About Data -- I take an Amazon data set , 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.

**Column Overview:**

Product\_id-Unique identifier for each product

product-name: Name of the product

category-Categorization of products, Computers & Accessories like this

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

**Problem Statement –**

Imagine that the Amazon sales manager is not able to know how many sales are happening in our company and which product is selling more and who is the top customer, So I think these are some problem statements.

1. What are the top-rated and most-reviewed products in each category
2. Which products are the top sellers in each category
3. Which products have low customer engagement or poor ratings despite high pricing or substantial discounts
4. Total sales
5. Top 5 customer and top 5 product

**Data Cleaning Steps –**

1. Handle Missing Values
2. Convert Data Types
3. Remove Duplicates
4. Standardize Data Formats - Standardization is especially important when dealing with data types like dates, currencies, categories , in this data actual\_price column include currency symbols, commas.
5. Handle Outliers
6. Save clean file for make deshbord