**Assignment Report**

**Task 1: Problem Analysis and Data Preprocessing**

**1. Problem Analysis**

**1.1 Objective**

The aim is to analyse the NewChic dataset to identify the top 10 products across selected categories and determine the best category based on key metrics: customer engagement (likes\_count), pricing (current\_price), discounts (discount), and product novelty (is\_new). This analysis will provide insights into product and category performance.

**1.2 Data Selection**

* **Comprehensive Coverage:** The analysis includes data from all provided categories (accessories, bags, beauty, house, jewellery, kids, men, shoes, women) to ensure a thorough evaluation. This approach captures a wide variety of products, helping to identify top performers.
* **Focused Metrics:** By prioritising metrics like likes\_count and current\_price, the analysis highlights products that are both popular and strategically priced, leading to more relevant insights.

**1.3 Defining “Top 10” and “Best”**

* **Top 10 Products:** Products are ranked based on likes\_count, current\_price, discount, and is\_new. This ranking ensures that the top 10 products are popular, strategically priced, attractively discounted, and potentially new.
* **Best Categories:** Categories are evaluated by their average current\_price, with the highest averages indicating the best categories. These categories likely represent premium segments that attract customers despite higher prices.

**1.4 Column Selection for Clustering, Classification, and Result Discussion**

* **Columns Chosen for Clustering and Classification:**
  + **current\_price:** Central to understanding pricing strategies. Used in clustering to group products with similar price ranges and in classification to categorize products based on price-related factors.
  + **likes\_count:** Reflects customer engagement, essential for clustering products by popularity and for classification to predict product success.
  + **discount:** Indicates the effectiveness of pricing strategies. Used in both clustering and classification to understand the impact of discounts on product categorization.
  + **is\_new:** Captures product novelty, which can be a significant factor in customer preferences and market success. Useful in clustering to identify groups of new versus established products and in classification to determine the likelihood of success based on novelty.
* **Columns for Result Discussion:**

**category:** Critical for grouping products and comparing performance across different categories. While it’s not used directly in clustering or classification algorithms, it is essential for discussing the results and understanding which categories perform best.

**name:** Identifies individual products, providing context in the result discussion. It allows for detailed analysis of top-performing items but is not used in algorithmic processing since it doesn't contribute to pattern recognition or model training.

**2. Program Data Preprocessing**

**2.1 Combining CSV Files**

* **Process:** Datasets from different categories were merged into a single DataFrame for consistent and comprehensive analysis.

**Code:**  
data\_frame\_combined = pd.concat([data\_frame\_accessories, data\_frame\_bags, data\_frame\_beauty, data\_frame\_house, data\_frame\_jewelry, data\_frame\_kids, data\_frame\_men, data\_frame\_shoes, data\_frame\_women], ignore\_index=True)

**2.2 Selecting Relevant Columns**

* **Process:** Only columns directly contributing to product performance and customer preferences were retained (category, name, current\_price, likes\_count, discount, is\_new).

**Code:**  
chosen\_columns = ['category', 'name', 'current\_price', 'likes\_count', 'discount', 'is\_new']

data\_frame\_preprocessed = data\_frame\_combined[chosen\_columns]

**2.3 Handling Missing Data**

* **Process:** Missing values in likes\_count and discount were filled with 0, and current\_price was filled with the median to maintain data reliability.

**Code:**  
data\_frame\_preprocessed['likes\_count'].fillna(0, inplace=True)

data\_frame\_preprocessed['discount'].fillna(0, inplace=True)

data\_frame\_preprocessed['current\_price'].fillna(data\_frame\_preprocessed['current\_price'].median(), inplace=True)

**2.4 Filtering by likes\_count**

* **Process:** Products with fewer than 50 likes were filtered out to focus on those with significant customer engagement.

**Code:**  
data\_frame\_preprocessed = data\_frame\_preprocessed[data\_frame\_preprocessed['likes\_count'] > 50]

**2.5 Identifying Top 10 Products**

* **Process:** Categories were ranked by average current\_price, and the top 7 categories were identified. Within these categories, the top 10 products were determined based on likes\_count, current\_price, discount, and is\_new.

**Code:**  
premium\_categories = data\_frame\_preprocessed.groupby('category')['current\_price'].mean().sort\_values(ascending=False)

best\_7\_categories = premium\_categories.head(7).index.tolist()

data\_frame\_filtered = data\_frame\_preprocessed[data\_frame\_preprocessed['category'].isin(best\_7\_categories)]

best\_10\_products = data\_frame\_filtered.sort\_values(

    by=['likes\_count', 'current\_price', 'discount', 'is\_new'],

    ascending=[False, False, False, False]

).head(10)

**2.6 Summary of Matched and Removed Columns**

* **Matched Columns:** Retained columns (category, name, current\_price, likes\_count, discount, is\_new) were essential for analysing customer behaviour and product performance.
* **Removed Columns:** Irrelevant columns like URLs and metadata were excluded to streamline the dataset.