# Technical Report on Data Cleaning and Title Optimization

**Introduction**

The dataset used for this analysis contains product information, including key columns such as product_id, title, bullet_points, description, product_type_id, and product_length. The objective was to clean the dataset by addressing issues like missing values, duplicates, outliers, and non-standardized column names, as well as to optimize the product titles for improved searchability and readability.

## Data Cleaning

**Issues Identified**

During the initial exploration of the dataset, several data quality issues were identified and resolved as follows;

- **Missing Values:** Found in bullet_points, description, product_type_id, and product_length.
- **Cleaning step:** Numerical columns (product_length and product_type_id): Missing values were filled using the mean and mode, while Categorical columns (bullet_points, description) were filled using fillna function
- **Duplicates:** 217 duplicate rows were detected,

  **Cleaning step:** duplicates were deleted
- **Outliers:** The product_length column contained extreme values, with a maximum of 96,000.
  **Cleaning step:** the Interquartile Range (IQR) method was used and 364 outliers were identified and removed to improve data accuracy.
- **Non-standardized Column Names**: Inconsistencies in column names
  **Cleaning step**: All column names were standardized for uniformity.

## Short Title Creation

**Methodology**

The objective of title optimization was to create concise, clear, and search-friendly titles. This was achieved through a combination of Excel functions (LEFT, RIGHT, SUBSTITUTE, CONCATENATE) to shorten the titles while retaining key product details.

The methodology involved:

- Truncating long descriptions.
- Retaining essential product identifiers.
- Removing unnecessary words to create a more compact, yet descriptive title.

**Examples of Title Optimization**

Here are a few examples of the title and short title:

| Title | Short_title |
|---|---|
| Marks & Spencer Girls' Pyjama Sets T86_2561C_Navy Mix_9-10Y | Marks & Spencer T86_2561C_Navy Mix_9-10Y |
| PRIKNIK Horn Red Electric Air Horn Compressor Interior Dual Tone Trumpet Loud Compatible with SX4 | PRIKNIK Horn Red with SX4 |
| The United Empire Loyalists: A Chronicle of the Great Migration | The United Empire Great Migration |

## Clean Dataset Overview

**Summary Statistics After Cleaning**

Here's an overview of key statistics after the cleaning process:

- **The final dataset:** contains **3,177 records**, with all identified issues successfully resolved
- **Missing Values**: Initially, bullet_points, description, product_type_id, and product_length had missing values.

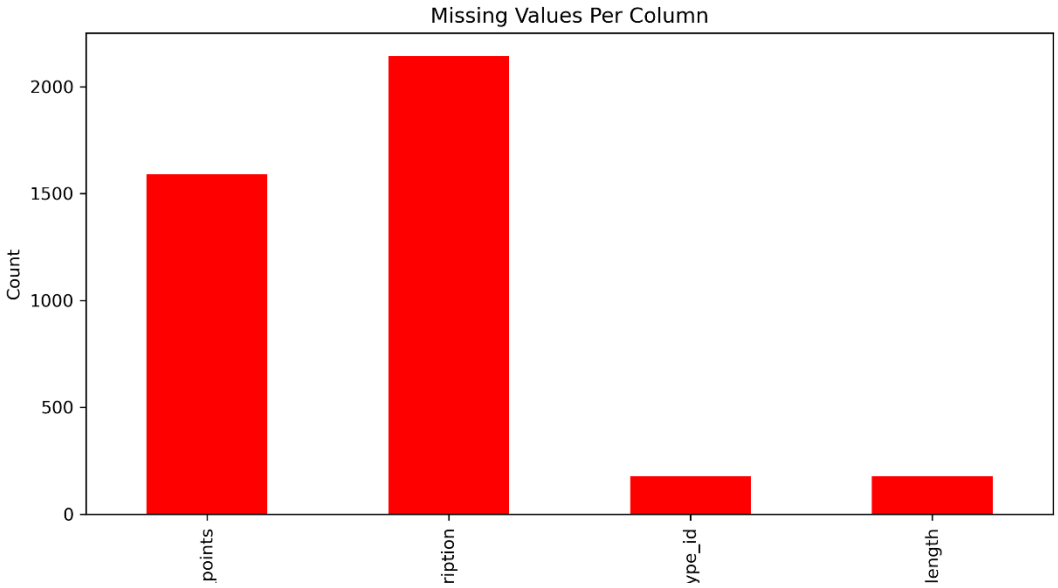
  **Now:** 0 missing values.
- **Duplicates:** 217 duplicate rows were found and removed.

  **Now:** No duplicates.
- **Outliers:** 364 extreme outliers in product_length were removed.

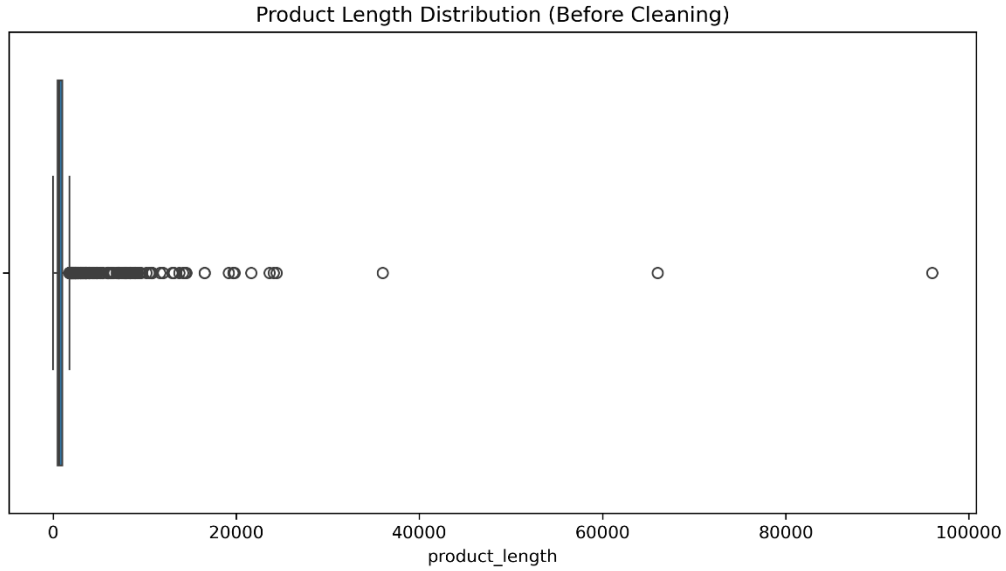
  **Now:** No extreme outliers.
- **Standardized Column Names**: All column names have been made consistent.

Below are the visualizations illustrating the impact of data cleaning and title optimization.

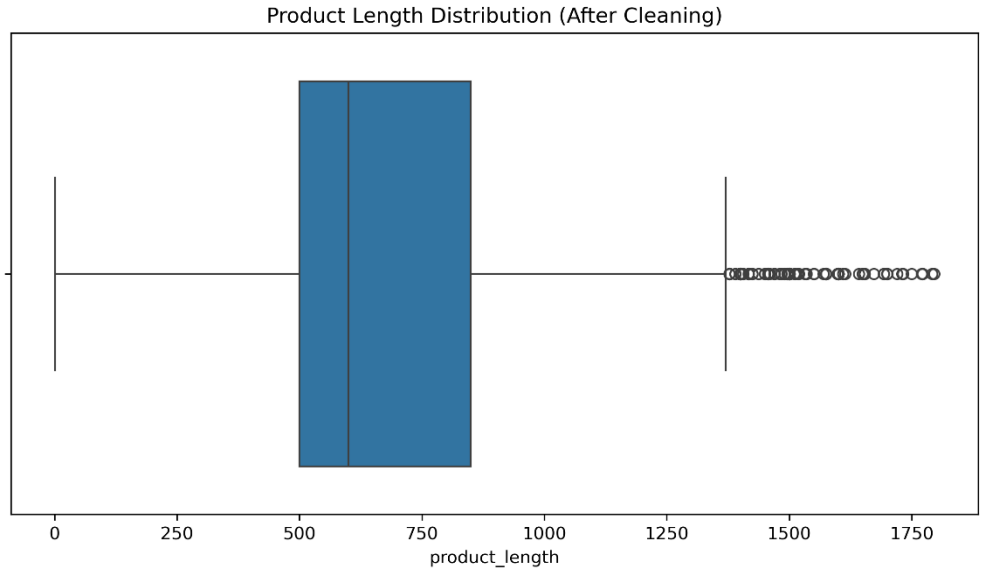
**Bar chart showing missing values in the dataset**



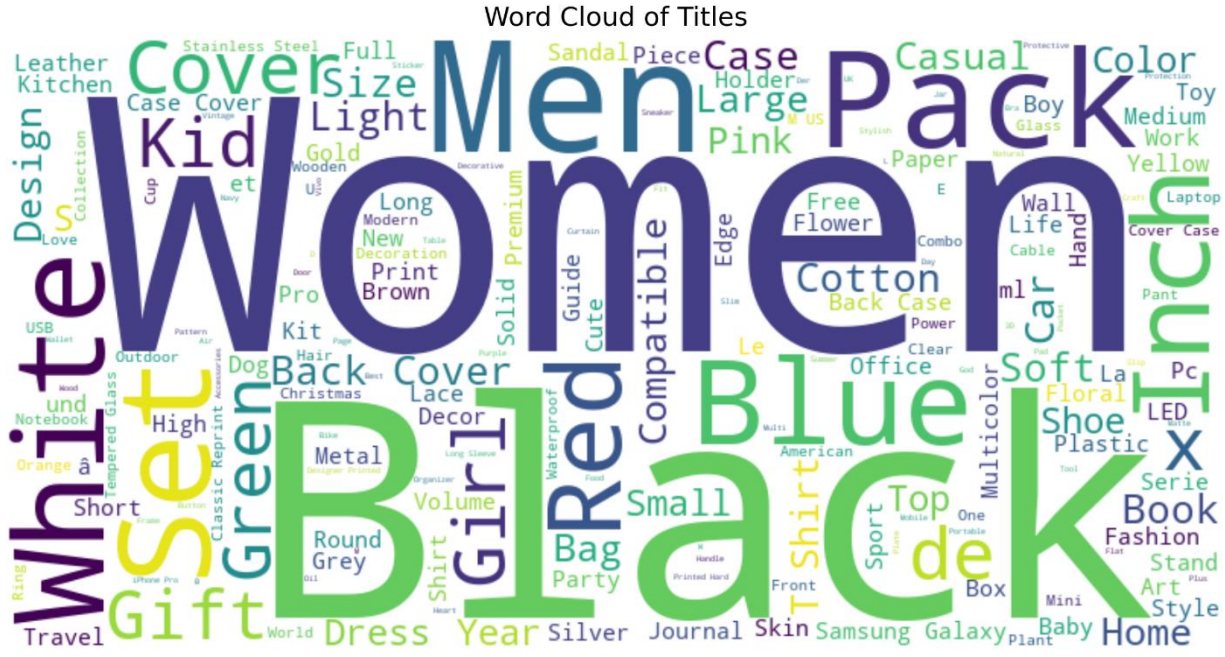

**A box plot showing outliers (before cleaning)**



**Boxplot showing the dataset after removing outliers**




**A word cloud showing title (before cleaning)**



**A bar chart showing the top ten short_title**