

## ◆ Data Cleaning Steps Using Power Query (Power BI)

Power Query provides a **visual and step-by-step transformation environment**, where each action is recorded automatically.

---

### 1 Importing Dataset

- Data imported from Excel / CSV / online sources into Power BI
  - Dataset opened in **Power Query Editor**
  - Initial inspection of rows, columns, and data types
- 

### 2 Removing Unnecessary Columns

- Identified columns not required for analysis
- Removed them to reduce complexity and improve performance

**Purpose:**

- Focus only on relevant attributes
  - Reduce memory usage
- 

### 3 Handling Missing Values

- Detected null values in multiple columns
  - Replaced nulls with:
    - "Unknown" for categorical columns
    - Default values where appropriate
  - In some cases, rows with excessive missing data were removed
- 

### 4 Changing Data Types

- Converted incorrect data types:
  - Text → Number
  - Text → Date
- Ensured numeric fields and dates were correctly formatted

**Importance:**

- Prevents calculation and visualization errors
- 

### Cleaning Text Data

- Used **Trim** to remove leading and trailing spaces
  - Used **Clean** to remove non-printable characters
  - Standardized text case and formatting
- 

### Removing Duplicate Records

- Identified duplicate rows
  - Removed duplicates to maintain data accuracy
- 

### Splitting Columns

- Split columns containing multiple values (e.g., address, duration)
- Created separate, meaningful columns

#### Example:

- Duration → Duration Value + Duration Type
- 

### Standardizing Categorical Values

- Corrected inconsistent labels (e.g., Yes/yes/Y)
  - Ensured uniform category names
- 

### Filtering Invalid Records

- Removed rows with missing key values
  - Filtered out unwanted or invalid entries
- 

### Final Review and Apply Changes

- Verified all transformations
  - Applied changes and loaded clean data into Power BI model
-

## ◆ Data Cleaning Steps Using Python (Pandas)

Python provides a **flexible, automated, and repeatable approach** to data cleaning using code.

---

### Importing Libraries and Dataset

- Imported Pandas and NumPy
- Loaded dataset from Excel or CSV file

**Purpose:**

- Prepare environment for data manipulation
- 

### Understanding Dataset Structure

- Checked:
  - Shape of dataset
  - Column names
  - Data types
  - Summary statistics

**Benefit:**

- Helps identify data quality issues early
- 

### Removing Duplicate Records

- Identified duplicate rows
- Removed duplicates using Pandas functions

**Importance:**

- Prevents repeated values from skewing analysis
- 

### Handling Missing Values

- Detected null values in each column
- Applied different strategies:
  - Forward fill / backward fill

- Replacing with blanks or default values
  - Dropping rows if necessary
- 

### Dropping Unnecessary Columns

- Removed irrelevant or unused columns
  - Reduced dataset size and noise
- 

### Cleaning Text Columns

- Removed special characters
  - Trimmed extra spaces
  - Standardized text formatting
- 

### Standardizing Categorical Data

- Replaced inconsistent values
- Converted multiple representations into a single standard value

#### Example:

- Yes / Y → Yes
  - No / N → No
- 

### Filtering Records

- Removed records based on conditions
  - Excluded invalid or restricted entries
- 

### Splitting Columns

- Split composite columns into multiple meaningful columns
  - Improved data structure and readability
-