ASSIGNMENT : TRANSFORMATION IN ETL

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

SOLUTION: Data Transformation is the stage of the ETL (Extract–Transform–Load) process where raw, inconsistent, incomplete, or unstructured data is converted into clean, standardized, and analysis-ready datasets.

While extraction collects data from various sources, transformation applies rules, cleaning techniques, business logic, and structural changes to make the data meaningful, reliable, and usable.

Transformation includes:

- Cleaning missing and duplicate records
- Normalizing and standardizing values
- Fixing inconsistent formats
- Mapping fields across datasets
- Encoding categorical fields
- Creating derived metrics
- Detecting and handling outliers
- Extracting and restructuring dates and text

IMPORTANCE;

Transformation directly impacts the quality of:

- Dashboards and reports
- Machine learning models
- Forecasting and trend analysis
- Customer segmentation
- Financial reconciliation
- HR analytics such as attrition, performance, and tenure

● Marketing data where inconsistencies are common

Organizations depend on clean and transformed data to make confident decisions.

A poorly transformed dataset leads to incorrect KPIs, flawed predictions, and unreliable insights.
That is why transformation is often called:
“The heart of ETL and the backbone of analytics.”

Question 2 : List any four common activities involved in Data Cleaning.
SOLUTION: THE FOUR COMMON ACTIVITIES INVOLVED IN DATA CLEANING;
1; Handling Missing Values
Why Missing Values Occur

● Manual data entry errors

● System failures

● Sensor issues

● Optional fields in forms

Techniques to Handle Missing Values:

1. Delete missing rows when % of missing rows is very small
2. Replace missing values with mean or median when data is numerical and with mode when there is categorical data.
3. For Time series use either forward fill or backward fill
4. We can also use custom values as 0 (Numerical) or “Unknown” (Categorical).

2;Removing Duplicate Records
Why Duplicates Occur

● Same customer submitting form twice

- System synchronization issues
- Data merged from multiple files

Steps to Remove Duplicates:

1. Identify duplicates and then identify which record to keep like First occurrence, Last updated timestamp.

2. After filtering the valid row, delete the rest.

3;Correcting Data Inconsistencies

Different formats of the same information due to:

- Human input variations
- Different system exports
- Missing validation rules

Examples of Inconsistencies:

Text / Category Issues Date Format Issues

- "Male", "MALE", "male", "M" 12-02-23
- “united states”, “USA”, “US”, “U.S.A.” 2023/02/12

4;Outlier Detection

Values that deviate significantly from most other observations.

Examples:

- A customer aged 200
- Monthly salary = ₹10 crore
- Negative sales quantity

Techniques for Detecting Outliers:

1. Z-Score Method

Measures how far a value is from mean (in terms of standard deviations).Typically, values beyond ±3 are outliers.

2. IQR Method (Interquartile Range)

IQR = Q3 – Q1

Outliers are:

- Less than Q1 – 1.5 × IQR
- Greater than Q3 + 1.5 × IQR

Outliers should be kept if valid and removed if error. Also we can cap or transform extreme values with upper / lower thresholds.

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

SOLUTION:

1. Normalization

Normalization is a scaling technique in which values are shifted and rescaled so that they lie within a fixed range, usually **0 to 1**.

Purpose:

- To bring all features to the same scale
- Prevent attributes with large values from dominating smaller ones
- Useful when data does **not follow normal distribution**

Formula (Min–Max Scaling):

Xnorm = $X - X_{min} / X_{max} \_X_{min}$

Characteristics:

- Range becomes [0,1] or sometimes [-1,1]
- Preserves the original distribution shape
- Sensitive to outliers

Commonly Used In:

- K-Nearest Neighbors (KNN)
- Neural Networks
- Image processing
- Distance-based algorithms

2. Standardization

Standardization transforms data so that it has:

- **Mean = 0**
- **Standard Deviation = 1**

This is also called **Z-Score Scaling**.

Formula:

$Xstd = X - \mu / \sigma$

where
μ = mean of feature
σ = standard deviation

Characteristics:

- Does not bound values to a fixed range
- Works well when data follows Gaussian (normal) distribution
- Less affected by outliers compared to normalization

Commonly Used In:

- Linear & Logistic Regression
- Principal Component Analysis (PCA)
- Support Vector Machines (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.

SOLUTION: Handling Missing Values in the “Age” Column

Two common techniques to handle missing values in the **Age** column are:

1.Mean/Median Imputation

- The missing age values are replaced with the **mean or median** of the available age data.
- **When to use:**
  - Use **mean** when the age data is normally distributed.

    - Use **median** when the data contains outliers or is skewed, as median is less affected by extreme values.
- **Advantage:**
    - Simple and quick to implement without losing records.
- **Disadvantage:**
    - May reduce data variability and introduce slight bias.

2. Deleting Rows with Missing Values

- Remove records where the Age value is missing.
- **When to use:**
    - When the number of missing values is very small (e.g., less than 5–10% of the dataset).
    - When sufficient data is available and deleting rows will not affect analysis.
- **Advantage:**
    - Ensures only complete and accurate data is used.
- **Disadvantage:**
    - Can lead to loss of important information if many rows are removed.

Question 5 : Convert the following inconsistent "Gender" entries into a standardized format ("Male", "Female"):

["M" , "male" , " F" , "Female" , "MALE" , "f"]

SOLUTION :

Given Data:

```
["M", "male", " F", "Female", "MALE", "f"]
```

The gender column contains inconsistent values due to:

- Different letter cases (male, MALE)
- Short forms (M, F)
- Extra spaces (" F")
- Mixed representations

These inconsistencies must be converted into a **standard format: "Male" and "Female".**

Steps to Standardize the Data

1. Remove Extra Spaces

Some values contain leading/trailing spaces such as **“ F”**.

Apply trimming to clean them → “F”

2. Convert to a Common Case

To avoid mismatch between:

- “male”, “MALE”, “Male”

Convert all values to **lowercase**:

```
[“m”, “male”, “f”, “female”, “male”, “f”]
```

3. Map Short Forms to Full Forms

Create mapping rules:

- {“m”, “male”} → **Male**
- {“f”, “female”} → **Female**

Original Value	After Cleaning	StandardValue
“M”	M → Male	**Male**
“male”	male → Male	**Male**
“ F”	F → Female	**Female**
“Female”	Already correct	**Female**
“MALE”	MALE → Male	**Male**
“f”	f → Female	**Female**

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

SOLUTION : One-Hot Encoding *i*s a technique used to convert categorical (text) data into numerical form so that machine learning algorithms can understand it.
Each category is converted into a separate binary column (0 or 1).

- Value **1** → category is present
- Value **0** → category is absent

It is mainly used for **nominal data** where there is no natural order.

Example with Categories: “Red, Blue, Green”

Original column:

Color
Red
Blue
Green

After One-Hot Encoding:

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

EXPLANATION;

- “Red” → [1, 0, 0]
- “Blue” → [0, 1, 0]
- “Green” → [0, 0, 1]

Each color gets its own column, and only one column has value **1** at a time.

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

SOLUTION:

DATA INTEGRATION:

Data Integration is the process of **combining data from multiple different sources** into a single, unified view for analysis and reporting.

- Sources may include databases, Excel files, APIs, cloud systems, etc.
- The goal is to create **consistent, complete, and centralized data**.
- It is a broader ETL activity that includes extraction, transformation, and loading.

Example:
Combining student data from admission system, attendance system, and exam portal into one data warehouse.

DATA MAPPING IN ETL;

Data Mapping defines **how fields from the source system correspond to fields in the target system**.

- It specifies which column goes where
- Defines data type conversion
- Sets transformation rules

Example:
Source column → Target column

- `fname` → `First_Name`
- `dob` → `Date_of_Birth`
- `marks` → `Total_Score`

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

SOLUTION :

Scaling is used to bring numerical features to a similar range. The two common methods are **Min-Max Scaling** and **Z-score Standardization**. When a dataset contains outliers, **Z-score standardization is preferred** for the following reasons:

1. Effect of Outliers in Min-Max Scaling

Min-Max scaling uses the formula:

$Xnorm = X - X_{min} / X_{max} \_X_{min}$

It depends completely on the **minimum and maximum values**.
If an extreme outlier is present, the max or min value changes drastically and all normal data points get compressed into a very small range. This distorts the original distribution and reduces the model's ability to learn patterns.

2. Z-score is Less Sensitive to Outliers

Z-score uses:

$Z=X-\mu / \sigma$

It depends on **mean and standard deviation**, not on extreme values.
Outliers may slightly affect the mean, but they do not control the entire scaling process. Therefore, the relative distance between normal observations is preserved.

3. Better for Statistical Algorithms

Many algorithms such as **PCA, Logistic Regression, SVM, and Linear Regression** assume data is centered around zero with unit variance. Z-score satisfies this assumption, while Min-Max does not.

4. Example

For data: [10, 12, 11, 13, 500]

- Min-Max squeezes 10–13 near 0 due to 500 (outlier).
- Z-score keeps 10–13 close together and treats 500 as a separate extreme value.