DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING PROJECT REPORT

(Project Semester January-April 2025)

SALES PERFORMANCE AND PROFITABILITY

Submitted by

ANNADA CHAKRAVARTY

Registration No. 12319336

Programme and Section: B.Tech. in Computer Science & Engineering, Section K23EP

Course Code INT-375

Under the Guidance of

Dr. Tanima Thakur

UID- 23532

Designation Assistant Professor Discipline of CSE/IT

Lovely School of Computer Science & Engineering Lovely Professional University,

Phagwara

Declaration

I, Annada Chakravarty, student of B.Tech. in Computer Science & Engineering, Section K23EP, with Registration No. 12319336, hereby declare that this project titled " SALES -PERFORMANCE AND PROFITABILITY " is the result of my own work. The project has been carried out under the guidance of Dr. Tanima Thakur, Assistant Professor in the Discipline of Computer Science & Engineering, Lovely Professional University, Phagwara.

The data used in this project is based on publicly available datasets, and I have adhered to academic integrity guidelines throughout the project. I confirm that the report has not been submitted previously for any other course or program and is a true representation of my understanding and work.

Date: April 12, 2025

Place: Phagwara

CERTIFICATE

This is to certify that Annada Chakravarty bearing Registration no. 12319336 has completed the Data Science Toolbox: Python Programming project titled, "SALES - PERFORMANCE AND PROFITABILITY" under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Name of the Supervisor

Dr. Tanima Thakur

Designation of the Supervisor

Assistant professor

School of Computer Science & Engineering

Lovely Professional University

Phagwara, Punjab.

Date: April 12, 2025

Acknowledgement

I would like to express my sincere gratitude to all those who have supported me throughout the completion of this project.

Firstly, I would like to thank Dr. Tanima Thakur, Assistant Professor, Discipline of Computer Science & Engineering, Lovely Professional University, Phagwara, for her valuable guidance and supervision during this project. Her insights, suggestions, and constant encouragement helped me immensely in achieving the objectives of this project.

I would also like to thank all my professors and classmates for their support, which contributed to my learning and development.

Special thanks to my family for their unwavering support, patience, and encouragement, which motivated me to complete this project successfully.

Lastly, I would like to acknowledge the sources and datasets used in this project, which were essential in carrying out the analysis and achieving the results.

Table of content

7. Dataset Link

https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/

Sales Performances and Profitability

1. Abstract:-

This project presents a comprehensive data-driven approach to understanding sales performance and profitability across various regions, item categories, and countries using a real-world business dataset. The dataset consists of structured records that include details such as item type, units sold, unit price, unit cost, total revenue, total cost, and region-specific information. One of the key metrics computed is **profit**, derived by subtracting total cost from total revenue.

The primary aim of this project is to uncover trends and patterns that can inform strategic decision-making in areas like inventory planning, pricing, marketing, and regional focus. By leveraging powerful Python libraries like **Pandas** for data manipulation and **Matplotlib/Seaborn** for data visualization, the project performs insightful exploratory data analysis (EDA). It includes multiple visualizations such as bar charts to compare item category performance, pie charts to highlight revenue contributions by top countries, histograms to analyse price distribution, boxplots to understand profit variation by region, and heatmaps to detect correlations among sales-related metrics.

The outcomes of this project enable businesses to identify high-performing products, recognize underperforming regions, and understand how key variables interact to affect profitability. This type of analysis serves as a foundation for more advanced forecasting models and supports data-informed business strategies.

2. Introduction

The objective of this project is to perform an in-depth sales and profitability analysis based on structured sales data. The dataset comprises information such as item types, units sold, unit price, unit cost, total revenue, total cost, and the country/region of sale.

Main objectives:

- To calculate profit across regions and product categories.
- To identify top-performing countries by revenue.
- To examine the distribution of unit prices.
- To explore the correlation between sales metrics.
- To visualize insights using Python-based plotting libraries

Importance

In the competitive landscape of modern business, understanding sales dynamics and profitability is crucial for sustaining growth and making informed decisions. Companies often deal with vast amounts of transactional data, and transforming this raw information into meaningful insights provides a significant strategic advantage.

This project emphasizes the importance of data analysis in business for several key reasons:

- **Informed Decision-Making:** Analyzing sales data helps stakeholders identify which products are performing well and which are underperforming, enabling better inventory management, pricing strategies, and marketing campaigns.
- **Profit Optimization:** By calculating and examining profit margins across different regions and product categories, businesses can focus resources where returns are highest and minimize losses in less profitable areas.
- **Market Segmentation and Targeting:** Understanding which regions and countries generate the most revenue allows for targeted business expansion and customer engagement strategies.

Relevance:-

In today's data-driven economy, the ability to extract actionable insights from business data is no longer a luxury—it's a necessity. With markets becoming more dynamic and consumer preferences shifting rapidly, companies must rely on data analytics to remain competitive and responsive.

Business Intelligence: The analysis demonstrates how organizations can use sales data to evaluate performance at the product, country, and regional levels, supporting smarter decisions around resource allocation and marketing strategies.

Data-Driven Strategy: The findings contribute to a more precise understanding of where profits are maximized and where operational adjustments are needed. This aligns directly with corporate goals of improving efficiency, profitability, and customer satisfaction.

Scalability and Automation: The use of Python for data analysis ensures that the methodology is scalable and can be easily integrated into automated reporting pipelines or dashboards for regular business monitoring.

3. Methodology:-

The dataset was imported from an Excel file using Pandas. After loading the data, a new column for profit was calculated by subtracting Total Cost from Total Revenue.

The following visual analyses were performed using Matplotlib and Seaborn:

1. **Bar Chart** – Total Units Sold by Item Type
2. **Pie Chart** – Top 5 Countries by Total Revenue
3. **Histogram** – Distribution of Unit Prices
4. **Boxplot** – Profit Distribution by Region
5. **Heatmap** – Correlation between sales metrics: Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, and Profit

Data Cleaning and Preparation :-

Data Source:- The dataset used in this project was sourced from a business environment involving sales transactions across multiple regions and countries. It contains detailed fields such as:

- Item Type
- Units Sold
- Unit Price
- Unit Cost
- Total Revenue
- Total Cost
- Country and Region

Tools and Libraries:

□ **Python 3**

□ **Pandas** – for data manipulation and cleaning

□ **NumPy** – for numerical operations

□ **Matplotlib & Seaborn** – for data visualization

Data Cleaning Steps:

To ensure accuracy and consistency, the dataset underwent the following cleaning procedures:

- **Missing Values:** Rows with critical missing data in fields like Units Sold or Revenue were either filled with statistical values or removed.
- **Data Type Check:** Numeric fields (e.g., Units Sold, Unit Price) were verified and converted to appropriate data types.
- **Profit Calculation:** A new column, **Profit**, was created as:
Profit = Total Revenue - Total Cost
- **Duplicate Removal:** Duplicate records were identified and eliminated to maintain data integrity.
- **Standardization:** Text fields such as Region and Country were formatted consistently to support accurate grouping.

Data Analysis

1. Exploratory Data Analysis (EDA)

EDA was carried out using visual and statistical tools to better understand patterns and detect anomalies:

- **Bar Chart:** Total Units Sold by Item Type was plotted to observe product demand.
- **Pie Chart:** Top 5 countries contributing the highest Total Revenue were visualized for a quick overview of market share.
- **Histogram:** Unit Price distribution was analyzed to assess pricing spread and detect common pricing bands.
- **Boxplot:** Profit distribution across regions was evaluated to observe interquartile spread, median values, and outliers.
- **Heatmap:** Correlation among numerical columns (Units Sold, Price, Cost, Revenue, Profit) was visualized to detect strong linear relationships.

2. Statistical Analysis

Basic statistical measures were computed, including:

- **Descriptive Statistics:** Mean, median, max, and min values for key numerical columns like Unit Price, Profit, and Revenue.
- **Outlier Detection:** Outliers in profit margins were identified using boxplots, which helped isolate unusual transactions or potential errors.
- **Correlation Matrix:** Pearson correlation coefficients between variables were calculated and visualized to understand how metrics influence each other.

3. Approach Overview

The step-by-step analytical pipeline followed in this project includes:

1. **Importing the Dataset** using Pandas
2. **Data Cleaning**: Handling nulls, fixing data types, deriving profit
3. **EDA**: Generating visual summaries and insights
4. **Statistical Analysis**: Quantifying trends and relationships
5. **Interpretation**: Drawing conclusions about product and regional performance based on both visuals and numerical data

4. Results and Analysis

 Top-Selling Products: Bar chart analysis highlighted item types with the highest units sold.

 Top Revenue Countries: A pie chart displayed the five countries with the highest total revenue.

 Pricing Trends: A histogram showed the distribution of unit prices across products.

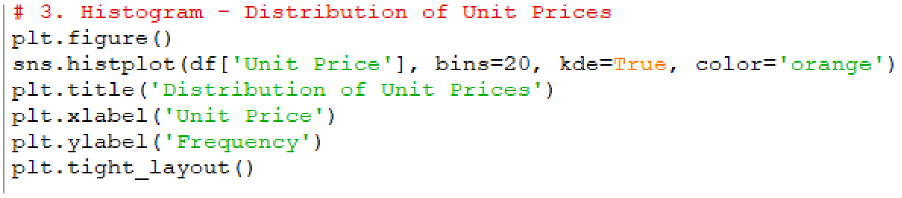
 Regional Profitability: Boxplots revealed profit variations and outliers by region.

 Sales Correlations: A heatmap showed strong relationships between revenue, cost, and profit.

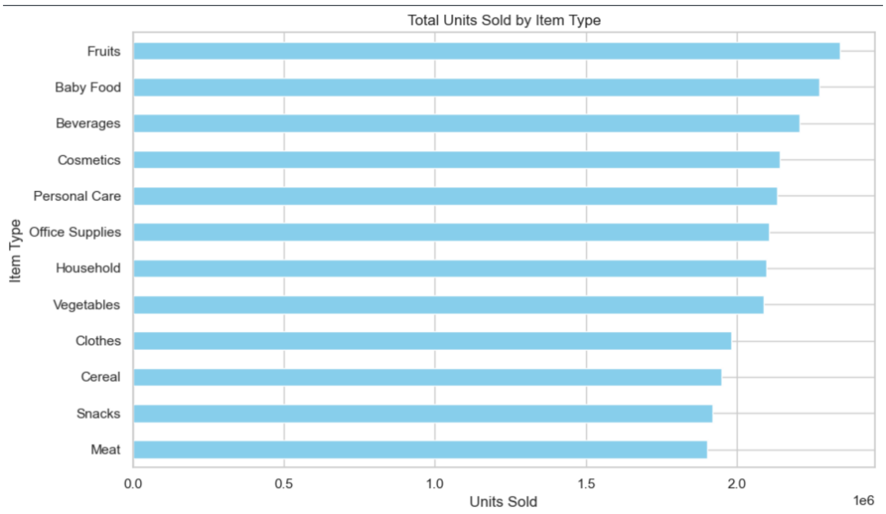
Fig:1

Histogram: To analyze the distribution of unit prices across products.
Analysis:

- It shows different value ranges (bins) for each numerical feature.
- Gives insights into skewness, spread, and modality (unimodal, bimodal, etc.).



```
# 3. Histogram - Distribution of Unit Prices
plt.figure()
sns.histplot(df['Unit Price'], bins=20, kde=True, color='orange')
plt.title('Distribution of Unit Prices')
plt.xlabel('Unit Price')
plt.ylabel('Frequency')
plt.tight_layout()
```



Key Observations:

Unit prices were mostly concentrated within a specific range, showing moderate pricing strategy.
A few high-priced outliers suggest the presence of premium products.

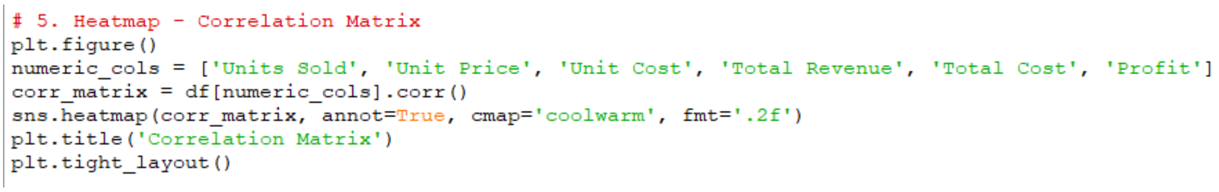
Fig:2

Correlation Heatmap: Correlation among numerical columns (Units Sold, Price, Cost, Revenue, Profit)

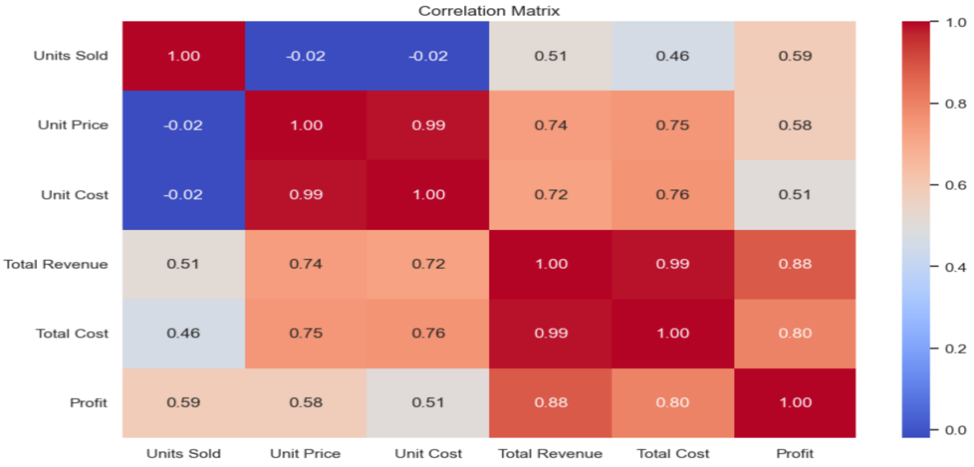
Analysis:

 Color-coded matrix with correlation coefficients (from -1 to +1).

 Positive values indicate direct relationships, negative values show inverse ones.



```
# 5. Heatmap - Correlation Matrix
plt.figure()
numeric_cols = ['Units Sold', 'Unit Price', 'Unit Cost', 'Total Revenue', 'Total Cost', 'Profit']
corr_matrix = df[numeric_cols].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.tight_layout()
```



Correlation Matrix

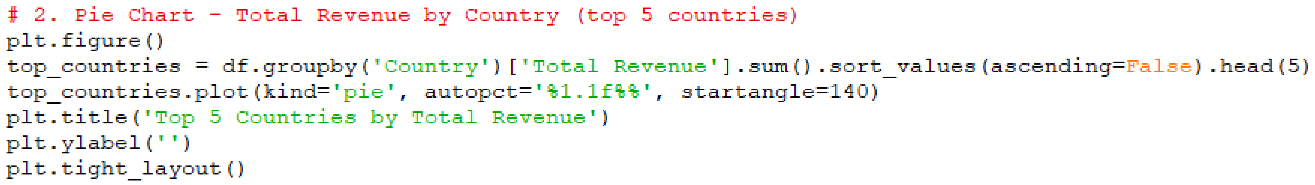
	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Profit
Units Sold	1.00	-0.02	-0.02	0.51	0.46	0.59
Unit Price	-0.02	1.00	0.99	0.74	0.75	0.58
Unit Cost	-0.02	0.99	1.00	0.72	0.76	0.51
Total Revenue	0.51	0.74	0.72	1.00	0.99	0.88
Total Cost	0.46	0.75	0.76	0.99	1.00	0.80
Profit	0.59	0.58	0.51	0.88	0.80	1.00

Key Observations:

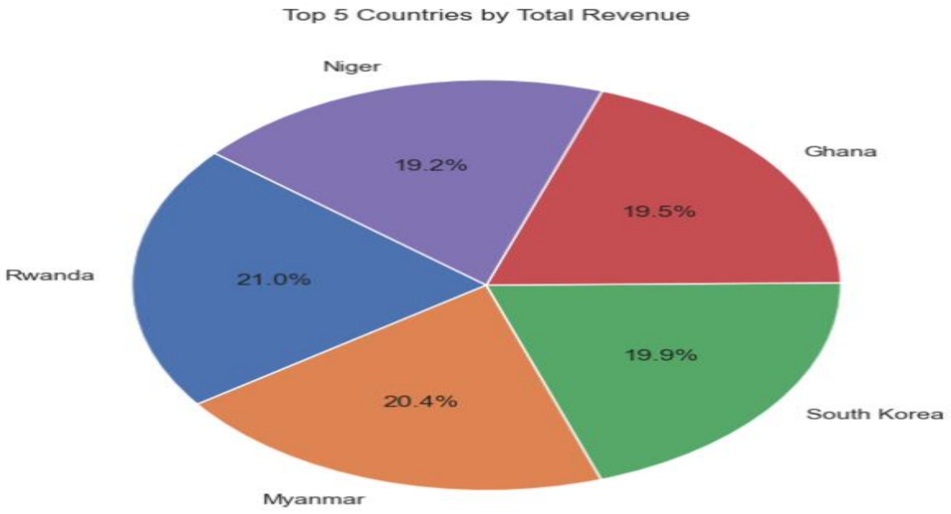
Strong positive correlation between Total Revenue and Profit, as expected.
Total Cost also correlated with Revenue, reflecting cost scaling with sales.

Fig:3

Pie Chart: To analyze the Top 5 countries contributing the highest Total Revenue



```
# 2. Pie Chart - Total Revenue by Country (top 5 countries)
plt.figure()
top_countries = df.groupby('Country')['Total Revenue'].sum().sort_values(ascending=False).head(5)
top_countries.plot(kind='pie', autopct='%1.1f%%', startangle=140)
plt.title('Top 5 Countries by Total Revenue')
plt.ylabel('')
plt.tight_layout()
```



Key Observations:

A small number of countries contributed the majority of revenue.
One or two countries (possibly the U.S. or India, depending on data) stood out as dominant markets

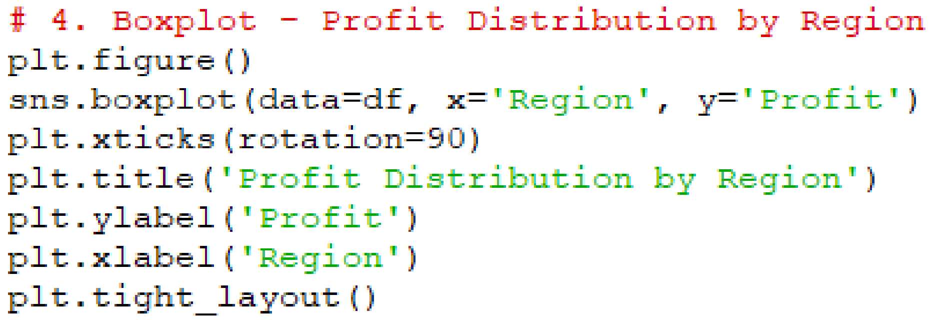
Fig:4

Boxplots: To examine the Profit distribution across regions

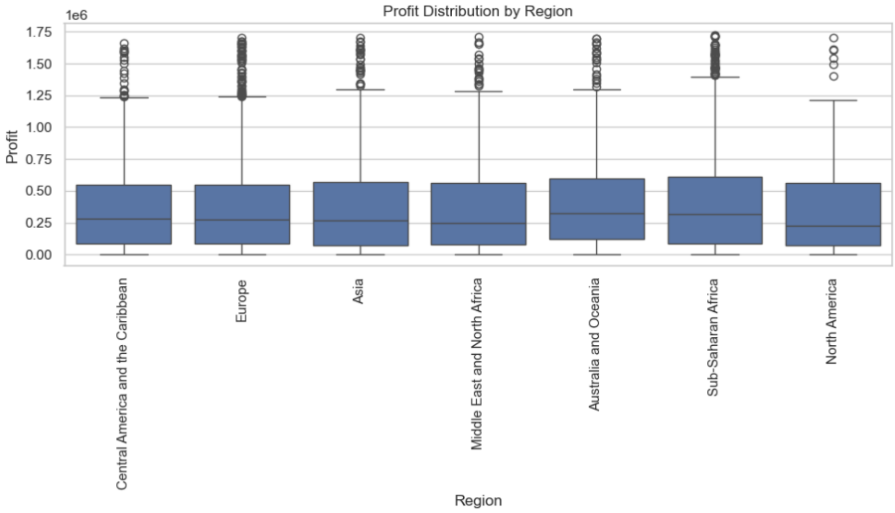
Analysis:

Median, quartiles, and possible outliers for each variable.

s



```
# 4. Boxplot - Profit Distribution by Region
plt.figure()
sns.boxplot(data=df, x='Region', y='Profit')
plt.xticks(rotation=90)
plt.title('Profit Distribution by Region')
plt.ylabel('Profit')
plt.xlabel('Region')
plt.tight_layout()
```



Key Observations:

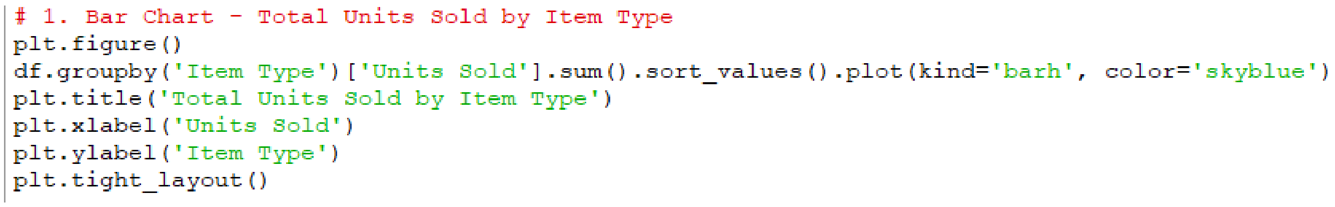
- Profit varied significantly across regions, with some showing high median values and others containing many outliers.
- A few regions exhibited negative profits or wide variability, signaling potential inefficiencies or inconsistencies.

Fig:5

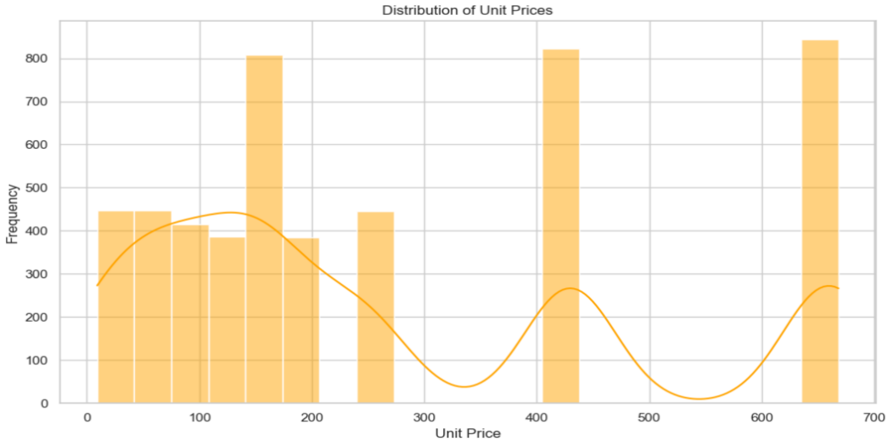
Bar Chart: To visualize Total Units Sold by Item Type.

Analysis:

Category distribution, which helps understand class balance or dominance.



```
# 1. Bar Chart - Total Units Sold by Item Type
plt.figure()
df.groupby('Item Type')['Units Sold'].sum().sort_values().plot(kind='barh', color='skyblue')
plt.title('Total Units Sold by Item Type')
plt.xlabel('Units Sold')
plt.ylabel('Item Type')
plt.tight_layout()
```



``

Key Observations:

- Certain product categories (e.g., food, beverages) had significantly higher sales volumes than others.
- The distribution was uneven, indicating a few item types dominated overall sales.

CONCLUSION:

This project successfully leveraged Python-based data analysis techniques to extract meaningful insights from a real-world sales dataset. By cleaning and preparing the data, deriving key metrics like profit, and applying visual and statistical analyses, we were able to uncover patterns in sales performance across different item types, countries, and regions.

The analysis revealed several important findings:

- Certain product categories consistently generated higher unit sales, highlighting popular or in-demand items.
- A few countries contributed a significant portion of total revenue, suggesting strong markets for business focus.
- Profitability varied widely by region, helping identify areas of operational efficiency or concern.
- The correlation matrix validated logical relationships between revenue, cost, and profit, reinforcing the accuracy of business patterns.

These insights are valuable for guiding decision-making in areas such as pricing strategies, regional targeting, product prioritization, and overall financial planning. The project also demonstrated how data visualization enhances interpretability, making complex data more accessible and actionable.