

Report Title

Data Optimization and cleaning

Name: Ochiobi Precious

# Table of Content

1. Introduction.

2. Data Cleaning.

3. Short Title Creation.

4. Clean Dataset

Project Overview.

During the HNG Internship, I conducted a data cleaning and title optimization process on a structured dataset to enhance its quality and usability. The dataset consists of 3,847 records, requiring cleaning to address inconsistencies, missing values, and formatting issues. Additionally, title optimization was performed to improve readability and relevance. This report outlines the key cleaning steps taken, the methodology for title optimization, and the overall improvements made to the dataset, ensuring better data integrity for analysis and decision-making.

**1.0 INTRODUCTION**

The Product Dataset comprises product-related information from various sources. It contains 3,847 records across multiple product categories. The dataset consists of six key fields**,** including:

∙ PRODUCTID – Unique identifier for each product

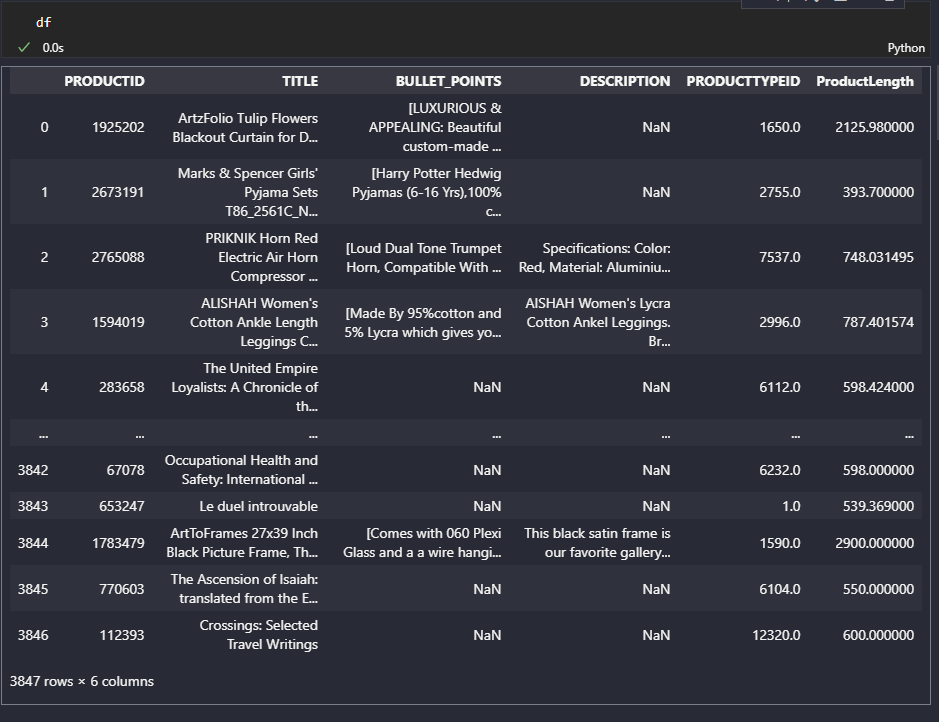
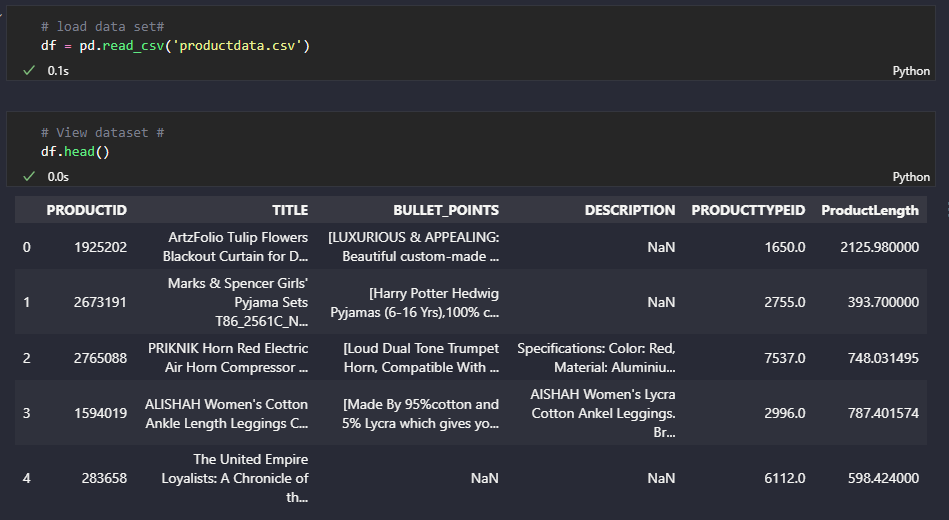
∙ TITLE – The name of the product

∙ BULLET\_POINTS – Key highlights or features of the product

∙ DESCRIPTION – A detailed explanation of the product

∙ PRODUCTTYPEID – Numeric identifier for product categories

∙ Product Length – A numerical measurement related to the product

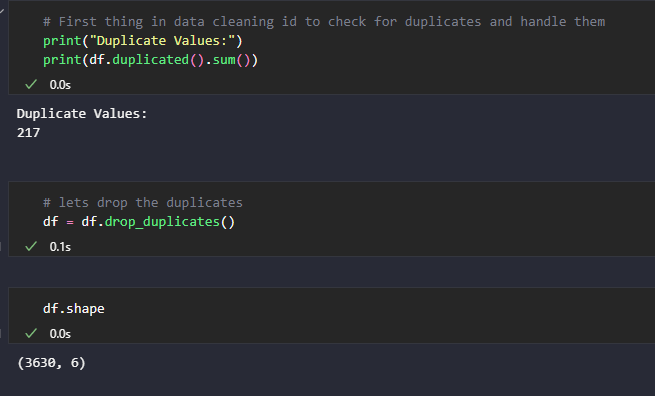
The primary goal of this task was to clean and optimize the dataset to enhance its quality and usability.  Specifically, the objectives included data cleaning and title optimization.

**2.0 DATA CLEANING**

The data cleaning and initial exploration phase was carried out using Python. Several data issues were identified and addressed as follows:

**Duplicate Values**

After checking the dataset, it was discovered that there were 217 duplicate records in the dataset.



The duplicates were dropped changing the number of records from 3,847 to 3,630.

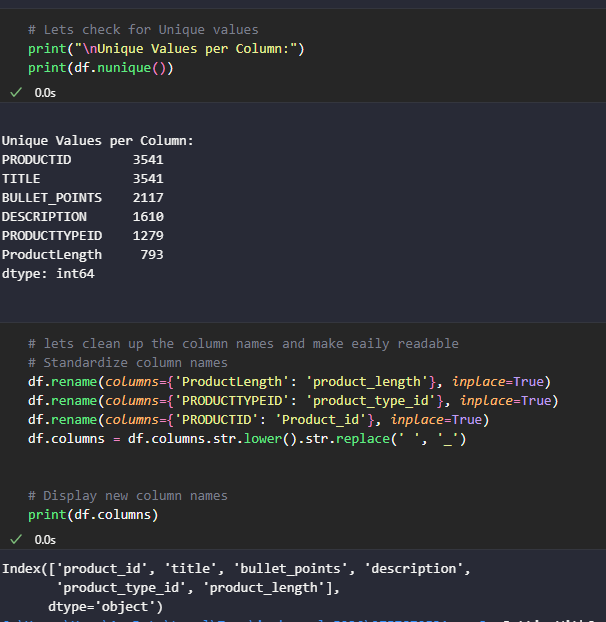
**Missing Values**

The missing values in the "Bullet Points" column were replaced with the placeholder "No details available" For the "Description" columns, missing values were filled with "No description available"  In the "Product Type ID" and "Product Length" columns, missing values were substituted with  "N/A."



∙ **Standardization**

* The "ProductLength" column was renamed to "PRODUCT\_LENGTH" to standardize the naming as spaces were replaced with underscores to ensure consistency with the other columns.
* After which converted text to lowercase for consistency



**Data Accuracy and Verification**

The describe() function in Pandas revealed an outlier in the "Product Length" column, as the  maximum value significantly exceeded the 75th percentile. However, given the context that the  "Product Length" represents the physical length of the product, this value was considered logical  and was not removed. Additionally, it was verified that no negative values were present in the  "Product Length" column.

**3.0 SHORT TITLE CREATION**

As part of the data preparation process, one of the key tasks was to optimize product titles by  making them shorter and more structured. The goal was to enhance readability, standardize  naming conventions, and ensure that each title conveyed essential product details concisely. To  achieve this, I applied a systematic approach to shortening titles while preserving their meaning  and key attributes.

1. **Initial Cleaning of Titles:** To begin the title optimization process, I implemented a function to  clean product titles by removing unnecessary elements while preserving essential details. The  function follows these key steps:

**-** If a title is missing (NaN), it is replaced with "No Title Available".

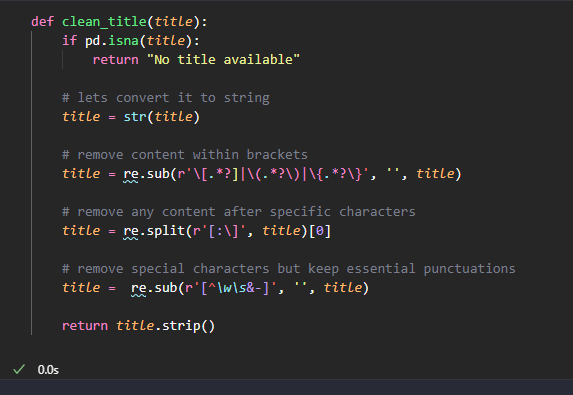
**-** Ensures all titles are treated as strings for consistency.

**-** Any information enclosed in square brackets [], parentheses (), or curly braces {} is  removed to keep the title concise.

**-** Titles are shortened by discarding text appearing after specific characters like colons (:)  and slashes (/), which often introduce unnecessary details.

**-** Unwanted special characters are removed, while essential symbols such as hyphens (-)  and ampersands (&) are retained for readability.

**-** The cleaned title is stripped of extra spaces to ensure a neat and uniform structure.



2. **Creating the short title column:** To further optimize product titles, I implemented a function  to generate concise and meaningful short titles. This function ensures that titles remain  informative, structured, and easy to analyze, while removing redundant words. The function  follows these key steps:

**-** The function first applies the clean\_title function to remove unnecessary characters and  standardize formatting.

**-** If the cleaned title is already 50 characters or shorter, it is returned unchanged. **-** Common stop words and stop words related to product naming are removed to keep the  title concise.

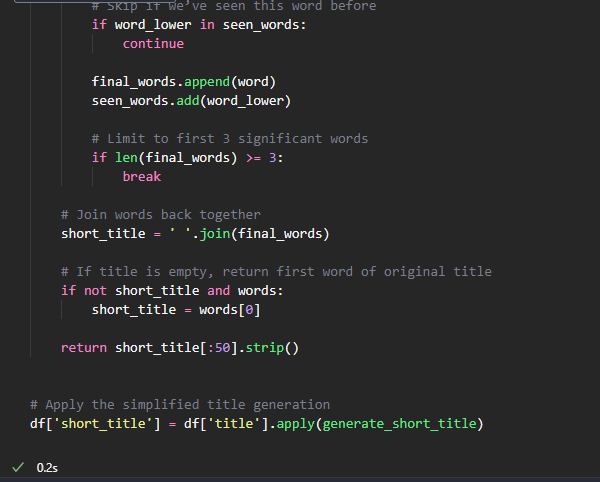
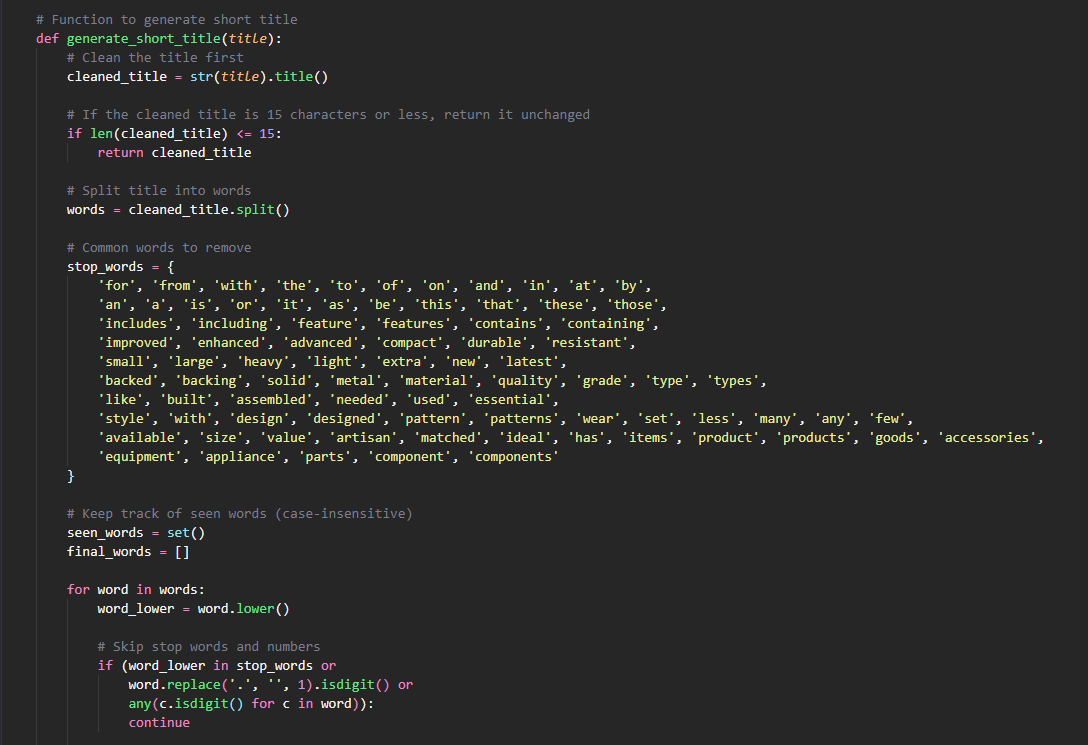
**-** The function ensures that no word is repeated and that numeric values are excluded  unless essential.

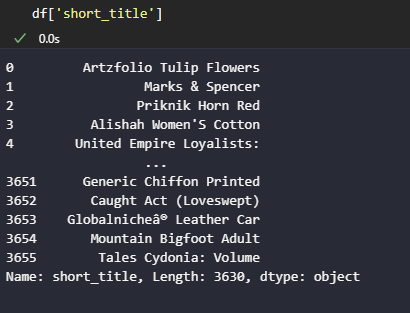
**-** The first five significant words are retained to maintain clarity.

**-** If the filtering process removes too much information, at least the first word of the  original title is kept.

**-** If necessary, the final title is truncated to ensure that it’s concise.

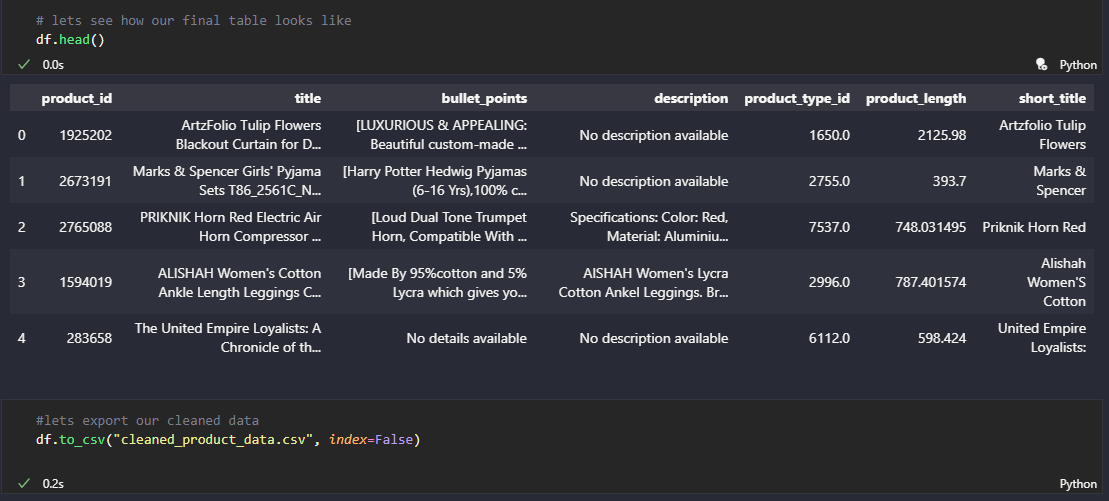
By implementing this method, product titles in the dataset became shorter, more structured, and  easier to analyse, improving overall readability and consistency





**4.0 CLEAN DATASET OVERVIEW**

With the completion of data cleaning and title optimization, the dataset is now more concise, structured, and easier to interpret**.** The main goal was to reduce inconsistencies, eliminate duplicates, fill in missing values and ensure product titles retained only the most relevant information.



The dataset is now well-organized, readable and optimized for  further analysis.