

DATA CLEANING PROCESS:

- Data cleaning is the most important stage in Data analysis.
 - It ensures that the dataset is accurate, consistent, and ready for further analysis and visualization.
- The following steps summarize the cleaning operations performed,

1. Importing Required Libraries

The analysis begins by loading essential Python libraries such as:

- Pandas — Data loading and cleaning
- NumPy — Handling numerical operations

2. Loading the Dataset

The raw dataset was imported using `pandas.read_csv()`. Initial inspection included:

- Displaying sample rows
- Checking column names and data structure

3. Identifying Missing Values

Missing and null values were detected using:

`df.isnull().sum()`

- Columns with fewer missing values were imputed while heavily missing ones were removed.

4. Handling Duplicates

Duplicates were checked and removed to avoid counting entries twice.

5. Data Type Corrections

Mismatched datatypes were corrected, including:

- Converting date strings to datetime
- Converting numeric columns using astype()

6. Removing Irrelevant Columns

Unnecessary fields that do not contribute to analysis were removed.

7. Standardizing and Cleaning Text Fields

Text inconsistencies were handled by trimming spaces, fixing cases, and removing special characters.

8. Final Validation

Summary statistics and shape checks were performed to ensure data quality.

9. Exporting Clean Data

The cleaned file was saved as a new CSV file ready for SQL querying and dashboard creation.