DAY - 7

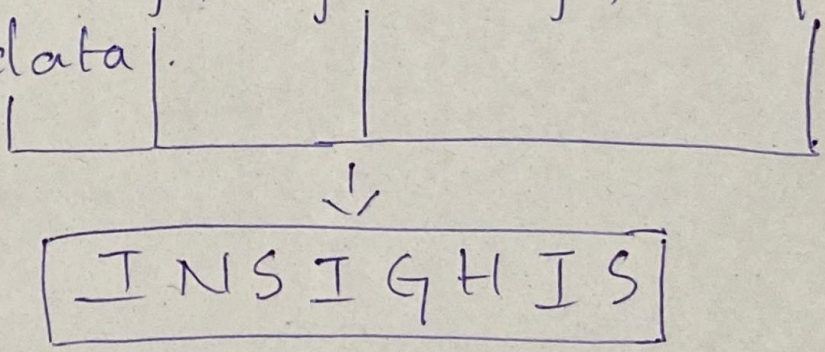
EDA and Feature Engineering:

Life cycle of Data-Science

1. Data ingestion (Data collecting)
2. EDA (Analysis over the data)
3. procesing (pre-procesing the data)
4. Model building (Algorithm building)(Math)
5. Validation & evalution of model.

These are the core machine learning pipeline.

Statistics : collecting, organising, Interpretion Analysis of data.



- For Analysing the data we need stats.
- In every domain EDA and procesing is required and Important.

Data collecting / Data Injestion can be done by : 1) Big data tools (+HDFS, Kafka)
2) Remote location (SQL/NO SQL)
3) file formats (CSV, TSV, XML, JSON)
4) Web Scrapping

Types of Data:

Tendency of data:

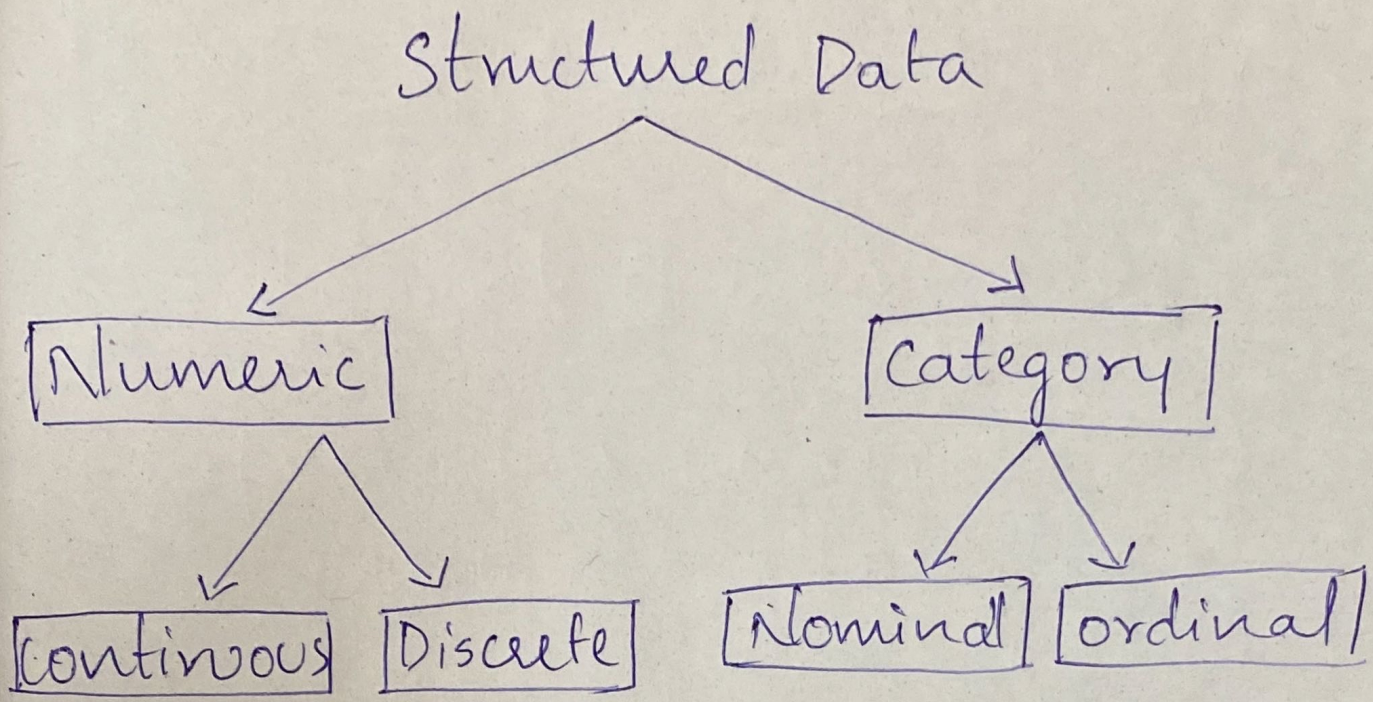
1) Batch Data: It is also known as historic data (comes more periodicly)
   i) Mini batch data (little more frequent)

2) Streaming Data: continuous / live data

These data can be divided into 3 parts:

1) structured data: Tabular form
2) unstructured data: Images, videos, voice, sound, text etc.
3) semi-structured data: xml, Json.

Machine learning → Structured data (Rows & Column



Univariate : 1 column

Bivariate : 2 columns

multivariate : more than 2 Columns

Independent / Dependent Analysis :

Dependent is checking on other variable while we are doing analysis is called as dependent Analysis

Independent is not checking or depending on other variable while we are doing analysis is called as Independent analysis.

EDA : Analysis of data based on a Given feature.

Pre processing / Feature engineering : cleaning, wrangling of data.

EDA STEPS : (Basic)

1) profile of the data
   - → Rows
   - → columns
   - → missing values
   - → categories
   - → numerical
   - → Duplicate values
   - → Data types
   - → RAM

2) Statistical Analysis (Interpretation)
   - → variance
   - → covariance
   - → standard deviation
   - → correlation
   - → chi Square test

→ t-test
→ z-test
→ Annova test.
→ mean/median/mode

3) Graphical Analysis
→ Box plot.
→ Scatter plot
→ pie chart
→ histogram
→ Kernal Density Estimator (KDE)
→ countbar
→ heat map.

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

FEATURE ENGEENIRING STEPS:

1) Handling missing values
2) Handling outliers
3) Scaling of data
4) Transformation (log, boxlog, Square, Cube)
5) Encoding categories
6) Imbalance data

7) Feauture Selection

8) Dimension reduction (PCA, TSNE)