BigBasket

India’s largest online supermarket

EXPLATORY DATA ANALYSIS

INTODUCTION:

- BigBasket is one of India's leading online grocery delivery services, offering a wide range of products, including fresh fruits, vegetables, dairy, packaged foods, personal care items, and household essentials. Founded in 2011, the company has grown rapidly, providing customers with a convenient and efficient way to shop for groceries from the comfort of their homes.

- BigBasket operates in multiple cities across India and has expanded its services through quick delivery options, scheduled deliveries, and express grocery delivery under its "BB Now" service. The platform also offers its own private-label products and works with local farmers and suppliers to ensure quality and affordability.

- In 2021, Tata Group acquired a majority stake in BigBasket, further strengthening its position in India's e-commerce and retail sector. The company continues to innovate with AI-driven inventory management, seamless app-based shopping, and partnerships with various brands to enhance the customer experience.

Objective:

1) Understanding Sales Trends:

- Analyze sales data to determine best-selling and least-selling products.
- Identify patterns in customer preferences over time.
- Use visualization tools (e.g., bar charts, line graphs) for better insights.

2) Pricing & Discounts Analysis:

- Calculate discount percentages on different products.
- Compare original vs. discounted prices to assess impact.
- Identify trends in discount effectiveness on sales performance.

3) Brand Performance Analysis:

- Assess brand popularity based on sales volume and revenue.
- Analyse customer ratings and reviews for sentiment insights.
- Compare competing brands within similar categories.

4) Category-Wise Product Distribution:

- Count the number of products in each category.
- Identify overrepresented or underrepresented product categories.
- Visualize data using pie charts or bar graphs for distribution clarity.

5) Data Cleaning & Preparation:

- Handle missing values (e.g., imputation or removal).
- Detect and treat outliers affecting analysis.
- Standardize data formats for consistency across all records.

IMPORTING PYTHON LIBRARY:

Pandas --> A powerful data manipulation and analysis library in Python. It provides data structures likeDataFrames and Series to handle structured data efficiently. It is widely used for data cleaning, transformation, and analysis.

Numpy –> A fundamental package for numerical computing in Python. It supports multi-dimensional arrays andmathematical functions for fast numerical computations. Essential for scientific computing and data analysis.

Seaborn –> A visualization library built on matplotlib, designed for statistical data visualization. It provides beautiful and informative graphs, such as histograms, box plots, and correlation heatmaps, making it easier to explore data patterns.

Matplotlib.pyplot –> A core plotting library in Python, allowing users to create a variety of static, animated, and interactive visualizations, such as line charts, bar charts, and scatter plots.

plotly.express –> A high-level interface for interactive and dynamic visualizations. It is particularly useful for creating visually appealing charts with zooming, hovering, and filtering capabilities.

warnings(warnings.filterwarnings("ignore")) –> The warnings module helps manage warning messages in Python. The filterwarnings("ignore") function suppresses unnecessary warnings, improving readability when running scripts.

READ THE DATASET

pd.read_csv() – This function loads a CSV file into a Pandas DataFrame.

Read the dataset and use head to see the first 13 row of the dataset.

```
df = pd.read_csv('/content/drive/MyDrive/BigBasket Products (1).csv')
```

```
df.head(12)
```

	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

✓ 0s completed at 9:47 PM

INFORMATION ABOUT THE DATASET:

As you can see that the dataset contains the **27,555 rows** and **10 columns**.

The dataset consumes **2.1 MB** of memory.

Index (int64): A unique identifier for each row.

product (object): Name of the product (1 missing value).

category (object): The main category of the product.

Sub_category (object): A more detailed classification within the category.

brand (object): The brand of the product (1 missing value).

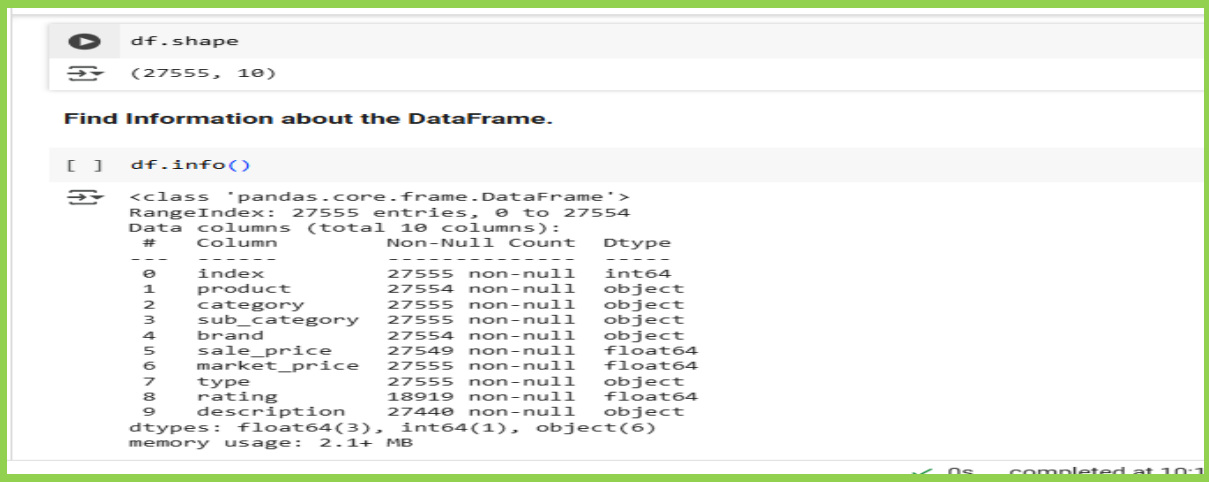
sale_price (float64): The discounted price (6 missing values).

market_price (float64): The original price before discount.

type (object): Likely indicates the type of product.

rating (float64): Customer rating (significant missing values).

description (object): Product description (115 missing values).



```
df.shape
(27555, 10)
```

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
```

DESCRIPTION ABOUT THE DATASET

Describe() function is used to description about the data.

Like the mean, median, sum, count, standard deviation, min, max of the numerical columns.

```
df.describe()
```

	index	sale_price	market_price
count	27555.00000	27555.000000	27555.000000
mean	13778.00000	334.600168	382.056664
std	7954.58767	1201.976472	581.730717
min	1.00000	2.450000	3.000000
25%	6889.50000	95.000000	100.000000
50%	13778.00000	190.010000	220.000000
75%	20666.50000	359.000000	425.000000
max	27555.00000	112475.000000	12500.000000

MISSING VALUES FIND OUT:

Is.null() is used to check the missing values in all the columns.

Is.null().sum() is used to give sum of all the missing columns in the data.

In Both the picture you can see the sum of all the null value present in the columns and also shows how much percentage of null values in the columns.

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

	0
index	0
product	1
category	0
sub_category	0
brand	1
sale_price	6
market_price	0
type	0
rating	8636
description	115

dtype: int64

```
df.isnull().sum()/df.shape[0]*100
```

	0
index	0.000000
product	0.003629
category	0.000000
sub_category	0.000000
brand	0.003629
sale_price	0.021775
market_price	0.000000
type	0.000000
rating	31.340954
description	0.417347

dtype: float64

HANDLING THE OUTLIERS:

Using IQR (Inter quantile range) method to detect the outliers.

- q1=df[col].quantile(0.25)
- q3=df[col].quantile(0.75)
- iqr = q3-q1
- lw = q1 - 1.5*iqr
- uw = q3 + 1.5*iqr

q1,q2(quantile1,quantile2),

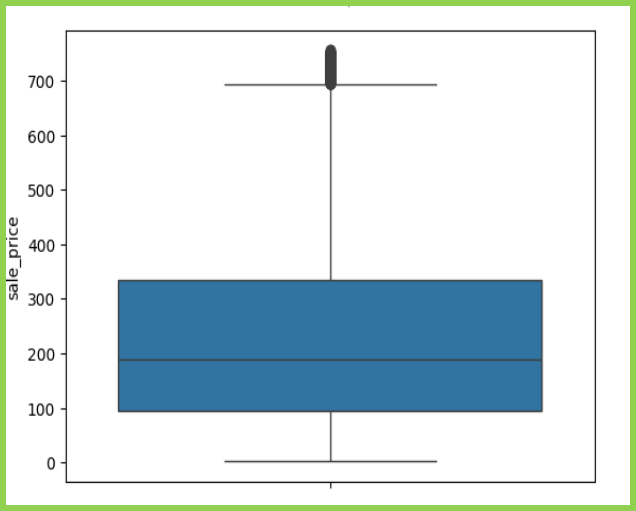
lw,uw (lower whisker, upper whisker)

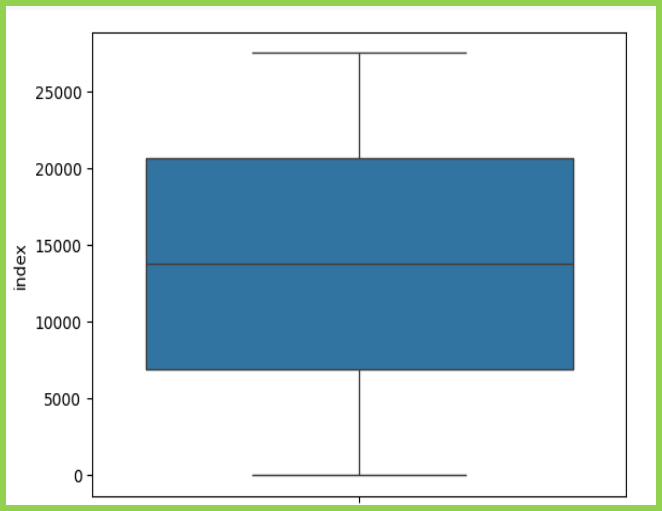
Handle & Fill the outliers by Mean

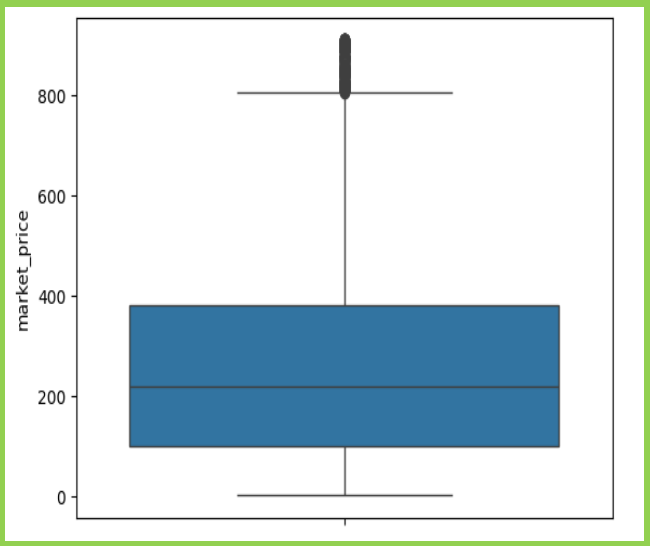
```
columns = ['sale_price','market_price','rating']
def outliers(df,columns):
  df = df.copy()
  for col in columns:
    q1=df[col].quantile(0.25)
    q3=df[col].quantile(0.75)
    iqr = q3-q1
    lw = q1 - 1.5*iqr
    uw = q3 + 1.5*iqr

    print(f'Lower_bound for {col}: {lw}')
    print(f'Upper_bound for {col}: {uw}')

    median_value = df[col].median()
    print(f'median of {col}: {median_value}')

    df[col] = np.where((df[col] < lw) | (df[col] > uw), median_value, df[col])
  return df


df_new = outliers(df,columns)
df_new
```

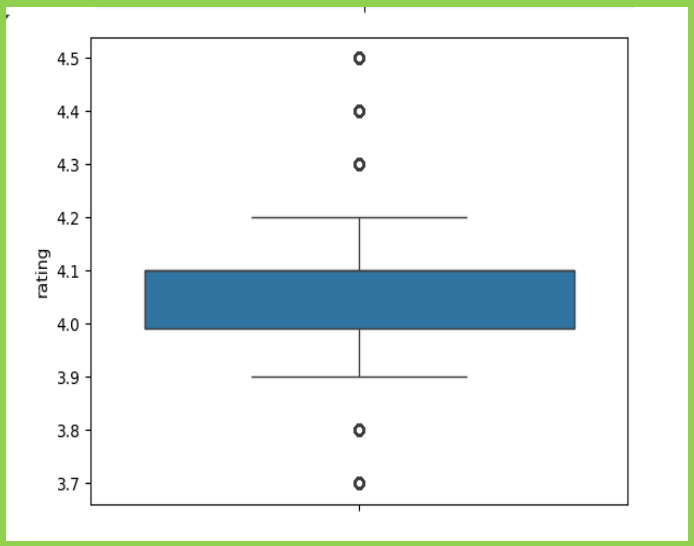
HANDLING THE OUTLIERS:

```
for i in df_new.select_dtypes(include='number'):
  sns.boxplot(y=df_new[i])
  plt.show()
```









ANALYSIS:

Task: Find out Top & least sold products.

High-value products like **Beard Kit** and **Olive Oil - Extra Virgin** contribute significantly to total revenue.

Some low-cost or niche products, such as **Serum** and **Hand Wash - Moisture Shield**, have minimal sales.

This analysis helps in identifying the best and worst-performing products based on total sales revenue, which can be useful for inventory management, marketing strategies, and sales optimization.

Task 1:-- Find out the top & least sold product

```
top_sold_products = df.groupby('product')['sale_price'].sum().sort_values(ascending=False).head(5)
print(top_sold_products)
```

```
product
Beard Kit                                          112475.00
4mm Aluminium Induction Base Chapati Roti Tawa - Silver    112178.00
Balloon - Polka Dot, 12 Inch                        88899.00
Extra Virgin Olive Oil                              24808.53
Olive Oil - Extra Virgin                            22568.22
Name: sale_price, dtype: float64
```

```
least_sold_products = df.groupby('product')['sale_price'].sum().sort_values(ascending=True).head(5)
print(least_sold_products)
```

```
product
Serum                                      3.0
Polo - The Mint With The Hole              5.0
Orbit Sugar-Free Chewing Gum - Lemon & Lime    5.0
Sugar Coated Chocolate                     5.0
Hand Wash - Moisture Shield                5.0
Name: sale_price, dtype: float64
```

Task: Measuring discount on a certain item.

Calculating Discount Percentage: The formula used to compute the discount percentage is:

Discount_percentage = (Market price – Sale price) / Market price *100

This formula determines the percentage difference between the market price and the sale price for each product.

Task 2 : Measuring discount on a certain item.

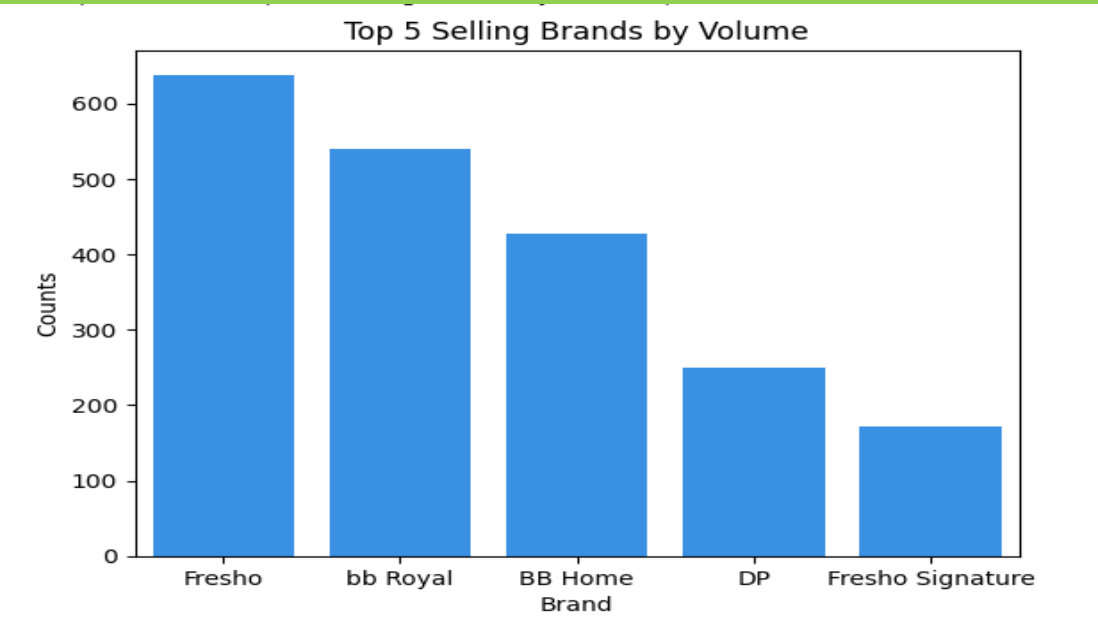
```
df['discount_percentage'] = ((df['market_price'] - df['sale_price']) / df['market_price']) * 100

df[['product', 'market_price', 'sale_price', 'discount_percentage']].head(10)
```

	product	market_price	sale_price	discount_percentage
0	Garlic Oil - Vegetarian Capsule 500 mg	220.0	220.0	0.000000
1	Water Bottle - Orange	180.0	180.0	0.000000
2	Brass Angle Deep - Plain, No.2	250.0	119.0	52.400000
3	Cereal Flip Lid Container/Storage Jar - Assort...	176.0	149.0	15.340909
4	Creme Soft Soap - For Hands & Body	162.0	162.0	0.000000
5	Germ - Removal Multipurpose Wipes	199.0	169.0	15.075377
6	Multani Mati	58.0	58.0	0.000000
7	Hand Sanitizer - 70% Alcohol Base	250.0	250.0	0.000000
8	Biotin & Collagen Volumizing Hair Shampoo + Bi...	1098.0	1098.0	0.000000
9	Scrub Pad - Anti- Bacterial, Regular	20.0	20.0	0.000000

Visualizations the graph

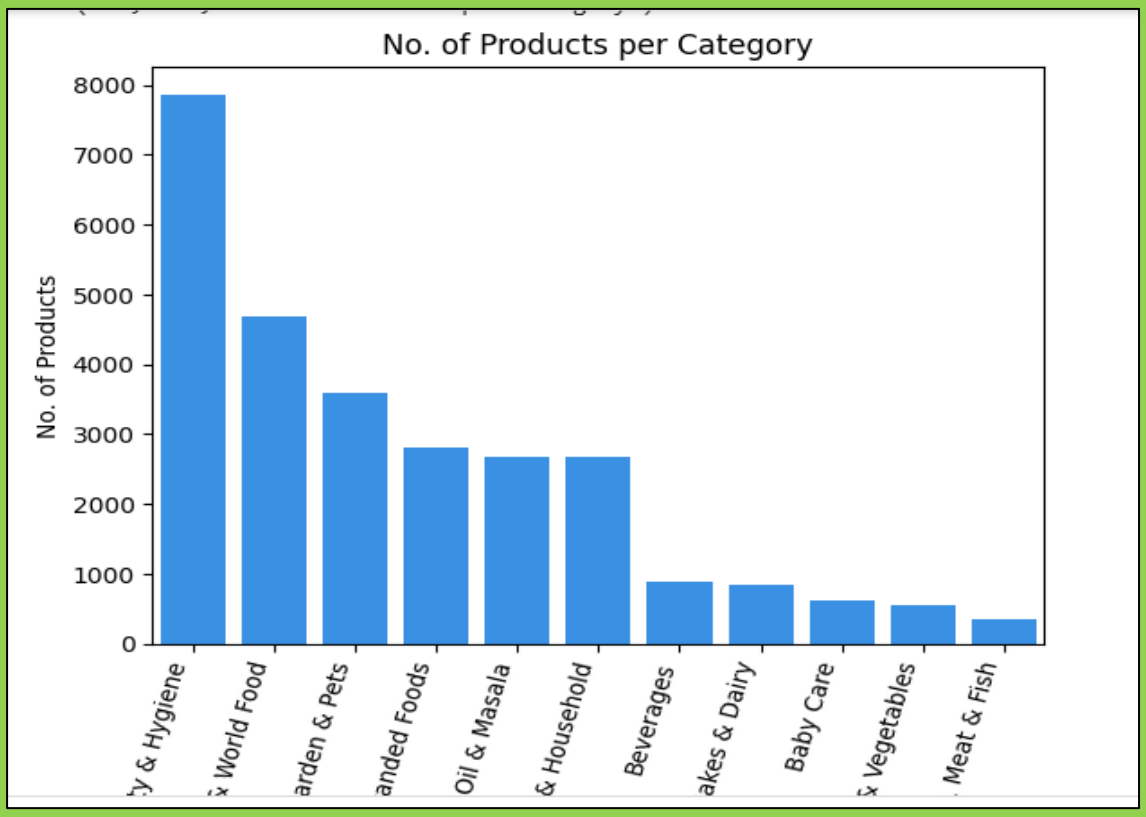
Create Plots of first 5 brand with the Sale price.

- A bar chart displaying the top 5 selling brands based on the number of products sold.
- The brand with the highest sales volume will have the tallest bar.
- The visualization helps identify the most popular brands in the dataset.



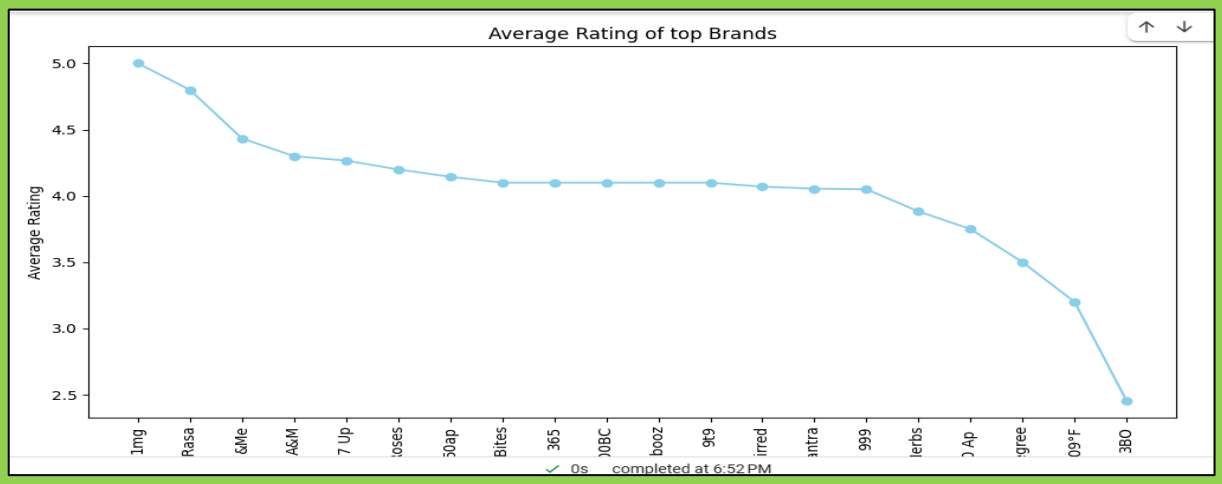
Number of Products Per Category.

- A **bar chart** displaying the number of products in each category.
- Categories with a higher number of products will have taller bars.
- The x-axis labels are rotated to avoid overlap and improve readability



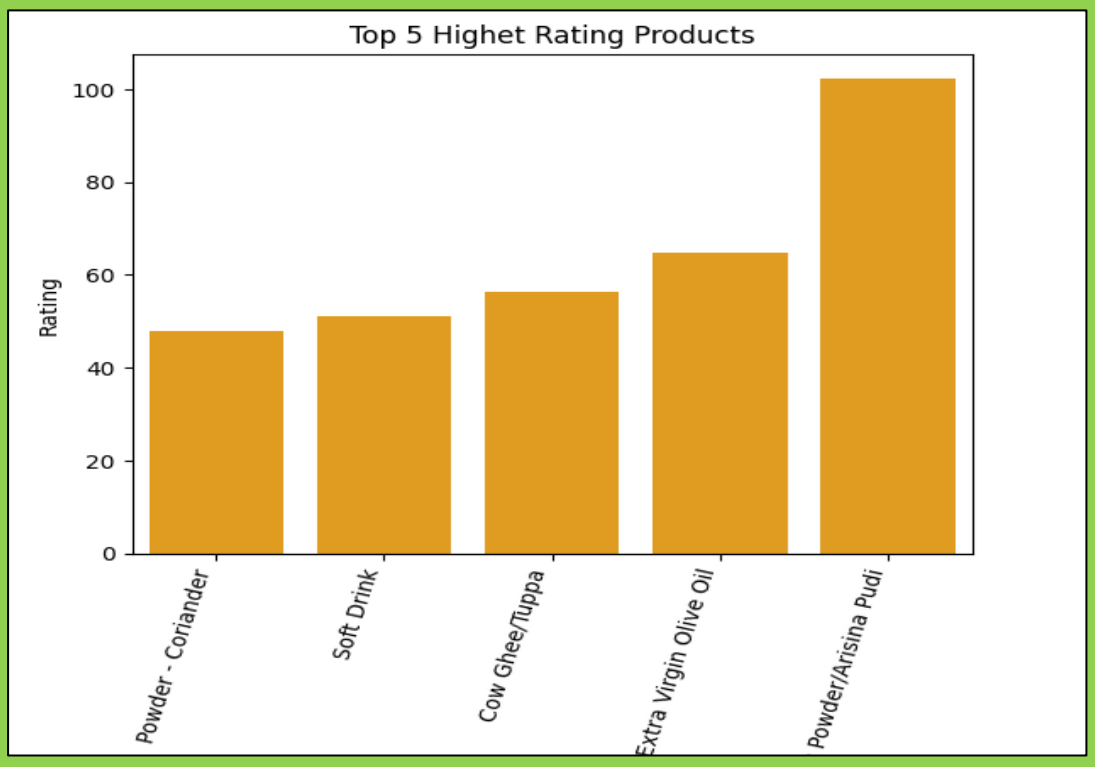
Average rating of Top Brands.

- A **line plot** displaying the average rating of the **top 20 brands**.
- Brands with **higher average ratings** will appear towards the top.
- The x-axis will show brand names, and the y-axis will show their corresponding average ratings.
- It helps identify which brands have the highest customer satisfaction based on ratings.
- It Useful for brand performance evaluation and market positioning.



Top Five Highest Rating Products

- In the graph we can see the Top 5 rating products:
- **Turmeric Powder/Arisina Pudi** is the most well-received product, possibly due to high customer satisfaction.
- Other products like **Extra Virgin Olive Oil** and **Cow Ghee/Tuppa** also have strong ratings, indicating customer preference for health-related items.
- **Soft Drink** making it into the top 5 suggests that it is a popular choice among consumers.

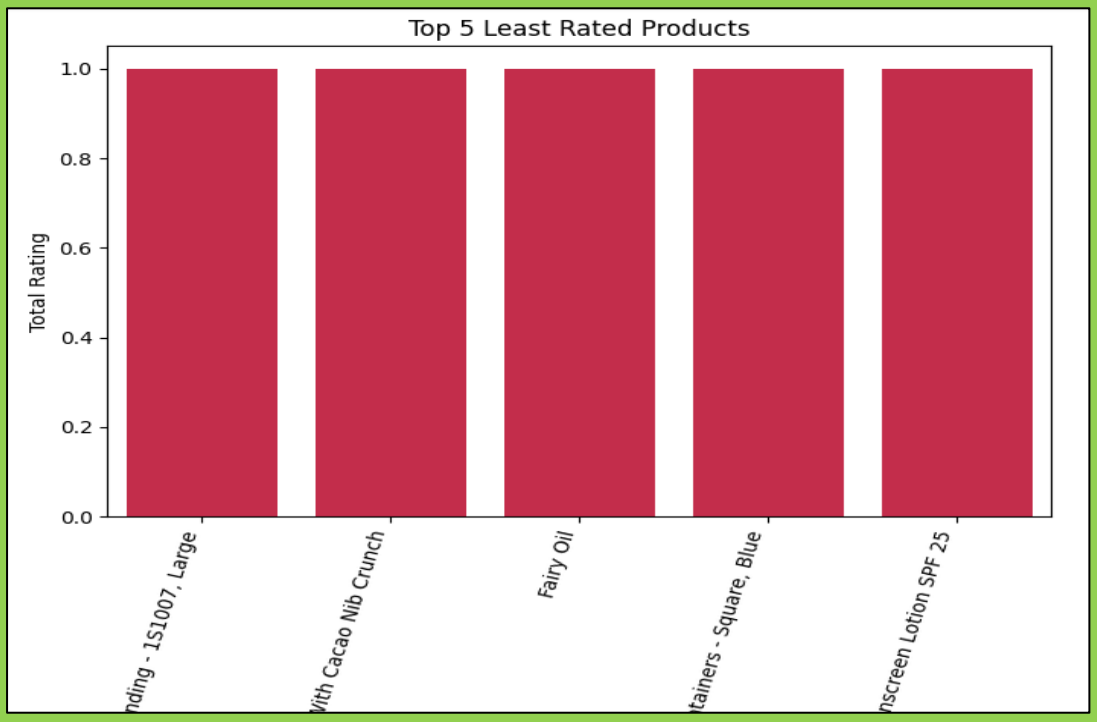


Top Five Least Rating Products

In this graph we can shown the top 5 least rating product.

Each of the five products has a **total rating of just 1.0**, suggesting minimal customer engagement.

This shows the lack of customers reviews rather than poor quality.



THANK YOU

PROJECT LINK :
https://colab.research.google.com/drive/13fsiiobww05zL2F1e1p7N7qrNpZQuclo