Problem Statement: Optimizing E-Commerce Strategy through Data-Driven Insights

1. Background

In the highly competitive e-commerce landscape, Amazon hosts millions of products across diverse categories, from Electronics to Home & Kitchen. For sellers and market analysts, simply having a great product is no longer enough. Success depends on understanding the delicate balance between **pricing**, **discount strategies**, and **customer sentiment** (ratings).

2. The Problem

Currently, raw sales data is overwhelming and siloed. Without a structured analysis pipeline, it is difficult to answer critical business questions, such as:

- **The Discount Trap:** Are heavy discounts actually leading to better customer ratings, or are they perceived as "cheap" or lower quality?
- **Category Saturation:** Which categories are over-saturated with products but have low consumer engagement?
- **Reliability Gaps:** How can we distinguish between a product with a "lucky" 5-star rating (few reviews) and a truly "reliable" top-seller (thousands of reviews)?

3. Proposed Solution

This project develops an end-to-end data pipeline to transform raw Amazon CSV data into a strategic decision-making tool. The solution involves:

1. **Data Engineering (Python):** Cleaning inconsistent text data and handling missing values to create a "Single Source of Truth."
2. **Database Management (SQL):** Moving data into a relational environment (MySQL) to allow for complex, multi-dimensional querying.
3. **Business Intelligence (Power BI):** Visualizing the results through interactive dashboards that allow stakeholders to identify "Value Gems"—products with high ratings and high discounts.

4. Target Objectives

- **Identify** the top 4 most discounted categories to understand market trends.
- **Visualize** the distribution of products across "High," "Medium," and "Low" discount tiers.
- **Segment** products by category and rating count to highlight the most "reliable" inventory.