Data Cleaning Report

Project: Amazon E-Commerce Sales Data Preparation
Intern: Khushi Bhatt

1. Introduction
The raw dataset contained 1,465 product entries from Amazon India, with 16 initial columns related to product details, pricing, discounts, and customer reviews. The goal of cleaning was to convert text-based numerical data into proper formats, handle inconsistencies, and create new features to enable robust sales and customer analysis.

2. Issues Identified & Actions Taken

Issue Category	Specific Problem	Action Taken	Business Rationale
Incorrect Data Types	discounted_price, actual_price stored as text (e.g., "₹1,299")	Removed '₹' and commas, converted to float64.	Enables mathematical operations (sum, average) for revenue and price analysis.
	discount_percentage stored as text (e.g., "64%")	Removed '%' symbol, converted to float64.	Allows for calculations of average discount and correlation analysis.

Issue Category	Specific Problem	Action Taken	Business Rationale
	rating stored as text, rating_count with commas (e.g., "24,269")	Converted rating to float64. Removed commas from rating_count and converted to int64.	Enables sorting, filtering, and statistical analysis on product popularity and quality.
Missing Values	2 missing values in rating_count.	Filled with 0, assuming no reviews were recorded.	Preserves the product record for pricing analysis while giving a logical value for review count.
	1 missing value in rating (after conversion).	Filled with the **column median (4.1)**.	Prevents loss of the product record; using the median minimizes bias in the rating

Issue Category	Specific Problem	Action Taken	Business Rationale
			distribution.
Data Inconsistency	category column contained long, pipe-separated strings.	Extracted the first category into a new main_category column.	Simplifies high-level analysis and visualization by product category.
Potential Outliers	Unrealistic discount_percentage values over 100%.	Identified and **capped all values at 100%**.	Ensures data integrity for analysis, as a discount cannot logically exceed the original price.

3. New Columns Created for Enhanced Analysis

- main_category: The primary product category for segmenting analysis.
- discount_amount: The absolute monetary saving (actual_price - discounted_price).
- is_high_impact: A Boolean flag identifying products with a rating >= 4.0 *and* a review count in the top 25%. Highlights top-performing products.

- value_score: A heuristic metric ((discount_percentage/100) * (rating/5)) to rank products by perceived value, combining savings and customer satisfaction.

4. Assumptions & Limitations

- rating_count is used as a proxy for sales volume, which is generally correlated but not perfectly accurate.
- The value_score is a simple heuristic. Real-world "value" is influenced by other factors like brand, features, and individual preferences.
- Filling the single missing rating with the median is a neutral approach but may slightly reduce variance in the rating data.

📌 **Task : Summary of Key Insights from the Cleaning Process**

- **Dataset Ready for Analysis:** The cleaned dataset now has **1,465 products** and **20 columns**, with **zero missing values** and all data types correctly formatted.
- **Critical Data Transformation:** The most significant cleaning effort was converting **price, discount, and rating columns from text to numeric formats**, which was the fundamental blocker for any quantitative analysis.
- **High-Impact Products Identified:** The new is_high_impact column automatically flags products that are both highly rated and widely reviewed. An initial check shows **approximately [You can run** df['is_high_impact'].sum() **to fill this in] products** meet this criteria, providing a direct list of market leaders to study.
- **New Analytical Dimensions Added:** The creation of discount_amount and value_score provides ready-made metrics for analysis, allowing immediate exploration of questions like "Do steeper discounts lead to higher perceived value?" without further data manipulation.
- **Data Integrity Assured:** Capping discounts at 100% and standardizing categories ensures that subsequent trend analysis and aggregations are based on reliable and logical figures.
