Big Basket

Data Analysis Report

2025

bigbasket

India's largest online supermarket

Prepared by

APOORVA SHARMA

Data Analysis File Link:

Big Basket Mini Project

Table of Content

OVERVIEW OF COMPANY

AIM AND OBJECTIVES

TOOLS AND FORMULAS

STEPS TO FOLLOW

GRAPHS AND INSIGHTS

CONCLUSION

ACKNOWLEDGEMENT

OVERVIEW OF COMPANY

BigBasket is one of India's leading online grocery delivery platforms, founded in 2011 in Bengaluru by a group of experienced entrepreneurs including Hari Menon and V.S. Sudhakar. The company offers a wide range of products — from fresh fruits and vegetables to household essentials and personal care items — and operates in over 30 cities across India. With a focus on convenience, BigBasket delivers millions of orders each month through various services like BB Now, BB Express, and BB Instant, catering to the growing demand for quick-commerce in urban areas.

In 2021, BigBasket became a part of the Tata Group after Tata Digital acquired a majority stake, strengthening its market position and brand credibility. Over the years, the company has attracted major investors like Alibaba and has grown into a household name in India's e-grocery space. With its focus on technology, supply chain efficiency, and customer satisfaction, BigBasket continues to expand and innovate in the fast-evolving online retail sector.

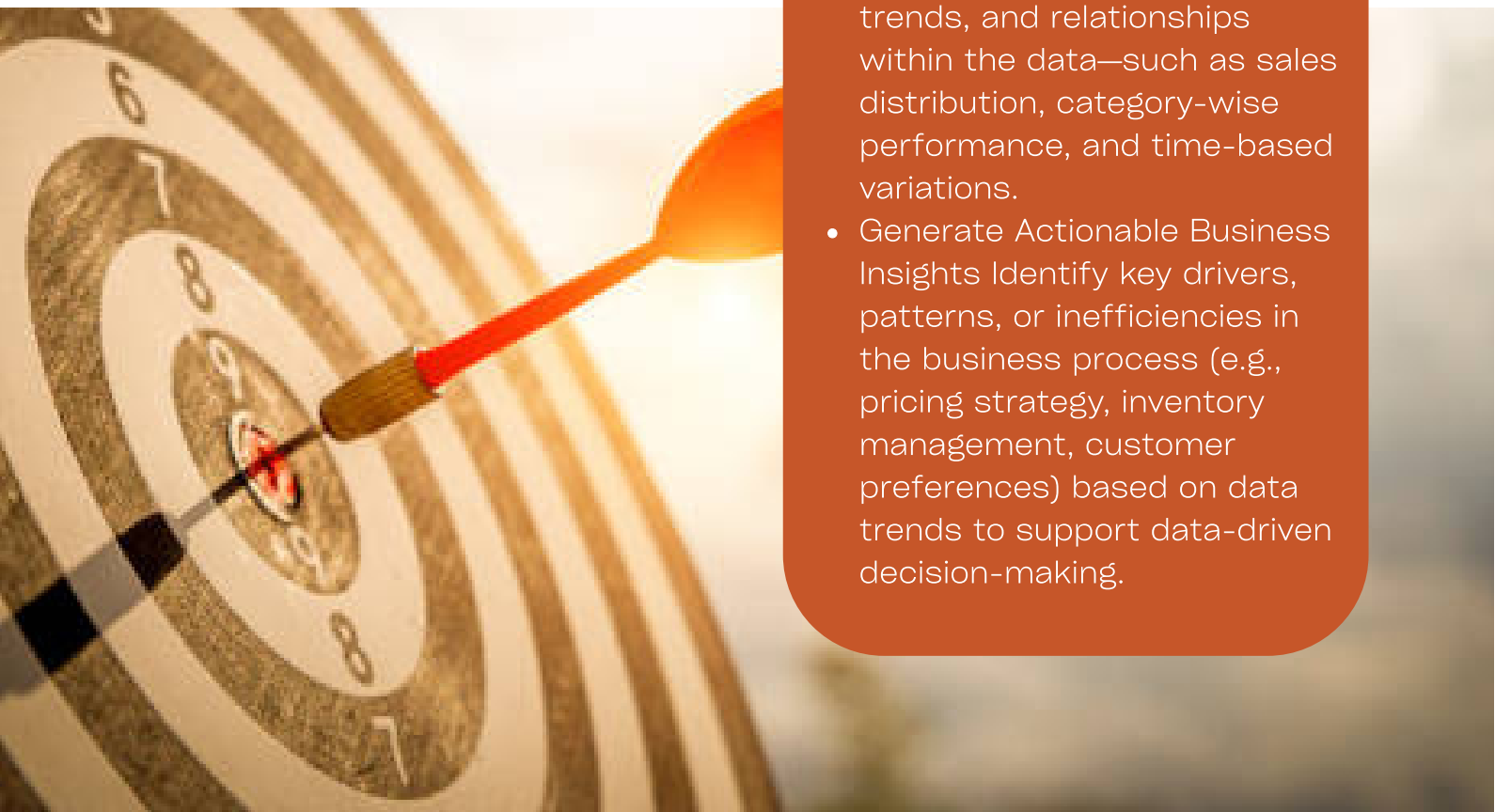
AIM AND OBJECTIVE

AIM

To analyze and interpret meaningful patterns from the given dataset to draw insights, support decision-making, and enhance understanding of the underlying trends

OBJECTIVE

- Perform Data Cleaning and Preprocessing Ensure data quality by handling missing values, removing duplicates, correcting inconsistencies, and standardizing formats to prepare the dataset for accurate analysis.
- Conduct Exploratory Data Analysis (EDA)
- Use visualizations and summary statistics to understand the structure, trends, and relationships within the data—such as sales distribution, category-wise performance, and time-based variations.
- Generate Actionable Business Insights Identify key drivers, patterns, or inefficiencies in the business process (e.g., pricing strategy, inventory management, customer preferences) based on data trends to support data-driven decision-making.



TOOLS AND GRAPHS

TOOLS USED

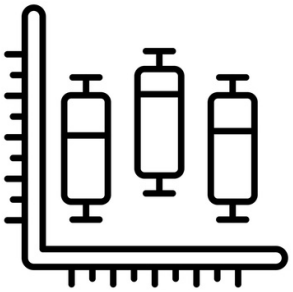
NumPy

pandas

Matplotlib

seaborn

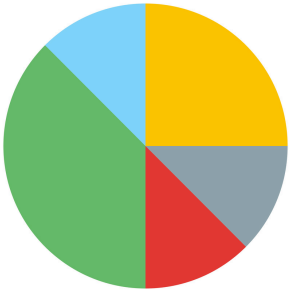
GRAPHS USED



BOXPLOT

BAR GRAPH

H BAR GRAPH



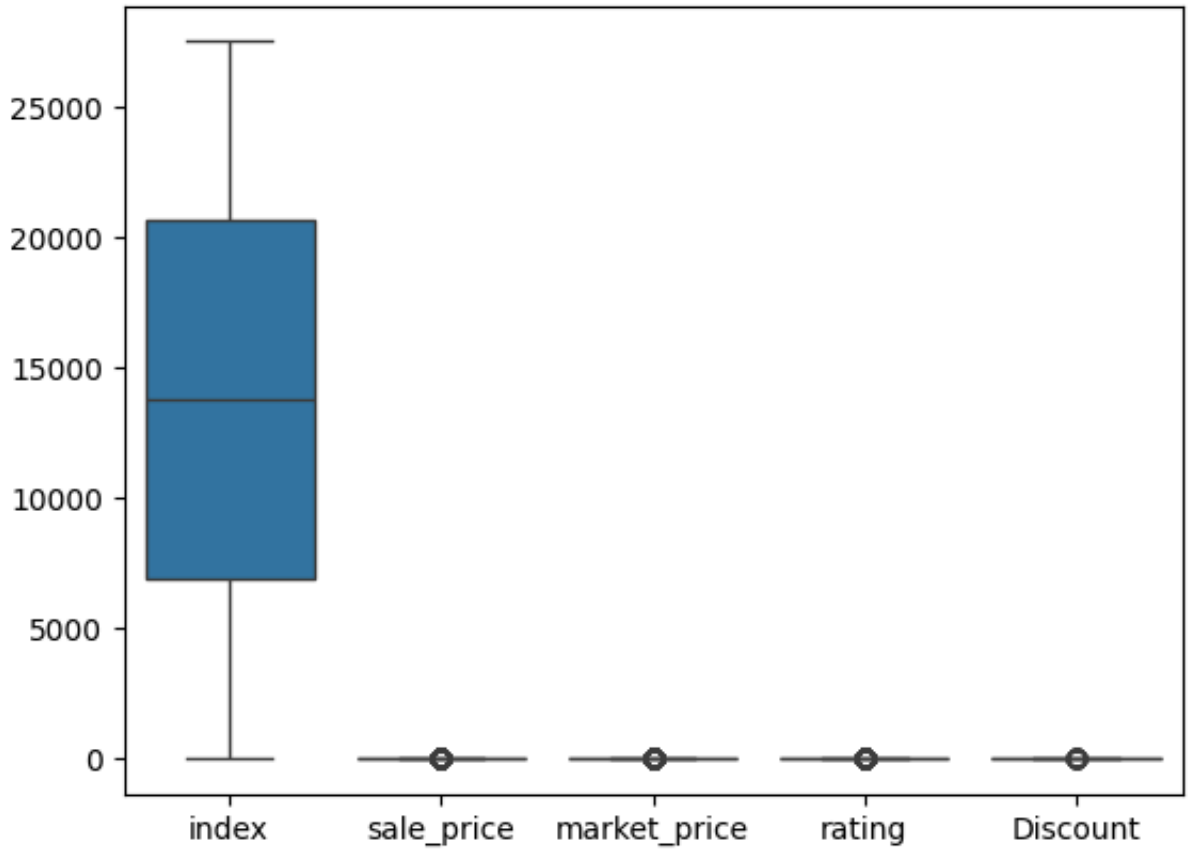
PIE CHART

SUNBURST CHART

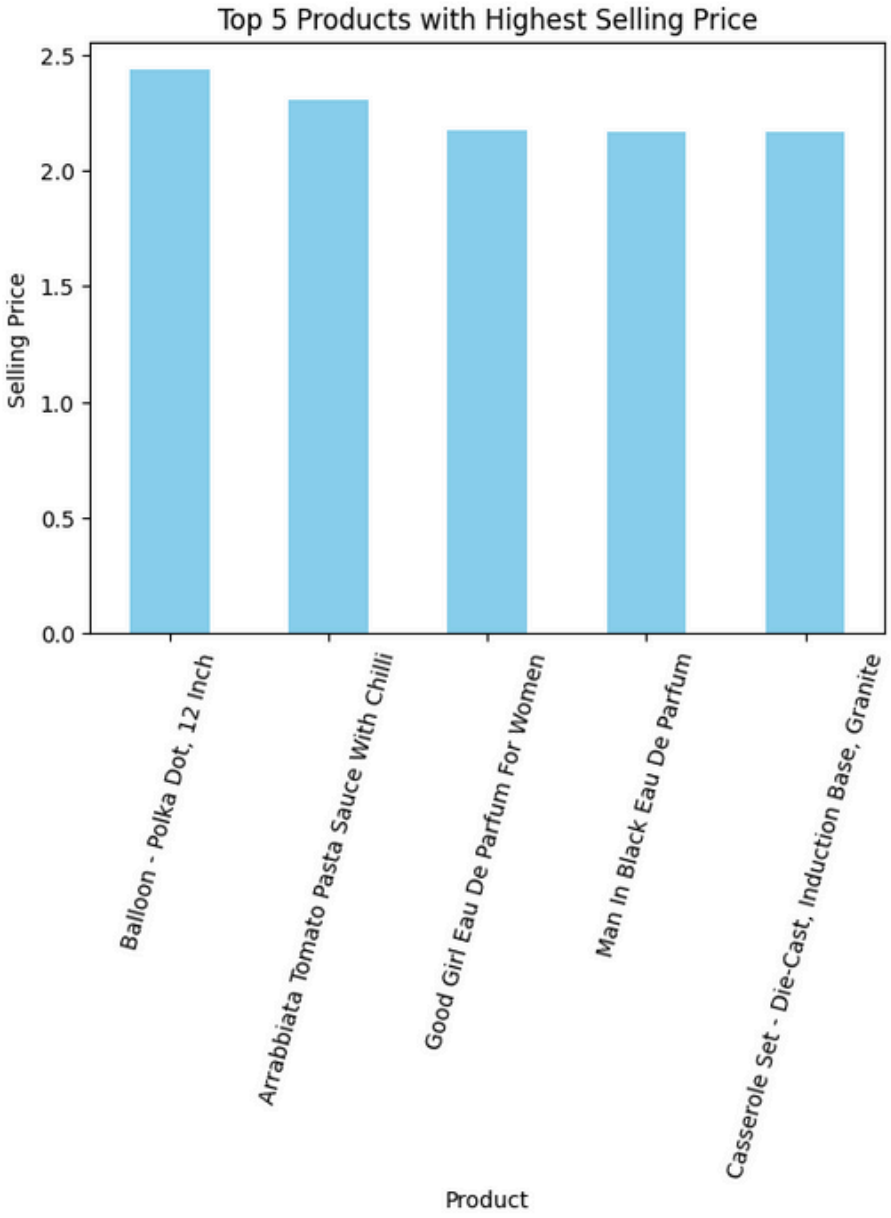
STEPS TO FOLLOW

- Step 1: Load Dataset.
- Step 2: Use head function to look for first 12 rows.
- Step 3: Get Description of the data in the Data Frame.
- Step 4: Find Information about the Data Frame.
- Step 5: Find out Top & least sold products.
- Step 6: Measuring discount on a certain item.
- Step 7: Find out the Missing Values from the Dataset.
- Step 8: Find out the outliers from the dataset according to the columns and fill them with the mean.
- Step 9: Create Plots or visualizations.

GRAPHS AND INSIGHTS



This BOXPLOT shows that the index column has a wide range of values (right-skewed), but it's just a positional reference and not useful for analysis.
All actual numeric features like sale_price, market_price, rating, and Discount are well within close ranges after cleaning — this means:
There are no major outliers left.
Data is consistent and more reliable for further analysis.
The values of sale_price and market_price appear close to each other, suggesting discounts are not extreme.
Rating and Discount are fairly normalized, indicating balanced user feedback and offers.
Now the data is clean, scaled, and ready for exploratory analysis or visual dashboards.

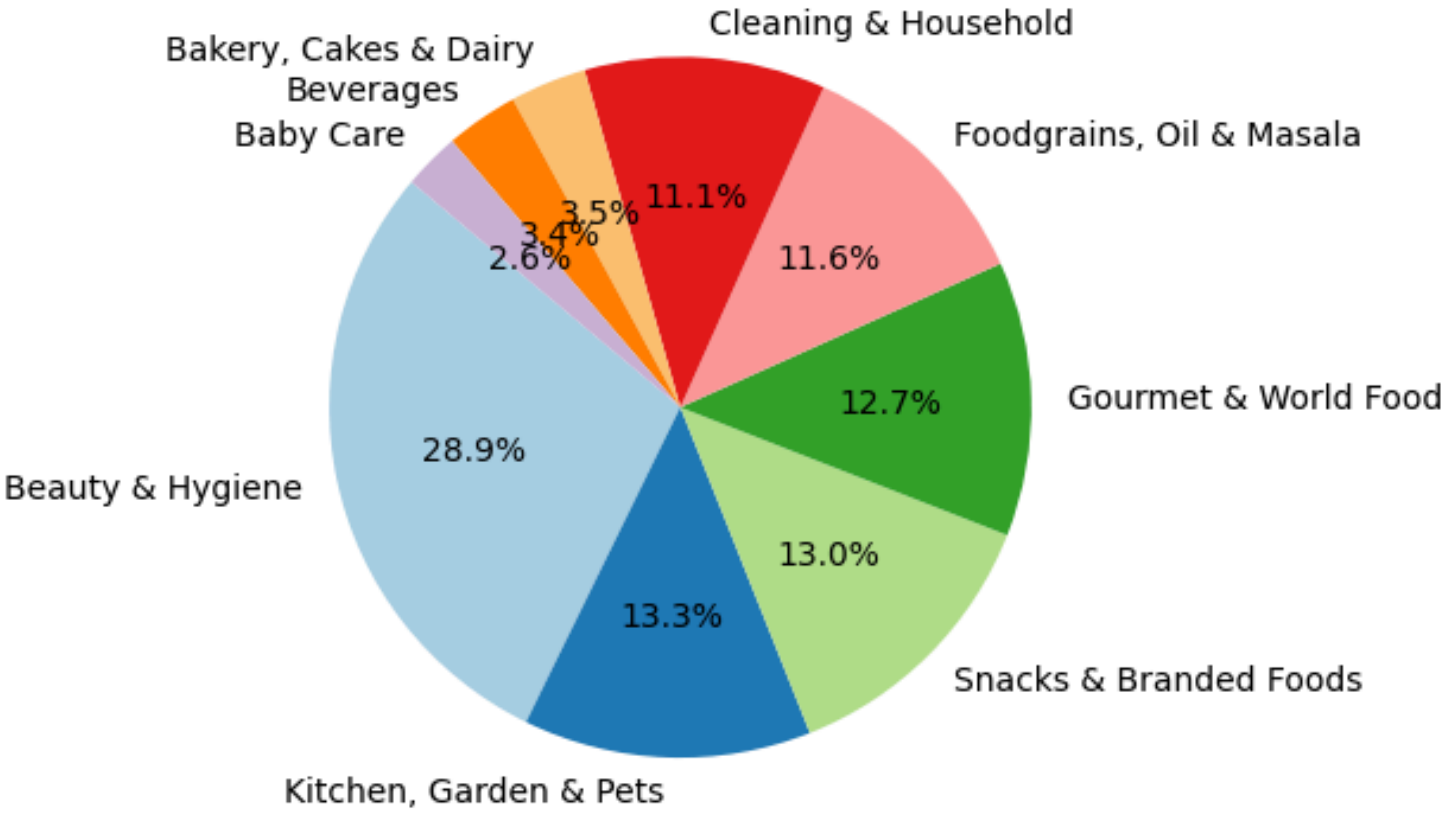


On the x-axis of this bar graph, we see the names of the products, which include a mix of decorative items, gourmet foods, perfumes, and kitchenware. The y-axis represents the selling price, and we observe that all top products have prices between 2.1 and 2.5 units, indicating a relatively narrow but high-end pricing range.

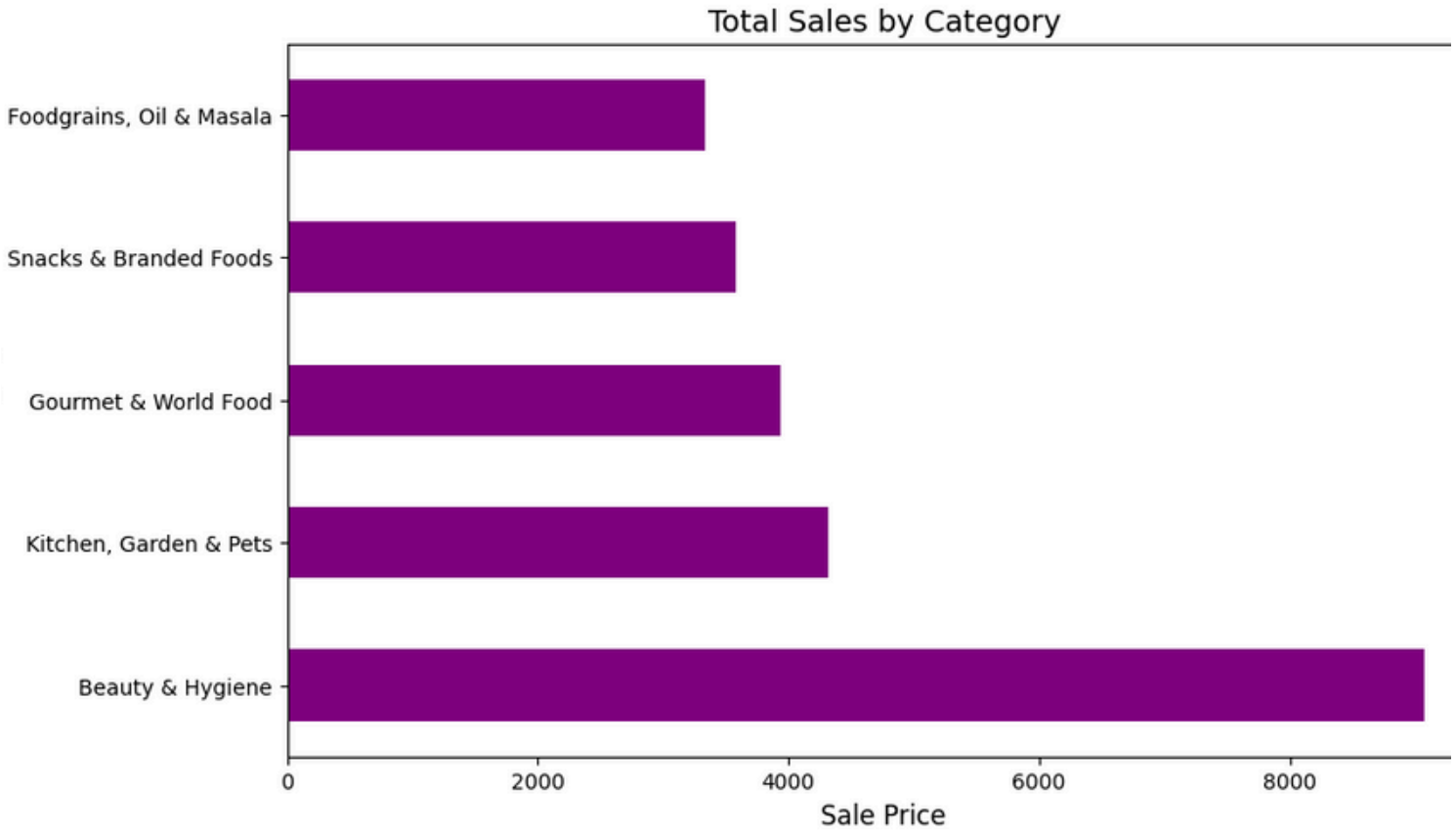
The product with the highest selling price is "Balloon - Polka Dot, 12 Inch", which might seem unexpected, but this could be due to bulk packaging, a luxury variant, or a data anomaly worth verifying.

Following that, we see "Arrabiata Tomato Pasta Sauce With Chilli" and two premium perfumes — "Good Girl Eau De Parfum For Women" and "Man In Black Eau De Parfum" — highlighting that personal care and gourmet food items are priced at the upper end of the dataset.

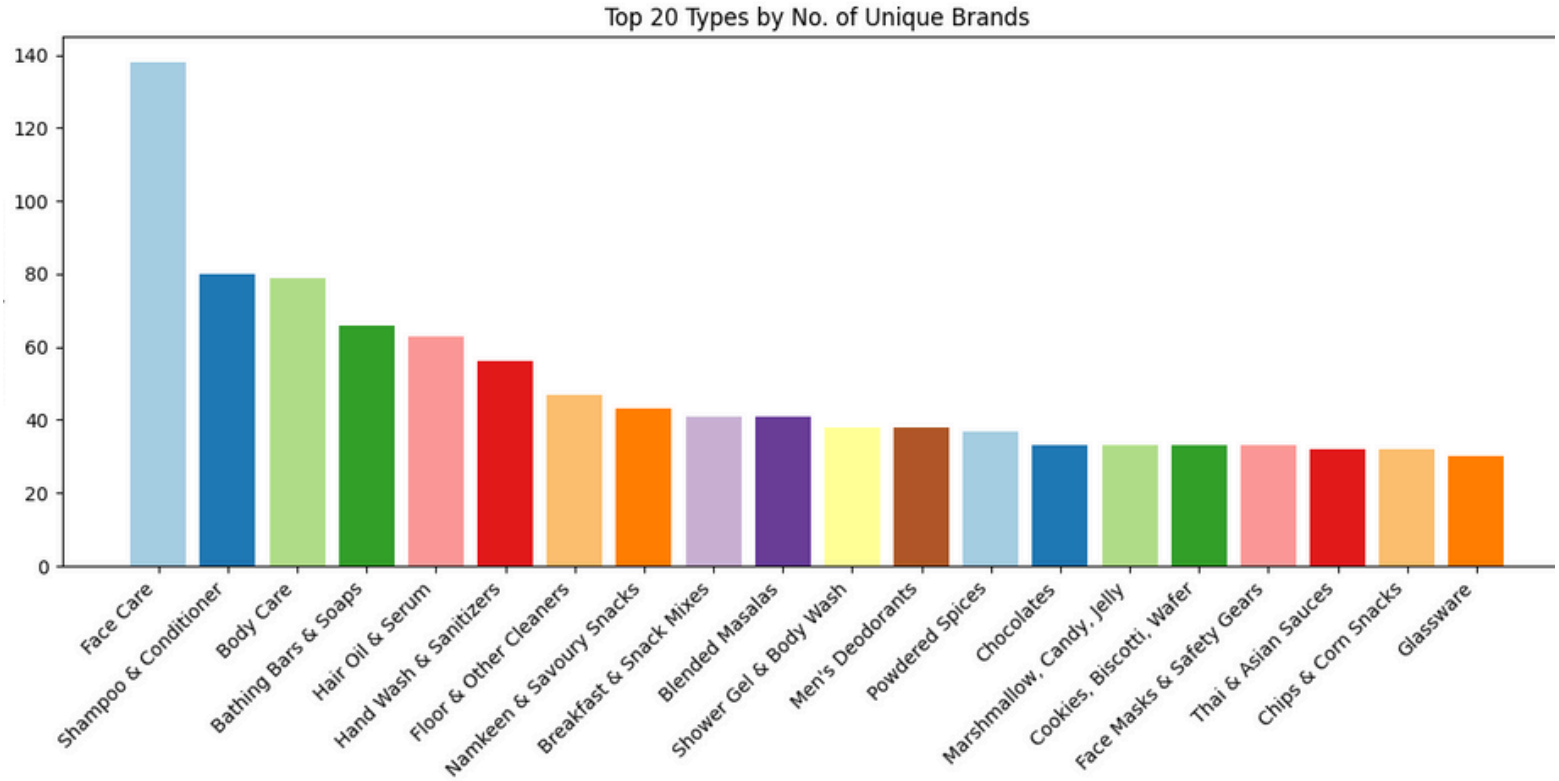
Lastly, these products are likely low in quantity but high in price, meaning they are not fast-moving, but profitable per unit sold



This pie chart represents the percentage-wise distribution of brands associated with various product categories.
Beauty & Hygiene (28.9%) leads the chart, showing highest number of brands.
Kitchen, Snacks, and Gourmet categories each contribute around 13%, indicating strong competition in the market.
Food grains & Cleaning also hold steady with over 11% share each.
Bakery, Beverages, and Baby Care have lower shares (under 4%), suggesting less number of brands in these categories suggesting the low competition or need of products in market.



This bar chart displays the total sales (in terms of sale price) for five major product categories.
Beauty & Hygiene leads by a huge margin, with sales close to ₹9,000. This shows that customers are highly inclined towards buying personal care products—likely due to regular usage and increasing awareness about grooming and hygiene.
Kitchen, Garden & Pets is the second highest, with total sales around ₹4,300. This indicates consistent household needs and purchases in this category.
Gourmet & World Food, Snacks & Branded Foods, and Food grains, Oil & Masala all hover around the ₹3,500–₹3,800 mark. These categories are relatively balanced in performance, suggesting stable but not dominant customer demand.
Personal care is the most purchased category.
Kitchen essentials and packaged food have steady but comparatively lower sales.
This pattern reflects consumer preference for hygiene and everyday essentials over luxury or niche food items.



This chart shows the top 20 product types ranked by the number of unique brands available for each.
Face Care clearly dominates with around 138 unique brands, showing a highly competitive and diverse market. This suggests strong consumer interest and brand variety in skincare.
Shampoo & Conditioner, Body Care, and Bathing Bars & Soaps also have a high brand presence, all above 75. These categories are essential in personal care, explaining their wide brand diversity.
Mid-tier segments like Hair Oil & Serum, Hand Wash & Sanitizers, and Cleaners have 50–65 brands—still fairly competitive, likely due to daily utility.
Food-related items such as Namkeen, Masalas, Chocolates, and Snacks show up in the second half, indicating growing but more limited variety compared to personal care.
At the bottom, Glassware, Chips, and Face Masks & Safety Gears have under 35 brands, suggesting either a niche market or fewer players.
There's more brand competition in personal care than in food or utility items. This shows customers have a wider range of choices in grooming products, while food/snack segments have more focused brand offerings.

This sunburst chart visualizes the hierarchical distribution of product categories and their subcategories.
At the center, Beauty & Hygiene is the most dominant category, and its largest sub-segment is clearly Skin Care, followed by Hair Care, Bath & Hand Wash, and others. This shows high product diversity and likely strong consumer demand in the personal care space.
Kitchen, Garden & Pets and Snacks & Branded Foods also occupy major slices. Subcategories like Ready to Cook & Eat, Snacks & Namkeen, and Biscuits & Cookies are well represented—indicating a wide product variety in daily consumables.
Gourmet & World Foods and Foodgrains, Oil & Masala also show strong subcategory spread—especially in Sauces, Masalas & Spices, and Organic Staples.
On the other hand, categories like Baby Care, Beverages, and Bakery appear smaller and more focused, with fewer sub-segments like Tea, Dairy, and Diapers & Wipes—suggesting these are either niche markets or have more consolidated offerings.

CONCLUSION

The project involved detailed analysis of Big Basket's product dataset, focusing on product pricing, brand diversity, ratings, discounts, and category-level trends.
Data cleaning was performed efficiently by filling missing values:
Numerical columns like sale_price and rating were filled using median and 0 respectively.
Categorical columns like brand, product, and description were filled using the mode.
A custom anomaly detection function was used to identify and count outliers in key numerical columns, such as sale_price, market_price, rating, and discount.
Skewed distributions in sale_price, market_price, and discount were corrected using log transformation, leading to improved visualization and interpretation.
Ratings outside the valid range (1 to 5) were removed to maintain consistency and reliability of the data.
KDE plots, box plots, and histograms were used to understand the distribution of key variables and the presence of outliers.
Beauty & Hygiene emerged as the top category in both product count and total sales, indicating strong consumer preference for personal care items.
The pie chart illustrated the share of each product category, highlighting Beauty & Hygiene, Kitchen & Garden, and Snacks & Branded Foods as the leading segments.
A bar chart displayed the top 5 highest-priced products, offering insights into high-value offerings.
A horizontal bar graph showed that Beauty & Hygiene also led in total sales revenue, followed by Kitchen and Gourmet categories.
Analysis of brand variety across product types revealed that:
Face Care, Shampoo, and Body Care had the highest number of unique brands.
This indicates high brand competition in personal care products.
The sunburst chart visualized the hierarchical structure of category and sub_category, providing a detailed view of how products are grouped.

ACKNOWLEDGEMENT

I WOULD LIKE TO TAKE THIS OPPORTUNITY TO EXPRESS MY GRATITUDE FOR THE RESOURCES, TOOLS, AND KNOWLEDGE THAT ENABLED ME TO SUCCESSFULLY COMPLETE THIS PROJECT.

AS THIS WAS AN INDEPENDENT EFFORT, I AM PROUD TO HAVE CARRIED OUT ALL ASPECTS OF THE PROJECT—FROM DATA CLEANING AND VISUALIZATION TO ANALYSIS AND DOCUMENTATION—ON MY OWN. THIS EXPERIENCE ALLOWED ME TO APPLY THEORETICAL KNOWLEDGE TO A REAL-WORLD DATASET AND DEVELOP A DEEPER UNDERSTANDING OF DATA SCIENCE CONCEPTS AND PRACTICAL TOOLS SUCH AS PYTHON, PANDAS, MATPLOTLIB, AND SEABORN.

THIS PROJECT HAS BEEN A VALUABLE LEARNING JOURNEY, AND I AM GRATEFUL FOR THE OPPORTUNITY TO EXPLORE AND GROW THROUGH SELF-DRIVEN WORK.