BIGBASKET

EXPLORATORY DATA ANALYSIS CAPSTONE PROJECT

INTRODUCTION

In the ever-evolving landscape of online grocery shopping, BigBasket has emerged as a dominant player, offering a vast and diverse array of products to cater to the needs of modern consumers. Understanding customer preferences, sales trends, and product performance is paramount for Big Basket's continued success and market leadership.

This project delves into a comprehensive analysis of Big Basket's product data, aiming to uncover valuable insights that can drive strategic decision-making and optimize business operations. By examining key metrics such as product categories, brands, pricing, discounts, and ratings, we seek to illuminate the dynamics of Big Basket's product portfolio and its impact on customer behavior.

OBJECTIVE

1.Identify Top-Performing Products and Categories:

- Assess the impact of discounts on sales volume and customer purchase behavior.
- Analyze the relationship between discount levels and product ratings.
- Determine the optimal discount strategies to maximize sales and customer engagement.

2.Evaluate Discount Effectiveness:

- Determine the products and categories that generate the highest sales and revenue for BigBasket.
- Analyze sales trends and identify potential growth opportunities within specific product segments.
- Uncover customer preferences for different product types and categories.

OBJECTIVE

3.Understand Customer Preferences and Brand Loyalty:

- Identify the most popular brands among BigBasket customers.
- Analyze customer preferences for specific brands within different product categories.
- Evaluate brand loyalty and customer retention rates.

4.Gain Insights for Strategic Decision-Making:

- Provide actionable insights to inform pricing strategies, inventory management, and marketing campaigns.
- Identify opportunities to optimize product offerings and improve customer experiences.
- Support data-driven decision-making to enhance Big Basket's business performance and market competitiveness.

PYTHON IMPORT LYBRARY

NumPy

NumPy (Numerical Python) is a powerful library for working with arrays and numerical data.
It is used for mathematical operations, statistical analysis, and handling large datasets efficiently.

Pandas

Pandas is a Python library used for data analysis and manipulation.
It is mainly used for working with structured data like tables, spreadsheets, or databases.
It provides Data Frames, which are like Excel tables but with much more flexibility and power.

Seaborn

Seaborn is a data visualization library used to create beautiful and professional charts.
It allows you to create Bar Charts, Line Charts, Scatter Plots, Box Plots, and more.
Seaborn makes data representation easier to understand.

Matplotlib

Matplotlib is another charting library but gives more customization options.
It is often used with Seaborn to modify or save graphs.
Seaborn internally uses Matplotlib to generate graphs.

LOADING DATASET

Loading Dataset

```
file_path = '/content/drive/MyDrive/bigbasket practice/Copy of BigBasket Products.csv'
```

I've load dataset named ‘Bigbasket Products.csv’ into a variable name ‘df’ in order to import the datasets, I’ve used pandas Libraries in which I’ve used ‘pd.read_csv’ command to import the respective datasets.

```
df.columns
```

```
Index(['index', 'product', 'category', 'sub_category', 'brand', 'sale_price',
       'market_price', 'type', 'rating', 'description', 'discount'],
      dtype='object')
```

I’ve accessed and displayed the column names of the loaded dataset using columns. These columns encompass various aspects of ‘Indian restaurants’ information, providing detailed overview of the dataset.

DATA OVERVIEW

DATA OVERVIEW

```
df.describe()
```

	index	sale_price	market_price	rating
count	27555.00000	27549.000000	27555.000000	18919.000000
mean	13778.00000	334.648391	382.056664	3.943295
std	7954.58767	1202.102113	581.730717	0.739217
min	1.00000	2.450000	3.000000	1.000000
25%	6889.50000	95.000000	100.000000	3.700000
50%	13778.00000	190.320000	220.000000	4.100000
75%	20666.50000	359.000000	425.000000	4.300000
max	27555.00000	112475.000000	12500.000000	5.000000

I've generated descriptive static for the loaded dataset stored in the variable 'df' using the describe method. This Pandas function provides a summary of statical measure , minimum , 25th percentile, median(50th percentile), 75th percentile and maximum values for each numeric column in the dataset . This summary aids in in understanding the central tendency ,dispersion , and distribution of the numeric feature in the dataset.

```
df.shape
(27555, 10)
```

df. Shape , revealing that it consists of 27555 rows and 10column

DATA OVERVIEW

CHECKING INFORMATION

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27555 entries, 0 to 27554
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   index          27555 non-null  int64  
 1   product        27554 non-null  object 
 2   category       27555 non-null  object 
 3   sub_category   27555 non-null  object 
 4   brand          27554 non-null  object 
 5   sale_price     27549 non-null  float64
 6   market_price   27555 non-null  float64
 7   type           27555 non-null  object 
 8   rating         18919 non-null  float64
 9   description    27440 non-null  object 
dtypes: float64(3), int64(1), object(6)
memory usage: 2.1+ MB
```

I've obtained information about the dataset using df.info(), This method provide a concise summary , including the total number of entries ,the data types of each column ,and the count of non-nun values

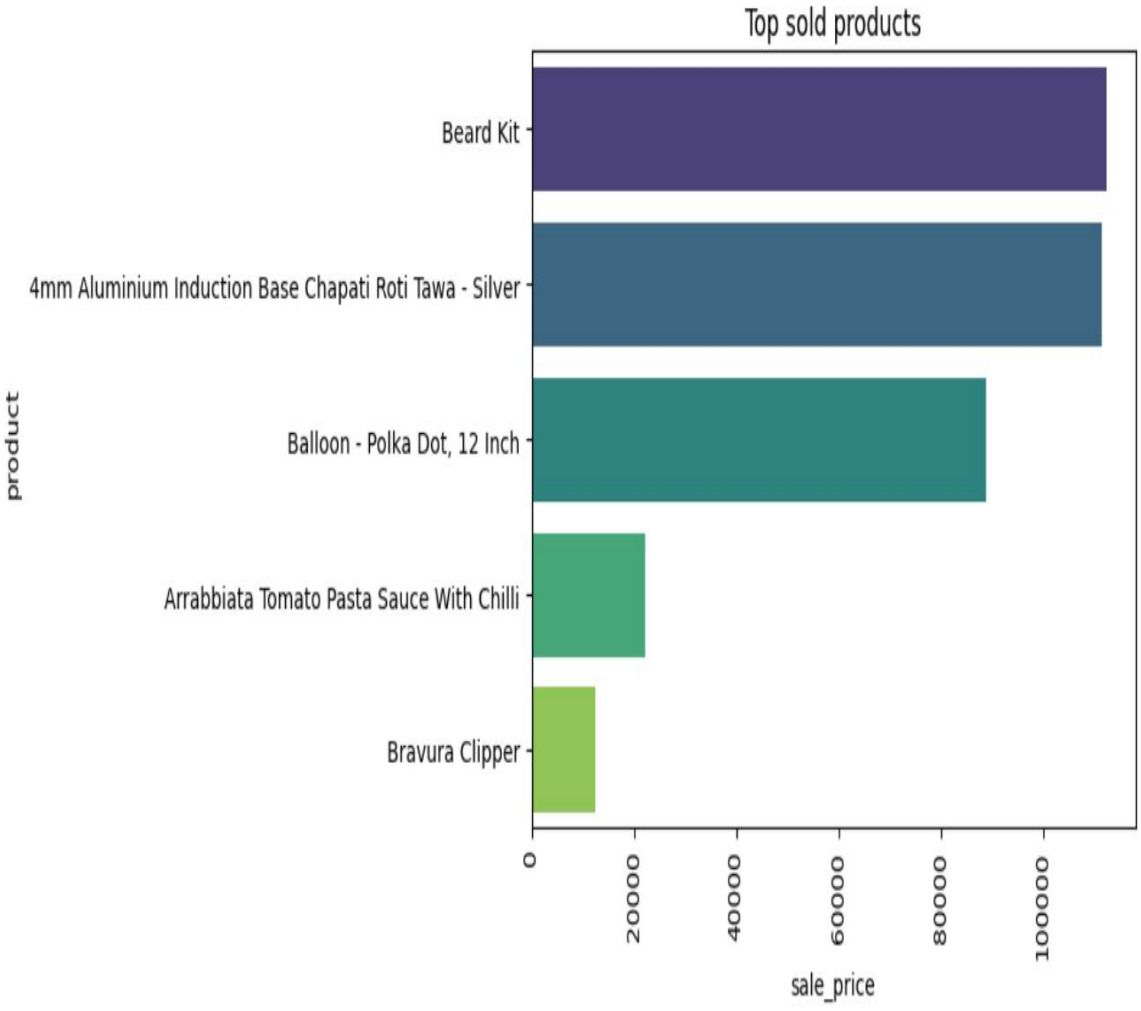
```
df.duplicated().sum()
```

```
0
```

Check my dataset no duplicate value present

DATA VISUALIIZATION

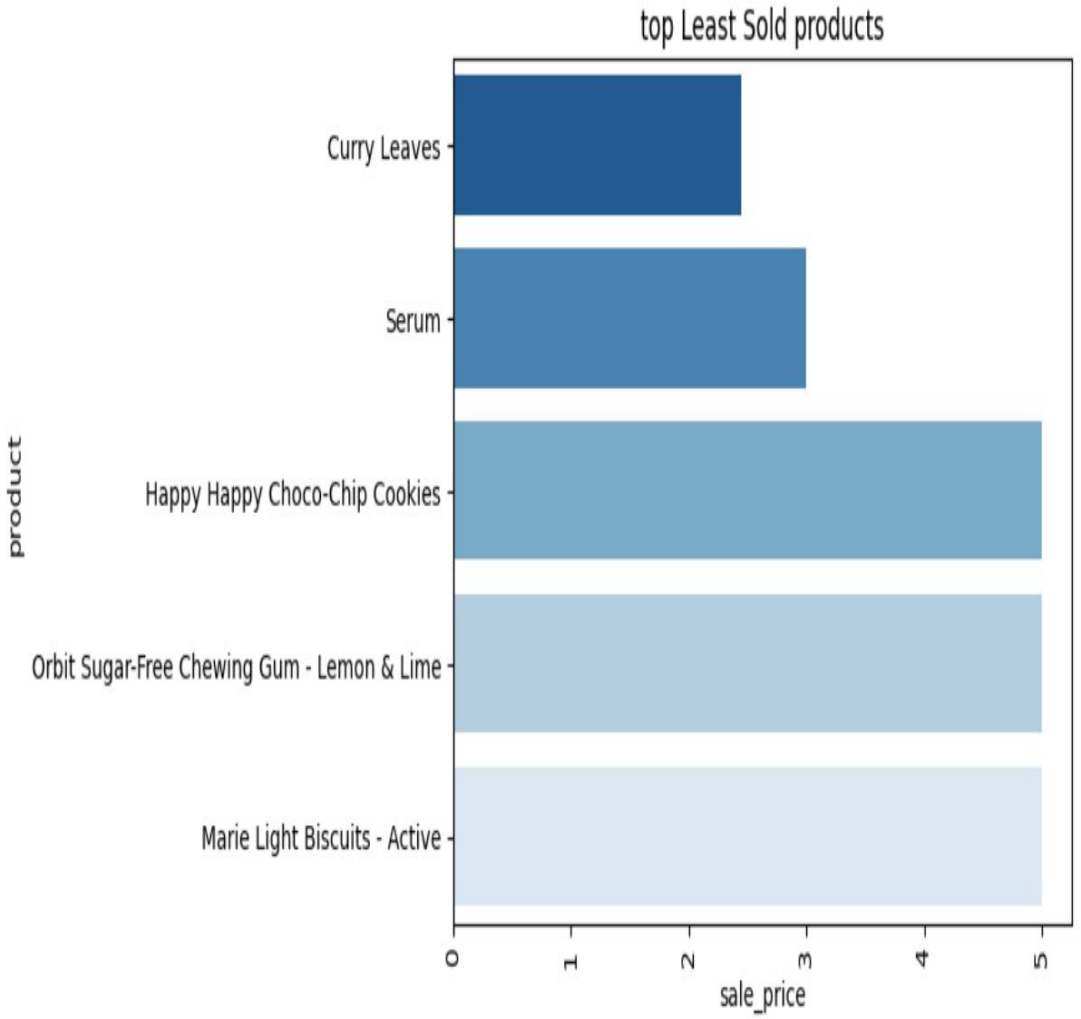
Top Sold products



- **Top-selling product:** Beard Kit
- **Other popular products:** Chapati Tawa, Balloons
- **Low-selling products:** Pasta Sauce, Clipper

DATA VISUALIZATION

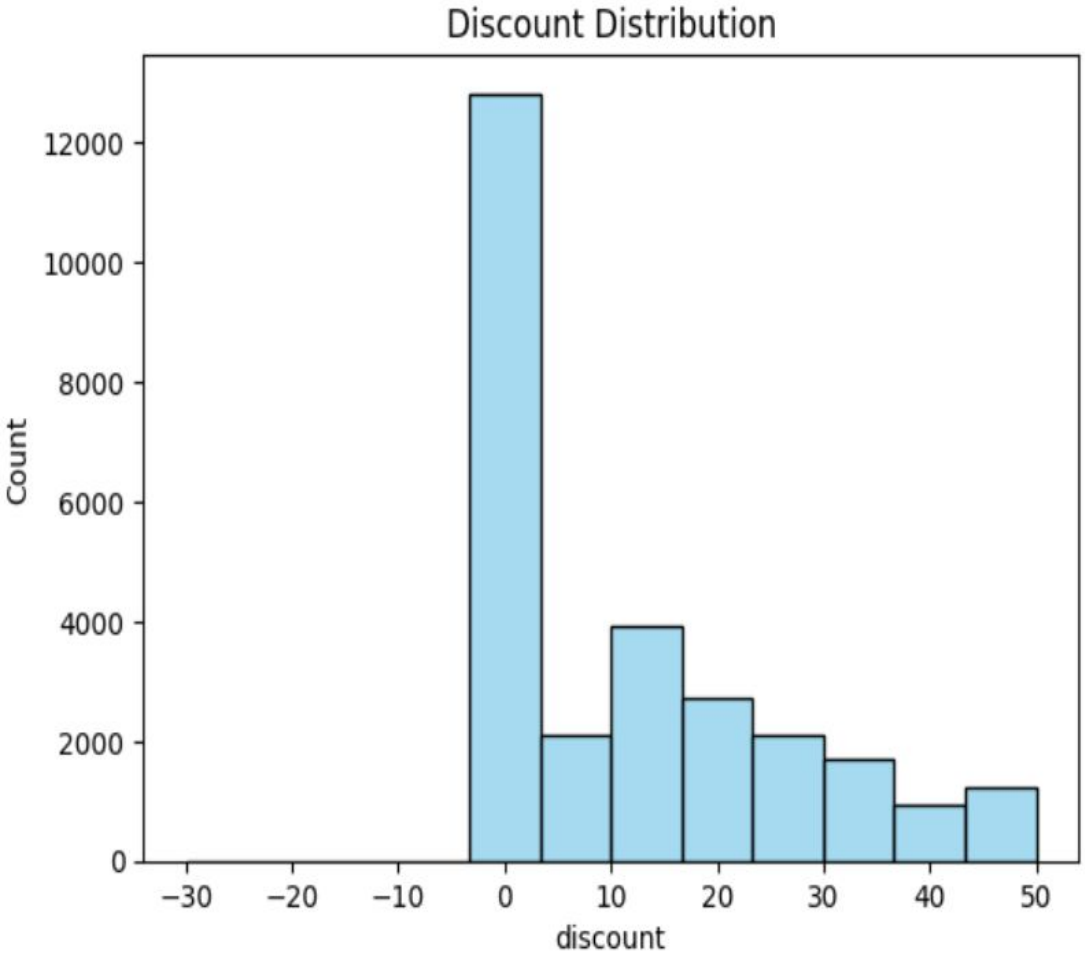
Top Least Sold products



- **Least sold product:** Curry Leaves
- **Other low-selling products:** Serum, Cookies, Chewing Gum, Biscuits

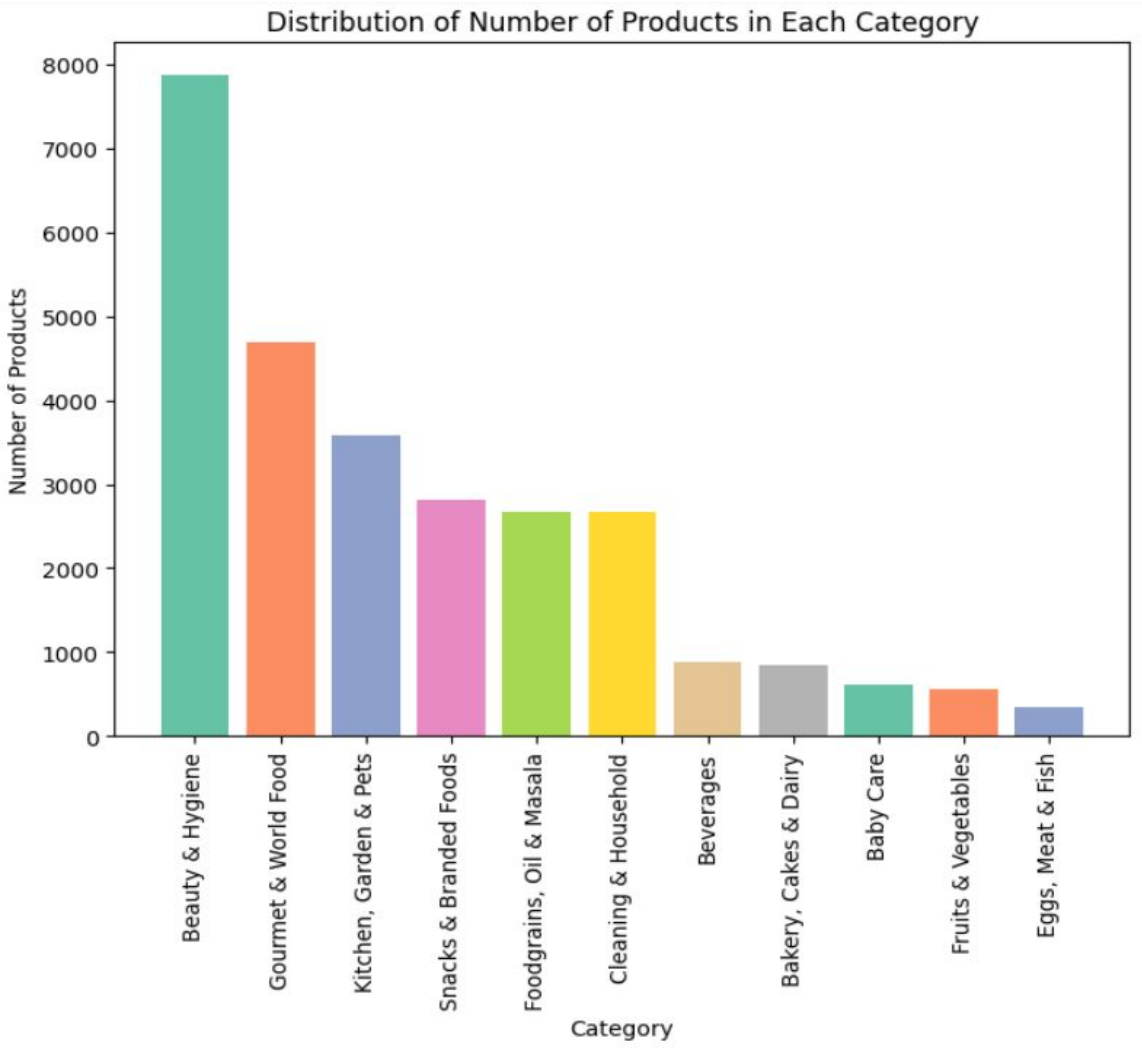
DATA VISUALIZATION

Discount distribution



- **Most discounts** are around 0%.
- **Higher discounts (10-50%)** have lower frequency.
- **Negative discounts** (returns or errors) exist but are rare.

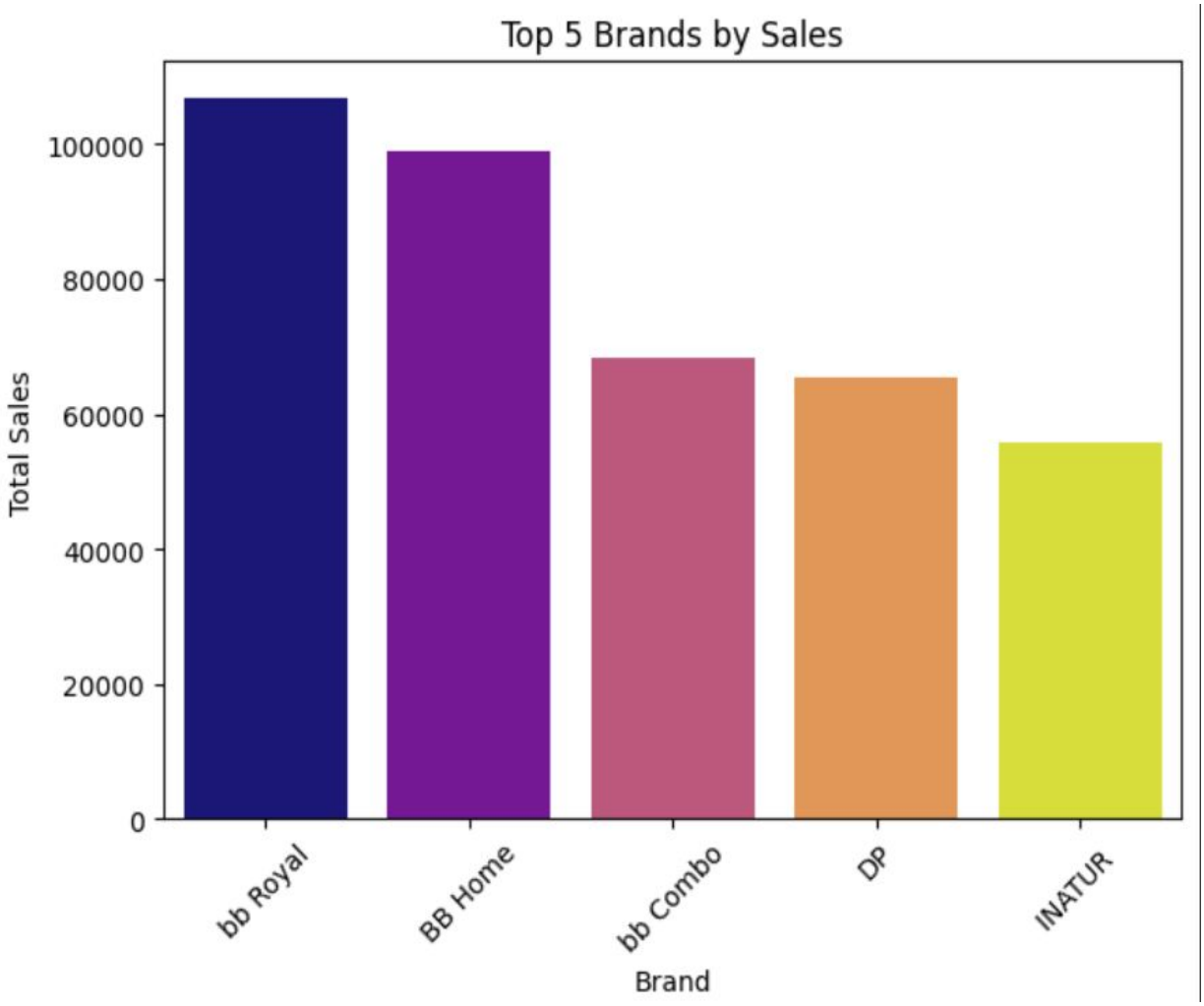
Distribution of Number of product of Each Category



- **Beauty & Hygiene dominates** with the highest number of products.
- **Gourmet & World Food is second,** showing strong variety.
- **Eggs, Meat & Fish have the least,** indicating limited options.

DATA VISUAZATION

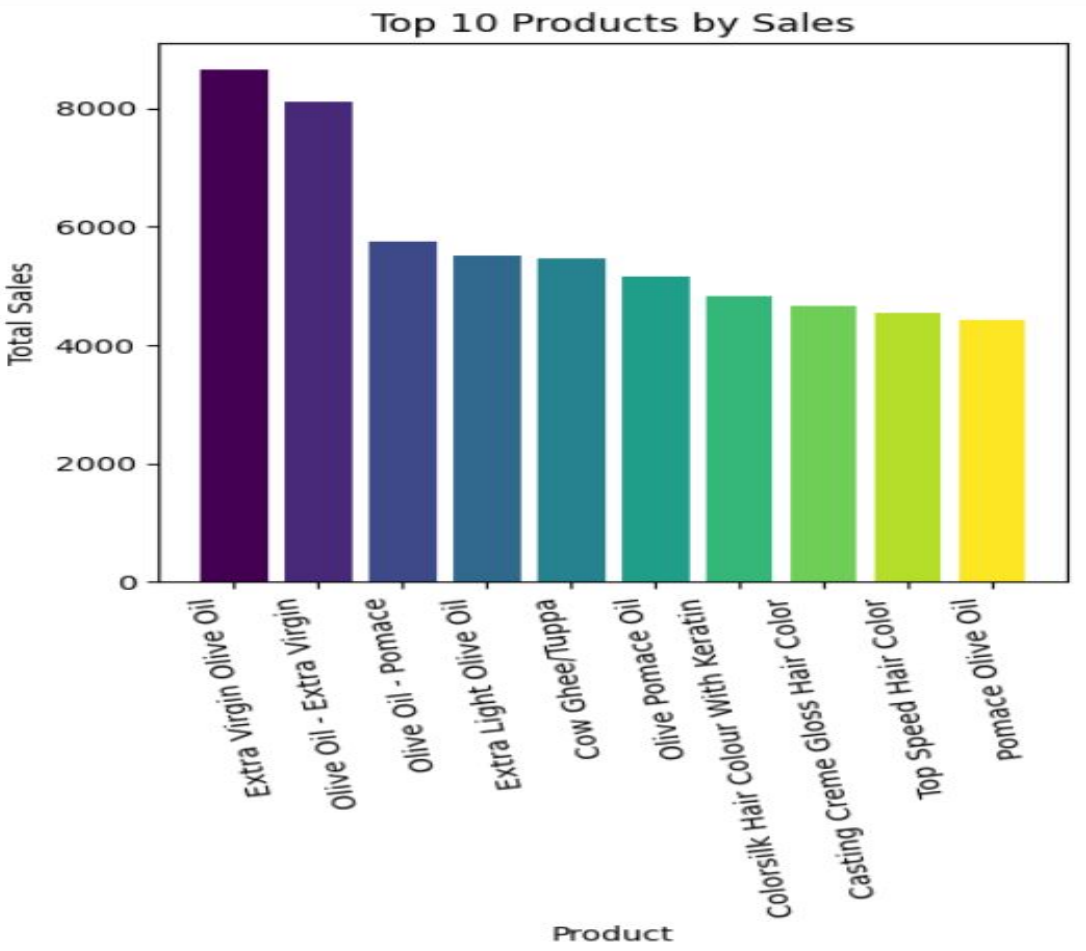
Top 5 Brands by Sale



- **bb Royal** leads in sales, followed closely by **BB Home**.
- **bb Combo** and **DP** have moderate sales.
- **INATUR** has the lowest sales among the top 5.
- A significant gap exists between the top 2 and the rest.

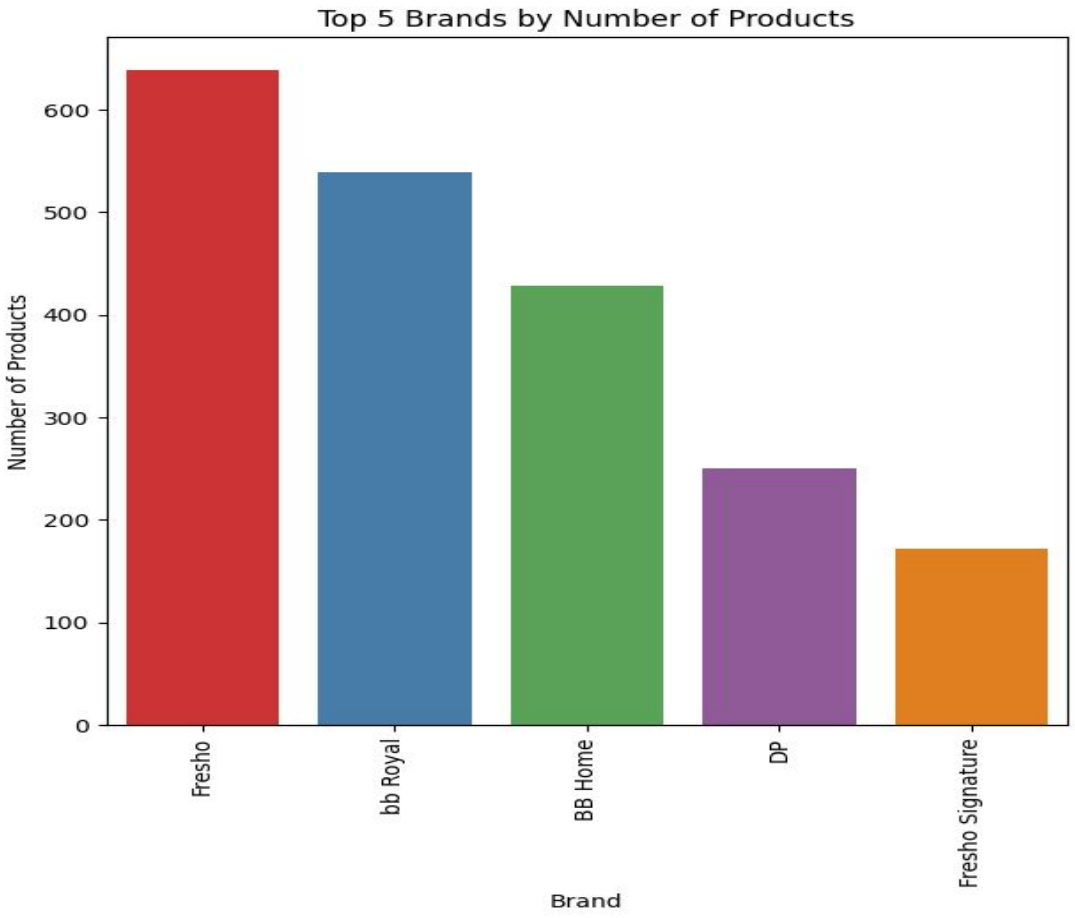
DATA VISUALIZATION

Top 10 products BY Sales



- **Extra Virgin Olive Oil** has the **highest sales**.
- **Olive Oil & Ghee** dominate the top-selling products.
- **Hair color products** also perform well.
- **Pomace Olive Oil** has the lowest sales in the top 10.
- **Oil products are leading in demand**.

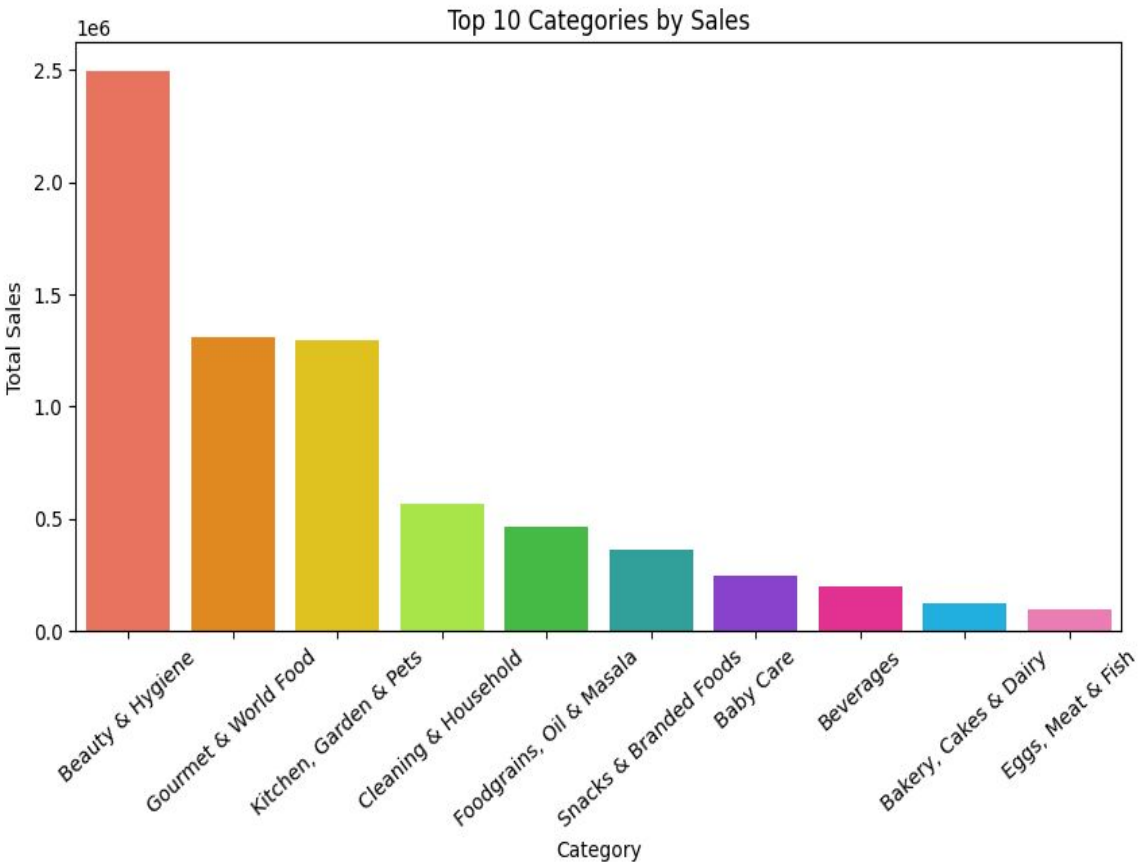
DATA VISUALIZATION



- **Fresho** has the highest number of products (600+).
- **bb Royal** is in second place (500+ products).
- **BB Home** ranks third (400+ products).
- **DP** and **Fresho Signature** have relatively fewer products.
- Overall, **Fresho, bb Royal, and BB Home** are the top 3 brands.

DATA VISUALIZATION

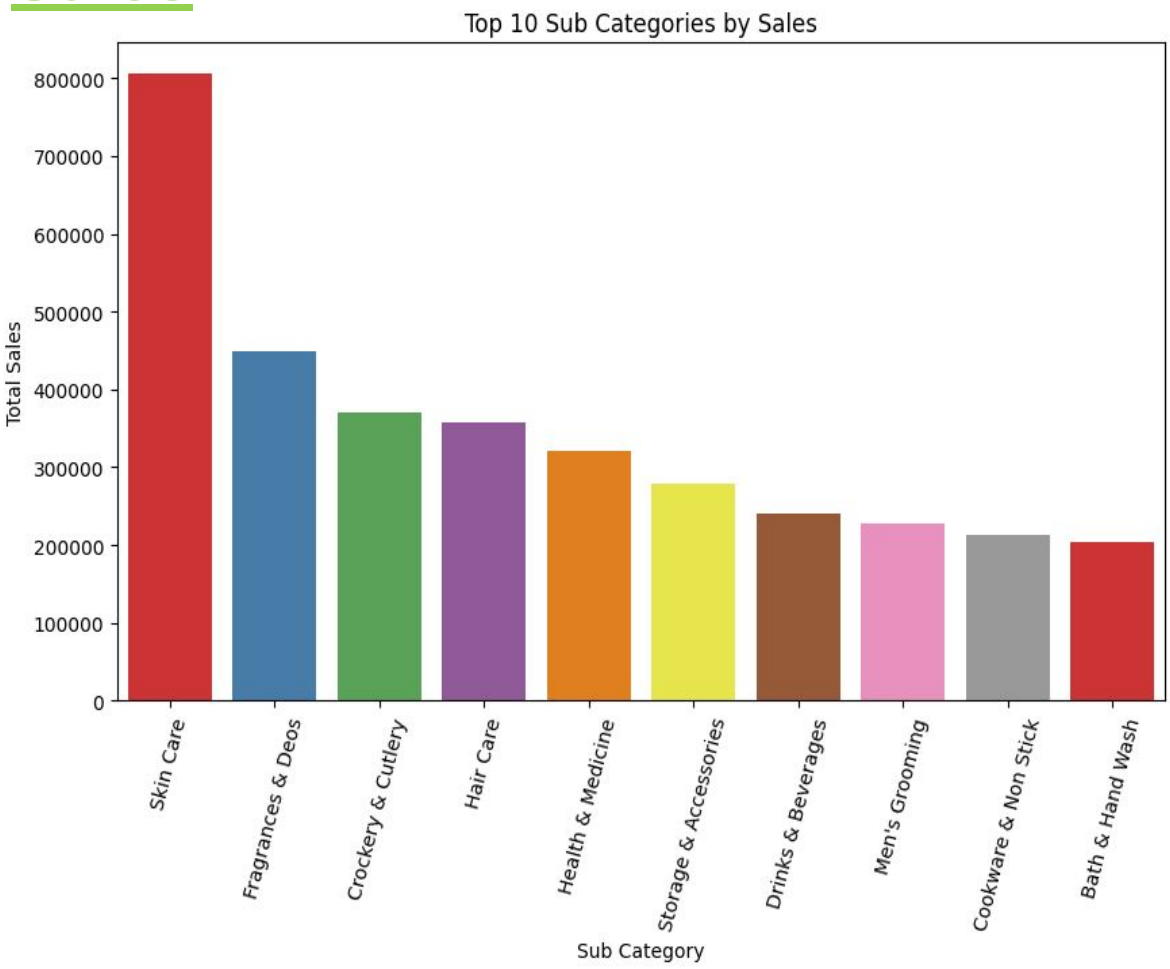
Top 10 Categories BY Sales



- **Beauty & Hygiene** has **highest sales.**
- **Gourmet & World Food** and **Kitchen, Garden & Pets** follow closely.
- **Cleaning & Household** performs moderately.
- **Eggs, Meat & Fish** have **lowest sales.**
- **Strategic promotions** can boost low-performing categories.

DATA VISUALIZATION

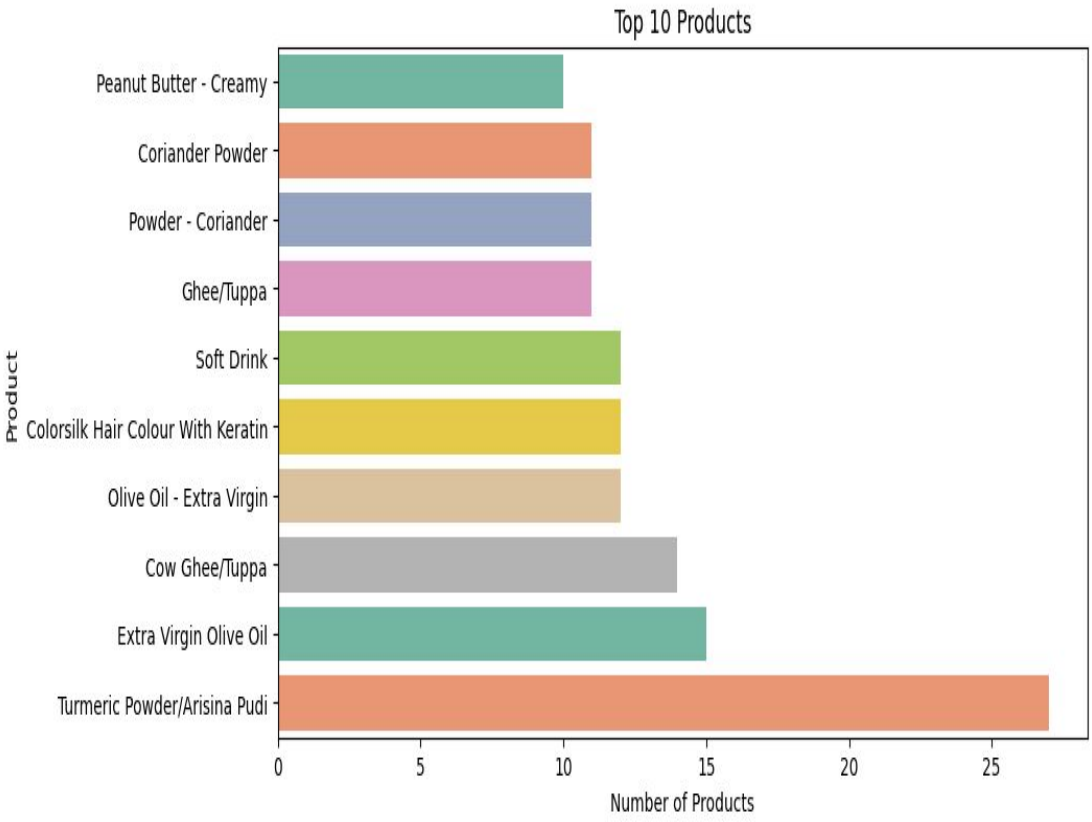
Top 10 Sub Categories BY Sales



- **Skin Care** has the highest sales (800K+).
- **Fragrances & Does** is the second top-selling sub-category.
- **Crockery & Cutlery** and **Hair Care** have similar sales.
- **Health & Medicine** follows closely behind.
- **Bath & Hand Wash** has the lowest sales among the top 10 sub-categories.

DATA VISUALIZATION

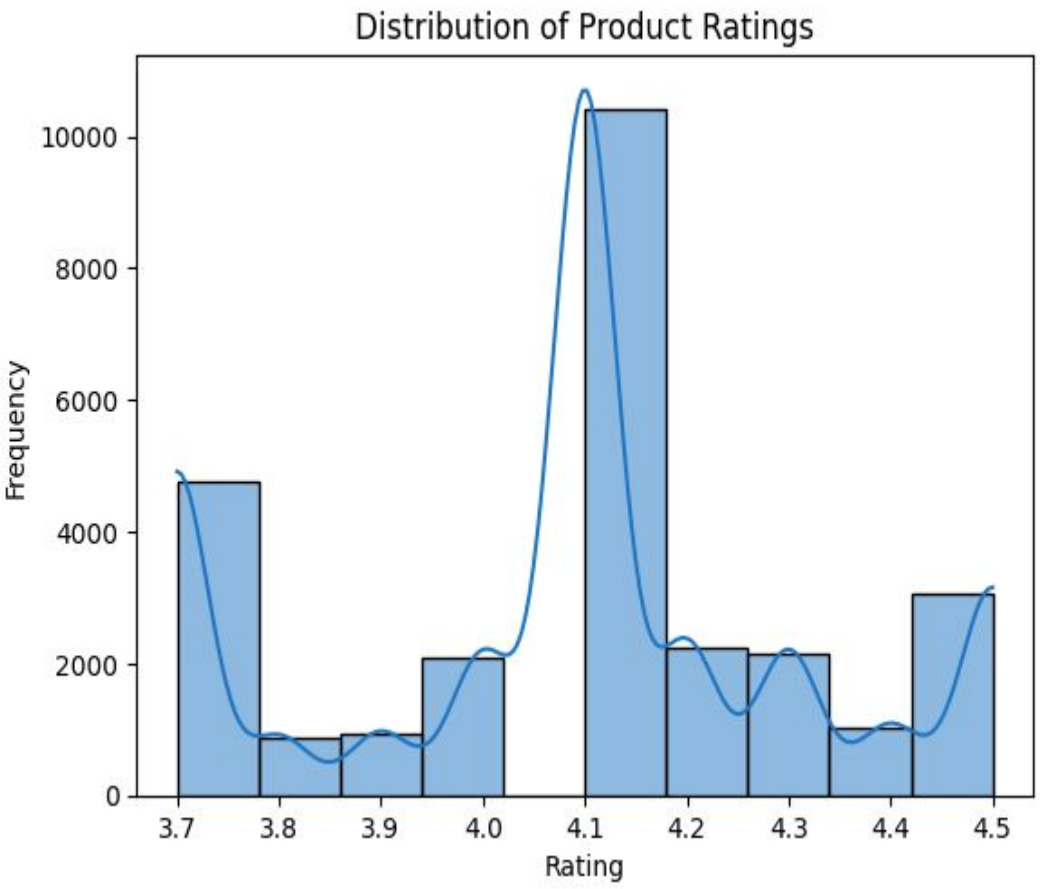
Top 10 products



- **Turmeric Powder** has the highest count.
- **Extra Virgin Olive Oil** and **Cow Ghee** are also in high numbers.
- Other products have relatively lower counts.
- The graph clearly indicates that some products are in higher demand.

DATA VISUALIZATION

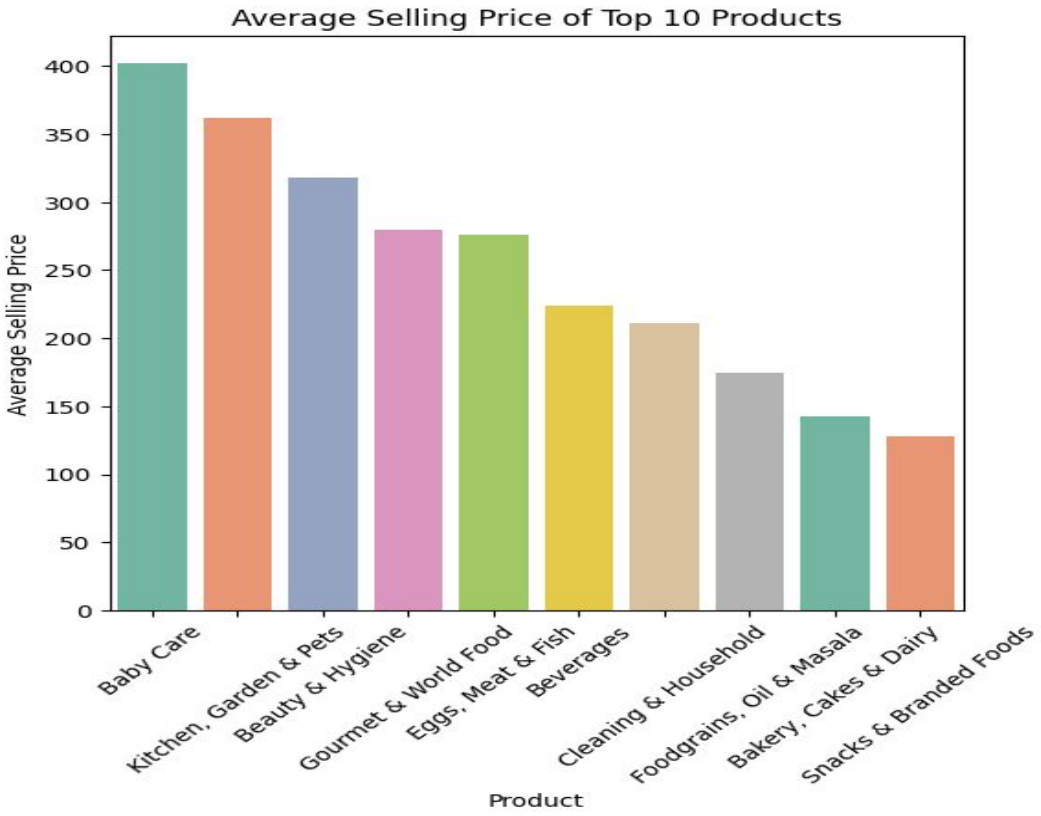
Distribution of Product Ratings



- **Most Ratings:** Around 4.1 is the peak.
- **Low Ratings:** Fewer below 3.9.
- **High Ratings:** Some near 4.5.
- **Uneven Spread:** Gaps in frequency.
- **Peak & Dips:** Sharp rise at 4.1, dips around 4.2.

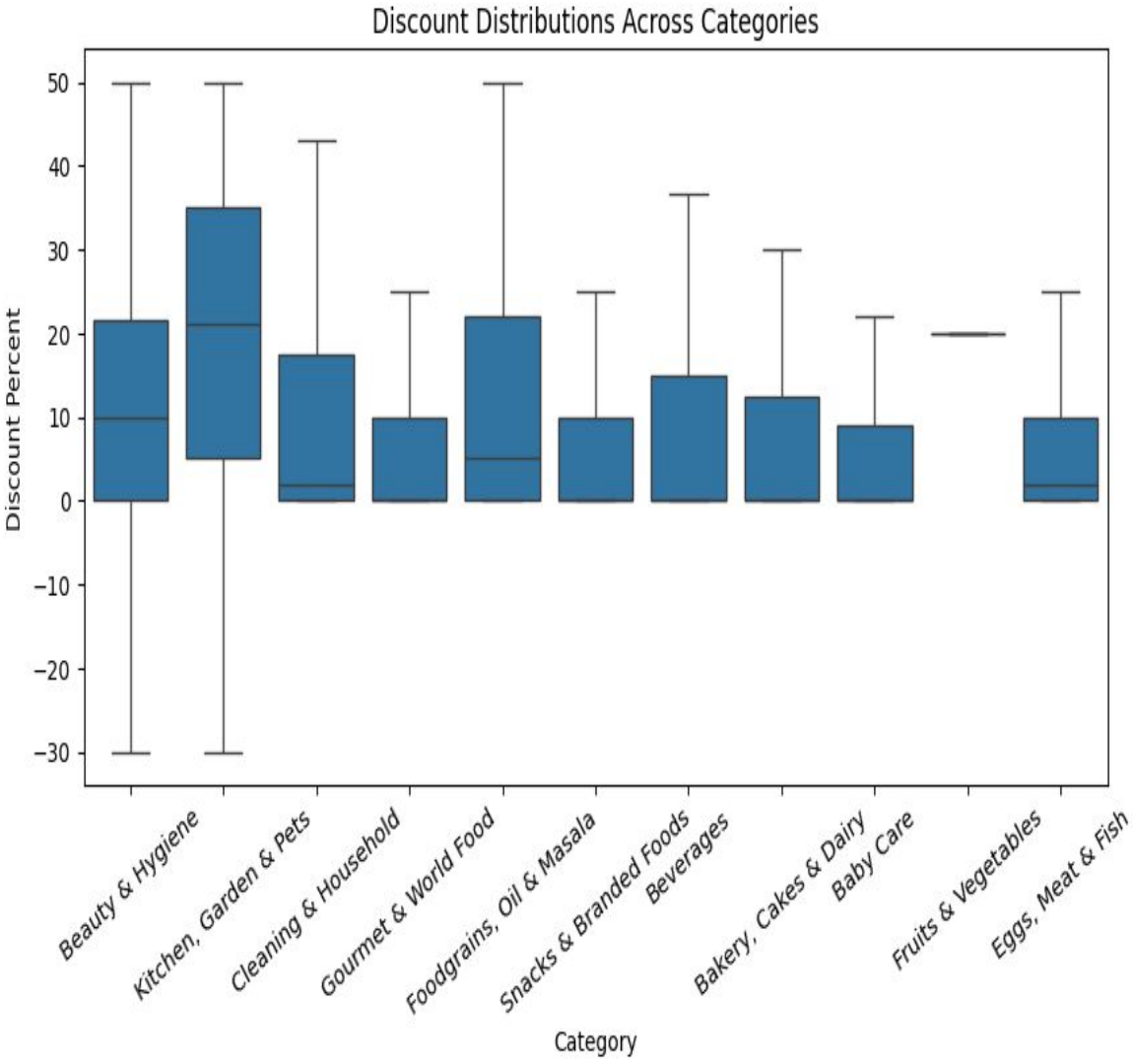
DATA VISUALIZATION

Average Selling Price of Top 10 Products



- **Highest Price:** Baby Care (~400).
- **Next Expensive:** Kitchen & Beauty items.
- **Mid-Range:** Eggs, Beverages, Cleaning.
- **Lowest Price:** Snacks & Bakery.
- **Clear Price Drop:** From left to right.

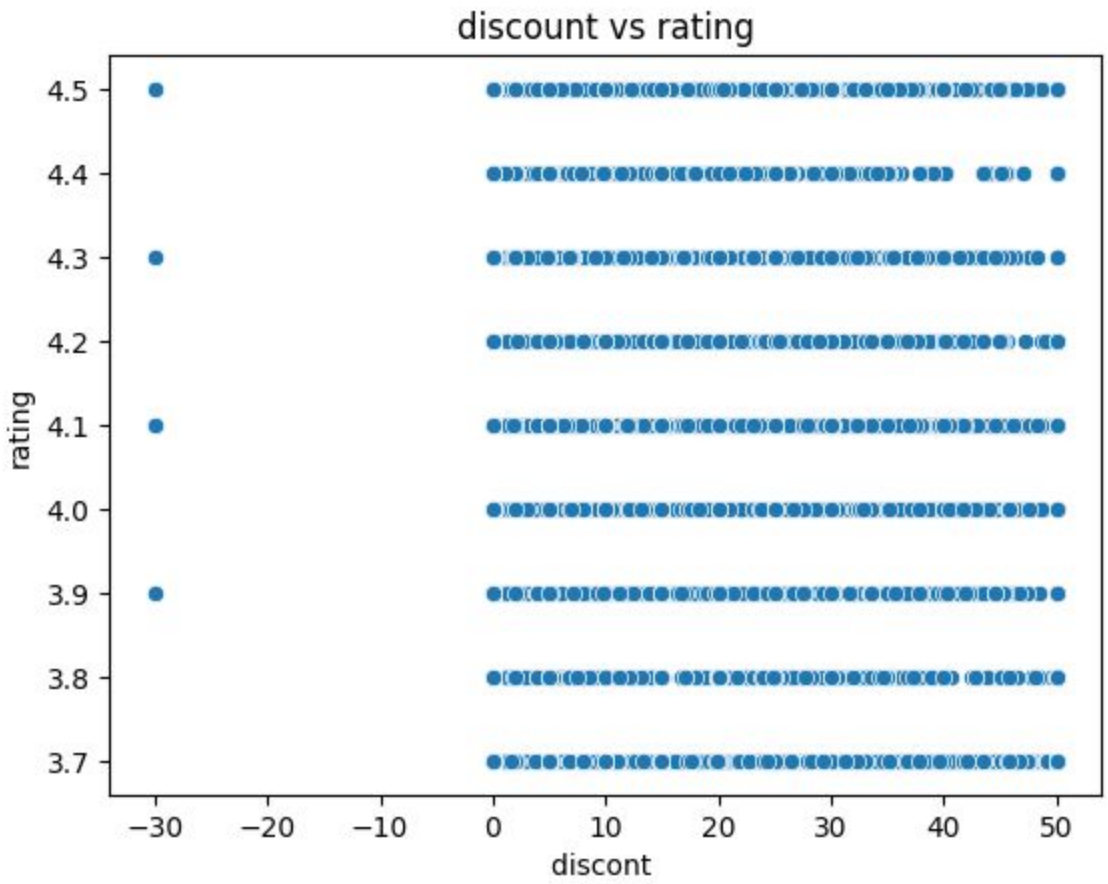
DATA VISUALIZATION



- **Kitchen, Garden & Pets** has the highest discount range.
- **Beauty & Hygiene** and **Gourmet & World Food** also offer high discounts.
- **Fruits & Vegetables** have the least discount variation.

DATA VISUALIZATION

Discount vs Rating



- **No Clear Trend -** Discounts don't strongly impact ratings.
- **Ratings Clustered -** Most ratings range from 3.7 to 4.5.
- **Negative Discounts Exist** - Some discounts are below zero.
- **Higher Discounts Are Common -** Many products have 10-50% discounts.
- **Outliers Present -** Some extreme discounts with varying ratings.

RECOMMANDATIONS

1.Sales Optimization

- **Beauty & Hygiene** has the highest sales, invest more in marketing and stock.
- **Gourmet & World Food** is performing well, expand product variety.
- **Low Sales Categories** like Meat, Bakery, and Beverages need discounts or promotions.

2. Pricing & Discounts

- **Higher Discount = More Sales**, focus on categories benefiting most.
- **Optimize Discounts** based on category performance (e.g., Foodgrains need better discounts).

3. Brand Strategy

- **Olive Oils & Hair Colors** are top-selling, expand similar products.
- **Check Customer Feedback** on least sold items to improve

RECOMMANDATIONS

4. Customer Preferences

- **Olive Oils & Hair Colors** are top-selling, expand similar products.
- **Check Customer Feedback** on least sold items to improve.

5. Rating & Discount Relation

- **Higher Discounts Get Higher Ratings,** use this insight for product pricing.
- **Analyze Low-Rated Products,** improve quality or offer better deals.

6. Category Expansion

- **Kitchen & Household** is performing well, expand this range.
- **Introduce New Products** in low-performing categories with high-potential demand.

7. Stock & Inventory Management

- **Top 10 products should be always in stock** to avoid losing customers.
- **Low-demand products should be limited** to reduce wastage.

CONCLUSION

- **Beauty & Hygiene leads sales –** This category has the highest revenue, indicating strong consumer demand.
- **Gourmet & World Food and Kitchen Essentials perform well –** These categories also generate high sales, suggesting a focus on premium and daily-use items.
- **Foodgrains, Oil & Masala show moderate sales –** These essentials perform decently but have scope for growth.
- **Beverages, Bakery, and Meat categories need attention –** Sales in these categories are lower, requiring better marketing and product variety.
- **Strategic discounts can drive sales –** Products with higher discounts tend to perform better, emphasizing the importance of price strategies.

THANX FOR READING

FOR CODE