**Technical Report: Product Data Cleaning**

**Introduction**

The objective of this project was to clean and optimize a product dataset to improve data quality and enhance SEO performance through title optimization. The dataset contained inconsistencies such as missing values, duplicate records, and unstructured product titles, which were addressed using Microsoft Excel.

The dataset is a product dataset on product provided by HNG, with 3,848 rows. The data has 6 columns with variables including

1. Product ID: This describes the unique identification numbers of each product in the dataset.
2. Title: This is the name of each product on the product dataset.
3. Bullet Points: This is a brief description of what you get with the product and how to use product to optimization.
4. Description: This describes the values provided by the product purchased.
5. Product Type ID: This describes the identification given to the product due to its type, e.g dresses have generic identification different to say books.
6. Product Length: This describes the length of the products on the product dataset.

**2. Data Cleaning Process**

* **Dataset Review**
* I loaded the dataset into microsoft **Excel** and conducted a preliminary to understand its structure.
* Identified key columns the key columns that needed cleaning.
* **Handling Missing Values**
* **Numerical Columns**
* Used Excel’s **AVERAGE function** to compute the mean and fill missing values.
* **Removing Duplicate Entries**
* Applied Excel’s **Remove Duplicates** feature to eliminate duplicate product entries.
* **Standardizing Data Formats**
* Standardized **column names** for clarity
* Used PROPER() function to correct text formatting inconsistencies.
* Verified that numeric values were accurate.

**3. Product Title Optimization**

* **Objective**

To improve readability and SEO performance, a new column was created to provide a concise version of each product title.

* **Steps used for Title Optimization**

1. Extracted key elements from the original title while removing redundant words.
2. Ensured that short titles retained **essential details** while staying within **30–50 characters**.
3. Used Excel’s **LEFT() and LEN() functions** to limit text length.

**4. Final Data Validation and Quality Check**

After performing the data cleaning and optimization steps, the dataset was thoroughly reviewed:

* **Checked for any remaining missing values** using Excel’s **FILTER() function**.
* **Re-ran the duplicate check** to ensure all redundant records were removed.

**Conclusion**

**The dataset has been successfully cleaned and optimized using Excel. The new title feature has improved product title readability and SEO performance. This cleaned dataset is now ready for further data analysis and visualization. Looking at the data more improvements could be made to improve the quality of data like removing the ‘product length’ and ‘product type ID’ as they both are of no relevance to the data and contained a lot of missing values which could hamper the quality of the data.**[**https://hng.tech/hire/data-analysts**](https://hng.tech/hire/data-analysts)[**https://hng.tech/internship**](https://hng.tech/internship)