Explanatory Data Analysis Report

Contents

1. Introduction

The dataset under analysis is a marketing dataset originally provided 3847 rows of input data with six columns:

(*productid, title, bullet_points, description, producttypeid, productlength*)

The research objective concentrated on converting unprocessed data into an enhanced standardized format that would serve useful analytical purposes for marketing and Search Engine Optimization (SEO) tasks. A new feature *short_title* was created using Natural Language Processing (NLP) techniques to generate efficient product titles optimized for search engine optimization.

This report includes an explanation of data transformation goals together with the procedural details of cleaning activities which justifies the data integrity requirements for future applications.

2. Objectives

The main objectives are listed below:

1. Data Quality Improvement:
    - Identify and address missing values, duplicates, and inconsistent formatting.
    - Standardize column names for clarity and consistency.
    - Validate the accuracy of numerical columns by checking for anomalies such as extreme values or unexpected distributions.
2. Feature Engineering – Creating *short_title*:
    - Develop a concise and SEO-friendly product title by primarily processing the "*title"* column.
    - Supplement the title with key phrases extracted from *bullet_points* and *description* when they provide additional valuable information.
    - Ensure that the final *short_title* meets a target character limit (between 30–50 characters) without sacrificing essential product details.
3. Advanced NLP Techniques:
    - Implement more sophisticated NLP methods using spaCy to extract noun phrases and named entities from text.
    - Utilize these extracted key phrases to create a more informative and succinct short title.

3. Data Cleaning

3.1. Initial Exploration and Identified Issues

During the Exploratory Data Analysis (EDA) phase, the following issues were identified in the dataset:

i. Missing Values:
    - *productid* and *title*: No missing values were found, indicating these fields are complete.
    - *bullet_points*: Approximately 41.36% of the rows had missing values.

- *description*: Around 55.73% of the rows were missing data.
- *producttypeid* and *ProductLength*: Each had approximately 4.63% missing values.

ii. Numerical Anomalies: The descriptive statistics for the *Productlength* column revealed a right-skewed distribution. With a median of 640, a mean of 1,126.91, and an extreme maximum value of 96,000, it was evident that there were significant outliers which might indicate data entry errors or inconsistencies in measurement units.

iii. Column Name Inconsistency: While most column names were in uppercase (e.g., PRODUCTID, TITLE), one column (*Productlength*) used mixed-case formatting. Standardizing column names was necessary for consistency and ease of use.

iv. 217 duplicated rows were identified and subsequently removed.

3.2. Data Cleaning Steps

Based on the initial EDA, the following data cleaning steps were implemented:

i. Standardizing Column Names

The column names were standardized to follow a consistent naming convention (lowercase with underscores) for clarity and ease of manipulation. **A Pandas rename function was used to effect these changes.**

ii. Handling Missing Values

- For Numerical Columns (*product_type_id* and *product_length*): Missing values were evaluated using descriptive statistics. Given that these columns had relatively low missingness (~4.63%), imputation was performed using robust measures such as the median. **This choice was made due to the presence of outliers which would have skewed the mean**.

- For Textual Columns (*bullet_points* and *description*):
  Hypothesis:
  To determine the value of including *bullet_points* and description for generating a concise *short_title*, we hypothesized that the usefulness of these fields depends on how much unique information they provide compared to the title column. Specifically, we set out to measure the overlap of key entities between *title* and the additional textual fields. If more than 80% of the entities were shared, it would imply that these fields add little new information. However, our analysis revealed that both *bullet_points* and *description* often contain extra keywords or details that are not present in the *title*—suggesting that, when informative, they can enrich the final *short_title*.

  Missing Values Challenge:
  A significant proportion of entries in *bullet_points* (approximately 41.36%) and description (around 55.73%) were missing. To handle this, we employed the following strategy:

  Placeholder Assignment:
  Missing values in these columns were replaced with clear placeholder texts:

bullet_points: "No bullet points available"
description: "No description available"

During the generation of the *short_*title feature, we implemented logic to ignore any row where the textual field contained only the placeholder text. This ensured that only rows with informative content from *bullet_points* and *description* columns contributed additional keywords to the final *short_title*.

Outcome:
This approach allowed us to preserve the integrity of the dataset by retaining all rows while effectively filtering out non-informative entries during feature extraction. The strategy not only addressed the high missing rate but also ensured that our advanced NLP techniques leveraged only meaningful content—resulting in more robust and SEO-friendly *short_title* feature generation.

iii. Validating Numerical Data Accuracy

Descriptive Statistics:

The *product_length* column was examined in detail. The significant disparity between the median and the maximum value, along with a high standard deviation, suggested the presence of outliers.

4. **Feature Engineering – Generating the *short_title***

To create a concise and SEO-friendly *short_title*, the following advanced steps were implemented:

Primary Source – "*title*" column to extract key product information.

Supplementary Sources – "*bullet_points"* and *"description* "

Using advanced NLP techniques via spaCy, key phrases (noun chunks and named entities) were extracted from all three text fields. However, if *bullet_points* or *description* contained only the placeholder text, they were omitted from the extraction process.

Combining and Cleaning Key Phrases: The extracted key phrases were deduplicated and combined. Regular expressions were applied to remove redundant or extraneous words (such as "set of," "includes," and "features").

Enforcing Character Limits: The combined key phrases were then trimmed to fit within a target character limit (30–50 characters).

Advanced NLP Implementation: The use of spaCy allowed for a more nuanced understanding of the text, enabling the extraction of meaningful phrases that contribute to the SEO value and clarity of the *short_title*.

5. **Clean Dataset Overview:**

   Highlight key statistics and improvements in the dataset.

These statistics reveal a pronounced right-skew with a very high maximum value relative to the median and the interquartile range. This suggests that there are extreme outliers that may be affecting downstream analysis as shown in the fig below

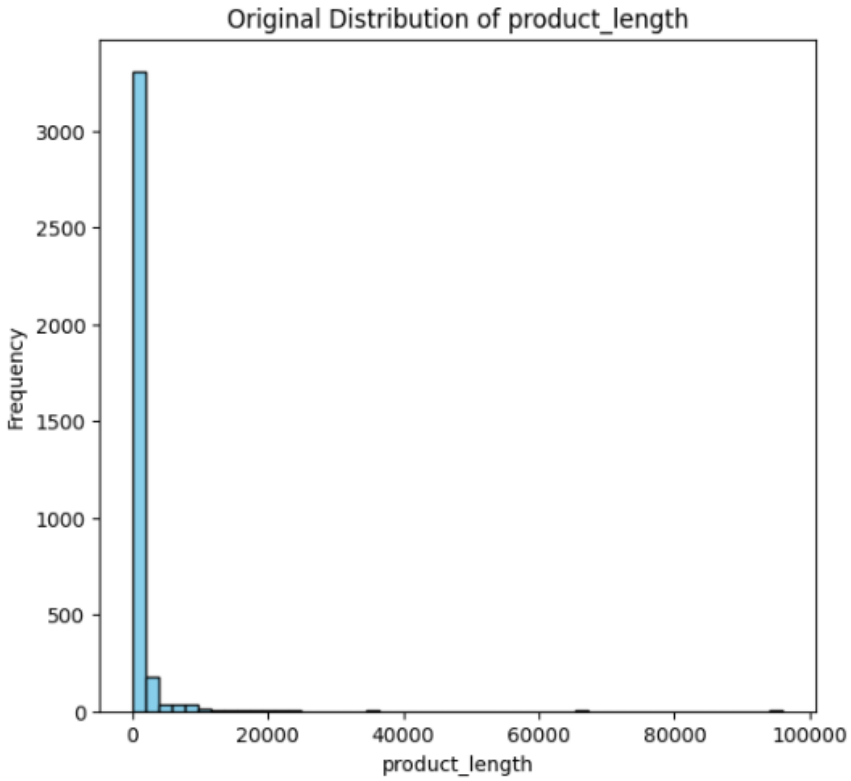

Figure 1: Distibution of the product_length feature

Suggested Improvement:

Transformation: Applying a logarithmic transformation to *product_length* may help normalize the distribution. This transformation is useful when modelling or performing statistical tests that assume a normal distribution.

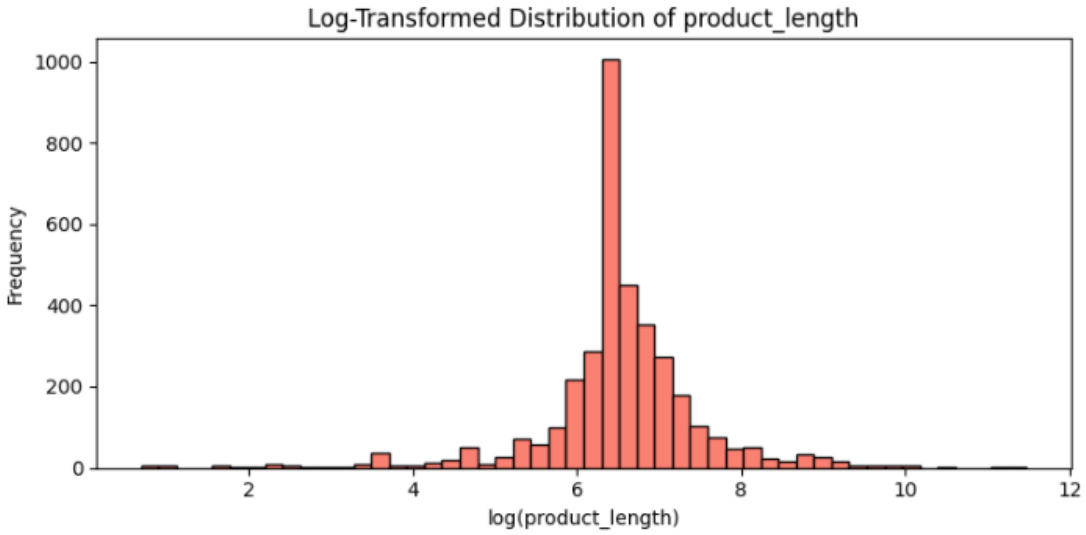

Figure 2: Log Transform of the product_length column

Impact of the Newly Created Feature: *short_title*

i. **SEO Optimization**: The *short_title* feature produces shortened product titles for search engine optimization (SEO) that maintain essential content between 30 to 50 characters. The *short_title* function both preserves essential keywords together with key product information and cuts away unneeded details.

ii. **Improved Readability and Marketing Impact**: Customers respond well to the shorter titles since they work well for marketing materials as well as online advertising content and product listings that need limited space.

iii. **Enhanced Data Utility**: The *short_title* integrates important phrases from the title with *bullet_points* when supplemented by description information to create a compressed data string. The complete dataset becomes more interpretable for future predictive tasks when grouping products or generating recommendations or identifying sentiments.

Summary of improvements, with examples of original and short titles.

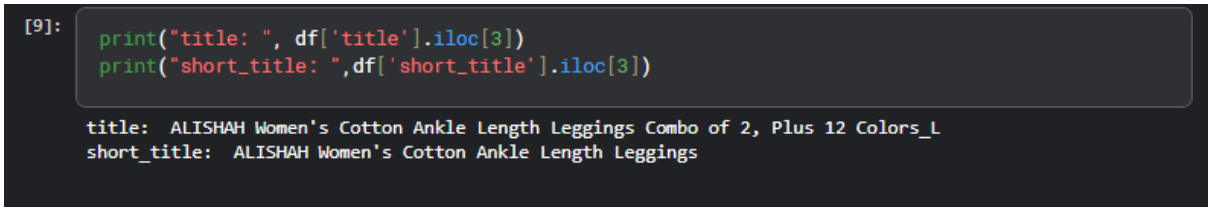

Figure 3: result of the new feature

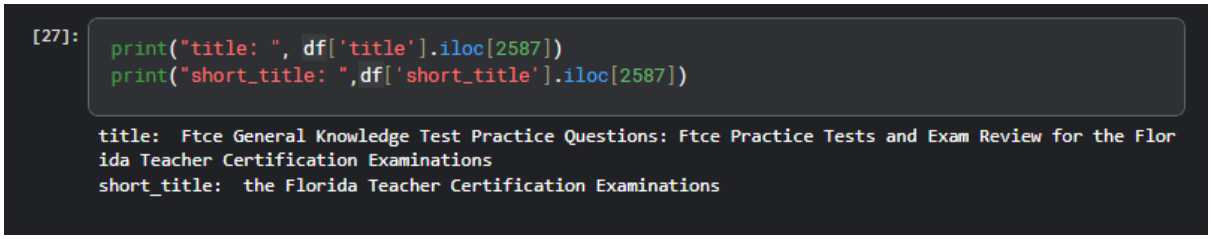

Figure 4: Result of the new feature

A word cloud of the *short_title* feature can quickly illustrate which keywords are most prevalent after the transformation.

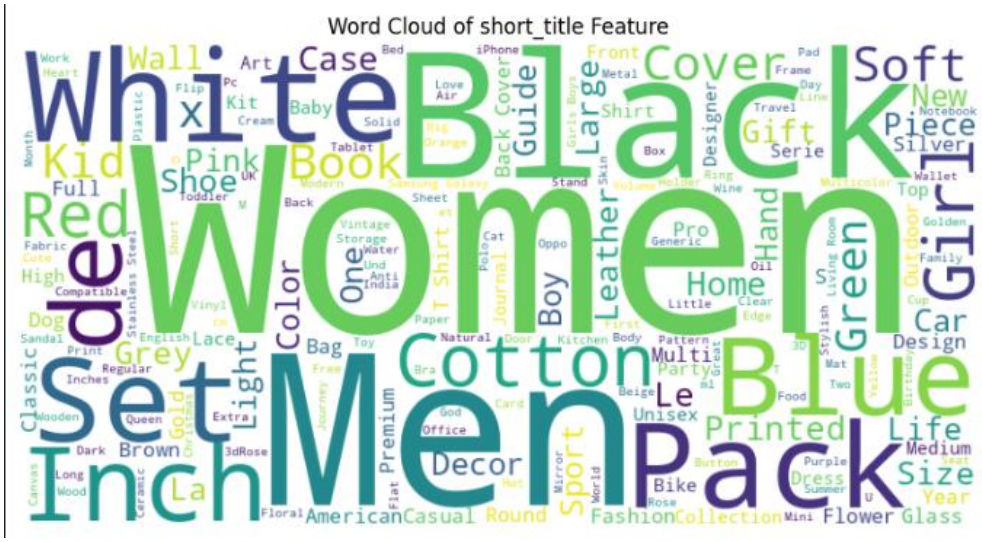

Figure 5: Word Cloud Visualization

Displayed below is the bar chart showing the frequency of top keywords in *short_title* can provide additional insights into the dominant features captured by this new column.

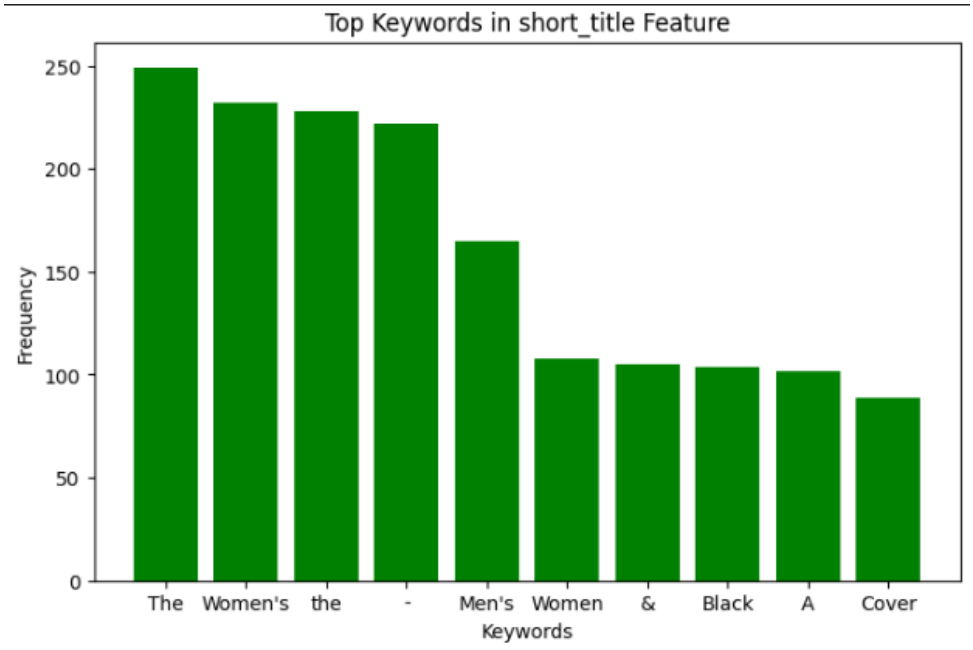

Figure 6: Frequency Analysis of Key Terms

Overall Impact on the Dataset

i. The data cleaning procedure established better data quality by making the dataset more consistent. Standardized columns together with managed missing data values and cleaned *product_length* numbers improve overall data quality.

ii. *Short_titles* benefit marketing through enhanced actionable insights because they are now generated robustly with advanced NLP methods. The short descriptive text created by marketers leads to better search engine visibility combined with increased user reaction.

iii. After appropriate transformations applied to *product_length* and the addition of *short_title* as a textual feature the dataset now provides better conditions for predictive

modeling as well as clustering analysis and other forms of exploratory and explanatory evaluation.

6. Conclusion

In summary, the data cleaning and feature engineering process has significantly improved the dataset's quality and utility. Standardized column names and consistent formatting now ensure clarity and ease of use. Missing values in *bullet_points* and *description* were addressed by imputing clear placeholders and filtering out non-informative entries during feature extraction. This allowed us to generate the new *short_title* feature using advanced NLP techniques, which extracts key phrases from the *title*—and, when valuable, supplements it with details from the other text fields—resulting in concise, SEO-friendly product titles.

Additionally, the *product_length* column was scrutinized; its right-skewed distribution and extreme outliers suggest further improvements, such as outlier capping or logarithmic transformation, could enhance analysis robustness.

Overall, these enhancements have created a consistent, enriched, and analysis-ready dataset that not only improves marketing effectiveness through actionable insights but also provides a strong foundation for further statistical modelling and advanced analyses.