MAHARISHI MARKANDESHWAR
(DEEMED TO BE UNIVERSITY)

Mullana,Ambala,Haryana.

An Internship Report

On

“Data Science and Analytics using python ”

Along with the project

“Sales Analysis and Forecasting for a Retail Store”

Submitted by:

Kiranbir singh

11222569

Under the Guidance of

Bikash bashyal,

Internship Co-coordinator at Invecareer

DECLARATION

I hereby declare that this project report titled "**Sales Analysis and Forecasting for a Retail Store**" is a genuine and original work completed by Kiranbir Singh, student of V semester in B.tech computer science at Maharishi Markandeshwar(Deemed to be University). This project was undertaken as part of the requirements for providing actionable insights and strategic recommendations to the given management of Retail Store Chain .

Acknowledgement

I would like to express my sincere gratitude to everyone who has supported and contributed to the successful completion of this project, titled "**Sales Analysis and Forecasting for a Retail Store.**"

First and foremost, I extend my heartfelt thanks to **Bikash bashyal**,Internship Co-coordinator at Invecareer, for their invaluable guidance, encouragement, and insightful feedback throughout the duration of this project. Their expertise and support have been instrumental in shaping the direction and scope of this work.

I am also grateful to the management team at Retail Store Chain Name for providing the opportunity to undertake this project and for granting access to the necessary resources and data. Their cooperation and trust have been crucial in making this project possible.

Special thanks to my colleagues and team members at Maharishi Markandeshwar (Deemed to be University) who have offered their assistance, shared their knowledge, and provided constructive feedback during the various stages of this project. Their collaboration and camaraderie have greatly enriched this experience

I would like to acknowledge the contributions of the academic and professional resources that have informed and inspired this work. The literature and methodologies referenced in this project have provided a strong foundation for the analysis and forecasting undertaken.

TABLE OF CONTENTS

PROBLEM STATEMENT:

Sales Analysis and Forecasting for a Retail Store

Introduction:

In the highly competitive retail industry, data-driven decision-making is pivotal for sustaining growth and maintaining a competitive edge. Retail stores accumulate vast amounts of data daily, encompassing sales transactions, customer behavior, inventory levels, and more. Properly analyzing this data can unveil valuable insights into sales trends, customer preferences, and market dynamics. By leveraging historical sales data, retail stores can not only optimize their current operations but also forecast future sales trends, enabling proactive management of inventory, pricing strategies, and marketing campaigns.

The primary goal of this project is to assist the management of a retail store chain in making informed decisions through comprehensive sales analysis and accurate forecasting. This project aims to delve into historical sales data to uncover patterns, identify top-selling products, and predict future sales. Such insights are crucial for optimizing inventory management, enhancing pricing strategies, and designing effective marketing campaigns. Ultimately, this data-driven approach will support the store's objective to maximize revenue, minimize costs, and improve customer satisfaction

Assumptions :

Accurate sales analysis and forecasting play a crucial role in the success of retail operations. By understanding sales patterns and predicting future trends, retail stores can achieve several benefits:

- **Data Assumptions:** Ensuring that the right amount of inventory is available at the right time helps in avoiding stockouts and overstock situations. This leads to improved customer satisfaction and reduced holding costs.

-**Business Assumptions:** Analyzing the relationship between sales and pricing helps in setting optimal prices that maximize revenue while remaining competitive in the market.

- **Forecasting Assumptions:** Identifying top-selling products and seasonal trends allows for the design of targeted marketing campaigns that can drive sales and attract more customers.

- **Informed Decision-Making:** Data-driven insights enable the management to make informed decisions, reducing reliance on intuition and guesswork.

Analysis:

The specific objectives of this project are outlined as follows:

1. Data Collection: Obtain or simulate a comprehensive dataset containing historical sales data. The datasets should include essential information such as the date of sale, product ID, quantity sold, price per unit, and product category.

2. Data Pre-processing: Clean and pre-process the collected data to ensure its accuracy and suitability for analysis. This involves handling missing values, removing duplicates, and performing necessary data transformations, such as calculating the total sales amount for each transaction.

3. Exploratory Data Analysis (EDA): Conduct an in-depth exploratory data analysis to gain insights into the sales data. This phase involves exploring sales trends over time, identifying seasonal patterns, examining the correlation between sales and other variables (e.g., price, product category), and determining the top-selling products or categories.

4. Visualization: Create visualizations to effectively communicate the findings from the EDA phase. Utilize plots and charts to represent sales trends, top-selling products, and relationships between different variables.Visualizations will aid in presenting data-driven insights to the management in a comprehensible manner.

5. Sales Forecasting: Develop a sales forecasting model to predict future sales based on historical data. Begin with basic forecasting methods such as moving averages and exponential smoothing. Evaluate the performance of the forecasting model using appropriate metrics to ensure accuracy.

6. Recommendations: Provide actionable recommendations to the management based on the analysis and forecasting results. These recommendations will focus on optimizing inventory management, refining pricing strategies, and planning targeted marketing campaigns to boost sales performance.

Findings:

The project will be executed through a structured approach, detailed as follows:

1. Data Collection:

- Simulation: In the absence of actual sales data, simulate a dataset containing historical sales information. Ensure the dataset includes key attributes such as the date of sale, product ID, quantity sold, price per unit, and product category.
- **Integrity:** Verify the dataset's integrity by ensuring it reflects realistic sales scenarios and contains no synthetic biases.

2. Data Preprocessing:

- **Cleaning:** Handle missing values appropriately by either imputing them or excluding the affected records. Remove duplicate records to maintain data quality.
- **Transformation:** Convert data types where necessary, such as ensuring dates are in the proper datetime format. Calculate the total sales amount for each transaction to facilitate further analysis.

3. Exploratory Data Analysis (EDA):

- **Trends and Patterns:** Analyze sales trends over time to identify growth patterns and seasonal variations.
- **Correlation Analysis:** Examine the relationship between sales and other variables, such as price and product category, to identify factors influencing sales performance.
- **Top-Selling Products:** Identify products and categories with the highest sales volumes, providing insights into customer preferences and demand.

4. Visualization:

- **Tools:** Utilize Matplotlib and Seaborn to create line plots, bar charts, and scatter plots.
- **Presentation:** Visualize sales trends, top-selling products, and correlations between variables to effectively communicate insights to the management team.

5. Sales Forecasting:

- **Model Selection:** Implement basic forecasting models such as moving averages and exponential smoothing to predict future sales.
- **Evaluation:** Assess the performance of the forecasting models using metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to ensure accuracy and reliability.

6. Recommendations:

- **Inventory Management:** Suggest strategies to optimize inventory levels based on sales forecasts, ensuring adequate stock availability and minimizing holding costs.
- **Pricing Strategies:** Recommend pricing adjustments based on the correlation between price and sales, aiming to maximize revenue.
- **Marketing Campaigns:** Propose targeted marketing campaigns focused on top-selling products and seasonal trends to boost sales and attract more customers.

Conclusion:

Upon completion of this project, the retail store management will receive a comprehensive report that includes:

- Key Insights: Detailed findings from the historical sales data analysis,
highlighting trends, patterns, and top-performing products.

- **Visual Representations:** Clear and informative visualizations that depict sales trends, top-selling products, and relationships between different variables.
- **Sales Forecasts:** Accurate predictions of future sales to assist in inventory planning and decision-making.
- **Strategic Recommendations:** Actionable insights and suggestions for optimizing inventory management, refining pricing strategies, and planning effective marketing campaigns.

These outcomes will empower the management team to make data-driven decisions, optimize operations, and ultimately enhance the store's overall sales performance. By leveraging the power of data analytics and forecasting, the retail store can improve its competitiveness and ensure sustained growth in the dynamic retail market.

Suggestions:

-Sales Forecasting

1. Model Selection:Simple Models: Start with simple models like Moving Average, Exponential Smoothing.Advanced Models: Use ARIMA, SARIMA, Prophet for more accurate forecasting.
2. Model Validation:Train-Test Split: Split the data into training and testing sets to validate the model.Cross-Validation: Use cross-validation techniques to ensure model robustness.
3. Model Evaluation:Error Metrics: Use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE) to evaluate model performance.Residual Analysis: Analyze residuals to check for any patterns or auto correlations.
4. Incorporating External Factors: Regressors: Include external factors (e.g., holidays, economic indicators) as regressors in the forecasting model.Scenario Analysis: Conduct scenario analysis to understand how different factors might impact future sales.

-Business Insights and Actions

1. Inventory Management:Demand Forecasting: Use sales forecasts to optimize inventory levels and reduce stockouts or overstock situations.Safety Stock: Calculate safety stock levels based on forecast accuracy and variability.
2. Marketing Strategy:Targeted Promotions: Use insights from customer segmentation and product performance to design targeted promotions.Seasonal Campaigns: Plan marketing campaigns around identified seasonal peaks.
3. Pricing Strategy:Dynamic Pricing: Implement dynamic pricing strategies based on price sensitivity analysis.Discount Strategies: Optimize discount strategies to maximize sales without eroding margins.
4. Customer Retention:Loyalty Programs: Design loyalty programs based on customer purchase behavior and segmentation analysis.Personalization: Use personalized marketing to enhance customer experience and retention.

Language and Tools Used:

For sales analysis and forecasting in Python, several libraries are typically used. Below is a detailed explanation of the language and libraries involved in this process:

Language: Python

Python is a versatile and widely-used programming language in data analysis, due to its ease of use, readability, and extensive support for data manipulation and analysis libraries.

Libraries Used:

1. Pandas: Purpose: data manipulation and analysis
2. Matplotlib: Purpose: Data visualization.
3. Matplotlib: Purpose: Data visualizatio.
4. Statsmodels: Purpose: Statistical modeling.

Platforms used:

1. Jupyter Notebook:

it is used to perform code and data information to analyse the data and best use for perfoming code online.

2. Kaggle:

it is used for large data set online for project. As it is a best platform for practising data analysis online. It is a prominent online community and platform for data scientist and machine learning practitioner.

Project:

-Data Collection:

Dataset (name = “store”)will be loaded with the help of kaggle

Code:

```
import pandas as pd
import numpy as np
store = pd.read_csv("/kaggle/input/superstore-sales/superstore_final_dataset (1).csv",encoding='latin-1')
store.head()
```

Output:

	Row_ID	Order_ID	Date	Ship_Date	Customer_ID	Customer_Name	Segment	Country	City	State	Postal_Code	Region	ProductID	Category	Product_Name	Price	Quantity	Totalsales
0	1	CA-2017-152156	8/11/2017	11/11/2017	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-BO-10001798	Furniture	Bush Somerset Collection Bookcase	261.9600	2	1047.840
1	2	CA-2017-152156	8/11/2017	11/11/2017	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-CH-10000454	Furniture	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400	2	2927.760
2	3	CA-2017-138688	12/6/2017	16/06/2017	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036.0	West	OFF-LA-10000240	Office Supplies	Self-Adhesive Address Labels for Typewriters b...	14.6200	2	58.480
3	4	US-2016-108966	11/10/2016	18/10/2016	SO-20335	Sean O Donnel	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	FUR-TA-10000577	Furniture	Bretford CR4500 Series Slim Rectangular Table	957.5775	2	3830.310
4	5	US-2016-108966	11/10/2016	18/10/2016	SO-20335	Sean O Donnel	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	OFF-ST-10000760	Office Supplies	Eldon Fold N Roll Cart System	22.3680	2	89.472

-Basic inspection in Data:

1.New(dataset name= “data”)will be extract from these:

Code:

```
data = store[['Date','ProductID',"Price",'Quantity','Totalsales']]
data.head()
```

	Date	ProductID	Price	Quantity	Totalsales
0	8/11/2017	FUR-BO-10001798	261.9600	2	261.9600
1	8/11/2017	FUR-CH-10000454	731.9400	2	731.9400
2	12/6/2017	OFF-LA-10000240	14.6200	2	14.6200
3	11/10/2016	FUR-TA-10000577	957.5775	2	957.5775
4	11/10/2016	OFF-ST-10000760	22.3680	2	22.3680

2.Attributes names in “data”

Code:

```
#columnsofdataframe
print(data.columns)
```

```
Index(['Date', 'ProductID', 'Quantity', 'Price', 'Totalsales'], dtype='object')
```

3.Number of rows and columns

Code:

```
#displaytheshapeofthedataframe
print(data.shape)
#intheformof(no.ofrows,no.ofcolumn)
```

```
(1552, 5)
```

4.Last 5 rows

Code:

```
#showsthelastfewrows
print(data.tail())
```

```
           Date  ProductID  Quantity      Price  Totalsales
1547 2024-03-27         60         2  12.677055   25.354109
1548 2024-03-28         18         4  85.190967  340.763867
1549 2024-03-29         69         1  31.603542   31.603542
1550 2024-03-30         61         2  39.981810   79.963621
1551 2024-03-31         26         1  30.964728   30.964728
```

5.information about dataset

Code:

```
#displaytheinforamtionabout the column
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1552 entries, 0 to 1551
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Date        1552 non-null   datetime64[ns]
 1   ProductID   1552 non-null   int32
 2   Quantity    1552 non-null   int32
 3   Price       1552 non-null   float64
 4   Totalsales  1552 non-null   float64
dtypes: datetime64[ns](1), float64(2), int32(2)
memory usage: 48.6 KB
None
```

-Data preprocessing:
Next, we'll clean the dataset. This includes handling missing values, removing duplicates, and converting data types if necessary.

Code:

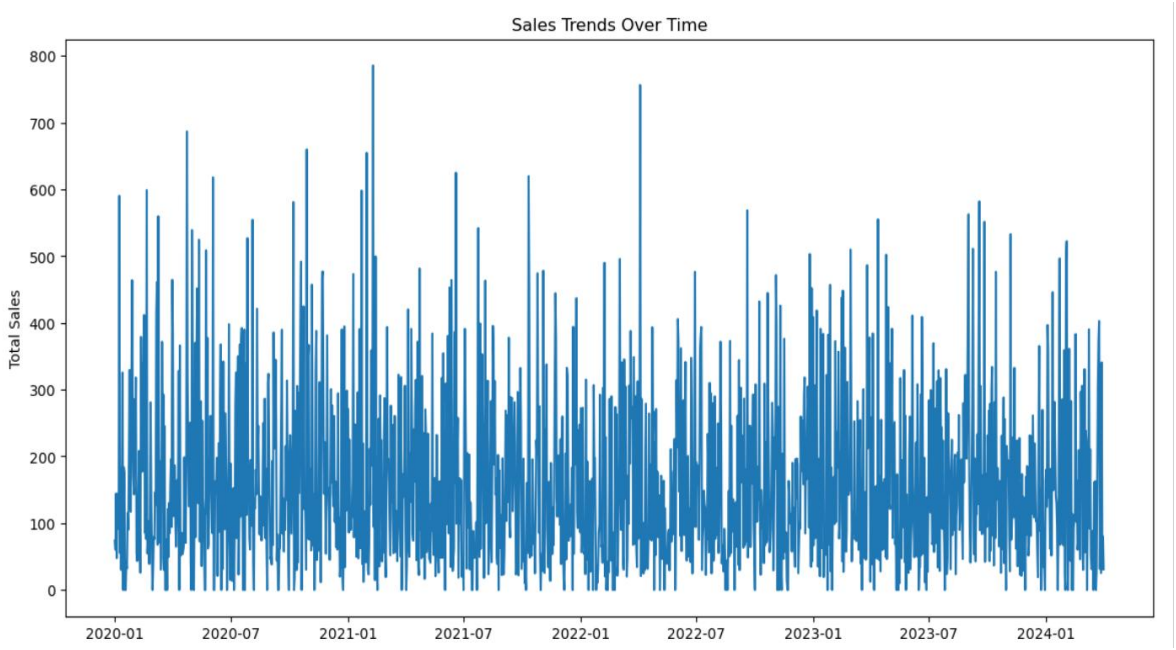
```
#dataprepocessing
#usedformissingvalue
data.isnull().sum()
#usedwhereduplicatevale
data.duplicated().sum()
```

-Exploratory Data Analysis (EDA):
We'll conduct EDA to gain insights into the sales data. This includes exploring trends over time, seasonality in sales, correlations, and identifying top-selling products or categories.
-Visualization:
Is to present findings from the EDA phase using Matplotlib.This Could include:

1. Line plots to visualize sales trends over time:
Code:

```
#sales over trends
import matplotlib.pyplot as plt
sales_trends=data.groupby('Date').agg({'Totalsales':'sum'}).reset_index()
# Line plot for sales trends over time
plt.figure(figsize=(14, 7))
plt.plot(sales_trends['Date'], sales_trends['Totalsales'])
plt.title('Sales Trends Over Time')
plt.xlabel('Date')
plt.ylabel('Total Sales')
plt.show()
```

Output:

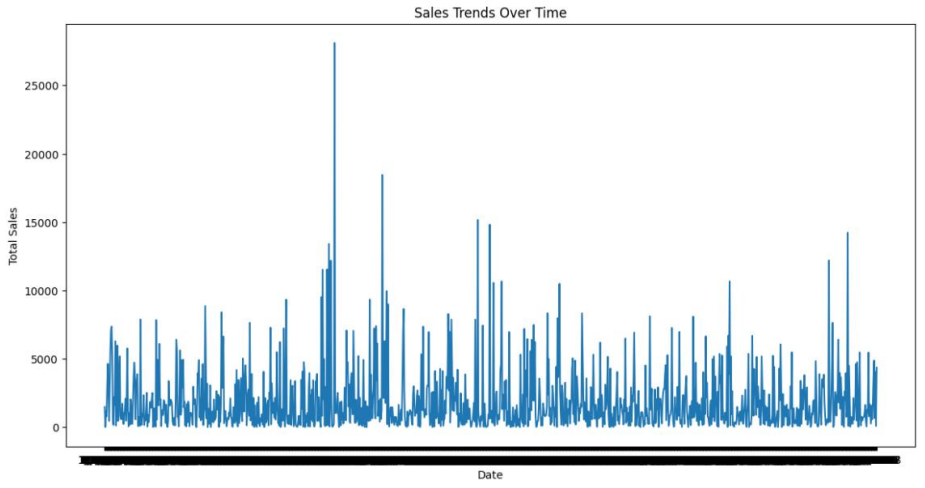


2. Line plots to show monthly sales trends

Code:

```
#monthly sales trends
# Aggregate total sales by month
data['Month'] = data['Date'].dt.to_period('M')
monthly_sales = data.groupby('Month').agg({'Totalsales': 'sum'}).reset_index()
# Convert Month to datetime for plotting purposes
monthly_sales['Month'] = monthly_sales['Month'].dt.to_timestamp()
# Line plot for monthly sales trends
plt.figure(figsize=(14, 7))
plt.plot(monthly_sales['Month'], monthly_sales['Totalsales'])
plt.title('Monthly Sales Trends')
plt.xlabel('Month')
plt.ylabel('Total Sales')
plt.show()
```

Output:

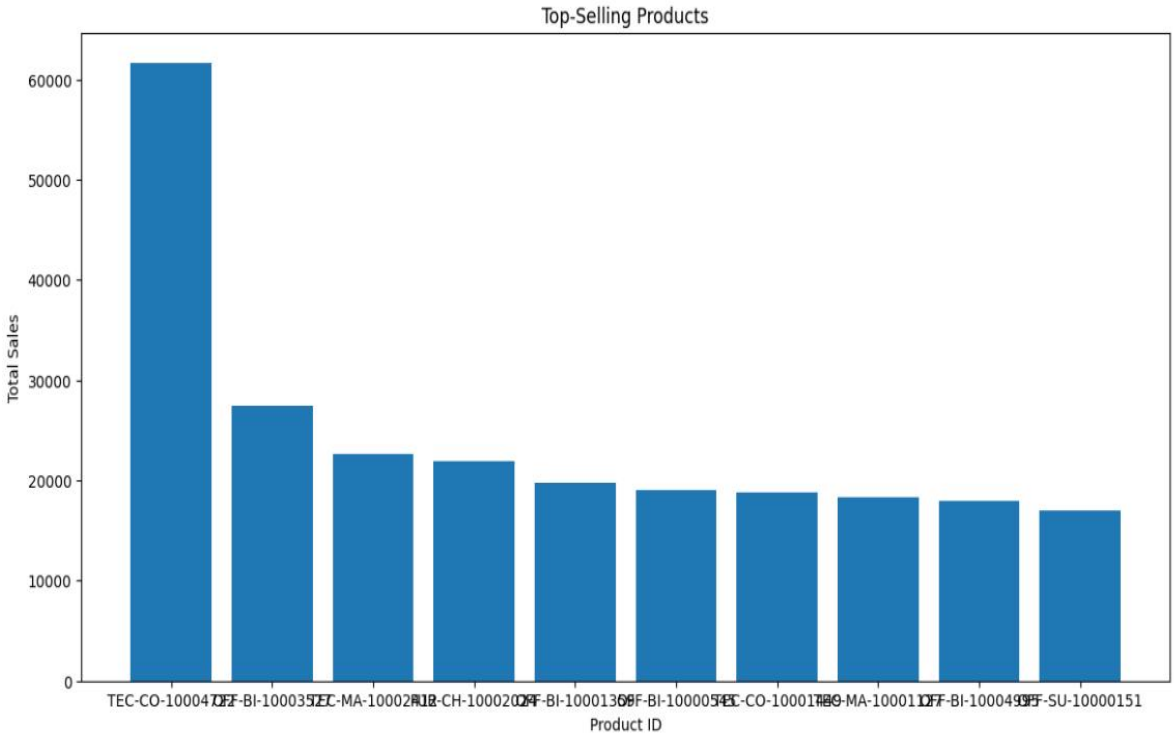


3. Bar plots to show top-selling products

Code:

```
# Aggregate total sales by product
top_products = data.groupby('ProductID').agg({'Totalsales': 'sum'}).sort_values(by='Totalsales', ascending=False).head(10).reset_index()
# Bar plot for top-selling products
plt.figure(figsize=(14, 7))
plt.bar(top_products['ProductID'], top_products['Totalsales'])
plt.title('Top-Selling Products')
plt.xlabel('Product ID')
plt.ylabel('Total Sales')
plt.show()
```

Output:

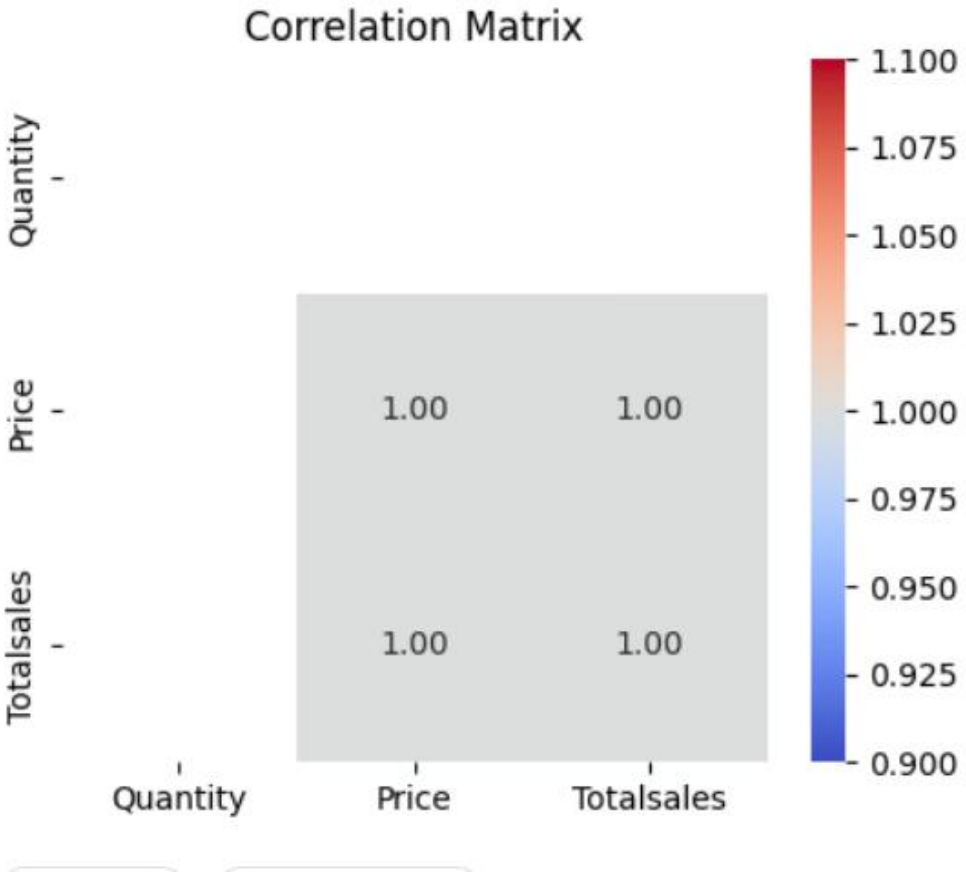


4. Scatter plots to explore relationship between variables.

Code:

```
import seaborn as sns
# Compute the correlation matrix
correlation_matrix = data[['Quantity', 'Price', 'Totalsales']].corr()
# Heatmap for the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```

Output:


Correlation Matrix
Quantity
Price
Totalsales
1.00
1.00
1.00
1.00
Quantity
Price
Totalsales
1.100
1.075
1.050
1.025
1.000
0.975
0.950
0.925
0.900