Project Title: Team Epsilon E-Commerce Entry Analysis

Team Members: Chris Bushelman, Chorr-Kin Chin, William Fetter, Rumani Kafle, Myatminn Khant, Jan Lelie, Molly Pfefferkorn, Xavier Walsh, Jacob Weiss, Sunil Williams

Project Outline: Examining trends in products listed on Amazon in 2023 for entry into e-commerce

Research Questions: See Below

Datasets Used: Amazon Product Dataset 2023: https://www.kaggle.com/datasets/asaniczka/amazon-products-dataset-2023-1-4m-products/data

Breakdown of Tasks: See Below

### **Project Proposal**

Before you start writing any code, your group should outline the scope and purpose of your project. This will help provide direction and safeguard against **scope creep** (the tendency for projects to become more complex after work begins).

**Background**: Team Epsilon has been approached by an entrepreneur looking to expand into e-commerce.   
 We have been tasked with analyzing sales data from Amazon.com to come up with an entry   
 strategy, including data on categories, reviews, pricing, and other data as available.

**Data**: Team Epsilon obtained sales data from 2023(US) from Kaggle.com. The dataset includes   
 categorical data as well as reviews counts, sales counts, and pricing among a few others.

**The questions you’ll ask of the data:  
1.** Which categories have the most products? The least products?- Jan  
**2.** Which categories have the most best-sellers? The least number of best-sellers?- Myat  
**3.** Which categories have the highest average ratings? The lowest average ratings?- Rumani  
**4.** Which categories have the highest number of reviews? The lowest number of reviews?- (Xavier)  
**5.** What is the correlation between number of reviews and star ratings? Full business and category- Molly  
**6.** What is the correlation between # of and # of best sellers? Maybe look at the categories with   
 the highest numbers of these items?(Chris)  
**7.** Statistics summary/average price for each category- (Sunil)  
**8.** Which categories dominate the top 10% of the sales rank?- (Sunil)  
**9.** Performance of “Men’s” vs “Women’s” categories: ratings, #reviews, #best sellers- Chorrkin  
**10.** What is the correlation between product price and the number of reviews? (Will)  
11. Do products with higher ratings tend to have more reviews? (Jake W)

12. Compare list price to current price and see which categories had the most sales

13. Make a new column, call it profitability (price x boughtInLastMonth) and then sort categories by profitability so we can speak on which categories hold the most profit.

Use the following example for guidance:

The aim of our project is to uncover patterns in credit card fraud. We’ll examine relationships between transaction types and location, purchase prices and times of day, purchase trends over the course of a year, and other related relationships derived from the data.

Other questions we could answer:

**1. Product Performance Analysis**

• Which product categories have the highest average ratings?

• What is the distribution of ratings across all products?

• Are there certain product categories that are overrepresented in terms of the number of products?

• What is the average price of products within each category?

• Which product categories tend to have the highest-priced products?

**2. Customer Behavior Analysis**

• Do products with higher ratings tend to have more reviews?

• What is the correlation between product price and the number of reviews?

• How do ratings vary by price range (e.g., low-cost, mid-tier, premium)?

• Which product categories tend to receive the most reviews on average?

**3. Sales and Market Trends**

*(if sales rank or sales data is included)*

• Which categories dominate the top 10% of the sales rank?

• Is there a relationship between sales rank and product price?

• How does the average sales rank differ between highly rated and poorly rated products? (Jake W interested but if someone else wants it feel free)

**4. Pricing Strategies**

• What is the price distribution across different categories?

• Do certain categories have a wider range of prices than others?

• Are there any outlier products with extremely high or low prices?

• How does the price vary for products with similar ratings?

**5. Competitive Analysis**

• Are there any dominant brands in specific categories?

• What percentage of products are from top brands versus smaller brands?

• Which categories show the most diversity in terms of brands?

**6. Text Analysis (if product descriptions or reviews are included)**

• What are the most frequently used keywords in product descriptions?

• Are there specific keywords that appear more often in top-rated products?

• Can sentiment analysis be used to detect patterns in review comments (positive vs. negative)?

**7. User Recommendations**

• Based on ratings, price, and reviews, what products would you recommend in each category?

• What features (price, rating, number of reviews) correlate most strongly with customer satisfaction?

• Can you build a model to predict sales rank based on price, reviews, and ratings?

**8. Visualizations for Key Insights**

• Create visualizations like:

• A bar chart of average ratings by category.

• A scatterplot of price vs. rating.

• A heatmap showing correlation between numerical features (price, reviews, ratings, etc.).

• A pie chart of product representation across categories.