TITLE OPTIMIZATION AND DATA CLEANING REPORT

INTRODUCTION

This task involves preparing product data for further marketing analysis by addressing data quality issues and creating a new feature for SEO-optimized, concise product titles. This report outlines the steps taken to clean the dataset, resolve data quality issues, and generate a short title feature for improved SEO and readability.

OBJECTIVES

- Ensure the dataset is clean, reliable, and ready for analysis.
- Create a concise and SEO-friendly short_title for each product.
- Resolve common data issues, including missing values, duplicates, and inconsistencies.

DATA EXPLORATION AND CLEANING

Key Variables in the Dataset: The dataset contains 3,847 rows and 6 columns:

- title (Long Title)
- product_type_id
- product_length
- bullet_point
- description

Missing Values: A scan using `df.isnull().sum()` revealed:

- bullet_point: 1,452 missing
- description: 1,985 missing
- product_type_id and product_length: 89 missing each

To handle missing values:

- bullet_point and description were filled with "NA" (Not Available) to avoid null-related errors.
- To address the missing values in product_type_id and product_length (about 2.45% of the dataset), I used boxplots to visually inspect the data for distribution and outliers. Since both columns were right-skewed with several outliers, I chose to fill the missing values using the median instead of the mean. The median is less sensitive to extreme values and helps maintain data integrity.

Below are the boxplots used for this analysis:

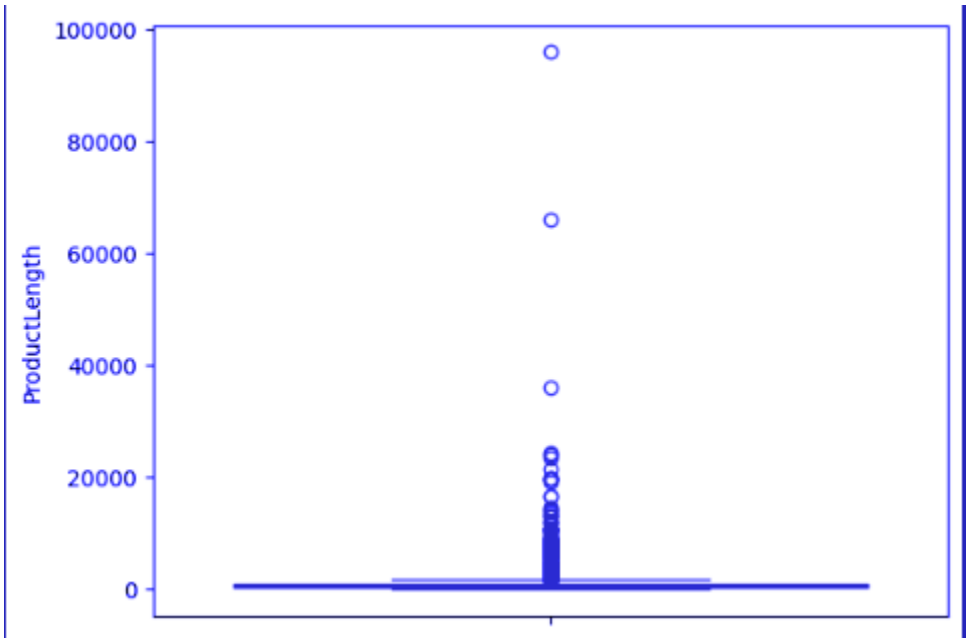


FIG1: PRODUCT LENGTH

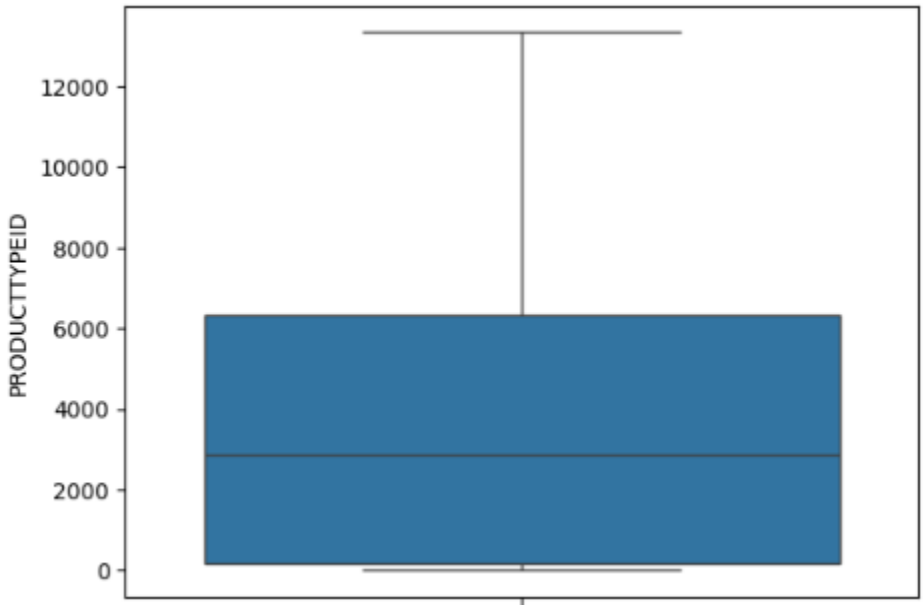


FIG 2: PRODUCT TYPE ID

Removing Duplicates: 217 duplicate rows were found and removed to maintain data quality and integrity.

Standardizing Column Names: Column names were renamed for clarity and consistency. For example:

- PRODUCTIDTYPE → product_id_type
- ProductLength → product_length

Creating the Short Title Feature

Objective: The short_title feature was created to provide concise, SEO-friendly product titles that retain important product details within a 50-character limit.

Methodology (Using Basic NLP)

To make the product titles more readable and suitable for search engines, I created a new column called short_title. This was done using some basic natural language processing (NLP) steps:

- Cleaned the text by removing symbols and unnecessary words like "Set", "Mens", and "Pack".
- Took the first few important words from the original title to keep it short and meaningful.
- Used regular expressions to pull out useful details like quantity (e.g., "3 PCS"), size (e.g., "2-3Y"), and color (e.g., "Red", "Navy").
- Added these details to the short title if they fit within the 50-character limit.
- Made sure all titles looked neat by fixing extra spaces and making the text Title Case (capitalizing each word).

This approach helped create short, clear, and helpful titles for all products in the dataset without needing manual editing.

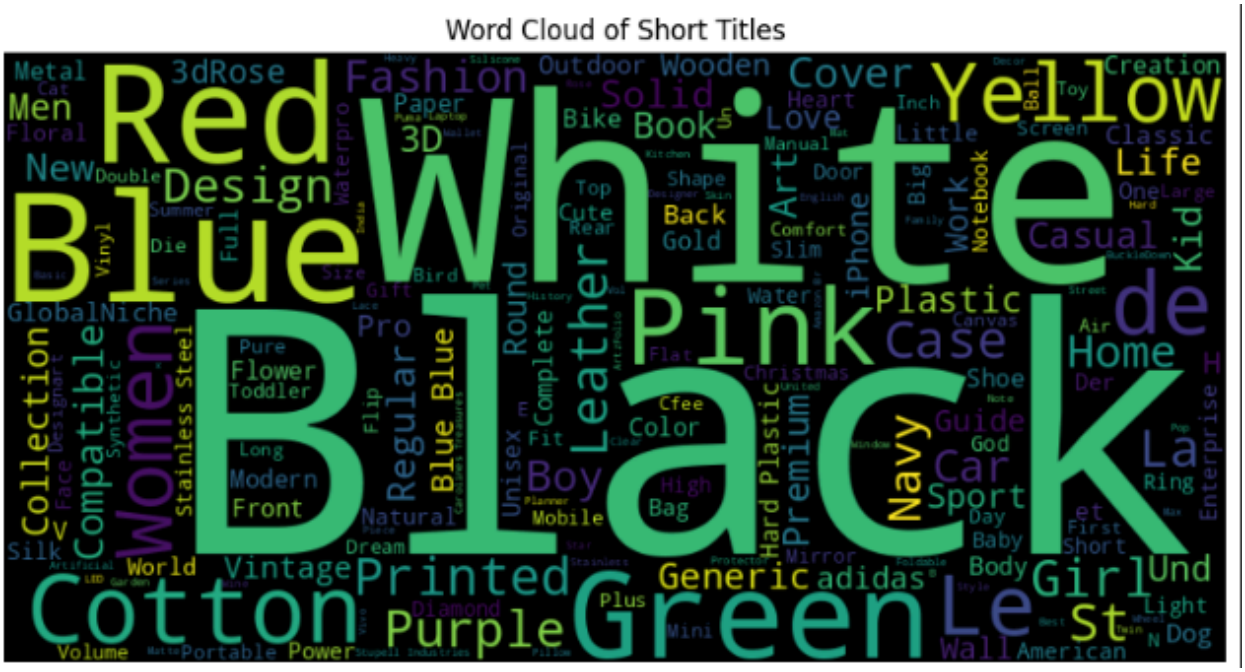
- The formula extracts the first 5 words from the product title, ensuring essential product information is retained.
- The result is then adjusted to ensure the length is between 30 and 50 characters, optimizing the title for search engines (SEO).

Example:

- **Original Title:** "Artzfolio Tulip Flowers Blackout Curtain for Door, Window & Room | Eyelets & Tie Back | Canvas Fabric | Set of 2 PCS"
- **Short Title:** " Artzfolio Tulip Flowers Blackout - 2 "

Visualizing Short Titles (Word Cloud)

To understand the common terms used in the optimized short titles, a word cloud was generated. This visualization highlights the most frequently appearing words, giving a quick overview of prominent keywords in the dataset.



This visual representation helps confirm that key descriptive elements especially colors are consistently reflected in the short titles, aiding product clarity and potential discoverability.

Cleaned Dataset Overview

The dataset was cleaned and optimized by resolving missing values, removing duplicates, standardizing column names, and adding the short_title feature.

Key Statistics:

Rows Before Cleaning: 3847

Rows After Cleaning: 3542

Missing Values Before Cleaning: 4091

Missing Values After Cleaning: 0

CONCLUSION

This project successfully cleaned and optimized a raw e-commerce product dataset by resolving missing values, removing duplicates, and generating an SEO-friendly short title using simple NLP. The short_title feature improves clarity and discoverability, while the word cloud provides a useful summary of keyword consistency.

Next steps could include clustering similar products, implementing TF-IDF for deeper text analysis, or using machine learning to predict missing values or recommend product categories.