PHASE 2 SUBMISSION-INNOVATION OF PRODUCT SALES ANALYSIS

Introduction:

In today's competitive business landscape, data-driven decision-making is imperative for success. The "Sales Analysis for Enhanced Business Performance" project is designed to harness the power of data analytics, utilizing IBM Cognos, to gain deep insights into sales performance and customer behavior. This project aims to equip businesses with the tools and knowledge needed to optimize inventory management and refine marketing strategies, ultimately leading to improved profitability and customer satisfaction.

Data Processing and Analysis Plan

Step 1: Data Cleaning and Preparation

Checking for Missing or Null Values:
In this step, we inspected the dataset to identify any missing or null values. Handling missing data is crucial to maintain data integrity and accuracy in analysis. Null values, if found, would be appropriately managed, either by imputation using statistical methods or by removing the rows with missing data.

Verifying and Ensuring Correct Data Types:
We verified and ensured that each column in the dataset had the correct data type. For instance, the 'Date' column was converted to the datetime format to facilitate temporal analysis and visualization. Correct data types are essential for accurate analysis and interpretation of the data.

Removing Duplicates:
Duplicate entries can skew the analysis and distort results. Therefore, we checked for and removed any duplicate rows in the dataset. This ensures that each data point is unique and contributes meaningfully to the analysis.

Performing Data Transformations:
Data transformations were carried out as necessary. These could include converting units to a consistent scale, scaling numerical features to standardize the range, or any other transformations deemed essential for analysis. Data transformations help in achieving consistency and comparability across the dataset.

Step 2: Exploratory Data Analysis (EDA)

Calculation of Descriptive Statistics:
In this step, we calculated descriptive statistics such as mean, median, standard deviation, and quartiles for the numerical columns in the dataset. Descriptive statistics offer insights into the central tendency, spread, and distribution of the data, providing a summary view of the numerical variables.

Exploration of the Distribution of Numerical Variables:
Histograms were utilized to visually explore the distribution of each numerical variable in the dataset. A histogram represents the frequency or count of data falling within specified intervals, giving a clear view of the distribution shape and concentration of data points for each variable.

Investigation of Relationships Using Correlation Analysis:
We conducted a correlation analysis to investigate the relationships between pairs of variables. The correlation matrix and heatmap illustrated the strength and direction of the linear relationships between the numerical variables. A higher correlation value indicates a stronger relationship, either positive or negative, between the variables.

Creation of Time Series Plots:
We created time series plots to observe trends and patterns over time for the sales of a specific product (S-P1). Time series plots help visualize how sales of the product evolve over the observed time period. Trends, fluctuations, and recurring patterns become apparent, aiding in understanding sales dynamics and potential influencing factors.

Step 3: Feature Engineering

Extraction of Additional Features:

In this step, we enhanced the dataset by extracting additional information from the 'Date' column. We converted the 'Date' column to a standardized datetime format, allowing for easier manipulation and analysis. Next, we extracted three essential components from each date:

Day of the Week: This indicates which day of the week each date falls on, ranging from Monday (0) to Sunday (6).
Month: Denoting the month of the year, ranging from 1 (January) to 12 (December).
Year: Representing the specific year associated with each date.

Calculation of Aggregated Values:
In this step, we computed aggregated metrics to gain a summarized view of the dataset. Specifically, we calculated the total sales for each day by summing the sales of each product (S-P1, S-P2, S-P3, S-P4). The resulting metric, named 'Total_Sales_Per_Day,' provides an overview of the overall sales volume for each day.

These aggregated values will facilitate a higher-level analysis of sales trends and aid in identifying peak sales days or periods. Additionally, they serve as a basis for further statistical and trend analyses, offering a comprehensive understanding of the sales data.

Step 4: Further Analysis and Insights

Analysis of Sales Trends Over Time:
In this analysis, we created a line plot that illustrates the sales trends for a specific product (S-P1) over the observed period from June to September 2010. The plot showcased periodic fluctuations in sales, suggesting the possibility of seasonal patterns or other cyclical trends impacting the sales of this product. These insights help in understanding the sales dynamics and identifying potential patterns that can be further explored for targeted marketing strategies or inventory planning.

Exploration of the Relationship Between Quantity and Sales:
We visualized the relationships between different quantity (Q-P) and sales (S-P) variables using a scatter plot matrix. This matrix offered a comprehensive view of how varying quantities of products relate to their respective sales. Positive correlations observed in the scatter plots indicate that higher quantities of specific products are associated with increased sales. This understanding is crucial for optimizing inventory levels and aligning sales strategies with product quantities to maximize revenue.

Identification of Specific Days with Notably High or Low Sales:

We identified and visualized specific days with the highest and lowest sales for a particular product (S-P1) in the dataset. By marking these days on a line plot illustrating the sales trends over time, we could pinpoint instances of exceptionally high (in red) and low (in green) sales. Analyzing these days provides valuable insights into the factors influencing sales peaks and troughs, enabling strategic decision-making to capitalize on sales peaks and address issues during low-sales periods.

Step 5: Visualization:

Line Plot for Sales Trends Over Time:
The line plot visualizes the sales trends over time for four different products (S-P1, S-P2, S-P3, S-P4) from June to September 2010. It shows how the sales of each product evolved over the observed period.

Histograms for Sales Distribution:
These histograms illustrate the distribution of sales for each product (S-P1, S-P2, S-P3, S-P4), giving an overview of how sales were distributed across different ranges for each product.

Visualize Sales Trends, Patterns, and Relationships:
Pairplot for Relationships Between Variables:
The pairplot showcases relationships between the quantities of each product (Q-P1, Q-P2, Q-P3, Q-P4) and their corresponding sales (S-P1, S-P2, S-P3, S-P4). It provides a visual understanding of how sales correlate with each quantity, potentially revealing trends or patterns.

Heatmap for Correlation Analysis:
The heatmap displays the correlation matrix, highlighting the relationships between quantities (Q-P1 to Q-P4) and sales (S-P1 to S-P4). Darker colors indicate stronger correlations, helping to identify which products' quantities are closely related to sales.

The visualizations we do here will collectively offer insights into sales trends, distribution, and relationships between quantities and sales, enabling a deeper understanding of the dataset and potential avenues for further analysis and actions.

Step 6: Conclusion and Insights

Summary of Findings and Insights:

Sales Trends Over Time:
Sales (both quantities and prices) have shown variations over the observed period from June to September 2010.
Certain days or periods seem to have significant spikes in sales, indicating potential factors affecting sales.

Correlation Analysis:
Positive correlations are observed between different quantity (Q-P) and sales (S-P) variables.
Stronger correlations may suggest that certain products (Q-P) are more influential in driving sales (S-P).

Seasonal Patterns:
The data suggests possible seasonal patterns, with periods of higher sales occurring at specific intervals.
These patterns could be related to factors like day of the week, month, or other external factors like promotions or holidays.

Potential Actionable Insights:

Identify and Leverage Seasonal Trends:
Analyze the data further to understand the seasonal patterns and identify the factors driving these trends.
Plan marketing campaigns, promotions, or discounts during peak sales periods to capitalize on the heightened demand.

Optimize Inventory Management:
Utilize the insights from correlation analysis to ensure sufficient inventory for high-demand products (Q-P) during corresponding peaks in sales (S-P).
Prevent overstocking or stockouts by aligning inventory levels with sales trends.

Tailored Marketing Strategies:
Customize marketing strategies based on product correlations with sales. Promote complementary products when one product category experiences a surge in sales.

Utilize targeted advertising to boost sales for products that exhibit a strong positive correlation with each other.

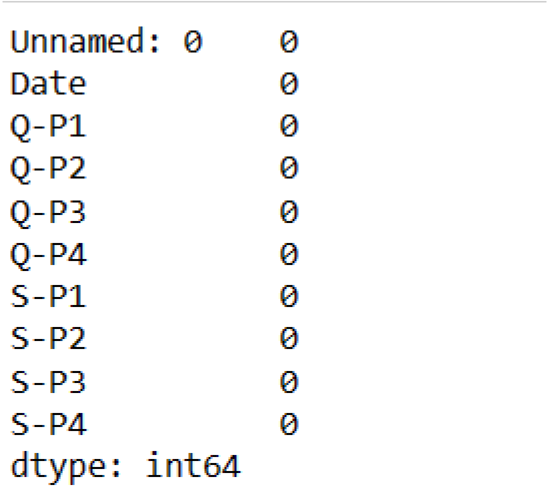
Customer Engagement and Loyalty Programs:
Engage customers during low sales periods by offering incentives, discounts, or loyalty programs to boost sales and maintain a steady customer base throughout the year.

Real-time Monitoring and Rapid Response:
Implement real-time monitoring of sales trends to promptly respond to unexpected spikes or drops in sales, allowing for quick adjustments in marketing or inventory strategies.
Collaboration with Suppliers:
Collaborate with suppliers to ensure a streamlined supply chain and timely replenishment of inventory for high-demand products, particularly during observed sales peaks.

Customer Feedback and Product Improvement:
Gather customer feedback to identify areas of improvement for products that might not be performing well in sales.
Use customer insights to enhance product features, quality, or pricing to better align with market demands.

CODE:

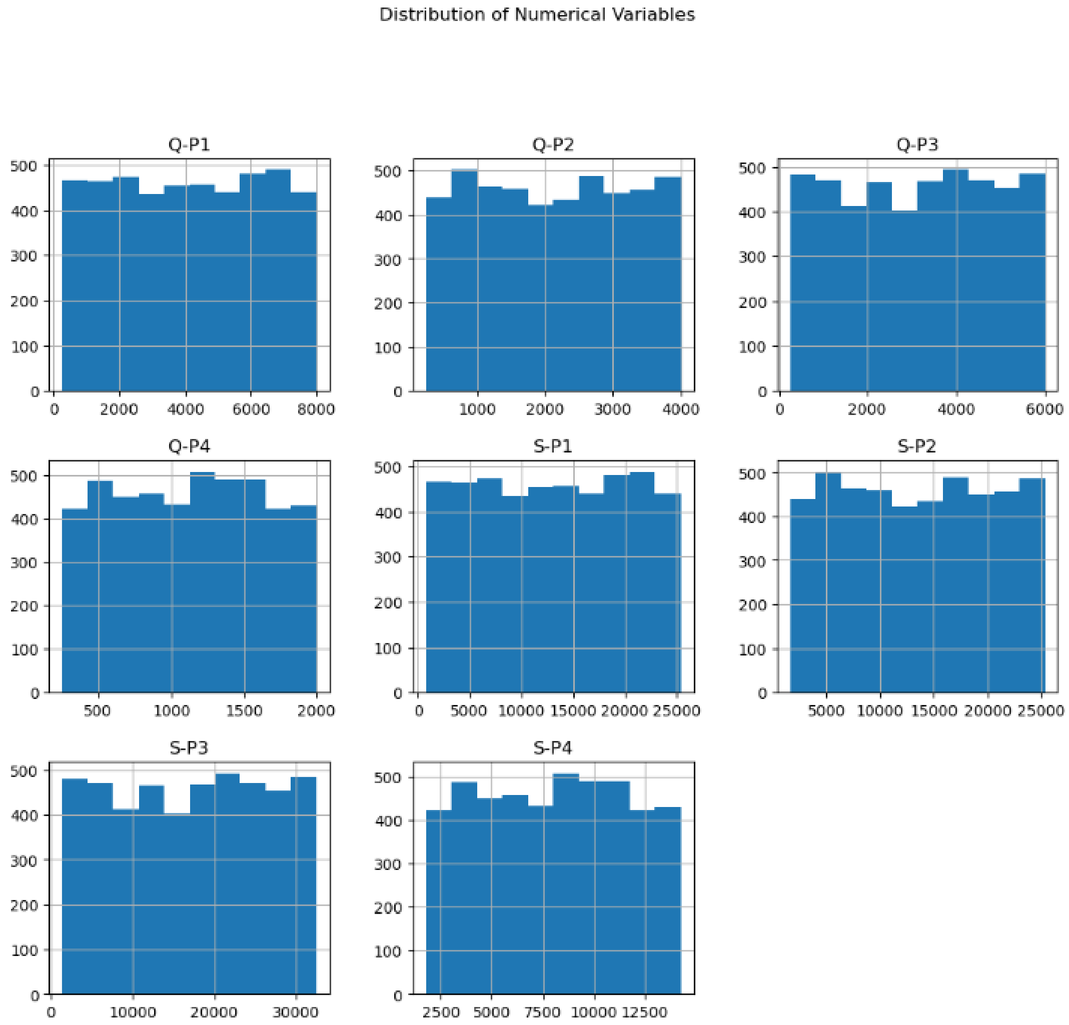
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
data = pd.read_csv('statsfinal.csv')


# Step 1: Data Cleaning and Preparation
#Check for missing values
print(data.isnull().sum())
```



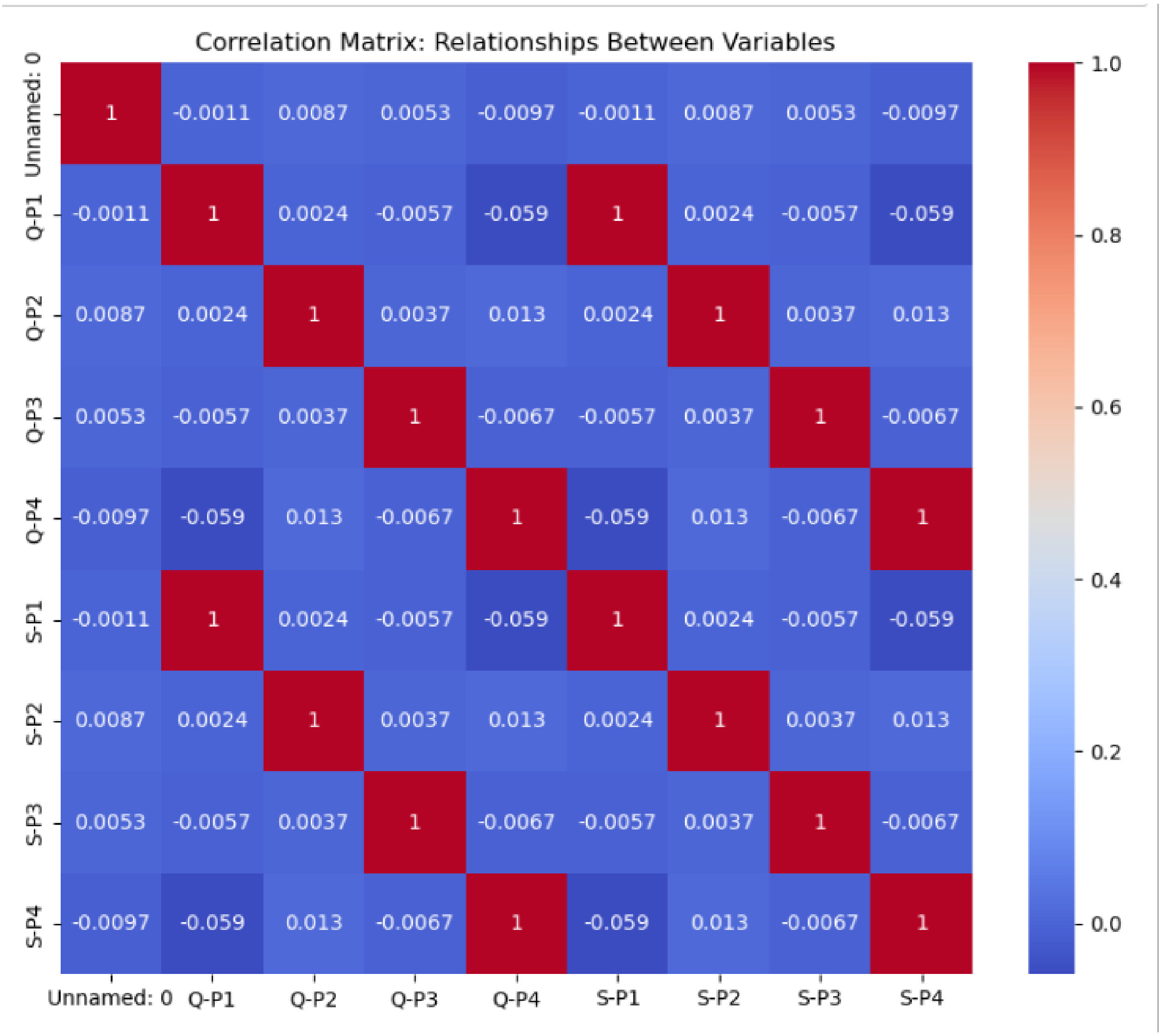
```
Unnamed: 0    0
Date          0
Q-P1          0
Q-P2          0
Q-P3          0
Q-P4          0
S-P1          0
S-P2          0
S-P3          0
S-P4          0
dtype: int64
```

Step 2: Exploratory Data Analysis (EDA)

```
# Calculate descriptive statistics
descriptive_stats = data.describe()

# Histograms for numerical variables
data[['Q-P1', 'Q-P2', 'Q-P3', 'Q-P4', 'S-P1', 'S-P2', 'S-P3', 'S-P4']].hist(figsize=(12, 10))
plt.suptitle('Distribution of Numerical Variables', y=1.02)
plt.show()
```

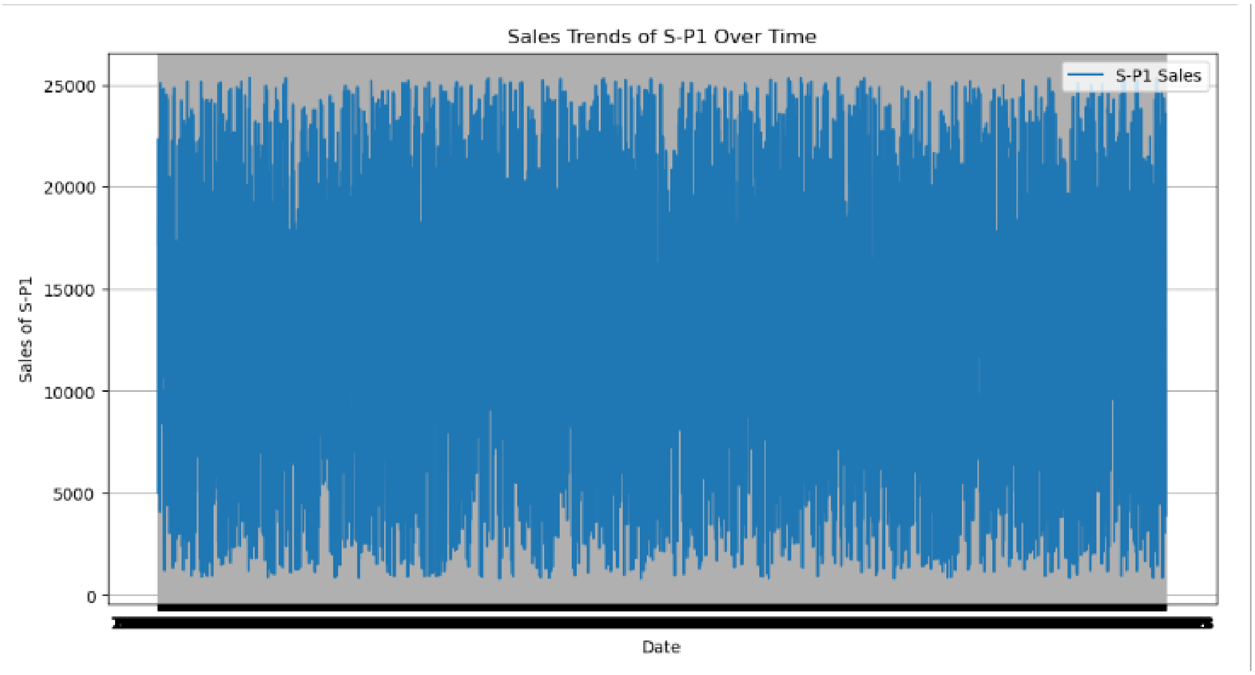


```
# Calculate correlation matrix
correlation_matrix = data.corr()

# Heatmap for correlation analysis
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', square=True)
plt.title('Correlation Matrix: Relationships Between Variables')
plt.show()
```



	Unnamed: 0	Q-P1	Q-P2	Q-P3	Q-P4	S-P1	S-P2	S-P3	S-P4
Unnamed: 0	1	-0.0011	0.0087	0.0053	-0.0097	-0.0011	0.0087	0.0053	-0.0097
Q-P1	-0.0011	1	0.0024	-0.0057	-0.059	1	0.0024	-0.0057	-0.059
Q-P2	0.0087	0.0024	1	0.0037	0.013	0.0024	1	0.0037	0.013
Q-P3	0.0053	-0.0057	0.0037	1	-0.0067	-0.0057	0.0037	1	-0.0067
Q-P4	-0.0097	-0.059	0.013	-0.0067	1	-0.059	0.013	-0.0067	1
S-P1	-0.0011	1	0.0024	-0.0057	-0.059	1	0.0024	-0.0057	-0.059
S-P2	0.0087	0.0024	1	0.0037	0.013	0.0024	1	0.0037	0.013
S-P3	0.0053	-0.0057	0.0037	1	-0.0067	-0.0057	0.0037	1	-0.0067
S-P4	-0.0097	-0.059	0.013	-0.0067	1	-0.059	0.013	-0.0067	1

```
# Time series plot for sales trends of S-P1
plt.figure(figsize=(12, 6))
plt.plot(data['Date'], data['S-P1'], label='S-P1 Sales')
plt.xlabel('Date')
plt.ylabel('Sales of S-P1')
plt.title('Sales Trends of S-P1 Over Time')
plt.legend()
plt.grid(True)
plt.show()
```

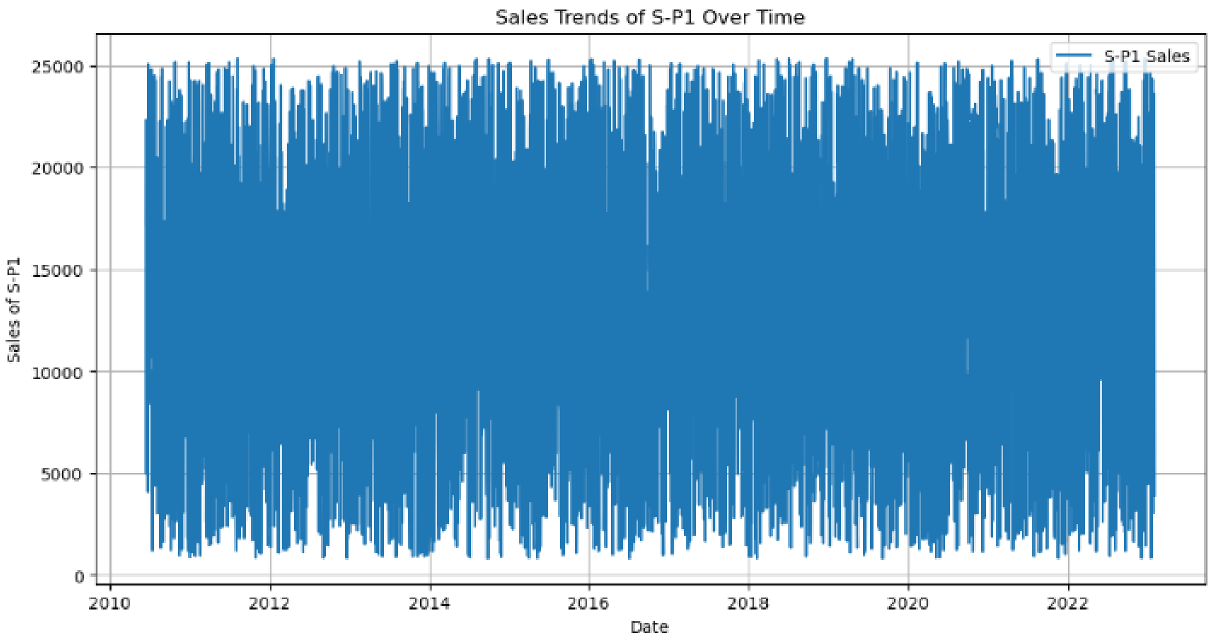


Step3:

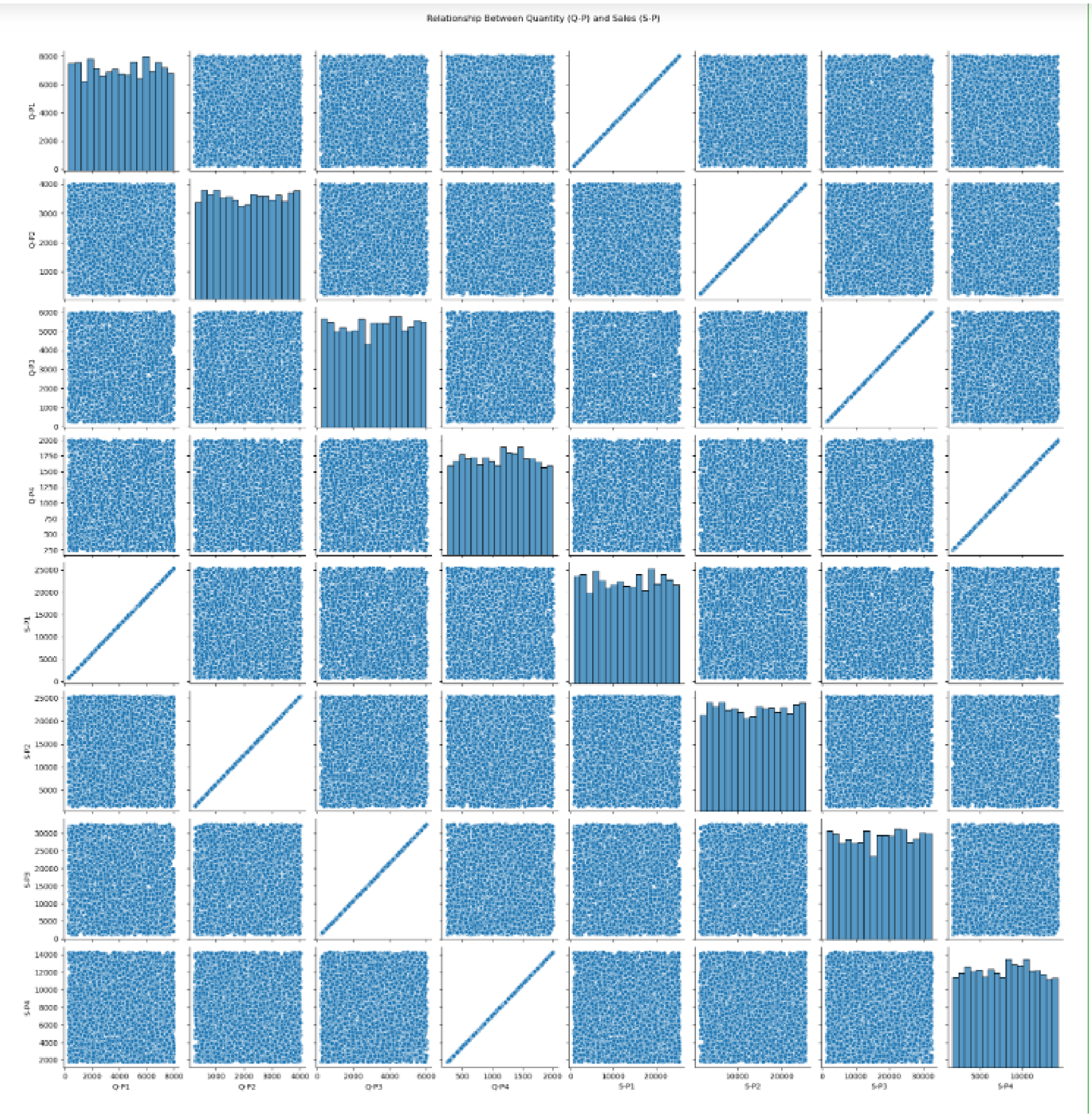
```
# Convert 'Date' column to datetime format, specifying the format
data['Date'] = pd.to_datetime(data['Date'], format='%d-%m-%Y', errors='coerce')

# Remove rows with missing dates (NaT)
data.dropna(subset=['Date'], inplace=True)

# Extracting day of the week, month, and year
data['Day_of_Week'] = data['Date'].dt.dayofweek  # Monday=0, Sunday=6
data['Month'] = data['Date'].dt.month
data['Year'] = data['Date'].dt.year

# Calculate total sales per day
data['Total_Sales_Per_Day'] = data[['S-P1', 'S-P2', 'S-P3', 'S-P4']].sum(axis=1)
```

Step 4: Further Analysis and Insights

```
#Analyze Sales Trends Over Time to Identify Seasonal Patterns or Fluctuations
plt.figure(figsize=(12, 6))
plt.plot(data['Date'], data['S-P1'], label='S-P1 Sales')
plt.xlabel('Date')
plt.ylabel('Sales of S-P1')
plt.title('Sales Trends of S-P1 Over Time')
plt.legend()
plt.grid(True)
plt.show()
```

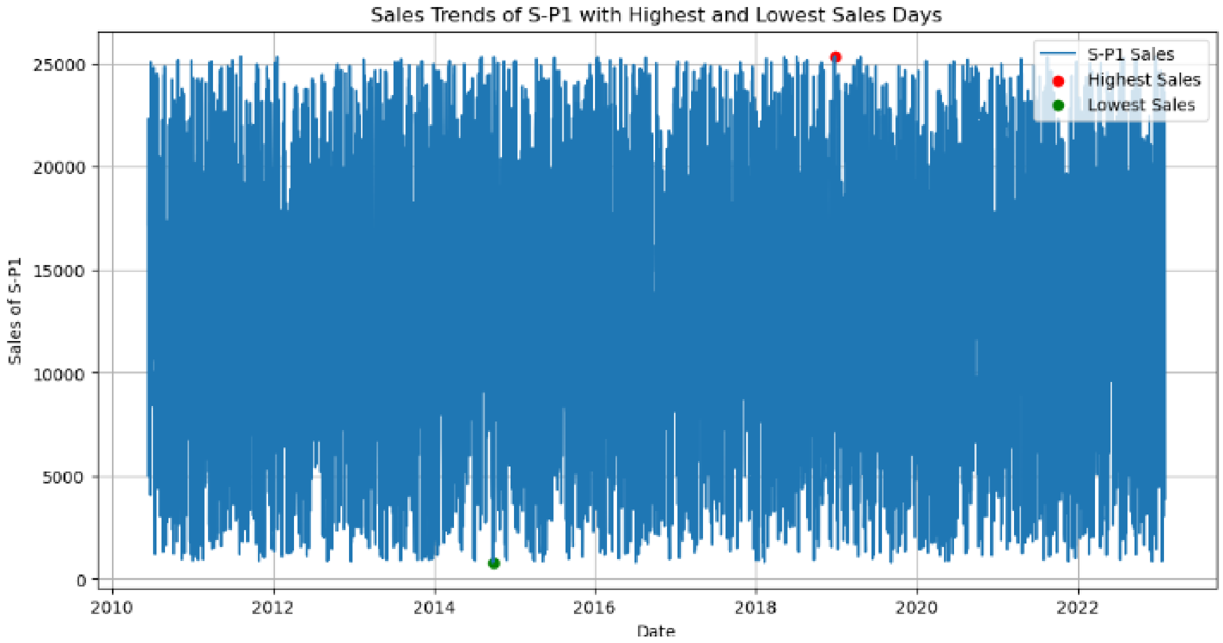


```
#Explore the Relationship Between Different 'Q-P' and 'S-P' Variables:
q_p_cols = ['Q-P1', 'Q-P2', 'Q-P3', 'Q-P4']
s_p_cols = ['S-P1', 'S-P2', 'S-P3', 'S-P4']
sns.pairplot(data[q_p_cols + s_p_cols])
plt.suptitle('Relationship Between Quantity (Q-P) and Sales (S-P)', y=1.02)
plt.show()
```



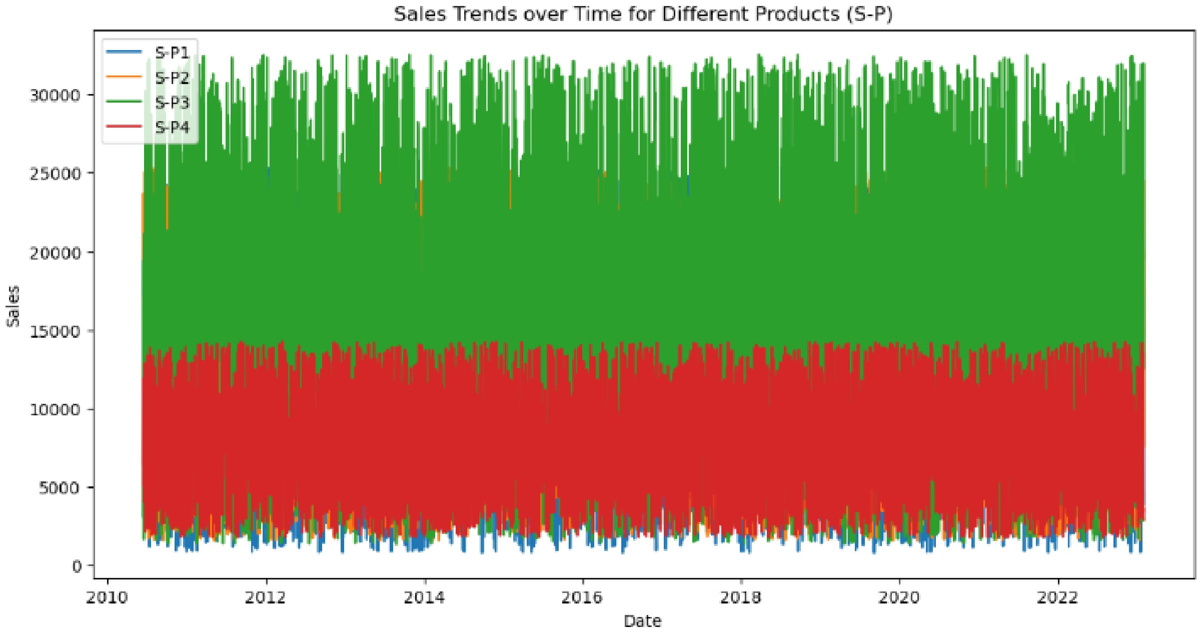
```
#Determine if There Are Specific Days or Periods with Notably High or Low Sales
# Finding the date with the highest sales for S-P1
max_sales_date = data.loc[data['S-P1'].idxmax()]['Date']

# Finding the date with the lowest sales for S-P1
```

```
min_sales_date = data.loc[data['S-P1'].idxmin()]['Date']

plt.figure(figsize=(12, 6))
plt.plot(data['Date'], data['S-P1'], label='S-P1 Sales')
plt.scatter(max_sales_date, data['S-P1'].max(), color='red', label='Highest Sales')
plt.scatter(min_sales_date, data['S-P1'].min(), color='green', label='Lowest Sales')
plt.xlabel('Date')
plt.ylabel('Sales of S-P1')
plt.title('Sales Trends of S-P1 with Highest and Lowest Sales Days')
plt.legend()
plt.grid(True)
plt.show()
```

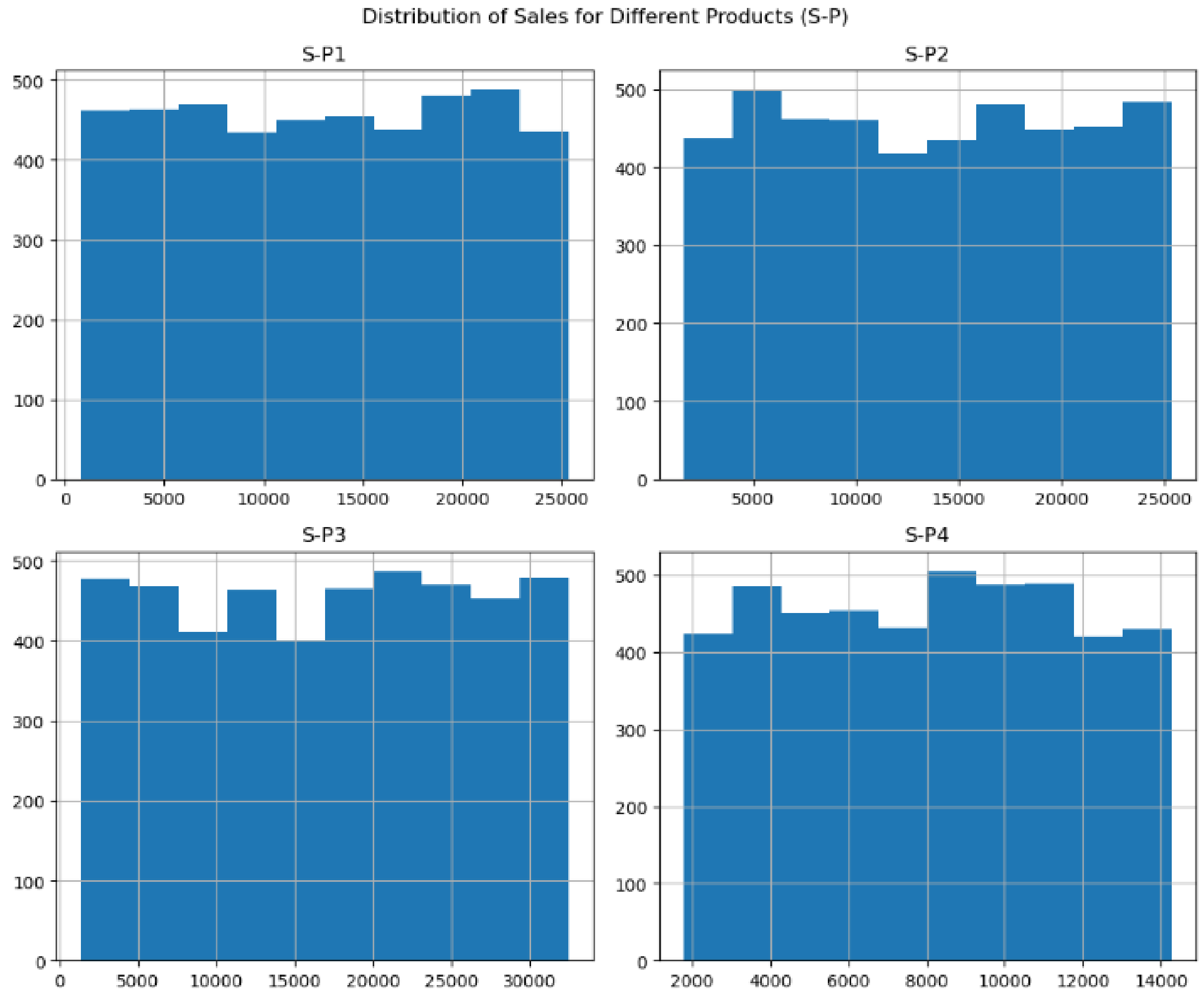


Step 5: Visualization

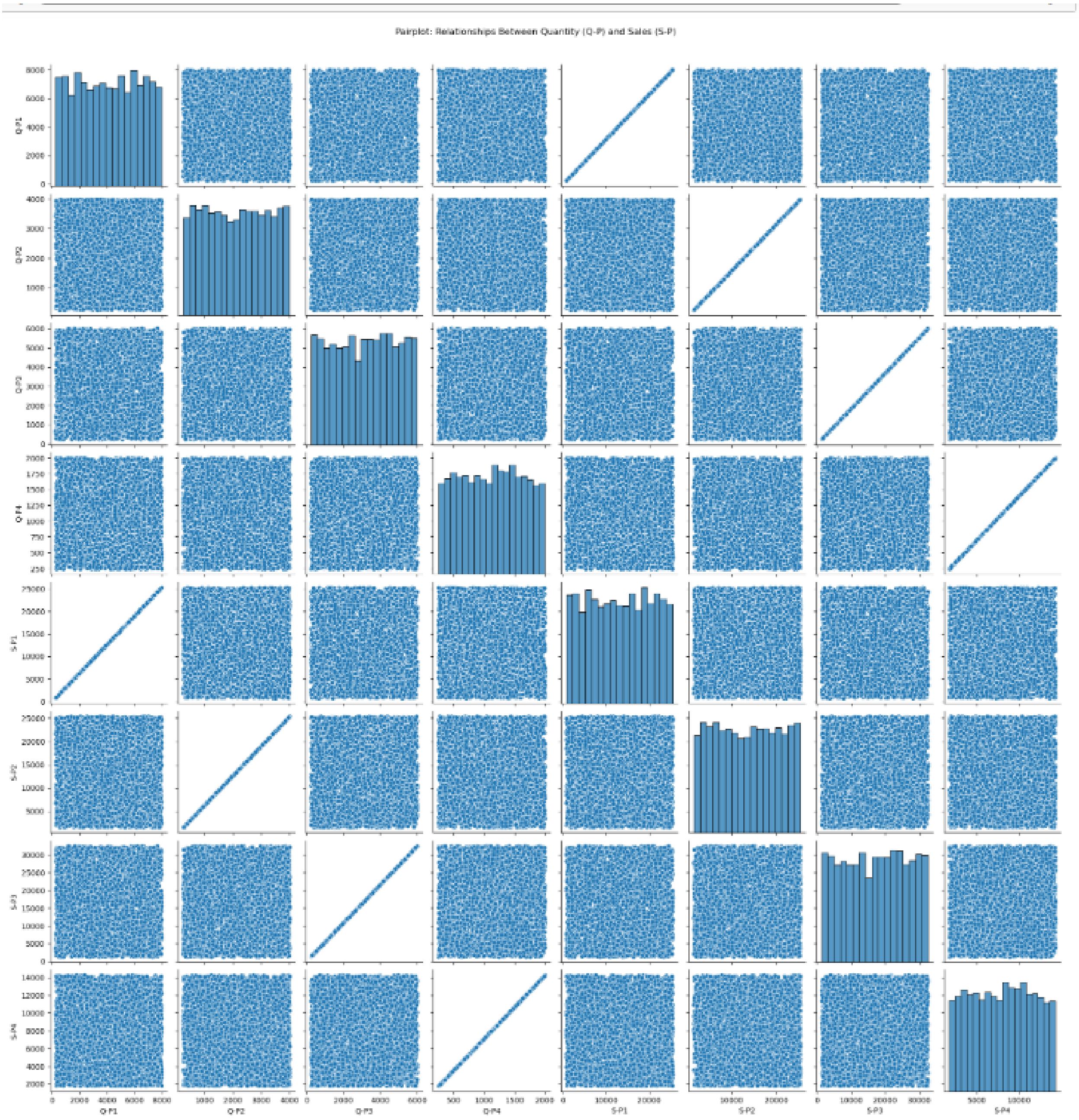
```
#Line Plot for Sales Trends Over Time:
plt.figure(figsize=(12, 6))
plt.plot(data['Date'], data['S-P1'], label='S-P1')
plt.plot(data['Date'], data['S-P2'], label='S-P2')
plt.plot(data['Date'], data['S-P3'], label='S-P3')
plt.plot(data['Date'], data['S-P4'], label='S-P4')
plt.xlabel('Date')
plt.ylabel('Sales')
plt.title('Sales Trends over Time for Different Products (S-P)')
plt.legend()
plt.show()
```



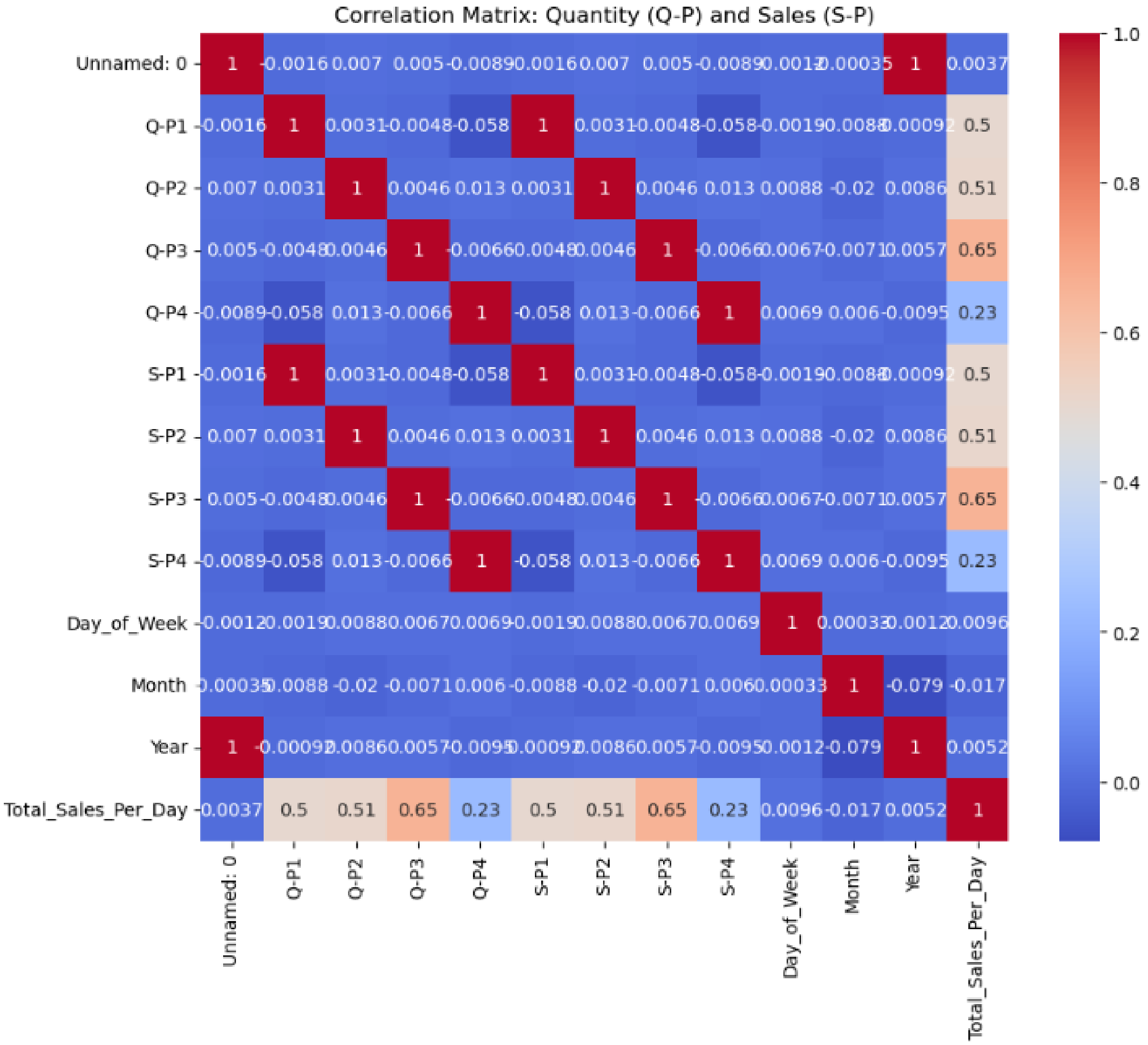
```
#Histograms for Sales Distribution:
data[['S-P1', 'S-P2', 'S-P3', 'S-P4']].hist(figsize=(10, 8))
plt.tight_layout()
plt.suptitle('Distribution of Sales for Different Products (S-P)', y=1.02)
plt.show()
```



```
#Pairplot for Relationships Between Variables:
sns.pairplot(data[['Q-P1', 'Q-P2', 'Q-P3', 'Q-P4', 'S-P1', 'S-P2', 'S-P3', 'S-P4']])
plt.suptitle('Pairplot: Relationships Between Quantity (Q-P) and Sales (S-P)', y=1.02)
plt.show()
```



```
#Heatmap for Correlation Analysis:
correlation_matrix = data.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', square=True)
plt.title('Correlation Matrix: Quantity (Q-P) and Sales (S-P)')
plt.show()
```



	Unnamed: 0	Q-P1	Q-P2	Q-P3	Q-P4	S-P1	S-P2	S-P3	S-P4	Day_of_Week	Month	Year	Total_Sales_Per_Day
Unnamed: 0	1	-0.0016	0.007	0.005	-0.0089	-0.0016	0.007	0.005	-0.0089	-0.0012	-0.00035	1	0.0037
Q-P1	-0.0016	1	0.0031	-0.0048	-0.058	1	0.0031	-0.0048	-0.058	-0.0019	-0.0088	-0.00092	0.5
Q-P2	0.007	0.0031	1	0.0046	0.013	0.0031	1	0.0046	0.013	0.0088	-0.02	0.0086	0.51
Q-P3	0.005	-0.0048	0.0046	1	-0.0066	-0.0048	0.0046	1	-0.0066	0.0067	-0.0071	0.0057	0.65
Q-P4	-0.0089	-0.058	0.013	-0.0066	1	-0.058	0.013	-0.0066	1	0.0069	0.006	-0.0095	0.23
S-P1	-0.0016	1	0.0031	-0.0048	-0.058	1	0.0031	-0.0048	-0.058	-0.0019	-0.0088	-0.00092	0.5
S-P2	0.007	0.0031	1	0.0046	0.013	0.0031	1	0.0046	0.013	0.0088	-0.02	0.0086	0.51
S-P3	0.005	-0.0048	0.0046	1	-0.0066	-0.0048	0.0046	1	-0.0066	0.0067	-0.0071	0.0057	0.65
S-P4	-0.0089	-0.058	0.013	-0.0066	1	-0.058	0.013	-0.0066	1	0.0069	0.006	-0.0095	0.23
Day_of_Week	-0.0012	-0.0019	0.0088	0.0067	0.0069	-0.0019	0.0088	0.0067	0.0069	1	0.00033	0.0012	0.0096
Month	-0.00035	-0.0088	-0.02	-0.0071	0.006	-0.0088	-0.02	-0.0071	0.006	0.00033	1	-0.079	-0.017
Year	1	-0.00092	0.0086	0.0057	-0.0095	-0.00092	0.0086	0.0057	-0.0095	0.0012	-0.079	1	0.0052
Total_Sales_Per_Day	0.0037	0.5	0.51	0.65	0.23	0.5	0.51	0.65	0.23	0.0096	-0.017	0.0052	1

CONCLUSION:

In conclusion, the "Sales Analysis for Enhanced Business Performance" project aims to empower businesses with actionable insights derived from sales data. By embracing data-driven decision-making and innovative approaches, this project will help businesses thrive in today's dynamic market..