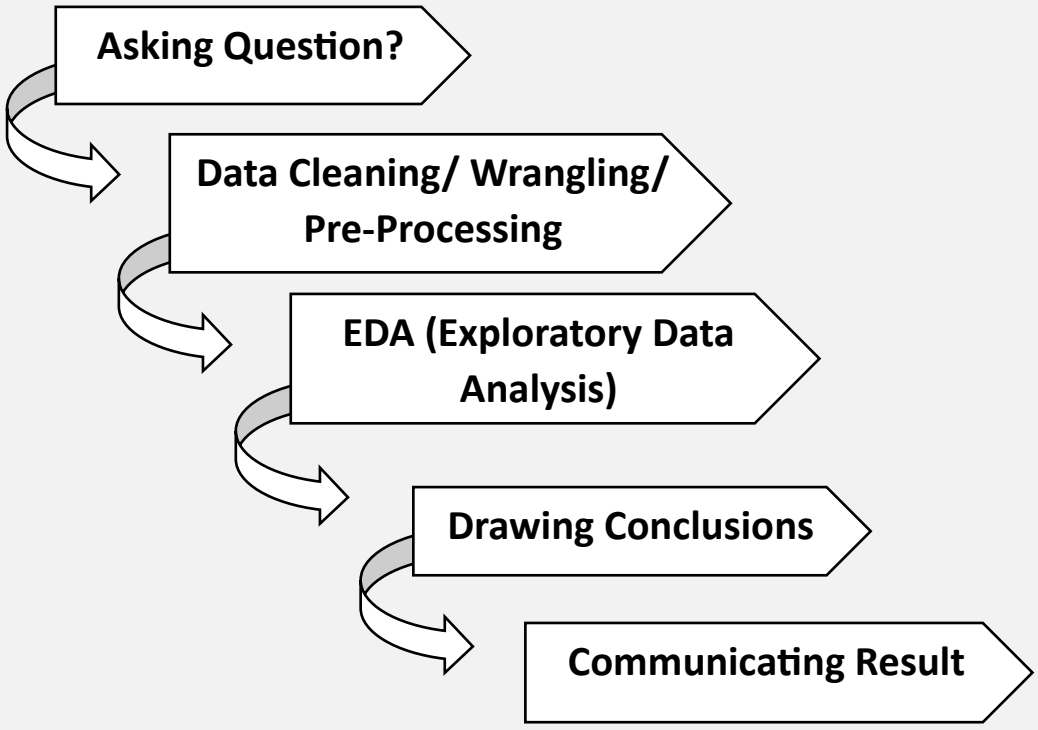
Data Analysis

Data analysis is a process to Inspecting, Cleaning, Transforming **and** Modeling Data with the goal of Discovering new and useful Information, Conclusion and Supporting Decision Making**.**

5 Steps for Data Analysis Process



Asking Question?

1. What feature will contribute to my analysis?
2. What feature will not contribute to my analysis?
3. Which of the feature have a strong correlation?
4. Do I need data pre-processing?
5. What kind of feature manipulation/engineering is required?

Data Wrangling/cleaning/ Munging

Data Wrangling is the process of transforming and mapping data from one **"Row"** data into **another format** with the **intent of making it more appropriate and valuable** for a variety of downstream purpose such as analytics.

In simple term if your data is not appropriate like: **having missing values, invalid column formatting, contains outliers, etc.** To process of **cleaning & formatting data** in a correct way is called **Data Wrangling**. There are 3 major point to perform wrangling data.

1. **Gathering Data**
2. **Accessing Data**
3. **Cleaning Data**

Gathering Data →

- CSV File
- API
- Web Scraping
- Excel
- Database

Accessing Data →

- Finding the number of rows and columns(**shape**)
- Data type of various columns(**info**)
- Check missing values (**info**)
- Check duplicate data (**is_unique**)
- Memory occupied by the dataset (**info**)
- High level mathematical overview of the data (**describe**)
- Check null value (**df.isnull().sum()**)

Cleaning Data →

- Filling Missing Data (**mean**)
- Remove Duplicate data (**drop_duplicates**)
- Incorrect data types (**astype**)

EDA

- Explore Data
- Augment Data (Manipulate data as per the requirement)

Explore Data →

- Finding correlation and covariance
- Doing univariate, bivariate and multivariate analysis
- Plotting Graphs

Finding correlation and covariance →

Covariance:

Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Where,

- xi = data value of x
- yi = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

Example:

X	**Y**
10	**40**
12	**48**
14	**56**
8	**32**

Step 1: Calculate Mean of X and Y

Mean of X (μx) : 10+12+14+8 / 4 = 11

Mean of Y(μy) = 40+48+56+32 = 44

Step 2: Substitute the values in the formula

$x_i - \bar{x}$	$y_i - \bar{y}$
10 – 11 = -1	40 – 44 = – 4
12 – 11 = 1	48 – 44 = 4
14 – 11 = 3	56 – 44 = 12
8 – 11 = -3	32 – 44 = 12

Substitute the above values in the formula

$$Cov(x,y) = \frac{(-1)(-4) + (1)(4) + (3)(12) + (-3)(12)}{4}$$

Cov(x,y) = 8/2 = **4**

Hence, Co-variance for the above data is 4

Types of Covariance:

Covariance can be classified under two types positive or negative:

- **Positive Covariance**: Indicates that two variables move in the same direction. If one variable increases, the other also increases, and vice versa.

- **Negative Covariance**: Indicates that two variables move in opposite directions. If one variable increases, the other decreases, and vice versa.

Applications of Covariance

1. Covariance is used in Biology such as in Genetics and Molecular Biology to measure certain DNAs.
2. Covariance is used in the prediction of amount investment on different assets in financial markets
3. Covariance is widely used to collate data obtained from astronomical /oceanographic studies to arrive at final conclusions
4. In Statistics to analyze a set of data with logical implications of principal component we can use covariance matrix
5. It is also used to study signals obtained in various forms.

Correlation:

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

Where:

Cov(x,y)= cov is the covariance

var(X) = standard deviation of X

var(Y) = standard deviation of Y

Example:

X	Y
10	40
12	48
14	56
8	32

Step 1: Calculate Mean of X and Y

Mean of X (μx) : 10+12+14+8 / 4 = 11

Mean of Y(μy) = 40+48+56+32/4 = 44

Step 2: Substitute the values in the formula

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

xi –$\bar{x}$	**yi – $\bar{y}$**
10 – 11 = -1	40 – 44 = – 4
12 – 11 = 1	48 – 44 = 4
14 – 11 = 3	56 – 44 = 12
8 – 11 = -3	32 – 44 = 12

Substitute the above values in the formula

Cov(x,y) = (-1) (-4) +(1)(4)+(3)(12)+(-3)(12)

_____

4

Cov(x,y) = 8/2 = **4**

Hence, Co-variance for the above data is 4

Step 3: Now substitute the obtained answer in Correlation formula

$$Correlation = \frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

Before substitution we have to find standard deviation of x and y

Lets take the data for X as mentioned in the table that is 10,12,14,8

To find standard deviation

$$s = \sqrt{\frac{\Sigma\,(X - \bar{x})^2}{n - 1}}$$

Step 1: Find the mean of x that is x̄

10+14+12+8 /4 = 11

Step 2: Find each number deviation: Subtract each score with mean to get mean deviation

10 – 11 = -1
12 – 11 = 1
14 – 11 = 3
8 – 11 = -3

Step 3: Square the mean deviation obtained

-1	**1**
1	**1**
3	**9**
-3	**9**

Step 4: Sum the squares

1+1+9+9 = 20

Step 5: Find the variance

Divide the sum of squares with n-1 that is 4-1 = 3

20 /3 = 6.6

Step 6: Find the square root

Sqrt of 6.6 = 2.581

Therefore,

Standard Deviation of x = 2.581

Find for Y using same method

The Standard Deviation of y = 10.29

Correlation = 4 / (2.581 x 10.29)

Correlation = 0.15065

Types of Correlation

1. Simple Correlation: In simple correlation, a single number expresses the degree to which two variables are related.
2. Partial Correlation: When one variable's effects are removed, the correlation between two variables is revealed in partial correlation.
3. Multiple correlation: A statistical technique that uses two or more variables to predict the value of one variable.

Applications of correlation

1. Time vs Money spent by a customer on online e-commerce websites
2. Comparison between the previous records of weather forecast to this current year.
3. Widely used in pattern recognition
4. Raise in temperature during summer v/s water consumption amongst family members is analyzed
5. The relationship between population and poverty is gauged

Methods of calculating the correlation

1. The graphic method
2. The scatter method
3. Co-relation Table
4. Karl Pearson Coefficient of Correlation
5. Coefficient of Concurrent deviation
6. Spearman's rank correlation coefficient

Doing univariate, bivariate and multivariate analysis →

- **Univariate analysis focuses on understanding individual variables.**

Purpose: Univariate analysis is primarily used to summarize and visualize the distribution of a single variable, assess its central tendency (mean, median, mode), dispersion (range, variance, standard deviation), and shape (e.g., histogram, box plot). It helps in understanding the characteristics of a single variable in isolation.

- **Bivariate analysis examines relationships between two variables.**

Purpose: Bivariate analysis is used to understand the relationship, correlation, or association between two variables. Common techniques for bivariate analysis include scatter plots, correlation coefficients (e.g., Pearson's correlation), and contingency tables (for categorical variables). It helps answer questions like, "Is there a relationship between a person's age and their income?"

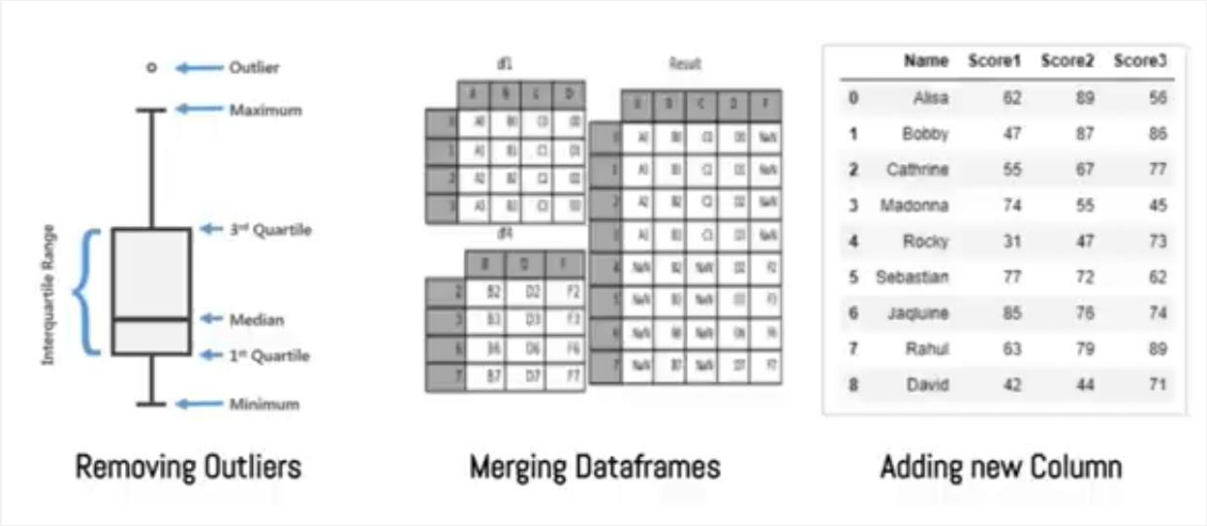
- **Multivariate analysis deals with the interactions and relationships among three or more variables.**

Purpose: Multivariate analysis is used when there are multiple variables at play, and researchers want to explore how these variables interact and influence each other. It encompasses various techniques, including multiple regression, principal component analysis (PCA), factor analysis, cluster analysis, and multivariate analysis of variance (MANOVA). Multivariate analysis is essential for understanding more complex relationships in data, such as predicting an outcome based on multiple predictor variables or clustering similar observations based on multiple characteristics.

Augmenting Data →

- Removing Outlier
- Merging Data Frame
- Adding new columns

These operations are collectively called as feature engineering.



	Name	Score1	Score2	Score3
0	Alisa	62	89	56
1	Bobby	47	87	86
2	Cathrine	55	67	77
3	Madonna	74	55	45
4	Rocky	31	47	73
5	Sebastian	77	72	62
6	Jaqluine	85	76	74
7	Rahul	63	79	89
8	David	42	44	71

Draw Conclusions

- Doing Prediction
- Draw conclusions on the basis of questions.

Some example conclusions based on Descriptive Statistics

1. Is Rohit Sharma a better batsman in 2nd innings (IPL Dataset)?
2. Does being a female increases your chances of Survival (Titanic Dataset)?
3. Is Delhi the most costly place for eating out(Zomato Dataset)?

Communicating Result / Data Storytelling

- **In person**
- **Reports**
- **Blog Post**
- **PPTs/Slide Decks**