EDA AND Feature engineering:-

Data Science Life Cycle:-
1. Data Ingestion.
2. EDA (Analysis)
3. Processing (Preprocessing)
4. Model Building.
5. Evaluate & Validate.

EDA:- Exploratory Data Analysis

Statistics:-
- Collect the Data.
- Organise the Data.
- Interpretation.
- Analysis of Data.

Kaggle is used to get different types of datasets.
- Titanic dataset.
- Diabetes dataset.

(1) Data Ingestion:-

Get data from:-

-> Big data tools
- Data can be at HDFS [Hadoop distributed File System]
- No SQL Database.
- Kafka [Streaming Data]
- Spark Streaming

-> Remote location - SQL, NoSQL

-> Some File Format :- csv, tsv, xmL, json, Excell

-> Scrap data from Website.

Types of data :- Tendency of Data ↓

Batch Data, Streaming data.

↓ Historic Data (Periodic Data)

↓ Continuous data

↓ Mini Batch Data [little more freq]

Data Wheather it is Batch Data or Streaming data can be divided into two parts.

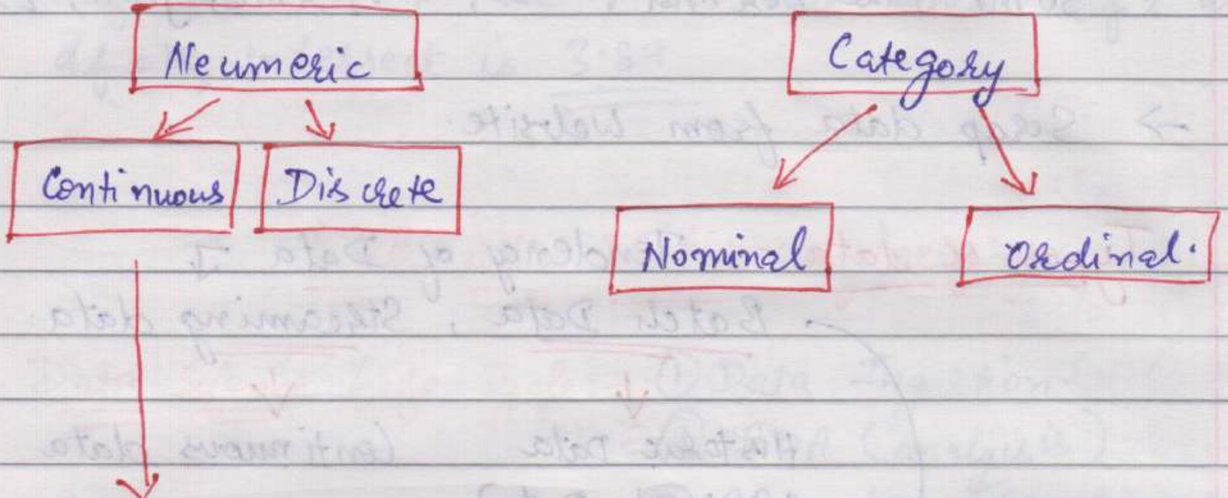
① Structured data. -> Table.
② Unstructured data. -> video, images, voice, text
③ Semi Structured data -> xml, json.

Structure Data :-

①	②	Feature ③
Weight	Height	BMI
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21

Continuous continuous Continuous.

Structure data can be divided into two parts -
-> Neumeric data.
-> Category data.



Continuous [means Neumerical data that can have decimal value]

Eg:- Height [160, 160.5, 160.55]

Discrete data means no decimal value. [Whole No.]
10, 100, 200 students in st class.

Category → male   Black
Female   White.

Nominal:→ Order does not matter.
male
Female } Order does not matter.

Ordinal :→ Order matters.
Example:- Degree:-

First we do 10th → 12th → Gradution → Post Graduation → Phd.

Dataset Student Performance:-

Name	Age	Height	Sex	Weight	Education.
Sunny	25	170	Male	70	UG
Arijit	30	1.80	Male	80	PG
Priyam	35	160	Male	60	UG
Priya	20	150	Female	55	PHD
Aditi	27	145	Female	58	PG

↑ ↑ ↑ ↑ ↑ ↑

Categorical Numerical Num Cat Num Cat.

↓ ↓ ↓ ↓ ↓ ↓

Nominal Continuous Continuous Nominal Continuous Ordinal.

Univariate:- Single Column - { If we want to check height then it is univariate }

Bivariante:- Two columns { If we want to check height with respect to Age then it is bivariante }

Multivariate:- More the two columns.

- If we want to check height and Age with respect to Sesc then it is multivariate.

Independant / Dependant

Suppose we have Age, height, Sex of a person and we can define weight by knowing Age, height, Sex.

So weight → Dependant

Age, Height, Sex → Independent.

Core ML Pipeline →

① Pata Ingestion
② EDA → Analysis.
③ Preprocessing → Feature Engineering
④ Model Building
⑤ Evaluation or Validation of model.

EDA → Preprocessing ⟹ model.

↑ Will impact

EDA:- Based on the given feature, we are going to perform the analysis of the data.

Preprocessing / Feature Engineering:-
- Cleaning of the Data.
- Rengling of the Data
- Preparing of the Data.

Is Preprocessing and Feature Engineering is same?
- Yes.

Name	AGE	Education	Salary	Experience.
Sunny	25	UG	25K	2
Deepak	30	PG	30K	3
Rushi	40	UG	40K	5
Priyam	50	PHD	50K	10
Shalini	20	UG	35K	1

EDA (Analysis) →
① Profile of the data.
② Statistical Analysis.
③ Graph based analysis.

Profile of the data:-

1. No. of Rows.
2. No. of Columns.
3. Missing values.
4. How many categorical column
5. How many Numerical column.
6. Is there Duplicate value.
7. Dtype.

Statistics based Analysis:- (Interpretation)

1. Variance of the column.
2. CoVariance of the column.
3. Standard Deviation.
4. Correlation of the data betn two column.
5. Perform Chi square test
6. Perform t-test.
7. Perform Z-Test.
8. Perform Anova Test.
9. mean / median / Mode.

Graph based analysis

1. Box plot.
2. Scatter plot
3. Pie chart.
4. Histogram.
5. KDE
6. Count bar.
7. Heat map.

Box Plot:- With the help of box plot we can find the outlier, distribution.

Count Bar:- Check how many rows and column is there

Heat Map:- We can check the correlation.

Histogram:- We can check the distribution.

Scatter Plot :- We can check the outlier of the data,
we can check data is linear or not.

By EDA We can [Preprocessing]
- Handle the missing value.
- Handle the outlier
- Scalling of data
- transformation (log, Box cox, square, Cube)
- encoding
- We can handle imbalance data.
- Feature Selection
- We can do Dimension reduction [PCA, tSNE)

Automated tool in Python For EDA.
- Pandas Profiling.
- mito
- Knime.

Books For Feature Engineering:-
① Feature Engineering and selection : A Practical Approach for Predictive Models.
② Python Feature Engineering Cookbook.
③ Feature engineering for machine learning.