**Product Sales Analysis with**

College code: 0001

**Machine Learning**

**Phase 4**

**Team Members:**

Diya Arshiya S (202115033) [diya.arshiya@gmail.com](mailto:(diya.arshiya@gmail.com))

Dhivyadharshini S K (2021115030) [dhivyadharshini0907@gmail.com](mailto:dhivyadharshini0907@gmail.com)

Mukesh Raja K (2021115065) [mukeshrajatmr2021@gmail.com](mailto:mukeshrajatmr2021@gmail.com)

Mukilarasan V (2021115066) [mukilarasan.v@gmail.com](mailto:mukilarasan.v@gmail.com)

Karthik V (2021115321) [karthiksk9360@gmail.com](mailto:karthiksk9360@gmail.com)

**Table of Contents**

1. Introduction
2. Data preprocessing

3. Visualization

4. Output

5. Insights

6. Conclusion

**Introduction**

The purpose of this report is to document the data analysing performed on the dataset contained in the "statsfinal.csv" file. Data analysing is a crucial step in data analysis. It gives the visualized form of data.

**Data preprocessing**

**Data Overview**

We begin by loading the dataset using the Python library `**pandas**`. The dataset is read from the *statsfinal.csv* file, and some initial information about the dataset is displayed using the `**info()**` method.

**Data Cleaning**

**i. Identifying the missing values:**

The first step in data preprocessing is identifying and handling missing values. Missing values can disrupt the analysis and modeling process. In this dataset, we identify missing values using the `**isnull().sum()`** method, which counts the number of missing values in each column.

**ii. Dropping Rows with Missing Values**

Even though there are no missing values, it is good practice to drop rows with missing data when necessary. This can be done using the `dropna()` method.

**iii. Removing Duplicates**

Duplicate rows can also affect the accuracy of analysis. To remove duplicate rows, the `drop\_duplicates()` method is used.

**iv. Data Formatting**

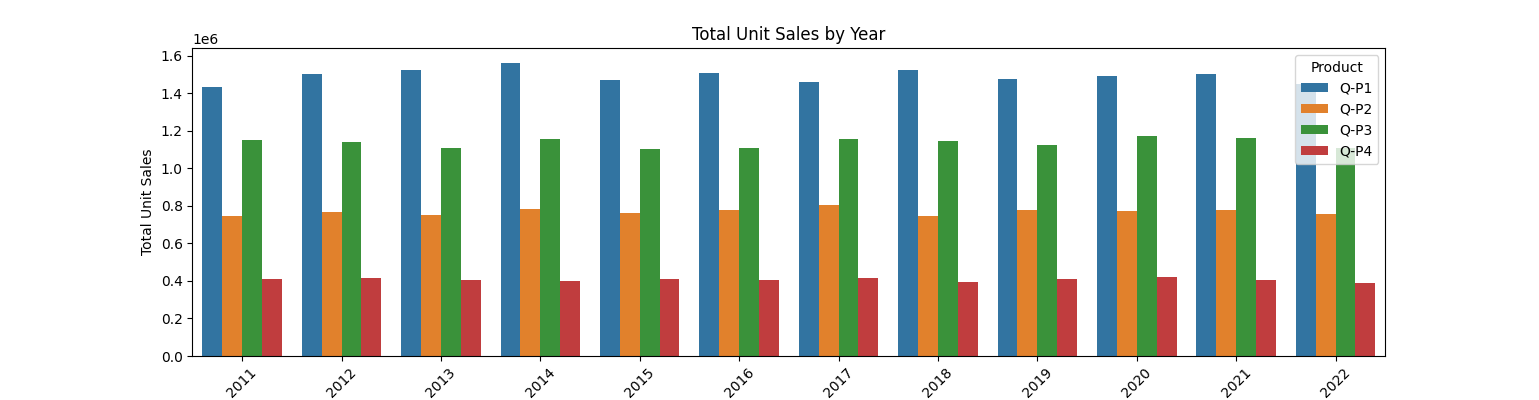
The next preprocessing step involves formatting the data, specifically by separating the date into separate columns for "Day," "Month," and "Year." This is achieved by applying a lambda function to split the "Date" column.

**v. Data Reduction**

In some cases, certain data points may need to be removed due to inconsistencies or insufficient data. In this dataset, data for the years 2010 and 2023 are removed as they have insufficient data. Additionally, incorrect date entries for September 31st and November 31st are also removed.

**Visualization**

**Total Unit sales of products P1, P2, P3, P4 from 2011 to 2022**

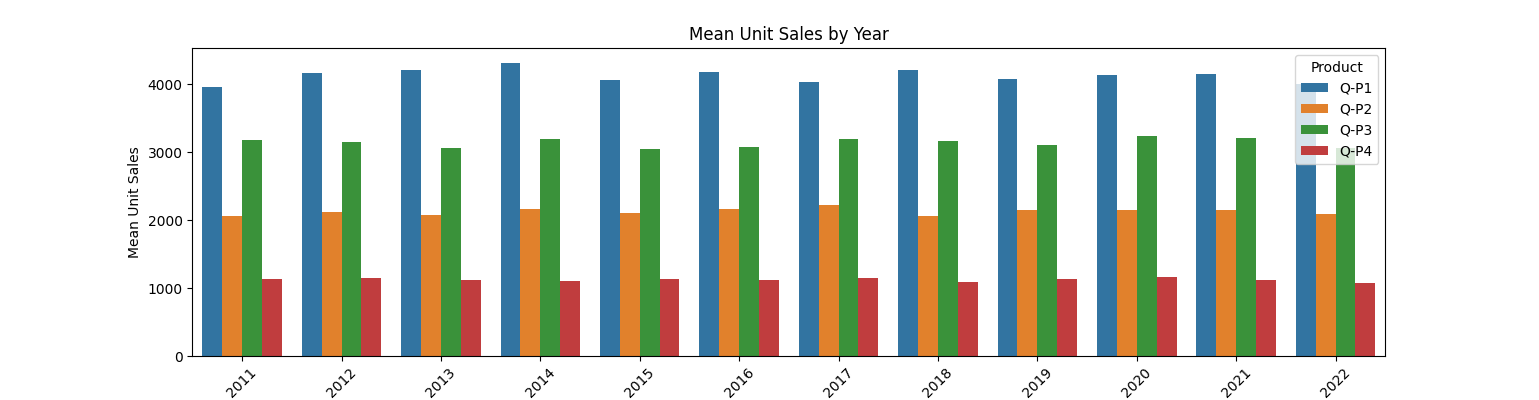
****

**Insights**

Total unit sales have been relatively consistent from 2011 to 2022.

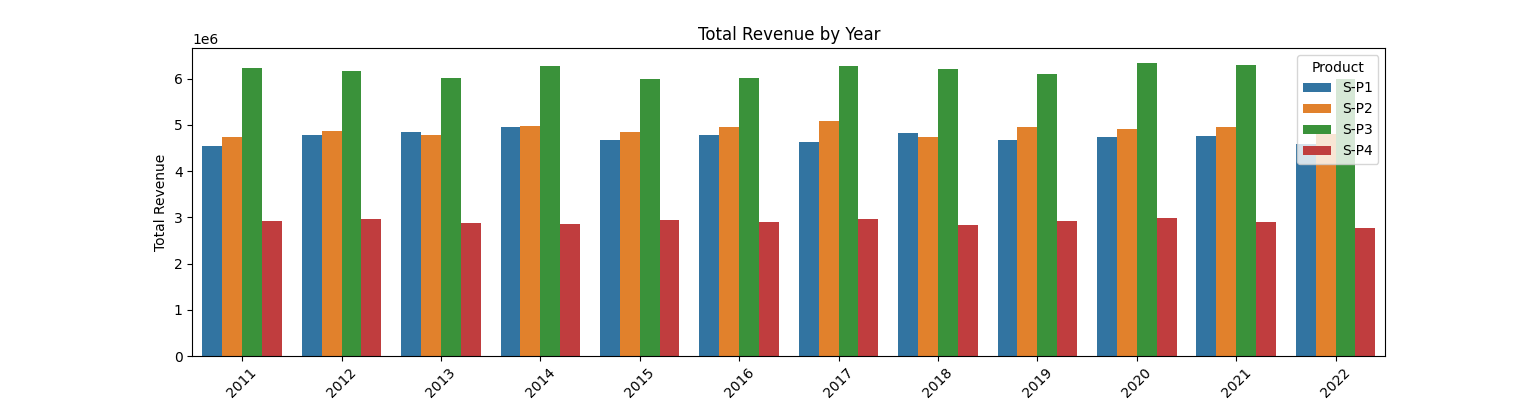
Product Q-P2 consistently leads in total unit sales.

**Mean Unit sales of products P1, P2, P3, P4 from 2011 to 2022**



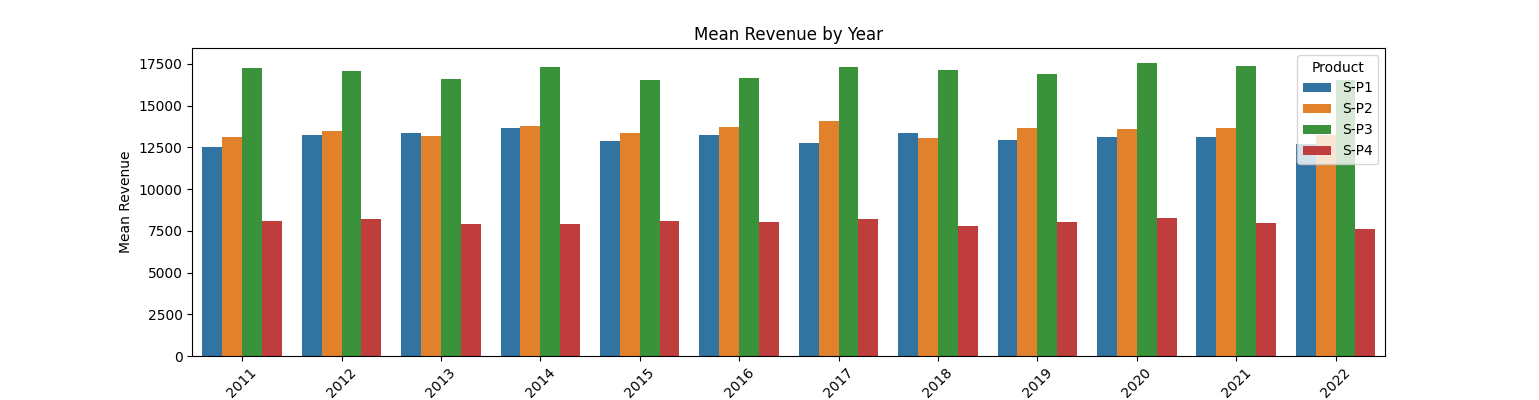
**Insights**

The mean unit sales for all products show a gradual increase over the years. Product Q-P4 has the highest mean unit sales in recent years.

**Total revenue earned from products P1, P2, P3, P4 from 2011 to 2022**

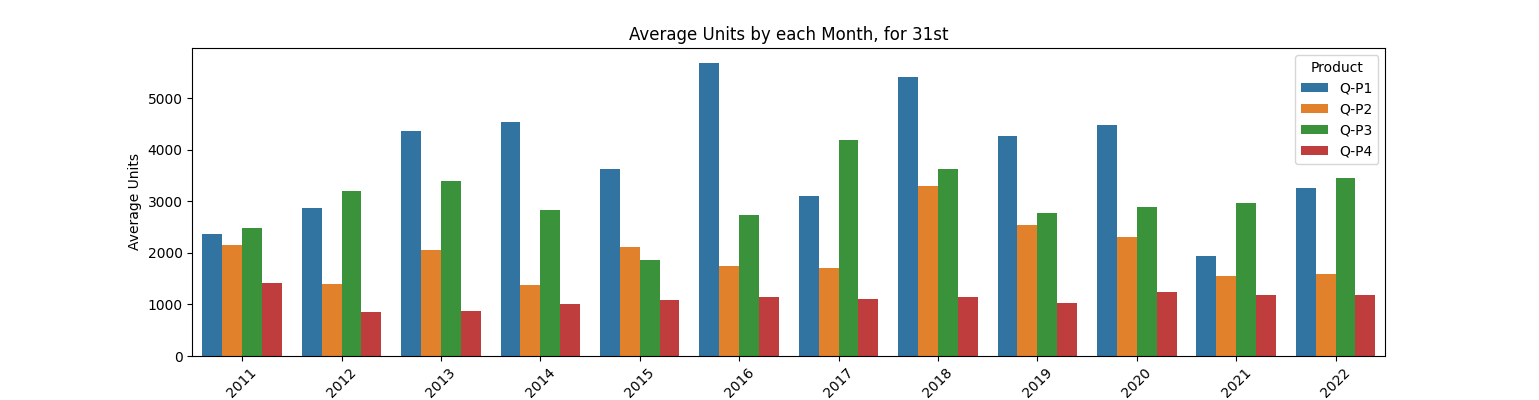
**Insights**

The revenue earned from product is consistent over the years. The product S-P1 earned more revenue

**Mean revenue earned from products P1, P2, P3, P4 from 2011 to 2022**

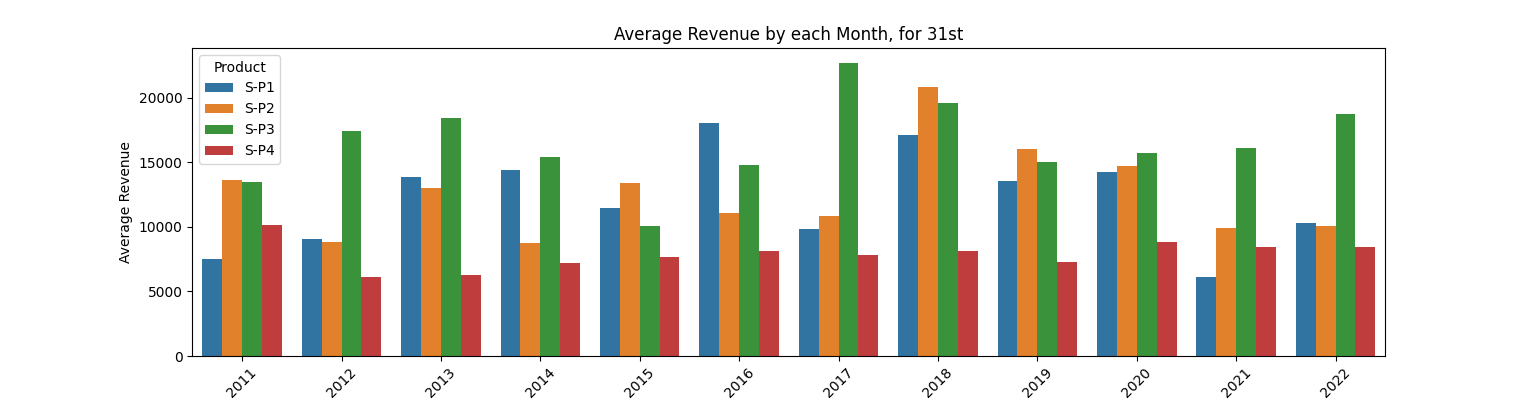
**Insights**

The revenue earned from product is consistent over the years. The product S-P1 earned more revenue

**Average units of product sold by each month, for 31st from 2011 to 2022**

**Insights**

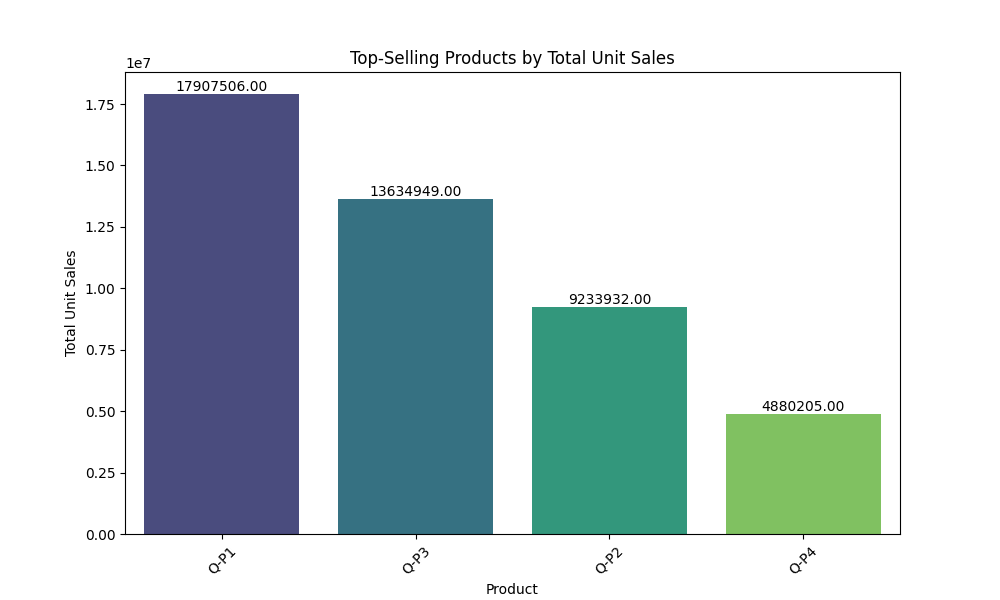
The units of product sold in each month is gradually increasing and decreasing over the years. The product Q-P1 is sold more

**Average revenue earned from product sold by each month, for 31st from 2011 to 2022**

**Insights**

The average revenue earned in each month is gradually increasing and decreasing over the years. The product S-P3 earned much.

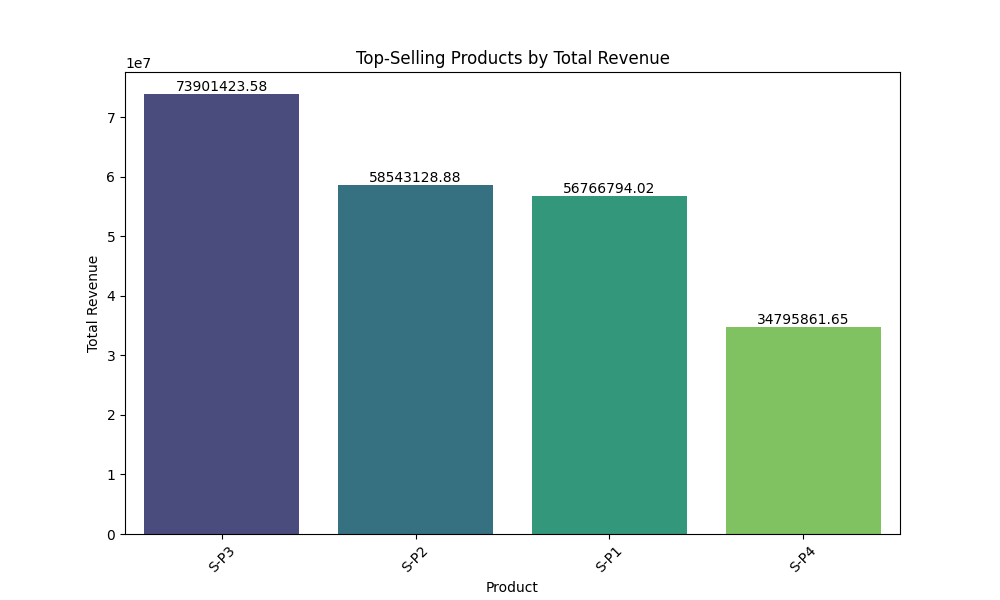
**Top selling products by total unit sales**



**Insights**

The product Q-P1 is the top selling product and Q-P4 is the least selling product based on unit sales

**Top selling products by total revenue**



**Insights**

The product S-P3 is the top selling product and S-P4 is the least selling product based on total revenue

**Conclusion**

The given product dataset is analysed and visualized