Product Data Analysis: Data Cleaning and Optimization Report

This report contains a documentation of the cleaning, preparation and optimization of a product data. The dataset includes product information such as titles, descriptions, and attributes. This task aims to ensure the dataset is clean, reliable, and ready for further marketing analysis. Additionally, a new column called 'short_title' was added to deliver SEO-optimized and concise product titles.

Data Cleaning/Preparation

- The column names were standardized for consistency (e.g., PRODUCTID was renamed to product_id).
- Duplicate entries were removed using Power Query to ensure data accuracy and avoid redundancy.
- Descriptive statistics were calculated for product_length to determine the following metrics:

Metric	Value
Average length	1155.41
Median length	638
Minimum length	1
Maximum length	96000
Standard deviation	2699.30

- Missing values in the following columns were handled as follows:
  description: Replaced with "No description."
  bullet_point: Replaced with "No bullet point."
- No negative or invalid values were found in essential columns such as product_length and product_type_id

Short Title Creation

The short titles were created using Power Query to generate concise and SEO-optimized titles that enhance marketing effectiveness without losing essential product information.

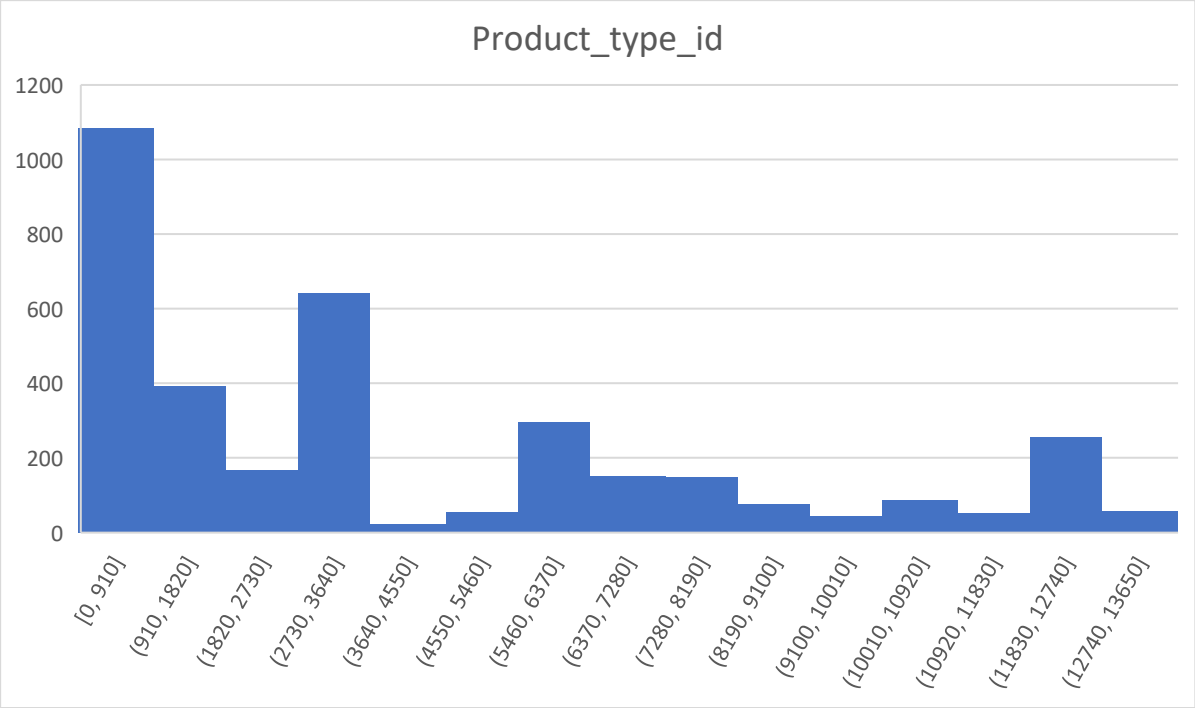
Original title	Short/Optimized title
PRIKNIK Horn Red Electric Air Horn Compressor Interior Dual Tone Trumpet Loud Compatible with SX4	PRIKNIK Horn Red Electric Air Horn Compressor
LILLUSORY Women's Turtleneck Oversized Sweaters 2022 Fall Long Batwing Sleeve Spilt Hem Tunic Pullover Sweater Knit Tops, Dark Apricot, X-Small	LILLUSORY Women's Turtleneck Oversized Sweaters 2022 Fall

Clean Dataset Overview

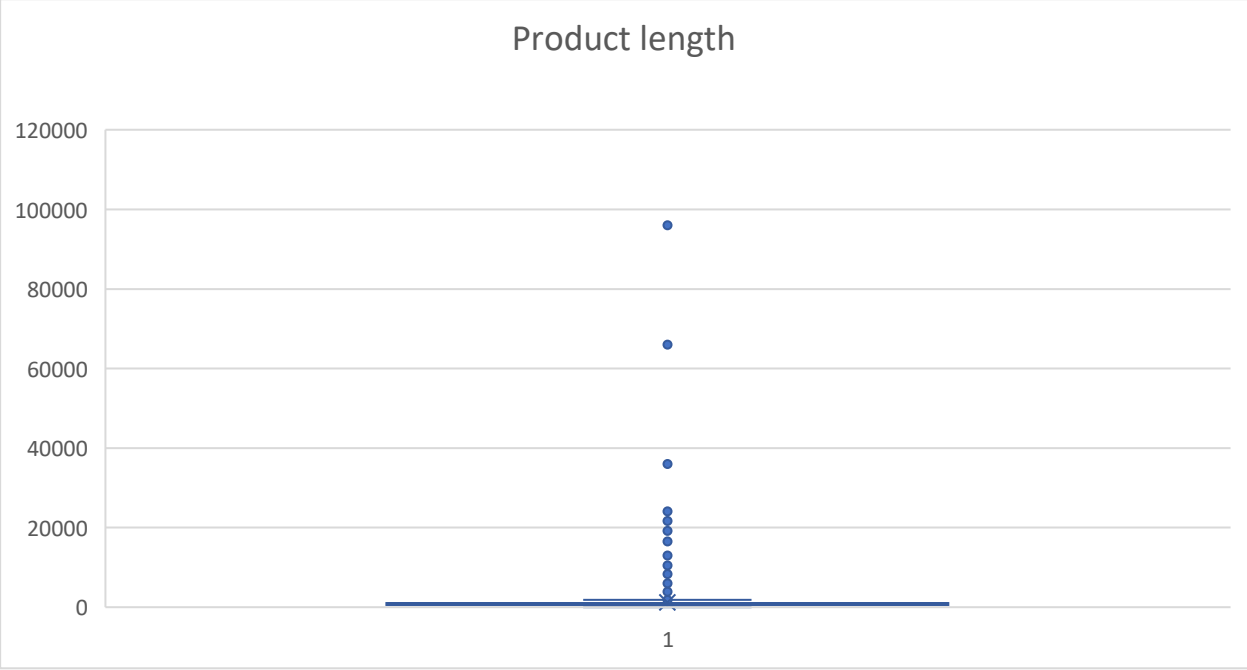
⇨ **Final Dataset Size**: 3,541 rows

- Missing values were fully addressed.
- Duplicate records were eliminated.
- Column names and data formats were standardized.
- A new column, short_title, was added, enhancing the dataset's utility for marketing purposes.

The dataset was successfully cleaned and optimized, ensuring high data quality and introducing a new feature for improved marketing impact. This process resolved all identified issues and provided a reliable foundation for further analysis.



There is a high concentration of products within the product_type_id range of **[0, 910)**. This suggests that a substantial portion of the products falls into this specific type or category.



The plot shows a few extreme values far above the whiskers (e.g., around 100,000 and 80,000) which might represent unusual entries in the dataset. Also, most of the product lengths cluster closely around the median, indicating low variability for the majority of the data.