MINI PROJECT

bigbasket

Exploratory Data Analysis

Presented by
Fardeen Ansari

fardeenansari203@gmail.com

LinkedIn

INDEX

About BigBasket

BigBasket is one of India's largest and most reliable online grocery platforms, established in 2011. The company offers a broad assortment of products such as fresh fruits and vegetables, household essentials, personal care items, beverages, and packaged foods. With a strong delivery network and a simple, user-friendly app and website, BigBasket enables customers to shop conveniently from home. It emphasizes product quality, on-time delivery, and competitive pricing to enhance the customer experience. Its business model focuses on fresh produce, a wide range of categories, and frequent discounts to attract and retain users.

BigBasket continues to expand its presence across major Indian cities, strengthening its position as a leading player in India's fast-growing online grocery market.

Tools

Objectives

- To understand the structure and main features of the BigBasket product dataset.
- To clean the data by handling missing values, outliers, and inconsistent records.
- To analyze product prices, discounts, and discount percentages.
- To identify top-rated and least-rated products based on customer ratings.
- To study product categories, sub-categories, brands, and pricing patterns.
- To visualize important insights using charts and graphs.
- To extract meaningful insights that help understand customer preferences and product performance.

Aim

The aim of this project is to analyze the BigBasket product dataset to gain a clear understanding of pricing patterns, discounts, product categories, brands, and customer ratings in the online grocery market. This project focuses on cleaning the dataset by handling missing values, outliers, and inconsistent data to ensure accurate analysis. Using data visualization techniques, the project seeks to identify popular and low-performing products, study customer preferences, and observe trends across different categories. The overall goal is to extract meaningful business insights that can help understand product performance, customer behavior, and the role of data-driven decision-making in e-commerce platforms like BigBasket.

Load DataSet

Load Data Set

```
[3]: a = 'BigBasket Products.csv'

[4]: df = pd.read_csv(a)
     df
```

[4]:

	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description
0	1	Garlic Oil - Vegetarian Capsule 500 mg	Beauty & Hygiene	Hair Care	Sri Sri Ayurveda	220.00	220.0	Hair Oil & Serum	4.1	This Product contains Garlic Oil that is known...
1	2	Water Bottle - Orange	Kitchen, Garden & Pets	Storage & Accessories	Mastercook	180.00	180.0	Water & Fridge Bottles	2.3	Each product is microwave safe (without lid), ...
2	3	Brass Angle Deep - Plain, No.2	Cleaning & Household	Pooja Needs	Trm	119.00	250.0	Lamp & Lamp Oil	3.4	A perfect gift for all occasions, be it your m...
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	Cleaning & Household	Bins & Bathroom Ware	Nakoda	149.00	176.0	Laundry, Storage Baskets	3.7	Multipurpose container with an attractive desi...
4	5	Creme Soft Soap - For Hands & Body	Beauty & Hygiene	Bath & Hand Wash	Nivea	162.00	162.0	Bathing Bars & Soaps	4.4	Nivea Creme Soft Soap gives your skin the best...
...	...	...	...	...	...	...	...	...	...	...
27550	27551	Wottagirl Perfume Spray - Heaven, Classic	Beauty & Hygiene	Fragrances & Deos	Layerr	199.20	249.0	Perfume	3.9	Layerr brings you Wottagirl Classic fragrant b...
27551	27552	Rosemary	Gourmet & World Food	Cooking & Baking Needs	Puramate	67.50	75.0	Herbs, Seasonings & Rubs	4.0	Puramate rosemary is enough to transform a dis...
27552	27553	Peri-Peri Sweet Potato Chips	Gourmet & World Food	Snacks, Dry Fruits, Nuts	FabBox	200.00	200.0	Nachos & Chips	3.8	We have taken the richness of Sweet

In this step, the BigBasket product dataset is loaded into the Python environment for analysis. The dataset is stored in a CSV file named "BigBasket Products.csv" and is imported using the Pandas library with the read_csv() function. After loading, the dataset is assigned to a DataFrame called df, which allows easy data handling and analysis. The DataFrame contains 27,555 rows and 10 columns, including product name, category, sub-category, brand, sale price, market price, product type, rating, and description.

Head Function

Look at first 12 rows.

```
print('First 12 Rows :')
df.head(12)
```

First 12 Rows :

	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description
0	1	Garlic Oil - Vegetarian Capsule 500 mg	Beauty & Hygiene	Hair Care	Sri Sri Ayurveda	220.0	220.0	Hair Oil & Serum	4.1	This Product contains Garlic Oil that is known...
1	2	Water Bottle - Orange	Kitchen, Garden & Pets	Storage & Accessories	Mastercook	180.0	180.0	Water & Fridge Bottles	2.3	Each product is microwave safe (without lid), ...
2	3	Brass Angle Deep - Plain, No.2	Cleaning & Household	Pooja Needs	Trm	119.0	250.0	Lamp & Lamp Oil	3.4	A perfect gift for all occasions, be it your m...
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	Cleaning & Household	Bins & Bathroom Ware	Nakoda	149.0	176.0	Laundry, Storage Baskets	3.7	Multipurpose container with an attractive desi...
4	5	Creme Soft Soap - For Hands & Body	Beauty & Hygiene	Bath & Hand Wash	Nivea	162.0	162.0	Bathing Bars & Soaps	4.4	Nivea Creme Soft Soap gives your skin the best...
5	6	Germ - Removal Multipurpose Wipes	Cleaning & Household	All Purpose Cleaners	Nature Protect	169.0	199.0	Disinfectant Spray & Cleaners	3.3	Stay protected from contamination with Multipu...
6	7	Multani Mati	Beauty & Hygiene	Skin Care	Satinance	58.0	58.0	Face Care	3.6	Satinance multani matti is an excellent skin t...
7	8	Hand Sanitizer - 70% Alcohol Base	Beauty & Hygiene	Bath & Hand Wash	Bionova	250.0	250.0	Hand Wash & Sanitizers	4.0	70%Alcohol based is gentle of hand leaves skin...
8	9	Biotin & Collagen Volumizing Hair Shampoo + Bi...	Beauty & Hygiene	Hair Care	StBotanica	1098.0	1098.0	Shampoo & Conditioner	3.5	An exclusive blend with Vitamin B7 Biotin, Hyd...
9	10	Scrub Pad - Anti- Bacterial, Regular	Cleaning & Household	Mops, Brushes & Scrubs	Scotch brite	20.0	20.0	Utensil Scrub-Pad, Glove	4.3	Scotch Brite Anti- Bacterial Scrub Pad thoroug...
10	11	Wheat Grass Powder - Raw	Gourmet & World Food	Cooking & Baking Needs	NUTRASHIL	261.0	290.0	Flours & Pre-Mixes	4.0	Wheatgrass is a superfood potent health food w...
11	12	Butter Cookies Gold Collection	Gourmet & World Food	Chocolates & Biscuits	Sapphire	600.0	600.0	Luxury Chocolates, Gifts	2.2	Enjoy a tin full of delicious butter cookies m...

After loading the dataset, the head(12) function is used to display the first 12 rows of the BigBasket product dataset. This step helps in getting an initial understanding of the data structure, column names, and sample values present in the dataset. By examining these rows, we can observe important details such as product names, categories, sub-categories, brands, sale prices, market prices, product types, ratings, and descriptions. Viewing the first few records is an essential step in data analysis, as it helps identify data quality issues, understand variable formats.

Description

Get Description of the data

```
print('\n Description :')
df.describe(include="all")
df.describe()
```

Description :

	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

the describe() function is used to generate descriptive statistics for the BigBasket product dataset. This provides a statistical summary of numerical columns such as sale_price, market_price, and rating. The output includes key measures like count, mean, standard deviation, minimum, maximum, and quartiles (25%, 50%, and 75%).

Info

Find Information about the DataFrame

```
]: 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
```

The df.info() function is used to get a summary of the dataset structure. It shows the total number of rows and columns, column names, data types, and non-null values for each column.

From the output, we can see that the dataset contains 27,555 rows and 10 columns. It includes numerical columns such as sale_price, market_price, and rating, and categorical columns like product, category, sub_category, brand, type, and description.

Some columns, such as rating, sale_price, brand, and description, have missing values, which indicates that data cleaning is required

Top & Least Sold Product

Find out Top & least sold products

```
rating_counts =df.groupby(['product'])['rating'].count().reset_index()
top_sold = rating_counts.sort_values(by='rating', ascending=False).head(5)

# Least 5 sold products
least_sold = rating_counts.sort_values(by='rating', ascending=True).head(5)

top_sold.head(), least_sold.head()
```

```
(                          product  rating
 22257  Turmeric Powder/Arisina Pudi      23
 5365                Cow Ghee/Tuppa      12
 19971                   Soft Drink      12
 8431                    Ghee/Tuppa      11
 17011           Powder - Coriander      11,
                                        product  rating
 7311              Face Scrub - Lavender With Chamomile       0
 7312                                Face Scrub - Nutri       0
 7313                    Face Scrub - Strawberry Extract       0
 7301  Face Pack - Bio Fruit Spot Lightening, Bxl Cel...       0
 7321                              Face Wash & Day Cream       0)
```

This step identifies the most sold and least sold products based on the count of customer ratings. Products with higher rating counts are considered top-selling, while those with zero or very low ratings are considered least sold.

The results show that items like Turmeric Powder/Arisina Pudi and Cow Ghee/Tuppa are among the top sold products, whereas some face scrubs and personal care items appear in the least sold category. This analysis helps understand product demand and customer interest across different items.

Discount

Measuring discount on a certain item.

```
df['discount']=df['market_price']-df['sale_price']
```

```
df['discount']=(df['discount']/df['market_price'])*100
```

```
df.head(5)
```

	index	product	category	sub_category	brand	sale_price	market_price	type	rating	description	discount
0	1	Garlic Oil - Vegetarian Capsule 500 mg	Beauty & Hygiene	Hair Care	Sri Sri Ayurveda	220.0	220.0	Hair Oil & Serum	4.1	This Product contains Garlic Oil that is known...	0.000000
1	2	Water Bottle - Orange	Kitchen, Garden & Pets	Storage & Accessories	Mastercook	180.0	180.0	Water & Fridge Bottles	2.3	Each product is microwave safe (without lid), ...	0.000000
2	3	Brass Angle Deep - Plain, No.2	Cleaning & Household	Pooja Needs	Trm	119.0	250.0	Lamp & Lamp Oil	3.4	A perfect gift for all occasions, be it your m...	52.400000
3	4	Cereal Flip Lid Container/Storage Jar - Assort...	Cleaning & Household	Bins & Bathroom Ware	Nakoda	149.0	176.0	Laundry, Storage Baskets	3.7	Multipurpose container with an attractive desi...	15.340909
4	5	Creme Soft Soap - For	Beauty &	Bath & Hand	Nivea	162.0	162.0	Bathing Bars	4.4	Nivea Creme Soft Soap	0.000000

In this step, discounts are calculated for each product by finding the difference between the market price and the sale price. The discount value is then converted into a percentage to clearly understand the level of price reduction offered on each item. This analysis helps in identifying products with high, low, or no discounts and comparing pricing strategies across different categories and brands. Studying discount percentages also provides insights into promotional strategies and customer attraction methods. Overall, discount analysis supports a better understanding of pricing behavior and its impact on product demand and customer purchasing decisions in the online grocery market.

Missing Values

Find out the Missing Values from the Dataset

```
df.isnull().sum()
```

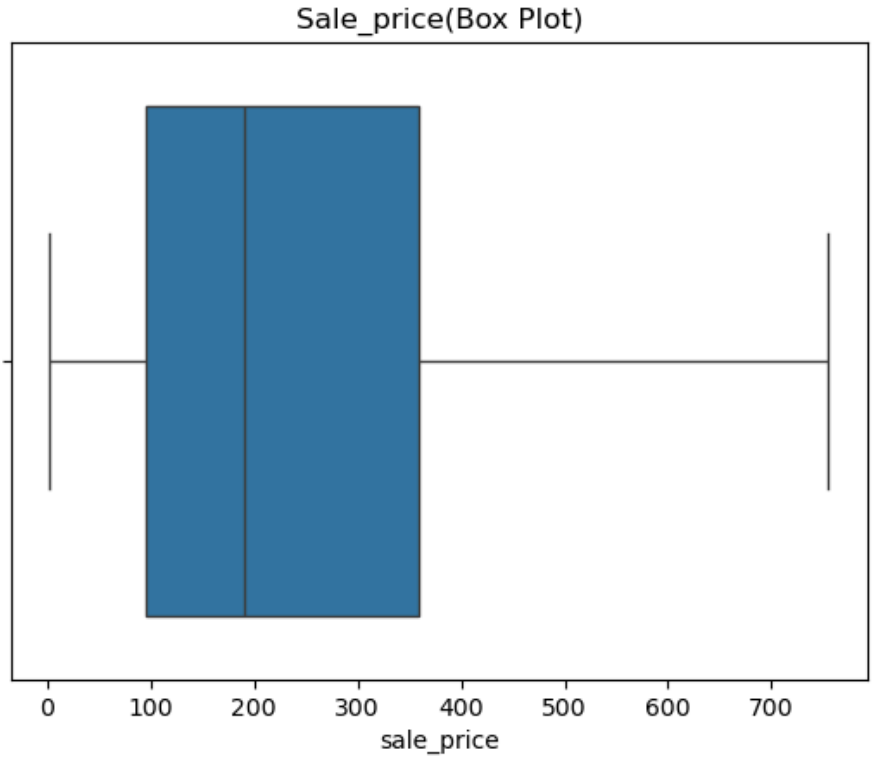
```
index             0
product           1
category          0
sub_category      0
brand             1
sale_price        6
market_price      0
type              0
rating         8636
description     115
discount          6
dtype: int64
```

```
df['rating'] = df['rating'].fillna(df['rating'].mean())
df['description'] = df['description'].fillna('Unknown')
df['discount'] = df['discount'].fillna(df['discount'].mean())
df.isnull().sum()
```

```
index          0
product        1
category       0
sub_category   0
brand          1
sale_price     6
market_price   0
type           0
rating         0
description    0
discount       0
dtype: int64
```

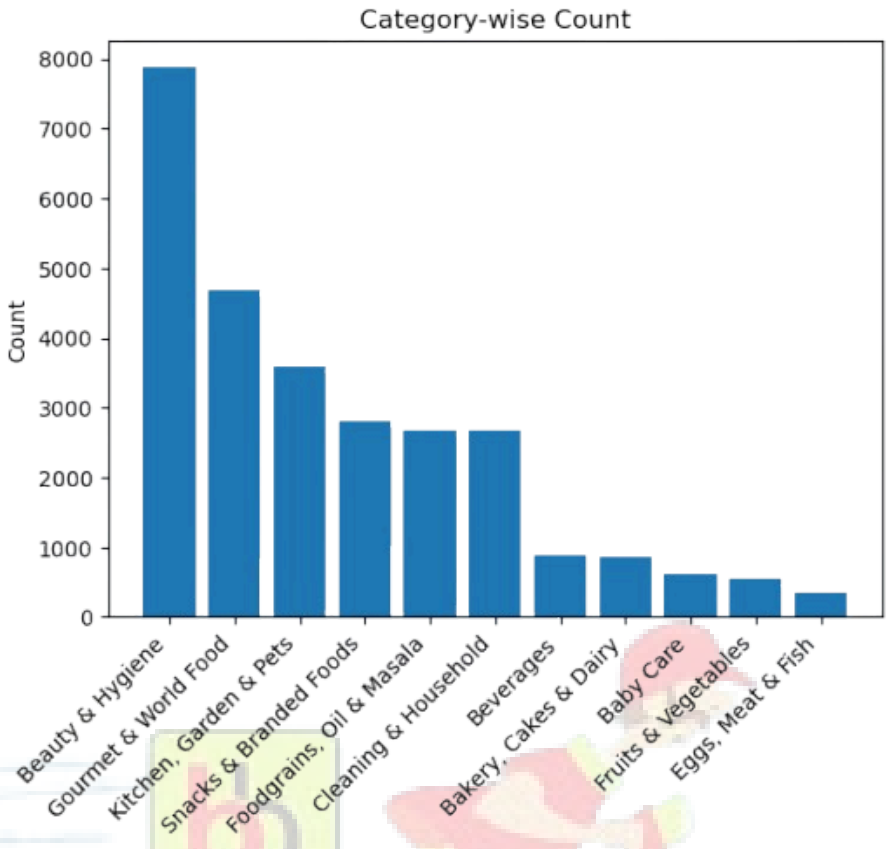
In this step, missing values in the BigBasket dataset are identified using the isnull().sum() function. The results show missing entries in columns such as rating, description, discount, product, brand, and sale_price. To handle these, numerical columns like rating and discount are filled with their respective mean values, while the description column is filled with the label "Unknown". After applying these methods, the dataset no longer contains missing values. This data cleaning step is essential to ensure accuracy, consistency, and reliability in further analysis and visualization.

Outliers



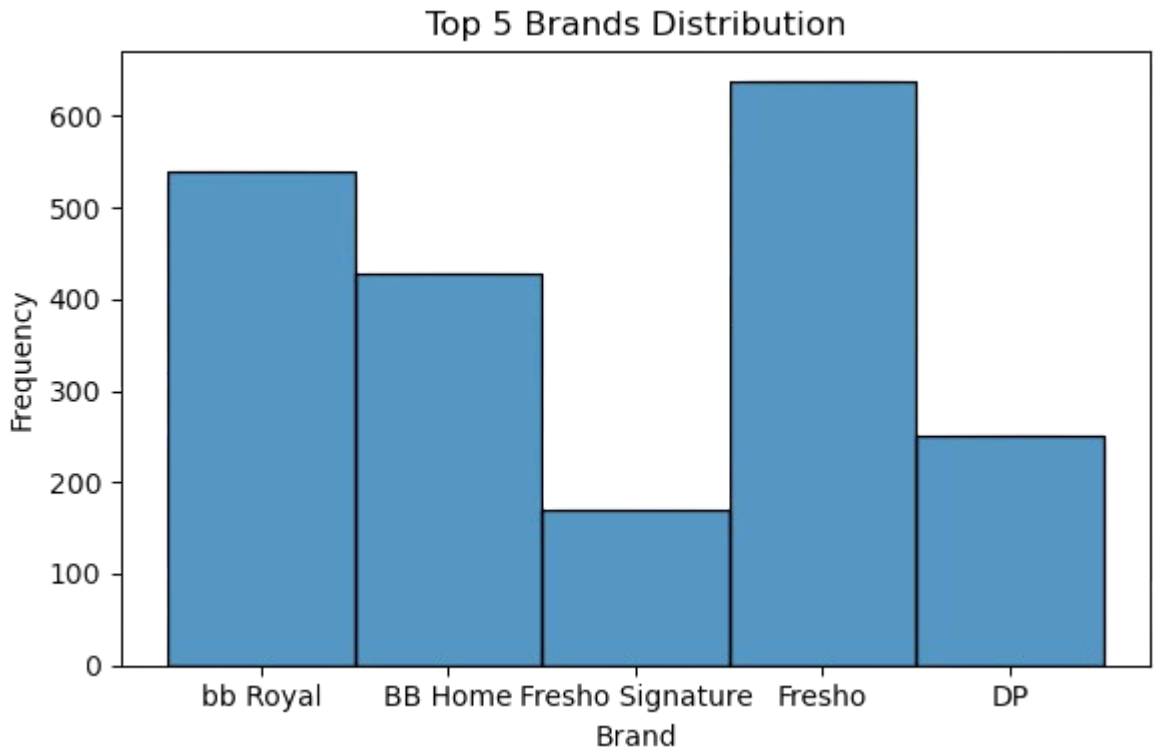
The box plot represents the distribution of sale prices of products in the BigBasket dataset. It shows the median price, interquartile range, and spread of the data. Most product prices lie within a moderate range, while a few high-priced items appear as outliers. This visualization helps identify price variation, detect extreme values, and understand overall pricing behavior, which is useful for pricing analysis and outlier detection in the dataset.

Visualization(Category Wise Counts)



This bar chart shows the distribution of products across different categories in the BigBasket dataset.
Categories like Beauty & Hygiene, Gourmet & World Food, and Kitchen, Garden & Pets have the highest number of products, indicating a wide variety in these segments.
On the other hand, categories such as Eggs, Meat & Fish and Fruits & Vegetables have fewer products.
This analysis helps understand product diversity and focus areas across categories.

Top 5 Brands Distribution

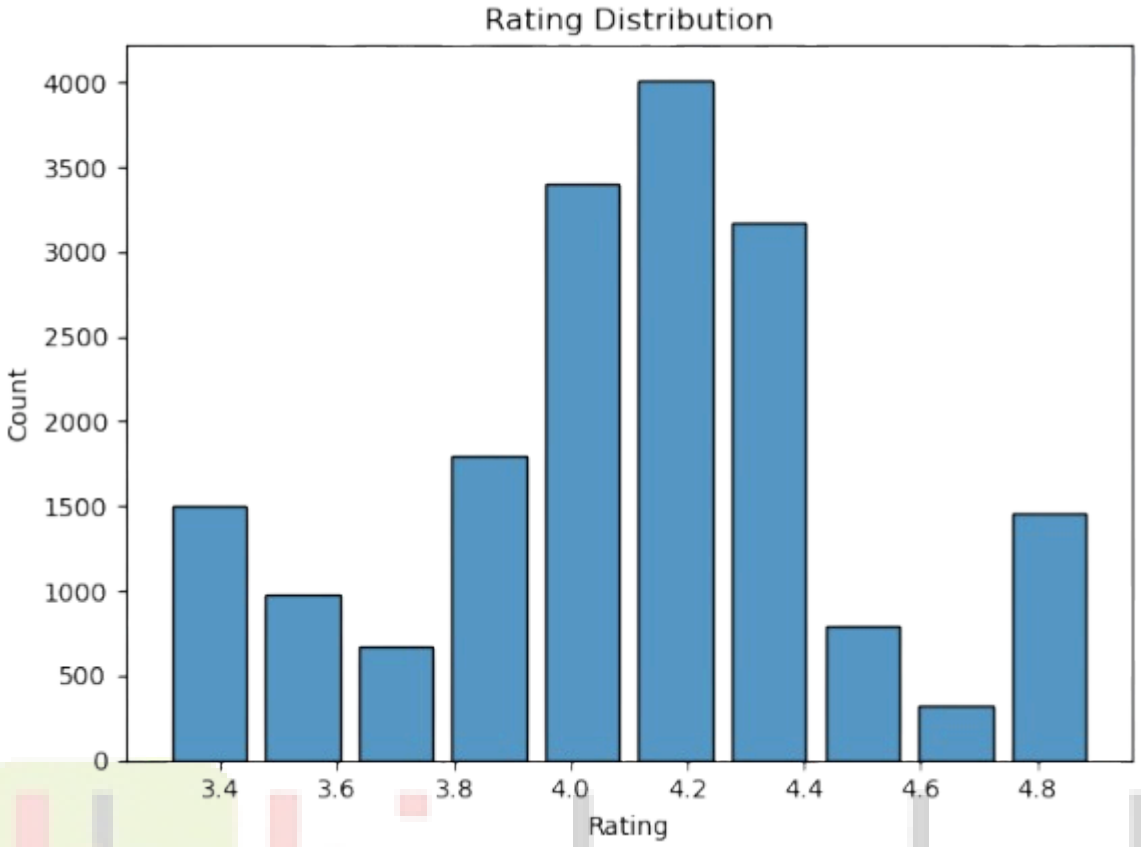


The chart shows the distribution of products among the top five brands on BigBasket. Brands like Fresho and bb Royal have the highest number of products, showing strong brand presence and wide product variety. Other brands such as BB Home, Fresho Signature, and DP have fewer products, indicating a more limited range. This analysis helps understand brand dominance, product availability, and BigBasket's focus on promoting its popular in-house brands.

Maket Price Vs Sales Price

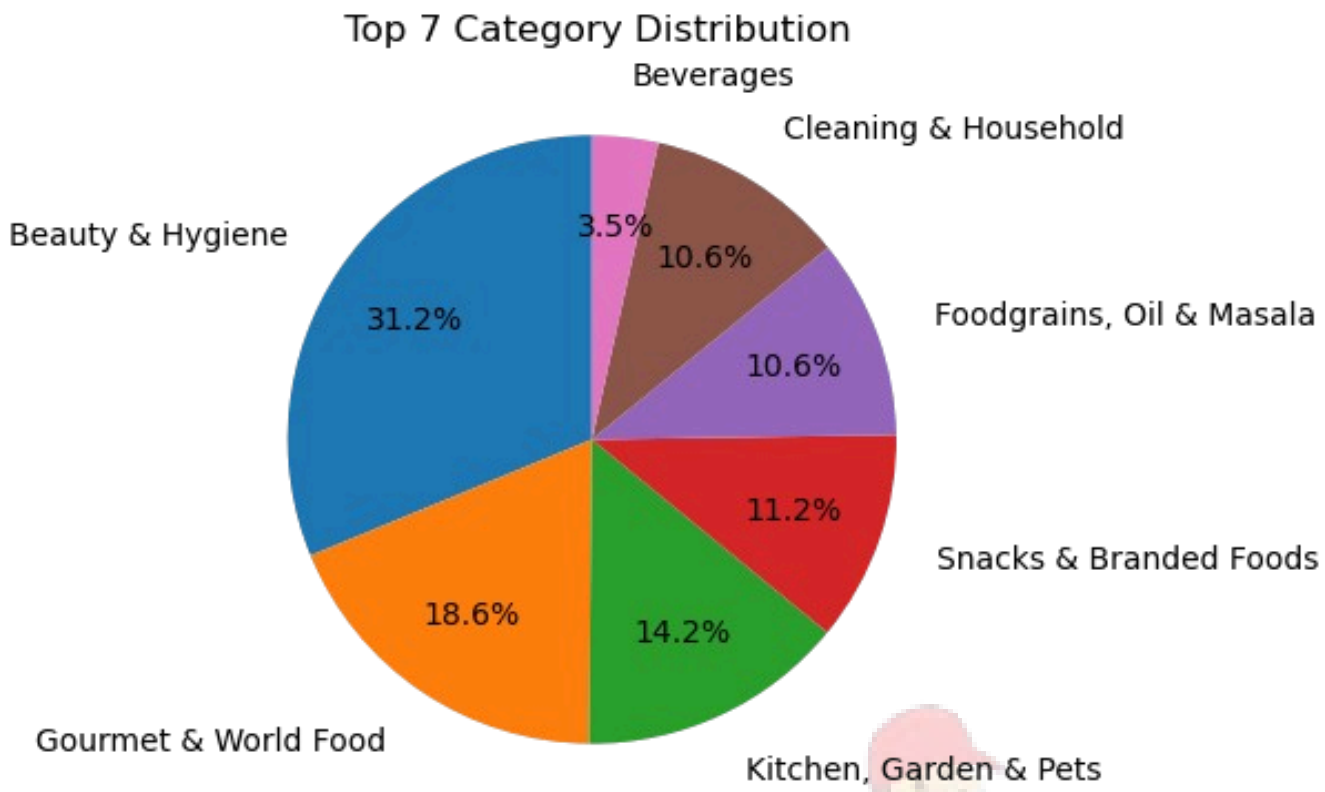
This scatter plot illustrates the relationship between market price and sale price of products in the BigBasket dataset. The points show a clear upward trend, indicating that as the market price increases, the sale price also increases. Most products are clustered along a diagonal line, suggesting consistent pricing with limited variation. Some points form horizontal and upper clusters, which may indicate fixed pricing, capped discounts, or premium products. Overall, the plot helps in understanding pricing patterns, discount limits, and the correlation between market and sale prices across different products.

Rating Distribution



This histogram shows the distribution of customer ratings for products in the BigBasket dataset. Most ratings are concentrated between 4.0 and 4.5, indicating that the majority of products receive positive feedback from customers. Fewer products fall in the lower rating range, while very high ratings (above 4.6) are also less frequent. Overall, this distribution suggests high customer satisfaction with most products available on the platform and helps identify general trends in product quality and customer experience.

Top 7 Category Distributio



This pie chart shows the percentage distribution of the top 7 product categories in the BigBasket dataset. Beauty & Hygiene holds the largest share, indicating a wide range of products in this category. It is followed by Gourmet & World Food and Kitchen, Garden & Pets, which also contribute significantly. Categories like Snacks & Branded Foods, Foodgrains, Oil & Masala, and Cleaning & Household have moderate shares, while Beverages has the smallest portion. This visualization helps understand category-wise product focus and inventory distribution on the platform.

Final Findings & Insights

- The dataset contains 27,000+ products, showing a wide and diverse product range.
- Beauty & Hygiene is the largest category, followed by Gourmet & World Food and Kitchen, Garden & Pets.
- Categories like Beverages and Eggs, Meat & Fish have fewer products, indicating selective inventory.
- Fresho and bb Royal are the top brands, highlighting BigBasket's strong private-label strategy.
- Most products are priced within a moderate range, making them affordable for customers.
- A small number of high-priced products act as outliers and represent premium items.
- There is a strong positive relationship between market price and sale price, showing consistent pricing.
- Discounts are applied selectively, mainly for promotional and marketing purposes.
- Customer ratings are mostly between 4.0 and 4.5, indicating high customer satisfaction.
- Low-rated products are limited, reflecting good overall product quality.
- Missing values and outliers were handled to improve data accuracy and reliability.
- The analysis shows how data-driven insights help understand customer preferences, pricing strategies, and product performance in the online grocery marke

Conclusion

This project successfully analyzed the BigBasket product dataset to understand product distribution, pricing strategies, discounts, brand presence, and customer ratings. The analysis shows that BigBasket offers a wide variety of products across multiple categories, with a strong focus on private-label brands such as Fresho and bb Royal. Most products are priced affordably, and discounts are used strategically to attract customers. Customer ratings indicate high overall satisfaction, reflecting good product quality and service. Data cleaning and outlier handling improved the reliability of results. Overall, the project highlights the importance of data analysis in understanding customer behavior, improving business strategies, and supporting informed decision-making in the online grocery industry.

Thank You

bigbasket

Python File Link