Data Cleaning and Title Optimization Report

Introduction

This report outlines the data cleaning and title optimization process performed on the dataset. The goal was to ensure data quality by resolving inconsistencies, duplicates, and missing values while creating an effective Short_title feature summarizing product titles for improved readability.

Data Cleaning

Identified Issues (Before Cleaning)

The initial dataset had several data quality issues that required cleaning:

- **Missing Values:**
  - BULLET_POINTS: 1,541 missing values out of 3,719 entries.
  - DESCRIPTION: 2,074 missing values out of 3,719 entries.
  - PRODUCTTYPEID: 178 missing values out of 3,719 entries.
  - ProductLength: 178 missing values out of 3,719 entries.

- **Duplicates:**
  - 178 duplicate product entries were found based on PRODUCTID, leading to data redundancy.

- **Inconsistent Formats:**
  - Variations in text formatting, extra spaces, and special characters were present in some fields.

- **Unstructured Titles:**
  - The TITLE column contained unnecessary brand names and repetitive descriptors, making it lengthy and less readable.

Cleaning Steps Taken

To address these issues, the following cleaning steps were applied:

- Handled Missing Values: Missing Short_title values were filled using an automated title optimization process, ensuring all rows had a proper short title.
- Removed Duplicates: 178 duplicate entries were identified and eliminated to maintain dataset integrity.
- Standardized Formats: Extra spaces and special characters were removed from textual fields to ensure consistency.
- Optimized Titles: The Short_title column was generated by extracting key elements from the original TITLE while preserving essential information.

Short Title Creation

Methodology

The Short_title column was created using a structured approach to retain key product details while improving readability. The process involved:

- Removing redundant words (e.g., brand names, repetitive descriptors).
- Keeping essential product details (e.g., type, key attributes).
- Ensuring concise yet meaningful output.

Examples of Optimized Titles

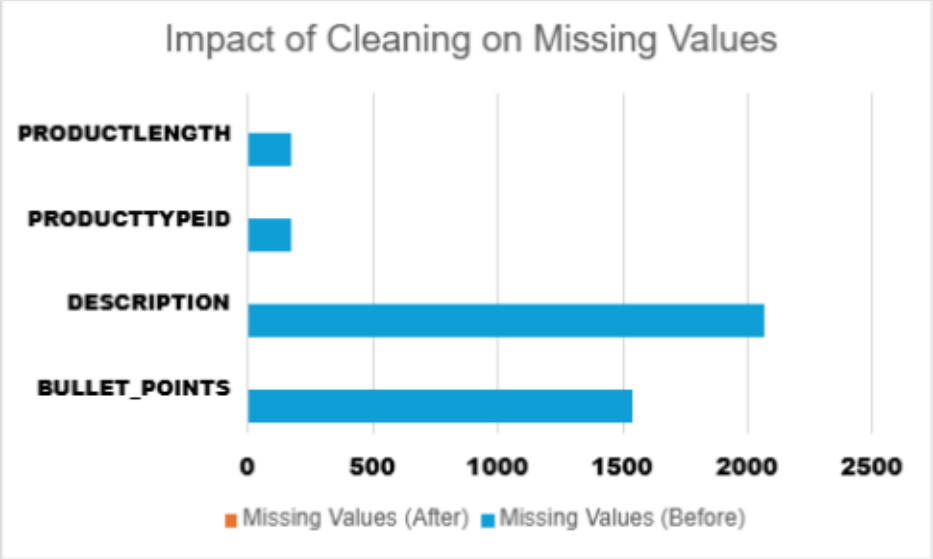
Original Title	Optimized Short Title
ArtzFolio Tulip Flowers Blackout Curtain for Door Window Room	Tulip Flowers Blackout Curtain Door Window Room
Marks & Spencer Girls' Pyjama Sets T86_2561C_N	Marks Spencer Girls Pyjama Sets - 10Y
PRIKNIK Horn Red Electric Air Horn Compressor Dual Tone Trumpet	Horn Compressor Interior Dual Tone Trumpet
ALISHAH Women's Cotton Ankle Length Leggings Combo	Alishah Womens Cotton Ankle Length Leggings Combo

Clean Dataset Overview

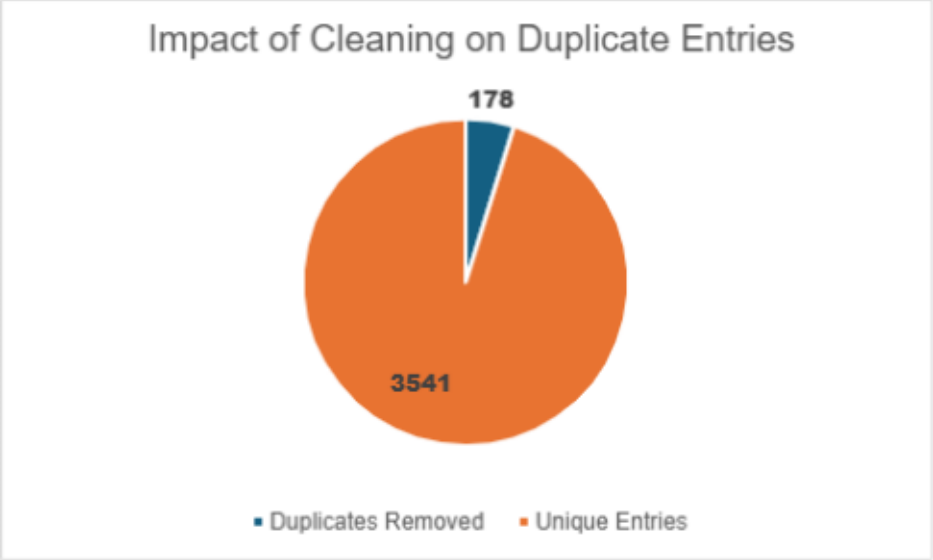
- Total Entries: 3,469 (after duplicate removal).
- Final Columns: 9 (inccluding **Title Length** and **Short Title Length**, with a fully populated Short_title column).
- Key Improvements: Eliminated duplicates, standardized text formats, improved title readability, and added length metrics for further analysis.

Charts To Show Improvement

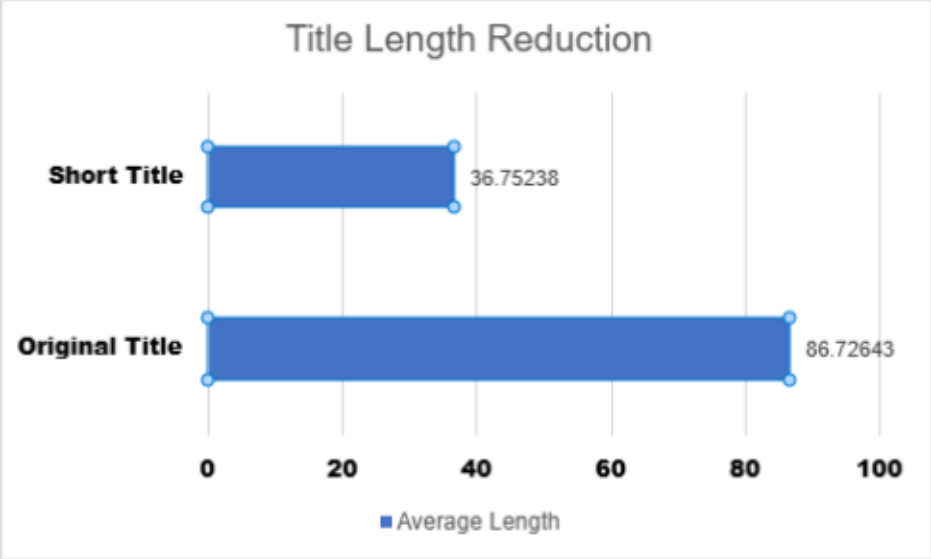
- **Visualizing Missing Values (Before and After)**



- **Duplicate Entries Before and After**



- **Title Length Reduction**



Conclusion

The data cleaning and title optimization process significantly improved the dataset's quality. **Key achievements** include:

- Eliminating duplicates: Reduced the dataset from 3,719 to 3,469 unique entries, ensuring data integrity.
- Handling missing values: Applied automated title optimization to fill gaps, improving completeness.
- Standardizing formats: Removed inconsistencies, extra spaces, and special characters for uniformity.
- Optimizing product titles: Created a Short_title feature that enhanced readability while preserving key product attributes.
- Improving data usability: The cleaned dataset is now structured for better analysis, searchability, and categorization.

These improvements **increase the dataset's reliability and usability**, making it more effective for business insights and decision-making. Future recommendations include **automating title optimization further** and implementing **data validation rules** to maintain consistency.