# Data Wrangling/Preprocessing: (Data Cleaning)

**Notes by Mannan Ul Haq (BDS-3C)**

Data in the real world often arrives in a less-than-ideal state, being dirty in various ways: it can be incomplete, containing missing values or lacking important attributes; noisy, with errors and outliers; or inconsistent, with discrepancies in codes or names. This is why the data cleaning process is crucial for refining and preparing data for analysis.

## 1. Handling Missing Values:

Sometimes, your dataset may have gaps or missing information.

- **Identifying Missing Values**: Start by identifying where your data has missing values. These are usually represented as blanks, **"NaN"** (Not-a-Number), or other placeholders.
- **Handling Strategies**: There are several ways to handle missing values:
    - **Imputation**: Fill in missing values with appropriate replacements. This can be done using the mean, median, mode, or even more complex imputation methods based on the nature of your data.
    - **Deletion**: In some cases, if missing values are few and won't impact your analysis significantly, you can delete rows or columns with missing values.

## 2. Removing Duplicate Values:

Duplicates in your data can skew your analysis.

- **Identifying Duplicates**: Detect duplicate records by comparing rows to see if they have identical values across all or specific columns.
- **Handling Strategies**:
    - **Dropping Duplicates**: Remove duplicate records, keeping only the first occurrence.

## 3. Identifying Outliers:

Outliers are data points that are significantly different from the rest of the data.

- **Identifying Outliers**: Visualizations like box plots, scatter plots, or statistical methods can help identify them. We can also identify outliers using mathematical methods like:

  1. **Z-Score Method**: Any data point with a Z-Score greater than **3** or less than **-3** is considered an outlier.
  2. **IQR Method**: Determine the lower and upper bounds for potential outliers:

  - Lower Bound: **Q1 - 1.5 * IQR**
  - Upper Bound: **Q3 + 1.5 * IQR**
  - Any data point below lower bound or above upper bound would be considered an outlier.

- **Handling Strategies**:
  - **Remove Outliers**: In some cases, outliers may be data entry errors or anomalies. Removing them might be appropriate.
  - **Smooth Data:** Change outliers with appropriate replacements. This can be done using the median or mode etc.

## 4. Correcting Inconsistent Data:

- **Identifying Inconsistencies**: Look for inconsistencies in data, such as variations in formatting, spelling errors, or units of measurement.
- **Handling Strategies**: Correcting inconsistencies involves:
  - **Standardization**: Ensure consistent formatting for text data (e.g., capitalization) and dates.
  - **Data Validation**: Validate data against predefined rules or patterns to catch inconsistencies.
  - **Conversion**: Convert units of measurement to a consistent format.
  - **Imputing Correct Data**: Replace incorrect data with the correct values when possible.

## 5. Handling Noisy Data:

- **Identifying Noisy Data**: Noisy data contains random variations or errors that can affect analysis.
- **Handling Strategies**: Strategies for handling noisy data include:
  - **Smoothing**: Apply smoothing techniques (e.g., moving averages) to reduce noise.
  - **Outlier Detection**: Use statistical methods to identify and handle outliers.
  - **Data Binning**: Group data into bins to reduce noise and identify patterns.