# How to do Data Cleaning

## 1. Handle Missing Data

- Identify Missing Values: Use methods like `isnull()` in Python’s Pandas to locate missing values in your dataset.  
 - Remove Rows or Columns with Missing Values: If missing data is minimal, you can simply drop those rows or columns using `dropna()`.  
 - Impute Missing Values: When missing data is substantial, fill in values:  
 - Numerical Data: Use mean, median, or mode imputation.  
 - Categorical Data: Use the mode (most frequent value) or create a new category for missing values.  
 - Advanced Imputation Techniques: Use model-based techniques like k-Nearest Neighbors (k-NN), multivariate imputation, or regression models to predict missing values.

## 2. Remove Duplicates

- Detect Duplicates: Identify duplicate rows with methods like `duplicated()` in Pandas.  
 - Remove Duplicates: Use `drop\_duplicates()` to remove duplicate entries while keeping the first occurrence, or customize by retaining the last occurrence or removing duplicates based on specific columns.

## 3. Handle Outliers

- Visualize Outliers: Use box plots, scatter plots, or histograms to spot unusual data points in continuous variables.  
 - Remove Outliers: Outliers can sometimes be removed if they are clearly erroneous (e.g., impossible values).  
 - Cap Outliers: Cap outliers at a certain threshold, such as the 95th percentile.  
 - Transform Data: Apply transformations like log or square root to reduce the effect of outliers if they’re not errors but naturally extreme.

## 4. Fix Structural Errors

- Standardize Values: Clean up inconsistencies in categorical data (e.g., “N/A”, “na”, “-” as placeholders for missing data) by replacing with `NaN` or a single standard value.  
 - Correct Misspellings and Typo Errors: Standardize similar values (e.g., “Male” and “M” should be consistent).  
 - Consistent Formatting: Ensure consistent data types (e.g., dates in `YYYY-MM-DD` format or phone numbers with consistent formatting).

## 5. Filter Irrelevant Data

- Drop Unnecessary Columns: Identify columns that won’t contribute to the analysis or modeling and drop them.  
 - Remove Unnecessary Rows: Exclude rows based on criteria like filtering out unneeded categories or irrelevant values.

## 6. Standardize Units and Scales

- Convert Units: Make sure all units are consistent, such as converting all distances to meters or weights to kilograms.  
 - Convert Scales: If certain features are on different scales (e.g., some are in percentage and others in ratio), standardize them for consistency.

## 7. Validate Data Integrity

- Check Range Validity: Ensure numerical values fall within a reasonable range (e.g., age should be between 0 and 120).  
 - Cross-Field Validation: Ensure values make sense when compared across fields (e.g., a `Start Date` should precede an `End Date`).

## 8. Encoding Categorical Variables

- Encode Text Data: Convert categorical text data into numerical form if needed, like label encoding or one-hot encoding.

## 9. Save the Cleaned Dataset

- Export the Cleaned Data: Save the cleaned dataset for further processing and keep an original version as a backup.