Technical report By

Olamilekan Koyi

On data cleaning

Using SQL

06 – February -

1. Introduction

The dataset under review, productdata, contains product-related information, including product titles, descriptions, and bullet points, which are critical for marketing and SEO purposes. This task involved preparing the dataset for further analysis by addressing data quality issues, cleaning unnecessary characters, removing duplicate entries, and generating optimized product titles for improved SEO and readability.

---

2. Data Cleaning

2.1 Issues Identified

Upon examining the dataset, several data quality issues were found:

- **Missing Values**: There were missing values in key columns, particularly in the Bullet_Points and Description fields.
- **Redundant Data**: Duplicate entries were identified based on the combination of Bullet_Points, Product_Length, and Product_ID.
- **Unnecessary Characters**: Unwanted words like "Set of," "Includes," and "Features" appeared in the Title, Bullet_Points, and Description fields, which needed to be removed for better readability.

2.2 Cleaning Steps Taken

- **Handling Missing Values**:
    - Replaced NULL values in the Bullet_Points column with 'None' and NULL values in the Description column with 'empty' using COALESCE().
    - The SQL code used for this step:
- UPDATE productdata
- SET Bullet_Points = 'None'
- WHERE Bullet_Points IS NULL;
-
- UPDATE productdata
- SET Description = 'Empty'
- WHERE Description IS NULL;

- **Removing Unnecessary Characters**:
    - Redundant words such as "Set of," "Includes," and "Features" were removed from the Title, Bullet_Points, and Description columns using REPLACE() and TRANSLATE() functions.
    - The SQL code for cleaning unnecessary characters:

```
UPDATE productdata
SET Title = TRIM(SUBSTRING(
   REPLACE(REPLACE(REPLACE(Title, 'Set of', ''), 'Includes', ''), 'Features', ''), 1, 50));

UPDATE productdata
SET Bullet_points = TRIM(SUBSTRING(
   REPLACE(REPLACE(REPLACE(Bullet_points, 'Set of', ''), 'Includes', ''), 'Features', ''), 1, 50));

UPDATE productdata
SET Description = TRIM(SUBSTRING(
   REPLACE(REPLACE(REPLACE(Description, 'Set of', ''), 'Includes', ''), 'Features', ''), 1, 50));
```

- **Removing Duplicate Entries**:
    - Duplicate entries were identified and removed by using a ROW_NUMBER() partitioning strategy to keep only one instance of each duplicated record.
    - SQL code used:

```
WITH CTE AS (
   SELECT *,
      ROW_NUMBER() OVER (PARTITION BY Bullet_Points, Product_Length ORDER BY Product_ID) AS row_num
   FROM productdata
)
DELETE FROM productdata
WHERE Product_ID IN (SELECT Product_ID FROM CTE WHERE row_num > 1);
```

---

3. Short Title Creation

3.1 Objective
The goal was to create concise and SEO-optimized product titles, focusing on essential details and retaining key information for readability. The titles were shortened to 30-50 characters, removing unnecessary words and keeping only the most important details.

3.2 Methodology

- Redundant words and phrases, such as "Set of," "Includes," and "Features," were removed from the Title, Bullet_Points, and Description fields using the REPLACE() function.
- Titles were trimmed to ensure that they do not exceed the specified character limit (30-50 characters).
- Example:
    - **Original Title**: "Tulip Flowers Blackout Curtain for Door, Window & Room | Eyelets & Tie Back | Canvas Fabric | Set of 2 PCS"
    - **Short Title**: "Tulip Blackout Curtain - 2 PCS"
    - **Original Title**: "Marks & Spencer Girls' Pyjama Sets T86_2561C_Navy Mix_9-10Y"
    - **Short Title**: "Girls' Navy Pyjama Set - 9-10Y"

3.3 SQL Implementation for Short Title Creation
The Title and Bullet_Points columns were trimmed using REPLACE() and SUBSTRING() to generate concise product titles:

```
UPDATE productdata
SET Title = TRIM(SUBSTRING(
  REPLACE(REPLACE(REPLACE(Title, 'Set of', ''), 'Includes', ''), 'Features', ''), 1, 50));
```

---

4. Clean Dataset Overview

After the cleaning and optimization process, the dataset was significantly improved:

- **Missing Values**: All missing Bullet_Points and Description values were replaced with default placeholders ('None' and 'Empty').
- **Duplicate Entries**: Duplicate records were removed, ensuring a unique dataset.
- **Title Optimization**: The titles were cleaned and shortened, retaining only essential details for SEO and readability.

Key Improvements:

- **Before Cleaning**:
    - NULL values were present in critical columns.
    - Duplicate records were found, leading to inaccurate data representation.

    - Product titles contained unnecessary words that hindered readability and SEO performance.

- **After Cleaning**:
    - The dataset is now free of missing values and duplicates.
    - Product titles are optimized and concise, adhering to SEO best practices.

---

5. Conclusion

The data cleaning and title optimization process successfully prepared the productdata dataset for further marketing analysis. By addressing missing values, removing redundant entries, and optimizing product titles, we ensured the dataset is reliable, clean, and ready for enhanced marketing efforts.

This concludes the report on the data cleaning and title optimization process. The steps taken have ensured that the dataset is ready for further analysis, and the optimized titles will contribute to better marketing and SEO outcomes.