Perfect! You're turning this **Amazon Sales Dataset** into a **real-world industry-style case study** – this is a fantastic move for interviews, portfolios, or training programs.

🧾 Title: Amazon Product Recommendation System — Industry Case Study

> 🔍 **Objective**: Build and evaluate a hybrid recommendation engine (content-based + collaborative filtering) for an e-commerce platform using product metadata and user behavior data.

📊 **Dataset**:
https://drive.google.com/file/d/1OA0wOG1epBxHAMNXun7kwNr_QsupUMzj/view?usp=sharing

🧠 **Total Marks**: 100
🎯 **Skill Areas**: Data cleaning, NLP, EDA, recommender systems, model evaluation, practical business thinking

---

📄 Structured Case Study Questions with Marks

Section A: Data Understanding & Cleaning (15 marks)

> Understand the dataset like a data scientist in a product analytics team.

Q1. (5 marks) Summarize the dataset:

- Number of unique users, products, and reviews
- Top 5 categories by number of products
- Price range and discount insights

Q2. (5 marks) Clean and preprocess the data:

- Convert prices to numeric
- Parse categories into hierarchy levels

- Normalize rating scores and count outliers or Create derived features like `price_difference`, `value_for_money_score`, weighted ratings

Q3. (5 marks) Handle missing values or anomalies

- Remove duplicates, invalid records
- Handle missing ratings/reviews with appropriate strategy

Section B: Exploratory Data Analysis (10 marks)

Think like a product analyst trying to identify buying patterns.

Q4. (5 marks) Visualize:

- Most reviewed products Visualize top 10 categories by number of products.
- Average rating per category
- Discounts vs actual price correlation
- User Engagement Insights (5 marks)
    - Which products have high ratings but low review counts?
    - Are highly rated products also heavily reviewed?

Q5. (5 marks) Create 3 actionable insights for Amazon's product strategy based on EDA.

Section C: Content-Based Filtering (20 marks)

Act like a content engineer personalizing user feeds based on product metadata.

Q6. (5 marks) Vectorize product text (`about_product` + `product_name`) using:

- TF-IDF or embeddings
- Build a product similarity matrix

Q7. (5 marks) Recommend top 5 similar products to:

- A **new product** with no reviews
- A product with high user dropout (bad ratings)

Q8. (5 marks) Add category, price, and discount to enhance content vectors

Q9. (5 marks) Evaluate recommendations:

- How diverse and relevant are the content-based results?

Section D: Collaborative Filtering (User–Item) (30 marks)

> Now you're a machine learning engineer building smart recommendations using user behavior.

Q10. (5 marks) Create a user-item matrix using `user_id`, `product_id`, and `rating`.

Q11. (20 marks) Apply:

- User-User Collaborative Filtering (cosine similarity)

Or

- Item-Item Collaborative Filtering (cosine or Pearson)
- Recommend top 5 unseen products per user

Q14. (5 marks) Recommend top 5 unseen products per user

Section E: Hybrid Recommender (Content + Collaborative) (20 marks)

> Step into the role of a senior ML engineer combining models for better performance.

Q15. (5 marks) Design a hybrid strategy:

- Score fusion: `0.6 * CF_score + 0.4 * Content_score`

Q16. (5 marks) Compare recommendation quality of hybrid vs individual methods.

Q17. (5 marks) Evaluate hybrid system on:

- A cold-start product (new product)
- A cold-start user (few reviews)

Q18. (5 marks) Suggest how to improve hybrid performance further using real-world constraints like:

- Popularity
- Recent purchases
- Product availability

Bonus Challenge

Section F: Business Strategy & Deployment – 10 Marks

1. Based on your findings:
    - Which model works best for new users ? (2 marks)
    - Which model works best for returning users ? (2 marks)
    - How can we recommend products with no ratings ? (2 marks)
2. How would you deploy this system in production? Mention tools/technologies. (2 marks)
3. What KPIs should Amazon track to measure success? (2 marks)