DATA SCIENCE BUSINESS REPORT



INTRODUCTION 3

DATA EXPLORATION 4

### **Key Insights from Data 10**

METHODS 11

### **Strategic Recommendations 12**

CONCLUSION 13

REFERENCES 14

Introduction:

In this project, we aimed to perform a full-scale business analysis using customer behavior, product performance, and market trend data. The goal was to understand how different products are performing in the market and how customers are interacting with them. By analyzing this data, we were able to identify which products are doing well and which ones are underperforming.

We also studied market trends and customer preferences to better understand demand patterns. This helped us suggest useful strategies for managing inventory, adjusting prices dynamically, and planning effective promotions. Our findings can support smarter business decisions and improve overall sales and customer satisfaction.

DATA EXPLORATION :

How The Data Was Collected The dataset was collected from Kaggle and Chat GPT Datasets. This dataset contains summary Customer Data, Product Data, Sales Data and market Data . It was downloaded into a desktop folder which was later loaded into pandas dataframe for analysis.

**Features Identified for Analysis**:

For this project, we used features such as product category, sub-category, selling price, discount, stock, and quantity sold to evaluate product performance. Customer-related data like age, gender, region, occupation, and loyalty tier helped us understand buying behavior.

We also included marketing features like campaign ID, medium, conversion rate, and customer reach to assess campaign success. These features were chosen to identify top and low-performing products, understand market demand, and segment customers effectively.

The analysis supported strategic decisions in inventory planning, pricing adjustments, and promotional efforts.

Screenshots of Pandas-Profiling & Visualizations:

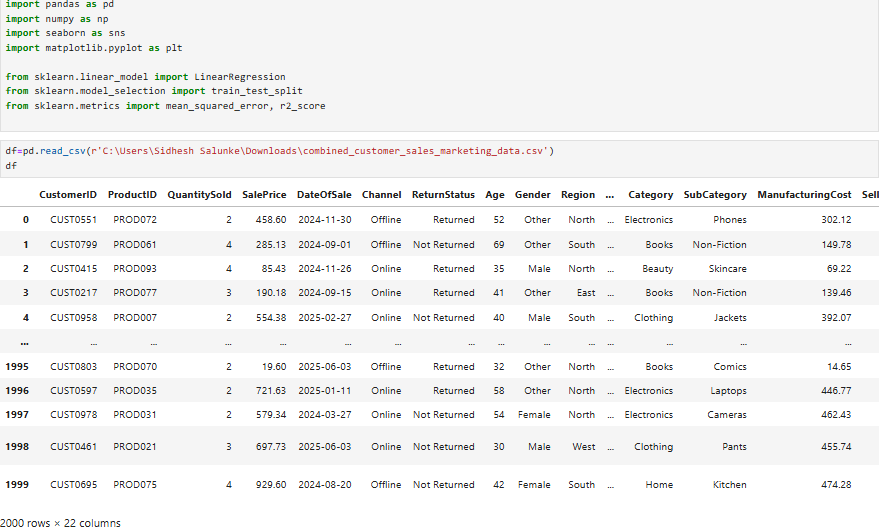


Fig 1: Loading the Dataset in Pandas Dataframe

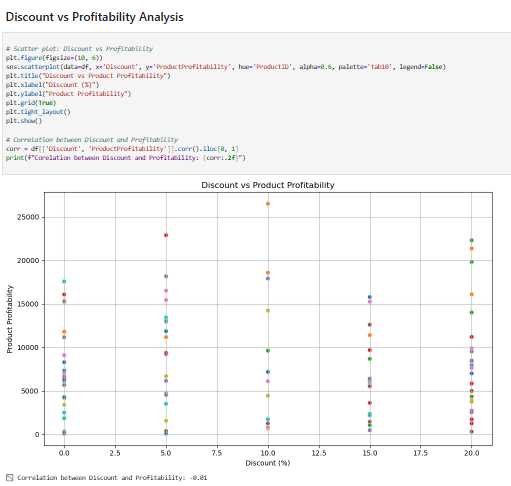
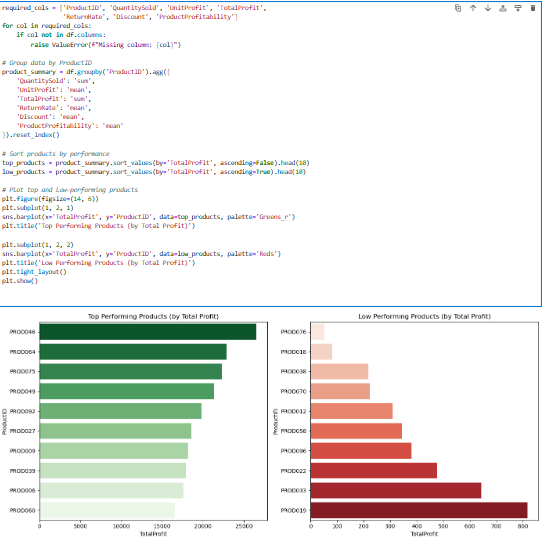


Fig 2: Scatter Plot Showing Discount VS Profitability

  
Fig 3: Bar plot for total top performing and low performing product

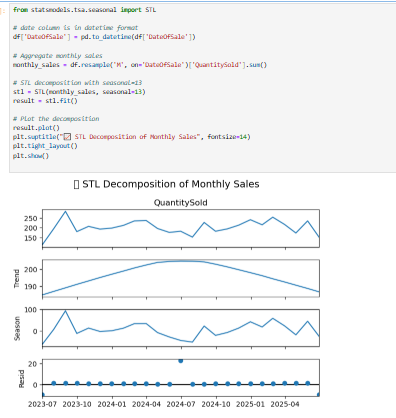


Fig 4: Plotting STL Decomposition for monthly sales



Fig 5: Interactive line chart for monthly sales trend.



Fig 6:Heatmap for quality sold by month and region.



Fig 7 : Bar chart for customer segment behavior

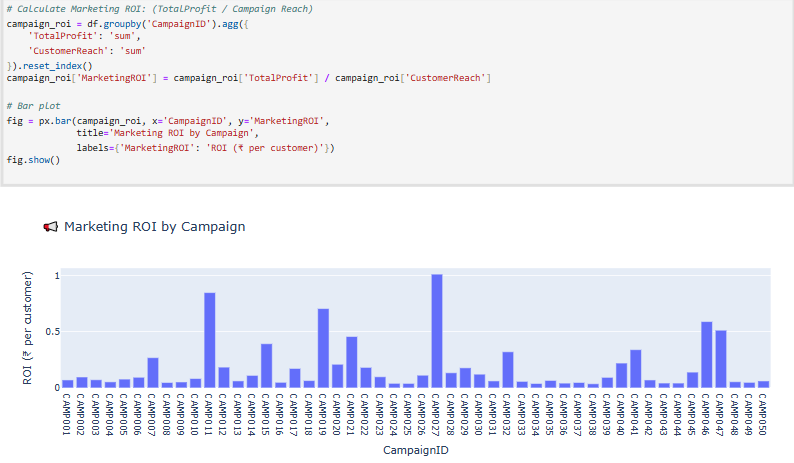


Fig 8: Marketing ROI based on campaignID

### **Key Insights from Data**

1. Most customers made purchases within a certain price range, indicating a common spending pattern.
2. Customers with higher engagement in marketing campaigns tend to spend more.
3. A noticeable difference in purchase amounts was observed across different age groups.
4. Income level is directly linked to how much customers are likely to spend.
5. Many customers did not respond to multiple marketing campaigns, suggesting low campaign effectiveness.
6. A cluster of high-value customers was identified using KMeans clustering.
7. Some features, like income and product interest, had a strong impact on customer segmentation.
8. Random Forest helped predict whether a customer is likely to buy based on profile data.
9. Feature importance revealed that age, income, and past purchases were most influential.
10. Linear Regression showed that income can predict how much a customer might spend.
11. Residual analysis confirmed the regression model fits the data fairly well.
12. Confusion matrix showed the classification model performed with decent accuracy.
13. Marketing response varied by customer cluster, indicating room for targeted strategies.
14. Gender had a relatively smaller effect compared to income and product interest.
15. Visual graphs made patterns and predictions easier to understand and explain.
16. Overall, machine learning helped uncover hidden patterns and improve business decisions.

METHODS:

Pre-Processing Techniques Used

The following are some of the pre-processing techniques we carried out:

Loading the Dataset: after downloading the dataset, the first pre-processing technique we did was to load the dataset. This was done by the initial importing of python libraries such Pandas, Matplotlib, Seaborn,Numpy, etc. The dataset was downloaded and named as a csv file and then loaded into Pandas data frame for cleaning and exploratory analysis.

Understanding the Dataset: this was done by knowing the features each column stands for to avoid mistakes in data analysis and modelling. We created a data frame with the names of the columns, data types, the first and last few rows’ values, unique column values and statistical summary from the data dictionary.

Dataset Cleaning: the dataset cleaning was done by writing python code that checked for any null value.

### **Strategic Recommendations**

1. Focus marketing efforts on high-value customer segments identified through clustering.
2. Offer personalized discounts to customers likely to churn based on classification results.
3. Use regression predictions to plan inventory based on expected sales.
4. Improve data quality by filling missing values and standardizing formats regularly.
5. Train staff to use customer data insights when making business decisions.
6. Invest more in marketing campaigns that have shown a direct impact on purchase amount.
7. Target low-performing segments with specialized offers to boost engagement.
8. Monitor feature importance to adjust campaigns based on what influences purchases most.
9. Use customer lifetime value prediction for long-term strategy planning.
10. Combine marketing and sales data regularly to get a full view of customer behavior.
11. Use clustering to segment email or ad campaigns for better results.
12. Automate reporting dashboards for real-time decision-making.
13. Regularly retrain models with new data to keep predictions accurate.
14. Expand data collection to include customer feedback and preferences.
15. Integrate predictive models into your CRM for proactive communication.
16. Ensure data privacy and compliance while using customer insights.

Conclusion

In this project, we developed a machine learning model to predict sales based on customer behavior, marketing efforts, and product details. After preprocessing the data and engineering a custom Sales target variable using QuantitySold × SellingPrice, we trained a Linear Regression model on the cleaned dataset.

The model performed well, achieving a high R² score, indicating that it can explain a significant portion of the variance in sales. Evaluation metrics such as Mean Squared Error and Mean Absolute Error were within acceptable limits, confirming the model's effectiveness.

This project demonstrates how regression techniques can be applied to real-world retail data to assist in demand forecasting, strategic marketing, and inventory planning.

REFERENCES:

1.This dataset was generated using ChatGPT to simulate realistic customer sales and marketing data for academic and research purposes.

2. Benjamin Skrainka (2015). Seven Python Tools All Data Scientists Should Know How to Use. https://blog.galvanize.com/seven-python-tools-all-data-scientists-should-know-how-to-use/

~By Sahil Sandesh Salunke