Data Collection and Preprocessing Phase

Date	15 March 2025
Team ID	LTVIP2025TMID25168
Project Title	Cosmatic 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	**Short description about dataset**: • **Shape:** (1472 rows, 16 columns) • **Columns:** • Label, Brand, Name, Price, Rank, Ingredients, Combination, Dry, Normal, Oily, Sensitive, Sensitive Skin Suitability, Dry Skin Suitability, Normal Skin Suitability, No. of Records, Oily Skin Suitability. • **Missing Values:** None • **Duplicate Rows:** None
Data Cleaning	**Data Cleaning Summary:** • **Duplicates Removed:** 0 (No duplicates found) • **Invalid Price Entries:** 0 (All prices are valid) • **Invalid Rank Entries:** 0 (All ranks are within the 0-5 range) • **Final Shape After Cleaning:** (1472 rows, 16 columns)
Data Transformation	**Data Transformation Summary** • **Sorting:** Sorted products by price in descending order. • **Filtering:** Extracted **666 products** suitable for sensitive skin. • **Calculated Field:** Added a new column **"Price per**

<table>
<tr>
<td></td>
<td>**Rating"** (Price divided by Rank).
- **Pivot Table:** Shows the **average price** of different product types based on suitability for sensitive skin.
Pivot Table - Average Price by Product Type & Skin Suitability
<table>
<tr><th>Product Type</th><th>Suitable ($)</th><th>Not Suitable ($)</th></tr>
<tr><td>Cleanser</td><td>34.28</td><td>31.64</td></tr>
<tr><td>Eye Cream</td><td>67.55</td><td>59.04</td></tr>
<tr><td>Face Mask</td><td>45.93</td><td>40.36</td></tr>
<tr><td>Moisturizer</td><td>74.55</td><td>63.32</td></tr>
<tr><td>Sun Protect</td><td>47.03</td><td>45.02</td></tr>
<tr><td>Treatment</td><td>78.96</td><td>79.37</td></tr>
</table>
</td>
</tr>
<tr>
<td>Data Type Conversion</td>
<td>**Data Type Rectification Summary:**
- **No changes were needed** for most columns since they already had appropriate data types.
- **Key numeric columns verified:**
  ○ **Price:** int64
  ○ **Rank:** float64
  ○ **No. of Records:** int64
  ○ **Price per Rating:** float64</td>
</tr>
<tr>
<td>Column Splitting and Merging</td>
<td>**Column Splitting & Merging Summary**
- **Splitting:** Extracted the **first ingredient** into a new column **"Primary Ingredient"**.
- **Merging:** Created a **"Full Product Name"** column by combining **Brand** and **Name**.</td>
</tr>
<tr>
<td>Data Modeling</td>
<td>**Data Modelling Overview:**
Since we have a **single dataset**, we can define relationships for a structured database if we plan to integrate it with other tables. Below are possible relationships:
Potential Tables & Relationships
1. **Products Table** (Main Table)
  ○ **Primary Key:** Product ID (can be created if needed)
  ○ Contains product details: Label, Brand, Name, Price, Rank, Ingredients, Full Product Name
2. **Skin Suitability Table** (One-to-Many Relationship)
  ○ **Foreign Key:** Product ID
  ○ Columns: Sensitive Skin Suitability, Dry Skin</td>
</tr>
</table>

	Suitability, Normal Skin Suitability, Oily Skin Suitability 3. **Ingredients Table** (One-to-Many Relationship) ○ **Foreign Key:** Product ID ○ Columns: Primary Ingredient, Full Ingredient List 4. **Reviews Table** (If data is available) ○ **Foreign Key:** Product ID ○ Columns: No. of Records, Rank, Price per Rating
Save Processed Data	Save the cleaned and processed data for future use. cosmetics.csv