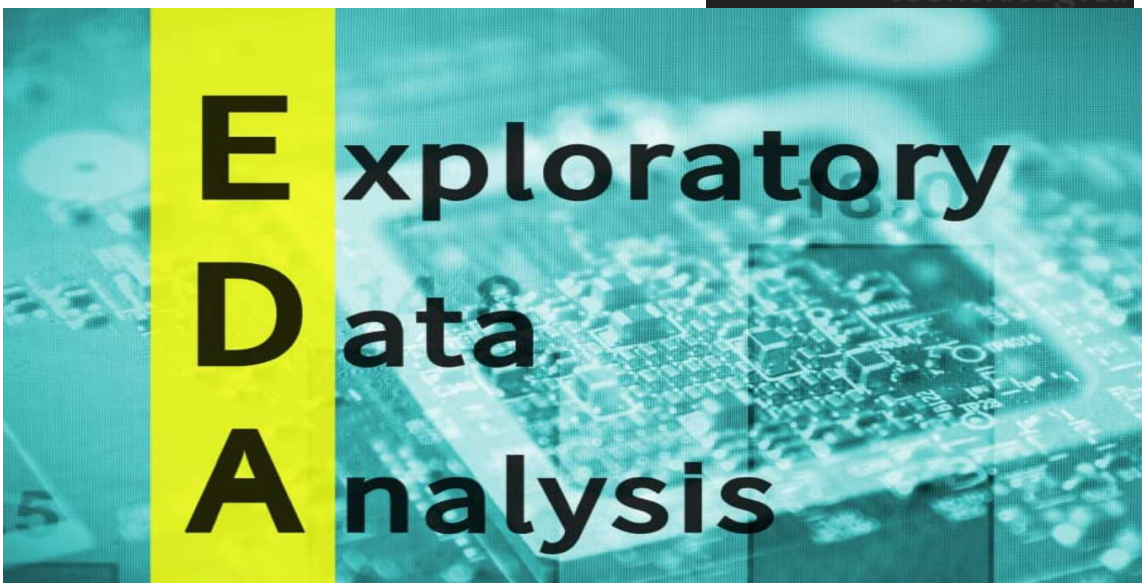


WORKSHOP ON EDA (EXPLORATORY DATA ANALYSIS)

AGENDA:

1- What is EDA & Deep understand about EDA
2- What is Raw Data
3- Missing Value treatment
4- What is Clean Data
5- Create Clean dataset for predictive model
6- Variable Identification
7- Univariate Analysis
8- Bivariate Analysis
9- Outlier Detection
10- Variable Creation

PRACTICLE:

11- Select Dataset
12- We will perform Data preparation
13- We will do Data cleaning using PANDAS, NUMPY
14- Perform Exploratory Data Analysis using MATPLOTLIB, SEABORN
15- Summarization & Documentation

What is EDA →

- EXPLARATORY DATA ANLYSIS (EDA) it is a data exploration technique to understand the various aspect of the data.
- Analyse the data various way – Excel, Python, R, BI tool – Tableau
- Excel data visualization is limited & for bigdata you can't visualize the data in excel
- That's why we need to go for one programming language
- Today we build EDA project end-to-end with practical
- Tableau you can visualize when the data is cleaned
- BI tools you can't visualize RAW DATA
- Let's discuss What is Raw data

RAW DATA →

- The data that has not been processed before.
- Sometime it also called as source data.
- Data directly comes from customer, weblog, html, xml etc.
- Data that is collected directly from the source and hasn't been processed, organized, cleaned or visually presented is considered raw data

Name	Domain	Age	Location	Salary	Exp
Mike	Datascience#$	34 years	Mumbai	5^00#0	2+
Teddy^	Testing	45' yr	Bangalore	10%%000	<3
Uma#r	Dataanalyst^^#			1$5%000	4> yrs
Jane	Ana^^lytics		Hyderbad	2000^0	
Uttam*	Statistics	67-yr		30000-	5+ year
Kim	NLP	55yr	Delhi	6000^$0	10+

MISSING VALUE TREATMENT →

- Can't visualize the data if the data has many missing values.
- Data categorized into 2 type – Numerical Data & Categorical Data
- If numerical Data is missing – then we will implement MEAN, MODE, MEDIAN strategy
- If categorical data is missing – then we will implement Mode Strategy

Numerical Data	
Missing Data	Fill Missing Data
10	10
20	20
	20
30	30

Categorical Data	
Missing Data	Fill Missing Data
Summer	Summer
Winter	Winter
	Winter
Rainy	Rainy
Winter	Winter

CLEAN DATA →

- Clean data technique also called as DATA CLEANSING which is very important in DATA ANALYST & DATA SCEINTIST journey.
- Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Name	Domain	Age	Location	Salary	Exp
Mike	Datascience	34	Mumbai	5000	2
Teddy	Testing	45	Bangalore	10000	3
Umar	Dataanalyst	50	Bangalore	15000	4

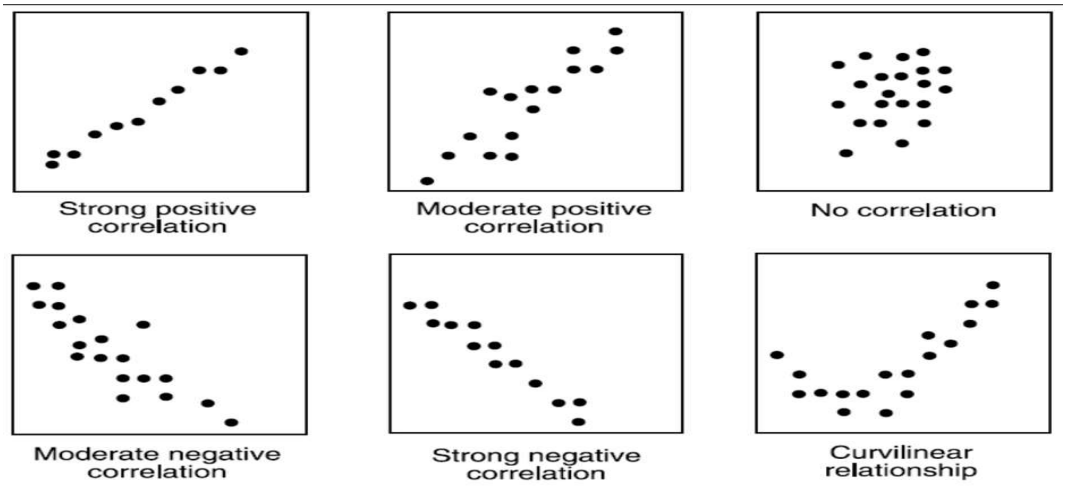
VARIABLE IDENTIFICATION →

- Variable identification in the dataset is very very important to choose right machine learning algorithm. Variable or attribute or feature are split into 2 types →
  - Independent Variable | Non target variable | Non predicted variable
  - Dependent Variable | target variable | Predicted variable

I.V	I.V	I.V	I.V	I.V	I.V	I.V	I.V	I.V	D.V
NAME	SFT	OFFICE	SHOPING	SCHOOLS	METROS	PLAYGROUP	SALARY	HOME	Purchase
XYZ	100							Z	Y
BZC	200	IY	N	N		Y	N	Y	N
ASDF	300		Y	Y	Y		200		y

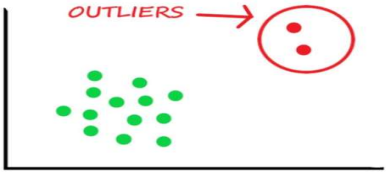
UNIVARIATE & BIVARIATE ANALYSIS →

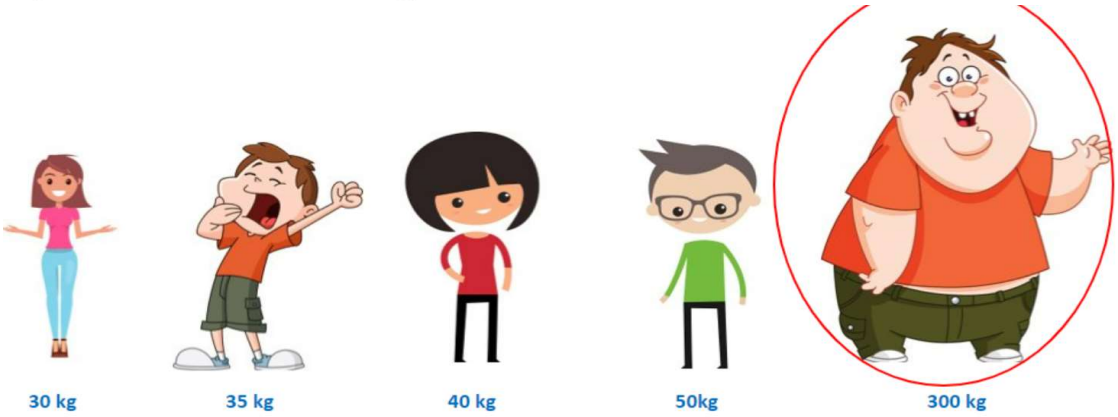
- Visualize the graph using one variable is called Univariate Analysis
- Visualize the graph using 2 variable is Bivariate Analysis
- Visualize the graph using more than 2 variable or many variables is Multivariate analysis
- Relation Between 2 variable – **CORELATION**
- Below is the pattern of corelation. Corelation is ranging from -1 to 1
- 0 to 1 → Positive corelation
- -1 to 0 → Negative Corelation
- 0 → No Corelation



OUTLIER DETECTION →

- Outlier detection also called as Anomaly Detection. This is very important in model building
- Outlier will impact the many classifications algorithm. Eg: Logistic Regression & KNN algorithm
- In statistics, an outlier is a data point that differs significantly from other observations.

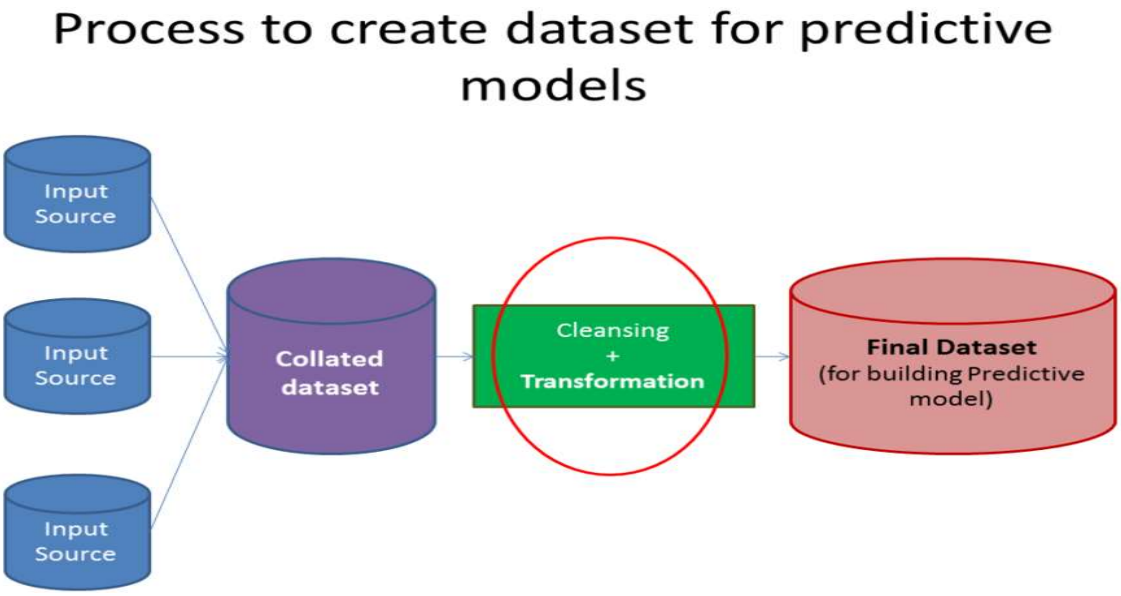




VARIABLE CREATION →

- While we build Machine learning model machine does not understand Sunny, Winter, Rainy.
- Let's impute categorical data to Numerical data before we training the data.
- Variable Creation also we called Encoding Technique
- Dummy variable is one of encoding Technique

DUMMY VARIABLE			
SEASON	SUNNY	RAINY	WINTER
SUNNY	1	0	0
RAINY	0	1	0
WINTER	0	0	1
SUNNY	1	0	0
WINTER	0	0	1
RAINY	0	1	0
RAINY	0	1	0



WHAT IS THE NEXT STEP →

- Next step is Machine Learning model building
- Then test the model accuracy
- Pass the future data to the model & model generate future prediction
- 1st level of test case to compare future prediction with live data
- Continue 3 level of test
- If all test cases are pass then we need to go for deployment
- After deployment we need to retrain the model with new data
- Final step is to Automize the ML model

SUMMARIZATION →

In this workshop we understand

- What is raw data
- How to clean data using python package - Pandas, NumPy
- How to visualize the data using package - matplotlib, seaborn
- How to convert one data type to other data type
- Univariate analysis
- Bivariate analysis
- Variable identification -- Independent variable & Dependent variable
- Outlier treatment
- How to fill missing numerical value & categorical treatment
- Variable creation with the help of dummy variable
- Learned live coding technique for data cleaning
- Above steps must require before to build machine learning model building
- We created cleaned dataset.

****

- **To understand in-depth & detailed concept please enroll my live classes.**
- **Thank you, Team, for joining today's workshop.**

****