Data Cleaning Checklist & AI Prompts: Prepare Your Data for Reliable Insights

Your Definitive Guide to Using ChatGPT and Microsoft Copilot for Efficient Data Cleaning

Introduction:

Clean data is the foundation of accurate analysis, reliable insights, and effective decision-making. The principle of "Garbage In, Garbage Out" (GIGO) is paramount: flawed data will always lead to flawed results, no matter how sophisticated your analysis tools are. This checklist provides a comprehensive, step-by-step guide to data cleaning, including practical example prompts for AI assistants like ChatGPT and Microsoft Copilot to accelerate your workflow significantly.

Important: Data Privacy and Security

- **Never** upload sensitive data (Personally Identifiable Information (PII), financial records, health information, etc.) to a general-purpose AI tool like the public ChatGPT interface without proper *anonymization* and *security measures*.
- Copilot, within a secure Microsoft 365 environment, generally offers stronger data protection *but always adheres to your organization's data governance policies and any relevant legal regulations (e.g., GDPR, CCPA, HIPAA)*.
- Consider using self-hosted or enterprise-level AI solutions for sensitive data, where data remains within your organization's secure environment.

I. Data Cleaning Checklist:

Use the checkboxes ([ ]) to track your progress.

- **[ ] 1. Initial Data Inspection (Understanding Your Data):**
    - **Task:** Before making *any* changes, develop a thorough understanding of your dataset's structure, content, and potential issues.
    - **Check For:**
        - Number of rows and columns.
        - Data types of each column (text, numeric, date, categorical, etc.).
        - *Statistical Summary*: Mean, median, standard deviation, min, max, and quartiles for numerical columns.
        - *Unique Values*: Number of unique values in categorical/text columns.
        - Obvious errors or inconsistencies (e.g., text in a numeric column, illogical values).
        - Range of values (to understand the scale and spread of the data).
        - *Visualizations*: Basic histograms or box plots to visually inspect distributions (optional but highly recommended).
        - **Example Prompt (ChatGPT - Comprehensive Overview):** *"I've uploaded a dataset [or: 'Here is a sample of my data: [paste sample]']. Can you provide a comprehensive overview of the data? Include:*
            - *The number of rows and columns.*
            - *The data type of each column.*
            - *A basic statistical summary (mean, median, standard deviation, min, max, quartiles) for each numerical column.*

        - *The number of unique values for each categorical or text column.*
        - *Any immediately obvious errors, inconsistencies, or potential data quality concerns you identify.*
    - *Explain your findings in a clear and concise manner."*

    - **Example Prompt (Copilot - Excel):**
      "Provide a comprehensive data summary for this table, including descriptive statistics for all numerical columns and the number of unique values for categorical columns. Highlight any potential data quality issues."

- **[ ] 2. Handling Missing Values:**
    - **Task:** Decide on the most appropriate strategy for dealing with missing data (represented by blanks, "N/A," "NULL," etc.).
    - **Options:**
        - **Deletion:** Remove rows or columns with missing values. *Use with extreme caution*: Only if the missing data is irrelevant or a very small percentage of the dataset. Deletion can introduce bias.
        - **Imputation:** Fill in missing values with estimated values. Common methods include:
            - *Mean/Median Imputation*: For numerical data.
            - *Mode Imputation*: For categorical data (most frequent value).
            - *AI-Predicted Imputation*: Use AI to predict missing values based on other columns (more advanced).
        - **Flagging:** Add a new column (e.g., "Missing_Age") to indicate that a value was originally missing. This preserves the information about missingness.
        - **Multiple Imputation (Advanced):** Create multiple datasets with different imputed values to account for uncertainty (briefly mention for awareness).
    - **Considerations:**
        - *Missingness Pattern*: Is the data Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR)? Understanding the pattern helps choose the best approach.
        - *Impact on Analysis*: How will each method (deletion, imputation) affect downstream analysis?
        - *Context is Key*: The best approach depends heavily on the nature of your data and your analysis goals.
        - If a column has more than 50% missing data, consider removing it
    - **Example Prompt (ChatGPT - AI-Predicted Imputation):**

      "For the missing values in the 'Income' column of the attached dataset, can you suggest an appropriate imputation method using other relevant columns to predict the missing income? Explain your approach, suggest a prompt to implement it, and provide the reasoning behind your choice."

- **Example Prompt (ChatGPT - Median Imputation with Reasoning):**

  "In the attached dataset, fill in any missing values in the 'Age' column with the median age. Explain the reasoning behind using the median instead of the mean in this case."

- **Example Prompt (Copilot - Excel - Simple Fill):**

  "Fill in the blank cells in the 'Region' column by copying the value from the cell directly above." (For cases where this is appropriate)

- **Example Prompt (Copilot - Excel):**

  "Fill empty cells in 'Date Sold' to be the same as 'Date Made'"

- **[ ] 3. Standardizing Formats:**
  - **Task:** Ensure consistent formatting across all data types.
  - **Check For:**
    - **Dates:** Different date formats (e.g., MM/DD/YYYY, DD-MM-YYYY, YYYY-MM-DD). Convert to a single, consistent format (ISO 8601: YYYY-MM-DD is generally recommended).
    - **Numbers:** Inconsistent decimal separators (comma vs. period), thousands of separators. Choose one standard and apply it consistently.
    - **Currencies:** Different currency symbols and inconsistent placement of symbols. Standardize to a single currency format, if applicable.
    - **Text:** Inconsistent capitalization, extra spaces, special characters.
    - **Categorical Data:** Inconsistent category names (e.g., "US," "USA," "United States"). Standardize to a single, preferred term.
  - **Example Prompt (ChatGPT - Date Standardization):**

    "Convert all dates in the 'Order Date' column to the ISO 8601 format (YYYY-MM-DD). Handle any variations in date formats that might be present."

  - **Example Prompt (Copilot - Excel - Date Formatting):**

    "Change the format of the 'Date' column to YYYY-MM-DD."

  - **Example Prompt (ChatGPT - Categorical Standardization):**

    "In the 'Region' column, standardize all entries representing the United States to 'USA.' Identify any variations that need to be standardized."

- **Example Prompt (ChatGPT - Number Formatting):**

  "Ensure all numbers in the 'Revenue' column use a period as the decimal separator and a comma as the thousands separator. Handle any currency symbols present."

- **[ ] 4. Cleaning Text Fields:**
  - **Task:** Correct errors, inconsistencies, and unwanted characters in text data.
  - **Check For:**
    - Typos and spelling errors.
    - Extra spaces (leading, trailing, or multiple spaces between words).
    - Inconsistent capitalization (e.g., "usa," "USA," "U.S.A.").
    - Special characters that should be removed or replaced.
    - Inconsistent abbreviations or terminology.
    - Character Encoding Issues (e.g., UTF-8 vs. other encodings - briefly mention).
    - **Example Prompt (ChatGPT - Comprehensive Text Cleaning):**

      "Clean the 'Customer Name' column in the following ways:

      - *Remove any leading, trailing, or multiple spaces between words.*
      - *Standardize capitalization to Title Case (e.g., 'John Smith').*
      - *Identify and flag any potential typos or spelling errors.*
      - *Remove any special characters."*
  - **Example Prompt (Copilot - Excel - Trimming Spaces):**

    "Remove extra spaces from the 'Product Name' column."

  - **(Optional Advanced) Example Prompt (ChatGPT - Regular Expression):**

    "Using regular expressions, extract all email addresses from the 'Contact Info' column and create a new column called 'Email Address.'"

- **[ ] 5. Handling Duplicate Entries:**
  - **Task:** Identify and address duplicate rows, considering whether they are errors or meaningful data points.
  - **Consider:**
    - *Type of Duplicates*: Are they exact duplicates (entire row identical) or near duplicates (e.g., slight variations in text fields)?
    - *Meaning of Duplicates*: Are duplicates genuine re-occurrences (e.g., multiple transactions by the same customer), or are they data entry errors?
    - *Granularity of Duplicates*: Should you check for duplicates across the entire row or only based on specific key columns (e.g., Order ID, Customer ID)?

        - *Action on Duplicates*:
            - **Deletion:** Remove all but one instance of the duplicate.
            - **Aggregation/Consolidation:** Combine information from duplicate rows (e.g., summing values for duplicate orders).
            - **Flagging:** Add a column to indicate duplicate entries for further investigation.
    - **Example Prompt (ChatGPT - Duplicate Removal Based on ID):**

      "Identify and remove any duplicate rows from the dataset. Base the identification of duplicates solely on the 'Order ID' column. Keep the first occurrence of each unique Order ID."

    - **Example Prompt (Copilot - Excel - Duplicate Removal):**

      "Remove duplicate rows based on the 'Order ID' column."

    - **(Optional Advanced) Example Prompt (ChatGPT - Deduplication/Fuzzy Matching):**

      "Identify near-duplicate entries in the Company Name column. Consider company names to be near-duplicates if they are similar but not identical, accounting for typos, extra spaces, or slight variations in wording (e.g., "Acme Corp" vs. "Acme Corporation"). Suggest potential merges and a prompt to standardize."

- **[ ] 6. Identifying and Handling Outliers/Anomalies:**

    - **Task:** Detect data points that deviate significantly from the norm and determine their validity.
    - **Consider:**
        - *Business Context*: Are the outliers plausible within the context of the data and the business domain? (e.g., a very large order might be legitimate).
        - *Potential Causes*: Could outliers be data entry errors, system glitches, or genuine (but unusual) events?
        - *Action*:
            - **Keep:** If the outlier is a valid data point.
            - **Correct:** If the outlier is due to a data entry error.
            - **Remove:** If the outlier is an invalid data point and cannot be corrected.
            - **Investigate Further:** If the cause or validity of the outlier is unclear.
    - **Methods for Outlier Detection:**
        - *Statistical Thresholds*:
            - Standard Deviation: Flag values beyond a certain number of standard deviations from the mean (e.g., 3 standard deviations).
            - IQR (Interquartile Range): A more robust method that is less sensitive to extreme values.

    - *Visual Methods*: Box plots, scatter plots.
    - *AI-based Anomaly Detection (Advanced)*: Use AI models for more sophisticated anomaly detection, particularly in datasets with many columns.
  - **Example Prompt (ChatGPT - IQR Outlier Detection):**

    "Identify any outliers in the 'Sales Amount' column using the IQR (Interquartile Range) method. Explain the IQR method in simple terms and list the identified outliers, along with their corresponding row numbers."

  - **Example Prompt (Copilot - Excel - Conditional Formatting):**

    *"Use conditional formatting to highlight cells in the 'Quantity' column where the value is greater than (AVERAGE('Quantity') + 3 * STDEV.P('Quantity')) or less than (AVERAGE('Quantity') - 3 * STDEV.P('Quantity'))." (Use Excel's Conditional Formatting feature.)*

  - **(Optional Advanced) Example Prompt (ChatGPT- AI-Based Anomaly Detection**

    "Suggest an AI-based anomaly detection method suitable for this dataset with [number] columns and an example prompt to implement it."

- **[ ] 7. Data Type Consistency:**

  - **Task:** Ensure that all values within a column conform to the expected data type.
  - **Check For:**
    - Numbers are stored as text (very common).
    - Dates are stored as text.
    - Erroneous entries (e.g., text in a numeric column, invalid date formats).
  - **Example Prompt (ChatGPT - Data Type Check):**

    "Check the 'Price' column for any non-numeric entries. List any such entries along with their row numbers, and suggest how to convert them to numeric values if possible."

  - **Example Prompt (Copilot - Excel - Data Type Check):**

    "Highlight cells in the 'Price' column that are not recognized as numbers."

II. AI Prompting Best Practices (for Data Cleaning):

- **Be Specific and Explicit:** Clearly describe the task, the specific column(s) involved, and the *precise* desired outcome. Avoid ambiguous wording.
- **Chain-of-Thought Prompting:** For complex tasks, ask the AI to explain its reasoning step-by-step. This helps you understand the AI's process and identify potential errors.

- **Iterative Prompting:** Start with a broader prompt, then refine it based on the AI's response. Be prepared to experiment with different phrasings to achieve the desired result.
- **Double-Check Everything!:** *Always* manually review any changes made by AI to your data. AI can make mistakes, especially with nuanced data.

Task	Description	Example Prompt (ChatGPT)	Example Prompt (Copilot/Excel)
Data Inspection	Get a comprehensive overview of data structure, statistics, and potential issues.	"Provide a data overview: rows, columns, data types, statistics (mean, median, etc. for numeric), unique values (categorical), and any obvious issues."	"Summarize this data, including descriptive statistics and potential issues."
Missing Values	Fill in, remove, or flag missing data points.	"Calculate the median age from 'Age' and fill in missing values with it. Explain your reasoning."	"Fill in blank cells in 'Region' by copying the value above."
Format Dates	Convert dates to a consistent format.	"Convert all dates in 'Order Date' to YYYY-MM-DD."	"Change the format of 'Date' to YYYY-MM-DD."
Clean Text	Remove extra spaces, standardize capitalization, and correct typos.	"Remove leading, trailing, and multiple spaces in 'Customer Name.' Standardize the Title Case."	"Remove extra spaces from 'Product Name.'"
Remove Duplicates	Identify and remove duplicate rows based on specific criteria.	"Remove duplicate rows based on 'Order ID,' keeping the first occurrence."	"Remove duplicate rows based on 'Order ID.'"
Find Outliers	Identify unusual data points that might be errors or require further investigation.	"Identify outliers in 'Price' using the IQR method. Explain your reasoning and list the outliers with row numbers."	"Highlight values in 'Quantity' that are more than three standard deviations from the mean."
Check Data Types	Ensure data types are consistent within each column.	"Check the 'Price' column for any non-numeric entries and list them along with their row indices."	"Highlight cells in the 'Price' column that are not recognized as numbers."

IV. Advanced Techniques (Beyond Basic Cleaning):

- **Data Transformation:** AI can help you create new columns based on existing data, perform calculations, and engineer new features. This is often called "feature engineering."
  - *Example: "Create a new column called 'Total Price' by multiplying the 'Quantity' and 'Unit Price' columns."*
  - *Example: "Extract the year from the 'Order Date' column and create a new column called 'Order Year.'"*
  - *Example: "Create a new column 'High-Value Customer' which is 'Yes' if the 'Total Spend' is above $1000 and 'No' otherwise."*
  - *Example: "Calculate the delivery time in days by subtracting 'Order Date' from 'Delivery Date,' and store it in a new column called 'Delivery Time'."*
  - *Example (ChatGPT - Feature Scaling): "Suggest and apply appropriate feature scaling or normalization techniques to the numerical columns in the dataset, preparing it for a machine learning model. Explain your choices."*
  - *Example (ChatGPT - Categorical Encoding): "Encode the 'Region' column using one-hot encoding. Explain the process."*
- **Data Aggregation:** AI can group and summarize data to reveal higher-level insights and trends.
  - *Example: "Calculate the total sales, average order value, and number of orders for each region, and display the results in a table."*
  - *Example: "Create a pivot table showing the average customer rating for each product category."*
  - *Example: "Generate a monthly summary of website traffic, including total visits, unique visitors, and average session duration."*
  - *Example (Copilot/Excel): Create a pivot table showing total sales by product category and region.*
- **Data Imputation (Advanced):** AI can use more sophisticated methods to fill in missing values based on patterns in the rest of your data. *Use with caution and always validate.*
  - *Example (ChatGPT): "For the missing values in the 'Customer Age' column, use other relevant columns (e.g., 'Purchase History,' 'Location') to predict and impute the missing ages. Explain the method used for prediction."*
- **Text Feature Extraction (Advanced):** If you have text data, AI can extract useful information from it.
  - *Example:* "From the 'Product Description' column, extract keywords related to product features and create new columns for each feature."
  - *Example:* "From the customer review, extract the sentiment."
- **Bias Detection (Advanced):** AI can help identify potential biases in your data. *This is an exploratory step, and human judgment is crucial.*
  - *Example: "Analyze the 'Customer Demographics' data. Are there any significant imbalances in the distribution of customers across different demographic groups (e.g., age, gender, location)? Report any potential biases that could affect analysis results."*

V. Data Documentation & Next Steps:

- **[ ] Data Cleaning Log:** *Crucially*, create a separate document (or a section within your analysis document) to record *all* data cleaning steps taken. This is essential for reproducibility and transparency. Include:
    - Date of cleaning.
    - Description of each cleaning step (e.g., "Removed rows with missing values in 'Order ID' column").
    - The rationale for each step (e.g., "Missing Order IDs would prevent accurate order tracking").
    - The AI prompt used (if applicable).
    - The number of rows/values affected (e.g., "5 rows removed").
- **[ ] Save Cleaned Data:** Save the cleaned data as a *new file* or a *new version* to preserve the original, raw data. *Never* overwrite your original data source.
- **[ ] Prepare for Analysis:** The cleaned data is now ready for:
    - Exploratory Data Analysis (EDA) - see Video 12.
    - Data Visualization.
    - Statistical Analysis.
    - Machine Learning Model Training (if applicable).

Key Reminders:

- **AI is a Tool, Not a Replacement:** Use AI to *accelerate* the data cleaning process, but always apply your own judgment and domain expertise.
- **Iterative Process:** Data cleaning is often iterative. You may need to repeat steps or try different approaches.
- **Document Everything:** Keep a detailed record of your cleaning steps for reproducibility.
- **Security and Privacy First!** Never compromise sensitive data.

Disclaimer:

All content and material on the upGrad website is copyrighted material, either belonging to upGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access, print, and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copies of this document, in part or full, saved to disc or to any other storage medium may only be used for subsequent self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of the content for any other commercial/unauthorized purposes in any way which could infringe the intellectual property rights of upGrad or its contributors, is strictly prohibited.
- No graphics, images, or photographs from any accompanying text in this document will be used separately for unauthorized purposes.
- No material in this document will be modified, adapted, or altered in any way.
- No part of this document or upGrad content may be reproduced or stored on any other website or included in any public or private electronic retrieval system or service without upGrad's prior written permission.
- Any rights not expressly granted in these terms are reserved.