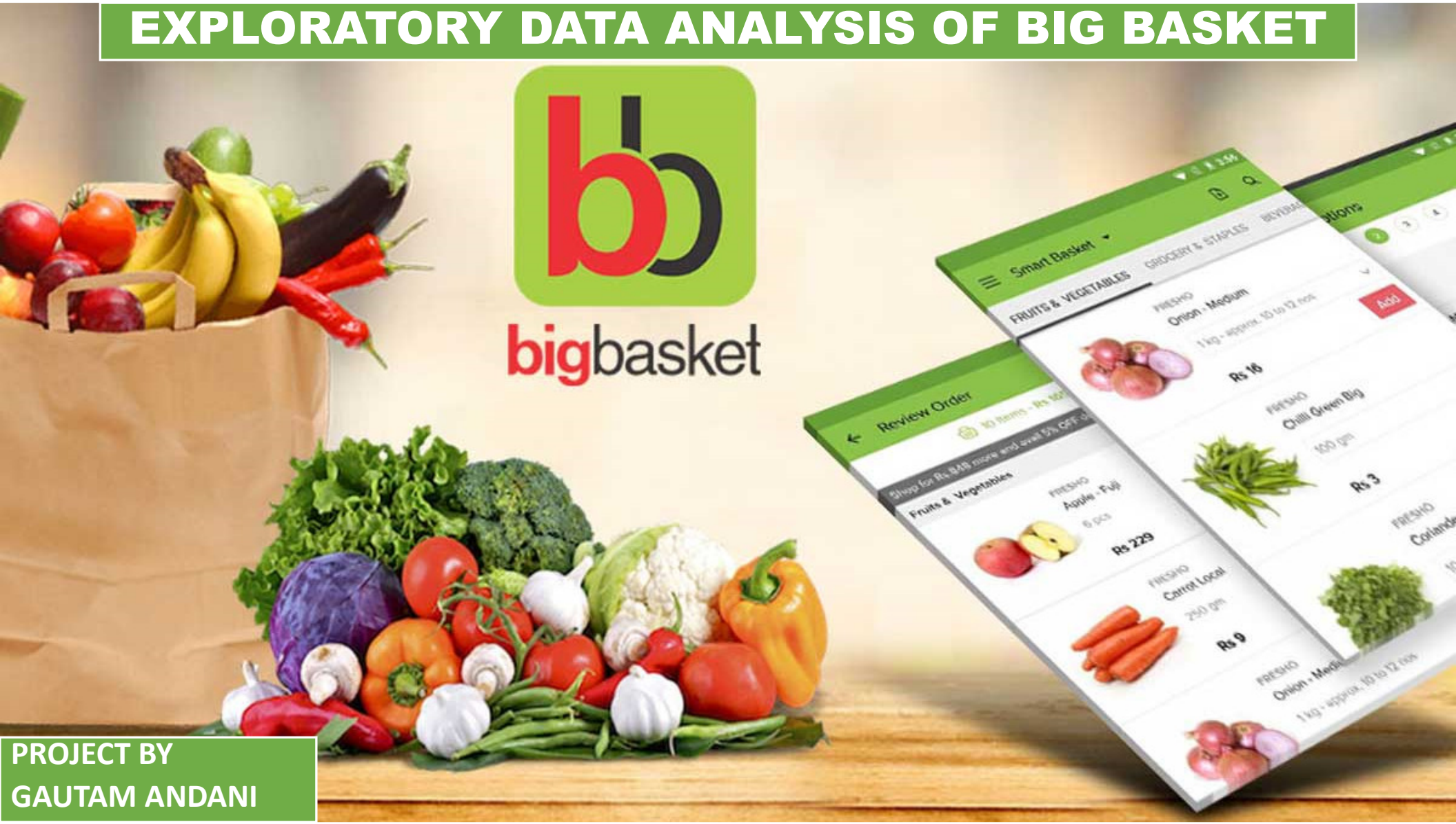
EXPLORATORY DATA ANALYSIS OF BIG BASKET



PROJECT BY
GAUTAM ANDANI

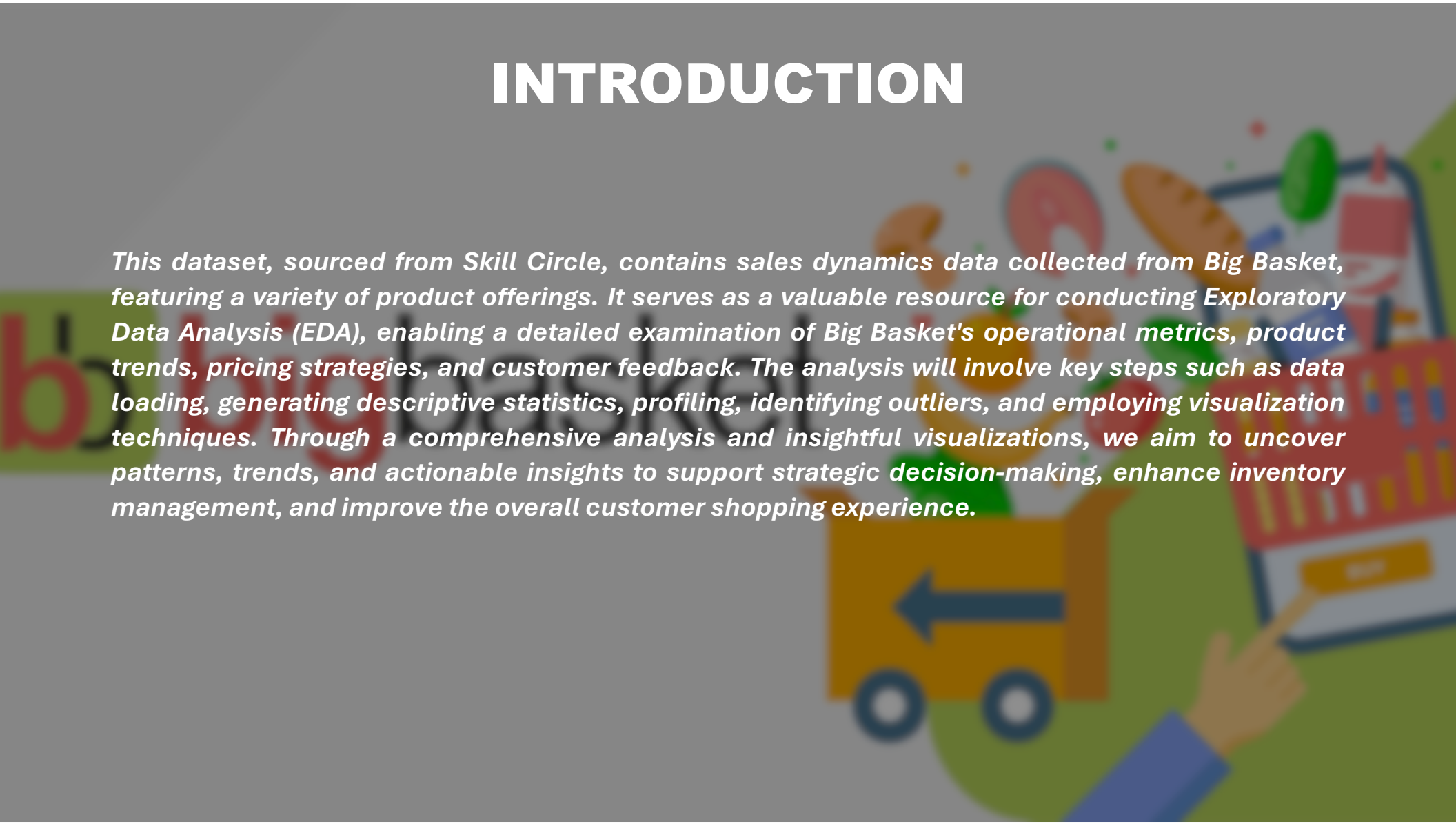
ABOUT BIG BASKET

- *Big Basket, founded in 2011 by Hari Menon, Vipul Parekh, V.S. Sudhakar, Abhinay Choudhari, and V.S. Ramesh, is one of India's largest online grocery delivery platforms. Headquartered in Bengaluru, it operates under the Tata Group, which acquired a majority stake in 2021.*

- *Big Basket offers a wide range of products, including fresh fruits and vegetables, pantry staples, personal care items, and household essentials, catering to millions of customers across India. The company follows a hybrid business model that combines an inventory-led approach, where products are stocked in their own warehouses to ensure quality and efficiency, and a marketplace model, where local vendors and farmers supply products.*

- *With its user-friendly app and website, flexible delivery slots, and focus on customer satisfaction, Big Basket has become a trusted name in the online grocery space.*

INTRODUCTION



This dataset, sourced from Skill Circle, contains sales dynamics data collected from Big Basket, featuring a variety of product offerings. It serves as a valuable resource for conducting Exploratory Data Analysis (EDA), enabling a detailed examination of Big Basket's operational metrics, product trends, pricing strategies, and customer feedback. The analysis will involve key steps such as data loading, generating descriptive statistics, profiling, identifying outliers, and employing visualization techniques. Through a comprehensive analysis and insightful visualizations, we aim to uncover patterns, trends, and actionable insights to support strategic decision-making, enhance inventory management, and improve the overall customer shopping experience.

OBJECTIVES OF THE PROJECT:

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 Collaboratory Notebook.

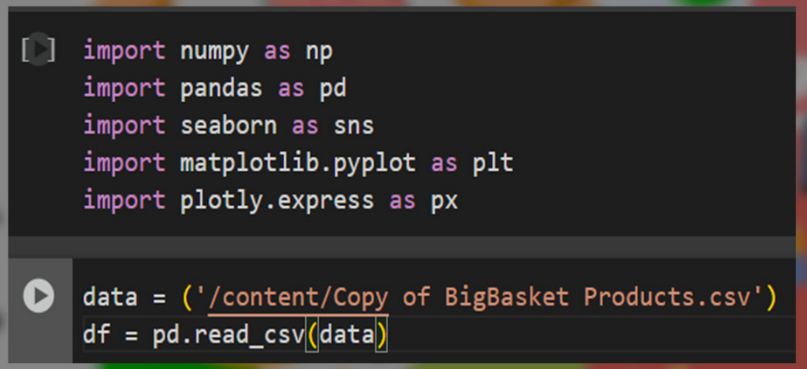
As we begin our Exploratory Data Analysis (EDA), I've named the dataset 'df'.

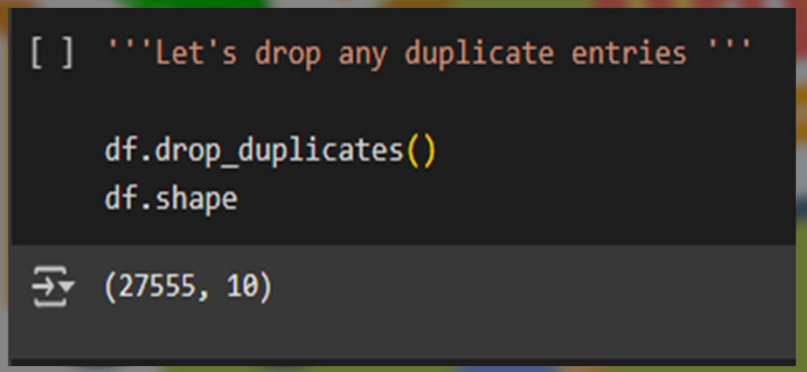
The dataset comprises of 27,555 Rows and 10 Columns.

For Data cleaning/visualization, I have utilized libraries like Numpy, Pandas, Seaborn, Matplotlib.

Any duplicate entries that were found have also been removed.

•





DESCRIPTION OF DATASET:

The dataset under analysis offers a detailed overview of BigBasket's product range and sales patterns. It includes 10 crucial attributes that highlight different aspects of the business:

Key Features:

***Index**: This attribute acts as a unique identifier for each record in the dataset.*

***Product**: Refers to the name or title of the products available on the BigBasket platform.*

***Category**: Groups products into broader classifications such as fruits, vegetables, dairy, beverages, and more.*

***Sub-Category**: Provides a more granular classification of products within each main category.*

```
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:

- ***Brand**: This attribute identifies the brand or manufacturer linked to each product.*
- ***Sale Price**: Represents the price at which the product is sold to customers.*
- ***Market Price**: Indicates the regular market price or standard pricing of the product.*
- ***Type**: Classifies products based on their specific nature or characteristics.*
- ***Rating**: Displays customer feedback or ratings given to each product on the BigBasket platform.*
- ***Description**: Offers a detailed explanation of the dataset, including its scope and the context in which it was created.*

```
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 & PREPROCESSING

The dataset contains a total of 8,759 missing values, distributed across both categorical and numerical features. Out of these, 117 null values are present in categorical features, while 8,642 belong to numerical features.

1. ***Brand:*** *The 'Brand' attribute has a single missing value within the categorical data. To maintain data consistency, this null value can be replaced with 'No Brand Provided.'*

2. ***Product:*** *The 'Product' attribute also has 1 missing value. Filling it with 'Product is not specified' ensures data completeness.*

3. ***Description:*** *This attribute has the highest number of missing values (115) among categorical features. As it primarily serves as a narrative and does not contribute significant insights, the entire column will be dropped.*

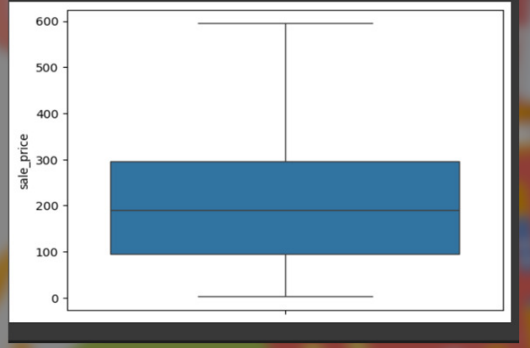
```
#For Categorical features like

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

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

# 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 & PREPROCESSING

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*

```
# So Filling the Null values in this feature by Median.

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

# The feature 'sale_price' is positively skewed.

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

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

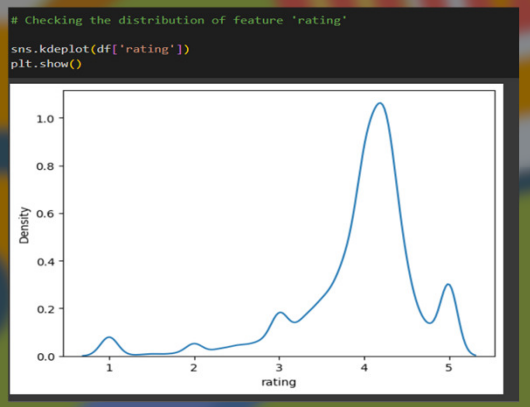


5) Rating : *Since this feature has no Outliers and is Negatively skewed, filling its 8,636 Null values with the median would be straightforward*

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

```
4.1
```

```
# Checking the distribution of feature 'rating'

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



DATA CLEANING & PREPROCESSING

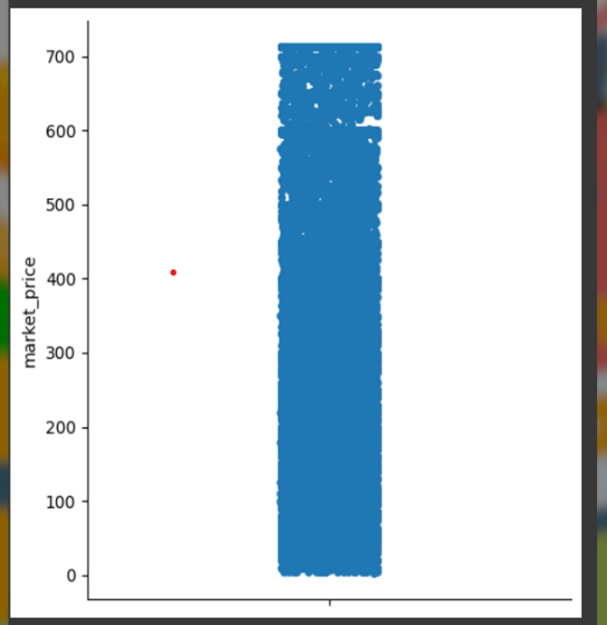
- *After fixing all missing values, the 'Market Price' feature needed to be checked for outliers to keep the data balanced.*
- *MarketPrice: Outliers in this feature were found using the IQR method. These outliers were replaced with the median value to ensure the data remains accurate and reliable.*

```
# Checking median

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

```
220.0
```

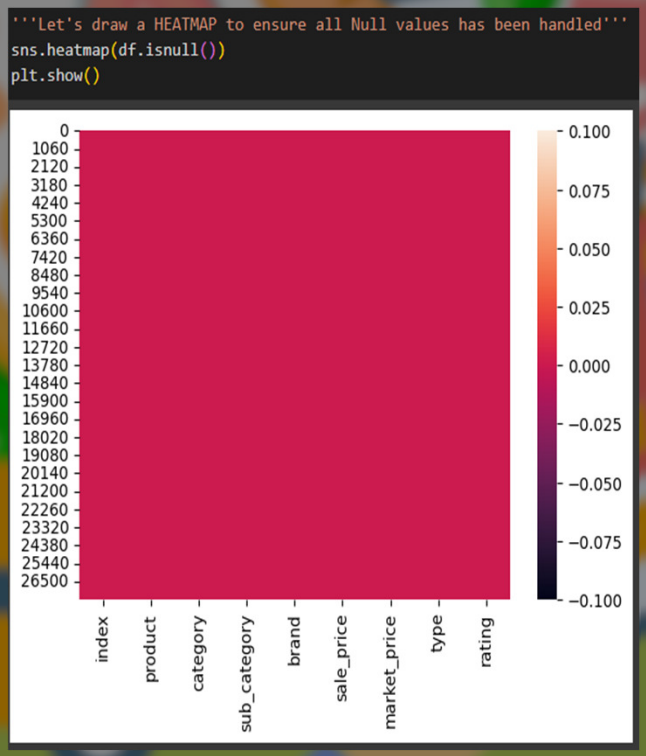
```
# Replacing Outliers in 'market_price' with Median

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

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



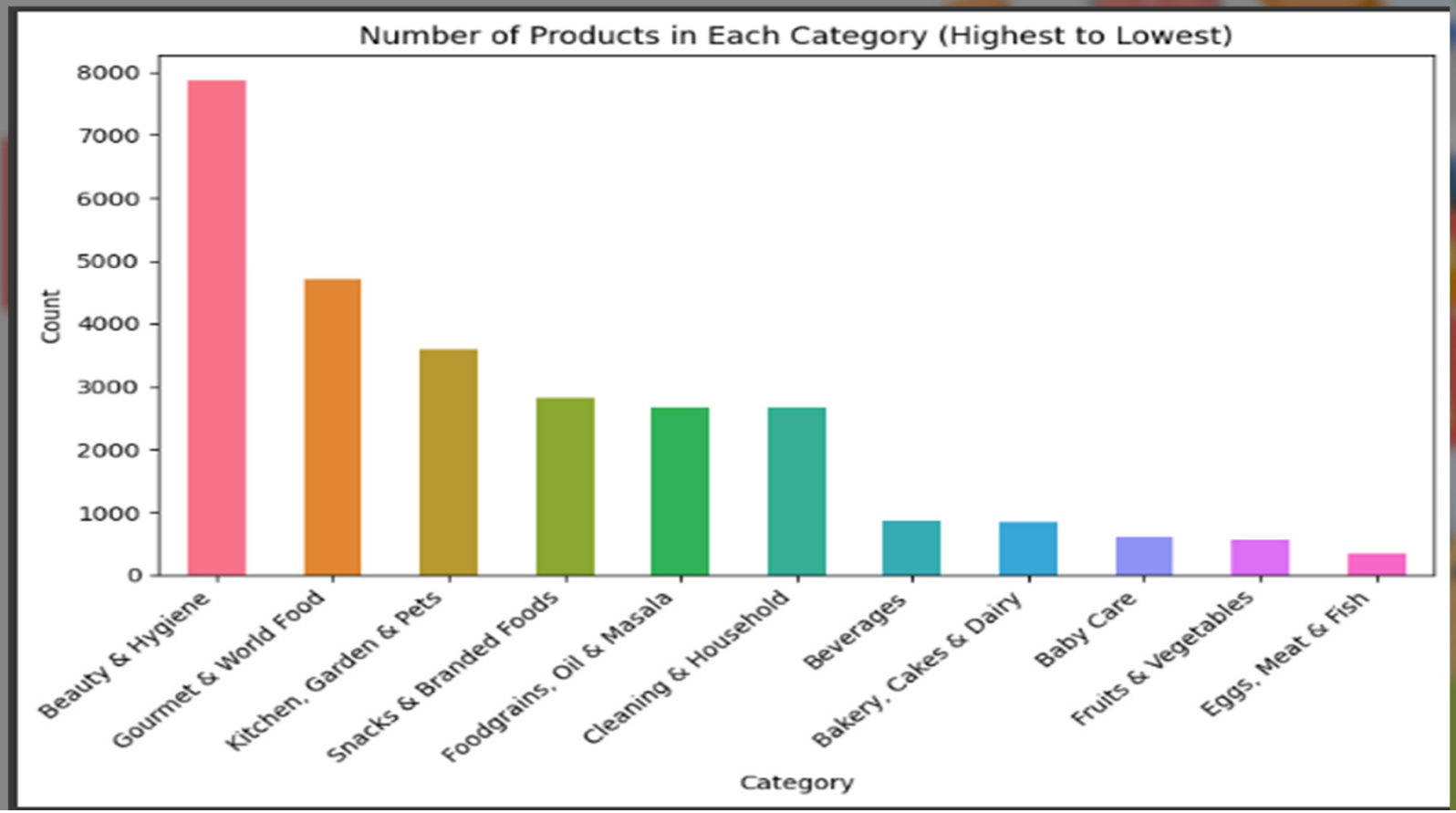
SUMMARY

- *Addressing missing values and outliers requires a systematic approach tailored to the dataset's features. Data cleaning and outlier handling are essential steps for ensuring accurate and meaningful analysis.*
- *The dataset had missing values in the 'product', 'brand', 'sale price', 'rating', and 'description' features. These were managed by filling missing values with the median/mode and removing irrelevant columns like 'description'.*
- *Outliers were found in the 'sale price' and 'market price' features. These were handled using the IQR method, with capping applied to keep values within acceptable boundaries.*
- *After resolving missing, invalid, and outlier values, the dataset is now clean and ready for analysis.*

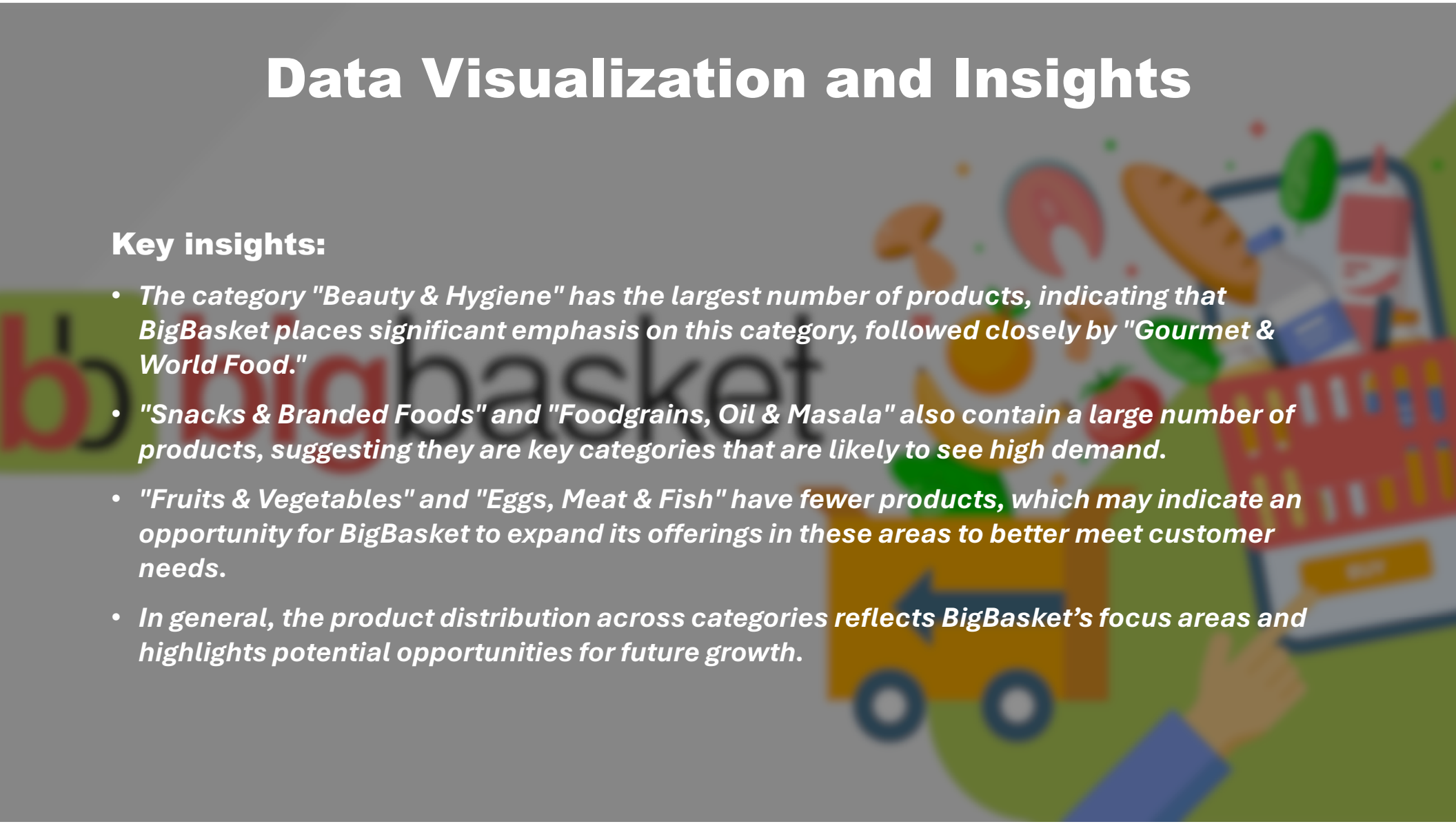


Data Visualization and Insights

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



Data Visualization and Insights

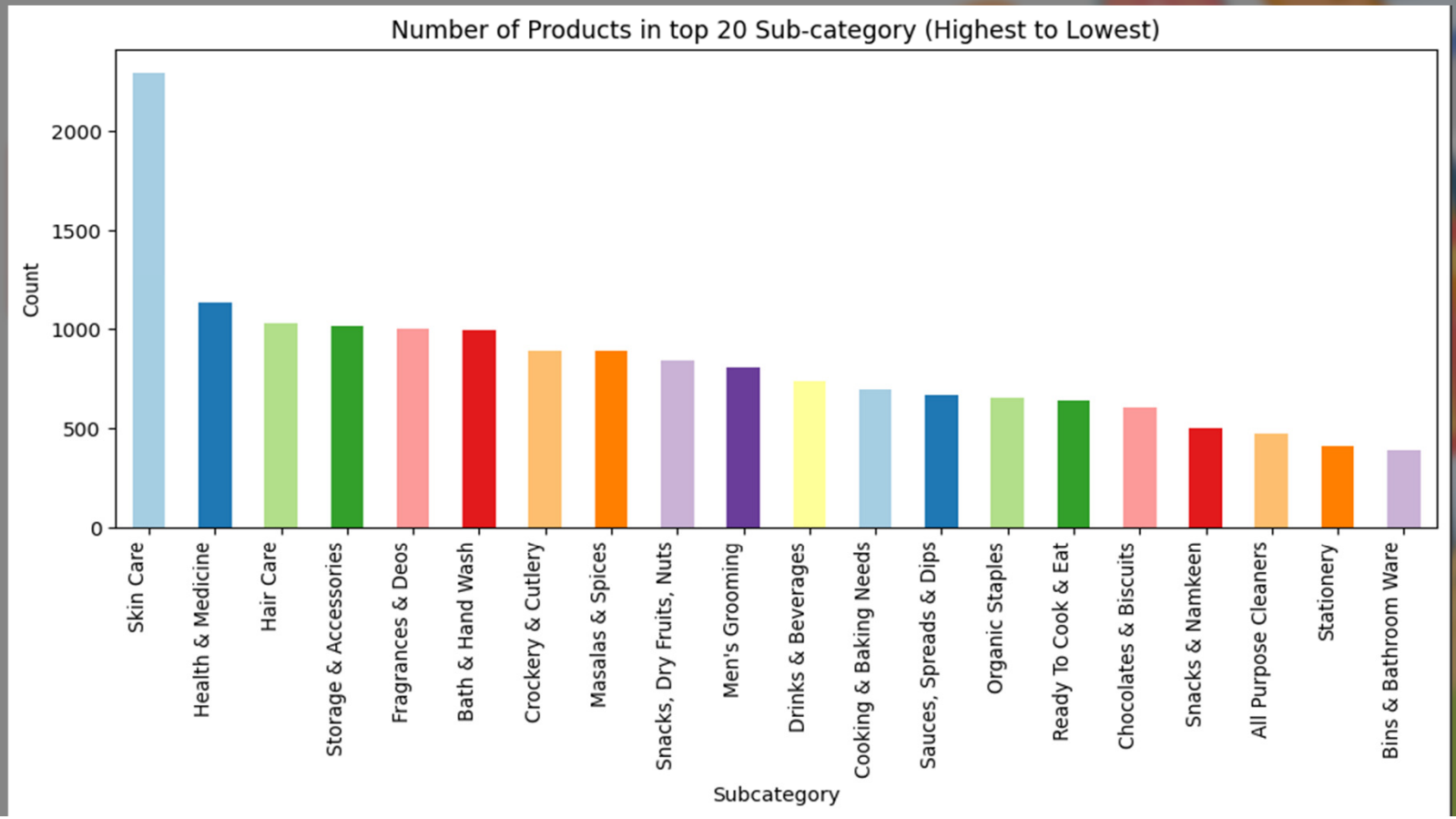


Key insights:

- *The category "Beauty & Hygiene" has the largest number of products, indicating that BigBasket places significant emphasis on this category, followed closely by "Gourmet & World Food."*
- *"Snacks & Branded Foods" and "Foodgrains, Oil & Masala" also contain a large number of products, suggesting they are key categories that are likely to see high demand.*
- *"Fruits & Vegetables" and "Eggs, Meat & Fish" have fewer products, which may indicate an opportunity for BigBasket to expand its offerings in these areas to better meet customer needs.*
- *In general, the product distribution across categories reflects BigBasket's focus areas and highlights potential opportunities for future growth.*

Data Visualization and Insights

Barchart : Plot the distribution of number of products in Top 20 sub-category.



Data Visualization and Insights

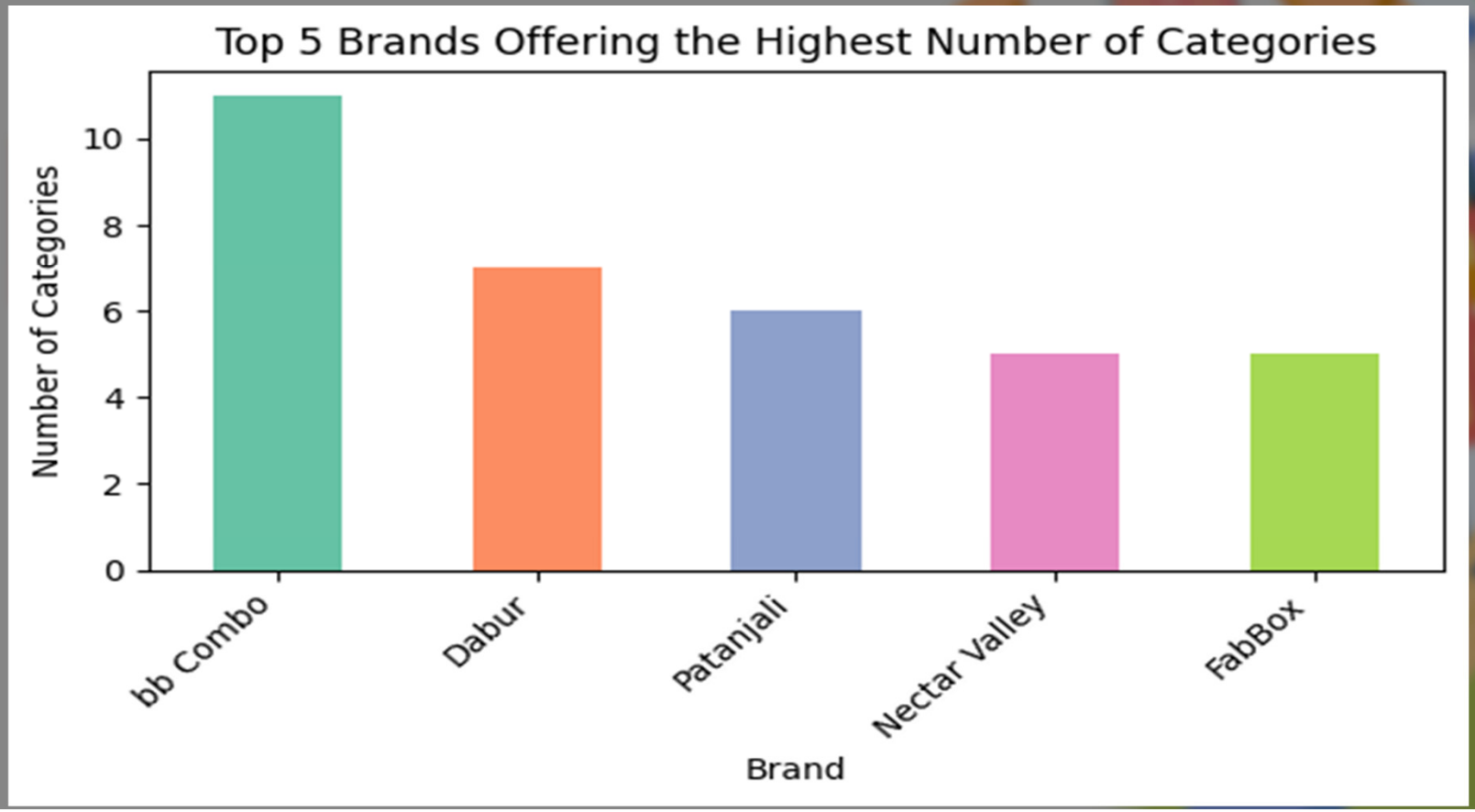
Barchart : Plot the distribution of number of products in Top 20 sub-category.

Key insights:

- *"Skin Care" is the top sub-category, offering the highest number of products, followed closely by "Health & Medicine."*
- *After the top three sub-categories ("Skin Care," "Health & Medicine," and "Hair Care"), there is a noticeable drop in product count.*
- *It's important to note that all three of the leading sub-categories belong to the "Beauty & Hygiene" category, suggesting that BigBasket focuses heavily on this area.*
- *The other sub-categories have fairly consistent product counts, with some minor variations.*

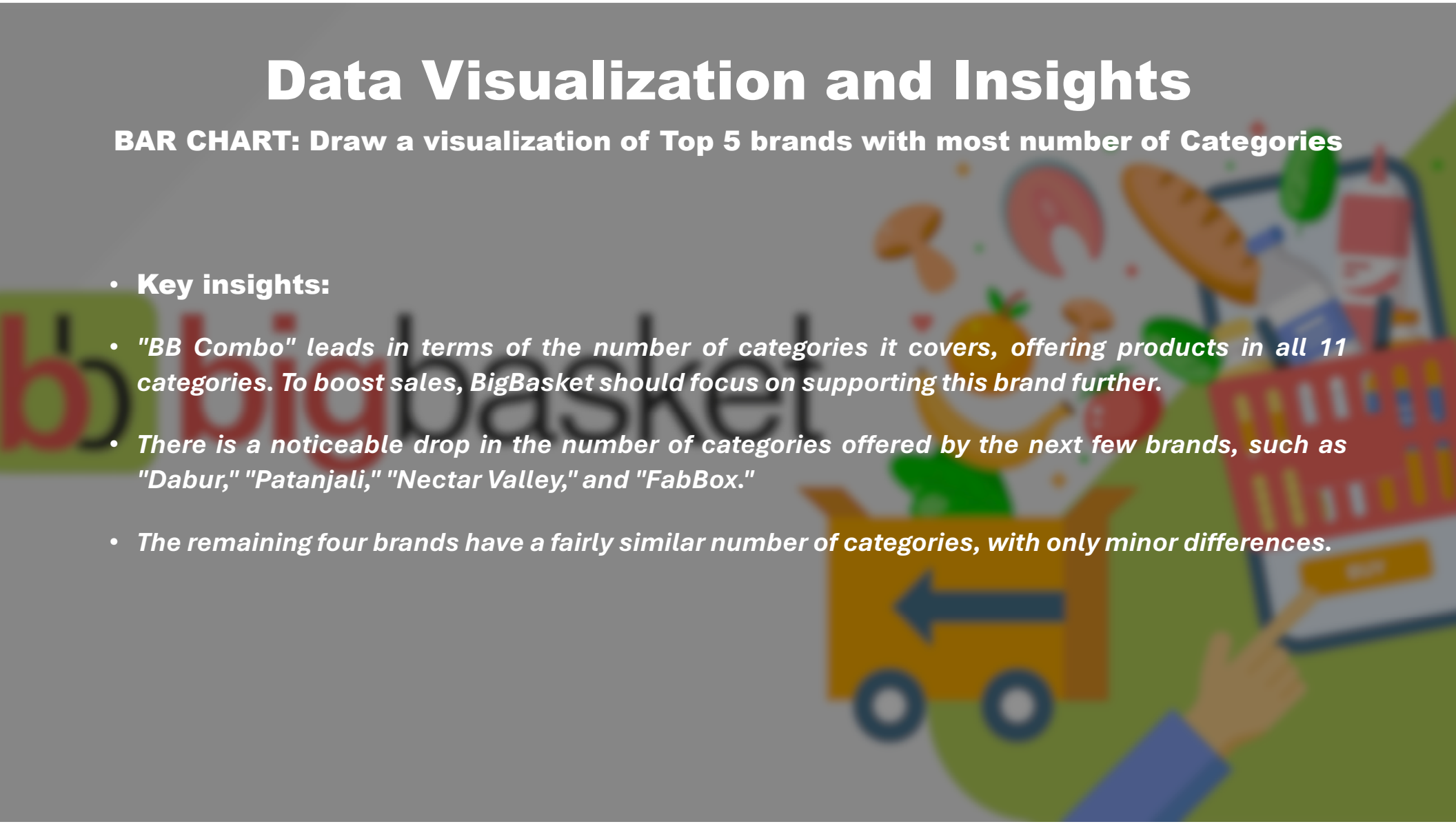
Data Visualization and Insights

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



Data Visualization and Insights

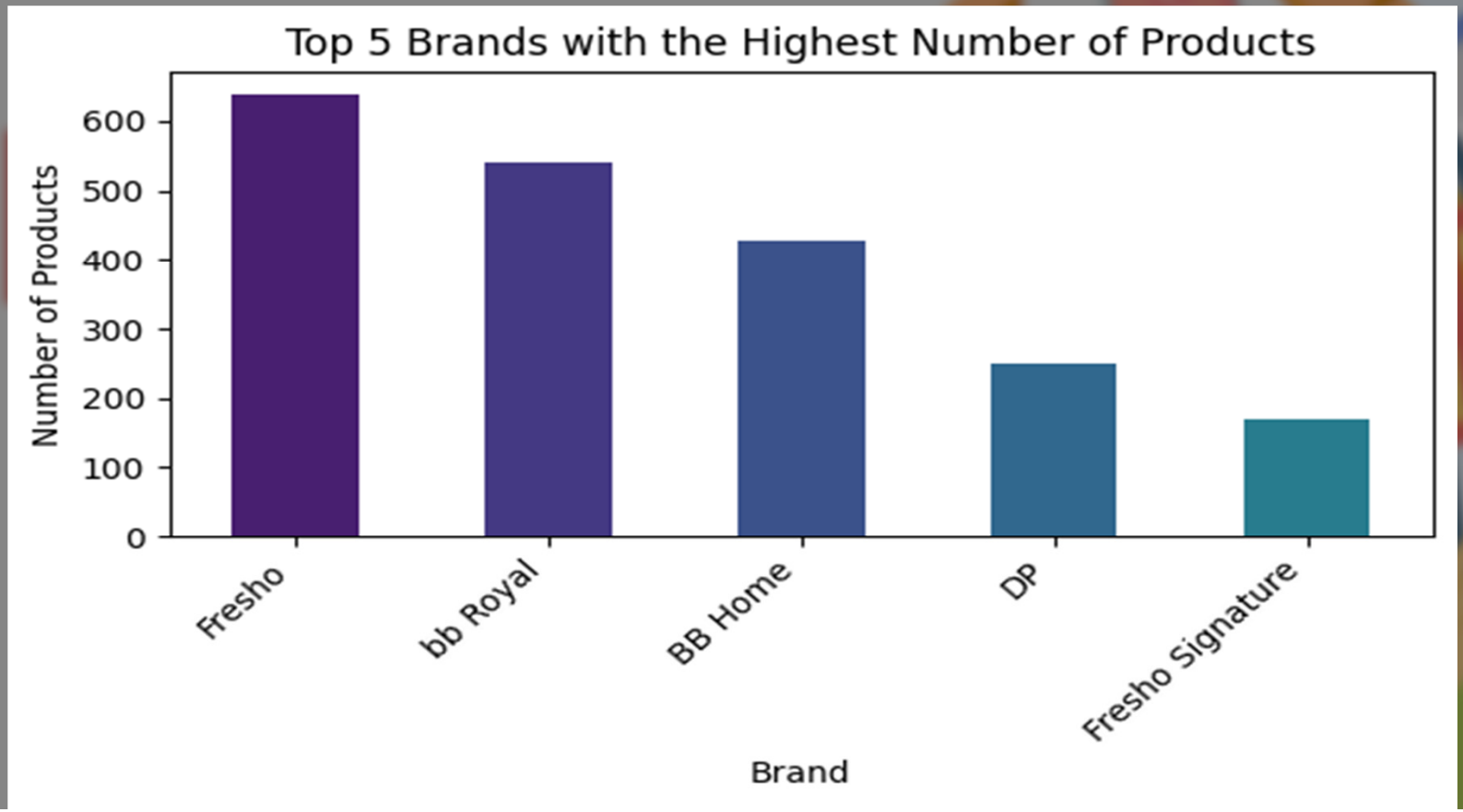
BAR CHART: Draw a visualization of Top 5 brands with most number of Categories



- **Key insights:**
- *"BB Combo" leads in terms of the number of categories it covers, offering products in all 11 categories. To boost sales, BigBasket should focus on supporting this brand further.*
- *There is a noticeable drop in the number of categories offered by the next few brands, such as "Dabur," "Patanjali," "Nectar Valley," and "FabBox."*
- *The remaining four brands have a fairly similar number of categories, with only minor differences.*

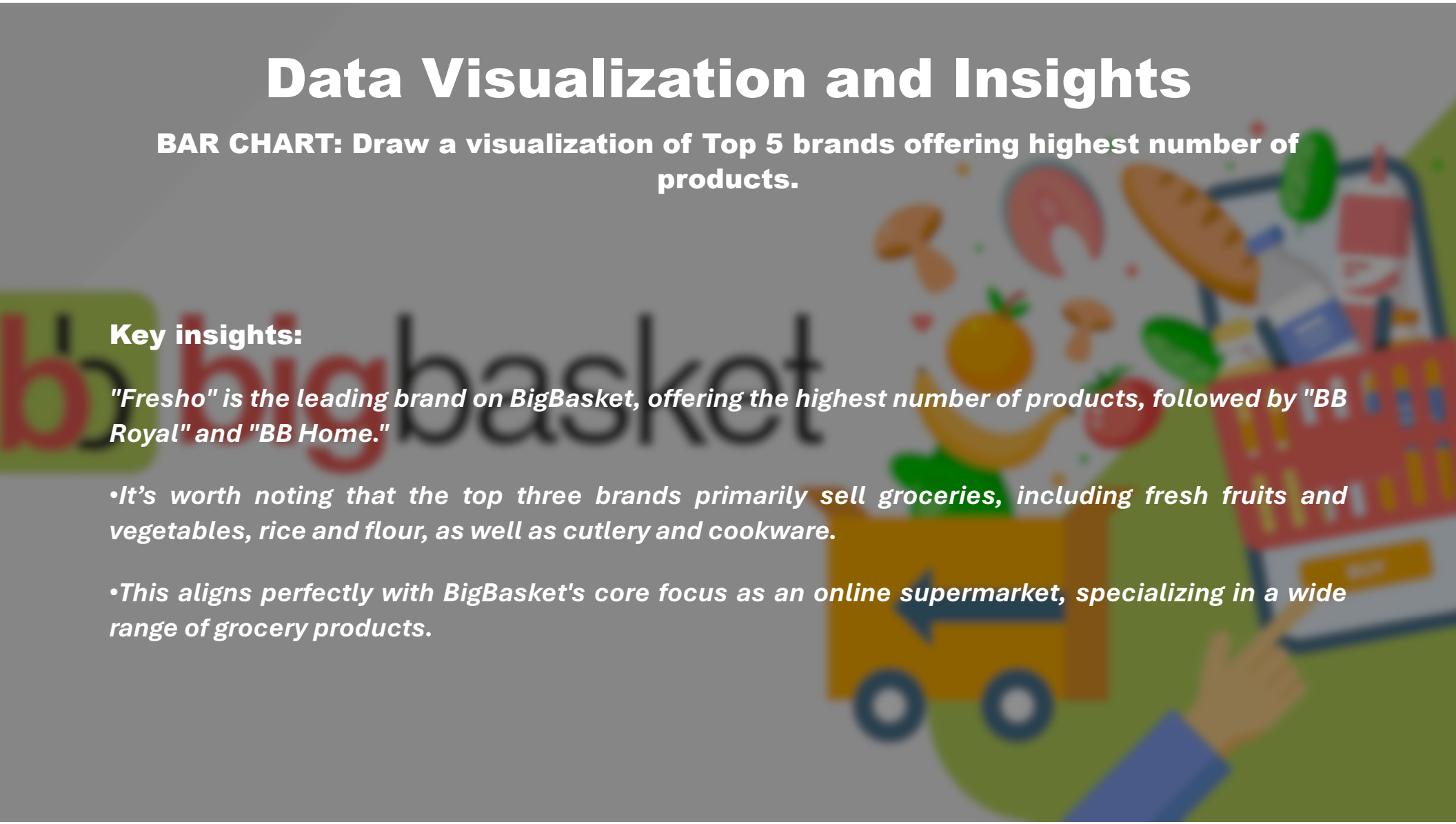
Data Visualization and Insights

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:

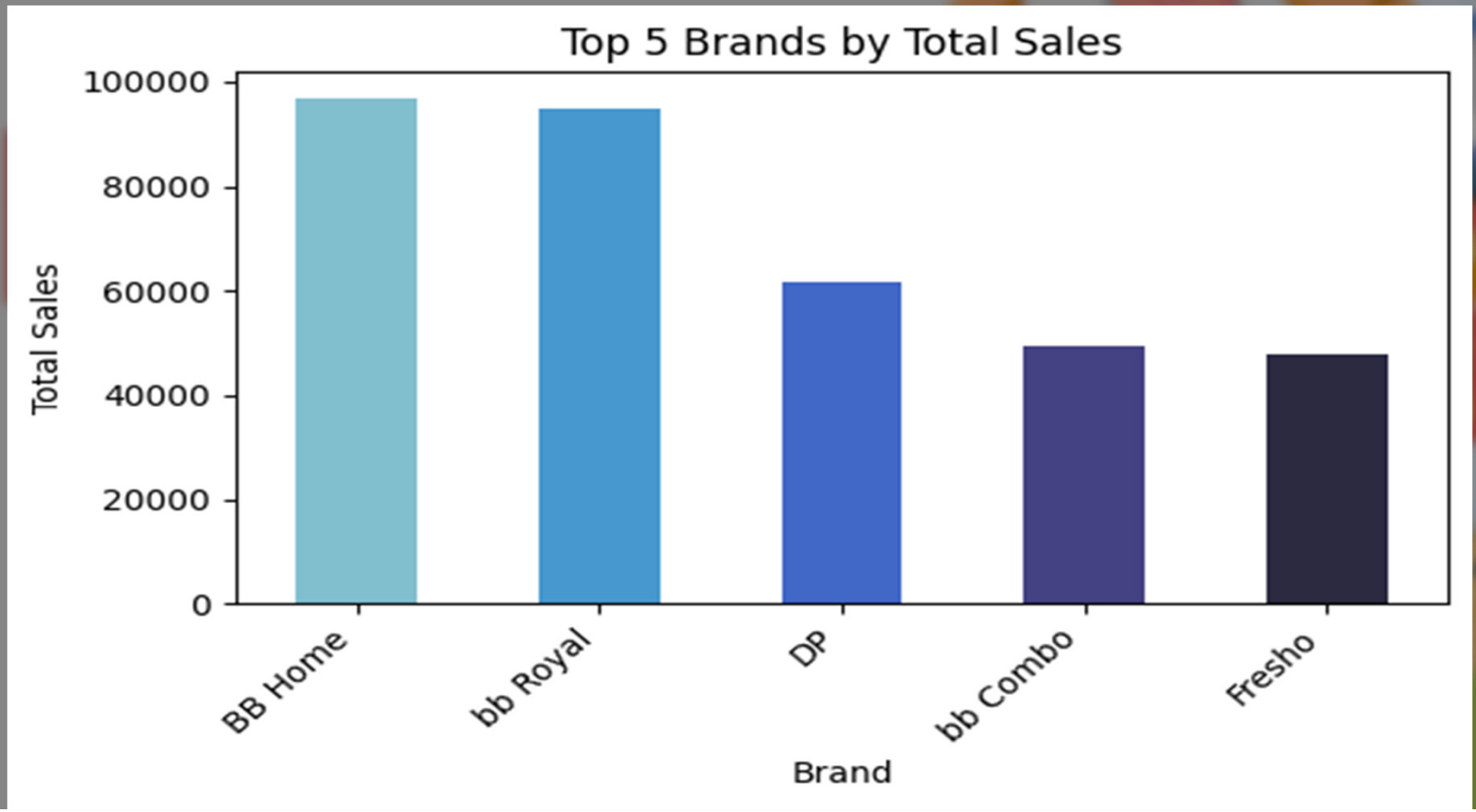
"Fresho" is the leading brand on BigBasket, offering the highest number of products, followed by "BB Royal" and "BB Home."

•It's worth noting that the top three brands primarily sell groceries, including fresh fruits and vegetables, rice and flour, as well as cutlery and cookware.

•This aligns perfectly with BigBasket's core focus as an online supermarket, specializing in a wide range of grocery products.

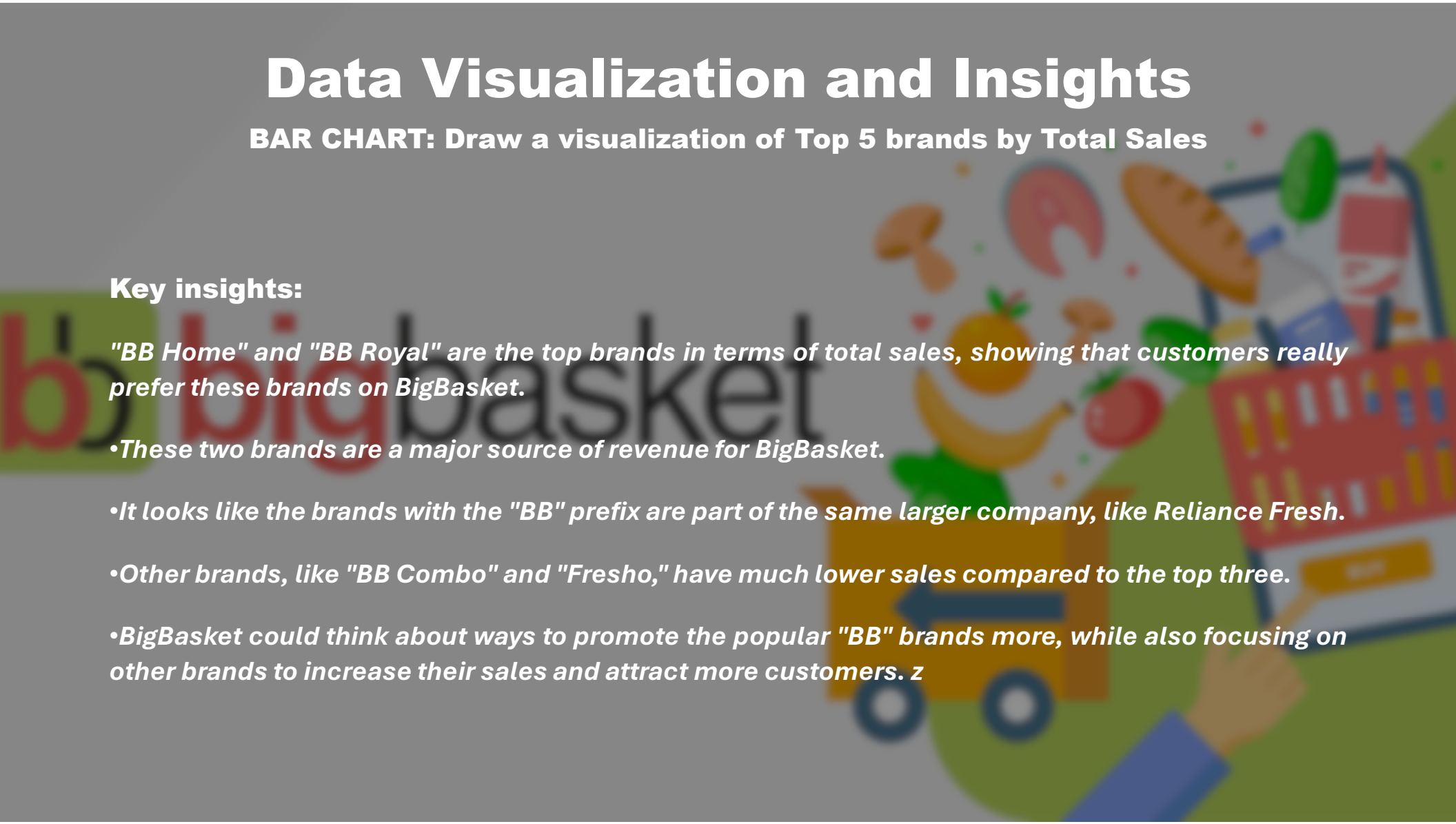
Data Visualization and Insights

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" are the top brands in terms of total sales, showing that customers really prefer these brands on BigBasket.

•These two brands are a major source of revenue for BigBasket.

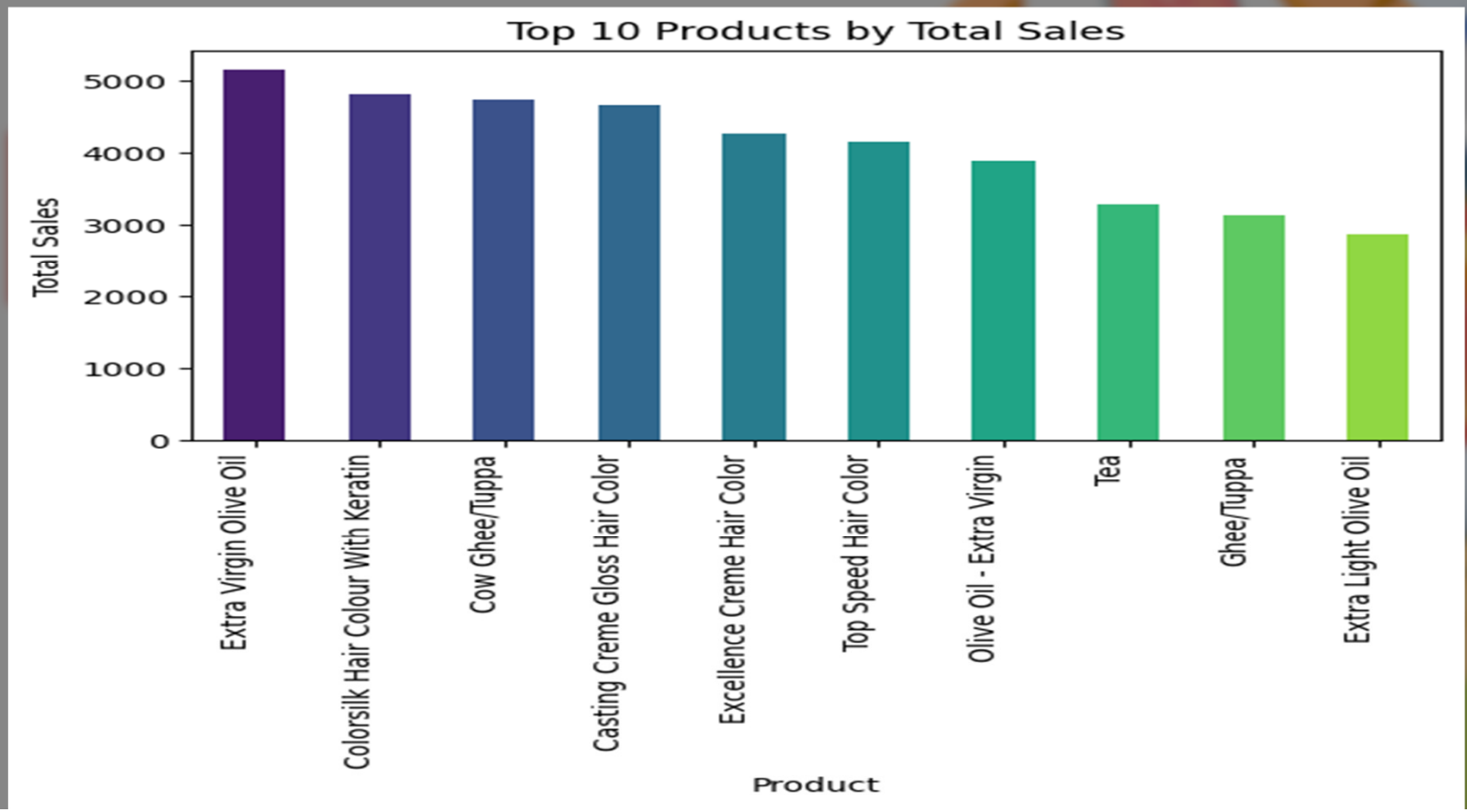
•It looks like the brands with the "BB" prefix are part of the same larger company, like Reliance Fresh.

•Other brands, like "BB Combo" and "Fresho," have much lower sales compared to the top three.

•BigBasket could think about ways to promote the popular "BB" brands more, while also focusing on other brands to increase their sales and attract more customers. z

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:

Key insights:

"Extra Virgin Olive Oil" is the best-selling product, with sales much higher than the other products.

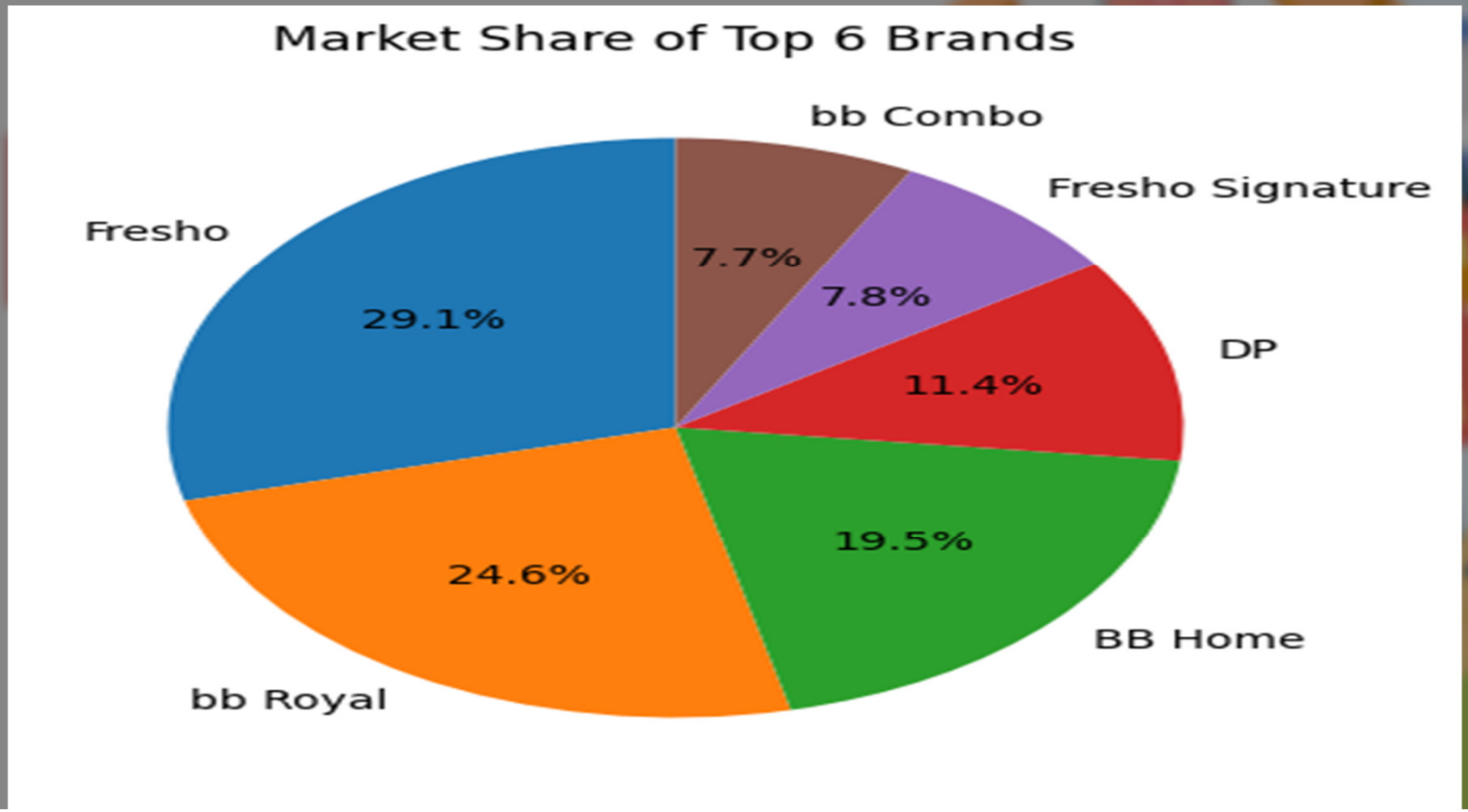
•"Colorsilk Hair Colour with Keratin" is the second best-seller, followed by "Cow Ghee/Tuppa" and "Casting Creme Gloss Hair Color."

•The next three products—"Excellence Creme Hair Color," "Top Speed Hair Color," and "Olive Oil - Extra Virgin"—have similar sales, with "Excellence Creme Hair Color" slightly ahead.

•The top 10 best-selling products are mostly from the "Beauty" (4 products) and "Foodgrains/Gourmet" (5 products) categories. This matches the findings from our earlier analysis in the 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" holds the largest market share at 29.1% among the top 6 brands, showing its strong popularity on BigBasket.

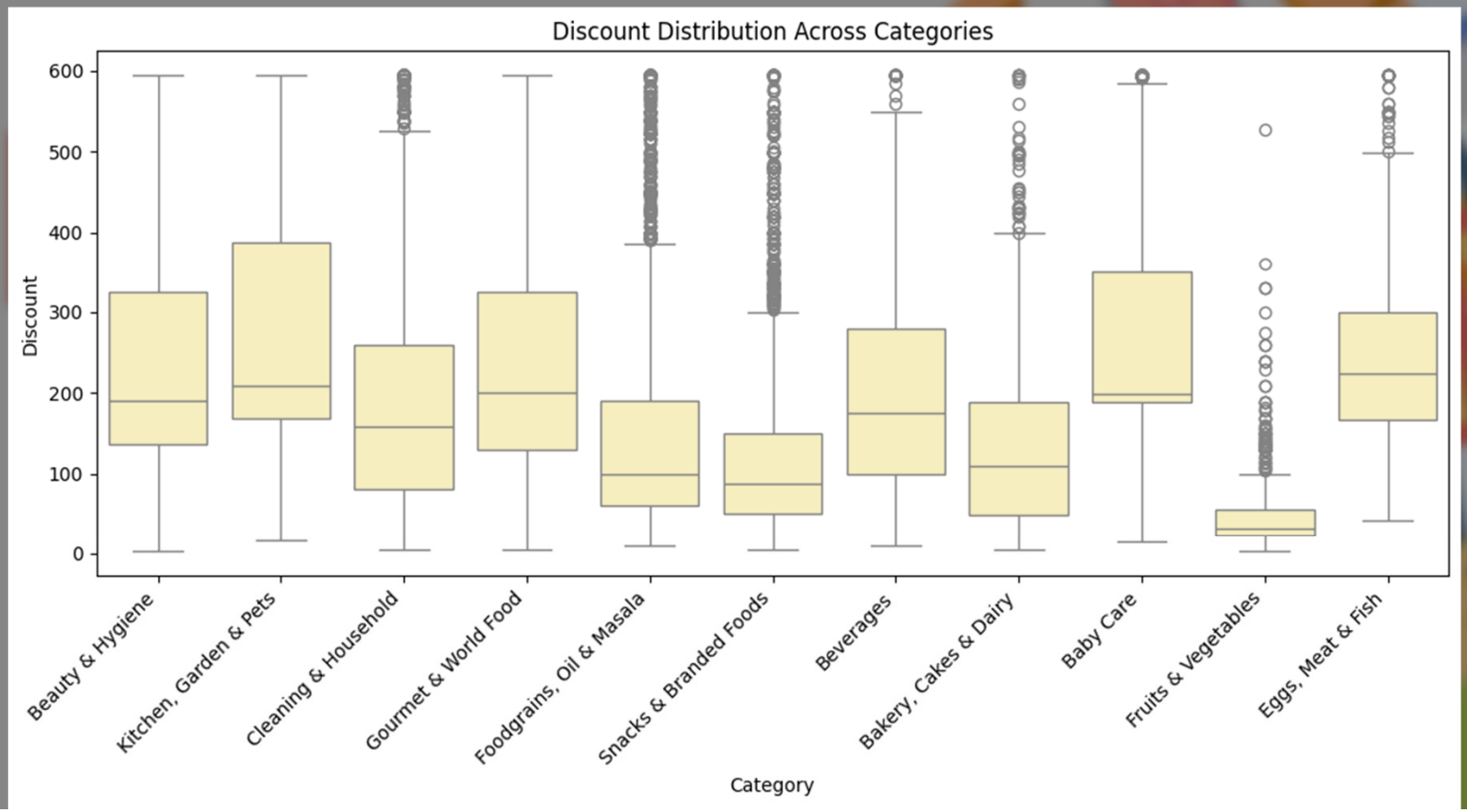
•"BB Royal" and "BB Home" also have significant market shares of 24.6% and 19.5%, indicating good brand recognition and customer loyalty.

•The chart shows that BigBasket offers a wide range of products, including categories like Baby Care ("bb Combo"), Cleaning & Household ("DP"), and Garden & Pets ("BB Home").

•BigBasket could focus on strengthening the market position of "Fresho," "BB Royal," and "BB Home" while also finding ways to grow the market share of other brands.

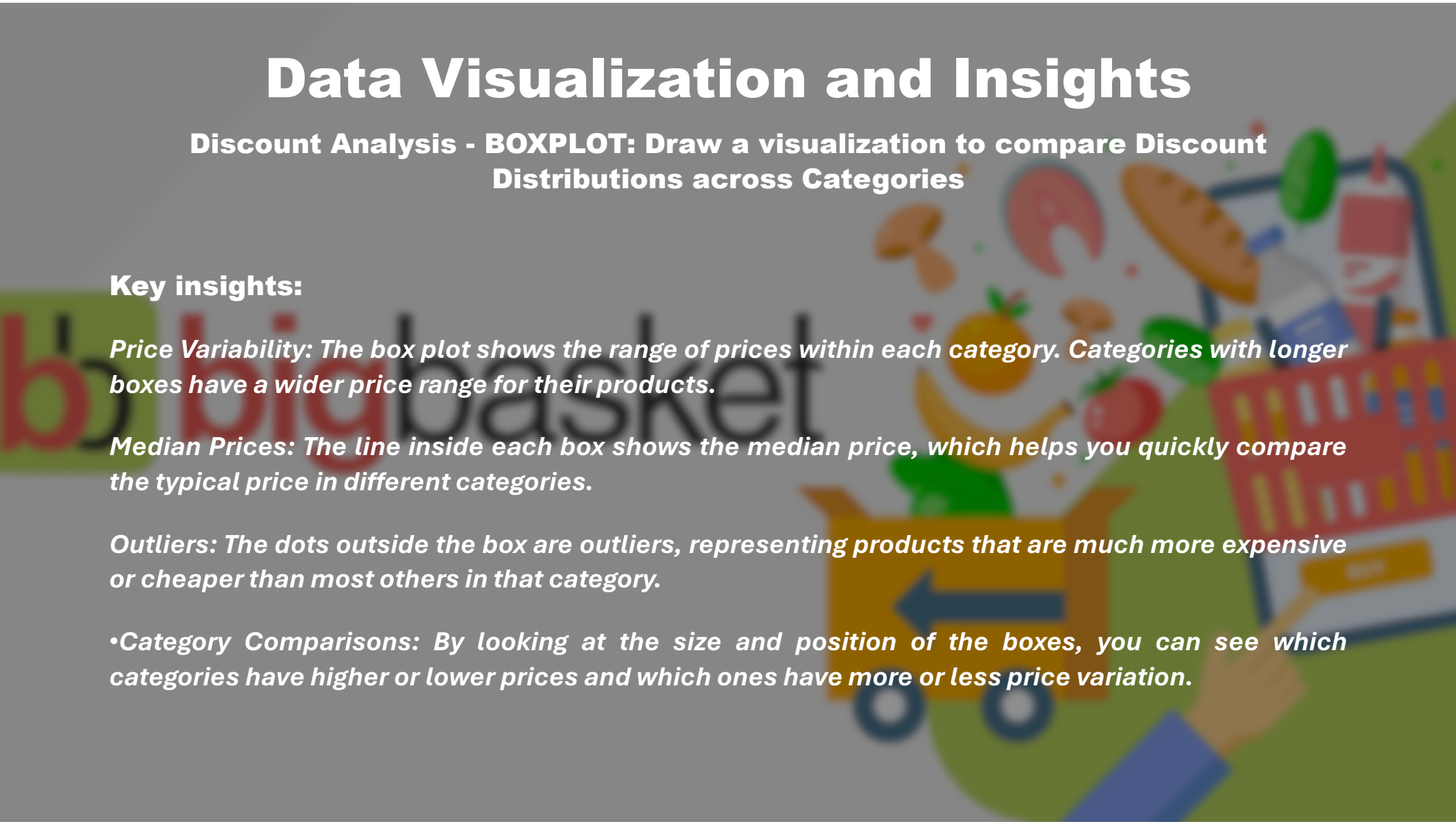
Data Visualization and Insights

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 shows the range of prices within each category. Categories with longer boxes have a wider price range for their products.

Median Prices: The line inside each box shows the median price, which helps you quickly compare the typical price in different categories.

Outliers: The dots outside the box are outliers, representing products that are much more expensive or cheaper than most others in that category.

•Category Comparisons: By looking at the size and position of the boxes, you can see which categories have higher or lower prices and which ones have more or less price variation.

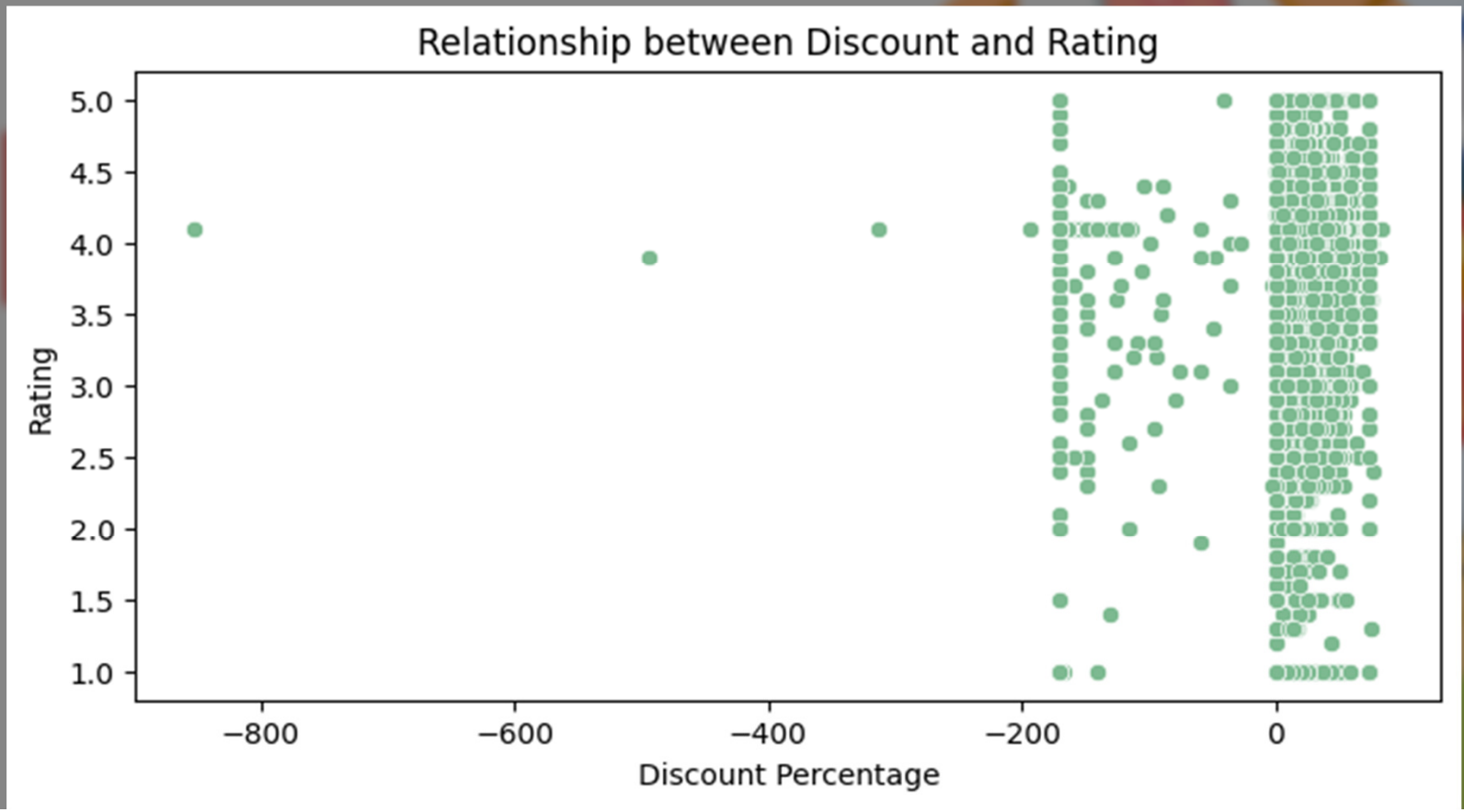
Data Visualization and Insights

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

- ***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

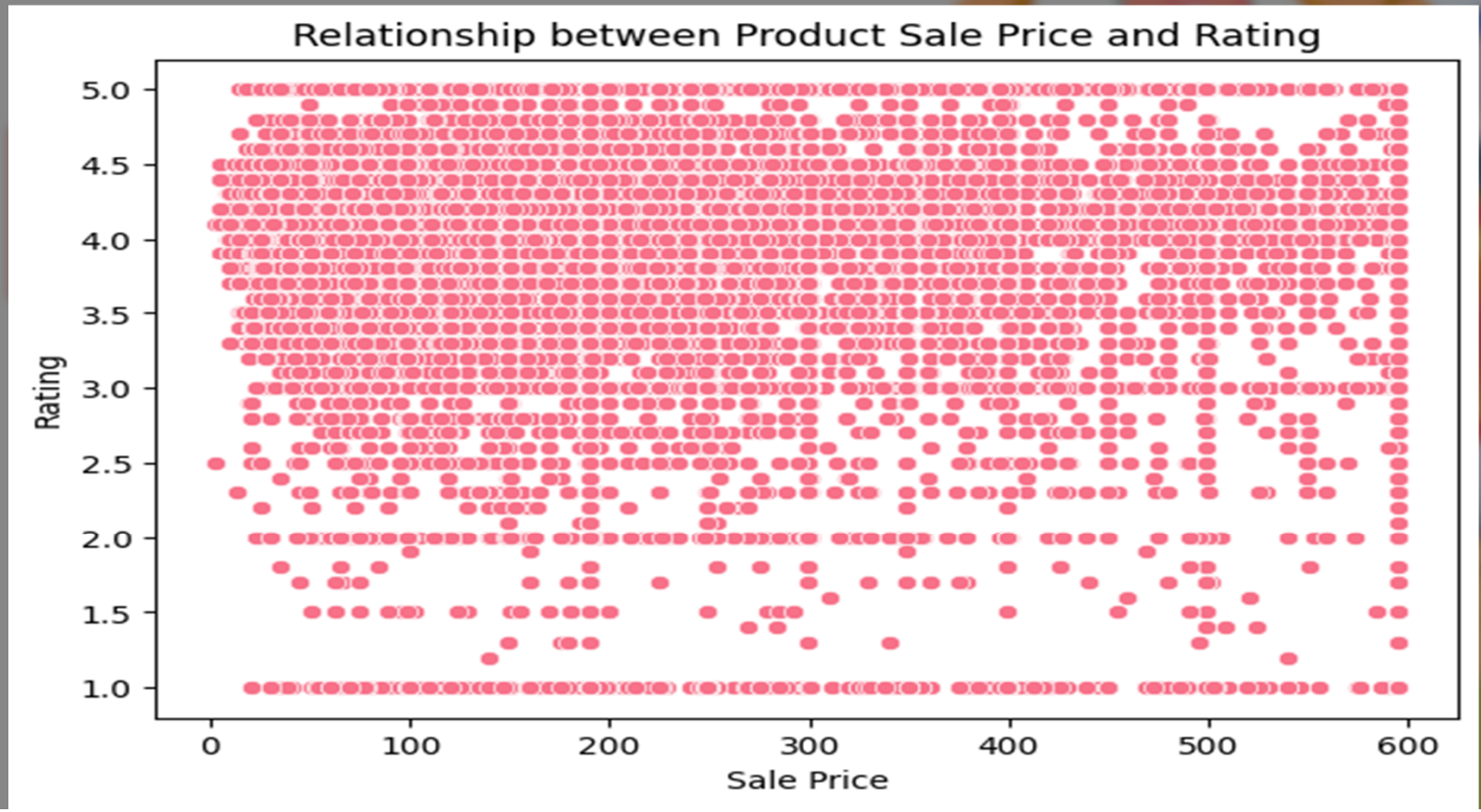
Discount Analysis - SCATTER PLOT: Draw a visualization to see if there's any relationship between Discount and Rating.

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*

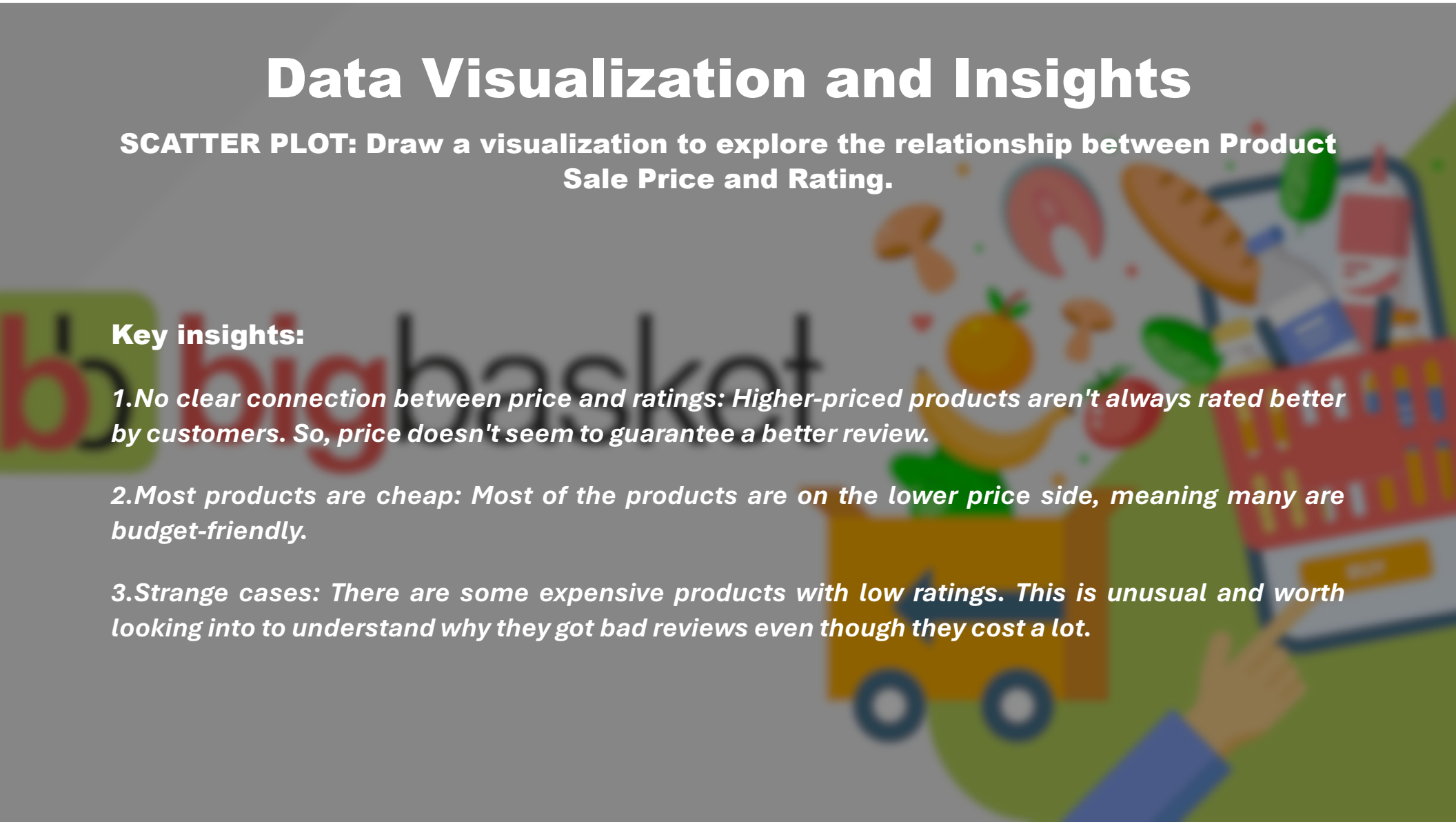
Data Visualization and Insights

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:

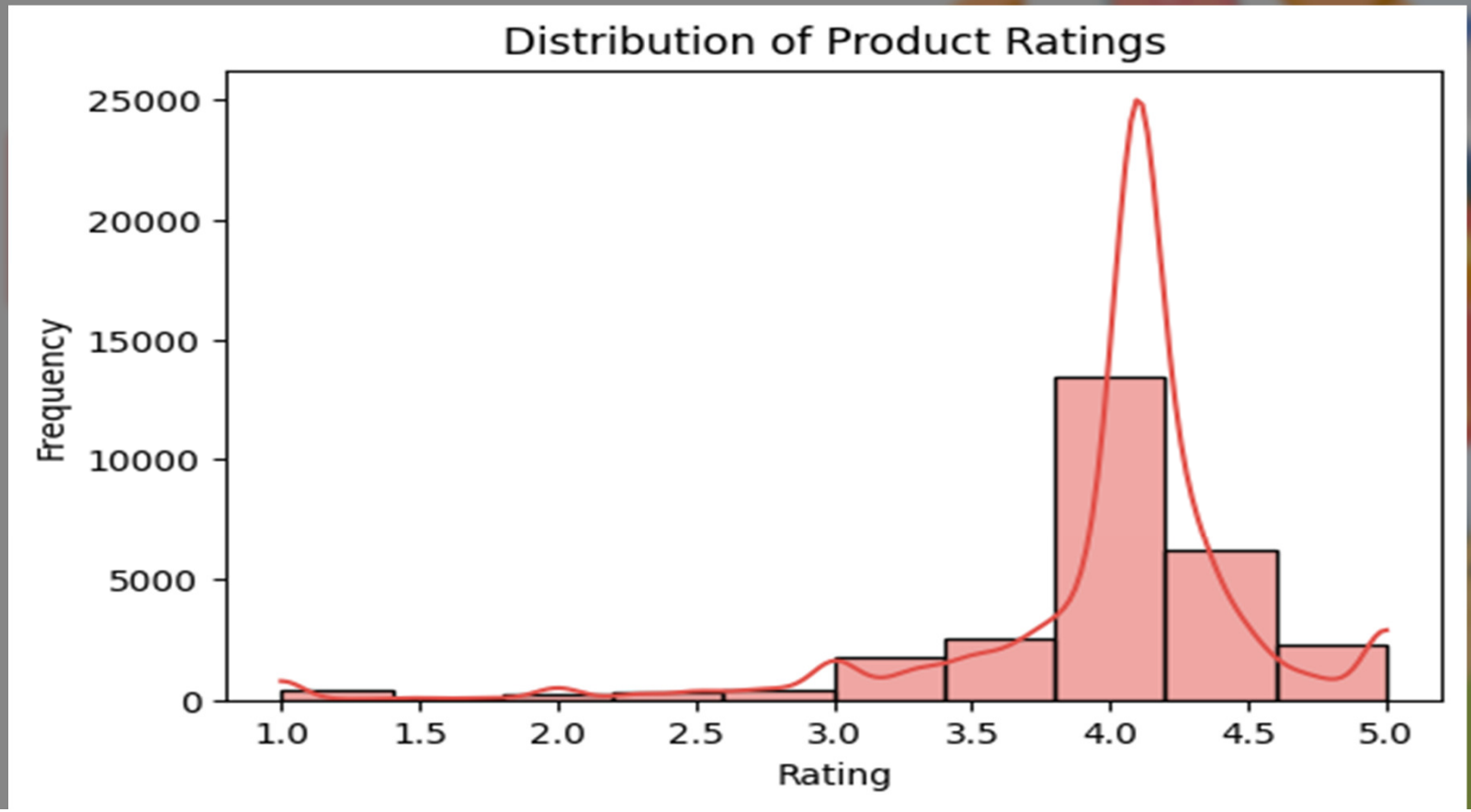
1.No clear connection between price and ratings: Higher-priced products aren't always rated better by customers. So, price doesn't seem to guarantee a better review.

2.Most products are cheap: Most of the products are on the lower price side, meaning many are budget-friendly.

3.Strange cases: There are some expensive products with low ratings. This is unusual and worth looking into to understand why they got bad reviews even though they cost a lot.

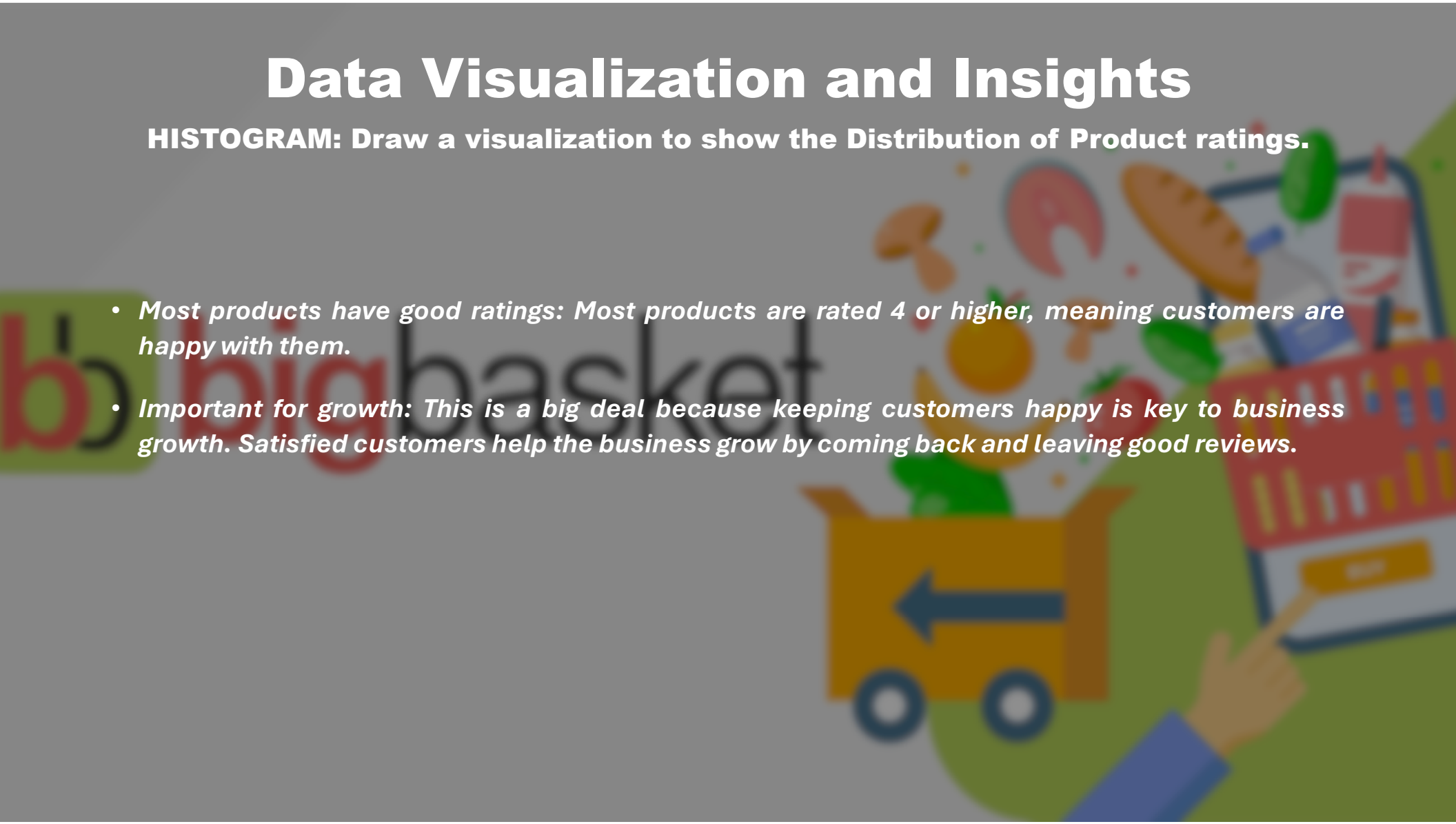
Data Visualization and Insights

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



Data Visualization and Insights

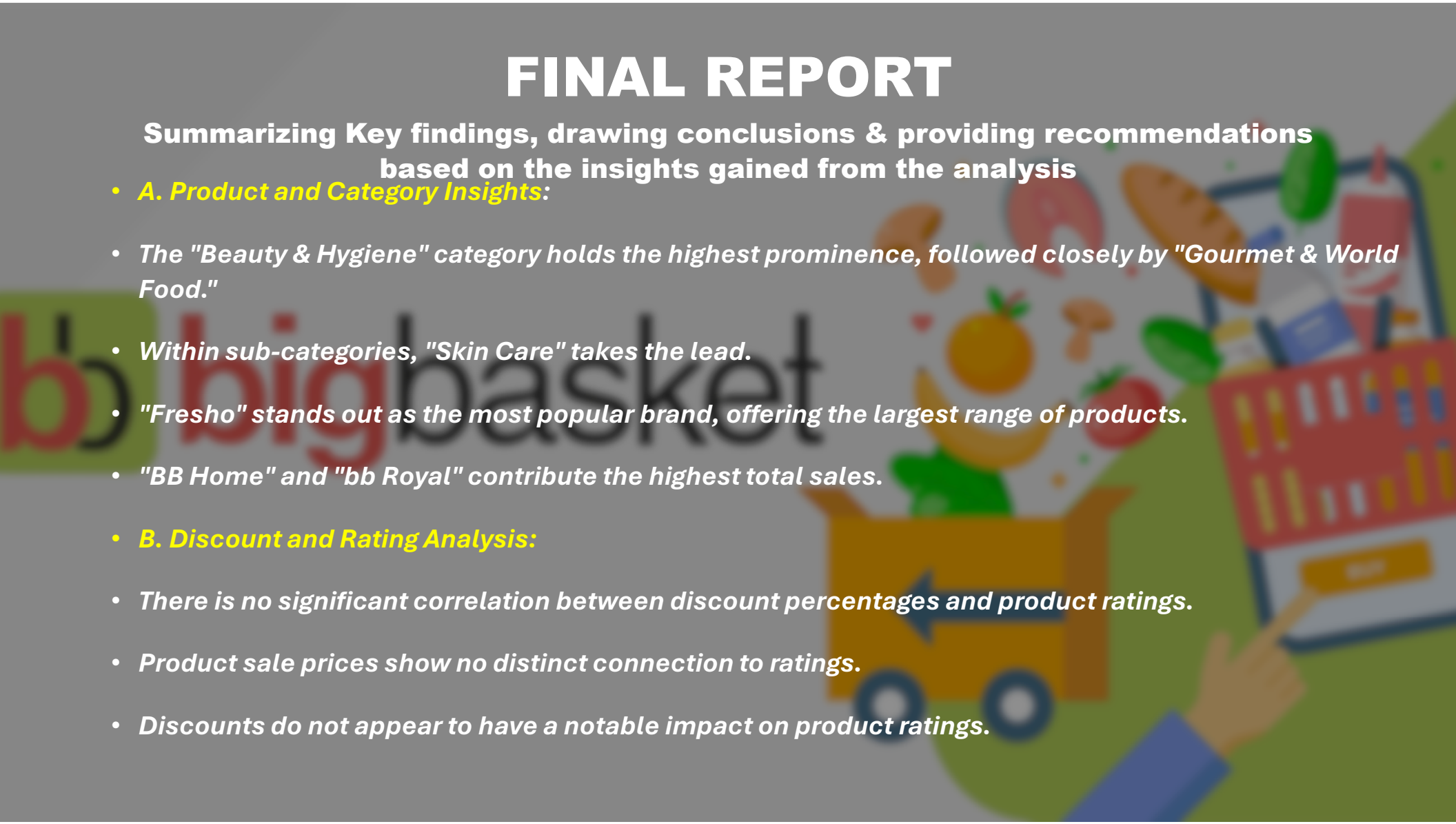
HISTOGRAM: Draw a visualization to show the Distribution of Product ratings.



- *Most products have good ratings: Most products are rated 4 or higher, meaning customers are happy with them.*
- *Important for growth: This is a big deal because keeping customers happy is key to business growth. Satisfied customers help the business grow by coming back and leaving good reviews.*

FINAL REPORT

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



- ***A. Product and Category Insights:***
- *The "Beauty & Hygiene" category holds the highest prominence, followed closely by "Gourmet & World Food."*
- *Within sub-categories, "Skin Care" takes the lead.*
- *"Fresho" stands out as the most popular brand, offering the largest range of products.*
- *"BB Home" and "bb Royal" contribute the highest total sales.*
- ***B. Discount and Rating Analysis:***
- *There is no significant correlation between discount percentages and product ratings.*
- *Product sale prices show no distinct connection to ratings.*
- *Discounts do not appear to have a notable impact on product ratings.*

FINAL REPORT

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

- *C. Rating Insights:*
- *Product ratings are predominantly skewed toward higher ratings (4 or above), indicating strong customer satisfaction.*
- *Customers tend to give favorable ratings regardless of the product's price, suggesting that quality outweighs pricing concerns.*
- *General Summary of Key Findings:*
- *Product Distribution: An analysis of product distribution across categories and sub-categories highlighted the most and least popular product types. Certain categories and sub-categories show significantly higher popularity compared to others.*

FINAL REPORT

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

- *General Summary of Key Findings:*
- *Brand Analysis: Leading brands were identified based on their presence across categories, the number of products offered, and total sales. A pie chart was used to visualize their market share.*
- *Discount and Rating Relationship: A scatter plot was used to examine the connection between discounts and ratings, revealing no significant correlation.*
- *Price and Rating Relationship: The relationship between product sale price and ratings was visualized, showing no evident trend.*
- *Rating Distribution: A histogram of product ratings revealed a strong concentration of higher ratings, indicating overall customer satisfaction.*

FINAL REPORT

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

- *Conclusions:*
- *Big Basket places significant emphasis on the "Beauty & Hygiene" and "Gourmet & World Food" categories, with a notable focus on the "Skin Care" sub-category.*
- *"Fresho" is a key brand for Big Basket, while "BB Home" and "bb Royal" serve as major contributors to revenue.*
- *These leading brands dominate the market by offering a wide variety of products, achieving high sales, and securing substantial market share.*
- *Discounts do not necessarily lead to higher ratings; instead, product quality and customer satisfaction are the primary drivers of positive feedback.*
- *Customers are largely satisfied with Big Basket's offerings, reflecting a positive overall shopping experience.*

FINAL REPORT

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

- *Recommendations:*
- *Big Basket should prioritize promoting products in popular categories and sub-categories, as these drive significant revenue for the brand.*
- *Expanding the product range in categories such as "Fruits & Vegetables" and "Eggs, Meat & Fish" would help attract a broader customer base.*
- *Leveraging the strong market presence of "Fresho," "BB Home," and "bb Royal" can further accelerate growth and reinforce customer loyalty.*
- *To enhance profitability, Big Basket should strategically partner with new brands and focus on marketing efforts through platforms like YouTube ads and TV commercials to strengthen its market presence.*

FINAL REPORT

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

- *Recommendations:*
- *Maintain a strong focus on product quality and customer experience to sustain high ratings and satisfaction levels.*
- *Develop strategies to enhance ratings for lower-rated products by addressing customer feedback and improving quality where needed.*
- *Implement targeted discount campaigns tailored to customer preferences and specific product categories to maximize engagement and sales.*
- *Regularly monitor customer satisfaction metrics and take prompt action to resolve any emerging issues to maintain a positive shopping experience.*


bigbasket
THANK YOU FOR READING