DAB501 -PROJECT #1 -PROBLEM STATEMENT

INTRODUCTION:

Group number- 11, Section- 001, Group Members-

- Vidya Biradar
- Sujata Biswas

BACKGROUND/ MOTIVATION:

Amazon being one of the largest e-commerce sites in the world provides plenty of clients a wide range of products at the convenience of customers' doorstep. Being an online retail tycoon, Amazon needs to understand the constantly changing customer behavior, needs and preferences to keep up with the consumers and stay on top of all the others in the field.

Apart from that, understanding and analysing Amazon's data of product reviews and ratings is crucial for other businesses and e-commerce organizations to gain insights and understand the pattern this platform follows so as to make better business decisions and to be able to compete and gain better outcomes.

PROBLEM STATEMENT:

The online retail industry is growing swiftly as more customers want to shop online for convenience and accessibility. As a result, to stay abreast of changing consumer preferences, internet retailers like Amazon must constantly improve their product selections. To do this, however, retailers must be knowledgeable about the preferences and purchasing practises of their customers.

The product category, price, discount, rating, and reviews are just a few of the many variables that affect consumer purchasing. After comparing and evaluating these variables, it is possible to draw important conclusions that will benefit both customers and shops like Amazon.

PROJECT PROPOSAL:

Our team will be analysing the Amazon dataset that consists of information on various categories such as products, product categories, costs, discounts, reviews, ratings and customer name and ID. We will be using various statistical techniques to visualize the underlying patterns, trends and relationships between factors such as product category, rating, discount percentage & rating count, rating & product category.

Our group aims to achieve these goals by doing the following steps-

1. Exploring & Studying Dataset- Studying the dataset and determining the factors based on which we can analyse trends to visualize them.
2. Cleaning the Data & Filling Missing Values (if any)- Cleaning using R and finding missing values(if any) to replace them using statistical methods.
3. Predictive Analysis- Checking which factors can be compared amongst each other to turn into graphs to check for useful insights.
4. Visualization- Converting the predictive analysis into graphs to show the analysis in a much more understandable manner conveying the right information without distorting facts.

ANALYSIS QUESTIONS:

The project will try to address the analysis questions listed below:

1. What are the best-selling product categories on Amazon, and what elements support their dominance?
2. Which product category has the highest rating and rating count?
3. What is the level of difference between discounted price and the actual price?
4. What is the relationship between product category, discount percentage and rating count?

DATASET DESCRIPTION:

This analysis will be based on the dataset that contains information from more than 1,000 Amazon products' ratings and reviews, as reported on the company's official website. This dataset includes various variables including-

- product_id - Product ID
- product_name - Name of the Product
- category - Category of the Product
- discounted_price - Discounted Price of the Product
- actual_price - Actual Price of the Product
- discount_percentage - Percentage of Discount for the Product
- rating - Rating of the Product
- rating_count - Number of people who voted for the Amazon rating
- about_product - Description about the Product
- user_id - ID of the user who wrote review for the Product
- user_name - Name of the user who wrote review for the Product
- review_id - ID of the user review
- review_title - Short review
- review_content - Long review
- img_link - Image Link of the Product
- product_link - Official Website Link of the Product

The dataset shows various categorical as well as continuous variables which are quite useful to make proper comparisons and to check for trends. The dataset gives information about the rating of the different products. This dataset in all has 1465 rows and 17 columns.