**Data Cleaning and Data Preprocessing**

"Data cleaning" and "data preprocessing" are two terms often used interchangeably, but they represent slightly different stages and processes within the overall data preparation pipeline:

1. **Data Cleaning:**

* Data cleaning refers to the process of identifying and correcting errors or inconsistencies in a dataset before it is analyzed or used for modeling.
* This process involves handling missing data, dealing with outliers, correcting inaccuracies, removing duplicates, and resolving inconsistencies in the data.
* Data cleaning aims to ensure that the dataset is accurate, complete, and consistent for further analysis or processing.

**Example: Customer Dataset Cleaning**



You receive a raw dataset from a retail store. It contains customer details such as name, age, email, and purchase amount. But the data is inconsistent and contains several issues.



**Raw Dataset (Before Cleaning)**



| **ID** | **Name** | **Age** | **Email** | **Purchase Amount** |
| --- | --- | --- | --- | --- |
| 1 | Alice | 25 | alice@example.com | 500 |
| 2 | Bob | - | bob[at]example.com | 250 |
| 3 | Charlie | 999 | charlie@example.com | NULL |
| 4 | Alice | 25 | alice@example.com | 500 |
| 5 |  | 30 | dave@example.com | -100 |
| 6 | Emma | NULL | emma@example.com | 300 |

**Issues Identified**

* Missing values in **Age**, **Name**, and **Purchase Amount**



* Invalid email format for **Bob**
* Outliers: **Age = 999**, **Purchase Amount = -100**
* Duplicate row (Row 1 and 4 are identical)
* NULL entries

**Cleaned Dataset (After Cleaning)**

| **ID** | **Name** | **Age** | **Email** | **Purchase Amount** |
| --- | --- | --- | --- | --- |
| 1 | Alice | 25 | alice@example.com | 500 |
| 2 | Bob | 28 | bob@example.com | 250 |
| 3 | Charlie | 35 | charlie@example.com | 0 |
| 5 | Dave | 30 | dave@example.com | 0 |
| 6 | Emma | 32 | emma@example.com | 300 |

**Cleaning Actions Taken**

| **Problem** | **Action Taken** |
| --- | --- |
| Missing age for Emma | Imputed with median age (e.g., 32) |
| Age = 999 | Replaced with estimated valid age (e.g., 35) |
| Missing name (Row 5) | Replaced with placeholder “Dave” from email |
| Purchase Amount = NULL / -100 | Replaced with 0 or imputed |
| Invalid email format | Corrected format: "bob[at]" → "bob@" |
| Duplicate row (Alice) | Removed duplicate |

1. **Data Preprocessing:**

* Data preprocessing is a broader term that encompasses various transformations applied to raw data to prepare it for analysis or modeling.
* This includes data cleaning but also involves additional steps such as feature scaling, feature selection, dimensionality reduction, and data transformation (e.g., normalization, standardization).
* The goal of data preprocessing is to transform the raw data into a format that is suitable for the specific requirements of the analysis or modeling techniques being applied.
* Data preprocessing can also involve tasks like encoding categorical variables, handling imbalanced datasets, and splitting the data into training and testing sets.

**Example: Customer Churn Dataset Preprocessing**

You have a raw dataset from a telecom company to predict customer churn. The dataset includes various types of features: numerical, categorical, and missing values.

**Raw Dataset (Before Preprocessing)**

| **CustomerID** | **Gender** | **Age** | **MonthlyCharges** | **Contract Type** | **Churn** |
| --- | --- | --- | --- | --- | --- |
| C001 | Male | 25 | 29.85 | Month-to-month | Yes |
| C002 | Female | 45 | 56.95 | One year | No |
| C003 | Male | 999 | 42.30 | Two year | No |
| C004 | Female |  | 89.10 | Month-to-month | Yes |
| C005 | Male | 33 | 0 | One year | No |

**Issues Identified**

* Age = 999 is likely an outlier
* Missing Age value (C004)
* MonthlyCharges has 0, which might be an anomaly
* “Contract Type” is categorical
* “Churn” is a categorical label
* All features are in different scales

**Preprocessing Actions**

| **Step** | **Description / Technique** |
| --- | --- |
| **Impute Missing Values** | Fill missing Age using median or mean |
| **Fix Outliers** | Replace Age = 999 with reasonable value (e.g., 40) |
| **Encode Categorical Data** | One-hot encode Contract Type, label encode Churn |
| **Feature Scaling** | Normalize or standardize Age, MonthlyCharges |
| **Split Dataset** | Split into train/test sets (e.g., 80%/20%) |

**Cleaned + Preprocessed Dataset (Numeric)**

| **CustomerID** | **Gender\_Male** | **Gender\_Female** | **Age** | **MonthlyCharges** | **Contract\_1Y** | **Contract\_2Y** | **Contract\_M2M** | **Churn (0=No, 1=Yes)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C001 | 1 | 0 | 25 | 29.85 | 0 | 0 | 1 | 1 |
| C002 | 0 | 1 | 45 | 56.95 | 1 | 0 | 0 | 0 |
| C003 | 1 | 0 | 40 | 42.30 | 0 | 1 | 0 | 0 |
| C004 | 0 | 1 | 36 | 89.10 | 0 | 0 | 1 | 1 |
| C005 | 1 | 0 | 33 | 15.00 | 1 | 0 | 0 | 0 |

✱ Note: Gender is encoded as binary columns, Contract type is one-hot encoded, Age and MonthlyCharges may be standardized/scaled before model training.

In summary, while data cleaning specifically focuses on identifying and correcting errors or inconsistencies in the dataset, data preprocessing includes a broader set of transformations aimed at preparing the data for analysis or modeling. Data preprocessing typically includes data cleaning as one of its components.