**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**

**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.

#### 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.

**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

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
|  |  | **Page No.** |
|  | DFD Level 0 all over process | **14** |
|  | DFD Level 1 - Data Collection and Preprocessing Module | **15** |
|  | DFD Level 1 – Exploratory Data Analysis Module | **16** |
|  | DFD Level 1 – Statistical Analysis Module | **17** |
|  | Co-relation matrix | **23** |
|  | Visualization of numerical columns | **25** |
|  | 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**

**CHAPTER 1**

**INTRODUCTION**

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. **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 SURVEYCHAPTER 2**

**LITERATURE SURVEY**

1. **Paper-1**

**Sales Forecasting and Its Impact on Business Performance**

* 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.

* 1. **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.

1. **Paper-2**

**Inventory Optimization in Retail**

* 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.

* 1. **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.

1. **Paper-3**

**Customer Behavior Analysis in E-Commerce**

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.

1. **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.

1. **Paper-4**

**Marketing Effectiveness in Digital Channels**

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.

1. **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 METHODOLOGYCHAPTER 3**

**PROPOSED METHODOLOGY**

* 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.

1. 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.

1. 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.

1. Model Development: Applying machine learning models (e.g., regression analysis, classification models) for performance prediction.

Identifying key factors influencing product success.

1. Evaluation and Insights: Assessing model accuracy and extracting actionable insights.
   1. **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.

* 1. **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).

* + 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

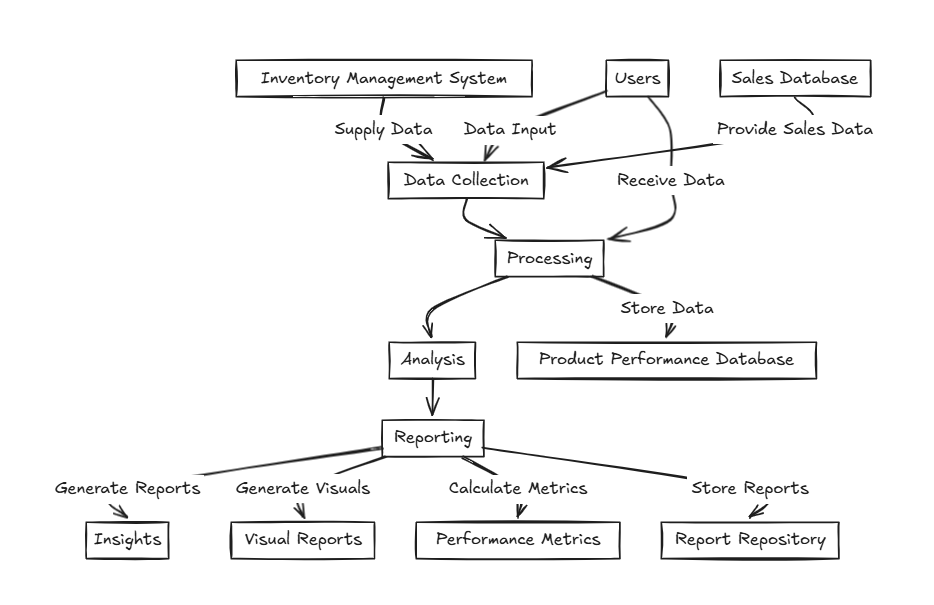


Figure1

Fig1 all over process

* + 1. **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.

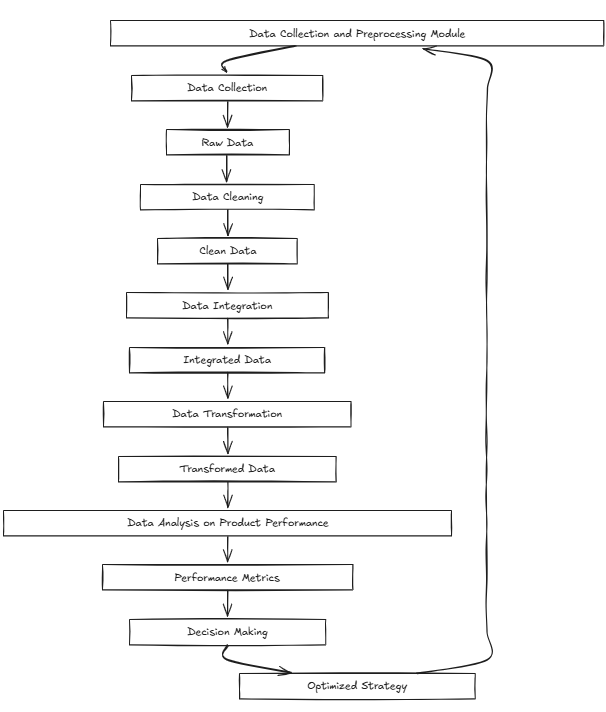


Figure2

Fig2 Data Collection and Preprocessing Module

* + 1. **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.

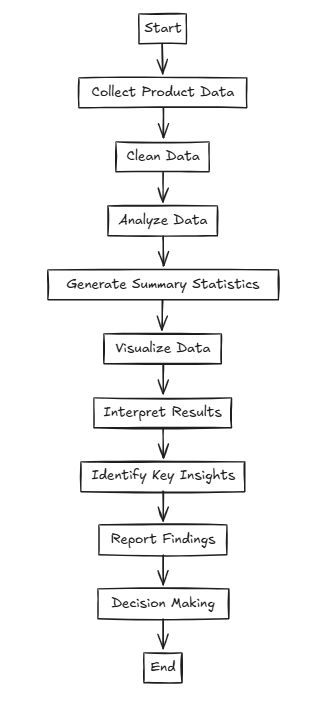


Figure3

Fig3 Exploratory Data Analysis Module

* + 1. **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.

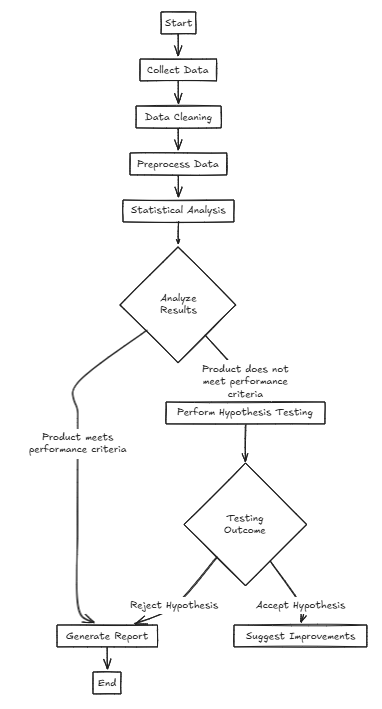
****

Figure4

**Fig4** Statistical Analysis Module

* 1. **Requirement Specification:**
     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(Colunms)

|  |  |
| --- | --- |
| 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.

Description of Columns :

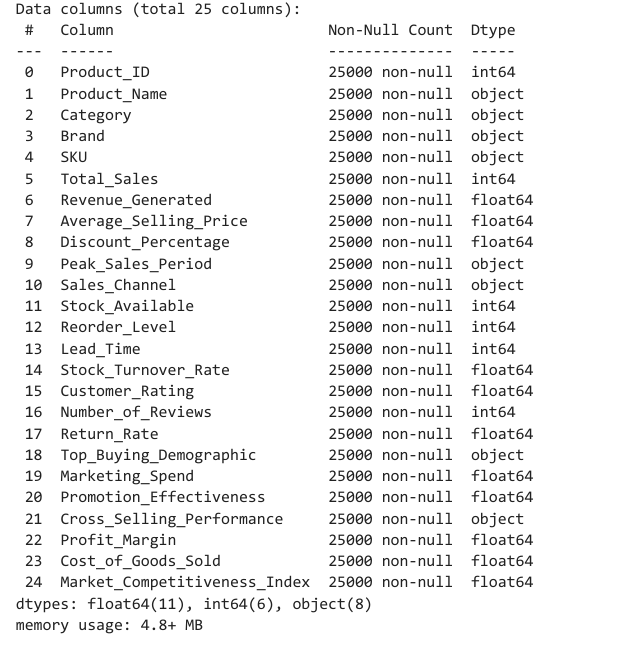


Table 2 description of columns

**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**

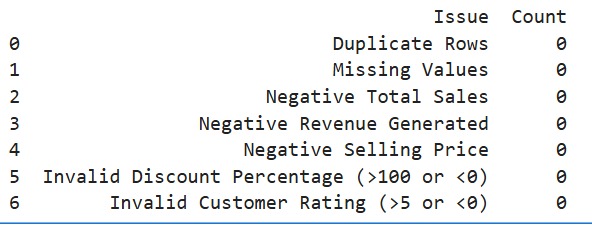
* Some products have **high return rates** and **low ratings**, indicating potential quality or customer satisfaction issues.
* Marketing spend effectiveness varies, influencing product visibility and demand.



Table 3 data description

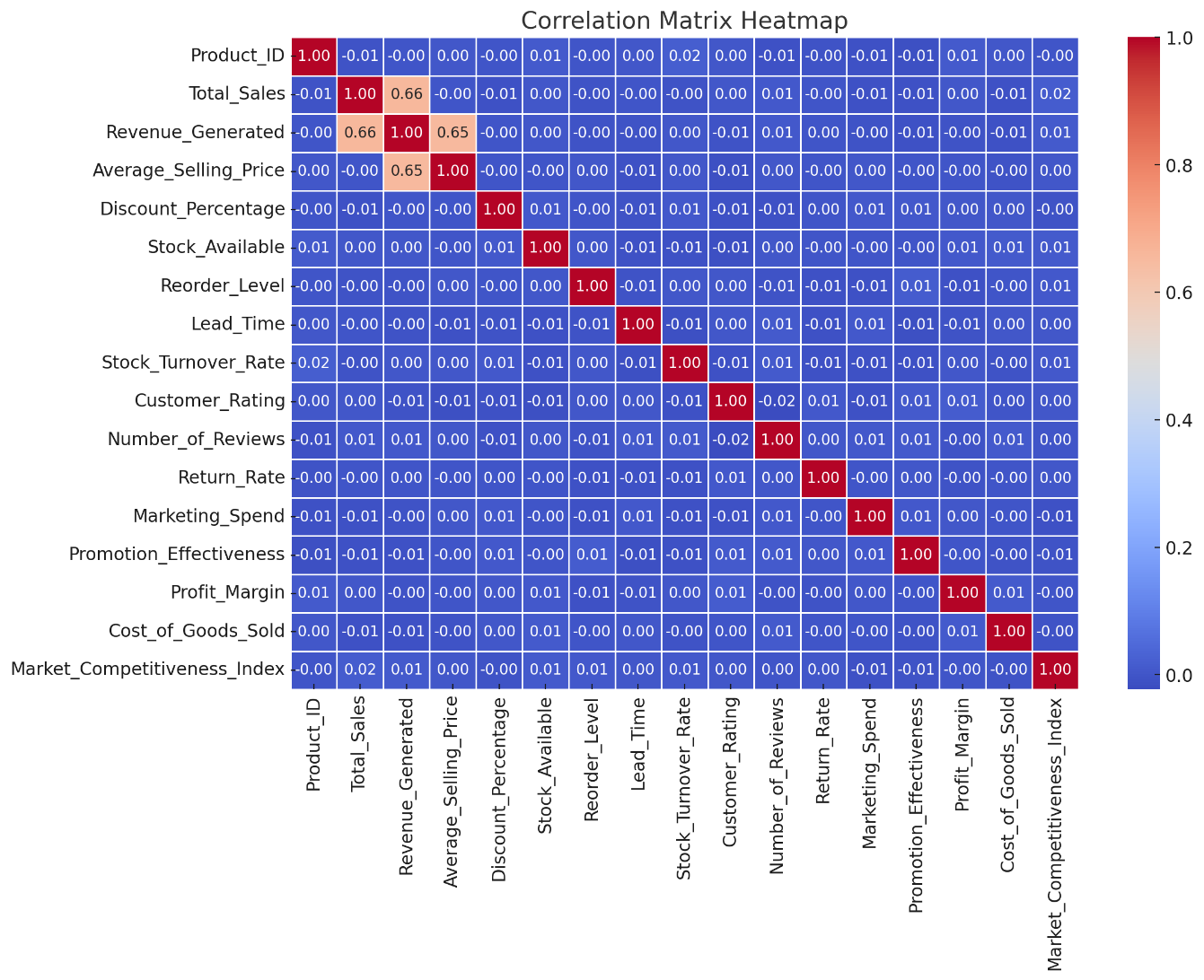
**4.2 Results of Data Cleaning:**

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

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.

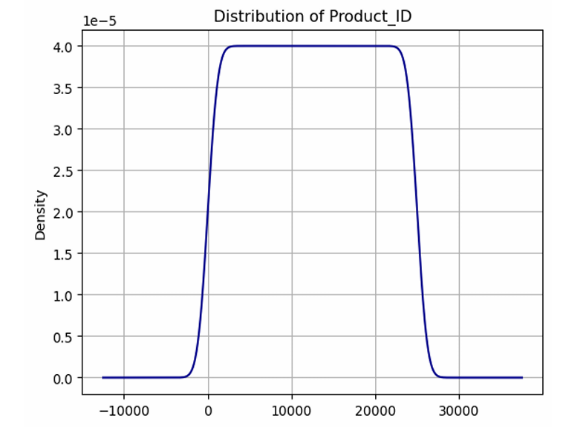
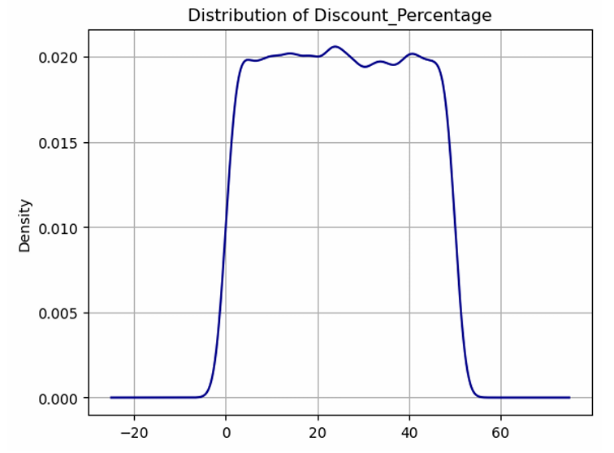
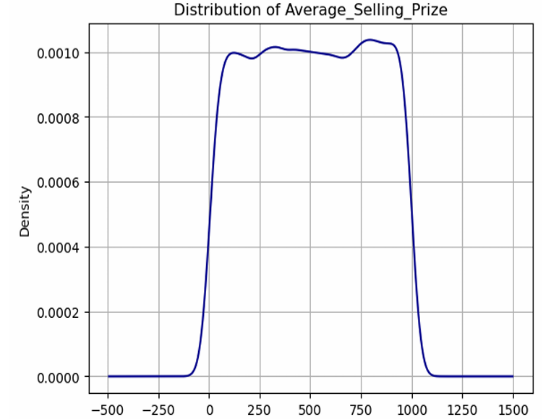
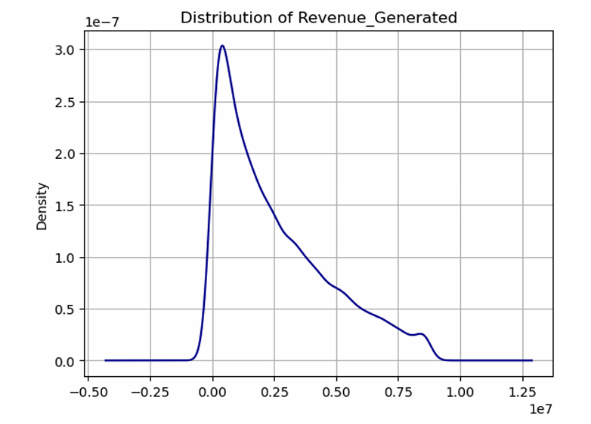
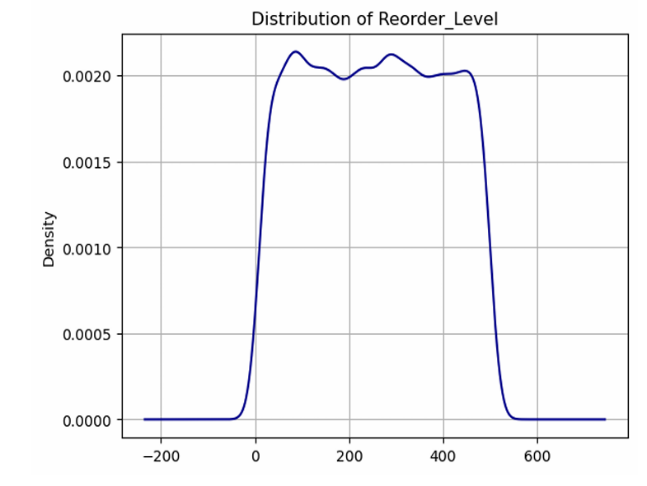
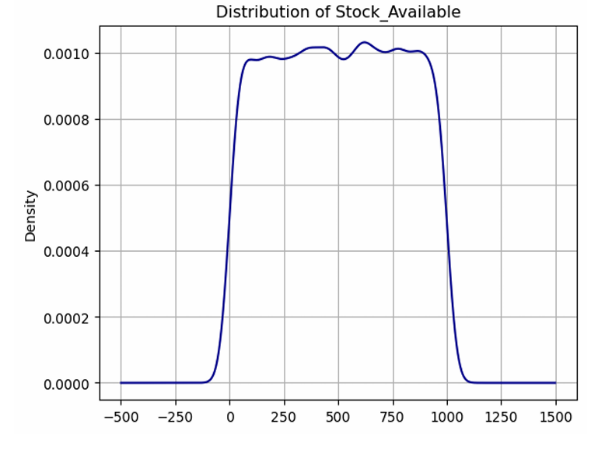
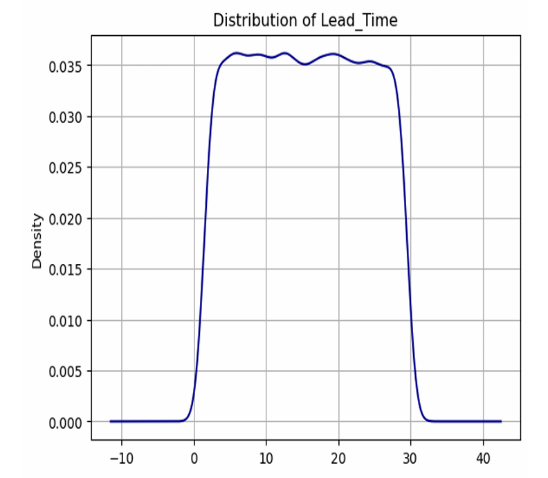
Fig 5 co-relation matrix

**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.
* **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.

A graph with a line

AI-generated content may be incorrect.   

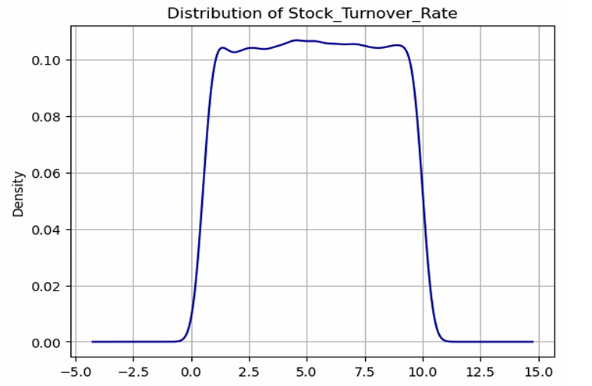
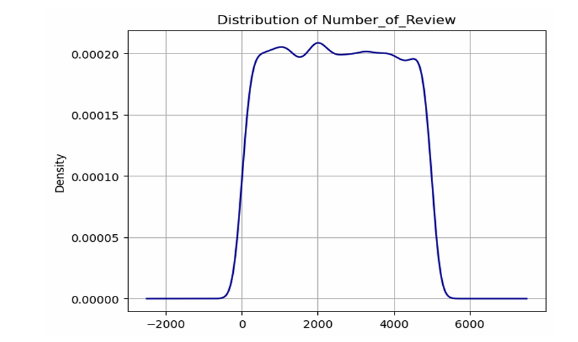
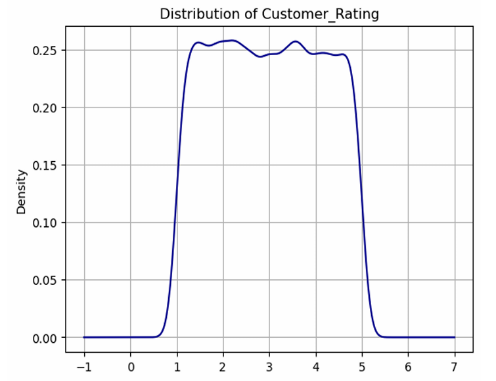
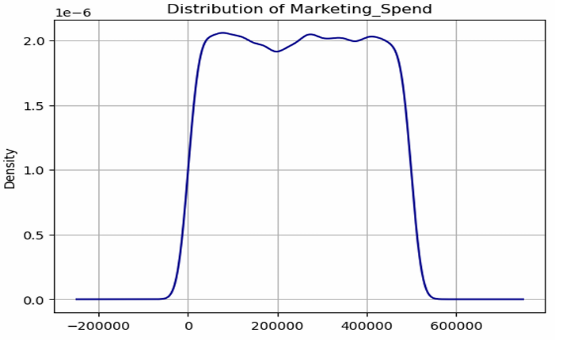
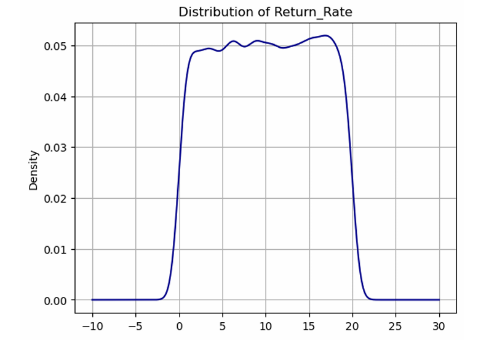
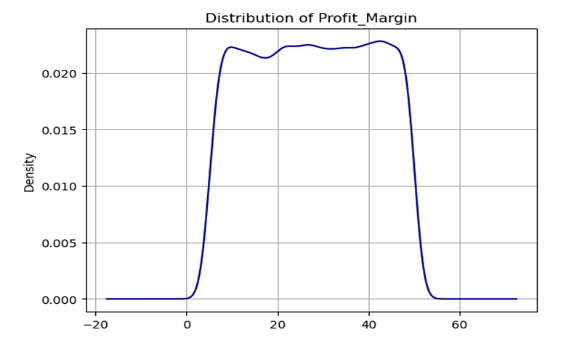
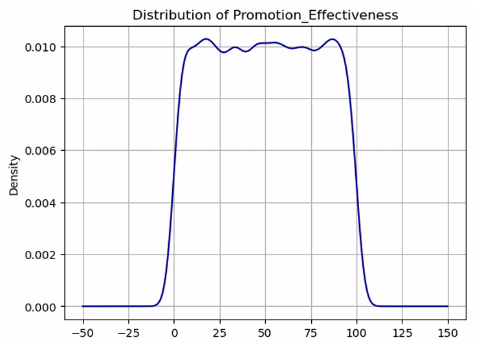
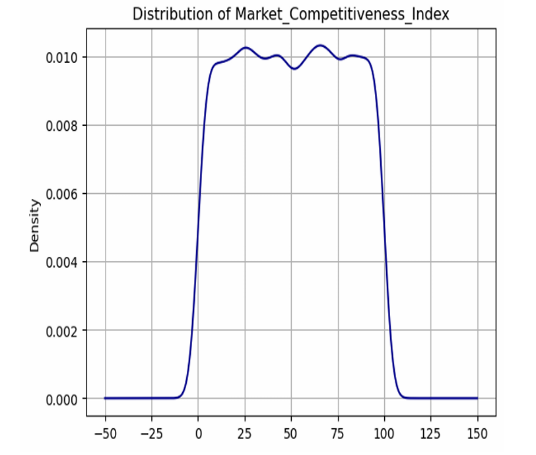
    

Fig 6 visualization of numerical columns

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).
* **Cross\_Selling\_Performance** – Indicates how well the product is sold alongside other related products.

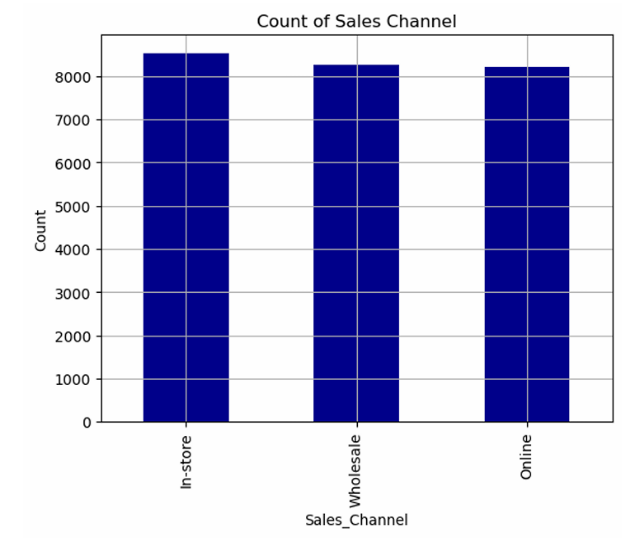
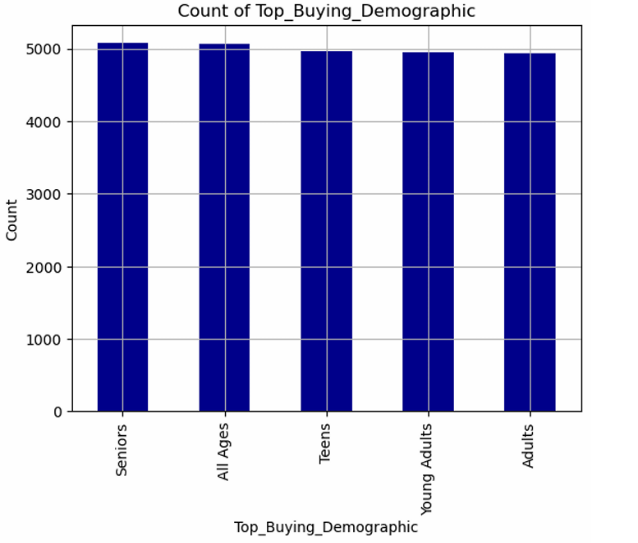
A graph of a number of products

AI-generated content may be incorrect. A graph with blue squares

AI-generated content may be incorrect.A graph of brand names

AI-generated content may be incorrect. A graph with blue bars

AI-generated content may be incorrect.

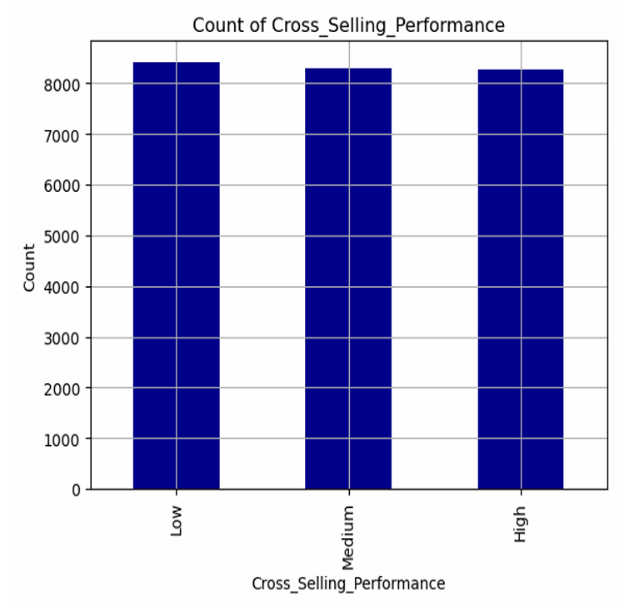


Fig 7 visualization of Categorical Columns

**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**

**CONCLUSIONCHAPTER 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

**Github Link**:

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

**Video Link**:

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

**REFERENCES**

1. **Reference analysis file** : Vacation Data – “Travel and Behavior Analysis” Provided by Prof. Praful sir , 2025.
2. Generated Dataset: Product Performance Dataset generated via ChatGPT, 2025.
3. Pandas Documentation: “Pandas: Data Analysis Library." Retrieved from https://pandas.pydata.org/pandas-docs/stable/
4. **Matplotlib Documentation**: “Matplotlib: Python Plotting Library." Retrieved from https://matplotlib.org/stable/contents.html
5. **Seaborn Documentation:** “Seaborn: Statistical Data Visualization." Retrieved from https://seaborn.pydata.org/
6. **NumPy Documentation**: " NumPy: The Fundamental Package for Scientific Computing with Python." Retrieved from https://numpy.org/doc/