BigBasket Product Assortment Analysis

Project Category: Exploratory Data Analysis

Tools & Technologies Used: Python, NumPy, Pandas, Matplotlib, Seaborn, Jupyter Notebooks

Introduction

BigBasket, India's leading online grocery platform, offers an extensive range of products spanning across multiple categories, subcategories, and brands. As the platform continues to grow and expand its catalog—now encompassing nearly 28,000 products—it faces increasing challenges in maintaining an optimal assortment, offering competitive pricing, and curating a brand portfolio that is both diverse and aligned with customer expectations.

In the fiercely competitive landscape of online grocery retail, staying ahead requires more than just variety. It demands a strategic balance of product availability, price segmentation, and brand representation. Without consistent evaluation, the assortment risks becoming bloated, imbalanced, or misaligned with market demands—potentially leading to customer dissatisfaction and missed revenue opportunities.

This project undertakes a comprehensive analysis of BigBasket's current product lineup to uncover insights that can drive smarter decision-making in assortment planning, pricing strategies, and brand positioning. Using a rich dataset of approximately 28,000 unique product entries, the analysis aims to identify gaps, inefficiencies, and strategic opportunities that can support data-driven growth initiatives.

🎯 Business Objectives

This analysis is structured around four core strategic pillars:

1. **Product Assortment Optimization**
    - Evaluate product coverage across categories and subcategories.
    - Identify underrepresented or overrepresented segments.
    - Detect redundant or potentially outdated items.
2. **Pricing Strategy Analysis**
    - Examine price distributions and outliers within and across segments.
    - Analyze brand-wise price positioning and consistency.
3. **Brand and Category Positioning**
    - Assess brand dominance or fragmentation within each category.
    - Identify risks from low brand diversity or over-concentration.
    - Spot opportunities for private labels or third-party collaborations.
4. **Opportunity and Gap Identification**
    - Highlight high-performing but low-assortment categories.
    - Detect opportunities for tiered pricing strategies and product bundling.

- Pinpoint gaps in offerings that could enhance customer value perception.

Project Overview

Dataset Description:

The dataset used in this analysis comprises **28,000+ products** listed on BigBasket's platform. Each record represents a unique product offering and includes a variety of attributes related to product identity, categorization, pricing, and brand information.

BigBasket Products List.csv

🔍 Key Features:

- **Product Name** – The official title of the product listed on the platform.
- **Category & Subcategory** – Hierarchical classification of the product (e.g., Beverages → Juices).
- **Brand** – The brand under which the product is marketed.
- **Pack Size** – Describes the quantity or weight of the product (e.g., 1L, 500g).
- **MRP (Maximum Retail Price)** – The listed price before any discounts.
- **Selling Price** – The price after discounts or promotions.
- **Discount (%)** – The discount percentage applied, if any.
- **Rating** – Customer rating score (where available).
- **Number of Ratings** – The total number of user-submitted ratings.
- **Description/Tags** – Textual product descriptors (e.g., organic, sugar-free, etc.)

📦 Structure:

- **Format:** CSV (Comma Separated Values)
- **Rows:** ~28,000
- **Columns:** ~10–12 core attributes, depending on availability

This dataset provides a rich foundation for analyzing category depth, price strategies, brand presence, and identifying gaps in the product portfolio.

Preprocessing Steps:

Effective data preprocessing was critical to ensure accuracy and consistency before analysis. The following steps were applied:

🔍 Handling Missing Values

- **Critical Fields**: Rows with missing `product` or `brand` names (only one each) were dropped to maintain data integrity.
- **Descriptive Text**: Missing values in the `description` column (~0.4%) were filled with a placeholder ("No description available") to retain completeness.
- **Ratings**: Approximately 31% of ratings were missing. These were retained to avoid significant data loss, as missing values could represent unrated or new products.

🧹 Dropping Redundant Columns

- The original `index` column was removed as it had no analytical value and duplicated Pandas' inherent indexing.

⚙️ Standardizing Column Names

- All column names were standardized to lowercase and underscores (e.g., `sale_price`, `sub_category`) to enhance readability and prevent code-related errors.

🧱 Data Type Optimization

- Categorical fields (`category`, `sub_category`, `brand`, `type`) were converted to the `category` data type to improve memory efficiency and processing speed.

🛠️ Feature Engineering

- **Discount Amount** and **Discount Percentage** were derived using `market_price` and `sale_price`, enabling in-depth pricing and promotional analysis.
- These features became essential in evaluating pricing strategies across categories and brands.

🧪 Data Validation

- Checked for and confirmed the absence of:
  - Zero or negative prices (none found)
  - Invalid ratings (none found; all between 1.0–5.0)

🧼 Categorical Normalization

- All string values in key categorical columns (`product`, `category`, `sub_category`, `brand`, `type`) were lowercased and stripped of extra spaces to eliminate inconsistencies and duplicates.

📈 Outlier Handling

- Outliers in pricing were retained intentionally:
  - **Rationale**: Outliers in sale or market prices reflect actual high-end or niche products and align with real-world assortment strategies.
  - **Usefulness**: They provide insight into premium tiers and help identify pricing spread across product types.

Data Exploration

During the initial exploration phase, the dataset was examined to understand its structure, distribution, and potential irregularities. Key steps included:

- **Structural Inspection**: Reviewed data types, column completeness, and overall dataset shape.
- **Value Counts & Uniqueness**: Assessed the number of unique categories, subcategories, and brands to understand breadth and diversity.
- **Descriptive Statistics**: Generated summary statistics (mean, median, min, max, standard deviation) for numerical fields like `Price`, `Rating`, and `Discount`.
- **Distribution Checks**:
  - Analyzed product distribution across categories and subcategories.
  - Explored rating and pricing distributions to detect skewness or anomalies.

- **Missing & Zero Values**: Identified nulls or zero values in key fields to flag potential data quality issues.
- **Outlier Detection**: Flagged unusually high or low values in price and ratings using quantiles and visual plots.
- **Sample Review**: Manually inspected random records for real-world inconsistencies (e.g., incorrect pack sizes or duplicate descriptions).

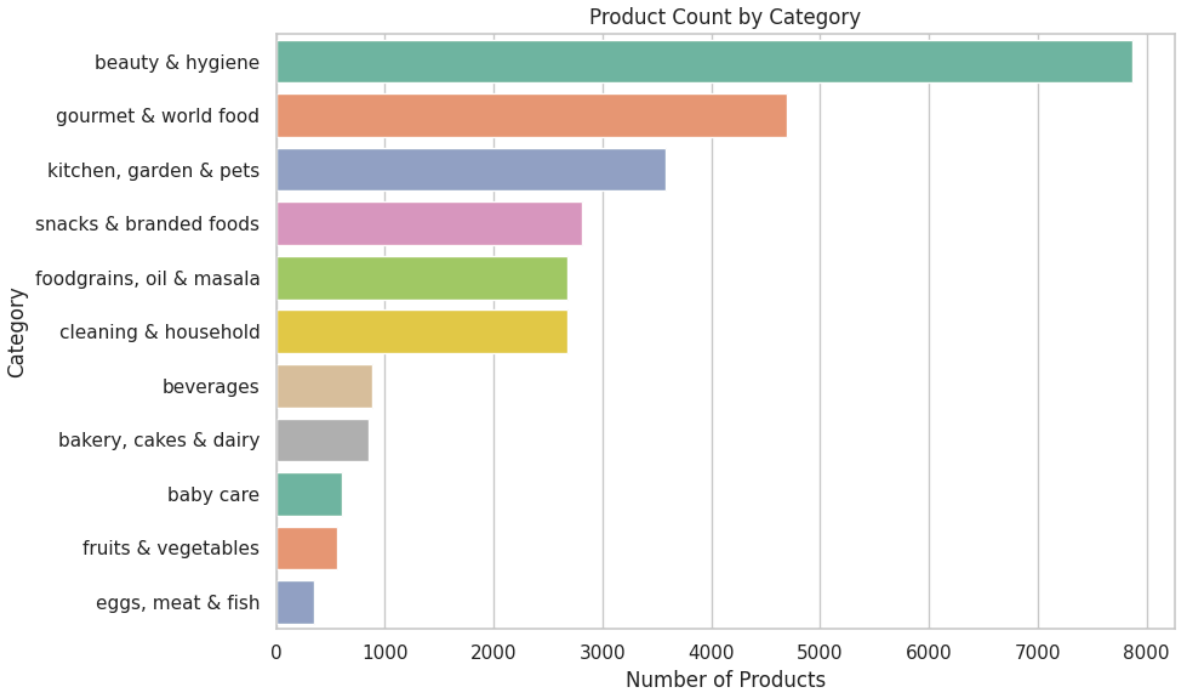
This foundational analysis provided valuable context for identifying trends, cleaning requirements, and forming early hypotheses for deeper insights.

Analysis Methods and Key Insights

To uncover strategic insights aligned with business objectives, the project utilized a range of exploratory and comparative data analysis methods. These included:

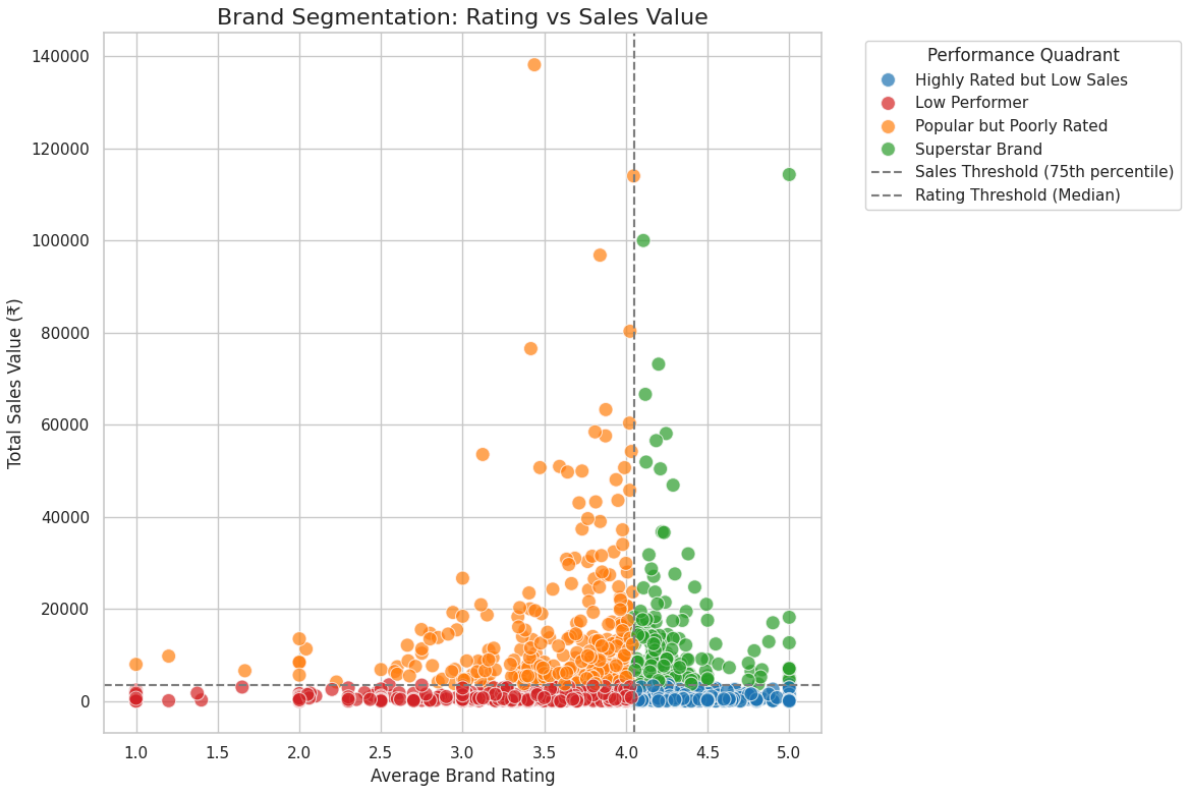
🔎 1. Univariate Analysis

- **Purpose:** Understand the distribution of individual variables.
- **Methods Used:** Histograms, bar plots, value counts, descriptive statistics.
- **Key Insights:**
  - Certain categories (e.g., Beauty & Hygiene) dominated the product lineup.
  - A large proportion of products had no ratings, indicating customer engagement gaps.
  - Prices were heavily right-skewed, with most products under ₹500.
- **Visualizations**:



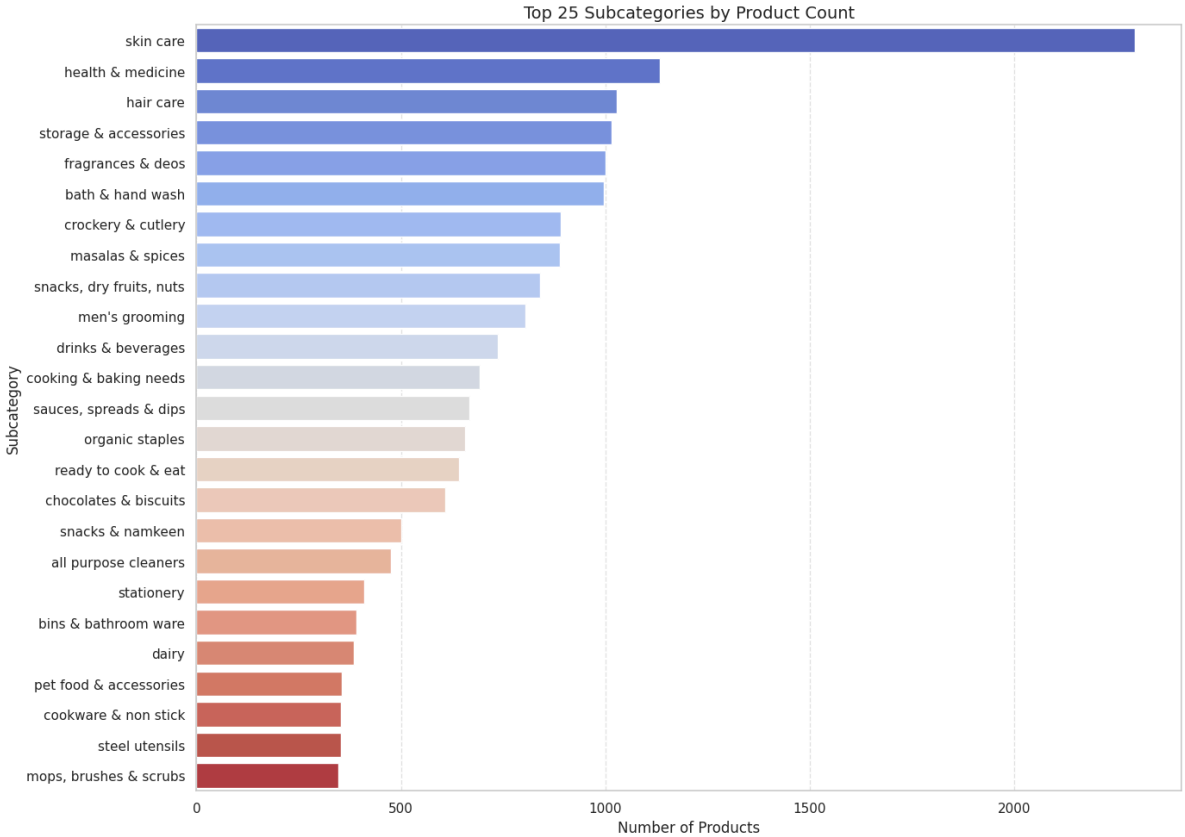
〓 2. Bivariate & Multivariate Analysis

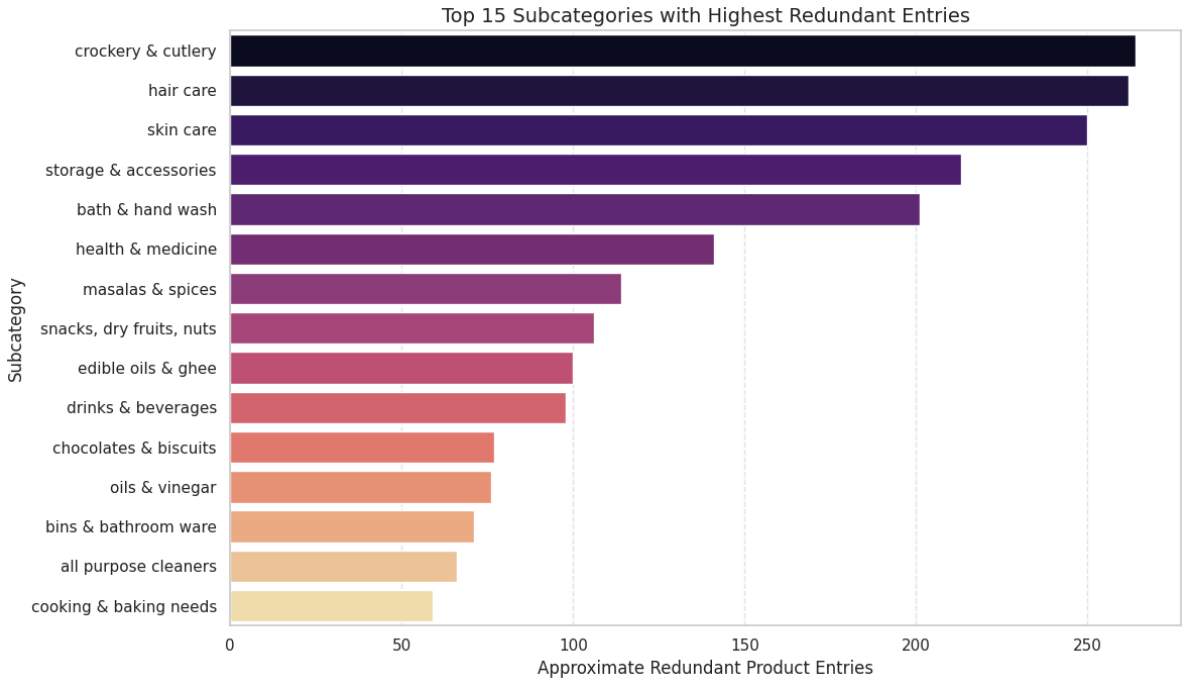
- **Purpose:** Explore relationships between variables like category, price, brand, and rating.
- **Methods Used:** Box plots, group-by aggregations, scatter plots, correlation checks.
- **Key Insights:**
  - Some brands positioned themselves with consistently higher pricing across subcategories.
  - A mismatch was found between high ratings and low assortment in certain segments.
  - Subcategories with wide price ranges suggested opportunities for **tiered pricing strategies**.
- **Visualizations**:



3. Product Assortment Analysis

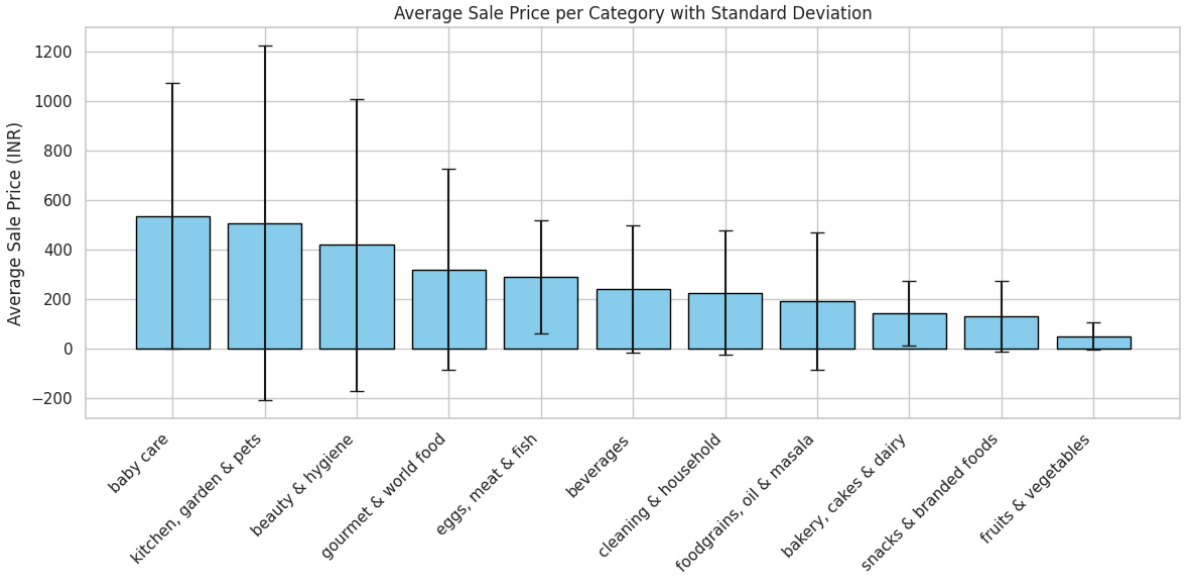
- **Purpose:** Evaluate depth and coverage across product categories and subcategories.
- **Methods Used:** Product counts, pivot tables, brand counts per subcategory.
- **Key Insights:**
  - Overrepresented subcategories could lead to internal competition and customer overwhelm.
  - Certain subcategories lacked brand diversity, raising potential **monopoly risks**.
  - Redundancy was detected in pack sizes and product variants (especially in categories like crockery & cutlery).
- **Visualizations**:

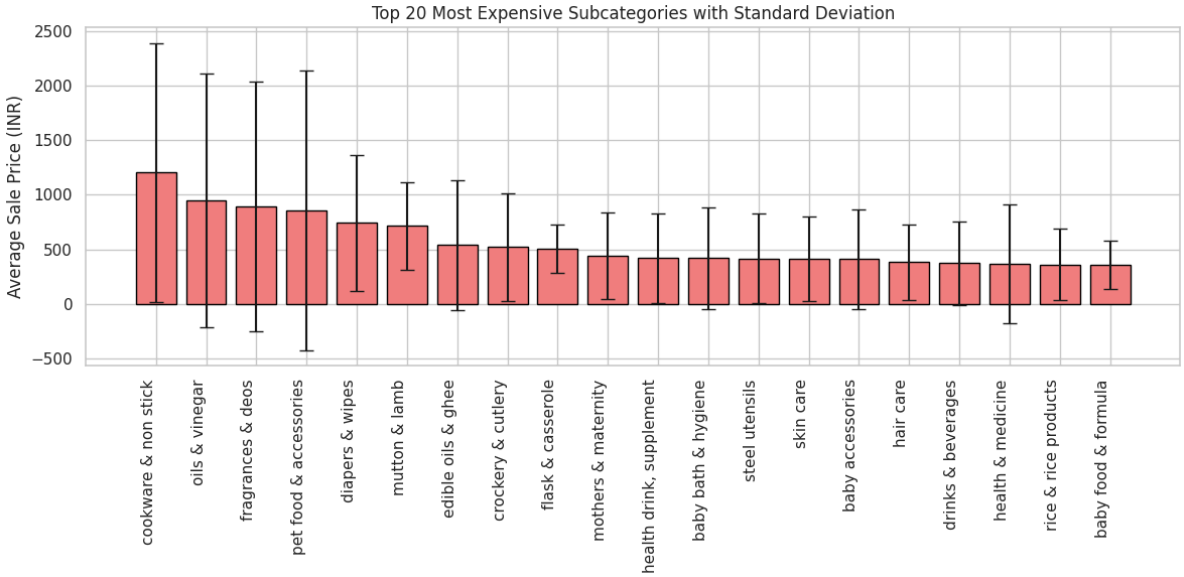


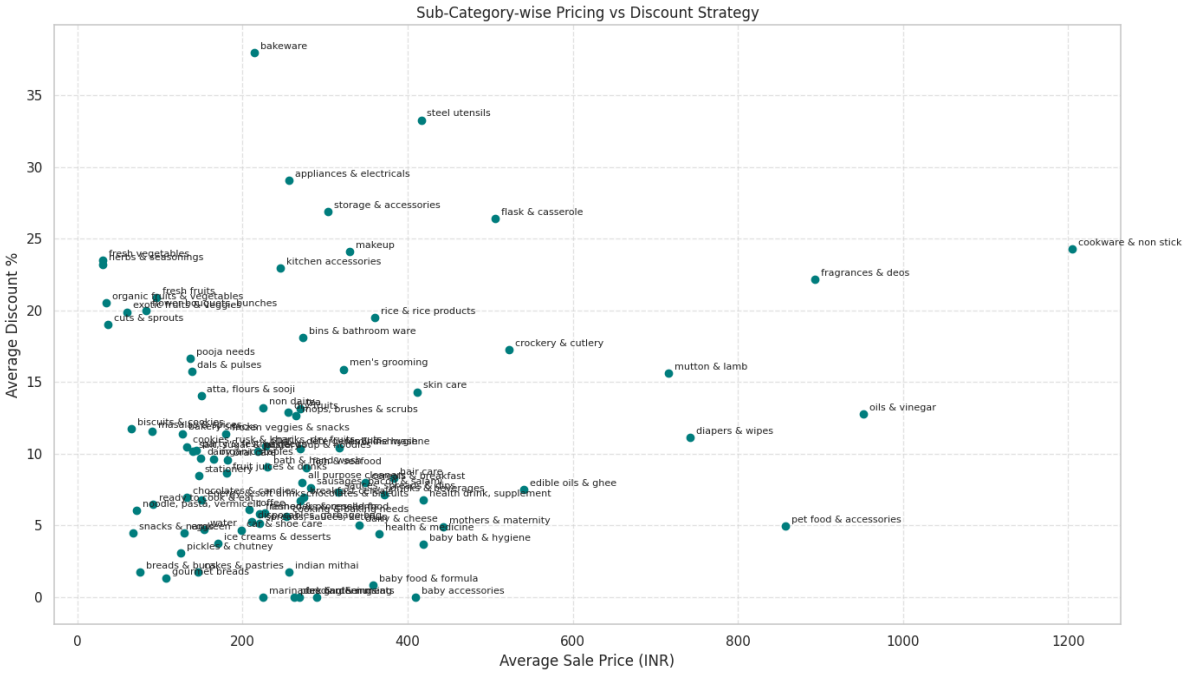


💰 4. Pricing Strategy Analysis

- **Purpose:** Analyze pricing trends, anomalies, and brand-level price positioning.
- **Methods Used:** Box plots by category/brand, discount analysis, MRP vs. Selling Price comparison.
- **Key Insights:**
  - Some categories showed pricing inconsistencies—similar products with vastly different prices.
  - High discount rates in specific subcategories suggested **margin pressure or promotional overuse**.
  - Private label brands were often priced lower than national brands in the same category, with comparable ratings.
- **Visualizations**:

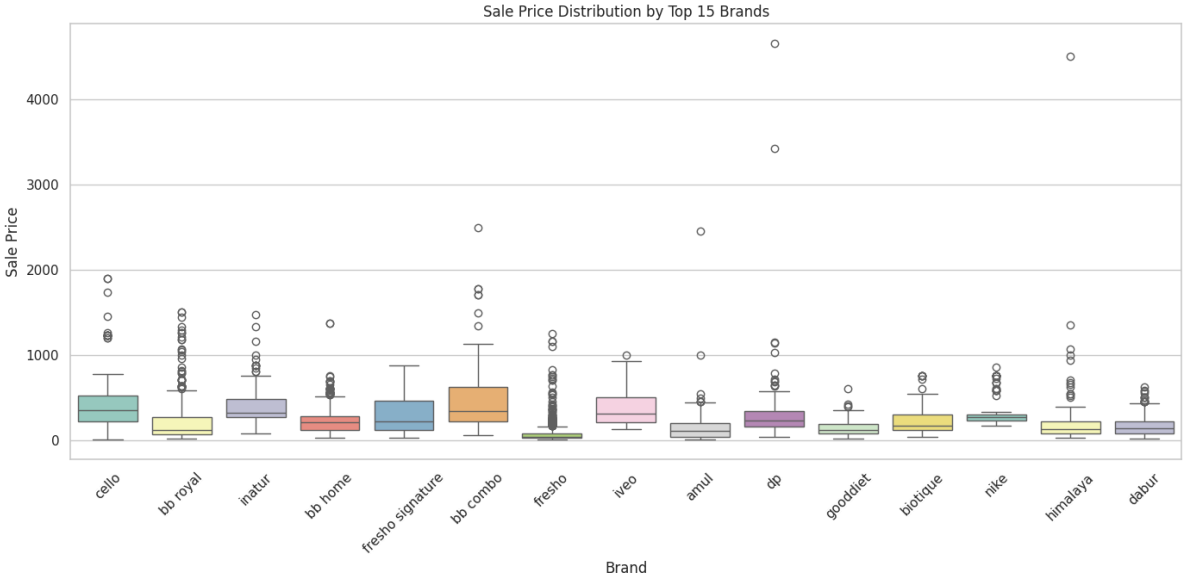

Average Sale Price per Category with Standard Deviation
Average Sale Price (INR)
1200
1000
800
600
400
200
0
−200
baby care
kitchen, garden & pets
beauty & hygiene
gourmet & world food
eggs, meat & fish
beverages
cleaning & household
foodgrains, oil & masala
bakery, cakes & dairy
snacks & branded foods
fruits & vegetables

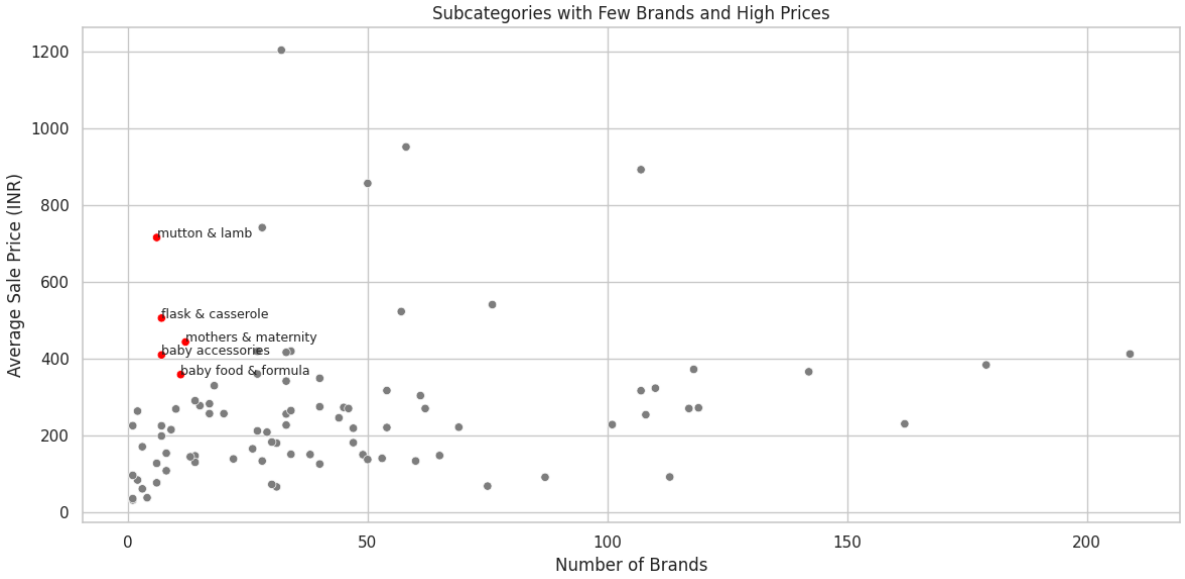

Top 20 Most Expensive Subcategories with Standard Deviation
Average Sale Price (INR)
2500
2000
1500
1000
500
0
−500
cookware & non stick
oils & vinegar
fragrances & deos
pet food & accessories
diapers & wipes
mutton & lamb
edible oils & ghee
crockery & cutlery
flask & casserole
mothers & maternity
health drink, supplement
baby bath & hygiene
steel utensils
skin care
baby accessories
hair care
drinks & beverages
health & medicine
rice & rice products
baby food & formula



5. Brand & Category Positioning

- **Purpose:** Examine brand presence and influence within categories.
- **Methods Used:** Brand frequency analysis, dominance ratio, rating averages by brand.
- **Key Insights:**
  - Several categories had high brand fragmentation, possibly diluting customer loyalty.
  - Some emerging brands were highly rated but had limited shelf space—indicating **growth opportunities**.
  - Opportunities were identified for **private label expansion** in underpenetrated categories.
- **Visualizations**:

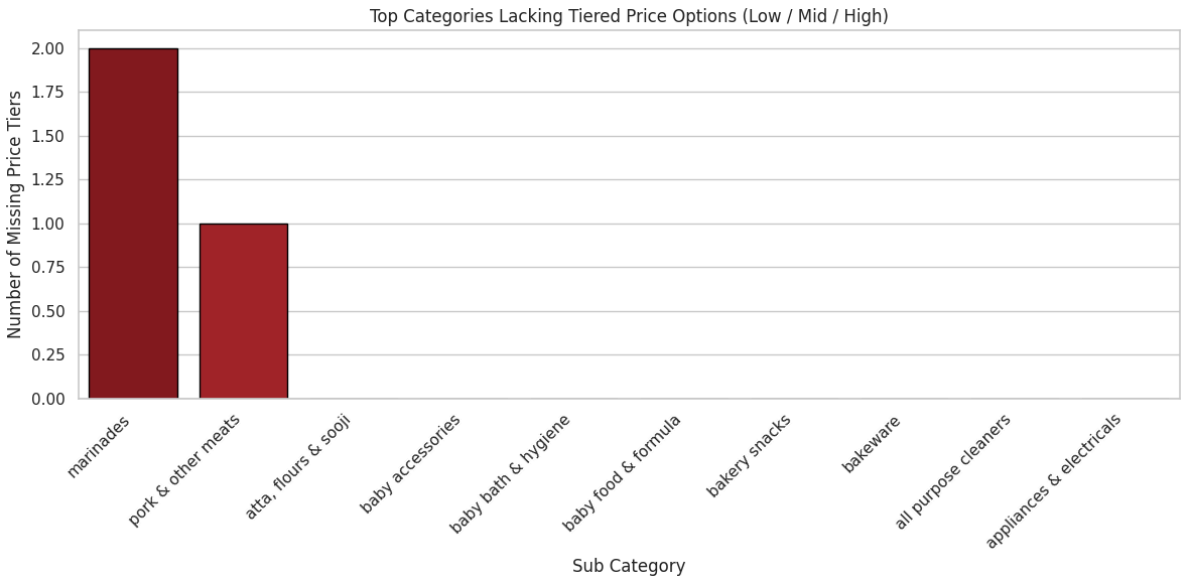


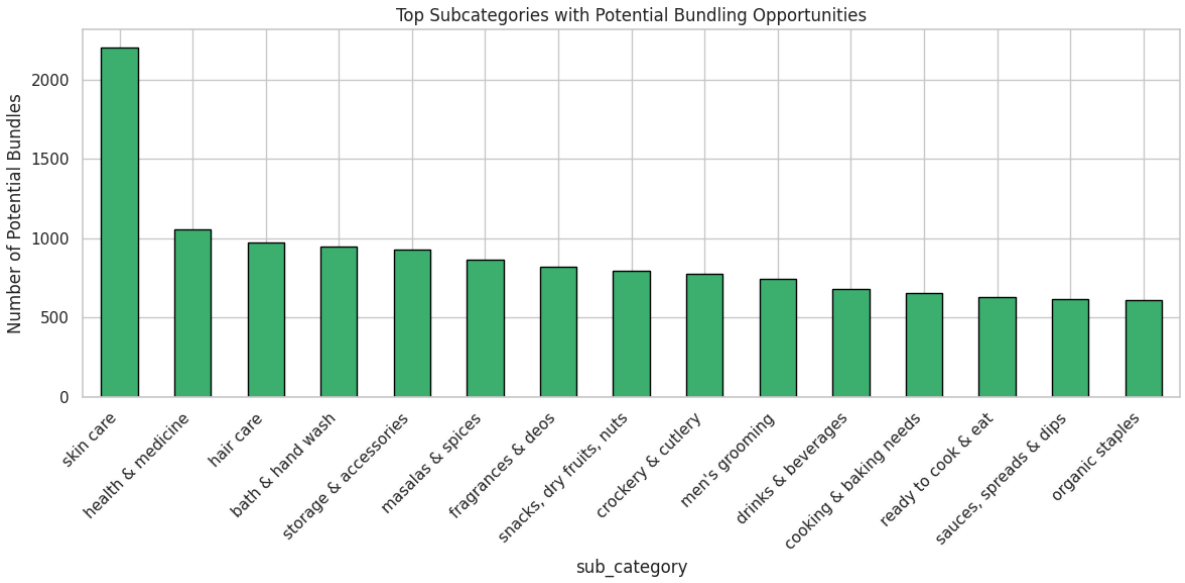


🔍 6. Gap & Opportunity Analysis

- **Purpose:** Identify where assortment, pricing, or branding fell short of market expectations.

- **Methods Used:** Custom filtering logic, high-rating/low-assortment cross-checks, bundling logic.
- **Key Insights:**
  - Several high-demand subcategories lacked tiered product offerings (e.g., premium vs. budget).
  - Bundling potential was observed in complementary products with high co-occurrence.
  - Specific categories showed high average prices but had very few products —signaling underexplored premium niches.
- **Visualizations**:





Conclusion

This comprehensive analysis of BigBasket's product assortment has revealed valuable insights into the platform's current market positioning, pricing strategy, and brand diversity. Through methodical data exploration and targeted evaluations aligned with business objectives, the project has identified several key areas of strength as well as opportunities for strategic improvement.

The analysis confirmed that while BigBasket offers a broad and diverse product catalog, this very scale introduces challenges—such as category imbalances, inconsistent pricing, brand dominance issues, and missed opportunities in high-demand yet low-assortment segments. Certain subcategories are oversaturated, leading to internal competition, while others are underrepresented despite strong customer interest and high ratings. Similarly, price variations and discounting patterns across brands and subcategories indicate a need for tighter pricing governance and structured tiering.

Brand-wise, the presence of monopolistic or fragmented brand landscapes across categories suggests that a deliberate approach to portfolio management—incorporating both private labels and strategic partnerships—can strengthen BigBasket's competitive edge.

Overall, this analysis offers a data-driven foundation for optimizing BigBasket's product strategy. By implementing insights derived from the study—such as enhancing brand assortment in select categories, standardizing price bands,

leveraging bundling opportunities, and expanding into underdeveloped premium niches—BigBasket can enhance customer satisfaction, improve operational efficiency, and solidify its leadership in India's online grocery market.

Recommendations

Based on the analysis findings, the following strategic actions are recommended to enhance BigBasket's product assortment, pricing structure, and brand portfolio performance:

🧺 1. Product Assortment Optimization

- **Reduce redundancy** in overrepresented subcategories by consolidating similar SKUs and pack sizes.
- **Expand assortment** in high-performing, underrepresented categories.
- **Audit outdated products** with low ratings and sales potential for potential phase-out.

💸 2. Pricing Strategy Enhancement

- **Introduce clear tiered pricing** across major subcategories to cater to budget, mid-range, and premium customer segments.
- **Standardize pricing logic** within categories to reduce internal price conflicts and customer confusion.
- **Monitor deep discounting trends** in certain categories that may erode margins; optimize promotional spending.

🏷️ 3. Brand and Category Positioning

- **Promote emerging high-rated brands** with limited presence by increasing visibility and distribution.
- **Balance brand concentration** in monopolistic categories to mitigate supply risks and promote healthy competition.
- **Invest in private label development** in high-margin or brand-fragmented categories to improve profitability and control.

🔍 4. Gap and Opportunity Identification

- **Develop bundled offerings** for complementary products (e.g., cereals + milk, snacks + beverages) to increase basket size and perceived value.
- **Identify high-price, low-assortment subcategories** for premium product expansion.
- **Target underpenetrated brand niches** in fast-moving subcategories to capture untapped demand.

📊 5. Data-Driven Portfolio Management

- **Implement ongoing assortment monitoring dashboards** to track category saturation, brand share, and pricing trends.
- **Leverage customer feedback and engagement data** (ratings, reviews, repeat purchases) for assortment decisions.
- **Integrate pricing and assortment insights** with marketing and inventory planning teams for aligned execution.

These actions are designed to not only address current inefficiencies but also to unlock new growth opportunities for BigBasket through a more strategic, customer-aligned, and competitive product catalog.

Limitations and Future Work

⚠️ Limitations

While the analysis provided valuable insights, several limitations should be acknowledged:

- **Lack of Sales Data:** The dataset does not include actual sales figures (e.g., units sold, revenue per product), limiting the ability to directly correlate assortment or pricing with performance.
- **No Customer Demographics or Behavior Data:** Without insights into customer segments, preferences, or buying patterns, the recommendations remain generalized rather than customer-specific.

- **Static Snapshot:** The dataset represents a static snapshot in time. Trends, seasonality, and time-dependent variations in demand or pricing cannot be captured without longitudinal data.
- **Limited Product Metadata:** Product tags and descriptions were not standardized or leveraged due to inconsistency, reducing the ability to perform in-depth NLP-based tagging or feature extraction.
- **Subjective Ratings Coverage:** A large number of products lack customer ratings or have very few, making it difficult to generalize sentiment or quality perception accurately across the catalog.

🔮 Future Scope

To build on the current work and enable deeper strategic insights, the following extensions are recommended:

- **Incorporate Transactional Sales Data:** Adding order-level data will allow for sales-driven analysis of product performance, pricing elasticity, and bundling effectiveness.
- **Customer Segmentation Analysis:** Integrating demographic or behavioral data can help tailor assortment and pricing strategies to different customer groups.
- **Time Series and Seasonal Trends:** Analyzing data over time will help detect product lifecycle trends, promotional effectiveness, and seasonal category demand shifts.
- **Competitor Benchmarking:** Comparing BigBasket's pricing and assortment with key competitors can highlight areas of differentiation or gaps in market coverage.
- **Advanced NLP on Product Texts:** Standardizing and mining product descriptions and tags can reveal latent attributes (e.g., "organic", "gluten-free", "baby-safe") that influence buying behavior.
- **Dynamic Assortment Models:** Implementing machine learning models to predict demand, optimize shelf space, and recommend additions or deletions from the catalog.

Addressing these areas will help evolve this analysis into a more dynamic, real-time, and customer-centered decision-making framework for BigBasket's merchandising

strategy.

References and Citations

- Dataset Source: **Download Dataset**
- Python Libraries Used: `pandas`, `numpy`, `matplotlib`, `seaborn`
- Kaggle - https://www.kaggle.com/datasets/chinmayshanbhag/big-basket-products
- https://www.theproductfolks.com/product-management-case-studies/improve-the-userbase-of-bigbasket
- https://www.bbmatrix.ai/case-studies/big-basket

Code Documentation

- All analysis steps were conducted in Python Jupyter Notebook.
- Each cell is documented with markdown explanations for clarity.
- Instructions:
  - Required: Python ≥3.7, Jupyter Notebook
  - Install required packages with: `pip install pandas numpy matplotlib seaborn`

Materials & Resources

IPYNB Notebook: https://colab.research.google.com/drive/1_yLgcjAHwjWokom_zBbConsCq27GFX53?usp=sharing

GitHub Link: https://github.com/Rudrajit12/BigBasket-Product-Assortment-Optimization

Credits:

Author: Rudrajit Bhattacharyya

Email ID: rudrajitb24@gmail.com

LinkedIn: https://www.linkedin.com/in/rudrajitb/

GitHub: https://github.com/Rudrajit12