TECHNICAL REPORT

INTRODUCTION

In today’s data-driven world, the quality of data plays an important role in decision-making and analysis. The dataset consisted of 3,847 entries, containing attributes such as PRODUCT_ID, TITLE, DESCRIPTION, BULLET_POINTS, PRODUCT_TYPE_ID, and various PRODUCT_LENGHTS. Upon initial inspection, several data quality issues were identified, including duplicate entries, missing values, inconsistent column formatting, and excessively long product titles.

The objective of this task was to clean and prepare a raw product dataset to ensure it is free from inconsistencies and ready for further marketing analysis. Additionally, a new feature, `short_title`, was created to optimize product titles for better SEO performance and readability.

NECESSITY OF DATA CLEANING AND TITLE OPTIMIZATION

Data cleaning and title optimizaton is neccessary because data inconsistencies can lead to inaccurate insights, which may affect product visibility, customer engagement, and business strategy.

The key challenges observed in the dataset include:

1. Duplicate Product Entries. The presence of multiple records with the same PRODUCT_ID can inflate product counts and distort analysis.
2. Missing Values: A significant percentage of entries lacked crucial information, particularly in DESCRIPTION and BULLET_POINTS.
3. Inconsistent Column Naming: The use of mixed-case formatting and spaces in column names reduced readability and analytical efficiency.
4. Lengthy Product Titles: Overly long titles can impact searchability and customer comprehension.

DATA CLEANING PROCESS

1. Duplicate Removal

An initial scan revealed 306 duplicate entries with identical PRODUCTID values. These duplicates were identified and removed, reducing the dataset size from ***3,847 to 3,541 unique entries***. This step ensured that each product was represented only once.

2. Handling Missing Values

A significant portion of the dataset contained missing values, particularly in BULLET_POINTS, DESCRIPTION, PRODUCTTYPEID, and ProductLength. The following strategies were implemented to address these gaps:

Column	**Missing Values (%)**	**Cleaning Approach**
BULLET_POINTS	41.4%	Replaced with "No detail" to maintain data completeness.
DESCRIPTION	55.7%	Replaced with "No description" to ensure consistency.
PRODUCTTYPEID	4.6%	Blanks were removed.
ProductLength	4.6%	Blanks were removed.

3. Column Name Standardization

To improve readability and analytical efficiency, column names were standardized using the following principles:

I. All column names were converted to uppercase for uniformity.
II. Spaces were replaced with underscores (e.g., Product Length → PRODUCT_LENGTH).
III. Congested sentences were seperated (e.g., ProductTypeId → PRODUCT_TYPE_ID).

PRODUCT TITLE OPTIMIZATION

Product titles have an important role in search engine rankings, customer engagement, and conversion rates. Many product names in the dataset were excessively long, making them difficult to scan and process. Optimizing these titles enhances product discoverability and improves marketing performance.

METHODOLOGY FOR TITLE OPTIMIZATION

A structured approach was applied to shorten and refine product titles in the dataset. TRIM function was used to remove leading and trailing spaces. PROPER function was used to ensure consistent capitalization. Redundant words like "includes," "set of," and "features." were removed using SUBSTITUTE function and titles were reduced to 30–50 characters using LEFT function. Short titles were cross-checked manually to ensure it retains vital information.

Before-and-After Examples

Original Title	Optimized Short Title
ALISHAH Women's Cotton Ankle Length Leggings Combo of 2, Plus 12 Colors_L	Women's Leggings Combo of 2, Plus 12 Colors_L
Delavala Self Adhesive Kitchen Backsplash Wallpaper, Oil Proof Aluminum Foil Kitchen Sticker (Sliver 5(Mtr))	Kitchen Wallpaper, Kitchen Sticker (Sliver 5(Mtr))
Hexwell Essential oil for Home Fragrance Oil Aroma Diffuser oil Set of 2 Rajnigandha Oil & TeaTree Oil -10ML Each	Essential, Diffuser & TeaTree Oil -10ML Each

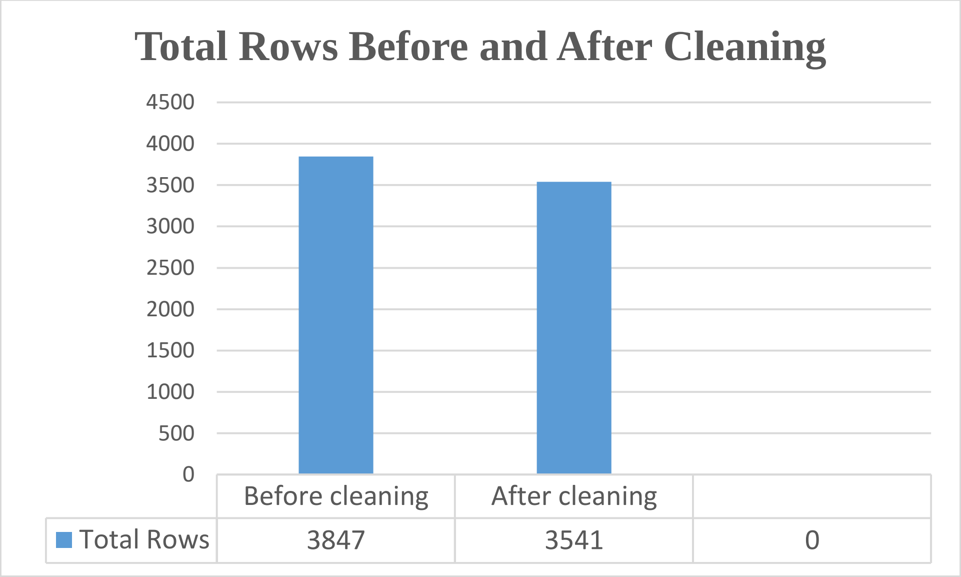
CLEAN DATASET OVERVIEW

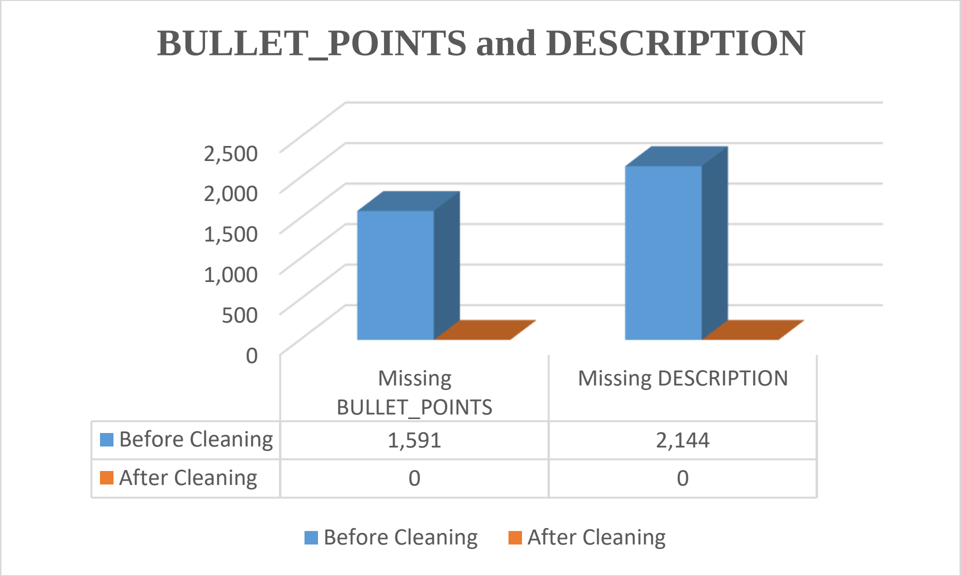
Key Metrics Before vs. After Cleaning

The table below summarizes the impact of the cleaning and optimization process.

Cleaning Step	Before Cleaning	After Cleaning
Total Entries	3,847	3,541 (Duplicates Removed)
Missing BULLET_POINTS	1,591 (41.4%)	0
Missing DESCRIPTION	2,144 (55.7%)	0
Missing PRODUCT_TYPE_ID	178 (4.6%)	0
Missing PRODUCT_LENGHT	178 (4.6%)	0

GRAPHICAL REPRESENTATION OF KEY METRICS





IMPACT OF DATA CLEANING AND OPTIMIZATION

When data are cleaned, it improves data completeness. In this dataset, missing values were addressed, ensuring that all product entries contain essential details. Alao, when column names are standardized and titles formatted, it makes the dataset more easy to understand. Shortened and structured titles improve product visibility and engagement.

CONCLUSION

The data cleaning and title optimization process significantly enhanced dataset quality, ensuring that it is ready for analysis. The following key improvements were achieved: Duplicates were removed, missing values were addressed, titles were optimized and columns were formatted. This cleaned dataset can be used to draw insights from customers, identify trend and also evaluate how optimized product titles affect search rankings.