Steps for Data Preprocessing and Data Cleaning

1. Understand the Data

- **Inspect the Dataset**: Examine the structure, size, and types of data in the dataset.
    - Reference: Check the first few rows (head), column names, and basic statistics.
- **Define the Objective**: Understand how each feature relates to the target variable.

---

2. Handle Missing Data

- **Identify Missing Values**: Analyze missing values column-wise and row-wise.
    - Reference: Summarize missing data as percentages or counts.
- **Handle Missing Data**:
    - Remove rows/columns with excessive missing data.
    - Impute missing values using mean, median, or mode for numerical features or the most frequent category for categorical features.
    - For time-series data, consider forward-fill or backward-fill techniques.

---

3. Remove Duplicates

- **Identify Duplicates**: Detect rows that are repeated in the dataset.
    - Reference: Use unique identifiers or all columns for comparison.
- **Remove Duplicates**: Keep the first occurrence and remove additional duplicates.

---

4. Handle Outliers

- **Detect Outliers**: Identify anomalies in numerical features using statistical methods:
    - Reference: Use interquartile range (IQR) or standard deviation to flag outliers.
- **Decide on Outlier Handling**:
    - Remove: For extreme outliers that result from data entry errors.
    - Transform: Apply log or square-root transformations to reduce the effect of extreme values.
    - Cap: Limit outliers to a maximum or minimum threshold.

---

5. Fix Data Types

- **Convert Data Types**: Ensure features are in their correct formats:
    - Convert date strings to datetime objects.
    - Convert numerical strings (e.g., "123.45") to float or integer.

    - Reference: Use metadata or domain knowledge to verify formats.

---

6. Handle Inconsistent Data

- **Standardize Text Data**: Fix inconsistent spelling, capitalization, or abbreviations in categorical variables.
    - Example: Normalize "NYC," "New York City," and "new york city" into one category.
- **Fix Formatting Issues**:
    - Ensure consistent units for numerical features (e.g., meters vs. kilometers).

---

7. Encode Categorical Data

- **Identify Categorical Features**: List features containing textual or categorical values.
- **Transform to Numeric**: Use encoding methods based on data and model requirements:
    - One-hot encoding for nominal variables (unordered categories).
    - Ordinal encoding for ordered categories.

---

8. Standardize/Normalize Data

- **Standardization**: Scale numerical features to have a mean of 0 and a standard deviation of 1.
    - Reference: Use for algorithms sensitive to feature scales (e.g., SVM, logistic regression).
- **Normalization**: Scale numerical features to a range of 0 to 1.
    - Reference: Use for features with widely varying scales.

---

9. Feature Engineering

- **Create New Features**: Combine existing features to derive new ones.
    - Reference: Extract useful insights from date-time (e.g., day of week, month).
- **Transform Features**: Apply transformations (e.g., log, exponential) to reduce skewness in data.
- **Remove Irrelevant Features**: Drop features that do not contribute to the target variable.

---

10. Split the Data

- **Train-Test Split**: Partition the dataset into training and testing subsets.

    - Reference: Typically 70% for training and 30% for testing, depending on the dataset size.
- **Stratify Data**: Maintain class distribution in classification tasks to avoid biased splits.

---

11. Check for Multicollinearity

- **Correlation Analysis**: Identify highly correlated features using correlation matrices.
    - Reference: Features with high correlation (e.g., >0.85) may cause multicollinearity.
- **Feature Removal**: Drop one of the correlated features to reduce redundancy.

---

12. Final Verification

- **Validate Dataset**: Ensure no missing values, duplicates, or inconsistencies remain.
- **Visualize Cleaned Data**: Use box plots, histograms, and scatter plots to confirm data quality.
- **Document Cleaning Steps**: Maintain a log of all changes made to the dataset for reproducibility.