What Is Data Cleaning?

Data Cleaning (also called **Data Preprocessing** or **Data Wrangling**) is the process of **detecting, correcting, and handling errors or inconsistencies** in raw data to make it suitable for analysis or modeling.

When you collect data from multiple sources (Excel, databases, APIs, web scraping, surveys, etc.), it often contains:

- Missing values
- Duplicates
- Incorrect data types
- Inconsistent formatting
- Outliers
- Typographical or human errors

Before any analysis or visualization, you must clean and prepare this data so that your results are accurate and reliable.

Why Is Data Cleaning Important?

1. **Ensures Accuracy:**
   Dirty data → Wrong insights → Bad business decisions.
   Clean data ensures the insights you derive actually reflect reality.

2. **Improves Efficiency:**
   A clean dataset saves time during analysis. You won't need to fix problems repeatedly.

3. **Improves Model Performance (in ML):**
   Clean data leads to more stable, high-performing models.

4. **Increases Trust in Data:**
   Stakeholders are more likely to trust and use your analysis if data quality is ensured.

5. **Supports Better Decision-Making:**
   Reliable insights come from reliable data — this is the foundation of data-driven decisions.

The Data Cleaning Process (Step-by-Step)

Step 1: Data Collection and Inspection

Before cleaning, you need to understand your data:

- How many rows and columns are there?
- What type of data do you have in each column (numeric, categorical, datetime, text)?
- What do missing values or anomalies look like?
- Are there any obvious duplicates or format inconsistencies?

Techniques:

- Data types and null value counts
- Summary statistics (`mean`, `median`, `min`, `max`)
- Distribution plots for numerical data
- Frequency counts for categorical data
- Visual inspection (head/tail of dataset)

Step 2: Handling Missing Data

- **Delete Rows/Columns:** If too much data is missing or unimportant.

- **Imputation (Filling values):**

    - Mean/Median/Mode for numeric
    - Mode/ Unknown for categorical
    - Forward/Backward fill for time series
    - Predictive imputation using ML models

Step 3: Fixing Inconsistent Formatting

In real-world datasets:

- Names may have inconsistent casing (e.g., "India", "india", "INDIA") or spelling errors (e.g., "Male", "male ", "M")
- Dates may have different formats
- Currencies may vary
- Extra spaces or typos in categorical fields

Fixes:

- Standardize casing (lowercase, titlecase)
- Unify date and time formats

- Trim whitespace
- Replace symbols or unwanted characters

Step 4: Handling Duplicates

- Identify exact duplicates.
- Check for near-duplicates (same name, different spelling).
- Remove or consolidate them

Step 5: Correcting Data Types

Each column must have the **correct data type**:

- Numeric → Integer/Float
- Text → String/Object
- Dates → Datetime format
- Boolean → True/False

Step 6: Handling Outliers

Types of Outliers:

- Genuine (e.g., a billionaire in income data)
- Errors (e.g., extra zero: 50000 instead of 5000)

Handling Techniques:

- Remove extreme outliers with Statistical methods (Z-score, IQR)
- Domain knowledge (is that value possible?)
- Transform data (log/sqrt)