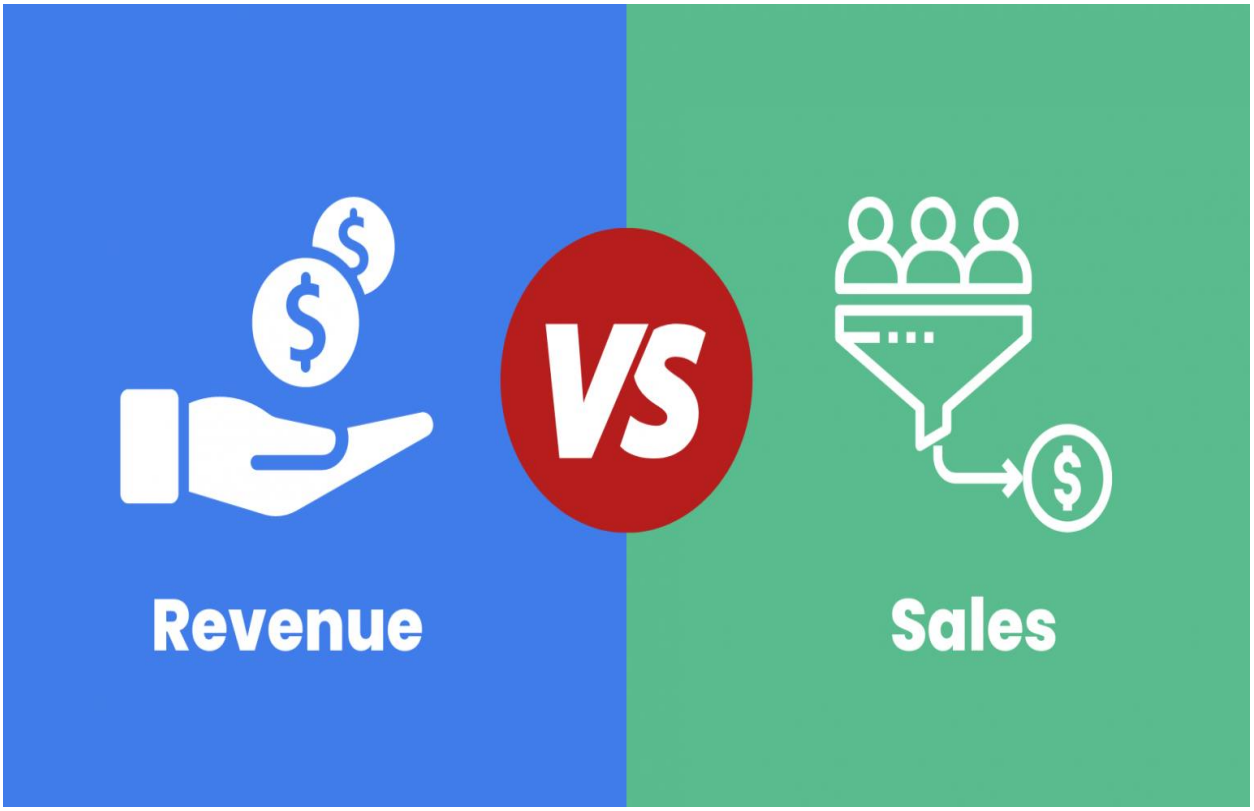
1. Project Proposal:

Predicting the Sales and Revenues of four products of the REC Corp LTD company for 2024



Source:
https://www.kaggle.com/datasets/ksabishek/product-sales-data/data

Predicting the sales and revenues for the year 2024

REC Corp LTD. is small-scaled business venture established in India. They have been selling FOUR PRODUCTS for OVER TEN YEARS.
The products are P1, P2, P3 and P4. They have collected data from their retail centers and organized it into a csv file, which has been given to you. **The excel file contains about 8 numerical parameters: **
Q1- Total unit sales of product 1
Q2- Total unit sales of product 2
Q3- Total unit sales of product 3
Q4- Total unit sales of product 4
S1- Total revenue from product 1
S2- Total revenue from product 2
S3- Total revenue from product 3
S4- Total revenue from product 4
Example: On 13-06-2010, product 1 had been brought by 5422 people and INR 17187.74 had been generated in revenue from product 1.
**Now, REC Corp needs you to solve the following questions: **
1) Is there any trend in the sales of all four products during certain months?
2) Out of all four products, which product has seen the highest sales in all the given years?
3) The CEO is considering an idea to drop the production of any one of the products. He wants you to analyze this data and

suggest whether his idea would result in a massive setback for the company.
4) The CEO would also like to predict the sales and revenues for the year 2024. He wants you to give a yearly estimate with the best possible accuracy.
Can you help REC Corp with your analytical and data science skills?

I will answer first three questions in Data Wrangling and EDA parts.

The most important question is last one. I will try to create a good model that it can perfectly predict sales and revenue in 2024.

I will try to use AR| ARMA| ARIMA| SARIMA| SARIMAX concepts to predict data.
Data wrangling and EDA are the parts that I can get insights and watch their patterns that I can use later.
Preprocessing and Modeling are the most important part that I have to create model and I will try to draw the best prediction, when I can compare actual data with my prediction.
In modeling part, I select the best P, Q, R that it has the lowest AIC and after that I will calculate RMSE. Then I will compare them, I select the specific P, Q, R. Finally, I will draw the conclusions that I can compare them correctly.

2. Data Wrangling and Exploratory Data Analysis

2.1 Contents

2.5 Data Wrangling

In this part, I try to clean data and prepare it for EDA (Exploratory Data Analysis). I have 9 columns that are date, QP1, QP2, QP3, QP4, SP1, SP2, SP3, and SP4. The types of data are normal, but date column is object. I will try to convert it to datetime. I will put date time as index later.

After converting the date column, I found out I have 26 NA values. I delete them.

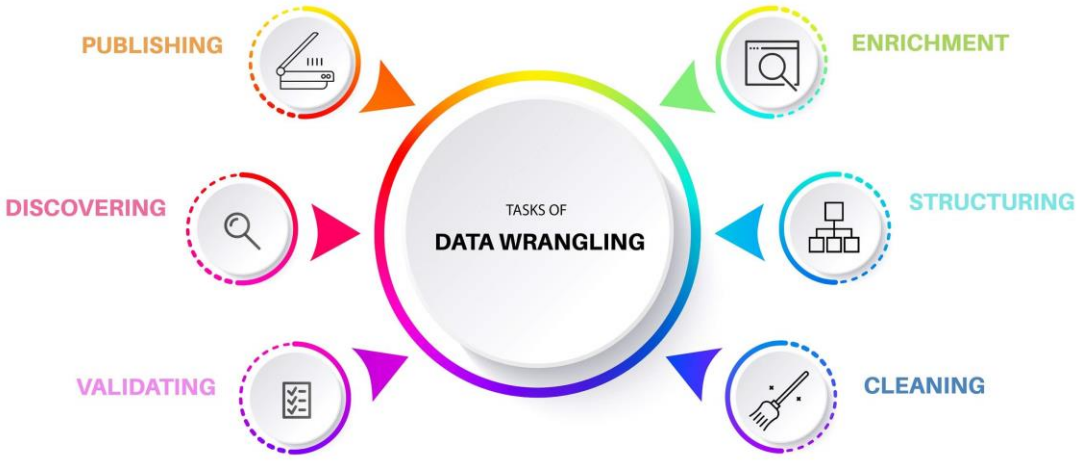
I fortunately don't have null values in our data. The date column has limitations values that are different in columns.

Our dates are between 2010-07 to 2023-12. When I compare all data, I will figure out the best data that I can use are between from 2011 to end of 2022. So, I can create a model that can predict the data in 2023 and 2024. I will explain it later.

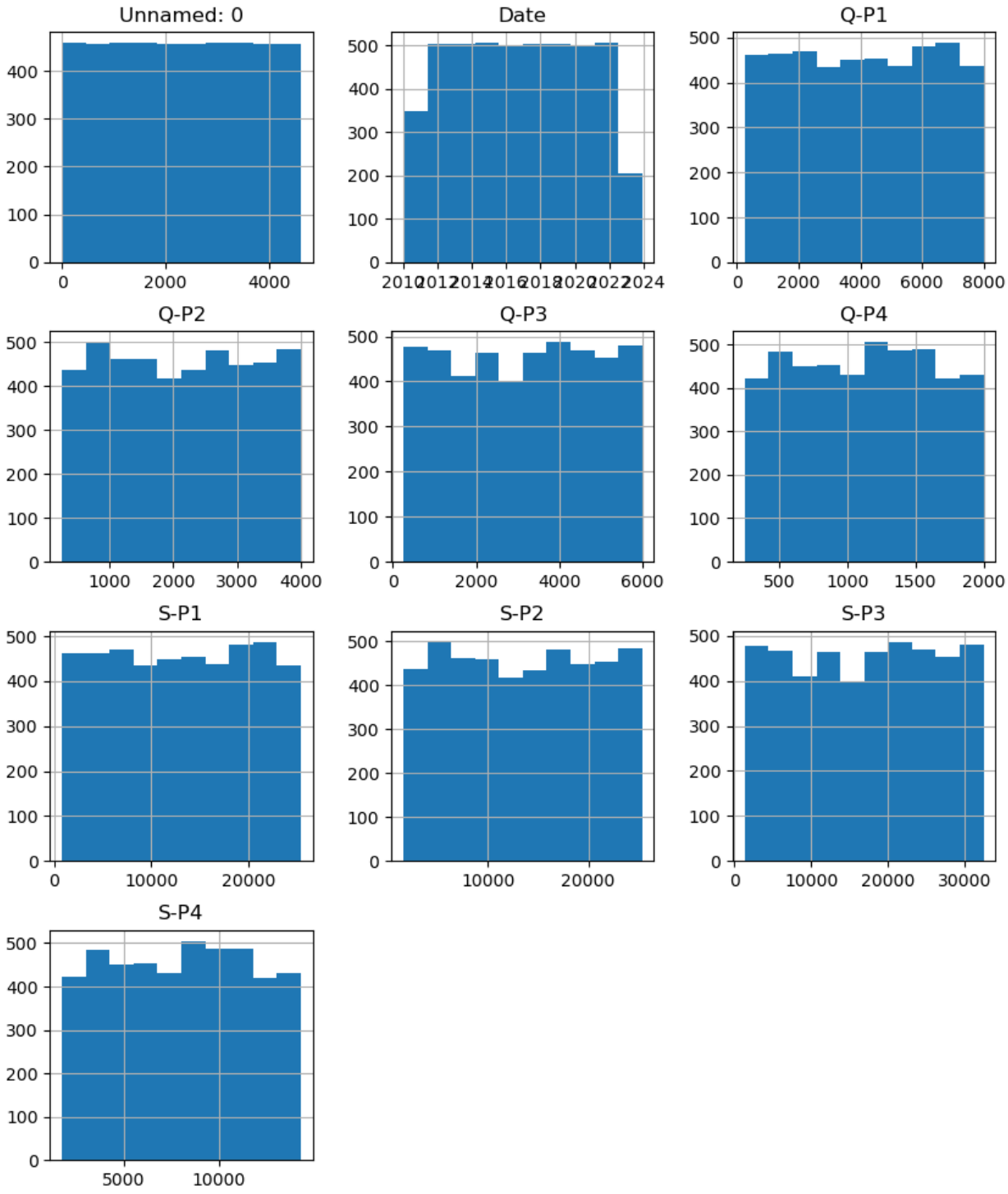
The dates start from first day of month that it is easy to find out the trends. Our trends are continuous, when I want to try time series.

The thing that I should notice is that changing is trendy and seasonal.

As we can see in next page, the distributions of our data aren't too specific. In the EDA part I will show the trend of data.



I tried to find out the distribution of our columns.



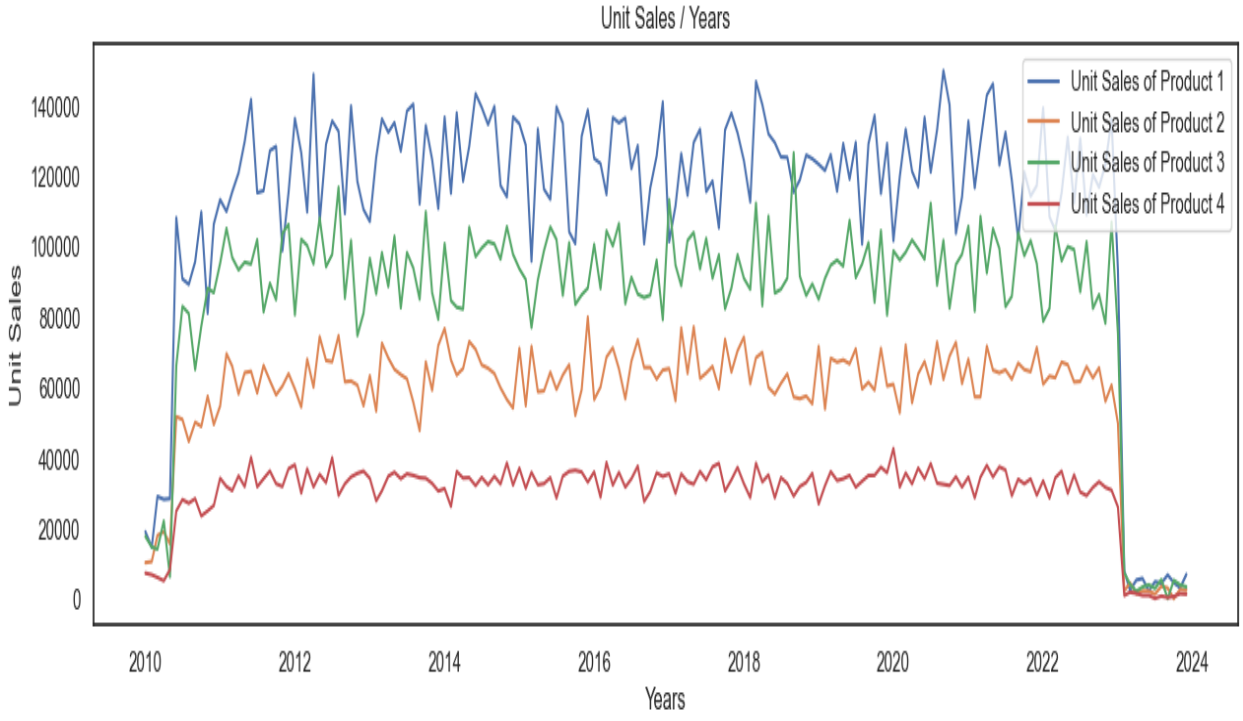
2.6 Exploratory Data Analysis (EDA)

We observe both trend and multiplicative seasonality from the plot shown below.

I create this plot bellow that is for first 4 columns that they are related to sales product 1, 2, 3, and 4.

Here I decided to remove the data in 2010 and 2024, because they are incomplete. They are incomplete in revenues columns too in next page for revenue product 1, 2, 3, and 4.

I'm going to show the trends and changes of unit sales of product 1, 2, 3, and 4. I find out Product 1 has the highest unit sales. Product 4 has the lowest unit sales. As we can see, there are trends and seasonal changes.

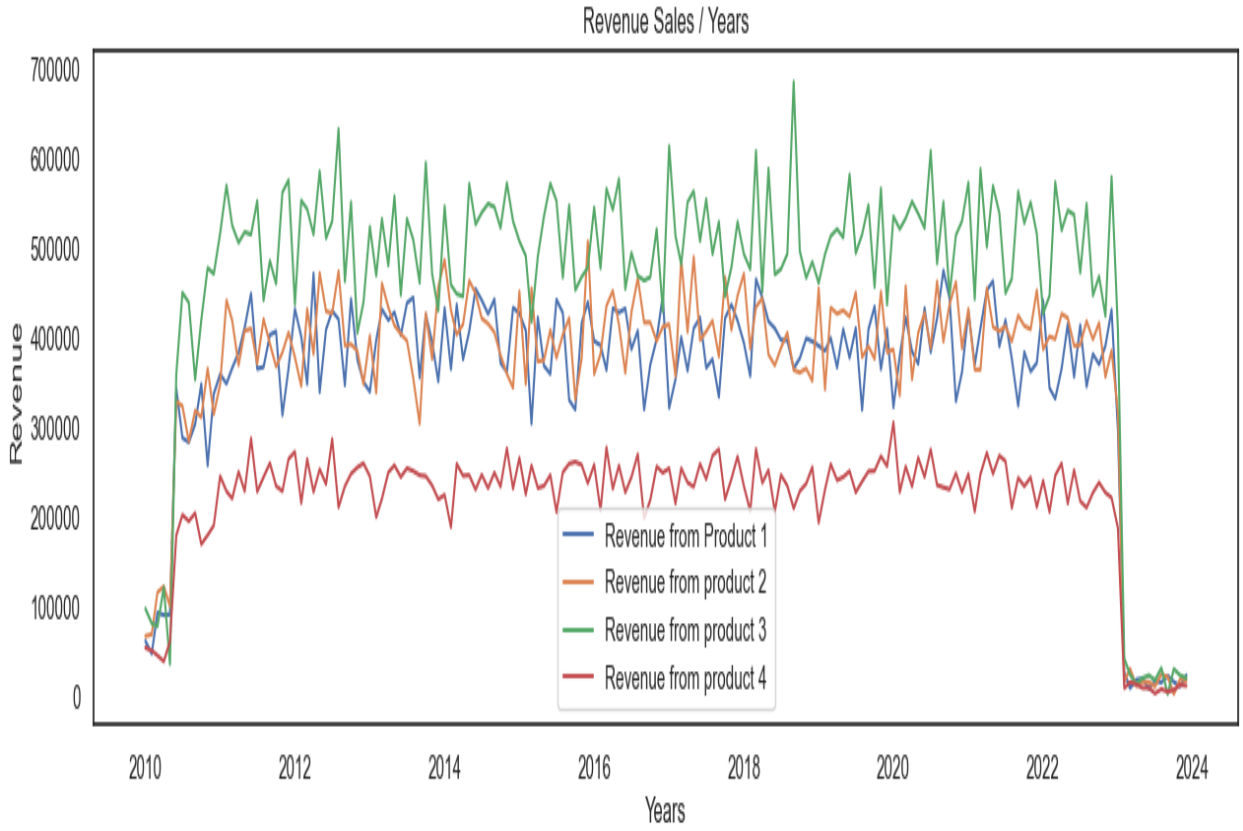


The chart below shows us the revenue of product 1, 2, 3, and 4. The chart tells that product 1 has the highest revenue and product 4 has the lowest revenue.

This chart is different in comparing to previous page. The revenues of product 2 and 3 are changing the same, but the sales of these products are not changing the same.

As we can see, there are seasonally changes in both charts.

As it's obvious, 2010th and 2023 don't have the complete data. So, in the next part, I will remove them.



2.6.1 How to check for Stationarity?

In this part, I try to give the good insight of stationarity.

Dickey-Fuller Test - Dicky Fuller Test on the timeseries is run to check for stationarity of data.

- Null Hypothesis H0: Time Series is non-stationary.
- Alternate Hypothesis Ha: Time Series is stationary. So ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected.

p-value for Time Series 1: 0.30
p-value for Time Series 2: 0.50
p-value for Time Series 3: 0.21
p-value for Time Series 4: 0.33
p-value for Time Series 5: 0.30
p-value for Time Series 6: 0.50
p-value for Time Series 7: 0.21
p-value for Time Series 8: 0.33

As we can see, all p-values are greater than 0.05. It means that they are not stationary series.

Performing the decomposition of data if there is an existence of seasonality and split the data accordingly.

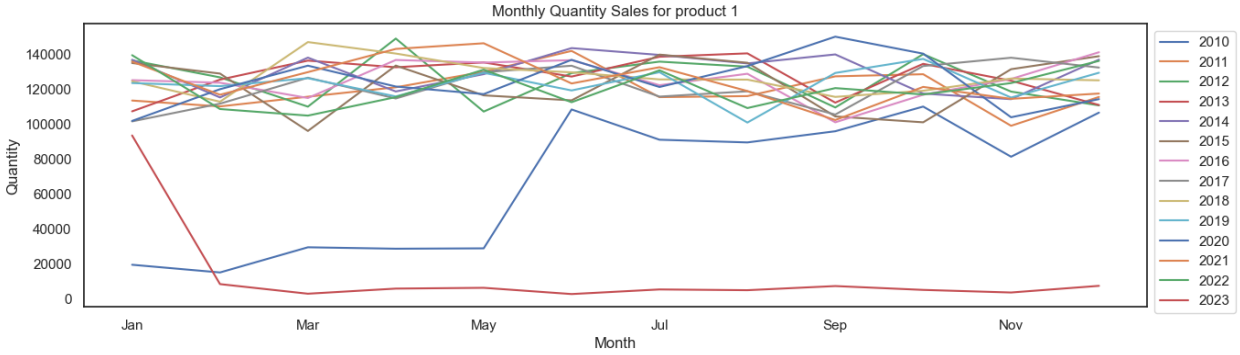
Observe how number of sales and revenues vary on a month-on-month basis. A stacked plot for every year will give us a clear pattern of any seasonality over the many years and those changes will be clearly reflected in the plots.

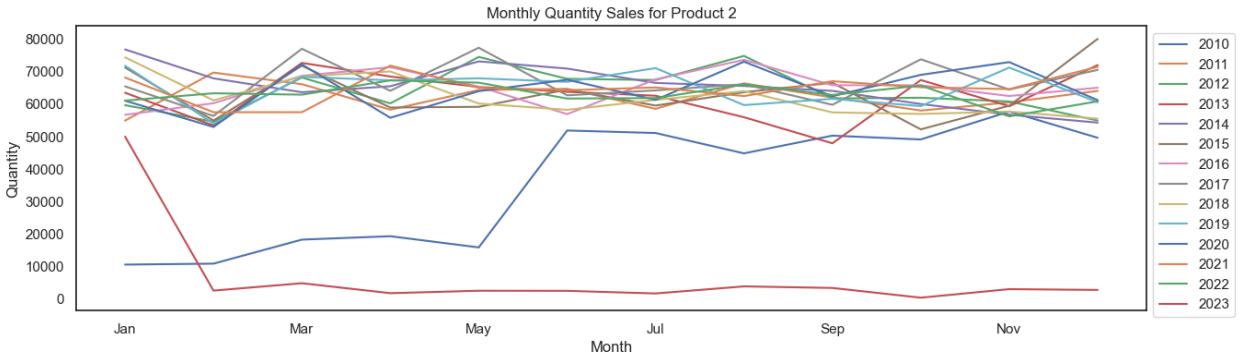
2.6.2 Categorizing Monthly and Yearly

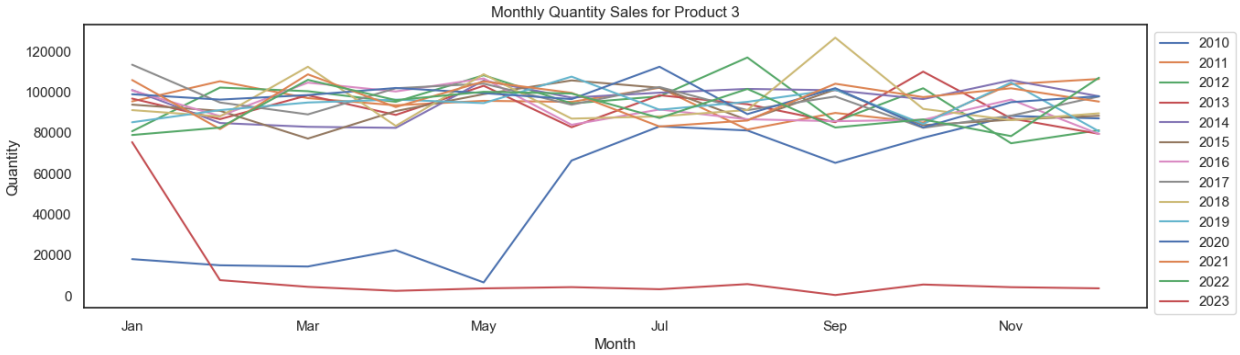
a) Categorizing Monthly

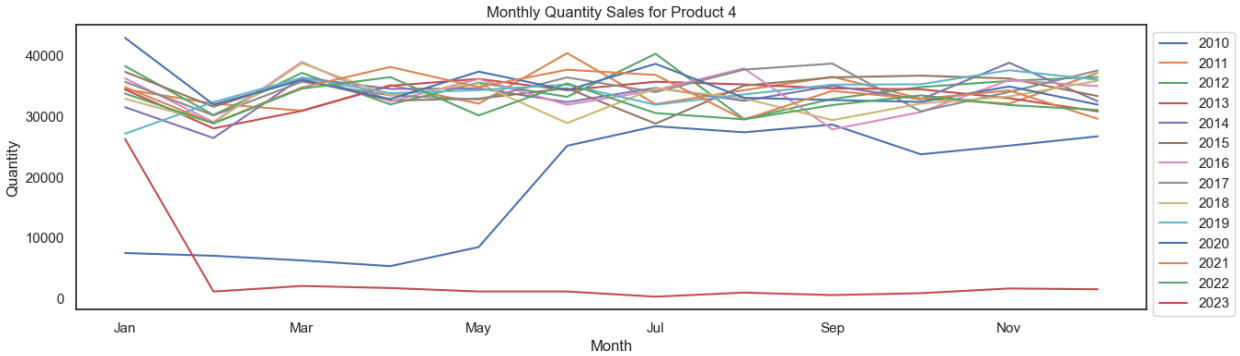
We can see the results of quantity sales for product 1, 2, 3, and 4.

As I noticed before, 2010 and 2023 have the incomplete data.

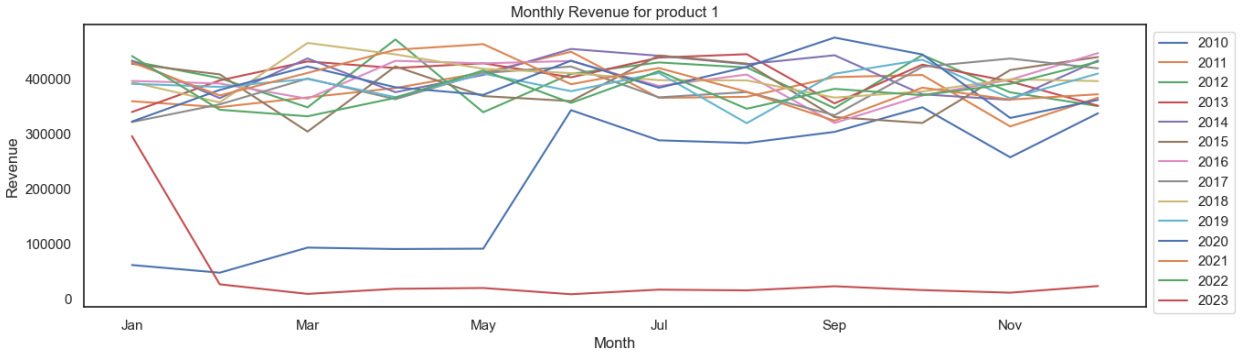


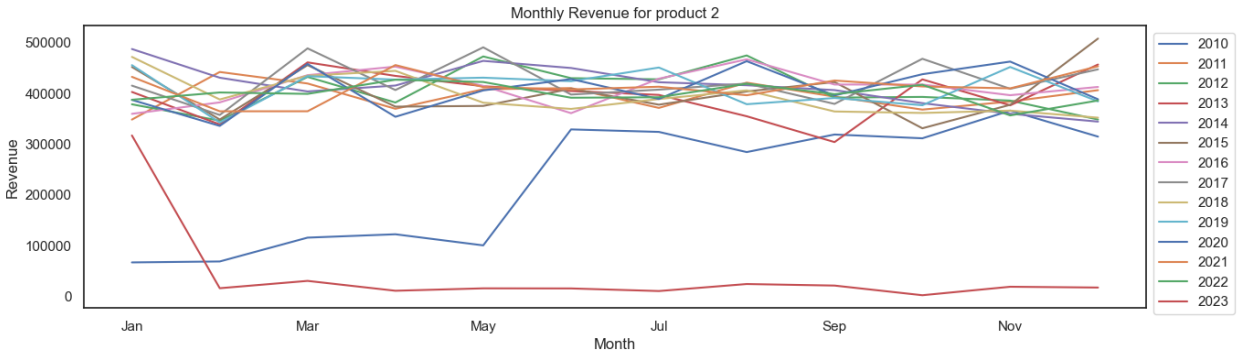


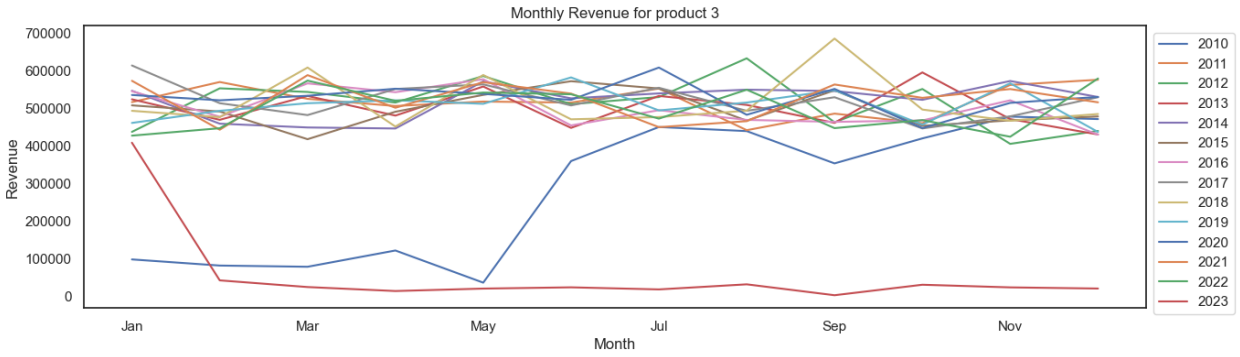


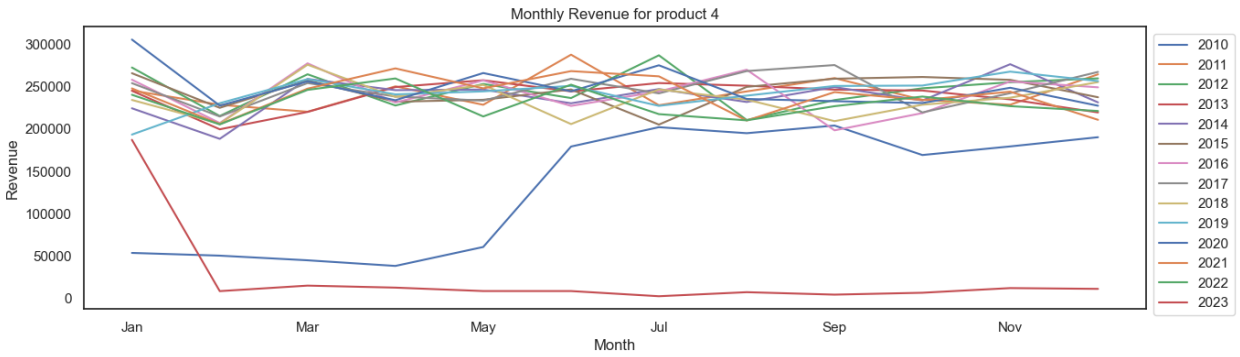


We can see the results of revenue sales for product 1, 2, 3, and 4.
As I noticed before, 2010 and 2023 have the incomplete data.



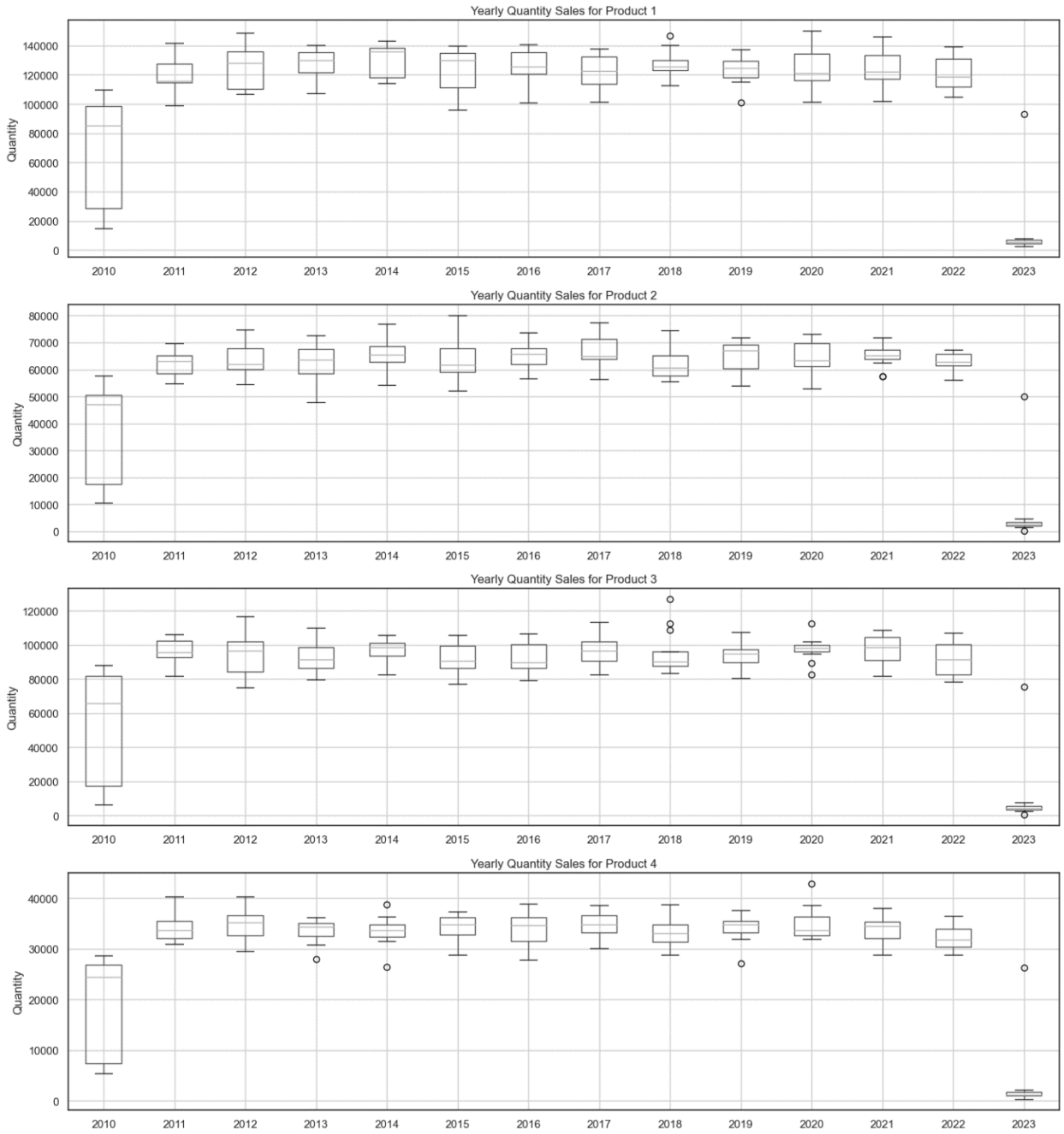




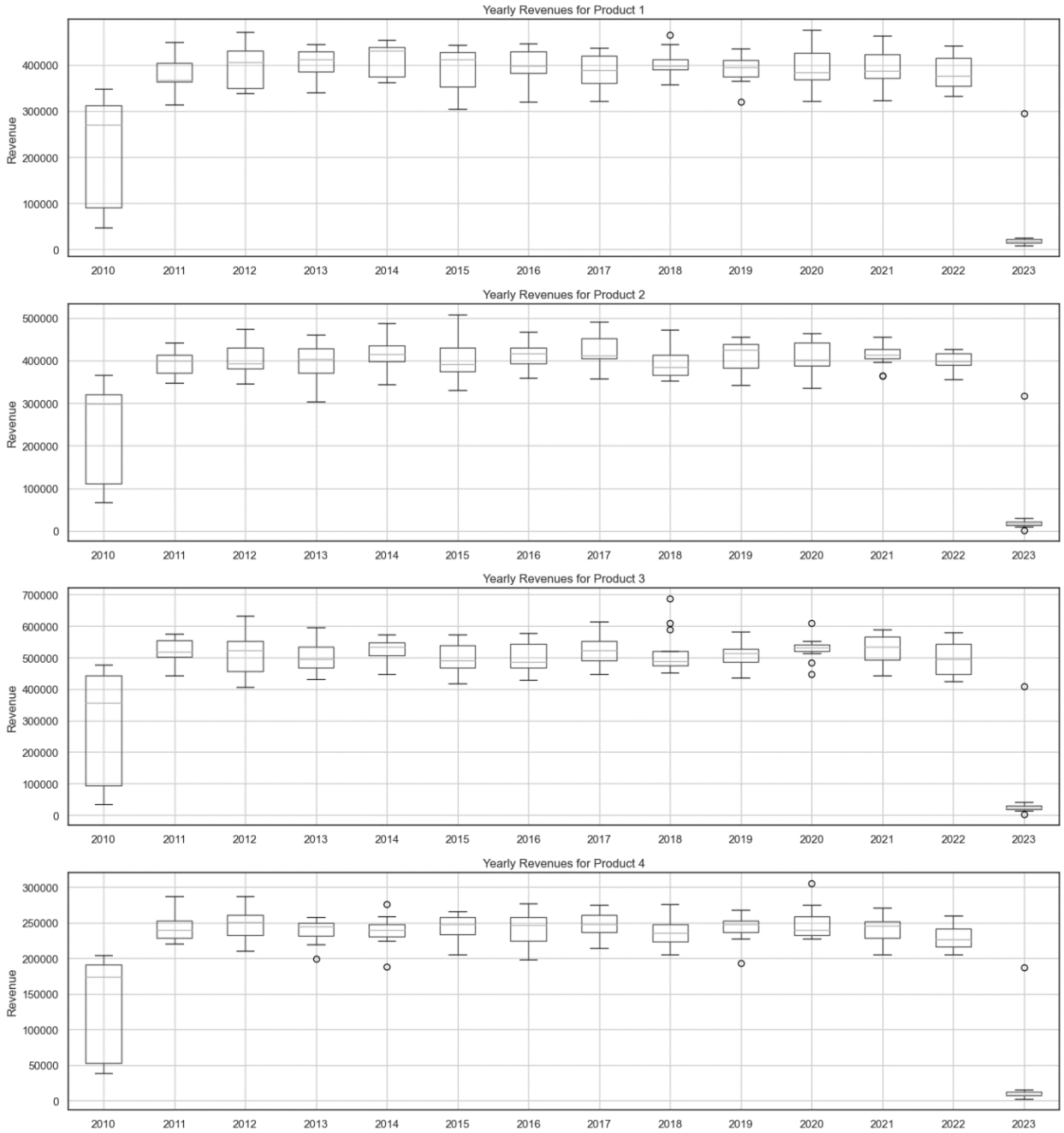


b) Categorizing Yearly

We can see the results of revenue sales for product 1, 2, 3, and 4.
As I noticed before, 2010 and 2023 have the incomplete data.



We can see the results of revenue sales for product 1, 2, 3, and 4.
As I noticed before, 2010 and 2023 have the incomplete data.



2.7 Summary:

Here is the question that CEO asked before:

**Now, REC Corp needs you to solve the following questions: **

1. Is there any trend in the sales of all four products during certain months?
2. Out of all four products, which product has seen the highest sales in all the given years?
3. The CEO is considering an idea to drop the production of any one of the products. He wants you to analyze this data and suggest whether his idea would result in a massive setback for the company.
4. The CEO would also like to predict the sales and revenues for the year 2024. He wants you to give a yearly estimate with the best possible accuracy.

Answers:

1. There is not any special trend in our data except that in 2010 and 2023 we have some specific trend for quantity and revenue sales. The reason is that our data in 2010 and 2023 are incomplete. I will remove them later.
2. Product 1 has the highest quantity and revenue sales. Product 4 has the lowest quantity and revenue sales. product 2 has greater quantity sales in comparing to product 3, however their revenues sales trend and amount are approximately equal.
3. If CEO decided to drop one product, I would suggest to drop product 4. because product 4 has the lowest quantity and revenue sales.
4. There is no answer in Data Wrangling and EDA. I will answer this question later. It needs to create a model for this question.

Other Conclusions:

- Product 1 has the highest unit sales and revenue between 2011 to 2022. (it's the answer of second question)
- Product 4 has the lowest unit sales and revenue between 2011 to 2022.
- product 2 has the higher unit sales than product 3, however their revenues are almost equal.
- their linear trends seasonally are changing.

3 Preprocessing and Modeling¶

3.1 Contents

3.2 Introduction

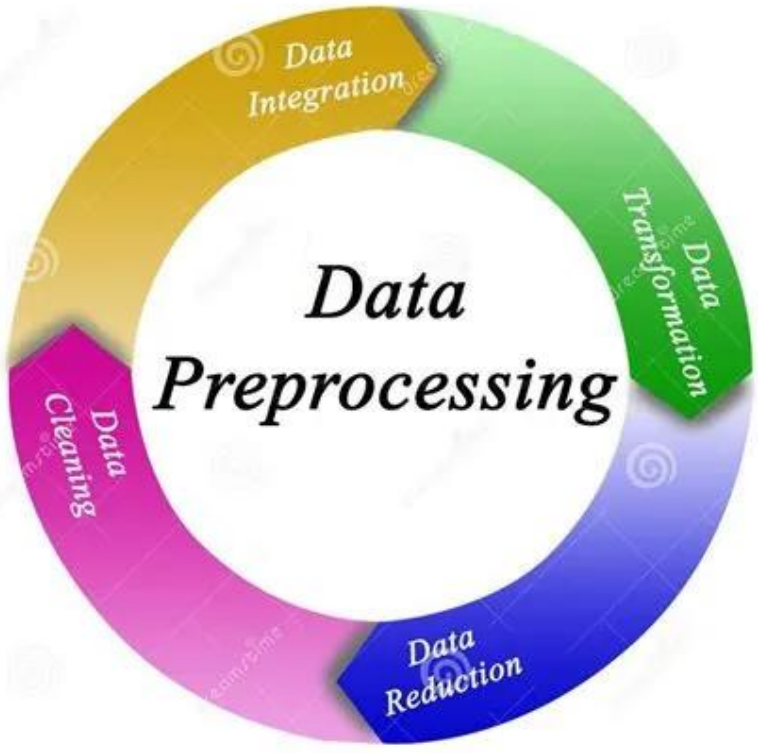
There is a request that I have to answer it. This request needs to predict the data, so I have to create a model that can predict it.

The request is below (4th question):

4th) The CEO would also like to predict the sales and revenues for the year 2023 and 2024. He wants you to give a yearly estimate with the best possible accuracy.

3.5 Select the Data

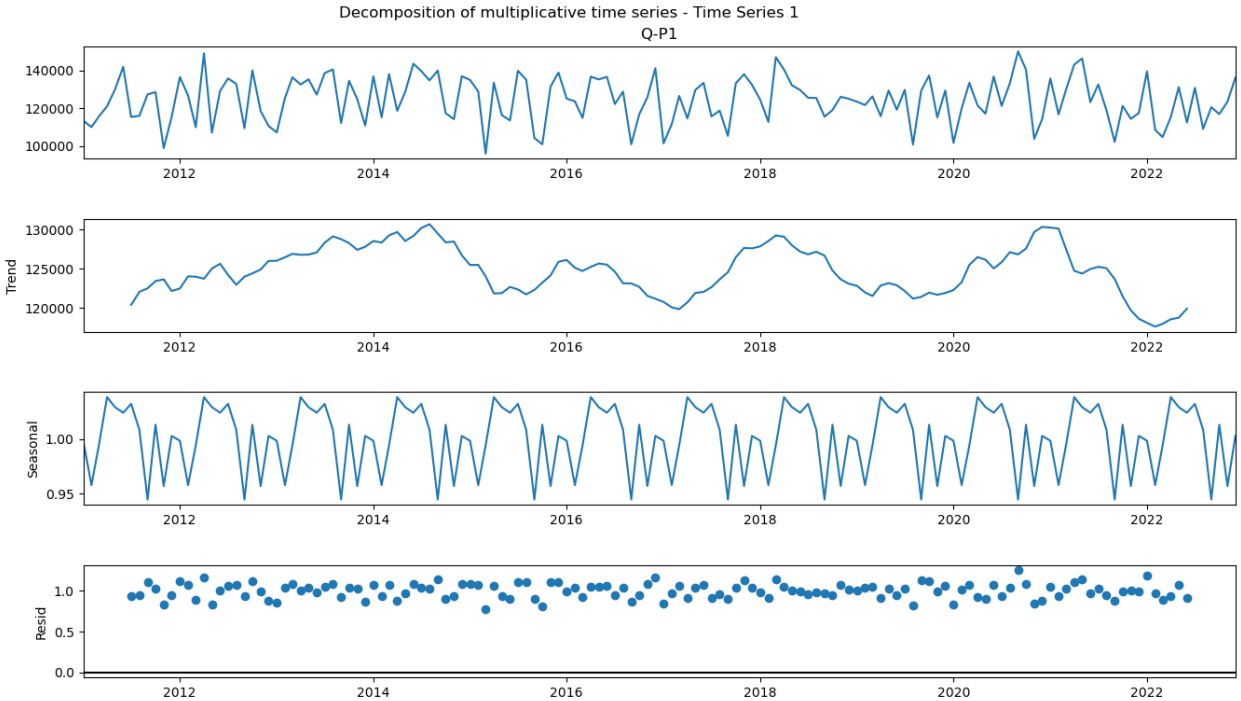
In previous part when I wanted to wrangle the data and explore data analysis, I found out the dates that are incomplete in 2010, 2023. Here, I decided to remove them. After that I will select the data between 2011 to 2019 as a train data. I will select the data between 2019 to 2022 as a test data. Test data size is 27% (3/11).

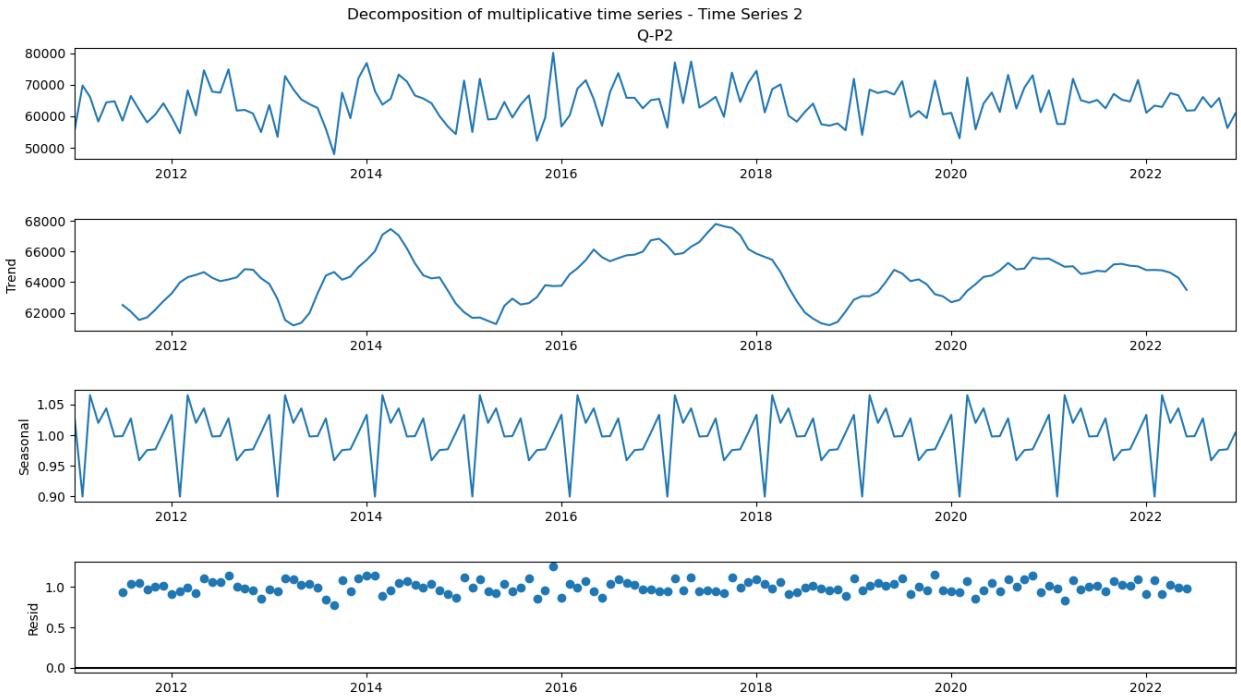


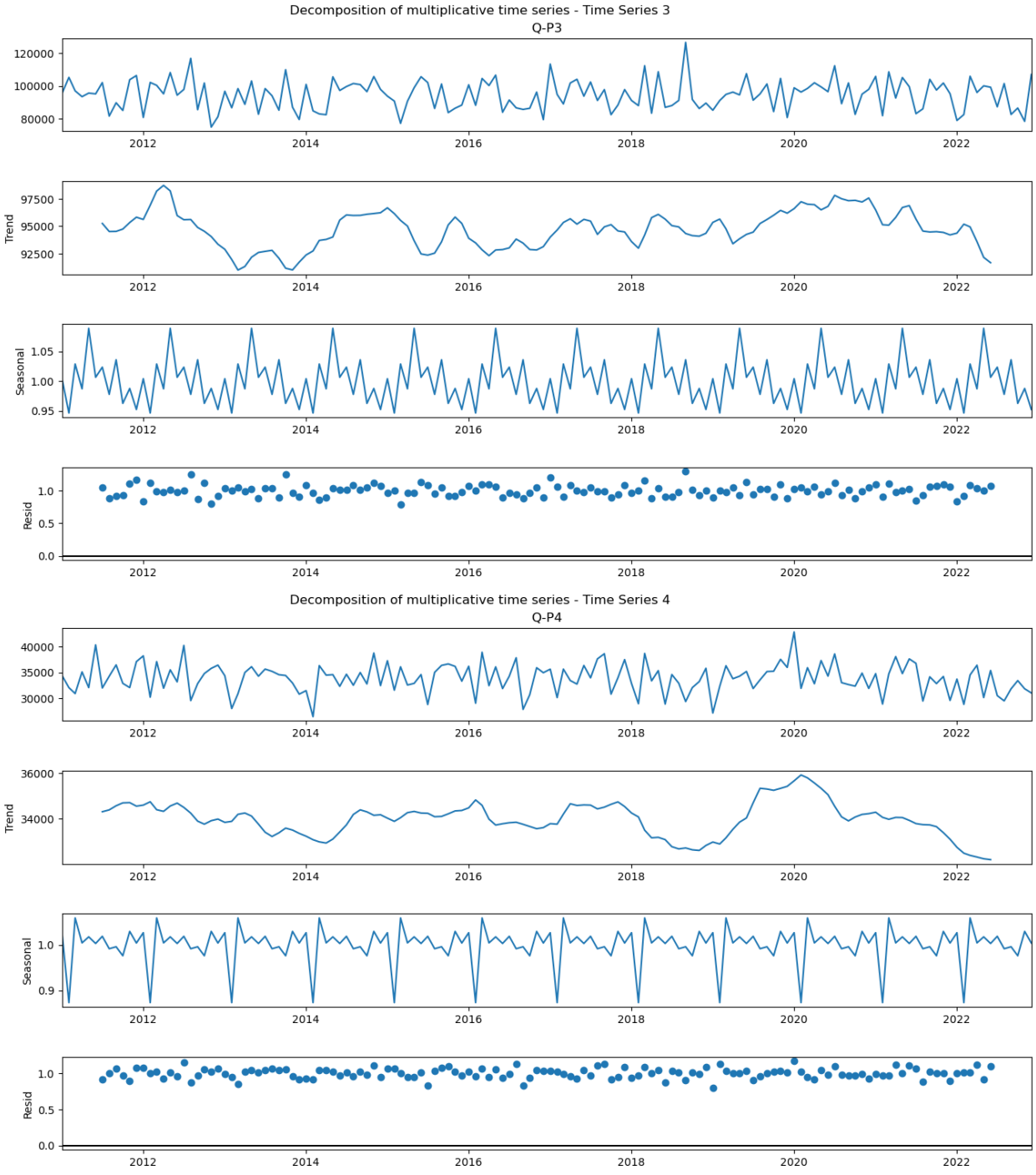
3.6 Time Series Decomposition

Here, I want to define the trend, seasonal, residual, and observed of changes to show what the data tell me.

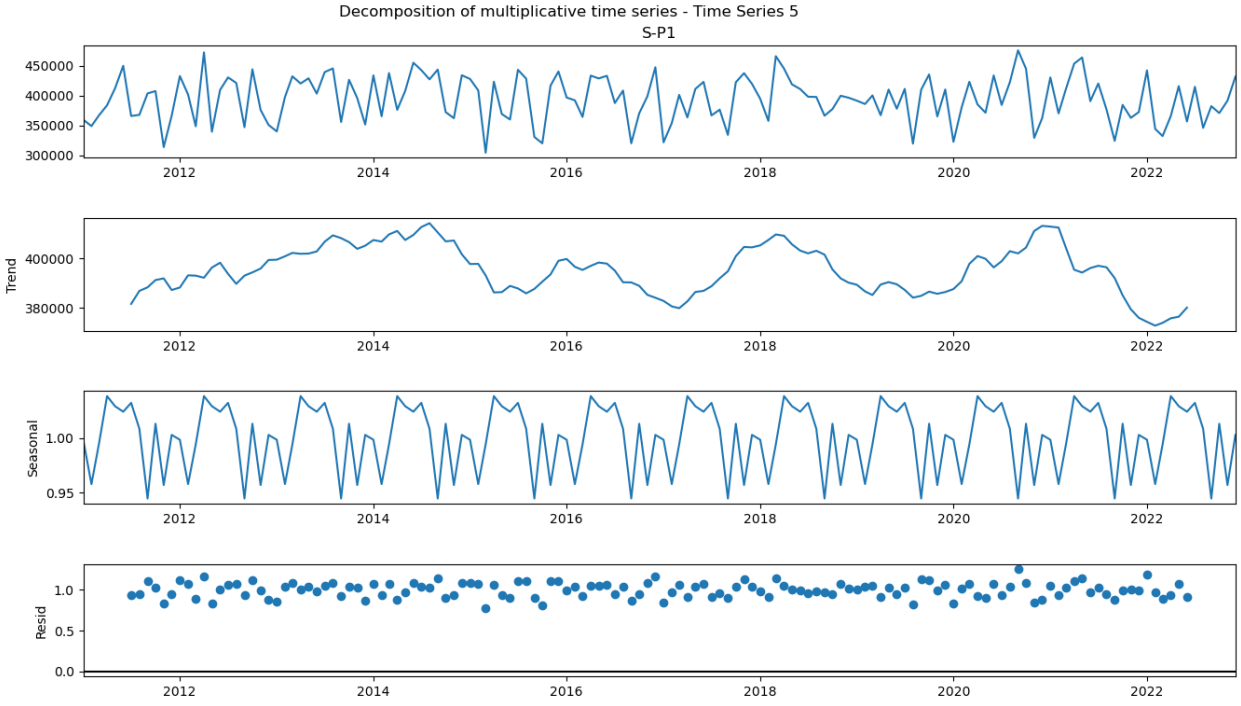
I want to create the charts to find out these concepts of my data.

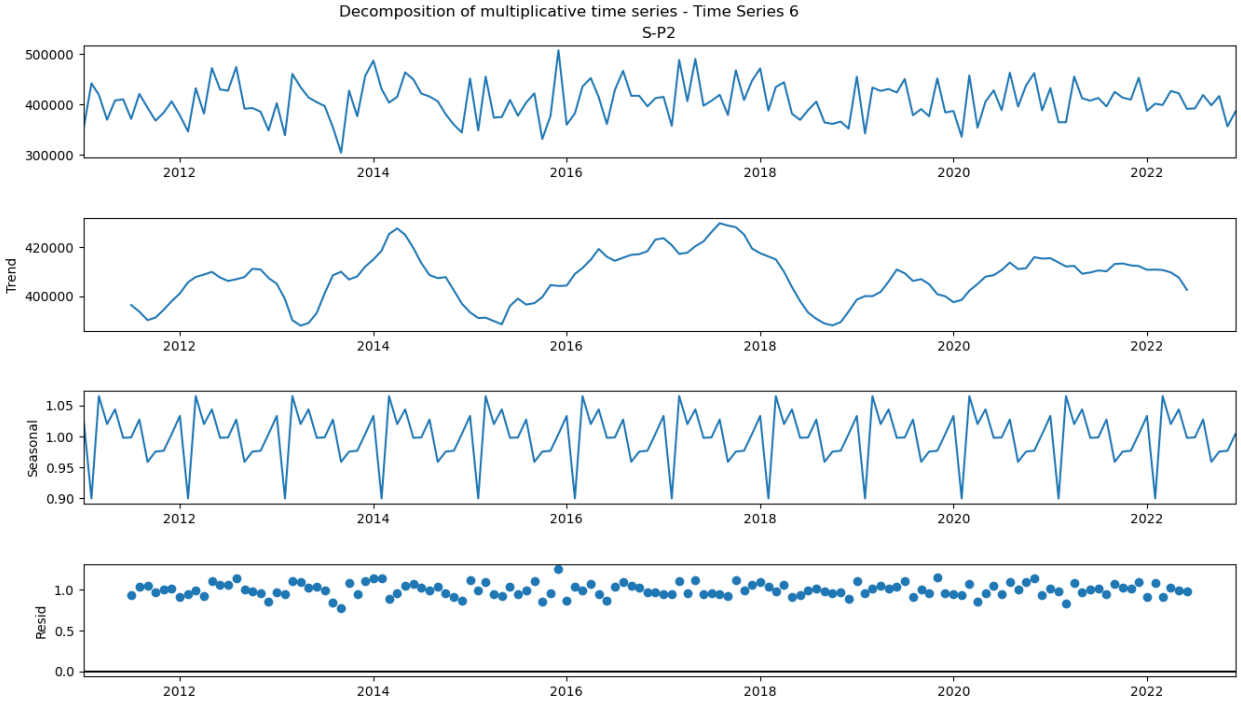


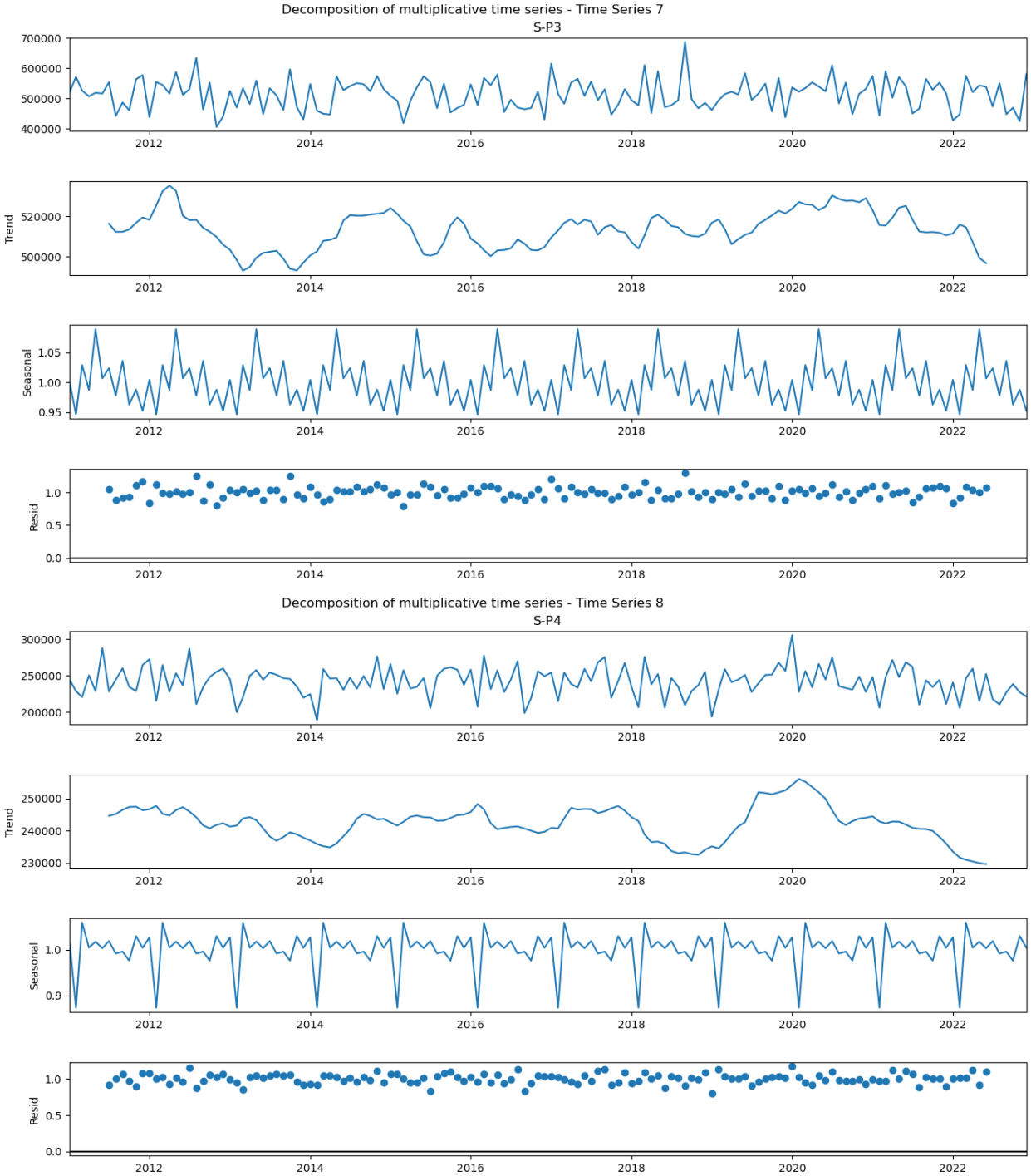




These are for quantity sales of product 1, 2, 3, and 4.


Decomposition of multiplicative time series - Time Series 5
S-P1
Trend
Seasonal
Resid
2012
2014
2016
2018
2020
2022


Decomposition of multiplicative time series - Time Series 6
S-P2
Trend
Seasonal
Resid
2012
2014
2016
2018
2020
2022



These are for revenue sales of product 1, 2, 3, and 4.

There are some differences in details, when I compare all 8-time series. But they are look similar. It means trends and patterns are the same as each other.

Observations:

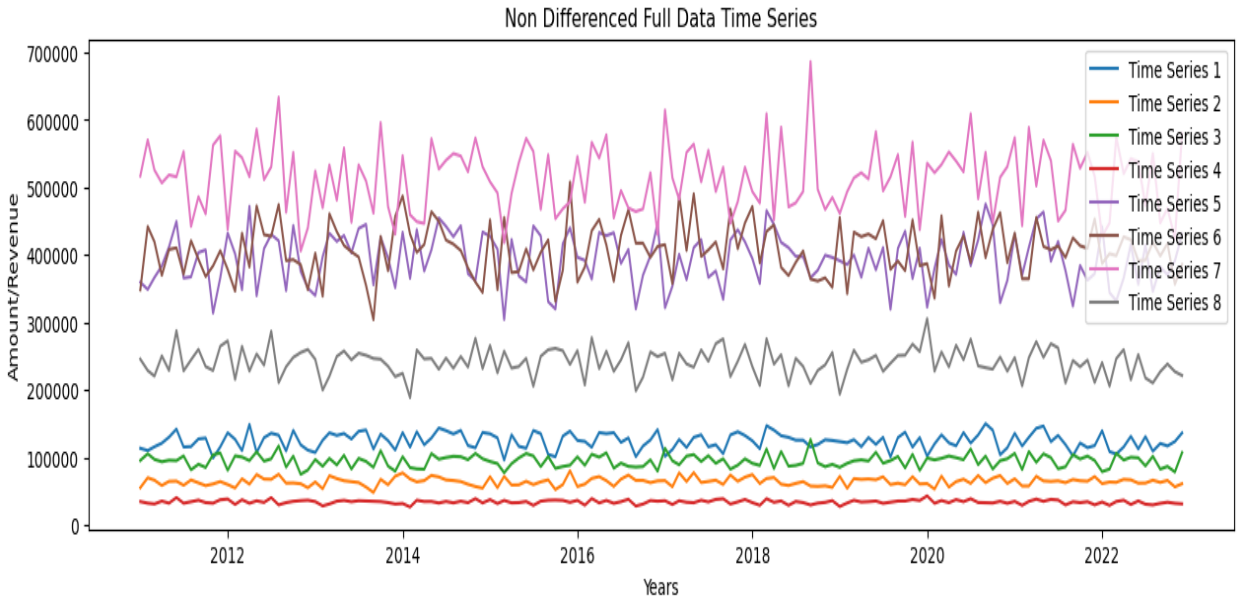
1.Trend: 12-months MA is a fairly straight line indicating a linear trend. Increasing at the first and decreasing at the end

are clear. As we notice before, we prefer to work on 2011-2022, because we have missing values in 2011, 2023, and 2024. and I decide to predict 2023 to 2024.

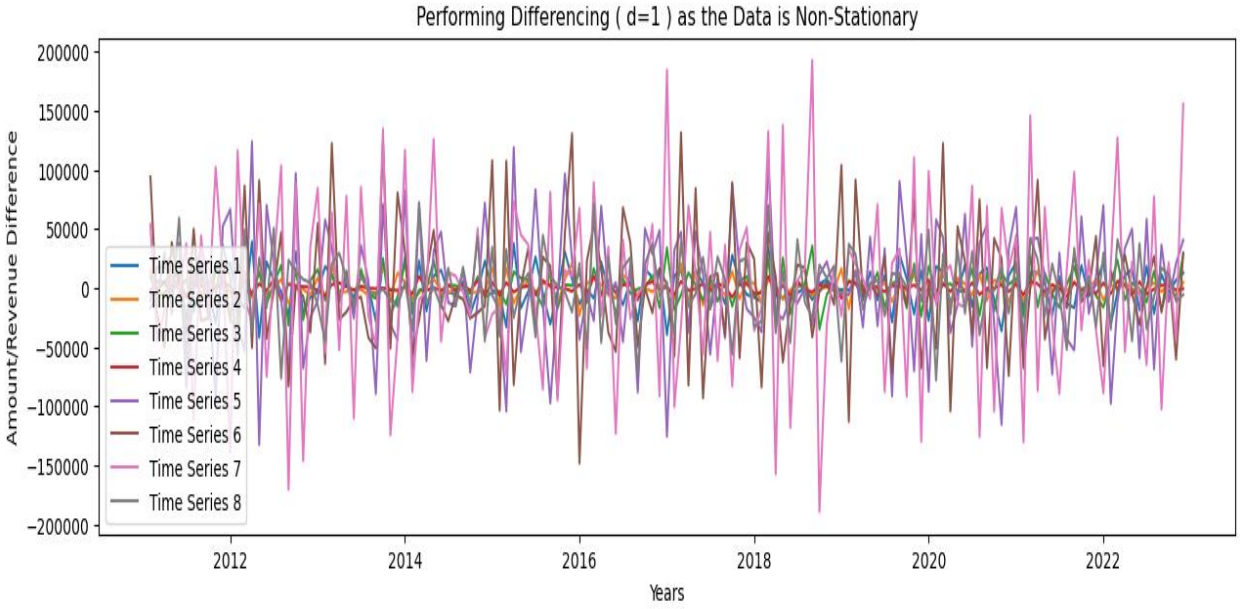
2.Seasonality: seasonality of 12 months is clearly visible.

3 Irregular Remainder (random): The multiplicative model works as there are no patterns in the residuals.

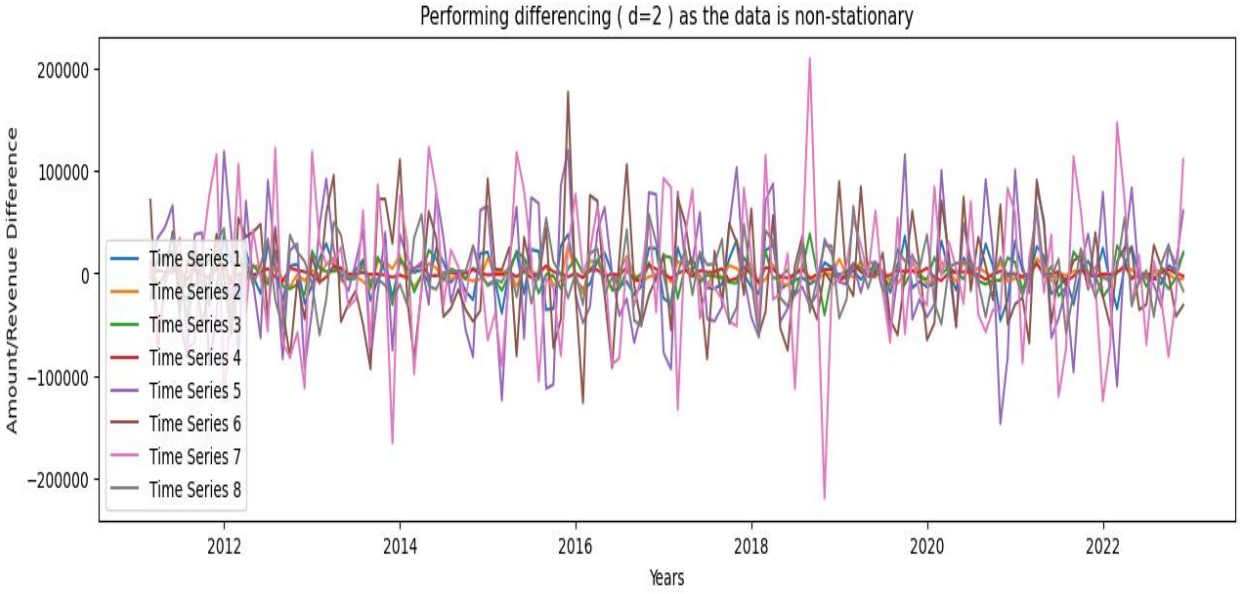
3.6.2 Non-Differenced Full Data Time Series



3.6.3 Performing Differencing (d=1) as the Data is non-stationary

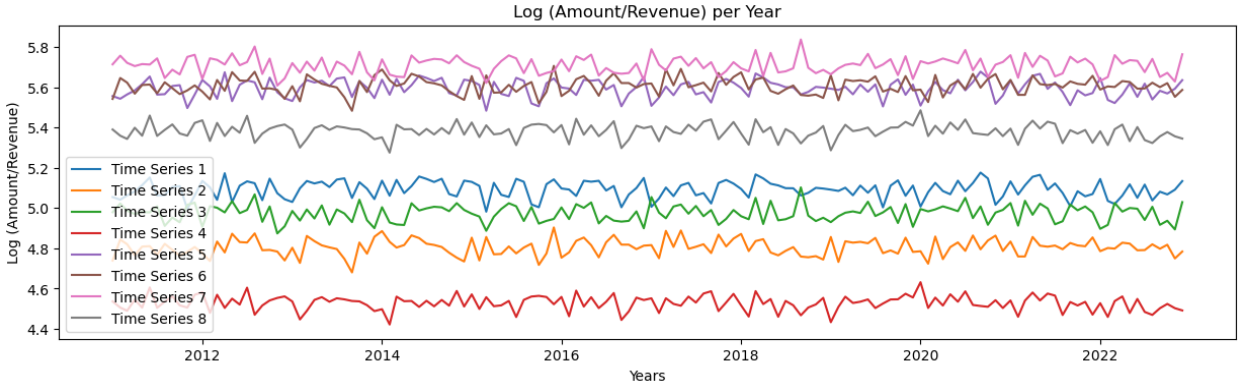


3.6.4 Performing differencing (d=2) as the data is non-stationary

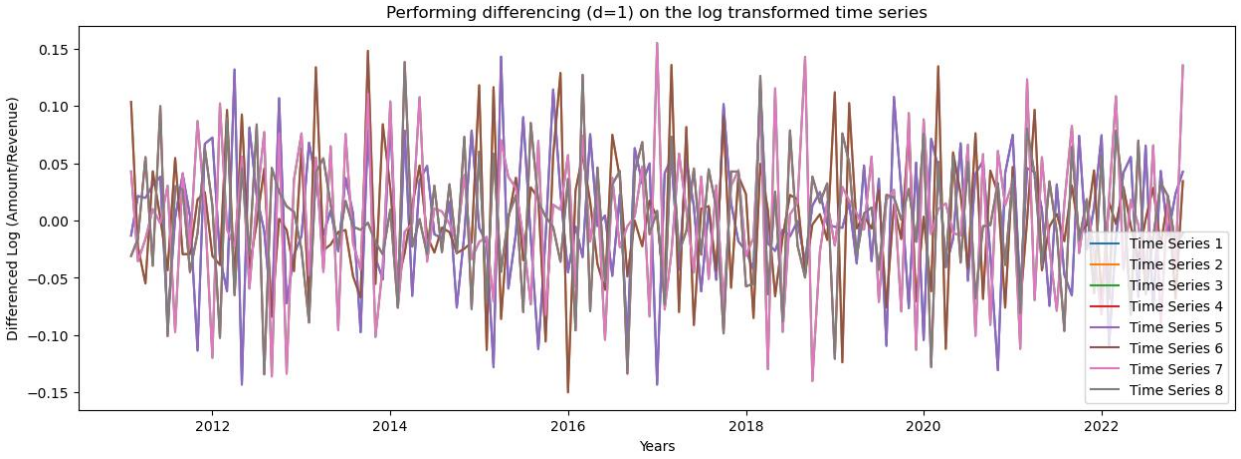


We observe seasonality even after differencing. This suggests a log transformation of the data.

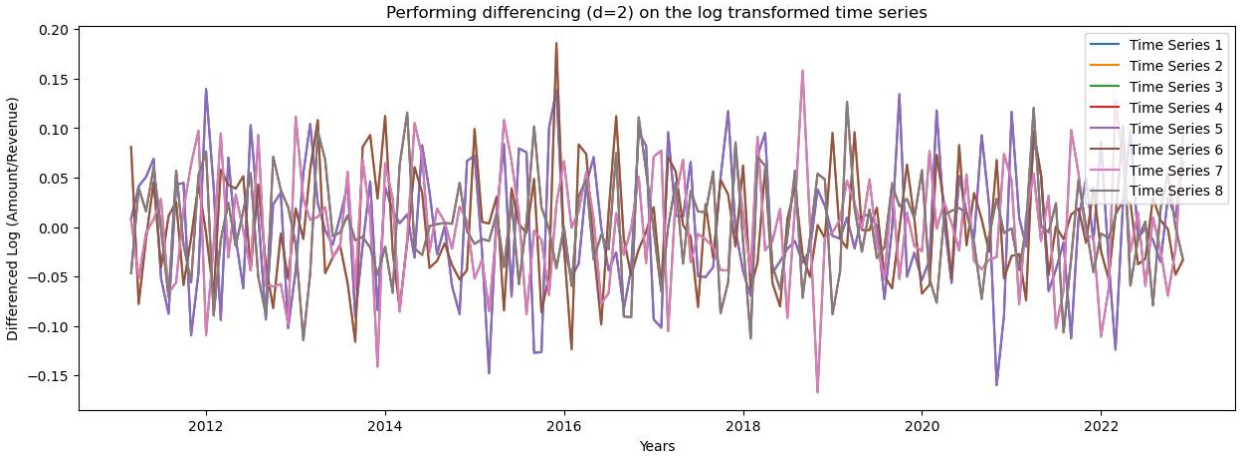
Log Transformed Time Series



3.6.5 Performing differencing (d=1) on the log transformed time series



3.6.6 Performing differencing (d=2) on the log transformed time series



3.7.2 Train Test Split

As I decided before I want to use AR, ARMA, ARIMA, and SARIMA model for prediction.

These models have some parameters that are below:

- p and seasonal P: indicate number of autoregressive terms (lags of the stationaries series)
- d and seasonal D: ...
- q and seasonal Q: indicate number of moving average terms (lags of the forecast errors)
- s: indicates seasonal length in the data.

I define these parameters here below:

p = 1, 2, 3

d = 0, 1

q = 1, 2, ,3

Here I try to select train data 2011 to 2019. I select test data 2019 to 2022.

P-values are approximately zero. We know it can't be zero, but it can be close to zero. Using the log transformed series as there is variance in the data.

3.8 Modeling for Q-P1 (The quantity of product 1)

Here I want to select quantity sales of product 1 (Ignore other columns). Then I will try to create and predict the data. If I wanted to work on all columns, the process will be too complicated.

As I selected, I will try AR, ARMA, ARIMA, and SARIMA for product 1 here.

3.8.1 AR Model for Q-P1 (The quantity of product 1)

AR Model: Autoregressive

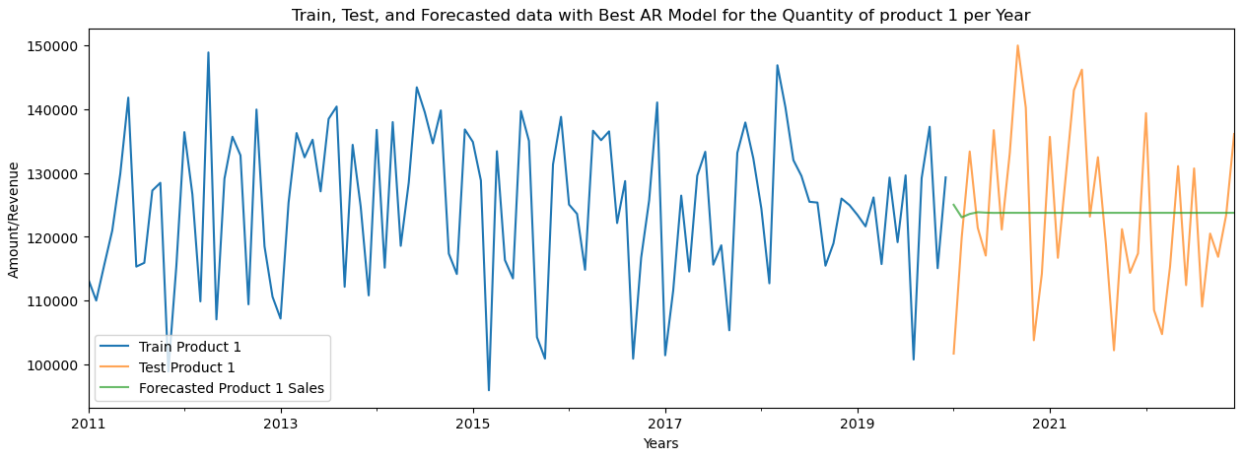
Use previous time period values to predict the current time period values AR Model building to estimate best 'p' (Lowest AIC Approach)

The best one with lowest AIC (Akaike Information Criteria) for quantity sales of product 1 is ARIMA (2, 0, 0).

3.8.1.1 Calculating RMSE with the best AR model for Q-P1 (The quantity of product 1)

The Root Mean Squared Error of our forecasts is 12674.089

3.8.1.2 Drawing Train, Test, and Forecasted data with Best AR Model for the Quantity of product 1 per Year:



	RMSE
Best AR Model Product 1: AR (2,0,0)	12674.088666

3.8.2 ARMA Model for Q-P1 (The quantity of product 1)

- Improving Autoregressive Models through Moving Average Forecasts.
- ARMA models consist of 2 components: -
- AR model: The data is modeled based on past observations.
- MA model: Previous forecast errors are incorporated into the model.

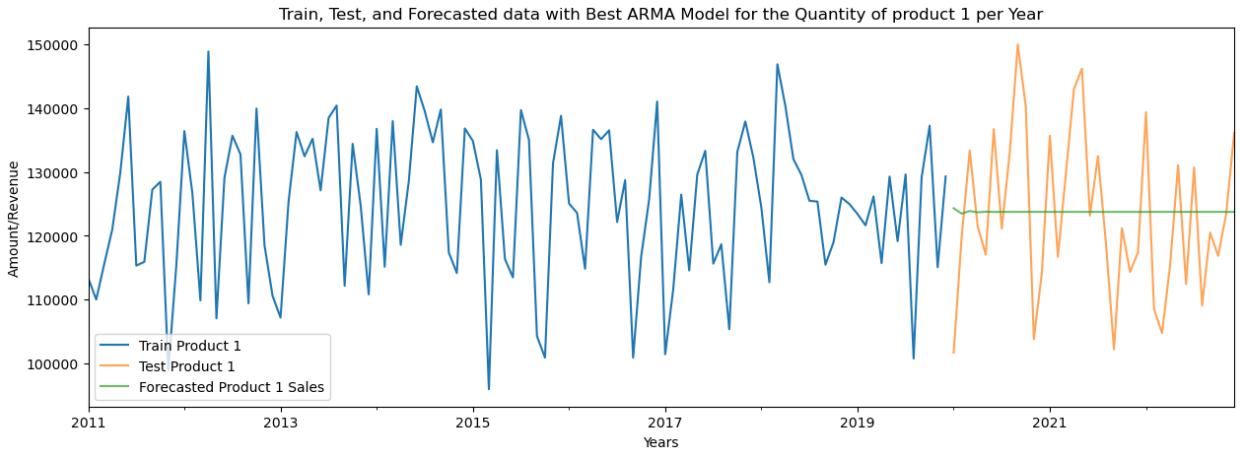
ARMA Model building to estimate best 'p', 'q' (Lowest AIC Approach)

The best one with lowest AIC for quantity sales of product 1 is ARMA (1, 0, 1).

3.8.2.1 Calculating RMSE with best MA model for Q-P1 (The quantity of product 1)

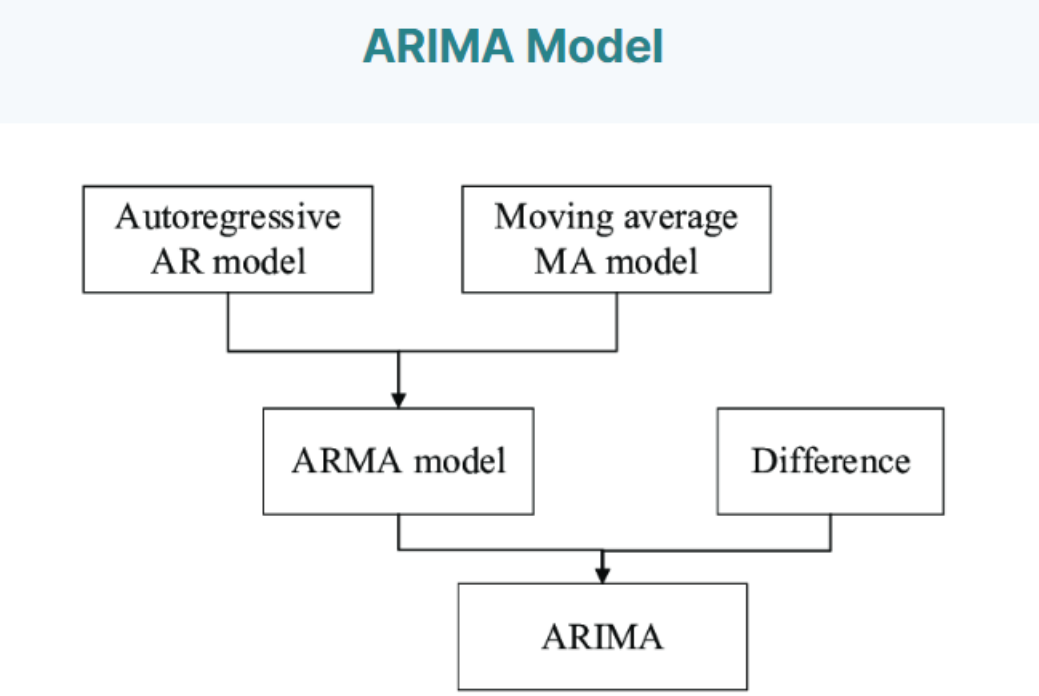
The Root Mean Squared Error of our forecasts is 12634.902

3.8.2.2 Drawing Train, Test, and Forecasted data with Best ARMA Model for the Quantity of product 1 per Year:



	RMSE
Best AR Model Product 1 : AR(2,0,0)	12674.088666
Best ARMA Model Product 1: ARMA (1, 0, 1)	12634.901547

3.8.3 ARIMA Model for Q-P1 (The quantity of product 1)



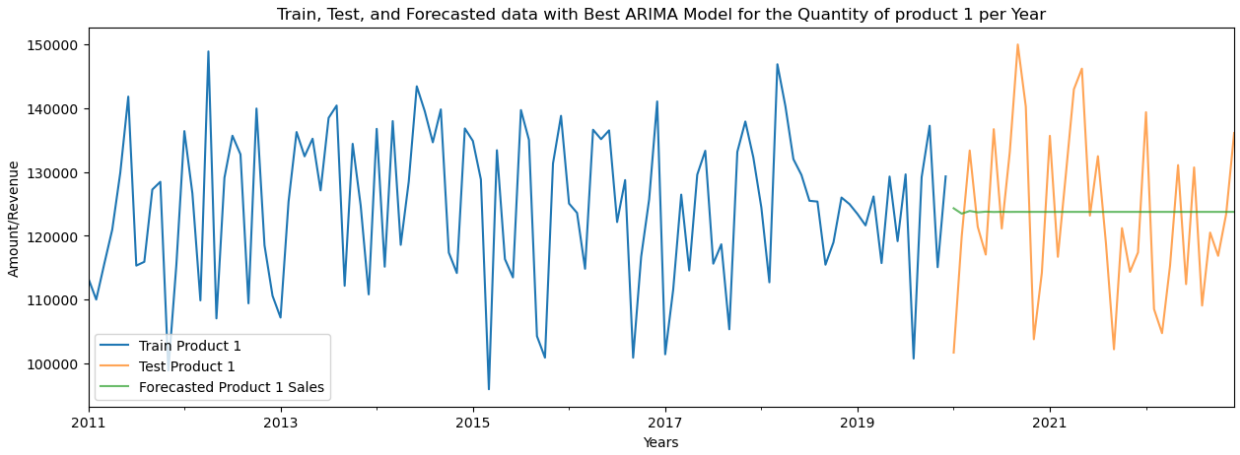
- ARIMA: Auto Regressive Integrated Moving Average is a way of modeling time series data for forecasting or predicting future data points.
- Improving AR Models by making Time Series stationary through Moving Average Forecasts
- ARIMA models consist of 3 components
- AR model: The data is modeled based on past observations.
- Integrated component: Whether the data needs to be differenced/transformed.
- MA model: Previous forecast errors are incorporated into the model.

The best one with lowest AIC for quantity sales of product 1 is ARIMA (1, 0, 1). The results of ARMA and ARIMA are exactly the same. So, we predict the results of them for quantity sales of product 1 should be the same, as well.

3.8.3.1 Calculating RMSE with best ARIMA model for Q-P1 (The quantity of product 1)

The Root Mean Squared Error of our ARIMA forecasts is 12634.902

3.8.3.2 Drawing Train, Test, and Forecasted data with Best ARIMA Model for the Quantity of product 1 per Year:



	RMSE
Best AR Model Product 1: AR (2,0,0)	12674.088666
Best ARMA Model Product 1: ARMA (1, 0, 1)	12634.901547
Best ARIMA Model Product 1: ARIMA (1,0,1)	12634.901547

As I predicted before the results of ARMA and ARIMA are the same.

3.8.4 SARIMA Model for Q-P1 (The quantity of product 1)

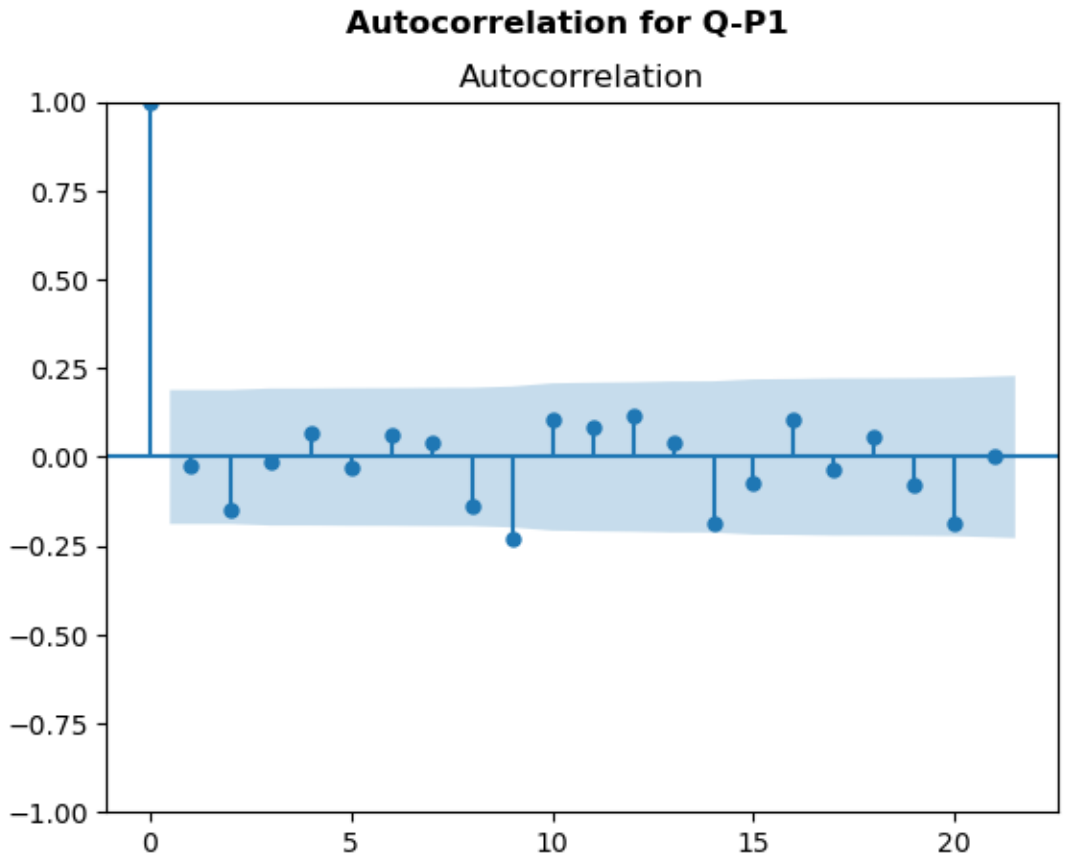
- The ARIMA models can be extended/improved to handle seasonal components of a data series
- The seasonal autoregressive moving average model is given by

SARIMA (p, d, q) (P, D, Q) m non-seasonal seasonal

- The above model consists of:
- Autoregressive and moving average components (p, q)
- Seasonal autoregressive and moving average components (P, Q)
- The ordinary and seasonal difference components of order ‘d’ and ‘D’
- Seasonal frequency ‘F’
- The value for the parameters (p, d, q) and (P, D, Q) can be decided by comparing different values for each and taking the lowest AIC value for the model build.
- The value for F can be consolidated by ACF plot

Finding Seasonality = 12 from ACF/PACF plots

Here I create auto correlation for Q-P1



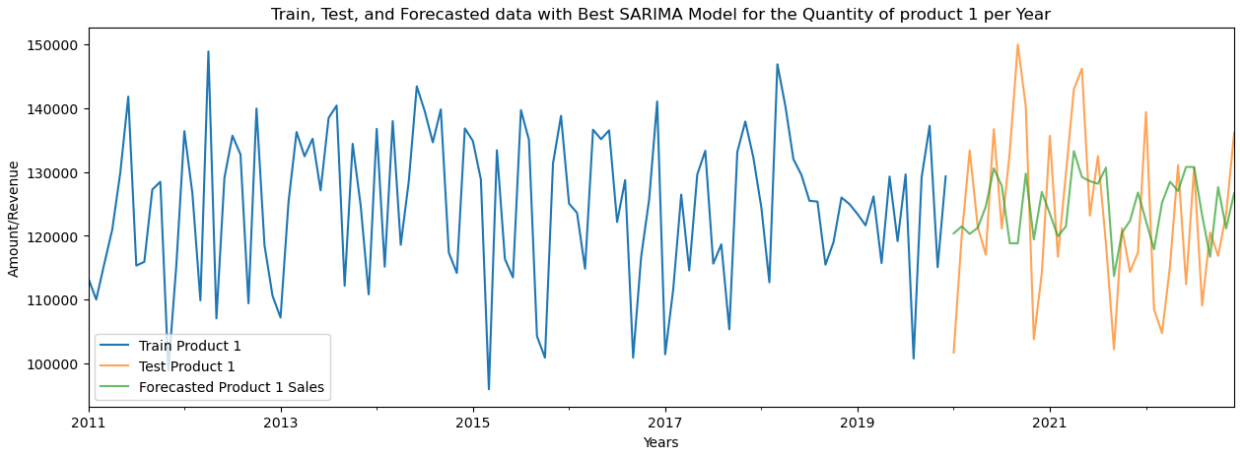
Inference * Criteria to choose the best fit model is the lowest/minimum AIC value For ARIMA (p, d, q) × (P, D, Q) S, we got SARIMAX (3, 1, 3) x (3, 1, 3, 12) model with the least AIC of 3984.33757 Here,

- p = non-seasonal AR order = 3,
- d = non-seasonal differencing = 0,
- q = non-seasonal MA order = 3,
- P = seasonal AR order = 3,
- D = seasonal differencing = 1,
- Q = seasonal MA order = 3,
- S = time span of repeating seasonal pattern = 12 Building SARIMA model with the best parameters

3.8.4.1 RMSE with the best SARIMA Model for Q-P1 (The quantity of product 1)

The Root Mean Squared Error of our forecasts SARIMA_0 is 12001.957

3.8.4.2 Drawing Train, Test, and Forecasted data with Best SARIMA Model for the Quantity of product 1 per Year



	RMSE
Best AR Model Product 1: AR (2,0,0)	12674.088666
Best ARMA Model Product 1: ARMA (1, 0, 1)	12634.901547
Best ARIMA Model Product 1: ARIMA (1,0,1)	12634.901547
Best SARIMA Model product 1: SARIMA (3, 1, 3) x (3, 1, 3, 12)	12001.956713

3.8.5 Conclusion for Q-P1 (The quantity of product 1):

The best model for product 1 is SARIMAX (3, 1, 3) x (3, 1, 3, 12) that its RMSE is 12001.956713

3.8.5.1 Forecast with the best model with the lowest RMSE

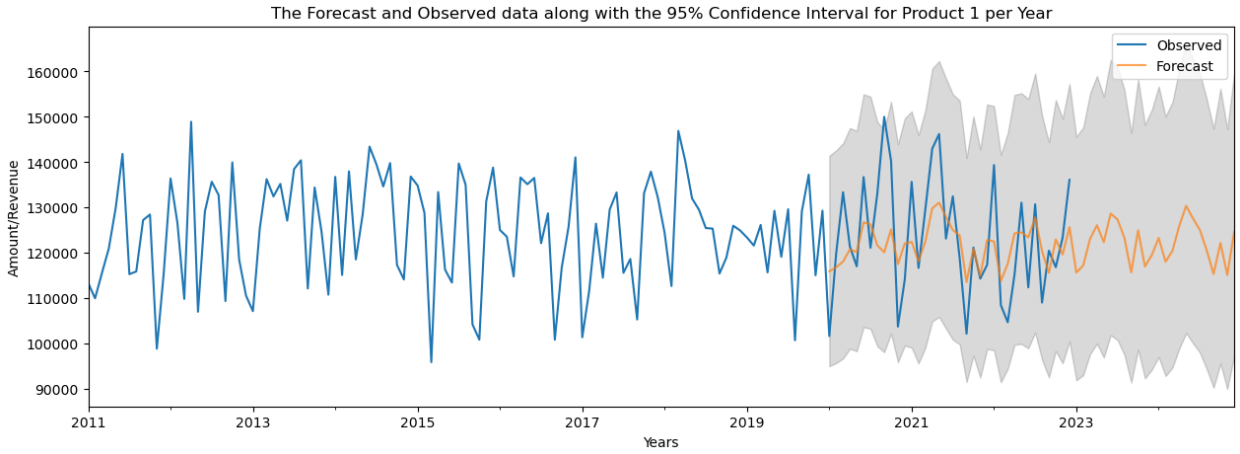
When plotting a forecast along with confidence bands of 99% and 95%, we're visualizing the uncertainty inherent in the prediction. The forecast itself represents the most likely outcome, while the confidence bands indicate the range within which we expect the actual outcomes to fall with a certain level of certainty. The 99% confidence band is wider than the 95% band, reflecting a higher level of confidence in capturing the true outcome within that range. Essentially, if we were to repeat the forecasting process numerous times, we'd expect about 99% (or 95%) of the actual outcomes to fall within the respective confidence bands. This visualization is crucial for decision-makers, as it helps them understand the range of potential outcomes and make informed choices considering the associated uncertainty.

	forecast	lower_ci_95	upper_ci_95	lower_ci_99	upper_ci_99
2020-01-01	115953.424199	95057.551431	141442.698460	95057.551431	141442.698460
2020-02-01	116854.802732	95744.508305	142619.615093	95744.508305	142619.615093
2020-03-01	118097.324744	96741.991966	144166.745260	96741.991966	144166.745260
2020-04-01	120802.468866	98913.279432	147535.665261	98913.279432	147535.665261
2020-05-01	120216.044846	98364.897144	146921.288570	98364.897144	146921.288570

3.8.5.2 Draw the Forecast Observed data along with the 95% Confidence Interval for Product 1 per Year:

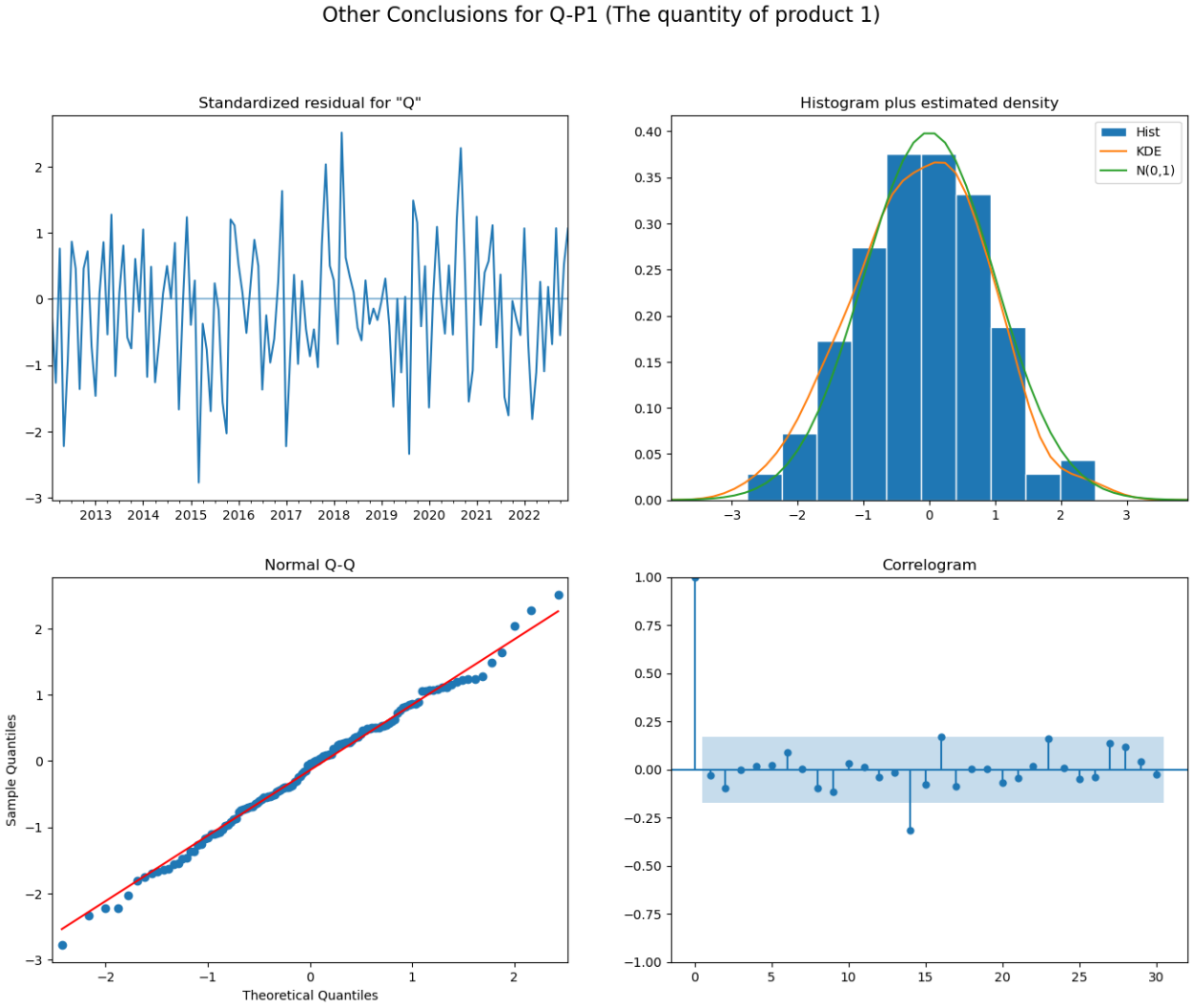
I decide to visualize with the 95% confidence band.

The final result of forecasting sales of product 1 is here.



So, we can see the 5th request of the CEO. As we can see we can use the best model (SARIMAX (3, 1, 3) x (3, 1, 3, 12)) for product 1, and we can predict the amount of product 1 for 2023, 2024 and etc.

3.8.5.3 Other Conclusions for Q-P1 (The quantity of product 1):



Plot ACF and PACF for residuals of the best model to ensure no more information is left for extraction.

Inference Note: 4 plots in the residuals diagnostic plots tell us:

- Standardized residuals plot the top left plot shows 1-step-ahead standardized residuals. If model is working correctly, then no pattern should be obvious in the residuals which is clearly not visible from the plot as well.
- Histogram plus estimated density plot This plot shows the distribution of the residuals. The orange line shows a smoothed version of this histogram, and the green line shows a normal distribution. If the model is good these two lines should be the same. Here there are small differences between them, which indicate that our model is doing just well enough.
- Normal Q-Q plot the Q-Q plot compare the distribution of residuals to normal distribution. If the distribution of the residuals is normal, then all the points should lie along the red line, except for some values at the end, which is exactly happening in this case.

- Correlogram plot the correlogram plot is the ACF plot of the residuals rather than the data. 95% of the correlations for lag >0 should not be significant (within the blue shades). If there is a significant correlation in the residuals, it means that there is information in the data that was not captured by the model, which is clearly not in this case.

This introduction of time series models explains ARMA, ARIMA and SARIMA models in a progressive order.

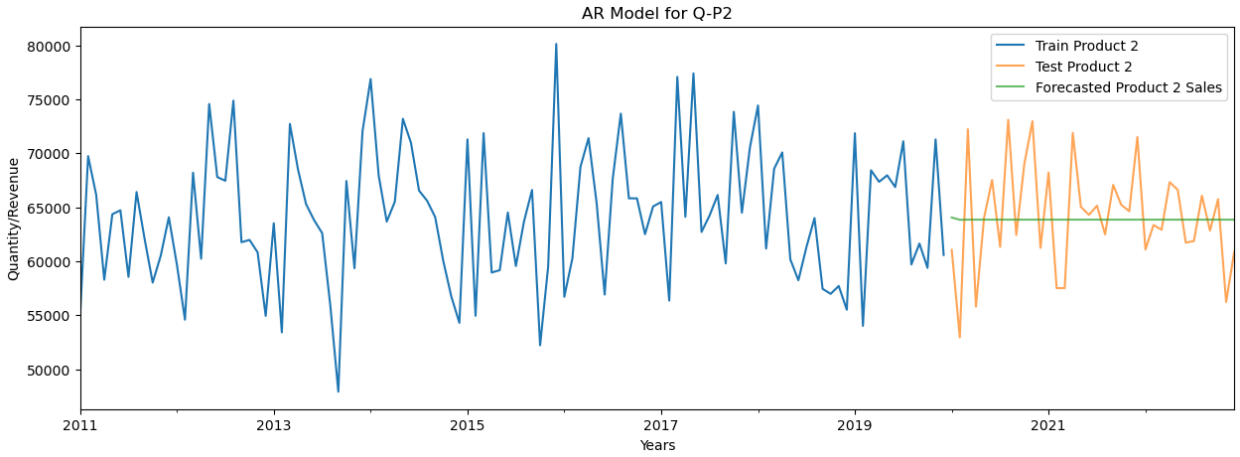
- ARMA: Autoregressive + Moving Average
- ARIMA: Autoregressive + Moving Average + Trend Differencing
- SARIMA: Autoregressive + Moving Average + Trend Differencing + Seasonal Differencing Furthermore, we explored concepts and techniques related to time series data, such as Stationarity, ADF test, ACF/PACF plot and AIC.

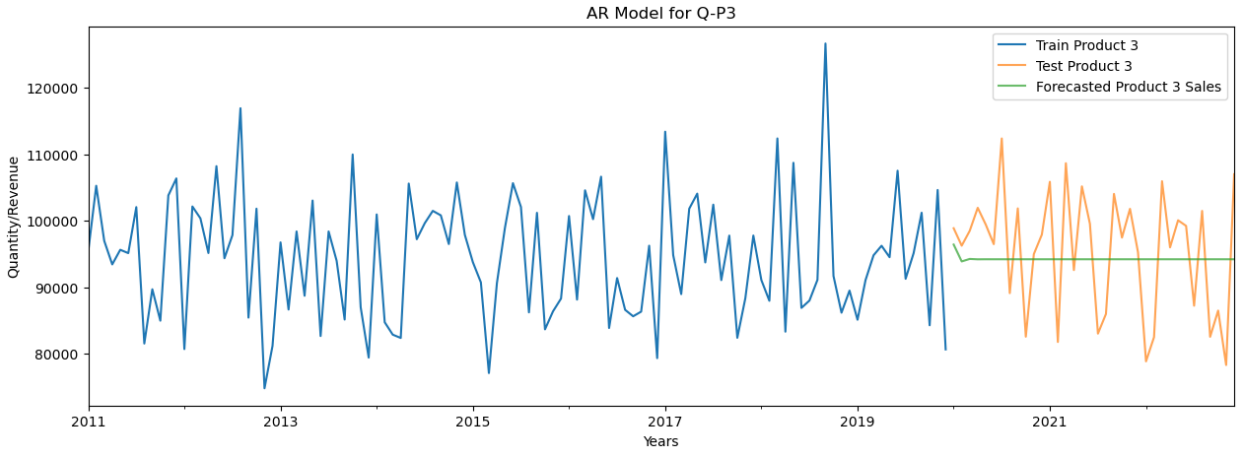
I create the best model with this pattern for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4, as well.

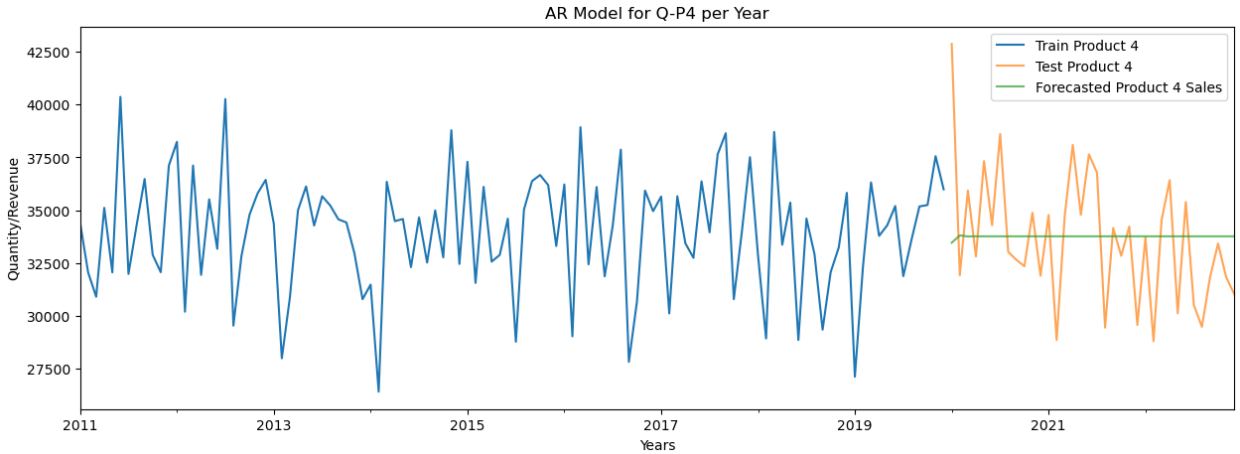
3.9 Modeling for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

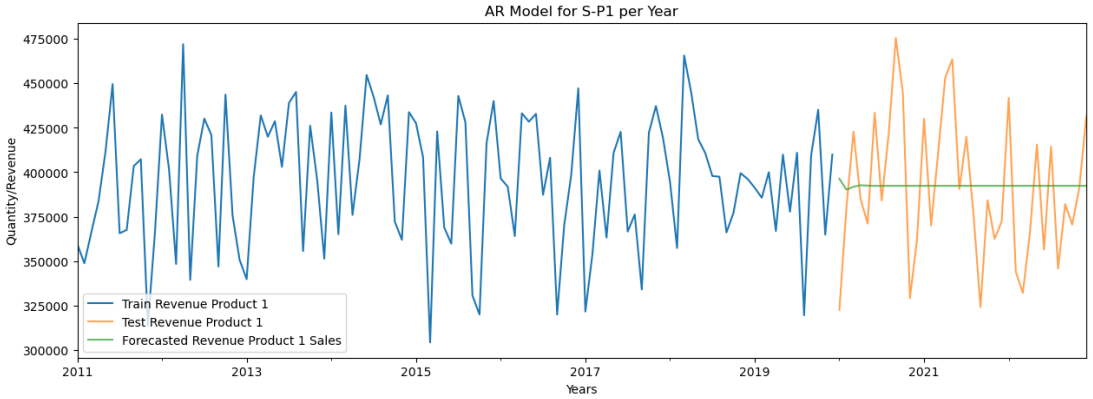
Here I will try to write main results (not details), because it becomes shorter.

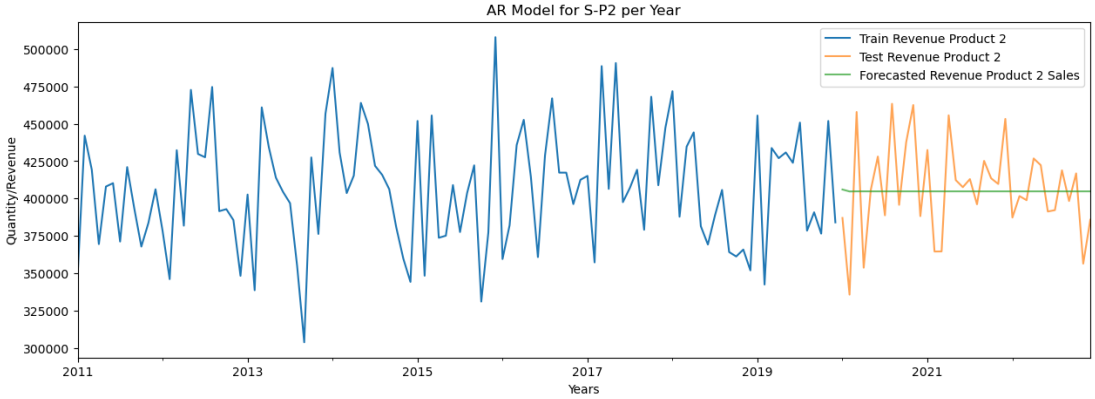
3.9.1.2 Drawing Train, Test, and Forecasted data with Best AR Model for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4 per Year

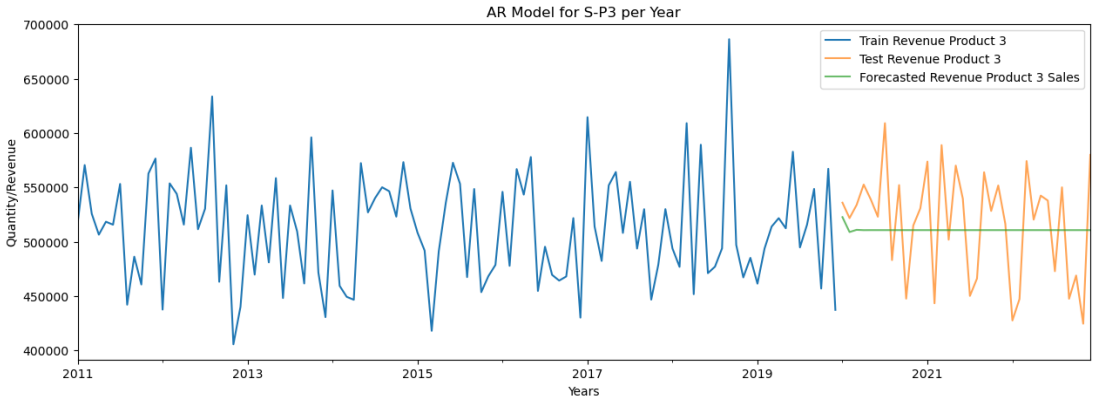


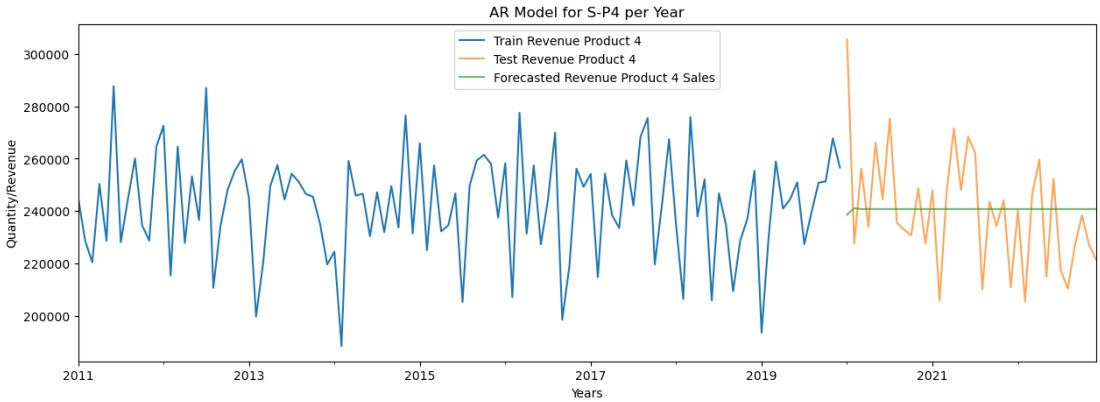




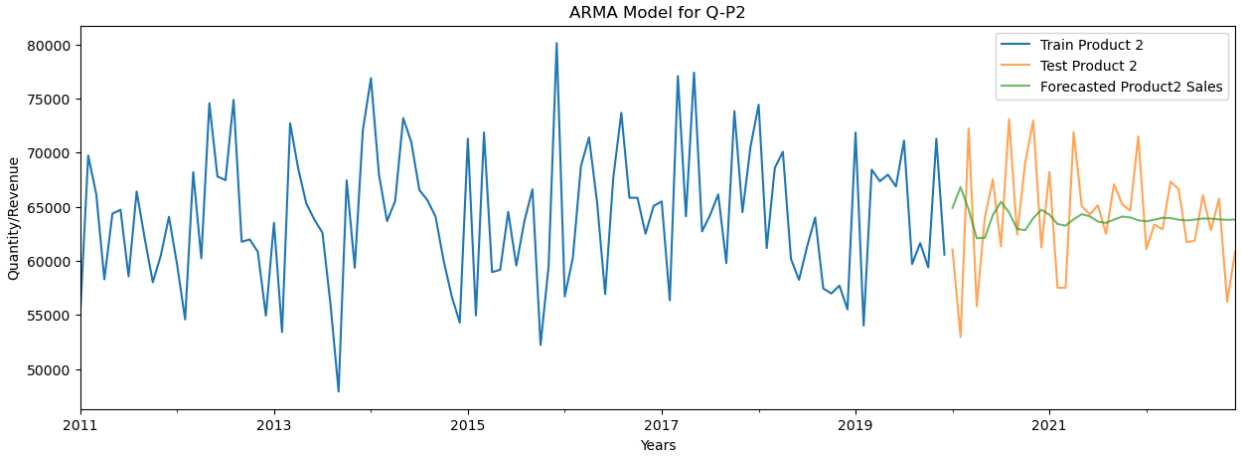

AR Model for S-P1 per Year
Train Revenue Product 1
Test Revenue Product 1
Forecasted Revenue Product 1 Sales
Quantity/Revenue
Years

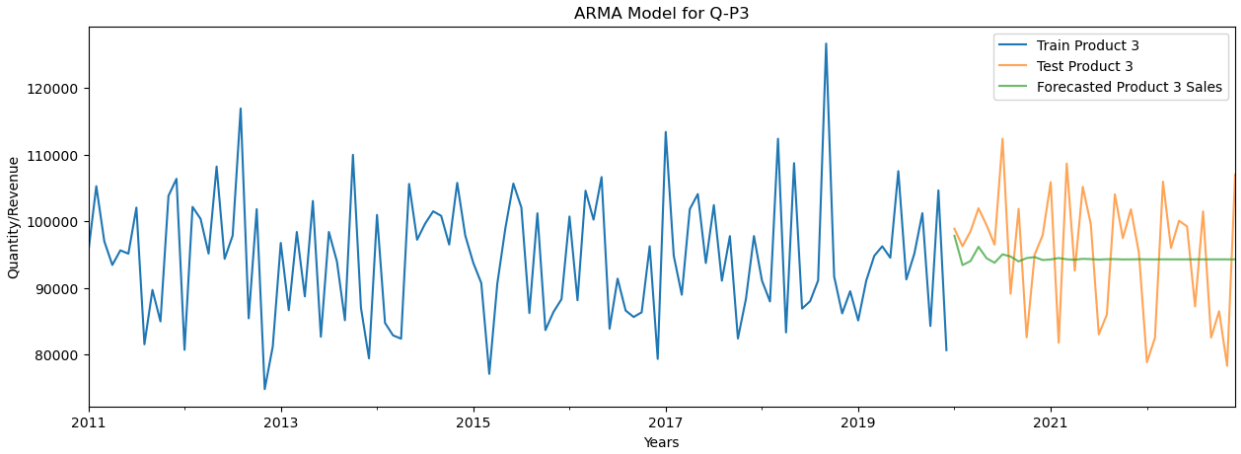

AR Model for S-P2 per Year
Train Revenue Product 2
Test Revenue Product 2
Forecasted Revenue Product 2 Sales
Quantity/Revenue
Years

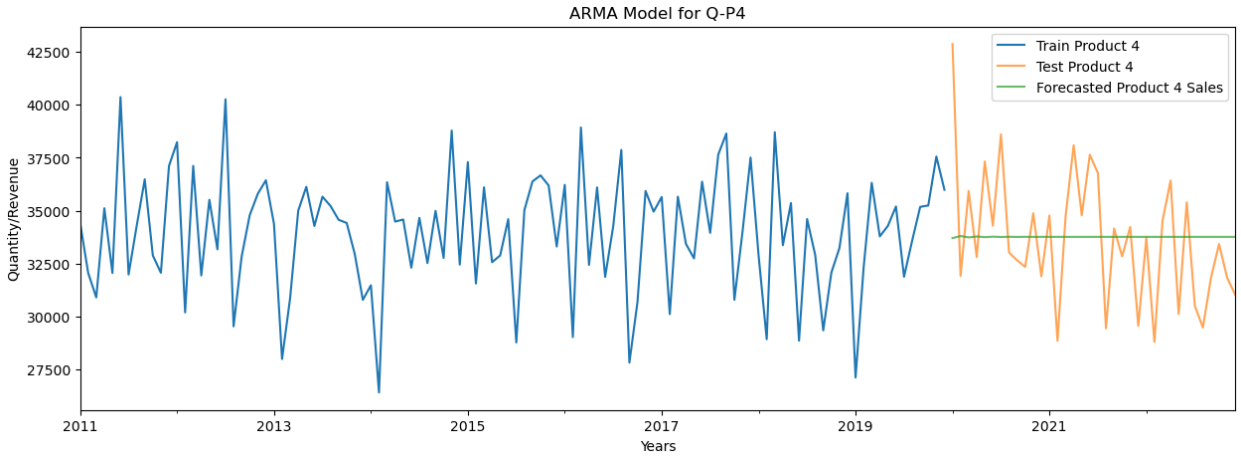

AR Model for S-P3 per Year
Train Revenue Product 3
Test Revenue Product 3
Forecasted Revenue Product 3 Sales
Quantity/Revenue
Years

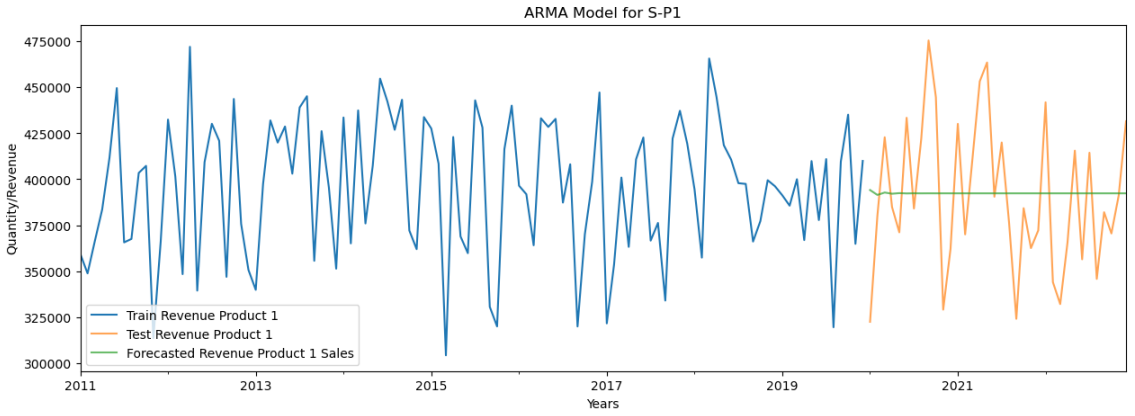

AR Model for S-P4 per Year
Train Revenue Product 4
Test Revenue Product 4
Forecasted Revenue Product 4 Sales
Quantity/Revenue
Years

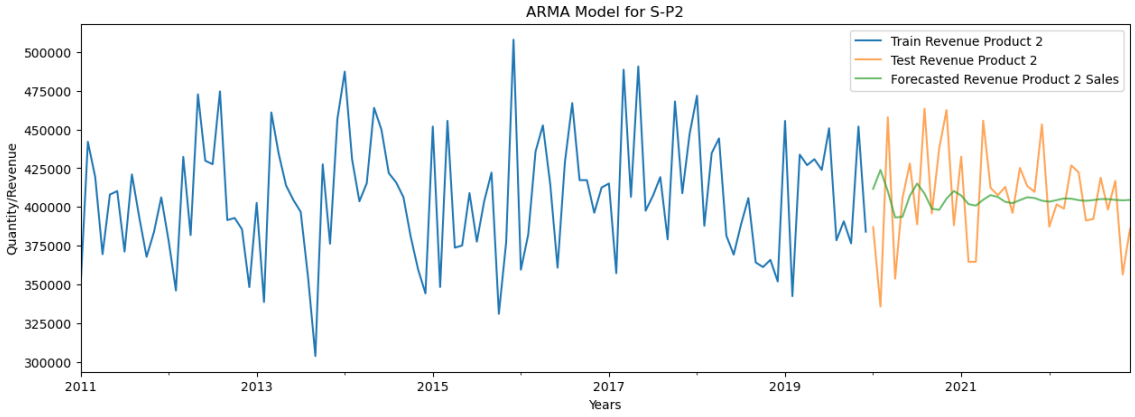
3.9.2.2 Drawing with best ARMA Model for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

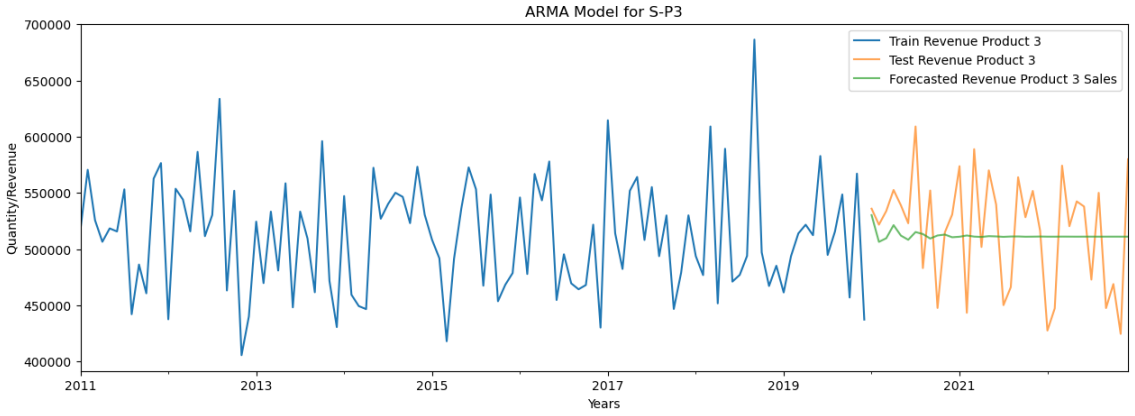


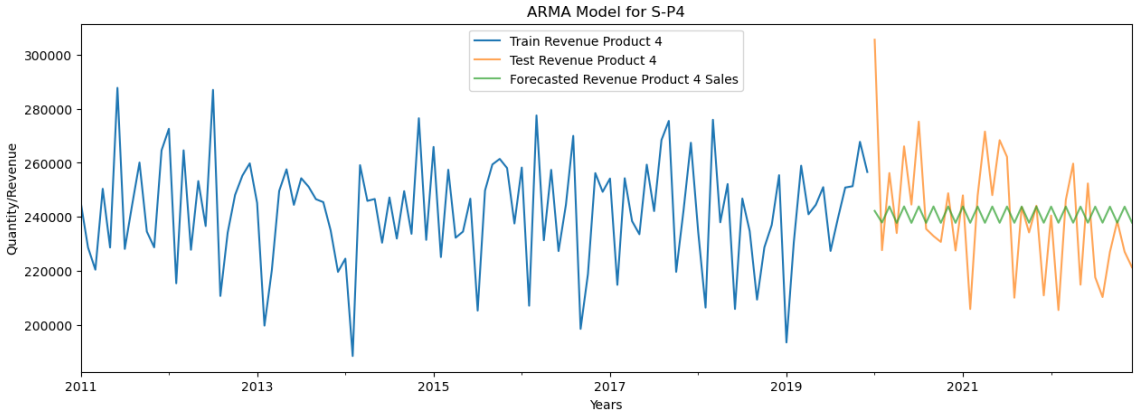




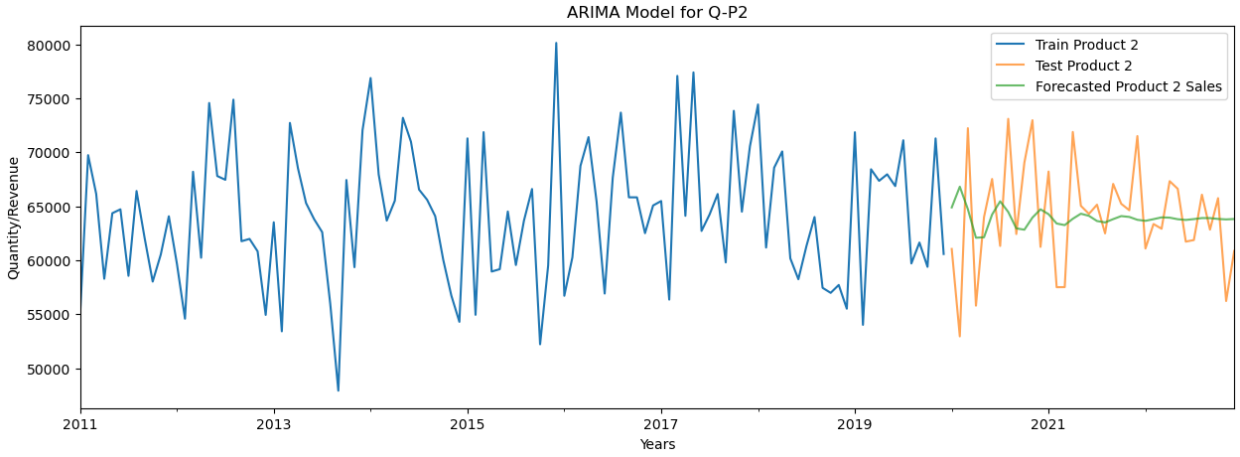

ARMA Model for S-P1
Quantity/Revenue
Years
Train Revenue Product 1
Test Revenue Product 1
Forecasted Revenue Product 1 Sales

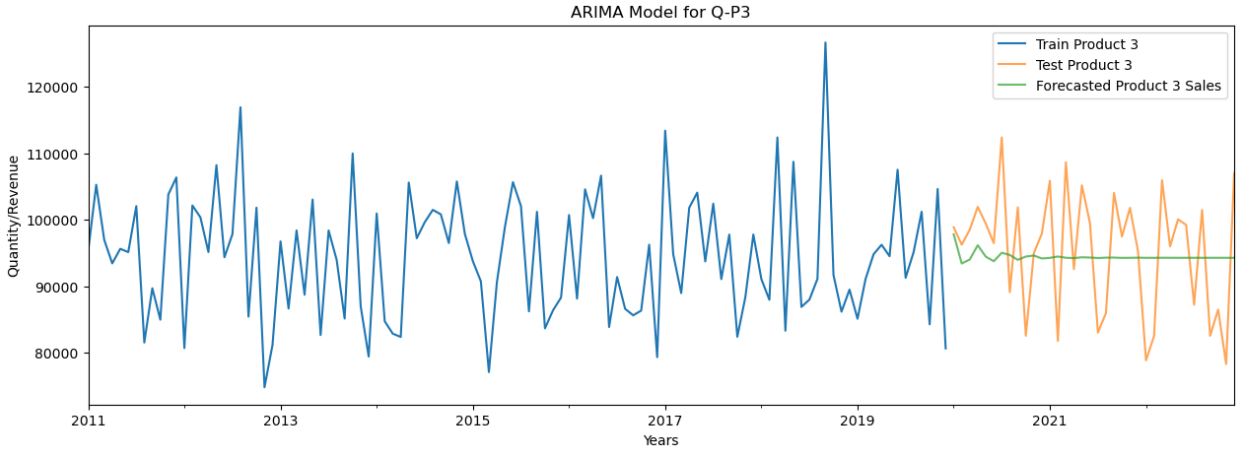

ARMA Model for S-P2
Quantity/Revenue
Years
Train Revenue Product 2
Test Revenue Product 2
Forecasted Revenue Product 2 Sales

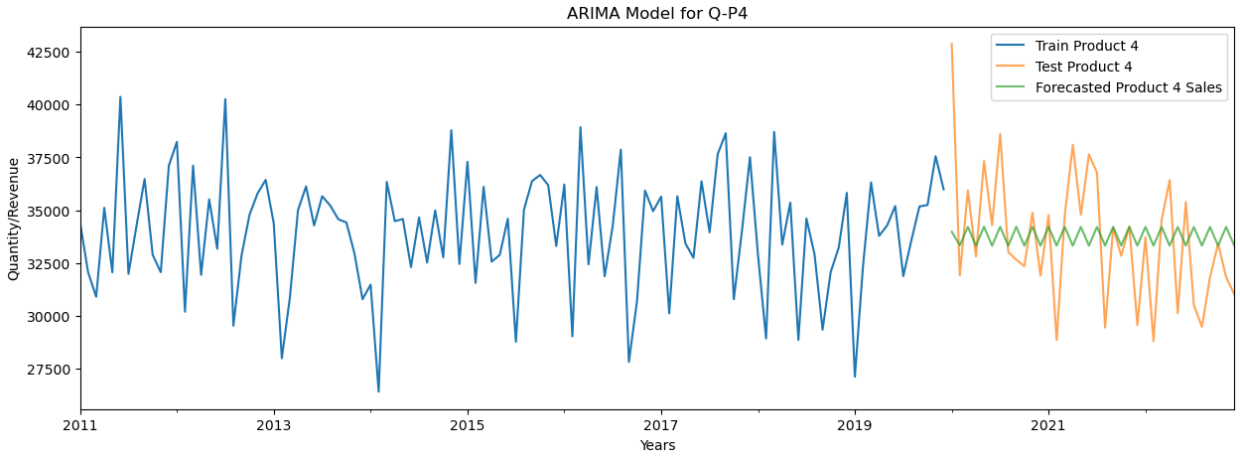

ARMA Model for S-P3
Quantity/Revenue
Years
Train Revenue Product 3
Test Revenue Product 3
Forecasted Revenue Product 3 Sales

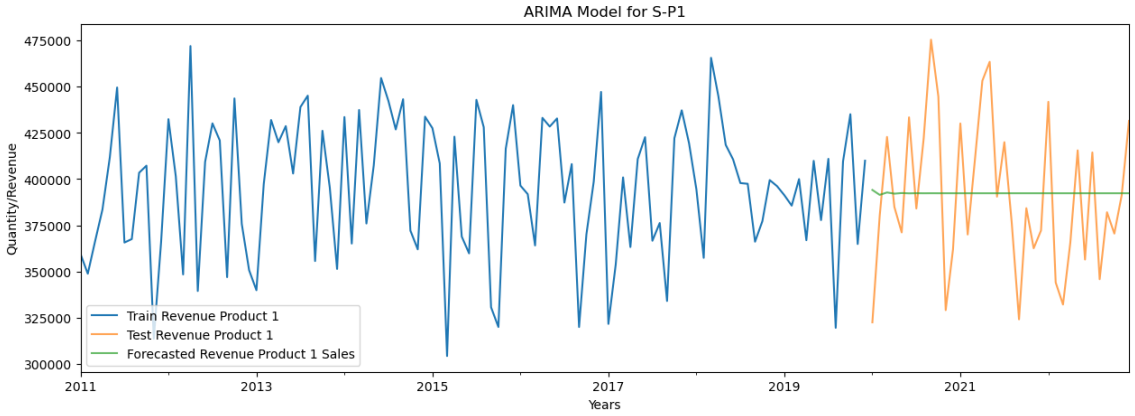

ARMA Model for S-P4
Quantity/Revenue
Years
Train Revenue Product 4
Test Revenue Product 4
Forecasted Revenue Product 4 Sales

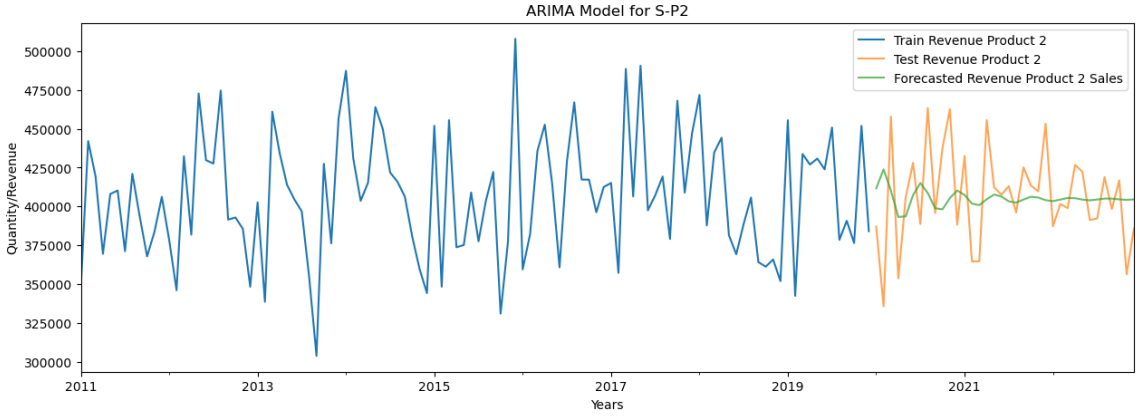
3.9.3.2 Drawing with best ARIMA Model for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

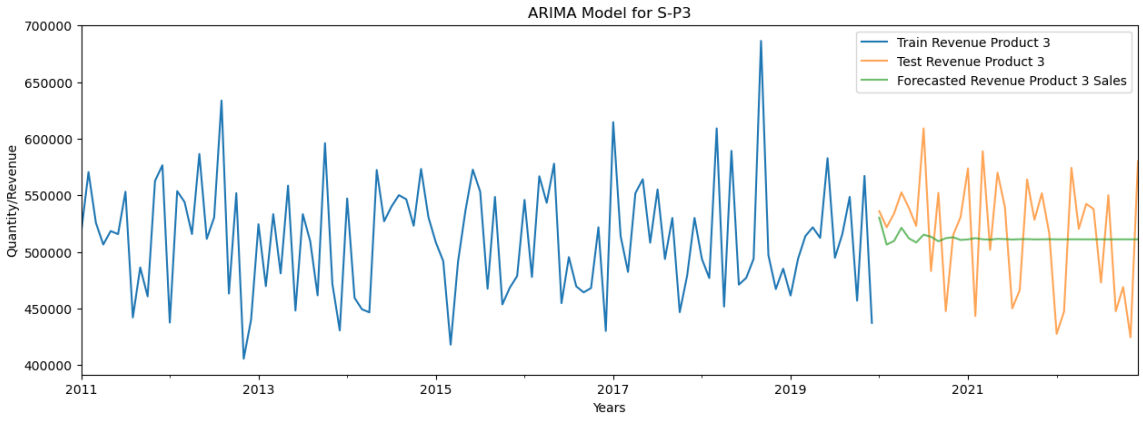


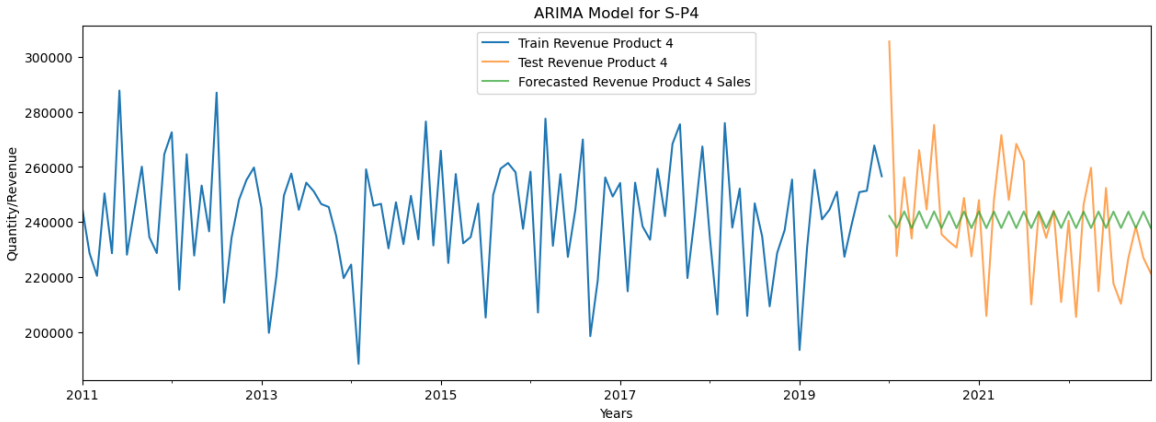




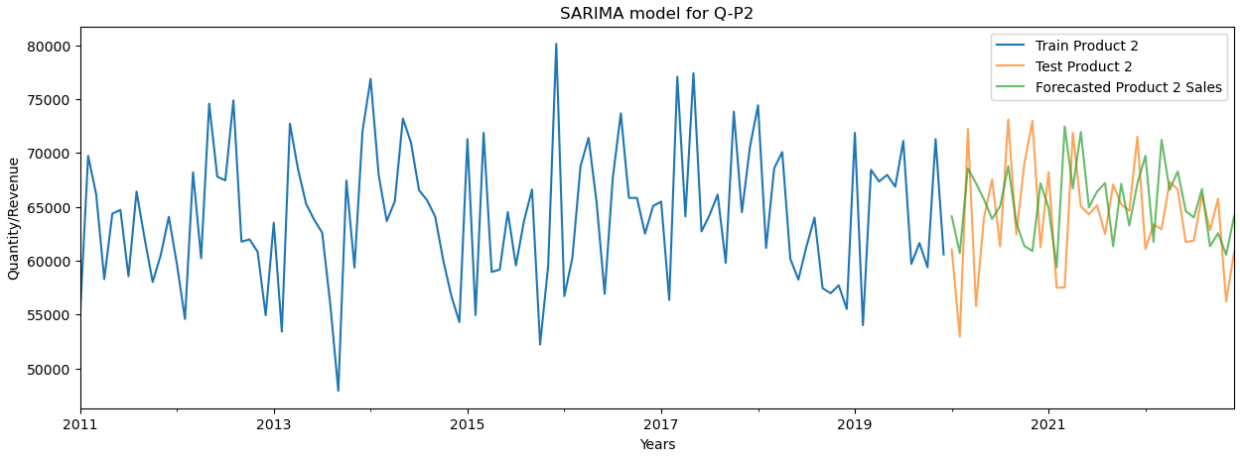

ARIMA Model for S-P1
Train Revenue Product 1
Test Revenue Product 1
Forecasted Revenue Product 1 Sales
Quantity/Revenue
Years

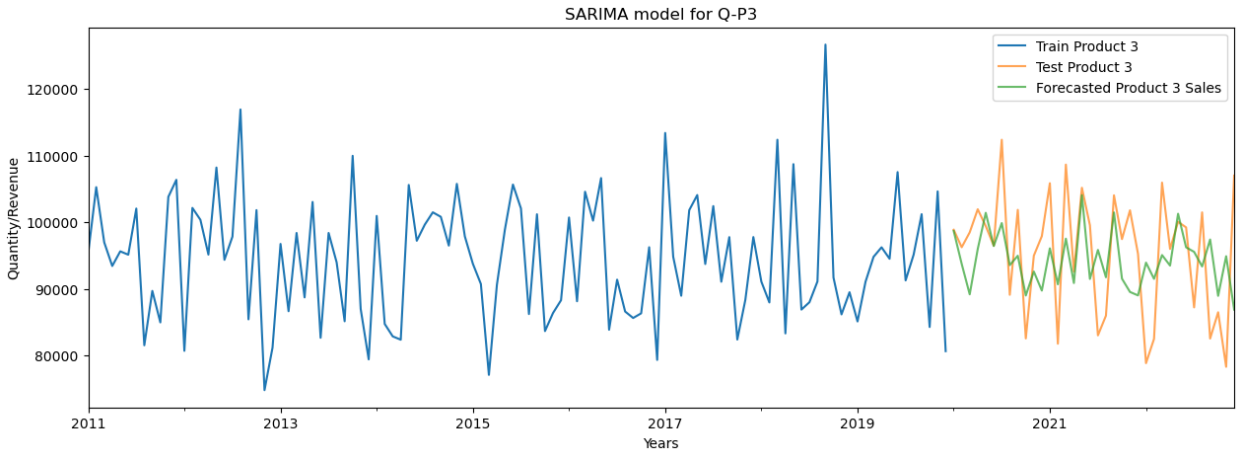

ARIMA Model for S-P2
Train Revenue Product 2
Test Revenue Product 2
Forecasted Revenue Product 2 Sales
Quantity/Revenue
Years

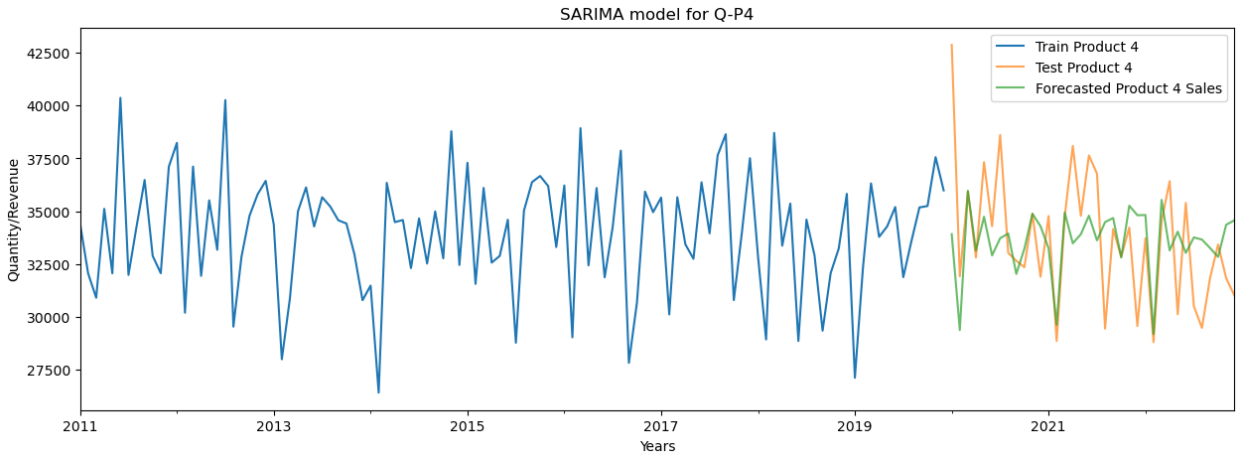

ARIMA Model for S-P3
Train Revenue Product 3
Test Revenue Product 3
Forecasted Revenue Product 3 Sales
Quantity/Revenue
Years

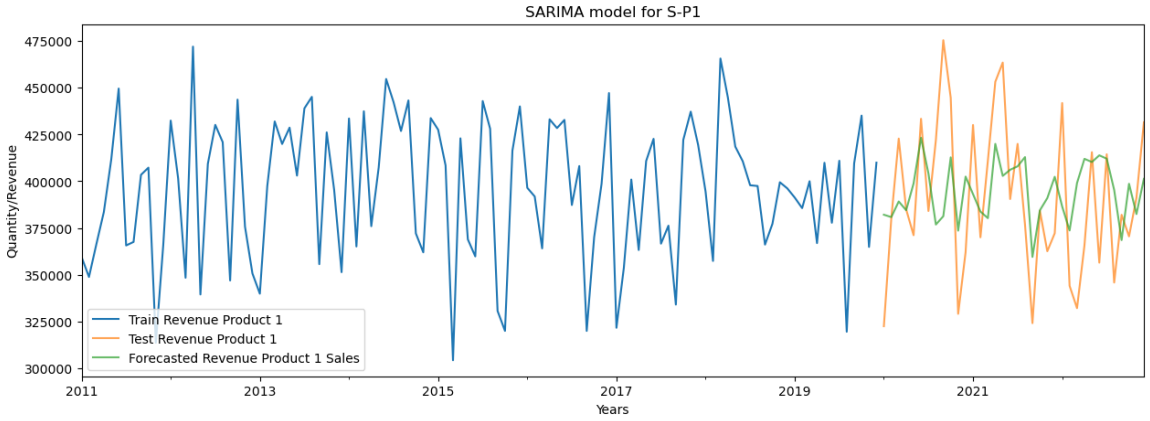

ARIMA Model for S-P4
Train Revenue Product 4
Test Revenue Product 4
Forecasted Revenue Product 4 Sales
Quantity/Revenue
Years

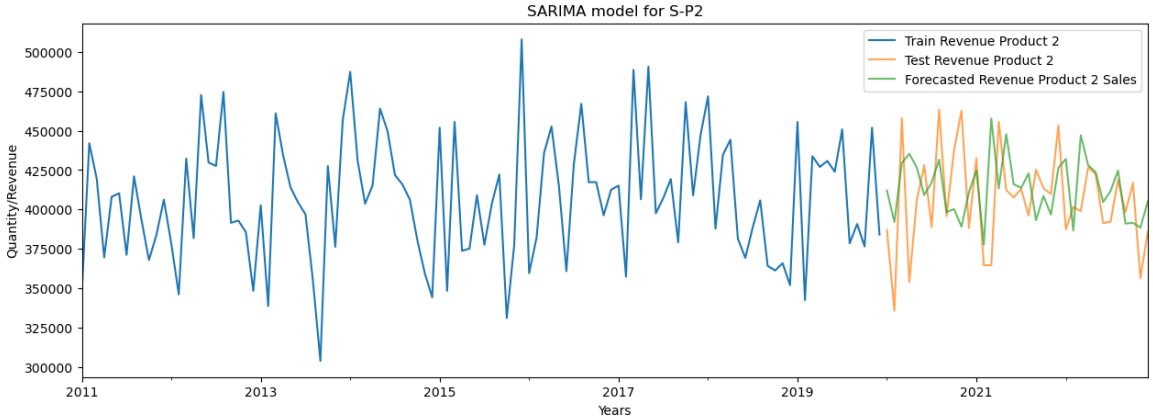
3.9.4.2 Drawing with best SARIMA model for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

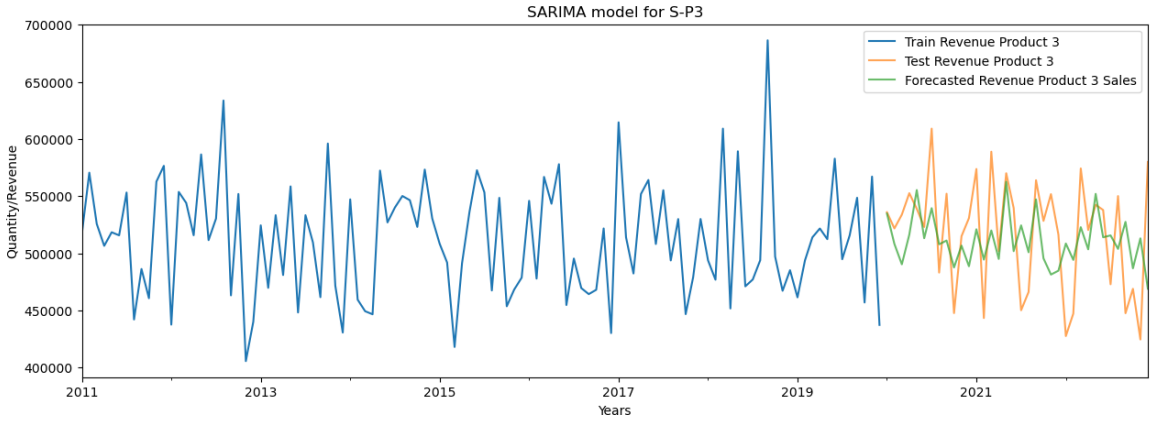


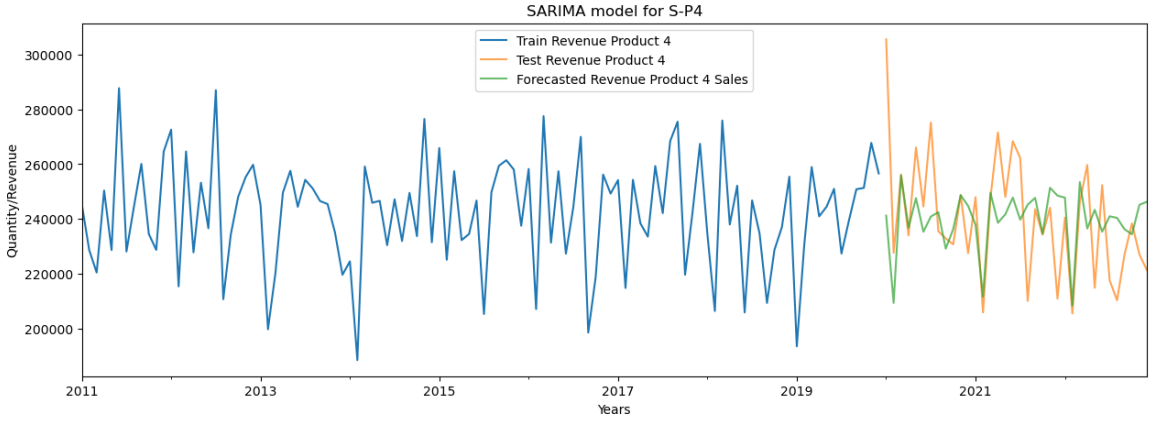





SARIMA model for S-P1
Quantity/Revenue
Years
Train Revenue Product 1
Test Revenue Product 1
Forecasted Revenue Product 1 Sales


SARIMA model for S-P2
Quantity/Revenue
Years
Train Revenue Product 2
Test Revenue Product 2
Forecasted Revenue Product 2 Sales


SARIMA model for S-P3
Quantity/Revenue
Years
Train Revenue Product 3
Test Revenue Product 3
Forecasted Revenue Product 3 Sales


SARIMA model for S-P4
Quantity/Revenue
Years
Train Revenue Product 4
Test Revenue Product 4
Forecasted Revenue Product 4 Sales

3.9.5 Conclusion for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

	RMSE
Best AR Model Product 2 : AR(1,0,0)	4831.925740
Best ARMA Model Product 2: ARMA (2, 0, 2)	4930.656694
Best ARIMA Model Product 2: ARMA (2, 0, 2)	4930.656694
Best SARIMA Model product 2: SARIMA (3, 1, 3) x (3, 1, 3, 12)	5529.360298

The best model for product 2 is AR (1,0,0) that its RMSE is 4831.925740.

	RMSE
Best AR Model Product 3: AR (1,0,0)	9097.457688
Best ARMA Model Product 3: ARMA (3, 0, 3)	9034.622200
Best ARIMA Model Product 3: ARMA (3, 0, 3)	9034.622200
Best SARIMA Model product 3: SARIMA (3, 1, 3) x (3, 1, 3, 12)	8739.050509

The best model for product 3 is SARIMAX (3, 1, 3) x (3, 1, 3, 12) that its RMSE is 8739.050509.

	RMSE
Best AR Model Product 4: AR (1,0,0)	3080.921856
Best ARMA Model Product 4: ARMA (1, 0, 1)	3061.521141
Best ARIMA Model Product 4: ARMA (2, 0, 1)	2954.061145
Best SARIMA Model product 4: SARIMA (3, 1, 3) x (3, 1, 3, 12)	2920.618921

The best model for product 4 is SARIMAX (3, 1, 3) x (3, 1, 3, 12) that its RMSE is 2920.618921.

	RMSE
Best AR Model Revenue Product 1: AR (2,0,0)	40176.861072
Best ARMA Model Revenue Product 1: ARMA (1, 0, 1)	40052.637944
Best ARIMA Model Revenue Product 1: ARMA (1, 0, 1)	40052.637944
Best SARIMA Model Revenue product 1: SARIMA (3, 1, 3) x (3, 1, 3, 12)	37954.147761

The best model for Revenue product 1 is SARIMAX (3, 1, 3) x (3, 1, 3, 12) that its RMSE is 37954.147761.

	RMSE
Best AR Model Revenue Product 2: AR (1,0,0)	30634.409195
Best ARMA Model Revenue Product 2: ARMA (2, 0, 2)	31293.444292
Best ARIMA Model Revenue Product 2: ARMA (2, 0, 2)	31293.444292
Best SARIMA Model Revenue product 2: SARIMA (3, 1, 3) x (3, 1, 3, 12)	34607.294546

The best model for Revenue product 2 is AR (1,0,0) that its RMSE is 30634.409195.

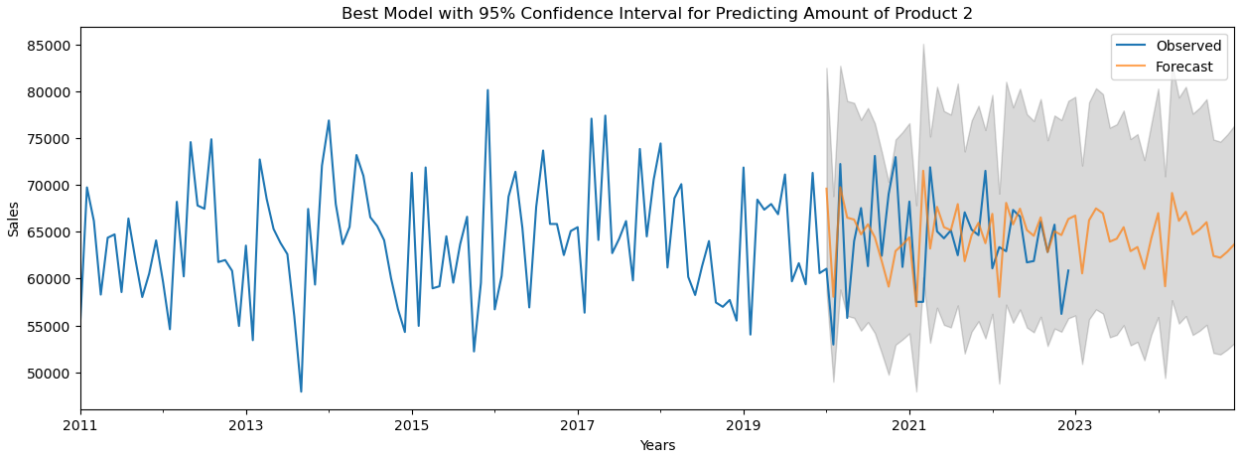
	RMSE
Best AR Model Revenue Product 3: AR (1,0,0)	49308.220668
Best ARMA Model Revenue Product 3: ARMA (3, 0, 3)	48968.143077
Best ARIMA Model Revenue Product 3: ARMA (3, 0, 3)	48968.143077
Best SARIMA Model Revenue product 3: SARIMA (3, 1, 3) x (3, 1, 3, 12)	48102.761913

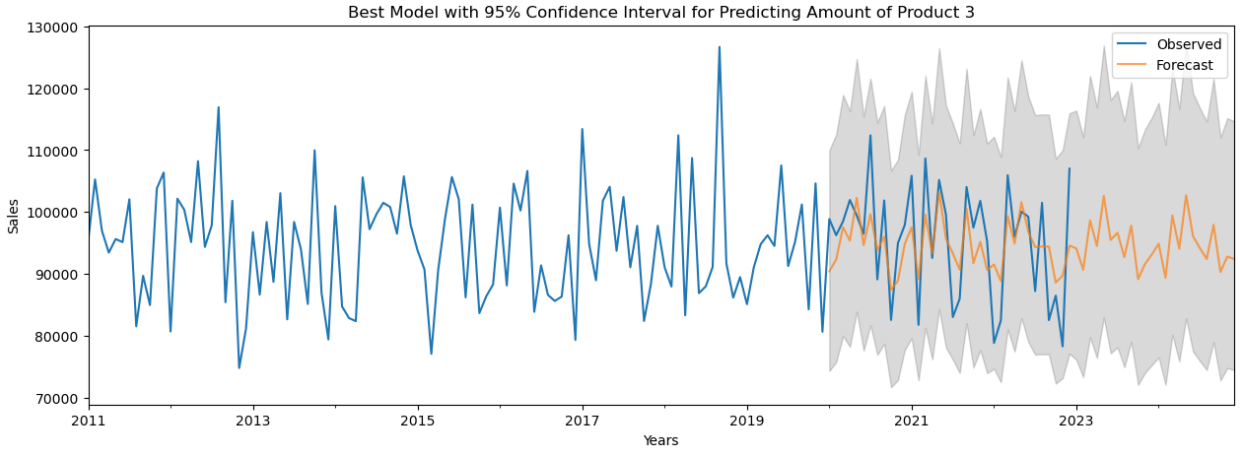
The best model for Revenue product 3 is SARIMAX (3, 1, 3) x (3, 1, 3, 12) that its RMSE is 48102.761913.

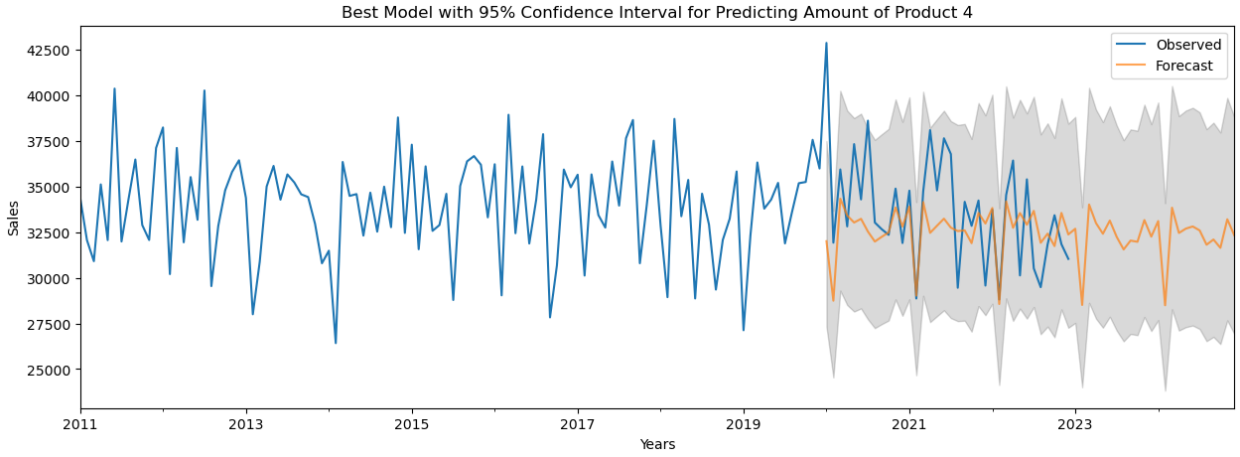
	RMSE
Best AR Model Revenue Product 4: AR (1,0,0)	21966.972834
Best ARMA Model Revenue Product 4: ARMA (2, 0, 1)	21090.605765
Best ARIMA Model Revenue Product 4: ARMA (2, 0, 1)	21090.605765
Best SARIMA Model Revenue product 4: SARIMA (3, 1, 3) x (3, 1, 3, 12)	20879.880150

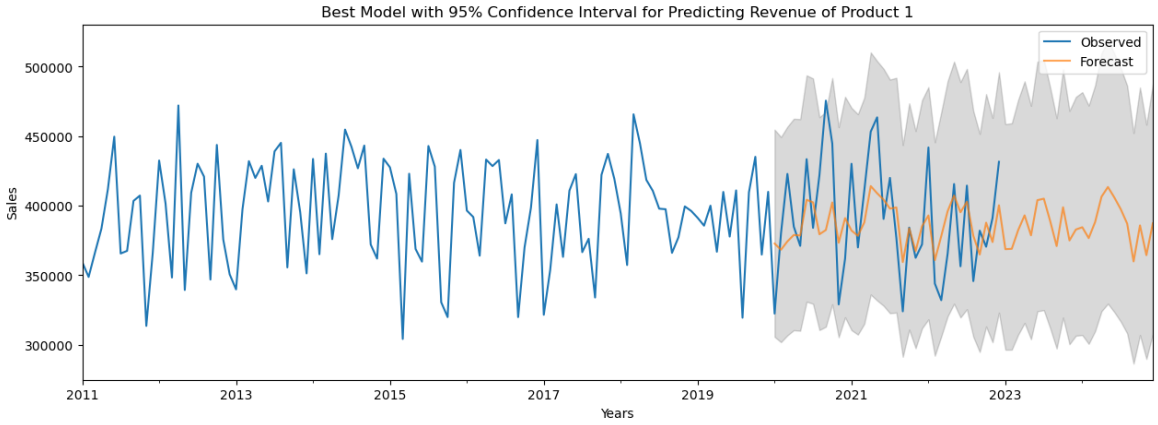
The best model for Revenue product 4 is SARIMAX (3, 1, 3) x (3, 1, 3, 12) that its RMSE is 20879.880150.

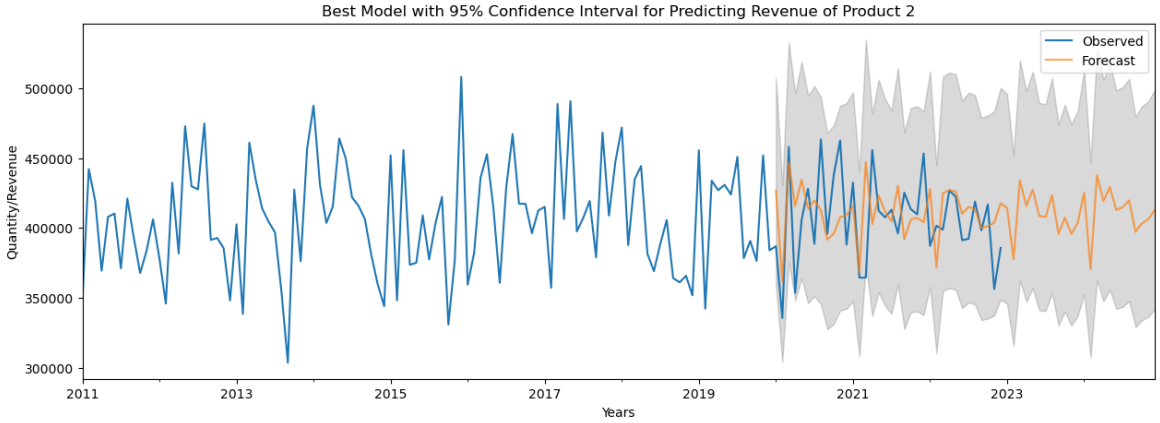
3.9.5.1 Forecast Quantity and Revenue with the best Model along with the 95% Confidence interval for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

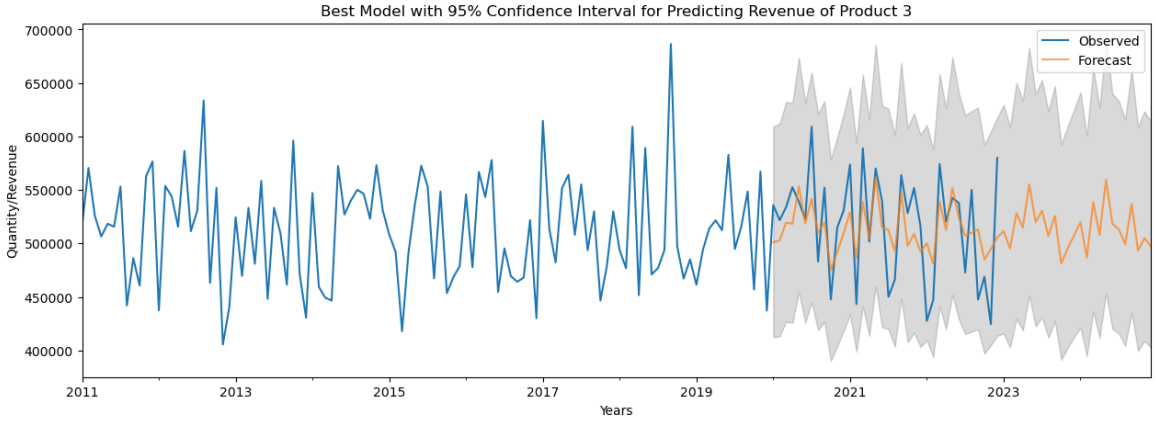


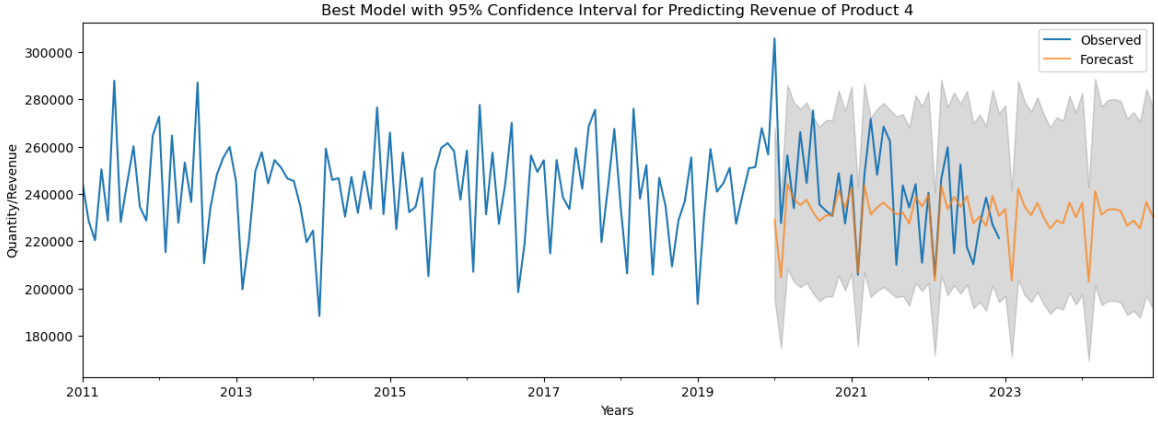





Best Model with 95% Confidence Interval for Predicting Revenue of Product 1
Observed
Forecast
Sales
Years


Best Model with 95% Confidence Interval for Predicting Revenue of Product 2
Observed
Forecast
Quantity/Revenue
Years

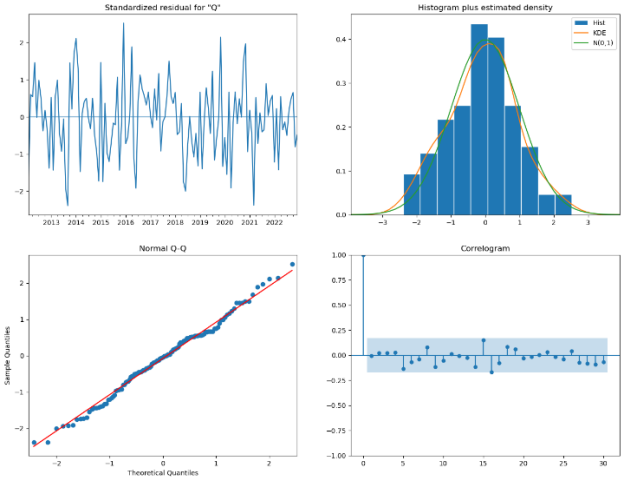

Best Model with 95% Confidence Interval for Predicting Revenue of Product 3
Observed
Forecast
Quantity/Revenue
Years


Best Model with 95% Confidence Interval for Predicting Revenue of Product 4
Observed
Forecast
Quantity/Revenue
Years

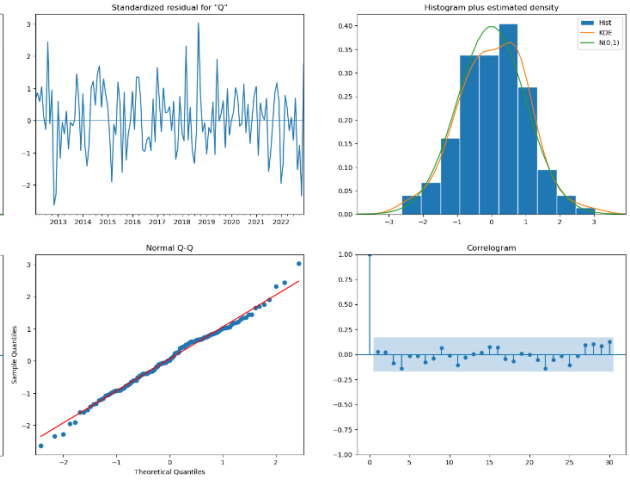
Finally, we can see the 5th question of the CEO. As we can use the best model of sales and revenue of product 1, 2, 3, and 4. And we can forecast them for 2023, 2024 and etc.

3.9.5.2 Other Conclusions for Q-P2, Q-P3, Q-P4, S-P1, S-P2, S-P3, S-P4

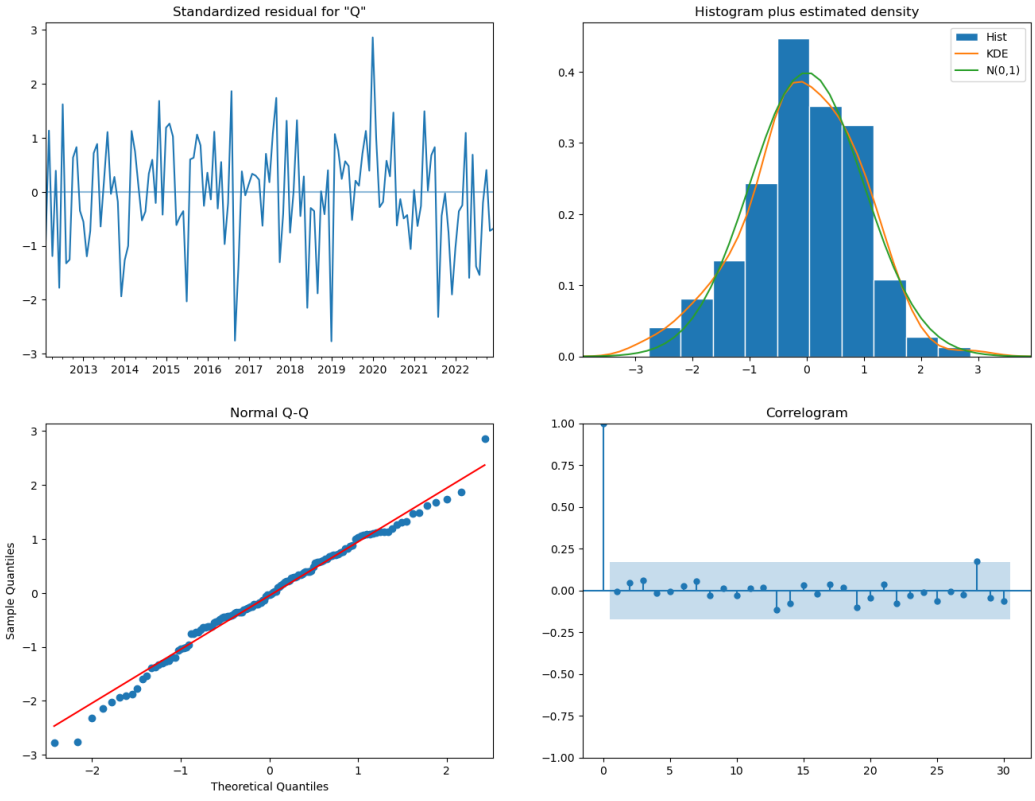
Other Conclusions for Q-P2 (The quantity of product 2)

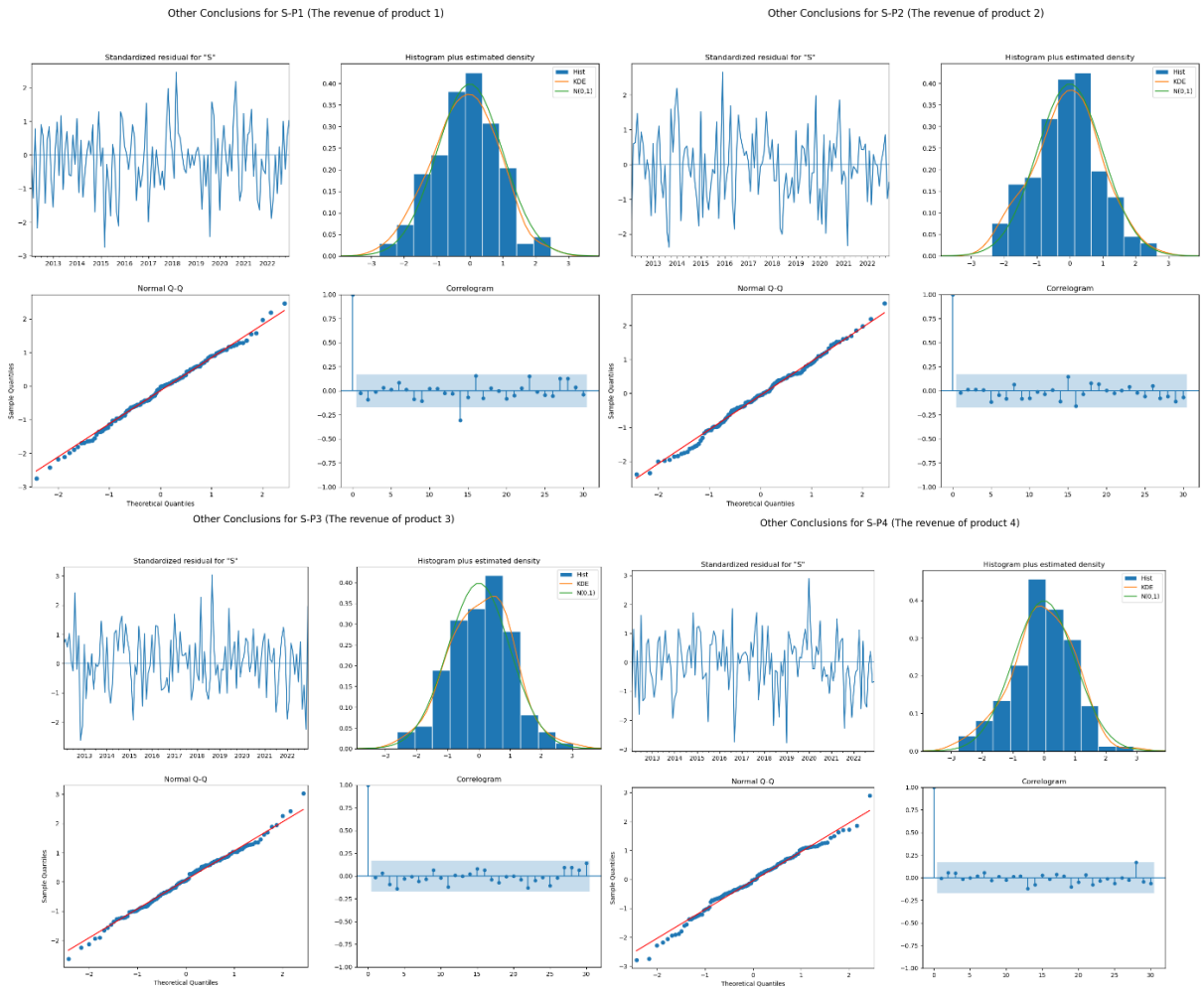


Other Conclusions for Q-P3 (The quantity of product 3)



Other Conclusions for Q-P4 (The quantity of product 4)





.

Inference Note: 4 plots in the residuals diagnostic plots tell us:

- Standardized residuals plot the top left plot shows 1-step-ahead standardized residuals. If model is working correctly, then no pattern should be obvious in the residuals which is clearly not visible from the plot as well.
- Histogram plus estimated density plot This plot shows the distribution of the residuals. The orange line shows a smoothed version of this histogram, and the green line shows a normal distribution. If the model is good these two lines should be the same. Here there are small differences between them, which indicate that our model is doing just well enough.
- Normal Q-Q plot the Q-Q plot compare the distribution of residuals to normal distribution. If the distribution of the residuals is normal, then all the points should lie along the red line, except for some values at the end, which is exactly happening in this case.
- Correlogram plot the correlogram plot is the ACF plot of the residuals rather than the data. 95% of the correlations for lag >0 should not be significant (within the blue shades). If there is a significant correlation in the residuals, it means that there is information in the data that was not captured by the model, which is clearly not in this case.