**PHASE 5**

**FINAL PROJECT DOCUMENTATION**

|  |  |
| --- | --- |
| **DATE** | 01 NOVEMBER 2023 |
| **TEAM ID** | 715 |
| **DOMAIN NAME** | DATA ANALYTICS |
| **PROJECT NAME** | PRODUCT SALES ANALYSIS |
| **MAXIMUM MARKS** |  |

**INTRODUCTION:**

The main objective of this phase is to explain project’s objective, design thinking process, data collection, cleaning and preprocessing and visualization of data using IBM Cognos that is used for predicting the future sales and customer preferences in product sales analysis.

**PROBLEM STATEMENT:**

Predict future sales performance for each product, region depending on top-selling products, peak sales analysis and customer preferences.

**LITERATURE SURVEY**

**Best Selling Product and Category Prediction Using Sales Analysis**

A sales analysis is a detailed report that tells about more profound understanding of a business’s sales performance, customer data, and the revenue. This tells you which deals are worth chasing and which are better left behind. Also, for the deals your sales team does decide to pursue, they’ll have a good approach ready to make the lead or customer more receptive to the sale. Using Sales Analysis helps to take retailers towards profit in this world of competition. Nowadays shopping malls keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. A predictive model is build using different algorithms. In this paper, we investigate forecasting sales for a Big Mart, with four machine learning algorithms (Random Forest, Linear Regression, Decision Tree and XG Booster Algorithms). The results show that the Random Forest algorithm performs better than the other two models.

**Pricing Strategies for Different Periods During Subsequent Selling Season for Seasonal Products**

This study applies a two-period pricing model to investigate the optimal pricing strategy for different periods during the subsequent selling season for seasonal products. The model assumes that the market is populated by two types of consumers, namely, myopic and strategic, and analyzes three optimal pricing strategies: one price (OP), preannounced slash price (PSP), and preannounced small price reduction (PSPR). Several propositions are derived by comparing these three strategies. Results show that the PSP strategy is superior to PSPR and OP strategies only when certain conditions are satisfied. Otherwise, the OP or the PSPR is consistently superior to the PSP. When retailers adopt a markdown price, they should reduce the price early to ensure a long second period. Finally, this study provides several numerical examples to illustrate the propositions derived from the theoretical analysis.

**How Online Product Reviews Affect Retail Sales: A Meta-analysis**

A growing body of research has emerged on online product reviews and their ability to elicit performance outcomes desired by retailers; yet, a common understanding of the performance implications of online product reviews has eluded us. Scholars continue to navigate an array of studies assessing different design elements of online product reviews, and various research settings and data sources. We undertake a meta-analysis of 26 empirical studies yielding 443 sales elasticities to examine how these variables relate to retail sales. Building on well-established meta-analytical methods, we address the following questions: How does review valence influence the elasticity of retailer sales? What about review volume? For which product types and usage situations do online product reviews have a greater impact on retailer sales elasticity? Which types of online reviewers and websites exert the greatest influence on retailer sales elasticity?

**An application for sales data analysis and visualization using Python and Django**

In the meantime, sales of the businesses are rising more and more. Companies want to trace their sales and their advancement. The entire data must be found and stored within the database. They have to see their company's expansion. The company wants to understand where they stand in comparison to different companies. Since this data can be in huge amounts, estimating the sales prediction is nearly difficult. Soon these terms, the physical estimation may become time consuming. Considering this problem, we've proposed the analysis and visualization of the data. Visualization is the graphical depiction of data with the help of graphs, tabulations, charts by which we could realize company sales. Using data visualization, we could form an opinion on a company's sales and sales by different salespeople. So, we've used a web application to integrate both the data analyzed as well as present that data to the end user with a simple user interface.

**Data Analysis and Visualization of Sales Data**

Data is being generated very rapidly due to increase in information in everyday life. Huge amount of data get accumulated from various organizations that is difficult to analyze and exploit. Data created by an expanding number of sensors in the environment such as traffic cameras and satellites, internet activities on social networking sites, healthcare database, government database, sales data etc., are example of huge data. Processing, analyzing and communicating this data are a challenge. Online shopping websites get flooded with voluminous amount of sales data every day. Analyzing and visualizing this data for information retrieval is a difficult task. Therefore a system is required which will effectively analyze and visualize data. This paper focuses on a system which will visualize sales data which will help users in applying intelligence in business,revenue generation, and decision making, managing business operation and tracking progress of tasks.

**PHASE – 1**

**DESIGN THINKING APPROACH:**

**Empathize:**

Understanding users plays a vital role while empathizing the problem. Preferences of customers will be changing time to time. So it is important to perform different tasks on customers to get a better result.

Actions to be performed:

* Keen observation
* Interviews
* Survey
* Questionnaries
* Active Listening

**Define:**

Defining a particular problem for the issue after analysing the customers through the tasks/actions performed during empathizing process.

The problem that our team is trying to focus on the prediction of future sales performance, since this is the main problem faced by the sales department due to inaccuracy and incompletion. The problem should be framed according to customer’s perspective.

Objectives:

* Pain Points
* Identify challenges
* Identify obstacles

**Ideate:**

Since the problem faced by department of sales is due to inaccuracy, the ideation comes with new ways to collect, analyse and using datasets.

Brainstorming among team members can bring out better solutions and open a new path for predicting data. Brainstorming on top selling products, peak sales periods, customer preferences and inventory management will be focussed by our team members and comes out with better solution.

Focussing on machine learning algorithms that gives better accuracy and using more data sources to gain more insights.

**Prototype:**

Prototyping the problem will help in identifying the best solution. Since the problem the predicting the future sales analysis, it is important to try out the identified machine learning algorithms for identifying better accuracy and best solution.

Actions to be performed:

* Collect historical sales data. This data should include information on product sales, region, channel, and time period.
* Clean and prepare the data. This involves removing errors and inconsistencies from the data, and formatting the data in a way that can be easily analysed.
* Choose a machine learning algorithm. There are a variety of machine learning algorithms that can be used for sales forecasting. Some popular algorithms include time series models, regression models, and machine learning ensembles.
* Train the model. The machine learning algorithm will need to be trained on the historical sales data. This will allow the model to learn the patterns and trends in your data.
* Generate forecasts. Once the model is trained, you can generate forecasts for future sales performance. The machine learning algorithm will be able to predict future sales based on the patterns and trends that it has learned from the historical data.

**Test:**

Since our team will provide some of the data for training and use remaining for testing, any type of machine learning algorithms will provide better accuracy and the best machine learning algorithm will be chosen.

Visualization tools like Power BI, Tableau, Cognos in order to produce charts, graphs and other visualizations that makes future predictions to be realized easier.

**Actions to be done with Cognos:**

Use reliable tools and software for better solution

Conduct statistical analysis

Iterate and refine based on the findings to enhanced accuracy and effectiveness.

**PHASE – 2**

The steps that will be followed for prediction of future sales and customer preferences by our team will be:

1. Load the dataset (available in skillup portal)
2. Prepare the data
3. Choose machine learning algorithm for prediction
4. Train the model
5. Test the model
6. Make predictions

Prepare the data

Import/Load the dataset into Jupyter Notebook

Getting sales dataset from Kaggle(Provided by IBM)

Separation of training and test data

Identifying ML algorithm-Classification algorithm

Predict the sales of product using algorithms

Test the model using remaining data.

Train the model using data.

Visualization of Data using Tools.

**LOAD THE DATASET:**

The dataset that is provided by IBM will be downloaded from Kaggle and the dataset will be imported. Jupyter Notebook will be used by our team for prediction of future sales and customer preferences. Python libraries like Numpy and Pandas will be used here.

**PREPARE THE DATA:**

The data which is imported in Jupyter Notebook will be cleaned and transformed for better accuracy. The raw data that is imported will be shaped according to the necessity. Similarly an analysis need to be done in order to define whether the data comes under classification or regression analysis.

**CHOOSE THE MACHINE LEARNING ALGORITHM:**

According to the dataset provided by IBM, the data are categorized. The dataset of sales data contain information like order priority which is categorized into high, low, critical and not specified. Therefore classification algorithms can be used for predicting future sales and customer preferences.

* **Logistic regression:** Logistic regression is a simple but powerful algorithm that can be used to model the probability of an event occurring. It is often used to predict whether a customer is likely to make a purchase, or not, whether or not a product is likely to sell.
* **Decision trees:** Decision trees can also be used for classification tasks. They work by constructing a tree-like model that represents the decisions and relationships between variables. Decision trees can be used to predict sales and customer preferences by considering multiple factors that impact these outcomes, such as customer demographics, product popularity, and past purchase history.
* **Support vector machines (SVMs):** SVMs are a type of classification algorithm that can be used to find a hyperplane that separates the data into different classes. SVMs are often used for classification tasks where the data is high-dimensional and sparse.
* **Random forests:**Random forests are an ensemble learning algorithm that combines multiple decision trees to produce a more accurate prediction. They are often used for classification tasks, such as predicting customer preferences or product demand.

**TRAIN THE MODEL:**

A model will be trained according to the data using python libraries like sklearn. Our team will train the model according to classification algorithms. After feeding the model that contain input and output data, the model learns to predict the output data for new input data.

**TEST THE MODEL AND MAKE PREDICTIONS:**

The data which is not used for training a model will be used for testing. Prediction of data will be done by using the metrices like accuracy, precision etc. When the model is performing well ,this model will be deployed and ensure that the prediction model is accurate and reliable.

**PHASE – 3**

**CLEANING AND PREPROCESSING:**

Cleaning and preprocessing of data are done using Jupyter Notebook by importing necessary libraries.

The data that is given in the dataset:

Q-P1: Total unit sales of Product 1

Q-P2: Total unit sales of Product 2

Q-P3: Total unit sales of Product 3

Q-P4: Total unit sales of Product 4

S-P1: Total revenue from Product 1

S-P2: Total revenue from Product 2

S-P3: Total revenue from Product 3

S-P4: Total revenue from Product 4

**Code with Output:**

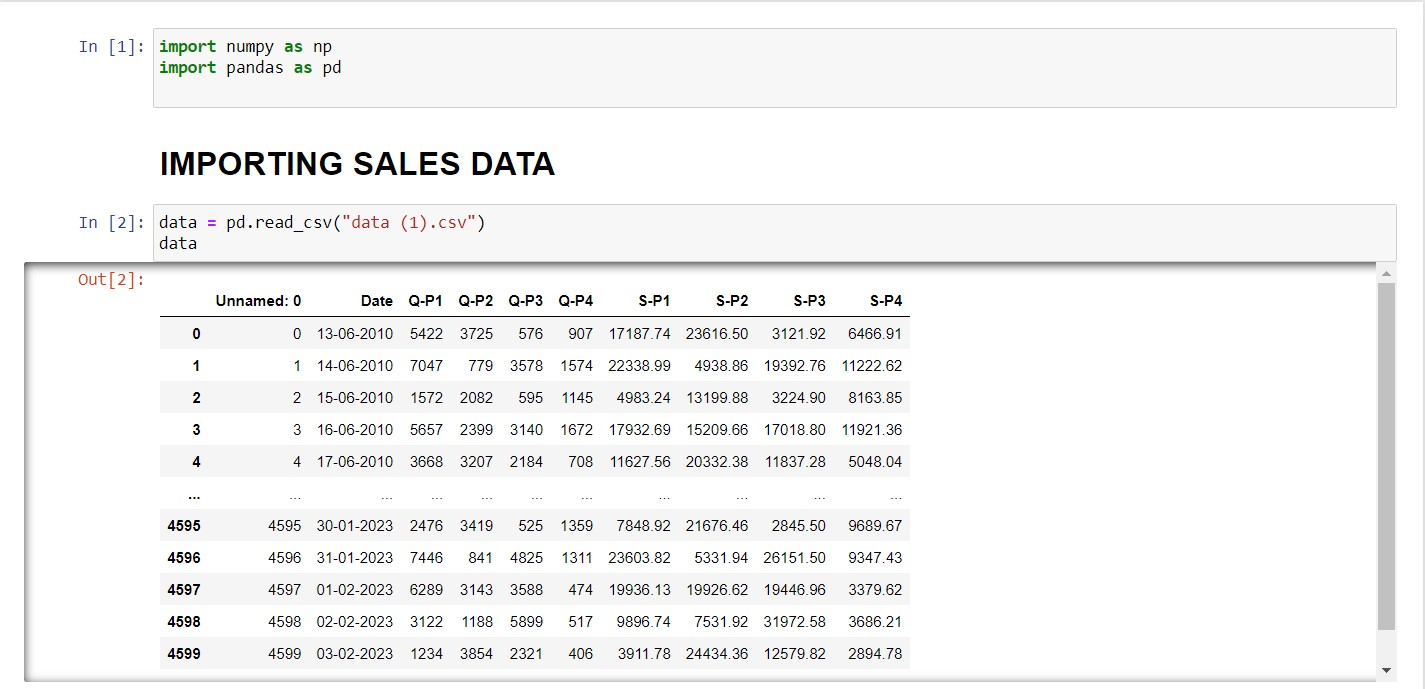
import numpy as np

import pandas as pd

**Import data**

data = pd.read\_csv("data.csv")

data

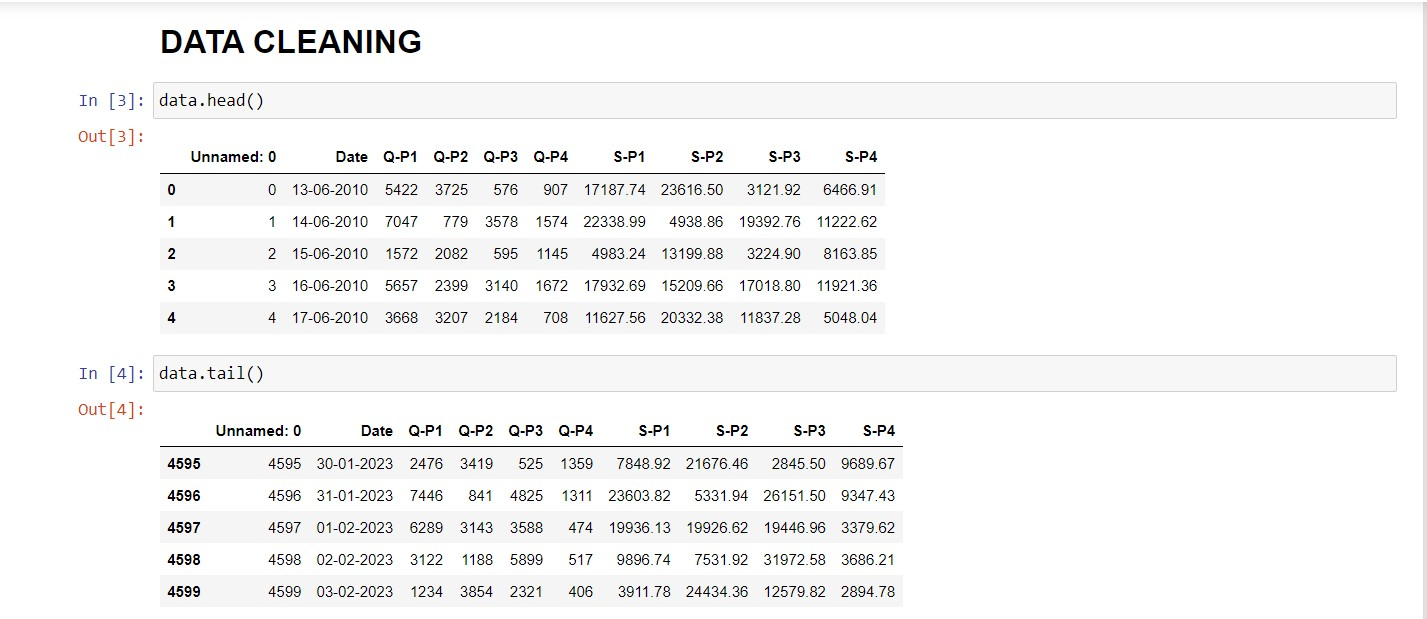


**Fig 5.1: Importing sales data along with numpy and pandas library**

**Cleaning the data**

data.head()

data.tail()



**Fig 5.2.1: Data Cleaning using data.head() and data.tail()**

data.describe()

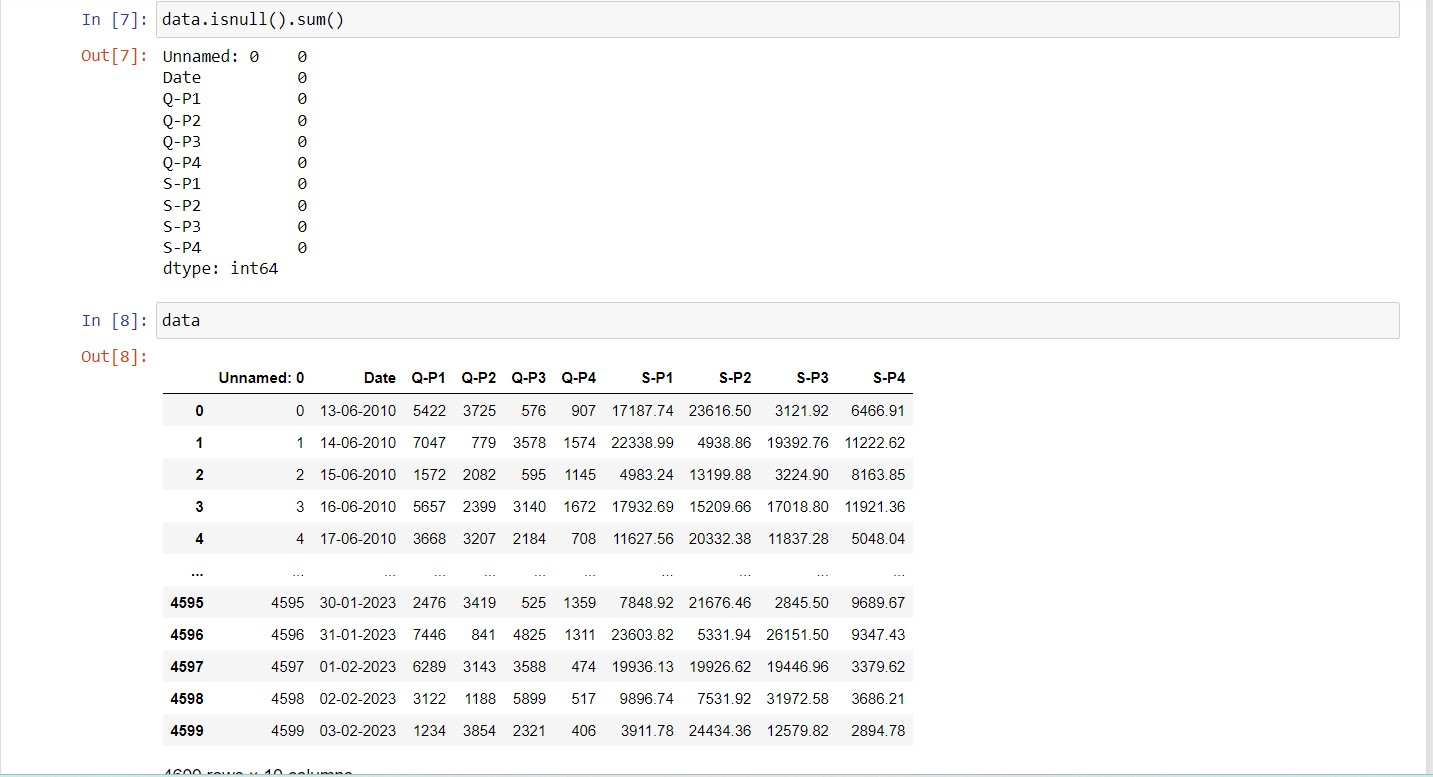
data.info()



**Fig 5.2.2: Data Cleaning using data.describe() and data.info()**

data.isnull().sum()

data



**Fig 5.2.3: Data Cleaning using data.isnull()sum() in order to check null values**

data.dropna(how="any").shape

data.dropna(how="all").shape

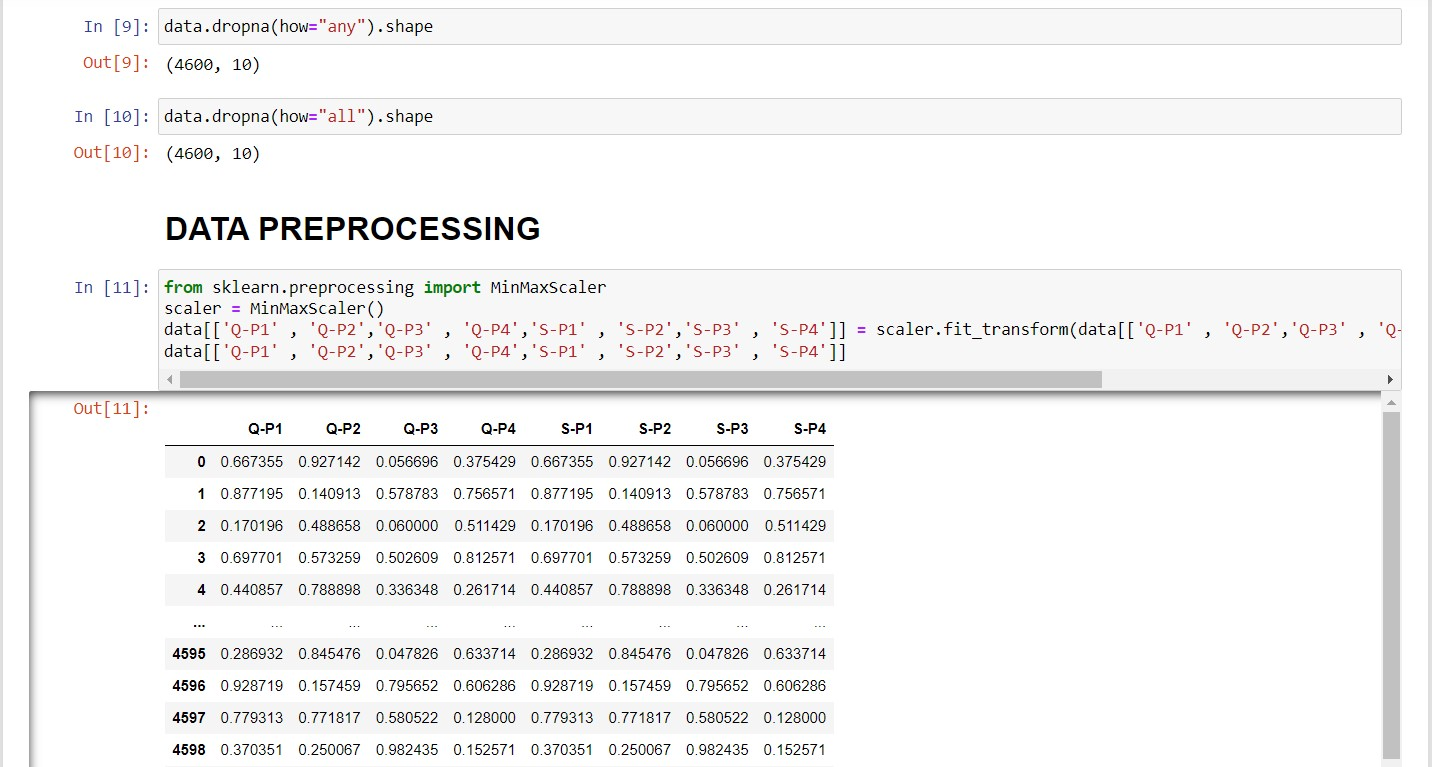
**Preprocessing of data**

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

data[['Q-P1' , 'Q-P2','Q-P3' , 'Q-P4','S-P1' , 'S-P2','S-P3' , 'S-P4']] = scaler.fit\_transform(data[['Q-P1' , 'Q-P2','Q-P3' , 'Q-P4','S-P1' , 'S-P2','S-P3' , 'S-P4']])

data[['Q-P1' , 'Q-P2','Q-P3' , 'Q-P4','S-P1' , 'S-P2','S-P3' , 'S-P4']]



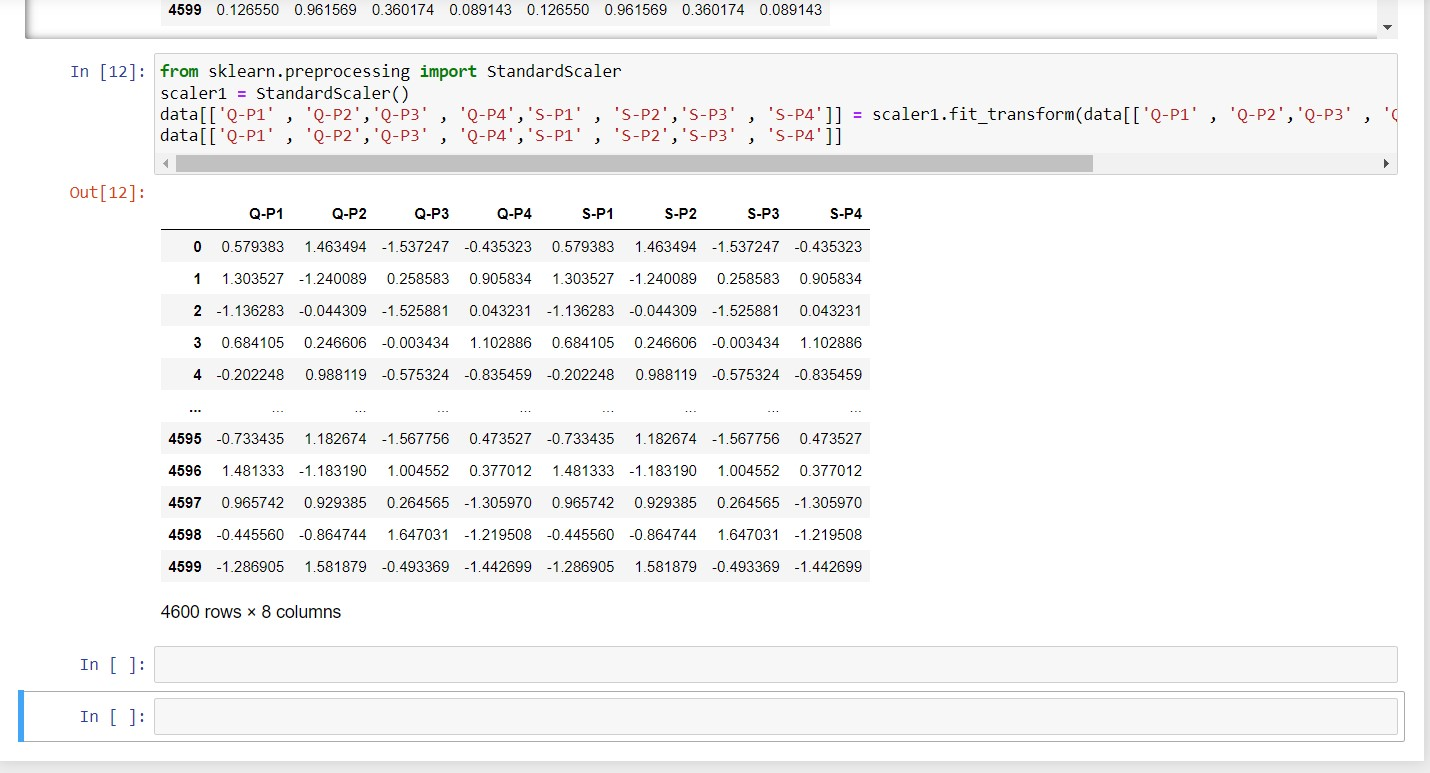
**Fig 5.2.4: Data Cleaning using data.dropna() and preprocessing the data by importing MinMaxScaler**

from sklearn.preprocessing import StandardScaler

scaler1 = StandardScaler()

data[['Q-P1' , 'Q-P2','Q-P3' , 'Q-P4','S-P1' , 'S-P2','S-P3' , 'S-P4']] = scaler1.fit\_transform(data[['Q-P1' , 'Q-P2','Q-P3' , 'Q-P4','S-P1' , 'S-P2','S-P3' , 'S-P4']])

data[['Q-P1' , 'Q-P2','Q-P3' , 'Q-P4','S-P1' , 'S-P2','S-P3' , 'S-P4']]



**Fig 5.2.5: Preprocessing the data by importing StandardScaler**

**PHASE – 4**

**VISUALIZATION OF DATA:**

**Calculating the average of unit sales and revenue of products:**

Q-P1: Total unit sales of Product 1

Q-P2: Total unit sales of Product 2

Q-P3: Total unit sales of Product 3

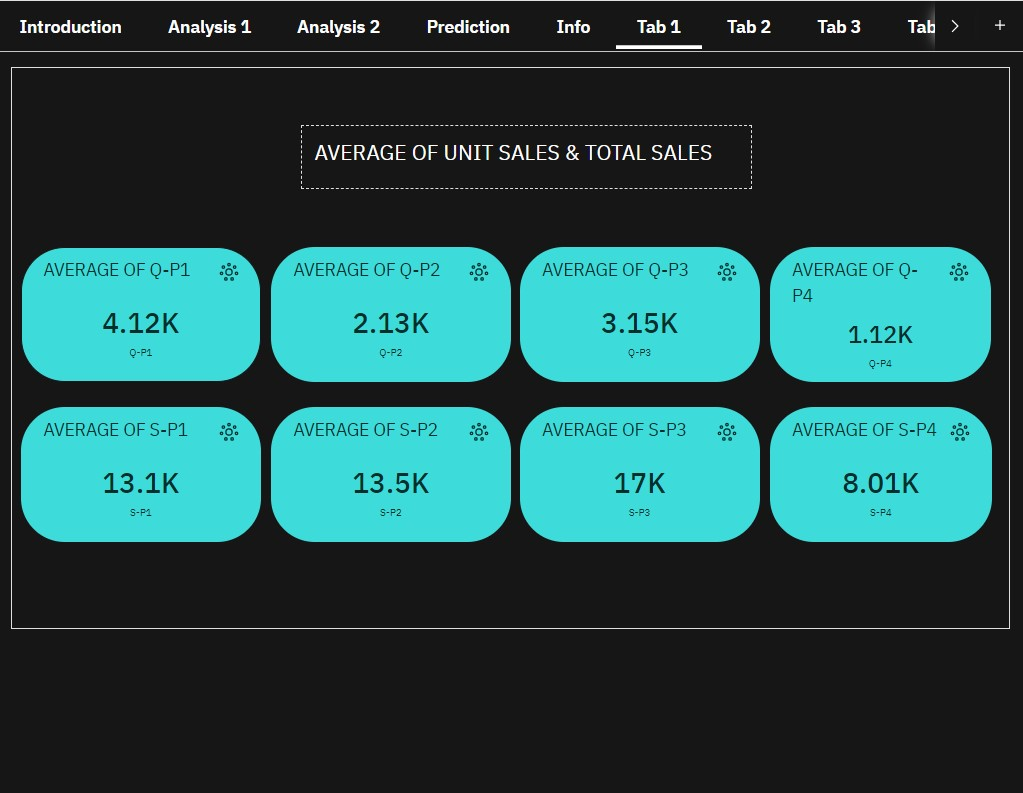
Q-P4: Total unit sales of Product 4

S-P1: Total revenue from Product 1

S-P2: Total revenue from Product 2

S-P3: Total revenue from Product 3

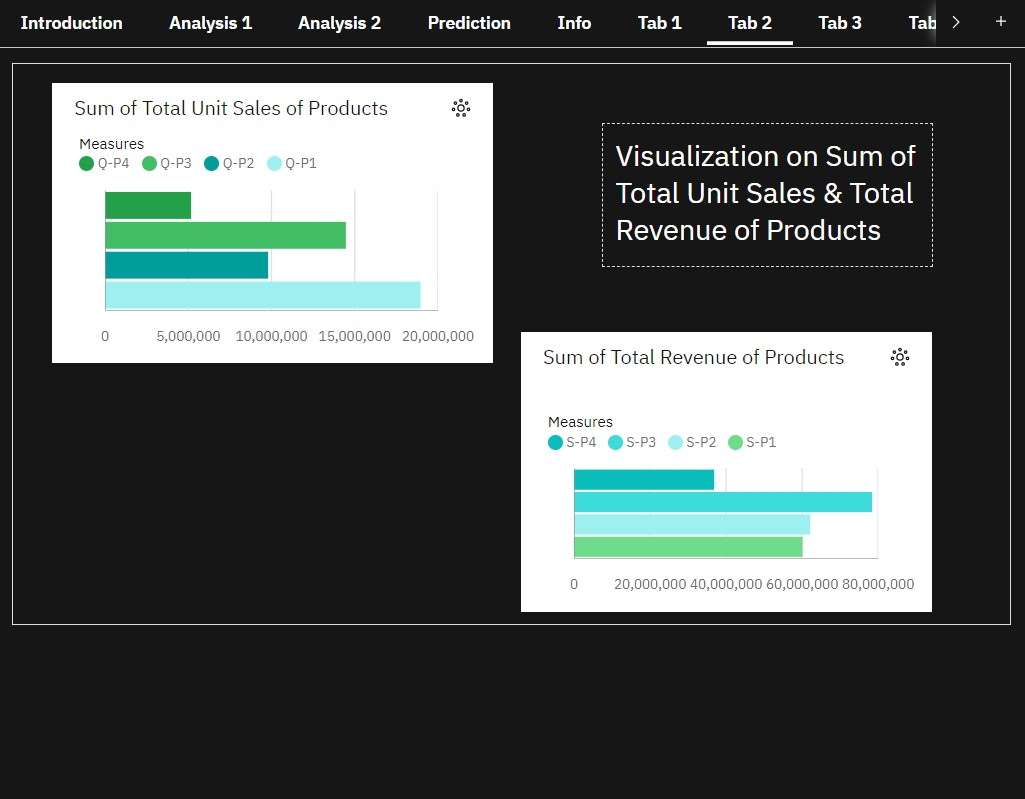
S-P4: Total revenue from Product 4



**Fig 5.3.1: Calculating the average of unit sales and total sales(revenue)**

**Visualization on sum of total unit sales & total revenue of products:**

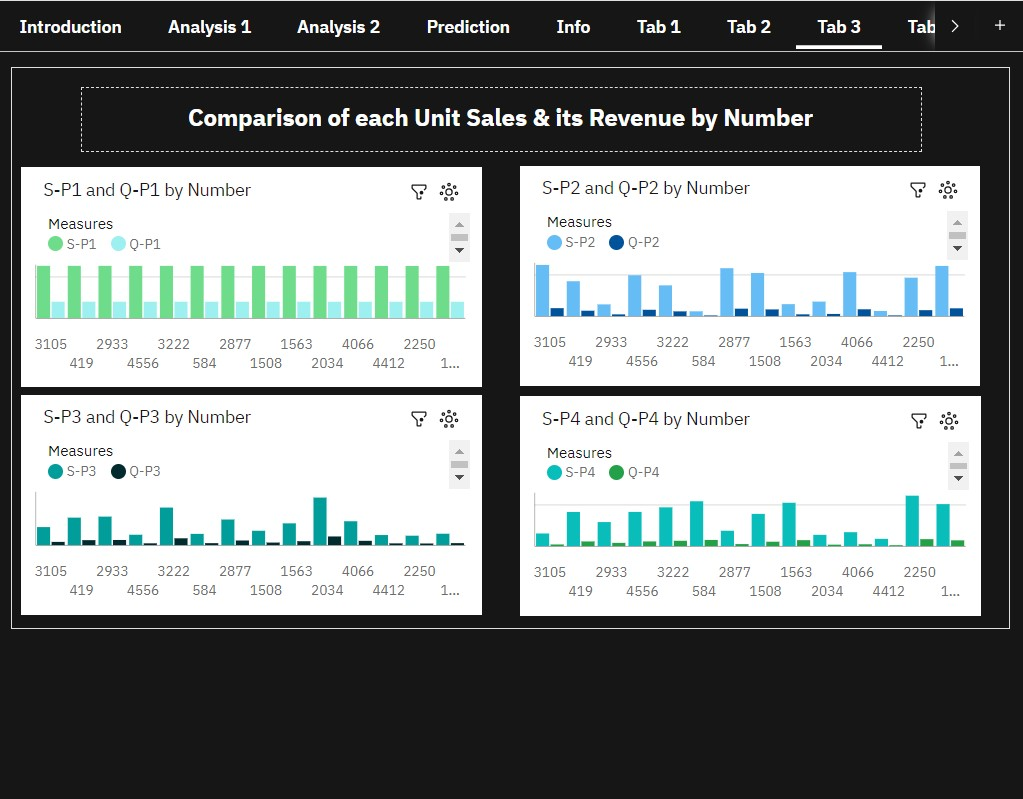
There are four products(P1, P2, P3, P4) sold by REC Corp Ltd. So it is necessary to visualize the range of products according to Unit sales and its revenue.



**Fig 5.3.2: Visualization on sum of total unit sales and revenue of products**

**Comparing each unit sales & its revenue:**

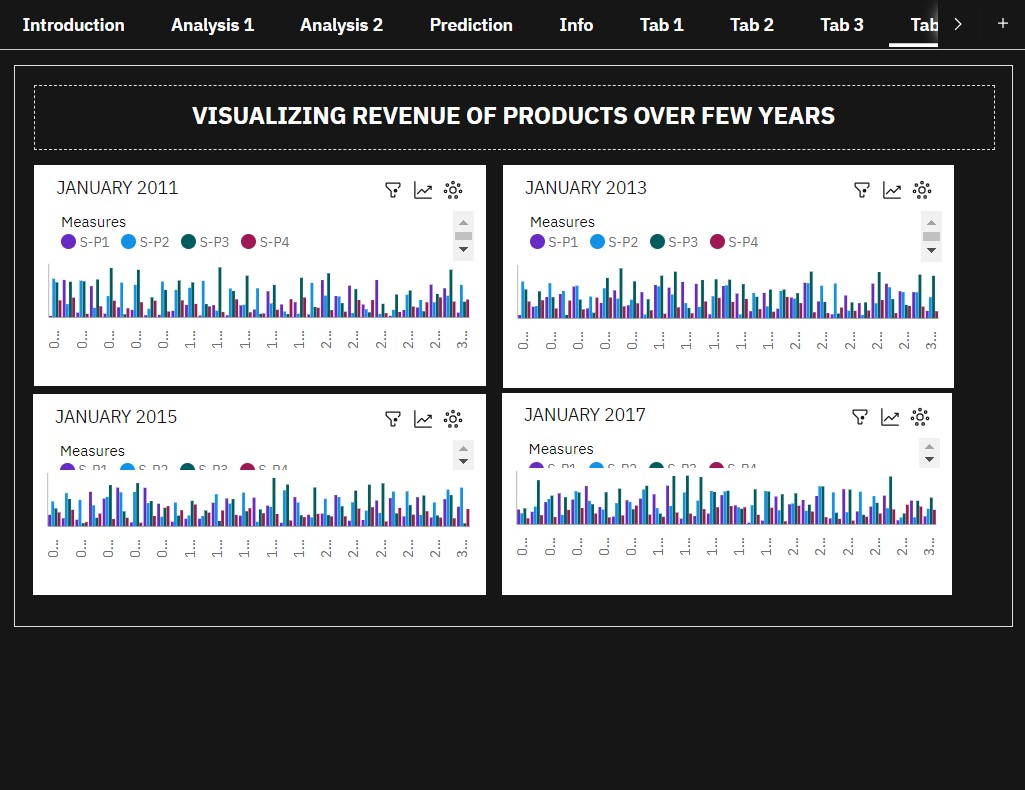
It is necessary to compare the total unit sales and total revenue of each product.



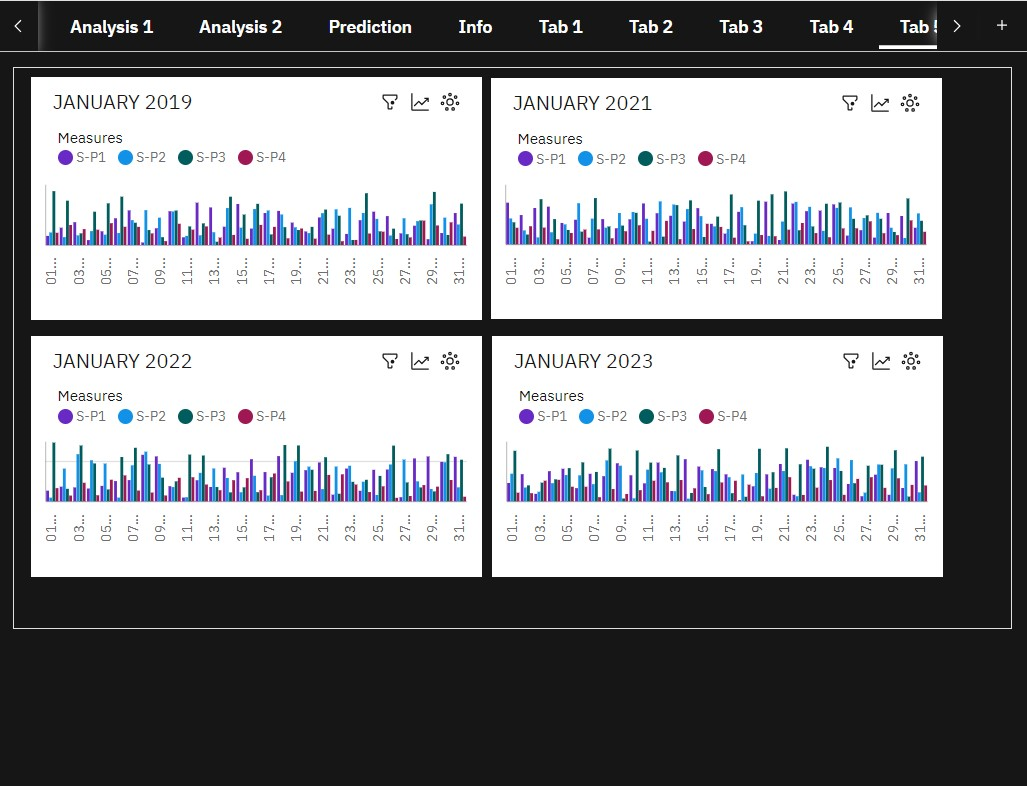
**Fig 5.3.3: Comparison of each unit sales and its revenue by number**

**Visualizing revenue of products over few years:**

The unit sales and its revenue changes each and every year. Visualizing revenue of products happened over years helps in predicting future sales.



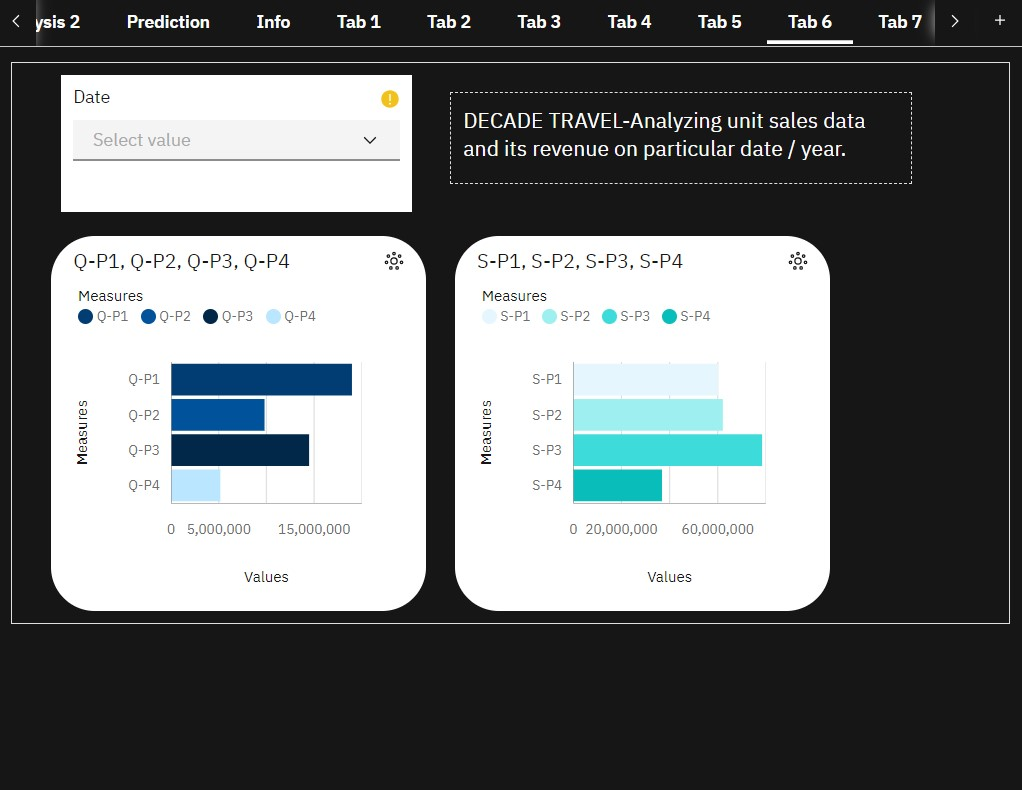
**Fig 5.3.4.1: Visualizing revenue of products over few years**



**Fig 5.3.4.2: Visualizing revenue of products over few years**

**DECADE TRAVEL:**

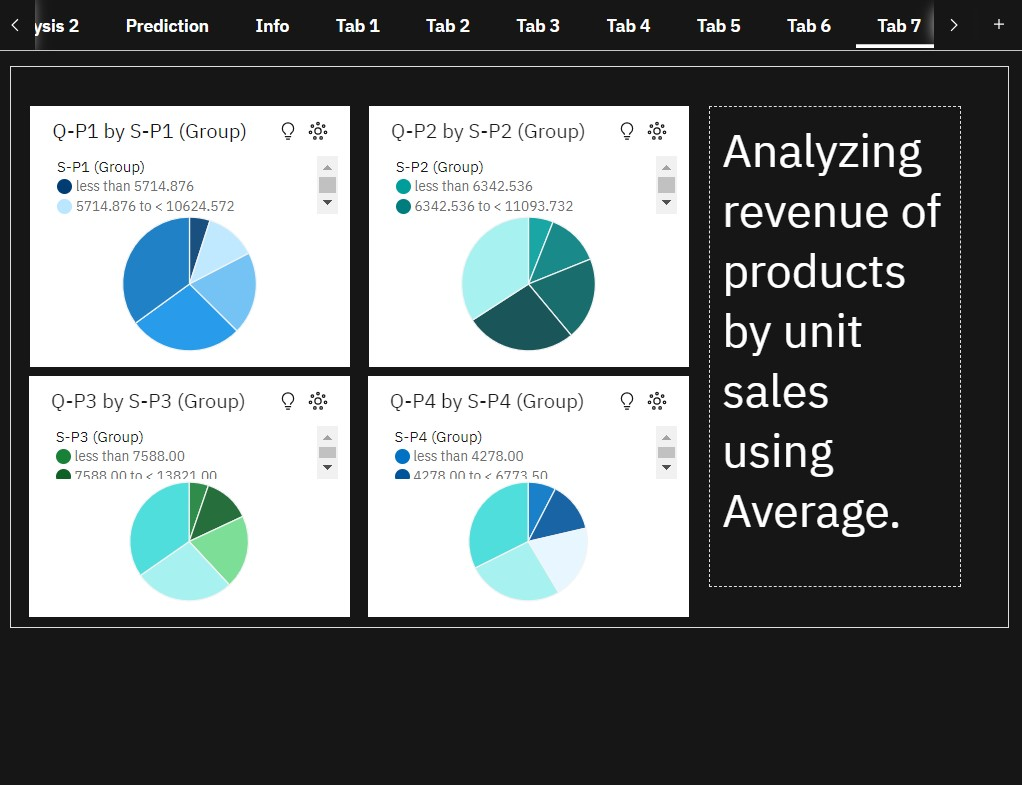
Since it is very difficult to analyse each data, it is good to have a drop down list that contain date and year. Using this date and year it is easy to analyse unit sales and its revenue over decades.



**Fig 5.3.5: Decade Travel**

**Analysing revenue of products by unit sales using average:**

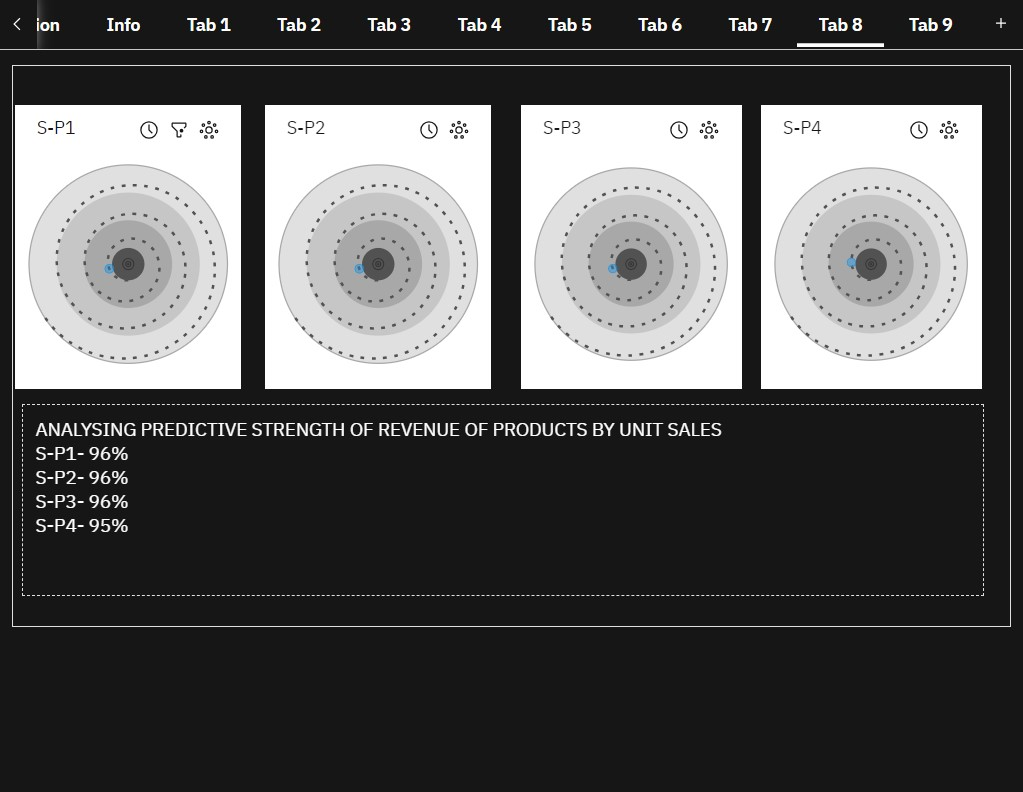
Revenue of products is analysed by using the average calculated.



**Fig 5.3.6: Analyzing revenue of products by unit sales using Average.**

**Predicting the strength of revenue of products:**

In order the predict the sales in future, it is necessary to predict the revenue status happened over decades.



**Fig 5.3.7: Predicting the strength of revenue of products**

According to this analysis

Revenue of P1, P2, P3 and P4 are almost near to 100%, which means the revenue of P1, P2, P3 and P4 will be higher in future.

(i.e) Total Revenue of P1 – 96%

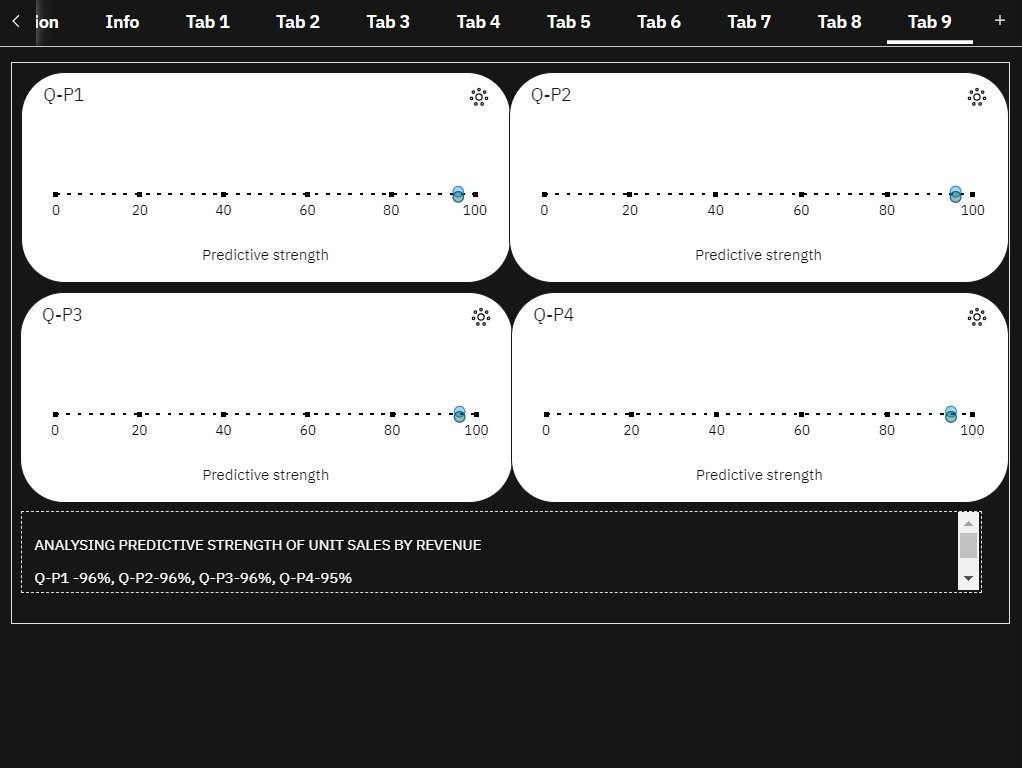
Total Revenue of P2 – 96%

Total Revenue of P3 – 96%

Total Revenue of P4 – 95%

**Predicting the unit sales of products**

Similarly it is necessary to predict the status of unit sales of products over decades.



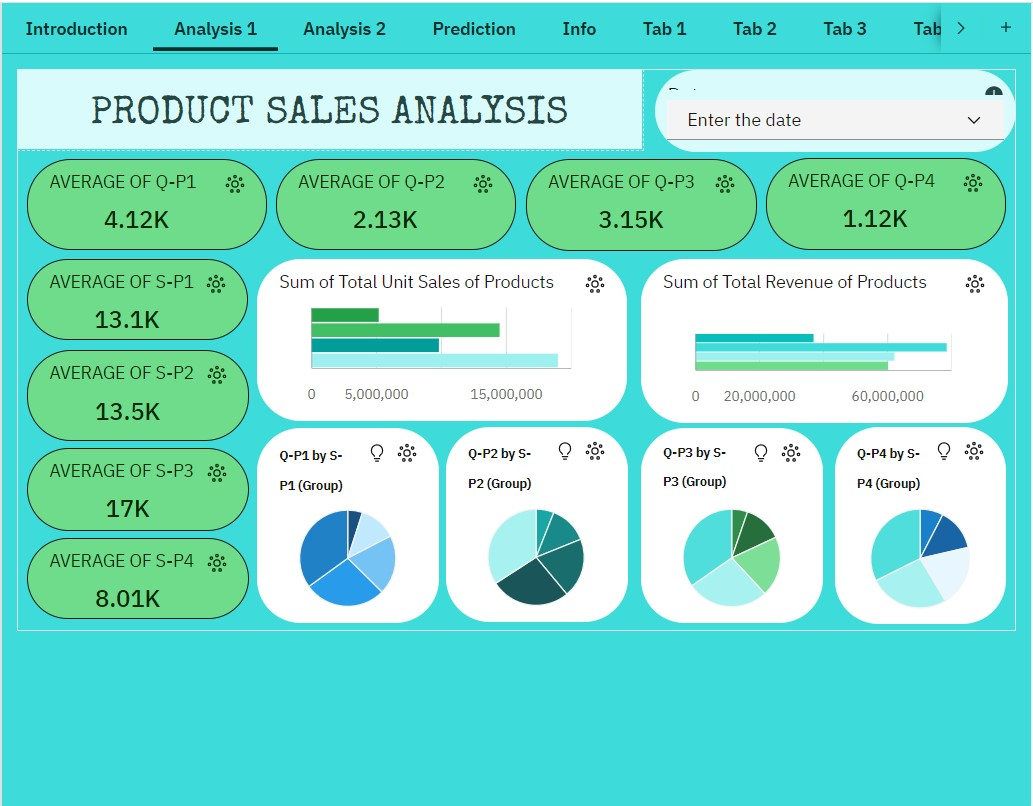
**Fig 5.3.8: Predicting the unit sales of products**

According to this analysis

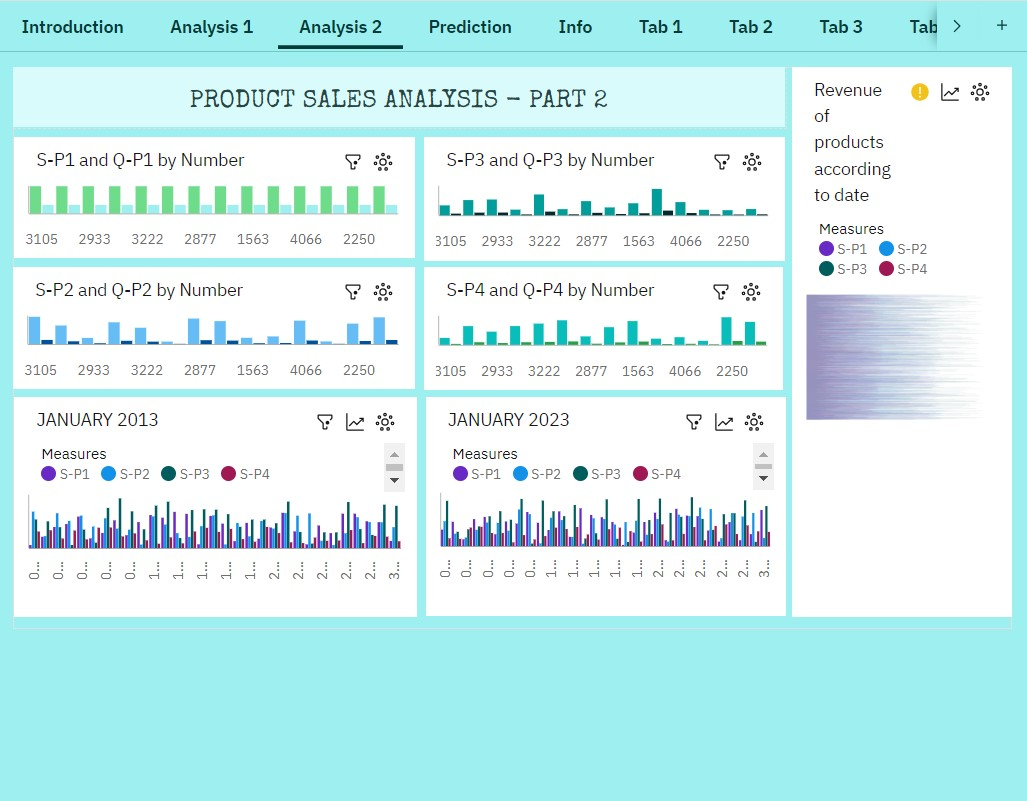
Unit sales of P1, P2, P3 and P4 are almost near to 100%,

(i.e)Unit Sales of P1 – 96%, P2 – 96%, P3 – 96%, P4 – 95%, which means the unit sales of P1, P2, P3 and P4 will be higher in future.

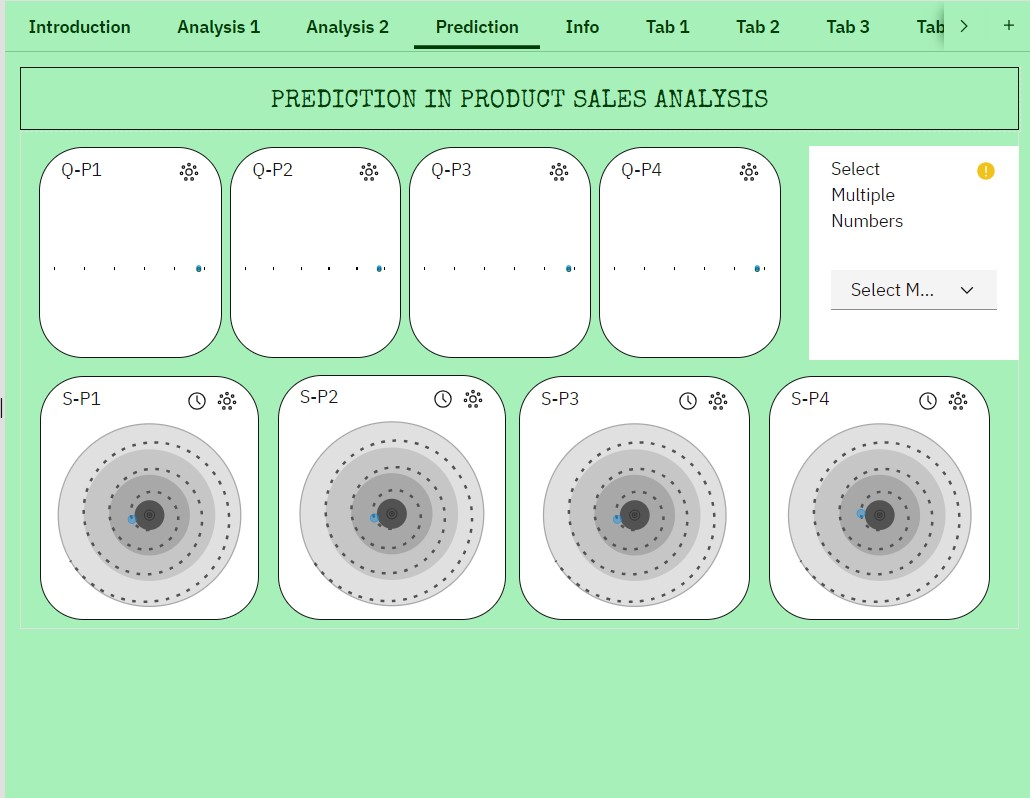
**FINAL DASHBOARD**



**Fig 5.4.1: Dashboard Part 1**



**Fig 5.4.2: Dashboard Part 2**



**Fig 5.4.3: Dashboard Part 3-Prediction**

**CONCLUSION:**

REC Corp Ltd has been provided the dataset regarding their unit sales and the revenue of each product. REC Corp Ltd has been selling four products and the data of those selling products from 2011 to 2023 has been added in the dataset. The product sales analysis helps in predicting future sales and customer preferences of the product sold by REC Corp Ltd. According to the analysis, the revenue of each product was higher and there were some variation within decades but did not affect their sales. After cleaning and preprocessing, our team received the perfect data which were helpful in visualizing the future sales of the products sold.

According to the analysis after visualization, the predictive analysis gave the accuracy of 96% and from this result it is clear that the **future sales of products will be higher and customer preferences over the products sold by REC Corp Ltd will also be higher.**