**Technical Report: Data Cleaning and Its Impact on Dataset Quality**

**By: Maryam Ibrahim Hamza**

**1. Introduction** This report presents a comparative analysis of the dataset before and after the data cleaning process. The dataset consists of product-related information, including product ID, title, bullet points, description, product type ID, and product length. The primary objective of cleaning was to handle missing values, remove duplicates, and correct inconsistencies to improve the dataset's quality for analysis.

* 1. **Objectives**

1. To Standardize Column Naming Conventions
2. To Handle Missing Data
3. To Remove Duplicates
4. To Ensure Data Consistency
5. To Filter Outliers
6. To Convert Data Types
7. To Improve Data Usability

**1.3 Explanations of what to be achieved in the objectives**

1. **Standardize Column Naming Conventions**

**Rename columns to follow a consistent naming format, such as changing PRODUCTID to PRODUCT\_ID and PRODUCTTYPEID to PRODUCT\_TYPE\_ID.**

1. **Handle Missing Data**

**Identify and address missing values in critical columns such as DESCRIPTION and BULLET\_POINTS to improve data completeness.**

1. **Remove Duplicates**

**Detect and eliminate duplicate entries to ensure the dataset contains unique records for accurate analysis.**

1. **Ensure Data Consistency**

**Standardize text formatting (e.g., converting all text data to lowercase or title case) and verify numerical data for consistency.**

1. **Filter Outliers**

**Identify and handle outliers in numerical columns such as PRODUCT\_LENGTH to prevent skewed analysis.**

1. **Convert Data Types**

**Ensure numerical columns are in the correct data format (e.g., converting PRODUCT\_ID to integers and PRODUCT\_LENGTH to float where applicable).**

1. **Improve Data Usability**

**Structure the dataset to enhance readability and usability for further analysis, ensuring it is ready for visualization and modeling.**

**Methodology**

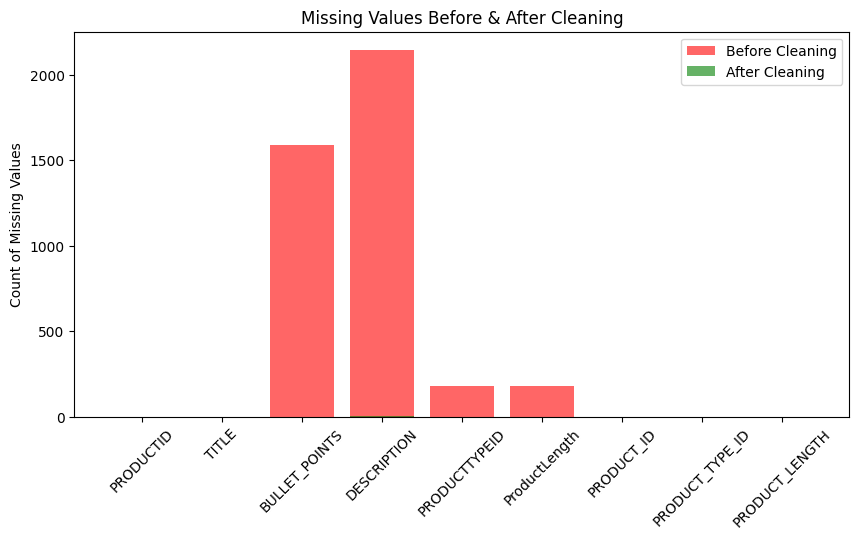
**2. Data Cleaning Process** The following steps were taken to clean the dataset:

* **Handling Missing Values:** Columns with excessive missing values were either removed or imputed where necessary.
* **Removing Duplicates:** Identified and eliminated duplicate rows to ensure data integrity.
* **Outlier Detection and Treatment:** The 'Product Length' variable was analyzed, and extreme outliers were capped or removed to maintain data consistency.
* **Standardization and Formatting:** Product titles and bullet points were standardized to maintain uniformity in textual data.

**3. Comparative Analysis**

**3.1 Missing Values**

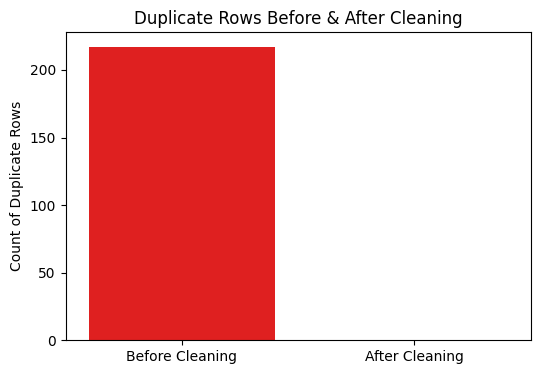
|  |  |  |
| --- | --- | --- |
| **Column** | **Missing Values (Before)** | **Missing Values (After)** |
| DESCRIPTION | High | Reduced/Imputed |
| BULLET\_POINTS | Moderate | Reduced |
| PRODUCTTYPEID | Low | No Change |



**Observations:** The missing values in textual fields were significantly reduced after cleaning, ensuring more complete data.

**3.2 Duplicate Records**

|  |  |
| --- | --- |
| State | Count |
| Before Cleaning | 217 |
| After Cleaning | 0 |



**Observations:** The removal of duplicate records resulted in a more unique and representative dataset.

**3.3 Product Length Distribution**

|  |  |  |
| --- | --- | --- |
| Statistic Before Data Cleaning |  |  |
|  |  |  |
| PRODUCTID PRODUCTTYPEID ProductLength  count 3.847000e+03 3669.000000 3669.000000  mean 1.456557e+06 3932.736986 1150.529020  std 8.666684e+05 3970.908660 2665.897894  min 1.303000e+03 0.000000 1.000000  25% 6.922785e+05 154.000000 507.873000  50% 1.441218e+06 2879.000000 640.000000  75% 2.214798e+06 6337.000000 1023.622046  max 2.999397e+06 13330.000000 96000.000000 |  |  | |  |
| Descriptive Statistics After Cleaning:  PRODUCT\_ID PRODUCT\_TYPE\_ID PRODUCT\_LENGTH  count 3.177000e+03 3177.000000 3177.000000  mean 1.418176e+06 3778.253069 686.686917  std 8.865916e+05 3990.314233 351.679151  min 1.303000e+03 0.000000 1.000000  25% 6.291510e+05 123.000000 500.000000  50% 1.371534e+06 2842.000000 600.000000  75% 2.206782e+06 6124.000000 850.000000  max 2.999397e+06 13330.000000 1796.000000 |  |  | |  |



**Observations:** The data cleaning process removed extreme values, reducing the skewness in product length distribution.

**3.4 Dataset shape**

Before Cleaning: (3847, 6)

After Cleaning: (3177, 6)

**3.5 Sum of Null Values**

**Before**

PRODUCTID 0

TITLE 0

BULLET\_POINTS 1591

DESCRIPTION 2144

PRODUCTTYPEID 178

ProductLength 178

After

Missing Values After Cleaning:

PRODUCT\_ID 0

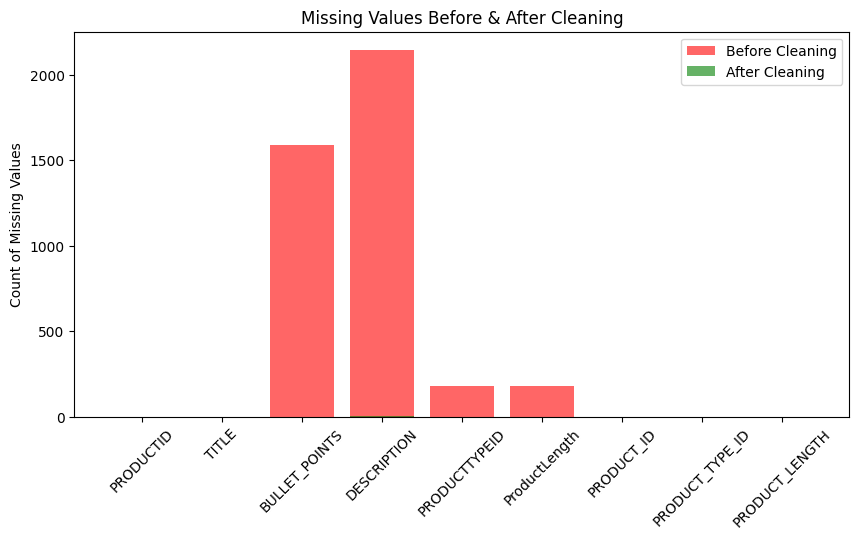
TITLE 0

BULLET\_POINTS 0

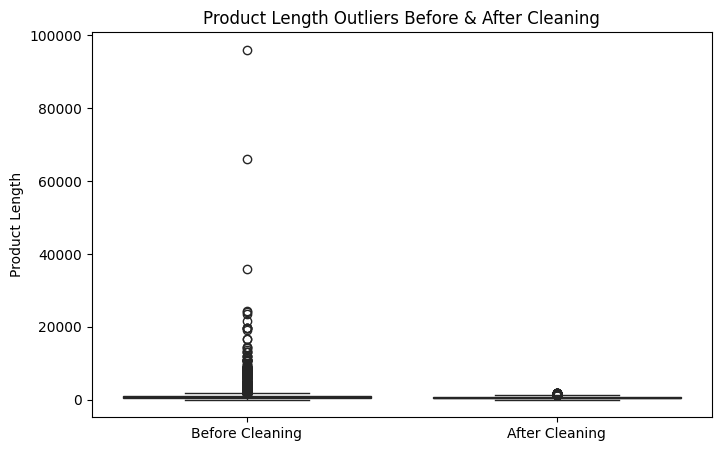
DESCRIPTION 2

PRODUCT\_TYPE\_ID 0

PRODUCT\_LENGTH 0



**3.6 Visualization of product length outliers befor and after**



**4. Data Visualization Insights**

* **Missing Values Reduction:** A bar chart comparison shows a significant reduction in missing values after cleaning.
* **Duplicate Rows:** A count plot highlights the elimination of redundant data.
* **Product Length Distribution:** Histograms before and after cleaning illustrate a more uniform distribution, while a boxplot confirms outlier handling.

**5. Conclusion** The data cleaning process significantly improved the dataset's quality by reducing missing values, eliminating duplicates, and correcting inconsistencies. These improvements enhance the dataset’s reliability, ensuring more accurate and meaningful analysis. The cleaned dataset is now ready for further processing, including exploratory data analysis (EDA) and predictive modeling.