# Transformation in ETL

**QUESTION 1**

**Define Data Transformation in ETL and explain why it is important**

**ANSWER**

Data Transformation is the process of converting raw, extracted data into a clean, structured,
and meaningful format that is suitable for analysis, reporting, and loading into a target system such as a data warehouse.

E (Extract) → Collect raw data from sources
T (Transform) → Clean, standardize, enrich, and reshape the data
L (Load) → Store the transformed data in the target system

Data cleaning – removing nulls, duplicates, and errors
Data standardization – formatting dates, currencies, text cases
Data type conversion – string → date, int → decimal
Filtering – keeping only required records
Aggregation – sum, count, average, max, min
Derivation – creating new columns (e.g., profit = revenue − cost)
Joining / merging – combining data from multiple sources
Encoding / mapping – replacing codes with meaningful values

Why Data Transformation Is Important

**Improves Data Quality**

Raw data is often inconsistent and messy.
Transformation ensures:

accuracy
consistency
completeness

**Makes Data Analysis Possible**
**Supports Business Rules**
**Improves Performance**

| | |
|---|---|
| **QUESTION 2** | **List any four common activities involved in Data Cleaning.** |
| **ANSWER** | **1. Handling Missing Values** |
| | **2. Removing Duplicate Records** |
| | **3. Correcting Inconsistent Data** |
| | **4. Validating and Filtering Data** |
| | |
| | **1. Handling Missing Values** |
| | Identify null or blank values |
| | Fill them using mean/median/mode, default values, or remove the records |
| | Prevents incorrect analysis due to incomplete data. |
| | |
| | **2. Removing Duplicate Records** |
| | Detect repeated rows or duplicate entries |
| | Keep only one correct version |
| | Avoids double counting and inaccurate results. |
| | |
| | **3. Correcting Inconsistent Data** |
| | Standardize formats (dates, text case, units) |
| | Fix spelling mistakes and inconsistent naming |
| | Ensures uniform and reliable data. |
| | |
| | **4. Validating and Filtering Data** |
| | Remove invalid or out-of-range values (e.g., negative age, future dates) |
| | Apply business rules |
| | Improves accuracy and data integrity. |

| | |
|---|---|
| **QUESTION 3** | **What is the difference between Normalization and Standardization?** |
| | **Normalization** |
| **ANSWER** | Normalization is a data preprocessing technique used to scale numeric data into a fixed range, usually 0 to 1, so that all features contribute equally to analysis or model training. |
| | Different features may have different scales |
| | Large values can dominate calculations, especially distance-based methods |
| | Normalization ensures fair comparison between features |
| | **Standardization** |
| | Standardization is a data preprocessing technique used to rescale numeric data so that it has a mean of 0 and a standard deviation of 1. |
| | It is also called Z-score normalization. |
| | Mean becomes 0 |
| | Standard deviation becomes 1 |
| | No fixed output range (values can be negative or >1) |
| | Where Standardization Is Used |
| | Linear Regression |
| | Logistic Regression |
| | Support Vector Machines (SVM) |
| | Principal Component Analysis (PCA) |

| | |
|---|---|
| **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** | **Mean / Median Imputation** |
| | Replace missing values in the Age column with: |
| | **Mean age or** |
| | **Median age** |
| | When to use: |
| | Data is numerical |
| | Missing values are few |
| | Median is preferred when data has outliers |
| | Distribution is roughly symmetric (for mean) |
| | **Deletion (Removing Rows)** |
| | Remove rows where Age is missing |
| | Missing values are very few |
| | Dataset is large |
| | Missing data appears random |
| | Age is not a critical feature |

| **Example:** |
|---|
| **average age = 35** |
| **Missing Age = 35** |

| | |
|---|---|
| **QUESTION 5** | **Convert the following inconsistent “Gender” entries into a standardized format (“Male”, “Female”):** |
| | ["M", "male", "F", "Female", "MALE", "f"] |
| **ANSWER** | **Standardization Rule** |
| | **F, M, male, MALE → Male** |
| | **M, F, f, Female → Female** |

| **convert the output** | **OUTPUT** |
|---|---|
| m | **MALE** |
| f | **FEMALE** |
| male | **MALE** |
| female | **FEMALE** |
| MALE | **MALE** |
| F | **FEMALE** |

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

**ANSWER**

One-Hot Encoding is a data preprocessing technique used to convert categorical variables into a numerical format that machine-learning models can understand.

| | RED | BLUE | GREEN |
|---|---|---|---|
| RED | 1 | 0 | 0 |
| BLUE | 0 | 1 | 0 |
| GREEN | 0 | 0 | 1 |

**Why One-Hot Encoding Is Used**

Machine-learning models cannot work directly with text

Prevents models from assuming an order between categories

| | |
|---|---|
| **QUESTION 7** | **Explain the difference between Data Integration and Data Mapping in ETL.** |
| **ANSWER** | Data Integration is the process of collecting and combining data from multiple heterogeneous sources such as databases, files and APIs into a single, unified view, usually in a data warehouse. Its main goal is to ensure that data from different systems works together consistently and can be used for reporting and analysis.<br><br>Data Mapping, on the other hand, is the process of defining how individual fields from the source systems correspond to fields in the target system. It specifies the rules for data movement and transformation, such as mapping cust_id to customer_id or converting date formats.<br><br>**Data integration focuses on bringing data together, while data mapping focuses on correctly matching and transforming each data element.** |

| | |
|---|---|
| **QUESTION 8** | **Explain why Z-score Standardization is preferred over Min-Max Scaling when outliers exist.** |
| **ANSWER** | Z-score standardization is preferred over Min-Max scaling when outliers exist because it is less sensitive to extreme values. |
| | Min-Max scaling rescales data using only the minimum and maximum values. When outliers are present, these extreme values stretch the range |
| | Z-score standardization, on the other hand, rescales data based on the mean and standard deviation. Outliers do affect the mean and standard deviation |
| | but they do not force all other values into a narrow range. Instead, outliers naturally receive large positive or negative Z-scores, while normal values remain well distributed around zero. |
| | Preserves relative differences between typical data points |
| | Makes outliers clearly identifiable |
| | Works better for models that assume normally distributed data |