AMAZON SALES DATA ANALYSIS

This dataset is having the data of 1K+ Amazon Product's Ratings and Reviews as per their details listed on the official website of Amazon

Features

- product_id - Product ID
- product_name - Name of the Product
- category - Category of the Product
- discounted_price - Discounted Price of the Product
- actual_price - Actual Price of the Product
- discount_percentage - Percentage of Discount for the Product
- rating - Rating of the Product
- rating_count - Number of people who voted for the Amazon rating
- about_product - Description about the Product
- user_id - ID of the user who wrote review for the Product
- user_name - Name of the user who wrote review for the Product
- review_id - ID of the user review
- review_title - Short review
- review_content - Long review
- img_link - Image Link of the Product
- product_link - Official Website Link of the Product

1) Data Import and Initial Setup

- The analysis begins by importing essential libraries:
- pandas and numpy for data manipulation.
- matplotlib and seaborn for visualization.
- The data is loaded from a CSV file, “amazon.csv” which contains information about Amazon product sales**.**

2) Exploratory Data Analysis (EDA)

Shape of Data: The dataset's shape (number of rows and columns) is checked, helping to understand its scale.

- The datset contains 1465 rows and 16 columms.

Initial Inspection:

- data.head() is used to display the first few rows, providing a quick look at the data.
- data.columns lists all column names, giving an overview of the available attributes.
- data.info() gives details on data types and identifies columns with missing values.

Statistical Summary:

- data.describe() provides basic statistics (mean, min, max, etc.) for numeric columns, aiding in initial insights.

3) Missing Values

- data.isnull().sum() identifies columns with missing values. The rating_count column, for example, has missing data.
- Rows with missing values in rating_count are removed to ensure data integrity in subsequent analysis.
- Now after removing the null values the dataset has 1463 rows and 16 columns.

4) Data Cleaning and Transformation

Price Columns:

- discounted_price and actual_price contain currency symbols and commas. These are removed, and values are converted to numeric types for accurate calculations.

Discount Percentage:

- The % symbol in discount_percentage is removed, and the values are converted to integers for simpler computations.

5) Duplicate Data Handling

- Duplicate rows are identified and removed if necessary. This step confirms data uniqueness, which is essential for accurate analysis.

Unique columnmn Analysis:

- The product_id column is investigated for unique entries. Some product IDs repeat, indicating multiple entries for the same product.
- The product_link column has no duplicate values so I just found out that unique col is product_link col which means this data is based on the product_links column.

6) Data EXtraction

- Top 5 Highest Rated Products.
- Bottom 5 Lowest Rated Products
- Product Count per Category
- Top 5 Most Expensive Products
- Top 5 Least Expensive Products
- Top 5 Highest Discounted Products

7) Data Visualization

- Visualizing Price Distributions
- visualizing the frequency of different product ratings.
- visualizing Prices by Category using boxplot
- Visualizing the count of products in each category.
- visualizing Comparison of Actual vs Discounted Prices
- visualizing Discount Percentage by Category
- Visualizing the distribution of actual product prices.
- Visualizing Relationship between Rating and Discounted price
- visualizing Top 5 Most Expensive Products by Category
- Visualizing average Dicount by category