BigBasket Data Analysis Project

Presented by: Shubham_shivyi

Date: 02 June 2025

SUBTITLE: A COMPREHENSIVE E-COMMERCE PRODUCT DATA STUDY

Introduction

About the Project:

This project focuses on analyzing a product dataset from **BigBasket**, one of India's
largest online grocery platforms. The goal is to uncover patterns in pricing, brand
presence, product categories, and customer ratings.

Objectives:

- Clean and preprocess real-world e-commerce data
- Explore trends in product pricing and popularity
- Visualize insights using Python tools like Pandas, Seaborn, and Plotly
- Identify outliers and understand market dynamics

Data Load Link

Df = pd.read_csv ("/content/BigBasketProducts(1sv”)

Data Overview

- Columns: Product, Category, Brand, Sale Price, Market Price, Rating, Description

- Missing Values: Found in product, brand, sale_price, rating, and description

- Cleaning performed using .fillna() and manual imputation

Methods Used

- **df.head()**

Displays the first 5 rows of the dataset. Helps in getting a quick look at the structure and sample data.

- **df.shape**

Returns the dimensions of the DataFrame — (rows, columns).

- **df.info()**

Gives a summary of the dataset including column names, non-null counts, and data types.

- **df.describe()**

Provides statistical summary (mean, min, max, quartiles) for numerical columns.

- **df.isnull().sum()**

Shows the total number of missing (null) values in each column.

- **df.loc()**

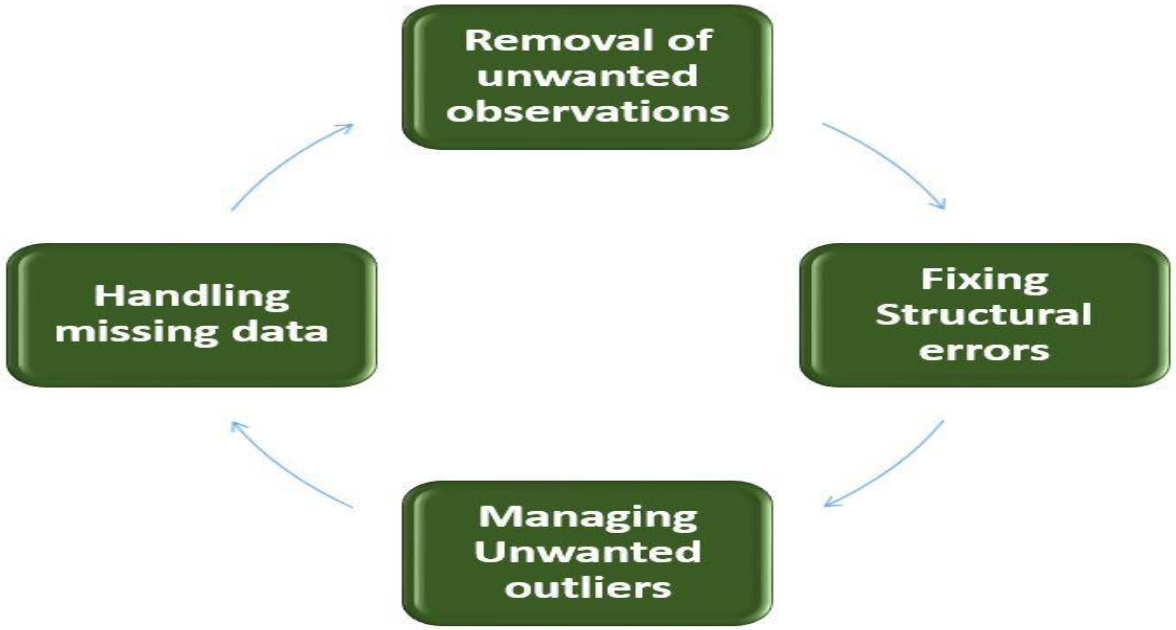
Used to access and modify specific rows and columns in the DataFrame by label/index.

- **groupby()**

Groups the data based on one or more columns for aggregation or analysis.

- **fillna()**

Used to fill missing values with a specific value (like mean, median, or mode).



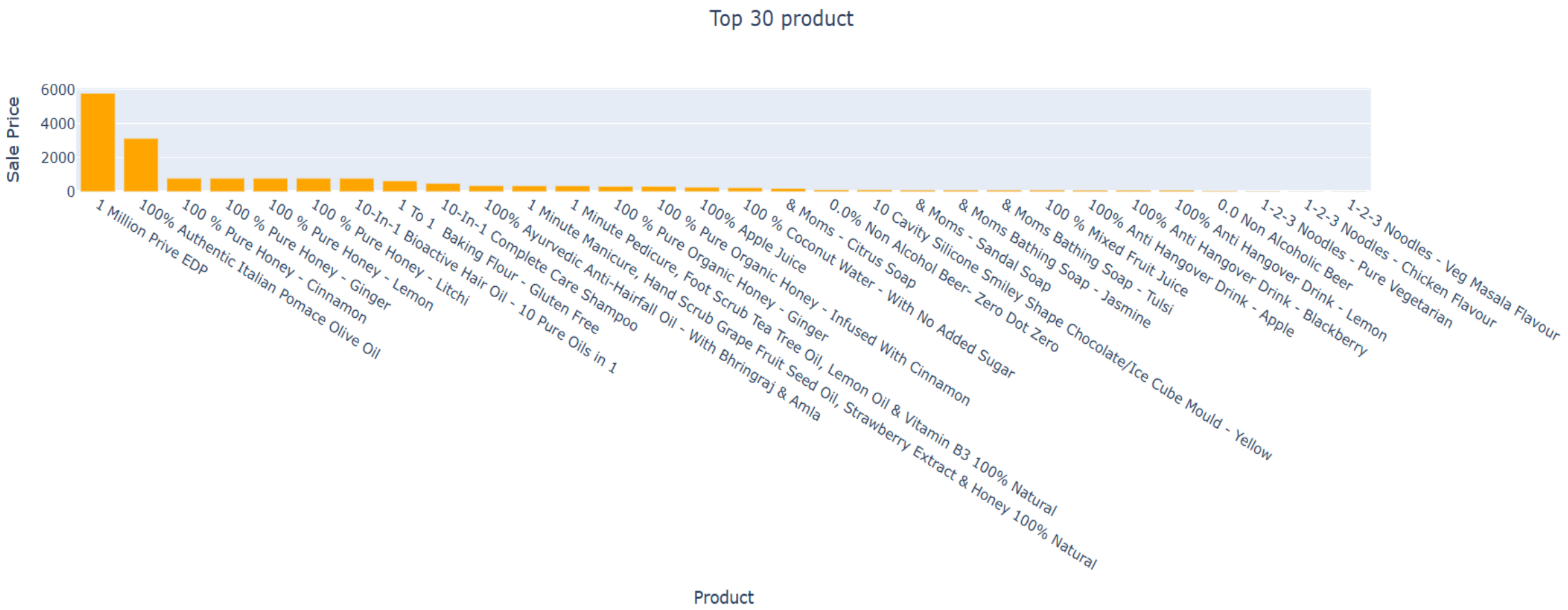
Data Cleaning Steps

- Filled missing product and brand values manually
- Filled missing sale_price with logical assumptions
- Filled rating with mode value (most common)
- Replaced missing description with “No Description”

Top 30 Products by Sale Price

Bar graph showing highest priced products

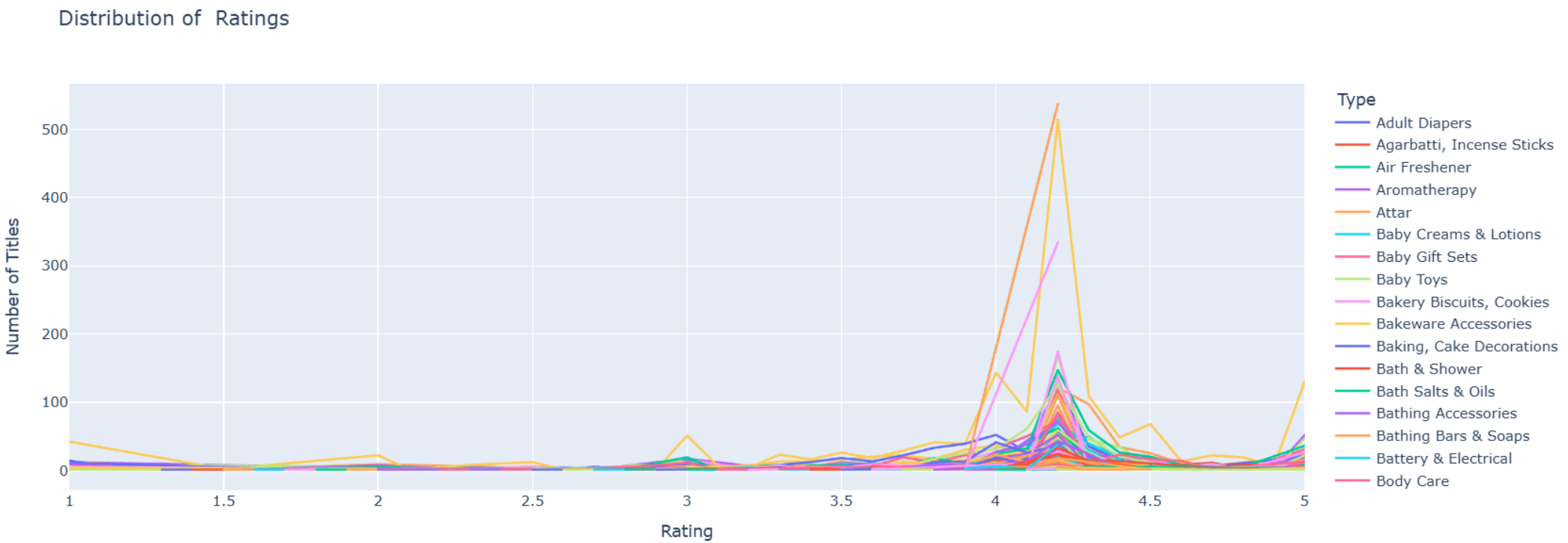
Notable insight: Some items priced extremely high—possible luxury products or errors.



Ratings Distribution by Type

Line plot shows how product types are distributed across different rating levels

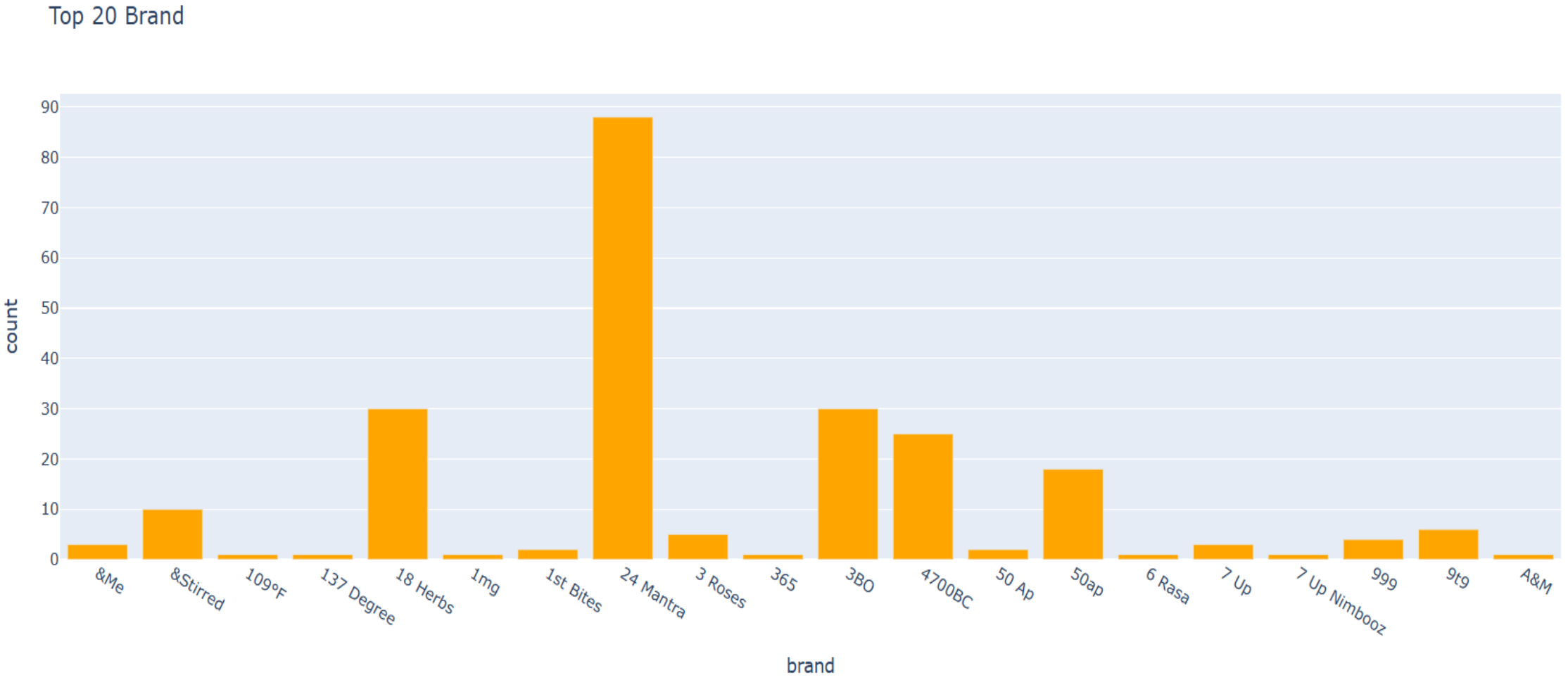
Ratings mostly range between 3.5 and 5



Top 20 Brands

Bar graph showing brands with most products listed

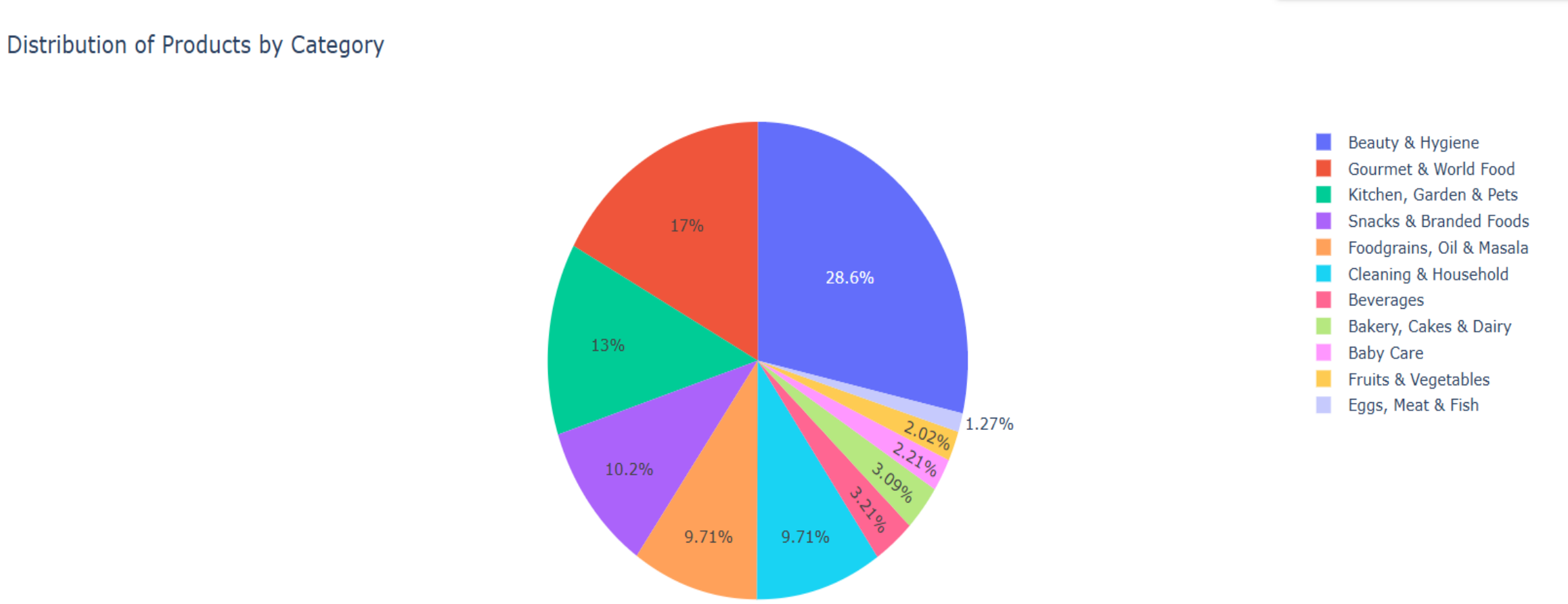
Examples: BB Home, 365, Nivea, etc.



Product Categories

Pie chart showing product distribution across categories

Top categories: Beauty & Hygiene, Gourmet Food, Household items



1 Data Size:
the dataset contains 27,555 product entries.

2 Sale Price Overview:
.Mean Sale Price : 334.56 Rupees
.Meadin (50%) :190.01 Rupees
.Min-Max: 2.45 to 112,475
The extremely high max sugests possible outliers or luxury/incorrectly entere products.

3 Market Prjice Overview
.Mean Market Price : 382.06 Rupees
.Median (50%) : 220 Rupees
.Min-Max : 3 to 12,500 Rupees
.Also shows a wide range, again indicating outliers or diverse product categories

4 Insights:
. Many product are priced relatively low (25% of products have sale price below rupees 95).
.A few high-priced items are signigicantly inflating the average.
.it might be useful to analyze and possibly remove or separately handle outliers for batter insights.

Conclusion

Key Findings from the Analysis:

- **Improved Data Quality:**

Missing and inconsistent data were cleaned and filled, making the dataset reliable for analysis.

- **Price Insights:**

Most products are low-priced, but a few very high-priced items indicate outliers or premium listings.

- **Ratings Overview:**

The majority of products have positive ratings (mostly 4 and above), reflecting customer satisfaction.

- **Brand & Category Trends:**

Some brands like BB Home and 365 are dominant, and categories like *Beauty & Hygiene* are most common.

Business Insights:

- BigBasket has a vast and diverse product range.
- Data-driven decisions can enhance pricing strategy, brand focus, and customer targeting.

Future Scope:

- Use of modeling and machine learning for deeper customer behavior analysis.
- Trend analysis over time for sales forecasting.
- Quality control based on product ratings and reviews.