**Amazon Sales Dataset**

**About Dataset**

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

**Inspiration**

Amazon is an American Tech Multi-National Company whose business interests include E-commerce, where they buy and store the inventory, and take care of everything from shipping and pricing to customer service and returns. I've created this dataset so that people can play with this dataset and do a lot of things as mentioned below

* Dataset Walkthrough
* Understanding Dataset Hierarchy
* Data Preprocessing
* Exploratory Data Analysis
* Data Visualization
* Making Recommendation System  
  This is a list of some of that things that you can do on this dataset. It's not definitely limited to the one that is mentioned there but a lot more other things can also be done.

**Questions**

**🔹 Basic Descriptive Analysis**

1. What is the distribution of product categories? Which category has the most products listed?
2. What is the average actual price vs. discounted price across categories?
3. Which product has the highest discount percentage?
4. How does the average rating vary by product category?
5. Which product categories have the most reviews (rating\_count)?

**🔹 Price & Discount Insights**

1. Do higher discounts always lead to higher ratings or more reviews?
2. Compare the **average discount percentage** across categories. Which category tends to have the biggest discounts?
3. What is the relationship between **actual price** and **rating**? Do cheaper products tend to get better ratings?
4. Are there outliers in pricing (very high or very low products compared to the category average)?

**🔹 Customer Reviews & Sentiment**

1. What are the most common words in product review titles and review content?
2. Do products with more reviews (rating\_count) tend to have higher average ratings?
3. Perform a sentiment analysis of review\_content (positive vs. negative). Which categories have the most positive vs. negative reviews?
4. Are short reviews (review\_title) more likely to be positive or negative compared to long review\_content?
5. Who are the top reviewers (users with the most reviews)?

**🔹 Product Popularity**

1. Which product has the highest number of reviews?
2. Which products have the best rating with the least number of reviews (hidden gems)?
3. Which products are overpriced (high actual price, low rating)?
4. Which products are underrated bargains (low price, high rating, high discount)?

**🔹 Comparative & Correlation**

1. Correlate **discount percentage** with **rating** — do higher discounts influence better customer satisfaction?
2. Correlate **rating\_count** with **discounted\_price** — do cheaper products get more reviews?
3. Does the **review sentiment** correlate with numerical ratings? (e.g., text says positive but rating is low).

**🔹 Visualization Tasks**

1. Create a bar chart showing **average rating per category**.
2. Plot a boxplot of **discount percentage per category**.
3. Create a scatterplot of **discounted\_price vs rating**, color-coded by category.
4. Build a word cloud from review\_content.
5. Create a correlation heatmap of numerical features (price, discount, rating, rating\_count).