EDA (Exploratory data Analysis)

+ FE (Feature Engineering)

Data Science lifecycle.

1. Data ingestion
   - big data tools. (HDFS, NoSQL, Kafka, spark)
   - remote location (sql, nosql)
   - Some file format (csv, xml, xlsx)
   - Website.

2. EDA (analysis)

3. Pre-processing

Involve Stats.

Machine Learning

4. Model building. ⟶ Mathematical process.

5. Evaluate & validate.

Statistics : Collect, organize, interpret, analyse data.

Insights.

↳ scientific, healthcare, social problem.

↳ Sales of Product ⟶ Sales is going down.

↓

Product, Paying attention to customer.

leadership, marketing.

competitor

↳ dataset(s) created.

1. Project Manager.
2. Buriness Analyst. Analysis. ⟶ Conclusion (Why sale is down?)
3. Data Scientist

↳ Domain expert.

Types of data : Batch data, streaming data.

Batch data → mini batch data. (little move - frequent)

Batch data → Historic data. (Periodic)

streaming data → Continuous data → LIVE.

(Machine Learning) → ML

DL. (Deep Learning)

1. Structured data. → table
2. Unstructured data. → videos, images, voice, text.
3. Semi-structured data. → xml, json

Structured data

- Numeric
  - Continuous
  - Discrete
- Category
  - Nominal
  - Ordinal

(Cont.)	(Cont.)	(Cont.)
kg.	ht.	BMI.
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21.
feature =>	1	2	3.

Continuous data ⟶ Continuous in nature.

eg. Height, weight.

Discrete data ⟶ Only whole numbers.

eg. students in a class.

Nominal ⟶ Order does not matter.

eg Male/Female.

Ordinal ⟶ Order matters.

eg. 10th,
12th,
UG,
PG

Student Performance dataset.

→ Features.

Name	Age	Height	Sex	Weight	Education.
Sunny	25	170	Male	70	PG.
Arijit	30	180	Male	80	PG
Priyam.	35	160	Male	60	UG.
Priya	20	150	female	55	Phd.
Aditi	27	145	female	58	PG.
Cat.	Num	Num.	Cat	Num	Cat.
Nominal	Cont.	Cont.	Nominal	Cont.	Ordinal

EDA. ⟶ type of data.

UG - 0

PG - 1.

Phd - 2.

Univariate. ⟶ One column (feature).

e.g Height.

Bivariate. ⟶ Two columns.

e.g Height vs Age.

Multivariate. ⟶ >2. columns.

e.g Sex vs Height, Weight.

Independent / Dependent variables.

Age, Height, Sex.

Independent variable.

define.

Height.

Dependent variable

Dependent on other variable

Pipeline of DS:

1. Data Ingestion.
2. EDA.
3. Preprocessing. (-feature engineering)
4. Model building.
5. Evaluation or validation of model.

-feature / column.

1. Missing value handling
2. Outlier.
3. Scaling of data.

} involves changing data. (-feature engi.)

Q: First EDA is required or FE ?

1

2

EDA

Pre-processing }.

eg -

Biryani. (Model)

Row data.

1. Chicken.
2. Rice.
3. Onion.
4. Oil.
5. Spices.

EDA.

1kg → clean.
1kg → Preprocess.
1kg → Cut.
1 ltr.
250 gm.

ML in real life

Validation. ← Model.

Tasting.

Biryani.

Dataset:

NAME	AGE	EDUCATION	SALARY	Exp.
Sonny	25	UG	25K	2
Deepak	30	PG	30K	3.
Ruchi	40	UG	40K	5.
Aman	50	Phd.	50K	10.
Shalini	20	UG	35K	1.

(1.) EDA (analysis).

Steps.

(i) Profile of the data.

(ii) Statistical analysis.

(iii) Graph based analysis.

Profile of data.

1. No. of rows.
2. No. of cols.
3. No. of missing.
4. Variable type.
   - Cat
   - Num.
5. How many duplicates
6. Data types.
7. RAM used.

Stats based interpretation.

i. Variation.
ii. Covariance.
iii. Standard deviation.
iv. Correlation.
v. Chi-square.
vi. t-test.
vii. Z-test.
ix. Anova-test.
x. Mean/Median/Mode.
xi. Skewness.
xii. Kurtosis.

Graph based analysis. (Plotting)

i. Box plot. → outliers, distribution.
ii. Scatter plot → Outliers, linearity.
iii. Histogram. → distribution.
iv. KDE. → distribution.
v. Count bar. → Rows, columns.
vi. Heatmap. → Correlation.

* Based on EDA we can do processing of data.

Preprocessing / Feature Engineering steps:

1. Missing data handling.
2. Outlier handling.
3. Scaling of data.
4. Transformation (log, boxcox, square, cube)
5. Encoding of categorical data.
6. Handle imbalance data.
7. feature selection.
8. Dimension reduction. (PCA, tsnE)

Automated tool for EDA in python:

1. Pandas profiling
2. Mito.
3. knime.