Data Collection & Cleaning

Overview:
Data cleaning is the process of identifying and correcting (or removing) inaccurate records from a dataset. It ensures the quality, consistency, and accuracy of data used for analysis and model training.

Key Concepts:

- **Removing Duplicates:**
  Duplicate rows can distort results and bias model outcomes. Using Pandas (df.drop_duplicates()), we can remove repeated entries and keep only unique records.
- **Handling Missing Values:**
  Missing data can occur due to collection errors or incomplete entries.
  - **Options to handle:**
    - Remove rows/columns with too many missing values (df.dropna()).
    - Fill missing values with the mean, median, or mode (df.fillna()).
- **Treating Outliers:**
  Outliers are extreme values that differ significantly from other observations.
  They can be detected using **boxplots** or statistical methods like the **IQR (Interquartile Range)** rule, and treated by capping, transforming, or removing them.

Importance:
Clean data ensures reliable and valid analysis. Without cleaning, even the most powerful models will produce misleading results.