PRODUCT SALES ANALYSIS USING MACHINE LEARNING

Phase 4 Submission Document

Project Name: Product Sales Analysis

Phase 4: Development Part 2

In this part we will continue building your project.

- Continue building the analysis by creating visualizations using IBM Cognos and generating actionable insights.
- Use IBM Cognos to design interactive dashboards and reports that display insights such as top-selling products, sales trends, and customer preferences.
- Derive insights from the visualizations, such as identifying products with the highest sales, peak sales periods, and customer preferences for specific products.

Step 1:

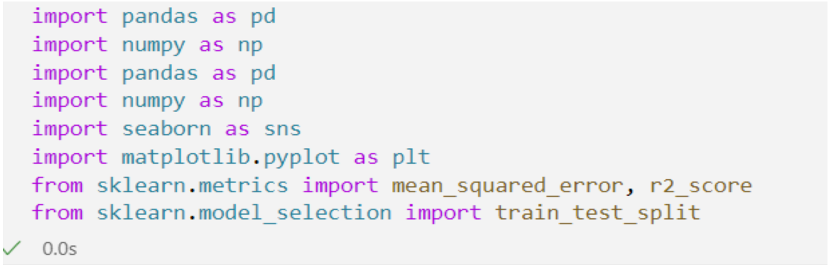
Dataset Loading and inspect

1. Load the Provided Dataset:

Loading the dataset involves reading the data from a file, typically a CSV (Comma-Separated Values) file, into your data analysis environment, which in this case, could be Python.

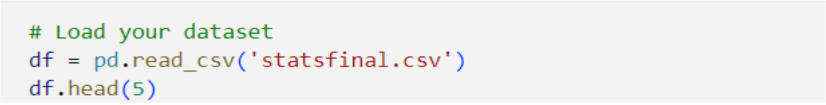
You can use libraries like Pandas to accomplish this. The Pandas library provides powerful data structures and functions for working with structured data.

Example Code to Load the Dataset:

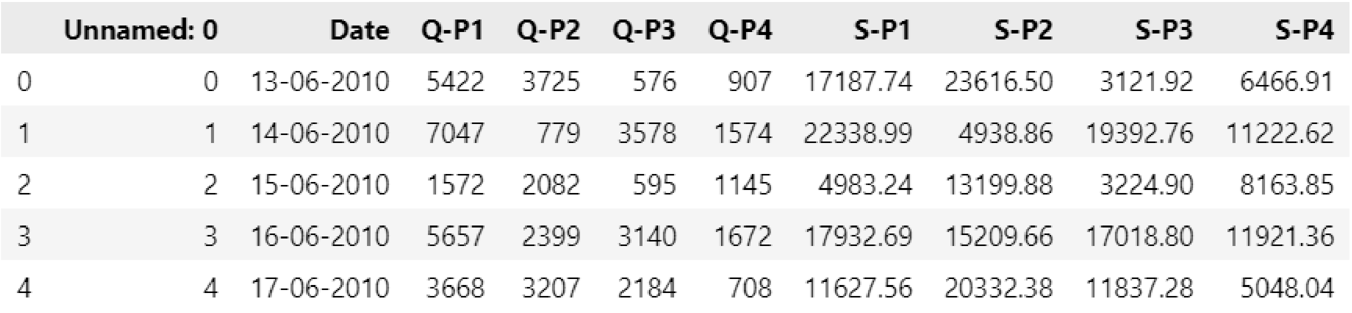


```
import pandas as pd
import numpy as np
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
```

✓ 0.0s



```
# Load your dataset
df = pd.read_csv('statsfinal.csv')
df.head(5)
```

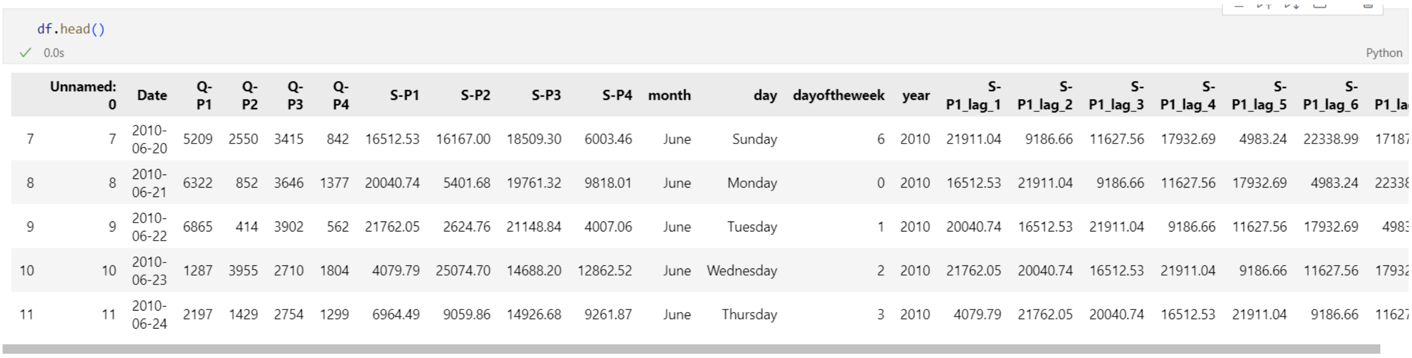


	Unnamed: 0	Date	Q-P1	Q-P2	Q-P3	Q-P4	S-P1	S-P2	S-P3	S-P4
0	0	13-06-2010	5422	3725	576	907	17187.74	23616.50	3121.92	6466.91
1	1	14-06-2010	7047	779	3578	1574	22338.99	4938.86	19392.76	11222.62
2	2	15-06-2010	1572	2082	595	1145	4983.24	13199.88	3224.90	8163.85
3	3	16-06-2010	5657	2399	3140	1672	17932.69	15209.66	17018.80	11921.36
4	4	17-06-2010	3668	3207	2184	708	11627.56	20332.38	11837.28	5048.04

This code reads the dataset from the "your_dataset.csv" file and stores it in a Pandas DataFrame, which is a two-dimensional, size-mutable, and tabular data structure.

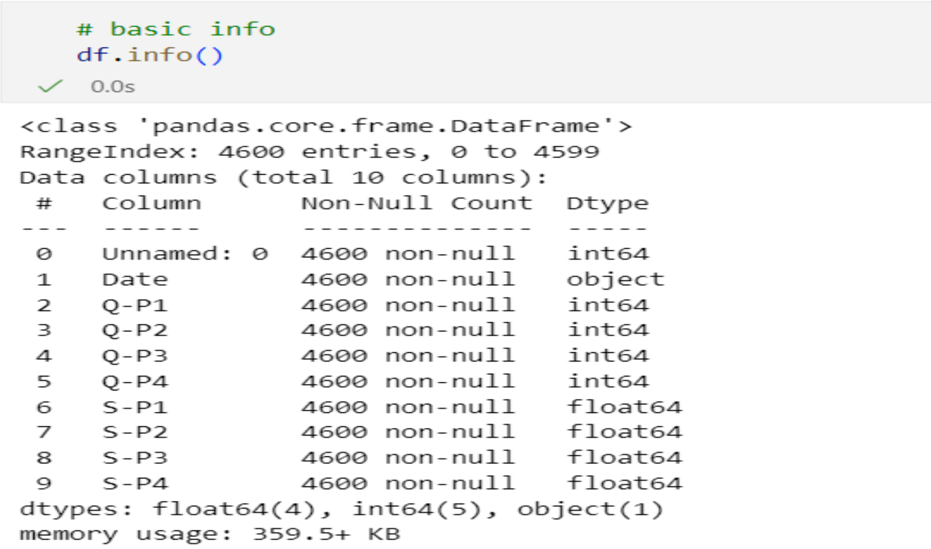
2. Inspect the Dataset:

- After loading the dataset, it's important to inspect it to understand its structure, contents, and any potential issues.
- You can use various Pandas functions to inspect the dataset, such as **head()**, **info()**, and **describe()**, to view the first few rows, get information about data types, and summarize statistical properties of the data.



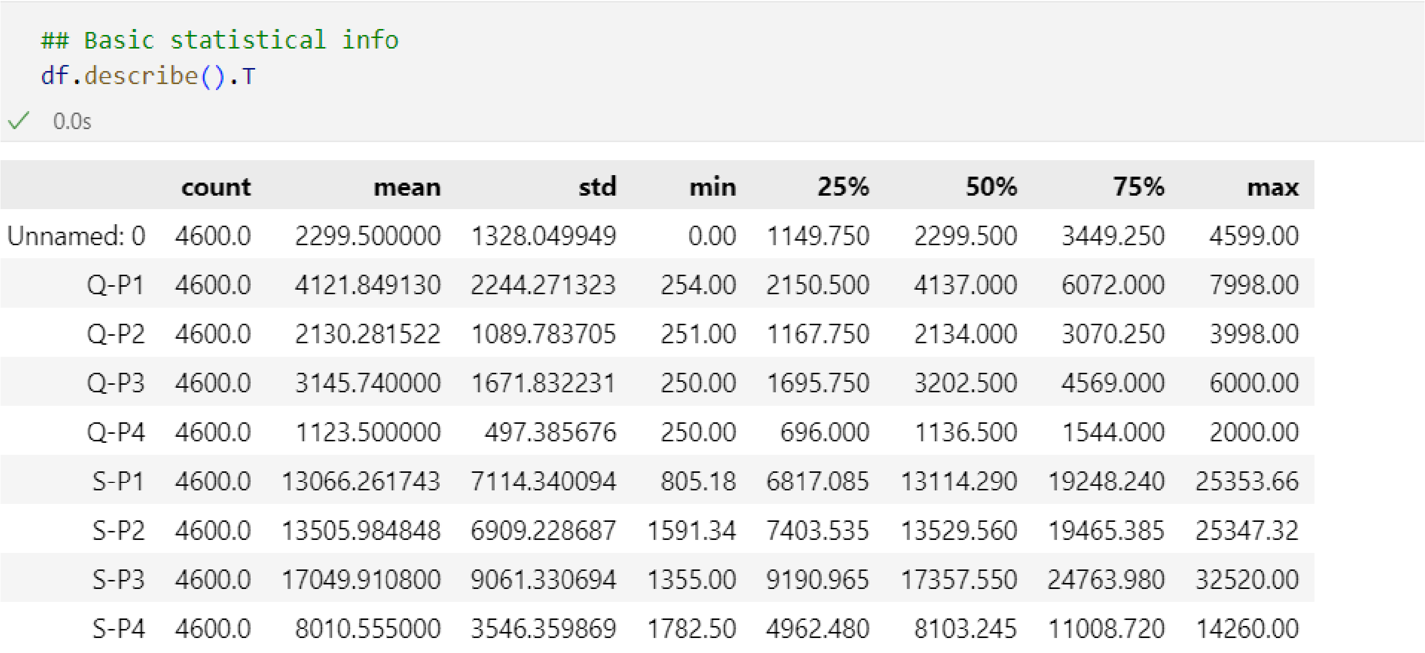
```
df.head()
```

	Unnamed: 0	Date	Q-P1	Q-P2	Q-P3	Q-P4	S-P1	S-P2	S-P3	S-P4	month	day	dayoftheweek	year	S-P1_lag_1	S-P1_lag_2	S-P1_lag_3	S-P1_lag_4	S-P1_lag_5	S-P1_lag_6	P1_la
7	7	2010-06-20	5209	2550	3415	842	16512.53	16167.00	18509.30	6003.46	June	Sunday	6	2010	21911.04	9186.66	11627.56	17932.69	4983.24	22338.99	17187
8	8	2010-06-21	6322	852	3646	1377	20040.74	5401.68	19761.32	9818.01	June	Monday	0	2010	16512.53	21911.04	9186.66	11627.56	17932.69	4983.24	22338
9	9	2010-06-22	6865	414	3902	562	21762.05	2624.76	21148.84	4007.06	June	Tuesday	1	2010	20040.74	16512.53	21911.04	9186.66	11627.56	17932.69	4983
10	10	2010-06-23	1287	3955	2710	1804	4079.79	25074.70	14688.20	12862.52	June	Wednesday	2	2010	21762.05	20040.74	16512.53	21911.04	9186.66	11627.56	17932
11	11	2010-06-24	2197	1429	2754	1299	6964.49	9059.86	14926.68	9261.87	June	Thursday	3	2010	4079.79	21762.05	20040.74	16512.53	21911.04	9186.66	11627



```
# basic info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Unnamed: 0   4600 non-null   int64  
 1   Date         4600 non-null   object 
 2   Q-P1         4600 non-null   int64  
 3   Q-P2         4600 non-null   int64  
 4   Q-P3         4600 non-null   int64  
 5   Q-P4         4600 non-null   int64  
 6   S-P1         4600 non-null   float64
 7   S-P2         4600 non-null   float64
 8   S-P3         4600 non-null   float64
 9   S-P4         4600 non-null   float64
dtypes: float64(4), int64(5), object(1)
memory usage: 359.5+ KB
```

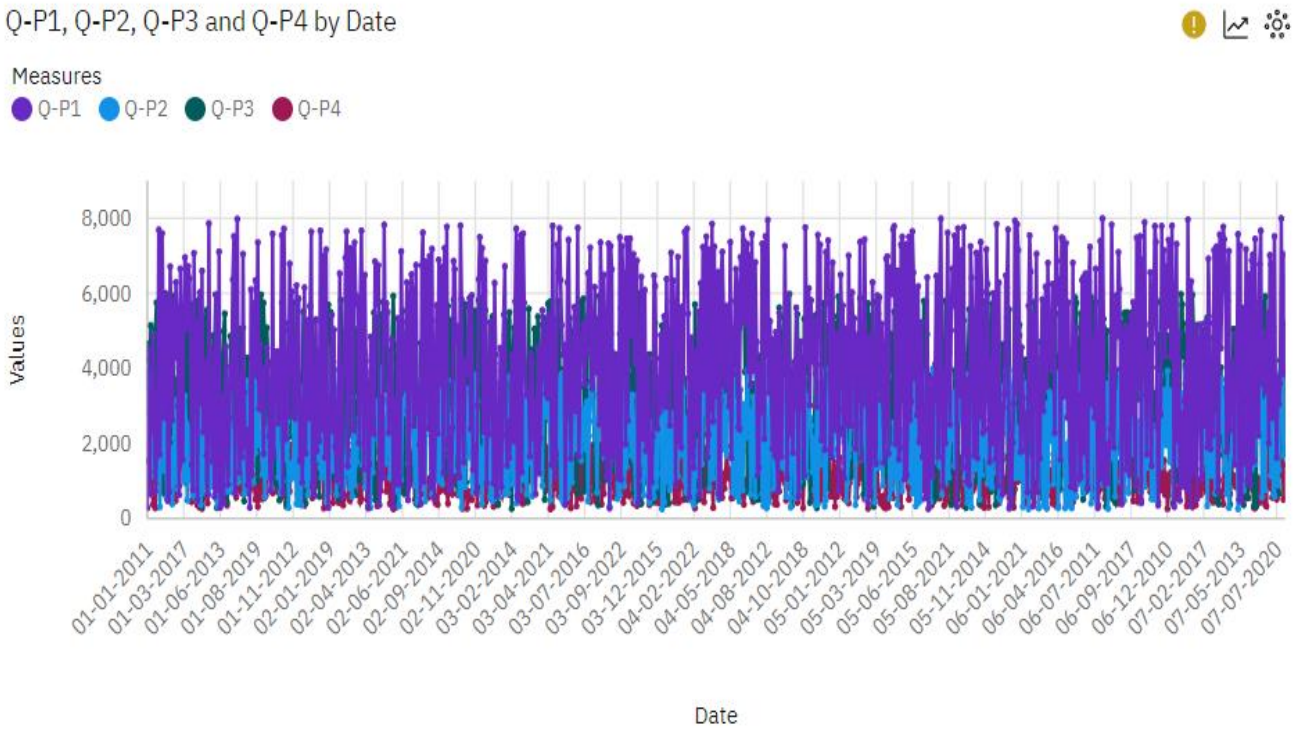


```
## Basic statistical info
df.describe().T
```

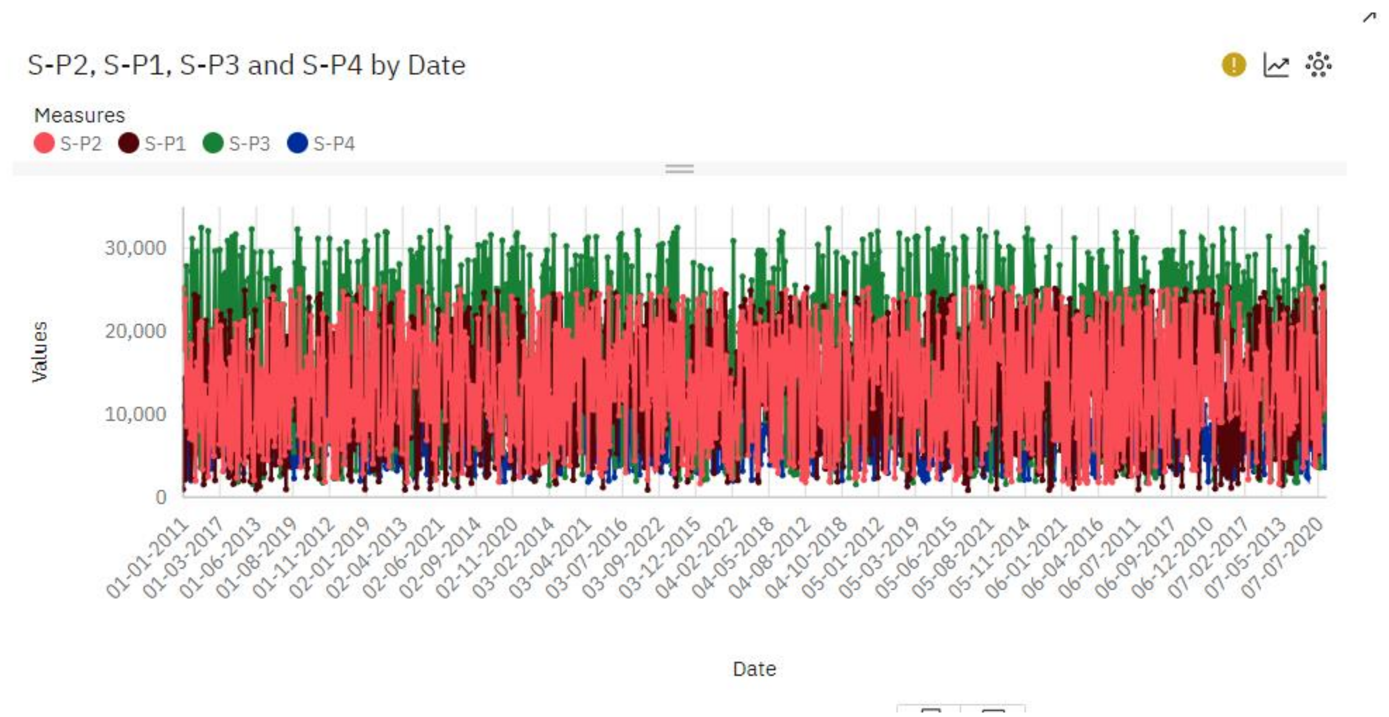
✓ 0.0s

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	4600.0	2299.500000	1328.049949	0.00	1149.750	2299.500	3449.250	4599.00
Q-P1	4600.0	4121.849130	2244.271323	254.00	2150.500	4137.000	6072.000	7998.00
Q-P2	4600.0	2130.281522	1089.783705	251.00	1167.750	2134.000	3070.250	3998.00
Q-P3	4600.0	3145.740000	1671.832231	250.00	1695.750	3202.500	4569.000	6000.00
Q-P4	4600.0	1123.500000	497.385676	250.00	696.000	1136.500	1544.000	2000.00
S-P1	4600.0	13066.261743	7114.340094	805.18	6817.085	13114.290	19248.240	25353.66
S-P2	4600.0	13505.984848	6909.228687	1591.34	7403.535	13529.560	19465.385	25347.32
S-P3	4600.0	17049.910800	9061.330694	1355.00	9190.965	17357.550	24763.980	32520.00
S-P4	4600.0	8010.555000	3546.359869	1782.50	4962.480	8103.245	11008.720	14260.00

Step 2:
Building the analysis by creating visualizations using IBM Cognos and generating actionable insights.

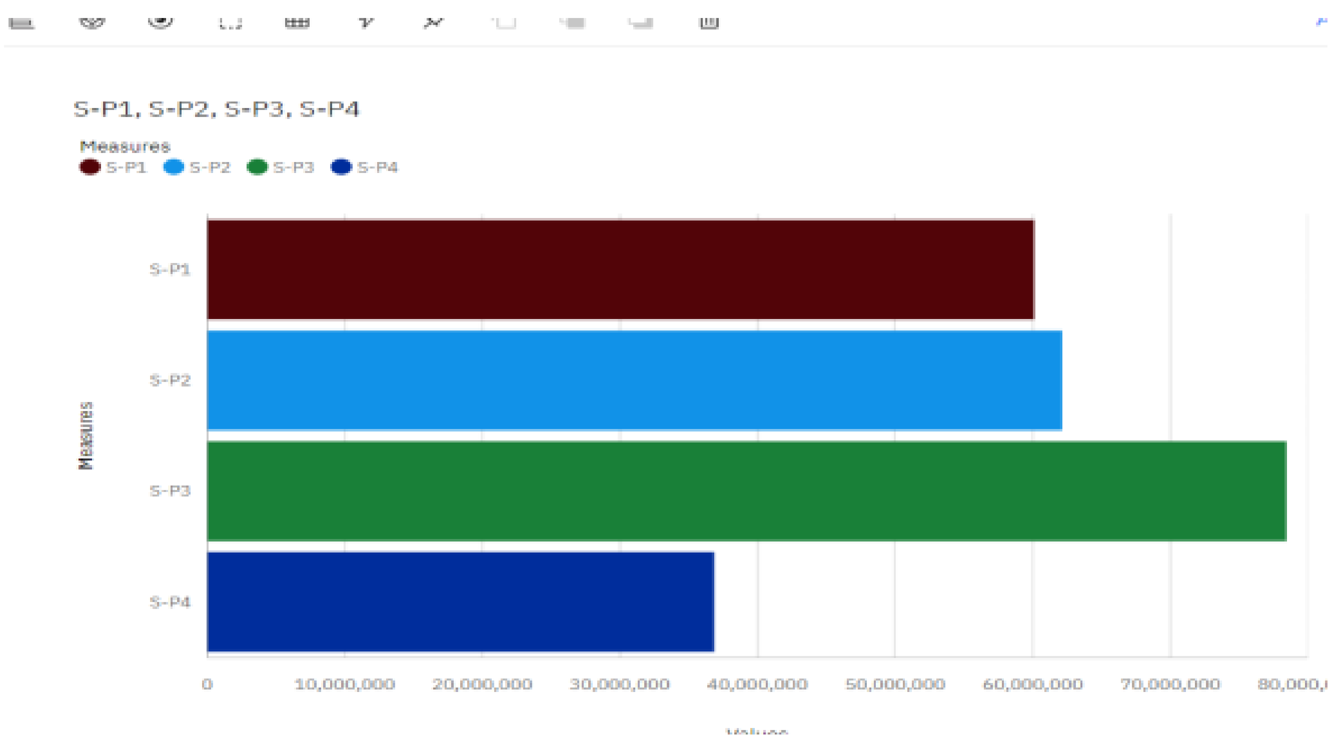


Visualizing the unit sales using line chart by plotting date in the x-axis and plotting unit sales in y-axis.Clearly we can observe the unit sales of product which is high among all the products.



Visualizing the revenue of all the products using line chart by plotting date in the x-axis and plotting revenue in y-axis.Clearly we can observe the revenue of producr which is high among all the product.

#Plotting bar chart to analyse which product has highest revenue



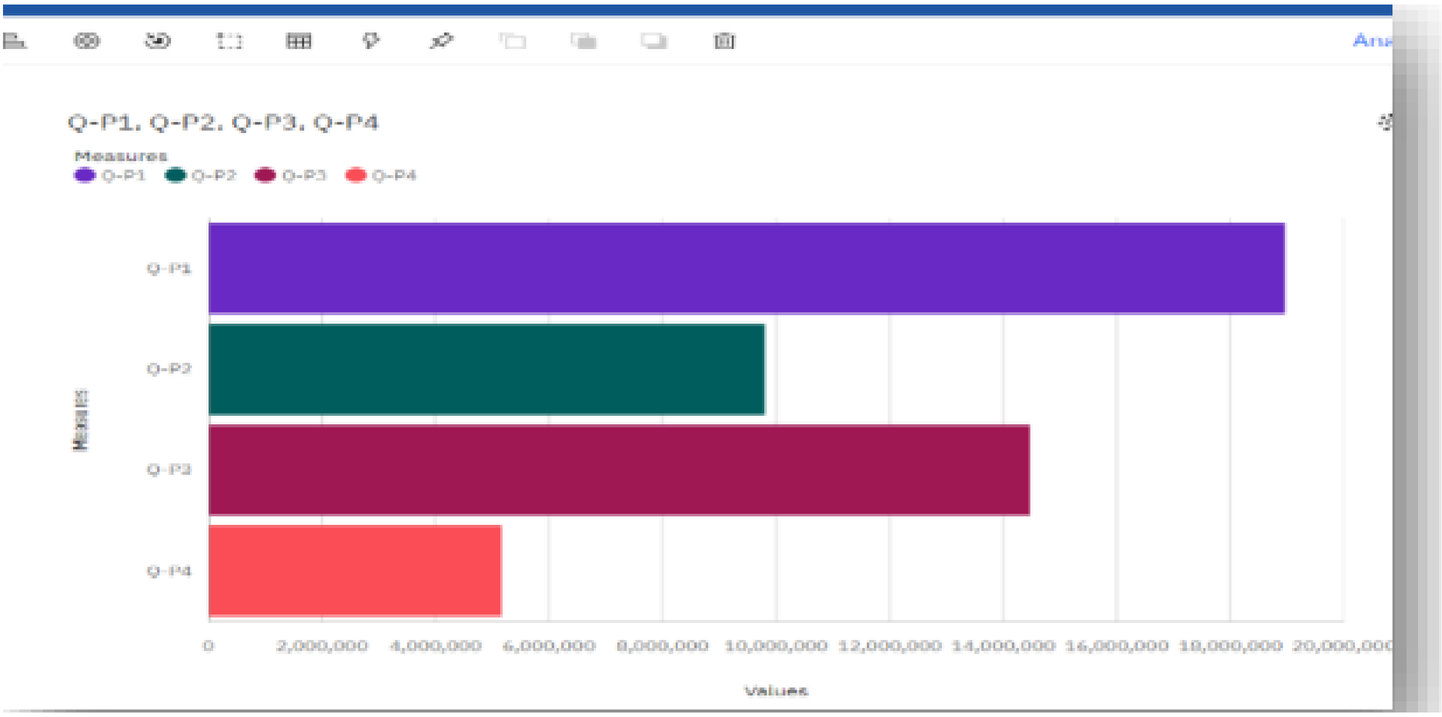
Here we can observe that the s-p3 has highest revenue among all the products. And S-P4 has the least revenue.

Revenue Analysis: In our analysis of the dataset, we focused on the total revenue generated by different products, including S-P1, S-P2, S-P3, and S-P4, over a specific period. This examination allowed us to assess the financial performance of each product.

S-P3's Revenue Leadership: It became evident from the data that S-P3 stands out as the product with the highest revenue. This indicates that, among the products in the dataset, S-P3 has consistently been the top revenue generator. Several factors may contribute to this, including a high price point, strong customer demand, or effective marketing strategies.

S-P4's Revenue Challenges: On the other end of the spectrum, S-P4 emerges as the product with the least revenue. This suggests that S-P4 has faced challenges in generating revenue compared to the other products in the dataset. Possible reasons for this could be lower demand, competitive pricing pressures, or other market dynamics.

#Plotting bar chart to analyse which product has highest unit sales



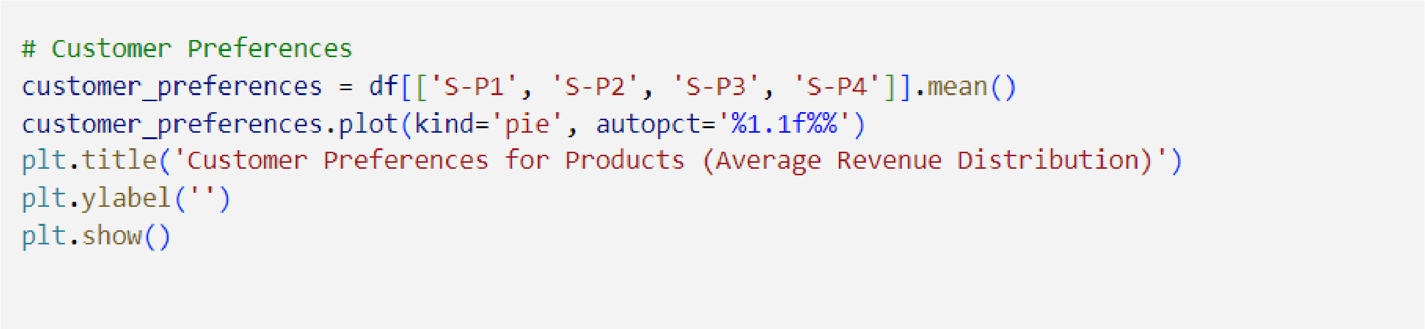
- Here we can observe that the Q-P1 has highest unit sales among all the products.
- And Q-P4 has the least unit sales.
- Our analysis of the dataset clearly demonstrates that Q-P1 stands out as the product with the highest unit sales. This means that, over the given period,

more customers have purchased Q-P1 compared to any other product. The reasons behind this could vary - it might have a strong market demand, effective marketing, competitive pricing, or superior product quality. Identifying these factors can help businesses further leverage Q-P1's success.

Step 3:
Display insights such as top-selling products, sales trends, and customer preferences.

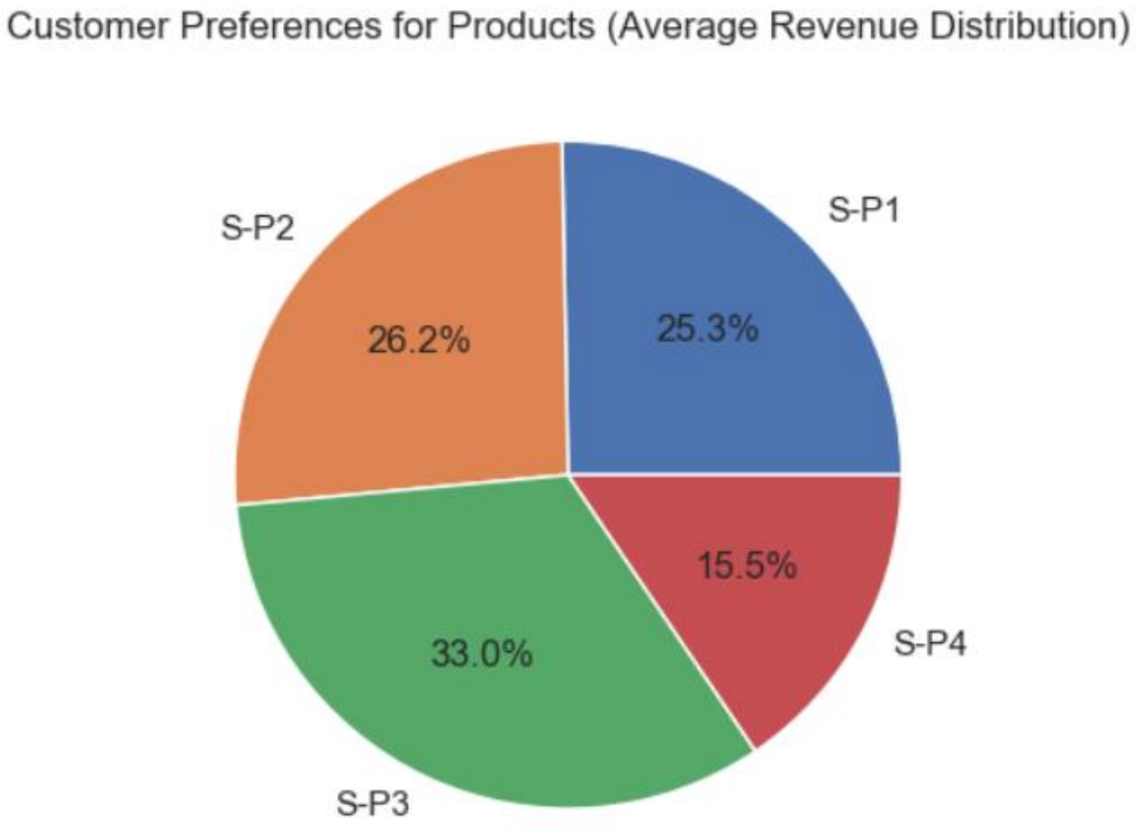
# Finding Customer preferences

By using the following python code we can findout the customer preference that is the product which is liked or prefered by the customer atmost.



```python
# Customer Preferences
customer_preferences = df[['S-P1', 'S-P2', 'S-P3', 'S-P4']].mean()
customer_preferences.plot(kind='pie', autopct='%1.1f%%')
plt.title('Customer Preferences for Products (Average Revenue Distribution)')
plt.ylabel('')
plt.show()
```

Output:



- From the output(pie chart) we can see that the product which is prefered by most of the people is S-P3 -it has 33 percentage of revenue.
- And the least prefered product by the people is S-P4 with 15.5 percentage of revenue.

- **Data Analysis:** we aimed to identify customer preferences for different products. We explored a dataset containing information on total unit sales and revenue generated by four different products, namely S-P1, S-P2, S-P3, and S-P4, over a specific period.

- **Pie Chart Visualization**: To determine customer preferences, we turned to data visualization. A key insight emerged from a pie chart we created using the data. This chart clearly displayed the market share of each product, making it evident which product was preferred by customers.

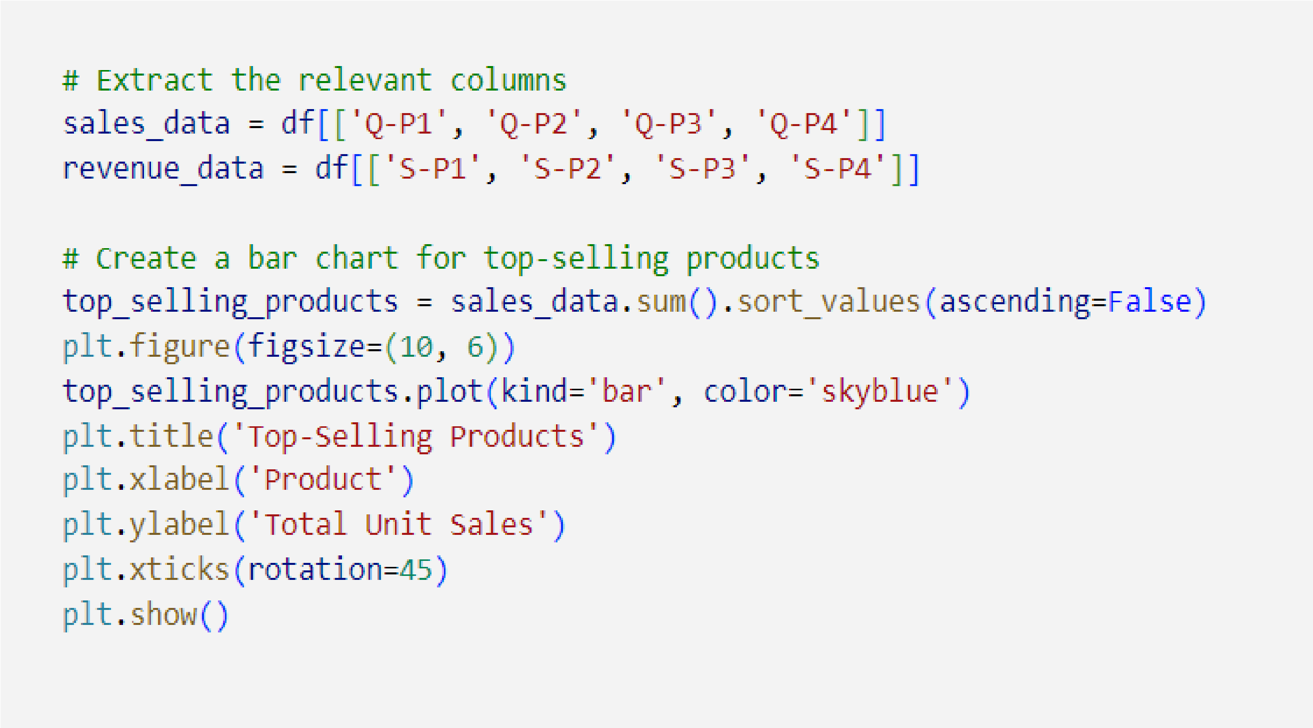
- **Identification of Customer Preference:** The pie chart unmistakably indicated that product S-P3 held the largest slice of the market share. This finding signifies that customer preference leaned significantly toward S-P3, making it the most popular product among customers.

- **Actionable Insight:** Armed with this insight, businesses can focus their efforts on promoting and optimizing the sales of S-P3. This actionable insight allows for more targeted marketing strategies and product improvements to capitalize on the product's popularity and boost overall sales and revenue.

In summary, a pie chart was employed to visualize customer preferences, and it revealed that product S-P3 stands out as the most preferred item, which can serve as a valuable guide for business decisions and strategies.

# Finding top selling Product

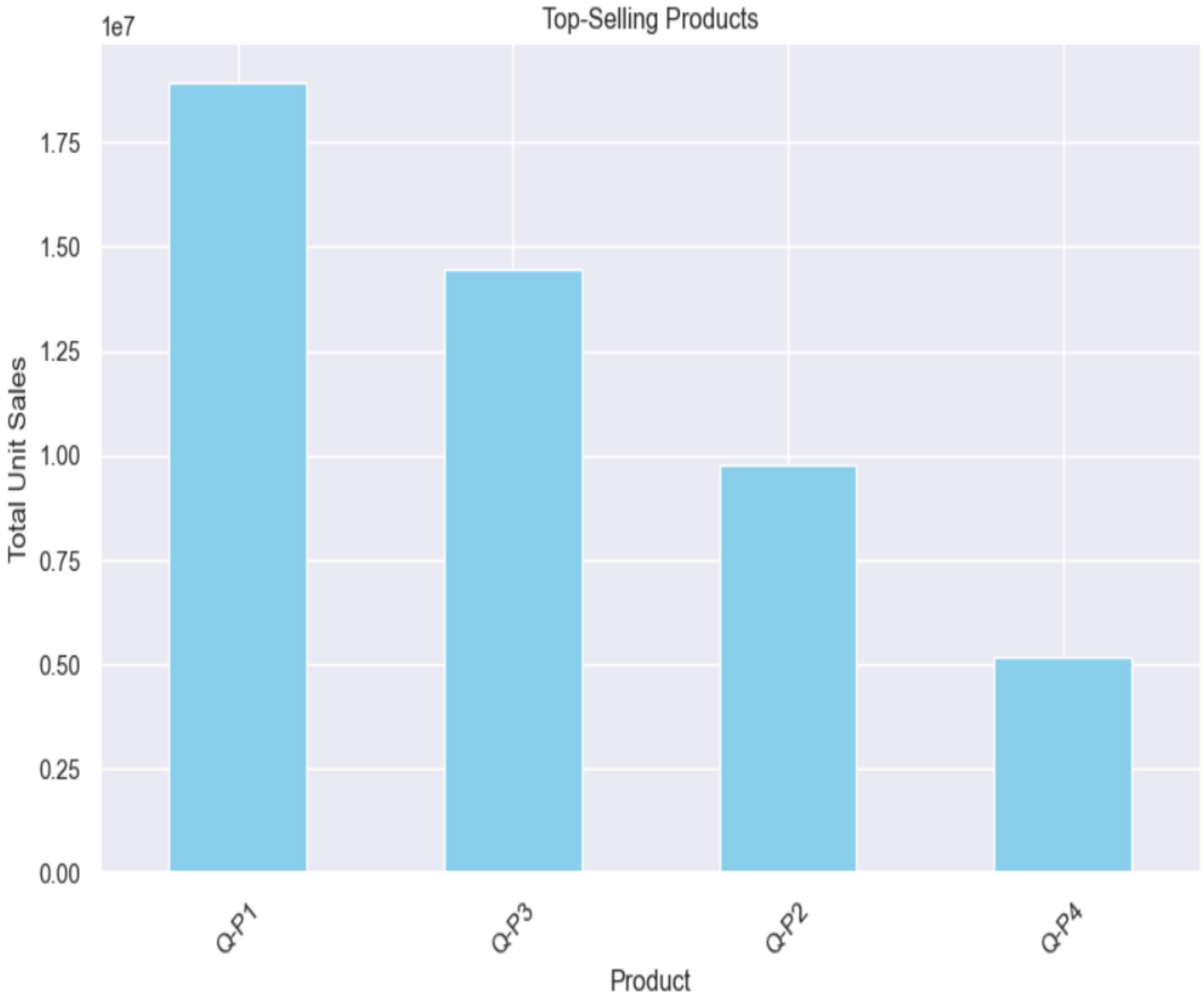
By using the following python code we can finfout the top selling product (the product which is sold high more than others)



```
# Extract the relevant columns
sales_data = df[['Q-P1', 'Q-P2', 'Q-P3', 'Q-P4']]
revenue_data = df[['S-P1', 'S-P2', 'S-P3', 'S-P4']]

# Create a bar chart for top-selling products
top_selling_products = sales_data.sum().sort_values(ascending=False)
plt.figure(figsize=(10, 6))
top_selling_products.plot(kind='bar', color='skyblue')
plt.title('Top-Selling Products')
plt.xlabel('Product')
plt.ylabel('Total Unit Sales')
plt.xticks(rotation=45)
plt.show()
```

Output:



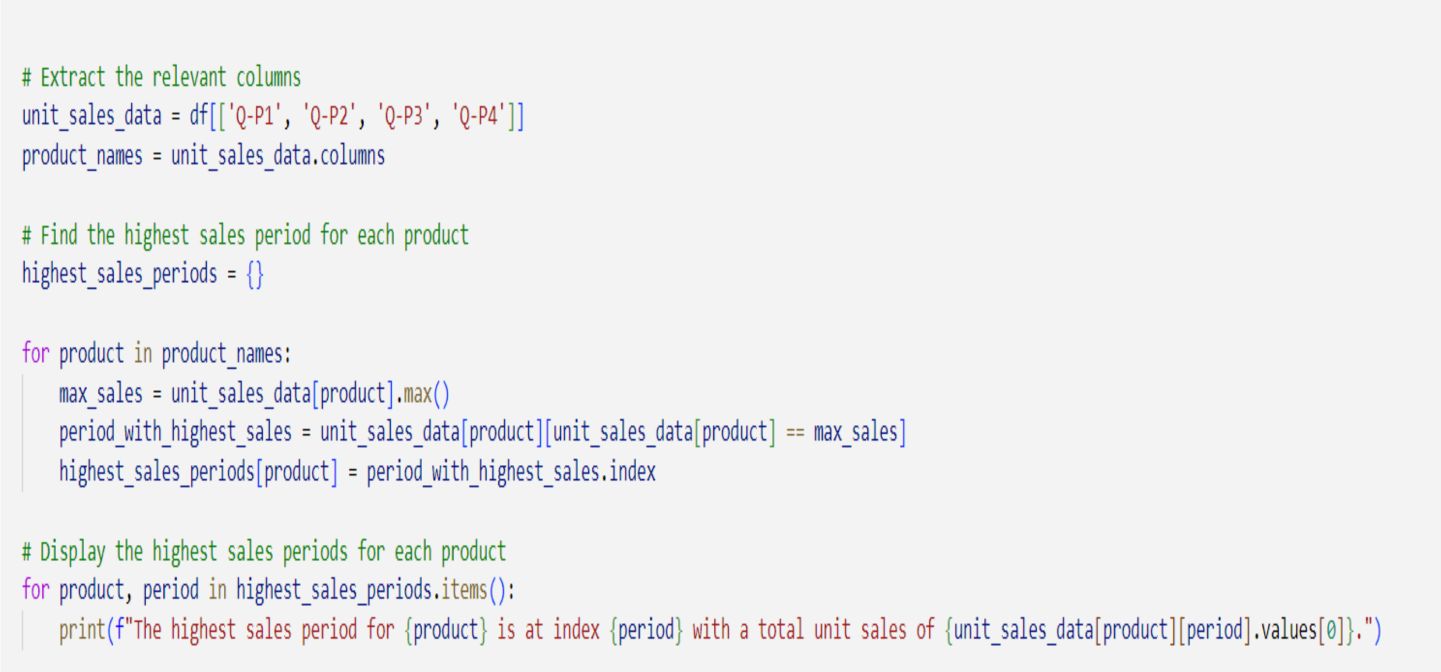
From the output we can see that the product with the top selling is Q-P1
And the product with low selling is Q-P4.

Bar Chart Representation: In our data analysis, we used a bar chart to visually represent the total unit sales of each product. Each product was represented as a bar, and the height of each bar corresponds to the total unit sales of that product. This graphical representation allowed for a quick and clear comparison of sales figures across different products.

Q-P1's Sales Leadership: It became evident from the bar chart that Q-P1 outshines the other products in terms of unit sales. The bar representing Q-P1 is noticeably taller than the bars of the other products. This indicates that Q-P1 has consistently achieved the highest sales figures, signifying its strong market presence.

# peak sales periods

We can also find the products peak sales without using visualization we can use the following python code to find that each products peak sales periods.



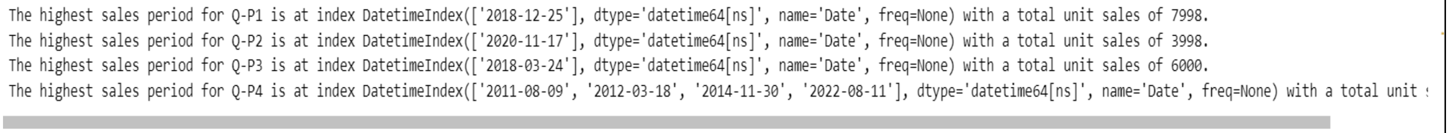
```
# Extract the relevant columns
unit_sales_data = df[['Q-P1', 'Q-P2', 'Q-P3', 'Q-P4']]
product_names = unit_sales_data.columns

# Find the highest sales period for each product
highest_sales_periods = {}

for product in product_names:
    max_sales = unit_sales_data[product].max()
    period_with_highest_sales = unit_sales_data[product][unit_sales_data[product] == max_sales]
    highest_sales_periods[product] = period_with_highest_sales.index

# Display the highest sales periods for each product
for product, period in highest_sales_periods.items():
    print(f"The highest sales period for {product} is at index {period} with a total unit sales of {unit_sales_data[product][period].values[0]}.")
```

Output:



```
The highest sales period for Q-P1 is at index DatetimeIndex(['2018-12-25'], dtype='datetime64[ns]', name='Date', freq=None) with a total unit sales of 7998.
The highest sales period for Q-P2 is at index DatetimeIndex(['2020-11-17'], dtype='datetime64[ns]', name='Date', freq=None) with a total unit sales of 3998.
The highest sales period for Q-P3 is at index DatetimeIndex(['2018-03-24'], dtype='datetime64[ns]', name='Date', freq=None) with a total unit sales of 6000.
The highest sales period for Q-P4 is at index DatetimeIndex(['2011-08-09', '2012-03-18', '2014-11-30', '2022-08-11'], dtype='datetime64[ns]', name='Date', freq=None) with a total unit
```

1. **Q-P1**: The highest sales period for Q-P1 occurred on December 25, 2018. During this period, a total of 7,998 units of Q-P1 were sold. This date seems to coincide with a significant spike in demand or a successful sales campaign for Q-P1, resulting in this high sales figure.

2. **Q-P2**: The peak sales period for Q-P2 was on November 17, 2020, with a total of 3,998 units sold. This suggests that Q-P2 experienced a surge in demand or achieved remarkable sales performance on this specific date.

3. **Q-P3**: The highest sales period for Q-P3 took place on March 24, 2018, with a total of 6,000 units sold. This date likely corresponds to a time when Q-P3 was in high demand, possibly due to factors like promotions, product launches, or market trends.

4. **Q-P4**: Interestingly, Q-P4 had multiple periods with the highest unit sales. These occurred on August 9, 2011, March 18, 2012, November 30, 2014, and August 11, 2022, with a total of 2,000 units sold during each of these periods. The occurrence of multiple peaks for Q-P4 indicates that its sales performance had several successful phases throughout the years. This could be due to various factors, including marketing strategies, seasonal trends, or unique product features.

In summary, each product has experienced its highest sales periods at different times, likely influenced by various factors that drove customer demand and sales performance. Understanding these peak sales periods can help businesses better plan their marketing and product promotion strategies to capitalize on these successful periods in the future.

Conclusion:

In summary, the analysis points out that Q-P1 has been consistently the top-selling product, whereas Q-P4 has faced difficulties in achieving high sales figures. These findings suggest a need for focused marketing and sales strategies for both products. Additionally, Q-P2 and Q-P3 have also had their successful sales periods, indicating opportunities to capitalize on these achievements in the future. Business decisions should be guided by these insights to maximize sales and revenue.