Transformation_in_ETL

Q1 : Define Data Transformation in ETL and explain why it is important.

Data transformation is the process of converting extracted data into a suitable format by cleaning, standardizing, and applying business rules. It is important because it improves data quality, ensures consistency, and prepares data for analysis.

Q2 : List any four common activities involved in Data Cleaning.

Four Common Activities Involved in Data Cleaning

1. Handling Missing Values - Filling, removing, or flagging missing data.
2. Removing Duplicate Records - Identifying and eliminating repeated entries.
3. Correcting Inconsistent Data - Standardizing formats such as dates, units, and text values.
4. Detecting and Handling Outliers - Identifying extreme values and treating them appropriately.

Q3 : What is the difference between Normalization and Standardization?
Difference Between Normalization and Standardization

Aspect	Normalization	Standardization
Definition	Rescales data to a fixed range	Rescales data to have mean 0 and standard deviation 1
Formula	($x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$)	($x' = \frac{x - \mu}{\sigma}$)
Resulting Range	Usually **0 to 1**	No fixed range
Sensitivity to Outliers	Sensitive	Less sensitive
Best Used When	Data has bounded range, no extreme outliers	Data follows (or is close to) normal distribution
Common Use	Distance-based algorithms (KNN, NN)	Algorithms assuming normality (SVM, Linear Regression)

Q4 : A dataset has missing values in the “Age” column. Suggest two techniques to handle this and explain when they should be used.

Missing age values can be handled using mean/median imputation or by deletion/flagging, depending on data distribution and the significance of missingness.
1) Mean / Median Imputation
Replace missing age values with the mean or median of the column.

When to Use:

- Mean: When age data is normally distributed and has no extreme outliers
- Median: When age data is skewed or contains outliers

2) Deletion or Flagging

- Deletion: Remove rows with missing age values
- Flagging: Add a new column indicating whether age was missing

When to Use:

- Deletion: When very few values are missing and dataset is large
- Flagging: When missingness itself carries useful information

Q5 : Convert the following inconsistent “Gender” entries into a standardized format (“Male”, “Female”): ["M", "male", "F", "Female", "MALE", "f"]

By using SQL we can standardized format as -

```
SELECT
  CASE
    WHEN LOWER(gender) IN ('m', 'male') THEN 'Male'
    WHEN LOWER(gender) IN ('f', 'female') THEN 'Female'
    ELSE gender
  END AS standardized_gender
FROM employees;
```

Q6 : What is One-Hot Encoding? Give an example with the categories: “Red, Blue, Green”.

One-Hot Encoding is a data transformation technique used to convert categorical variables into binary (0/1) columns, so that they can be used in data analysis and machine learning models.

Color
Red
Blue
Green

Color_Red	Color_Blue	Color_Green
1	0	0
0	1	0
0	0	1

Each category becomes a separate column, 1 indicates presence of the category, 0 indicates absence.

Q7 : Explain the difference between Data Integration and Data Mapping in ETL.

Difference Between Data Integration and Data Mapping in ETL:

Data Integration	Data Mapping
Combines data from multiple sources into a unified dataset	Defines how source fields correspond to target fields
To create a single, consistent view of data	To ensure correct transformation and loading
Focuses on broad process involving multiple systems	Focuses on specific step within integration/transformation
Produces integrated dataset	Produces mapping rules or logic
Merging sales data from different databases	Mapping cust_id → customer_id

Q8 : Explain why Z-score Standardization is preferred over Min-Max Scaling when outliers exist.

Z-score standardization rescales data based on the mean and standard deviation, while Min–Max scaling rescales data using the minimum and maximum values.
When outliers exist:

- Min–Max scaling is highly affected because extreme values stretch the range, compressing most data points into a narrow interval.
- Z-score standardization is less sensitive to outliers, as it measures how far a value is from the mean in terms of standard deviations.

Therefore, Z-score standardization preserves the relative distribution of the data better in the presence of outliers.