



Exploratory Data Analysis

PROJECT BY SHIVAM RAWAT



About Big Basket

- ▶ **Big Basket** is an Indian online grocer headquartered in [Bangalore](#), India, and currently owned by [Tata Digital](#). It was the first online grocer in India, set up in 2011. It is a registered company with the name **Supermarket Grocery Supplies Pvt. Ltd.**
- ▶ As of January 2023, the company operates in more than 30 cities in India and processes around 15 million orders per month.
- ▶ According to [Bloomberg News](#), Big Basket is considering an [initial public offering](#) by 2025 as part of its growth plans.

INTRODUCTION:

His Dataset is sourced from Skill Circle and contains data collected from Big Basket. After a quick view of the Dataset, it looks like Sales dynamics data frame with multiple Product offerings. The dataset is a crucial asset for Exploratory Data Analysis (EDA), allowing us to explore Big Basket's operational metrics, product popularity, pricing strategies, and customer feedback in detail.

This will involve steps such as loading the data, generating descriptive statistics, profiling the data, identifying outliers, and using visualization techniques.

By conducting thorough analysis and creating visualizations, we seek to identify patterns, trends, and insights that can guide strategic decisions, improve inventory management, and enhance the shopping experience for customers

OBJECTIVES OF THE PROJECT:

- ▶ □ The goals of this assessment is to -
- ▶ □ Sales Data Analysis : Understanding of General Sales performance and patterns.
- ▶ □ Top Selling Products : Identify which products are driving High Sales for the brand.
- ▶ □ Discount Analysis : Measure Discounts offered on products and analyze their impact on Sales.
- ▶ □ Handling Missing Values : Ensuring data quality by identifying and Handling Missing Values appropriately.
- ▶ □ Anomaly Detection and Handling : Identify and manage Anomalies to maintain data integrity.
- ▶ □ Consumer Insights : Ratings and product reviews provide valuable feedback that can guide product improvements and marketing efforts.
- ▶ □ Data Visualization : Create visual representations of data to better understand trends and insights.

DESCRIPTION OF DATASET:

- • The Dataset has been imported from Google Drive.
- • I have performed my work using Google Collaboratory Notebook.
- • As we begin our Exploratory Data Analysis (EDA), I've named the dataset 'df'.
- • The dataset comprises of 27,555 Rows and 10 Columns.
- • For Data cleaning/visualization, I have utilized libraries like NumPy, Pandas, Seaborn, Matplotlib.
- • Any duplicate entries that were found have also been removed.

```
[4] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

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

```
↗ Mounted at /content/drive
```

```
[6] df = pd.read_csv("/content/drive/MyDrive/Copy of BigBasket Products.csv")
```

```
▶ df.drop_duplicates(inplace=True)
```

```
[10] df.shape
```

```
↗ (27555, 10)
```

DESCRIPTION OF DATASET:

▫ The dataset under examination provides a comprehensive insight into Big Basket's product offerings and sales dynamics. It encompasses 10 key attributes that shed light on various facets of the business:

▫ Key Features include: - Index: This attribute serves as a unique identifier for each entry in the dataset.

▫ Product: The 'Product' attribute represents the title or name of the products listed on the Big Basket platform.

▫ Category: The 'Category' attribute classifies the products into broader categories, such as fruits, vegetables, dairy products, beverages, etc.

▫ Sub Category: Within each broad category are further classified into more categories.

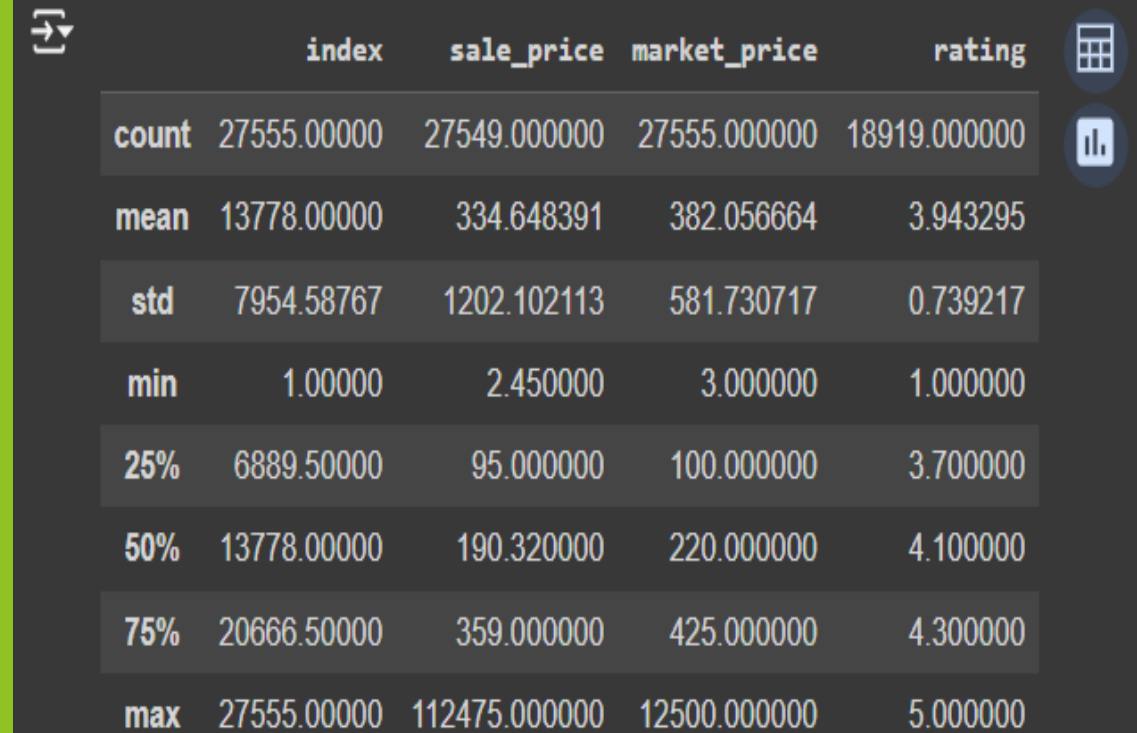
```
df.info()
```

```
<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
```

DESCRIPTION OF DATASET:

- Brand: The 'Brand' attribute indicates the brand or manufacturer associated with each product.
- Sale Price: The 'Sale Price' attribute denotes the price at which each product is offered to consumers.
- Market Price: The 'Market Price' attribute specifies the standard market price of each product.
- Type: The 'Type' attribute categorizes the products based on their nature or characteristics.
- Rating: The 'Rating' attribute represents the consumer rating or feedback received by each product on the Big Basket platform.
- 6. □ Description: The 'Description' attribute provides a detailed narrative describing the dataset, its scope, and the context in which it was compiled

```
[12] df.describe()
```



The image shows a Jupyter Notebook interface with a dark theme. On the left, there is a sidebar with icons for file operations, a table view, and a bar chart. The main area displays a table of statistics for a dataset. The table has five columns: 'index', 'sale_price', 'market_price', and 'rating'. The rows represent different statistical measures: 'count', 'mean', 'std', 'min', '25%', '50%', '75%', and 'max'. The values are displayed in a light blue font on a dark background.

	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

DATA CLEANING & PRE-PROCESSING:

▫ The Dataset contains a total of 8,759 Null values. Of these, 117 are found in categorical features, while 8,642 are in numerical features.

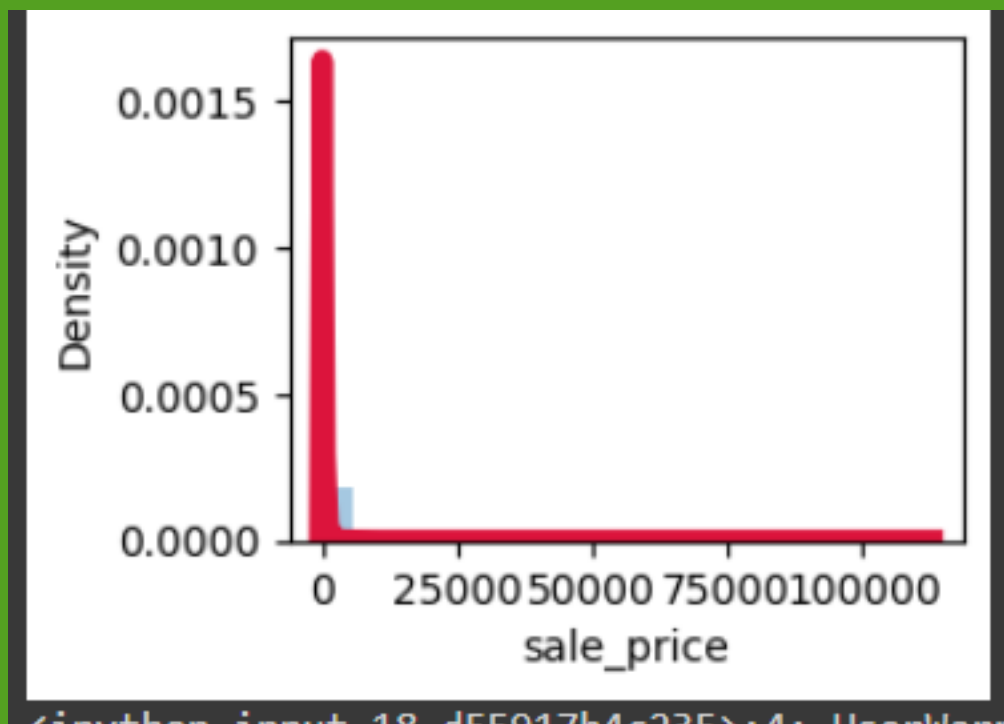
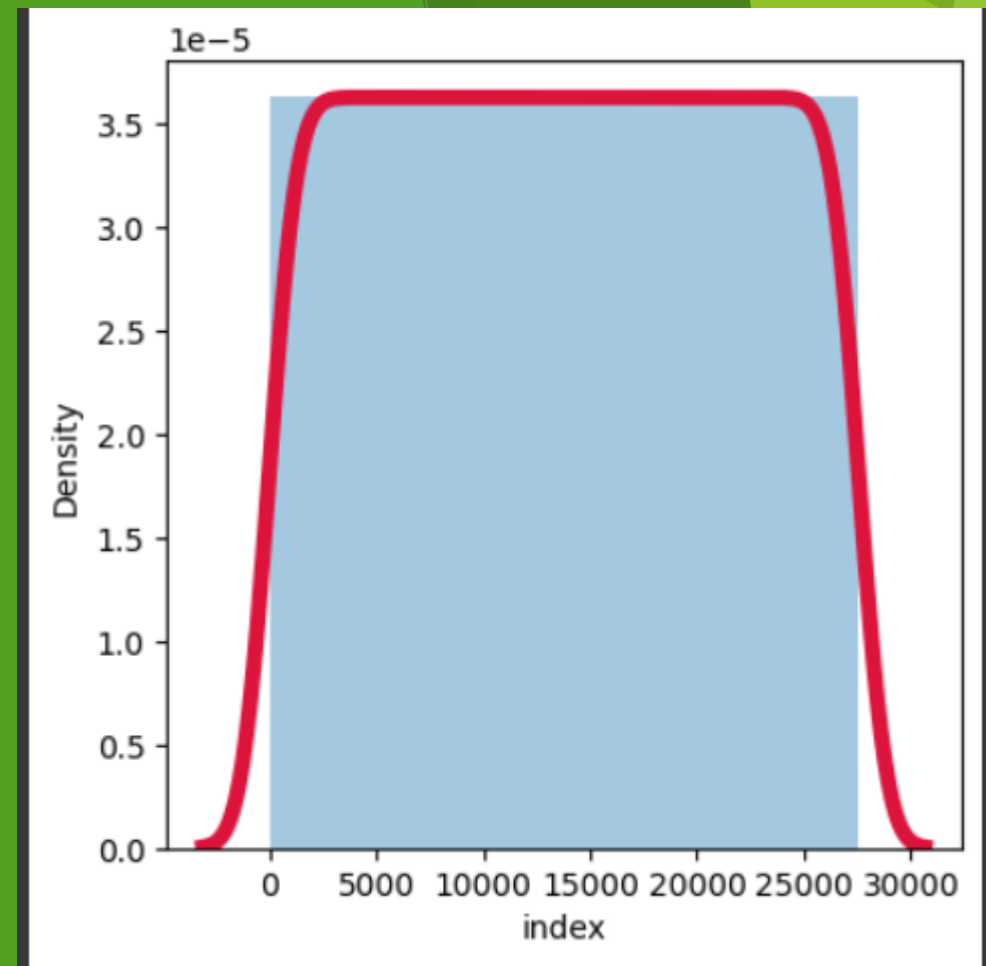
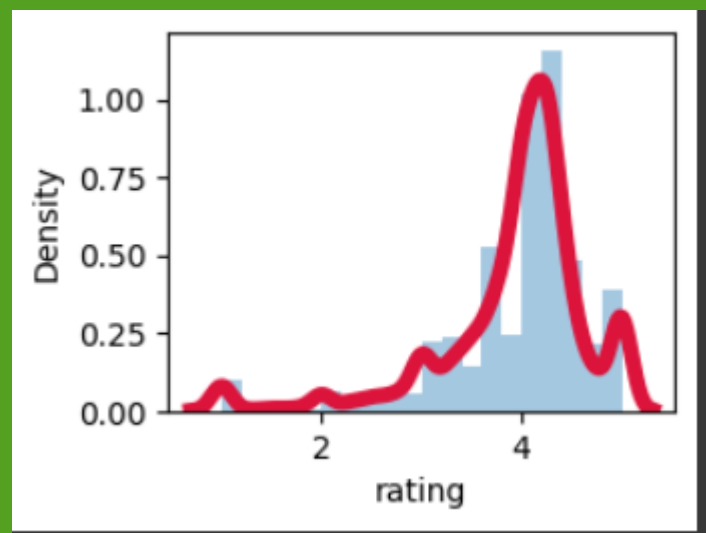
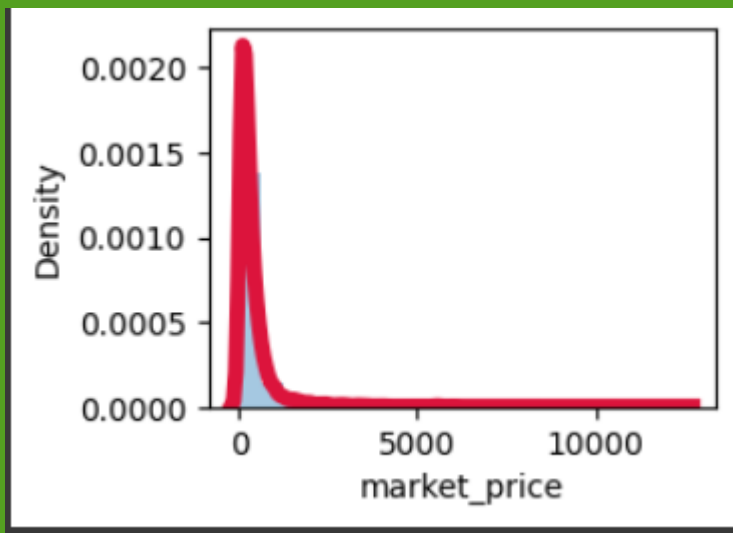
▫ First handling missing value in numerical column

##Mean: Whenever your data is numeric and normally distributed, in this case you will impute missing values with Mean.

##Median: Whenever your data is numeric and skewed, in this case you will impute missing values with Median.

▫ For the 'sale-price' attribute, which has 6 null value filling in the missing entries with the 'median' will help ensure data completeness

▫ For the 'rating' attribute, which has 8636 null value filling in the missing entries with the 'median' will help ensure data completeness.



DATA CLEANING & PRE-PROCESSING:

```
median_sale_price = df['sale_price'].median()  
median_sale_price
```

```
190.32
```

```
[20] df['sale_price'].fillna(median_sale_price,inplace = True)
```

```
[21] median_rating = df['rating'].median()  
median_rating
```

```
4.1
```

```
df['rating'].fillna(median_rating,inplace = True)
```

- The Dataset contains a total of 8,759 Null values. Of these, 117 are found in categorical features, while 8,642 are in numerical features.
- **Brand**: The 'Brand' attribute has only 1 null value in the categorical data. To ensure data completeness, this value can be filled with **'No Brand Provided'**.
- **Product**: For another categorical attribute 'Product' which has again 1 null value, using **'Product is not specified'** to fill in the missing value is a viable solution.

```
df['brand'].fillna('brand not be provided',inplace = True)
```

```
<ipython-input-25-f69b0b84f798>:1: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.  
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.
```

```
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original
```

```
df['brand'].fillna('brand not be provided',inplace = True)
```

- For the 'description' attribute, which has 115 null value filling in the missing entries with the 'No description' will help ensure data completeness

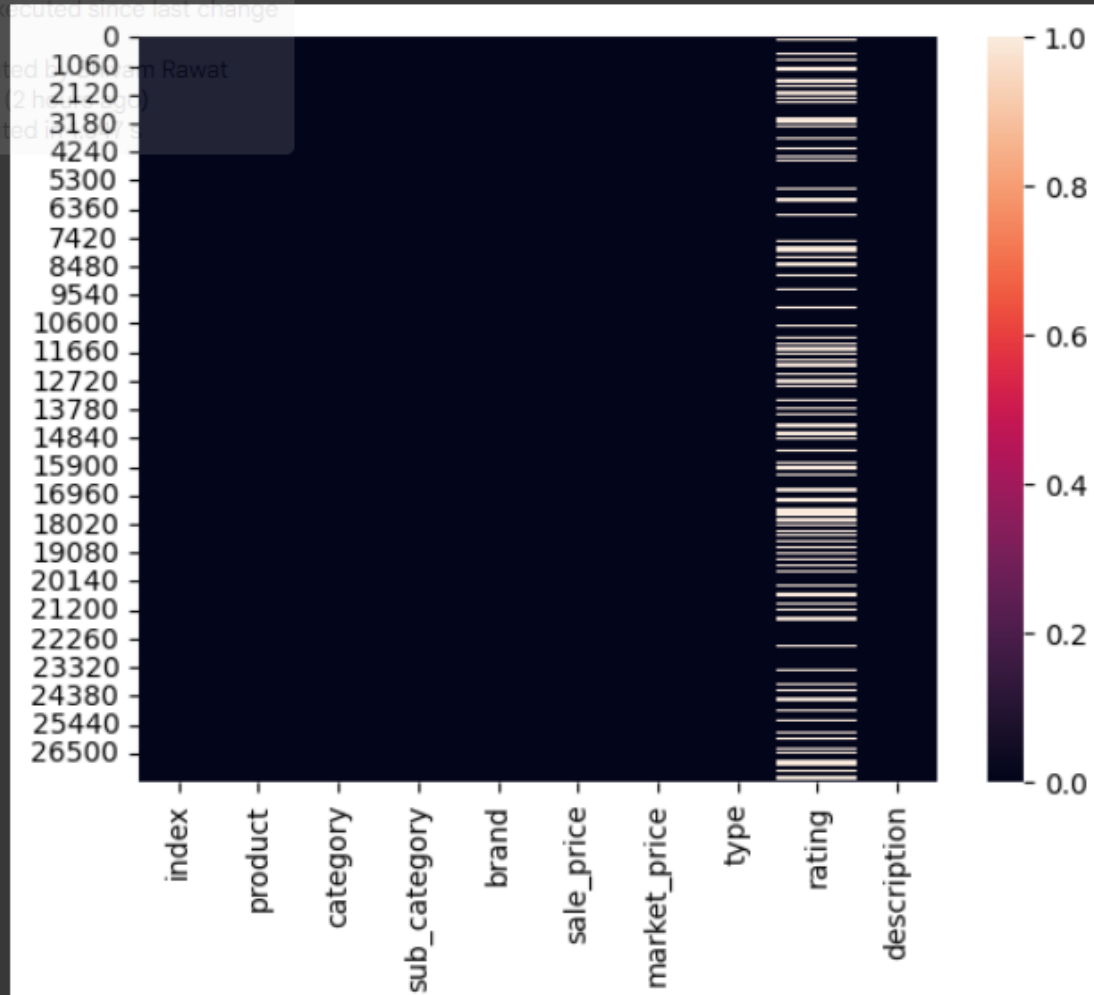
```
[24] df['description'].fillna('No description',inplace = True)
```

HEATMAPS

~BEFORE CLEANING

```
sns.heatmap(df.isnull())
```

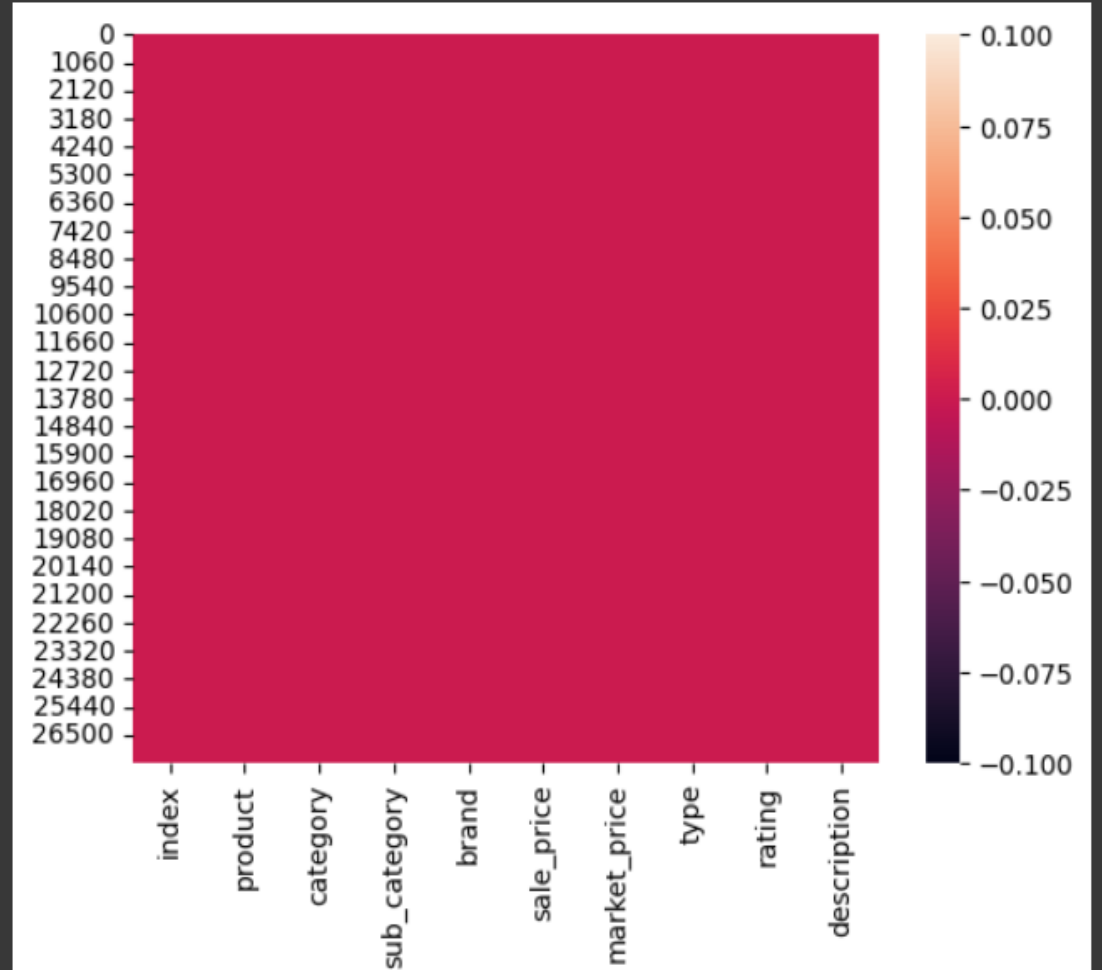
<Axes: >



~AFTER CLEANING

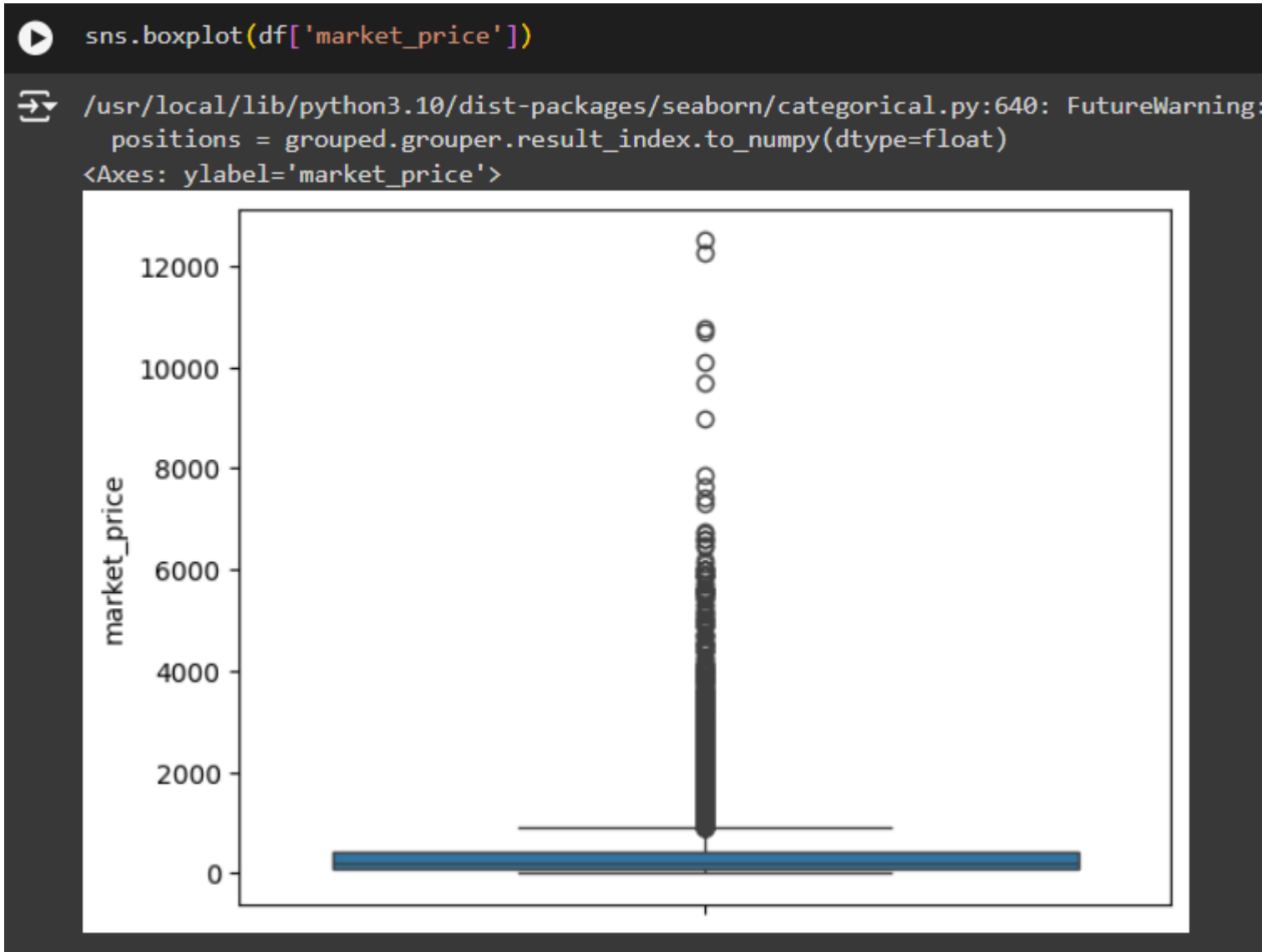
```
[27] sns.heatmap(df.isnull())
```

<Axes: >



REMOVING OUTLIERS:

- ▶ □ After generating a box plot for the 'market_price', we identified the presence of outliers in this column.



REMOVING OUTLIERS :

- ▶ To address these outliers, we will apply the IQR method.

```
▶ Q1 = df['market_price'].quantile(0.25)
  print(Q1)
  Q3 = df['market_price'].quantile(0.75)
  print(Q3)
  IQR = Q3 - Q1
  print(IQR)
```

```
↵ 100.0
   425.0
   325.0
```

```
[35] lower_bound = Q1 - 1.5 * IQR
      upper_bound = Q3 + 1.5 * IQR

      print(lower_bound)
      print(upper_bound)
```

```
↵ -387.5
   912.5
```

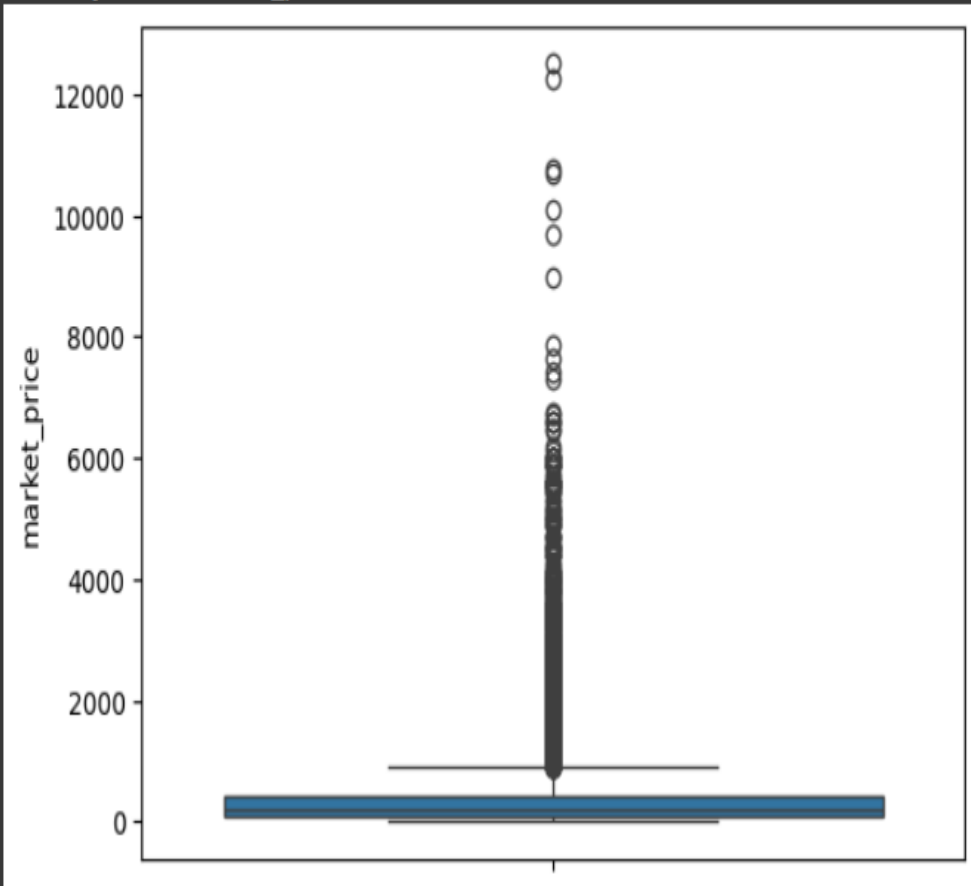
```
[36] df['market_price'] = np.where(df['market_price'] > upper_bound, upper_bound, df['market_price'])
      df['market_price'] = np.where(df['market_price'] < lower_bound, lower_bound, df['market_price'])
```

REMOVING OUTLIERS :

BEFORE REMOVING

```
sns.boxplot(df['market_price'])
```

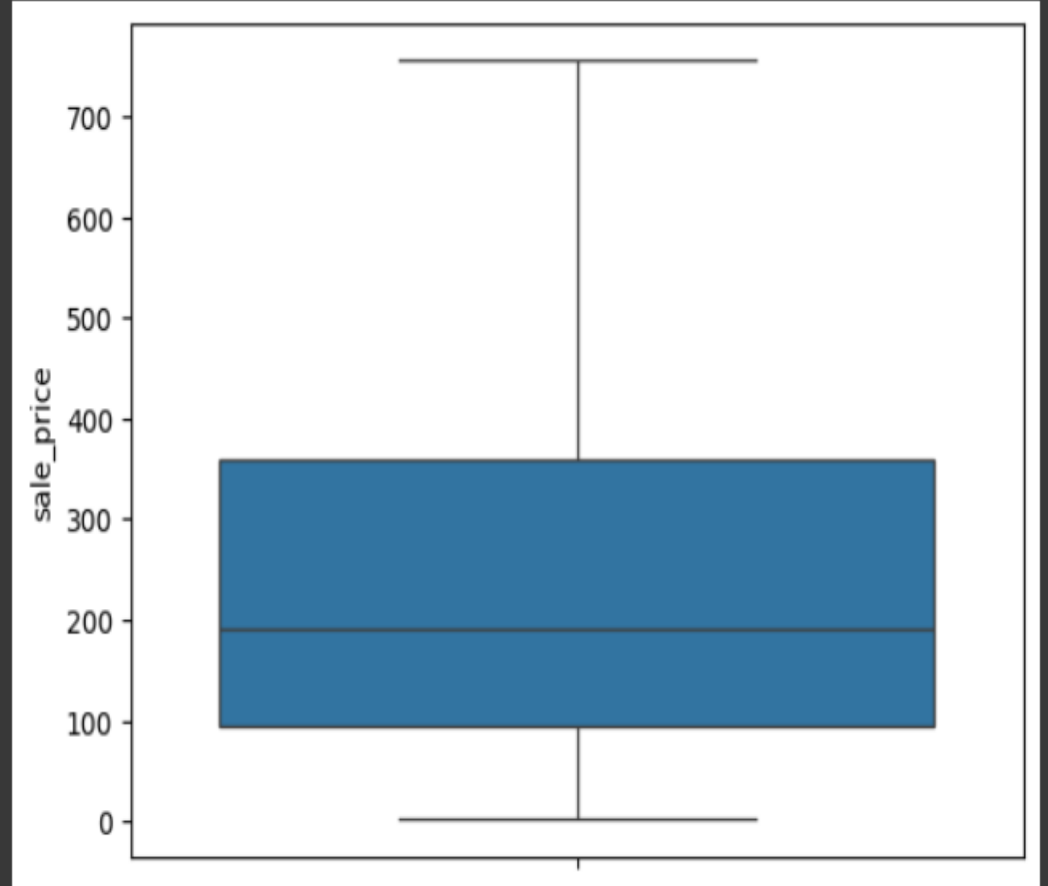
```
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:640: FutureWarning:  
positions = grouped.grouper.result_index.to_numpy(dtype=float)  
<Axes: ylabel='market_price'>
```



AFTER REMOVING

```
sns.boxplot(df['sale_price'])  
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:640: FutureWarning:  
positions = grouped.grouper.result_index.to_numpy(dtype=float)
```

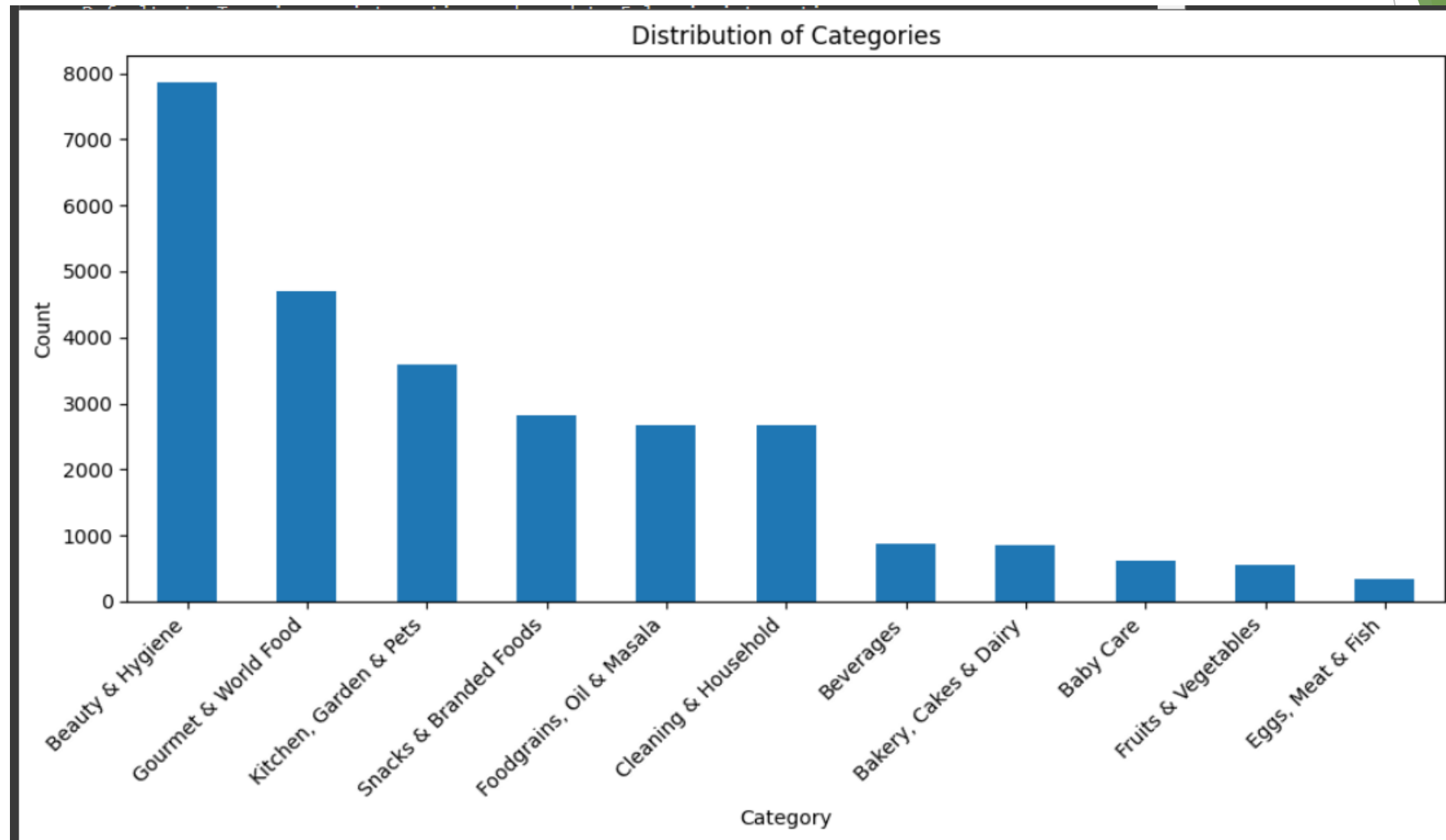


DATA CLEANING & PRE-PROCESSING:

- **Summary** - To summarize, addressing Null values and Outliers necessitates a methodical approach tailored to the data's characteristics and specific attributes. Data cleaning and Outlier handling are crucial steps for accurate analysis.
- ✓ The dataset contained Missing Values in 'product', 'brand', 'sale price', 'rating' and 'description' features. These were handled by imputation (filling with median/mode) and dropping irrelevant columns ('description').
- ✓ Outliers were present in 'sale price' and 'market price'. These were addressed using the IQR method and capping to boundary values.
- ✓ With these Null, Missing, and Invalid values appropriately addressed, we are now ready to move forward with analyzing the dataset.

DATA VISUALIZATION AND INSIGHTS

- **BAR CHART:** Plot the distribution of number of products in each Category

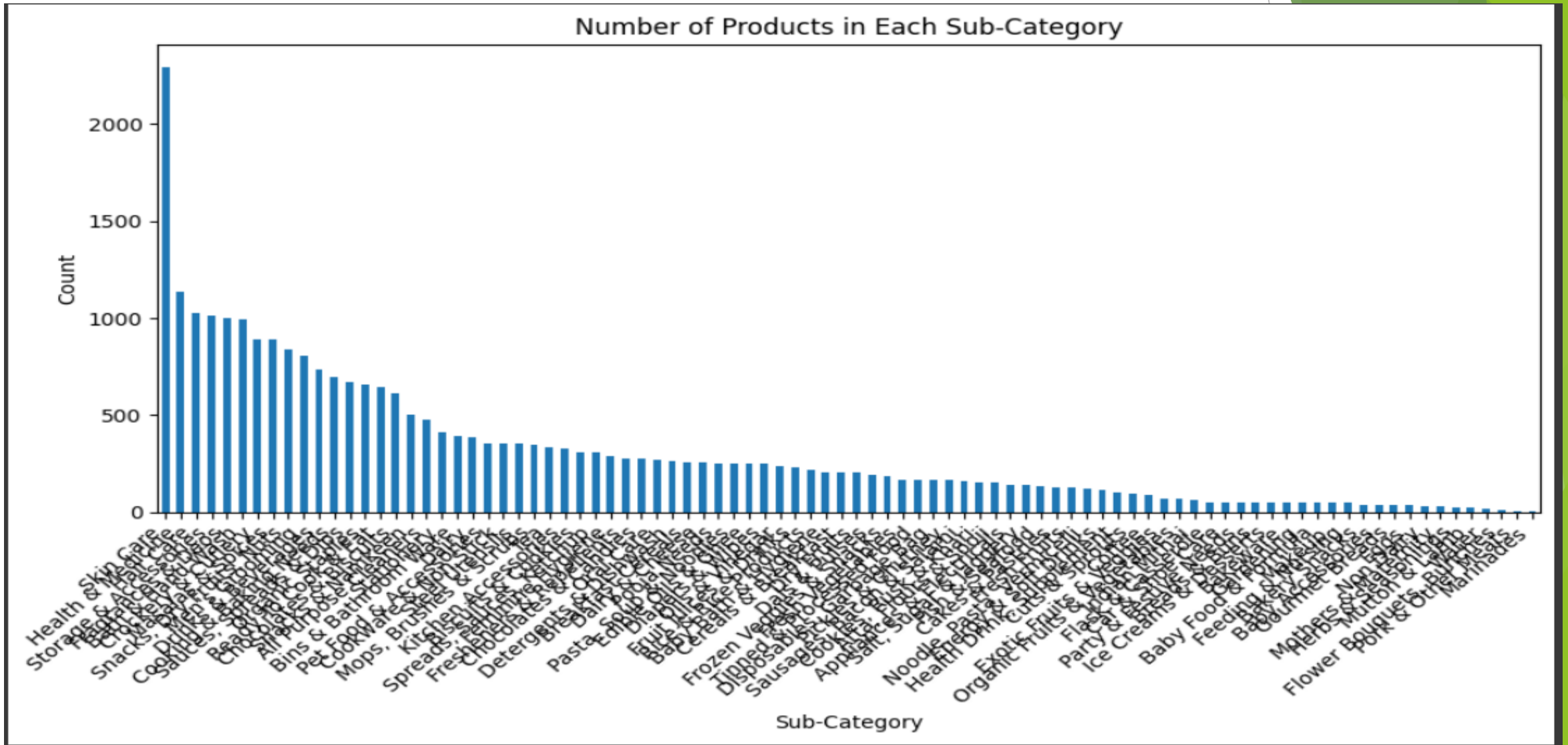


DATA VISUALIZATION AND INSIGHTS

- ▶ □ **BAR CHART:** Plot the distribution of number of products in each Category
- ▶ □ Key insights:
- ▶ The category "Beauty & Hygiene" has the highest number of products. This suggests that Big Basket has a strong focus on this category followed by "Gourmet & World Food".
- ▶ The categories "Snacks & Branded Foods" and "Foodgrains, Oil & Masala" also have a significant number of products. These are essential categories that are likely to be in high demand.
- ▶ The categories "Fruits & Vegetables" and "Eggs, Meat & Fish" have a relatively smaller number of products. Big Basket may want to consider expanding their offerings in these categories to cater to a wider range of customer needs.
- ▶ Overall, the distribution of products across categories provides insights into Big Basket's focus areas and potential areas for growth

DATA VISUALIZATION AND INSIGHTS

- **BAR CHART:** Plot the distribution of number of products in Top 15 Sub-category



DATA VISUALIZATION AND INSIGHTS

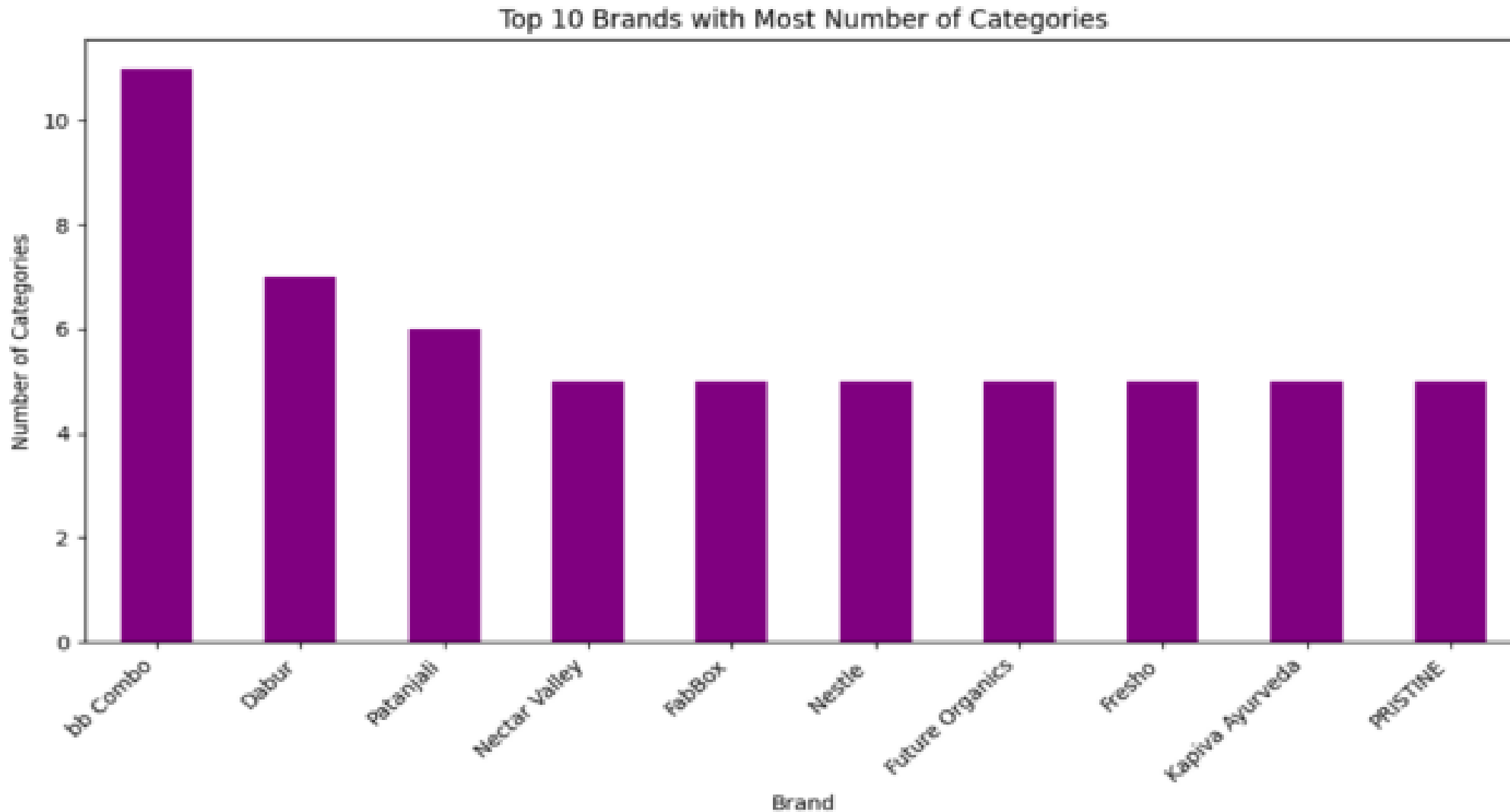
- BAR CHART: Plot the distribution of number of products in Top 20 Sub-category

- Key insights:

- ➤ "Skin Care" is the leading sub-category with the highest number of products. "Health & Medicine" follows closely behind "Skin Care" in terms of product count.
- ➤ There's a significant drop in product count after the top 3 categories ("Skincare", "Health & Medicine", and "Hair Care").
- ➤ It should be noted that all top 3 Sub-categories belongs to category "Beauty & Hygiene" estimating that Big Basket focus more on these categories.
- ➤ The remaining sub-categories have relatively similar product counts, with some fluctuations.

Data Visualization and Insights

- **BAR CHART:** Draw a visualization of Top 10 brands with most number of Categories



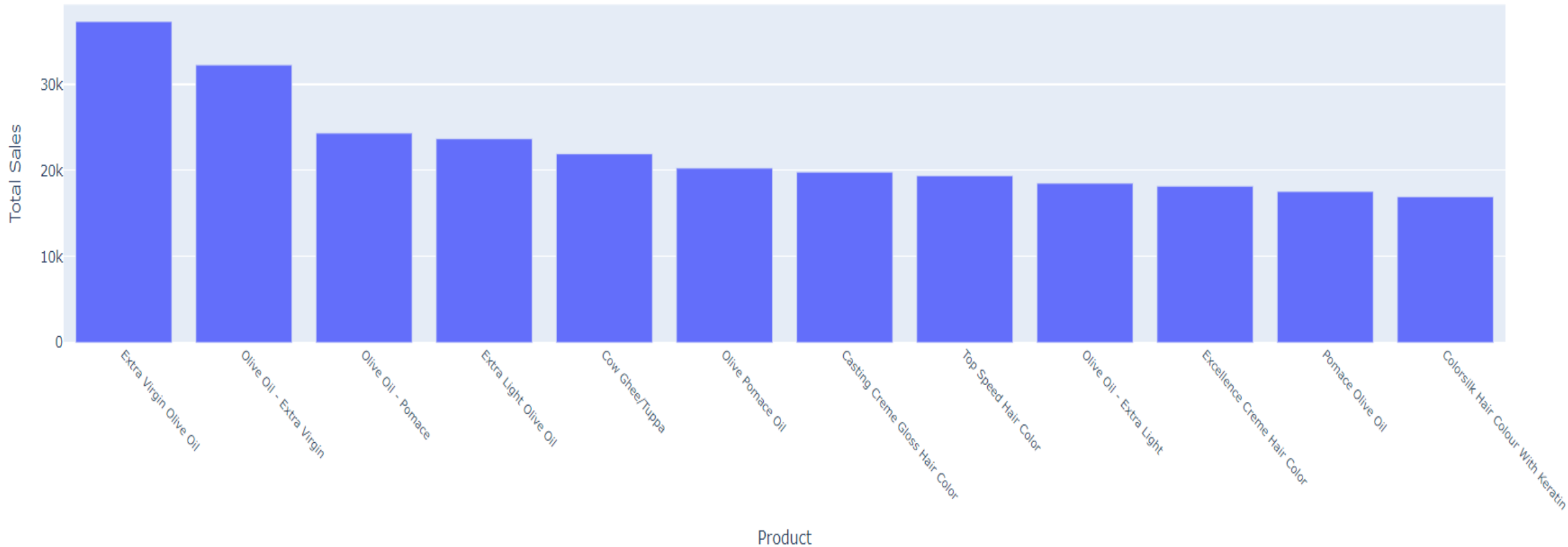
Data Visualization and Insights

- ▶ **BAR CHART:** Draw a visualization of Top 10 brands with most number of Categories.
- ▶ **Key insights:**
 - ▶ "bb Combo" is the clear leader in terms of the number of categories offered with offering products in all 11 categories. To increase sales, Big Basket should prioritize support for these brands.
 - ▶ There's a significant drop in the number of categories offered by the subsequent brands ("Dabur", "Patanjali", "Nectar Valley", and "FabBox").
 - ▶ The remaining four brands have a relatively similar number of categories, with slight variations

Data Visualization and Insights

- **BAR CHART:** Draw a visualization of Top 10 products by Total Sales.

Top 12 Products by Total Sales



Data Visualization and Insights

▣ **BAR CHART:** Draw a visualization of Top 10 products by Total Sales.

▣ **Key insights:**

▣ ➤ "Extra Virgin Olive Oil" is the top-selling product, with sales significantly higher than the other products.

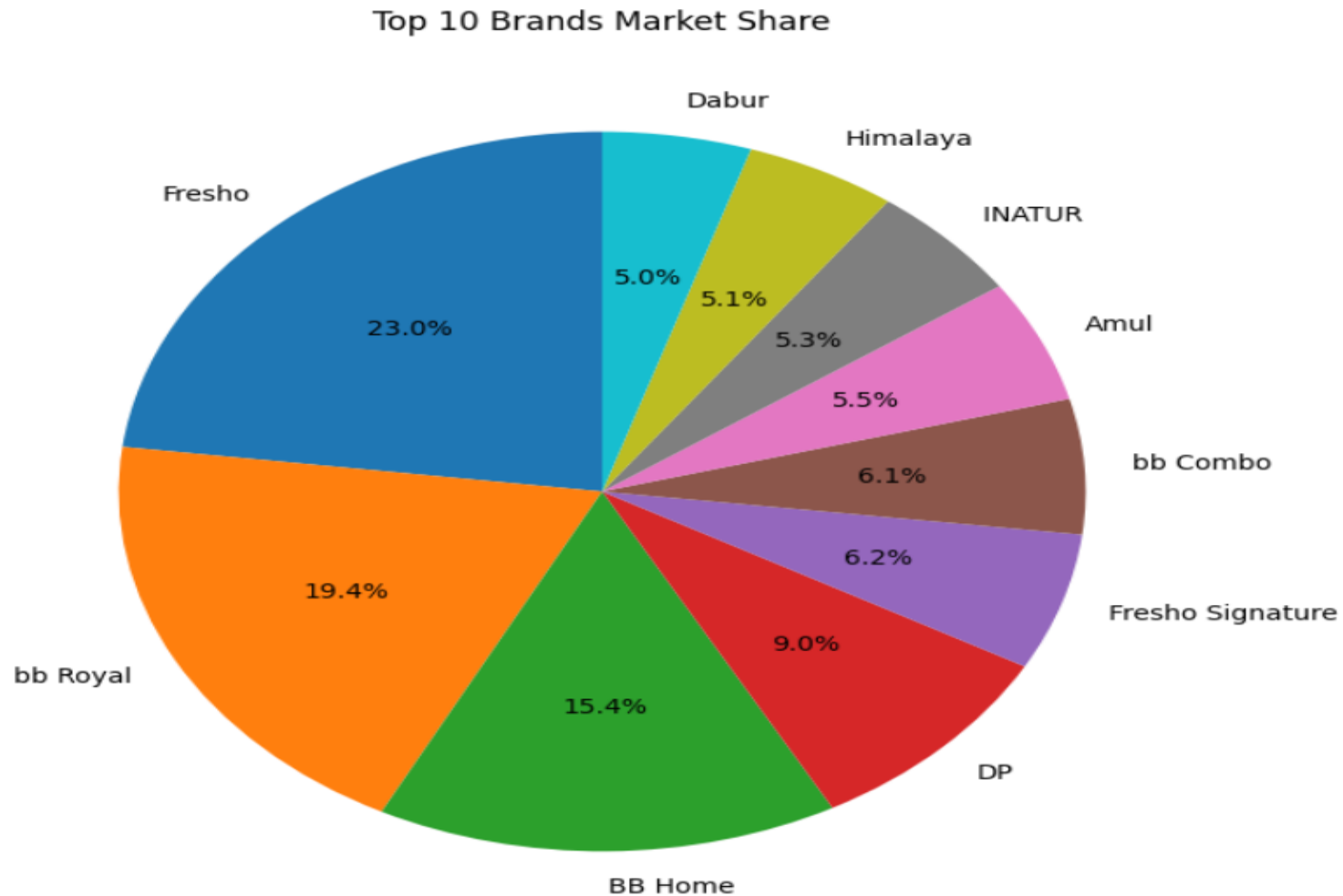
▣ ➤ "Color silk Hair Color With Keratin" is the second best-selling product, followed closely by "Cow Ghee/Tupac" and "Casting Creme Gloss Hair Color".

▣ ➤ The following three products, "Excellence Creme Hair Color", "Top Speed Hair Color", and "Olive Oil - Extra Virgin" have relatively similar sales figures, with "Excellence Creme Hair Color" being slightly ahead.

▣ ➤ Note that Top 10 products list is significantly dominated with "Beauty" (4 Products) and "Foodgrains/Gourmet" related items (5 Products). This finding aligns with our previous analysis in Bar chart of Question

Data Visualization and Insights

- ▶ Brand Analysis -
- ▶ **PIE CHART:** Draw a visualization of Top 10 Brands to show their Market Share



DATA VISUALIZATION AND INSIGHTS

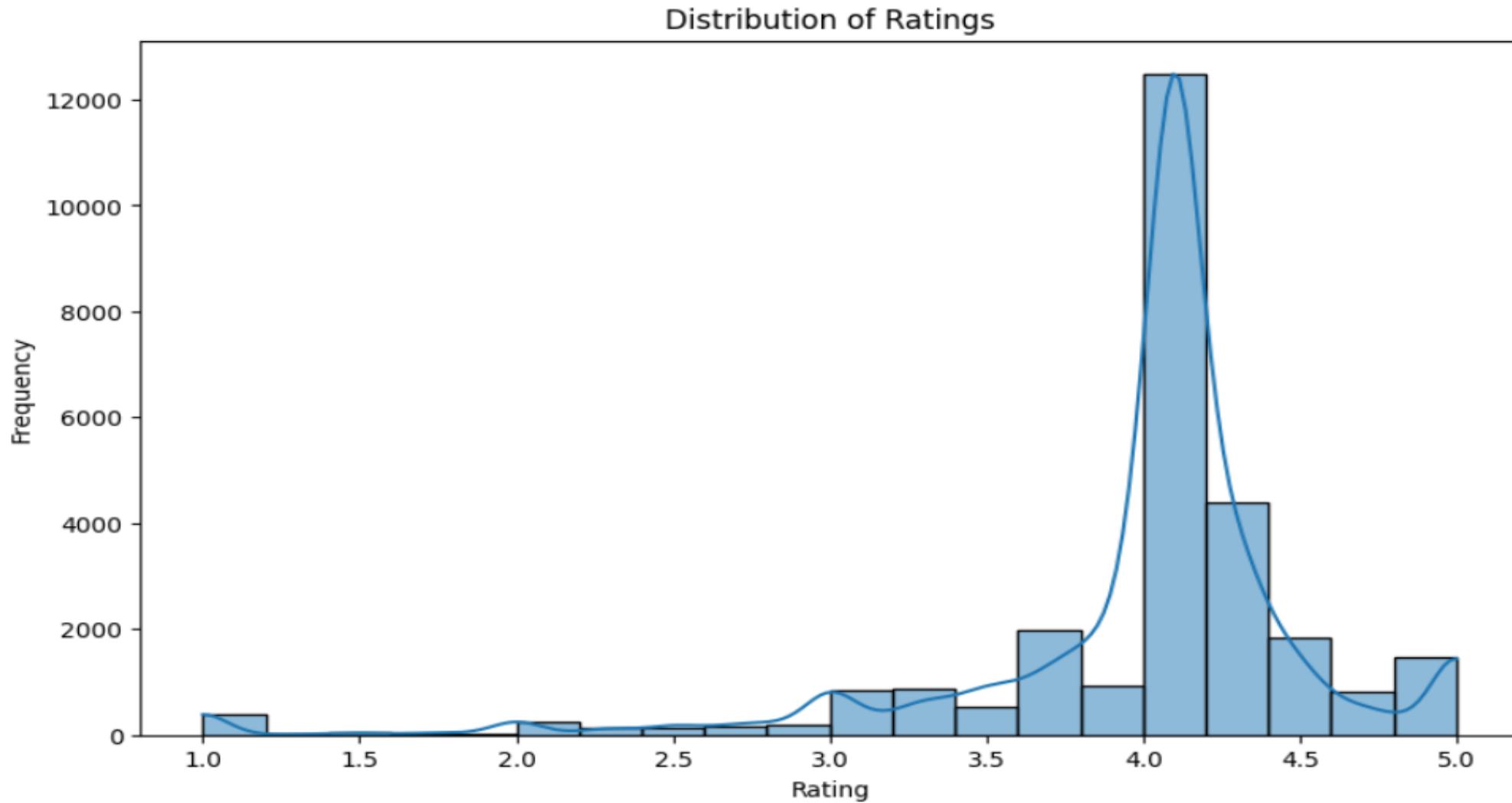
► Brand Analysis -

► **PIE CHART:**

- Draw a visualization of Top 10 Brands to show their Market Share. Key insights:
- "Fresh" commands the largest market share with 23.0% among the top 10 brands, indicating its strong presence and popularity on Big Basket.
- "bb Royal" and "BB Home" also hold significant market shares with 19.4% and 15.4%, suggesting their strong brand recognition and customer loyalty.
- The chart reveals that Big Basket offers a diverse product range, encompassing categories like Baby care ("bb Combo"), Cleaning & Household ("DP"), Garden & Pets ("BB Home").
- Big Basket could consider strategies to further strengthen the market position of "Fresh", "bb Royal", and "BB Home" while also exploring ways to increase the market share of other brands

DATA VISUALIZATIONS AND INSIGHTS

- ▢ Draw a visualization to show the Distribution of Product ratings.



DATA VISUALIZATION AND INSIGHTS

- HISTOGRAM: Draw a visualization to show the Distribution of Product ratings.

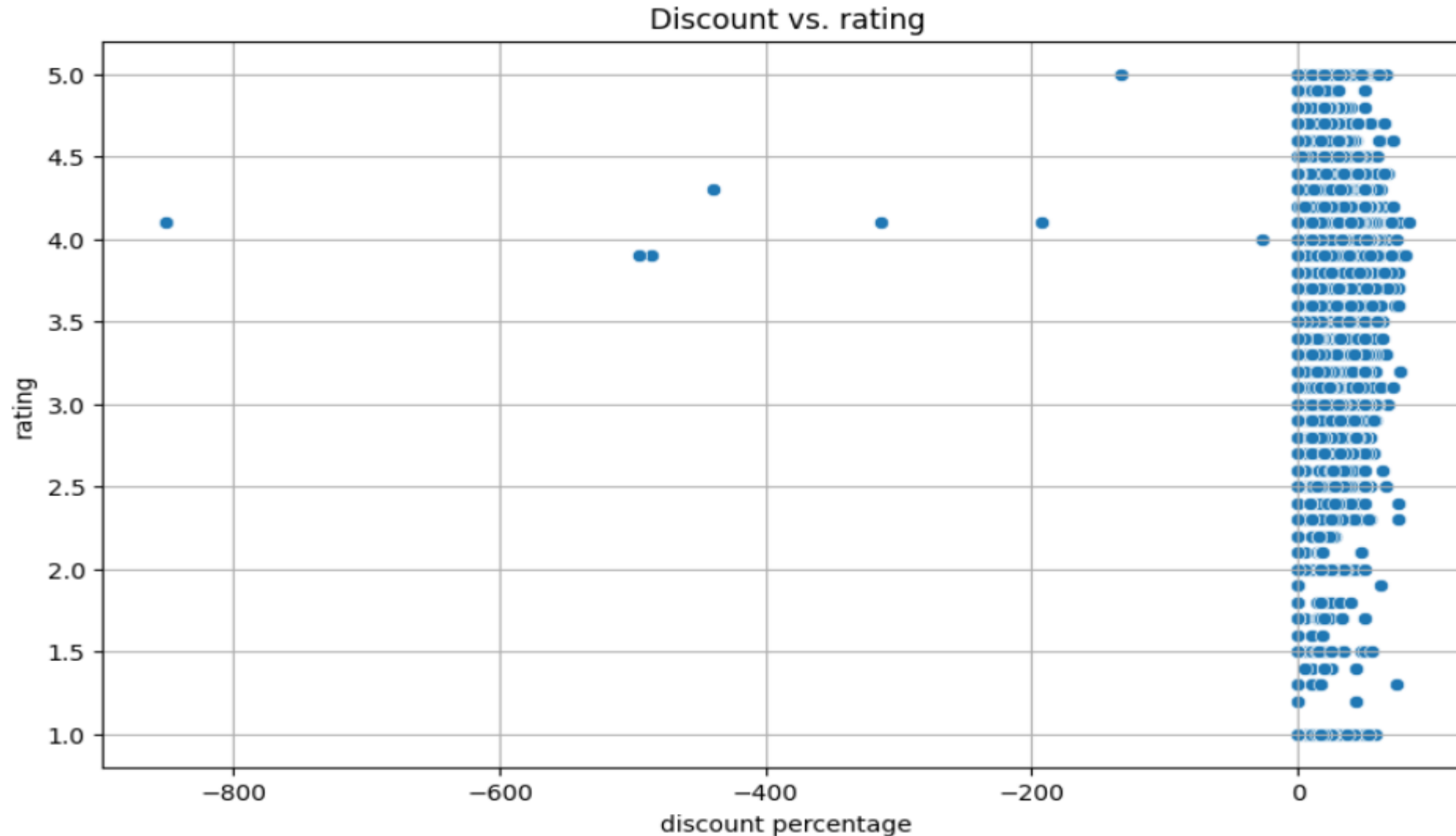
- Key insights:

- ➤ The distribution of product "Ratings" is heavily skewed towards higher ratings, with the majority of products receiving ratings of 4 or above. This suggests that customers are generally satisfied with the products being sold on the platform.

- ➤ I strongly believe that this is a significant finding as prioritizing customer satisfaction is a fundamental key to growth for service-oriented businesses

DATA VISUALIZATION AND INSIGHTS

- **SCATTER PLOT:** Draw a visualization to explore the relationship between Product Sale Price and Rating.

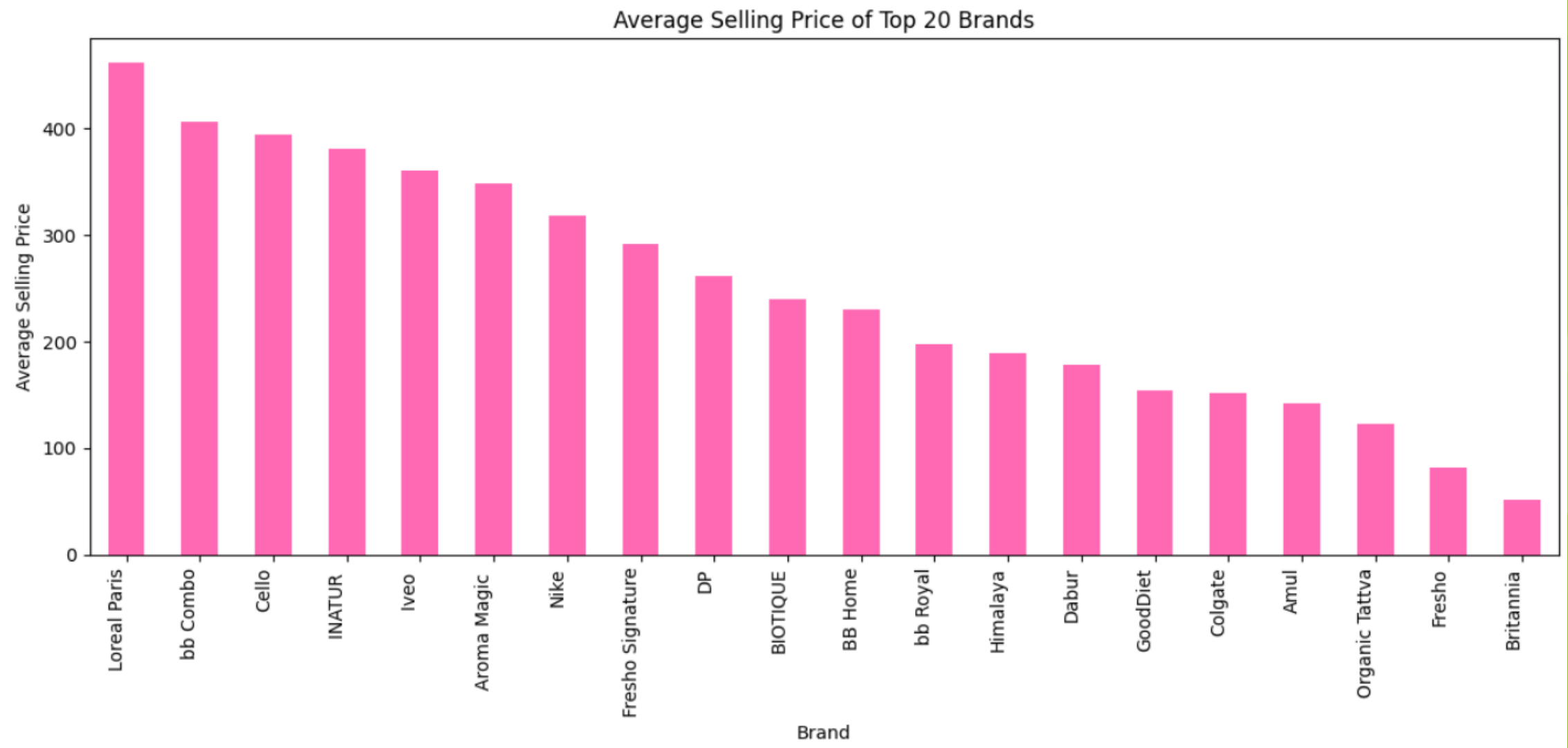


DATA VISUALIZATION AND INSIGHTS

- ▶ □ **Discount Analysis -**
- ▶ □ **SCATTER PLOT:** Draw a visualization to see if there's any relationship between Discount and Rating. Key insights:
 - ▶ □ ➤ **No Clear Correlation :** 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
- ▶ □ **Discount Strategy :** While discounts can attract customers, they may not be the primary driver of positive product ratings. Focus on overall product quality and customer experience to improve ratings.

DATA VISUALIZATION AND INSIGHTS

► Average Selling Price of Top 20 Brands



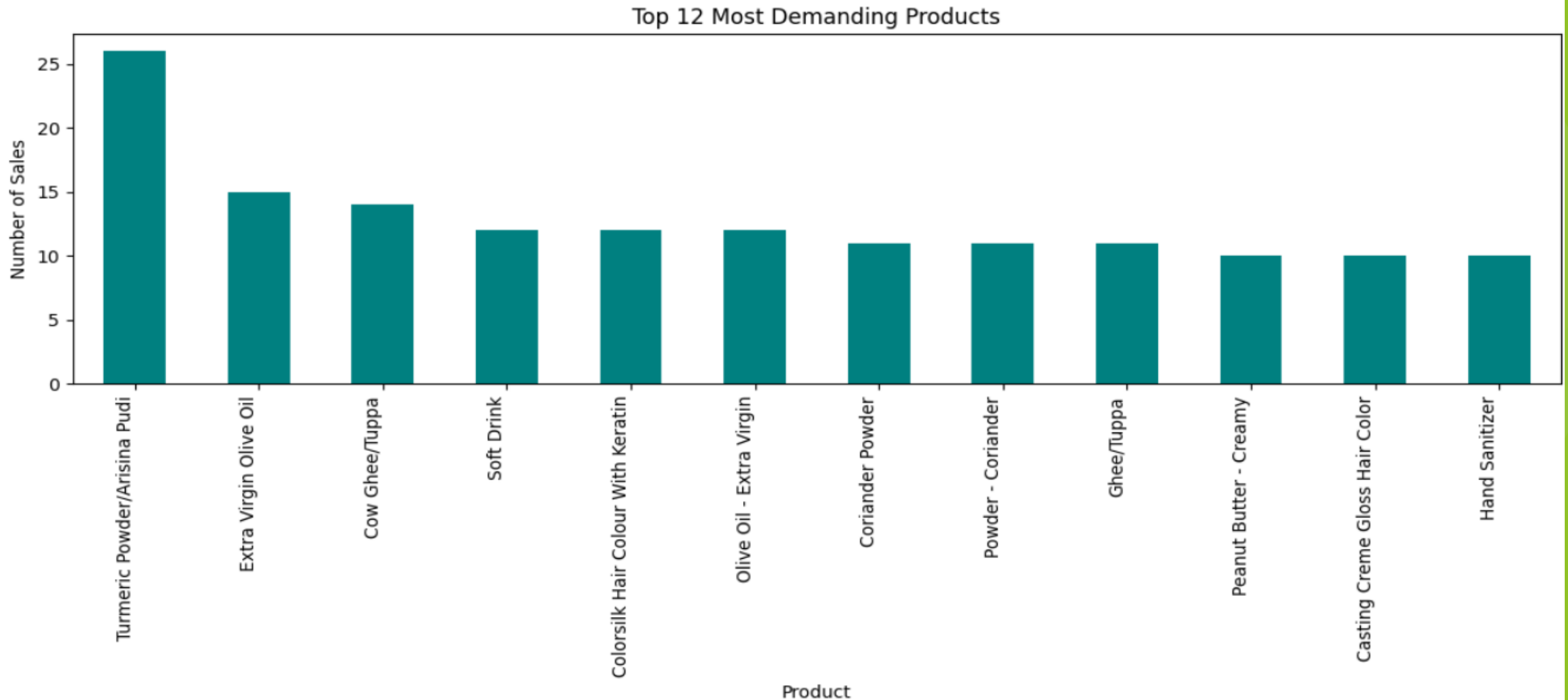
DATA VISUALIZATION AND INSIGHTS

□ Key Insights

- L'Oréal Paris : has the highest average selling price among the listed brands.
- Britannia : has the lowest average selling price. There's a wide range of average selling prices, indicating diversity in product categories and brand positioning.
- Beauty and personal care brands : like L'Oréal Paris, bb Combo, and INATUR generally have higher average selling prices.
- Food and FMCG brands : like Britannia, Fresho, and Amul generally have lower average selling prices.
- Nike : stands out as a non-beauty/FMCG brand with a relatively high average selling price.
- Overall, this chart provides a snapshot of the pricing landscape for top brands on the platform, which can be useful for competitive analysis and pricing strategy.

DATA VISUALIZATION AND INSIGHTS

► Top 12 most demanding products



DATA VISUALIZATION AND INSIGHTS

1. KEY INSIGHTS

2. Turmeric Powder Leads:

1. Turmeric Powder/Arisina Pudi is the most demanding product with over 25 sales, significantly higher than the others.

3. Even Distribution for Other Products:

1. The rest of the products show a relatively even distribution of demand, ranging between 10 and 15 sales.

4. Olive Oil and Ghee Are Popular:

1. Extra Virgin Olive Oil and Cow Ghee/Thuppa are among the top-selling products, highlighting demand for cooking and health-related items.

5. Hair Color and Skincare Items Included:

1. Colorsilk Hair Colour with Keratin and Casting Creme Gloss Hair Color are also popular, indicating a demand for beauty products.

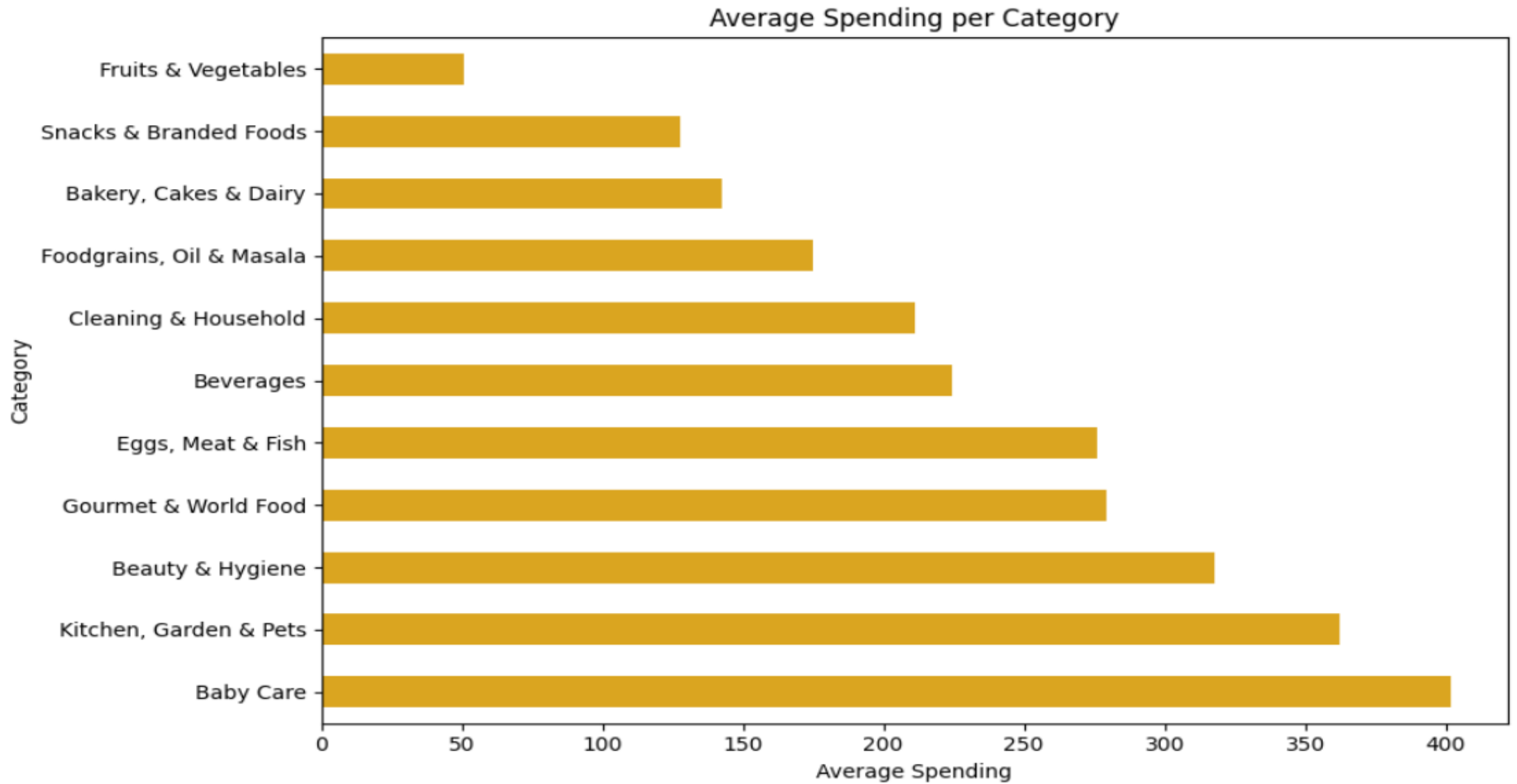
6. Variety in Top Products:

1. The top products include a mix of food items, beverages, and personal care products, showing a broad range of customer preferences.

- These insights could help prioritize stock management or promotions for these high-demand items.

DATA VISUALIZATION AND INSIGHTS

► Average spending per category



DATA VISUALIZATION AND INSIGHTS

KEY INSIGHTS

- 1. Identifies High-Spending Categories:** It determines which categories have the highest average spending, indicating potential areas of high customer demand and revenue generation.
- 2. Enables Category Comparison:** By visualizing the average spending across different categories, it helps to quickly compare their performance and relative importance in terms of customer expenditure.
- 3. Guides Business Strategies:** These insights can be valuable for informing various business strategies, such as pricing adjustments, inventory management, targeted marketing campaigns, and new product development decisions.

FINAL REPORT

- ▶ □ The exploratory data analysis (EDA) of the Big Basket e-commerce dataset has uncovered significant insights regarding product offerings, sales trends, pricing strategies, and customer feedback. By effectively managing missing data, eliminating outliers, and performing thorough data analysis, this study has generated actionable insights that can facilitate business growth, inspire innovation, and improve customer satisfaction within India's expanding online grocery market.
- ▶ □ The results of this EDA can provide a solid foundation for future research, strategic planning, and decision-making processes. This empowers Big Basket and other stakeholders in the online grocery industry to make informed choices, optimize their operations, and seize new opportunities in the fastevolving e-commerce landscape.

FINAL REPORT

▶ CONCLUSIONS :

- ▶ □ Big Basket's focus is on "Beauty & Hygiene" and "Gourmet & World Food" categories, with a strong emphasis on "Skin Care".
- ▶ □ "Fresho" is a key brand for Big Basket, while "BB Home" and "bb Royal" are major revenue drivers.
- ▶ □ These Top brands dominate the market in terms of product variety, sales, and market share.
- ▶ □ Discounts don't necessarily guarantee higher ratings; product quality and customer experience are crucial.
- ▶ □ Customers are mostly pleased with the products offered by Big Basket, and their overall experience is positive

FINAL REPORT

▶ RECOMMENDATIONS :

- ▶ □ Big Basket is supposed to leverage the popularity of "Fresho", "BB Home", and "bb Royal" for further growth.
- ▶ □ Big Basket must expand product offerings in categories like "Fruits & Vegetables" and "Eggs, Meat & Fish" to cater to a wider audience.
- ▶ □ Consider targeted discounts based on customer preferences and product categories. □ Prioritize product quality and customer experience to maintain high ratings.
- ▶ □ Firstly, Big Basket should typically concentrate on promoting products in popular categories and sub-categories, as these are significant revenue generators for the brand.



A TATA Enterprise

THANK YOU FOR READING

For coding part, kindly refer to below link :

https://colab.research.google.com/drive/1Ll6zIVWvVJSDDI6eENgNzDM1vL8_ZvaT?usp=sharing