**Stage 1 - Data Cleaning and Preparation**  
**Task: Data Cleaning, Preparation, and Title OptimizationObjective:**  
Prepare raw marketing data for analysis by addressing data quality issues and creating a new feature, short\_title, for improved SEO and readability. The task involves resolving data quality issues like missing values and duplicates, standardizing data formats, and generating concise product titles for better marketing impact.  
**Dataset:**  
Product Data ([Dataset Link](https://docs.google.com/spreadsheets/d/1z3Io4xtC2FMif2lskLbDDcRe5OTjoi4l/edit?usp=sharing&ouid=102606023409835819338&rtpof=true&sd=true" \t "https://app.slack.com/client/T086NG3N69Y/_blank))**Task Overview**  
This task focuses on ensuring the dataset is clean, reliable, and ready for further marketing analysis while introducing a new short\_title feature for SEO-optimized and concise product titles. You will explore, identify, and resolve common data issues using tools like Python (Pandas, NumPy) or Excel. Additionally, you'll implement logic to generate shorter product titles for each entry, ensuring they retain key information while being concise and prepare a detailed technical report.Steps to Complete the Task  
**1. Dataset Familiarization**

* Download and review the dataset to understand its structure and key variables.
* Identify significant variables for cleaning and for creating the short\_title feature.
* Check column names for clarity and consistency.

**2. Initial Data Exploration**

* Conduct a high-level review of the dataset using descriptive statistics.
* Identify data quality issues such as missing values, duplicates, and inconsistent formats.
* Use Python (Pandas, NumPy) or Excel for initial exploration.

**3. Data Cleaning Process**

* **Handling Missing Values:**Identify and evaluate columns with missing data.
* Apply appropriate strategies.
* **Removing Duplicate Entries:**Locate and remove duplicate records to ensure data accuracy.
* **Standardize column names** (e.g., sales\_amount instead of SalesAmount).
* **Verifying Data Accuracy:**Check critical variables for anomalies (e.g., negative prices or invalid ratings).

**4. Creating the short\_title Feature**

* **Objective:**Generate a concise version of product titles that retains essential information for SEO and readability.
* **Steps:**Analyze original product titles to identify key components (e.g., product name, category, attributes).
* Remove redundant phrases or words (e.g., "includes," "set of," "features").
* Limit titles to 30–50 characters, focusing on essential details and keywords.
* **Implementation 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"

**5. Documentation and Reporting**

* Prepare a professional report documenting the data cleaning and title optimization process:
  + **Introduction:** Briefly describe the dataset and task objectives.
  + **Data Cleaning:** Summarize issues identified and the cleaning steps taken.
  + **Short Title Creation:** Explain the methodology and examples of optimized titles.
  + **Clean Dataset Overview:** Highlight key statistics and improvements in the dataset.
  + Include screenshots or visualizations illustrating the impact of data cleaning and title optimization.

**6. Review and Submission**

* Proofread and format the report for clarity and professionalism.
* Submit the cleaned dataset, including the short\_title column, and the report.

Deliverables

* **Cleaned Dataset:**Updated dataset free of missing values, duplicates, and inconsistencies, with the short\_title feature added.
* **Technical Report:**A professional document including:
* Introduction and objectives.
* Detailed steps for data cleaning and title optimization.
* Summary of improvements, with examples of original and short titles.

1. **NOTE**!

Your links should be google drive links. Change default setting on each google drive file to enable anyone have access to view.

1. **Submission:**
2. Submission link: [Submit](https://docs.google.com/forms/d/e/1FAIpQLSc4y0aFfK2qSYM5NX5oXpcUHetCvwMydEl7PPzqlzzqn48QOw/viewform?usp=sharing" \t "https://app.slack.com/client/T086NG3N69Y/_blank).

Acceptance Criteria

* **Data Cleaning:**All key issues in the dataset are resolved, including missing values, duplicates, and inconsistent formats.
* A clean, standardized dataset is prepared with the short\_title feature added.
* **Title Optimization:**The short\_title feature effectively summarizes product titles, retaining essential information and improving readability.
* Examples of the optimized titles are included in the report.
* **Technical Report:**The report is concise, professional, and includes:
* Overview of data cleaning and short title creation.
* Examples and insights from the clean dataset.
* **Timely Submission:**The task is completed and submitted before deadline.
* Deadline:**6th February, 11:59 pm.**
* Total Points: **10 points. *You need to get at least 6 points to move to the next stage.***