Assignment 1

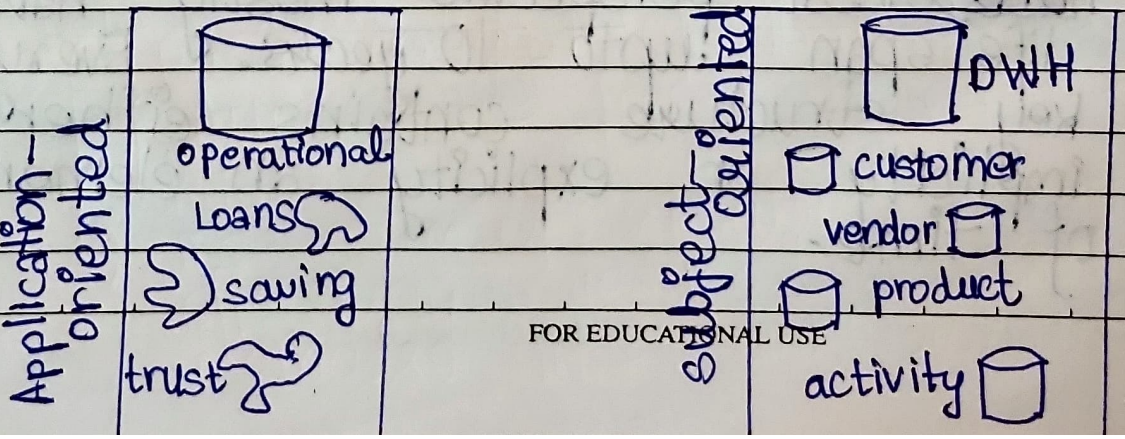
Chapter-1 : DWH & Introduction to Data Mining

Name: Atharva Lotankar
Class: D15-C RNo: 27
Batch: B ; DMBI

Q1. Explain the main characteristics of a Data Warehouses.

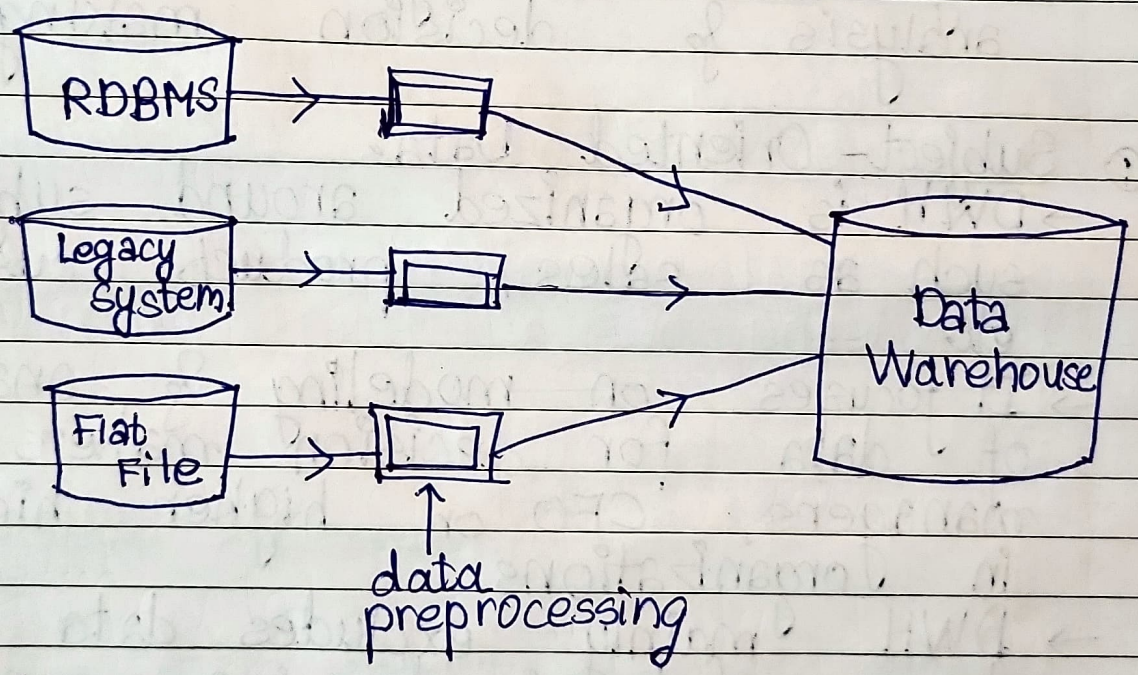
→ As per Bill Inmon, Data Warehouse has four main characteristics that distinguishes it from other data storage systems, making it ideal for analysis & decision making:

① Subject-Oriented Data:
→ DWH is organized around subjects such as sales, product, customers, etc.
→ It focuses on modeling & analysis of data for decision makers like managers, CEO or higher hierarchy in organizations.
→ DWH mainly excludes data which is not useful in decision support process.



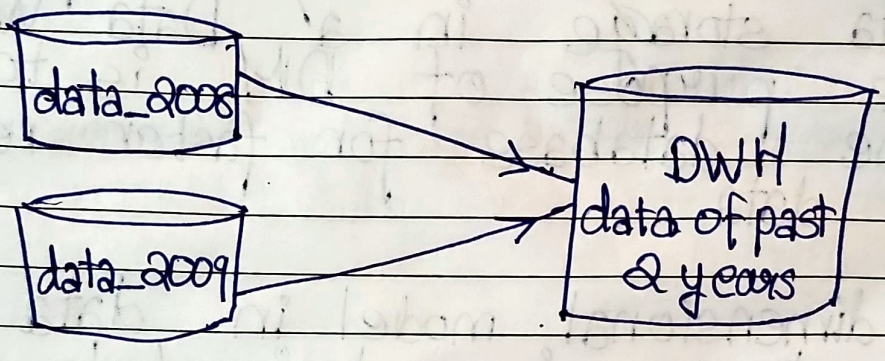
(ii) Integrated data: DWH is constructed by integrating data from multiple heterogeneous or different sources. (Like Oracle, DB2, Access,..)

Data Pre-processing are applied to ensure consistency, as data is coming from heterogeneous sources.
Before data from DWH loading, we have to remove all the inconsistencies.

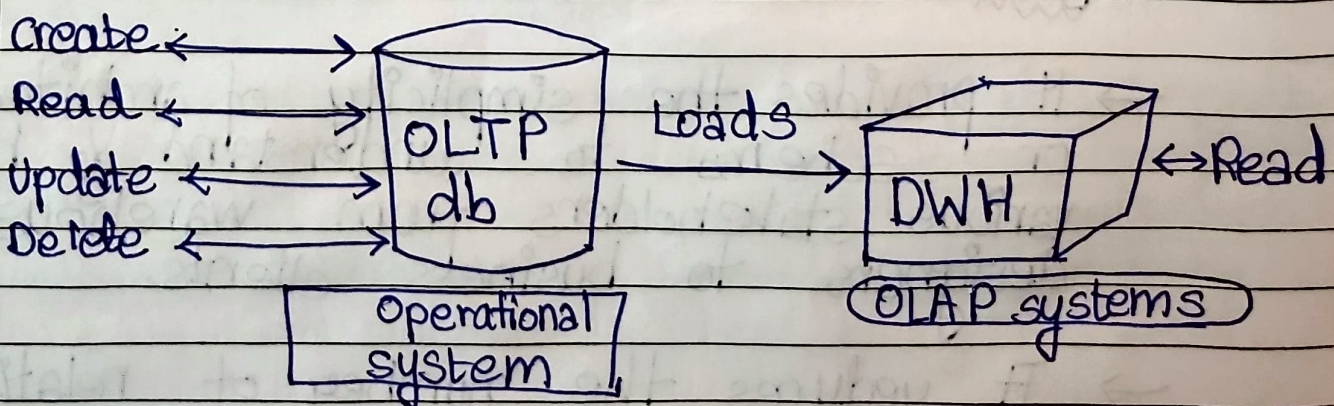


(iii) Time-variant data: provides info. from historical perspective mostly having life span upto 10 years. Every key structure contains either implicity or explicity an element of time.

The data in DWH is meant for analysis & decision making. If a user is looking at selling pattern of specific product, user needs data not only about the current sale, but on past sales as well.



(iv) Non-volatile data: Data once recorded in data ware house cannot be updated like OLTP system which allows updating data after recording also. Unless acted upon by regulatory or statutory regulations.



Q2. Define dimensional modelling & explain its importance in data warehouse design.

→ Dimensional Modelling (DM) is a logical data design technique optimized for data storage in a Data Warehouse. The purpose of DM is to optimize the database for faster retrieval of data.

A dimensional model in data warehouse is designed to read, summarize, analyze numeric info like values, balances, counts, weights, etc. in a data warehouse.

* Importance of Dim Modeling in DWH

→ It provides the simplicity of architecture or schema to understand & handle various stakeholders from warehouse designers to business clients.

→ It reduces the number of relationships between different varied data elements loaded in DWH.

→ It promotes data quality by enforcing foreign key constraints as a form of referential integrity check on a DWH. The dim-D modeling helps the db admins to maintain the reliability of the data.

→ The aggregated data functions used in schemas optimized the query performance posted by the customers. Since DWH size keeps on increasing & with this increased size, the optimization becomes the concern which dimension modeling makes it easy.

Q3. Differentiate between OLAP & OLTP with respect to : Purpose, Data Type, Query Complexity, Normalization.

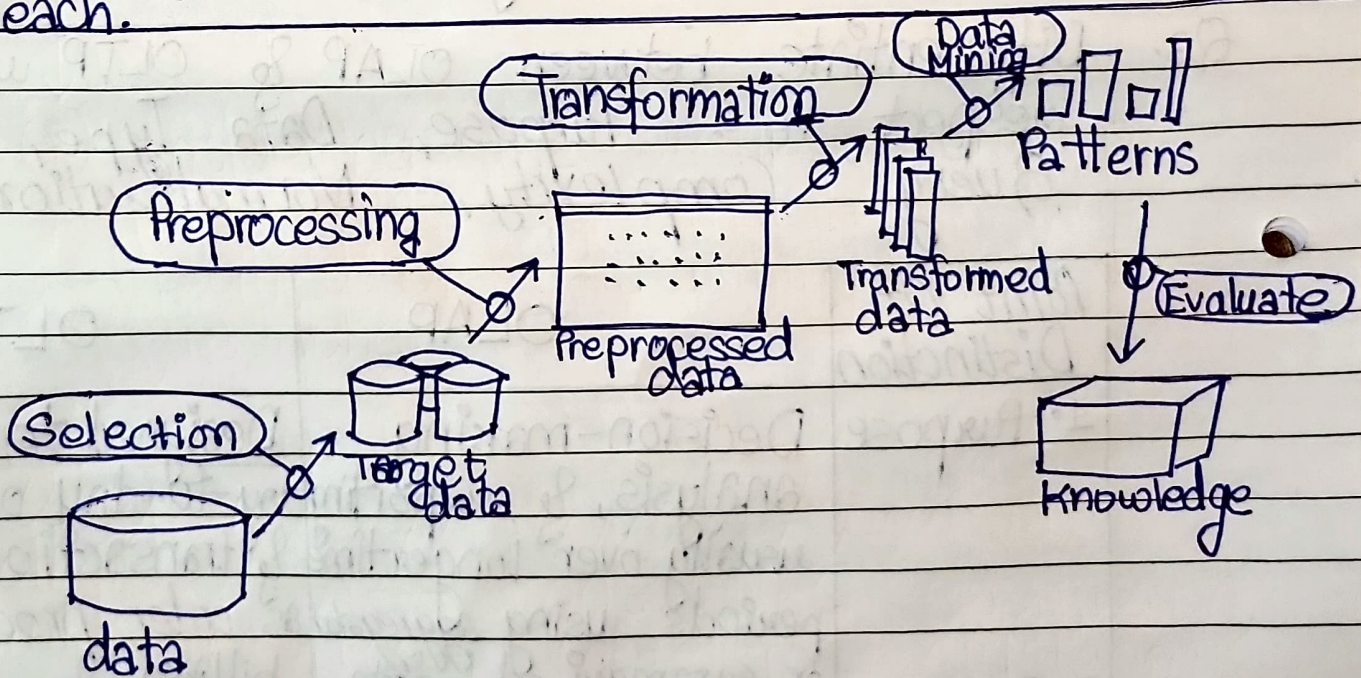
→

Point Of Distinction	OLAP	OLTP
1. Purpose	Decision-making analysis, & reporting usually over longer time periods using aggregated or summarized data.	Designed to support day-to-day operations & transactions such as order processing, billing & CRM.
2. Data Type	stores current & historical data integrated frm. multiple sources	stores only current (real-time data) that is continuously updated.

3. Query Complexity	It manages complex queries, aggregations, trends, drill downs	It handles simple queries that retrieve or modify specific records, often short & quick.
4. Normalization	Uses denormalized or star schema for fast querying & make analysis easier.	Uses Normalized schema (BCNF) for efficient updates & to reduces data redudancy.

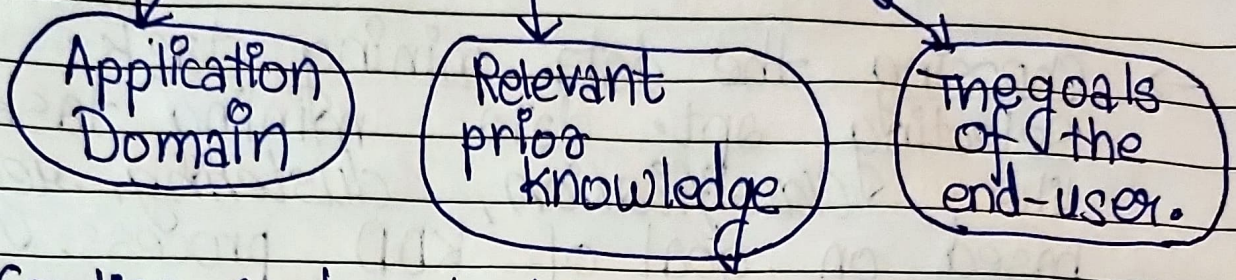
Q4. What are the main steps involved in the KDD process? Briefly explain each.

→



Steps involved in Knowledge Discovery in Databases:

→ Developing an understanding of:



→ Creating a target data set: Selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

→ Data Cleaning & pre-processing:
(1) Noise or outliers are removed.
(2) Essential information is collected for modeling or accounting for noise.
(3) Missing data fields are handled by using appropriate strategies.
(4) Time sequence information & changes are maintained.

→ Data reduction & projection:
(1) Based on goal of task, useful features are found to represent data.
(2) The num of variables may be effectively reduced using methodes like dimensionality

reduction or transformation.

→ Choosing the data mining task:
selecting apt. data mining tasks like classification, clustering, regression, based on goal of KDD process.

→ Choosing data mining algorithm(s):

(1) Pattern search is done using the apt. Data Mining method(s).

(2) A decision is taken on which models & parameters may be appropriate.

(3) Considering the overall criteria of KDD process a match for the particular data mining method is done.

→ Data Mining:
Using a representational form or other representations like classification, rules or trees, regression, clustering for searching patterns of interest.

→ Interpreting mined patterns & consolidating discovered knowledge.

Q5. Give five real-world applications of data mining & explain how mining helps in each.

→ Telecommunications Fraud detection:
Data Mining analyzes vast spatio-temporal data from calls & network usage to detect fraudulent behaviour patterns. By identifying outliers & unusal usage, companies can promptly flag & stop fraudulent users, preventing financial loss. Mining also optimizes network traffic & understand user behaviour for better service quality.

→ Retail Customer Insights:
Retailers mine purchase history, demographics, & promotion responsiveness to understand customer segments. This aids personalized marketing, product recommendations & improving customer satisfaction. Mining sales data also helps evaluate marketing campaigns' effectiveness & optimizes inventory based on buying patterns.

→ Healthcare Predictive Analysis:
Mining patient records & treatment histories enables early disease detection & personalized care plans.

Healthcare providers use data mining to predict high-risk patients, optimize resource allocation, & improve treatment outcomes, ultimately enhancing cost-effiency & patient care quality.

→ Smart City Traffic Management:
DM in processes real time traffic flow, public transport usage, & sensor data to dynamically control traffic lights & reduce congestion. This makes urban transportation more efficient, decreases travel time, & cuts pollution by optimizing traffic patterns based on mined data.

→ Criminal Investigation:
Law enforcement mines extensive crime reports & spatial data to detect crime patterns & links among activities. This aids faster crime solving by identifying hotspots & suspect behaviours, improving resource deployment & preventive policing strategies.

Assignment 1

Chapter 2: Data Exploration & Preprocessing

Name: Atharva Lotankar
Class: D15-C ; RNum: 27
Batch-B ; DMBI

Q1. Explain different types of attributes in data mining with examples: Nominal, Ordinal, & Interval, Ratio

→ **Nominal attributes**

Also known as 'categorical attributes' & allow for only qualitative classification. Every individual item has a certain distinct categories, but quantification or ranking the order of the categories is not possible.

Nominal attribute categories can be numbered arbitarily. Arithmetic & logical operations on nominal data can't be performed.

e.g. → Nationality = {American, Canadian, Indian, German, French, Russian, Turk}

→ Diagnosis_type = {Diabetes, Hypertension, Asthma, None}

→ Operating_System = {Windows, Linux, MacOS, Android, iOS}

→ Ordinal Attribute:
A discrete ordinal attribute is a nominal attribute, which have meaningful order or rank for its different states.

The interval between different states is uneven due to which arithmetic operations aren't possible, however logical opern^s may be applied.

e.g. & → education_level = {HighSchool, Bachelor's, Master's, PhD}

→ socioeconomic_status = {low, middle, high}

→ clothing_size = {XS, S, M, L, XL}

→ Numeric Attributes: They are quantifiable. Can be measured in terms of a quantity, which can either have an integer or real value.

They are of 2 types:

↝ Interval-scaled attributes

These are continuous measurement on a linear scale. They are numeric values (variables) where order of values matters. The difference between values is meaningful & consistent.

However, there is no true zero point (i.e. zero does not mean "none")

e.g. Temp in °C / °F = {20°C, 30°C, 0°C}
→ a 10°C difference is meaningful, but 0°C doesn't mean no temperature.

dates = {1990, 2000, 2020}
→ The diff. between years is meaningful, but year 0 AD doesn't mean "no time".

↝ Ratio-scaled attributes

These are continuous positive measurements on a non-linear scale. They have all properties of Interval-scaled attributes, plus a true (absolute) zero point.

We can order the values, calculate meaningful distances & calculate ratios because zero means none.

e.g. height = {170 cm, 180 cm, 0 cm}
→ 180 cm is twice as tall as 90 cm, & 0 cm means absence of height

income = {₹1,00,000, ₹500, ₹0}
→ ₹1 lakh is twice as much as ₹5K, & ₹0 means null income.

Q2. What is role of statistical measures in data summarization?

→ Statistical measures play a fundamental role in data summdurization by providing concise, meaningful & interpretable information about large & complex datasets.
Instead of analyzing raw data, which can be extensive & difficult to interpret, statistical measures help to simplify the data into a form that highlights key patterns, trends & distributions.

Firstly: Measures of Central Tendency such as mean, median, mode, midrange describe the central point around which data tends to cluster. These measures give a quick understanding of

typical or average value in the dataset, which is often crucial for making decisions or comparisons.

Secondly: Measures of Dispersion → such as variance, quartiles, std. deviation, percentiles quantify the spread or variability of data. These indicators show how much the data values deviate from the average, which helps in understanding the consistency, risk, or reliability of the dataset.

Moreover, statistical measures are essential in detecting outliers, data trends, & anomalies, which are vital for quality control, forecasting, strategic planning. They also for more advanced analyses such as hypothesis testing, regression analysis, & ML models.

Q3. Define & compute the following with examples: Mean, Median, Mode, Variance & Standard Deviation.

→ Mean: The sum of all values divided by number of values. It represents the central value of a dataset.

Formula → Mean (μ) = $\frac{\sum x_i}{n}$

$x_i \equiv$ data points ; n = num. of data pts.

Example → dataset = [4, 8, 6, 5, 3]

$\mu = \frac{4+8+6+5+3}{5} = \frac{26}{5} = 5.2$

→ Median: The middle value of an ordered dataset. The measure is suitable when data is skewed.
It has to be arranged in order of magnitude to formulate.

Formula →

if n is odd (n%2 != 0)

Median = $\frac{(n+1)^{th} \text{ Term}}{2}$

if n is even (n%2 == 0)

Median = $\frac{(\frac{n}{2})^{th} \text{ Term} + (\frac{n}{2}+1)^{th} \text{ Term}}{2}$

Example →

d1 = [7, 3, 9, 2, 6]
order = 2, 3, 6, 7, 9

Median = $\frac{(5+1)^{th} \text{ T}}{2} = 3^{rd}$ Term

6

d2 = [2, 3, 6, 7]
order = 2, 3, 6, 7

↓

Median = 3^{rd} T + 3^{rd} T / 2

↳ $\frac{3+6}{2} = 4.5$

→ **Mode:** The value that appears most frequently in dataset. A datset may have one mode (unimodal), more than one (bimodal) or no-mode.

Mode doesn't have formula as it is more in counting but does have an empirical formula

E. formula (approx.) → Mode = 3 * Median − 2 * Mean

Example → dataset = [4, 6, 2, 4, 3, 5, 5, 4, 5, 7, 1]

→ mode = 4 and 5 (appeared 3 times)

it is bi-modal

→ **Variance:** Measures the average squared deviation from mean. It reflects how spread out the data is.

Formula → $\text{Var}(\sigma^2) = \frac{\sum (x_i - \mu)^2}{n}$

$x_i \equiv$ data pts. $n \equiv$ num. of data pts.

$\mu \equiv$ mean

Example → dataset = [2, 4, 6, 8]

$$\mu = \frac{2+4+6+8}{4} = \frac{20}{4} = 5$$

$$\sigma^2 = \frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4}$$

$$= \frac{9+1+1+9}{4} = \frac{20}{4} = 5$$

Indicating data could be symmetric or skewed.

→ Standard Deviation: The square root of variance. It gives measure of spread in same units as the data.

Formula →

$$\text{std. deviation } (\sigma) = \sqrt{\text{Variance}}$$

$$= \sqrt{\frac{1}{n}(\Sigma x_i - \mu)^2}$$

Example →

dataset = [2, 4, 6, 8]

$\Rightarrow \sigma^2 = 5$

$$SD = \sqrt{\sigma^2} = \sqrt{5} = 2.236$$

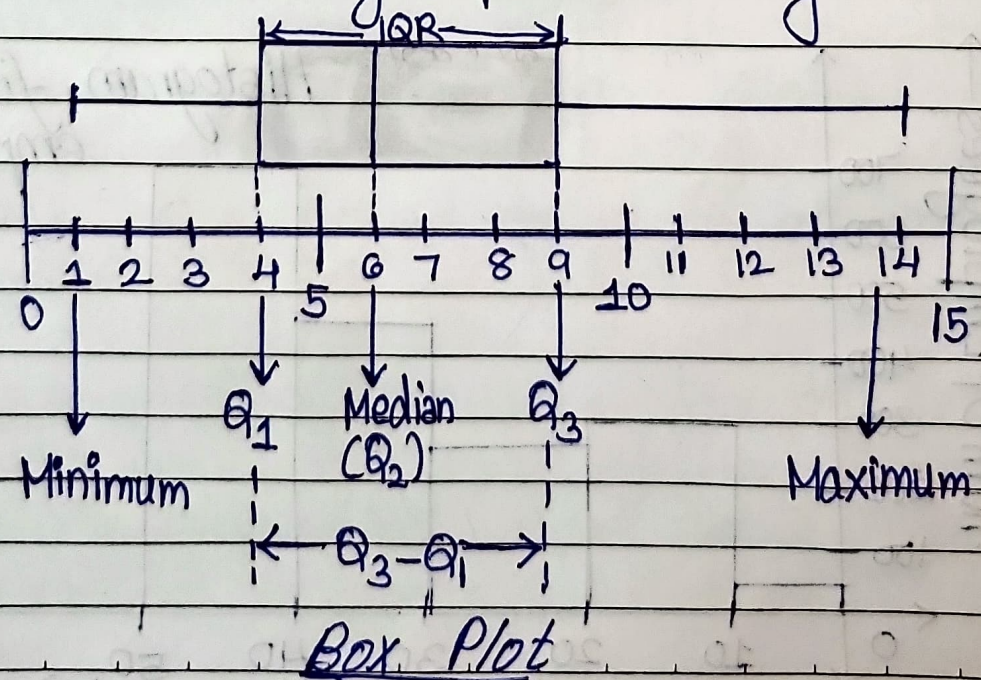
Q4: Explain the use of box plots & histograms for data summarization.

→ **Box Plots (Box-&-Whisker Plots)**

Box plots provide a visual summary of distribution of a dataset based on five key statistics:

- minimum
- first quartile (Q_1)
- median (Q_2)
- third quartile (Q_3)
- maximum.

They are especially useful for identify the central tendency, display data spread & variability via Interquartile range ($IQR = Q_3 - Q_1$), detecting outliers (pts. outside whiskers) & comparing distributions across different groups side-by-side.



Box Plot

→ Histograms display the frequency distribution of dataset by grouping data into bins or intervals & showing how many data points fall into each bin.

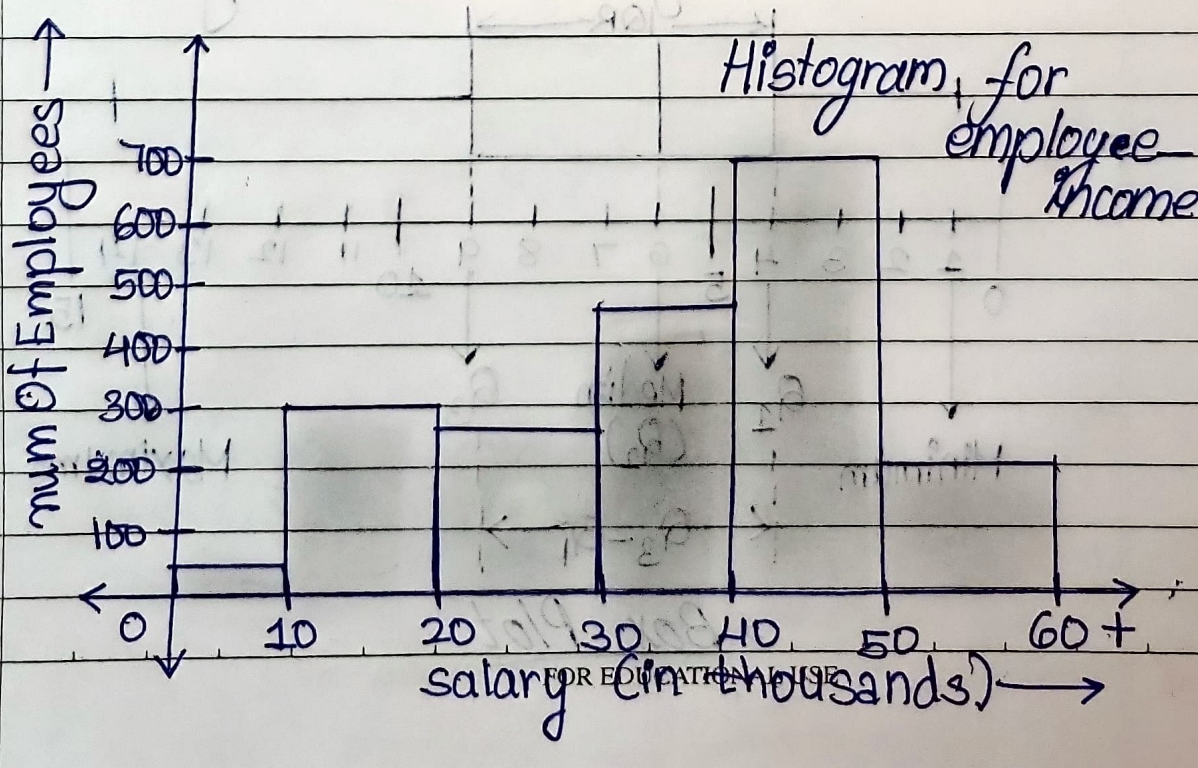
They are useful for:

→ Understanding shape & spread of data distribution.

→ Identifying modes (peaks) in the data.

→ Detecting skewness (whether data is symmetric or leans left / right).

→ Spotting gaps or unusual clusters in the data.



Q5. Discuss different techniques for handling: Missing values, Noisy data, Inconsistent data.

→ Missing Values Techniques to handle:

① → Ignore the data row: In case of classification suppose a class label is missing for a row, such a data row could be ignored, or many attr within a row could be ignored.

② → Fill the missing values manually: This isn't feasible for large dataset & also time consuming.

③ → Use a global constant to fill in for missing vals: Sometimes missing values are difficult to be predicted, then a global constant value like "null", "N/A", "_" can be used to fill all missing values.

④ → Use attribute mean: For missing values, mean or median of its discrete values may be used as a replacement.

⑤ → Use data mining algorithm to predict the most probable value: Missing values may also be filled by using regression, inferences from Bayesian formalism, decision trees, clustering algorithms

→ Noisy Data Techniques to handle:

① Binning Method - Considering neighbourhood of sorted data smoothening can be applied. The sorted data is placed into bins or buckets.

Smoothing by → bin means
→ bin medians
→ bin boundaries

e.g. noisy-data = [4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34]

→ bins n=3 ⇒ B_1: 4, 8, 9, 15
B_2: 21, 21, 24, 25
B_3: 26, 28, 29, 34

by means:
$B_{\mu 1}$: 9, 9, 9, 9
$B_{\mu 2}$: 23, 23, 23, 23
$B_{\mu 3}$: 29, 29, 29, 29

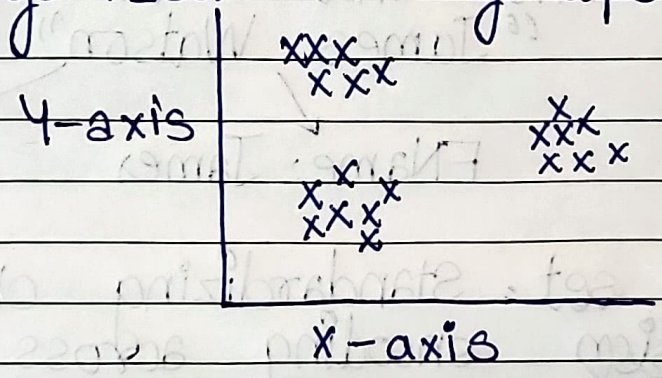
by medians:
B_{md1}: 8.5, 8.5, 8.5, 8.5
B_{md2}: 22.5, 22.5, 22.5, 22.5
B_{md3}: 28.5, 28.5, 28.5, 28.5

by bin boundaries (extreme-nearer):
B_{b1}: 4, 4, 4, 15
B_{b2}: 21, 21, 25, 25
B_{b3}: 26, 26, 26, 34

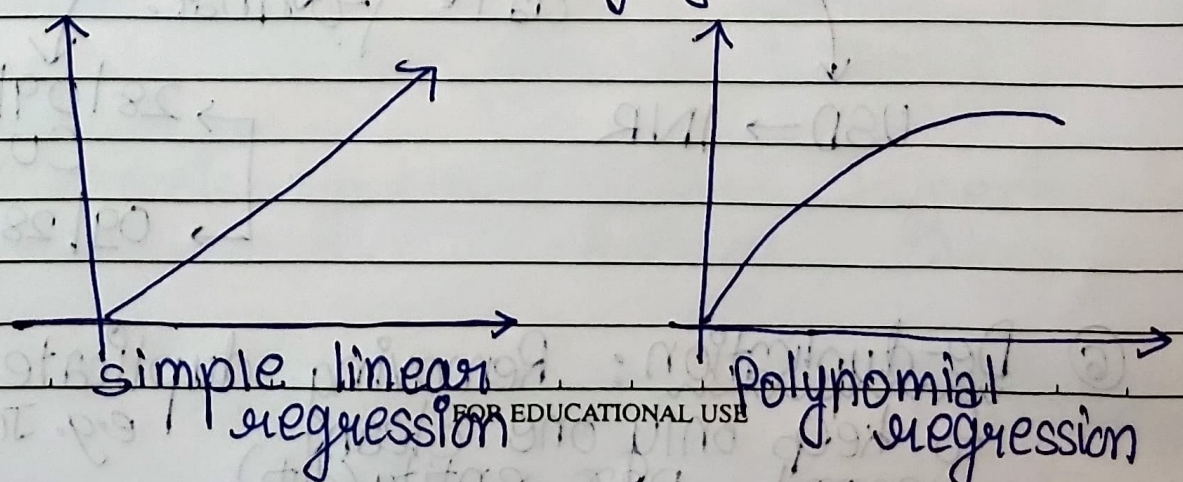
② Outlier Analysis by Clustering

Partition dataset into clusters & one can store cluster representation only, i.e. replace all values of cluster by that one value representing the cluster.

Outliers can be detected by using clustering techniques, where related values are organized into groups or clusters.



③ Regression: Statistical measure used to determine the strength of relationship between one dependent variable denoted by Y & series of independent changing variables.



→ Inconsistent data techniques to handle:

① Format revision: Changing datatypes or lengths to fit in standards

② Decoding of Fields: Translating codes into readable values

③ Splitting of fields: Breaking combined fields into individual components.
e.g. "James Watson"
↙ FName: James
↘ LName: Watson

④ Character set conversion: Standardizing character encoding across systems like EBCDIC → ASCII or Unicode

⑤ Conversion of Unit & Date: Standardizing units of measurement like currency or using consistent date format.
↙ USD → INR
↘ 28/09/2025 (UK)
↘ 09/28/2025 (US)

⑥ De-duplication: Removing duplicate entries to keep only one clean record per entity (data)
e.g. John Watson
J. Watson →