# Exploratory Data Analysis of Big Basket!

Project by
KUSHAGR SAXENA

# ABOUT BIG BASKET:

- *Big Basket is one of India's largest online grocery delivery services. Headquartered in Bangalore & launched in 2011, it offers a wide range of products including Fruits, Vegetables, Dairy products, Snacks, Beverages, and Household items. The platform allows customers to order groceries online and have them delivered to their doorstep, aiming to provide convenience and a wide selection of items.*
- *Big Basket was the first online grocer in India with Mr. Shahrukh Khan as brand ambassador, it has become a popular choice for people looking to shop for groceries without leaving their homes. In 2021, Big Basket was acquired by Tata Group, which has further integrated it into its retail operations.*
- *Over the years, despite facing competition from emerging players like JioMart and Blinkit, Big Basket has sustained its dominance, leveraging its expansive customer base and adept transition to online retail.*

# INTRODUCTION:

- *This 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 -**

***I. Sales Data Analysis:* Understanding of General Sales performance and patterns.**

***II. Top Selling Products:* Identify which products are driving High Sales for the brand.**

***III. Discount Analysis:* Measure Discounts offered on products and analyze their impact on Sales.**

***IV. Handling Missing Values:* Ensuring data quality by identifying and Handling Missing Values appropriately.**

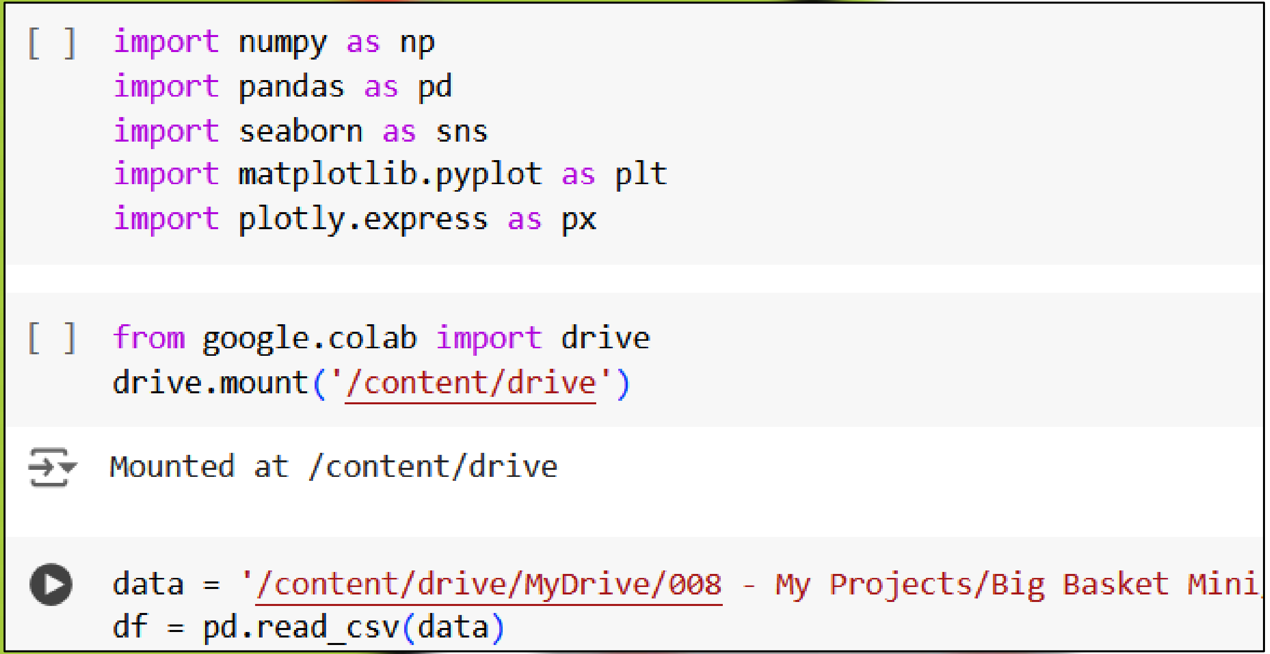
***V. Anomaly Detection and Handling:* Identify and manage Anomalies to maintain data integrity.**

***VI. Consumer Insights:* Ratings and product reviews provide valuable feedback that can guide product improvements and marketing efforts.**

***VII. 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 Colaboratory 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.*

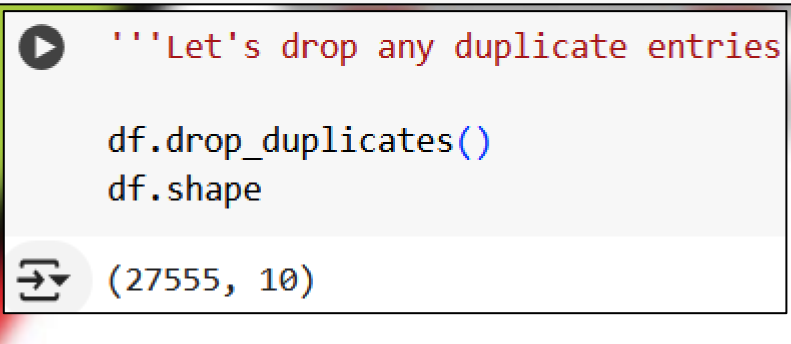


```
[ ] import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    import plotly.express as px
```

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

```
Mounted at /content/drive
```

```
data = '/content/drive/MyDrive/008 - My Projects/Big Basket Mini
df = pd.read_csv(data)
```



```
'''Let's drop any duplicate entries

df.drop_duplicates()
df.shape
```

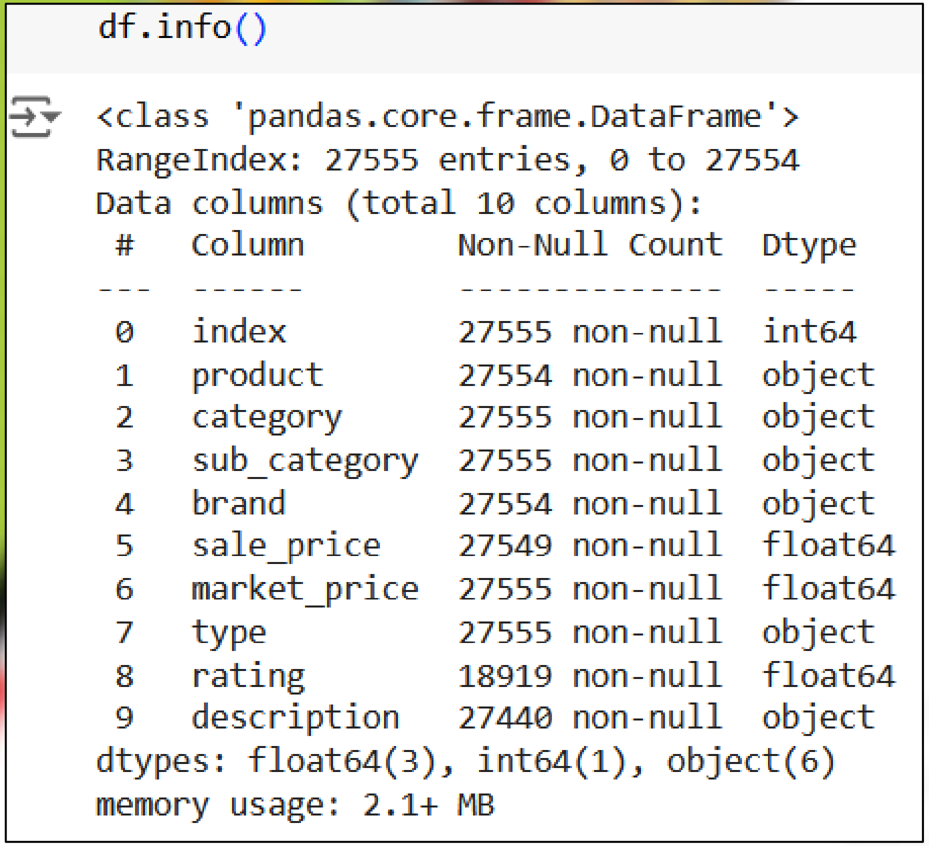
```
(27555, 10)
```

# 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, products are further classified into more specific sub-categories.*



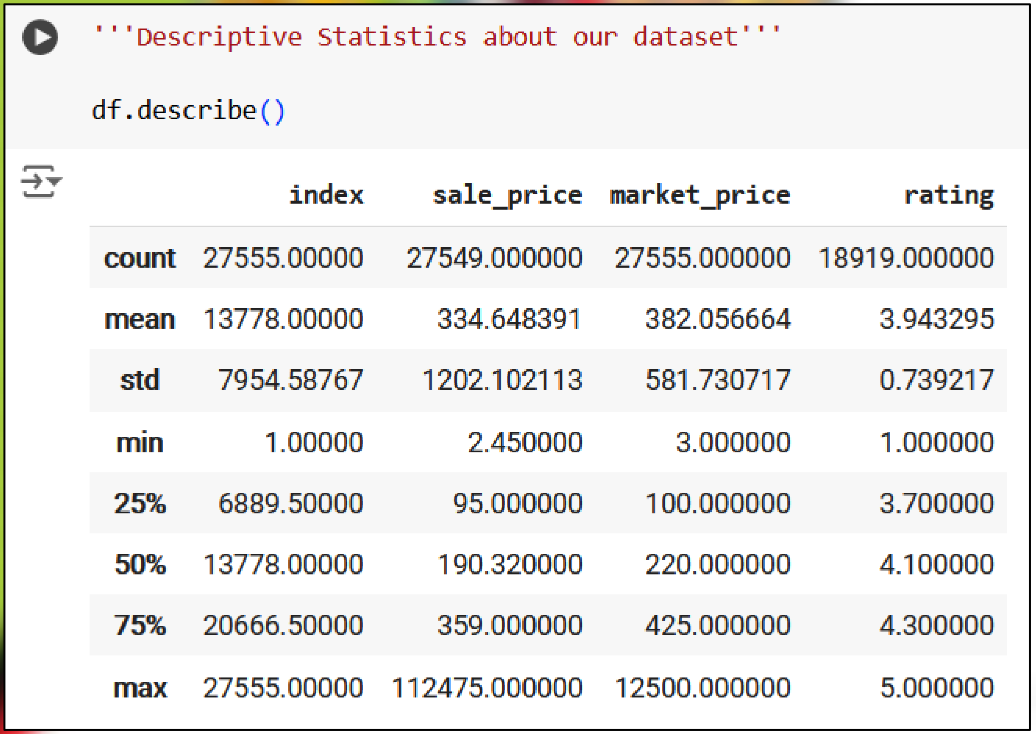
```
df.info()
```

```
<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:

*Key Features include:*

- ***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.*
- ***Description:** The 'Description' attribute provides a detailed narrative describing the dataset, its scope, and the context in which it was compiled.*



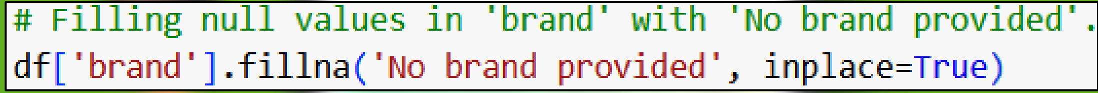
```
'''Descriptive Statistics about our dataset'''

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 |

# DATA CLEANING & PRE-PROCESSING:

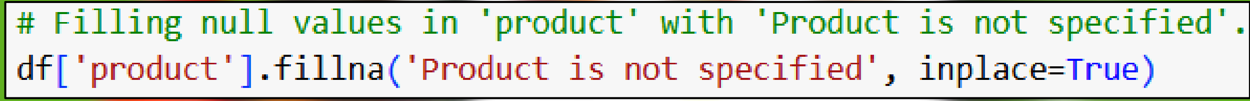
*The Dataset contains a total of 8,759 Null values. Of these, 117 are found in categorical features, while 8,642 are in numerical features.*

1) ***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'.*



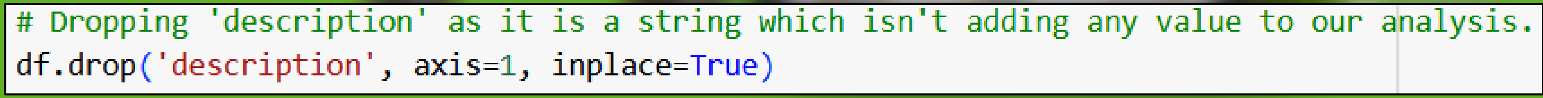
```
# Filling null values in 'brand' with 'No brand provided'.
df['brand'].fillna('No brand provided', inplace=True)
```

2) ***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.*



```
# Filling null values in 'product' with 'Product is not specified'.
df['product'].fillna('Product is not specified', inplace=True)
```

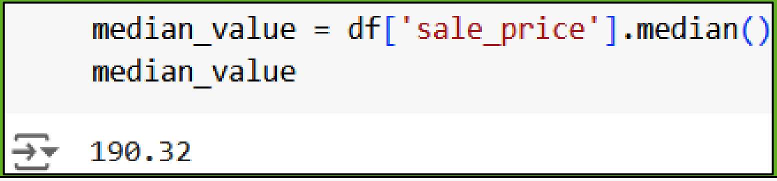
3) ***Description** : This attribute contains the highest number of categorical values (115). Since it does not contribute meaningful insights being a narrative description only, we will remove the entire column.*



```
# Dropping 'description' as it is a string which isn't adding any value to our analysis.
df.drop('description', axis=1, inplace=True)
```

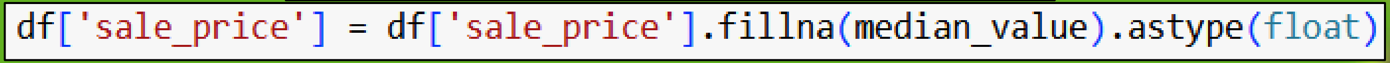
# DATA CLEANING & PRE-PROCESSING:

4) *Sale price : This feature had both Outliers and Null values present in it. Firstly, 6 Null values has been filled with 'Median' and then the Outliers has been handled by using Inter-Quartile Range(IQR) Method.*

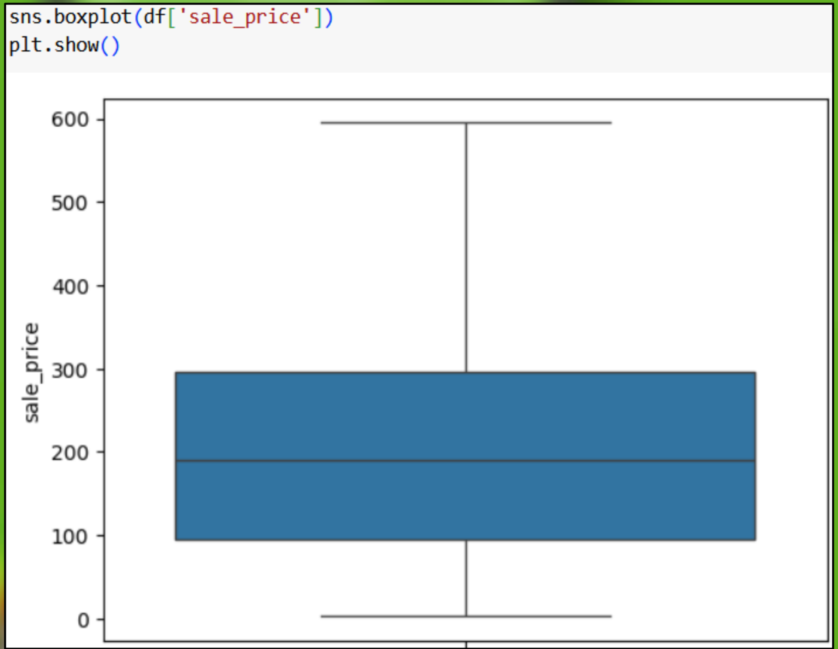


```
median_value = df['sale_price'].median()
median_value
```

```
190.32
```

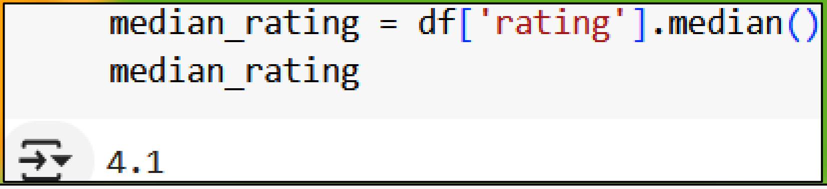


```
df['sale_price'] = df['sale_price'].fillna(median_value).astype(float)
```



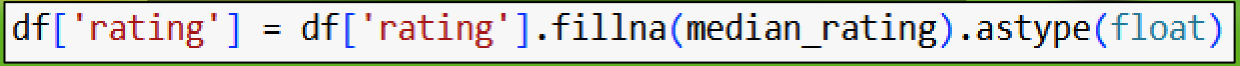
```
sns.boxplot(df['sale_price'])
plt.show()
```

5) *Rating : Since this feature has no Outliers and is Negatively skewed, filling its 8,636 Null values with the median would be straightforward.*

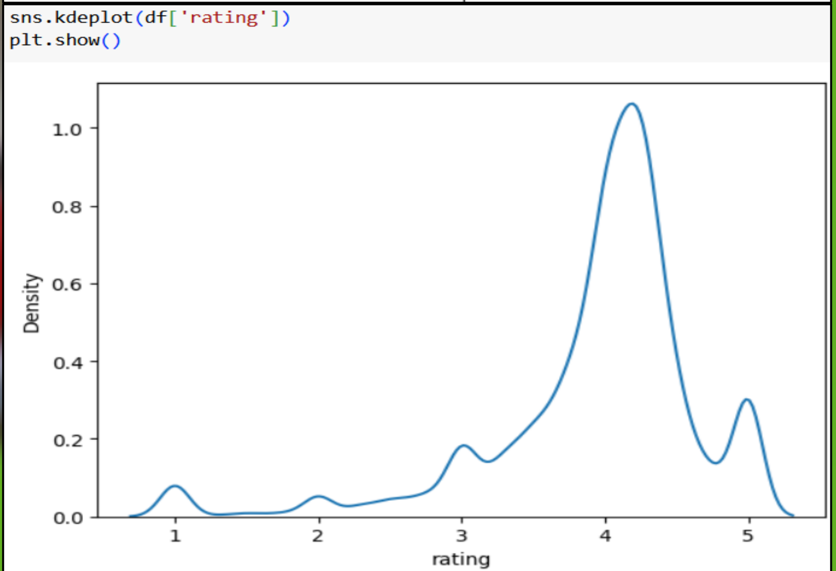


```
median_rating = df['rating'].median()
median_rating
```

```
4.1
```



```
df['rating'] = df['rating'].fillna(median_rating).astype(float)
```

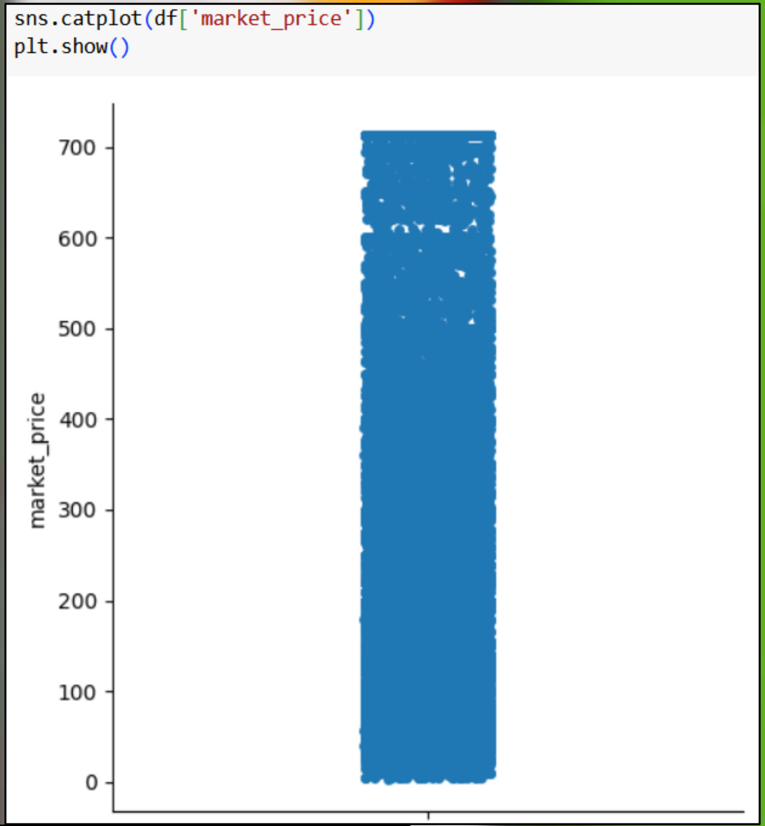


```
sns.kdeplot(df['rating'])
plt.show()
```

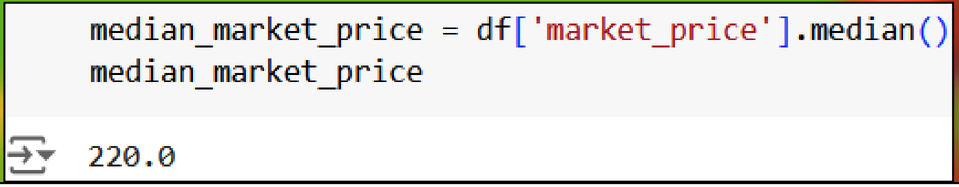
# DATA CLEANING & PRE-PROCESSING:

*Point to Ponder - As all the Null values has been handled, we still have one Numerical feature ('Market Price') left to check at least for Outliers to maintain data equilibrium for better insights.*

6) *Market Price : This feature did contain a few Outliers which has been identified by using the Inter-Quartile Range (IQR) Method. Then those Outliers has been replaced with 'Median' to ensure data accuracy.*

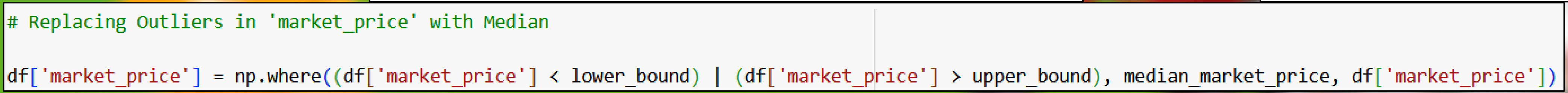


```
sns.catplot(df['market_price'])
plt.show()
```



```
median_market_price = df['market_price'].median()
median_market_price
```

```
220.0
```



```
# Replacing Outliers in 'market_price' with Median

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

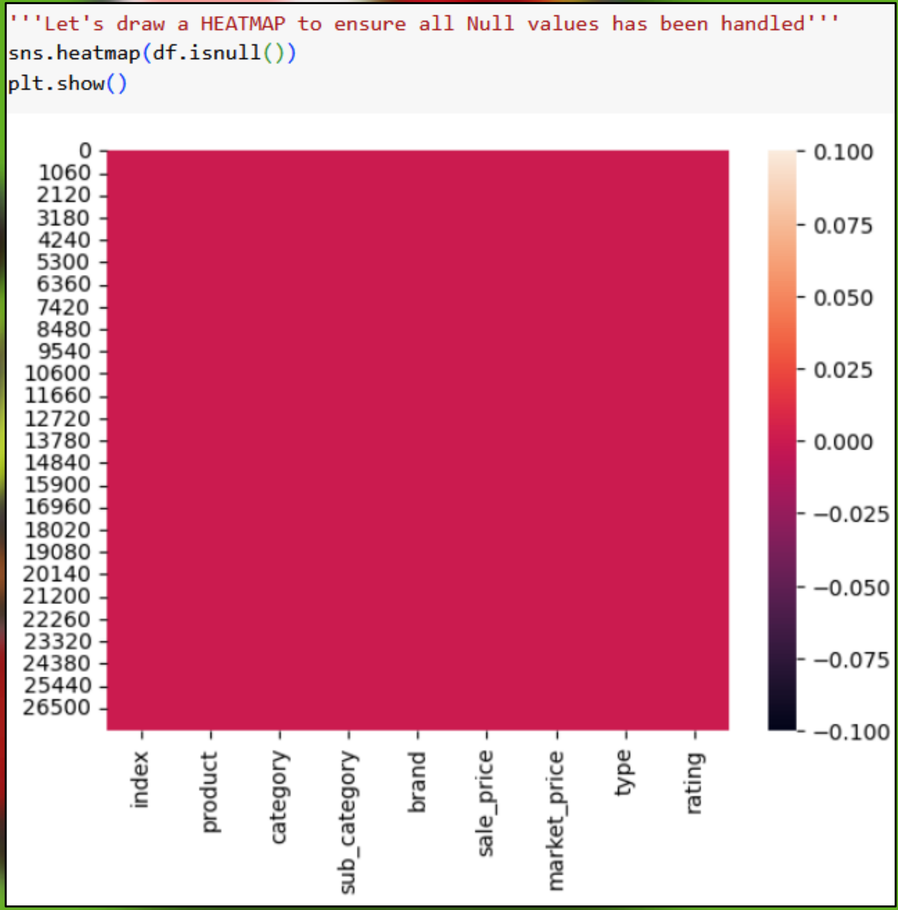
# DATA CLEANING & PRE-PROCESSING:

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.*

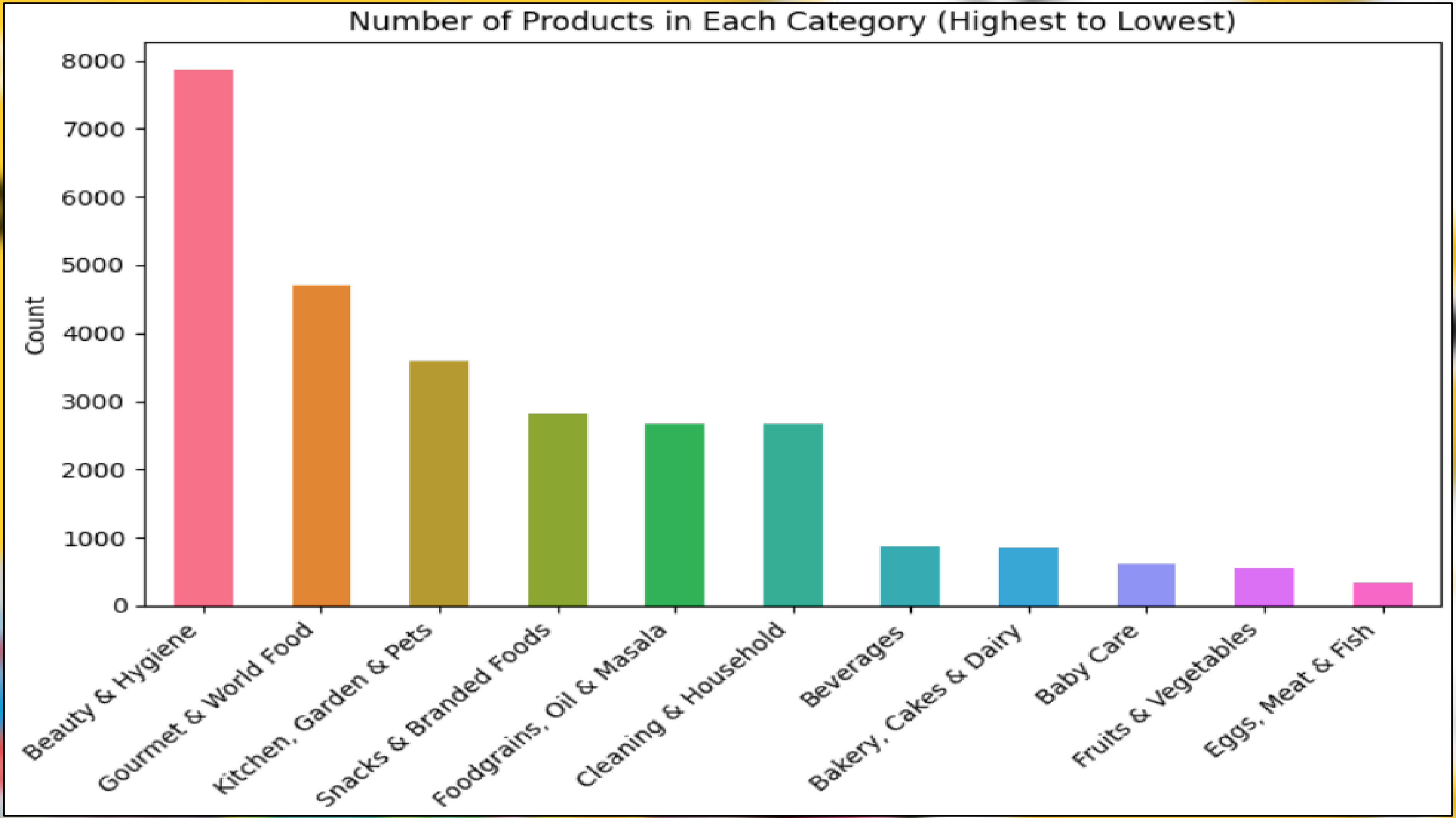
✓ *With these Null, Missing, and Invalid values appropriately addressed, we are now ready to move forward with analyzing the dataset.*



```
'''Let's draw a HEATMAP to ensure all Null values has been handled'''
sns.heatmap(df.isnull())
plt.show()
```

# 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".*

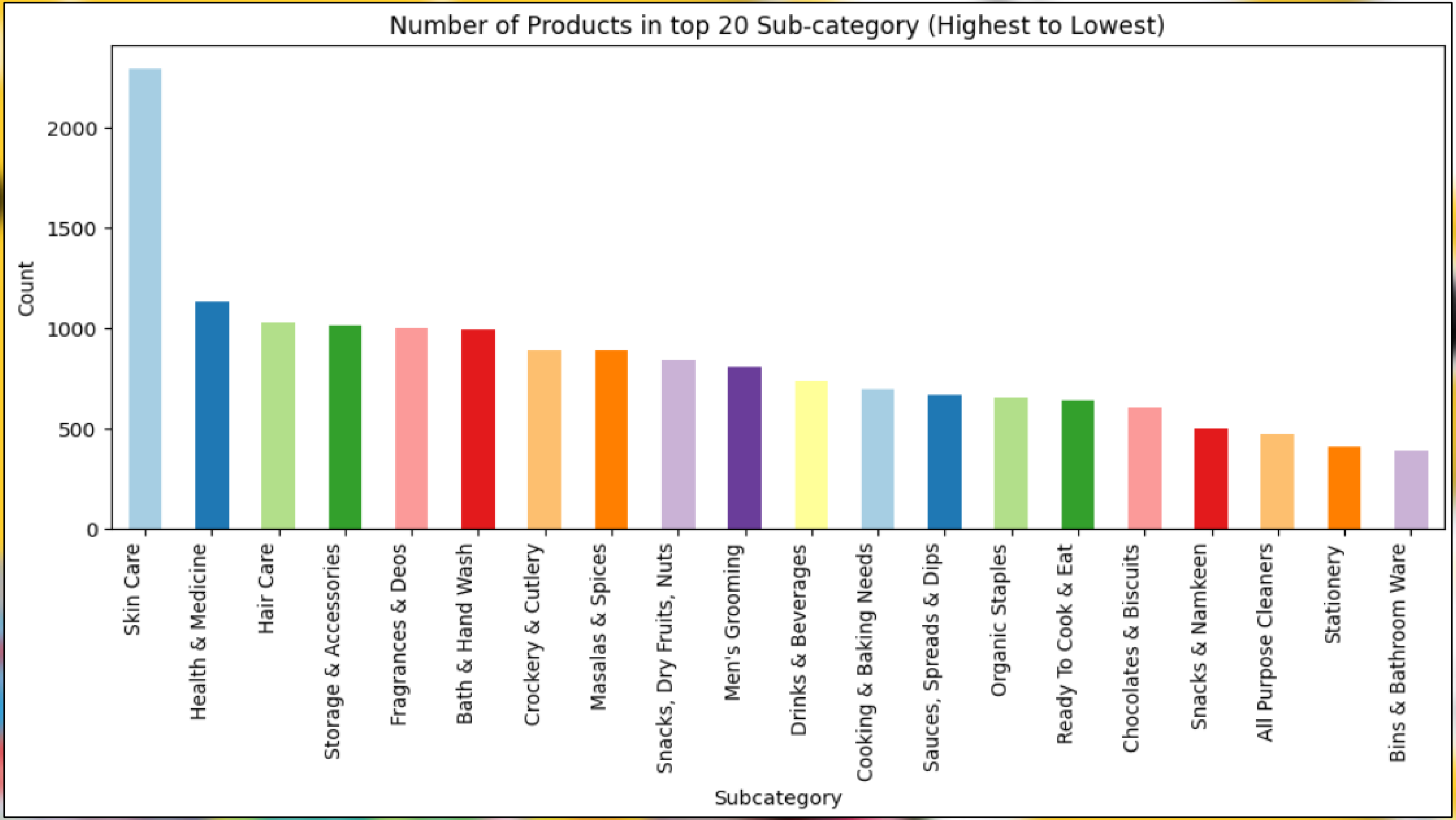
- *The categories "Snacks & Branded Foods" and "Foodgrains, Oil & Masala" also have a significant number of products. These are essential categories that are likely to be in high demand.*

- *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 20 Sub-category.*



# Data Visualization and Insights

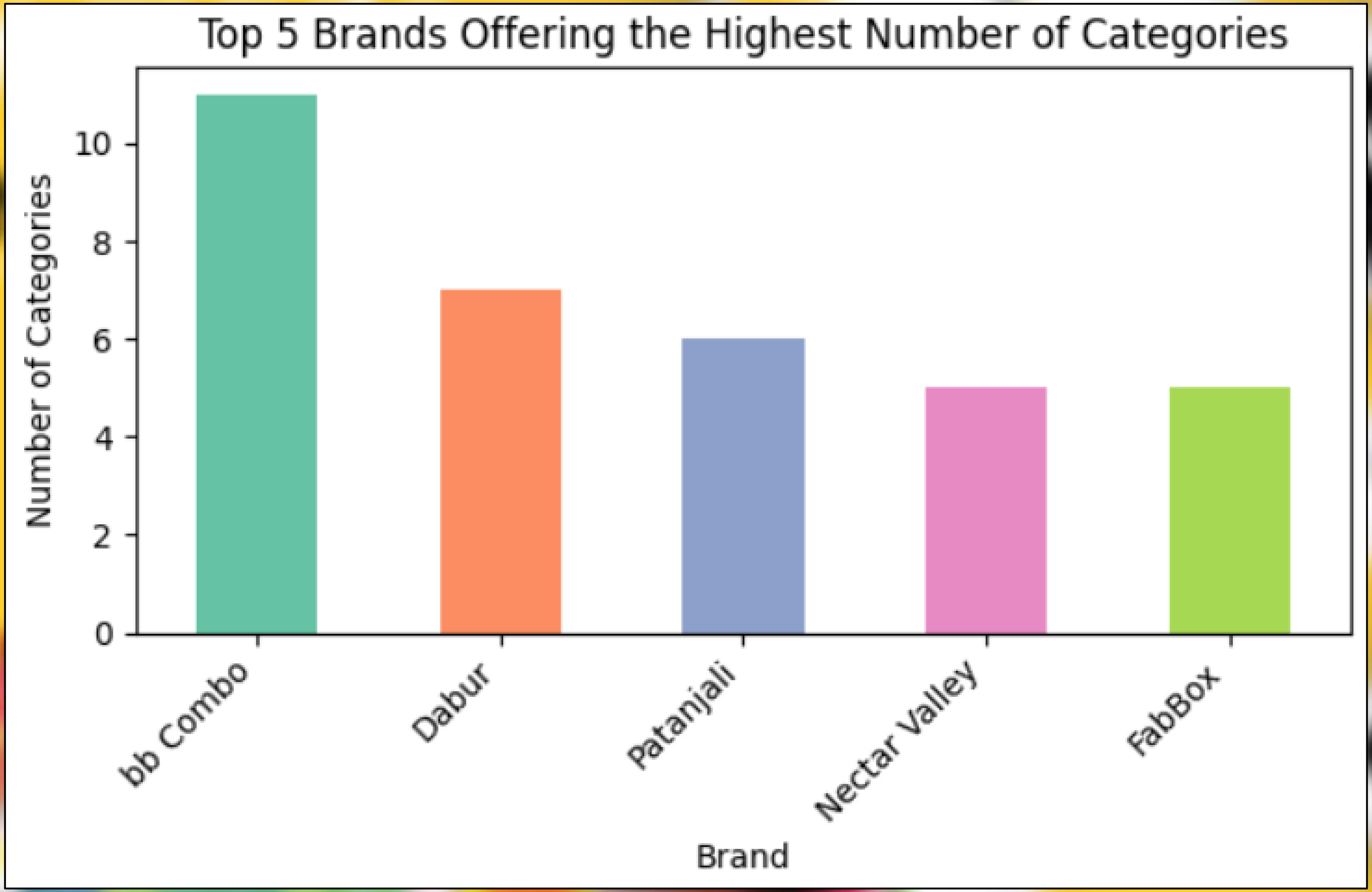
*BAR CHART: Plot the distribution of number of products in Top 20 Sub-category.*

## 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

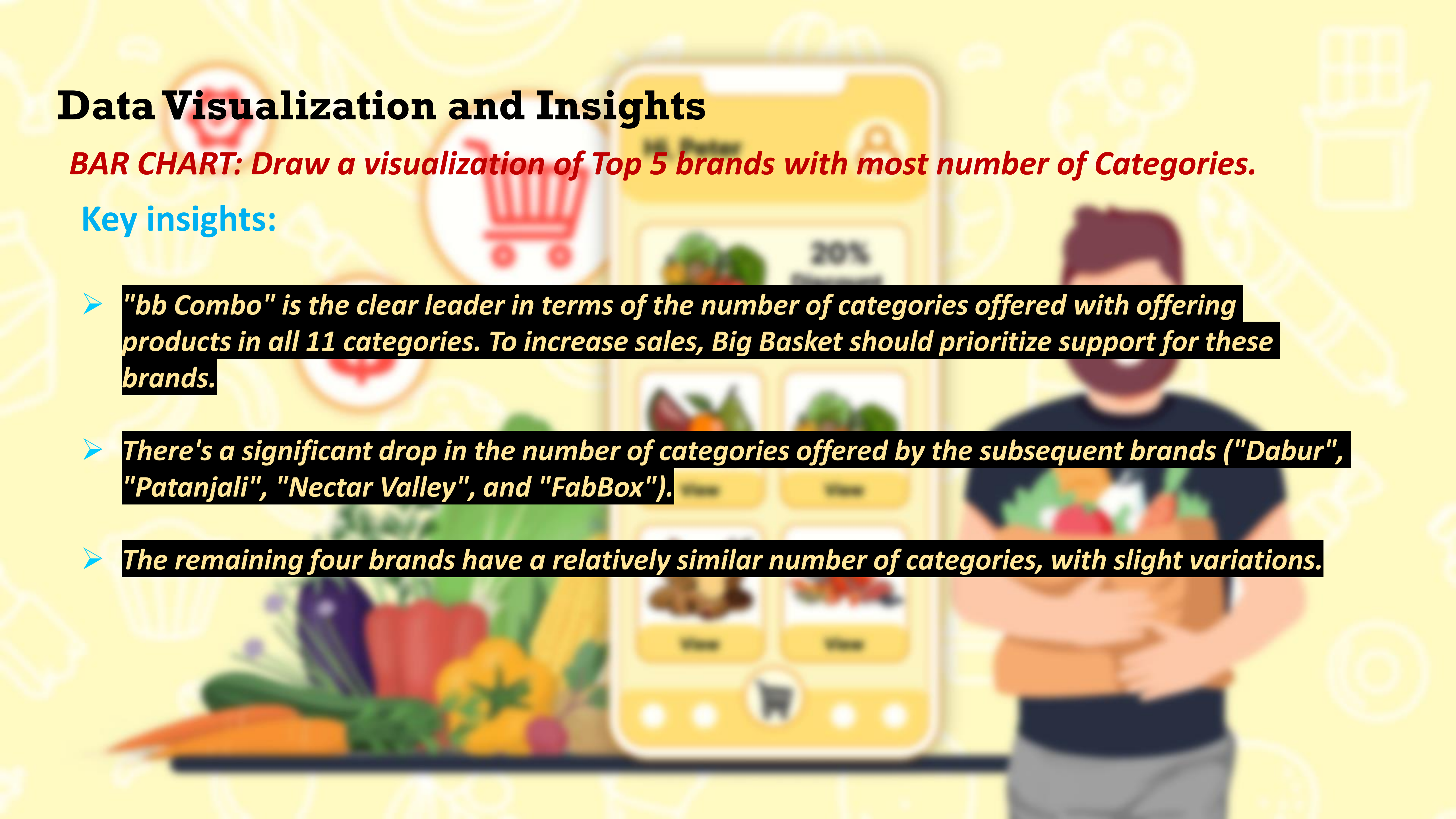
***BAR CHART: Draw a visualization of Top 5 brands with most number of Categories.***



# Data Visualization and Insights

*BAR CHART: Draw a visualization of Top 5 brands with most number of Categories.*

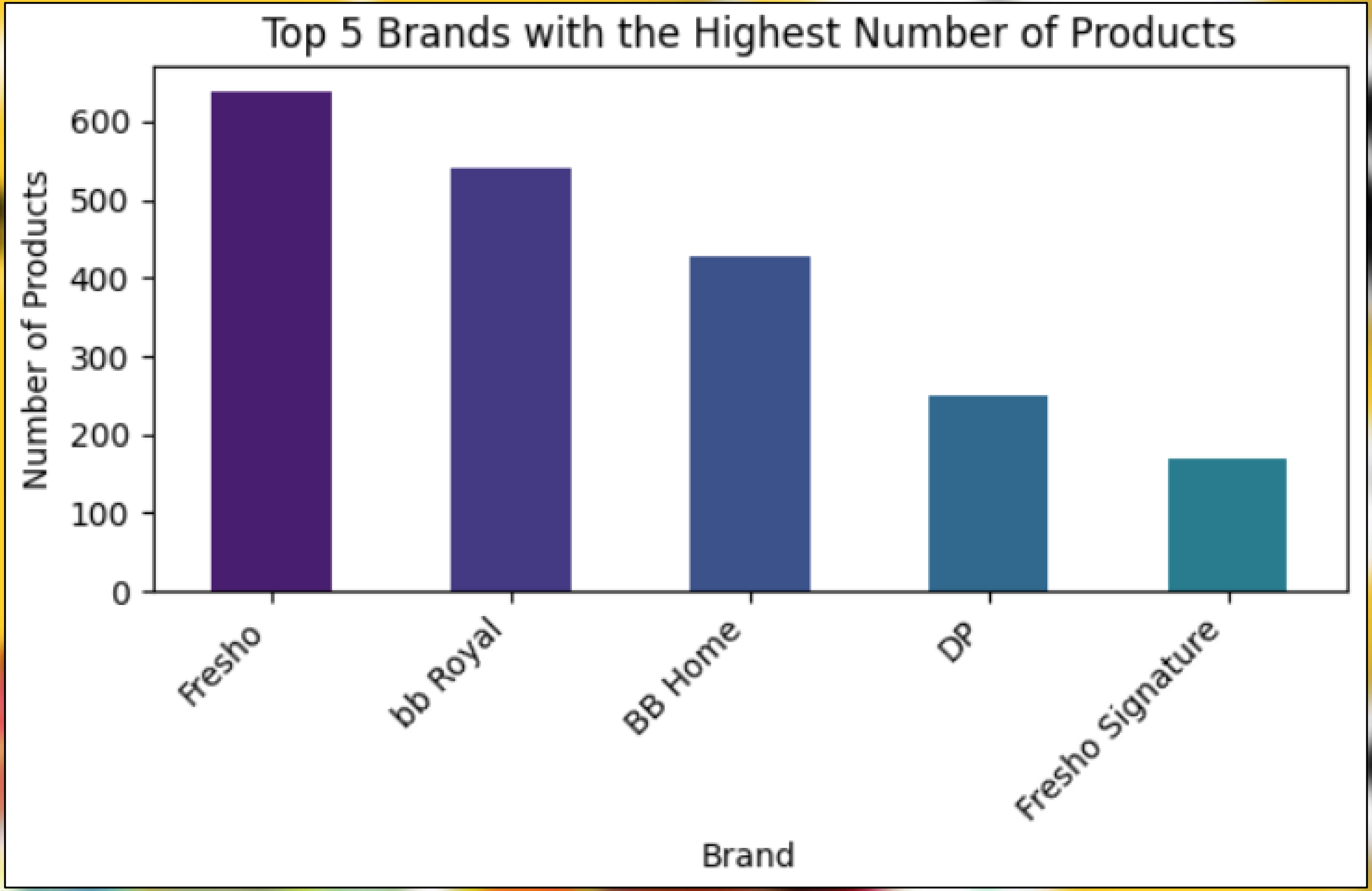
## Key insights:



- *"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 5 brands offering highest number of products.*



# Data Visualization and Insights

*BAR CHART: Draw a visualization of Top 5 brands offering highest number of products.*

## Key insights:

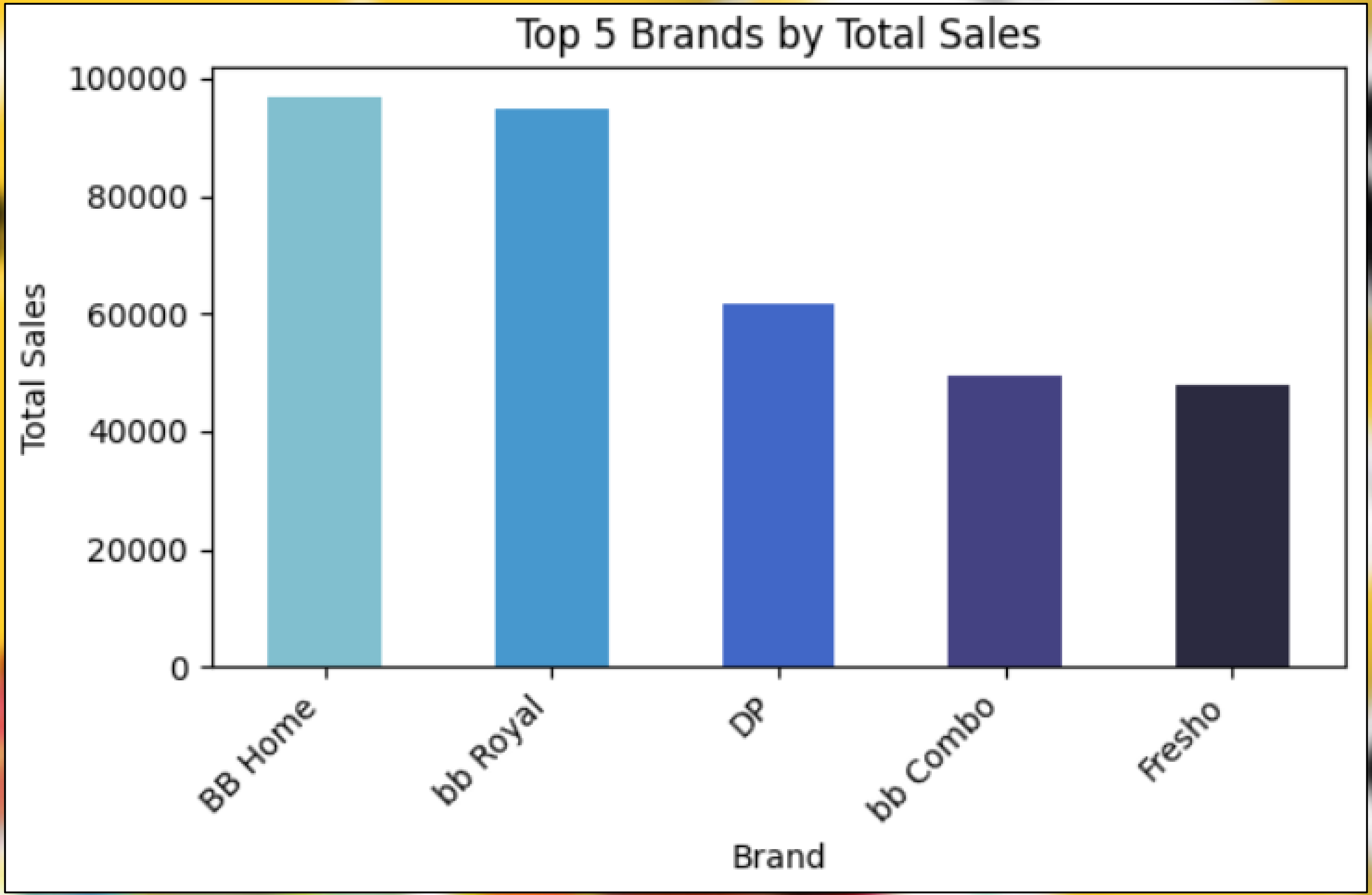
- *"Fresho" is the dominant brand with the highest number of products offered on Big Basket, followed by "bb Royal", "BB Home".*

- *Note that top 3 brands of this graph are selling Groceries, which are either fresh fruits/vegetables, Rice/Flour or Cutlery/Cookware.*

- *Which completely makes clear sense as Big Basket is all about online supermarket selling Groceries of various type.*

# Data Visualization and Insights

*BAR CHART: Draw a visualization of Top 5 brands by Total Sales.*



# Data Visualization and Insights

*BAR CHART: Draw a visualization of Top 5 brands by Total Sales.*

## Key insights:

- *"BB Home" and "bb Royal" is the clear leader in terms of total sales, indicating their popularity and strong customer preference among Big Basket users.*

- *This suggests that "BB Home" and "bb Royal" holds a dominant position in the market and is a key contributor to Big Basket's revenue.*

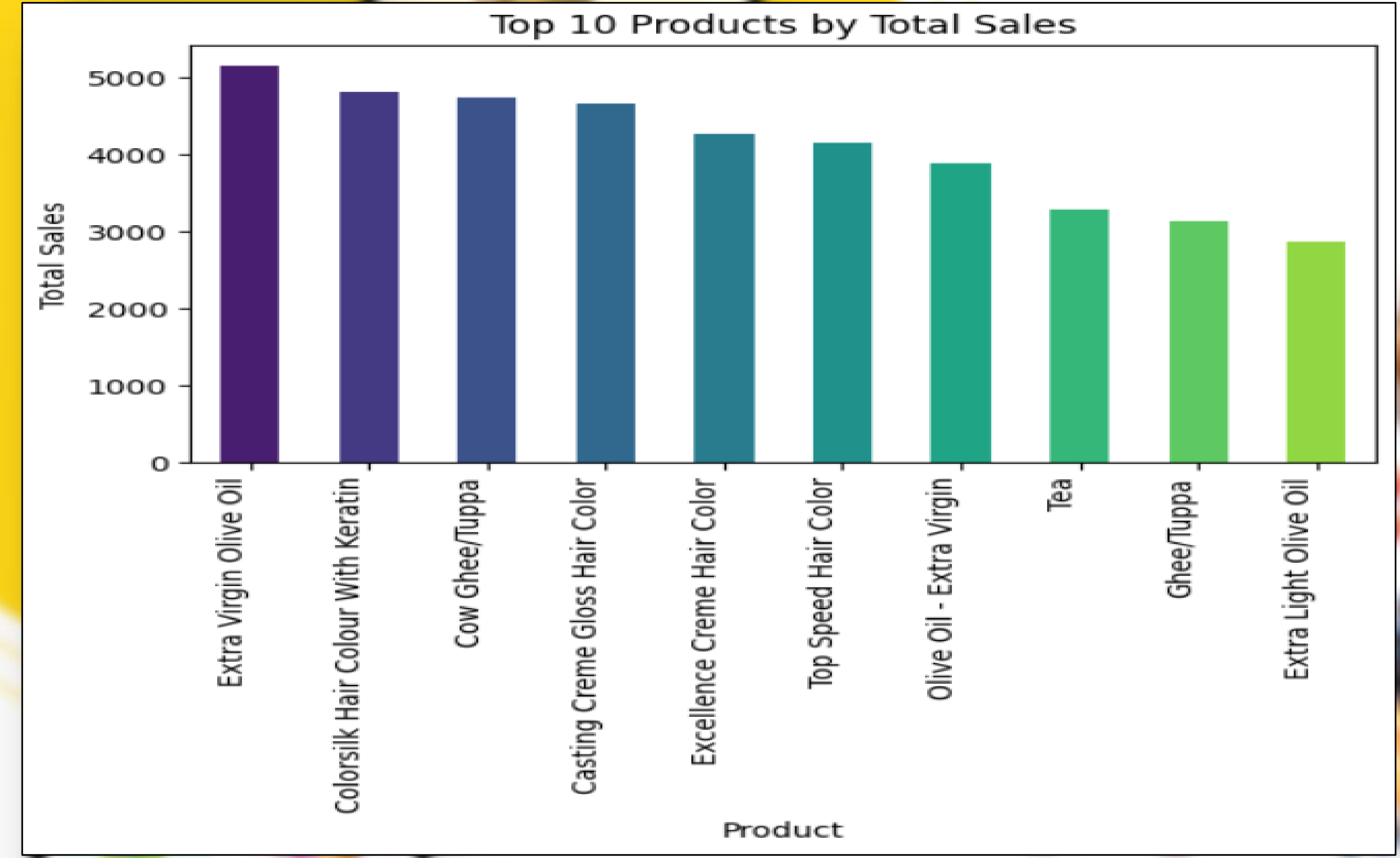
- *It also seems that brands with the prefix "BB" are affiliated companies of a single larger corporation, similar to Reliance Fresh.*

- *The remaining brands, "bb Combo" & "Fresho", have significantly lower total sales compared to Top 3 performers.*

- *Big Basket could consider strategies to further leverage the popularity of all "BB" Brands while also focusing on promoting other brands to diversify its sales and potentially capture a larger market share.*

# Data Visualization and Insights

***BAR CHART: Draw a visualization of Top 10 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.*

- *"Colorsilk Hair Colour With Keratin" is the second best-selling product, followed closely by "Cow Ghee/Tuppa" and "Casting Creme Gloss Hair Color".*

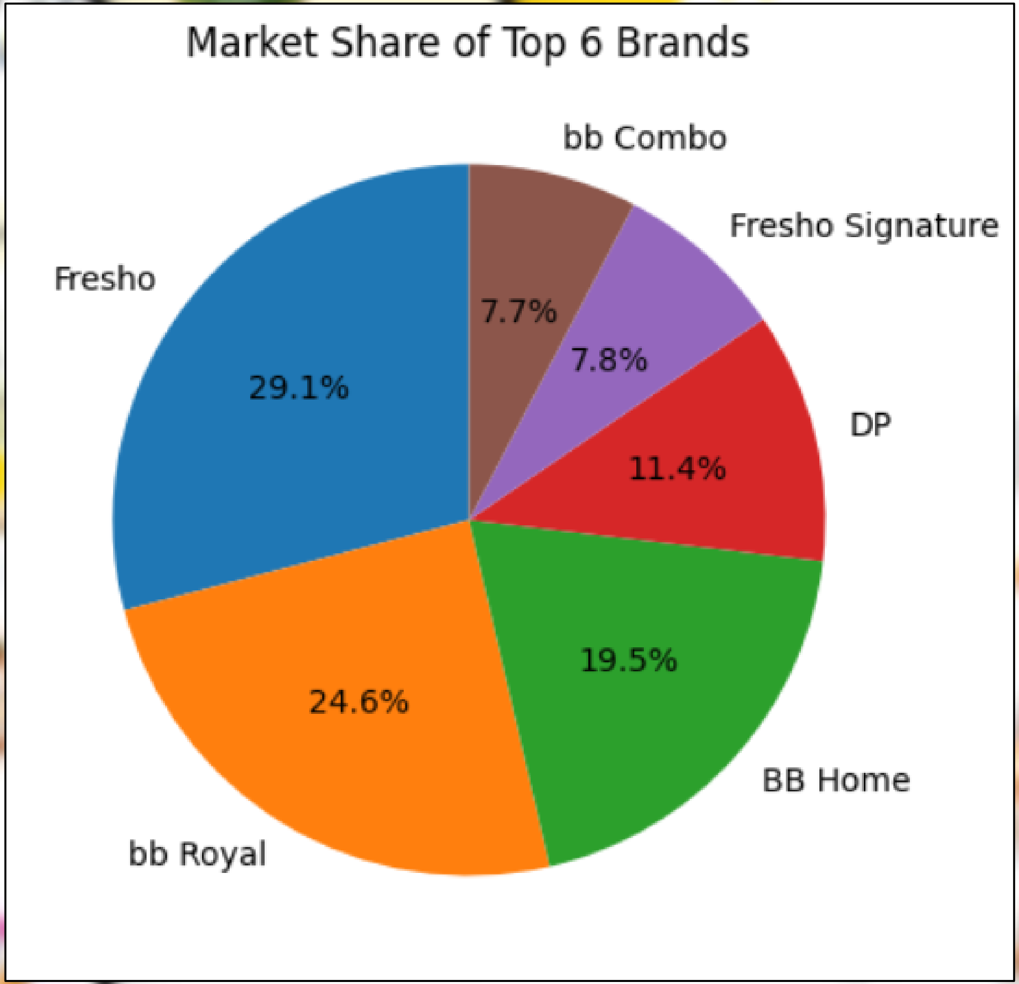
- *The following three products, "Excellence Creme Hair Color", "Top Speed Hair Color", and "Olive Oil - Extra Virgin" have relatively similar sales figures, with "Excellence Creme Hair Color" being slightly ahead.*

- *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 1.*

# Data Visualization and Insights

*Brand Analysis -*

*PIE CHART: Draw a visualization of Top 6 Brands to show their Market Share.*



# Data Visualization and Insights

*Brand Analysis -*

*PIE CHART: Draw a visualization of Top 6 Brands to show their Market Share.*

## Key insights:

- *"Fresho" commands the largest market share with 29.1% among the top 6 brands, indicating its strong presence and popularity on Big Basket.*

- *"bb Royal" and "BB Home" also hold significant market shares with 24.6% and 19.5%, suggesting their strong brand recognition and customer loyalty.*

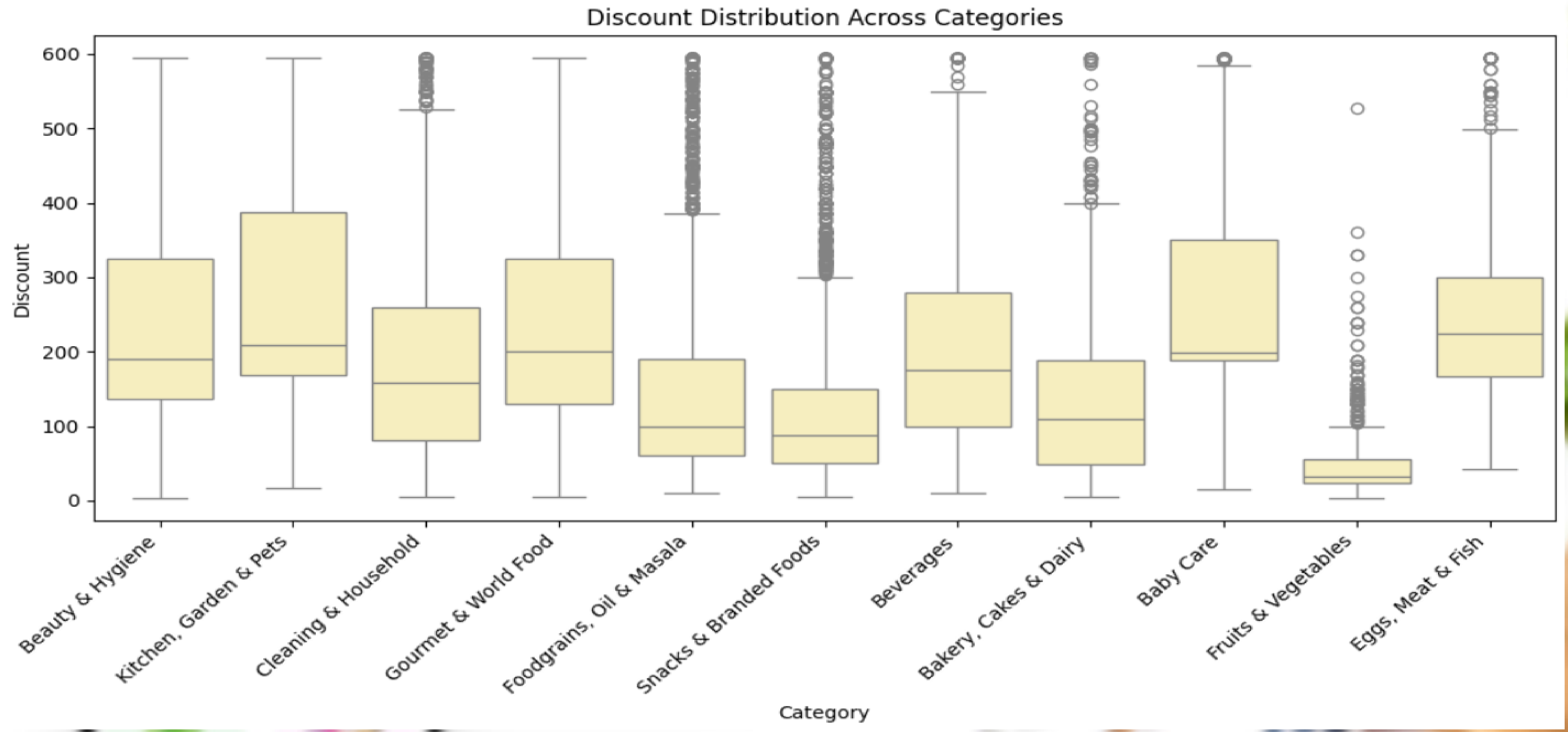
- *The chart reveals that Big Basket offers a diverse product range, encompassing categories like Baby care ("bb Combo"), Cleaning & Household ("DP"), Garden & Pets ("BB Home").*

- *Big Basket could consider strategies to further strengthen the market position of "Fresho", "bb Royal", and "BB Home" while also exploring ways to increase the market share of other brands.*

# Data Visualization and Insights

*Discount Analysis -*

*BOXPLOT: Draw a visualization to compare Discount Distributions across Categories.*



# Data Visualization and Insights

## *Discount Analysis -*

## *BOXPLOT: Draw a visualization to compare Discount Distributions across Categories.*

### Key insights:

- ***Price Variability** : The box plot reveals the spread of prices within each category. Categories with longer boxes indicate a wider range of prices for products within that category.*

- ***Median Prices** : The horizontal line within each box represents the median price. This allows you to quickly compare the typical price point of products across different categories.*

- ***Outliers** : The dots outside the whiskers of the box plot represent outlier prices. These are products that are significantly more expensive or cheaper than the majority of products within their category.*

- ***Category Comparisons** : By comparing the positions and sizes of the boxes, you can identify categories with higher or lower overall prices and those with greater or lesser price variability.*

# Data Visualization and Insights

## *Discount Analysis -*
## *BOXPLOT: Draw a visualization to compare Discount Distributions across Categories.*

### Key insights:

***Potential Business Implications***

- ***Pricing Strategy** : This visualization can inform pricing decisions for new products or adjustments to existing pricing. Like if we see a category with a high median price and low variability, we might consider introducing a lower-priced product to capture a different market segment.*

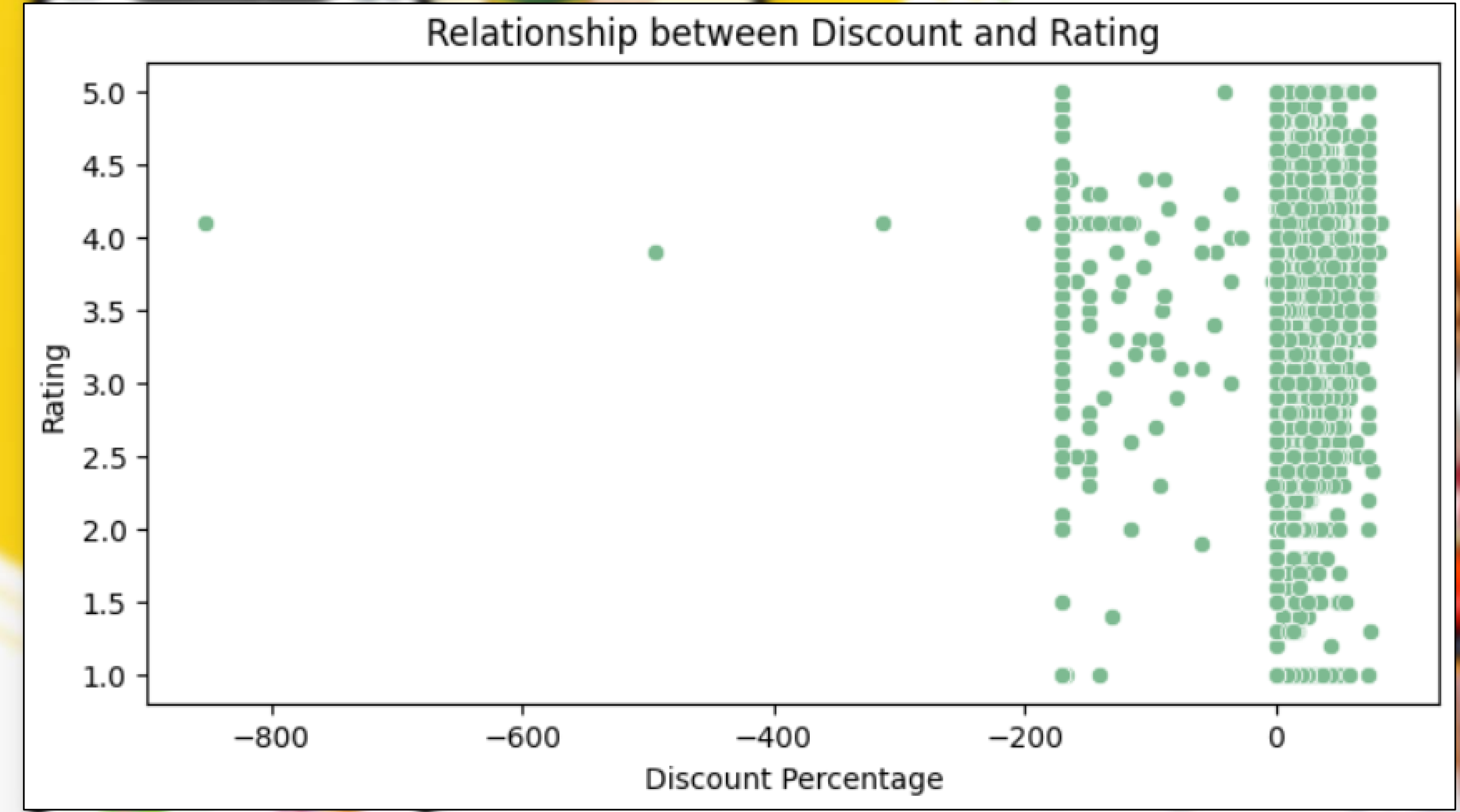
- ***Inventory Management** : Understanding price distributions can help with inventory management. Categories with a wide range of prices might require a more diverse inventory strategy compared to categories with a narrow price range.*

- ***Marketing and Promotions** : The insights from this plot can be used to tailor marketing and promotional efforts. For instance, we might focus discounts on categories with higher median prices to attract price-sensitive customers.*

# Data Visualization and Insights

*Discount Analysis -*

*SCATTER PLOT: Draw a visualization to see if there's any relationship between Discount 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.*

- ***Potential Factors** : Other factors, such as product quality, brand reputation, and customer expectations, likely play a more significant role in determining product ratings.*

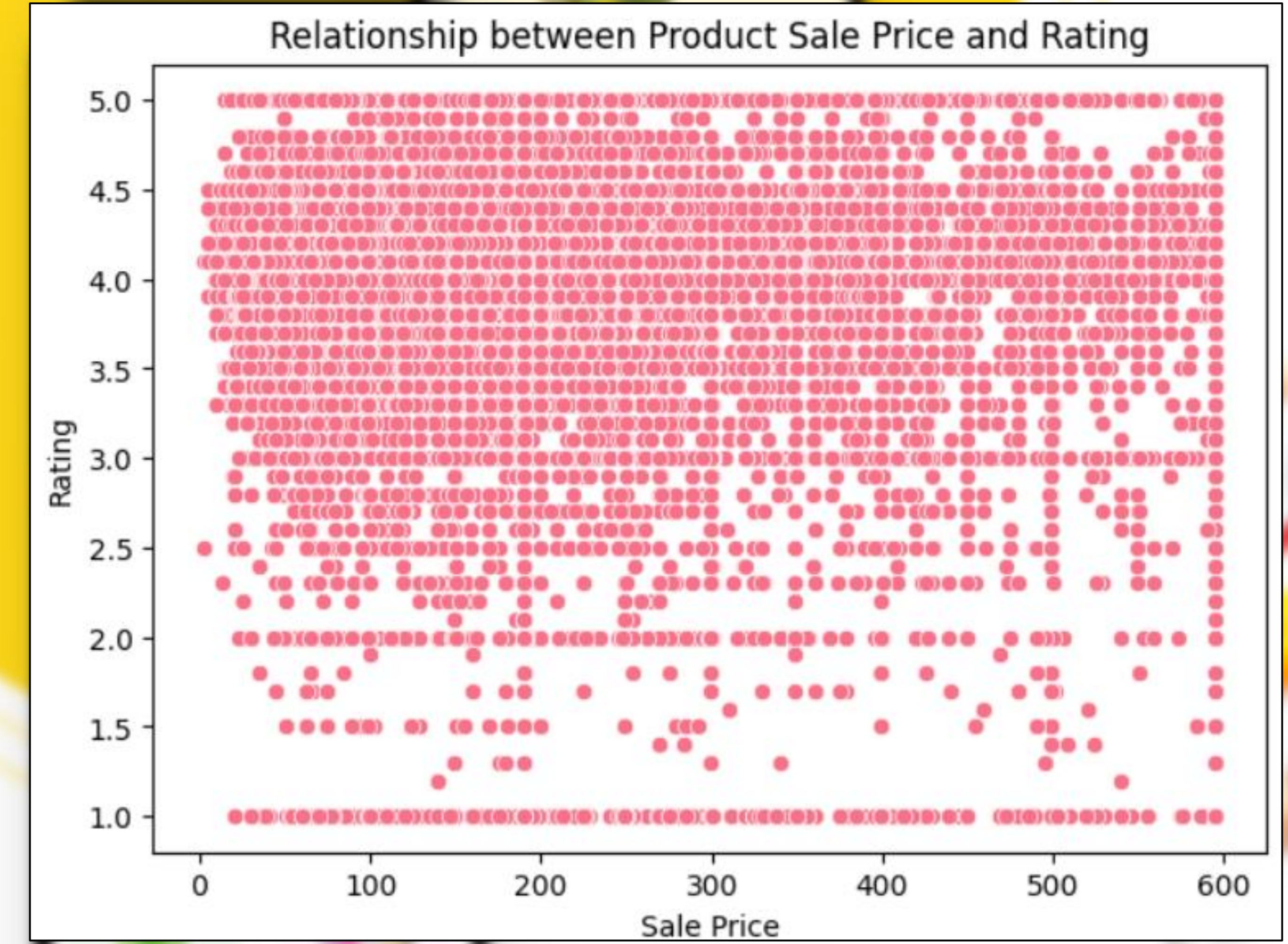
***Business Implications***

- ***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.*

- ***Targeted Promotions** : Consider offering targeted discounts based on customer preferences and product categories, rather than relying on blanket discounts.*

# Data Visualization and Insights

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



# Data Visualization and Insights

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

## Key insights:

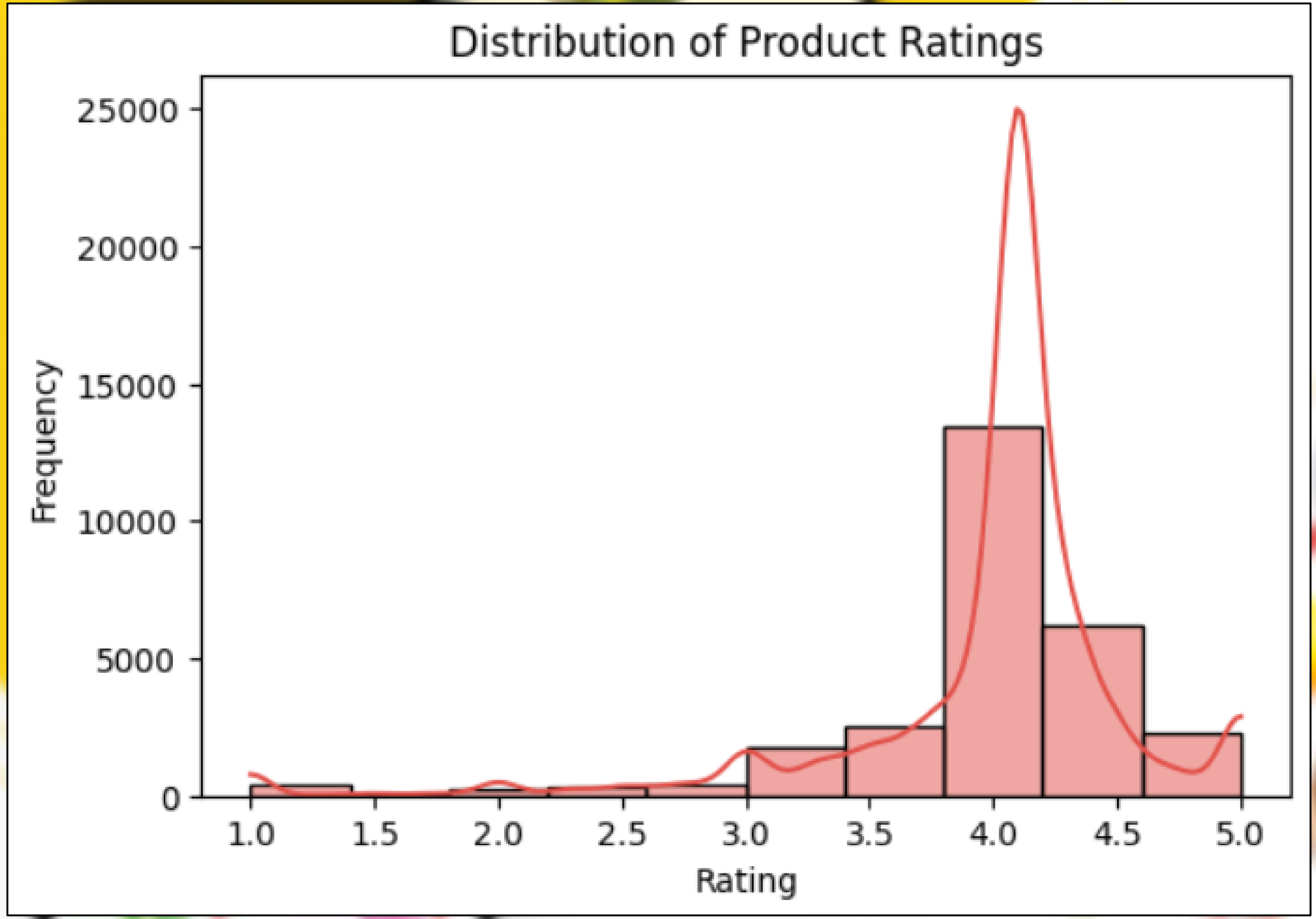
- *No clear correlation : There doesn't seem to be a strong linear relationship between product "Sale Price" and "Rating". This suggests that customers don't necessarily rate higher-priced products more favorably.*

- *Concentration of products : Most products are concentrated in the lower price range, regardless of their rating. This indicates that the majority of products offered are budget-friendly.*

- *Potential outliers : There are a few products with high prices and low ratings, which could be worth investigating further to understand why they didn't receive favorable reviews despite their cost.*

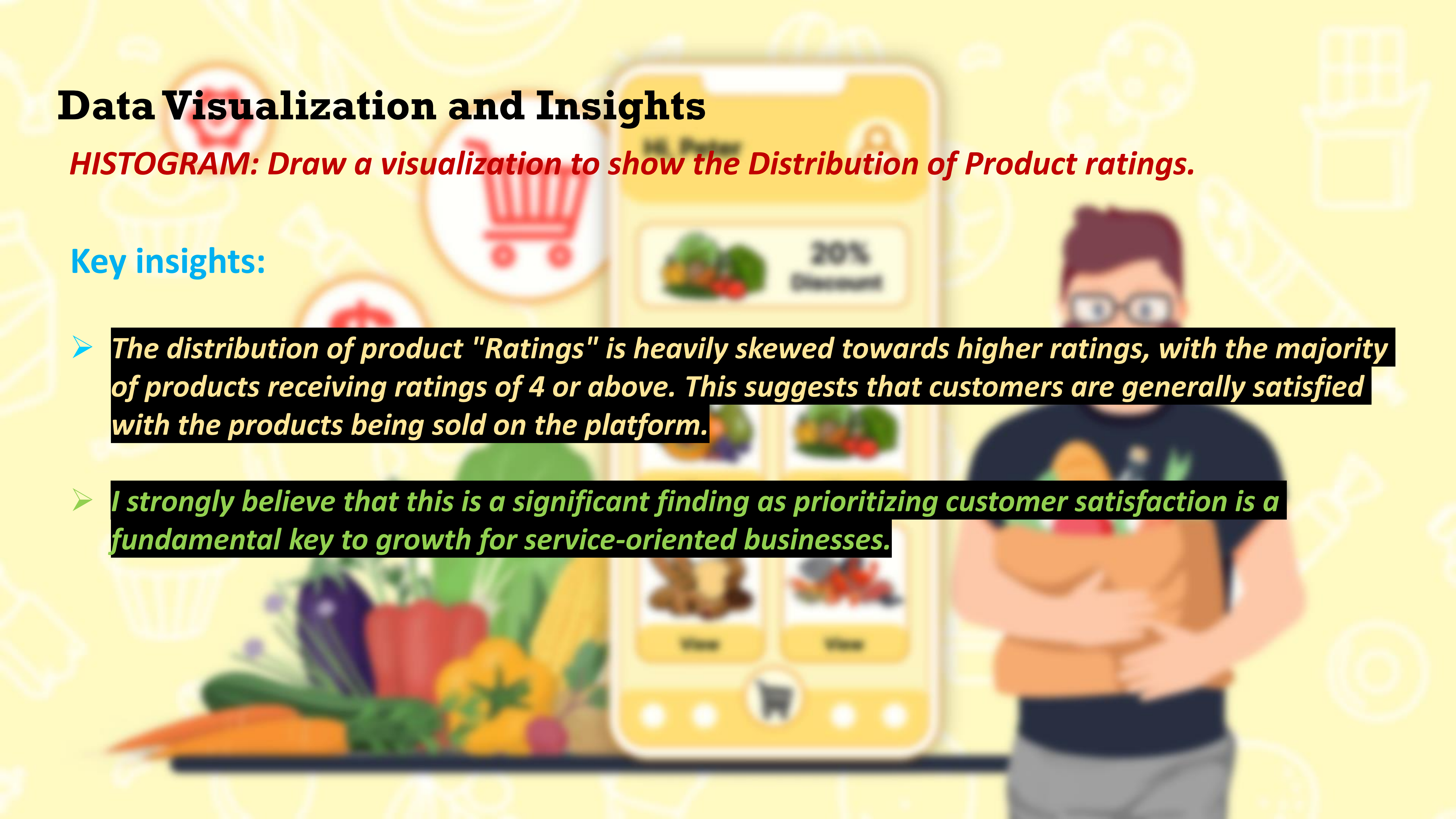
# Data Visualization and Insights

***HISTOGRAM: 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.*

# FINAL REPORT

*Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :*

## KEY FINDINGS :

### *A. Product and Category Analysis:*

- ***"Beauty & Hygiene" is the most dominant category, followed by "Gourmet & World Food".***
- ***"Skin Care" is the leading sub-category.***
- ***"Fresho" is the most popular brand with the highest number of products.***
- ***"BB Home" and "bb Royal" generate the highest total sales.***

### *B. Discount Analysis:*

- ***No strong correlation between Discount percentage and product rating.***
- ***Product Sale price does not have a clear relationship with rating.***
- ***Discounts do not appear to strongly influence product ratings.***

# FINAL REPORT

*Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :*

## KEY FINDINGS :

### C. *Rating Analysis:*

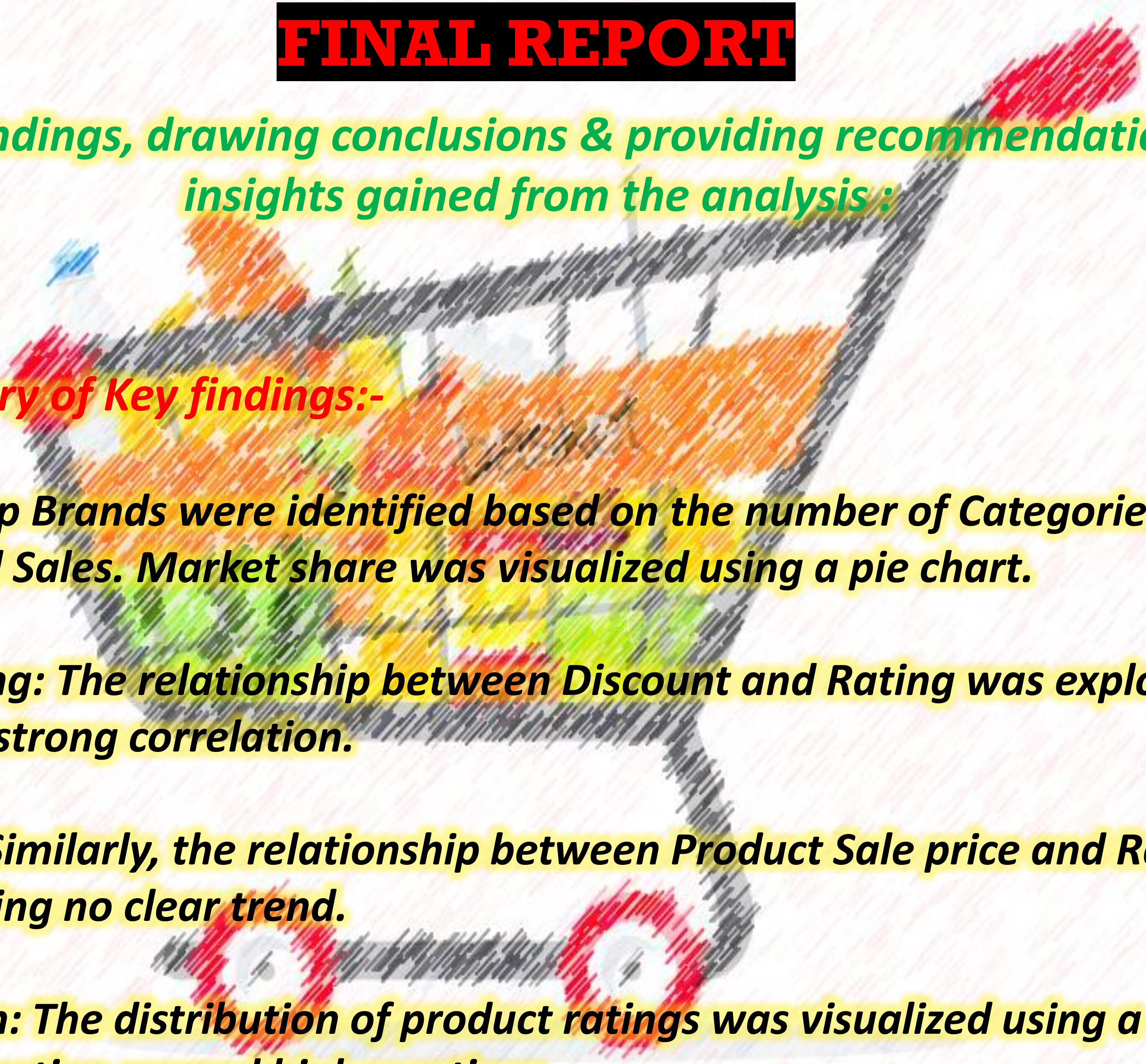
- ❖ ***Product ratings are heavily skewed towards higher ratings (4 or above). Which is a clear sign of customer satisfaction.***
- ❖ ***Customers usually gave higher ratings to products, indicating that the price, whether low or high, isn't a major concern for them. They are satisfied with the quality they receive.***

### ❑ *General Summary of Key findings:-*

- ❖ ***Product Distribution: The distribution of Products across Categories and Sub-categories was visualized, revealing the most and least popular Products types. Certain product categories and sub-categories are more popular than others.***

# FINAL REPORT

***Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :***



## KEY FINDINGS :

- ❑ ***General Summary of Key findings:-***

- ❖ ***Brand Analysis: Top Brands were identified based on the number of Categories, Number of products, and Total Sales. Market share was visualized using a pie chart.***

- ❖ ***Discount and Rating: The relationship between Discount and Rating was explored using a scatter plot, indicating no strong correlation.***

- ❖ ***Price and Rating: Similarly, the relationship between Product Sale price and Rating was visualized, suggesting no clear trend.***

- ❖ ***Rating Distribution: The distribution of product ratings was visualized using a histogram, showing a concentration around higher ratings.***

# FINAL REPORT

***Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :***

## 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

*Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :*

## RECOMMANDATIONS :

- ❖ ***Firstly, Big Basket should typically concentrate on promoting products in popular categories and sub-categories, as these are significant revenue generators for the brand.***

- ❖ ***Big Basket must expand product offerings in categories like "Fruits & Vegetables" and "Eggs, Meat & Fish" to cater to a wider audience.***

- ❖ ***Big Basket is supposed to leverage the popularity of "Fresho", "BB Home", and "bb Royal" for further growth.***

- ❖ ***It would be wise for Big Basket to identify and partner with brands to boost profits, focusing on marketing through YouTube ads and TV commercials to leverage their market presence.***

# FINAL REPORT

*Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :*

## RECOMMANDATIONS :

- ❖ ***Prioritize product quality and customer experience to maintain high ratings.***
- ❖ ***Big Basket is expected to consider strategies to improve ratings for products with lower ratings too.***
- ❖ ***Consider targeted discounts based on customer preferences and product categories.***
- ❖ ***Big Basket ought to continue monitoring customer satisfaction and address any potential issues promptly.***

*THANK YOU FOR READING*

*For coding part, kindly refer to below link :-*

*https://github.com/SaxenaKushagr/Big-Basket-Analysis.git*