**Level 1 (Regular Data Science Questions):**

1. Calculate the average Rating for each Category. Which category has the highest average Rating?

Highest average rating: Books, average rating: 4.8

1. What is the correlation coefficient between Price and Discount?

-0.049

1. Find the top 3 most common Colors in the dataset and the number of products associated with each.

Black: 16 products; Red: 10 products; Blue: 10 products

**Level 2 (Multiple Step Hard Data Science Questions):**

1. For each Category, compute the total revenue (Price \* Quantity) and determine which Category generates the most revenue.

Fashion: $83,239.20

1. Create a new column “PriceAfterDiscount” (Price \* (1 - Discount/100)). Calculate the mean PriceAfterDiscount for each Category.

Beauty: $9.86, Books: $21.91, Electronics: $71.99, Fashion: $69.37, Health & Personal Care: $23.91, Home & Kitchen: $65.03, Sports & Outdoors: $21.76, Toys & Games: $13.65

1. Group the data by Color and calculate the average Quantity and Discount. Identify the Color with the highest average Quantity and lowest average Discount.

Highest average Quantity: Pink (104.29 units), color with lowest average Discount: Gray (0.83%)

1. Perform a time series analysis on monthly sales data. Assume each row corresponds to a sale made in a consecutive month starting from January 2020. Identify the month with the highest total sales.

March 2021: $22,499.10

1. Determine the Price range (max - min) for each Category. Which Category has the widest Price range?

Fashion: $230.00

1. Identify the top 5 most rated products (by Rating) and calculate their average Price.

Adventure Book Collection: Rating 4.8, Price $49.99; Classic Novels Collection: Rating 4.8, Price $29.99; Mystery Novel Collection: Rating 4.8, Price $19.99; Travel Guide Book Collection: Rating 4.8, Price $29.99; Travel Memoir Collection: Rating 4.8, Price $19.99

**Level 3 (Multistep Data Analysis and Machine Learning Questions):**

1. Using a clustering algorithm, categorize the products based on Price and Rating. How many clusters are formed and what are the central values of Price and Rating in each cluster?

3 clusters, 0: Central Price: $26.15, Central Rating: 4.67; 1: Central Price: $129.99, Central Rating: 4.35; 2: Central Price: $23.28, Central Rating: 4.13

1. Employ a decision tree classifier to predict the Category of a product based on Price, Quantity, Discount, and Rating. What is the accuracy of this model?

100%

1. Use a regression model to predict the Quantity sold based on Price, Discount, and Rating. What is the R-squared value of this model?

0.306

1. Apply a neural network model to classify products into high-rated (Rating >= 4) and low-rated (Rating < 4). What is the precision and recall of this model?

Precision and Recall of 100%

1. Implement a random forest model to predict the Color of a product based on Price, Quantity, and Discount. What is the feature importance ranking from this model?

Price: Approximately 57.6% importance; Quantity: Approximately 25.0% importance; Discount: Approximately 17.3% importance

1. Conduct a principal component analysis (PCA) to reduce the dimensions of the dataset. How much variance is explained by the first two principal components?

The first principal component accounts for about 55.25% of the variance. The second principal component accounts for about 26.78% of the variance.