AI Data Engineer

Take Home Assignment: The Messy Review Dataset
Scenario: You are working with a team of data scientists to build a sentiment analysis model for product reviews. They have provided you with a raw, messy data set. Your task is to transform this into a clean, well-documented dataset ready for model training.

The Task:

1. **Data Ingestion & Cleaning:** Write a Python script to load a subset of the Amazon Customer Reviews dataset. Perform necessary data cleaning, including:
    a. Handling missing or empty reviews.
    b. Standardizing text (e.g., converting to lowercase, removing HTML tags if present).
    c. Identify and document at least two data quality issues you find in the dataset (e.g., duplicate reviews, suspicious patterns).
2. **Data Annotation:** The dataset has a star_rating column. Create a new sentiment column by mapping the ratings:
    a. 1-2 stars -> negative
    b. 3 stars -> neutral
    c. 4-5 stars -> positive
3. **Output:** Generate a final, clean CSV file named training_dataset.csv containing the cleaned review_body and the new sentiment label.

Dataset:

- **Amazon Customer Reviews Dataset:** (Attached to email)

Deliverables:

- A Python script for the entire process.
- The training_dataset.csv file.
- A README.md explaining your process, the data quality issues you found, and your labeling logic.

Skills Assessed:

- Data Creation & Annotation
- Data Quality & Evaluation
- Data Manipulation (Pandas)
- Documentation