Data Collection and Preprocessing Phase

Date	25-03-2025
Team ID	LTVIP2025TMID28445
Project Title	Cosmetic insights
Maximum Marks	10 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	The dataset contains cosmetic product details, including ingredients, brand information, user reviews, and sustainability ratings. It is sourced from Kaggle and other public cosmetic databases.
Data Cleaning	- Identified and handled missing values in product descriptions and ingredient lists. - Removed duplicate entries to avoid redundancy. - Standardized brand names and ingredient terms for consistency.
Data Transformation	Applied filtering to select relevant product categories. - Sorted data based on user preferences (e.g., organic, cruelty-free). - Used pivot tables to analyze ingredient frequency across products. - Created calculated fields for average product ratings and ingredient safety scores.
Data Type Conversion	- Converted price fields from text to numerical format. - Ensured date fields for product launch and review timestamps were correctly formatted. - Standardized categorical data for better analysis.
Column Splitting and Merging	- Split ingredient lists into separate components for better analysis. - Merged multiple datasets (product details, user reviews, sustainability scores) into a unified dataset.

Data Modeling	- Established relationships between tables: products, brands, reviews, and ingredients. - Created a relational schema to improve data retrieval efficiency.
Save Processed Data	- Stored the cleaned and processed dataset in a structured format (CSV/SQL). - Ensured data integrity for future machine learning and analytics use.