Question 1 : Define Data Transformation in ETL and explain why it is important

Sol: Data Transformation is the process of converting raw, extracted data into a clean, structured, and standardized format during the **Transform** phase of the ETL (Extract, Transform, Load) process.

ETL consists of three stages:

1. **Extract** – Collecting data from various sources such as databases, files, or APIs.
2. **Transform** – Cleaning, modifying, validating, and organizing the data.
3. **Load** – Storing the transformed data into a target system such as a data warehouse.

During the transformation stage, data is processed to make it suitable for analysis and reporting.

Importance of Data Transformation

1. Improves Data Quality
Removes incorrect, duplicate, or missing values to ensure accuracy.

2. Ensures Data Consistency
Standardizes formats across different systems (e.g., converting “M”, “male”, “MALE” into “Male”).

3. Enables Accurate Analysis
Clean and structured data helps in generating reliable reports and business insights.

4. Supports Data Integration
Combines data from multiple sources into a unified format.

5. Enhances System Performance
Optimized data structure improves query speed and overall system efficiency.

Question 2 : List any four common activities involved in Data Cleaning.

Sol: 1. Handling Missing Values

Missing data occurs when no value is stored for a variable in a dataset.
Techniques to handle missing values include:

- Replacing with mean, median, or mode
- Filling with a default value
- Removing rows or columns with excessive missing data

2. Removing Duplicate Records

- Duplicate data can lead to incorrect analysis and reporting.
  During cleaning, repeated records are identified and removed to maintain data accuracy and integrity.

3. Correcting Inconsistent Data

Data may be entered in different formats, such as:

- "M", "Male", "male"
- Different date formats (DD/MM/YYYY vs MM/DD/YYYY)

4. Fixing Incorrect or Invalid Data

Sometimes data may contain errors such as:

- Negative age values
- Invalid email formats
- Out-of-range numbers

Question 3 : What is the difference between Normalization and Standardization?

Sol: Normalization and Standardization are data transformation techniques used to scale numerical data. They help improve the performance of machine learning models and statistical analysis.

Normalization is a scaling technique that transforms data values into a fixed range, usually between **0 and 1**.

Formula (Min-Max Scaling):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Features of Normalization:

- Scales data within a specific range (0 to 1)
- Sensitive to outliers
- Useful when data does not follow a normal distribution
- Commonly used in algorithms like K-Nearest Neighbors (KNN) and Neural Networks

Standardization

Standardization is a scaling technique that transforms data so that it has a **mean of 0** and a **standard deviation of 1**.

Formula (Z-Score):

$$Z = \frac{X - \mu}{\sigma}$$

Features of Standardization:

- Data is centered around zero
- Less affected by outliers compared to normalization
- Used when data follows a normal distribution
- Commonly used in algorithms like Logistic Regression and SVM

Question 4 : A dataset has missing values in the “Age” column. Suggest two techniques to handle this and explain when they should be used.

Sol: Missing values are common in datasets and can affect the accuracy of analysis and machine learning models. When the “Age” column contains missing values, appropriate techniques must be applied to handle them effectively.

Technique 1: Mean or Median Imputation

In this method, missing values are replaced with the **mean** or **median** of the Age column.

When to Use Mean Imputation:

- When the data is normally distributed
- When there are no significant outliers

When to Use Median Imputation:

- When the data contains outliers
- When the distribution is skewed

Example:
If the average age of people in the dataset is 30, the missing Age values can be replaced with 30.

Advantage:

- Simple and quick method
- Maintains dataset size

Question 5 : Convert the following inconsistent “Gender” entries into a standardized format (“Male”, “Female”): ["M", "male", "F", "Female", "MALE", "f"]

Sol: Inconsistent data entries can cause errors in analysis and reporting. Therefore, it is important to standardize categorical values into a uniform format.

Step 1: Identify Inconsistent Values

The given values contain different formats such as:

- Uppercase and lowercase letters
- Abbreviations
- Full words

Examples:

- "M", "male", "MALE"
- "F", "f", "Female"

Step 2: Apply Standardization Rules

- Convert all variations of male-related entries ("M", "male", "MALE") into **"Male"**
- Convert all variations of female-related entries ("F", "f", "Female") into **"Female"**

Standardized Output

["Male", "Male", "Female", "Female", "Male", "Female"]

Question 6 : What is One-Hot Encoding? Give an example with the categories: “Red, Blue, Green”.

Sol: One-Hot Encoding is a data transformation technique used to convert categorical variables into a numerical format so that they can be used in machine learning models.

Machine learning algorithms cannot directly process categorical (text) data, so One-Hot Encoding converts each category into a separate binary (0 or 1) column.

One-Hot Encoding creates new columns for each unique category in a variable.
Each row will have:

- **1** in the column representing its category
- **0** in all other category columns

Example

Given Categories:

Color = ["Red", "Blue", "Green"]

After applying One-Hot Encoding, the data will be transformed as follows:

Color	**Red**	**Blue**	**Green**
Red	1	0	0
Blue	0	1	0
Green	0	0	1

- If the color is **Red**, then Red = 1 and others = 0
- If the color is **Blue**, then Blue = 1 and others = 0
- If the color is **Green**, then Green = 1 and others = 0

Advantages of One-Hot Encoding

- Prevents ordinal relationship between categories
- Makes categorical data suitable for machine learning models
- Easy to interpret

Question 7 : Explain the difference between Data Integration and Data Mapping in ETL.

Sol: In the ETL (Extract, Transform, Load) process, Data Integration and Data Mapping are important concepts that help in combining and organizing data from multiple sources. Although they are related, they serve different purposes.

Basis	**Data Integration**	**Data Mapping**
Meaning	Combining data from multiple sources	Matching source fields to target fields
Focus	Unifying datasets	Defining field-level relationships
Scope	Broader process	Part of the transformation process
Purpose	Create a single consolidated dataset	Ensure correct data transfer

Data Integration

Example:

Combining customer data from:

- CRM system
- Sales database
- Marketing platform

All data is merged into a central data warehouse.

Data Mapping

Example:

Source Field → Target Field

- “Cust_ID” → “Customer_ID”
- “DOB” → “Date_of_Birth”

Question 8 : Explain why Z-score Standardization is preferred over Min-Max Scaling when outliers exist.

Sol: Scaling techniques are used in data transformation to bring numerical values into a comparable range. Two commonly used scaling methods are **Min-Max Scaling (Normalization)** and **Z-score Standardization**. When outliers are present in the dataset, Z-score Standardization is generally preferred.

Min-Max Scaling (Normalization)

Formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Characteristics:

- Scales data between 0 and 1
- Uses minimum and maximum values
- Highly sensitive to outliers

Problem with Outliers:

If a dataset contains extreme values (very large or very small numbers), the range (Xmax - Xmin) increases significantly. This compresses most of the normal data points into a very small range, reducing model performance.

Z-score Standardization

Formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- μ= Mean
- σ= Standard Deviation

Characteristics:

- Centers data around mean (0)
- Standard deviation becomes 1
- Less affected by extreme values