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

Answer 1: Data Transformation in ETL is the process of converting raw data extracted from source systems into a clean, consistent, and usable format before loading it into a target system such as a data warehouse or data lake.

It involves operations like:

- Cleaning data (handling nulls, duplicates, incorrect values)
- Changing data types and formats (e.g., string to date, currency conversion)
- Applying business rules (calculations, aggregations, validations)
- Standardizing data (naming conventions, units of measurement)
- Joining or splitting datasets

Why Data Transformation is Important

1. **Ensures Data Quality**
   Raw data often contains errors, missing values, and inconsistencies. Transformation improves accuracy and reliability.
2. **Maintains Consistency Across Sources**
   Data from multiple systems may follow different formats or standards. Transformation makes data uniform.
3. **Supports Business Logic & Analytics**
   Business rules and calculations are applied during transformation so reports and dashboards reflect correct insights.
4. **Improves Query Performance**
   Transformed and structured data is optimized for faster reporting and analysis.
5. **Enables Decision-Making**
   Clean and meaningful data helps stakeholders make informed business decisions.

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

Answer 2: Four common activities involved in Data Cleaning are:

1. **Handling Missing Values**
   Filling missing data using methods like mean/median, default values, or removing incomplete records.
2. **Removing Duplicates**
   Identifying and deleting duplicate records to avoid incorrect counts and analysis errors.

3. **Correcting Inconsistent Data**
   Fixing inconsistencies such as different spellings, formats, or units (e.g., “NY”, “New York”).
4. **Validating and Correcting Data Types**
   Ensuring values match the expected data type (e.g., numbers stored as text, invalid dates).

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

Answer 3: **Normalization--** makes values comparable by bringing them into the same range.

- Scaling data to a fixed range, usually **0 to 1**
- **Formula-- (X-min)/(max-min)**
- Changes the scale of the distribution but not the shape.
- Sensitive to outliers
- Used in Distance-based algorithms (KNN, Neural Networks),i.e., when bounds are known.

Standardization-- makes data centered around zero with consistent variance.

- Rescaling data so it has **mean = 0** and **standard deviation = 1.**
- **Formula— (X-mean)/standard deviation**
- Centers data of the distributionand changes spread.
- Less sensitive to outliers than normalization.
- Used in Statistical models, ML algorithms like Linear Regression, SVM,i.e, when data follows normal distribution.

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 4: **1. Imputation using Mean or Median**

- **What it is:**
  Replace missing age values with the **mean** or **median** age of the dataset.
- **When to use:**
  - When the dataset is **large**
  - When age values are **numerical**
  - **Median** is preferred if the data has **outliers**

- **Mean** is suitable if data is normally distributed

Example:
If the median age is 28, replace all missing age values with 28.

Removing Records with Missing Age

- **What it is:**
  Delete rows where the age value is missing.
- **When to use:**
  - When only a **small number of records** have missing age values
  - When age is **not a critical attribute**
  - When removing rows will not affect overall data patterns

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

Answer 5: **Step 1: Convert to a common case**

Convert all values to lowercase (or uppercase) to remove case sensitivity.

Step 2: Map values to standard labels

Original Value	Standardized Value
M	Male
male	Male
MALE	Male
F	Female
f	Female
Female	Female

Final Standardized Output: ["Male", "Male", "Female", "Female", "Male", "Female"]

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

Answer 6: **One-Hot Encoding** is a data transformation technique used to convert **categorical variables** into a **numerical format** so they can be used by machine learning algorithms.

How it Works

- Each unique category becomes a **separate binary column**
- A value of **1** indicates the presence of that category
- A value of **0** indicates absence

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

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

Answer 7:

Aspect	Data Integration	Data Mapping
Definition	The process of **combining data** from multiple source systems into a single unified view	The process of **defining relationships** between source fields and target fields
Purpose		To ensure data flows correctly from source to target

	To bring data together for analysis and reporting	
Focus	Data consolidation	Field-to-field transformation rules
Scope	Broader process	A specific step within integration
Example	Merging CRM and Sales database data into a data warehouse	Mapping `cust_id → customer_id`, `dob → birth_date`
Occurs When	During the overall ETL pipeline	During transformation and load design

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

Answer 8: **Z-score Standardization** is preferred over **Min-Max Scaling** when outliers exist because it is **less sensitive to extreme values**.

Min-Max Scaling

- Scales values using the **minimum and maximum** values of the dataset.
- **Outliers strongly affect** the min and max values.
- As a result, most data points get **compressed into a narrow range**, reducing their usefulness.

Example:
If one age value is 120 while most ages are between 20–60, Min-Max scaling will shrink the majority of values close to 0.

Z-score Standardization

- Scales data based on **mean and standard deviation**.
- Outliers increase the standard deviation but **do not distort the entire scale**.
- Preserves relative distance between normal data points.