EXPLORATORY DATA ANALYSIS

bigbasket

INTRODUCTION

In the ever-evolving world of e-commerce, online grocery shopping has become a major sector, transforming how people purchase everyday necessities. Since its inception in 2011, Big Basket has been a leading player in this field, establishing itself as India's largest online grocery store. Even with the rise of competitors such as Blinkit , Big Basket has maintained its leading position by effectively leveraging its broad customer base and successfully adapting to the online retail environment.

In the following sections of this project, we will adopt a systematic approach to Exploratory Data Analysis (EDA), including steps such as data loading, generating descriptive statistics, data profiling, detecting anomalies, and employing visualization methods. Our goal is to uncover actionable insights that can propel business growth, stimulate innovation, and strengthen Big Basket's leadership in India's expanding online grocery market.

DESCRIPTION

- ❑ I have conducted my work using Google Colab Notebook.
- ❑ The dataset has been imported from Google Drive.
- ❑ As we begin our Exploratory Data Analysis (EDA), I've named the dataset "bb".
- ❑ The dataset comprises of 27555 rows and 10 columns.
- ❑ For data cleaning, I have utilized libraries like Numpy , Pandas , Matplotlib , Plotly and Seaborn .
- ❑ Any duplicate entries that were found have also been removed.

```
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
```

```
bb.drop_duplicates(inplace=True)
```

```
bb.shape
```

```
(27555, 10)
```

DESCRIPTION

The Big Basket e-commerce dataset offers a thorough overview of the product range and sales patterns of India's largest online grocery store, Big Basket. It includes ten specific attributes, each providing important insights into different aspects of the company's operations. The attributes included in the dataset are as follows:

- **Index: A distinct identifier assigned to each record in the dataset.**
- **Product: The name or title of the product as it appears on the platform.**
- **Category: The general classification under which the product falls.**
- **Sub-Category: A more detailed classification within the general category.**
- **Brand: The brand linked to the product.**
- **Sale-Price: The price at which the product is sold to customers on the platform.**
- **Market-Price: The typical price of the product in the market.**
- **Type: The classification or nature of the product.**
- **Rating: The feedback or rating given by consumers for the product.**
- **Description: An in-depth explanation offering context about the dataset.**

DESCRIPTION

```
bb.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

```
bb.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
```

Data Cleaning & Pre-Processing:

Our dataset has a total of 8759 null values. Of these, 117 are found in categorical features, while 8642 are in numerical features.

- For the 'product' attribute, which has 1 null value filling in the missing entries with the 'Unknown' will help ensure data completeness.

```
bb['product'].fillna('Unknown', inplace=True)
```

- For the 'brand' attribute, which has 1 null value filling in the missing entries with the 'Unknown' will help ensure data completeness.

```
[15] bb['brand'].fillna('Unknown', inplace=True)
```

```
bb.isnull().sum()
```

	0
index	0
product	1
category	0
sub_category	0
brand	1
sale_price	6
market_price	0
type	0
rating	8636
description	115

dtype: int64

Data Cleaning & Pre-Processing:

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

```
[16] median_sale_price = bb['sale_price'].median()
     median_sale_price

     190.32

[17] bb['sale_price'].fillna(median_sale_price, inplace=True)
```

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

```
[19] median_rating = bb['rating'].median()
     median_rating

     4.1

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

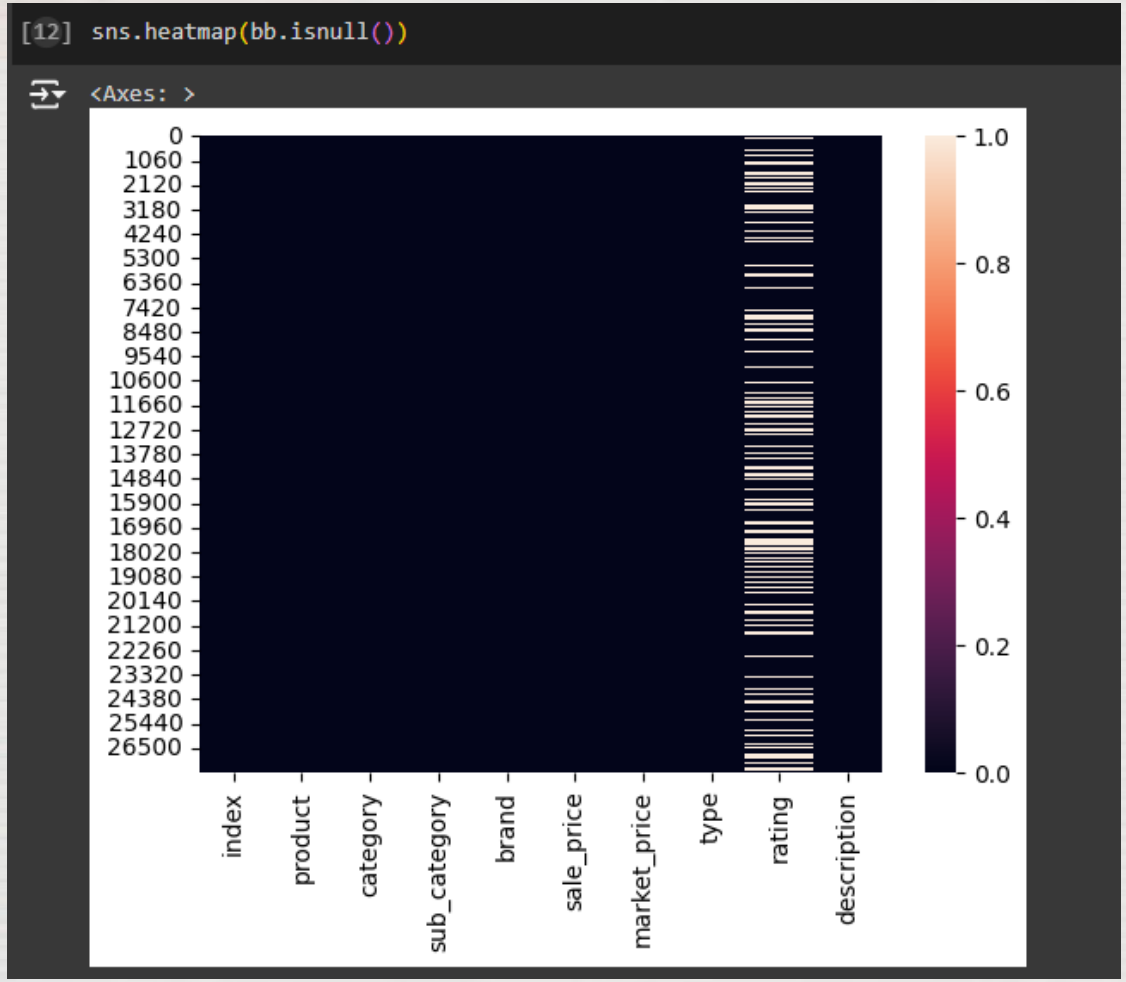
- For the 'description' attribute, which has 115 null value filling in the missing entries with the 'No description' will help ensure data completeness.

```
[22] bb['description'].fillna('No description', inplace=True)
```

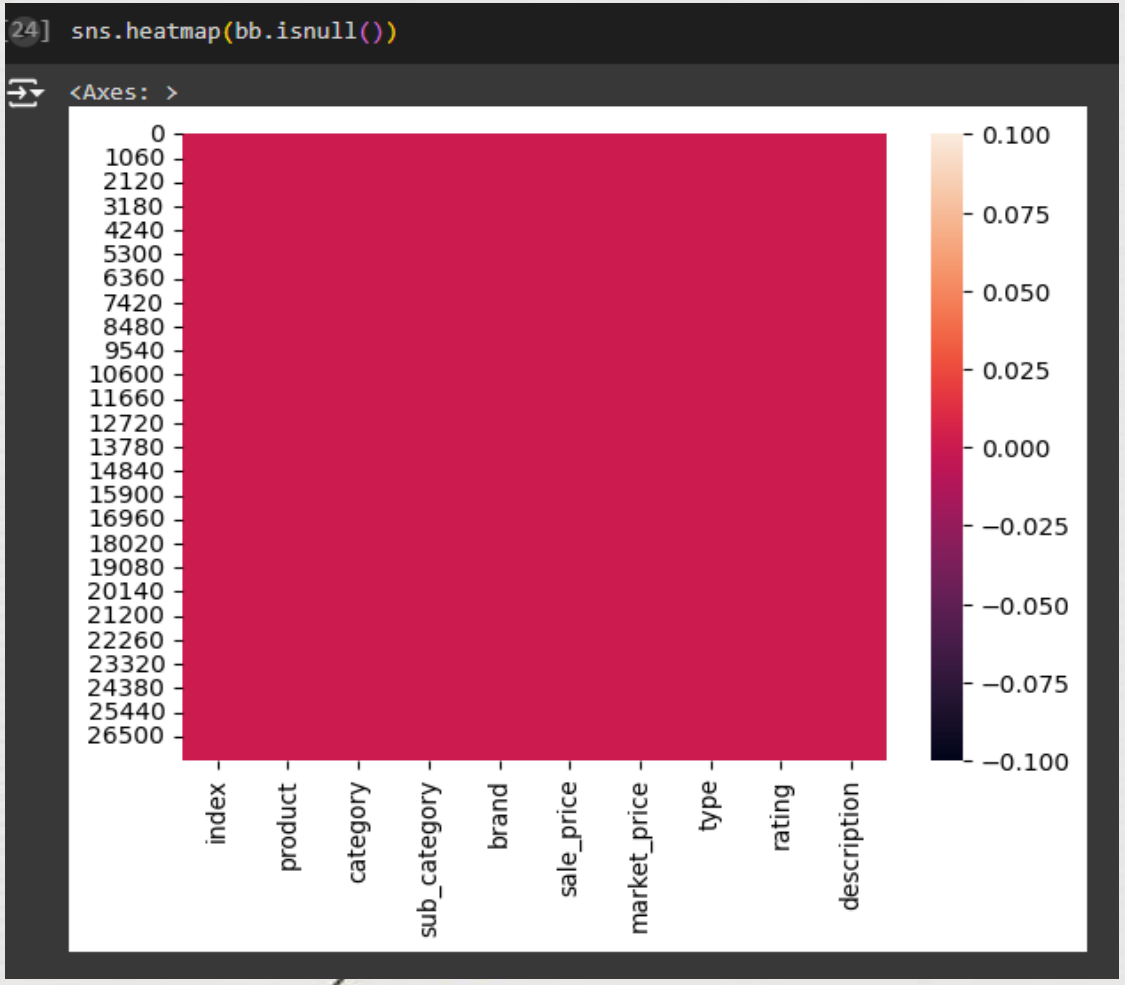
```
[23] bb.isnull().sum()

                    0
index               0
product             0
category            0
sub_category        0
brand               0
sale_price          0
market_price        0
type                0
rating              0
description         0

dtype: int64
```

HEATMAPS

Before Cleaning



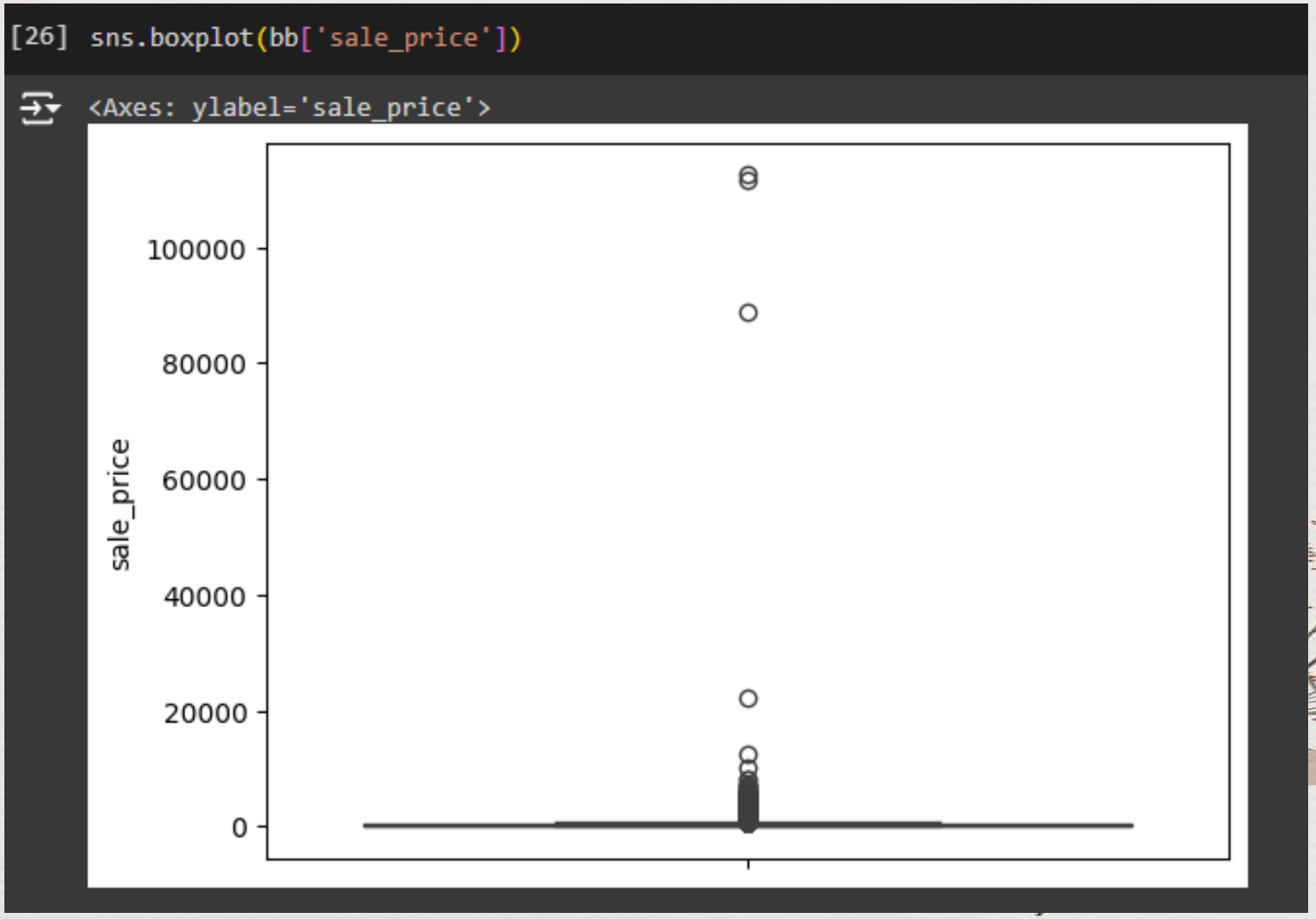
After Cleaning



REMOVING OUTLIERS

After generating a box plot for the 'sale-price' , we identified the presence of outliers in this column.

```
[26] sns.boxplot(bb['sale_price'])
```



REMOVING OUTLIERS

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

```
Q1 = bb['sale_price'].quantile(0.25)
print(f"Q1 is {Q1}")

Q3 = bb['sale_price'].quantile(0.75)
print(f"Q3 is {Q3}")
```

```
Q1 is 95.0
Q3 is 359.0
```

```
IQR = Q3 - Q1
print(f"IQR is {IQR}")
```

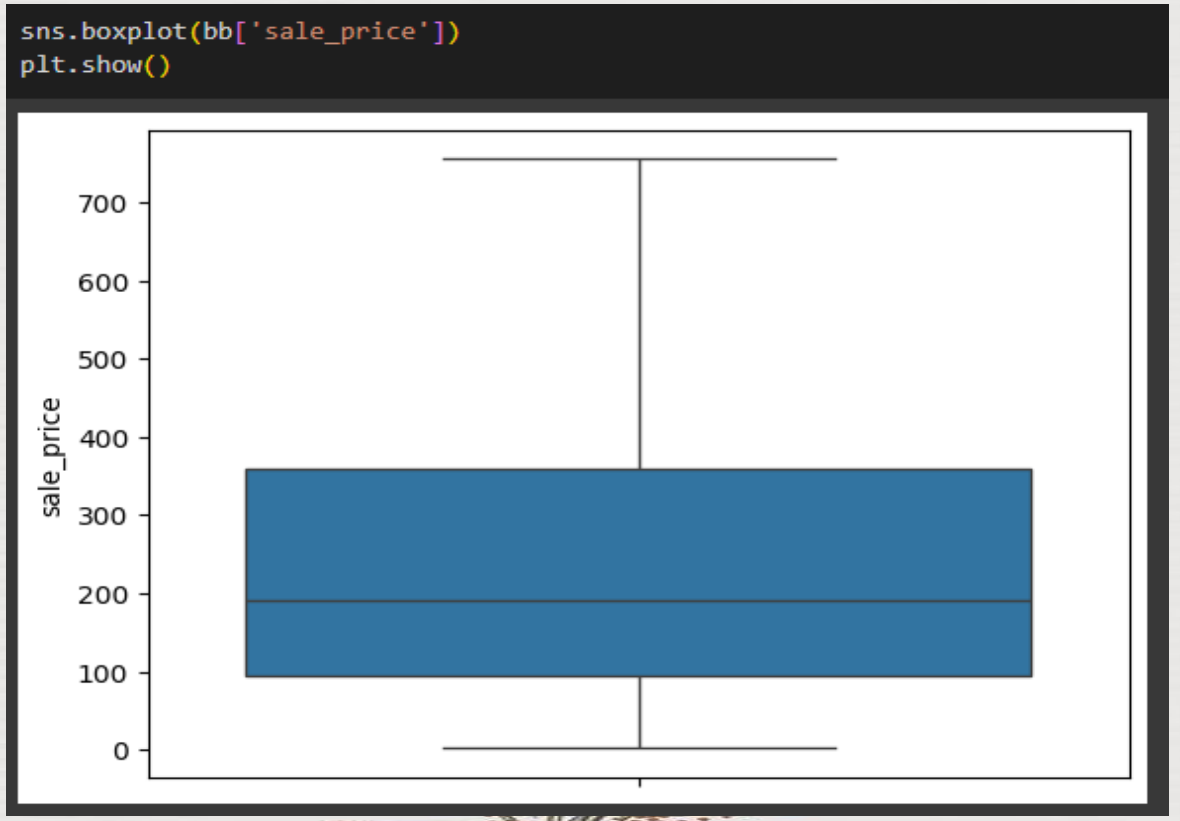
```
IQR is 264.0
```

```
lower_bound = Q1 - 1.5 * IQR
print(lower_bound)

upper_bound = Q3 + 1.5 * IQR
print(upper_bound)
```

```
-301.0
755.0
```

```
bb['sale_price'] = np.where(bb['sale_price'] < lower_bound, lower_bound, bb['sale_price'])
bb['sale_price'] = np.where(bb['sale_price'] > upper_bound, upper_bound, bb['sale_price'])
```

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

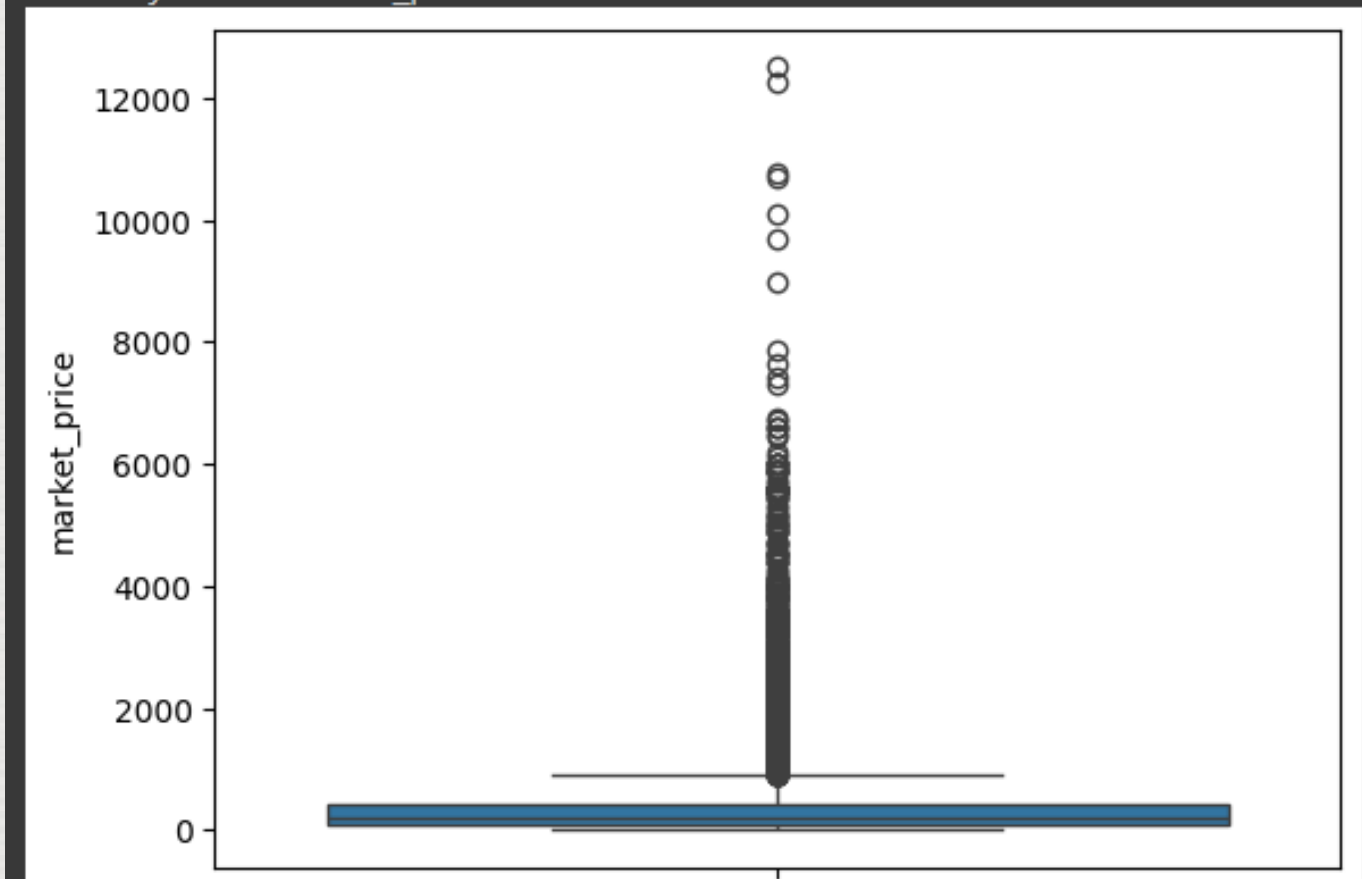


REMOVING OUTLIERS

After generating a box plot for the 'market-price' , we identified the presence of outliers in this column.

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

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



REMOVING OUTLIERS

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

```
Q1 = bb['market_price'].quantile(0.25)
print(f"Q1 is {Q1}")

Q3 = bb['market_price'].quantile(0.75)
print(f"Q3 is {Q3}")
```

```
Q1 is 100.0
Q3 is 425.0
```

```
IQR = Q3 - Q1
print(f"IQR is {IQR}")
```

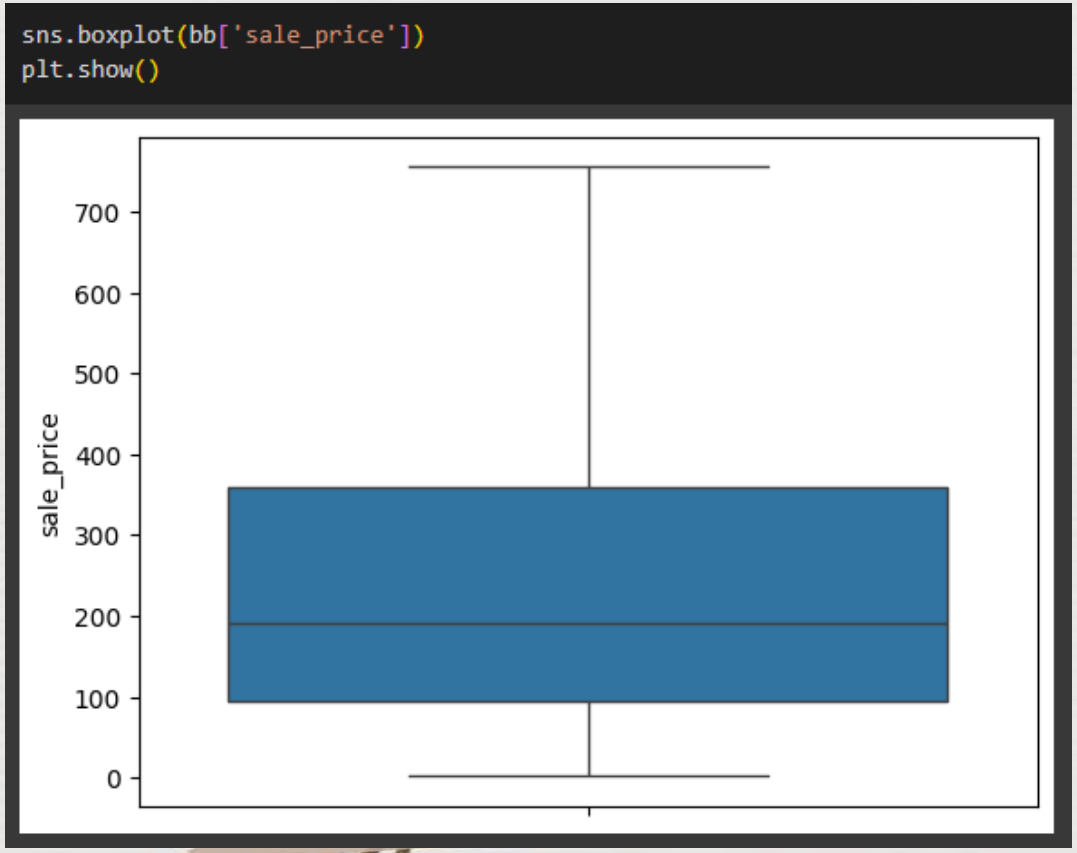
```
IQR is 325.0
```

```
lower_bound = Q1 - 1.5 * IQR
print(lower_bound)

upper_bound = Q3 + 1.5 * IQR
print(upper_bound)
```

```
-387.5
912.5
```

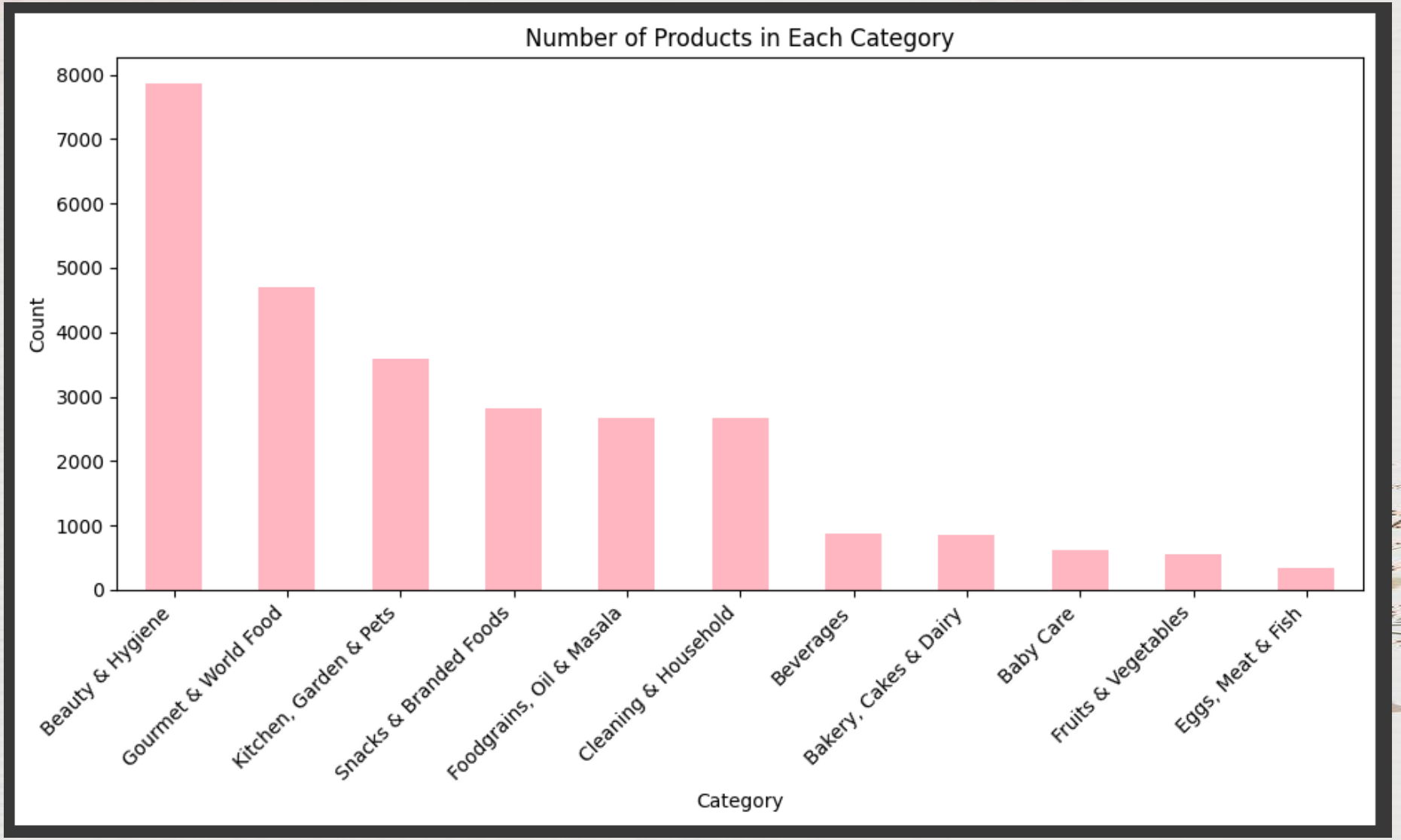
```
bb['market_pricev'] = np.where(bb['market_price'] < lower_bound, lower_bound, bb['market_price'])
bb['market_price'] = np.where(bb['market_price'] > upper_bound, upper_bound, bb['market_price'])
```

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



DATA VISUALIZATION

- **No of product in each category**

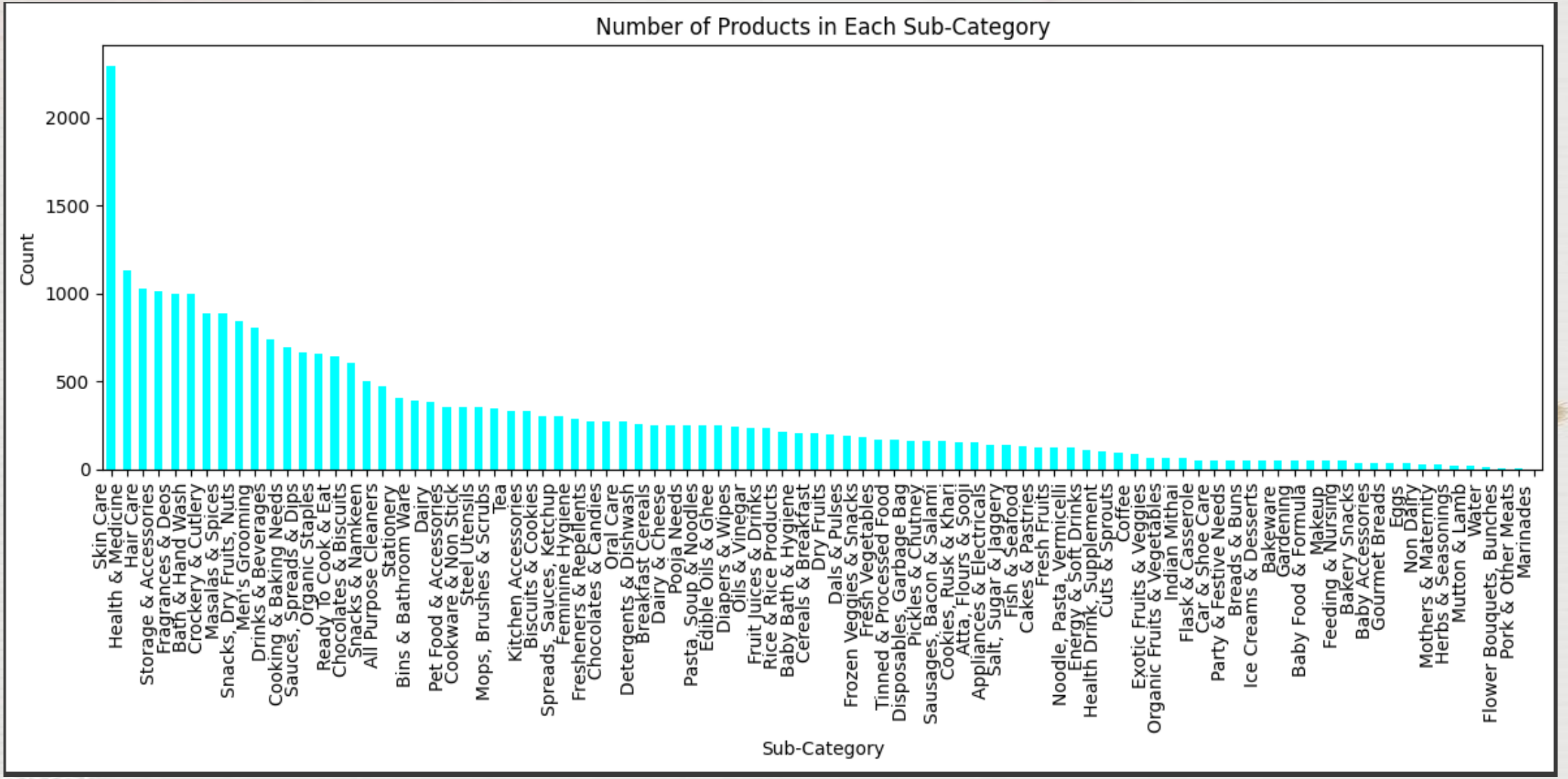


Key Insights

- 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.

DATA VISUALIZATION

❖ No of product in each sub-category



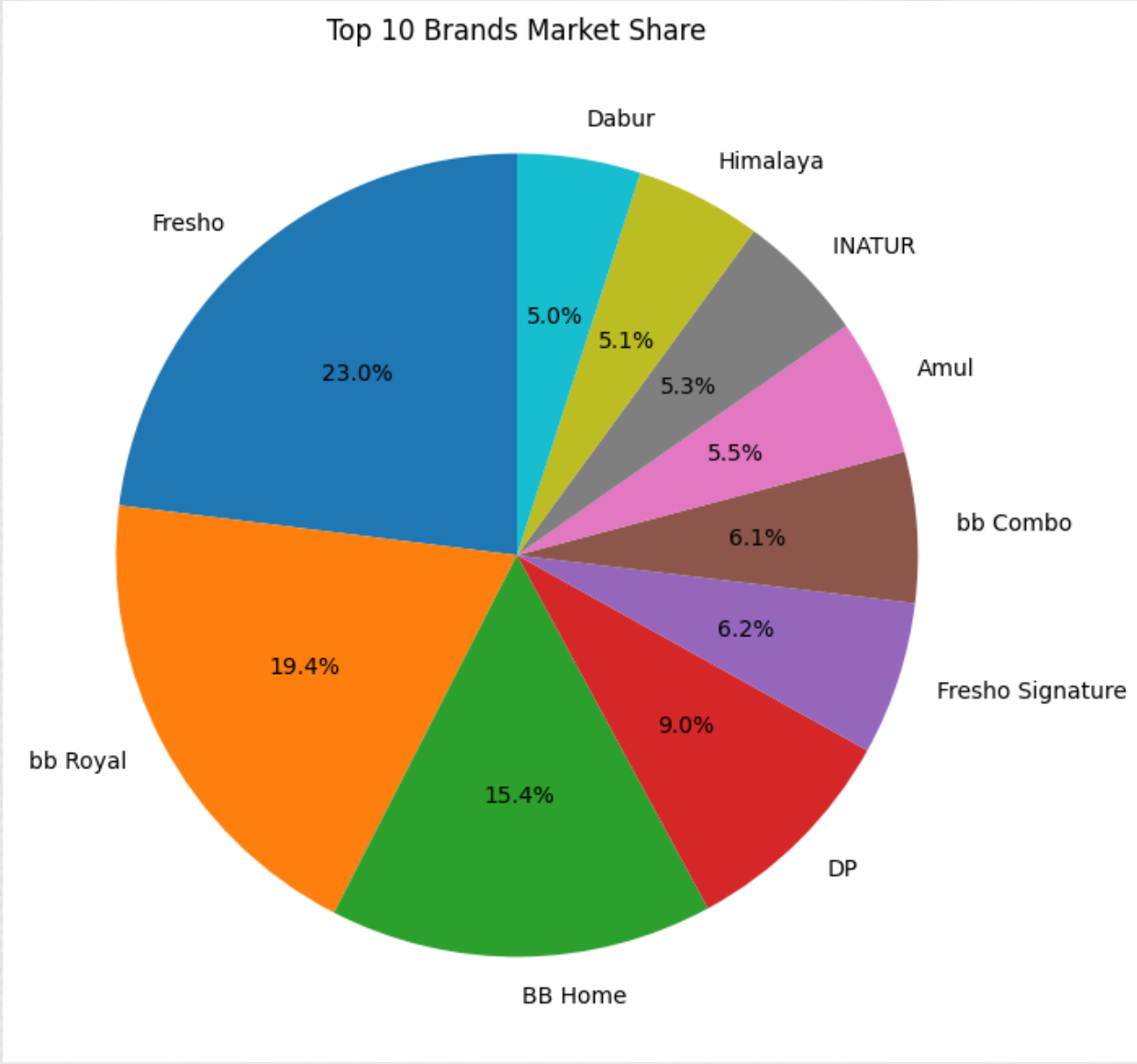
Key Insights

- **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.

- Overall, the graph shows that there is a wide range of products available in the grocery store, with a focus on personal care and health products.

DATA VISUALIZATION

❖ Market Share of Top 10 Brands



Key Insights

- **Fresho dominates :** Fresho , Big Basket's private label, has the largest market share at 23%. This indicates its popularity and success among customers.

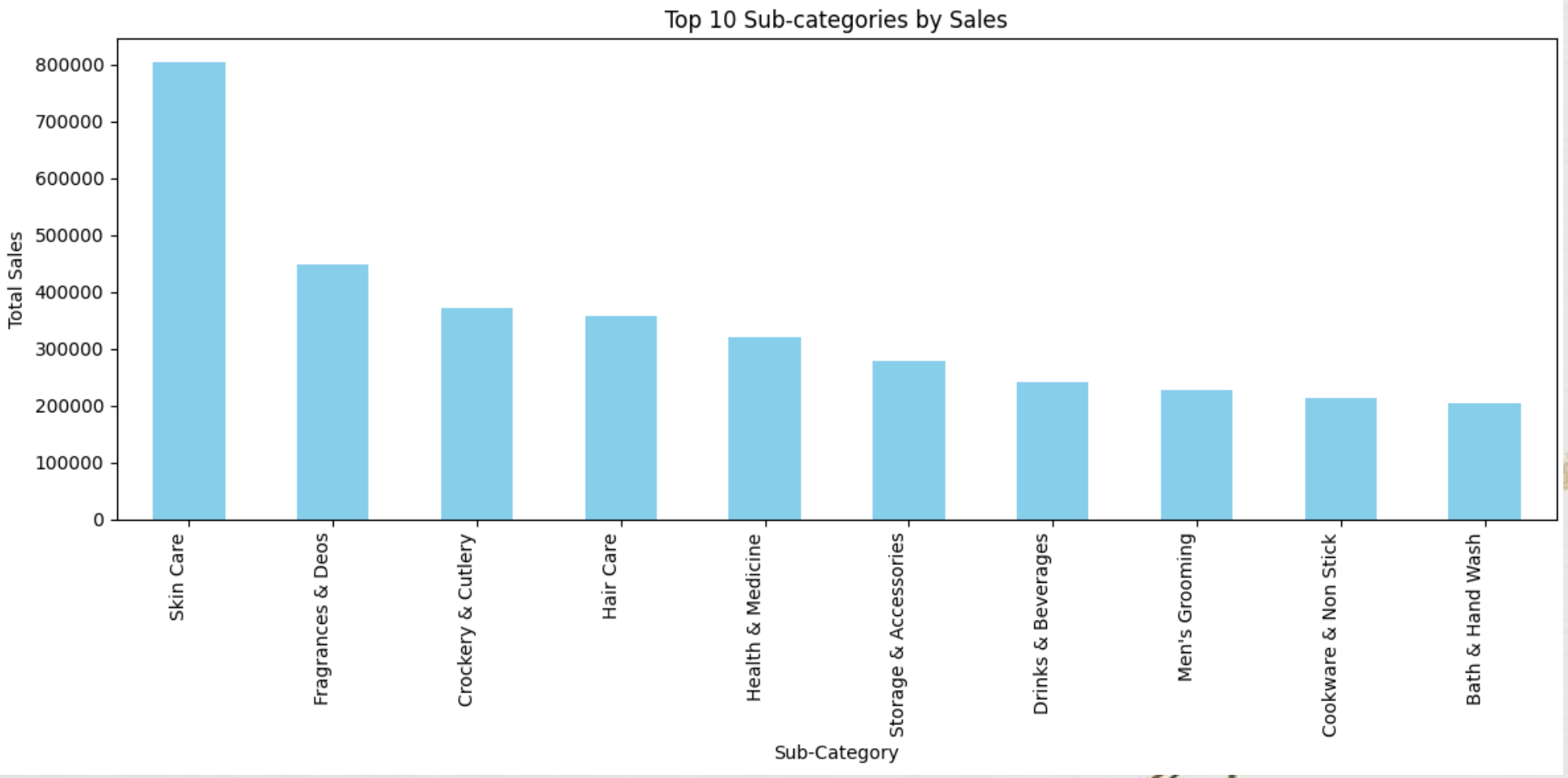
- **bb Royal and BB Home are strong performers :** These two brands, also likely Big Basket's own, hold significant market shares at 19.4% and 15.4% respectively. This further demonstrates the success of Big Basket's private label strategy.

- **Established brands have a presence :** Well-known brands like Dabur, Himalaya, and Amul also feature in the top 10, although with smaller shares compared to Big Basket's own brands.

- **Diverse product offerings :** The chart includes brands from various categories, such as personal care (Dabur, Himalaya), dairy (Amul), and packaged foods (Fresho Signature, bb Combo). This suggests Big Basket offers a wide range of products to cater to different customer needs.

DATA VISUALIZATION

❖ Top 10 Sub-categories by Scale



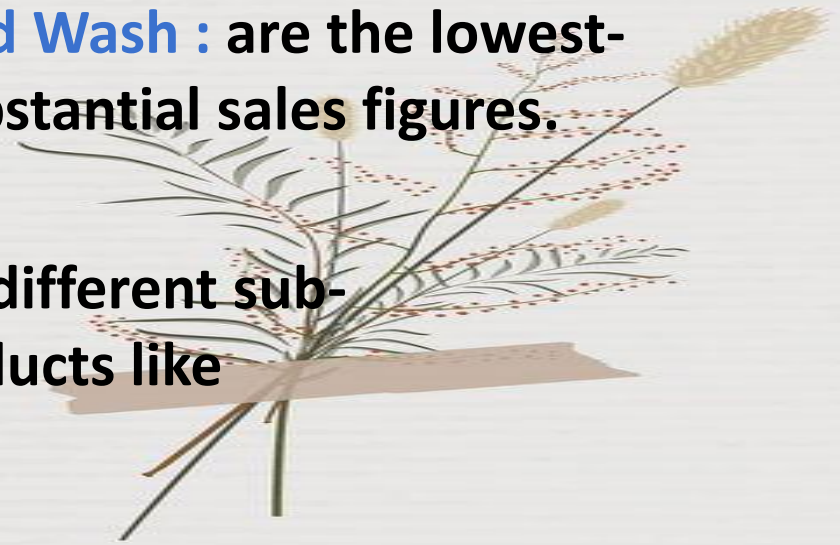
Key Insights

- **Skin Care :** is the top-selling sub-category, with significantly higher sales than any other category.

- **Fragrances & Deos and Crockery & Cutlery :** are the next best-selling categories, with similar sales figures.

- **Hair Care, Health & Medicine, Storage & Accessories, and Drinks & Beverages :** all have relatively similar sales, falling in the mid-range.

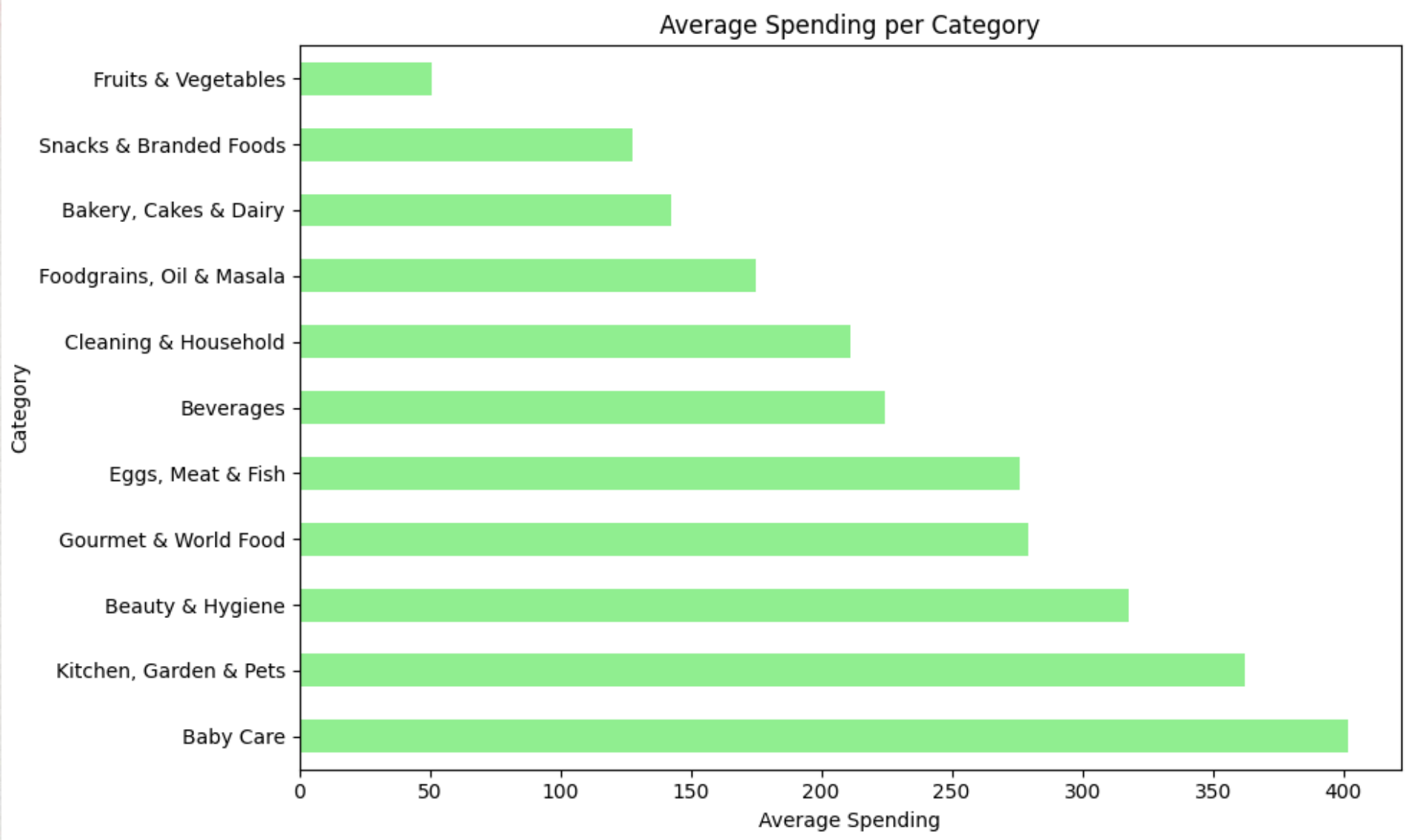
- **Men's Grooming, Cookware & Non Stick,and Bath & Hand Wash :** are the lowest-selling categories on the graph, though they still have substantial sales figures.



❖ Overall, the graph indicates a wide range of sales across different sub-categories, with a strong emphasis on personal care products like skincare and fragrances

DATA VISUALIZATION

❖ Average Spending per Category



Key Insights

- **Highest Spending :** The category with the highest average spending is "Fruits & Vegetables." This indicates that consumers prioritize fresh produce and are willing to allocate a larger portion of their budget to these items.

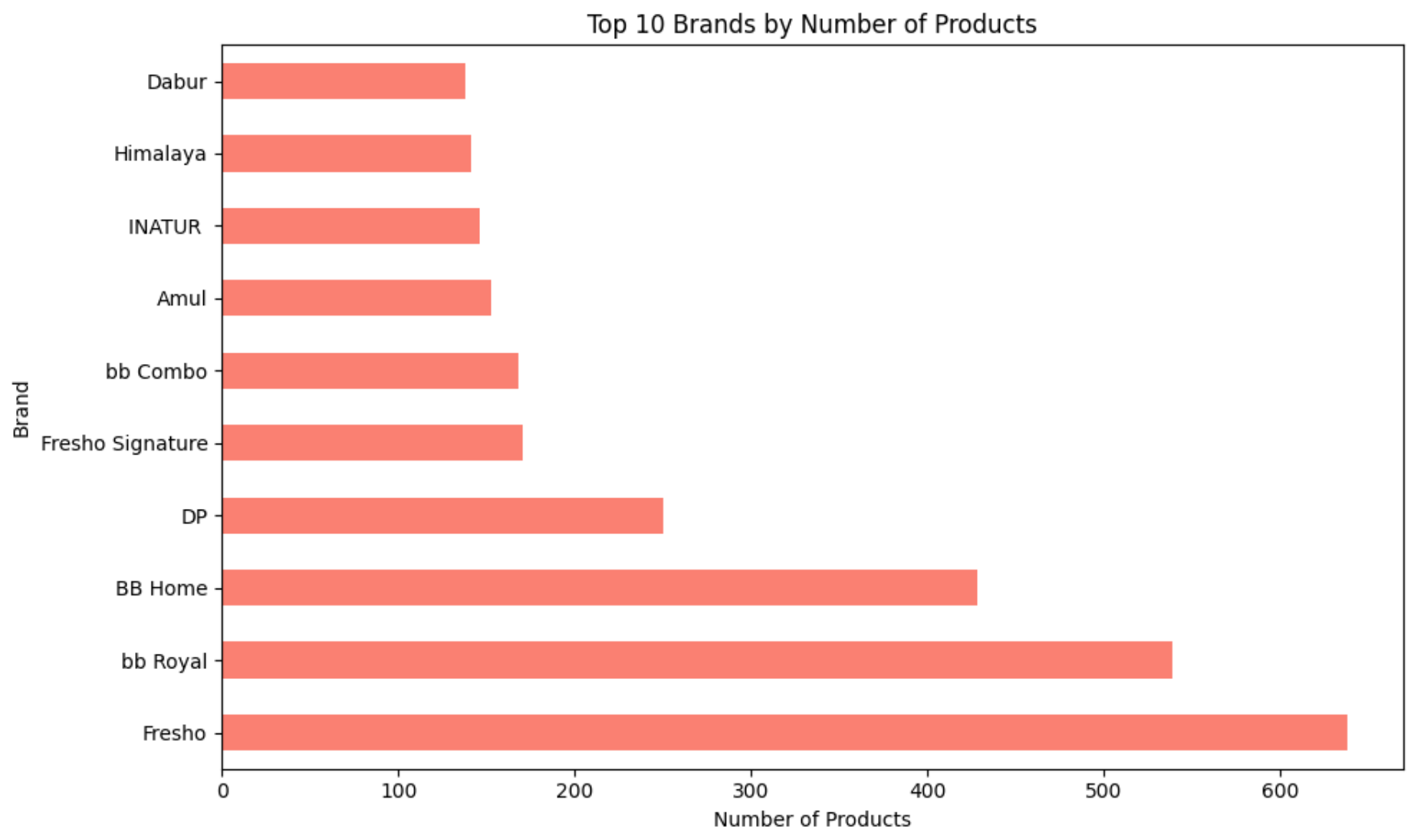
- **Moderate Spending :** Categories like "Snacks & Branded Foods," "Bakery, Cakes & Dairy," and "Food-grains", Oil & Masala" show moderate average spending. These are essential items in a typical household, reflecting consistent demand.

- **Lower Spending :** Categories such as "Baby Care," "Kitchen, Garden & Pets," and "Beauty & Hygiene" exhibit lower average spending. This could be due to less frequent purchases or a smaller proportion of the budget allocated to these items.

- Overall, the graph provides a clear picture of consumer spending patterns across different grocery categories. This information can be valuable for retailers and marketers to understand consumer preferences and tailor their strategies accordingly.

DATA VISUALIZATION

❖ Top 10 Brands by Number of Products



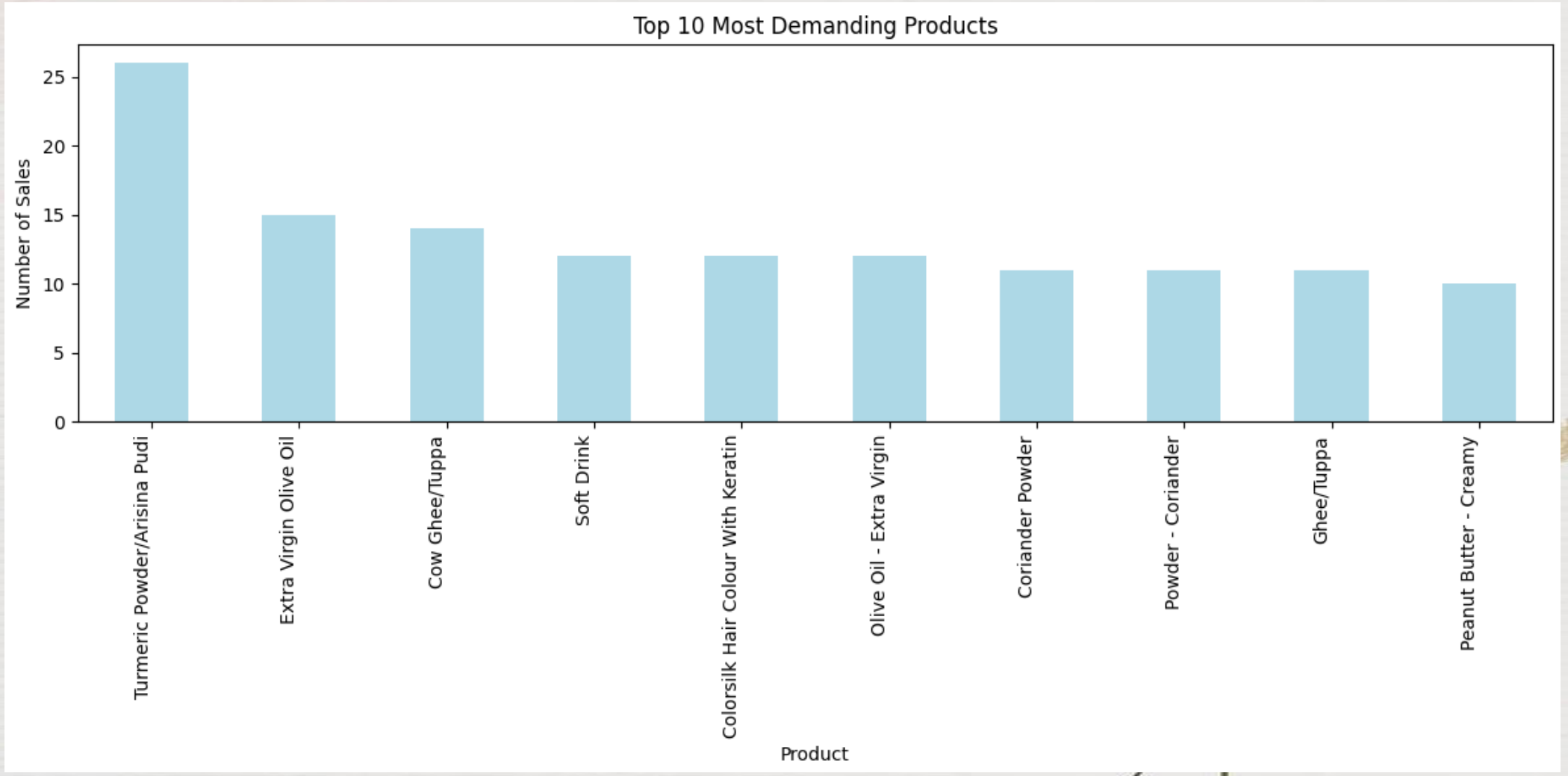
Key Insights

- **Fresho :** is the leading brand with the most extensive product range, offering over 600 different products.
- **BB Royal and BB Home :** follow, with a substantial number of products, exceeding 400.
- **DP and Fresho Signature :** have a similar product range, each offering around 300 products.
- **BB Combo, Amul, INATUR, and Himalaya :** have a comparable number of products, ranging from approximately 200 to 250.
- **Dabur :** has the least number of products among the top 10 brands, with just over 100 products.

- Overall, the graph highlights the diversity in product offerings among the top 10 brands, with Fresho being the dominant player.

DATA VISUALIZATION

❖ Top 10 Brands by Number of Products



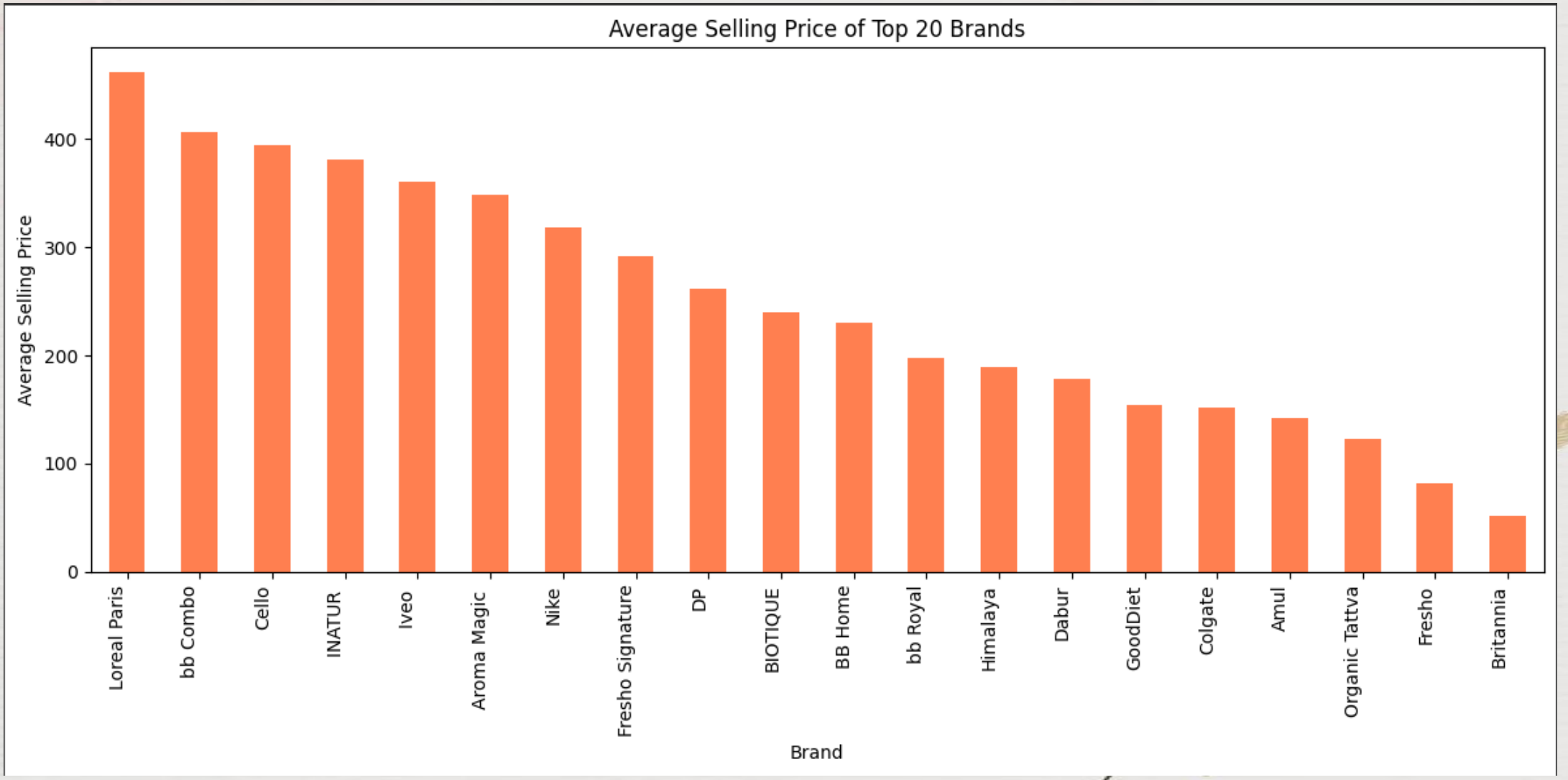
Key Insights

- **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 products**

DATA VISUALIZATION

❖ Average Selling Price of Top 20 Brands



Key Insights

- **Loreal 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 Loreal 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.

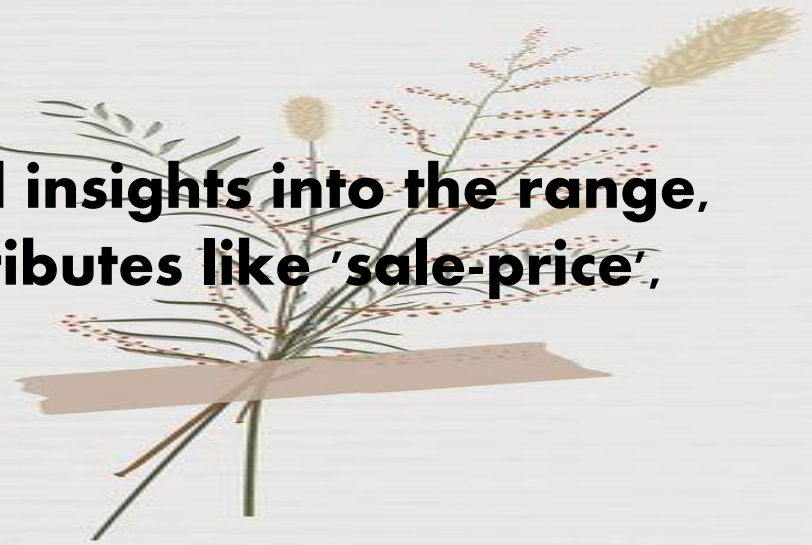
- **Nike :** stands out as a non-beauty/FMCG brand with a relatively high average selling price.

❖ 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.

FINAL REPORT

- ✓ **The Big Basket E-commerce dataset provides valuable insights into India's leading online grocery store's product range, sales trends, and customer feedback. With ten key attributes, this dataset offers a thorough understanding of Big Basket's operational metrics, pricing strategies, and customer preferences. This exploratory data analysis (EDA) seeks to reveal patterns, trends, and insights that can guide strategic decisions, improve inventory management, and enhance the overall shopping experience for customers.**

- ✓ **A summary of the dataset's key statistics provided insights into the range, distribution, and central tendencies of numeric attributes like 'sale-price', 'market-price' and 'rating**



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 fast-evolving e-commerce landscape.**

FINAL REPORT

- ✓ **The Big Basket E-commerce dataset provides a comprehensive overview of India's largest online grocery supermarket's product offerings, sales dynamics, and customer feedback. With ten key attributes encompassing product details, pricing information, brand categorization, and customer ratings, this dataset serves as a valuable resource for understanding the operational metrics and consumer preferences shaping the online grocery sector in India.**

- ✓ **ATTRIBUTES :**
- ❑ **Index:** Unique identifier for each product entry. Product: Title or name of the grocery items listed.
- ❑ **Category:** Broad classification of products such as fruits, vegetables, dairy products, etc.
- ❑ **Sub-Category:** Specific classification within each broader category.
- ❑ **Brand:** Brand or manufacturer associated with each product.
- ❑ **Sale-Price:** Price at which products are offered to consumers. Market-Price: Standard market price of the products.
- ❑ **Type:** Nature or characteristics of the products.
- ❑ **Rating:** Consumer ratings or feedback received by each product.
- ❑ **Description:** Detailed narrative describing the dataset and its context.

FINAL REPORT

- ✓ **The Big Basket e-commerce dataset provides a robust foundation for comprehending India's online grocery market, offering invaluable insights into product demand, pricing approaches, customer feedback, and industry trends. By meticulously preparing the data, eliminating outliers, and conducting thorough exploratory analysis, this dataset empowers stakeholders to make well-informed decisions, streamline operations, and seize emerging opportunities in the fast-paced e-commerce landscape**

- ✓ **In summary, the Big Basket e-commerce dataset not only serves as a critical resource for immediate operational improvements but also lays the groundwork for long-term strategic planning. By harnessing the power of data analytics, stakeholders can navigate the complexities of the online grocery sector, adapt to changing market dynamics, and ultimately drive sustainable growth in this rapidly evolving landscape.**

THANKS FOR READING--->>>

FOR CODING PART-->>>

https://colab.research.google.com/drive/1TLVJCD2WmtGhFVu3pd-6muCbYhAB_Lwy?usp=sharing