Name: **Gargi Devendra Shintre**
Div: **D15C** Roll No. **69**

Practical No. 3

Aim:

To perform Exploratory Data Analysis and Visualization Python.

Introduction:

Exploratory Data Analysis (EDA) is a crucial step in data science that involves summarizing, cleaning, and visualizing data to uncover hidden patterns, trends, and relationships. It provides insights that guide data preprocessing, feature engineering, and model building.
The given dataset consists of the top Amazon bestselling items for the year 2025, including product details such as price, star rating, number of ratings, and ranking. Through EDA, we aim to analyze product distributions, customer preferences, and correlations between price, popularity, and ratings.

Descriptive Analysis – Central Tendency

Definition:

Central tendency measures the "center" of a dataset using mean, median, and mode.

- **Mean** – Average value
- **Median** – Middle value when sorted
- **Mode** – Most frequent value

Execution:

We calculated mean, median, and mode for numerical attributes:

- Product Price
- Star Rating
- Number of Ratings

Inference:

- The **average product price** is around ₹732, while the **median price** is ~₹146. This suggests the presence of **outliers** (a few highly expensive products).
- The **average star rating** is ~4.1, indicating that most products are well-rated.
- The **median number of ratings** is ~218, while the mean is ~1,312, which again shows **skewness** due to a few very popular items.

Descriptive Analysis – Dispersion

Definition:

Dispersion shows how spread out the data is.

- **Range** = Max – Min
- **Variance** = Average squared deviation from mean
- **Standard Deviation** = Spread around mean

- **IQR (Interquartile Range)** = Q3 – Q1

Execution:
We computed dispersion for price, ratings, and number of ratings.

Inference:
- The **price range** is ₹2.99 to ₹12,739, indicating extreme variability.
- A **high standard deviation** in ratings count shows that while some products get very few reviews, others attract thousands.
- The **IQR of price** highlights that 50% of products are priced between ₹40 – ₹741.

Correlation Analysis

Definition:
Correlation measures how strongly two numerical variables are related.

Execution:
We checked correlation among:
- Price vs Number of Ratings
- Price vs Star Rating
- Ratings vs Number of Ratings

Inference:
- Weak correlation between **price and ratings** suggests that expensive products are not always higher rated.
- **Number of ratings and star rating** show only mild correlation, meaning popularity doesn't always imply quality.
- Rankings are mostly independent of price and ratings.

Data Visualization

1. **Histogram – Distribution of Price**
    - Most products are low-priced (<₹500).
    - A few expensive products act as **outliers**.

2. **Box Plot – Ratings by Country**
    - Ratings are consistently high across countries (majority above 3.5).
    - Few outliers show poorly rated products.

3. **Scatter Plot – Price vs Number of Ratings**
    - Most high-rated products are in the **affordable range**.

    - Some cheap products still dominate reviews, showing popularity isn't tied to cost.

4. **Bar Chart – Top 10 Products by Number of Ratings**
    - These products dominate customer attention.
    - Bestseller items generally belong to affordable categories.

5. **Heatmap – Correlation Matrix**
    - Confirms weak correlations between price and popularity.
    - Suggests other hidden factors (e.g., marketing, brand reputation) drive sales.

6. **Pie Chart – Market Share by Top 10 Countries**
    - Most entries in this dataset are from **India (IN)**, since it's region-specific data.

Conclusion:

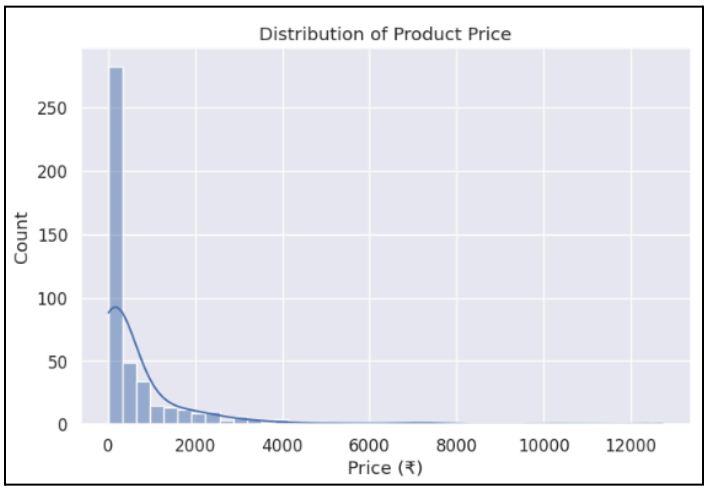
The EDA revealed that:

- Amazon bestsellers are mostly **affordable products**, while a few premium items act as outliers.
- **High customer ratings (4.0–4.5)** are common across products, showing that bestsellers tend to satisfy customers.
- **Popularity (number of ratings)** is not always correlated with price or rating, suggesting factors like accessibility, demand, and brand influence sales.
- Visualization helped identify skewness, outliers, and product trends, making EDA a critical step before any predictive modeling.

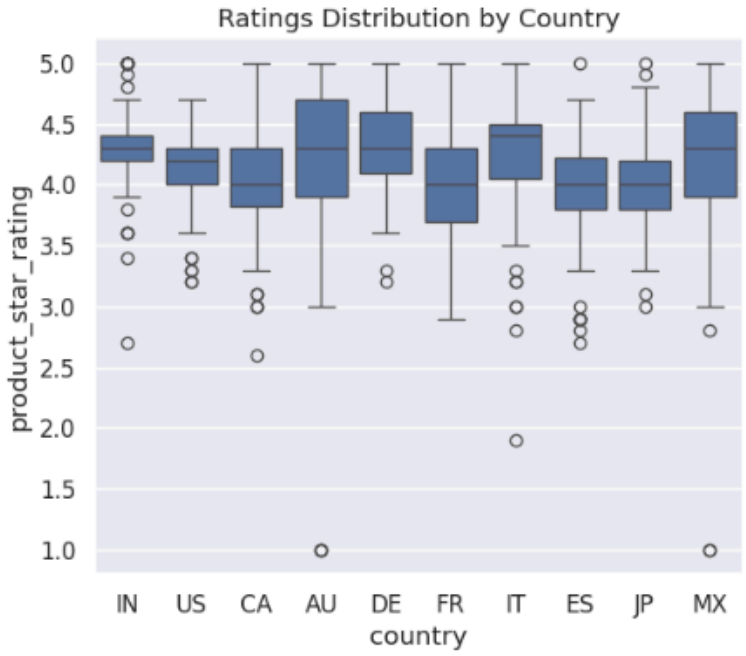
Output:

```
#Data Visualization

# Histogram - Distribution of Price
plt.figure(figsize=(8,5))
sns.histplot(df["product_price"].dropna(), bins=40, kde=True)
plt.title("Distribution of Product Price")
plt.xlabel("Price (₹)")
plt.ylabel("Count")
plt.show()
```

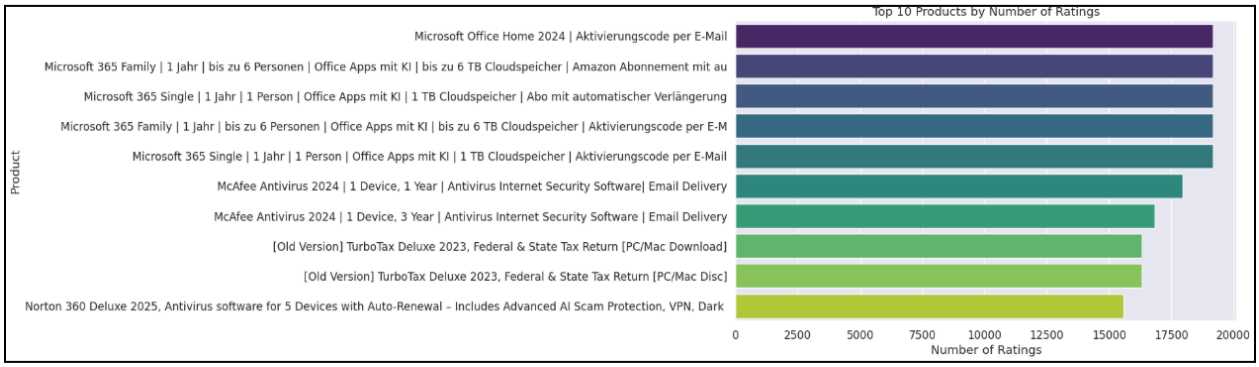


```
#Boxplot - Ratings by Country
plt.figure(figsize=(6,5))
sns.boxplot(x="country", y="product_star_rating", data=df)
plt.title("Ratings Distribution by Country")
plt.show()
```

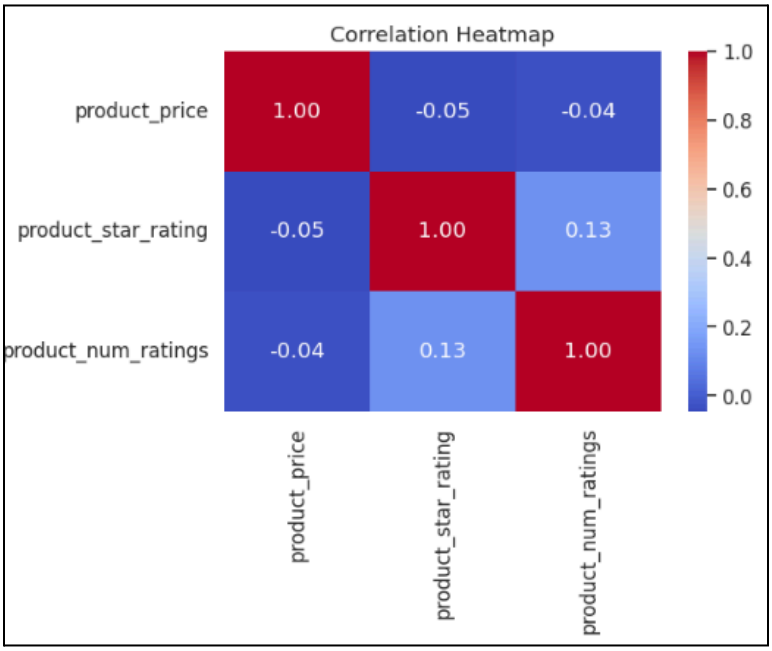



```
#Scatter Plot - Price vs Number of Ratings
plt.figure(figsize=(8,5))
sns.scatterplot(x="product_price", y="product_num_ratings", data=df, alpha=0.6)
plt.title("Price vs Number of Ratings")
plt.xlabel("Price (₹)")
plt.ylabel("Number of Ratings")
plt.show()
```

```
#Bar Chart - Top 10 Products by Number of Ratings
top10 = df.nlargest(10, "product_num_ratings")
plt.figure(figsize=(10,6))
sns.barplot(x="product_num_ratings", y="product_title", data=top10, palette="viridis")
plt.title("Top 10 Products by Number of Ratings")
plt.xlabel("Number of Ratings")
plt.ylabel("Product")
plt.show()
```




```
#Heatmap - Correlation Matrix
corr = df[["product_price", "product_star_rating", "product_num_ratings"]].corr()
plt.figure(figsize=(6,4))
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



```
#Pie Chart - Market share by country
country_counts = df["country"].value_counts().head(10)
plt.figure(figsize=(6,6))
plt.pie(country_counts, labels=country_counts.index, autopct="%1.1f%%", startangle=90)
plt.title("Market Share by Top 10 Countries")
plt.show()
```