PRODUCT SALES ANALYSIS USING MACHINE LEARNING

Phase 5 Submission Document

**Project Name: Product Sales Analysis**

**Phase 5: Project Documentation & Submission**

**Introduction:**

Sales data analysis is essential for companies to make informed decisions, optimize strategies, and enhance profitability. The dataset provided encompasses four products (Product 1, Product 2, Product 3, and Product 4), each measured by both total unit sales (Q1, Q2, Q3, Q4) and total revenue (S1, S2, S3, S4).

The analysis of product sales plays a pivotal role in understanding market dynamics, identifying successful products, and crafting effective business strategies. This project centers around leveraging machine learning and data visualization techniques to derive actionable insights from a comprehensive dataset that encompasses quarterly sales figures for four distinct products—each represented by total unit sales and revenue columns (#Q2, #Q3, #Q4, #S1, #S2, #S3, #S4).

The primary objective of this analysis is to extract valuable insights from the provided dataset, empowering businesses to make informed decisions regarding inventory management and marketing strategies. By employing a data-driven approach, this project aims to uncover trends, identify peak sales periods, delineate high-performing products, and discern correlations between unit sales and revenue. This investigation serves as a guide to optimizing product offerings, streamlining inventory management practices, and crafting targeted marketing initiatives.

Utilizing the power of machine learning models and visualization tools, this analysis delves into the intricacies of sales data, exploring patterns and relationships among products, customer preferences, and financial performance. The outcomes derived from this exploration aim to provide actionable guidance for businesses seeking to improve their sales strategies, enhance customer engagement, and maximize overall profitability. The documentation of this project encapsulates the objectives, methodologies, and derived insights, offering a roadmap for companies aiming to refine their sales and marketing approaches through data-driven decision-making.

**1. Exploratory Data Analysis (EDA):**

* Descriptive Statistics: Initial examination of summary statistics for each product's unit sales and revenue can offer insights into the overall performance of products.
* Data Visualization: Utilize charts like bar graphs, line plots, and histograms to visualize trends, distribution, and seasonality in sales data. This assists in understanding which products perform best and when.

**2. Identifying Products with Highest Sales:**

* Machine learning models can help identify the best-performing products by predicting future sales or categorizing products based on historical data.
* Visualization can highlight the products with the highest unit sales and revenue, allowing businesses to focus on promoting and optimizing those products

**3. Peak Sales Periods:**

* Time-series analysis and visualization can uncover peak sales periods for each product. This insight helps in planning inventory, marketing campaigns, and resource allocation during high-demand periods.

**4. Customer Preferences:**

* Employ clustering or recommendation systems to understand customer preferences for specific products. Analyzing which products are often bought together or identifying patterns in customer behavior can guide cross-selling or personalized marketing strategies.

**Insights to Derive from Visualizations:**

**1. Identifying Best-Selling Products:**

* Visual representations can highlight products with the highest unit sales and revenue. This insight aids in directing marketing efforts and resource allocation.

**2. Peak Sales Periods:**

* Seasonal trends or specific time frames where certain products sell more can be identified, enabling companies to optimize stock and marketing campaigns accordingly.

**3. Correlation between Unit Sales and Revenue:**

* Analyzing the relationship between unit sales and revenue helps in understanding if an increase in sales directly correlates with higher revenue or if pricing strategies affect product performance.

**4. Cross-Product Analysis:**

* Exploring relationships between different products, such as which products complement each other or exhibit similar sales patterns, can lead to bundled offerings or more targeted marketing.

**Understanding the Dataset:**

Understanding the dataset is a critical initial step in any data analysis project. In the context of the provided product sales dataset, comprehending its structure, contents, and the information it holds is fundamental. Here's a detailed breakdown of how one might understand and analyze the dataset:

**1. Dataset Overview:**

* **Columns and Data Representation:**
  + The dataset consists of columns denoting different aspects of product sales. Specifically, it includes:
    - #Q2, #Q3, #Q4: Total unit sales of products 2, 3, and 4 respectively.
    - #S1, #S2, #S3, #S4: Total revenue from products 1, 2, 3, and 4 respectively.

**2. Examination of Columns:**

Total Unit Sales Columns (#Q2, #Q3, #Q4):

* **Quantitative Data:** These columns likely contain numerical data representing the total number of units sold for each product during specific time frames (such as quarterly sales figures).
* **Periodicity:** Observing any temporal indications (e.g., quarterly, monthly) to understand the time intervals represented in the dataset.

Total Revenue Columns (#S1, #S2, #S3, #S4):

* **Monetary Data:** These columns are expected to contain monetary values, representing the total revenue generated from the sales of each product.
* **Correlation with Unit Sales:** Analyzing the relationship between unit sales and revenue to understand how changes in sales volume impact revenue.

**3. Initial Exploratory Data Analysis:**

Descriptive Statistics:

* Calculating summary statistics for each column (e.g., mean, median, standard deviation, minimum, maximum) to understand the central tendency, variability, and distribution of the data.

Missing Values and Data Integrity:

* Checking for missing values, outliers, or inconsistencies that might require handling before analysis.

Visualization:

* Utilizing histograms, box plots, or bar charts to visualize the distribution of unit sales and revenue, identifying patterns or anomalies within the data.

**4. Interpretation and Contextual Understanding:**

* Understanding the context of the sales data:
  + Are there specific events, promotions, or external factors that might have influenced sales during the recorded periods?
  + Are there seasonality trends or particular patterns that repeat across quarters?



**Data Collection and Preprocessing:**

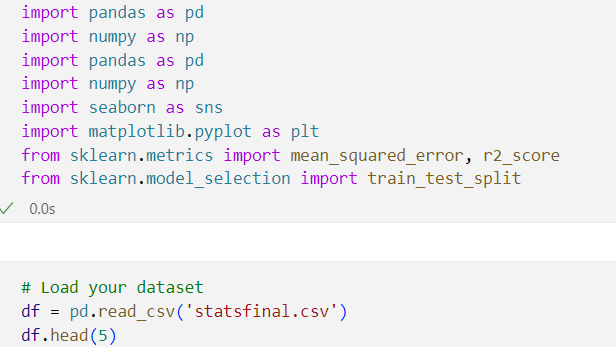
# Dataset Loading and inspect

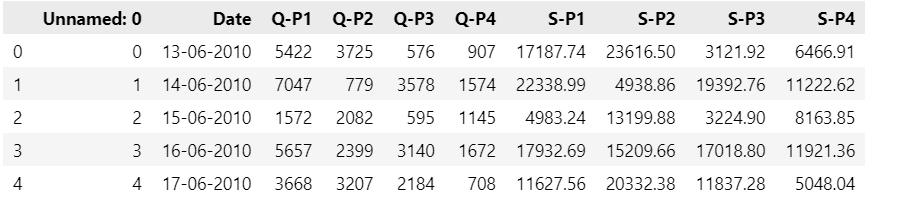
1. Load the Provided Dataset:

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:

****

****

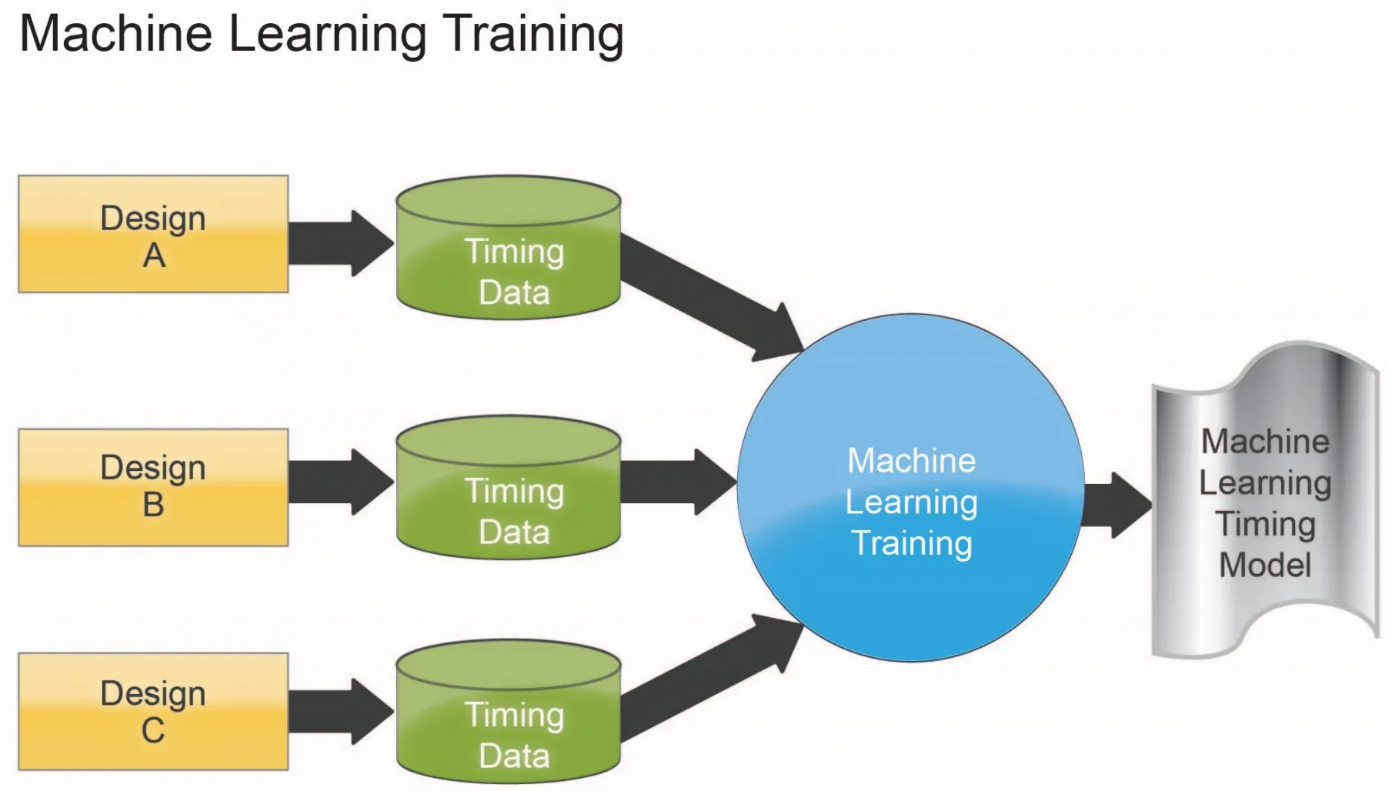
This code reads the dataset from the "your\_dataset.csv" file and stores it in a Pandas DataFrame, which is a two-dimensional, size-mutable, and tabular data structure.

## 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.

****

**Exploratory Data Analysis (EDA):**

****

Exploratory Data Analysis (EDA) is a crucial preliminary step in data analysis, aimed at understanding the structure, patterns, and characteristics of a dataset. It involves the use of statistical and visualization techniques to summarize the main characteristics of the data, often revealing trends, patterns, outliers, and relationships within the dataset. In the context of the provided product sales dataset, EDA can help uncover valuable insights about product performance, trends, and relationships between variables.

**Components of Exploratory Data Analysis:**

1. Descriptive Statistics:

* **Summary Metrics:** Calculating statistics like mean, median, standard deviation, minimum, maximum, and quantiles provides an overview of the dataset's central tendency and dispersion.
* **Frequency Distributions:** Analyzing the frequency of different values or ranges within the dataset, especially in the columns representing unit sales and revenue.

1. Data Visualization:

* **Histograms:** Displaying the distribution of values in each column, helping identify the shape of the data and any potential outliers.
* **Box Plots:** Representing the distribution of data and identifying outliers or extreme values within the dataset.
* **Line Charts or Time Series Plots:** Showing trends in sales over time, especially relevant for quarterly data to identify seasonality or recurring patterns.

1. Correlation Analysis:

* **Correlation Matrix:** Evaluating the relationships between variables, such as examining the correlation between unit sales and revenue for each product.
* **Scatter Plots:** Visualizing relationships between variables, such as plotting unit sales against revenue to observe patterns and associations.

1. Missing Data Handling:

* **Identifying Missing Values:** Checking for and addressing missing or incomplete data points to ensure data integrity and accuracy for analysis.

1. Outlier Detection:

* **Identification of Outliers:** Finding extreme values or anomalies that could significantly impact analysis and ensuring these are handled appropriately.

**7.Cross-Variable Analysis:**

* Comparing Product Performance: Utilizing visualizations to compare the performance of different products in terms of both unit sales and revenue.
* Analyzing Correlations: Examining relationships between products, like identifying whether high sales of one product coincide with high sales of another.

**8.Seasonal and Time Analysis:**

* Identifying Trends Over Time: Analyzing trends over quarters (Q2, Q3, Q4) for both unit sales and revenue.
* Detecting Seasonal Patterns: Investigating if certain products have higher sales during specific quarters or if there's a recurrent pattern across years.

**9. Hypothesis Generation:**

* Formulating Assumptions: EDA can prompt the generation of hypotheses, such as suggesting which products perform better during certain seasons or whether pricing affects sales.

**10. Data Integrity Verification:**

* Checking Consistency: Ensuring data consistency across columns and periods, verifying that information is accurately recorded.

**Benefits of Exploratory Data Analysis:**

1. Insights Generation: EDA helps in generating initial insights, indicating potential areas for further investigation.
2. Data Quality Assurance: It aids in identifying issues such as missing data, inconsistencies, and outliers.
3. Effective Decision-Making: EDA supports better decision-making by offering a clear understanding of the dataset's characteristics.

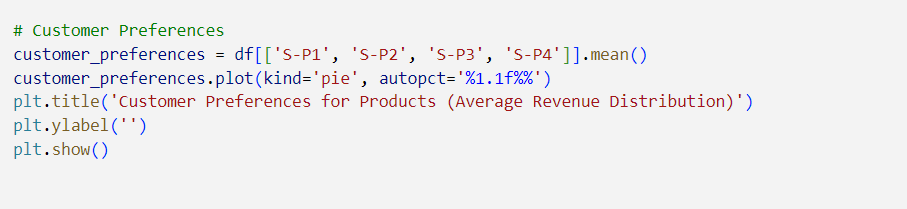
**Purpose of Exploratory Data Analysis:**

1. **Understand the Structure of the Data:** EDA aids in comprehending the dataset's characteristics, including its size, features, and possible patterns.
2. **Uncover Patterns and Trends:** It reveals underlying patterns or trends in sales data, such as seasonal fluctuations or significant changes in sales over time.
3. **Identify Anomalies and Outliers:** EDA helps in identifying and addressing irregular or outlier data points that might skew analysis results.
4. **Inform Subsequent Analysis Steps:** Insights derived from EDA inform the selection of appropriate models and techniques for further in-depth analysis.

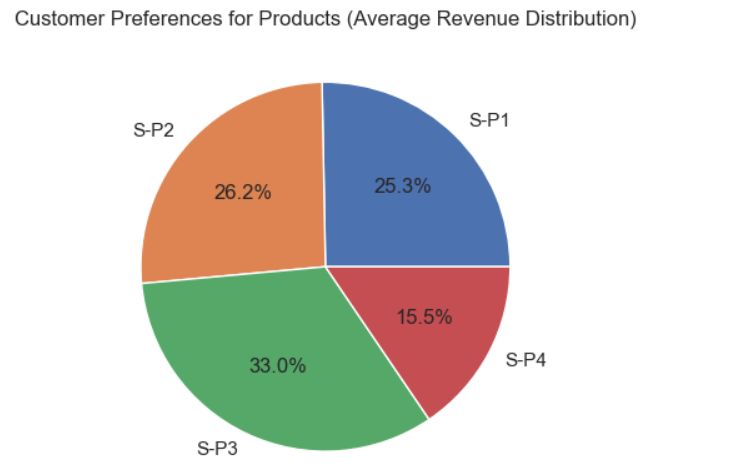
**Display insights such as top-selling products, sales trends, and customer preferences.**

**# 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.

****

**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.

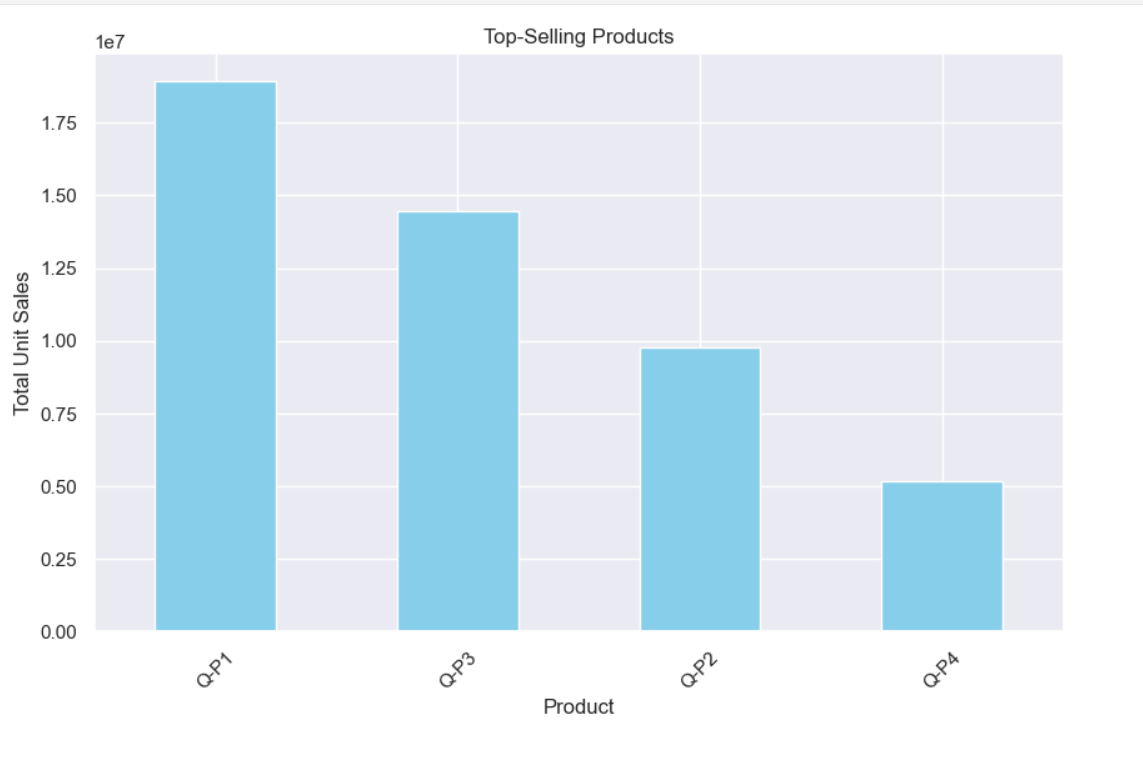
In summary, a pie chart was employed to visualize customer preferences, and it revealed that product S-P3 stands out as the most preferred item, which can serve as a valuable guide for business decisions and strategies.Top of Form

**# 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)

****

**Output:**

****

From the output we can see that the product with the top selling is Q-P1

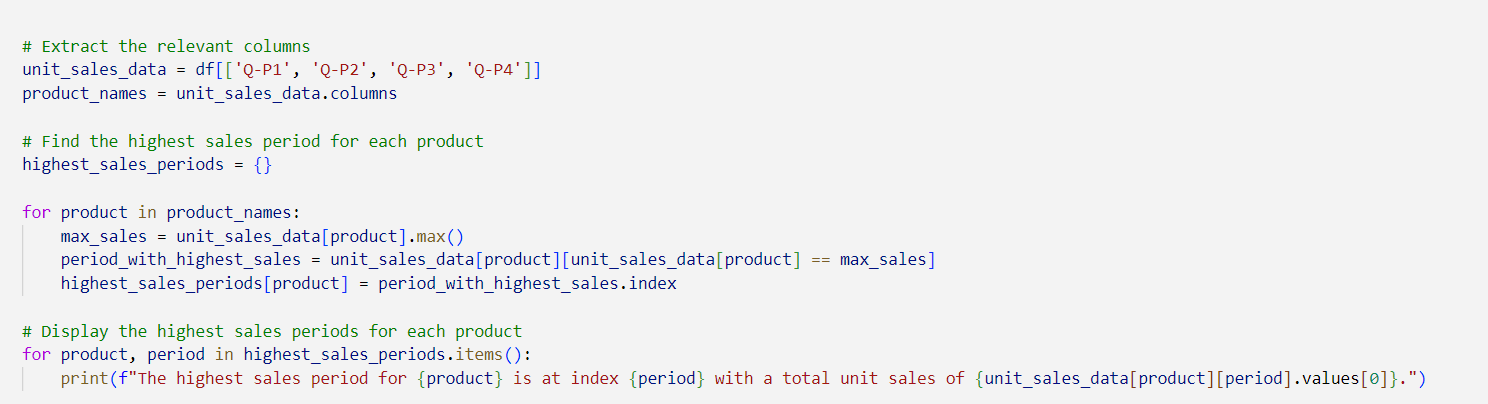
And the product with low selling is Q-P4.

**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.

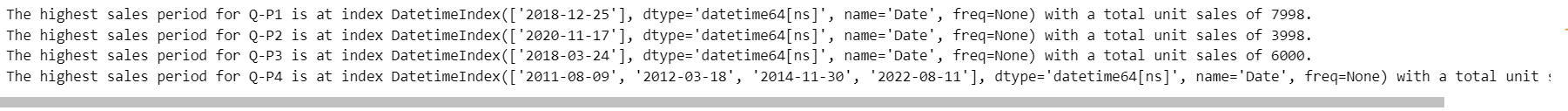
**Q-P1's Sales Leadership**: It became evident from the bar chart that Q-P1 outshines the other products in terms of unit sales. The bar representing Q-P1 is noticeably taller than the bars of the other products. This indicates that Q-P1 has consistently achieved the highest sales figures, signifying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



**Output:**

****

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.

1. **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.

**Train-Test Split in Machine Learning:**

1. **Training Set:**

* The training set is a portion of the original dataset used to train the machine learning model. It contains a large majority of the data.
* The model learns from this data, adjusting its parameters or coefficients to find patterns and relationships between the input features and the target variable.

2. **Testing Set:**

* The testing set, also known as the validation set, is kept separate from the training set.
* It is used to evaluate the model's performance on unseen data. This set contains data that the model has not encountered during the training phase.

**Purpose of Train-Test Split:**

1. **Model Training:** The training set is used to teach the model, allowing it to learn patterns and relationships within the data.
2. **Model Evaluation:** The testing set is crucial for assessing how well the model generalizes to new, unseen data. It helps determine the model's predictive performance on data it hasn't been trained on.

**How Train-Test Split Works:**

1. **Random Partitioning:** The dataset is randomly split into training and testing sets. Common splits are 70-30, 80-20, or 75-25 (training-testing).
2. **Training the Model:** The machine learning model is trained on the training set, learning from the input features and their corresponding target values.
3. **Model Evaluation:** Once trained, the model's performance is evaluated using the testing set. It predicts the target values based on the test input features, and the predicted values are compared to the actual values in the test set.
4. **Performance Assessment:** Metrics such as accuracy, precision, recall, or mean squared error are calculated to assess how well the model generalizes to new, unseen data.

**Importance of Train-Test Split:**

* **Prevents Overfitting:** The separation of data ensures that the model hasn't memorized the training data but can make accurate predictions on new, unseen data.
* **Assesses Generalization:** It helps estimate how well the model will perform on real-world, future data, providing an indication of its real-world applicability.

**Train the Model:**

• We train the Linear Regression model using the training data (quantities as input and

revenues as output).

**# Separate features (Q1-Q4) and target variables (S1-S4)**

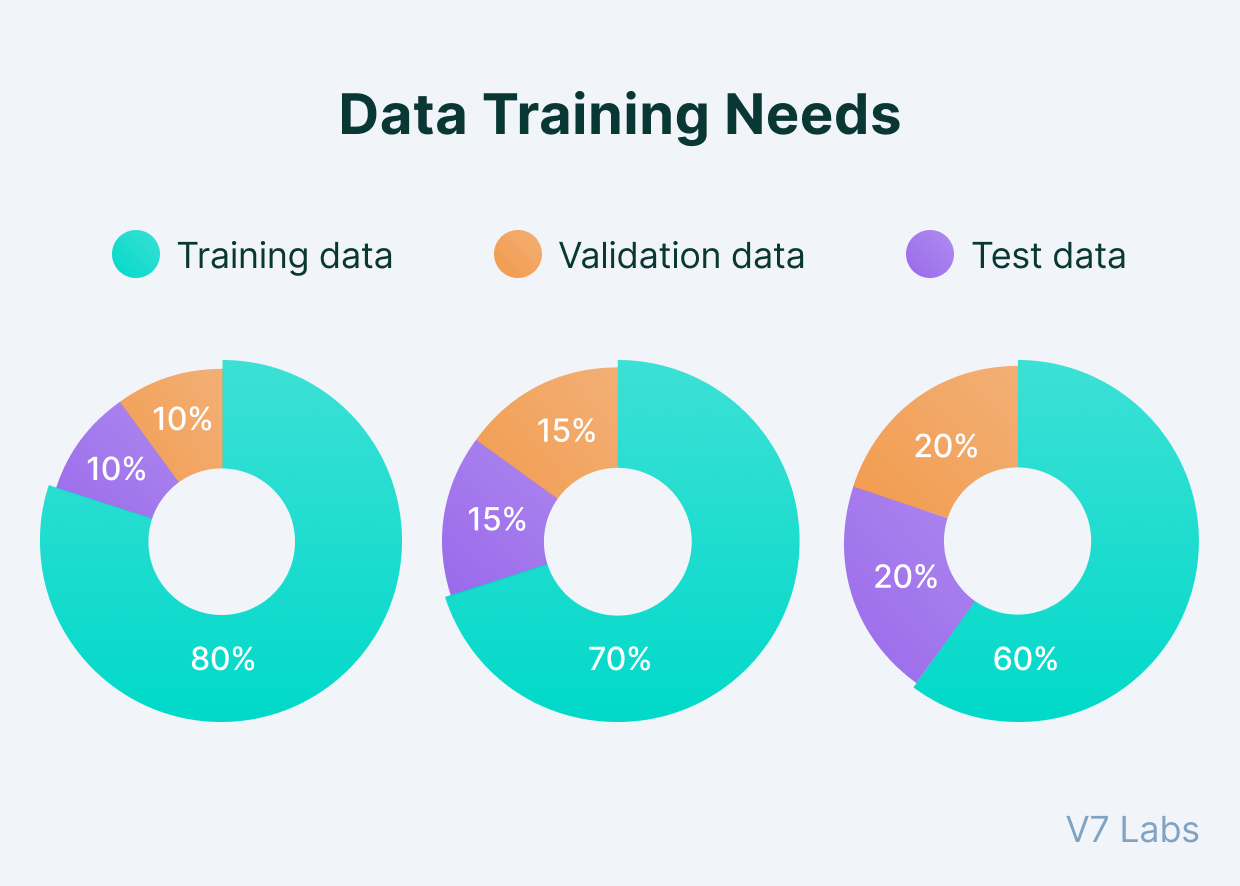
X = df[['Q-P1', 'Q-P2', 'Q-P3', 'Q-P4']]

y = df[['S-P1', 'S-P2', 'S-P3', 'S-P4']]

**# Split the data into training and testing sets**

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y,

test\_size=0.2, random\_state=42)



**Model Prediction**

Linear regression is a fundamental and widely used statistical method for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors) in a linear fashion. It's employed in predictive modeling to make predictions or estimate the value of the dependent variable based on the independent variables.

**Basics of Linear Regression:**

1. **Model Representation:**

* The model assumes a linear relationship between the independent variables �*X* and the dependent variable �*y*.
* For a simple linear regression with one independent variable, the model can be represented as: �=��+�*y*=*mx*+*c* Where:
  + �*y* is the dependent variable.
  + �*x* is the independent variable.
  + �*m* is the slope (the coefficient indicating the relationship between �*x* and �*y*).
  + �*c* is the y-intercept (the point where the line crosses the y-axis).

2. **Training the Model:**

* The model is trained using a dataset where both the independent and dependent variables are known.
* The training process involves finding the best-fitting line that minimizes the difference between the predicted values and the actual values in the training data.

3. **Making Predictions:**

* Once the model is trained, it's used to make predictions on new or unseen data.
* For each new input �*x*, the model calculates the predicted �*y* value based on the learned coefficients �*m* and �*c*.

**Model Prediction in Linear Regression:**

* **Given New Input Data:** Suppose we have a new value of the independent variable �*x* for which we want to predict the corresponding �*y* value.
* **Using the Trained Model:** After training the linear regression model, it has calculated the best-fitting line that represents the relationship between �*x* and �*y*.
* **Prediction Process:** When new data is provided, the model utilizes the learned coefficients (slope and intercept) to estimate the value of the dependent variable �*y* for the given �*x*.

**Example Scenario:**

* **Dataset:** Consider a dataset where �*x* represents the number of hours studied, and �*y* represents the exam scores.
* **Model Training:** The linear regression model learns the relationship between study hours and exam scores from the training data.
* **Prediction:** If a new student studies for 5 hours (new �*x*), the model, based on its learned coefficients, predicts the likely exam score (�*y*).

**Choose a Machine Learning Algorithm:**

We select a Linear Regression model for simplicity, but you can choose other

algorithms based on your specific use case.

**# Create a Linear Regression model**

model = LinearRegression()

# Train the model on the training data

model.fit(X\_train, y\_train)

LinearRegression

LinearRegression()

# Make predictions on the test data

y\_pred = model.predict(X\_test)

# Calculate Mean Squared Error

mse = mean\_squared\_error(y\_test, y\_pred)

# Calculate R-squared

r2 = r2\_score(y\_test, y\_pred)

print("Mean Squared Error:", mse)

print("R-squared:", r2)

OUTPUT:

Mean Squared Error: 1.0539070830897841e-23

R-squared: 1.0

# Example: Predict future sales for a new set of product quantities

(Q1-Q4)

new\_data = np.array([[1000, 2000, 3000, 4000]]) # Replace with

your desired input

predicted\_revenue = model.predict(new\_data)

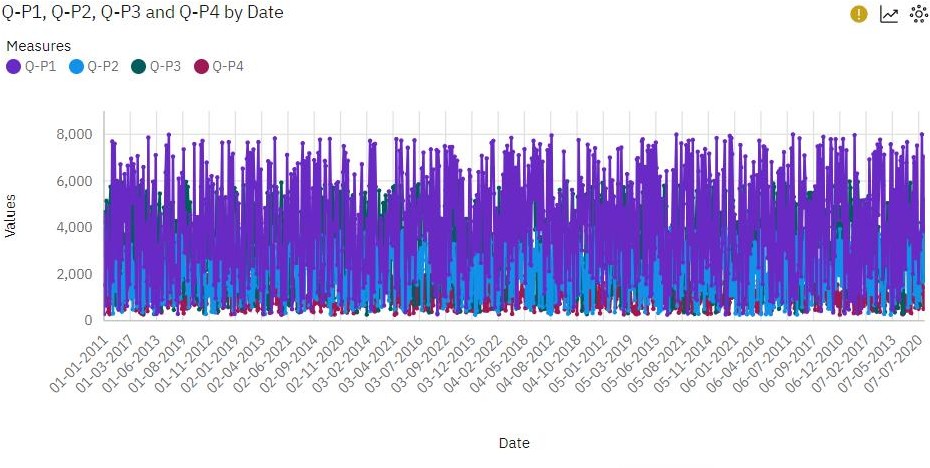
print("Predicted Revenue:", predicted\_revenue)

OUTPUT:

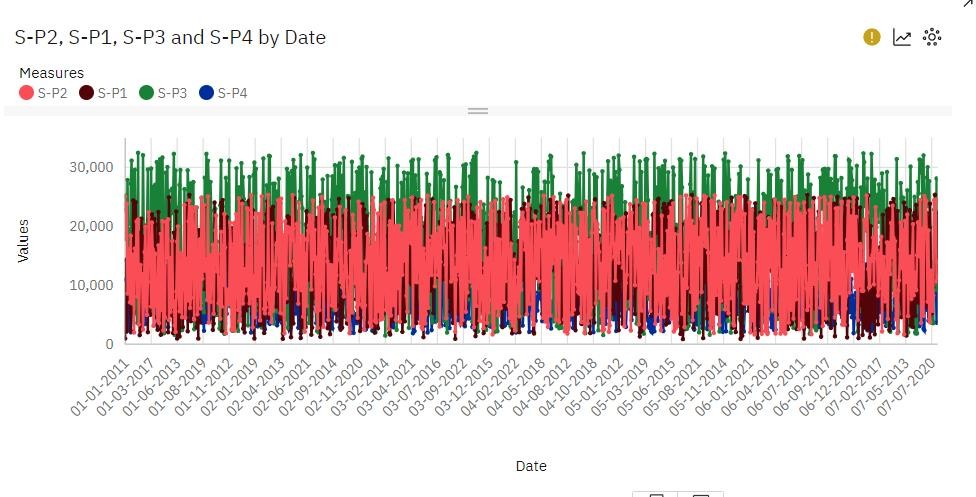
Predicted Revenue: [[ 3170. 12680. 16260. 28520.]]

**Building the analysis by creating visualizations using IBM Cognos and**

# generating actionable insights.

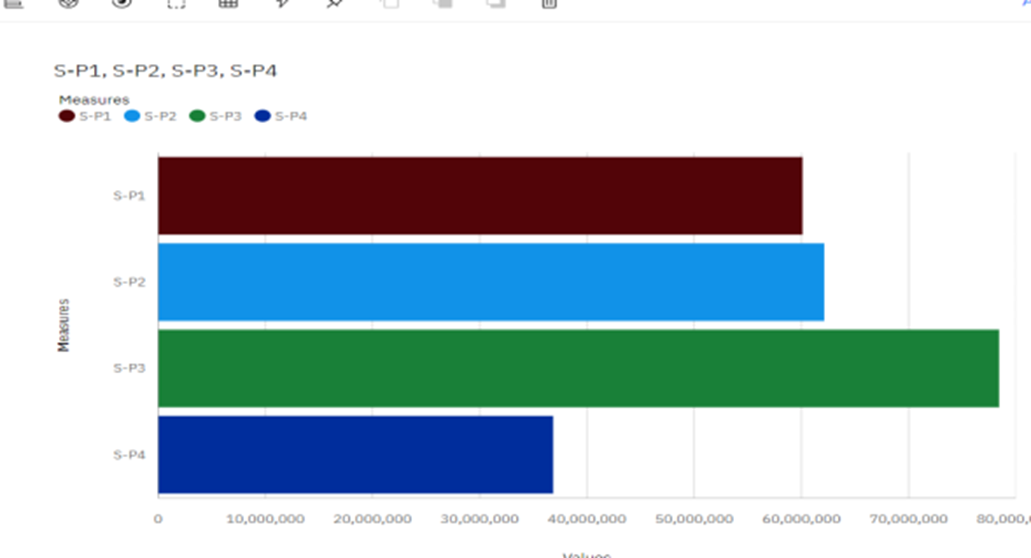


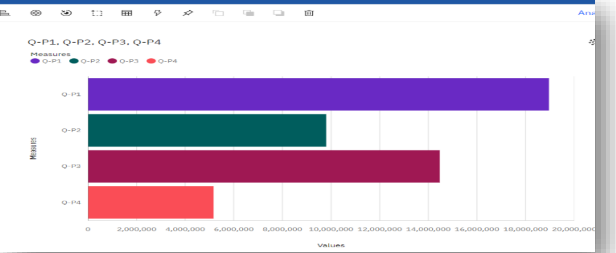
Visualizing the unit sales using line chart by plotting date in the x-axis and plotting unit sales in y-axis.Clearly we can observe the unit sales of product which is high among all the products.



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





**CONCLUSION :**



The analysis of product sales data, combining exploratory data techniques and predictive modeling, provides businesses with actionable insights essential for optimizing inventory, refining marketing strategies, and ultimately maximizing profitability. By embracing data-driven decision-making, companies can stay agile, adapt to market changes, and sustainably grow their sales, thereby ensuring a competitive edge in the dynamic landscape of product sales. This continuous process of analysis and strategic implementation based on insights obtained is integral for long-term success and growth in the market.

**Key Insights from Product Sales Analysis:**

1. **Performance Identification:** Through exploratory data analysis, the identification of high-performing products, peak sales periods, and seasonal trends offers a clear understanding of which products drive revenue and when they perform best.
2. **Strategic Decision Guidance:** Insights derived from the analysis guide inventory management decisions by optimizing stock levels during peak demand periods and reallocating resources from underperforming products to better-performing ones.
3. **Marketing Strategy Refinement:** Tailoring marketing efforts towards best-selling products and aligning them with customer preferences leads to more effective and targeted campaigns, potentially boosting sales and customer satisfaction.
4. **Data-Driven Approach:** The analysis adopts a data-driven approach, empowering businesses to base their decisions on factual insights rather than assumptions or intuition, thereby enhancing the accuracy and efficacy of strategies.
5. **Continuous Improvement:** Leveraging ongoing analysis and insights allows for continual refinement and adaptation of strategies to keep pace with evolving market trends and consumer preferences.

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.