Analytical Questions for E-commerce Data Exploration

Jayant Kumar M

Data Source: Amazon Sales Dataset (Kaggle)

November 5, 2025

Analytical Questions by Category

A. Pricing and Discount Analysis (P&D)

Q1. What is the **average discounted price** and the **average actual price** for all products in the dataset? *[Hint: `Pandas mean()`]*

Q2. What is the **standard deviation** of the `discount_percentage`? *[Hint: `Pandas/Numpy std()`]*

Q3. Which product (`product_name` and `product_id`) has the **highest discount percentage**? *[Hint: `Pandas idxmax()`]*

Q4. Calculate the **total potential savings** (sum of `actual_price`−`discounted_price`) across all unique products. *[Hint: `Pandas arithmetic and sum()`]*

Q5. What is the **average actual price** for products that have a `discount_percentage` **greater than** 50%? *[Hint: `Pandas filtering and mean()`]*

Q6. Create a **histogram** to visualize the distribution of `discount_percentage` across all products. *[Hint: `Matplotlib hist()`]*

Q7. Generate **box plots** to compare the distribution of `discounted_price` across the top 5 largest categories. *[Hint: `Matplotlib boxplot() or Pandas plotting with groupby()`]*

Q8. Identify the number of products where the calculated discount (based on `actual_price` and `discounted_price`) **does not match** the listed `discount_percentage` (requiring data consistency check).

B. Rating and Review Analysis (R&R)

Q9 What is the overall average `rating` of all products, **weighted** by `rating_count`? *[Hint: Numpy weighted average calculation.]*

Q10 Which product (`product_id`) has the highest `rating_count` (most reviews/votes)? *[Hint: Pandas max()]*

Q11 Calculate the total number of individual ratings/votes (`rating_count`) recorded across the entire dataset. *[Hint: Pandas sum()]*

Q12 Determine the **distribution of rating values** $(1.0, 2.0, 3.0, 4.0, 5.0)$ and visualize it using a **bar chart**. *[Hint: Pandas value_counts() and Matplotlib bar()]*

Q13 Find the **product** with the longest `review_content` and identify the user who wrote it (`user_name`). *[Hint: Pandas string length calculation and idxmax()]*

Q14 Calculate the **correlation coefficient** between `rating` and `discount_percentage`. *[Hint: Pandas corr()]*

Q15 Is there a difference in average `rating` between products with a high rating count (e.g., top 10%) and those with a low rating count? *[Hint: Pandas quantile() and mean()]*

C. Category and Product Analysis (C&P)

Q16 Which `category` has the highest number of unique products? *[Hint: Pandas value_counts() and nunique()]*

Q17 What is the average `discounted_price` for each `category`? *[Hint: Pandas groupby() and mean()]*

Q18 Identify the top 10 most frequently occurring `product_name` and their respective counts. *[Hint: Pandas value_counts() with head()]*

Q19 Plot a bar chart showing the total accumulated `rating_count` for the top 10 categories. *[Hint: Pandas groupby() and Matplotlib bar()]*

Q20 Calculate the **coefficient of variation** ($\frac{\text{Standard Deviation}}{\text{Mean}}$) for the `discounted_price` within the top 3 largest categories to measure price volatility. *[Hint: Pandas groupby() and Numpy division]*

Q21 Which categories have an average `rating` below 3.5? *[Hint: Pandas groupby() and filtering.]*

Q22 Extract the length of the `about_product` description for each product and find the average description length per category. *[Hint: Pandas string functions and groupby()]*

Q23 Create a stacked bar chart showing the percentage of products in each category that are rated **4.0** or higher versus those rated **3.0** or lower.

D. User and Distribution Analysis (U&D)

Q24 Find the top 10 users (`user_name`) who have written the most reviews and display their review count. *[Hint: Pandas value_counts()]*

Q25 How many unique users have reviewed products in the "Electronics" category? *[Hint: Pandas filtering and nunique()]*

Q26 Create a scatter plot of `rating_count` versus `discounted_price` to visualize if expensive or cheap products attract more attention. *[Hint: Matplotlib scatter()]*

Q27 Calculate the **skewness and kurtosis** of the `rating_count` distribution to understand its shape. *[Hint: Pandas/Numpy statistical methods]*

Q28 Group products into three equal price bins (low, medium, high) based on `discounted_price` and calculate the average `rating` for each bin. *[Hint: Pandas qcut() or cut() and groupby()]*

Q29 Calculate the **correlation matrix** between all relevant numerical columns (`discounted_price`, `actual_price`, `discount_percentage`, `rating`, `rating_count`). *[Hint: `Pandas corr()`]*

Q30 Identify the percentage of products that have both a high discount (e.g., $> 50\%$) AND a high rating (e.g., > 4.0). *[Hint: `Pandas filtering and counting`.]*