**Technical Report: Data Cleaning & Title Optimization**

**1. Introduction**

This report details the data cleaning and preparation process for a product marketing dataset. The primary goal was to resolve data quality issues such as missing values, duplicates, and inconsistencies while creating an SEO-optimized short\_title feature. The cleaned dataset ensures better accuracy and usability for further marketing analysis.

**2. Data Cleaning Process**

**Handling Missing Values:**

1. **PRODUCTTYPEID & ProductLength:**
   * Both columns had **4.63% missing values**.
   * Instead of dropping these rows, the missing values were filled with the median to maintain data consistency.
2. **BULLET\_POINTS:**
   * **41.36% of the values were missing**.
   * Missing values were replaced with an empty string to retain the structure and ensure the column remains usable.
3. **DESCRIPTION:**
   * Had the highest missing values (**55.73%**).
   * Since more than half the data was missing, the column was dropped to avoid unreliable analysis.

**Handling Duplicates:**

* Duplicate rows were identified and removed to prevent redundant data.

**Standardizing Column Names:**

* Columns were converted to lowercase with underscores replacing spaces for consistency and ease of access.

**3. Title Optimization (Short Title Creation)**

**Objective:**

To generate concise and SEO-friendly product titles by removing unnecessary words while retaining essential details.

**Methodology:**

1. Analyzed product titles to identify key components (e.g., brand, product type, size, color).
2. Removed redundant phrases like "set of," "includes," or unnecessary characters.
3. Limited the title to **30–50 characters** for improved readability and search impact.

**Examples:**

|  |  |
| --- | --- |
| **Original Title** | **Optimized Short Title** |
| ArtzFolio Tulip Flowers Blackout Curtain for Door, Window & Room | Eyelets & Tie Back | Canvas Fabric | Width 4.5feet (54inch) Height 5 feet (60 inch); Set of 2 PCS | ArtzFolio Tulip Flowers Blackout Curtain Door |
| Bloodlines: Race, Cross and the Christian | Bloodlines Race Cross Christian |
| Handbook of Classical and Modern Mandaic | Handbook of Classical Modern Mandaic |

**Impact of Short Titles:**

* **Enhanced SEO**: Concise titles improve search rankings and visibility.
* **Improved Readability**: Users can quickly understand product details.
* **Higher Engagement**: Short, clear titles attract more clicks and conversions.

**4. Cleaned Dataset Overview**

* **Duplicates removed**
* **Missing values handled**
* **Standardized column names**
* **Short titles added**

The final cleaned dataset is now ready for further analysis and marketing optimization.

**5. Conclusion**

Through structured data cleaning and title optimization, the dataset quality was improved, making it more reliable and effective for marketing use. The short titles enhance SEO and user engagement. This cleaned dataset will facilitate better decision-making and analytics.