AMAZON Sales Analysis: Uncovering key Insights for Business Growth

Analyzed by **Tahrin Ferdousi (** Data Analyst **)**

Introduction: This project focuses on analyzing Amazon sales data to discover trends, understand product performance , and identify key factors that influence sales. It aims to support better business decisions through data-driven insights.

Goals:

- Clean and explore sales data
- Identify top-selling products and categories
- Analyze factors affecting sales (e.g., price, ratings, reviews)
- Visualize key metrics for clarity
- Recommend strategies for sales improvement

Tools used:

- Python
- Pandas
- Matplotlib
- Seaborn
- Jupyter Notebook

Data Source: Kaggle

```python
import pandas as pd
```

```python
df = pd.read_csv(r'C:\Users\user\Downloads\archive (11).zip')
```

```python
print(df.head(2))
```

```
   product_id                                       product_name  \
0  B07JW9H4J1  Wayona Nylon Braided USB to Lightning Fast Cha...
1  B098NS6PVG  Ambrane Unbreakable 60W / 3A Fast Charging 1.5...

                                            category discounted_price
0  Computers&Accessories|Accessories&Peripherals|...            ₹399
1  Computers&Accessories|Accessories&Peripherals|...            ₹199

  actual_price discount_percentage rating rating_count  \
0       ₹1,099                 64%    4.2       24,269
1         ₹349                 43%    4.0       43,994

                                       about_product  \
0  High Compatibility : Compatible With iPhone 12...
1  Compatible with all Type C enabled devices, be...

                                                user_id  \
0  AG3D6O4STAQKAY2UVGEUV46KN35Q,AHMY5CWJMMK5BJRBB...
1  AECPFYFQVRUWC3KGNLJIOREFP5LQ,AGYYVPDD7YG7FYNBX...

                                           user_name  \
```

Loading and getting information of the dataset

Used Pandas to load a *.csv* file which containing e-commerce product data. After loading I displayed first two rows to get a quick overview of the dataset. It includes information such as product name, category, prices, etc.

```python
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   product_id           1465 non-null   object
 1   product_name         1465 non-null   object
 2   category             1465 non-null   object
 3   discounted_price     1465 non-null   object
 4   actual_price         1465 non-null   object
 5   discount_percentage  1465 non-null   object
 6   rating               1465 non-null   object
 7   rating_count         1463 non-null   object
 8   about_product        1465 non-null   object
 9   user_id              1465 non-null   object
 10  user_name            1465 non-null   object
 11  review_id            1465 non-null   object
 12  review_title         1465 non-null   object
 13  review_content       1465 non-null   object
 14  img_link             1465 non-null   object
 15  product_link         1465 non-null   object
dtypes: object(16)
memory usage: 183.3+ KB
None
```

Getting Information

It displays a concise summery including the number of rows, columns, data types, and not null counts for each column. this helps quickly understand the data structure and missing values.

```
print(df.describe())
```

```
        product_id                                       product_name  \
count         1465                                               1465
unique        1351                                               1337
top     B07JW9H4J1  Fire-Boltt Ninja Call Pro Plus 1.83" Smart Wat...
freq             3                                                  5

                                                 category discounted_price  \
count                                                1465             1465
unique                                                211              550
top     Computers&Accessories|Accessories&Peripherals|...             ₹199
freq                                                  233               53

       actual_price discount_percentage rating rating_count  \
count          1465                1465   1465         1463
unique          449                  92     28         1143
top            ₹999                 50%    4.1        9,378
freq            120                  56    244            9

                                            about_product  \
count                                                1465
unique                                               1293
top     [CHARGE & SYNC FUNCTION]- This cable comes wit...
freq                                                    6
```

Getting a proper description of the dataset

Used *df.describe()* to view basic states like mean, count, and standard deviation for numerical columns like rating and *rating_count*. It is useful for spotting outliners and understanding data distribution.

```python
print(df.isnull().sum())
```

```
product_id             0
product_name           0
category               0
discounted_price       0
actual_price           0
discount_percentage    0
rating                 0
rating_count           2
about_product          0
user_id                0
user_name              0
review_id              0
review_title           0
review_content         0
img_link               0
product_link           0
dtype: int64
```

Null value counts

Used *df.isnull().sum()* to find missing values in each columns which is an essential step before cleaning to ensure data quality.

```python
df[['discounted_price','actual_price','rating_count']] = df[['discounted_price','actual_price','rating_count']].apply(pd.to_numeric, errors='coerce')
```

```python
print(df.head(2))
```

```
   product_id                                       product_name  \
0  B07JW9H4J1  Wayona Nylon Braided USB to Lightning Fast Cha...
1  B098NS6PVG  Ambrane Unbreakable 60W / 3A Fast Charging 1.5...

                                          category discounted_price  \
0  Computers&Accessories|Accessories&Peripherals|...             ₹399
1  Computers&Accessories|Accessories&Peripherals|...             ₹199

  actual_price discount_percentage rating rating_count  \
0       ₹1,099                 64%    4.2       24,269
1         ₹349                 43%    4.0       43,994

                                       about_product  \
0  High Compatibility : Compatible With iPhone 12...
1  Compatible with all Type C enabled devices, be...

                                             user_id  \
0  AG3D6O4STAQKAY2UVGEUV46KN35Q,AHMY5CWJMMK5BJRBB...
1  AECPFYFQVRUWC3KGNLJIOREFP5LQ,AGYYVPDD7YG7FYNBX...

                                           user_name  \
0  Manav,Adarsh gupta,Sundeep,S.Sayeed Ahmed,jasp...
1  ArdKn,Nirbhay kumar,Sagar Viswanathan,Asp,Plac...
```

Converted *discounted_price, actual_price*, and *rating_count* column to numeric. Which will help to analyze the dataset. Then used *df.head()* to confirm the change

```python
print(df.dtypes)
```

```
product_id              object
product_name            object
category                object
discounted_price       float64
actual_price           float64
discount_percentage     object
rating                  object
rating_count           float64
about_product           object
user_id                 object
user_name               object
review_id               object
review_title            object
review_content          object
img_link                object
product_link            object
dtype: object
```

Preparing these columns for analysis and confirming the change

Rechecked *df.dtypes* to confirm that those columns were successfully converted to numeric types.

Top Selling Products

```
top_selling = df.sort_values(by='rating_count', ascending=False)
```

```
top_selling.head(10)
```

	product_id	product_name	category	discounted_price	actual_price	discount_percentage	rating	rating_count	about
1279	B08L12N5H1	Eureka Forbes car Vac 100 Watts Powerful Sucti...	Home&Kitchen|Kitchen&HomeAppliances|Vacuum,Cle...	NaN	NaN	16%	|	992.0	No Ins provid p
750	B00ZRBWPA0	Eveready Red 1012 AAA Batteries - Pack of 10	Electronics|GeneralPurposeBatteries&BatteryCha...	NaN	NaN	12%	4.3	989.0	Thinl Think E Evere
1059	B09Y5FZK9N	Pigeon 1.5 litre Hot Kettle and Stainless Stee...	Home&Kitchen|Kitchen&HomeAppliances|SmallKitch...	NaN	NaN	48%	3.7	976.0	Class Th mirror
151	B08QSDKFGQ	Zoul USB Type C Fast Charging 3A Nylon Braided...	Computers&Accessories|Accessories&Peripherals|...	NaN	NaN	69%	4.3	974.0	{3A/Q CHAR DA
32	B08QSC1XY8	Zoul USB C 60W Fast Charging 3A	Computers&Accessories|Accessories&Peripherals|	NaN	NaN	65%	4.3	974.0	{3A/Q CHAR

Identified Top_selling_Products by grouping and sorting based on *rating_count*. Displayed the top 10 using the *top_selling.head(10)* to highlight high-engagement items.

Top Categories by Revenue

```
df['Estimated_revenue']=df['discounted_price'] * df['rating_count']
top_categories_revenue = df.groupby('category')['Estimated_revenue'].sum().sort_values(ascending=False)
```

```
top_categories_revenue.head(10)
```

```
category
Car&Motorbike|CarAccessories|InteriorAccessories|AirPurifiers&Ionizers              0.0
Home&Kitchen|Kitchen&HomeAppliances|SmallKitchenAppliances|EggBoilers               0.0
Home&Kitchen|Heating,Cooling&AirQuality|RoomHeaters                                 0.0
Home&Kitchen|Heating,Cooling&AirQuality|RoomHeaters|ElectricHeaters                 0.0
Home&Kitchen|Heating,Cooling&AirQuality|RoomHeaters|FanHeaters                      0.0
Home&Kitchen|Heating,Cooling&AirQuality|RoomHeaters|HalogenHeaters                  0.0
Home&Kitchen|Heating,Cooling&AirQuality|RoomHeaters|HeatConvectors                  0.0
Home&Kitchen|Heating,Cooling&AirQuality|WaterHeaters&Geysers|ImmersionRods          0.0
Home&Kitchen|Heating,Cooling&AirQuality|WaterHeaters&Geysers|InstantWaterHeaters    0.0
Home&Kitchen|Heating,Cooling&AirQuality|WaterHeaters&Geysers|StorageWaterHeaters    0.0
Name: Estimated_revenue, dtype: float64
```

This code calculates estimated revenue for each product by multiplying *discounted_price* and *rating_count*. Then it sums these revenues by *category* and sorts them to find the highest-earning categories. Showing the top 10 helps identify the most profitable product groups.

Top Product Category by Rating and Review Engagement

```
rating_engagement = df.groupby('category')['rating_count'].sum().sort_values(ascending=False)
review_engagement = df.groupby('category')['review_id'].count().sort_values(ascending=False)
engagement_df = pd.DataFrame({'rating_engagement': rating_engagement,'review_engagement':review_engagement}).sort_values(by='rating_engagement',ascending
engagement_df.head(10)
```

category	rating_engagement	review_engagement
Computers&Accessories|Accessories&Peripherals|Cables&Accessories|Cables|USBCables	24503.0	233
Electronics|HomeTheater,TV&Video|Accessories|RemoteControls	12675.0	49
Home&Kitchen|Heating,Cooling&AirQuality|WaterHeaters&Geysers|InstantWaterHeaters	5208.0	23
Electronics|WearableTechnology|SmartWatches	4695.0	76
Home&Kitchen|Kitchen&HomeAppliances|Vacuum,Cleaning&Ironing|Irons,Steamers&Accessories|LintShavers	3802.0	22
Home&Kitchen|Kitchen&HomeAppliances|Vacuum,Cleaning&Ironing|Vacuums&FloorCare|Vacuums|HandheldVacuums	3471.0	8
Electronics|HomeTheater,TV&Video|Televisions|SmartTelevisions	3213.0	63
Home&Kitchen|Kitchen&HomeAppliances|SmallKitchenAppliances|Kettles&HotWaterDispensers|Kettle&ToasterSets	2969.0	13
Home&Kitchen|Kitchen&HomeAppliances|SmallKitchenAppliances|EggBoilers	2144.0	11
Home&Kitchen|Heating,Cooling&AirQuality|RoomHeaters|FanHeaters	2066.0	20

This code calculate engagement metrics by category. It first sums the total *rating_counts* per category (**rating_engagement**) and counts the number of reviews per *category* (**review_engagement**). These two metrics are combined into a new **DataFrame** and sorted by **rating engagement** in descending order. Displaying the top 10 rows shows which categories have the highest user interaction through ratings and reviews.

Preparing column for analysis (Removing special symbol)

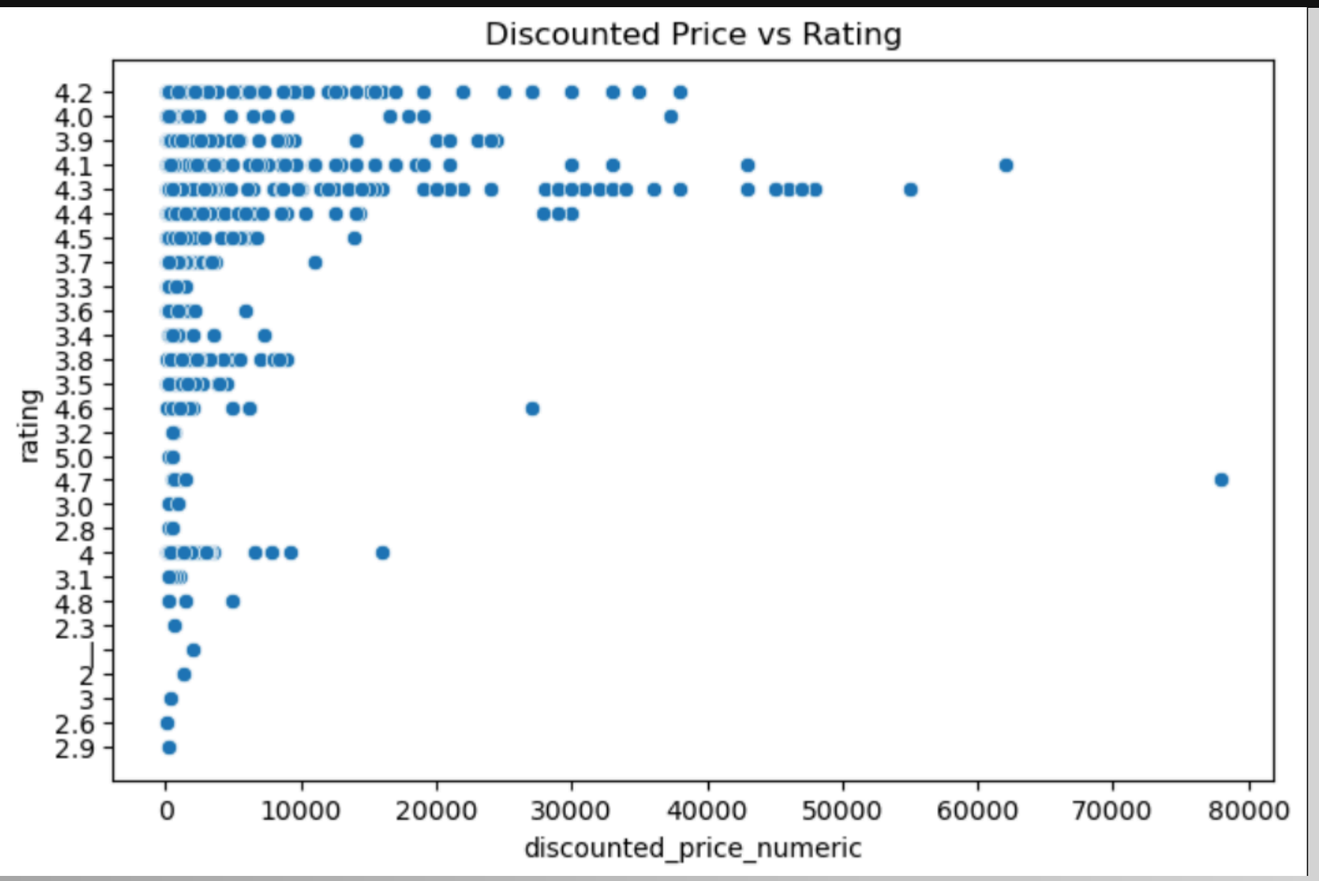
```
df[['discounted_price','actual_price','rating_count','rating']] = df[['discounted_price','actual_price','rating_count','rating']].apply(pd.to_numeric, er
```

```
import pandas as pd
df['discounted_price_clean'] = df['discounted_price'].astype(str).str.replace('₹', '', regex=False).str.replace(',', '', regex=False)


df['discounted_price_numeric'] = pd.to_numeric(df['discounted_price_clean'], errors='coerce')


print(df[['discounted_price', 'discounted_price_numeric']])
```

```
     discounted_price  discounted_price_numeric
0                ₹399                     399.0
1                ₹199                     199.0
2                ₹199                     199.0
3                ₹329                     329.0
4                ₹154                     154.0
...               ...                       ...
1460             ₹379                     379.0
1461           ₹2,280                    2280.0
1462           ₹2,219                    2219.0
1463           ₹1,399                    1399.0
1464           ₹2,863                    2863.0

[1465 rows x 2 columns]
```

This code clean and converts the *discounted_price* column from a string with currency symbols and commas into a numeric format. First it removes the **rupee symbol** and **commas**, then convert the cleaned string into numeric columns, handling any errors gracefully. Finally the original and cleaned numeric price columns are printed for comparison.

```python
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(8,5))
sns.scatterplot(data=df, x='discounted_price_numeric', y='rating')
plt.title('Discounted Price vs Rating')
plt.show()
```



Price vs Rating (Scatter Plot)

This scatter plot shows how *discounted_price* relates to product *rating,* helping visualize any trends between price and rating.

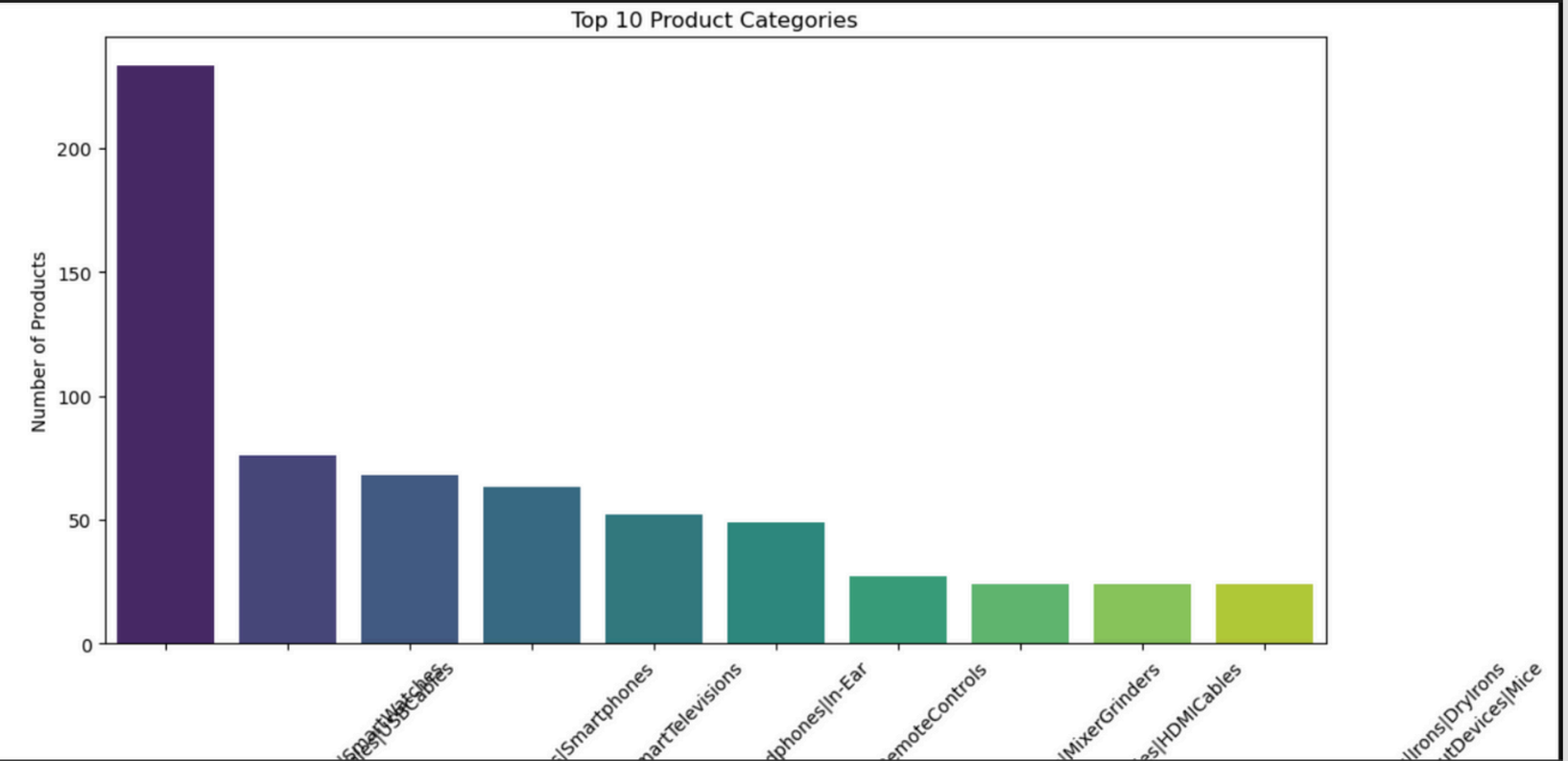
Correlation with Rating and Review count

```python
print("Correlation with Rating:")
print(df['discount_percentage'].corr(df['rating']))

print("\nCorrelation with review count:")
print(df['discount_percentage'].corr(df.groupby('product_id')['review_id'].transform('count')))
```

```
Correlation with Rating:
-0.15537543972693016

Correlation with review count:
0.20488211461384617
```

This code calculates the correlation between *discount_percentage* and *product_rating*, and also between *discount_percentage* and the number of reviews per product. It helps understand how discount levels relate to customer ratings and review activity.

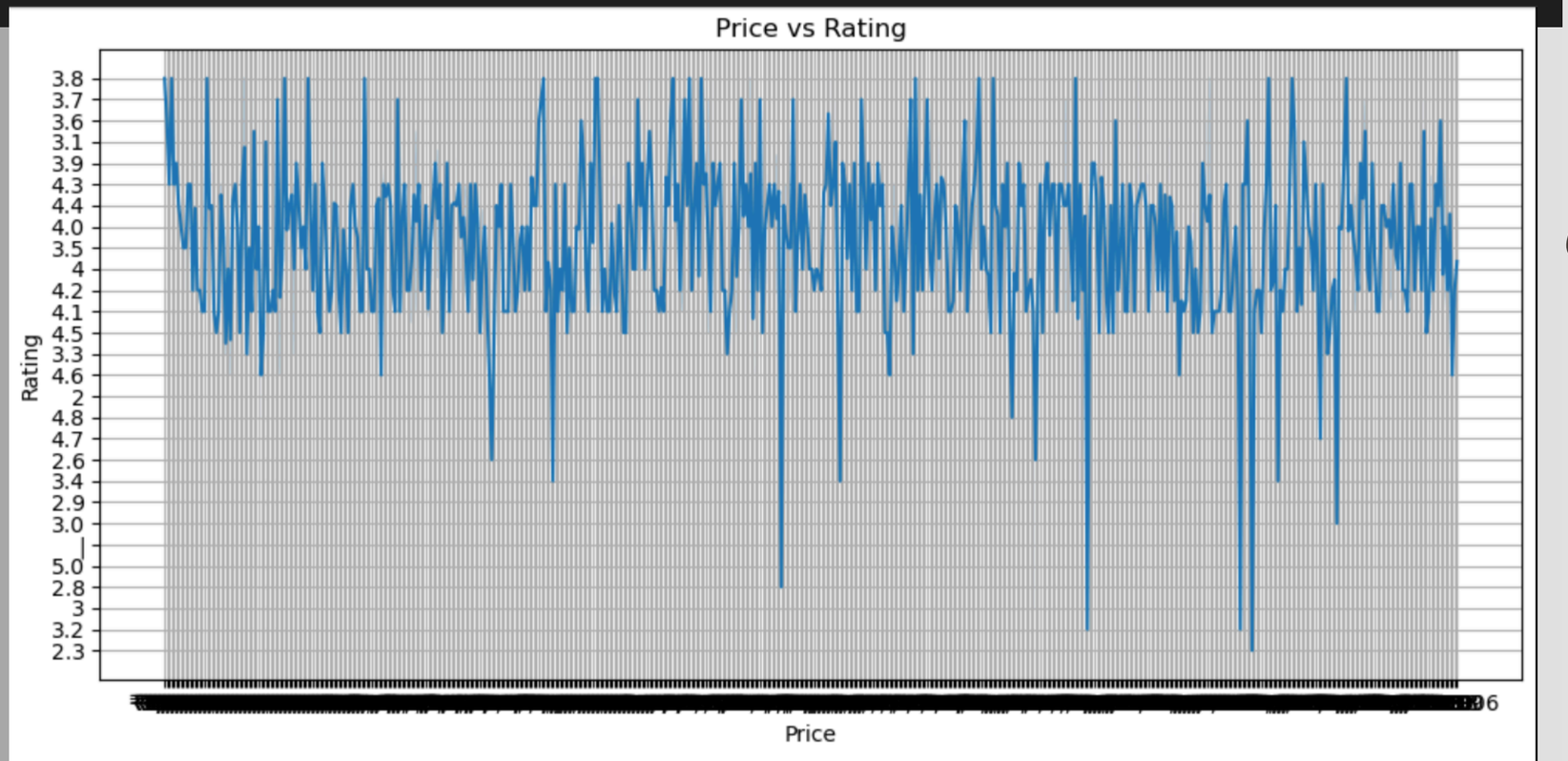
```python
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(12, 6))
top_categories = df['category'].value_counts().head(10)
sns.barplot(x=top_categories.index, y=top_categories.values, palette="viridis")
plt.title("Top 10 Product Categories")
plt.xticks(rotation=45)
plt.ylabel("Number of Products")
plt.tight_layout()
plt.show()
```



Top 10 Product Categories (Bar Plot)

This bar plot shows the top 10 categories by product count with clear labels and colors.

```python
df_sorted = df[['discounted_price', 'rating']].dropna().sort_values('discounted_price')

plt.figure(figsize=(10, 5))
sns.lineplot(x='discounted_price', y='rating', data=df_sorted)
plt.title("Price vs Rating")
plt.xlabel("Price")
plt.ylabel("Rating")
plt.grid(True)
plt.tight_layout()
plt.show()
```



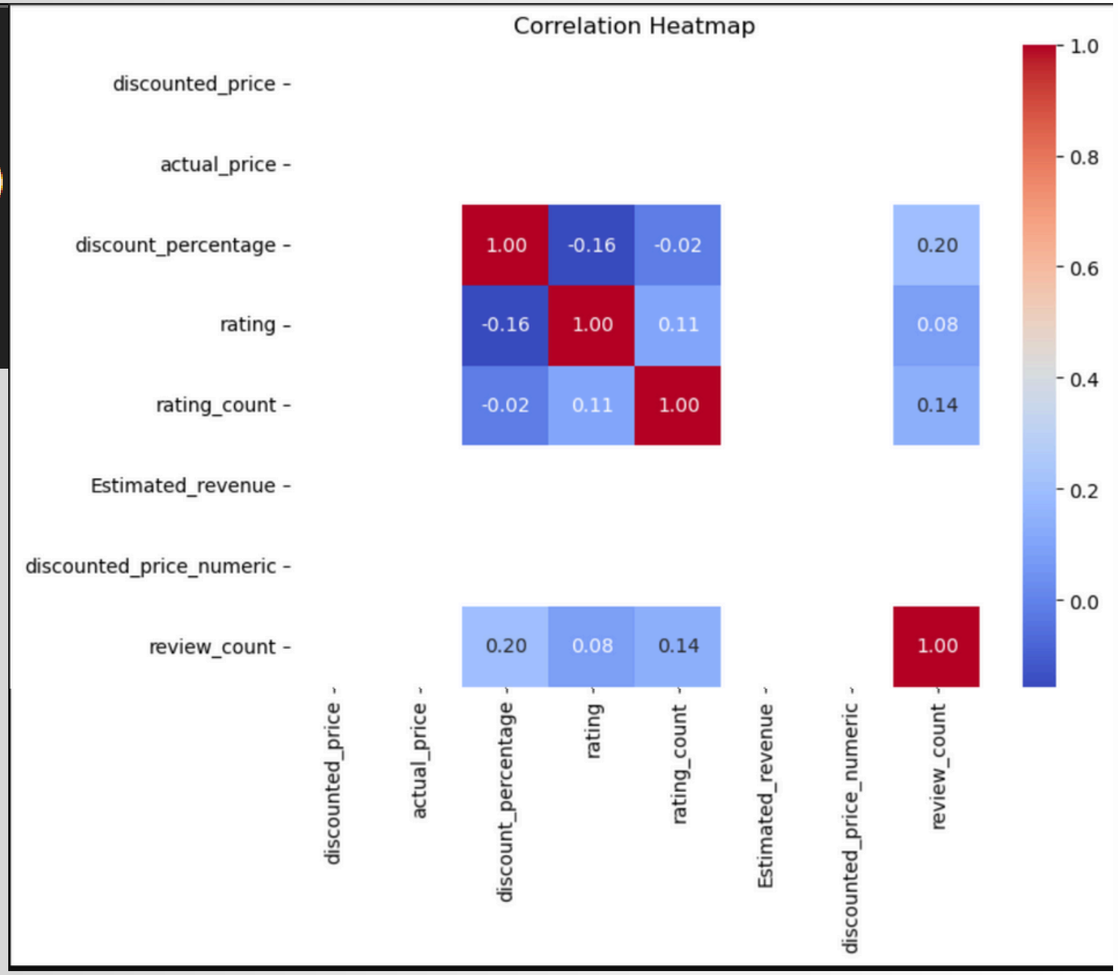
Price vs Rating (Line Graph)

This code plots a line graph showing how *rating* changes with *discounted price*. It sorts data by price and removes missing values to ensure a smooth trend line.

Correlation Heatmap

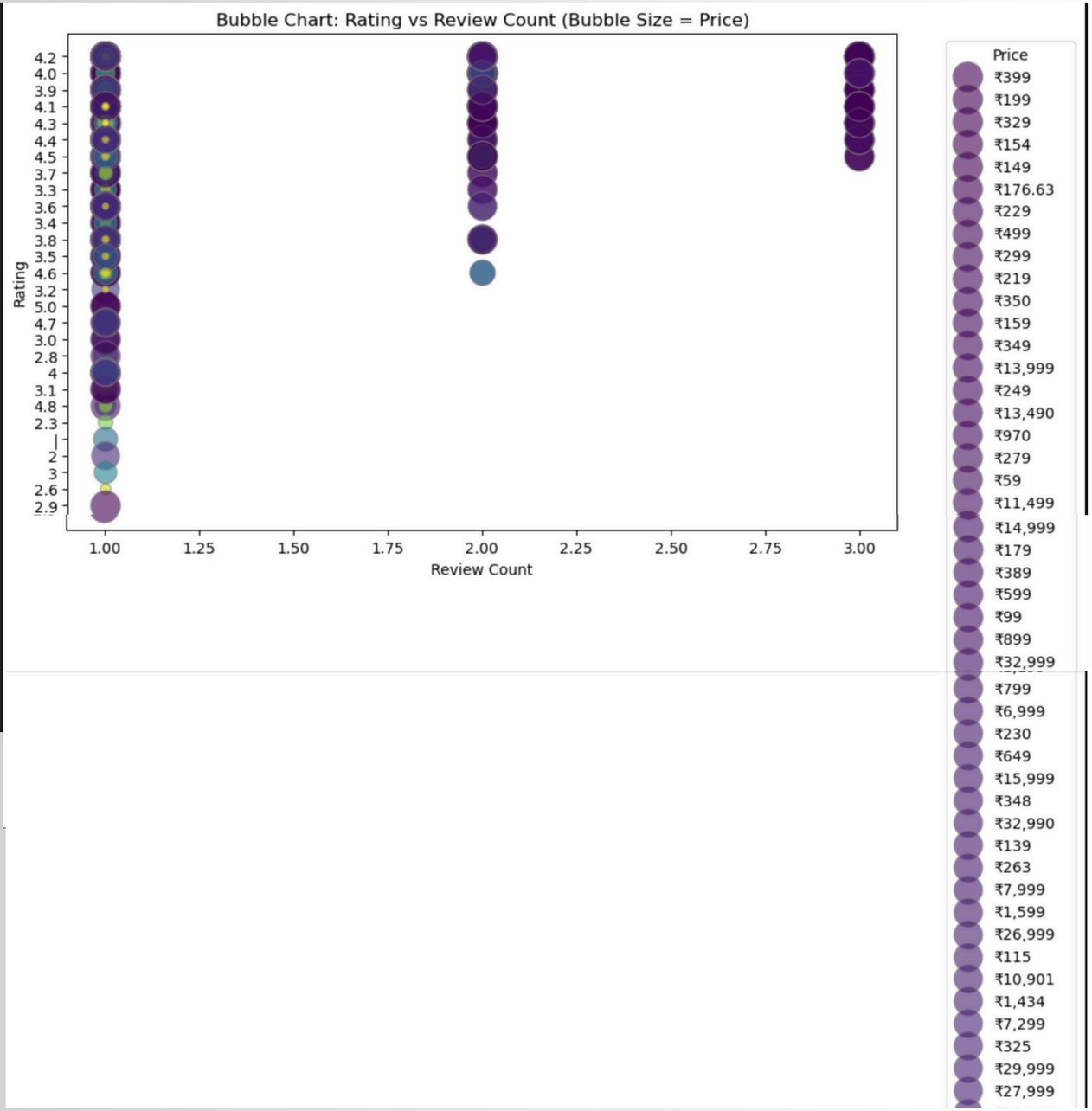
```
plt.figure(figsize=(8, 6))
corr = df.select_dtypes(include='number').corr()
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```

This heatmap displays the correlation coefficient between numeric columns, highlighting key relationships like how price and rating relate.



Rating vs Review count (Bubble Chart)

```python
df_bubble = df[['review_id', 'rating', 'discounted_price']].dropna()
df_bubble['review_count'] = df.groupby('product_id')['review_id'].transform('count')

plt.figure(figsize=(10, 6))
sns.scatterplot(
    data=df_bubble,
    x='review_count',
    y='rating',
    size='discounted_price',
    hue='discounted_price',
    sizes=(20, 400),
    palette='viridis',
    alpha=0.6,
    edgecolor='gray'
)
plt.title("Bubble Chart: Rating vs Review Count (Bubble Size = Price)")
plt.xlabel("Review Count")
plt.ylabel("Rating")
plt.legend(title='Price', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



This bubble chart shows the relationship between *review_count* and *rating*. Each bubble's size and color represent the discounted price influences rating and reviews.

Insights & Business Decisions

1. **Top Product Categories**

Insight: A few product categories dominate sales volume, while others have significantly lower representations.
Business Decision:
- Focus marketing and stock investment on the top performing categories.
- Explore reasons for low performance in other categories (e.g., lack of visibility, pricing, or customer need).
- Consider bundling low-performing items with top-seller to increase visibility.

2. Price vs Rating

Insight: There is no strong visible trend between price and rating, suggesting pricing does not heavily influence rating.
Business Decision:
- Maintain flexible pricing strategies across product tiers without fearing a negative impact on customer satisfaction.
- Consider focusing on product quality or customer experience to improve ratings.

3. Estimated Revenue by Category

Insight: Some categories shows 0 revenue, likely due to either lack of sales or data quality issues.
Business Decision:
- Audit product listings in those categories to fix data gaps (e.g., missing price or errors).
- Reassess whether to continue offering products in consistently zero-revenue categories.

4. Correlation Heatmap

Insight: Strong correlation between *discounted_price* and *Price* as expected but weak correlation between *rating-related* metrics and price variables.
Business Decision:
- Discounting has limited impact on rating. Use price promotions strategically for revenue, not necessarily for reputation.
- Consider using bundles or service-based value to improve ratings instead.

5. Top-Selling Products

Insight: Top-selling products have high review counts ang good ratings, indicating customer trust and popularity.
Business Decision:
- Prioritize these products in featured sections, email campaigns, and ads.
- Use them as benchmarks to improve weaker items (study description, price, review quality, etc.).

6.Rating vs Review count with Price Influence

Insight: Products with more reviews also tend tend to have higher ratings. Price seems to moderately influence both.
Business Decision:
- Encourage customer reviews, especially on higher-priced items, as they often reinforce high ratings.
- Use discounts to boost early traction on new or high-priced products.

7.Correlation with Review count & Rating

Insight: Weak correlation between discount percentage and rating/review count. This supports earlier insights that pricing alone isn't a strong driver of reputation.
Business Decision:
- Focus efforts on product quality, delivery experience, and after-sales service to boost ratings and reviews.

Overall Strategy Recommendations

1. **Double down** on top-performing categories.
2. **Clean up** zero-revenue data and review product listing.
3. **Improve Product Visibility** through reviews rather than just pricing.
4. **Design Campaigns** around high-rated, high-selling items.
5. **Focus on Quality and service** over discounts for improving reputation.

Thank you for exploring my project !
Feel free to leave feedback