+ Hare Krishna +

EDA & feature engineering:

Data Science Life cycle:

1. data injection (collection)
2. EDA (analysis) (Stats)
3. Processing (prc) (math)
4. Model building
5. evaluate & validate

machine learning → 4, 5

EDA:

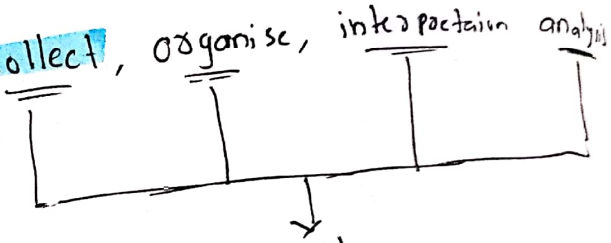
↓ Exploratory data analysis

it often involves the use

* graphs
* chats
* summary statistics
* and various data visualization

Statistics:

collect, organise, interpretation analysis



Problem statement:

Scientific, healthcare, Social Problem

Sales of product: → Sales is going down.

reasons may be ⇒ product, Price, marketing, competitor... etc

We have to prepar a → dataset → analysis → conclusion.

1. Project manager
2. Project Business analyst ↔ domain expert
3. data scientist.

data can be present in → hadoop, NoSql, kafka, Spark Streaming (HDFS) → hadoop distributed file system.
→ (Bigdata tools), remote location (Sql, NoSql)
→ some file formate (csv, tsv, xML, json, excel ... etc)
→ websites.

types of data:
→ Batch data, Streaming data
- Batch data → mini batch data (little more freq), historical data (periodic)
- Streaming data → continiues data

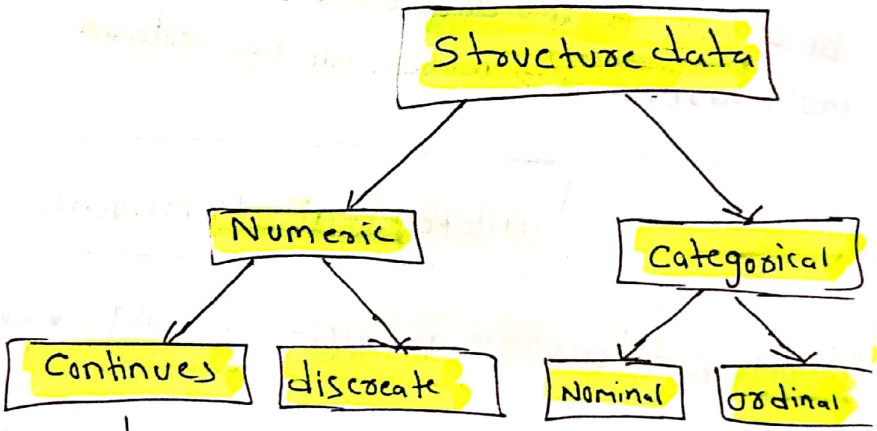
1. Structure data :- table formate
2. unstructure data :- video, image, voice, sound, text → deep learning
3. Semi structure data :- xML, json

Structure data:

weight (kg)	height (cm)	BMI
70	170	22
80	180	24
90	140	26
100	200	30
60	160	21

height → continues:
like 170, 17.5, 7.65 ... etc

discreate means in a school they are
10, students
20 students
45 students
but we cannot say 25.5, 30.5
so it is standard.



Categorical → male, female → black, whik
nominal → order doesnot matter
red, yellow, back, green
ordinal → meaning full order or rank.
Ex: class ranking: 1st class, 2nd class, 3rd class etc.

Ex:- data sets

assume that

→ Uni-variant
→ bi-variant
→ multivariant

Name	Age	weight	Sex	height	Education
Naveen	25	70	male	170	UG
Soikanth	24	60	male	160	PG
Varun	35	50	Male	150	UG
Sai	20	70	Male	170	Phd
Raghava	21	76	Male	176	PG
Priya	32	80	Male	180	PG

Catagorical → Nominal
Numerical → continues
Numerical → continues
Catagorical → nominal
Numerical → continues
Categorical → (ordinal)

types of data (we have to check)

UG → 0
PG - 1
Phd - 2
} level

Uni-variate → single column
bi-variate → double column
multivariate → more than two column.

independent / Dependent variable

independent variable:- if any variable or value are independent which means not dependent on any value or variable it called independent Ex:- Age, Sex

dependent variable:- opposite of independent variable

ex:- weigh, height

Core ML pipeline

1. Data ingestion
2. EDA (Exploratory data analysis)
3. Preprocessing ⇒ [nothing but feature engineering]
4. model building
5. Evalution or validation of model.

⇒ Data → we have to analysis

feature / column
- missing value
- outlier
- scaling

→ change

↓

feature engineering

first → EDA is required

↓

Feature engineering

↓

Pre process

again check the changed data.

again check the changed data.

Practical Example:

Suppose we are doing Party and we want to prepare Biryani

Biryani (Model)

Raw data
1. chicken
2. rice
3. oil
4. onion
5. Species

EDA
- 1-kg
- 1-kg
- 1-liter
- 1-kg
- 500gm

→ Pre-Processing ↓ cut & wash

→ model-building ↓ cook Biryani

↳ (test) ↓ Validation

⇒ EDA → analysis of data

⇒ Preprocessing (or) feature Engineering → We can perform as many times as we wish

⇒

uni-variate
bi-variant
multi-variant

Name	Age	education	Salary	Emp
Krishna	25	UG	90 lc	2
Rama	30	UG	80 lc	3
Hare	40	PG	96 lc	5
govinda	50	Phd	87 lc	10
jagannadh	20	UG	180 lc	6

① EDA

(analysis)

① Profile of the data
② statical analysis
③ Graph based analysis

Profile of the data

① ROW
② column
③ missing
④ catagorical
⑤ numarical
⑥ duplication
⑦ Dtype
⑧ RAM

graph based (analysis)

plotting
data analysis
dashboard

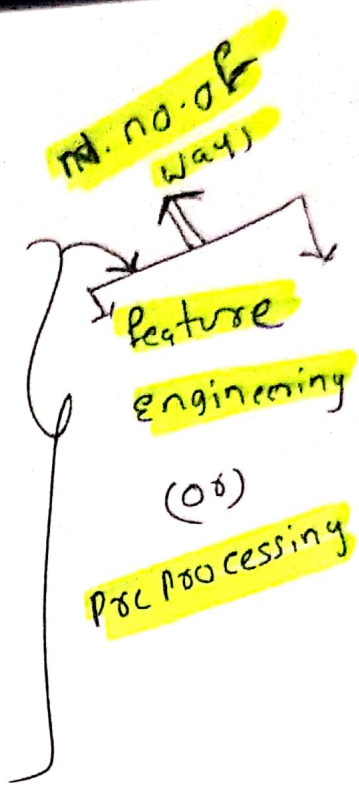
① Box Plot → outlier, distribution, statistical conclusion
② Scatter Plot → outliers
③ histogram → distribution
④ kernal density estimation (KDE)
⑤ count bar → Rows, Column
⑥ heat map → correlation

stats based (anterpration)

① Variance
② Co-variance
③ stand
④ Co-orelation
⑤ chi square - test
⑥ t-test
⑦ z-test
⑧ anova test
⑨ mean / median / mode

⇒ Based on a EDA we can do a processing of the data.

1. missing value handle
2. outlier handle
3. scaling of data
4. transformation (log, Box, square, cube)
5. en-coding
6. imbalance data
7. feature selection
8. Dimention reduction (PCA, ESNE)



missing null value → missing value

EDA → pre Processing

outlier → handling

catag (man, woman) → encoding

skewed range → scale (with a certain range)

5 → count of feature → handle imbalance
→ feature selection
→ dimension deduction.

Encoding:- Converting the catagorical values to numarical value is called en-coding.