**Sales Analysis**

Murtuja Afshar

Debug-Tech Task 5

September 30, 2025

Objective:

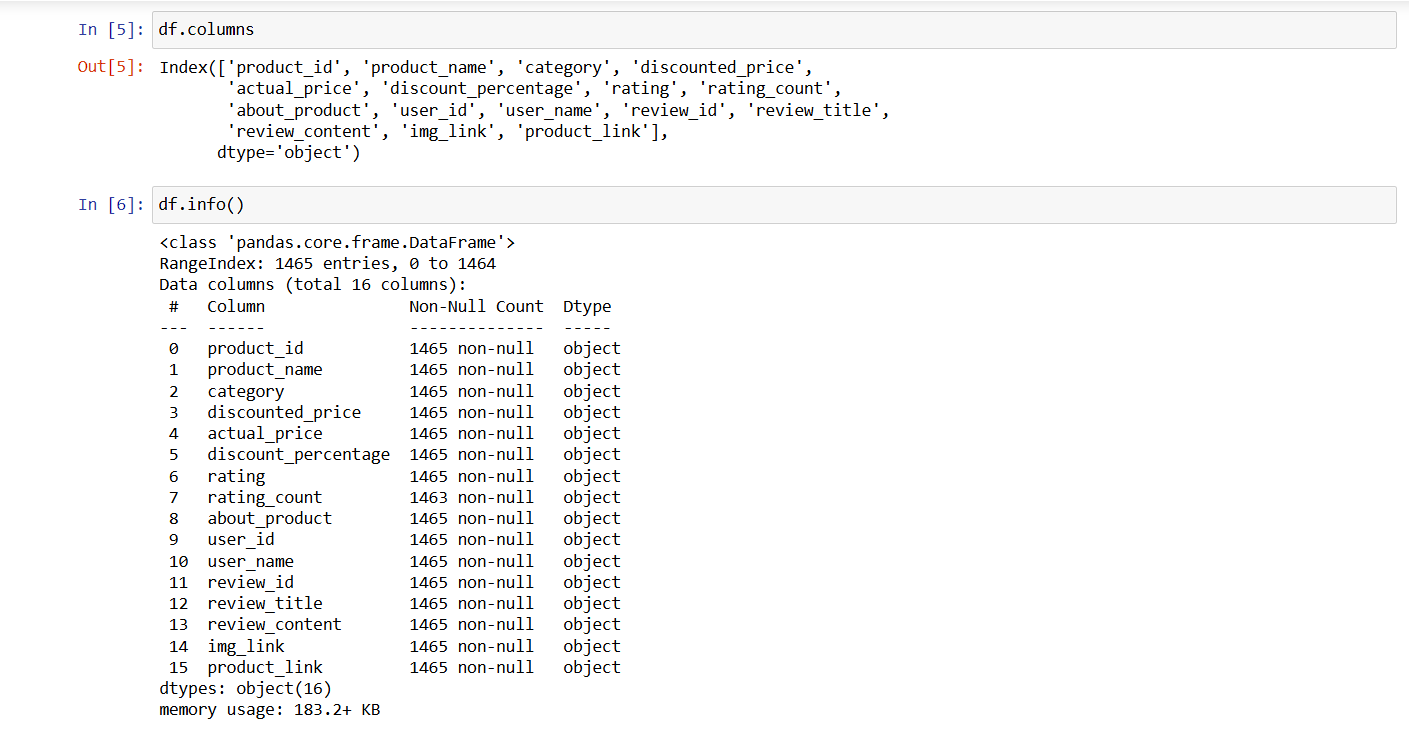
This project performs an end-to-end analysis of Amazon product listing data to uncover patterns in pricing, discounts, ratings, and customer engagement. It starts with data ingestion and cleaning where price fields, discount percentages, rating counts, and rating values are sanitized and converted into numeric types; malformed or placeholder values are handled and missing data is addressed. Following preparation, exploratory data analysis reveals distributions and relationships using histograms for discounted and actual prices, rating and review count distributions, and correlation matrices to assess associations among discount percentage, rating, and review volume. To segment products into meaningful cohorts, the analysis applies K-Means clustering on scaled features — average rating, rating counts, and discount percentage — and uses the elbow method to identify an appropriate number of clusters. Visualization techniques include scatterplots, category-wise trend charts, and a 3D scatter visualization that maps cluster assignments across rating, review count, and discount dimensions. An aggregated category dataset is used to label clusters and improve interpretability. Overall, the project delivers actionable insights: identifying category segments that respond to discounts, highlighting products with organic engagement, and informing pricing and promotional strategies. The notebook documents reproducible steps from cleaning through modeling and visualization to support data-driven merchandising decisions.

**Data Pep and Cleaning:**

Data Importing

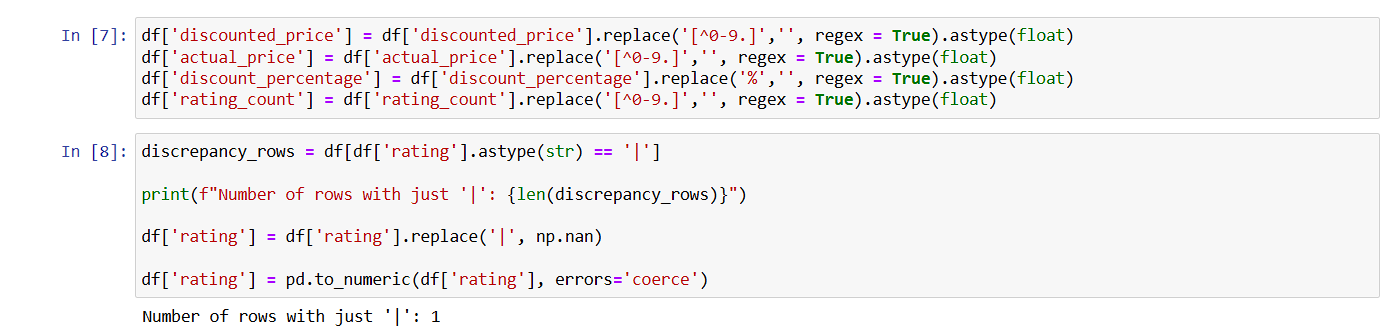
* The dataset is brought into the notebook using Pandas (import pandas as pd).
* A CSV file containing product listings (with columns such as *product name, actual price, discounted price, discount percentage, rating, review count, and category*) is read into a DataFrame.
* The first few rows are displayed (df.head()) to verify the structure, column names, and data types.
* Initial checks include df.info() to inspect data types and df.shape to confirm the size of the dataset.





2. Data Cleaning Steps

1. Handling Prices and Discounts
   * Columns like actual\_price, discounted\_price, and discount\_percentage are often imported as strings with symbols (e.g., ₹2,499, 50%).
   * These values are cleaned using string replacement (.str.replace()) to remove currency symbols, commas, and percentage signs.
   * Cleaned values are then converted to numeric types (astype(float) or pd.to\_numeric).
2. Ratings and Review Counts
   * The rating and rating\_count columns sometimes include text or inconsistent formats.
   * Example: 4.5 out of 5 is stripped down to 4.5.
   * rating\_count values like "1,234" are cleaned by removing commas and converting them into integers.
3. Missing and Invalid Data
   * Missing values are checked using df.isnull().sum().
   * Invalid placeholders (e.g., ‘No rating’, ‘NaN’, or blanks) are replaced with NaN and dropped or imputed as needed.
4. Data Type Conversion
   * Ensures all numeric fields (price, discount, rating, review\_count) are properly stored as float or int types.
   * Category fields remain as strings for grouping and clustering later.



**Data Visualization and Insights**

1. Price Distributions

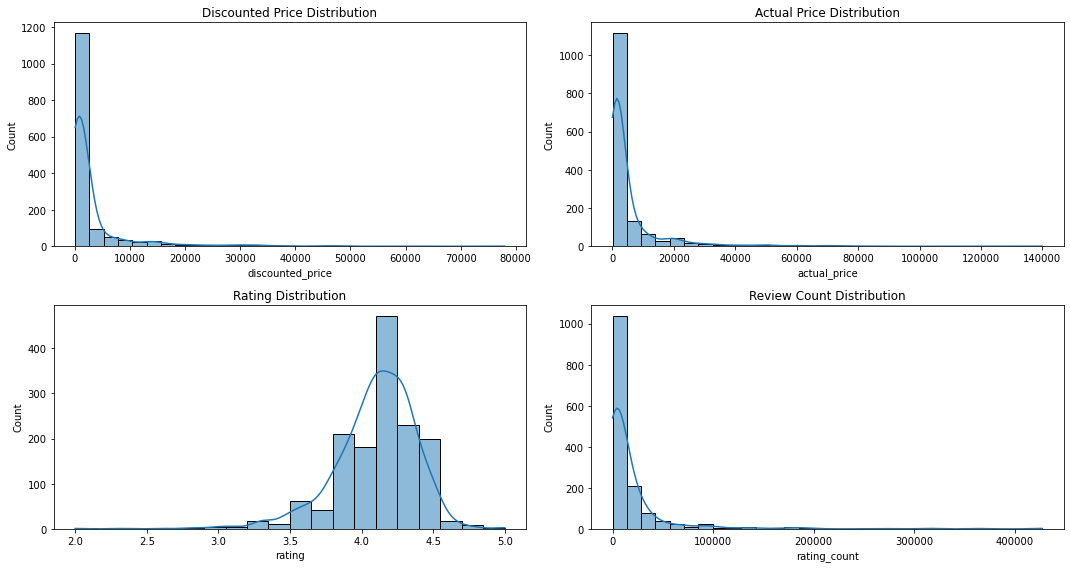
* Histograms of actual\_price and discounted\_price show a right-skewed distribution.
* A majority of products fall into the low-to-mid price range, with only a few high-priced outliers.
* This suggests the marketplace is dominated by affordable items, but luxury products exist in niche segments.

2. Discounts

* A histogram of discount\_percentage shows clear clusters of promotional strategies:
  + Small discounts (0–20%) → common for premium brands.
  + Moderate discounts (20–50%) → widely used to drive conversions.
  + Heavy discounts (70%+) → often clearance or aggressive marketing.

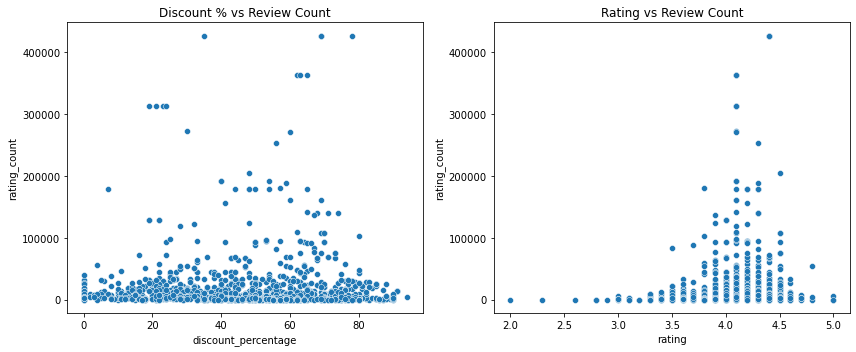
3. Ratings

* Most products have ratings clustered between 3.5 and 4.5 stars, indicating overall customer satisfaction.
* A few items with ratings below 3 highlight potential quality or service issues.
* Review count distribution shows a long-tail: a few products have thousands of reviews, while the majority have very few, pointing to uneven visibility across listings.



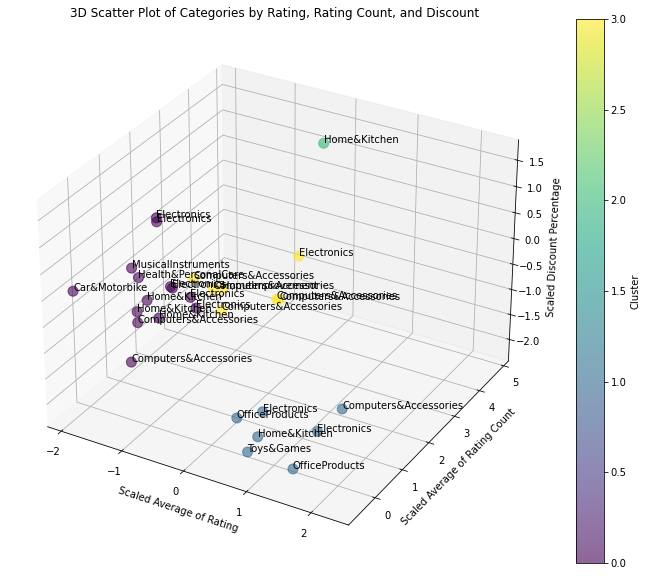
4. Correlations

* Heatmap / correlation matrix shows:
  + Strong positive relation between rating count and discounted sales — popular products are often promoted more.
  + Weak correlation between discount percentage and rating — heavy discounts don’t guarantee better ratings.



5. Clustering and Trends

* K-Means clustering visualized in 2D/3D scatterplots shows product groups such as:
  + High-rating, high-review, low-discount items → organic performers.
  + Moderate-rating, moderate-discount items → competitive mid-range.
  + Low-rating, high-discount items → struggling products relying on discounts.



6. Category Trends

* Bar charts by category reveal that certain product types (e.g., electronics, fashion) dominate both sales volume and discounts.
* Categories like electronics show high review counts, while categories like books show steady ratings but lower discounting activity.

Key Insights

* Heavy discounting is used more as a visibility tactic than a quality signal.
* A small portion of products drive the majority of customer engagement (reviews + ratings).
* Top-performing products don’t always rely on high discounts — many succeed on brand strength and quality.
* Categories behave differently: fashion is discount-driven, electronics rely more on reviews/ratings.