bigbasket

Exploratory Data Analysis

DIVYANSHI NIGAM

INTRODUCTION

E-commerce (electronic commerce) is the activity of electronically buying or selling of products on online services or over the Internet. E- commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems.

E-commerce, particularly online grocery shopping, is experiencing significant growth, driven by convenience, time-saving, and the increasing popularity of online platforms.

Ecommerce grocery is defined as the business of selling groceries online. It concentrates on customer value, the convenience of shopping, and also potentially quick and easy delivery.

About Big Basket

BigBasket, India's largest online grocery supermarket, was founded in 2011 and is headquartered in Bangalore. Acquired by Tata Digital in 2021, it has revolutionized grocery shopping by offering a seamless online experience through its website and mobile app (available on Android & iOS).

Serving over **10 million customers across 30+ cities**, BigBasket provides a vast selection of products, including **fruits, vegetables, dairy, beverages, snacks, personal care, and household essentials**, all delivered to customers' doorsteps. It was the **first online grocer in India** and gained further popularity with **Shahrukh Khan as its brand ambassador**.

Despite increasing competition from platforms like **Blinkit and JioMart**, BigBasket has maintained its dominance by leveraging its expansive customer base and smooth transition to online retail.

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is a method used to analyze datasets by summarizing their key characteristics through statistical graphics and visualization techniques.

- In this project, we systematically approach EDA by **loading data, generating descriptive statistics, profiling data, detecting anomalies, and applying visualization methods**. The dataset, contains sales dynamics and product offerings from **BigBasket**, providing insights into **operational metrics, product popularity, pricing strategies, and customer feedback**.

- Through **outlier detection, trend analysis, and visualizations**, EDA helps uncover valuable patterns that can drive **business growth, optimize inventory management, and enhance customer experience**, ultimately strengthening **BigBasket's leadership in India's online grocery market**.

Data dictionary

This dataset contains 10 attributes with simple meaning and which are described as follows:

1. index - Simply the Index as a unique identifier for each entry in the dataset
2. product - Title of the product (as they're listed) that is the title or name of the products listed on the Big Basket platform.
3. category - Category into which product has been classified into broader categories, such as fruits, vegetables, dairy products, beverages, etc.
4. sub_category- Subcategory into which product has been kept

Data dictionary

This dataset contains 10 attributes with simple meaning and which are described as follows:

5. brand - Brand of the product or manufacturer associated with each product.
6. sale_price - Price at which product is being sold on the site to the consumers
7. market_price – Standard market price of the product
8. type - Type into which product falls based on their nature or characteristics
9. rating - Rating the product has got from its consumers
10. description - Description of the dataset in detail

Tools

Development Environment	Libraries Used	Language used
Google Colab Notebook	Numpy Pandas Matplotlib Seaborn.	Python

Project Workflow

Step 1: Installation and Importing of Libraries

Step 2: Data Collection and Loading

Step 3: Basic Data Inspection

Step 4: Data Cleaning and Pre Processing

Step 5: Exploratory Data Analysis and Visualization

Step 6. Findings

Step 7:Recommendations

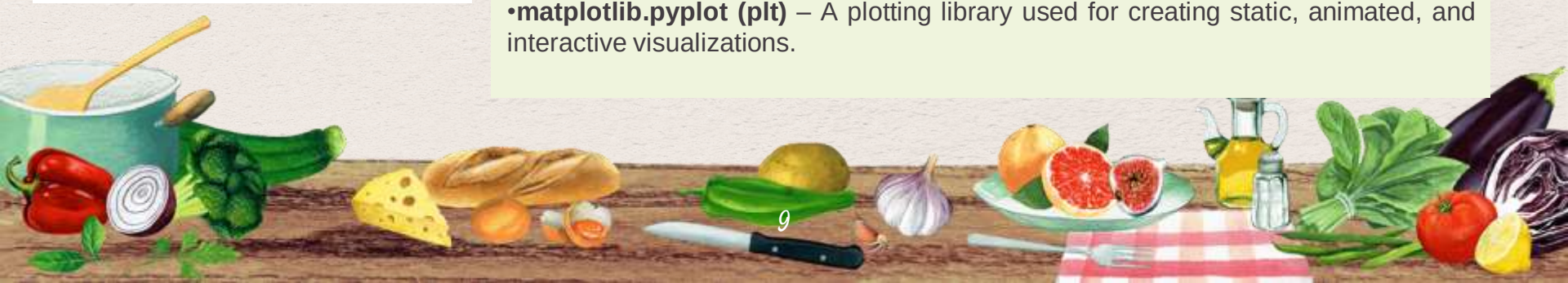
Step 1: Installation and Importing of Libraries

```
[3]  !pip install Numpy
     !pip install Pandas
     !pip install Seaborn
     !pip install Matplotlib
```

```
Requirement already satisfied: Numpy in /usr/local/lib/python3.11/dist-packages (2.0.2)
Requirement already satisfied: Pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-packages (from Pandas) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from Pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from Pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from Pandas) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->Pandas) (1.17.0)
```

```
[1]  ## Importing libraries
     import numpy as np
     import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
```

- **numpy (np)** – Used for numerical computations and handling arrays efficiently.
- **pandas (pd)** – Provides data manipulation and analysis tools, mainly for working with DataFrames.
- **seaborn (sns)** – Enhances data visualization by creating aesthetically pleasing statistical graphics.
- **matplotlib.pyplot (plt)** – A plotting library used for creating static, animated, and interactive visualizations.



Step 2: Data Collection and Loading

I have used **Google Colab** as my development environment for this project, as it provides a cloud-based Jupyter Notebook with powerful computing capabilities. To access files stored in Google Drive, I used the following command:

```
from google.colab import drive
drive.mount('/content/drive')
```

This command mounts Google Drive to the Colab environment, allowing seamless access to datasets stored in Drive. After mounting, I loaded the dataset into a **pandas DataFrame** for further analysis.

```
# loading dataset

data = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/EDA DATASETS/BigBasket Products.csv")
```

Step 3: Basic Data Insepection

Step 2: Use head function to look for first 12 rows

```
[6] data.head(12)
```

	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description
0	1	Garlic Oil - Vegetarian Capsule 500 mg	Beauty & Hygiene	Hair Care	Sri Sri Ayurveda	220.0	220.0	Hair Oil & Serum	4.1	This Product contains Garlic Oil that is known...
1	2	Water Bottle - Orange	Kitchen, Garden & Pets	Storage & Accessories	Mastercook	180.0	180.0	Water & Fridge Bottles	2.3	Each product is microwave safe (without lid), ...
2	3	Brass Angle Deep - Plain, No.2	Cleaning & Household	Pooja Needs	Trm	119.0	250.0	Lamp & Lamp Oil	3.4	A perfect gift for all occasions, be it your m...
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	Cleaning & Household	Bins & Bathroom Ware	Nakoda	149.0	176.0	Laundry, Storage Baskets	3.7	Multipurpose container with an attractive desi...
4	5	Creme Soft Soap - For Hands & Body	Beauty & Hygiene	Bath & Hand Wash	Nivea	162.0	162.0	Bathing Bars & Soaps	4.4	Nivea Creme Soft Soap gives your skin the best...
5	6	Germ - Removal Multipurpose Wipes	Cleaning & Household	All Purpose Cleaners	Nature Protect	169.0	199.0	Disinfectant Spray & Cleaners	3.3	Stay protected from contamination with Multipu...
6	7	Multani Mati	Beauty & Hygiene	Skin Care	Satinance	58.0	58.0	Face Care	3.6	Satinance multani matti is an excellent skin t...

Step 3: Basic Data Insepection

```
rows=data.shape[0]
columns=data.shape[1]
print(f"This data has {rows} and {columns}")
```

```
This data has 27555 and 10
```

```
[10] data.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
```

The dataset comprises of 27555 rows and 10 columns.

- **Categorical Columns:** product, category, sub_category, brand, type, description (Text-based).
- **Numerical Columns:** sale_price, market_price, rating (Float), index (Integer).

- **1 missing value** in the **product** column.
- **1 missing value** in the **brand** column.
- **6 missing values** in the **sale_price** column.
- **8,636 missing values** in the **rating** column (significant missing data).
- **115 missing values** in the **description** column.

Step 3: Basic Data Insepection

```
#summary statistic of data, by default includes all numerical columns
data.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

It gives us summary statistic of data set, by default return for only numerical columns and if we want for all columns including categorical, we can:

Data.describe(include="all")

Or can specify datatypes of preferred columns we want.

- Findings: We can see that ratings ranges from minimum 1 and to maximum 5.
- Minimum sale price of item is INR 2.45 and Maximum INR 112475(considering INR)

Step 4: Data Cleaning and Pre processing

index	0
product	1
category	0
sub_category	0
brand	1
sale_price	0
market_price	0
type	0
rating	8500
description	110
Discount_amount	0
Discount_percentage	0

dtype: int64

1. HANDLING MISSING VALUES

•If a numerical column **does not** have significant outliers, the **mean** can be used.

•However, in this case, since price and rating data may contain outliers (e.g., unusually high/low prices or extreme ratings), we use the **median** to avoid bias.

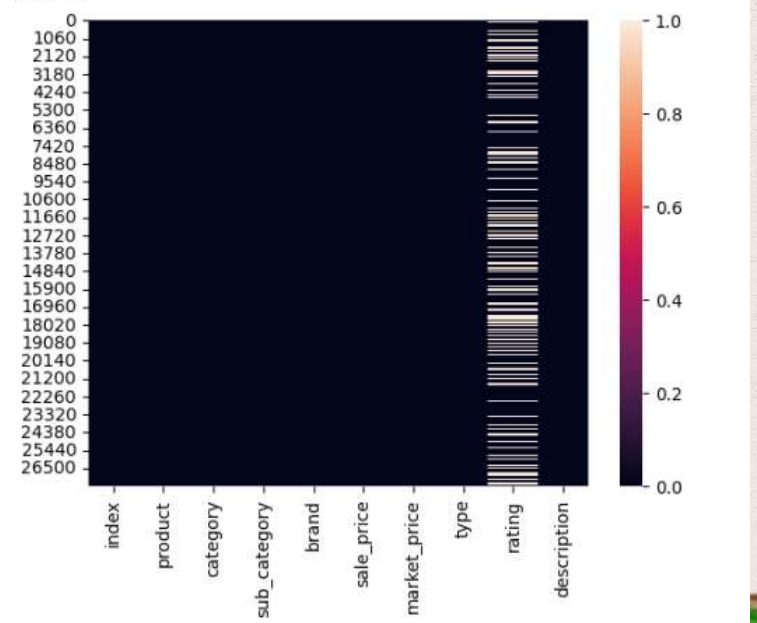
```
[62] data["product"] = data["product"].fillna(data["product"].mode()[0])
[63] data["brand"] = data["brand"].fillna(data["brand"].mode()[0])
[64] data["sale_price"] = data["sale_price"].fillna(data["sale_price"].median())
[65] data["rating"] = data["rating"].fillna(data["rating"].median())
```

For categorical columns, we replace missing values with the **mode** (most frequently occurring value) since categorical data does not have a meaningful numerical average.

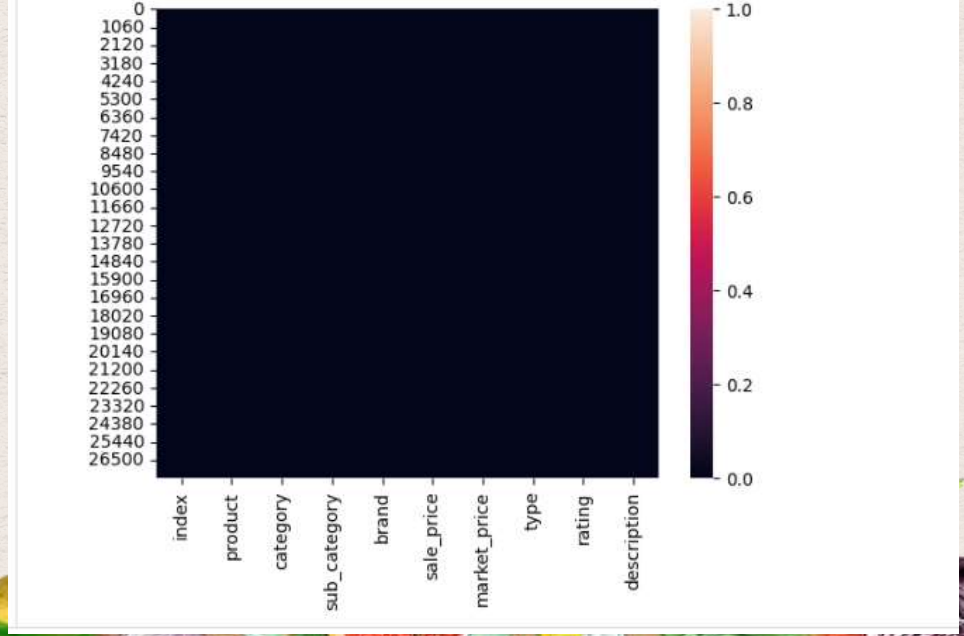
When a numerical column has **outliers**, using the **mean** can distort the data, as outliers significantly impact the average. Instead, we use the **median**, which is more robust to extreme values.

Before Vs After Handling Missing Values

```
[61] sns.heatmap(data.isnull())
```




```
sns.heatmap(data.isnull())
```



Step 4: Data Cleaning and Pre processing

Feature Extraction (Creating Features)

In this step, I created two new columns: **"Discount Amount"** and **"Discount Percentage"** to analyze pricing strategies and discount trends. The **Discount Amount** represents the difference between the **Market Price** and **Sale Price**, while **Discount Percentage** helps assess the relative discount given on each product. Additionally, I handled **missing values** by filling categorical data with the **mode** and numerical data with the **median** (where outliers were present). Outliers were identified and treated to ensure accurate discount analysis and meaningful insights.

```
[24] data["Dicount_amount"] =  data["market_price"] -data["sale_price"]
     data.head()
```

```
data["Discount_amount"] = data["Discount_amount"].fillna(data["Discount_amount"].median())
data["Discount_percentage"] = data["Discount_percentage"].fillna(data["Discount_percentage"].median())
```

```
[78] discount_data = data[["product","sale_price","market_price","Discount_amount"]].head(5)
```

discount_data

	product	sale_price	market_price	Discount_amount
0	Garlic Oil - Vegetarian Capsule 500 mg	220.0	220.0	0.0
1	Water Bottle - Orange	180.0	180.0	0.0
2	Brass Angle Deep - Plain, No.2	115.0	266.0	151.0
3	Cereal Flip Lid Container/Storage Jar - Assort...	149.0	176.0	27.0
4	Creme Soft Soap - For Hands & Body	162.0	162.0	0.0

Next steps: Generate code with discount_data | View recommended plots | New interactive sheet

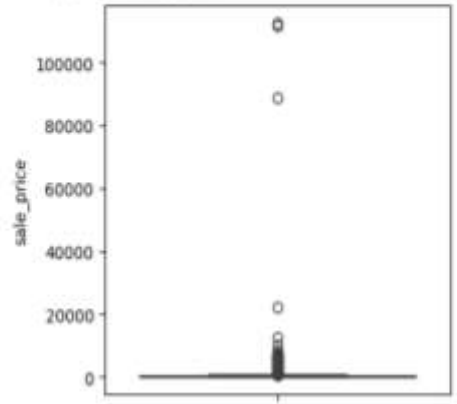
```
[27] data["Discount_percentage"] = (data["Discount_amount"]/data["market_price"])*100
     data.head()
```

	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description	Discount_amount	Discount_percentage
0	1	Garlic Oil - Vegetarian	Beauty &	Hair Care	Sri Sri	220.0	220.0	Hair Oil	4.1	This Product contains	0.0	0.000000

Step 3: Basic Data Insepection

```
plt.figure(figsize=(4,4))
sns.boxplot(data['sale_price'])
```

```
<Axes: ylabel='sale_price'>
```



HANDLING AND REMOVING OUTLIERS

Outliers are extreme values in a dataset that deviate significantly from the majority of the data.

After generating a box plot for the ‘sale-price’ , we identified the presence of outliers in this column .Hence, storing all the numerical columns in a variable and then handling outliers together using Inter Quartile Range method. Setting upper bound and lower bound.

```
[74] num_col = [i for i in data.select_dtypes(include = ["float64"]).columns]
```

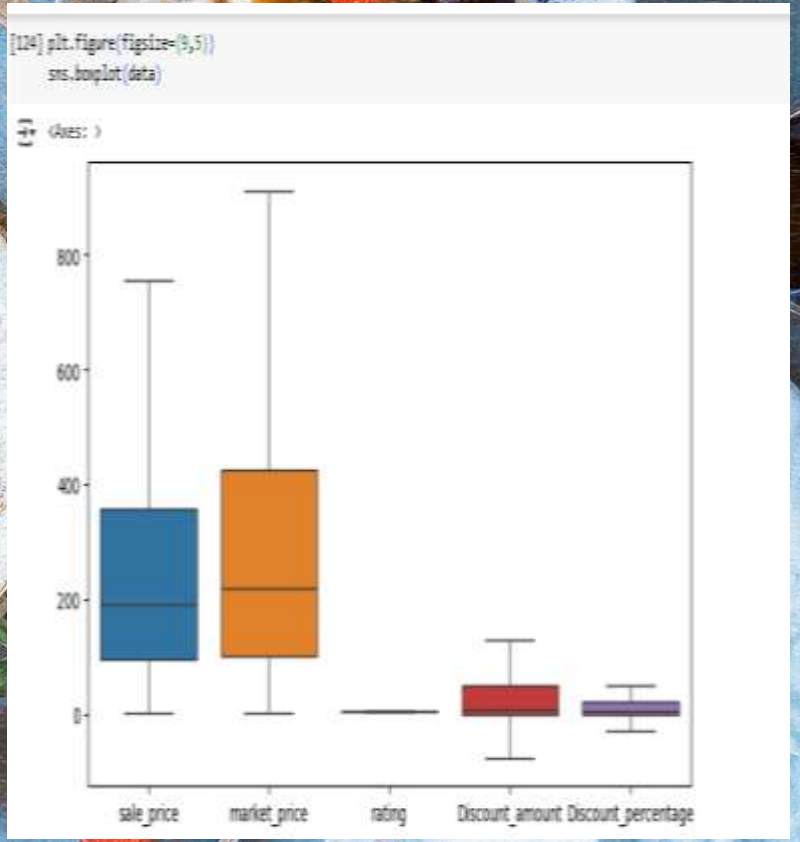
IQR

```
[75] for i in num_col:
        Q1 = data[i].quantile(0.25)
        Q3 = data[i].quantile(0.75)
        IQR = Q3-Q1
        UL = Q3 + 1.5 *IQR
        LL = Q1 - 1.5*IQR
        data[i] = np.where(data[i]>UL,UL,
                           np.where(data[i]<LL,LL,
                                    data[i]))
```

Outliers Removed

I removed all outliers using the Interquartile Range (IQR) method, which helps in identifying and eliminating extreme values that could distort analysis. IQR is the range between the 75th percentile (Q3) and the 25th percentile (Q1) of a dataset, and any values beyond 1.5 times the IQR from Q1 or Q3 are considered outliers.

By removing these outliers, the dataset becomes more reliable for analysis, as extreme values can skew averages, distort trends, and mislead insights. This ensures that key metrics like pricing, discounts, and ratings reflect true patterns, leading to better decision-making and business strategy formulation for BigBasket.

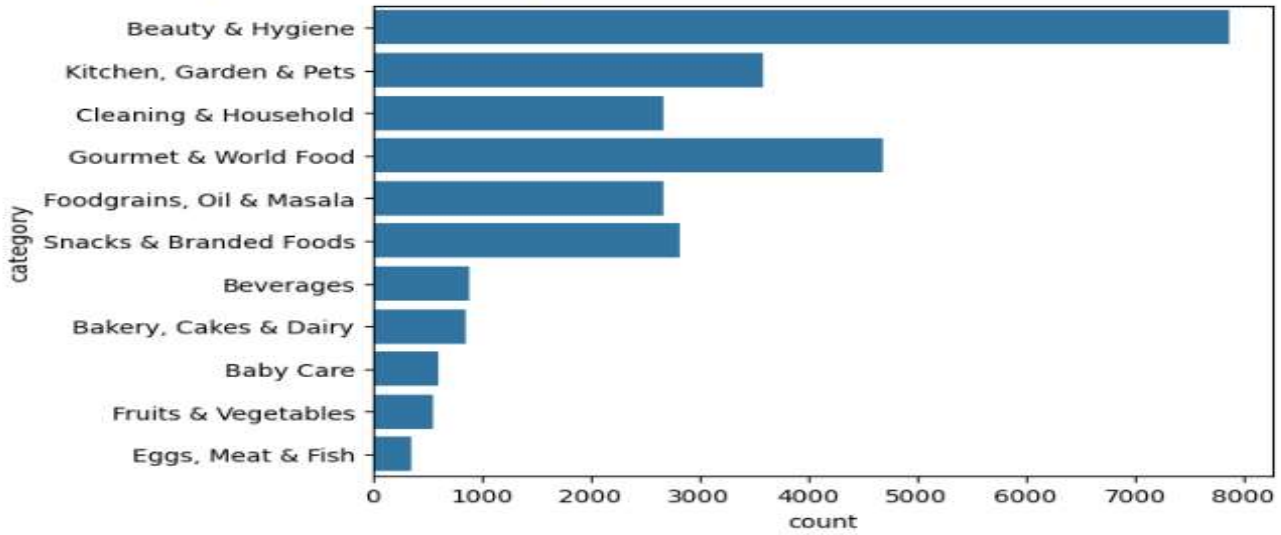


Step 5: Data Analysis and visualization

1. Plot the distribution of number of products in each category.

```
#BAR CHART: Plot the distribution of number of products in each Category.
sns.countplot(data['category'])
```

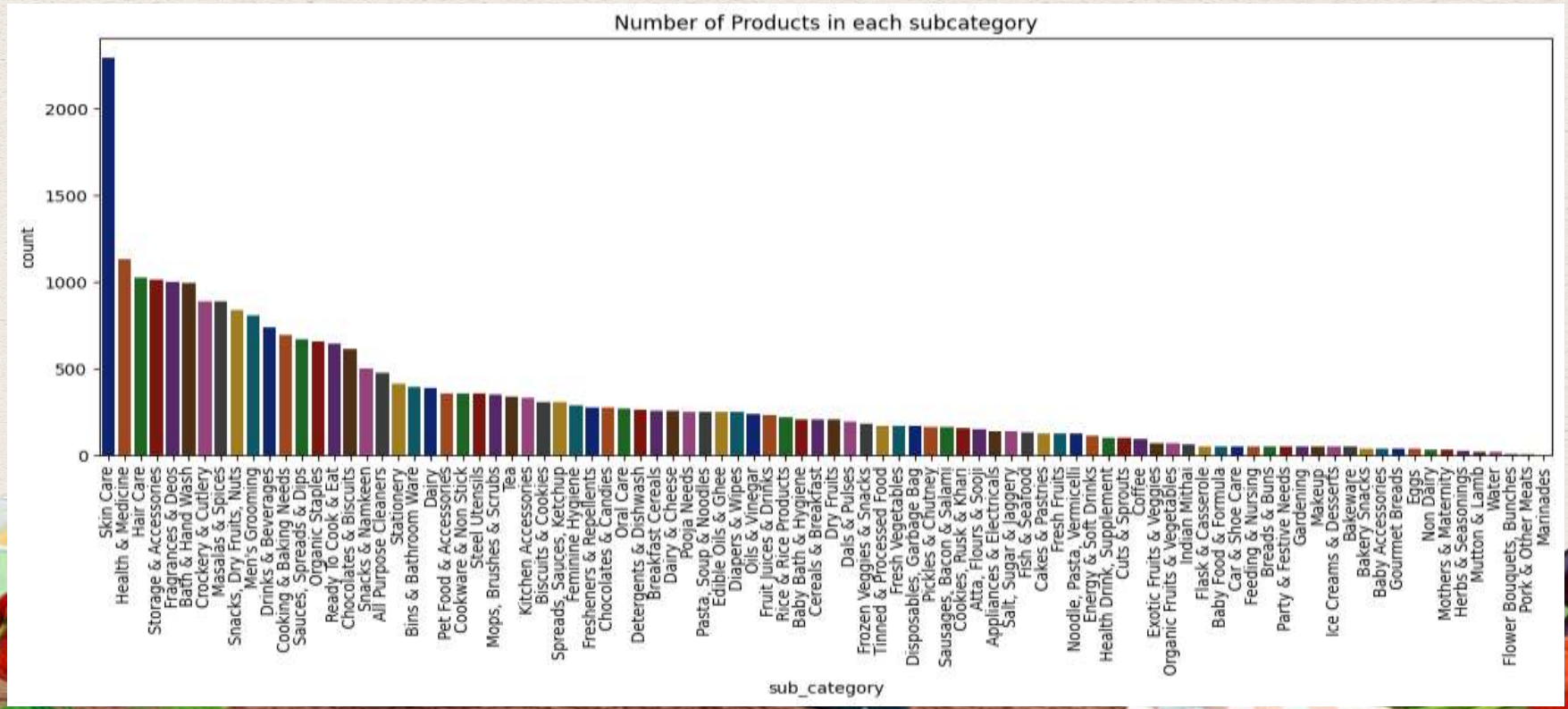
```
<Axes: xlabel='count', ylabel='category'>
```



Key Insights

- The most popular category is “Beauty & Hygiene” , “Gourment & World Food” , “Kitchen, Garden & Pets” .
- The least popular categories are “Egg , Meat and Fish" and “Fruits and Vegetables".
- The distribution of products across categories is not even, with some categories having significantly more products than others.

2. Plot the distribution of Number of products in each subcategory.



Key Insights

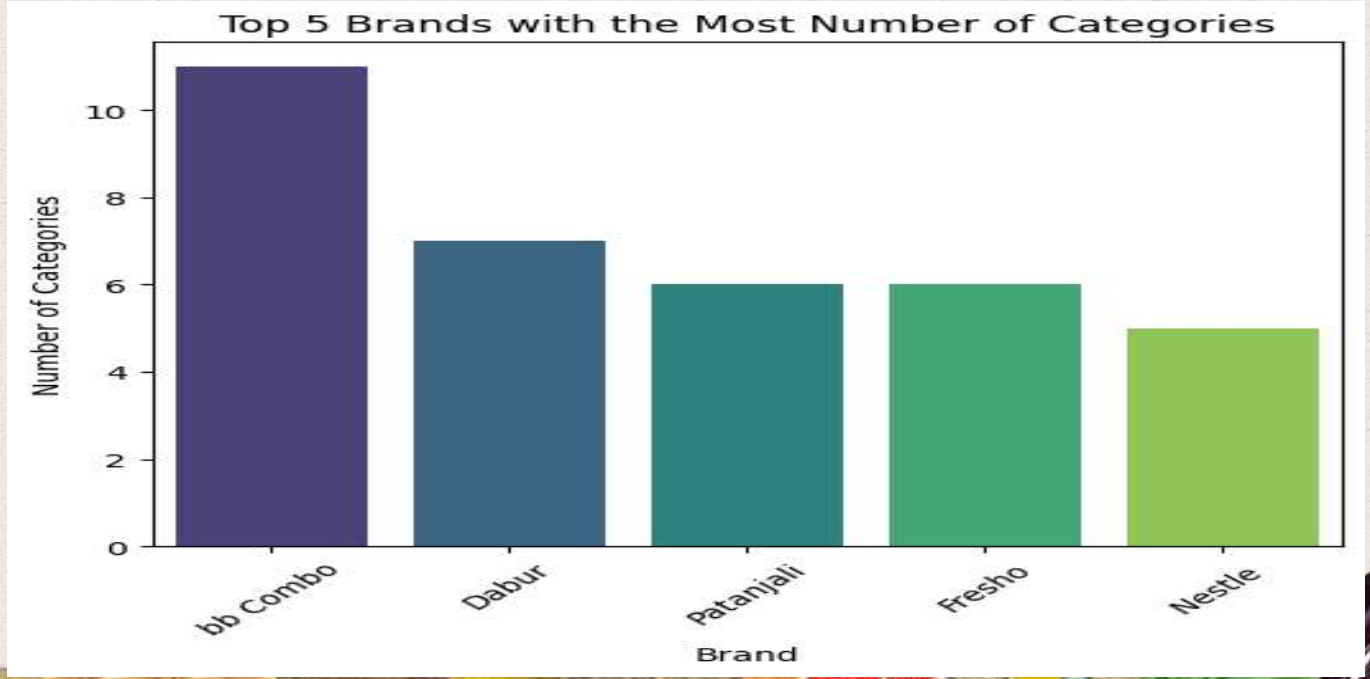
- Skin Care : is the most popular sub-category, with the highest number of products.
- Health & Medicine : is the second most popular sub-category.
- Hair Care : is the third most popular sub-category. The least popular sub-categories are Pork & Other Meats and Marinades.

- **Note: All the top 3 subcategories fall into Broader category of “Beauty and Hygiene”.**

3. Top 5 brands which offers the widest variety of categories,

brand	category
bb Combo	11
Dabur	7
Patanjali	6
Fresho	6
Nestle	5

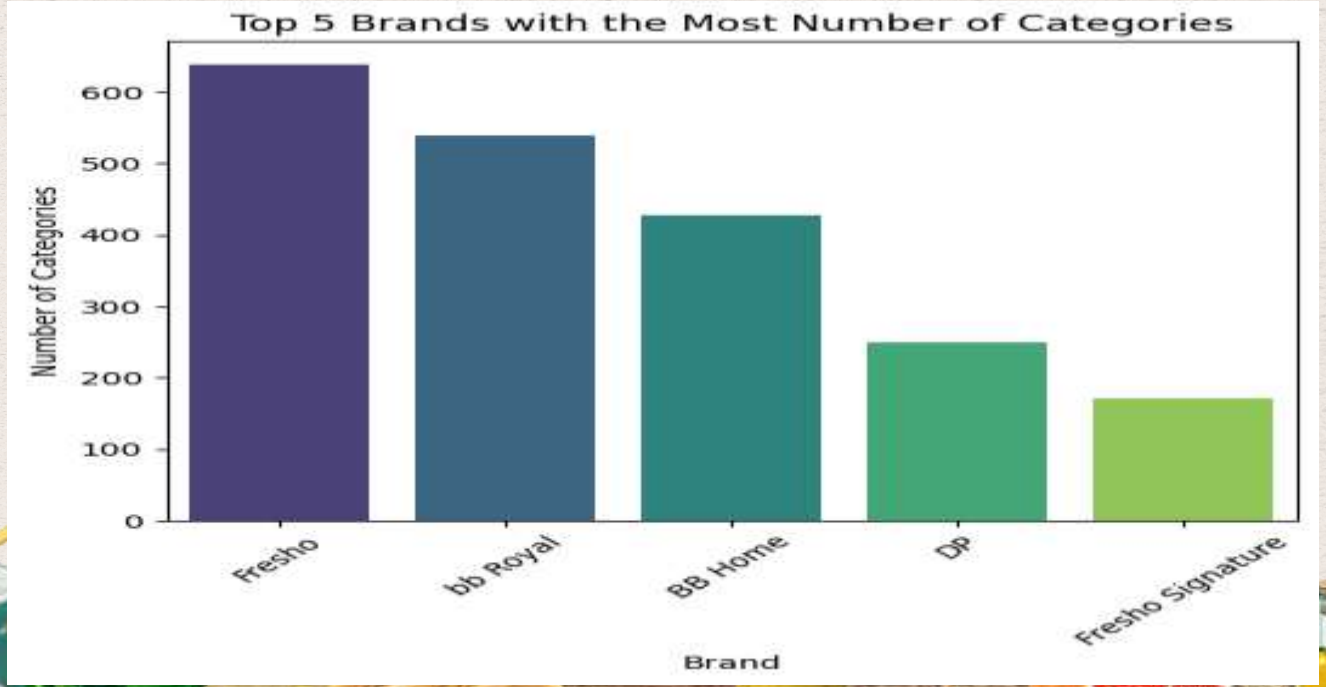
dtype: int64



Key Insights

- **"bb Combo" Leads:** Dominates with **11 categories**, making it the most diverse brand. Big Basket should prioritize it for promotions.
- **Significant Drop After "bb Combo":** "Dabur" follows with only **7 categories**, showing a major gap.
- "Patanjali," "Fresho," and "Nestle" offer products in nearly equal categories, focusing on specific niches.
- **4.** Smaller brands can expand their category range to enhance market presence.

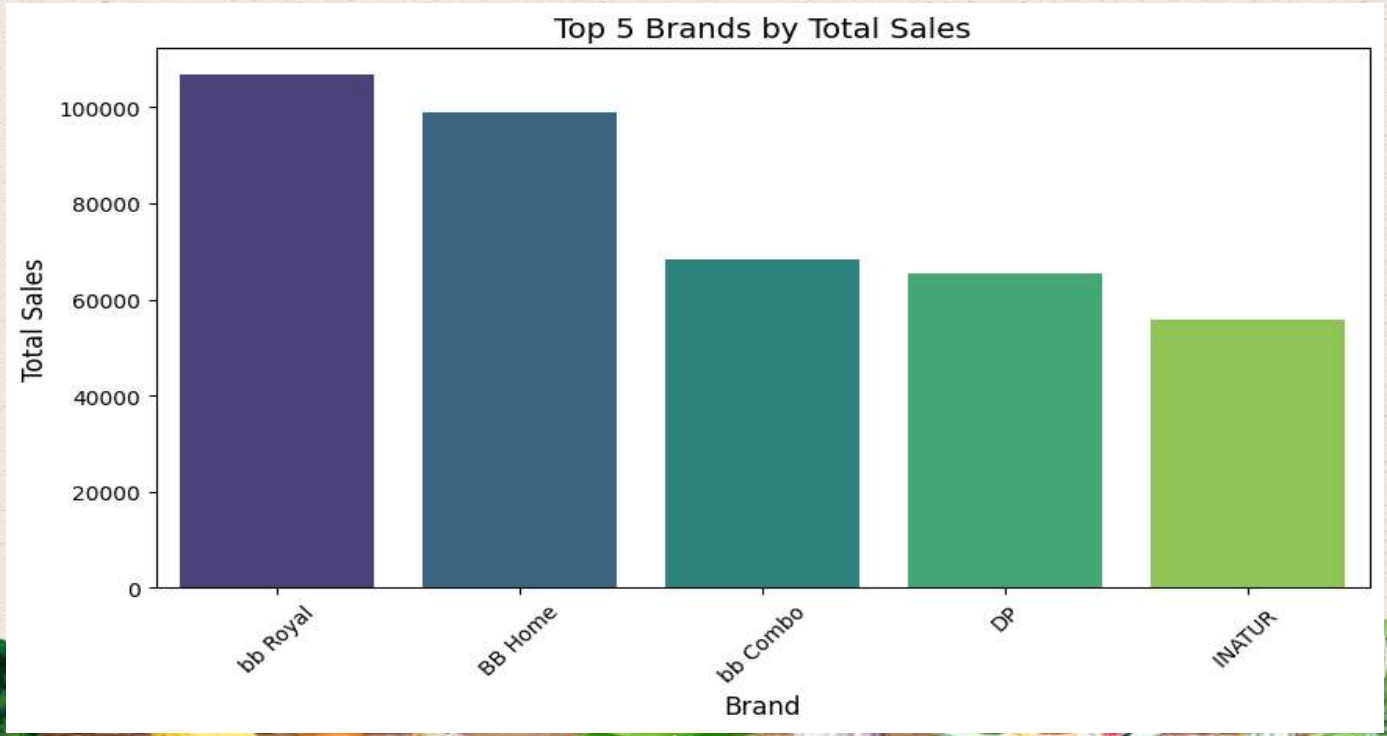
4. Top 5 brands offering highest number of products.,



Key Insights

- **"Fresho" Leads:** Dominates with the highest number of categories, indicating strong product diversity.
- **"bb Royal" & "BB Home":** Follow closely, showing significant presence but trailing behind "Fresho."
- **Big Drop After Top 3:** "DP" and "Fresho Signature" have far fewer categories, indicating a narrower product focus.

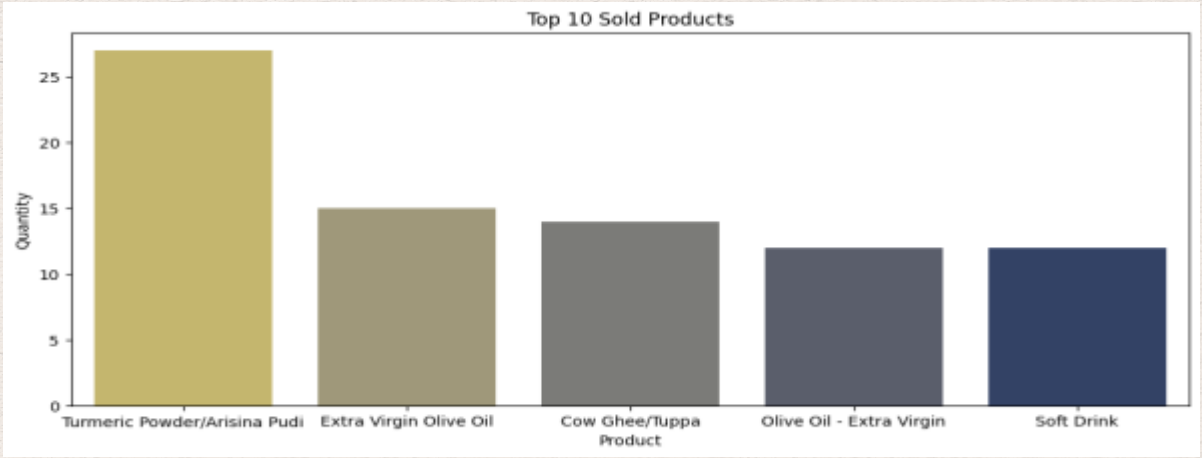
5. Top 5 brands by Total Sales.



Key Insights

"BB Royal" and "BB Home" are market leaders, contributing the highest sales, highlighting strong brand trust and customer preference.
These two brands **dominate Big Basket's revenue**, suggesting they are flagship brands.
Brands with the **"BB" prefix** appear to be in-house or affiliated with a parent corporation, similar to Reliance Fresh.
"bb Combo" and "Fresho" lag behind, with significantly lower sales than the top two, indicating a potential gap in demand or market reach.
Strategic opportunities: Big Basket could boost marketing for lower-performing brands or leverage the success of its top brands to enhance overall sales.

6. Find out top 5 and bottom 5 Sold products



Top 10 Sold Products:

	product	quantity
0	Turmeric Powder/Arisina Pudi	26
1	Extra Virgin Olive Oil	15
2	Cow Ghee/Tuppa	14
3	Olive Oil - Extra Virgin	12
4	Soft Drink	12
5	Colorsilk Hair Colour With Keratin	12
6	Ghee/Tuppa	11
7	Powder - Coriander	11
8	Coriander Powder	11
9	Peanut Butter - Creamy	10

****

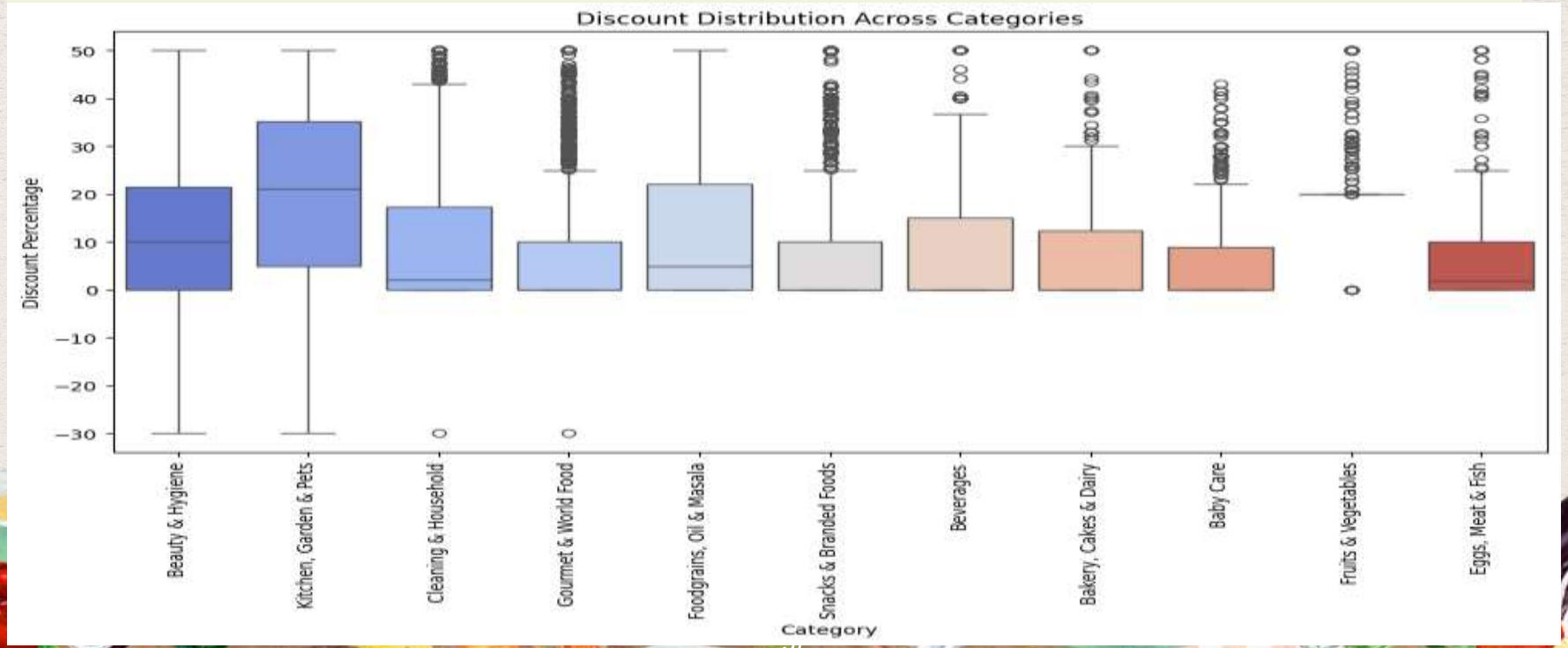
Bottom 10 Sold Products:

	product	quantity
23530	Sauteed Onion & Garlic Pasta Sauce	1
23531	Pepper & Herb Salami Chicken	1
23532	Nutmeg Powder	1
23533	Disney Mickey Mouse Plastic Kids Sipper Bottle...	1
23534	Chocolates-Roasted Peanut Chocolate	1
23535	Opalware Classique Serving Bowl - Medium, Roya...	1
23536	Lavangadi Vati - Respiratory Conditions, 300mg	1
23537	Pomegranate - Small	1
23538	Butter - Cashew, Smooth	1
23539	Powder - Pepper	1

Key Insights

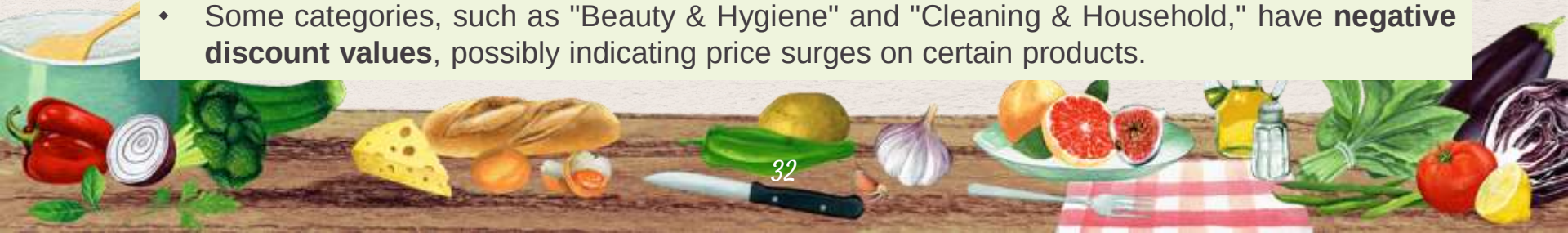
- Turmeric Powder/Arisina Pudi : is the most popular product with the highest number of sales.
- Extra Virgin Olive Oil : and Cow Ghee/Tuppa are also among the top-selling products.
- Soft drinks: are in the middle range of demand.
- Hair color, olive oil, coriander powder, and peanut butter: have relatively lower sales compared to the top-selling products.
- Overall, the graph indicates a high demand for essential cooking ingredients and personal care product

7. Draw a visualization to compare Discount Distributions across Categories

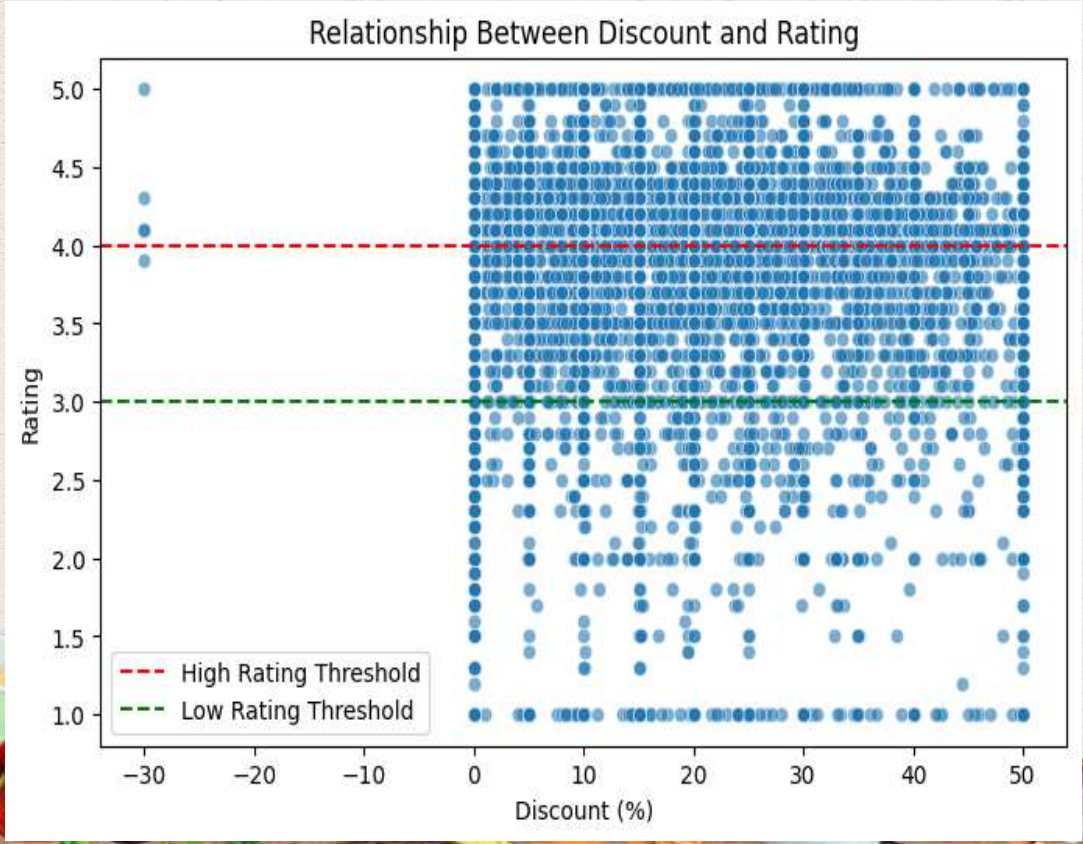


Key Insights

- Categories with **longer boxes** have a **wider range of discounts**, indicating that discounts vary significantly within these categories.
- The **horizontal line inside each box** represents the **median discount**, helping compare the typical discount percentage across categories.
- Multiple categories, including "Foodgrains, Oil & Masala" and "Beverages," show **significant outliers**, suggesting occasional extreme discounts or promotional offers.
- **Dots outside the whiskers** represent **outlier discounts**, showing products that receive unusually **high or low discounts** compared to others in the same category.
- Some categories, like **Kitchen, Garden & Pets and Beauty & Hygiene**, have **higher median discounts**, making them more attractive for customers looking for deals.
- "Eggs, Meat & Fish" and "Baby Care" show **lower median discounts** and **less variability**, suggesting stable pricing strategies.
- Some categories, such as "Beauty & Hygiene" and "Cleaning & Household," have **negative discount values**, possibly indicating price surges on certain products.



8. Draw a visualization of Top 6 Brands to show their Market Share



The scatter plot does not show a strong linear relationship between "Discount" percentage and "Rating". This suggests that offering a higher discount does not necessarily lead to a higher product rating.

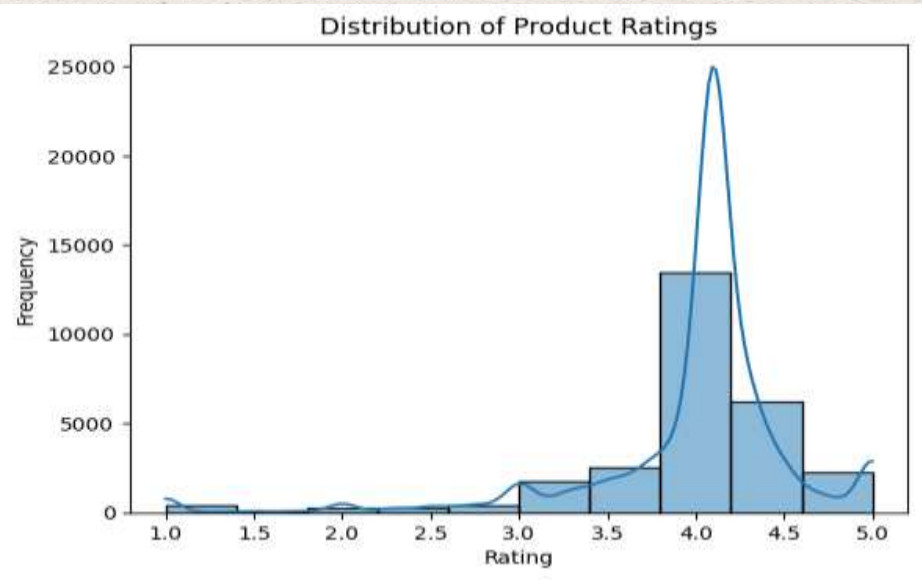
Other factors, such as product quality, brand reputation, and customer expectations, likely play a more significant role in determining product ratings

9. Visualization to explore the relationship between Product Sale Price and Rating

#If we do not handle outliers in rating column
num_col = [i for i in data.select_dtypes(include=["float64"]).columns if i != "rating"]

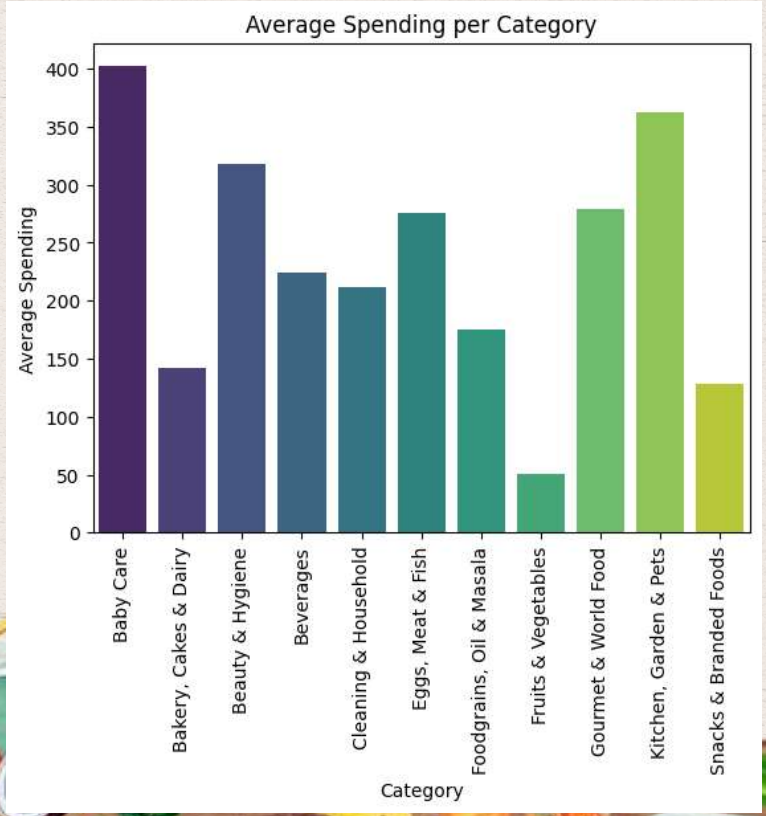
There doesn't seem to be a strong linear relationship between product "Sale Price" and "Rating". This suggests that customers don't necessarily rate higher-priced products more favorably.
Most products are concentrated in the lower price range, regardless of their rating. This indicates that the majority of products offered are budget-friendly.

10. Draw a visualization to show the Distribution of Product ratings.



The distribution of product "Ratings" is heavily left skewed towards higher ratings, with the majority of products receiving ratings of 4 or above. This actually shows that Big basket fulfills the purpose of customer satisfaction with respect to its services.

11. Average Spending per Category



category	sale_price
Baby Care	402.105885
Bakery, Cakes & Dairy	142.268637
Beauty & Hygiene	317.571865
Beverages	224.160644
Cleaning & Household	211.180138
Eggs, Meat & Fish	276.177200
Foodgrains, Oil & Masala	174.805688
Fruits & Vegetables	50.889336
Gourmet & World Food	279.295557
Kitchen, Garden & Pets	362.080106
Snacks & Branded Foods	127.905704

dtype: float64

Findings

- **"Beauty & Hygiene"** is the most dominant category, followed by **"Gourmet & World Food."**
- **"Skin Care"** leads as the most popular sub-category.
- **"Fresho"** has the highest number of products, while **"BB Home" and "BB Royal"** generate the highest total sales, indicating strong customer preference.
- **"BB Royal" and "BB Home"** are the key revenue drivers, suggesting strong brand trust and positioning.
- The top brands dominate the market in terms of **product variety, sales, and market share.**

Findings

- No strong correlation exists between **discount percentage and product ratings.**
- **Product sale price does not directly influence ratings**, suggesting price isn't the major concern for customers.
- Categories with wider discount ranges include **"Foodgrains, Oil & Masala" and "Beverages,"** indicating significant promotional activity.
- **"Eggs, Meat & Fish" and "Baby Care"** have lower median discounts, reflecting stable pricing strategies.
- Some categories like **"Beauty & Hygiene" and "Cleaning & Household"** exhibit negative discount values, possibly indicating **price surges** on select products.

Findings

- Ratings are **skewed towards higher values (4 or above)**, demonstrating strong customer satisfaction.
- **Higher discounts do not guarantee better ratings**, emphasizing the role of product quality and customer experience.
- Most products fall in the **lower price range**, yet receive high ratings, indicating a preference for **budget-friendly yet quality offerings.**

Findings

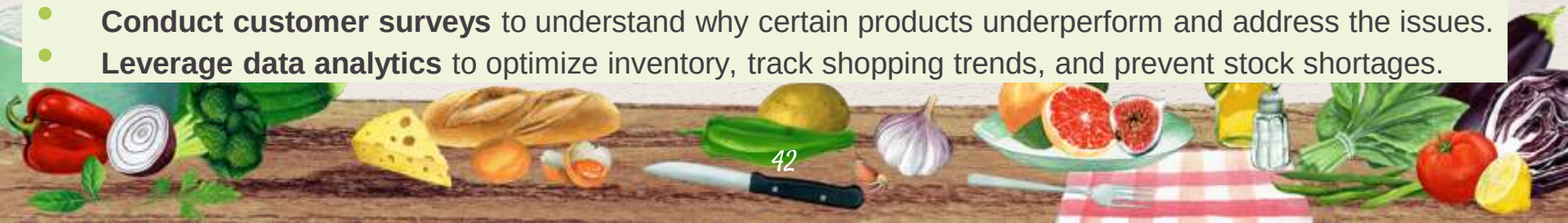
- **Top-selling products:**
  - **Turmeric Powder (Arisina Pudi)** has the highest sales.
  - **Extra Virgin Olive Oil and Cow Ghee** are also among the best sellers.
  - **Soft drinks** have moderate demand.
  - **Hair color, olive oil, coriander powder, and peanut butter** show lower sales.
- **"BB Combo" and "Fresho" brands** have significantly lower sales, suggesting a gap in demand or market reach.

Findings

- Highest average spending per category:
  - Baby Care (₹402.10 per product)
  - Bakery, Cakes, and Dairy
  - Beauty & Hygiene
  - Indicates essential and personal care items receive higher spending.

Recommendations

- **Invest more in top-performing categories** like "Beauty & Hygiene" and "Gourmet & World Food."
- **Promote "BB Royal" and "BB Home" more aggressively** to maintain their strong market presence.
- **Enhance marketing strategies for underperforming brands** like "BB Combo" and "Fresho."
- **Use cross-promotional strategies** to increase sales of lower-demand products (e.g., hair color, peanut butter).
- **Focus on product quality improvements** instead of heavy discounting to maintain high customer ratings.
- **Maintain stable pricing** for "Eggs, Meat & Fish" and "Baby Care" categories where discounts are less impactful.
- **Avoid frequent price surges** in "Beauty & Hygiene" and "Cleaning & Household" categories.
- **Improve customer engagement** through better loyalty programs and personalized recommendations.
- **Conduct customer surveys** to understand why certain products underperform and address the issues.
- **Leverage data analytics** to optimize inventory, track shopping trends, and prevent stock shortages.



Thankyou

Divyanshi Nigam

Aspiring Data Analyst