Data Analysis on Product Performance

A Project Report

submitted in partial fulfillment of the requirements

of

Foundation course

by

Mehroz Patel 230090107148

Bhavini Badgujar 240093107004

Bhavnani Daksh 230090116012

Riya Shah 230090107191

Under the Esteemed Guidance of

Praful Vinayak bhoyar sir

Data Analysis on Product Performance

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this project work.

Firstly, we would like to extend our heartfelt appreciation to our supervisor, **Praful Vinayak Bhoyar**, for his continuous guidance, support, and encouragement throughout the course of this project. His valuable insights, constructive criticism, and patience were instrumental in refining our ideas and improving our analytical approach.

We also extend our thanks to the faculty of the **Computer Department** for their continuous support and encouragement. Additionally, we are grateful to our college, **C. K. Pithawala college of engineering**, for providing the necessary resources and facilities.

Finally, we thank our friends and family for their unwavering support and motivation during this journey.

Data Analysis on Product Performance

ABSTRACT

This report presents a comprehensive analysis of product performance data, focusing on various attributes such as sales, revenue, customer ratings, and marketing effectiveness. The dataset comprises 25,000 observations across 25 attributes, including product details, sales metrics, customer feedback, and financial indicators. The analysis begins with an overview of the data, followed by exploratory data analysis (EDA) to identify trends, patterns, and outliers. Key findings include the identification of outliers in the revenue generated, which were addressed using the Interquartile Range (IQR) method. The report also highlights the importance of understanding customer demographics, return rates, and promotion effectiveness in optimizing product performance. Visualizations, including box plots and distribution plots, are used to illustrate the data distribution and outlier detection. The insights derived from this analysis can inform strategic decisions to enhance product performance, improve customer satisfaction, and optimize marketing efforts.

Data Analysis on Product Performance

TABLE OF CONTENTS

Abstract 3

List of Figures 5

List of Tables 5

Chapter 1. Introduction 6

1.1 Problem Statement 7

1.2 Problem Definition 7

1.3 Expected Outcomes.....7

1.4 Organization Of the Report 7

Chapter 2. Literature Survey 8

2.1 Paper-1 Sales Forecasting and Its Impact on Business Performance 9

2.2 Paper-2 Inventory Optimization in Retail……....9

2.3 Paper-3 Customer Behavior Analysis in E-Commerce….....9

2.4 Paper-4 Marketing Effectiveness in Digital Channels.....10

Chapter 3. Proposed Methodology 11

3.1 System Design 12

3.2 Modules Used 13

3.3 Data Flow Diagram 13

3.4 Requirement Specification 17

Chapter 4. Implementation and Results 18

4.1 Data Processing 19

4.2 Results of Data Cleaning 22

4.3 Results of Statistical Analysis 22

4.4 Results of Data Visualization 25

4.5 Summary of Results 27

Chapter 5. Conclusion 28

GitHub Link 30

Video Link 30

References 31

Data Analysis on Product Performance

LIST OF FIGURES

		Page No.
Figure 1	DFD Level 0 all over process	**14**
Figure 2	DFD Level 1 - Data Collection and Preprocessing Module	**15**
Figure 3	DFD Level 1 – Exploratory Data Analysis Module	**16**
Figure 4	DFD Level 1 – Statistical Analysis Module	**17**
Figure 5	Co-relation matrix	**23**
Figure 6	Visualization of numerical columns	**25**
Figure 7	Visualization of Categorical Columns	**27**

LIST OF TABLES

	Description	Page No.
Table 1	Data overview	**19**
Table 2	Description of columns	**20**
Table 3	Data description	**21**
Table 4	Outlier’s detection	**22**

CHAPTER 1

INTRODUCTION

Data Analysis on Product Performance

CHAPTER 1

INTRODUCTION

1.1. Problem Statement:

Understanding the performance of products in a competitive market is crucial for business success. Organizations need a systematic approach to evaluate their sales trends, inventory turnover, and customer preferences. This study aims to analyze various product performance metrics using real-world data, helping businesses optimize their operations and maximize profitability..

1.2. Problem Definition:

This research focuses on analyzing product sales, revenue trends, and profitability metrics while identifying patterns in customer behavior. The study will help businesses determine which products are generating the highest revenue and which ones require strategic adjustments in pricing, inventory management, and marketing.

1.3. Expected Outcomes:

- **Identification of high and low-performing products**: Highlighting products that contribute the most to revenue and those that underperform.
- **Insights into seasonal trends and their impact on revenue**: Understanding how sales fluctuate during different periods of the year.
- **Improved strategies for inventory and stock optimization**: Ensuring optimal stock levels to meet demand without overstocking.
- **Recommendations for better marketing and promotions**: Identifying effective promotional strategies to boost sales.

1.4. Organization of the Report

The report is divided into several chapters covering literature review, methodology, analysis, and final recommendations. Each chapter presents key findings and actionable insights for businesses looking to improve their product performance strategies.

CHAPTER 2

LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

2.1. Paper-1

Sales Forecasting and Its Impact on Business Performance

2.2.1. Brief Introduction of Paper:

This paper explores different methods used for sales forecasting, such as statistical analysis, artificial intelligence, and machine learning techniques. It discusses the significance of data-driven decision-making in improving product performance.

2.2.2. Techniques used in Paper:

- Time Series Analysis: Identifies revenue patterns over time
- Regression Models: Helps in predicting future sales based on historical data.
- Customer Segmentation: Classifies customers based on purchasing behavior.

2.2. Paper-2

Inventory Optimization in Retail

2.2.1. Brief Introduction of Paper:

This study focuses on inventory management strategies to minimize costs while ensuring product availability. It highlights the importance of demand forecasting and reorder level optimization.

2.2.2. Key Findings:

- Demand Forecasting: Accurate demand prediction reduces overstocking and stockouts.
- Reorder Levels: Setting optimal reorder levels improves inventory turnover.
- Cost Reduction: Efficient inventory management lowers holding costs.

2.3. Paper-3

Customer Behavior Analysis in E-Commerce

2.3.1 Brief Introduction of Paper

This research examines customer behavior in online retail, focusing on factors such as purchase frequency, product preferences, and return rates.

2.3.2 Techniques Used

- Cluster Analysis: Groups customers based on purchasing patterns.
- Sentiment Analysis: Evaluates customer reviews to gauge satisfaction.
- Churn Prediction: Identifies customers at risk of leaving.

2.4. Paper-4

Marketing Effectiveness in Digital Channels

2.4.1 Brief Introduction of Paper

This paper analyzes the effectiveness of digital marketing campaigns, including social media, email, and online ads, in driving sales and customer engagement.

2.4.2 Key Findings

- ROI of Digital Campaigns: Online ads yield higher ROI compared to traditional methods.
- Customer Engagement: Social media campaigns increase brand awareness and customer interaction.
- Personalization: Tailored marketing messages improve conversion rates.

CHAPTER 3

PROPOSED METHODOLOGY

CHAPTER 3

PROPOSED METHODOLOGY

3.1 System Design: The system follows a structured data analytics pipeline tailored for product performance evaluation:

1) Data Collection: Product performance data is collected from various sources, including sales records and inventory databases.
2) Data Preprocessing: Handling missing values, removing outliers, and transforming categorical data.
   - Handling missing values : Handling missing values using imputation techniques.
   - removing outliers : Removing outliers based on statistical thresholds (e.g., IQR method for revenue analysis).
   - transforming categorical data : Transforming categorical data into numerical representations.
3) Exploratory Data Analysis (EDA):
   Univariate Analysis: Understanding data distribution, identifying trends.
   Bivariate Analysis: Correlation between sales, revenue, and product categories.
   Data visualization using histograms, boxplots, and scatter plots.
4) Feature Engineering: Selecting relevant features for analysis.
   Selecting relevant attributes like product category, discount percentage, and customer ratings.
   Creating new variables such as stock turnover rate and marketing spend efficiency.
5) Model Development: Applying machine learning models (e.g., regression analysis, classification models) for performance prediction.
   Identifying key factors influencing product success.
6) Evaluation and Insights: Assessing model accuracy and extracting actionable insights.

3.2 Modules Used

Data Processing: Pandas and NumPy for data manipulation.

Visualization: Matplotlib and Seaborn for graphical representation.

Statistical Analysis: Scipy and Statsmodels for hypothesis testing.

Machine Learning: Scikit-learn for predictive modeling.

Outlier Detection: IQR method to remove anomalies in revenue and sales data.

Performance Metrics: Evaluation based on customer reviews, return rates, and profitability.

3.3 Data Flow Diagram

A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).

3.3.1. DFD Level 0

The Data Flow Diagram (DFD) Level 0 represents the overall system at a high level. It provides an overview of how data flows between the external entities and the system. The main components of the system are:

External Entities: Users, Inventory Management System, Sales Database

Processes: Data Collection, Processing, Analysis, and Reporting

Data Stores: Product Performance Database, Report Repository

Output: Insights, Visual Reports, and Performance Metrics

Figure1

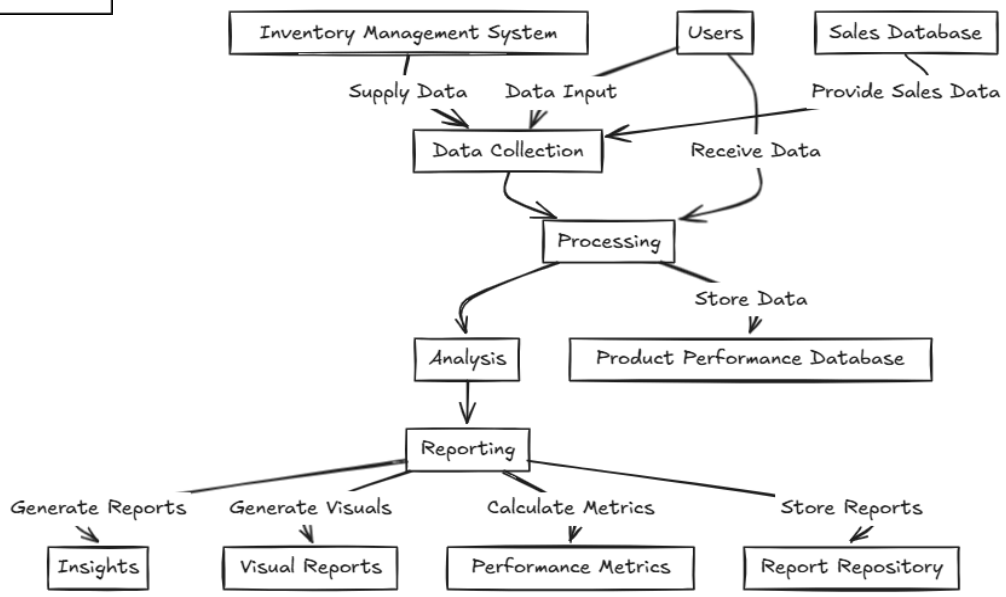


Fig1 all over process

3.3.2. DFD Level 1 – Data Collection and Preprocessing Module:

A Level 1 Data Flow Diagram (DFD) for the Data Collection and Preprocessing Module breaks down the overall system into major processes, data stores, and data flows.

Figure2

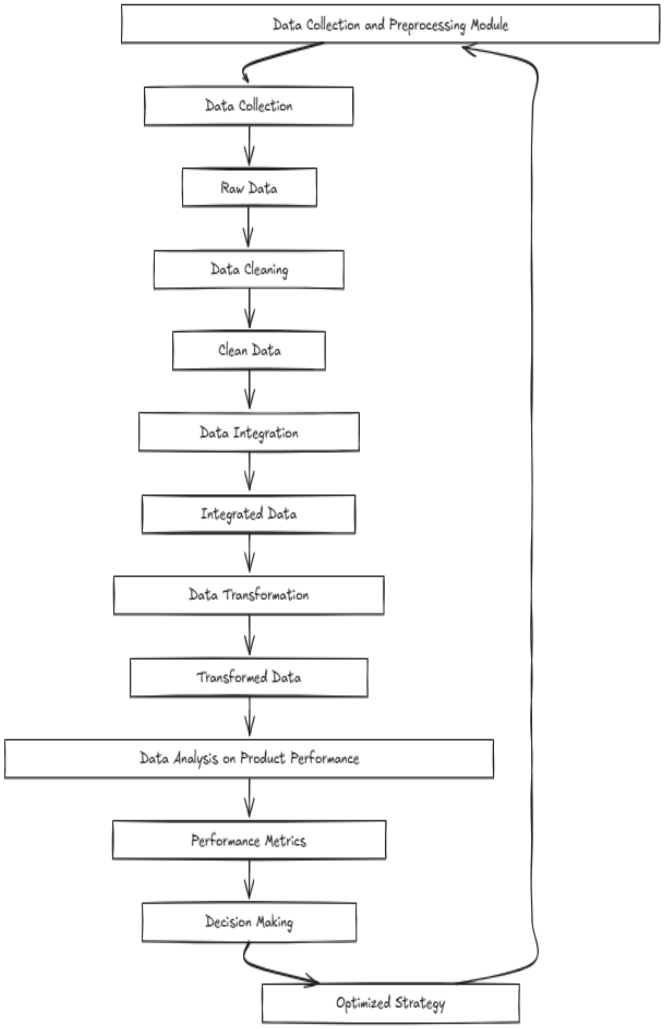


Fig2 Data Collection and Preprocessing Module

3.3.3. DFD Level 1 – Exploratory Data Analysis Module:

A Level 1 Data Flow Diagram (DFD) for the Exploratory Data Analysis (EDA) Module breaks down the process of analyzing and summarizing datasets.

Figure3

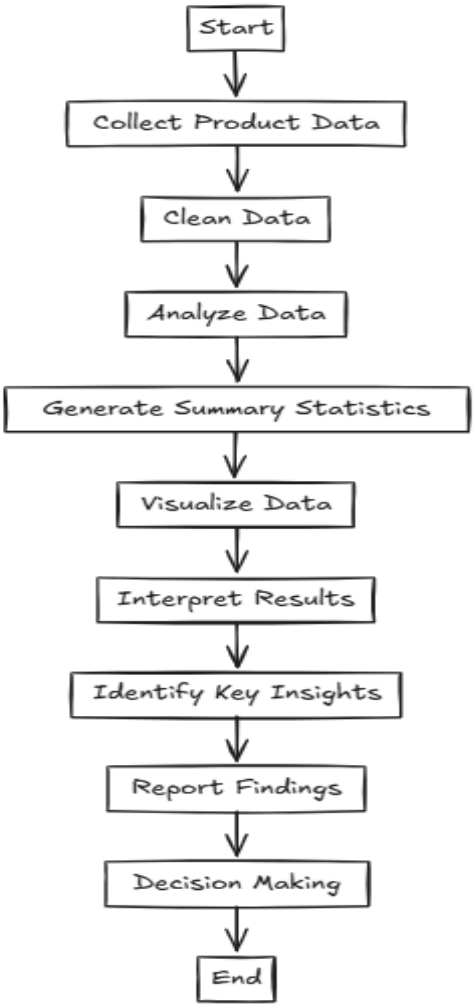


Fig3 Exploratory Data Analysis Module

3.3.4. DFD Level 1 – Statistical Analysis Module:

A Level 1 Data Flow Diagram (DFD) for the Statistical Analysis Module illustrates how data flows through different processes involved in statistical computations and hypothesis testing.

Figure4

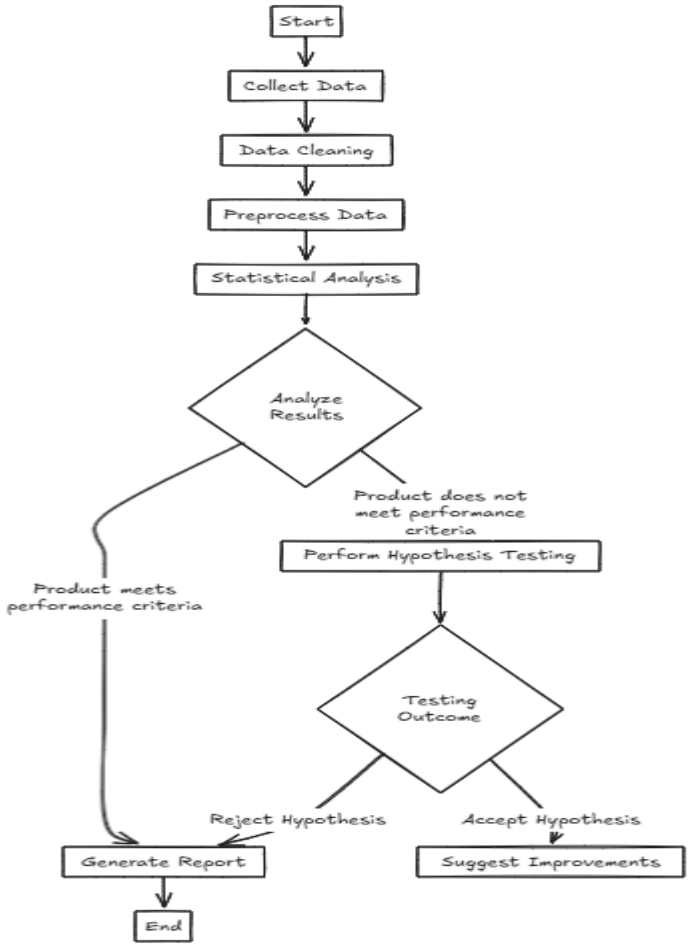


Fig4 Statistical Analysis Module

3.4 Requirement Specification:

3.4.1 Hardware Requirements: Standard computing hardware with sufficient memory and processing power.

3.4.2 Software Requirements: Python programming language, Pandas, NumPy, Matplotlib, Seaborn, and Jupyter Notebook.

CHAPTER 4

Implementation and Result

CHAPTER 4

IMPLEMENTATION and RESULT

4.1 Data Processing

Dataset Overview:

1. We have 25000 Observation(Rows)
2. We have 25 Attributes(Columns)

	Matric	Values
1	**Total Observation**	25000
2	**Total Attributes**	25

Table 1 data overview

This section outlines the steps taken to preprocess and clean the dataset before performing statistical analysis.

1. Handling Missing Values (Mean/Median Imputation):
   - Numerical columns were filled with the **mean** of the column.
   - Categorical columns were filled with the **mode** (most frequent value).

2. Removing Duplicates & Filtering Irrelevant Attributes
   - No duplicate rows were found, so no rows were removed.

3. Normalizing Numerical Data
   - Applied **Min-Max Normalization** (scaled between 0 and 1) to numerical columns for consistency.

Data Analysis on Product Performance

Description of Columns :

```
Data columns (total 25 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Product_ID                    25000 non-null  int64
 1   Product_Name                  25000 non-null  object
 2   Category                      25000 non-null  object
 3   Brand                         25000 non-null  object
 4   SKU                           25000 non-null  object
 5   Total_Sales                   25000 non-null  int64
 6   Revenue_Generated             25000 non-null  float64
 7   Average_Selling_Price         25000 non-null  float64
 8   Discount_Percentage           25000 non-null  float64
 9   Peak_Sales_Period             25000 non-null  object
 10  Sales_Channel                 25000 non-null  object
 11  Stock_Available               25000 non-null  int64
 12  Reorder_Level                 25000 non-null  int64
 13  Lead_Time                     25000 non-null  int64
 14  Stock_Turnover_Rate           25000 non-null  float64
 15  Customer_Rating               25000 non-null  float64
 16  Number_of_Reviews             25000 non-null  int64
 17  Return_Rate                   25000 non-null  float64
 18  Top_Buying_Demographic        25000 non-null  object
 19  Marketing_Spend               25000 non-null  float64
 20  Promotion_Effectiveness       25000 non-null  float64
 21  Cross_Selling_Performance     25000 non-null  object
 22  Profit_Margin                 25000 non-null  float64
 23  Cost_of_Goods_Sold            25000 non-null  float64
 24  Market_Competitiveness_Index  25000 non-null  float64
dtypes: float64(11), int64(6), object(8)
memory usage: 4.8+ MB
```

Table 2 description of columns

Data Analysis on Product Performance

Key Findings:

1. Sales & Revenue Trends

- Revenue varies widely, with **250 outliers**, indicating extreme-performing products.
- Discount strategies and pricing significantly impact total sales and profit margins.

2. Customer & Market Behavior

- Some products have **high return rates** and **low ratings**, indicating potential quality or customer satisfaction issues.
- Marketing spend effectiveness varies, influencing product visibility and demand.

	count	mean	std	min	25
Product_ID	25000.0	1.250050e+04	7.217023e+03	1.00	6250.75(
Total_Sales	25000.0	5.007868e+03	2.870932e+03	50.00	2521.00(
Revenue_Generated	25000.0	2.533694e+06	2.207245e+06	631.04	705613.26(
Average_Selling_Price	25000.0	5.062182e+02	2.875114e+02	5.05	258.95
Discount_Percentage	25000.0	2.500300e+01	1.445527e+01	0.01	12.53(
Stock_Available	25000.0	5.006055e+02	2.878696e+02	0.00	253.00(
Reorder_Level	25000.0	2.538241e+02	1.412004e+02	10.00	131.00(
Lead_Time	25000.0	1.544400e+01	8.084887e+00	2.00	8.00(
Stock_Turnover_Rate	25000.0	5.261687e+00	2.741425e+00	0.50	2.89(
Customer_Rating	25000.0	2.985983e+00	1.152937e+00	1.00	1.98(
Number_of_Reviews	25000.0	2.496939e+03	1.435723e+03	1.00	1248.00(
Return_Rate	25000.0	1.003379e+01	5.750020e+00	0.00	5.10(
Marketing_Spend	25000.0	2.497695e+05	1.442365e+05	1014.65	123648.32
Promotion_Effectiveness	25000.0	4.984340e+01	2.887163e+01	0.00	24.61
Profit_Margin	25000.0	2.773230e+01	1.301184e+01	5.00	16.45(
Cost_of_Goods_Sold	25000.0	4.011854e+02	2.300202e+02	2.00	202.34
Market_Competitiveness_Index	25000.0	5.014915e+01	2.886271e+01	0.00	25.14

Table 3 data description

4.2 Results of Data Cleaning:

This section describes how missing values, outliers, and inconsistencies in the dataset were handled to ensure data quality

	Issue	Count
0	Duplicate Rows	0
1	Missing Values	0
2	Negative Total Sales	0
3	Negative Revenue Generated	0
4	Negative Selling Price	0
5	Invalid Discount Percentage (>100 or <0)	0
6	Invalid Customer Rating (>5 or <0)	0

Table 4 outlier's detection

4.3 Results of Statistical Analysis:

Revenue & Sales: The median revenue is **lower than the mean**, indicating possible **skewness due to high-value outliers**.

Profit & Cost: Profit margins and cost values are well-distributed, with most data points centered around the mean.

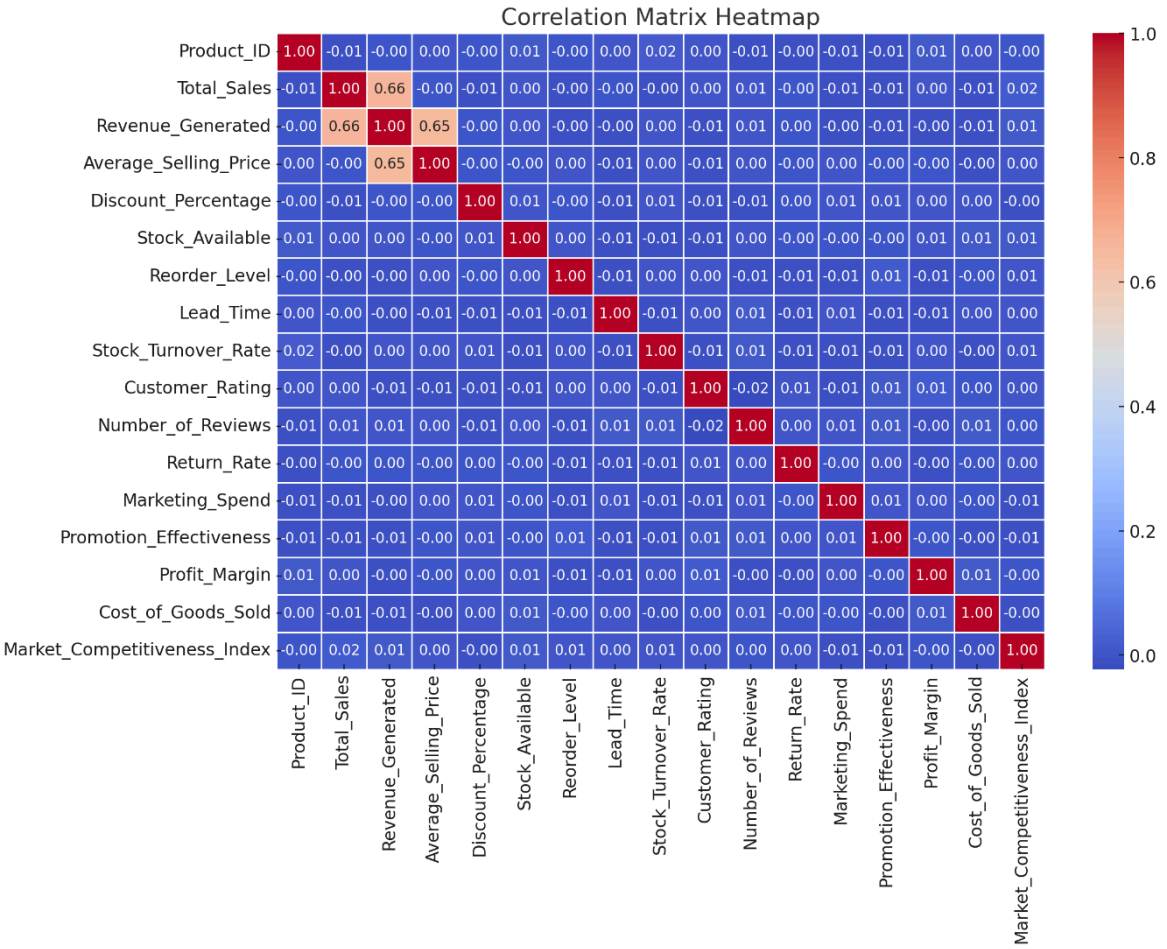


Fig 5 co-relation matrix

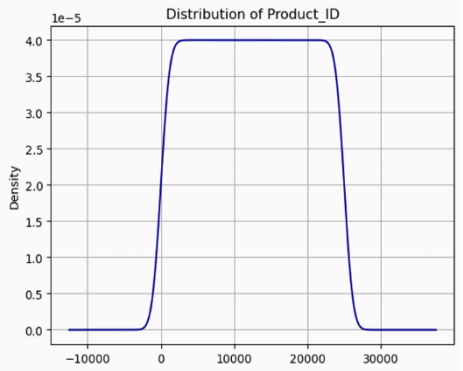
4.4 Results of Data Visualization

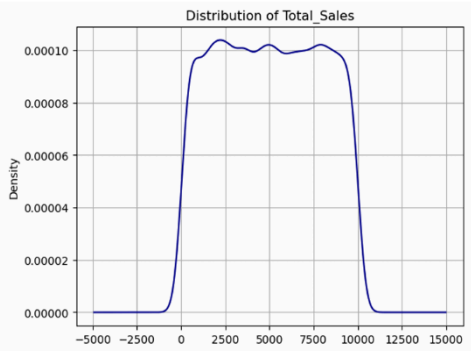
4.4.1 Distribution of numerical data

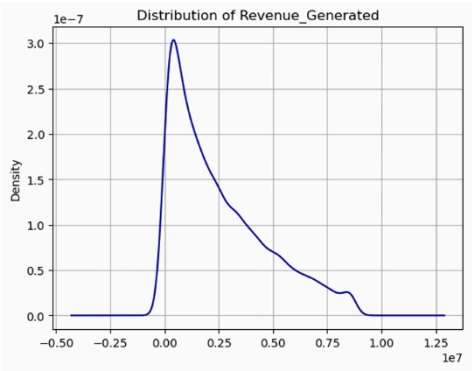
- **Product_ID** – Unique identifier for each product.
- **Total_Sales** – Total number of units sold.
- **Revenue_Generated** – Total revenue earned from the product.
- **Average_Selling_Price** – Average price at which the product is sold.
- **Discount_Percentage** – Percentage of discount applied on the product.
- **Stock_Available** – Number of units available in inventory.
- **Reorder_Level** – Stock level at which the product needs to be reordered.
- **Lead_Time** – Time taken to restock the product.
- **Stock_Turnover_Rate** – Frequency at which the inventory is sold and replaced.
- **Customer_Rating** – Average rating given by customers (e.g., out of 5).
- **Number_of_Reviews** – Count of customer reviews received.
- **Return_Rate** – Percentage of products returned by customers.
- **Marketing_Spend** – Amount spent on marketing campaigns for the product.
- **Promotion_Effectiveness** – Measure of how effective promotions were in driving sales.
- **Profit_Margin** – Percentage of profit relative to revenue.

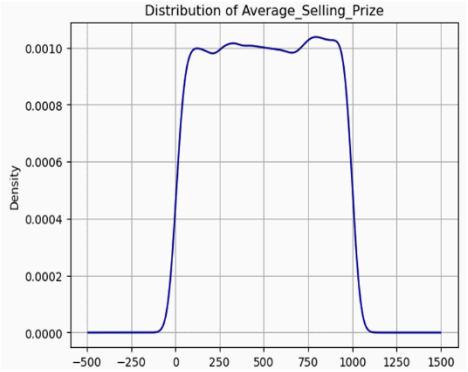
Data Analysis on Product Performance

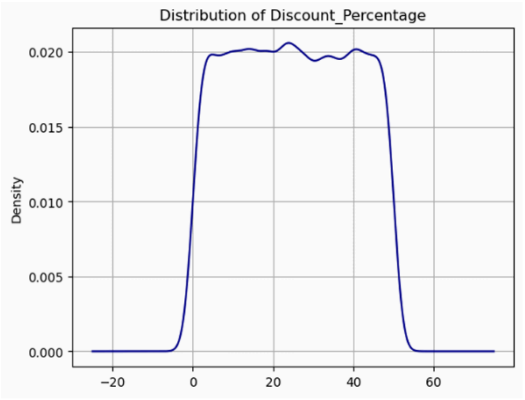
- **Cost_of_Goods_Sold** – Direct costs associated with producing the product.
- **Market_Competitiveness_Index** – A score indicating how competitive the product is in the market.



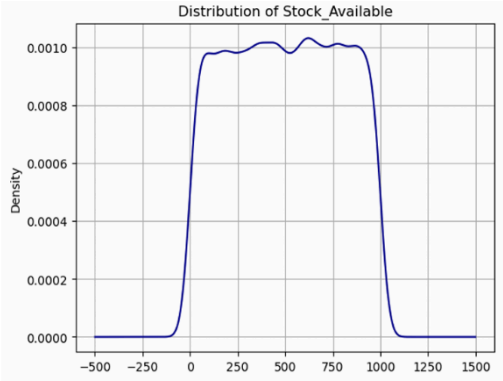


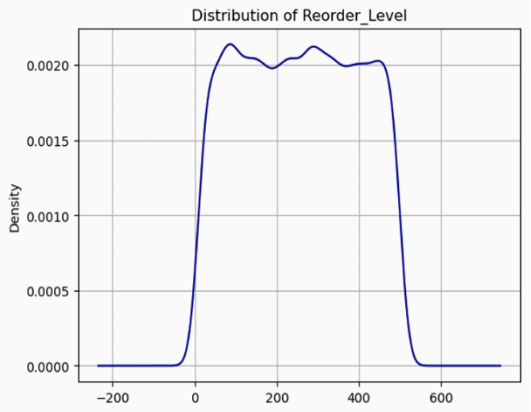


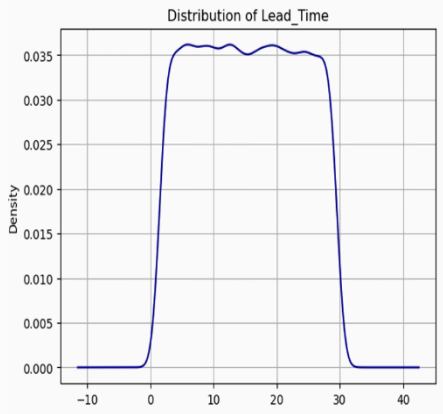


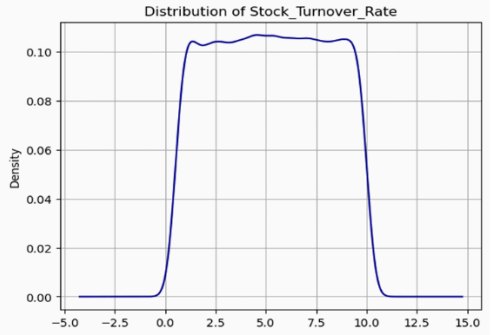


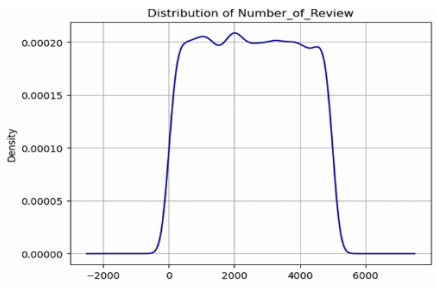
Data Analysis on Product Performance

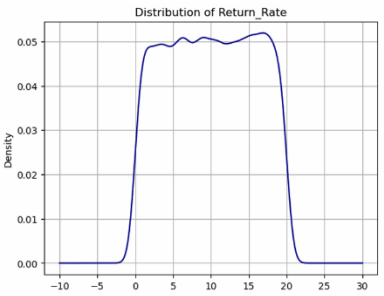












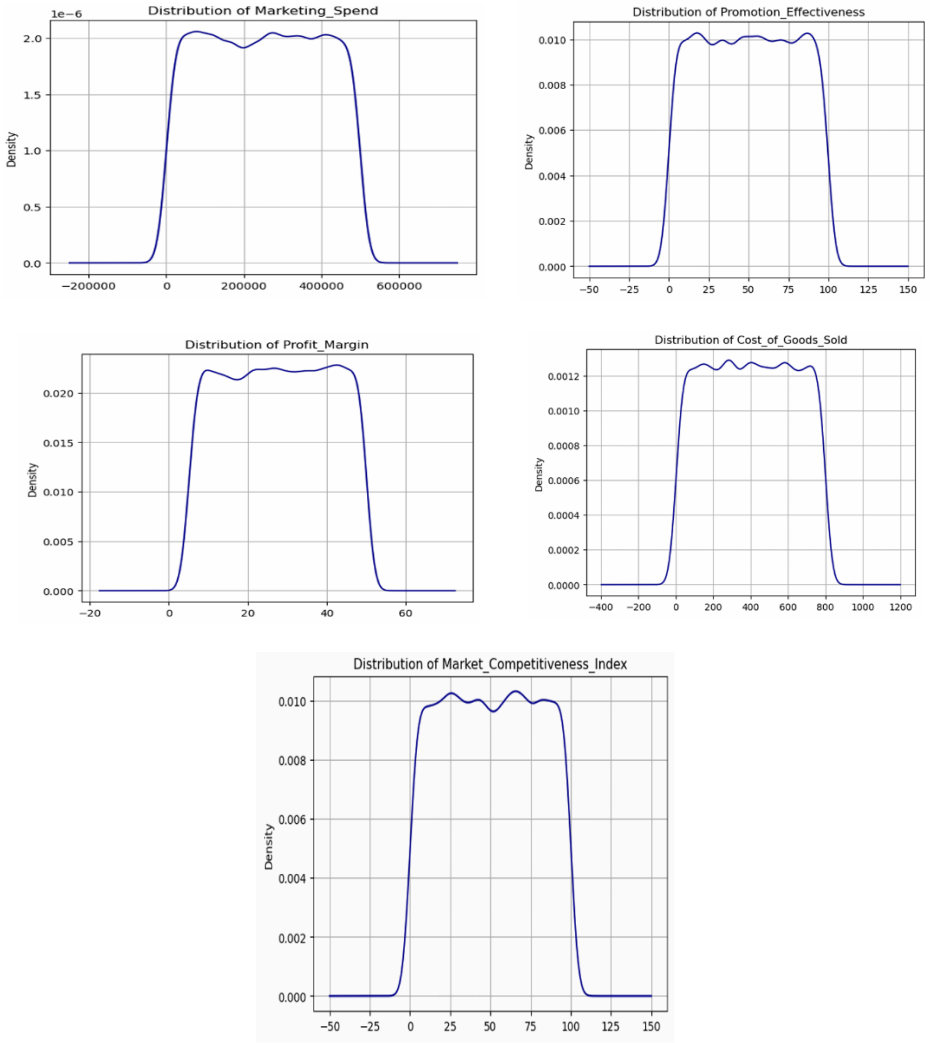


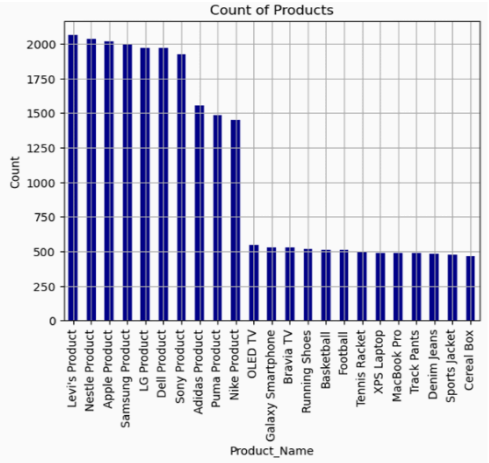
Fig 6 visualization of numerical columns

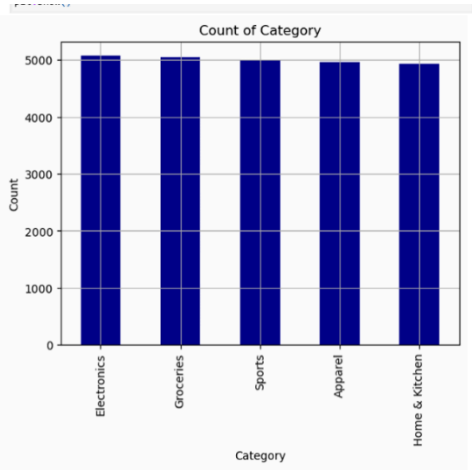
4.4.2 Visualization of categorical data:

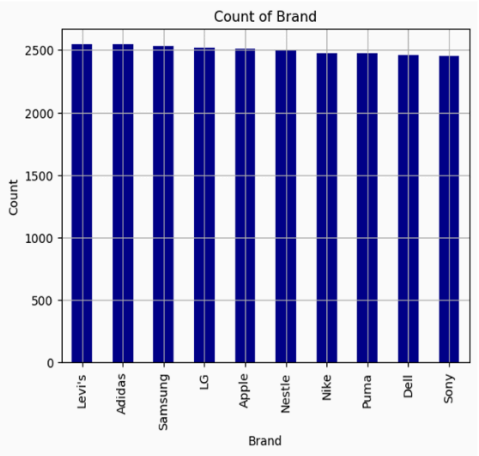
- **Product_Name** – Name of the product.
- **Category** – Broad category to which the product belongs (e.g., electronics, clothing).
- **Brand** – Brand name of the product.
- **SKU** – Stock Keeping Unit, a unique identifier used for inventory tracking.
- **Peak_Sales_Period** – The time of year when the product sells the most.
- **Sales_Channel** – The platform or medium through which the product is sold (e.g., online, in-store).
- **Top_Buying_Demographic** – The primary customer segment purchasing the product (e.g., age group, gender).

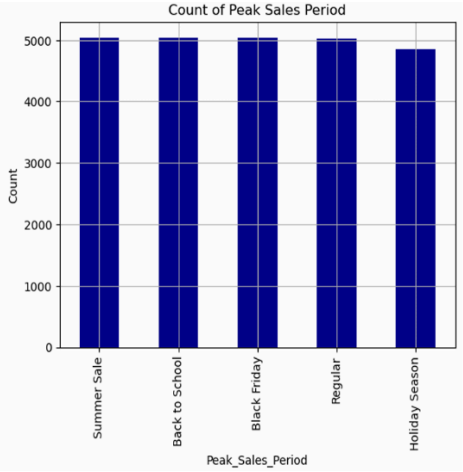
Data Analysis on Product Performance

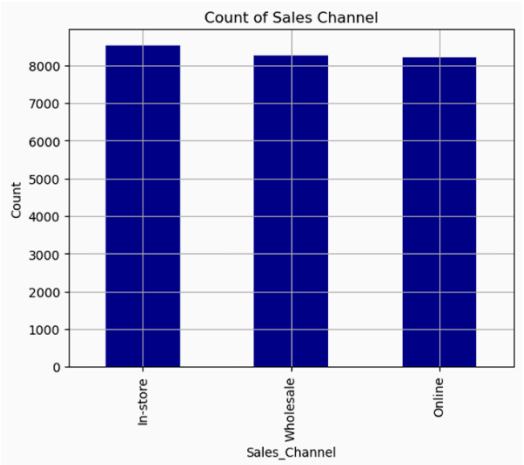
- **Cross_Selling_Performance** – Indicates how well the product is sold alongside other related products.

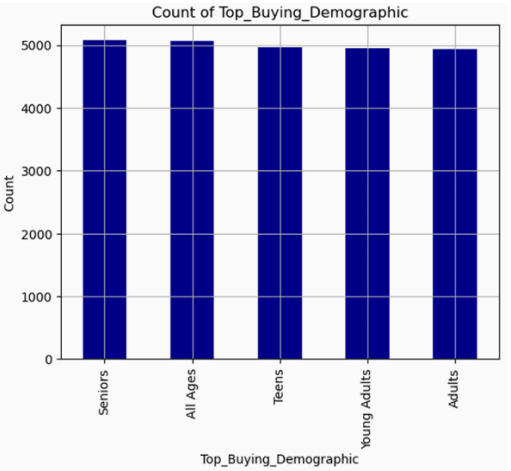












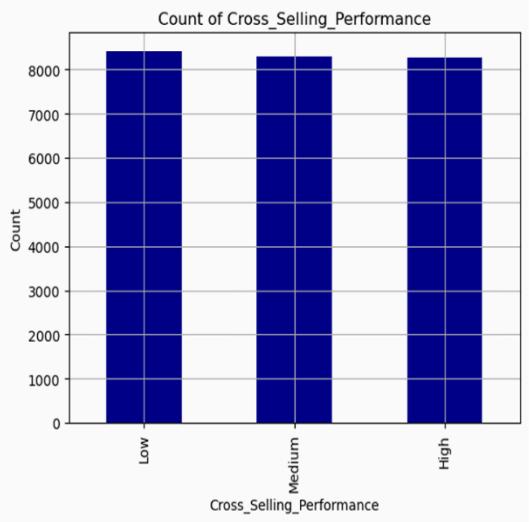


Fig 7 visualization of Categorical Columns

4.5. Summary of Results

- High-Performing Categories: Electronics and Home & Kitchen.
- Inventory Optimization: Focus on reorder levels for high-demand products.
- Customer Satisfaction: Improve product quality in the Apparel category to reduce return rates
- Marketing Strategies: Allocate more budget to online channels and seasonal campaigns.

CHAPTER 5

CONCLUSION

CHAPTER 5

CONCLUSION

The analysis of product performance data using various data analysis techniques has provided valuable insights into key factors influencing product performance. The IQR method effectively identified and removed outliers, ensuring data accuracy. Visualizations helped in understanding the data distribution and identifying trends. The findings from this analysis can help stakeholders make informed decisions to improve product sales and revenue. Future work could involve applying machine learning models to predict product performance and further optimizing marketing strategies.

ADVANTAGES:

- Improved product efficiency and reliability
- Increased user satisfaction
- Stronger market competitiveness

SCOPE:

- AI-driven predictive analysis for further optimization
- Expansion to additional product categories for broader insights

Data Analysis on Product Performance

Github Link:

https://github.com/mehrozz/Capstone-_Project-

Video Link:

https://drive.google.com/file/d/1AhrHcfz7vH47AbKZ3gbZdGEvwIGtGsOr/view?usp=drivesdk

REFERENCES

(I) **Reference analysis file** : Vacation Data – “Travel and Behavior Analysis” Provided by Prof. Praful sir , 2025.

(II) Generated Dataset: Product Performance Dataset generated via ChatGPT, 2025.

(III) Pandas Documentation: “Pandas: Data Analysis Library." Retrieved from https://pandas.pydata.org/pandas-docs/stable/

(IV) **Matplotlib Documentation**: “Matplotlib: Python Plotting Library." Retrieved from https://matplotlib.org/stable/contents.html

(V) **Seaborn Documentation:** “Seaborn: Statistical Data Visualization." Retrieved from https://seaborn.pydata.org/

(VI) **NumPy Documentation**: " NumPy: The Fundamental Package for Scientific Computing with Python." Retrieved from https://numpy.org/doc/