Exploratory Data Analysis

BigBasket Exploratory Data Analysis

About BigBasket,

BigBasket, a prominent online grocery platform in India, was founded in 2011 and gained popularity for its convenient home delivery service. Tata Digital, a subsidiary of the Tata Group, is a key player in digital businesses. If recent developments are accurate, Tata Digital has acquired BigBasket, marking Tata's entry into the online grocery and e-commerce sector. This move aligns with the broader trend of traditional businesses expanding their digital presence to stay competitive. For the latest information, it is advisable to check the most recent and reliable sources.

Goal of the Project:

The goal of my Exploratory Data Analysis (EDA) project on BigBasket is to gain valuable insights and a comprehensive understanding of the dataset related to the online grocery platform. Through detailed exploration, visualization, and statistical analysis, I aim to uncover patterns, trends, and relationships within the data. This exploration will help identify key factors influencing customer behavior, product preferences, and overall business performance on BigBasket. The insights derived from the EDA will not only provide a descriptive overview of the dataset but also serve as a foundation for making informed business decisions, optimizing operations, and enhancing the user experience on the platform.

Dataset Description

Dataset Description:

The dataset comprises information related to items listed on the BigBasket online grocery platform. It consists of the following columns:

1. **'index'**: A unique identifier for each item in the dataset.
2. **'product'**: The name or title of the product listed on Big Basket.
3. **'category'**: The overarching category to which the product belongs.
4. **'sub_category'**: A more specific classification under the main category.
5. **'brand'**: The brand associated with the product.
6. **'sale_price'**: The selling price of the product on BigBasket.
7. **'market_price'**: The regular market price of the product.
8. **'type'**: The type or variant of the product.
9. **'rating'**: The rating assigned to the product, indicating customer satisfaction.
10. **'description'**: A brief description or information about the product.

This dataset is valuable for conducting exploratory data analysis (EDA) to uncover patterns, trends, and relationships within the data. Analysis of these features can provide insights into cus- tomer preferences, pricing dynamics, product categorization, and overall market trends on the BigBasket platform. The 'index' column serves as a unique identifier for easy referencing and data management.

Dataset Description

Dataset Source:

The dataset has been scraped from the official website of Big-Basket, the leading online grocery platform. The information collected includes details about items listed on the website, such as product names, categories, sub-categories, brands, sale prices, market prices, product types, ratings, and descriptions.
The scraping process involved extracting relevant data from the web pages of BigBasket to create a comprehensive dataset for analysis.

Dataset Size:

The dataset under consideration is of substantial size, consisting of 27,555 rows and 10 columns. Each row corresponds to a unique entry, representing a specific item listed on the BigBas- ket online grocery platform. The columns provide a diverse set of information, including details about the product, category, sub-category, brand, pricing, type, rating, and a brief descrip- tion.

The significant number of rows allows for a comprehensive anal- ysis of the dataset, enabling a thorough exploration of trends, patterns, and insights related to the products available on Big- Basket. This sizeable dataset provides a rich source of informa- tion that can contribute to a detailed and meaningful Explorato- ry Data Analysis (EDA) to better understand customer prefer- ences, market dynamics, and other relevant factors.

Data Cleaning and Preparation

Dealing with Null Values:

In the dataset, we have identified the following instances of missing values:

- Null values in 'Brand' column
- Null values in 'Product’ column.
- Six Null values in 'Sale_Price' column.
- 8,636 Null values in 'Rating' column.
- 115 Null values in 'Description' column.

To address these missing values, we will undertake the following actions:

Handling Nulls in 'Product' and 'Brand':

- For the 'Product' column, no imputation will be performed, as product names are crucial and cannot be accurately guessed or replaced.
- For the 'Brand' column, Null values will be replaced with 'Unknown' to denote missing brand information.

Dealing with Nulls in 'Sale_Price':

- Given the authenticity of market prices, we plan to fill the Null values in the 'Sale_Price' column with the corresponding values from the 'Market_Price' column.

Data Cleaning and Preparation

Handling Nulls in 'Rating':

■ Null values in the 'Rating' column will be assigned a value of 0, indicating no available rating. This approach ensures that the absence of a rating does not result in the loss of valuable information.

Dealing with Nulls in 'Description':

■ Null values in the 'Description' column will be imputed with the placeholder 'No description' to enhance the completeness and clarity of the dataset.

Designating 'Coffee' for Null Values in a Specific Product Category:

■ As a specific case related to a coffee product, Null values in the 'Product' column will be replaced with 'Coffee' as a placeholder. This is based on the understanding that this particular entry pertains to a coffee product and contains only one product.

By implementing these strategies, we aim to ensure the dataset remains informative and suitable for analysis, minimizing the impact of missing values on the overall quality and utility of the data.

Data Cleaning and Preparation

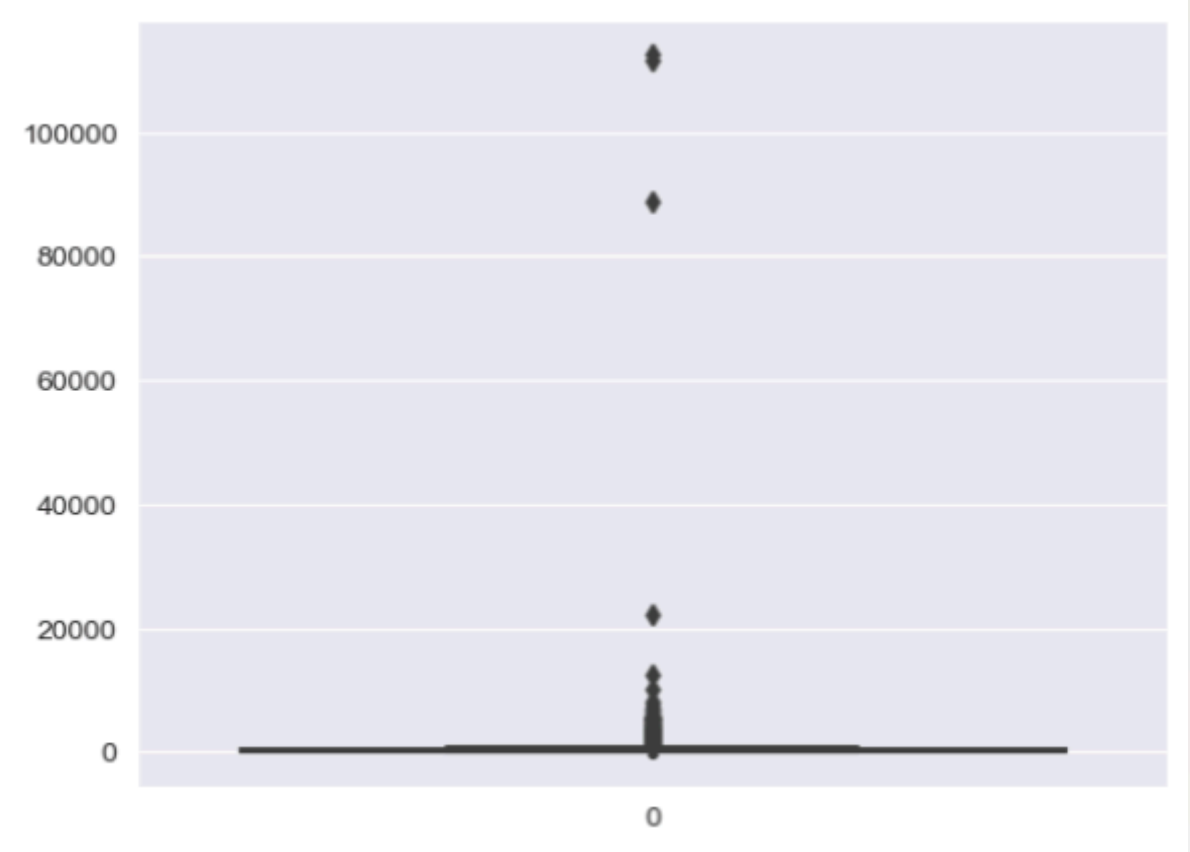
Outlier Detection:

Upon conducting an analysis of the dataset, several observations regarding outliers have been identified:

Prices Exceeding Market Price (MRP):

Four items have been observed with prices that exceed their corresponding Market Price (MRP). This is an unusual pattern and requires attention to ensure data accuracy and consistency.

Data Cleaning and Preparation



Identification of Outliers with High Values:

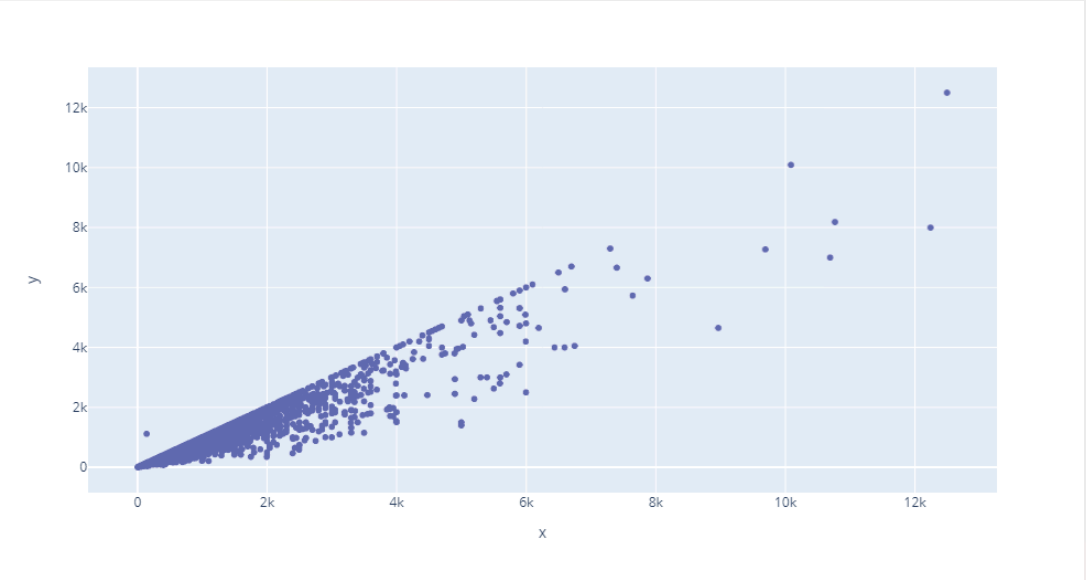
Outliers have been detected, with the majority of them having prices exceeding 20,000 INR. This extreme pricing raises concerns about the accuracy and validity of these entries.

False Entries in Sale Price:

False entries have been identified in the 'Sale_Price' field. To address this issue, further investigation was conducted, leading to the conclusion that Market Price is considered a more reliable metric in this context.

Data Cleaning and Preparation

Further Analysis and Data Rectification:



In addition to the previously mentioned observations, a specific anomaly has been identified where one product, with a market_- price below 500, is displaying a sale_price exceeding 1000. This unusual outlier is causing a deviation from the linear trend in the graph, impacting the overall data integrity.

To gain more precise insights, a scatter graph has been generated using Plotly. This visualization has facilitated the identification of a particular item with a market_price of 140 exhibiting an unexpect- ed sale_price of 1114.8.

Upon closer examination of this product, it has been determined that the listed sale_price of 1114.8 is a data entry error. The correct discounted price appears to be 114.8, but it has been erroneously recorded as 1114.8.

To rectify this discrepancy and ensure the accuracy of the dataset, the incorrect sale_price value is being substituted with the corre- sponding market_price. This correction aims to align the data with actual pricing information and eliminate inaccuracies that could affect the overall analysis and interpretation of the dataset.

Data Cleaning and Preparation

Feature Selection:

In the process of refining the dataset for analysis, it has been observed that the DataFrames already possess index numbers, rendering the inclusion of an additional index column unnecessary. To streamline the data and enhance simplicity, a decision has been made to eliminate the index column from the DataFrame.

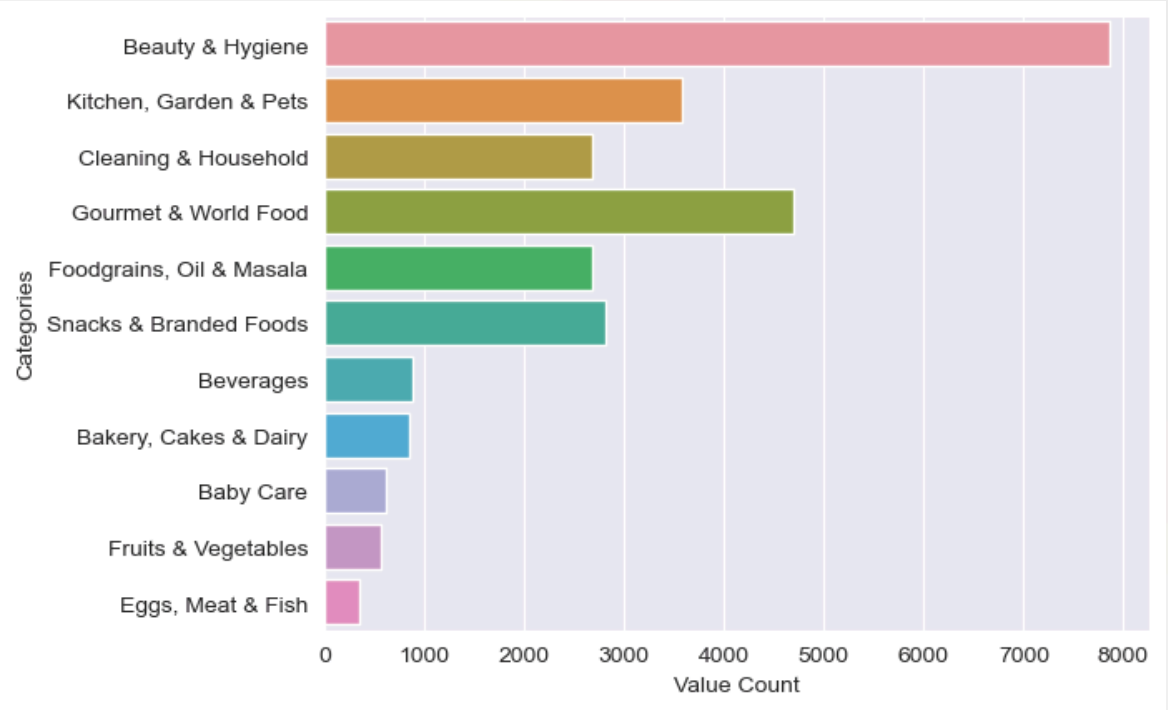
Action Taken:

■ The index column has been removed from the DataFrame as it does not provide additional meaningful information and only duplicates the existing index numbers.

By implementing this feature selection step, the dataset is now more concise and focused, eliminating redundant information and improving the efficiency of subsequent analyses. This practice adheres to the principle of maintaining data clarity and relevance, ensuring that only essential features are retained for meaningful insights.

Data Visualisations and Insights

An analysis of product quantities within different categories and subcategories has provided valuable insights into the distribution of items on the BigBasket platform.

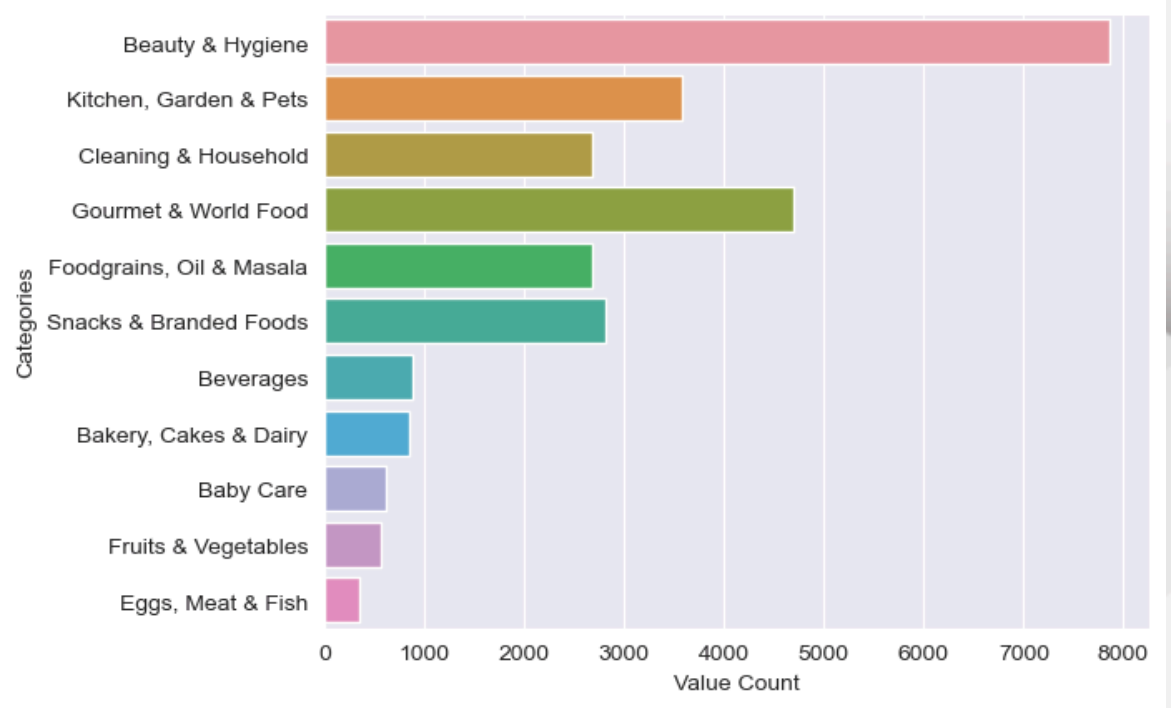


Observations and Key Insights:

'Beauty & Hygiene' Dominates:

The category 'Beauty & Hygiene' stands out with the highest product count, indicating that it is the top-selling category on the platform. This suggests a significant demand for personal care and hygiene products.
Food Industry Significance:

Data Visualisations and Insights



Food Industry Significance:

'Gourmet & World Food' is the second-highest in terms of product count, highlighting substantial demand in the food industry. The diversity and global appeal of gourmet and world food products contribute to their popularity.
Home and Pet Goods Demand:

Home and Pet Goods Demand:

The 'Kitchen, Garden, and Pet Goods' category ranks third, reflecting sustained demand. Consumer preferences for home improvement and the popularity of pet-related products contribute to its notable sales.
Widespread Appeal Categories:

Data Visualisations and Insights

Widespread Appeal Categories:

Similar quantities in 'Snack & Branded Food,' 'Foodgrains, Oil and Masala,' and 'Cleaning & Household' categories suggest widespread appeal, essential nature, and consistent usage, respectively.

Comparable Categories:

"Beverages" and "Bakery, Cakes & Dairy" consistently feature a comparable number of products, indicating a balanced assortment in these popular categories.
Niche Nature of Baby Care:

Niche Nature of Baby Care:

The "Baby Care" category ranks third in the least number of products, likely due to its niche market nature. Specialized products focusing on infant needs result in a smaller range.
Shelf Life Considerations:

Shelf Life Considerations:

The "Fruits & Vegetables" category has the second least quantity, primarily due to the shorter shelf life and continuous refrigeration requirements.
Freshness Demands in Protein:

Freshness Demands in Protein:

The "Eggs, Meat & Fish" category has the least number of products, likely due to customer demand for freshness. The complex supply chain, temperature sensitivity, and careful handling contribute to the limited product variety.
Notable Observation - Least Grocery Items:

Data Visualisations and Insights

Notable Observation - Least Grocery Items:

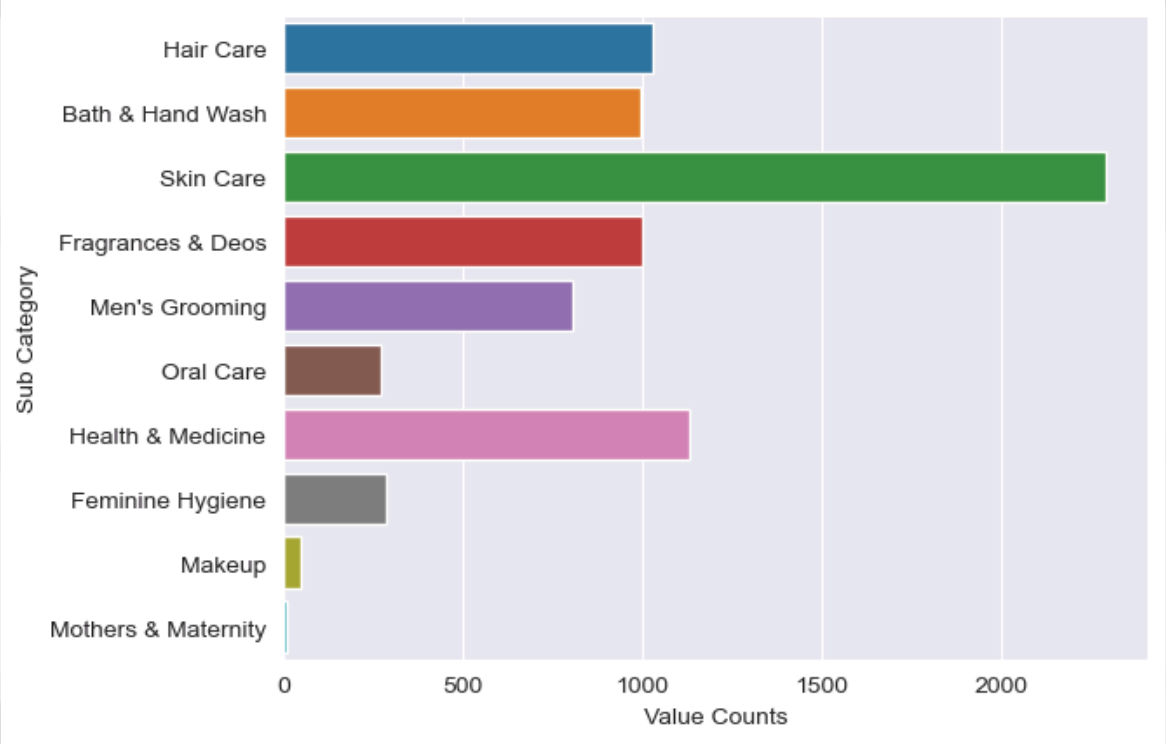
A significant insight is that BigBasket, renowned for grocery delivery, stocks the least number of grocery items among the analyzed categories. Further analysis will be conducted to identify the reasons behind this observation.

This comprehensive analysis provides a nuanced understanding of product distribution, enabling informed decision-making for both consumers and stakeholders in the online grocery industry.

Data Visualisations and Insights

Analysis of Best-Selling Category: Beauty & Hygiene

In our exploration of the best-selling category, Beauty & Hygiene, we delve into its subcategories, assigning the variable name 'sub_cat' for clarity. The primary aim is to scrutinize the prominence of subcategories, considering factors such as life cycle and sales trends.



Observations and Insights:

Extensive Stocking of "Skin Care":

"Skin care" products emerge as the most extensively stocked subcategory, suggesting their versatility, prolonged shelf life, and significant demand. The diverse range within this subcategory caters to various skin care needs.

Data Visualisations and Insights

Versatility and Demand in "Health and Medicine":

The "Health and Medicine" subcategory ranks as the second-highest in quantity, highlighting its versatility and substantial demand. Products in this subcategory likely cover a broad spectrum, including health supplements and medicinal items.

Toiletries Dominate with Similar Quantities:

Subsequent subcategories, namely "Hair Care," "Fragrances & Deos," "Bath & Hand Wash," and "Men's Grooming," exhibit similar quantities of products. These fall under the toiletries section, addressing essential daily needs and showcasing a competitive market.

Comparatively Lower Quantity in "Oral Care":

The "Oral Care" subcategory exhibits a comparatively lower quantity than the top categories. This could be attributed to limited competition in this sector, as customers may be less experimental with oral care products.

Low Quantity in "Makeup" Products:

The quantity of "Makeup" products is notably low, suggesting a possible preference among sellers for platforms other than Big Basket in this particular category.

Insight: *The company could enhance its presence in the "Makeup" category by implementing targeted advertising, introducing enticing offers for buyers, and attracting a greater number of sellers to its platform.*

Data Visualisations and Insights

Limited Versatility in "Mothers & Maternity":

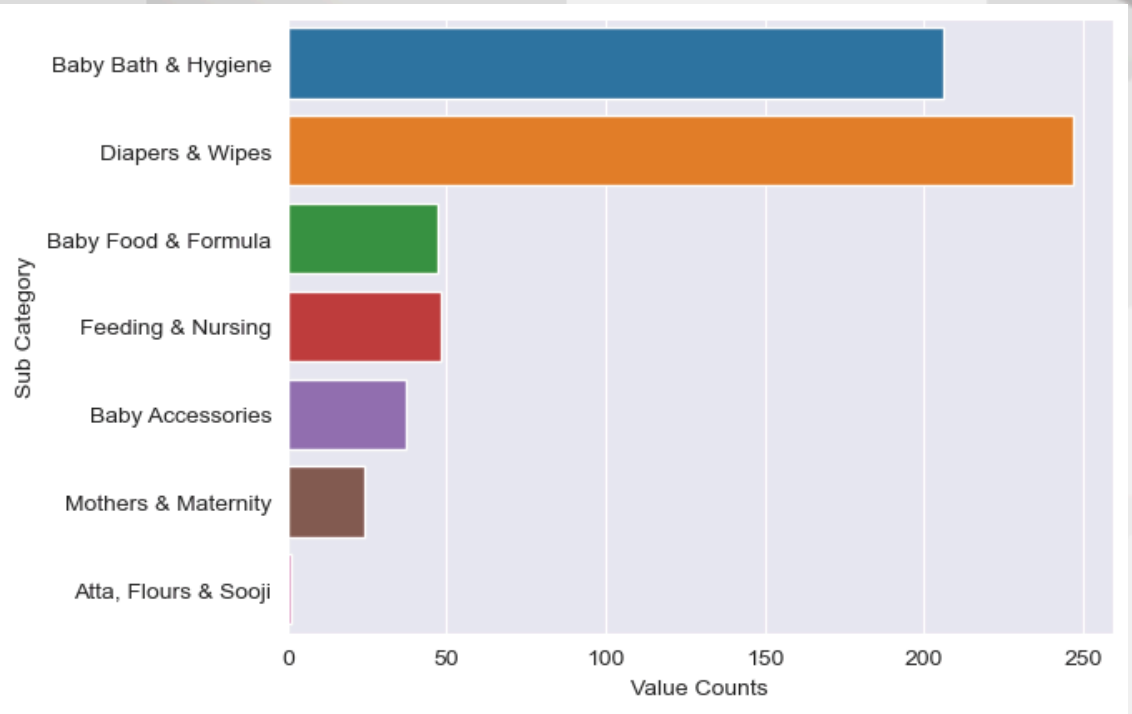
The "Mothers & Maternity" subcategory has the least quantity, indicating its limited versatility and application during a specific period. Products in this category likely cater to the needs of expectant mothers and those in the early stages of motherhood.

This detailed analysis provides valuable insights into the distribution of subcategories within the best-selling category, guiding strategic decisions for inventory management, marketing, and potential expansion into specific segments.

Data Visualisations and Insights

Analysis of the Third Least Stocked Category: Baby Care

In our examination of the third least stocked category, Baby Care, we aim to uncover insights into the lower quantity of baby products and the nuances within different subcategories.



Shorter Usage Period and Customer Sensitivity:

The lower quantity of baby products can be attributed to their shorter period of usage, as babies quickly outgrow certain items, and the heightened sensitivity of customers towards products for their infants.

Data Visualisations and Insights

Diapers and Wipes Dominate:

Subcategories like diapers and wipes exhibit higher quantities due to their daily use nature and increased market competition. These essential items are in constant demand, leading to a more extensive product range.
Extensive Range in Baby Bath and Hygiene:

Extensive Range in Baby Bath and Hygiene:

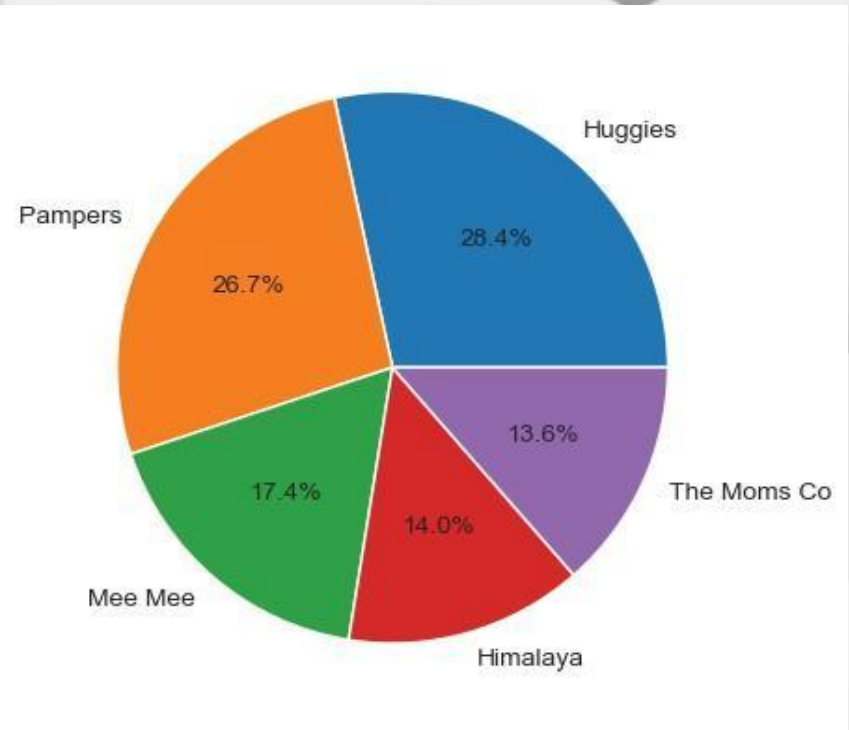
Baby bath and hygiene products show an extensive quantity of items, indicating their frequent usage and the availability of a wide range of products catering to different baby care needs.

Comparing "Mothers & Maternity" and "Beauty and Hygiene" Categories:

The "Mothers & Maternity" category comprises 24 items, while the "Beauty and Hygiene" category has 7 items. This discrepancy may suggest a greater specificity towards
baby-related products in the former and a focus on maternal hygiene in the latter.

Data Visualisations and Insights

Dominant Players in the Market:



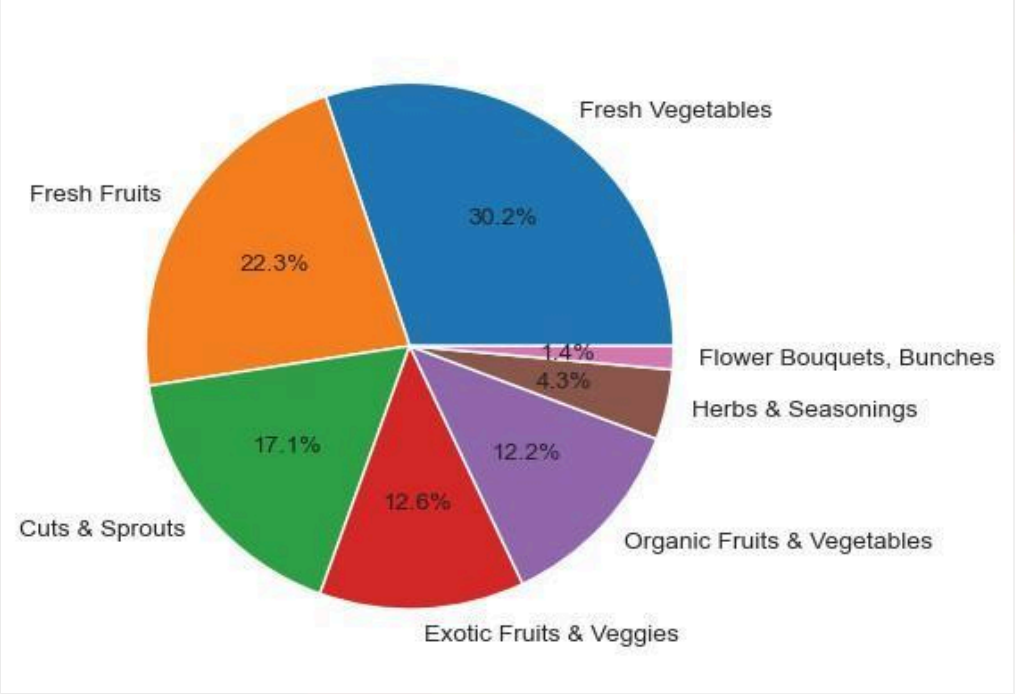
The baby care market is highly competitive, with brands like Huggies and Pampers as dominant players. Anticipated future performance suggests that brands like Mee Mee, Himalaya, and The Moms Co. may thrive in a more competitive market, emphasizing the importance of providing quality products.

This analysis provides valuable insights into the dynamics of the Baby Care category, shedding light on factors such as customer sensitivity, the competitive landscape, and the specific needs of both infants and mothers. These insights can guide marketing strategies, inventory management, and potential partnerships in the baby care segment.

Data Visualisations and Insights

Analysis of Food-Related Items: Fruits & Vegetables

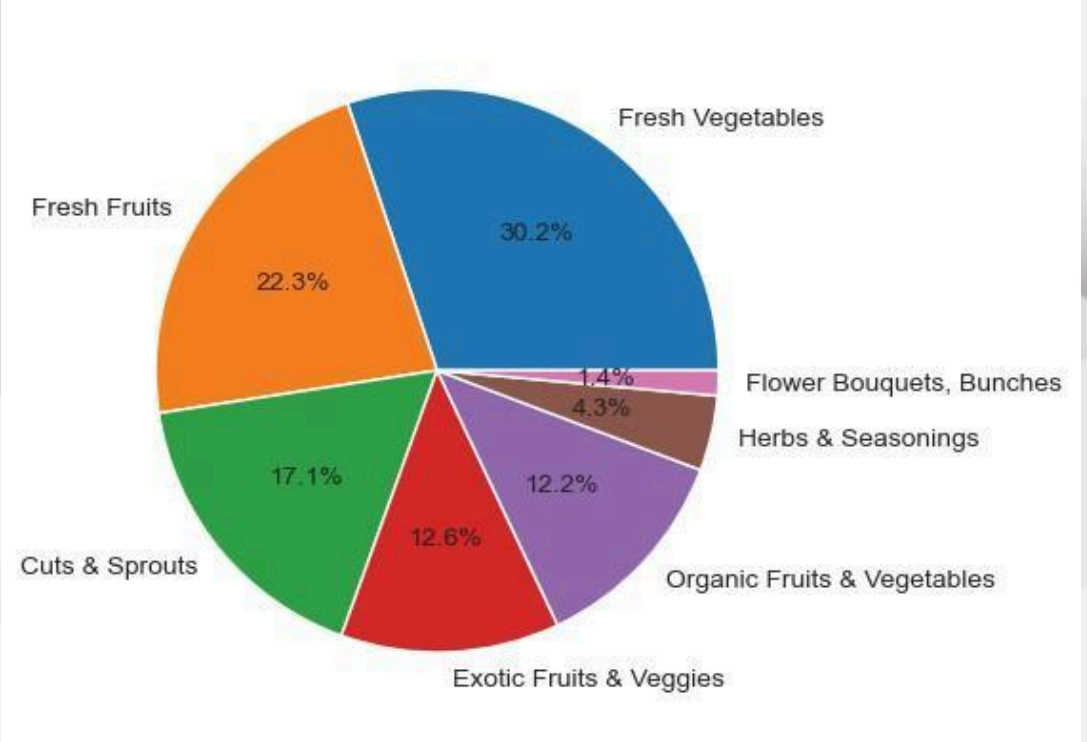
In our analysis of the food-related items listed on Big Basket, we focus specifically on the category of Fruits & Vegetables, which encompasses 557 items. The objective is to gain insights into the composition of this category and identify potential areas for improvement.



Dominance of Fresh Fruits and Vegetables:

Fresh fruits and vegetables constitute a significant share of approximately 65.1% within the category. This includes both common and exotic varieties, indicating a diverse offering to cater to varying customer preferences.

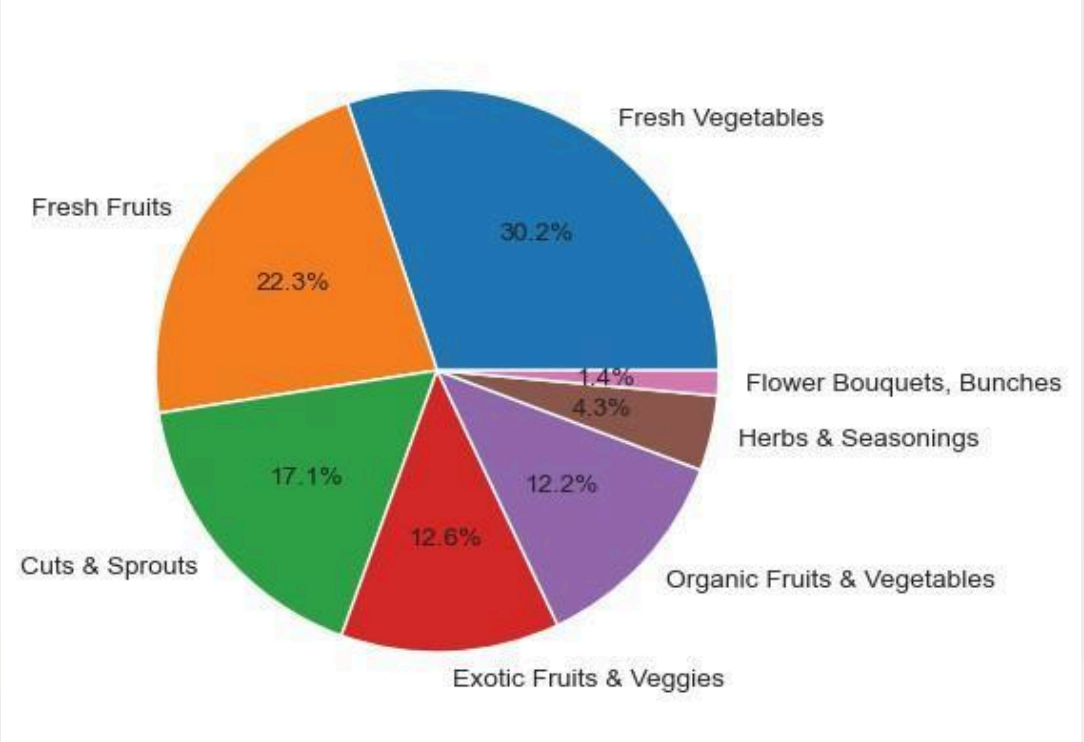
Data Visualisations and Insights



Limited Items in "Herbs and Seasonings" Subcategory:

The "Herbs and Seasonings" subcategory currently has a relatively limited list of items. This observation suggests that there is room for improvement in diversifying the range of herbs and seasonings available on the platform.
Insight: To enhance the variety and appeal of the "Herbs and Seasonings" subcategory, Big Basket could explore partnerships with additional suppliers, ensuring a broader selection for customers.

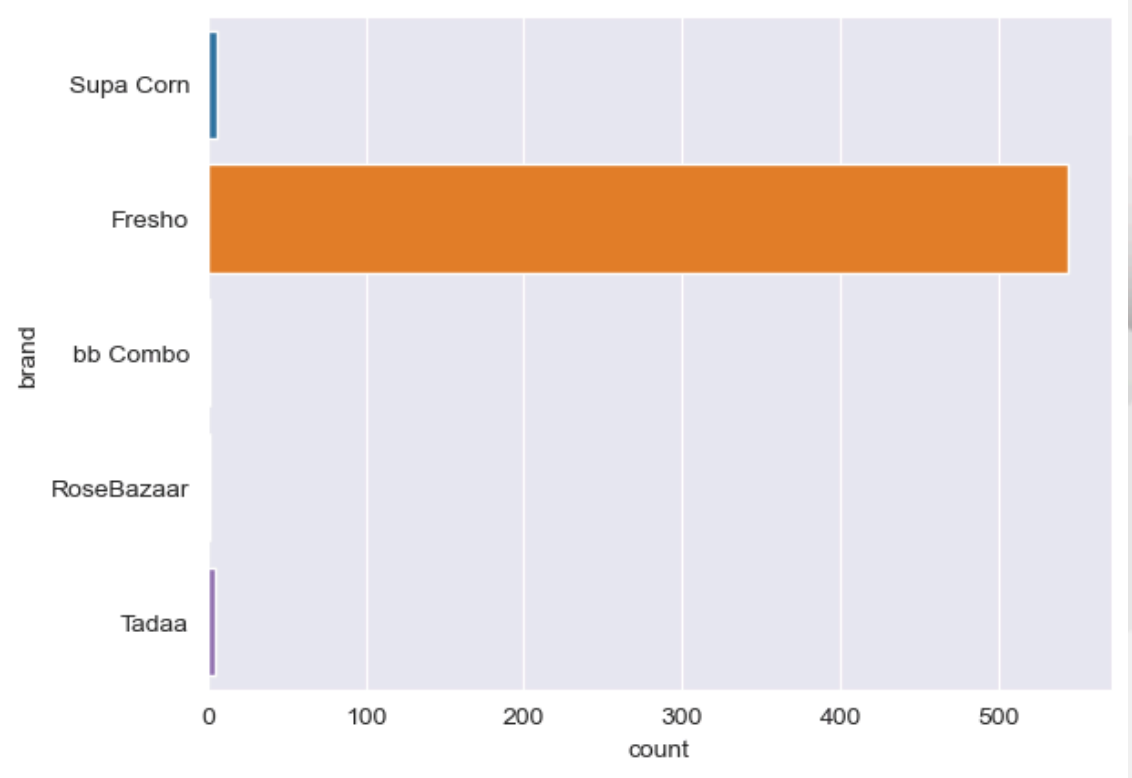
Data Visualisations and Insights



Limited Items in "Herbs and Seasonings" Subcategory:

The "Herbs and Seasonings" subcategory currently has a relatively limited list of items. This observation suggests that there is room for improvement in diversifying the range of herbs and seasonings available on the platform.
Insight: To enhance the variety and appeal of the "Herbs and Seasonings" subcategory, Big Basket could explore partnerships with additional suppliers, ensuring a broader selection for customers.

Data Visualisations and Insights

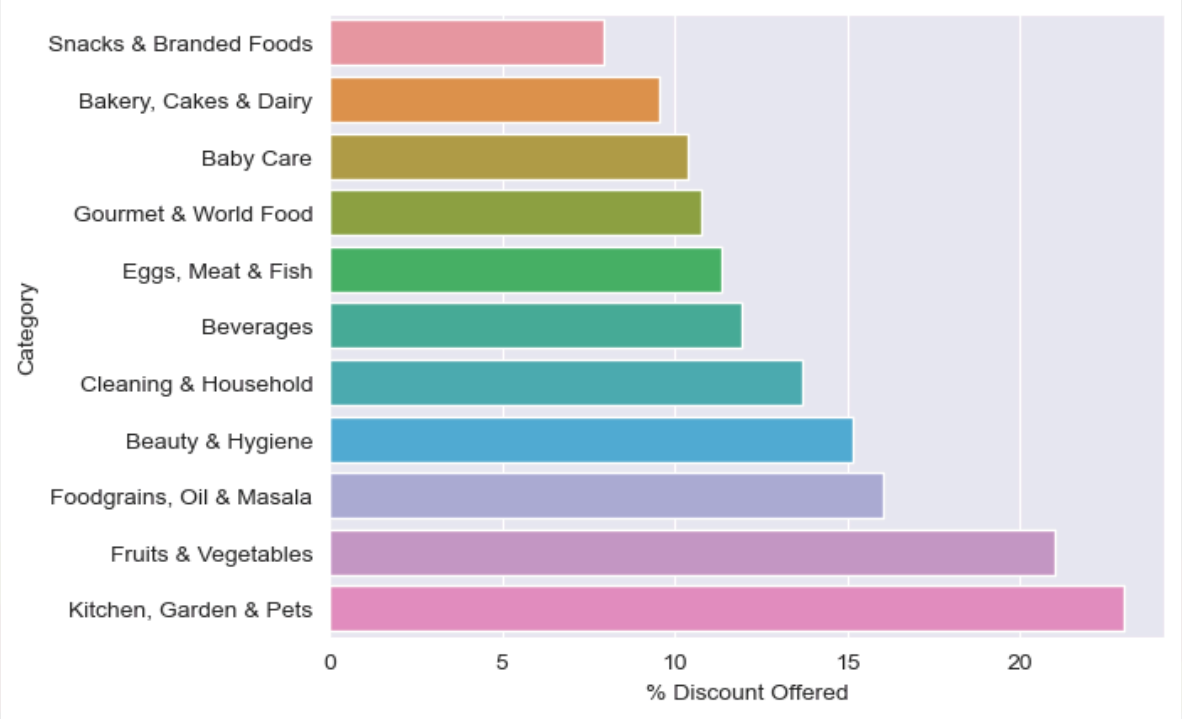


Brand Dominance:

The brand "Fresho" currently holds a dominant position in this category, with minimal competition. This could be an opportunity for Big Basket to enhance competitiveness by inviting more brands to its platform. Encouraging competitive pricing and ensuring the delivery of high-quality goods can further enrich the customer experience.

Insight: Strategic measures to attract diverse brands could contribute to a more vibrant marketplace, offering customers a wider array of choices and fostering healthy competition among suppliers.

Data Visualisations and Insights



Observations and Insights:

Highest Discount in "Kitchen, Garden & Pets":

The category "Kitchen, Garden & Pets" stands out with the highest discount of 23.03%. This could be attributed to various factors such as high profit margins, clearance sales, or seasonal discounts aimed at boosting sales and attracting customers in this specific segment.

Least Discount in "Snacks & Branded Foods":

Conversely, the category "Snacks & Branded Foods" has the least discount at 7.94%. This may be indicative of thinner profit margins in this competitive market, where retailers may not provide as substantial discounts compared to other categories.

Data Visualisations and Insights

Higher Discount in "Fruits & Vegetables":

The category "Fruits & Vegetables" exhibits a higher discount rate of 21.02%. This could be influenced by factors such as the perishable nature of these items, seasonal fluctuations, and the need to manage inventory with a shorter shelf life.

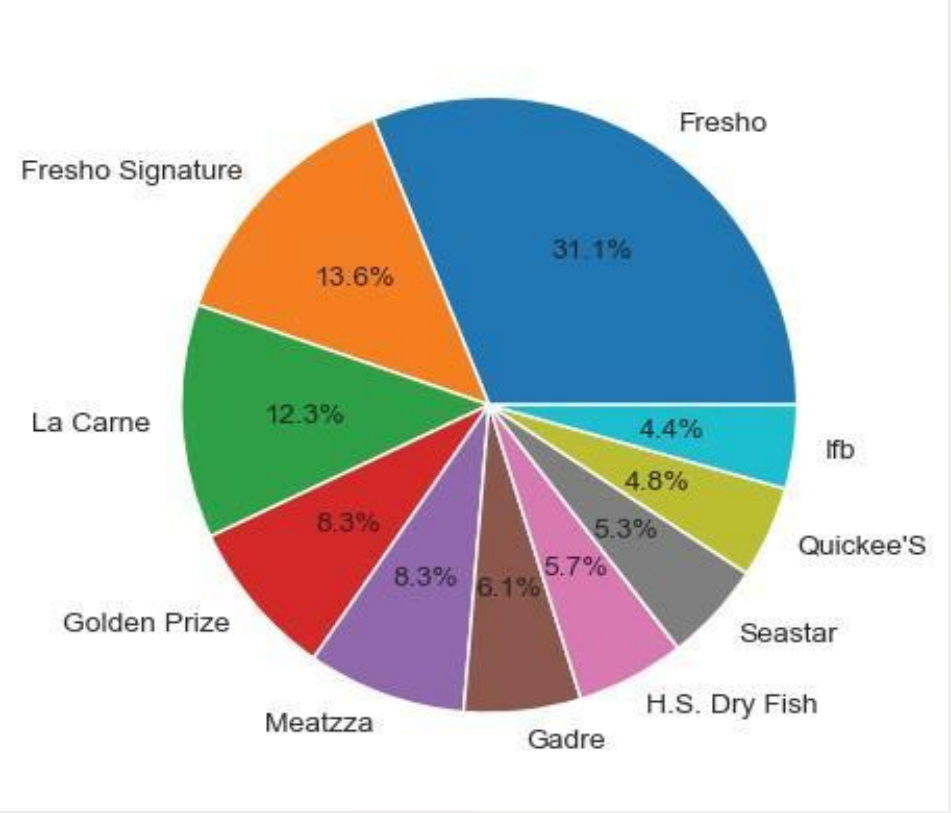
Insight: Understanding the varying discount rates across categories provides insights into pricing strategies, market dynamics, and customer preferences. For categories with higher discounts, exploring the underlying reasons for these rates can guide future marketing and sales strategies.

This analysis contributes to a better understanding of discounting practices across different product categories on BigBasket, enabling strategic considerations for pricing and promotions.

Data Visualisations and Insights

Analysis of Eggs, Meat & Fish Category: Market Share of Brands

In our examination of the Eggs, Meat & Fish category, the focus is on analyzing the market share of brands, with a specific consideration of the top 10 brands for convenience and clarity.



Observations and Insights:

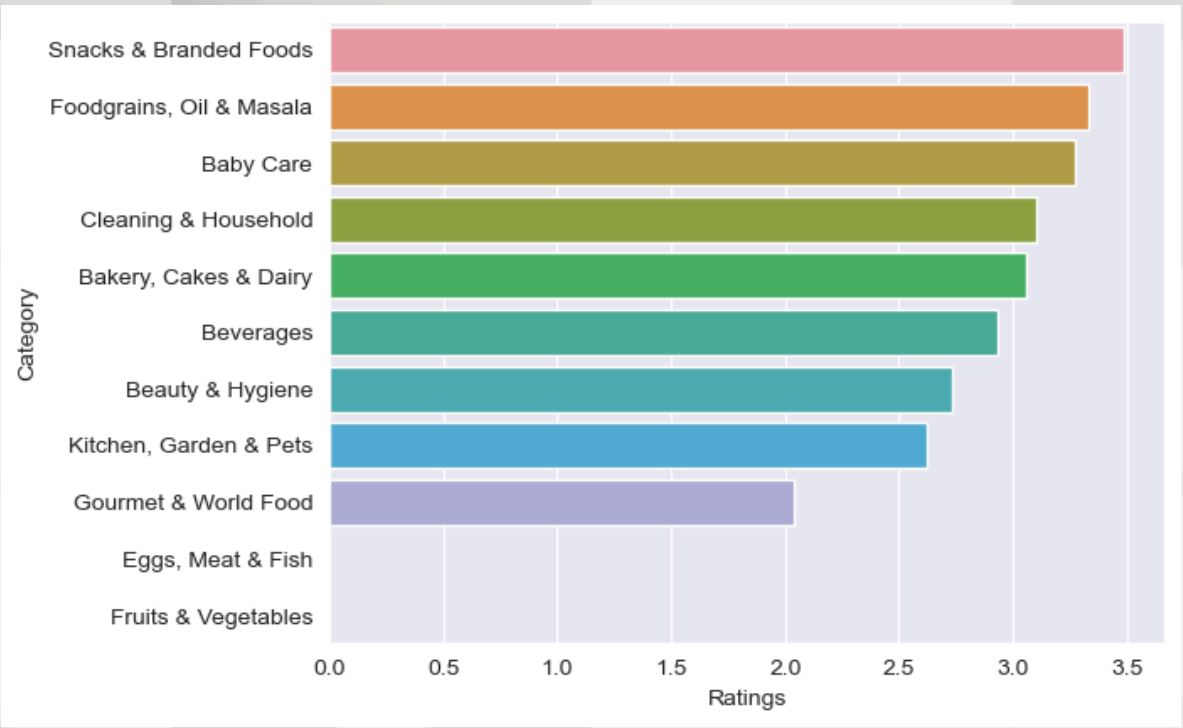
Dominance of Fresho:

Within the egg, meat, and fish category, Fresho emerges as the dominant brand, holding a significant market share. This establishes Fresho as a major player and a preferred choice among customers.

Data Visualisations and Insights

Analysis Based on Ratings

In our analysis based on ratings, we aim to derive insights into customer perceptions and satisfaction across different categories on BigBasket.



Observation and Insights:

Inconsistencies in Ratings for Certain Categories:

Noticing inconsistencies in the "Eggs, Meat & Fish" and "Fruits & Vegetables" categories with the absence of any ratings, it is currently challenging to identify the causes or assess opportunities for improvement. However, addressing this issue remains a potential area for enhancement in the future.

Data Visualisations and Insights

High Ratings in "Snacks and Branded Foods":

The highest ratings for the "Snacks and Branded Foods" category suggest strong customer appeal, high product quality, and the frequent consumption of these items. Positive experiences, coupled with consistent quality and a wide variety of options, contribute to the elevated ratings in this category.

Low Ratings in "Gourmet & World Food":

The low ratings for "Gourmet & World Food" could be attributed to factors such as limited customer familiarity, niche appeal, and potentially higher price points. This category may be less popular among a broader audience, resulting in lower overall customer satisfaction.

Insight: Addressing the challenges in rating consistency for certain categories, and potentially exploring ways to improve customer awareness and engagement in the "Gourmet & World Food" category, could lead to enhanced customer satisfaction and ratings.

This analysis provides a glimpse into customer sentiment across different categories on BigBasket. Understanding the factors influencing ratings allows for strategic considerations to enhance customer experiences and satisfaction in specific product segments.

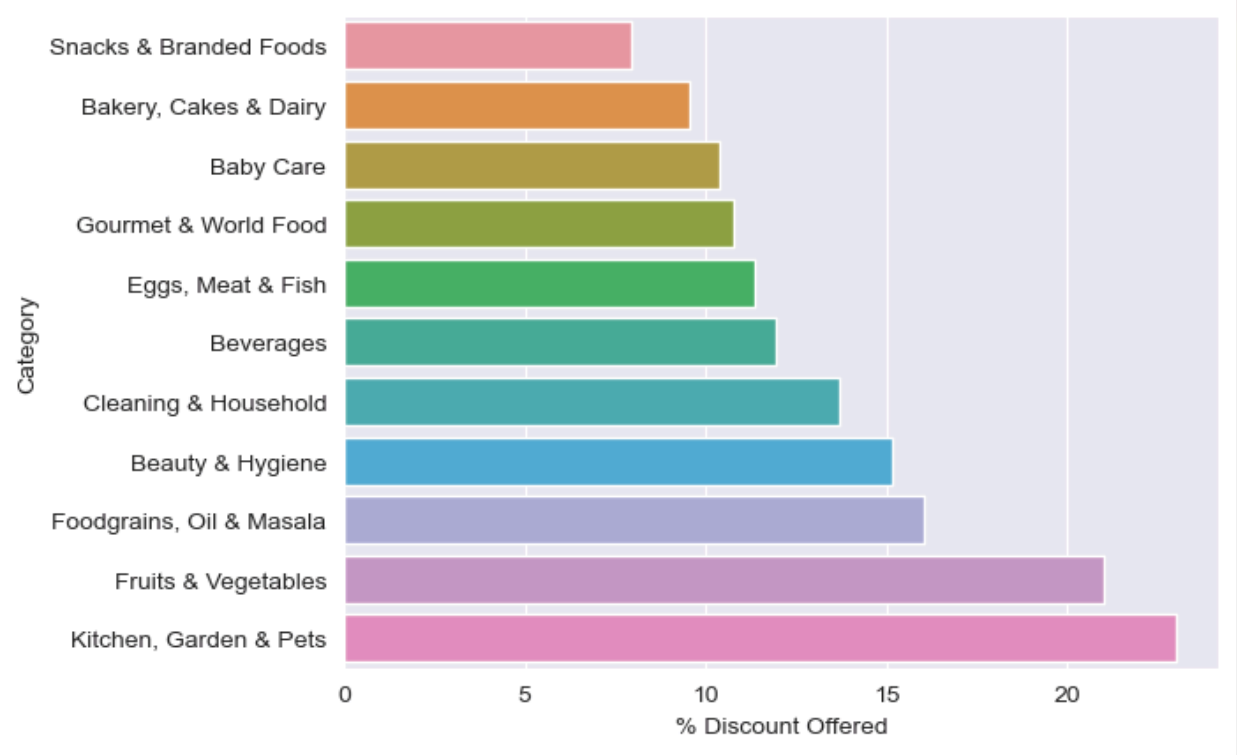
Data Visualisations and Insights

Reviewing Items Based on Discounts Offered

In the analysis of items based on the discounts offered, the focus is on understanding the variations in discount rates across different categories on BigBasket.

Methodology:

Assigning the Maximum Retail Price (MRP) to the variable "MRP" for ease of reference.
Calculating discounts for each category using the Maximum Retail Price (MRP) and the Selling Price (SRP).



Final Report: Comprehensive Analysis of BigBasket Dataset

Overview:

This report presents a comprehensive analysis of the BigBasket dataset, covering various aspects such as product categories, subcategories, brand performance, ratings, and discounts. The insights derived aim to inform strategic decision-making, identify opportunities for improvement, and enhance the overall customer experience on the platform.

1. Exploratory Data Analysis (EDA):

1.1 Overview of BigBasket:

- Background: BigBasket, a leading online grocery platform, has been analyzed to gain insights into its product offerings, market presence, and customer preferences.
- Ownership: Recent developments indicate that Tata Digital, a subsidiary of the Tata Group, has acquired BigBasket, marking Tata's entry into the online grocery and e-commerce sector.

1.2 Dataset Description:

- Dataset Content: The dataset includes information about items listed on BigBasket, featuring columns such as product name, category, subcategory, brand, sale price, market price, type, rating, and description.
- Size: The dataset comprises 27,555 rows and 10 columns.

Final Report

2. Data Preprocessing:

2.1 Dealing with Null Values:

- Null values were addressed in various columns by employing strategies such as imputation, substitution, and assigning placeholder values.

2.2 Outlier Detection:

- Outliers, especially in pricing, were identified and addressed by substituting sale prices with corresponding market prices where inaccuracies were observed.

2.3 Feature Selection:

- The index column was removed from the DataFrame to streamline the data and improve efficiency.

3. Category and Subcategory Analysis:

3.1 Top-Selling Categories:

- The "Beauty & Hygiene" category emerged as the top-selling, followed by "Gourmet & World Food" and "Kitchen, Garden, and Pet Goods."
- Insights into each category provided valuable information for decision-making.

3.3 Baby Care Analysis:

- Analysis of the Baby Care category highlighted the niche nature of this segment, with specific insights into different subcategories.

3.4 Fruits & Vegetables Analysis:

- Examination of the Fruits & Vegetables category focused on the diversity of products, offering insights for inventory management.

4. Ratings Analysis:

- Ratings analysis revealed inconsistencies in certain categories, and insights were provided into potential areas for improvement.

5. Discounts Analysis:

- The analysis of discounts across categories shed light on pricing strategies, market dynamics, and customer preferences.

6. Recommendations:

- **Invite Diverse Brands: Fostering competition by inviting new suppliers, especially in categories dominated by a single brand, can enhance product quality and variety.**

- **Enhance "Gourmet & World Food" Awareness: To address low ratings in this category, efforts to improve customer awareness and engagement could lead to increased satisfaction.**

- **Strategic Pricing in Competitive Categories:** Categories with thinner profit margins, such as "Snacks & Branded Foods," may benefit from strategic pricing strategies to attract customers.

- **Address Null Ratings:** Exploring ways to address null ratings in certain categories could contribute to more comprehensive customer feedback.

- **Diversify Discounts:** Considering the reasons behind varying discount rates can guide strategies to diversify discounts and meet customer expectations.

Conclusion:

This comprehensive analysis provides a detailed understanding of the BigBasket dataset, offering actionable insights for improving product offerings, customer satisfaction, and overall business performance. The recommendations provided aim to guide strategic decision-making for a more competitive and customer-centric online grocery platform.

Summary

Summary:

The analysis of the BigBasket dataset has provided valuable insights into various facets of the online grocery platform. Here are key highlights and takeaways:

1. Ownership and Dataset Overview:

- BigBasket, a leading online grocery platform, is now owned by Tata Digital, marking a significant entry for the Tata Group into the e-commerce sector.
- The dataset comprises 27,555 rows and 10 columns, offering a comprehensive view of listed items on BigBasket.

2. Data Preprocessing:

- Dealt with null values through imputation, substitution, and placeholder values.
- Detected and addressed outliers, particularly in pricing, ensuring data accuracy.
- Streamlined the dataset by removing the index column through feature selection.

3. Category and Subcategory Analysis:

- Identified top-selling categories, with "Beauty & Hygiene" leading the pack.
- Analyzed subcategories, providing insights into market share, product dominance, and potential areas for improvement.
- Explored specific categories like Baby Care, Fruits & Vegetables, and identified niche markets and inventory management strategies.

Summary

4. Ratings and Discounts Analysis:

- Observed inconsistencies in certain categories' ratings, suggesting potential areas for improvement.
- Analyzed discounts across categories, revealing insights into pricing strategies, market dynamics, and customer preferences.

5. Recommendations:

- Advocated for inviting diverse brands to enhance competition and product quality.
- Suggested strategies to improve awareness and engagement in specific categories with low ratings.
- Highlighted the importance of strategic pricing in competitive categories and addressing null ratings.

6. Conclusion:

- The final report provided a comprehensive view of BigBasket, offering actionable recommendations for strategic decision-making.

- The analysis sets the groundwork for enhancing customer experiences, improving product offerings, and maintaining competitiveness in the online grocery market.

This summary encapsulates the key findings and recommendations, providing stakeholders with a roadmap for future improvements and strategic initiatives on the BigBasket platform.

Thanks for Reading!

For Code and EDA files

https://github.com/Chiragkukreja73/Big-Basket-Data-Analytics-Project