DAY 2 Statistics
EDA & Feature engineering

EDA & FEATURE ENGINEERING

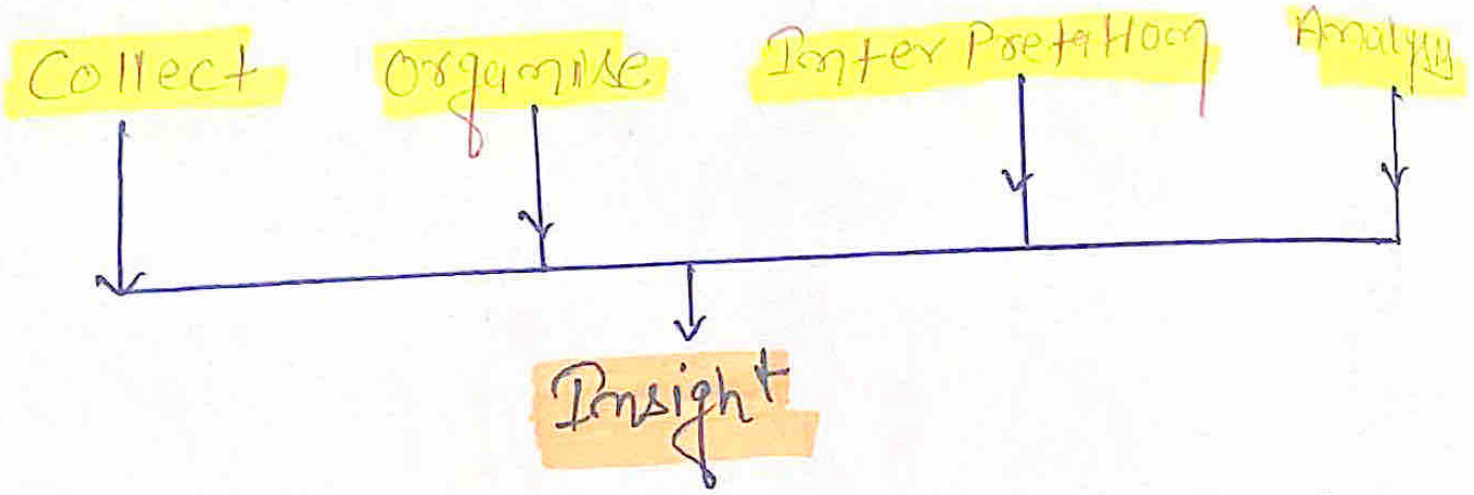
DATA Science life cycle : [ML Life cycle]

1. Data ingestion (Collection of DATA)
2. EDA (Analysis over DATA)
3. Pre-Processing or feature engineering
4. Model Building
5. Evaluate & validate model

} ⇒ CORE ML PIPELINE

Statistics:

Statistics is the Science of collecting, Organising & analysing the Data



Examples :- Scientific Problem, Healthcare
* every where Statistics use

Problem Statement :-

Sales of Product → Sales is going Down

Analysis → Product, Paying Attention to Customer leadership, Marketing, Competitor.

Dataset → Analysis → Conclusion

1. Project Manager
2. Business Analyst → Domain expert
3. Data Scientist → used to get conclusion

Data ingestion: Collecting Data

Examples: Big Data tools, Remote Location (SQL, NOSQL), web Scrapping

Some file Format: (CSV, XML, JSON, xls)

Types of DATA:-

BATCH DATA:- Historical Data, Mini batch Data (Periodic)

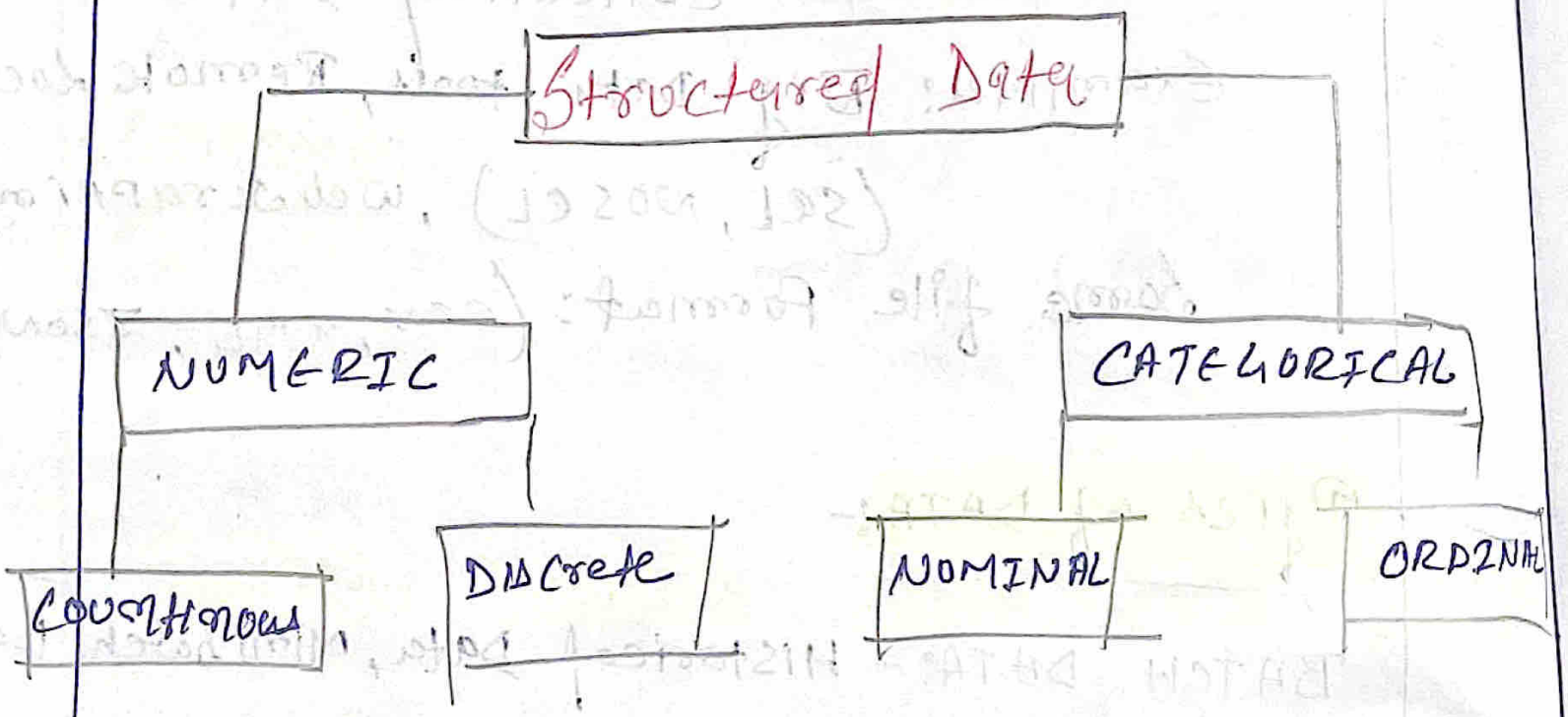
Streaming Data:- COUNTINOUS DATA (Live DATA)

1. Structure → Table (Row & Column) → ML
2. UNstructure → Video, Images, voice, Sound, Text } → DL
3. Semi-Structure → JSON, XML

* Example of Structured Data

Feature 1	Feature 2	Feature 3
Weight	Height	BMI
70	170	22
80	180	24
90	190	26
100	200	20
60	160	21

* In Table columns are also called Feature.



NOMINAL :- ORDER Doesnot Matter (MALE/FEMALE)

ORDINAL :- ORDER MATTER

Eg :- Degree
10th
12th
UG
PG
PHD

Student Performance (Dataset)

NAME	AGE	Height	Sex	weight	Educatio OM
SUNNY	25	170	Male	70	UG
ARJIT	30	160	Male	60	PG
Priyam	35	160	Male	60	UG
Priya	20	150	female	50	PG
Aditi	27	145	Female	55	PHD
ASHU	35	175	Male	85	PG
VIJAY	21	140	Male	65	PG

Category → NOMINAL
NUMERIC → COUNTING
NUM → COUNTING
CAT → NOMINAL
NUM → COUNTING
CAT → ORDINAL.

Used to Find Data Types

Univariate ⟶ Single column
Bivariate ⟶ Two column
Multivariate ⟶ More than Two column

INDIPENDENT AND DEPENDENT VARIABLE

[Age, Height, Sex] ⟶ weight
↳ Indipendent ↳ Dependent

Q. What Required first EDA or FE

=> First EDA
↓
Feature engineering

Example:-

RAW DATA
1. Chicken
2. Rice
3. onion
4. oil
5. Spices

→ EDA →

Preprocessing
1 kg
1 kg
1 kg
1 litre
500gm

→ Biryani Model

→ VALIDATION (TASTE)

*Above is example of ML Model

1. EDA
2. Preprocessing
3. Model
4. Validation

EDA :- Analysis of Data Based on the given Data.

Q. Is Preprocessing & feature engineering is same?

=> Yes, Both are same.

DATASET

NAME	AGE	EDUCATION	SALARY	Experience
Sunny	25	UG	25K	2
Deepak	30	PG	30 K	3
Rushi	40	UG	40K	15
Aman	50	PHD	50K	10
Shalini	20	UG	35K	1

EDA (Analysis)

1. Profile of DATA
2. Statistic Analysis
3. Graph Based Analysis

→ Profile of DATA

1. Row
2. Column
3. Missing value
4. Category
5. Numeric
6. Duplicate
7. Dtypes
8. Ram (Size)

⇒ Analysis

Statistic Analysis (Interpretation)

1. Variance
2. Co-variance
3. Standard deviation
4. Co-relation
5. Chi-square test
6. T-test
7. Z-test
8. Anova Test
9. Mean, Median, Mode

→ Univariate
→ Bivariate
→ Multi-variate

Graph Based Analysis (Plotting)

EDA

1. Box Plot → Outlier, Distribution, Statistical Profile
2. Scatter Plot → Outlier & Linear or not
3. Pie Chart →
4. Histogram → Distribution
5. KDE Plot
6. Count BAR ⇒ USED for counting (BAR chart)
7. Heat Map ⇒ Correlation

Observation Matter for Conclusion

Also import Dashboarding

Q. Based on EDA, can we do Pre Processing of Data?

⇒ True

Pre-Processing of DATA :-

1. Missing value Handle
2. outliers Handle
3. Scaling of Data
4. Transformation (Log, Box-Cox, Square, cube)
5. Encoding
6. Imbalance Data
7. Feature Selection
8. Dimension Reduction (PCA, tsne)
9. Duplicate value
10. Split / Merge / Drop / ADD (Column)

These are Steps of feature engineering

1. Missing Null value → Missing value handle (PP)
2. Outliyer → Handle
3. categorical (man, women) → Encoding
4. Skewed Range → Scale (within certain range)
5. count feature → Handle Imbalanced Data
   → Feature Selection
   → Dimension Reduction (PCA, tsne)

Encoding :- To change categorical Data into Numerical Data called Encoding

Some Automated Tools in Python For EDA

1. Pandas Profiling
2. Sweetviz
3. Autoviz
4. D-Tale
5. Mito
6. Knime
7. DataPrep

DAY-8
EDA & Feature engineering

DAY 2 OF EDA & F.E

Revision of Day-1:-

Core ML PipeLine

=> Data ingestion
-> EDA (Analysis)
-> Feature engineering or Pre-Processing
-> Model Building
-> Evaluation matrix or validation

EDA:

(1) Profile of the DATA
2. Statistical Based Analysis
3. Graph Based Analysis (important)

PreProcessing

1. Missing value handle
2. outliers handle
3. Scaling of Data
4. Transformation (Log, Cox-Box, Square, Cube)
5. Encoding
6. Handle Imbalanced Data
7. Feature Selection
8. Dimension Reduction
9. Duplicate value / Duplicate column
10. Split / Merge / Drop / Add (column)

Pre Processing → Model

Ways of Performing Feature engineering

1. Missing value Handle
   - ⇒ Fill with Random no.
   - ⇒ Forward / Backward filing.
   - ⇒ Statistical Approch (Mean, Median, Mode)
   - ⇒ with the help of end of Distribution, Fill the missing value.
   - ⇒ Drop the Row
   - ⇒ Impute with KNN (KNN-IMPUTER)
   - ⇒ ML Algorithm for Missing value
   - ⇒ Build own ML Model to Predict Missing Value.

2. Outlier Handling

Detect the outliyer

- Z-Score
- IQR
- Box
- Scatter-Plot
- Violin-Plot

Handle outlier

- Drop
- Fill with Median
- Replace / Binning

3. Transformation

1. Box Cox Transformation
2. Power Transformation
3. log
4. Square
5. cube

4. Scaling of Data

i) Standardization
ii) Min Max Scaler
iii) unit Scaling

5. Encoding various Method

1. one-hot encoding
2. Label - Encoding
3. Binary Encoding
4. Target Guided Encoding
5. Hash Encoding

6. Imbalanced Dataset Treatment various Method:-

1. Collect More Data
2. under Sampling
3. over Sampling
4. Cluster Based over Sampling

Q. How to find the Best Model Accuracy of various Method?

⇒ To increase the Accuracy, we need to change the Pre-Processing technique & use different Method or Step from the above.

⇒ we needs to use each & every Pre-Processing Step & find Best Accuracy.

Q: How do we transform the Data?

⇒ import numpy as np
np.log(df)
sns.distplot(df)