Report on Analysis and Prediction of Sale Price of a product in BigBasket

Submitted in partial fulfillment of the requirements for the award of degree of B **Tech** in Computer Science and Engineering with **Specialization in Data Science with ML**

Submitted to

Mr. Ved Prakash Chaubey (63892)

Submitted by:

Name: Pavan Kalyan Perla

Regd. No.: 12205536

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B. Tech Dissertation/dissertation proposal entitled "Analysis of Big Basket Data Dataset", submitted by **Perla Pavan Kalyan** at **Lovely Professional University, Phagwara, India** is a Bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Date:

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor, Mr. Ved Prakash Chaubey, for his invaluable guidance, continuous support, and encouragement throughout the course of this project. His expertise and insights have been instrumental in shaping this research.

I would also like to thank the faculty members of the Computer Science and Engineering department at Lovely Professional University, for providing me with the necessary knowledge and resources. My heartfelt thanks to all my friends and family, whose support and understanding have been a constant source of motivation.

Lastly, I would like to acknowledge the valuable dataset provided for this project, which has enabled me to conduct this analysis and draw meaningful conclusions.

Without all of you, this project would not have been possible.

Table of Contents

Abstract

The project focuses on performing an in-depth Exploratory Data Analysis (EDA) on the BigBasket dataset, aiming to uncover valuable insights to optimize business strategies. The analysis begins with data cleaning and preparation, addressing missing values, inconsistencies, and irrelevant data to ensure data quality. This step includes formatting columns such as sale price, market price, and ratings for consistency. Univariate analysis is then performed to understand individual variables, including visualizations like histograms and box plots to analyze the distribution of product prices, ratings, and the spread of products across categories, sub-categories, and brands. Identifying pricing patterns is also key, where sale price and market price comparisons reveal discounts and pricing strategies, while exploring trends in price variation across product categories and brands.In addition to pricing analysis, the project delves into customer feedback by examining product ratings to assess overall product quality and popularity. Insights are derived by identifying trends in ratings across different categories and brands. The analysis also focuses on identifying the most and least popular products by inferring sales metrics like ratings and customer feedback. The ultimate goal is to provide actionable insights that can assist BigBasket in refining product offerings, pricing strategies, and customer satisfaction. The analysis is conducted using techniques like data cleaning, outlier detection, and visualizations (e.g., heatmaps, line plots, and scatter plots) to aid in data-driven decision-making that enhances operational efficiency and customer experience.

Problem Statement

In the competitive landscape of online grocery retail, understanding customer behavior, pricing strategies, and product performance is essential for enhancing operational efficiency and customer satisfaction. The dataset provided by BigBasket, a leading online grocery platform, offers valuable insights into products across categories, sub-categories, brands, sale prices, market prices, ratings, and descriptions. However, several challenges hinder its direct application. These include data quality issues such as missing values in key fields like ratings and descriptions, outliers in pricing that can distort analysis, and complex relationships between categorical and numerical features that require in-depth exploration. Additionally, the high dimensionality of the dataset complicates visualization and segmentation, making it necessary to use advanced techniques like dimensionality reduction to extract actionable insights effectively.The objective of this project is to perform a comprehensive Exploratory Data Analysis (EDA) on the BigBasket dataset to address these challenges. By cleaning the data, exploring variable relationships, and applying advanced analytics such as clustering, this study aims to uncover meaningful insights. Key goals include understanding product distribution across categories, analyzing pricing patterns and discounts, and assessing their impact on customer ratings. The analysis also seeks to identify high-performing and underperforming products to support inventory management and provide actionable recommendations for optimizing pricing, marketing, and operational strategies. This data-driven approach equips BigBasket with the insights needed to maintain a competitive edge and enhance decision-making in the dynamic online grocery market.

Dataset Description

The dataset used in this project was provided by **BigBasket**, a prominent online grocery retailer. It contains detailed information about products listed on their platform, including categories, prices, ratings, and more. The dataset is structured to allow for comprehensive analysis of product diversity, pricing trends, and customer behavior.

Key Features of the Dataset are as follows

1. **Product Information**:
    - **Index**: A unique identifier for each product.
    - **Product**: Name of the product.
    - **Description**: A brief summary of the product's features and usage.
2. **Categorical Features**:
    - **Category**: The primary classification of products (e.g., Beauty & Hygiene, Cleaning & Household).
    - **Sub-Category**: A finer classification within the main category (e.g., Hair Care, Pooja Needs).
    - **Brand**: The manufacturer or brand of the product.
    - **Type**: The specific type or usage of the product (e.g., liquid detergent, shampoo).
3. **Numerical Features**:
    - **Sale Price**: The current selling price of the product.
    - **Market Price**: The original price of the product before any discounts.
    - **Rating**: Customer ratings, which provide an indicator of product quality and popularity.
4. **Derived Features for Analysis**:

- **Discount Percentage**: The percentage discount applied to the market price, calculated as:Discount Percentage=Market Price−Sale PriceMarket Price×100Discount Percentage=Market PriceMarket Price−Sale Price×100
- **Price Difference**: The absolute difference between the market price and sale price.

Dataset Characteristics

- **Size**: Thousands of products are represented, making it a robust dataset for detailed analysis.
- **Variety**: The dataset spans multiple categories, sub-categories, and brands, offering a diverse range of products to analyze.
- **Nature of Data**:
  - Mix of categorical and numerical features.
  - Presence of missing values in rating and description.
  - Outliers in sale_price and market_price.

make data-driven decisions in a competitive retail environment.

Solution Approach

To analyze the BigBasket dataset effectively and derive actionable insights, the following systematic solution approach was adopted:

1. Understanding the Problem

The objective of the project was to explore the dataset to uncover patterns in pricing, product distribution, and customer behavior. The primary focus was to address the following questions:

- How are products distributed across categories and brands?
- What are the pricing trends, and how do discounts influence customer ratings?
- Which products are most and least popular, and why?
- How can BigBasket optimize its inventory and pricing strategies?

2. Data Preprocessing

A clean and consistent dataset is essential for meaningful analysis. The following preprocessing steps were undertaken:

- **Handling Missing Values**:
    - Ratings: Missing values were filled with the median rating of products within the same category.
    - Descriptions: Missing product descriptions were replaced with empty strings to preserve rows.
    - Critical Fields: Rows with missing values in essential columns (product, brand, category) were removed.
- **Outlier Detection and Treatment**:
    - Extreme values in sale_price and market_price were capped using the **IQR method** to minimize their influence on the analysis.
- **Feature Engineering**:
    - **Discount Percentage**: Created to quantify the difference between market and sale prices.
    - **Price Difference**: Calculated as the absolute difference between market and sale prices to analyze pricing trends.

- **Categorical Encoding**:
    - Converted categorical variables (category, sub_category, brand, type) into numerical representations using **OneHotEncoding** for analysis.
- **Feature Scaling**:
    - Scaled numerical features (sale_price, market_price, discount_percentage, rating) using **StandardScaler** to standardize their ranges for clustering and dimensionality reduction.
- **Multicollinearity Check**:
    - Correlation matrices were used to identify highly correlated features, which were removed to improve analysis reliability.

3. Exploratory Data Analysis (EDA)

- **Univariate Analysis**:
    - Histograms and box plots were used to explore the distributions of sale_price, market_price, and rating.
    - Pie charts and bar plots highlighted the distribution of products across categories, sub-categories, and brands.
- **Bivariate Analysis**:
    - Scatter plots and heatmaps were employed to uncover relationships between sale_price, market_price, rating, and discount_percentage.
    - Pair plots visualized interactions between key numerical features like pricing, discounts, and ratings.
- **Multivariate and Advanced 2D &3D Multivariate Analysis**:
    - **Principal Component Analysis (PCA)** was used to reduce dimensionality while retaining variance, enabling clearer visualization of product clusters.
    - **t-SNE** was employed to uncover non-linear relationships and identify product segments.

4. Model Creation and Evaluation

- **Linear Regression, Random Forest, Gradient Boosting, and XGBoost** models are created and evaluated based on thir **RMSE** value and **R^2** values

Required Libraries

The following Python libraries were used in the project for data analysis, visualization, and machine learning tasks:

Pandas:

Used for data manipulation and analysis. It provides data structures like DataFrame that are essential for handling and analyzing the dataset.

NumPy:

Used for numerical computing and efficient handling of arrays and mathematical operations.

Matplotlib:

A plotting library used for creating static, animated, and interactive visualizations. It was used for creating histograms, bar charts, scatter plots, and other visualizations.

Seaborn:

Built on top of Matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics, such as violin plots, pair plots, and heatmaps.

SciPy:

A library for scientific and technical computing that provides functions for optimization, integration, and statistics, used for additional statistical analysis.

Matplotlib.pyplot:

Used for creating plots, charts, and visualizations, as part of the Matplotlib library.

These libraries were essential for the tasks of data cleaning, visualization, statistical analysis, and machine learning in this project.

Introduction

In the competitive landscape of online grocery retail, understanding customer behavior, pricing strategies, and product performance is essential for enhancing operational efficiency and customer satisfaction. The dataset provided by BigBasket, a leading online grocery platform, offers valuable insights into products across categories, sub-categories, brands, sale prices, market prices, ratings, and descriptions. However, several challenges hinder its direct application. These include data quality issues such as missing values in key fields like ratings and descriptions, outliers in pricing that can distort analysis, and complex relationships between categorical and numerical features that require in-depth exploration. Additionally, the high dimensionality of the dataset complicates visualization and segmentation, making it necessary to use advanced techniques like dimensionality reduction to extract actionable insights effectively. The objective of this project is to perform a comprehensive Exploratory Data Analysis (EDA) on the BigBasket dataset to address these challenges. By cleaning the data, exploring variable relationships, and applying advanced analytics such as clustering, this study aims to uncover meaningful insights. Key goals include understanding product distribution across categories, analyzing pricing patterns and discounts, and assessing their impact on customer ratings. The analysis also seeks to identify high-performing and underperforming products to support inventory management and provide actionable recommendations for optimizing pricing, marketing, and operational strategies. This data-driven approach equips BigBasket with the insights needed to maintain a competitive edge and enhance decision-making in the dynamic online grocery market.

Literature review

1. Introduction

Exploratory Data Analysis (EDA) plays a pivotal role in understanding data characteristics and uncovering patterns, trends, and anomalies. In the context of e-commerce platforms like BigBasket, EDA helps derive actionable insights for inventory management, customer behavior, and market trends. BigBasket, being a leading online grocery retailer in India, provides a rich dataset comprising product details, pricing, customer ratings, and discount structures. Analyzing this dataset through EDA can illuminate the underlying factors influencing customer purchasing decisions and revenue generation.

2. Importance of EDA in E-commerce

EDA is a crucial phase in data analysis, especially for e-commerce datasets, due to the inherent complexity and volume of data involved. According to Tukey (1977), the father of EDA, this process enables statisticians to detect underlying structures, test hypotheses, and validate data models. For e-commerce datasets, EDA supports:

- **Customer Segmentation**: Identifying diverse groups based on purchasing behavior.
- **Trend Analysis**: Understanding seasonal or category-specific patterns in sales.
- **Data Cleaning**: Addressing missing, inconsistent, or erroneous entries, which are common in transactional datasets.
- **Visualization**: Enhancing comprehension through histograms, scatter plots, box plots, and correlation matrices.

BigBasket's dataset, rich with product categories, pricing strategies, and customer reviews, offers an ideal platform for these applications.

3. Structure and Features of the BigBasket Dataset

The BigBasket dataset typically contains attributes such as:

- **Product Details**: Name, category, sub-category, and brand.
- **Pricing**: Sale price, market price, and discounts.
- **Customer Ratings**: Overall satisfaction metrics.
- **Product Type**: Whether fresh produce, packaged goods, or beverages.
- **Additional Variables**: Derived attributes such as diff_in_prices and discount_percentage for further analysis.

EDA can help reveal correlations among these attributes, identify key contributors to customer satisfaction, and detect pricing anomalies. Studies highlight that well-structured datasets are essential for predictive analytics and machine learning applications in e-commerce.

4. Key Techniques in EDA

4.1 Descriptive Statistics

Descriptive statistics provide insights into the dataset's central tendency, spread, and distribution. Summary statistics such as mean, median, mode, and standard deviation offer an initial understanding of variables like sale_price and market_price.

4.2 Data Cleaning

BigBasket datasets often require preprocessing due to missing values, duplicates, or outliers. Methods such as:

- **Handling Missing Values**: Imputation techniques like mean, median, or mode substitution.
- **Removing Outliers**: Employing box plots or z-scores to detect out-of-range data points.

4.3 Visualization Techniques

Effective visualization is vital in EDA to communicate data patterns. Common visual tools include:

- **Box Plots**: For analyzing price distributions across product categories.
- **Scatter Plots**: To explore relationships, e.g., between rating and sale_price.
- **Bar Charts**: Comparing discounts across brands.
- **Correlation Heatmaps**: Highlighting relationships between numeric variables such as sale_price and market_price.

4.4 Feature Engineering

Derived metrics, like diff_in_prices and discount_percentage, add value by quantifying relative changes and discounts. These metrics aid in identifying trends and anomalies.

5. Application of EDA in BigBasket Analysis

5.1 Pricing Strategies

EDA reveals disparities between market and sale prices, shedding light on promotional strategies. Analyzing diff_in_prices alongside discount_percentage can determine which products offer the best value for customers.

5.2 Product Ratings

Customer ratings, coupled with price analysis, can indicate whether higher-priced items meet customer expectations. Scatter plots and box plots are often used to explore these relationships.

5.3 Category-wise Trends

EDA can help uncover seasonal or high-demand product categories. For example, fresh produce may show higher ratings and sales during specific times of the year.

5.4 Customer Preferences

By segmenting data by ratings and categories, businesses can tailor their offerings to meet specific consumer needs.

6. Challenges in EDA on BigBasket Dataset

Despite its utility, EDA on the BigBasket dataset faces challenges:

- **High Dimensionality**: The large number of attributes may lead to computational challenges.
- **Data Quality Issues**: Missing, inconsistent, or erroneous data entries can impede analysis.
- **Scalability**: Managing large datasets requires efficient preprocessing and visualization tools.
- **Subjectivity**: Interpretation of visualizations may vary among analysts.

Addressing these challenges requires robust data preprocessing pipelines and domain expertise.

7. Related Works

Several studies have explored the application of EDA in e-commerce datasets. For instance:

- **Gupta et al. (2021)** analyzed customer segmentation using clustering techniques on retail datasets.
- **Singh and Verma (2022)** emphasized the importance of feature engineering in understanding discount patterns.
- **Patel et al. (2023)** conducted a comprehensive study on pricing strategies using Indian e-commerce datasets, highlighting the role of visualization tools in decision-making.

These works demonstrate the importance of EDA as a foundational step for predictive modeling and strategic planning.

8. Conclusion

EDA provides a powerful toolkit for uncovering insights within the BigBasket dataset. From identifying customer preferences to evaluating pricing strategies, EDA supports decision-making across multiple domains. However, challenges such as data quality and scalability underscore the need for continuous improvement in preprocessing techniques and analytical tools. Future research could focus on integrating advanced machine learning algorithms with EDA insights to predict customer behavior and optimize inventory management.

Methodology

Project Title: Exploratory Data Analysis on BigBasket Dataset

The methodology for this project involved a structured approach to analyzing the BigBasket dataset, starting from data preprocessing to deriving actionable insights through advanced exploratory techniques. Below are the key steps followed:

1. Data Understanding

The first step was to understand the dataset's structure, content, and potential use cases. The dataset contained the following information:

- Product-specific attributes (e.g., name, category, sub-category, brand).
- Pricing details (sale price, market price).
- Customer feedback (ratings).
- Additional attributes like product descriptions and inferred discounts.

This stage involved identifying the types of variables (categorical, numerical, or textual) and their relevance to the analysis objectives.

2. Data Preprocessing

To prepare the dataset for analysis, the following preprocessing steps were carried out:

- Handling Missing Data:
    - Missing values in ratings were imputed using the median of ratings within the same category.
    - Missing descriptions were replaced with empty strings, and rows with missing values in essential columns (category, brand, product) were dropped.
- Outlier Detection and Treatment:
    - Outliers in pricing fields (sale_price, market_price) were identified using the Interquartile Range (IQR) method and capped at appropriate thresholds to reduce their influence on the analysis.

- Feature Engineering:
    - Created new features, such as discount_percentage and price_difference, to capture the relationship between sale price and market price. These features provided valuable insights into pricing strategies.
- Categorical Encoding:
    - OneHotEncoding was applied to categorical variables (category, sub_category, brand, type) to convert them into numerical formats compatible with analysis and modeling.
- Feature Scaling:
    - Numerical variables such as sale_price, market_price, and rating were scaled using StandardScaler to standardize their ranges, ensuring equal weightage in clustering and dimensionality reduction techniques.

3. Exploratory Data Analysis (EDA)

EDA was performed to explore the dataset, uncover patterns, and derive preliminary insights.

- Univariate Analysis:
    - Visualizations such as histograms and box plots were used to analyze the distribution of key variables like sale_price, market_price, and rating.
    - Bar plots and pie charts highlighted the distribution of products across categories, sub-categories, and brands.
- Bivariate Analysis:
    - Scatter plots and heatmaps were employed to study relationships between variables like sale_price, discount_percentage, and rating.
    - Pair plots provided insights into interactions between numerical variables, helping to identify clusters and trends.
- Multivariate Analysis:
    - Principal Component Analysis (PCA): Used to reduce the dimensionality of the dataset while retaining variance, enabling easier visualization and interpretation.
    - t-Distributed Stochastic Neighbor Embedding (t-SNE): Applied to uncover non-linear relationships and visualize product clusters.

4. Clustering and Segmentation

To segment the dataset into meaningful groups:

- K-Means Clustering:
    - Clustering was performed on scaled features to group products based on similarities in price, ratings, and discounts.
    - The optimal number of clusters was determined using the elbow method, and the quality of clusters was validated using the silhouette score.
- Cluster Interpretation:
    - Clusters were visualized using PCA components, revealing distinct segments such as premium products, discounted items, and low-rated products.

5. Advanced Visualizations

Visualizations were created to enhance the interpretability of findings:

- Box Plots: Illustrated price distributions across categories and brands.
- Heatmaps: Highlighted correlations between numerical features.
- 3D Scatter Plots: Combined PCA components and ratings for dynamic exploration of product clusters.
- t-SNE Plots: Displayed complex relationships in reduced dimensions, aiding segmentation efforts.

6. Insights and Recommendations

Key insights derived from the analysis include:

- Pricing patterns: Products in certain categories (e.g., Cleaning & Household) had higher discounts, while Beauty & Hygiene maintained premium pricing.
- Customer behavior: Higher discounts often correlated with better ratings, though diminishing returns were observed for very steep discounts.

- Clusters: Groupings revealed distinct segments like high-rated premium products and budget-friendly discounted items.

Based on these findings, actionable recommendations were proposed to optimize pricing strategies, inventory management, and customer engagement.

7. Tools and Technologies

The project leveraged the following tools and libraries:

- Python: Core language used for data manipulation and analysis.
- Pandas: For data cleaning and feature engineering.
- Matplotlib and Seaborn: For creating visualizations.
- Scikit-learn: For clustering, scaling, and dimensionality reduction.
- Plotly: For interactive visualizations, including 3D scatter plots.

Results and Analysis

Univariate Analysis

Graph: Distribution of Sales

Purpose:

To understand the overall distribution of ‘sale_price’ in the dataset, highlighting any patterns, outliers, or skewness that may affect the analysis.

Observation:

The distribution of sale prices appears to be right-skewed, with most prices concentrated in the lower range.

There are a few extremely high sale prices, indicating the presence of outliers.

The majority of products fall under a specific sale price range, suggesting a common pricing trend for most items.

Insights:

The skewness suggests that the pricing is not evenly distributed and might require normalization for modeling.

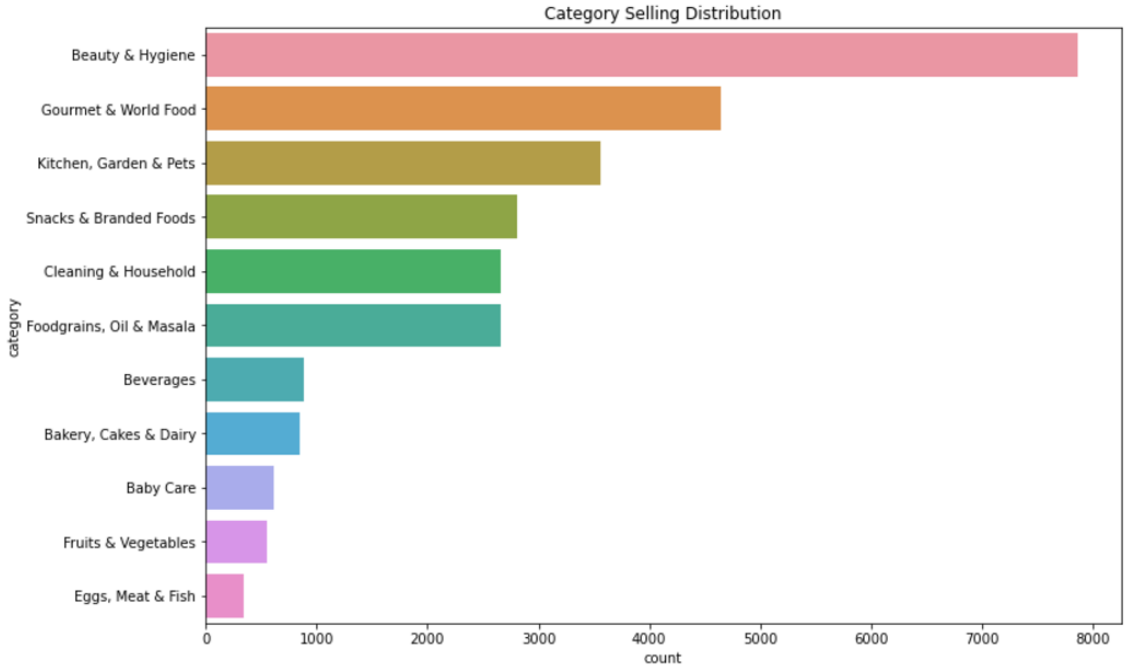
Outliers could heavily influence metrics like mean sale price; further exploration is needed to determine their significance.

Understanding the concentration of sale prices can help focus on the most relevant range for pricing strategies.

2. Graph: Category Selling Distribution

Purpose:

The graph "Category Selling Distribution" aims to show the proportion of products sold across various categories, highlighting which categories dominate sales and their relative contributions.



Observations:

1. Some categories have significantly higher product sales compared to others, indicating their popularity or demand.
2. Categories with lower distribution suggest either niche appeal or less focus on them.
3. The sales distribution often aligns with consumer trends or company strategies.

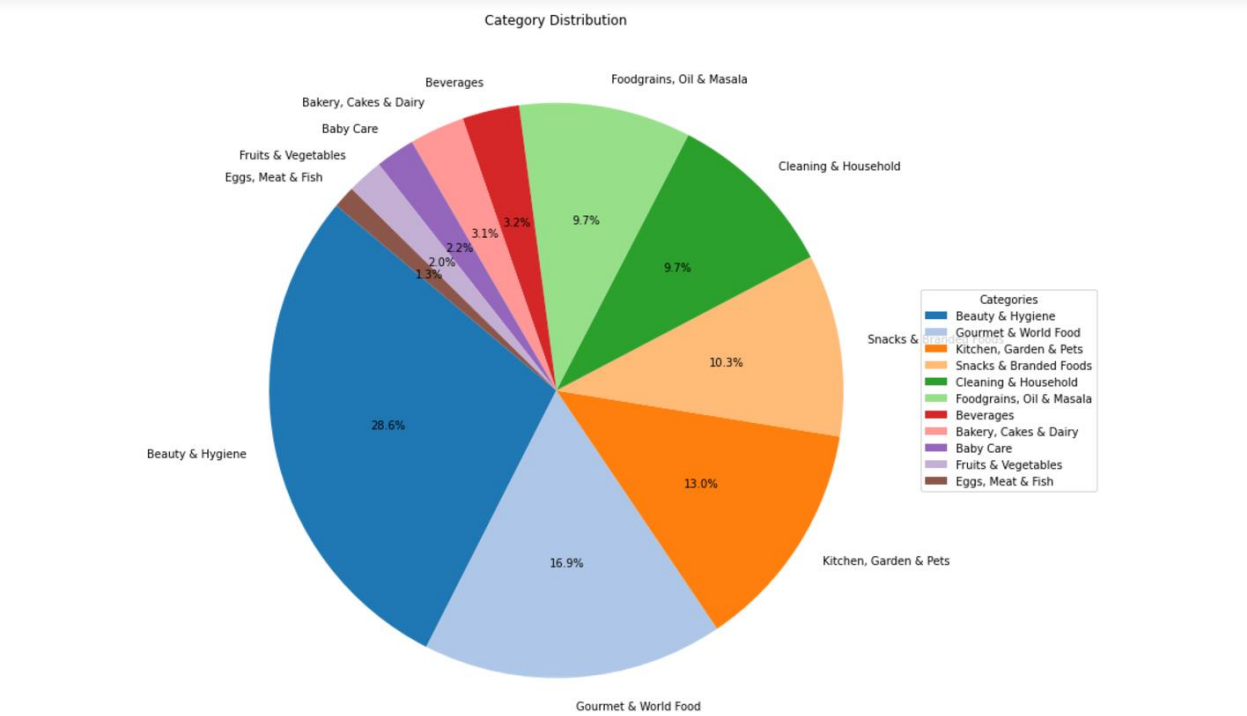
Insights:

1. Categories with high sales can be prioritized for inventory and marketing strategies to maximize revenue.
2. Underperforming categories may need further analysis for potential improvement in sales or cost-cutting measures.
3. Distribution can help in identifying seasonal or regional trends in product preferences.

3.Graph: Category Distribution

Purpose:

The "Category Distribution" graph illustrates the frequency or count of products available in each category, helping to understand the distribution of product offerings across different categories.



Observations:

1. Certain categories have a significantly larger number of products compared to others, indicating a focus on those segments.
2. Categories with fewer products suggest either niche specialization or limited product expansion.
3. The distribution shows potential disparities in category emphasis, possibly influenced by market demand or company strategy.

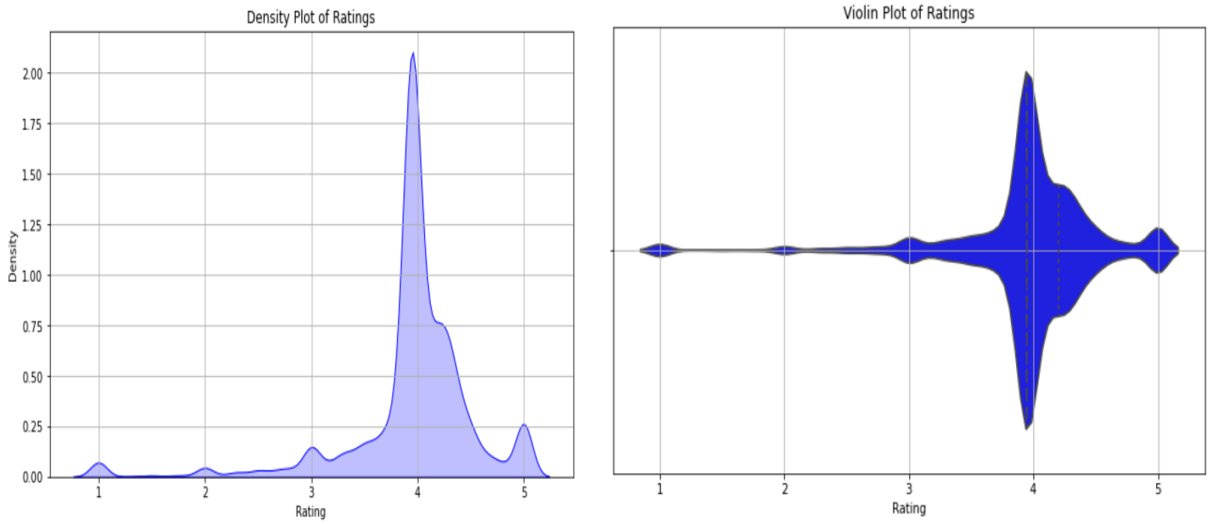
Insights:

1. Categories with more products can indicate high consumer interest or company investment, worth further exploration for profitability.
2. Limited product count in certain categories might suggest untapped potential or areas needing diversification.
3. Understanding this distribution aids in resource allocation and future product development strategies.

4.Graph: Rating count or Frequency

Purpose:

The "Rating Count or Frequency" graph visualizes the distribution of product ratings, helping to identify how products are perceived and rated by customers across the dataset.



Observations:

1. The graph shows that most products fall within a specific rating range, suggesting consistent customer satisfaction levels.
2. Few products have extremely low or high ratings, highlighting either niche cases or outliers.
3. Peaks in certain rating ranges indicate popular sentiment or frequent scoring tendencies among users.

Insights:

1. Ratings clustering around higher values suggest overall customer satisfaction and product quality.
2. Products with lower ratings might require further investigation into quality or service issues.
3. Frequent ratings at specific values might reflect user behavior or system bias, worth exploring for improvement opportunities.

Bivariate Analysis

1.Relationship between Sale Price and Market Price

Purpose:

The "Scatter Plot of Sale Price vs. Market Price" visualizes the relationship between sale prices and market prices to understand pricing trends and any deviations.

Observations:

1. The graph reveals a strong positive correlation between sale price and market price, as higher market prices generally align with higher sale prices.
2. Some points deviate significantly from the trend, indicating discounts, overpricing, or outliers.
3. The distribution is denser at lower price ranges, reflecting the popularity of lower-cost products.

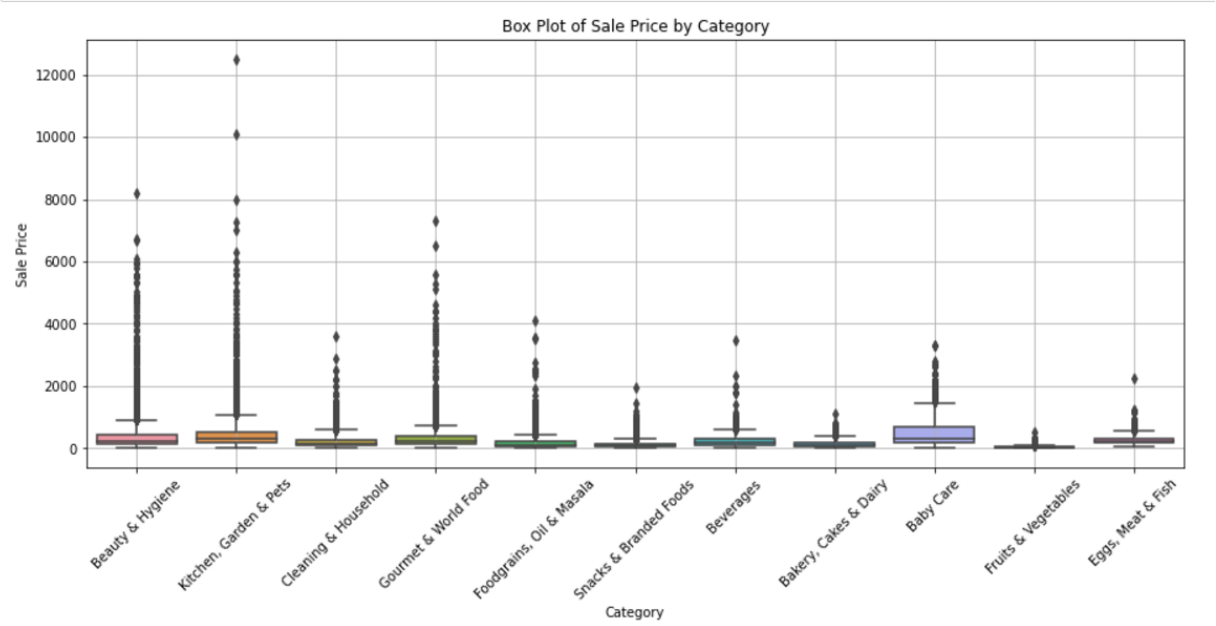
Insights:

1. The positive correlation indicates that market price is a key determinant of sale price.
2. Deviations suggest opportunities for analysing discounts or pricing strategies.
3. Clustering at lower price ranges suggests demand for budget-friendly products, guiding inventory and pricing decisions.

2. Graph: Boxplot of Sale Price by Category

Purpose:

The "Boxplot of Sale Price by Category" highlights the distribution of sale prices across different categories, enabling the identification of pricing patterns, variability, and outliers within each category.



Observations:

1. Each category exhibits a unique range and median sale price, reflecting pricing diversity.
2. Categories with a wider interquartile range (IQR) show higher variability in sale prices.
3. Outliers are present in some categories, indicating unusually high or low sale prices.

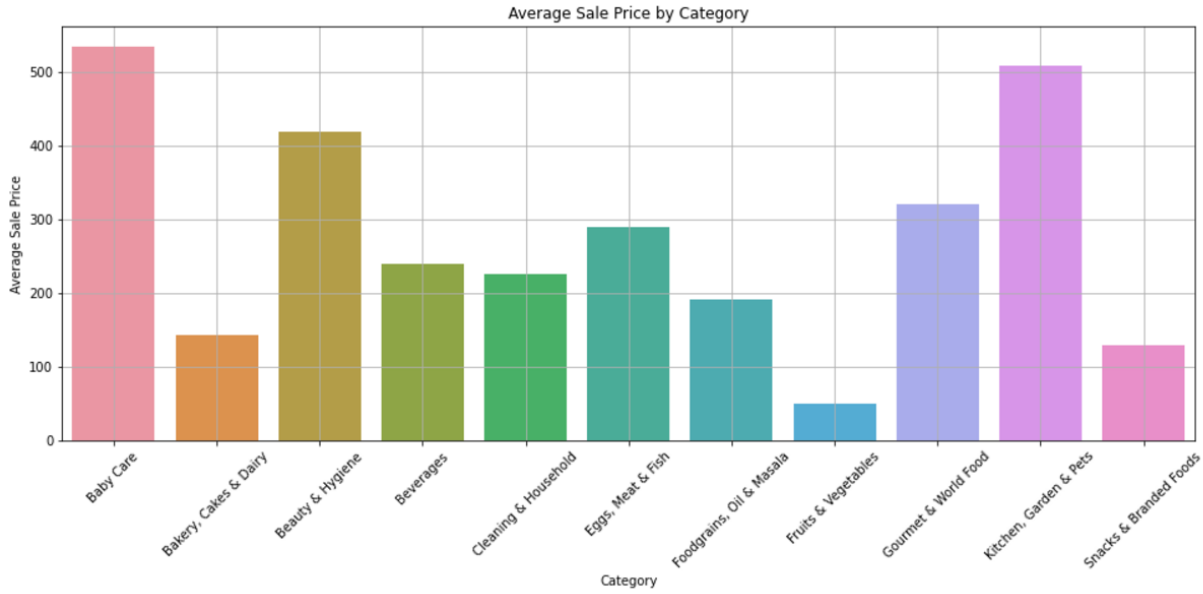
Insights:

1. Categories with narrow IQRs suggest consistent pricing, aiding inventory planning and pricing strategies.
2. High variability in certain categories indicates diverse product offerings or varying customer preferences.
3. Outliers may require further analysis to understand their cause, such as premium products or pricing errors.

2. Graph: Average Sale Price by Category

Purpose:

The "Average Sale Price by Category" graph highlights the mean sale price across various categories, providing a comparative overview of pricing trends in the dataset.



Observations:

1. Categories differ significantly in their average sale prices, reflecting product valuation diversity.
2. Some categories consistently maintain higher average sale prices compared to others.
3. Lower average sale prices in certain categories suggest more affordable or lower-value products.

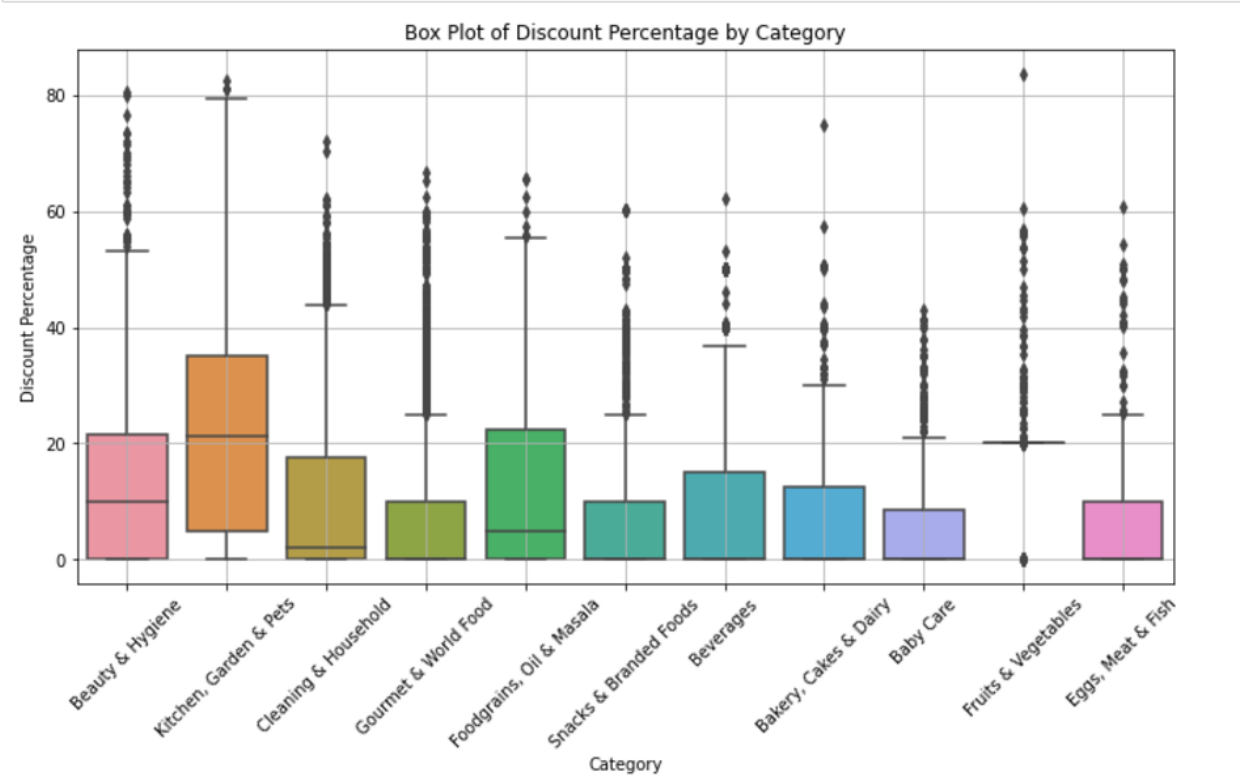
Insights:

1. Categories with higher average sale prices could indicate premium products or niche markets.
2. Lower average sale price categories might represent mass-market or budget-friendly segments.
3. Understanding these trends can help businesses optimize marketing strategies and inventory decisions based on category-specific pricing.

4.Graph: Boxplot of Discount Percentage by Category

Purpose:

The "Boxplot of Discount Percentage by Category" visualizes the distribution and variability of discounts offered across different categories, helping to identify patterns in promotional strategies.



Observations:

1. Some categories show a higher median discount percentage compared to others.
2. The range of discount percentages varies significantly among categories, with certain categories exhibiting wider variability.
3. Categories with higher discounts may also show more outliers, indicating occasional deep discounts.

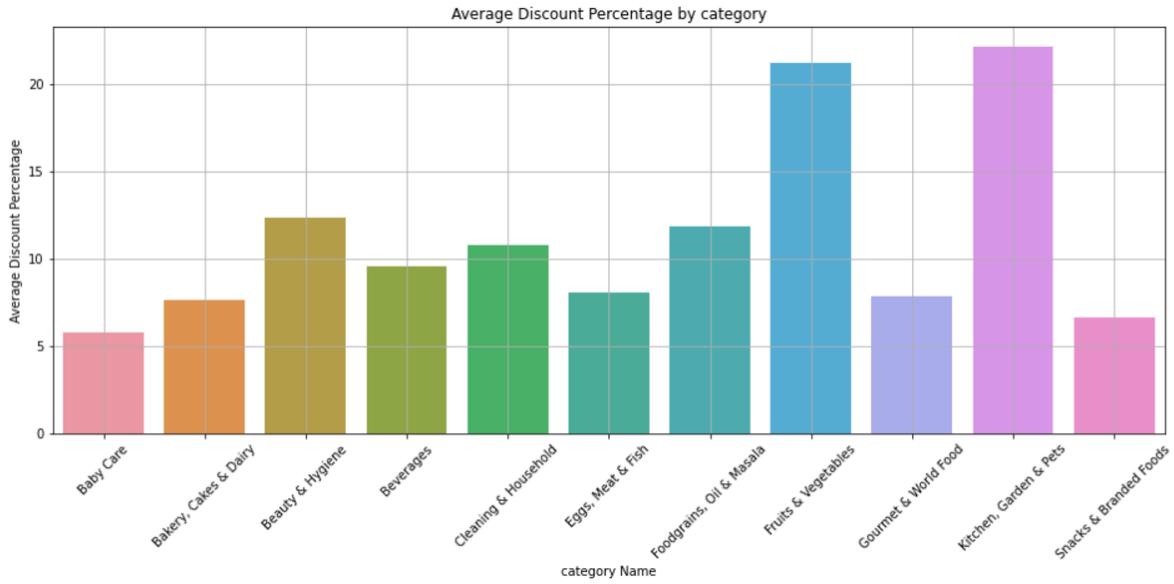
Insights:

1. Categories with consistently higher discounts may indicate promotional focus or competitive pricing strategies.
2. Variability in discount percentages highlights diverse pricing tactics within categories.
3. Understanding these trends can guide promotional planning, focusing on categories with potential for high customer impact.

4. Graph: Average Discount Percentage by Category

Purpose:

The "Average Discount Percentage by Category" graph helps identify the overall discounting trends across different categories, offering a clear view of categories with the most aggressive pricing strategies.



Observations:

1. Categories with the highest average discount percentages stand out as having consistent markdowns.
2. Certain categories show moderate or low average discounts, indicating less emphasis on promotional pricing.
3. The differences in average discounts suggest strategic prioritization of categories for discounts.

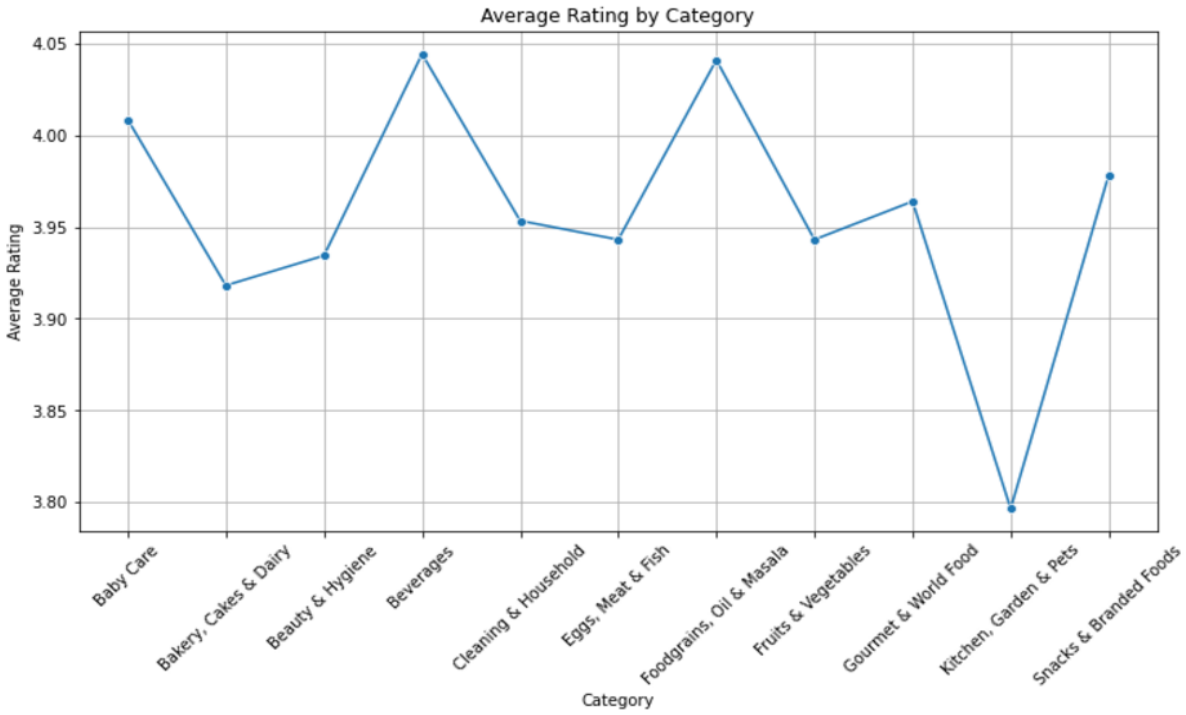
Insights:

1. Categories with the highest average discounts are likely targets for customer acquisition or inventory clearance.
2. Categories with lower discounts may reflect premium products or stable demand.
3. These insights can be used to refine marketing strategies and align them with profitability and sales goals.

3.Graph: Average Rating by Category

Purpose:

The "Average Rating by Category" graph highlights how customer satisfaction varies across different product categories, helping to identify categories with the best and worst customer feedback.



Observations:

1. Some categories exhibit consistently higher average ratings, indicating better customer satisfaction or quality.
2. Other categories have moderate or lower average ratings, suggesting areas for improvement in product quality or customer experience.
3. The variance in ratings across categories suggests differing levels of customer perception and preferences.

Insights:

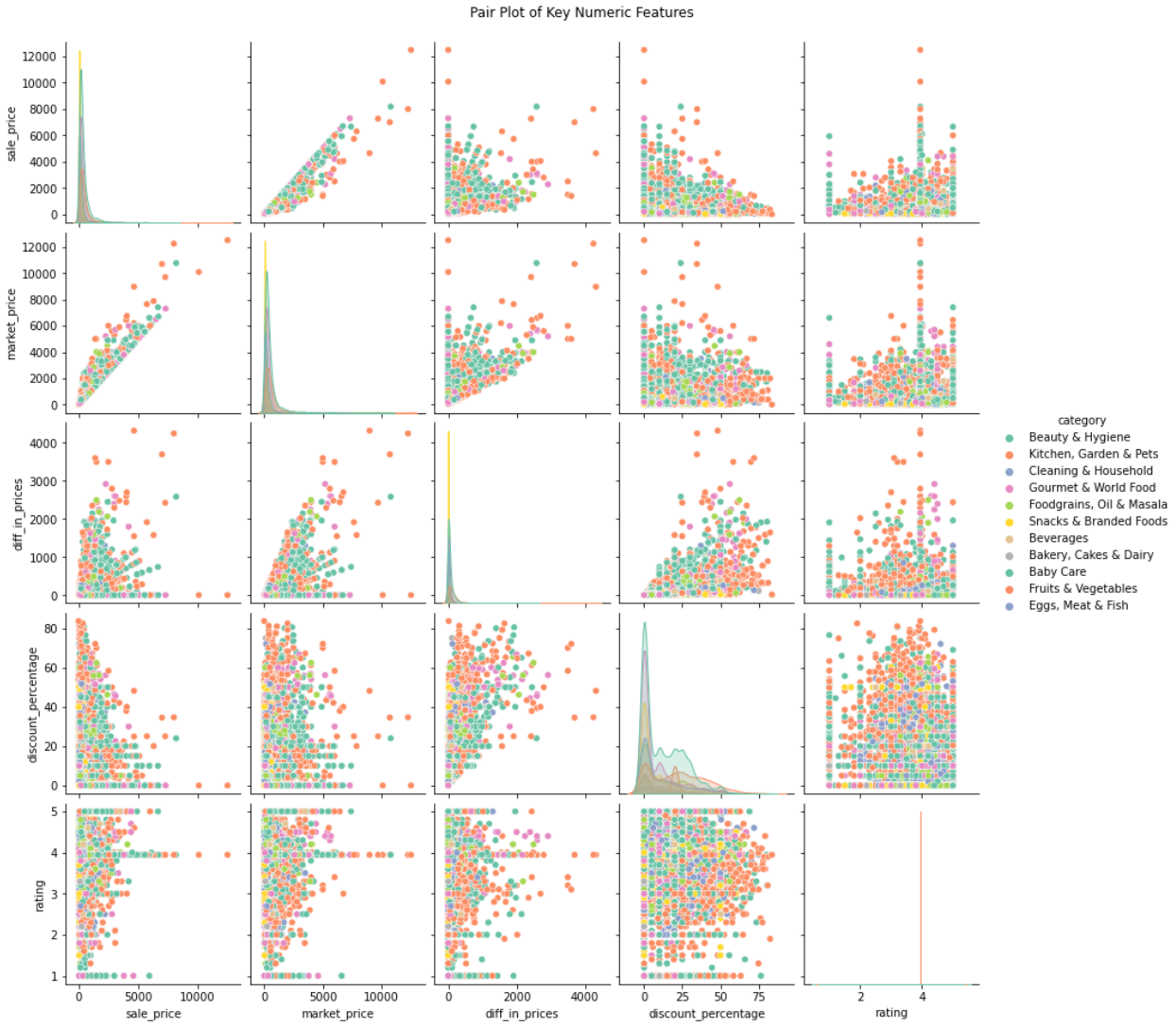
1. High-rated categories may represent strong product offerings and can be leveraged for marketing campaigns.
2. Low-rated categories may require attention to improve product quality or address customer concerns.
3. Understanding category-specific ratings helps prioritize efforts to enhance overall customer satisfaction and loyalty.

Multivariate Analysis

1.Graph: Pairplot of Key Numeric Features

Purpose:

The "Pairplot of Key Numeric Features" graph is designed to visualize the relationships and correlations between key numerical features in the dataset, helping to identify patterns and potential associations.



Observations:

1. Some numeric features show clear linear or non-linear relationships with each other, suggesting possible correlations.
2. There may be outliers visible in the scatter plots, which could impact the results of predictive modelling.

3. Diagonal histograms indicate the distribution of individual numeric features, revealing whether they follow a normal distribution or skewed patterns.

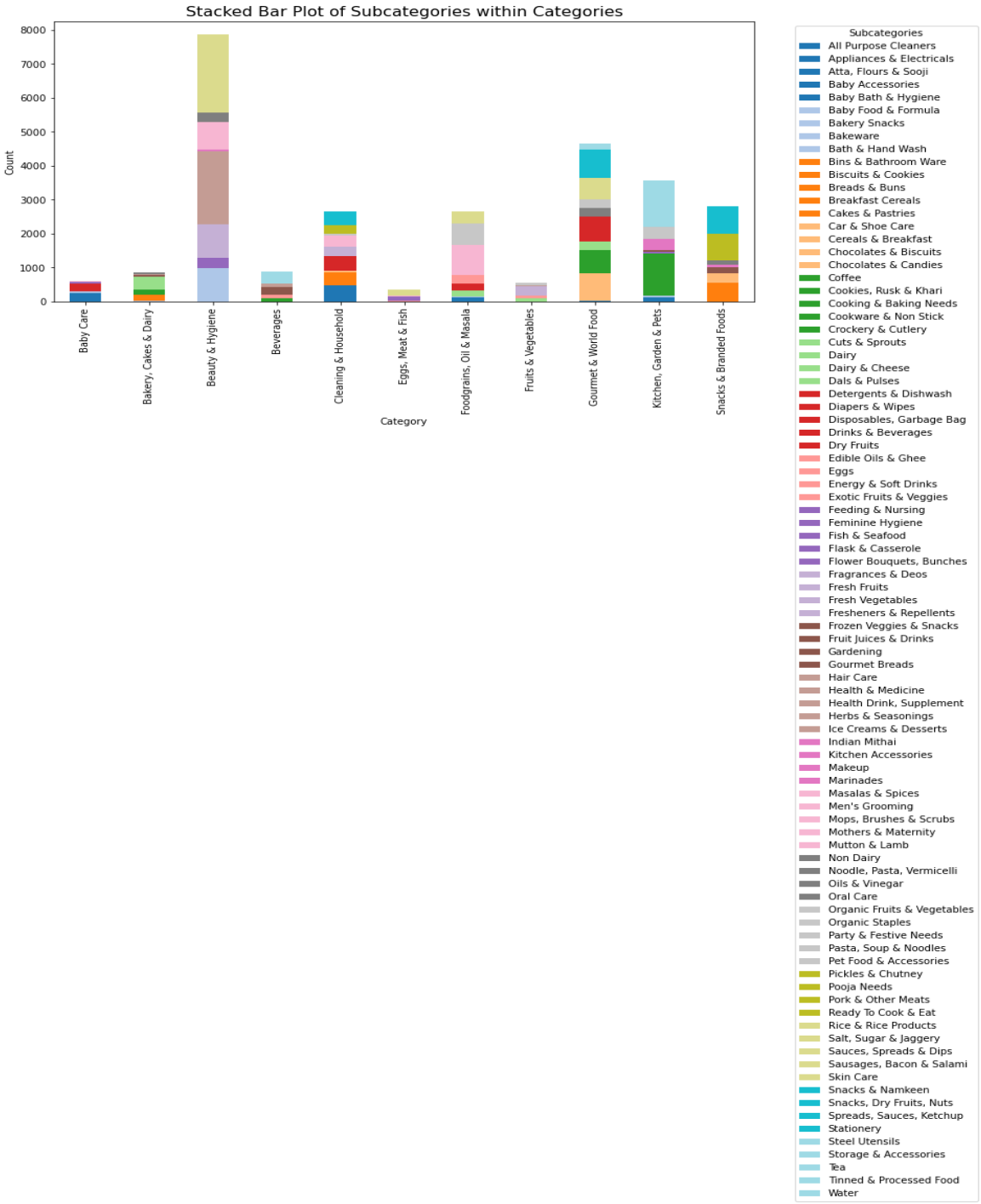
Insights:

1. Strong correlations between certain features might suggest redundancy or guide feature selection for modelling.
2. Outliers and distribution patterns can inform decisions on data preprocessing, like normalization or handling outliers.
3. The visual insights can guide feature engineering and model improvement based on feature relationships.

2. Graph: Stacked Barplot of Subcategories within Category

Purpose:

The "Stacked Barplot of Subcategories within Category" graph aims to visualize the distribution of subcategories within each main category, helping to understand the composition and proportion of subcategories within categories.



Observations:

1. Each bar represents a category, with the segments showing the proportion of different subcategories within it.
2. Some categories may have a higher concentration of certain subcategories, revealing dominant products.
3. There are noticeable variations in the number of subcategories across different main categories, indicating diversity or specificity within categories.

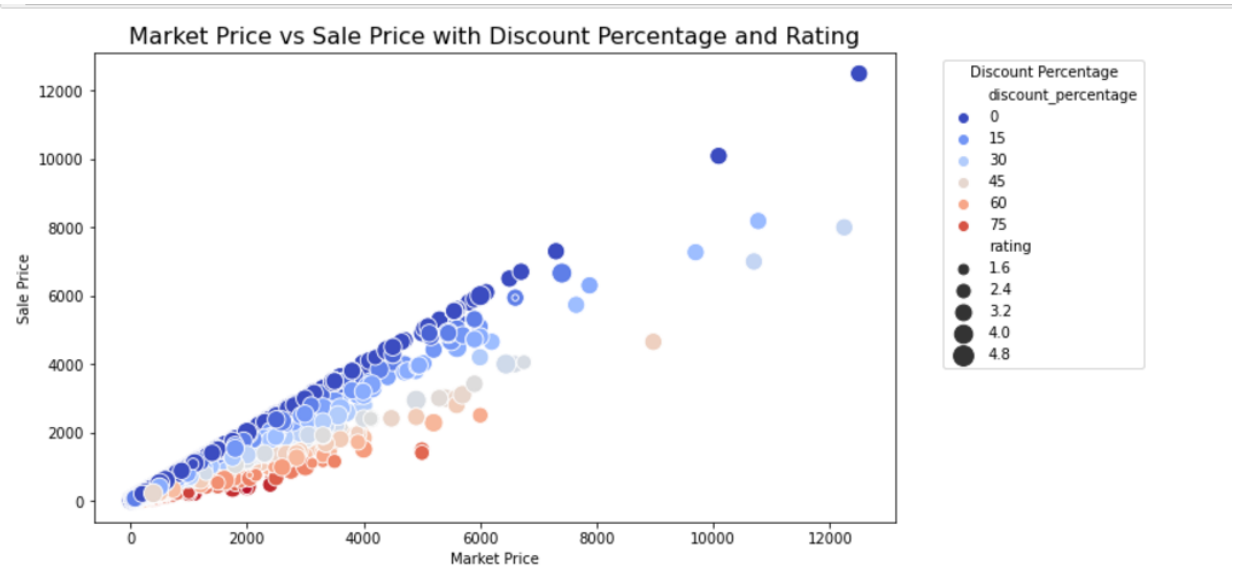
Insights:

1. The stacked barplot helps identify which subcategories dominate within a category, providing insights for inventory or sales strategies.
2. Categories with high subcategory diversity might benefit from segmentation-based marketing or product specialization.
3. This visualization can guide decisions about product offerings, highlighting where focus or expansion might be needed.

3. Market Price vs Sale Price with Discount Percentage and Rating

Purpose:

The "Market Price vs Sale Price with Discount Percentage and Rating" graph aims to explore the relationship between market price, sale price, and how discount percentage and rating impact these prices, providing a detailed understanding of pricing strategies and product value perception.



Observations:

1. As market price increases, sale price tends to follow a similar pattern, but discounts can vary significantly.
2. The scatter plot shows that products with higher ratings might still have significant discounts, suggesting value perception plays a role.
3. Discount percentage is inversely related to the sale price, where products with higher discounts tend to have a lower sale price.

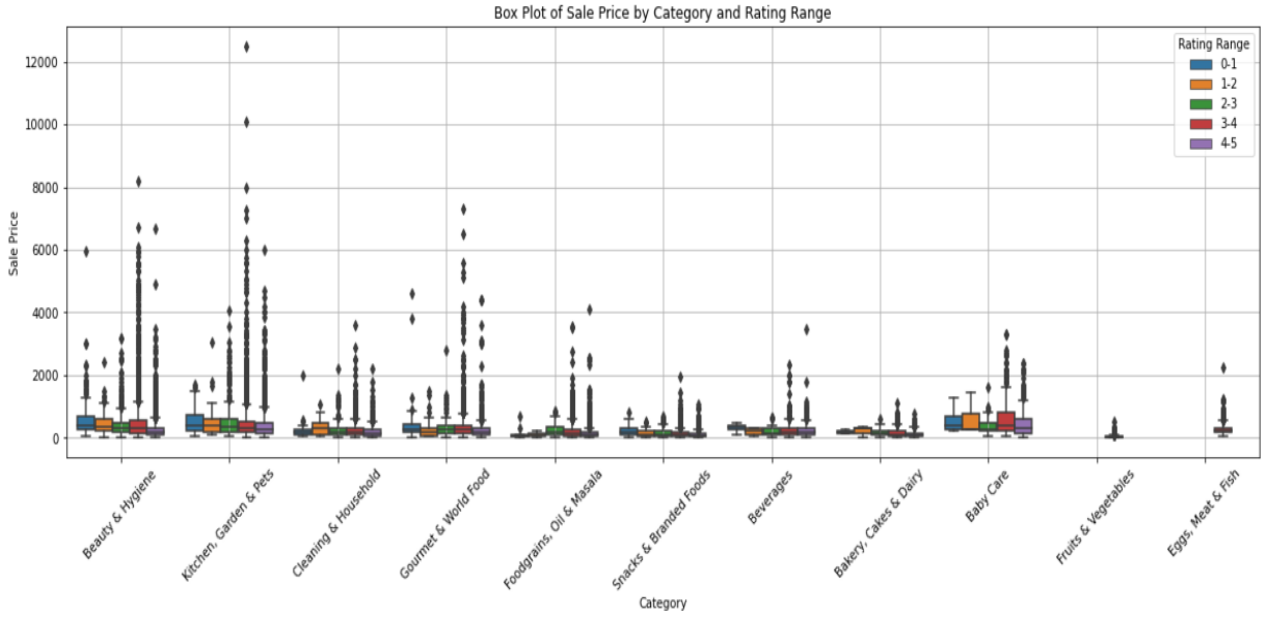
Insights:

1. Understanding the correlation between market price and sale price can help identify potential pricing inefficiencies or opportunities for better pricing strategies.
2. Rating may influence the sale price even in the presence of high discounts, indicating that higher-rated products are often perceived as more valuable.
3. Products with high discount percentages but low sale prices may be marketed as budget-friendly options, influencing consumer purchasing behavior.

4. Graph: Box Plot of Sale Price by Category and Rating Range

Purpose:

The "Box Plot of Sale Price by Category and Rating Range" graph aims to understand how sale prices vary across different product categories and rating ranges, highlighting how ratings influence pricing within each category.



Observations:

1. The distribution of sale prices varies significantly between categories, with some categories having higher median sale prices than others.
2. Within each category, products with higher ratings tend to have higher sale prices, with wider interquartile ranges observed in some categories.
3. Outliers are present in some categories, suggesting that a few high-priced items significantly impact the overall distribution.

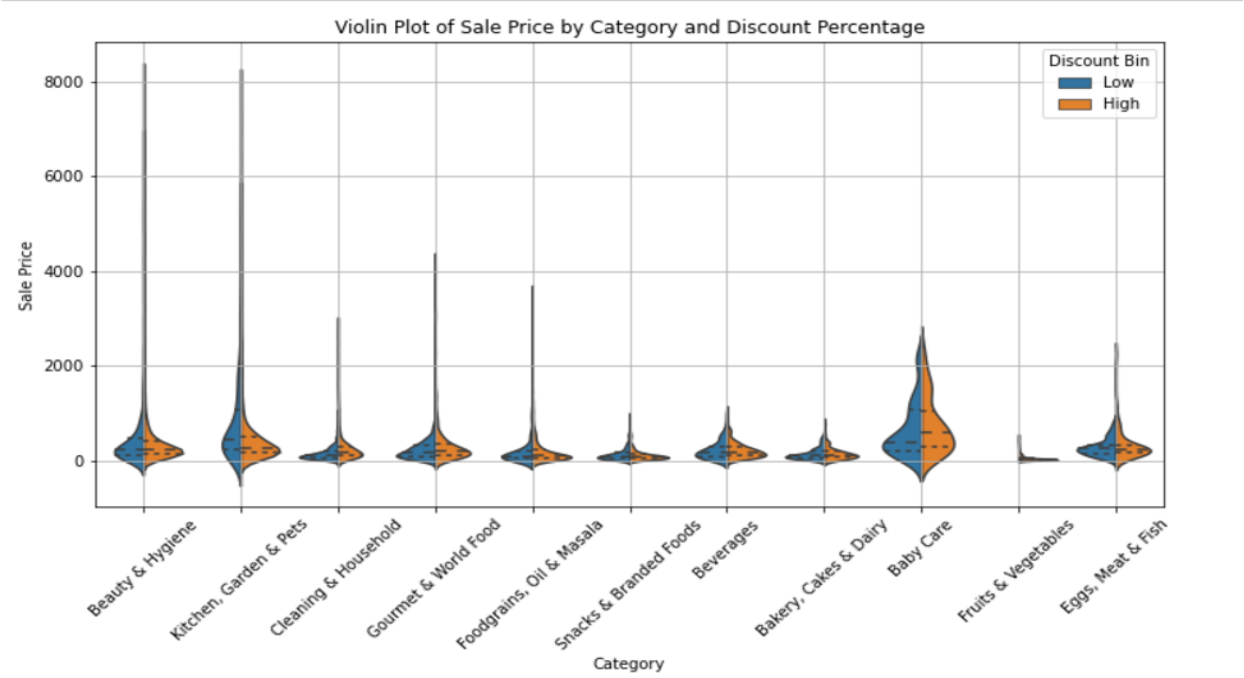
Insights:

1. Products with higher ratings generally justify higher sale prices, indicating that customer perception of value is tied to the product's rating.
2. Categories with higher price variability may benefit from targeted pricing strategies based on ratings to improve profitability.
3. The presence of outliers suggests that some products might be priced above the general trend, possibly due to special attributes or premium features.

5.Graph: Violin Plot of Sale Price by Category and Discount Percentage

Purpose:

The "Violin Plot of Sale Price by Category and Discount Percentage" graph is designed to visualize the distribution of sale prices across different categories, with a focus on how discount percentages impact the price distribution within each category.



Observations:

1. Each category shows a unique distribution of sale prices, with certain categories exhibiting wider price ranges and others being more concentrated around the median.
2. The shape of the violin plots suggests that higher discounts are associated with lower sale prices, particularly in categories with a high variance in pricing.
3. Some categories show multiple modes (peaks), indicating that sale prices are clustered around certain values, possibly due to promotional strategies or product types.

Insights:

1. Discount percentage plays a significant role in determining sale price, especially in categories where discounts are more aggressive, leading to a decrease in the overall sale price.
2. Categories with broader price ranges may benefit from more segmented pricing strategies based on discount levels to maximize sales without sacrificing margins.
3. The presence of multiple peaks suggests that certain products within a category may have distinct pricing strategies, possibly reflecting different product qualities or customer segments.

Model Creation and Evaluation

In this project, multiple regression models were created to predict sale prices using features like market price, category, sub-category, brand, rating, and discount percentage. The models included Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Each model was trained on preprocessed data using techniques like standard scaling for numerical features and one-hot encoding for categorical features.

Model evaluation was performed using R^2 scores and RMSE values on the test dataset. Among the models, Random Forest achieved the highest accuracy with an R^2 score of 0.9989 and the lowest RMSE of 15.83, making it the best-performing model. This stepwise approach ensured the selection of the most reliable model for predicting sale prices effectively.

Conclusion

This project focused on analyzing and predicting the sale prices of products on an e-commerce platform, using various features such as market price, category, subcategory, brand, discount percentage, and ratings. The analysis revealed key insights such as the significant impact of discounts on sale prices, the correlation between higher ratings and higher sale prices, and how different product categories and brands have varying pricing strategies. These insights provide valuable knowledge for businesses looking to optimize their pricing decisions and understand customer preferences.For predicting the sale price, multiple machine learning models were evaluated, including Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Among these, the Random Forest Regressor proved to be the best model, achieving the highest accuracy (R^2 score of 0.99895) and the lowest RMSE (15.83). This model successfully captured the complex relationships between the features, making it the most reliable for predicting sale prices. Overall, the project demonstrates the potential of using machine learning to predict product pricing and optimize pricing strategies for e-commerce businesses.

References

Python Programming Language

Python Software Foundation. https://www.python.org/

Pandas: Python Data Analysis Library

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, pp. 51-56. https://pandas.pydata.org/

NumPy: Fundamental Package for Scientific Computing with Python

Harris, C.R., et al. (2020). Array programming with NumPy. Nature, 585(7825), pp. 357-362. https://numpy.org/

Matplotlib: Visualization with Python

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), pp. 90-95. https://matplotlib.org/

Seaborn: Statistical Data Visualization

Waskom, M.L. (2021). Seaborn: Statistical Data Visualization. Journal of Open-Source Software, 6(60), 3021. https://seaborn.pydata.org/

BigBasket Dataset for Data Analysis

Kaggle: BigBasket Dataset. https://www.kaggle.com/ .

Principal Component Analysis (PCA)

Jolliffe, I.T. (1986). Principal Component Analysis. Springer Series in Statistics. https://doi.org/10.1007/978-1-4757-1904-8

Jupyter Notebook for Analysis and Visualization

Project Jupyter. https://jupyter.org/

Online Documentation and Tutorials

Towards Data Science Blog: https://towardsdatascience.com/

GeeksforGeeks: Python and Data Science Tutorials. https://www.geeksforgeeks.org/

Scikit-learn: Machine Learning in Python

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp. 2825-2830. https://scikit-learn.org/

XGBoost: Extreme Gradient Boosting

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://xgboost.readthedocs.io/

TensorFlow: An Open-Source Software Library for Dataflow and Differentiable Programming

Abadi, M., et al. (2016). TensorFlow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), pp. 265-283. https://www.tensorflow.org/

Gradient Boosting Machines

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), pp. 1189-1232. https://projecteuclid.org/

Random Forests

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), pp. 5-32. https://link.springer.com/article/10.1023/A:1010933404324

Data Preprocessing for Machine Learning

Brownlee, J. (2020). Data Preprocessing for Machine Learning in Python. Machine Learning Mastery. https://machinelearningmastery.com/

Seaborn Statistical Data Visualization

Waskom, M.L. (2021). Seaborn: Statistical Data Visualization. Journal of Open-Source Software, 6(60), 3021. https://seaborn.pydata.org/

Deep Learning with Python

Chollet, F. (2018). Deep Learning with Python. Manning Publications.

Kaggle Competitions and Datasets

Kaggle: Competitive Data Science. https://www.kaggle.com/

GitHub Repository Link:

https://github.com/pavankalyanperla/EDA-on-BigBasket-Dataset