Name: Priyanshu Sahu
Batch: Data Analytics Nov Live Batch

Assignment 19: Data Transformation

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

Answer:

Data Transformation:

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

In ETL:

- **Extract →** Get data from source
- **Transform →** Clean and modify data
- **Load →** Store into database/warehouse

Why it is important:

- Makes data consistent
- Removes errors and duplicates
- Converts data into proper format
- Improves data quality
- Makes data ready for analysis and reporting

Without transformation, raw data may be messy and unreliable.

Name: Priyanshu Sahu
Batch: Data Analytics Nov Live Batch

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

Answer:

Four common activities:

1. **Handling Missing Values** – Filling or removing null values
2. **Removing Duplicates** – Deleting repeated records
3. **Correcting Inconsistent Data** – Standardizing formats (e.g., M, Male → Male)
4. **Fixing Data Types** – Converting text to numbers, dates to proper format

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

Answer:

Normalization	Standardization
Scales data between 0 and 1	Scales data around mean = 0
Formula: (X - min) / (max - min)	Formula: (X - mean) / standard deviation
Sensitive to outliers	Handles outliers better
Used in Neural Networks	Used in statistical models

-

Simple difference:

- **Normalization = fixed range (0 to 1)**
- **Standardization = based on mean and standard deviation**

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

Answer:

Two techniques:

1. Mean/Median Imputation

- Replace missing age with **mean or median age**
- Use **median** if outliers exist
- Use **mean** if data is normally distributed

2. Delete Rows

- Remove rows where Age is missing
- Use when:
    - Missing values are very few
    - Data size is large

If many values are missing, do NOT delete rows (you may lose important data).

Name: Priyanshu Sahu
Batch: Data Analytics Nov Live Batch

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

Answer:

Example inconsistent data:

- M
- male
- Male
- F
- female
- FEMALE

Standardized Output:

- M, male → **Male**
- F, female → **Female**

Use mapping logic:

If value in ['M','male','Male'] → Male

If value in ['F','female','Female'] → Female

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

Name: Priyanshu Sahu
Batch: Data Analytics Nov Live Batch

Answer:

One-Hot Encoding converts categorical data into multiple binary (0/1) columns.

Example categories:
Red, Blue, Green

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

Each category gets its own column.

Used in Machine Learning because models understand numbers, not text.

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

Answer:

Data Integration

- Combining data from multiple sources into one system.
- Example: Combining sales data from website + mobile app.

Data Mapping

- Matching fields from source to destination.
- Example:
  - Source: cust_name
  - Target: customer_name

👉 Simple difference:

- **Integration = Combine data**
- **Mapping = Match columns**

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

Answer:

Min-Max Scaling:

- Uses min and max values
- Outliers affect min and max
- Data gets compressed

Z-score Standardization:

- Uses mean and standard deviation
- Less affected by extreme values
- Keeps distribution shape

Name: **Priyanshu Sahu**
Batch: **Data Analytics Nov Live Batch**

Example:

If one person has age = 200 (outlier),

- Min-Max will distort all values
- Z-score will handle it better

Therefore, **Z-score is preferred when outliers exist.**