TIME SERIES FORECASTING BUSINESS REPORT

Rose Wine Analysis & Sparkling Wine Analysis

MAY 4, 2025

HIYA SHAH
PGP DSBA

Table of Contents

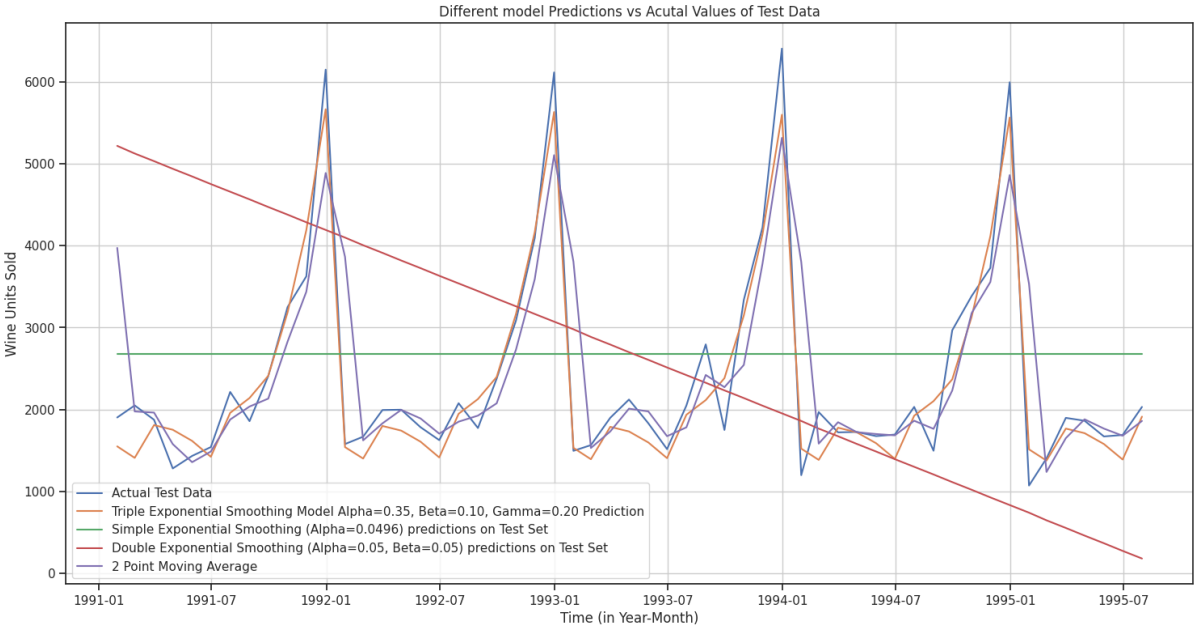
```
First few rows of Training Data
             Rose_Wine_Sales  Time
Time_Stamp
1980-01-31            112.0     1
1980-02-29            118.0     2
1980-03-31            129.0     3
1980-04-30             99.0     4
1980-05-31            116.0     5

Last few rows of Training Data
             Rose_Wine_Sales  Time
Time_Stamp
1990-08-31             70.0   128
1990-09-30             83.0   129
1990-10-31             65.0   130
1990-11-30            110.0   131
1990-12-31            132.0   132

First few rows of Test Data
             Rose_Wine_Sales  Time
Time_Stamp
1991-01-31             54.0   133
1991-02-28             55.0   134
1991-03-31             66.0   135
1991-04-30             65.0   136
1991-05-31             60.0   137

Last few rows of Test Data
             Rose_Wine_Sales  Time
Time_Stamp
1995-03-31             45.0   183
1995-04-30             52.0   184
1995-05-31             28.0   185
1995-06-30             40.0   186
1995-07-31             62.0   187
```


5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-



EXECUTIVE SUMMARY

This report presents a comprehensive analysis and forecast of wine sales trends throughout the 20th century based on historical data provided by ABC Estate Wines. The primary objective is to uncover actionable insights and develop predictive models that will support strategic decision-making and enhance future sales performance.

By employing robust data analytics techniques, we have examined historical sales patterns across various wine types. The analysis reveals key trends, seasonality, and external factors that have influenced wine consumption over the decades. Leveraging time-series forecasting methods, the report provides accurate predictions of future wine sales trajectories, enabling ABC Estate Wines to proactively align their production, marketing, and distribution strategies.

This data-driven approach positions ABC Estate Wines to identify emerging opportunities, address potential challenges, and maintain a competitive edge in an evolving market landscape. The insights derived will serve as a foundation for optimizing inventory management, market segmentation, and targeted promotions to drive sustained growth.

INTRODUCTION

ABC Estate Wines, a renowned name in wine production, has accumulated a rich dataset of wine sales spanning the 20th century. These records, detailing the performance of various wine types, offer a unique opportunity to analyze historical sales dynamics and forecast future trends.

In today's competitive and data-driven marketplace, understanding past performance is crucial for crafting strategies that resonate with consumer preferences and market conditions. This project aims to harness the power of data analytics to uncover hidden patterns, assess the impact of temporal and economic factors on wine sales, and provide forward-looking projections.

Through this report, ABC Estate Wines will gain critical insights into how consumer behavior has evolved over the century, which wine categories have demonstrated consistent growth, and how seasonal and macroeconomic elements influence demand. The findings will empower the company to refine its business strategies, adapt to shifting market trends, and ensure continued success in the wine industry.

Rose Wine Analysis

Sample of the dataset

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

Sample of first 5 rows of the dataset

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Sample of last 5 rows of the dataset

Dataset has 2 columns which captures the Year and Month of recorded data and the number of units sold on corresponding Year-Month respectively**.**

Read the data as an appropriate Time Series data and plot the data.

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='ME')
```

	YearMonth	Rose	Time_Stamp
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

Time Stamp created from 'YearMonth' column

The dataset has 2 variables and 187 rows in total. The "YearMonth" column can be deleted after creating a suitable time stamp column because it is not necessary for our modelling. The column Rose is of float type. Additionally, we can observe from the data above that Rose column has some missing values which needs to be imputed further as it's a time series.

Resulting dataset after removing the “Year-Month” column and appending Time_Stamp column

Time_Stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Time_Stamp	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Time_Stamp column has been set as index of the dataset and column Rose has been renamed as Rose_Wine_Sales.

Renaming the column

Time_Stamp	Rose_Wine_Sales
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

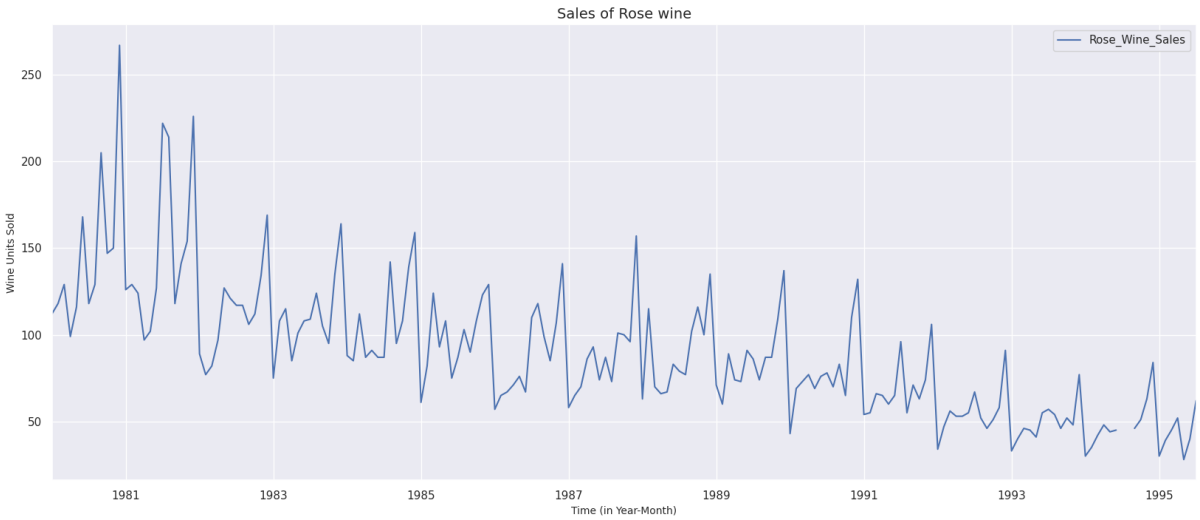
Time_Stamp column has been set as index of the dataset and column Rose has been renamed as Rose_Wine_Sales.

Checking null values in the dataset

Time_Stamp	Rose_Wine_Sales
1994-07-31	NaN
1994-08-31	NaN

Since its a time series, we cannot remove the null values and hence it must be imputed.

Plot the Time Series to understand the behaviour of the data.

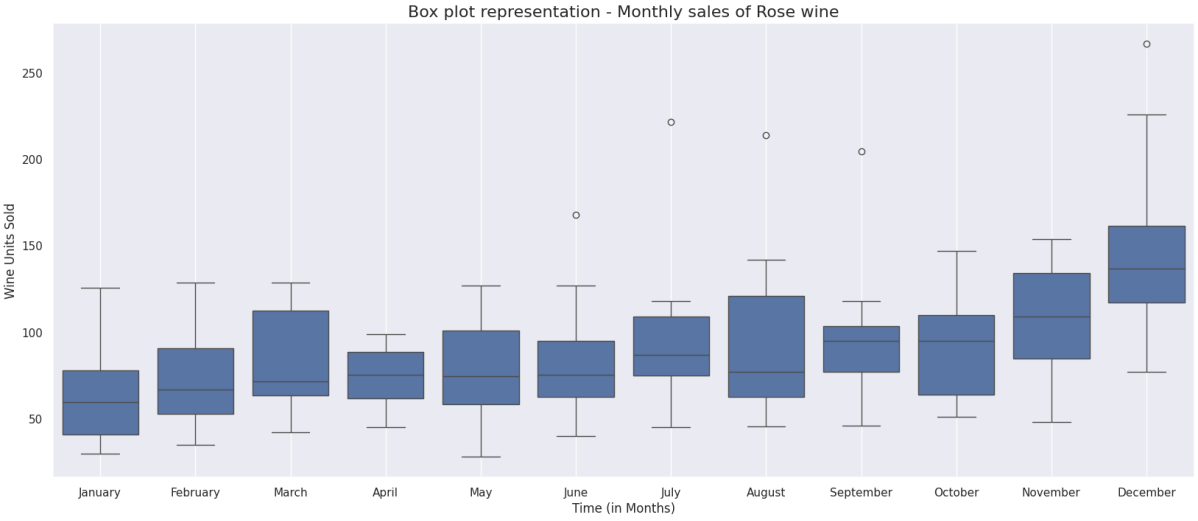


INSIGHTS:

1. **The average sales of Rose wine is around 90 units**
2. **The minimuim quantity of sales recorded is 28 and the max is 267 units**
3. **Around 50% of the sales recorded in a month are around 85 units**
4. **Only 25% of the sales recorded in a month are above 111 units**

The spread of sales across different years and within different months across years:

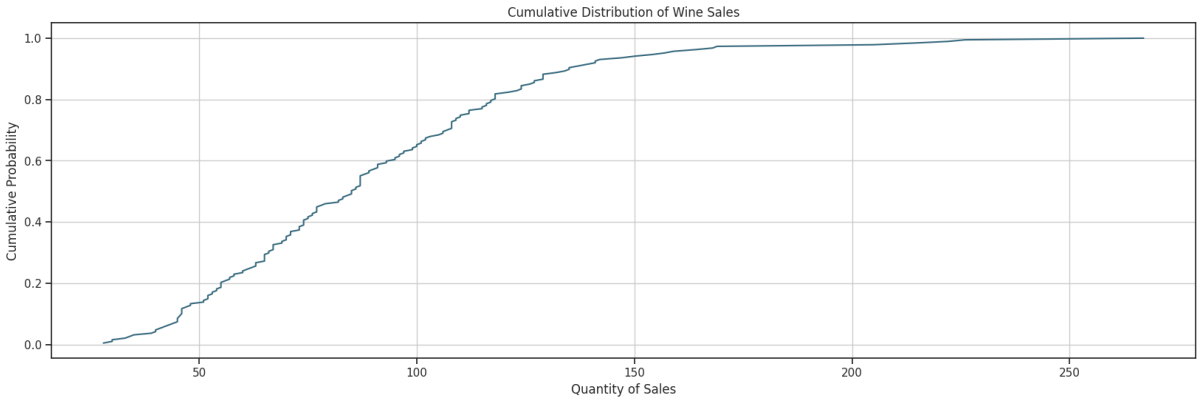
1. **The sales of rose wine have gradually declined over the years**
2. **At the initial years, the sales were high with maximum sales happening in 1981.**



1. **December month records the highest average sales**
2. **The sales seems to usually pick in the last 4 months**

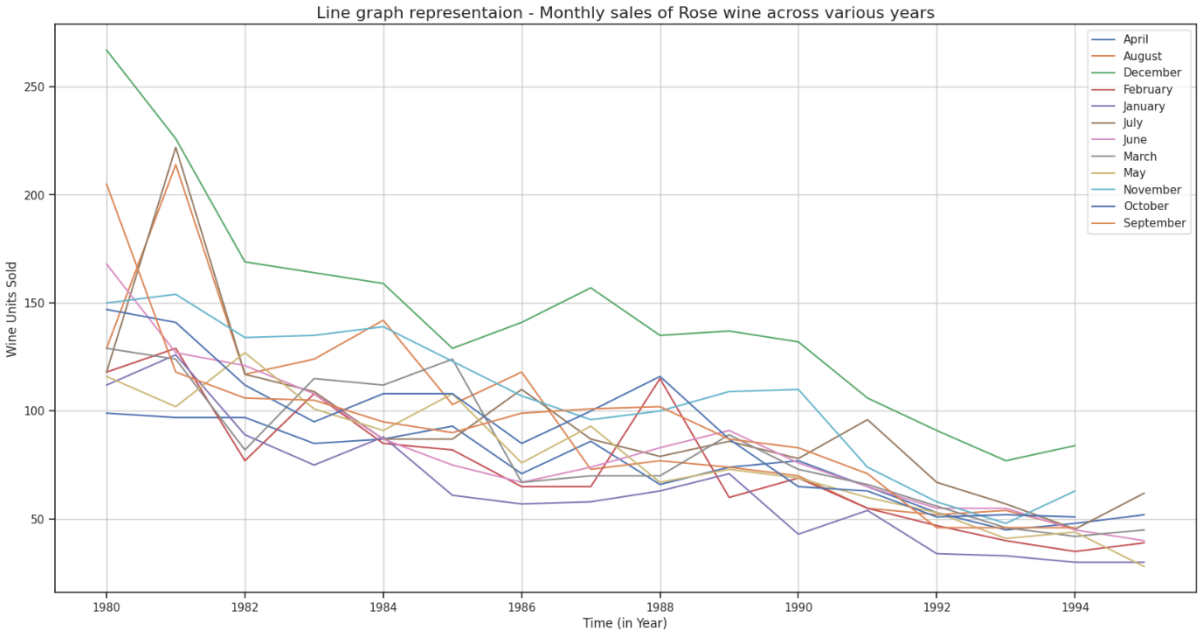
Average order is greater in december and lowest in january

Plot the empirical cumulative Distribution function



1. **Around 70 to 90% of the orders are within 100 to 150 units.**

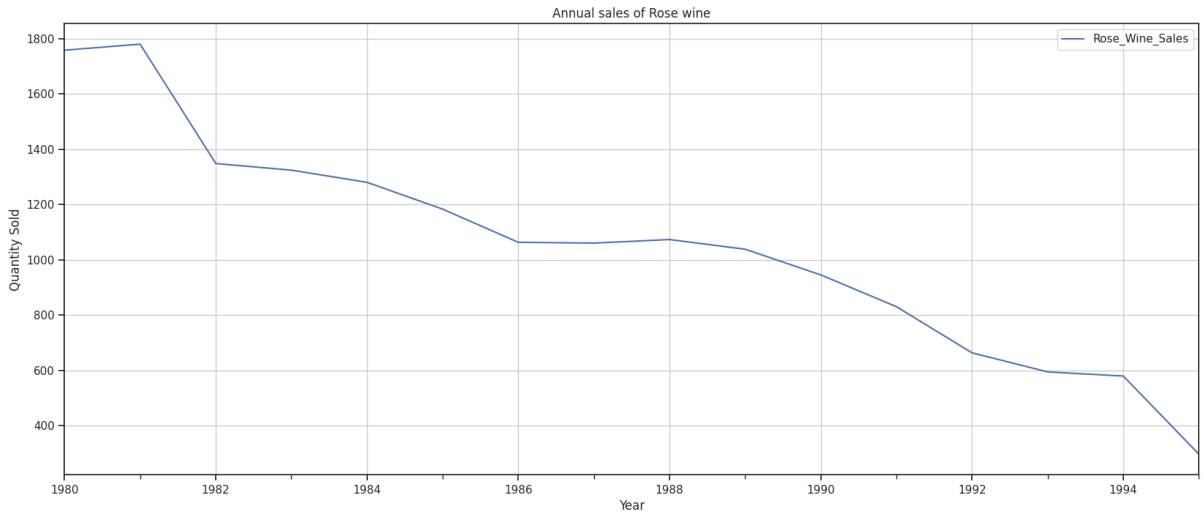
Monthly sales across years



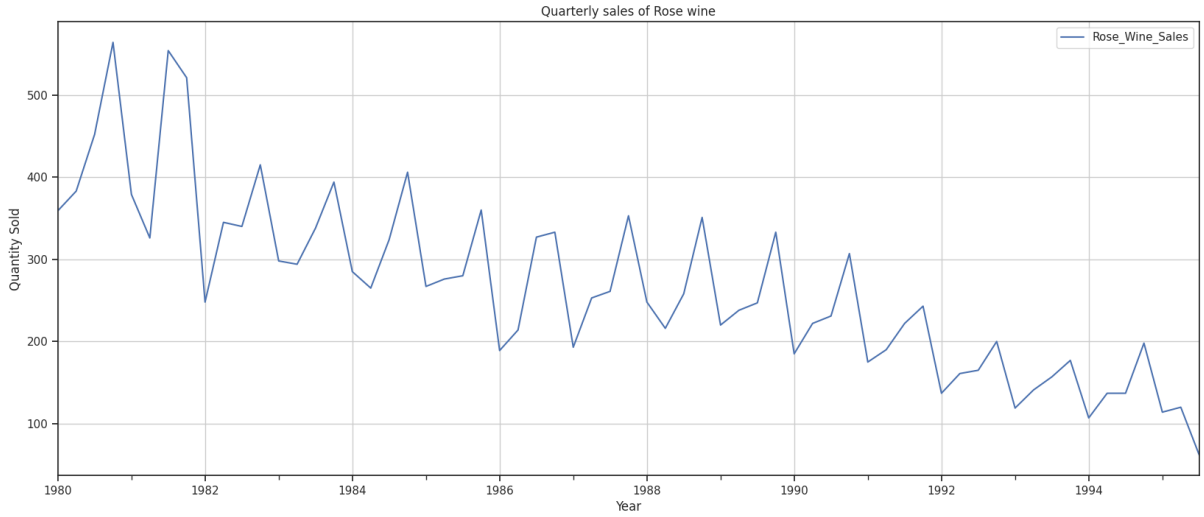
Observation:

- **The sales trajectory appears to be precisely the reverse of that seen in the yearly plot, increasing near the end of each year.**
- **January has the lowest wine sales while December sees the greatest. The sales modestly grow from January to August and then sharply climb after that.**
- **Additionally, we can see that there are outliers in the box plots.**

Annual Sales:

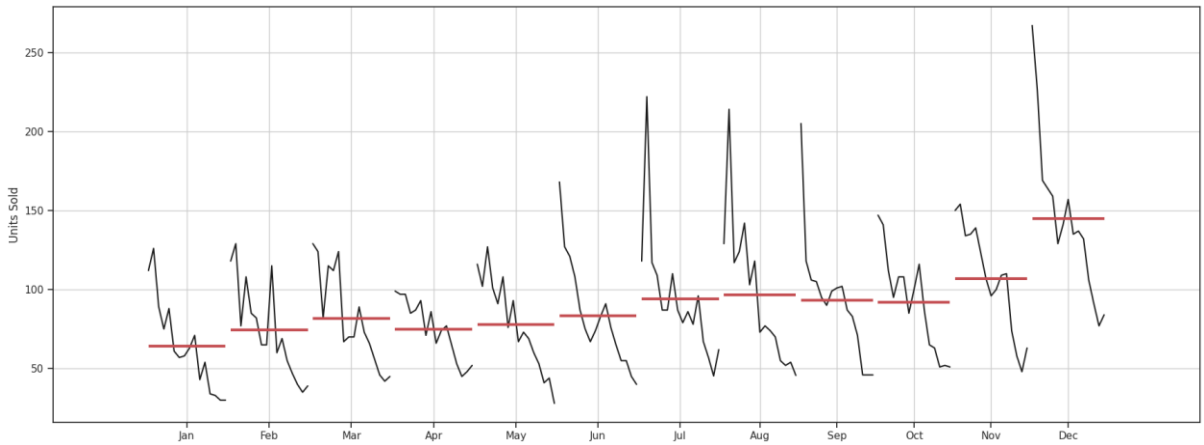


Quarterly Sales:



The sales are incresingly high in Q3 and Q4

Month plot of times series



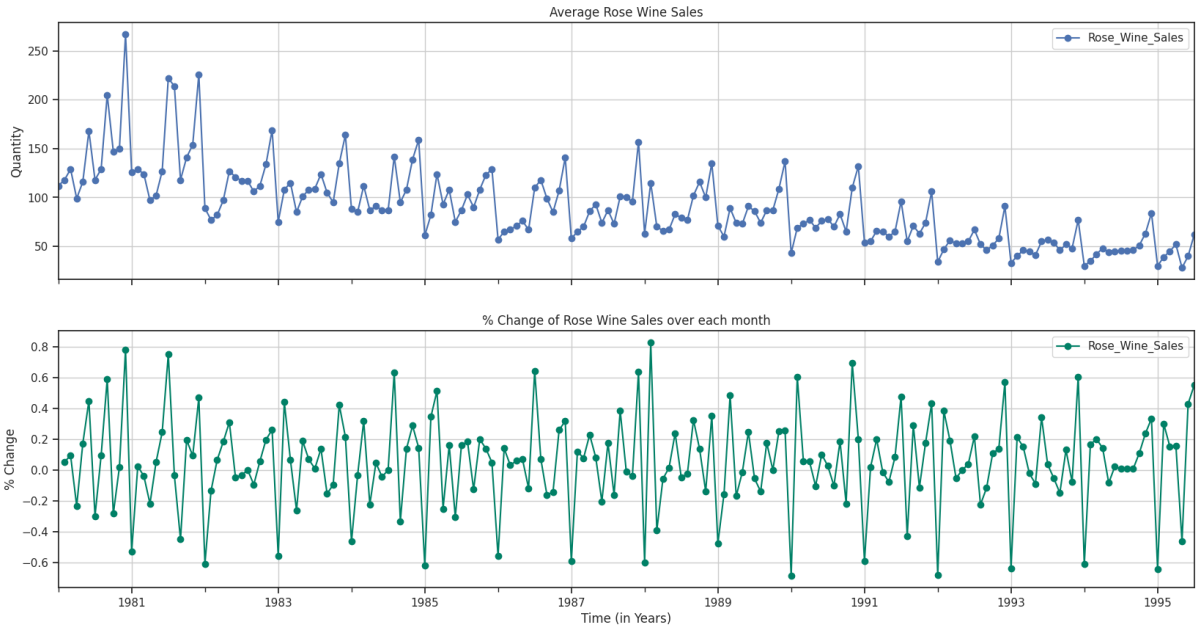
Observation:

• After 1981, the sales fell drastically. Sales are typically lowest in the first quarter and highest in the fourth quarter.

• Every year, December has the highest sales, followed by November and October. January had the lowest sales.

• From the cumulative distribution graph, we can observe that around 70 to 75 percent of the units sold are fewer than 100, and 90% of the units sold are less than 150. Only 15% of sales involved less than 50 items. Therefore, it is clear that the bulk of sales were in the range of 50 to 100 units.

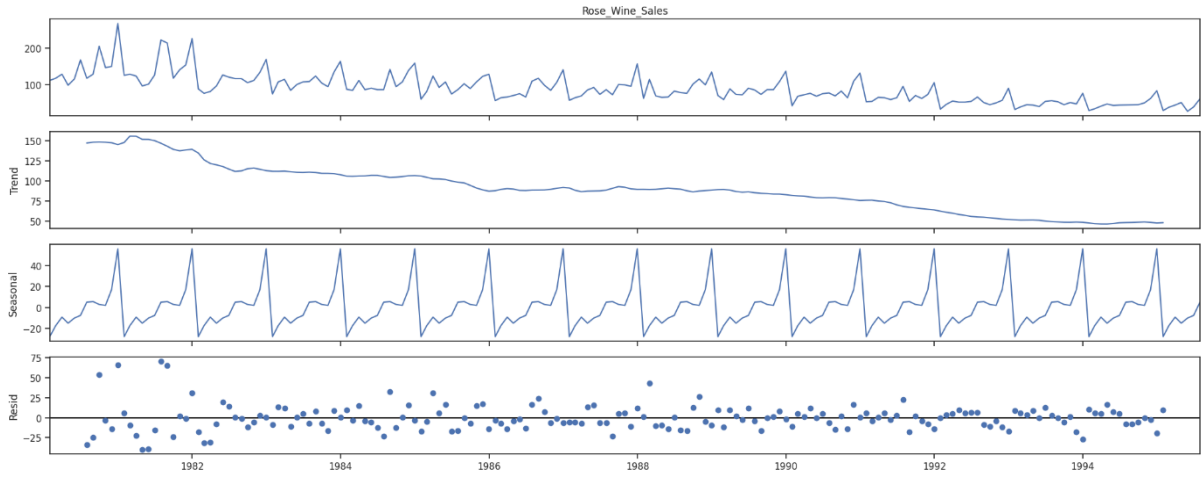
Average Wine sales per month & change percentage over each month



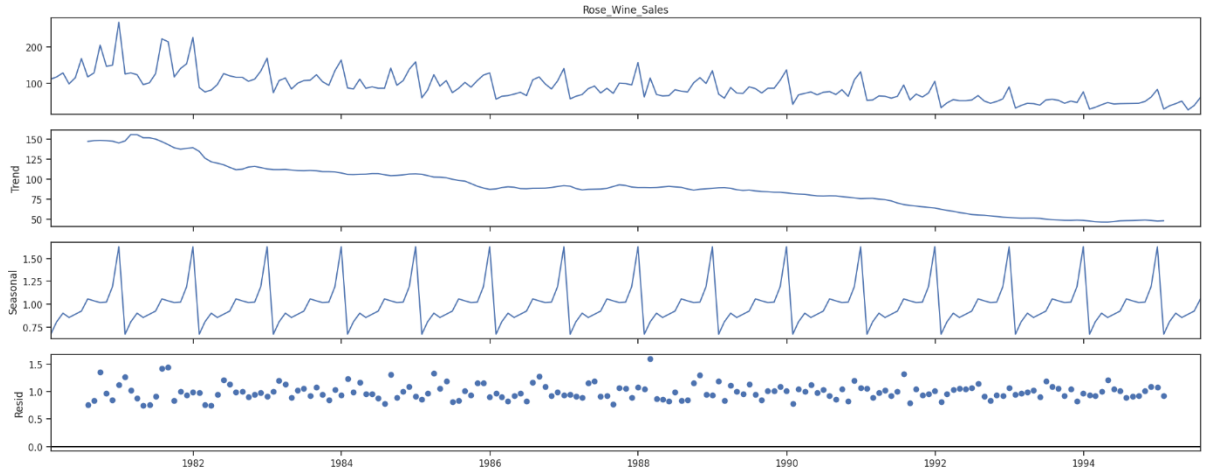
1. **Average Rose wine sales graph shows us a downward trend along with yearly seasonality present in it**
2. **% change graph shows us the seasonality of the change in sales to be constant throghout the lifetime of sales**

Decomposition of the Time Series

Additive decomposition



Multiplicative Decomposition



Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal. • The residual patterns after additive decomposition of the time series appear to represent the seasonal element and exhibit substantial variation.

• In the multiplicative decomposition of the time series, it has been observed that the seasonal fluctuation of residuals is under control.

• The size of the seasonal variations doesn't change on comparison, but the residuals are tightly controlled by the multiplicative decomposition. In addition to this, the residuals are not independent of seasonality thus we may assume that it is multiplicative.

3) Split the data into training and test. The test data should start in 1991

Train and test data are separated from the provided dataset. Sales data up to 1991 is included in the training data, while data from 1991 through 1995 is used for testing.

First few rows of Training Data

Time_Stamp	Rose_Wine_Sales
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0
1980-06-30	168.0
1980-07-31	118.0
1980-08-31	129.0
1980-09-30	205.0
1980-10-31	147.0

Last few rows of Training Data

Time_Stamp	Rose_Wine_Sales
1990-03-31	73.0
1990-04-30	77.0
1990-05-31	69.0
1990-06-30	76.0
1990-07-31	78.0
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

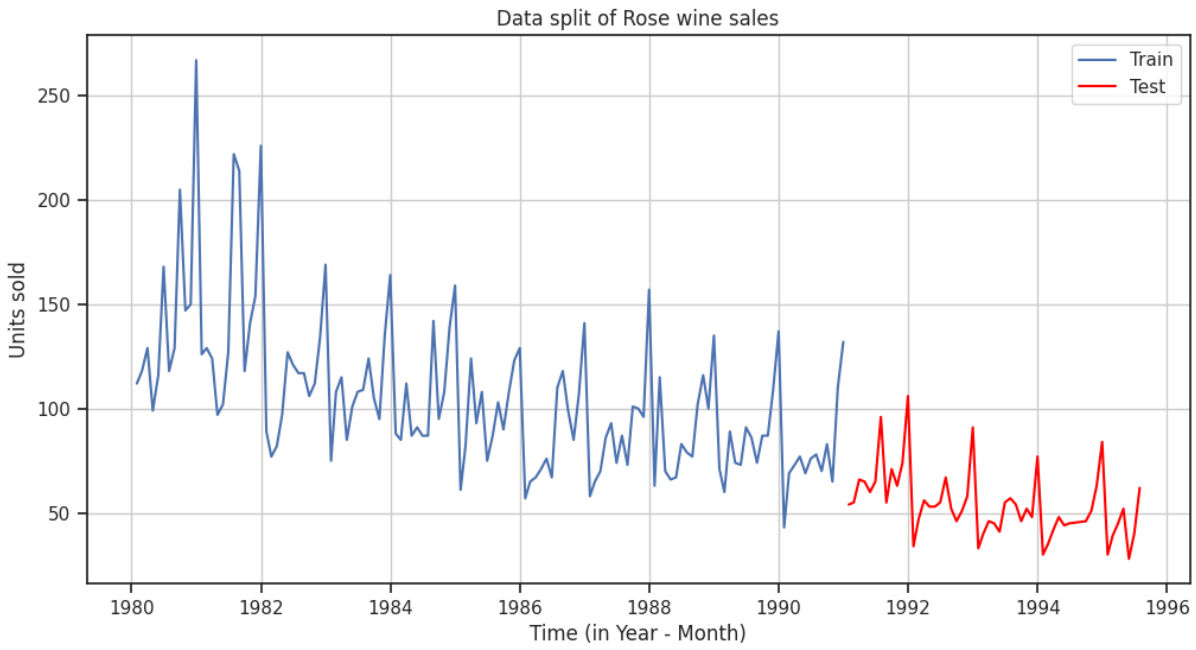
First few rows of Test Data

Time_Stamp	Rose_Wine_Sales
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0
1991-06-30	65.0
1991-07-31	96.0
1991-08-31	55.0
1991-09-30	71.0
1991-10-31	63.0

Last few rows of Test Data

Time_Stamp	Rose_Wine_Sales
1994-10-31	51.0
1994-11-30	63.0
1994-12-31	84.0
1995-01-31	30.0
1995-02-28	39.0
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Line Plot – Splitting of time series into Train & Test data



4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression

First few rows of Training Data

Time_Stamp	Rose_Wine_Sales	Time
1980-01-31	112.0	1
1980-02-29	118.0	2
1980-03-31	129.0	3
1980-04-30	99.0	4
1980-05-31	116.0	5

Last few rows of Training Data

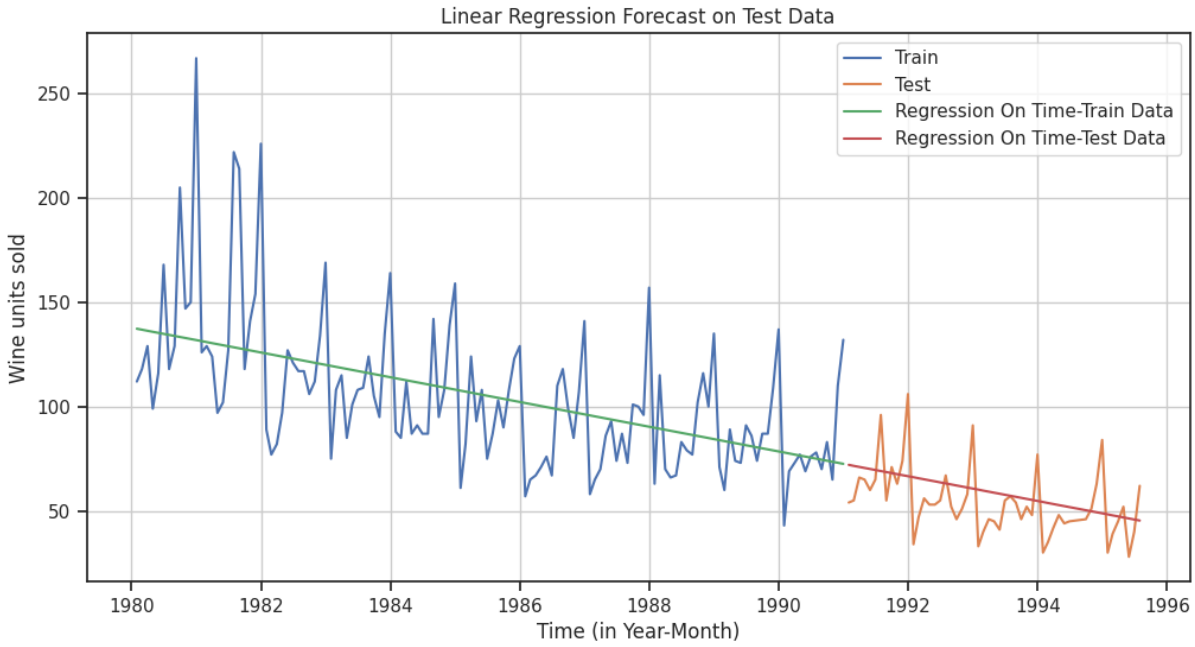
Time_Stamp	Rose_Wine_Sales	Time
1990-08-31	70.0	128
1990-09-30	83.0	129
1990-10-31	65.0	130
1990-11-30	110.0	131
1990-12-31	132.0	132

First few rows of Test Data

Time_Stamp	Rose_Wine_Sales	Time
1991-01-31	54.0	133
1991-02-28	55.0	134
1991-03-31	66.0	135
1991-04-30	65.0	136
1991-05-31	60.0	137

Last few rows of Test Data

Time_Stamp	Rose_Wine_Sales	Time
1995-03-31	45.0	183
1995-04-30	52.0	184
1995-05-31	28.0	185
1995-06-30	40.0	186
1995-07-31	62.0	187



Observation:

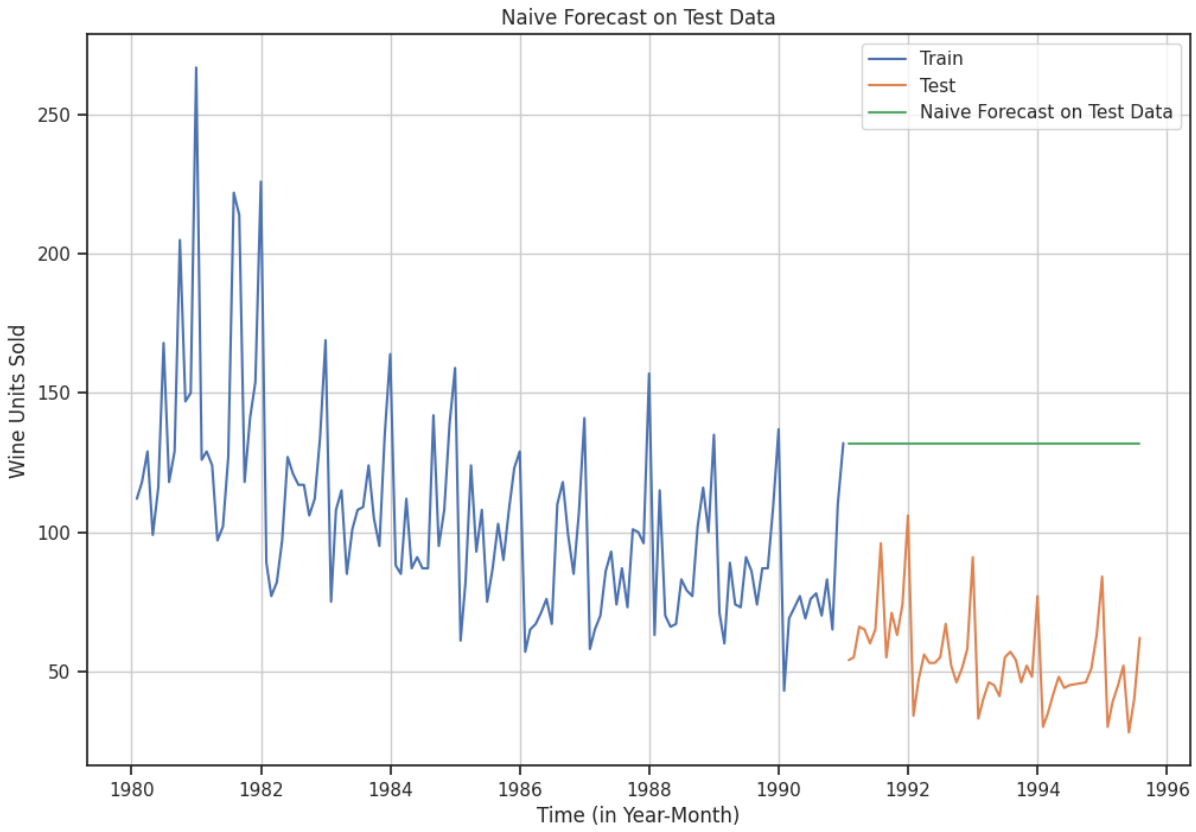
- **We can see from the graphs above that the time series has a falling trend and is seasonal**
- **The train and test data trends have been caught by the linear regression model however, it is unable to account for seasonality**
- **The root means squared error (RMSE) for the linear regression model is 15.268. The size of the seasonal**

Linear Regression: Model Evaluation

	Test RMSE
Linear Regression	15.268887

Model 2: Naive Forecast (y^t+1=yt)

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today,therefore the prediction for day after tomorrow is also today.



Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal

• The seasonality and trend of the time series data cannot be captured by the simple forecast model.

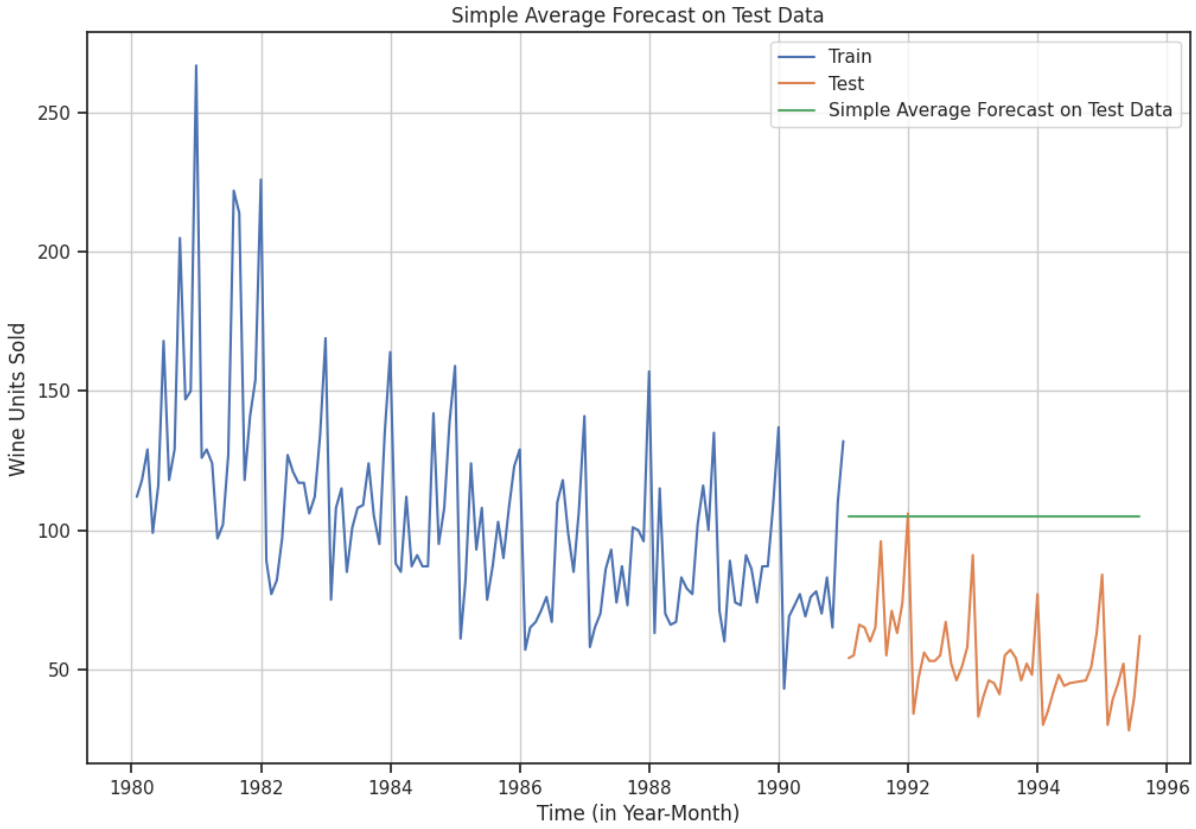
• The root means squared error (RMSE) for the naïve forecast model is 79.719 which is significantly higher than the regression model.

index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689

Method 3: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.

Time_Stamp	Rose_Wine_Sales	mean_forecast
1991-01-31 00:00:00	54.0	104.93939393939394
1991-02-28 00:00:00	55.0	104.93939393939394
1991-03-31 00:00:00	66.0	104.93939393939394
1991-04-30 00:00:00	65.0	104.93939393939394
1991-05-31 00:00:00	60.0	104.93939393939394



Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal

• The seasonality and trend of the time series data cannot be captured by the simple average model.
• The root means squared error (RMSE) for the simple average model is 53.46 which is significantly higher than the regression model but lower than naïve forecast model.

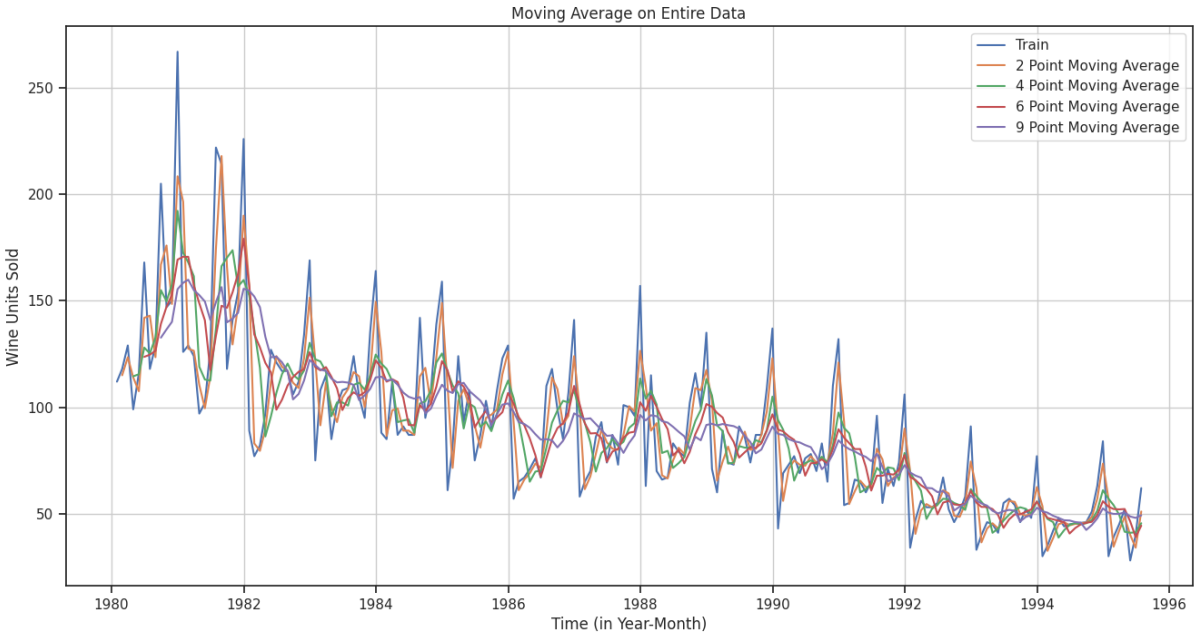
Simple Average: Model Evaluation

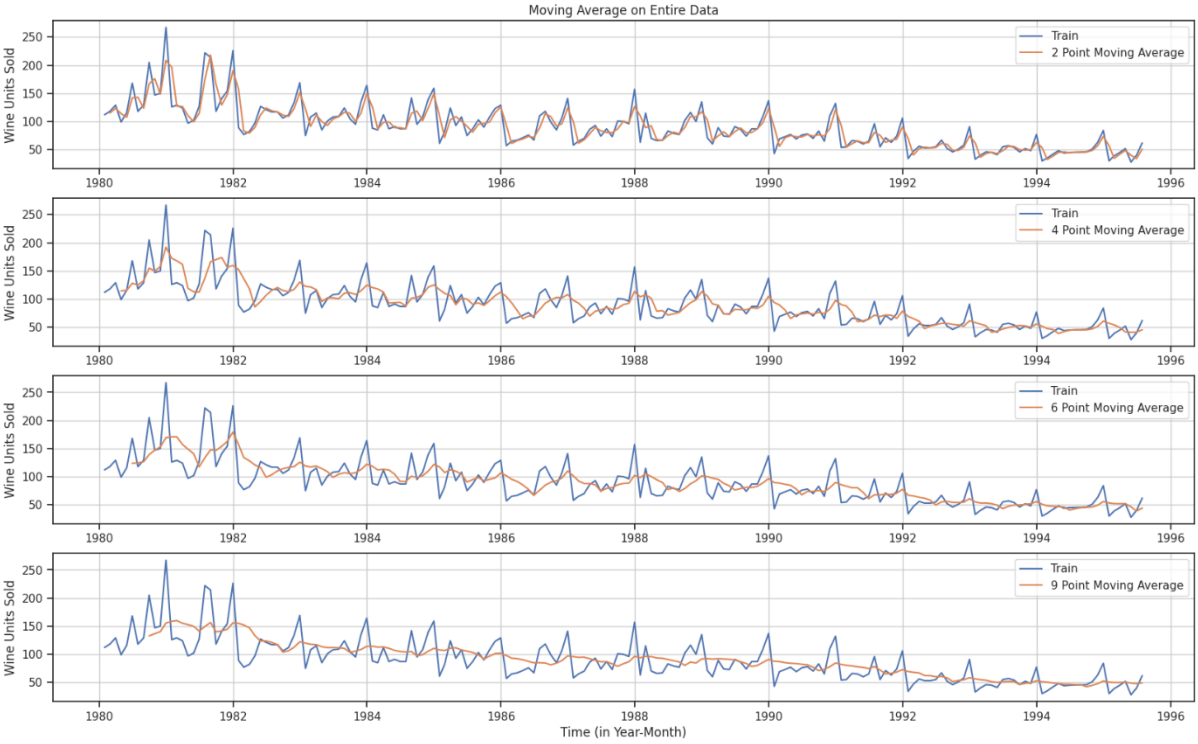
index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689
Simple Average	53.460367314208234

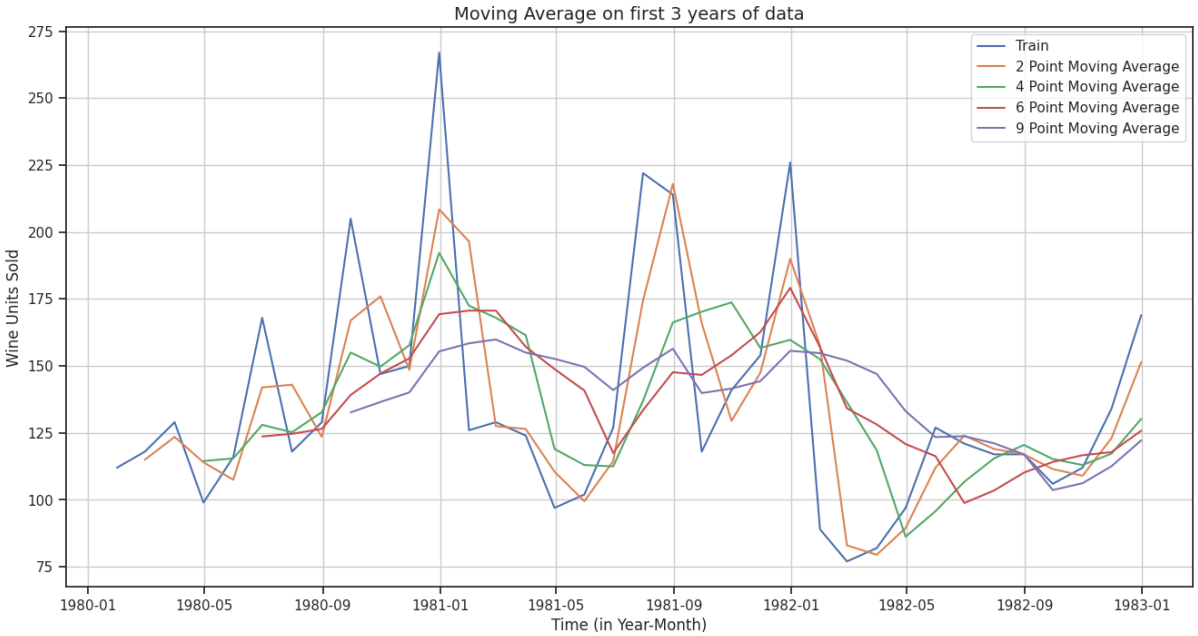
Model 4 – Moving Average (MA)

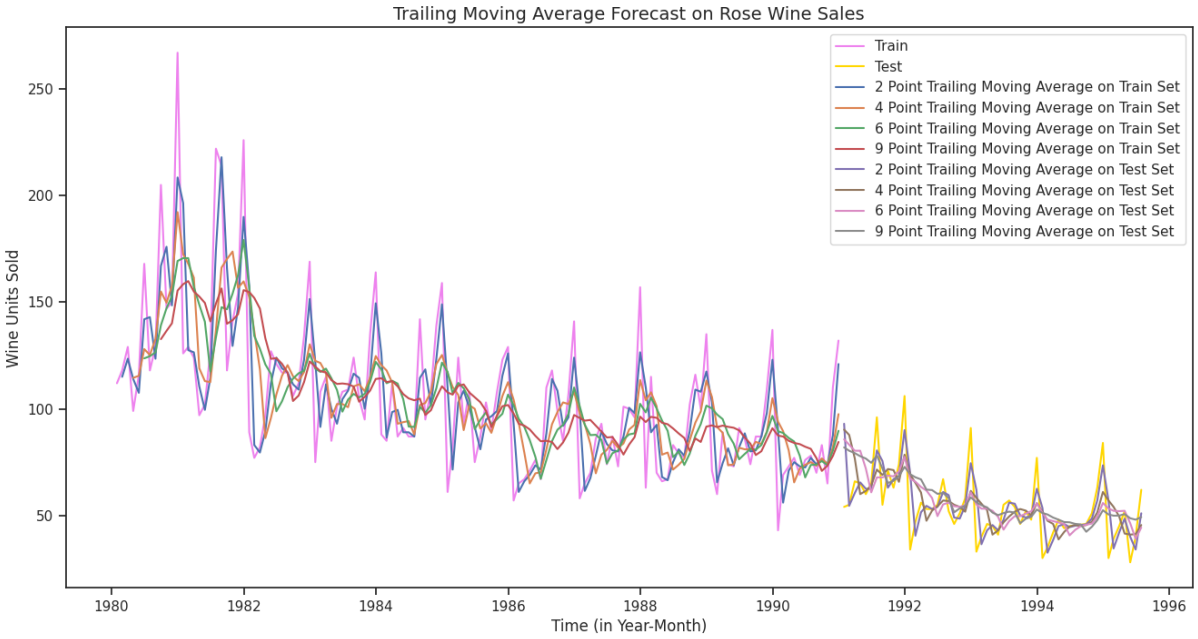
For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

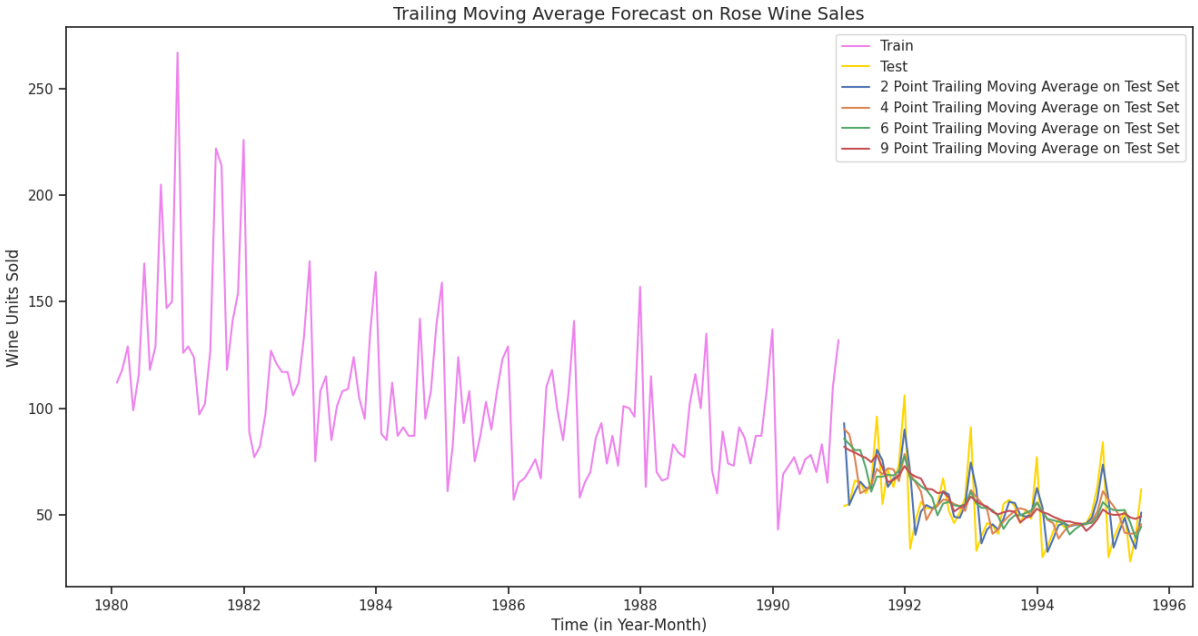
Time_Stamp	RoseWineSales	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31 00:00:00	112.0	NaN	NaN	NaN	NaN
1980-02-29 00:00:00	118.0	115.0	NaN	NaN	NaN
1980-03-31 00:00:00	129.0	123.5	NaN	NaN	NaN
1980-04-30 00:00:00	99.0	114.0	114.5	NaN	NaN
1980-05-31 00:00:00	116.0	107.5	115.5	NaN	NaN
1980-06-30 00:00:00	168.0	142.0	128.0	123.66	NaN
1980-07-31 00:00:00	118.0	143.0	125.25	124.66	NaN
1980-08-31 00:00:00	129.0	123.5	132.75	126.5	NaN
1980-09-30 00:00:00	205.0	167.0	155.0	139.16	132.666

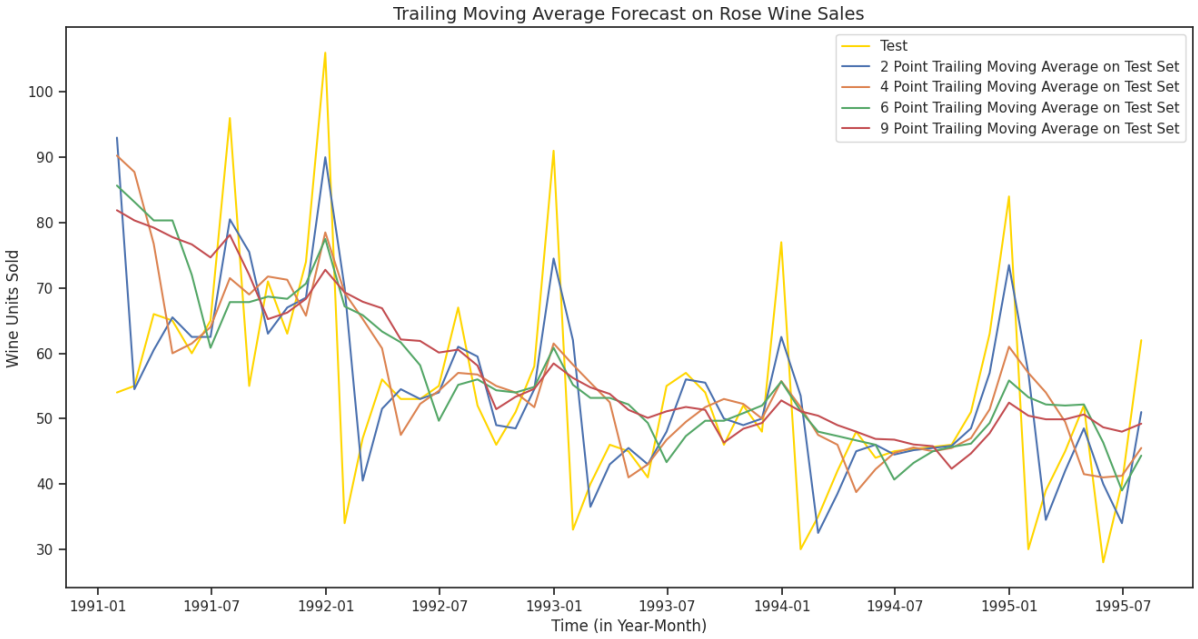

Moving Average on Entire Data
Train
2 Point Moving Average
4 Point Moving Average
6 Point Moving Average
9 Point Moving Average
Wine Units Sold
Time (in Year-Month)


Moving Average on Entire Data
Train
2 Point Moving Average
4 Point Moving Average
6 Point Moving Average
9 Point Moving Average
Wine Units Sold
Time (in Year-Month)


Moving Average on first 3 years of data
Train
2 Point Moving Average
4 Point Moving Average
6 Point Moving Average
9 Point Moving Average
Wine Units Sold
Time (in Year-Month)


Trailing Moving Average Forecast on Rose Wine Sales
Train
Test
2 Point Trailing Moving Average on Train Set
4 Point Trailing Moving Average on Train Set
6 Point Trailing Moving Average on Train Set
9 Point Trailing Moving Average on Train Set
2 Point Trailing Moving Average on Test Set
4 Point Trailing Moving Average on Test Set
6 Point Trailing Moving Average on Test Set
9 Point Trailing Moving Average on Test Set
Wine Units Sold
Time (in Year-Month)





Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal

• The seasonality and trend of the time series data may both be predicted using moving average models.

• We can see how the data smooth out as the number of observation points taken increases. The 2-point TMA has characteristics that are more similar to test results than the 9-point TMA.

• The root means squared error (RMSE) for the 2-point trailing average model is 11.529, which is lowest than all models build so far.

Moving Average: Model Evaluation

For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.529

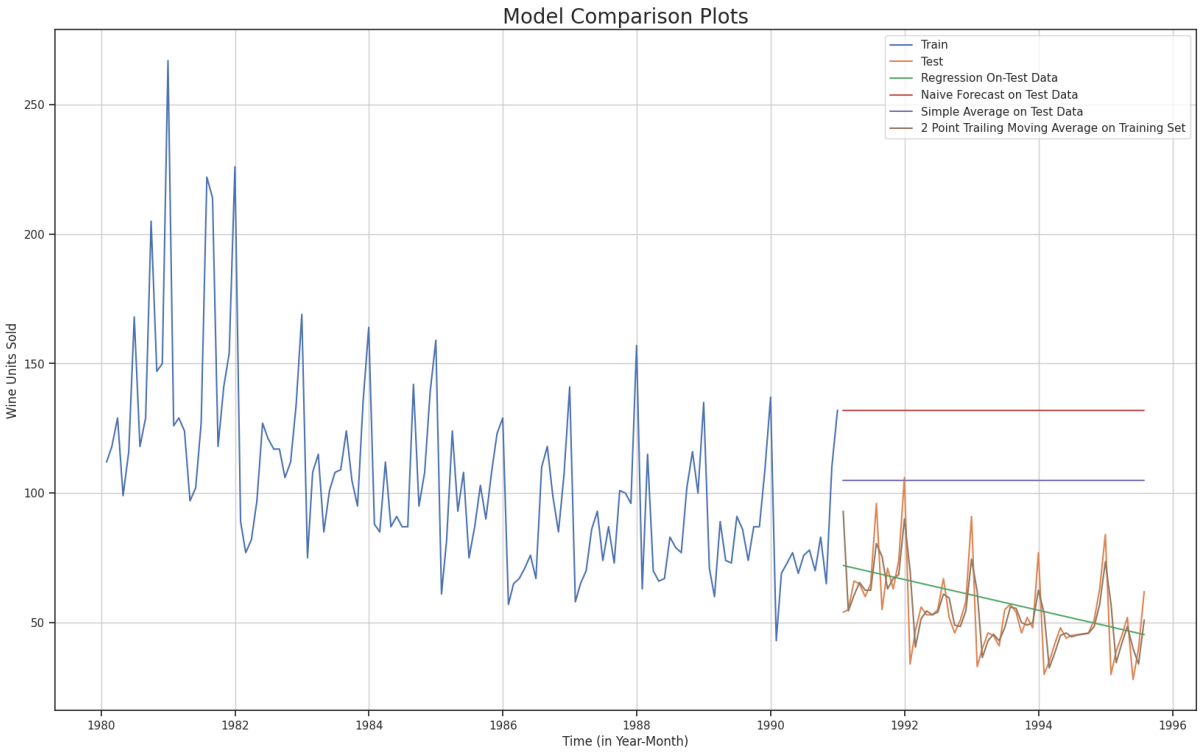
For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.451

For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.566

For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.728

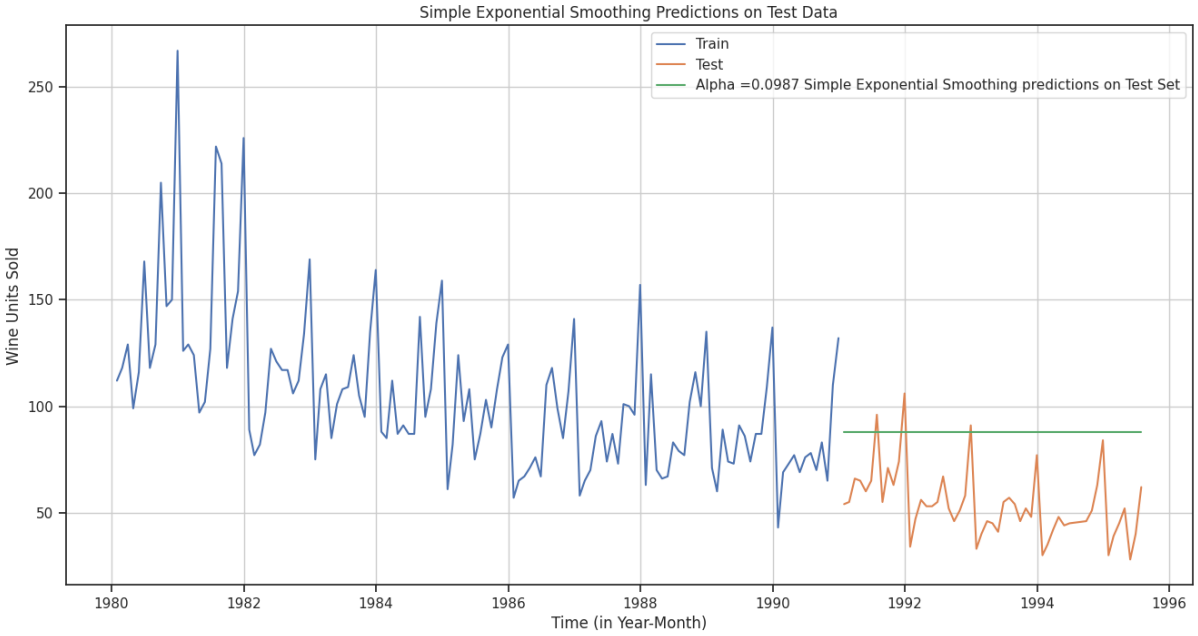
index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689
Simple Average	53.460367314208234
2 point TMA	11.529277632500342
4 point TMA	14.451376090500547
6 point TMA	14.566261685653924
9 point TMA	14.727595576680468

Let's compare the visualization of each model's predictions that we have constructed so far before investigating exponential smoothing methods.



Model 5: Simple Exponential Smoothing

Time_Stamp	Rose_Wine_Sales	predict
1991-01-31 00:00:00	54.0	87.98376547088112
1991-02-28 00:00:00	55.0	87.98376547088112
1991-03-31 00:00:00	66.0	87.98376547088112
1991-04-30 00:00:00	65.0	87.98376547088112
1991-05-31 00:00:00	60.0	87.98376547088112



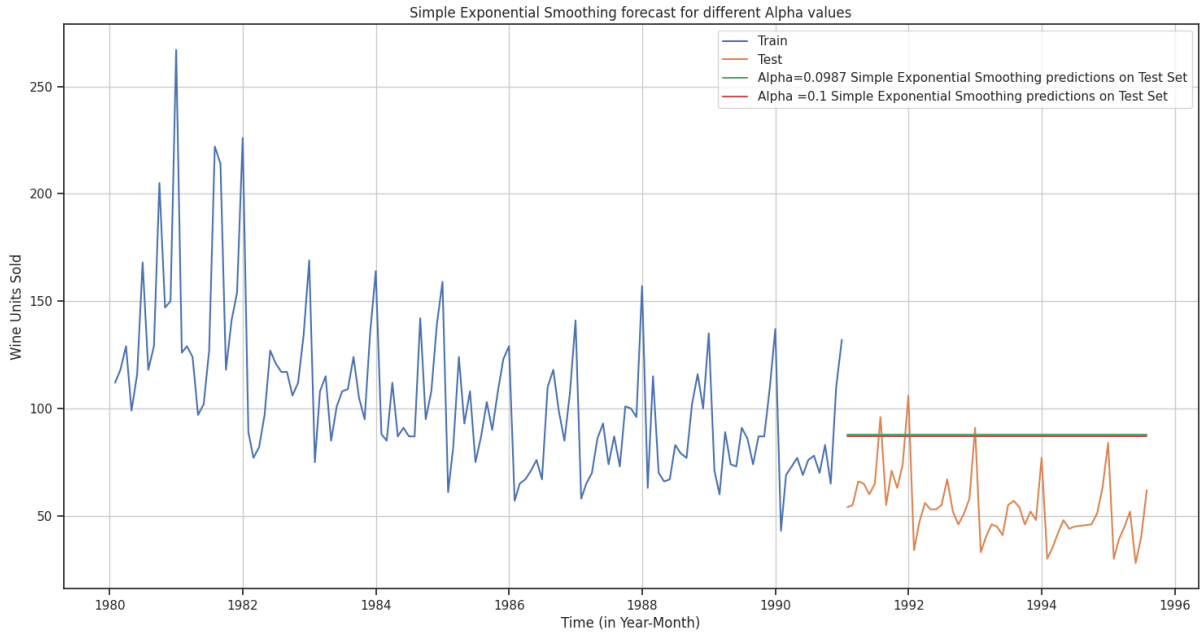
Model Evaluation for α = 0.0987 : Simple Exponential Smoothing

index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689
Simple Average	53.460367314208234
2 point TMA	11.529277632500342
4 point TMA	14.451376090500547
6 point TMA	14.566261685653924
9 point TMA	14.727595576680468
Alpha=0.0987,SimpleExponentialSmoothing	37.59200650745667

Setting different alpha values.

index	Alpha Values	Train RMSE	Test RMSE
0	0.1	31.81561020726095	36.82782707405664
1	0.15000000000000002	31.80984457017911	38.721919754361735
2	0.20000000000000004	31.97939065354775	41.361671124341825
3	0.25000000000000006	32.21187109382383	44.36059115088523
4	0.30000000000000004	32.47016356243125	47.50461726490155
5	0.3500000000000001	32.74434122709075	50.66546890623244
6	0.40000000000000013	33.0351297829429	53.76720357001585
7	0.45000000000000007	33.34657780662275	56.766931769940165
8	0.5000000000000001	33.68283896026435	59.64158456936708
9	0.5500000000000002	34.04704188415793	62.37878861594536
10	0.6000000000000002	34.44117100799307	64.97108770146639
11	0.6500000000000001	34.86635616650917	67.41270347070467
12	0.7000000000000002	35.32326089689815	69.69796313988307
13	0.7500000000000002	35.81243549483615	71.82065362777291

index	Alpha Values	Train RMSE	Test RMSE
14	0.8000000000000002	36.33459618723694	73.77379392930294
15	0.8500000000000002	36.89083533847241	75.54953831839295
16	0.9000000000000002	37.4827822463688	77.1390783950577
17	0.9500000000000003	38.11273526508124	78.53249839466851



Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal

• When there is neither a trend nor a seasonal component to the time series, simple exponential smoothing is typically used. It is due to this reason, it unable to capture the characteristics of the time series data.

• The root means squared error (RMSE) for the simple exponential smoothing model with Alpha=0.0987 is 36.796 and for Alpha=0.1, RMSE is 36.827.

• The Simple Exponential Smoothing with alpha=0.0987 is taken as the best model among two as it has the lowest test RMSE.

Method 6: Double Exponential Smoothing (Holt's Model)

This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters. Applicable when data has Trend but no seasonality. Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series. A second smoothing parameter β corresponds to the trend series.

Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short-term average value or level and the other for capturing the trend.

Intercept or Level equation, Lt is given by: $L_t = \alpha Y_t + (1-\alpha)F_t$

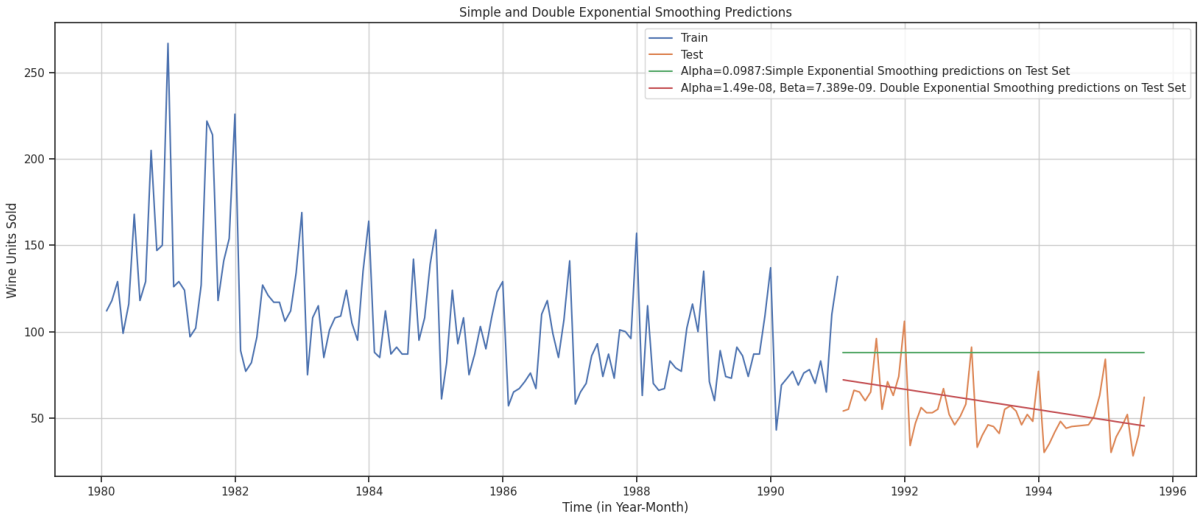
Trend equation is given by $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

Here, αα and ββ are the smoothing constants for level and trend, respectively, $0 < 1$ and $0 < \beta < 1$.

The forecast at time t + 1 is given by $F_{t+1} = L_t + T_t$

$F_{t+n} = L_t + nT_t$

Two parameters α and β are estimated in this model. Level and Trend are accounted for in this model.

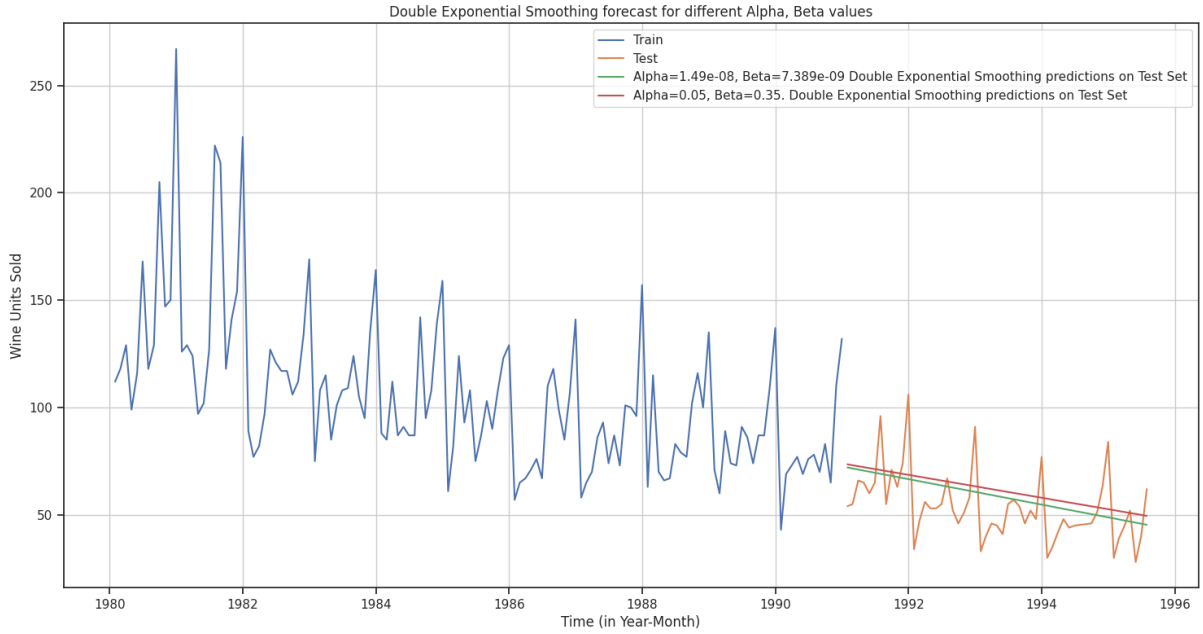


Model Evaluation- Double Exponential Smoothing(Holt's Model)

For DES forecast on the Rose Testing Data: RMSE is 15.271

index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689
Simple Average	53.460367314208234
2 point TMA	11.529277632500342
4 point TMA	14.451376090500547
6 point TMA	14.566261685653924
9 point TMA	14.727595576680468
Alpha=0.0987,SimpleExponentialSmoothing	37.59200650745667
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.270900637349518

index	Alpha	Beta	Train RMSE	Test RMSE
6	0.05	0.35000000000000003	36.23399687684336	16.328994356740804
5	0.05	0.3	36.61687732492835	18.624520481716626
2	0.05	0.15000000000000002	39.106562708807175	23.71678713439637
0	0.05	0.05	49.734056110453395	31.526698022684904
7	0.05	0.4	35.783736892787104	31.577953349652905



Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal

• When there is simply trend and no seasonality in the time series data, the double exponential smoothing model performs well. It is due to this reason it is only able to capture the trend characteristics of the data and seasonality is not accounted for.

• The root means squared error (RMSE) for the double exponential smoothing model with Alpha=1.49e-08, Beta=7.389e-09 is 15.268 and for Alpha=0.05, Beta=0.35 (Auto tuned model), RMSE is 16.328994.

• The Double Exponential Smoothing with Alpha=1.49e-08, Beta=7.389e-09 is taken as the best model among two as it has the lowest test RMSE.

• Additionally, it should be highlighted that compared to the simple exponential smoothing model, the double exponential smoothing model has almost halved the RMSE values.

Method 7: Triple Exponential Smoothing (Holt- Winter's Model)

This model is an extension of DES known as Triple Exponential Smoothing model which estimates three smoothing parameters. Applicable when data has both Trend and seasonality. Three separate components are considered: Level, Trend and Seasonality.

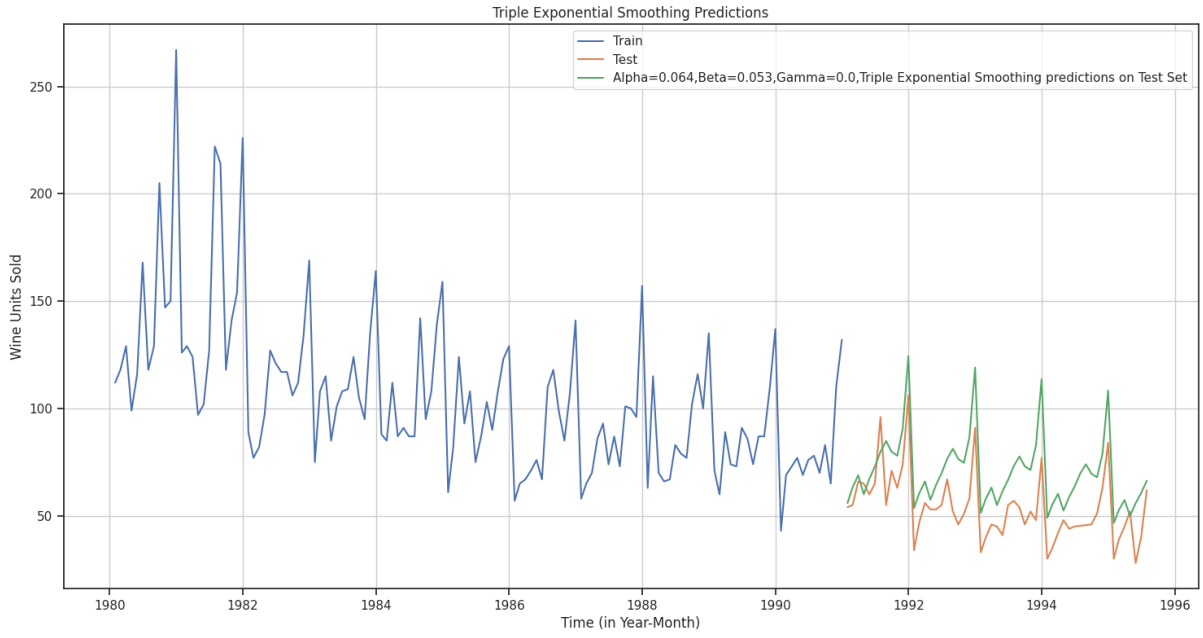
One smoothing parameter α corresponds to the level series.

A second smoothing parameter β corresponds to the trend series.

A third smoothing parameter γ corresponds to the seasonality series

where, $0 < \alpha < \beta < \gamma < 1$

Time_Stamp	Rose_Wine_Sales	auto_predict
1991-01-31 00:00:00	54.0	55.942245548807584
1991-02-28 00:00:00	55.0	63.240623582309155
1991-03-31 00:00:00	66.0	68.89967400895928
1991-04-30 00:00:00	65.0	60.007486067789735
1991-05-31 00:00:00	60.0	67.25714966965026



6 Rose Wine - TES predictions on Test data

The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set. 48 The range of alpha value is from 0.1 to 1.0 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.

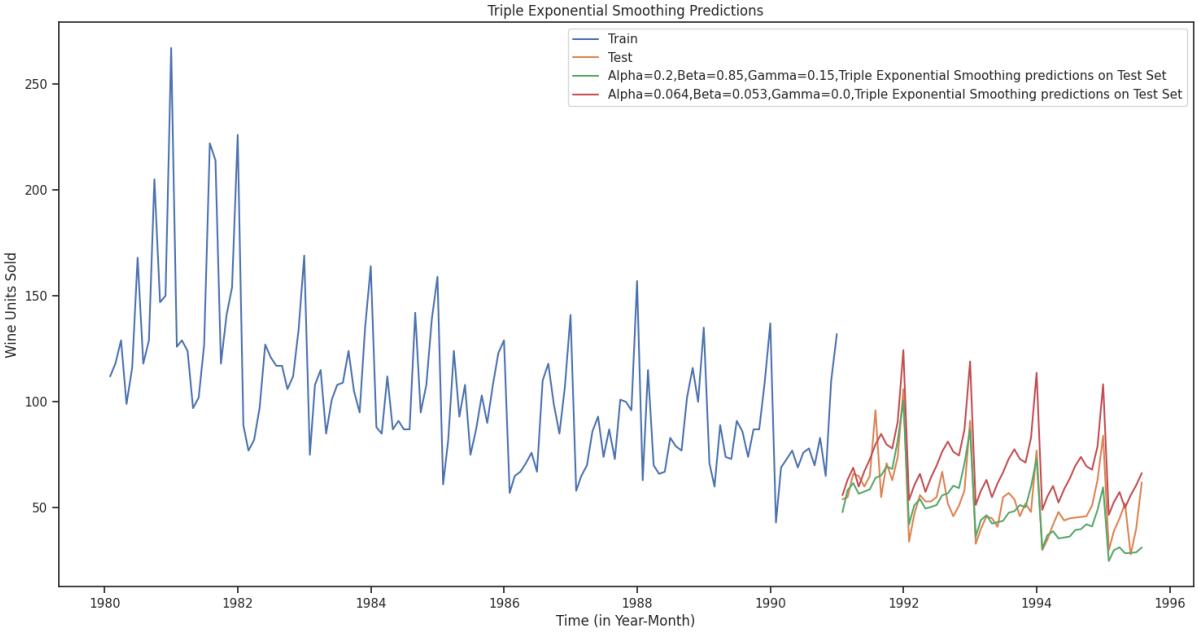
index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689
Simple Average	53.460367314208234
2 point TMA	11.529277632500342
4 point TMA	14.451376090500547
6 point TMA	14.566261685653924
9 point TMA	14.727595576680468
Alpha=0.0987,SimpleExponentialSmoothing	37.59200650745667
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.270900637349518
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	19.112865532946618

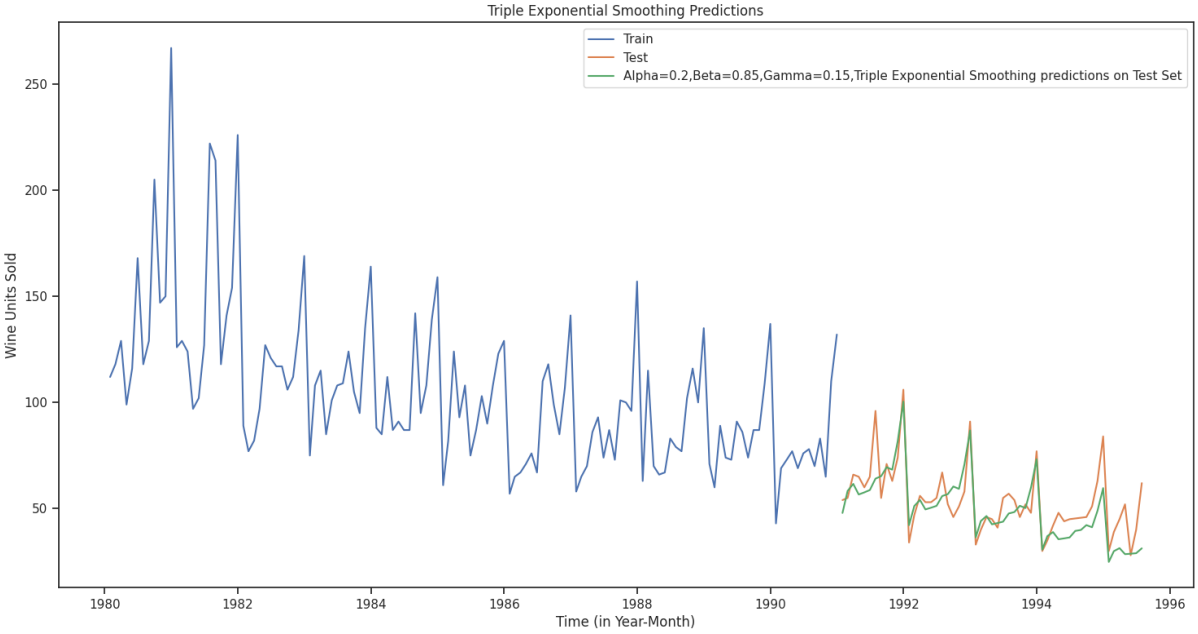
Show 102550100 per page

Calculating the performance metrics for different values of alpha, beta and gamma

1 to 5 of 5 entriesFilter

index	Alpha	Beta	Gamma	Train RMSE	Test RMSE
38	0.1	0.20000000000000004	0.1	19.770392082177274	9.223385307567227
19	0.1	0.15000000000000002	0.1	19.647564352663277	9.28063035843363
20	0.1	0.15000000000000002	0.150	19.879642435613896	9.313311136012944
39	0.1	0.20000000000000004	0.150	19.993052765354456	9.337491775323308
21	0.1	0.15000000000000002	0.200	20.148043178123995	9.406805470200052


Triple Exponential Smoothing Predictions
Train
Test
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing predictions on Test Set
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing predictions on Test Set
Wine Units Sold
Time (in Year-Month)


Triple Exponential Smoothing Predictions
Train
Test
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing predictions on Test Set
Wine Units Sold
Time (in Year-Month)

index	Test RMSE
Linear Regression	15.268887473349798
Naive Model	79.71857596912689
Simple Average	53.460367314208234
2 point TMA	11.529277632500342
4 point TMA	14.451376090500547
6 point TMA	14.566261685653924
9 point TMA	14.727595576680468
Alpha=0.0987,SimpleExponentialSmoothing	37.59200650745667
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.270900637349518
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	19.112865532946618
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing	10.27987620009132

Observation:

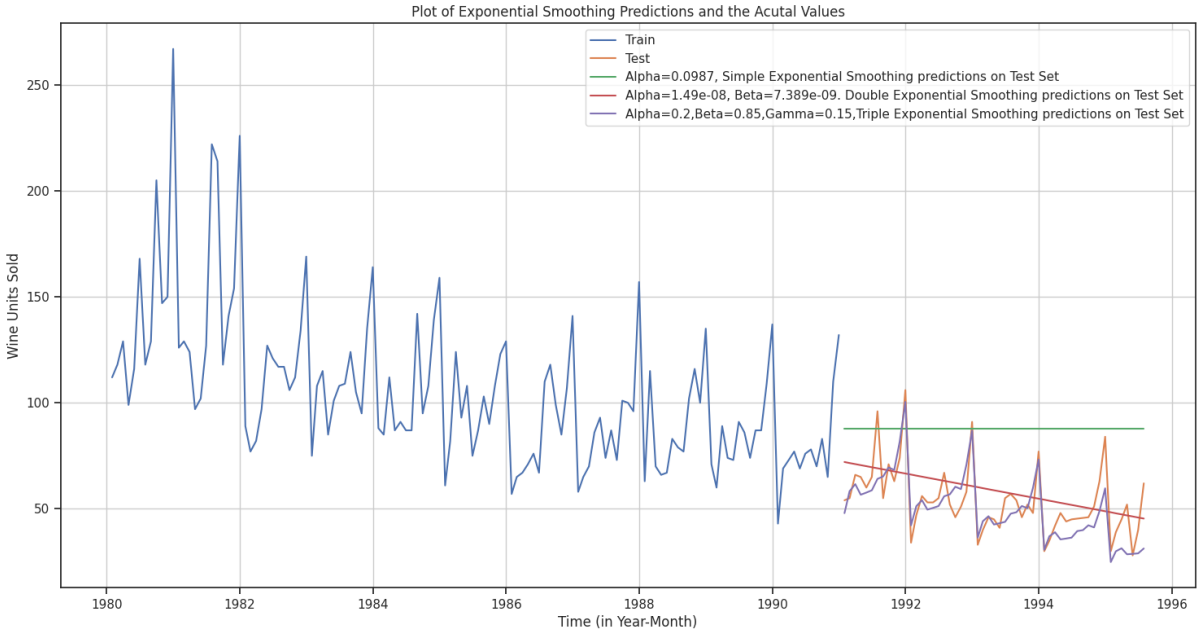
• We can see from the graphs above that the time series has a falling trend and is seasonal

• When there is both trend and seasonality in the time series data, the triple exponential model works well. It is due to this reason it able to capture both the trend and seasonal characteristics and nearly match the actual test data plot.

• The root means squared error (RMSE) for the double exponential smoothing model with Alpha=0.064, Beta=0.053, Gamma=0.0 is 21.154 and for Alpha=0.2, Beta=0.85, Gamma=0.15 (Auto tuned model), RMSE is 9.121.

• The Triple Exponential Smoothing with Alpha=0.2, Beta=0.85, Gamma=0.15 is taken as the best model among two as it has the lowest test RMSE.

• Additionally, it should be highlighted that compared to the double exponential smoothing model, the triple exponential smoothing model has almost reduced the RMSE value by 40%.



5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Check for stationarity of the whole Time Series data.

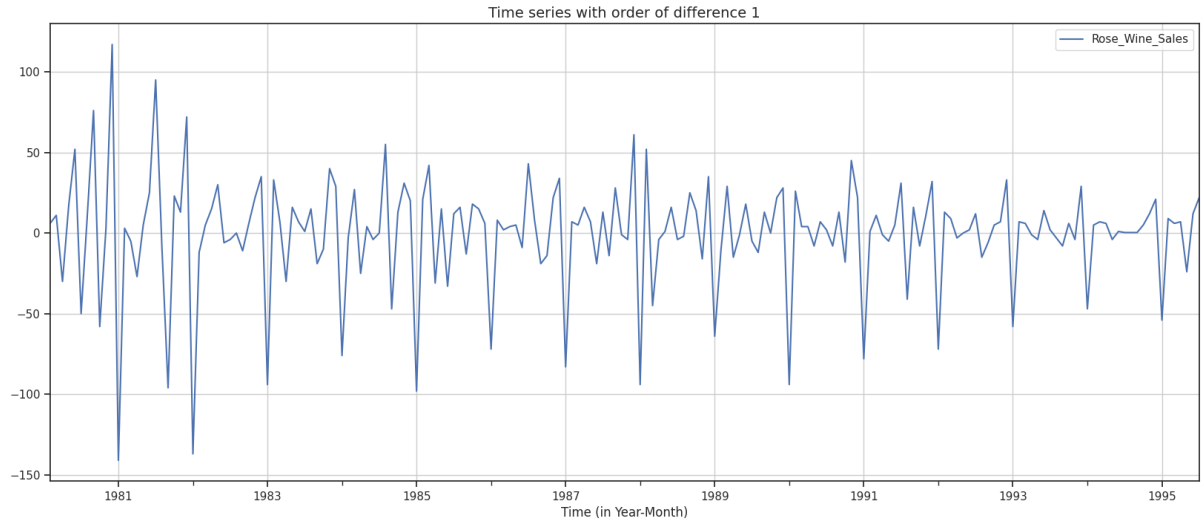
The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

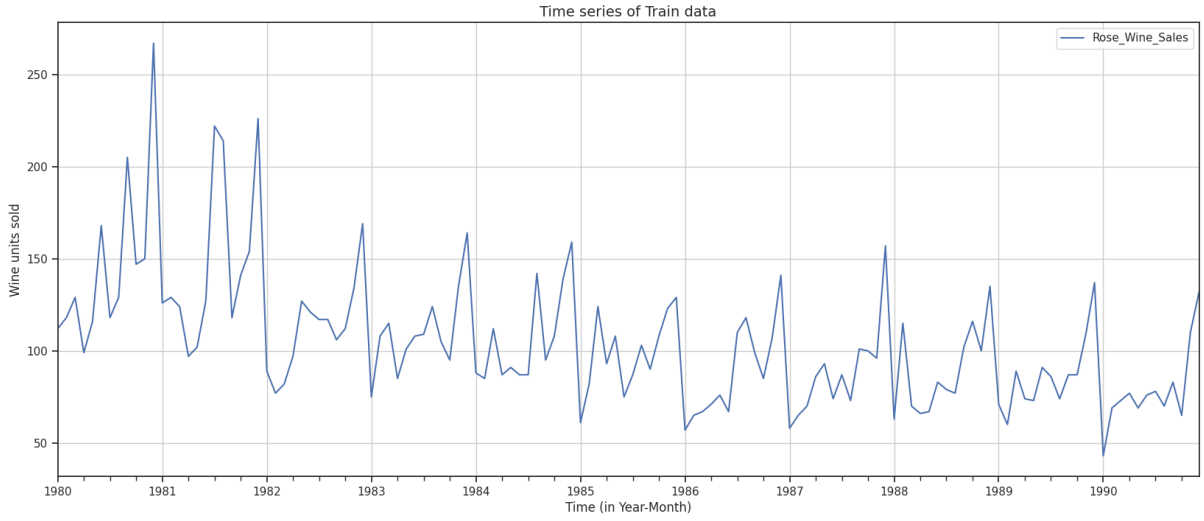
- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

We see that at 5% significant level the Time Series is non-stationary.Let us take one level of differencing to see whether the series becomes stationary.

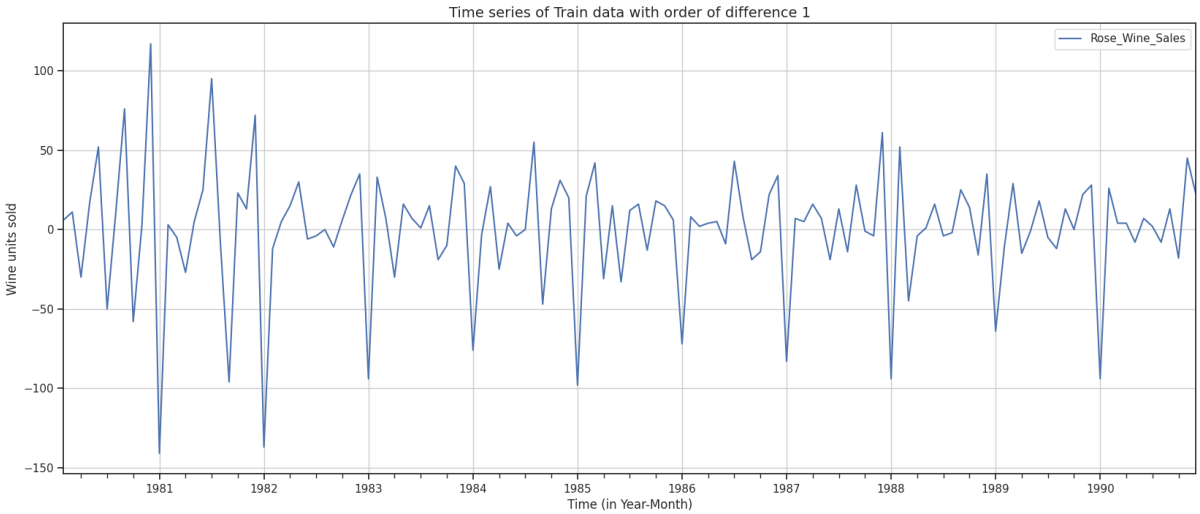


Check for stationarity of the Training Data Time Series.



Inference: We see that at 5% significant level the Time Series of training data is non-stationary as p-value is 0.756 which is more than alpha value (0.05), therefore we fail to reject the null hypothesis. Let us take one level of differencing to see whether the series becomes stationary

Inference: We see that at 5% significant level the Time Series of training data is non-stationary as p-value is 3.894e-08 which is less than alpha value (0.05), therefore we reject the null hypothesis. We can see that the provided training time series becomes stationary with differencing.



Time Series Plot of Training data with differencing

Observation:

• As per the Augmented Dicky-Fuller test, we observed that the time series data by itself is not stationary, however, it becomes stationary when differencing is done.

• The same thing is also observed with Training data. Therefore, for training the models, it can be built with order of difference d=1.

6) Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Model 8 – Auto-Regressive Integrated Moving Average (ARIMA)

Auto-regression means regression of a variable on itself. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR. ARIMA models may be used to represent any "non-seasonal" time series that has patterns and isn't just random noise.

An ARIMA model is characterized by 3 terms: p, d, q where, p is the order of the Auto Regressive (AR) term q is the order of the Moving Average (MA) term d is the number of differencing required to make the time series stationary

For the selection criteria of p,d,q the below ARIMA model is built by using automated model parameters with lowest Akaike Information Criteria.

Examples of the parameter combinations for the Model

Model: (0, 1, 0)

Model: (0, 1, 1)

Model: (0, 1, 2)

Model: (0, 1, 3)

Model: (0, 1, 4)

Model: (1, 1, 0)

Model: (1, 1, 1)

Model: (1, 1, 2)

Model: (1, 1, 3)

Model: (1, 1, 4)

Model: (2, 1, 0)

Model: (2, 1, 1)

Model: (2, 1, 2)

Model: (2, 1, 3)

Model: (2, 1, 4)

Model: (3, 1, 0)

Model: (3, 1, 1)

Model: (3, 1, 2)

Model: (3, 1, 3)

Model: (3, 1, 4)

Model: (4, 1, 0)

Model: (4, 1, 1)

Model: (4, 1, 2)

Model: (4, 1, 3)

Model: (4, 1, 4)

AIC values for different parameter combinations

ARIMA(0, 1, 0) - AIC:1333.1546729124348

ARIMA(0, 1, 1) - AIC:1282.309831974831

ARIMA(0, 1, 2) - AIC:1279.6715288535768

ARIMA(0, 1, 3) - AIC:1280.5453761734657

ARIMA(0, 1, 4) - AIC:1281.6766982143947

ARIMA(1, 1, 0) - AIC:1317.350310538146

ARIMA(1, 1, 1) - AIC:1280.5742295380066

ARIMA(1, 1, 2) - AIC:1279.8707234231927

ARIMA(1, 1, 3) - AIC:1281.870722330997

ARIMA(1, 1, 4) - AIC:1279.6052628202858

ARIMA(2, 1, 0) - AIC:1298.6110341604967

ARIMA(2, 1, 1) - AIC:1281.507862186856

ARIMA(2, 1, 2) - AIC:1281.8707222264661

ARIMA(2, 1, 3) - AIC:1274.6948343352535

ARIMA(2, 1, 4) - AIC:1278.7690096450388

ARIMA(3, 1, 0) - AIC:1297.4810917271684

ARIMA(3, 1, 1) - AIC:1282.419277627198

ARIMA(3, 1, 2) - AIC:1283.720740597713

ARIMA(3, 1, 3) - AIC:1278.6670248711753

ARIMA(4, 1, 0) - AIC:1287.7190768609312

ARIMA(4, 1, 0) - AIC:1296.3266569004463

ARIMA(4, 1, 1) - AIC:1283.7931715123077

ARIMA(4, 1, 2) - AIC:1285.7182485632707

ARIMA(4, 1, 3) - AIC:1278.4514096395753

ARIMA(4, 1, 4) - AIC:1282.3371724386477

We can see that among all the possible given combinations, the AIC is lowest for the combination (2,1,3). Hence, the model is built with these parameters to determine the RMSE value of test data.

SARIMAX Results

==========

Dep. Variable: Rose_Wine_Sales No. Observations: 132

Model: ARIMA(2, 1, 3) Log Likelihood -631.347

Date: Sun, 04 May 2025 AIC 1274.695

Time: 16:00:42 BIC 1291.946

Sample: 01-31-1980 HQIC 1281.705

- 12-31-1990

Covariance Type: opg

==========

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6781	0.084	-20.030	0.000	-1.842	-1.514
ar.L2	-0.7289	0.084	-8.700	0.000	-0.893	-0.565
ma.L1	1.0451	0.711	1.469	0.142	-0.349	2.439
ma.L2	-0.7716	0.139	-5.551	0.000	-1.044	-0.499
ma.L3	-0.9047	0.647	-1.399	0.162	-2.172	0.362

sigma2 858.2445 600.102 1.430 0.153 -317.935 2034.424

==========

=

Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 24.45

Prob(Q): 0.88 Prob(JB): 0.00

Heteroskedasticity (H): 0.40 Skew: 0.71

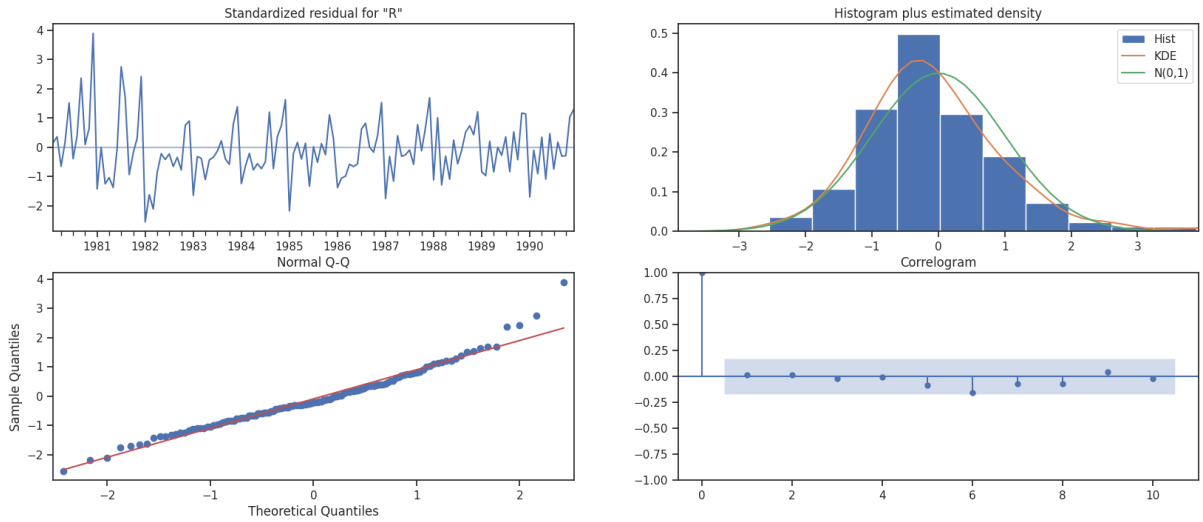
Prob(H) (two-sided): 0.00 Kurtosis: 4.57

==========

=

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).



Observation:

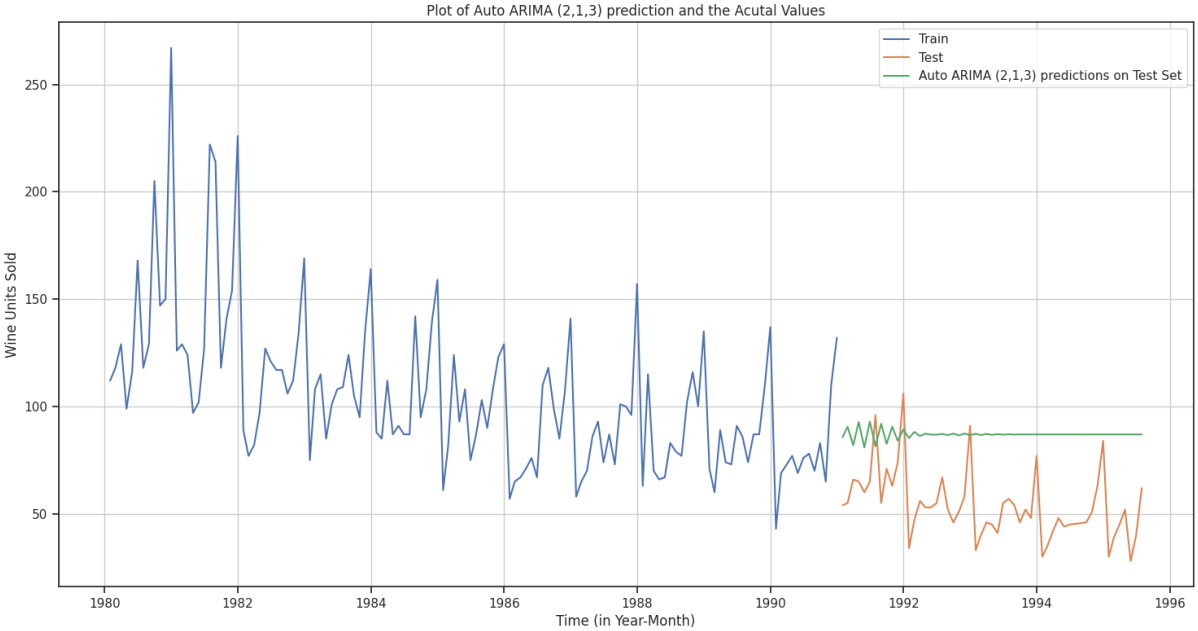
• The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. The AIC is lowest for the combination (2,1,3) as we see from the above results.

• From the Standardized residual plot above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.

• The histogram plus estimated density plot suggests a slightly uniform distribution with mean zero and slightly skewed to the right.

• In Normal Q-Q plot, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.

• The correlogram plot of residuals shows that the residuals are not auto correlated.



Plot of Automated ARIMA (2,1,3) predictions on Test data

Automated ARIMA: Model Evaluation

For evaluating the model's performance metrics, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)

index	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.8169146419244	75.84685520971229

Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal

• ARIMA models performs well on non-seasonal time series. It is due to this reason it is unable to capture the entire characteristics of the test data.

• The root means squared error (RMSE) of test data for the ARIMA model with (p=2, d=1, q=3) is 36.813.

• Not surprisingly, the RMSE of the aforementioned ARIMA model is greater than the majority of previously constructed models.

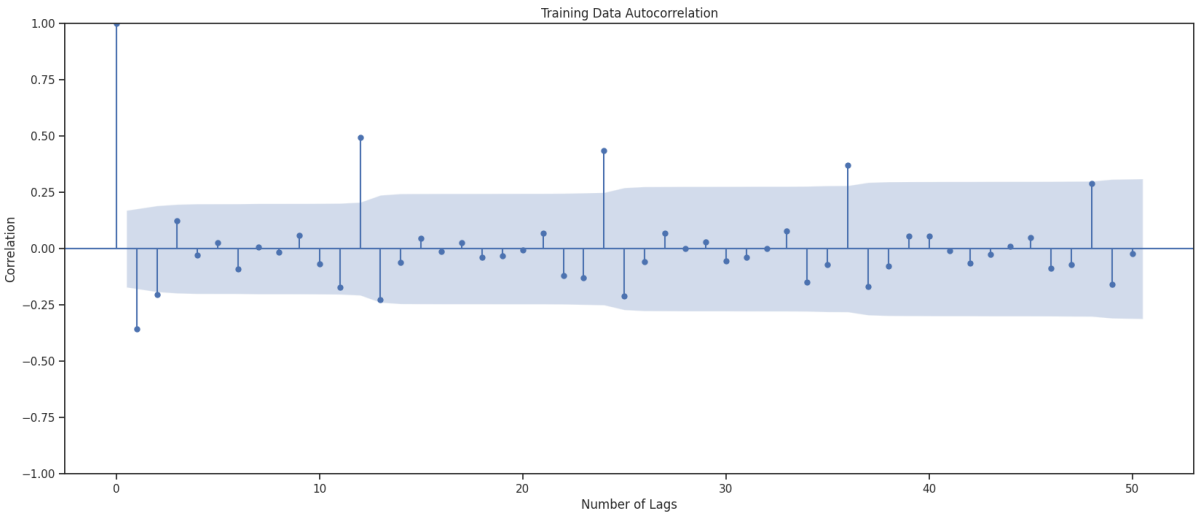
Model 9 – Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

SARIMA models or also known as Seasonal ARIMA is an extension of ARIMA for a time series data with defined seasonality. SARIMA models use seasonal differencing which is similar to regular differencing.

A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F where, p is the order of the Auto Regressive (AR) term q is the order of the Moving Average (MA) term d is the number of differencing required to make the time series

stationary P is the order of the Seasonal Auto Regressive (AR) term Q is the order of the Seasonal Moving Average (MA) term D is the number of seasonal differencing required to make the time series stationary F is the seasonal frequency of the time series

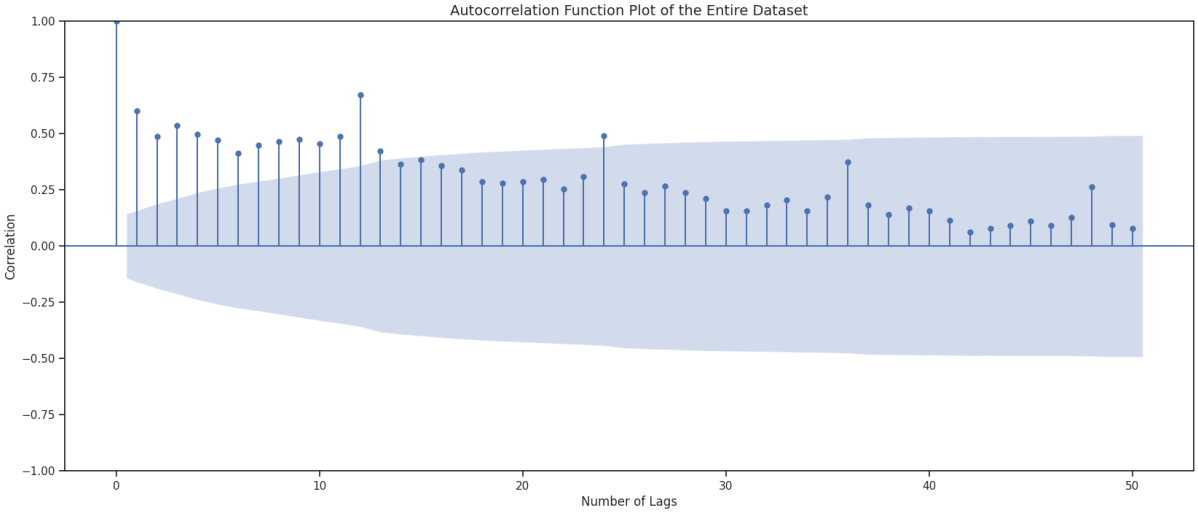
We must examine the PACF and ACF plots, respectively, at delays that are the multiple of "F" in order to determine the "P" and "Q" values, and determine where these cut-off values are (for appropriate confidence interval bands). By examining the lowest AIC values, we can also estimate "p," "q," "P," and "Q" for the SARIMA models. By examining the ACF plots, one may calculate the seasonal parameter 'F'. The existence of seasonality should be shown by a spike in the ACF plot at multiples of "F.

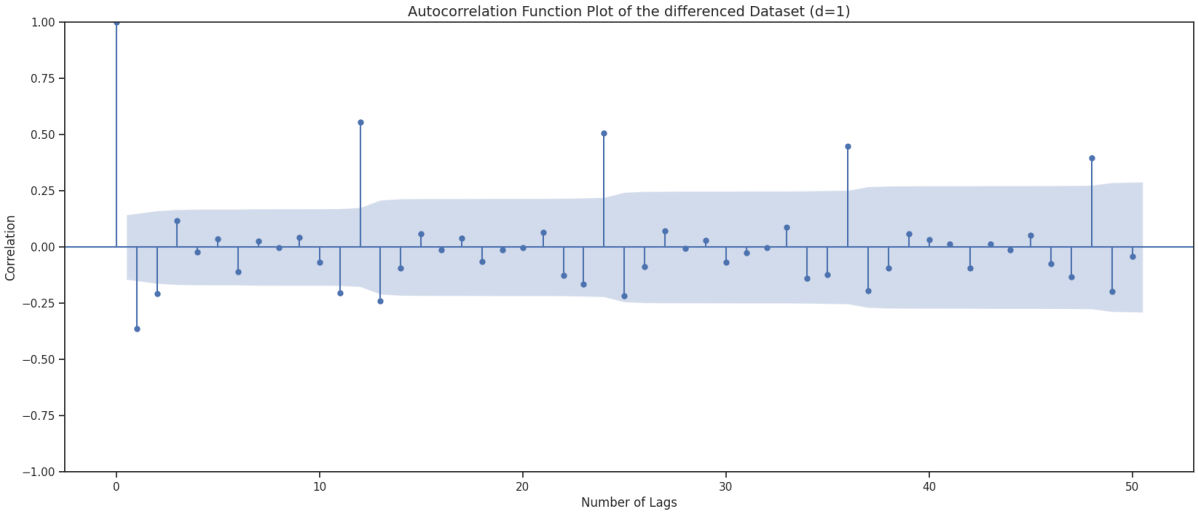


From the above ACF plot we can observe that at every 12th lag is significant indicating the presence of seasonality. Hence for our model building we will consider the term F=12.

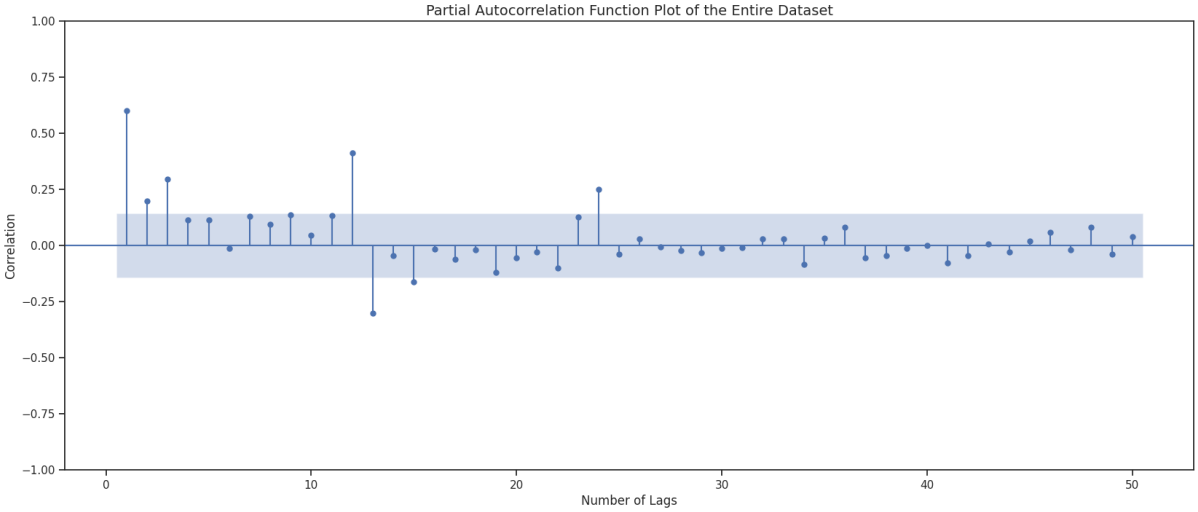
Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.

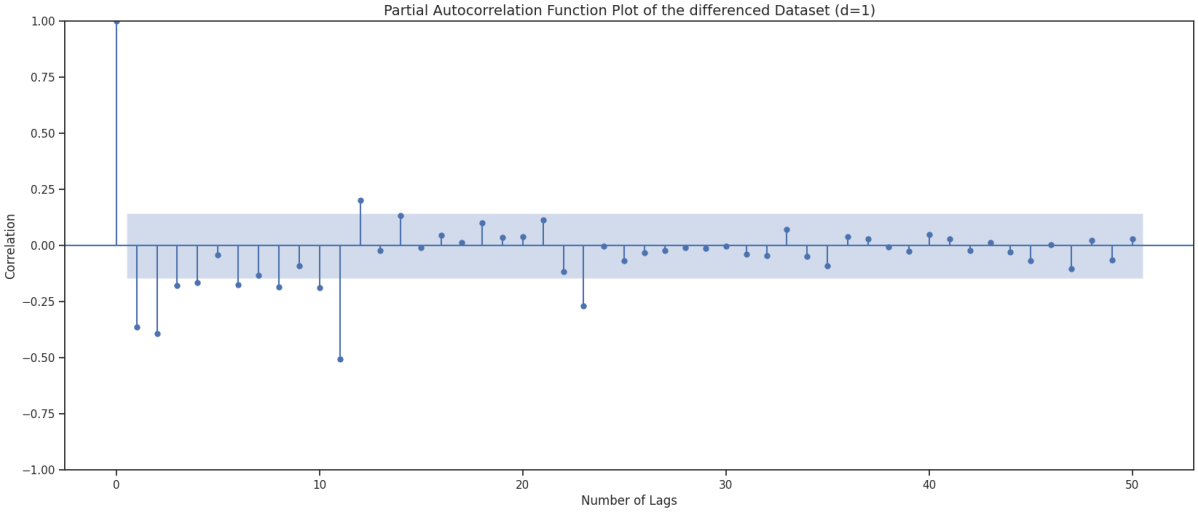
ACF plot





PACF plot


Partial Autocorrelation Function Plot of the Entire Dataset
Correlation
Number of Lags
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50


Partial Autocorrelation Function Plot of the differenced Dataset (d=1)
Correlation
Number of Lags
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50

For the selection criteria of p, d, q, P, D, Q & F the below SARIMA model is built by using automated model parameters with lowest Akaike Information Criteria.

Examples of the parameter combinations for the Model are

Model: (0, 1, 1)(0, 0, 1, 12)

Model: (0, 1, 2)(0, 0, 2, 12)

Model: (0, 1, 3)(0, 0, 3, 12)

Model: (1, 1, 0)(1, 0, 0, 12)

Model: (1, 1, 1)(1, 0, 1, 12)

Model: (1, 1, 2)(1, 0, 2, 12)

Model: (1, 1, 3)(1, 0, 3, 12)

Model: (2, 1, 0)(2, 0, 0, 12)

Model: (2, 1, 1)(2, 0, 1, 12)

Model: (2, 1, 2)(2, 0, 2, 12)

Model: (2, 1, 3)(2, 0, 3, 12)

Model: (3, 1, 0)(3, 0, 0, 12)

Model: (3, 1, 1)(3, 0, 1, 12)

Model: (3, 1, 2)(3, 0, 2, 12)

Model: (3, 1, 3)(3, 0, 3, 12)

index	param	seasonal	AIC
222	3,1,1	3,0,2,12	774.4002855257404
238	3,1,2	3,0,2,12	774.8809373813601
220	3,1,1	3,0,0,12	775.4266990180312
221	3,1,1	3,0,1,12	775.4953300789074
252	3,1,3	3,0,0,12	775.5610184773062

We can see that among all the possible given combinations, the AIC is lowest for the combination (3,1,1) (3,0,2,12). Hence, the model is built with these parameters to determine the RMSE value of test data.

SARIMAX Results

==========

Dep. Variable: Rose_Wine_Sales No. Observations: 132

Model: SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12) Log Likelihood -377.200

Date: Sun, 04 May 2025 AIC 774.400

Time: 16:08:27 BIC 799.618

Sample: 01-31-1980 HQIC 784.578

- 12-31-1990

Covariance Type: opg

==========

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0464	0.126	0.367	0.714	-0.202	0.294
ar.L2	-0.0060	0.120	-0.050	0.960	-0.241	0.229
ar.L3	-0.1808	0.098	-1.837	0.066	-0.374	0.012
ma.L1	-0.9370	0.067	-13.905	0.000	-1.069	-0.805
ar.S.L12	0.7639	0.165	4.639	0.000	0.441	1.087
ar.S.L24	0.0840	0.159	0.527	0.598	-0.229	0.397
ar.S.L36	0.0727	0.095	0.764	0.445	-0.114	0.259
ma.S.L12	-0.4969	0.250	-1.988	0.047	-0.987	-0.007
ma.S.L24	-0.2191	0.210	-1.044	0.296	-0.630	0.192
sigma2	192.1546	39.628	4.849	0.000	114.486	269.824

==========

Ljung-Box (L1) (Q): 0.30 Jarque-Bera (JB): 1.64

Prob(Q): 0.58 Prob(JB): 0.44

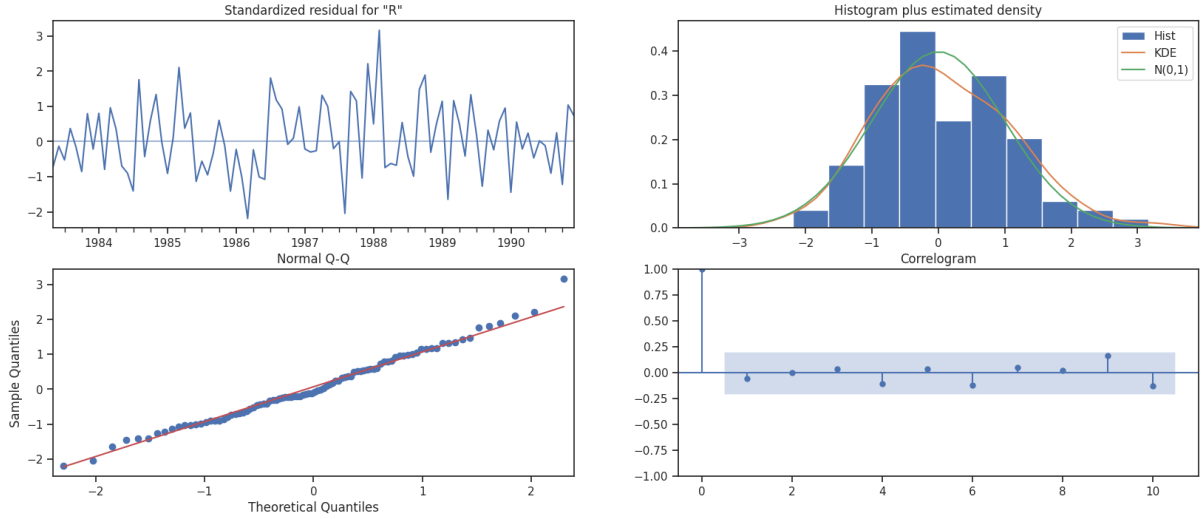
Heteroskedasticity (H): 1.11 Skew: 0.33

Prob(H) (two-sided): 0.77 Kurtosis: 3.03

==========

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).



Automated SARIMA – Diagnostics plot

Observation:

• The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. The AIC is lowest for the combination (3,1,1) (3,0,2,12) as we see from the above results.

• From the Standardized residual plot above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.

• The histogram plus estimated density plot suggests a slightly uniform distribution with mean zero and slightly skewed to the right.

• In Normal Q-Q plot, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.

• The correlogram plot of residuals shows that the residuals are not auto correlated.

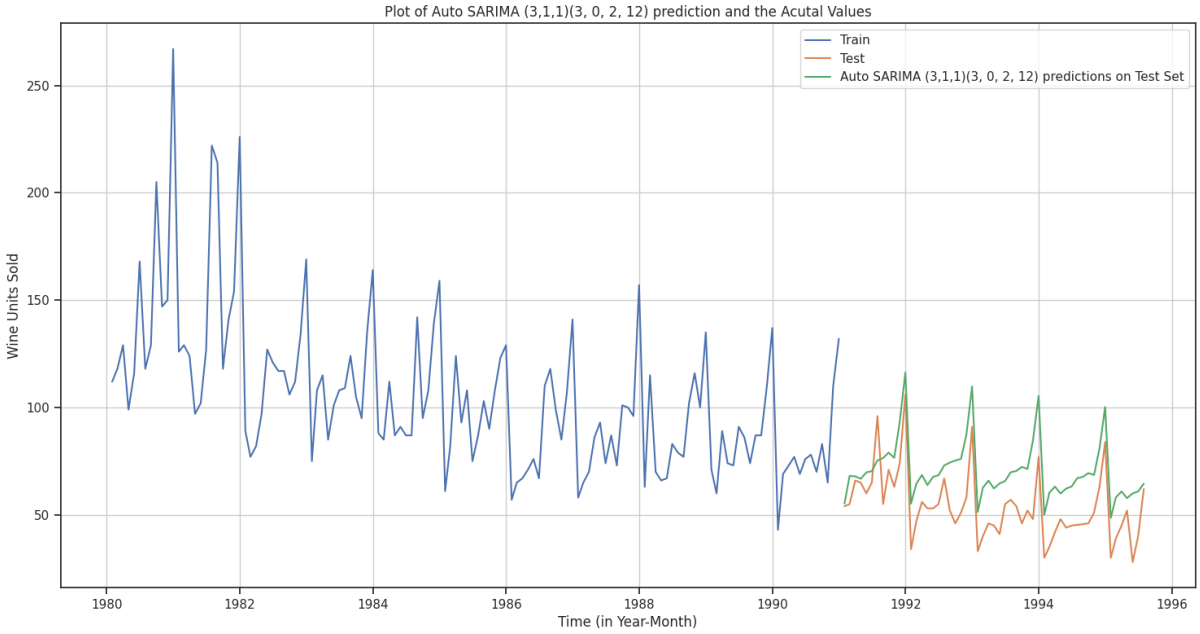
Predict on the Test Set using this model and evaluate the model.

	predicted_mean
1991-01-31	55.235804

	predicted_mean
1991-02-28	68.122717
1991-03-31	67.908796
1991-04-30	66.786275
1991-05-31	69.760396
1991-06-30	70.329027
1991-07-31	75.359554
1991-08-31	76.492004
1991-09-30	78.971473
1991-10-31	76.538751
1991-11-30	93.249097
1991-12-31	116.283322
1992-01-31	55.202512
1992-02-29	64.444133
1992-03-31	68.547786
1992-04-30	63.872415
1992-05-31	67.700119
1992-06-30	68.443663
1992-07-31	72.972131

	predicted_mean
1992-08-31	74.325289
1992-09-30	75.318018
1992-10-31	76.046923
1992-11-30	87.421391
1992-12-31	109.807383
1993-01-31	51.298317
1993-02-28	62.617128
1993-03-31	65.912098
1993-04-30	62.264427
1993-05-31	64.612125
1993-06-30	65.747624
1993-07-31	69.826334
1993-08-31	70.419917
1993-09-30	72.331879
1993-10-31	71.365764
1993-11-30	84.722363
1993-12-31	105.357994
1994-01-31	49.950026

	predicted_mean
1994-02-28	60.311826
1994-03-31	63.158237
1994-04-30	59.896978
1994-05-31	62.227966
1994-06-30	63.199166
1994-07-31	67.061279
1994-08-31	67.710829
1994-09-30	69.435109
1994-10-31	68.581419
1994-11-30	80.955444
1994-12-31	100.275049
1995-01-31	48.590836
1995-02-28	58.129339
1995-03-31	60.878993
1995-04-30	57.741226
1995-05-31	59.997485
1995-06-30	60.888866
1995-07-31	64.511161



Automated SARIMA: Model Evaluation

For evaluating the model performance, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)

index	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.8169146419244	75.84685520971229
Auto SARIMA (3,1,1)(3,0,2,12)	18.8818485219433	75.84685520971229

Observation:

• We can see from the graphs above that the time series has a falling trend and is seasonal • SARIMA model performs well on seasonal time series. It is due to this reason it is able to capture the entire characteristics of the test data.

• The root means squared error (RMSE) of test data for the SARIMA model with (p=3, d=1, q=1) (P=3, D=0, Q=2, F=12) is 18.881.

• Additionally, it should be highlighted that compared to the ARIMA model, the SARIMA model has almost halved the RMSE value.

7) Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Model 10 – Auto-Regressive Integrated Moving Average (ARIMA) - Manual

An ARIMA model is characterized by 3 terms: p, d, q where, p is the order of the Auto Regressive (AR) term q is the order of the Moving Average (MA) term d is the number of differencing required to make the time series stationary

Indicating which previous series values are most beneficial in forecasting future values, autocorrelation and partial autocorrelation are measures of relationship between present and past series values. You may identify the sequence of processes in an ARIMA model using this information.

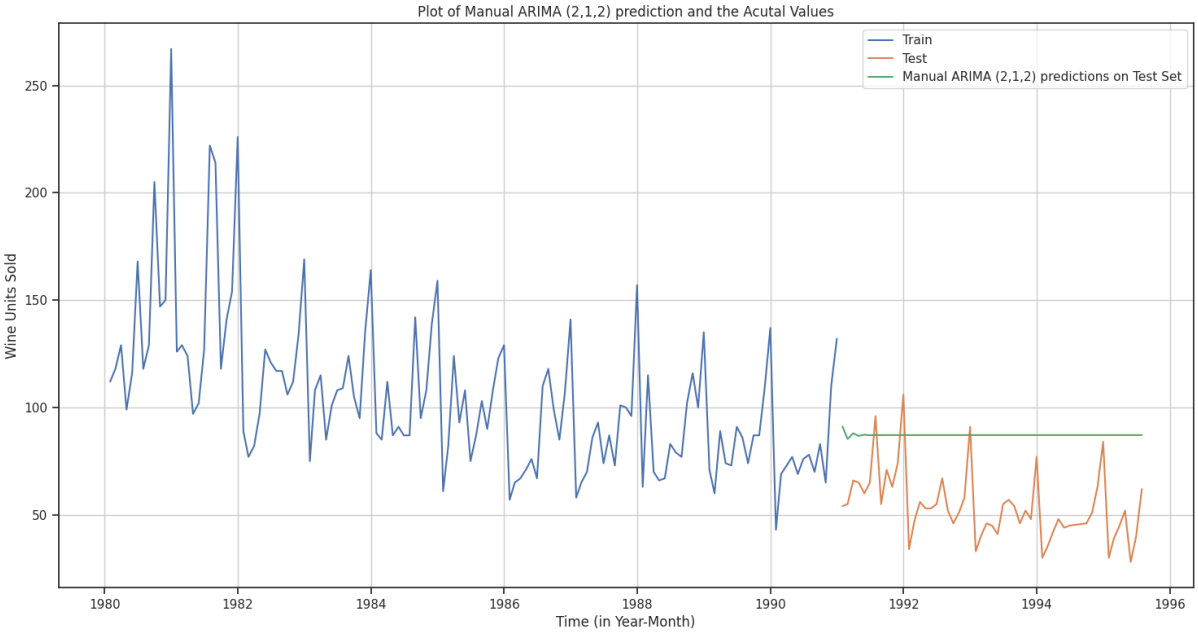
The parameters p & q can be determined by looking at the PACF & ACF plots respectively.

Autocorrelation function (ACF) - At lag k, this is the correlation between series values that are k intervals apart.

Partial autocorrelation function (PACF) - At lag k, this is the correlation between series values that are k intervals apart, accounting for the values of the intervals between.

In an ACF & PACF plots, each bar represents the size and direction of the connection. Bars that cross the red line are statistically significant.

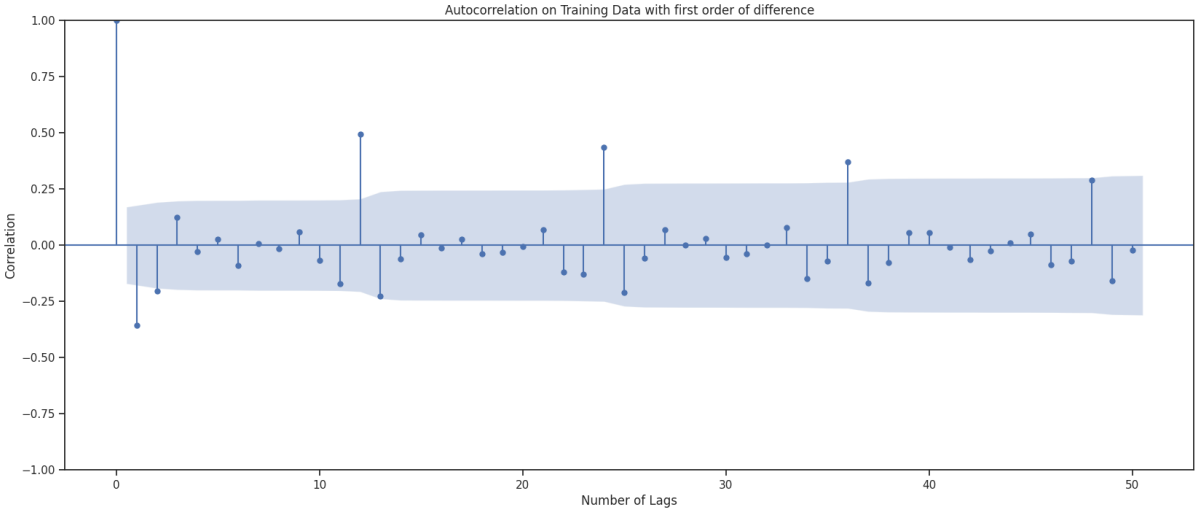
Predict on the Test Set using this model and evaluate the model.



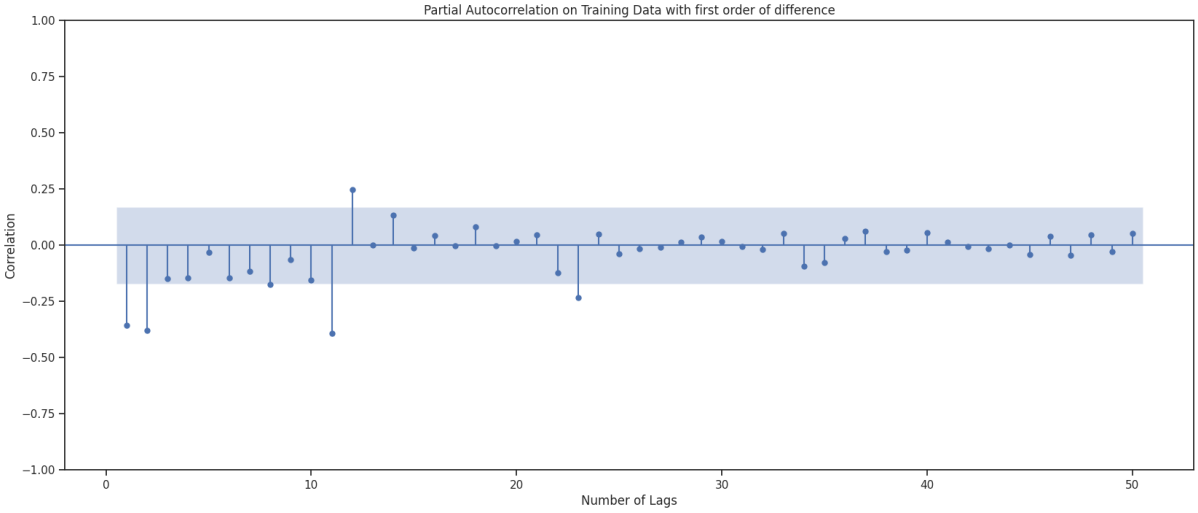
index	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.8169146419244	75.84685520971229
Auto SARIMA (3,1,1)(3,0,2,12)	18.8818485219433	75.84685520971229
Manual ARIMA(2,1,2)	36.870991041964494	75.84685520971229

Manual SARIMA Model

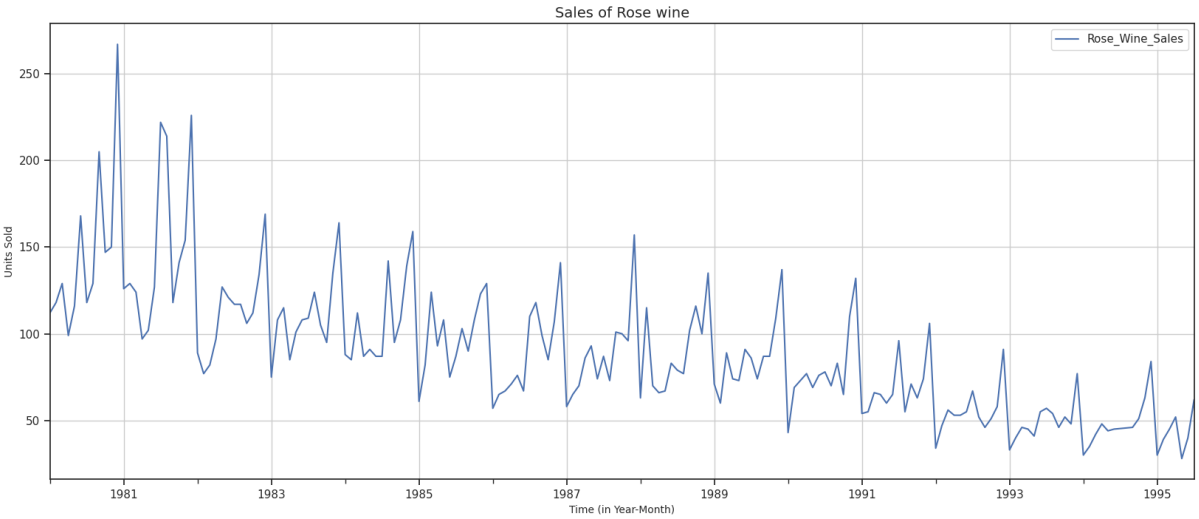
ACF plot

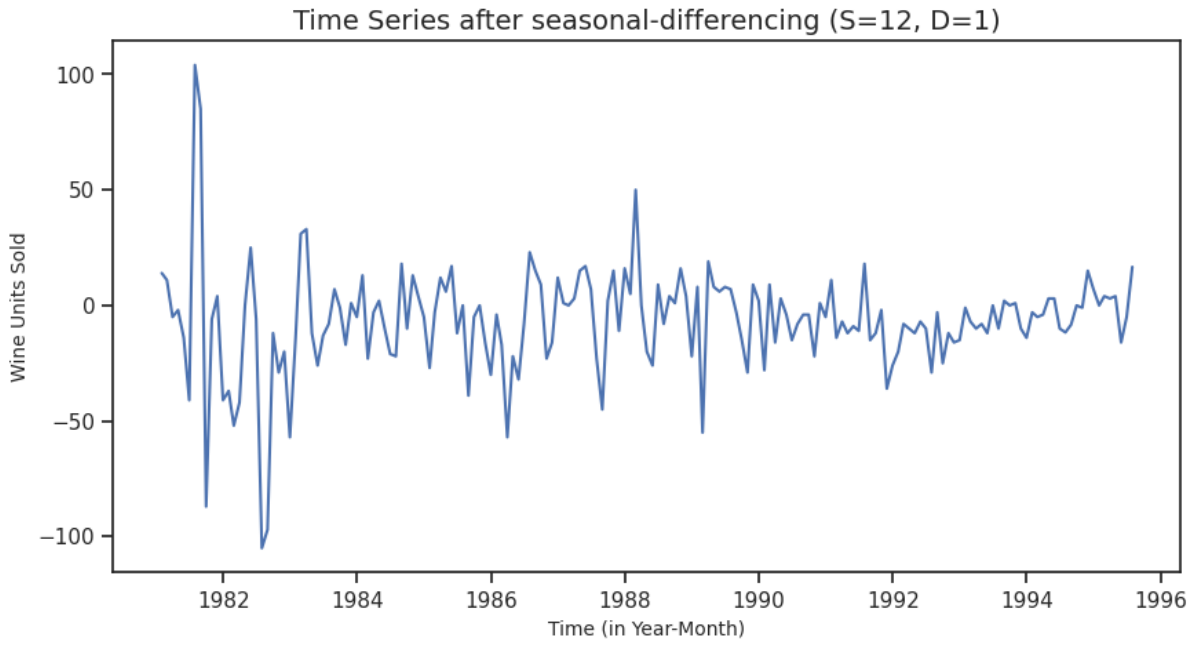


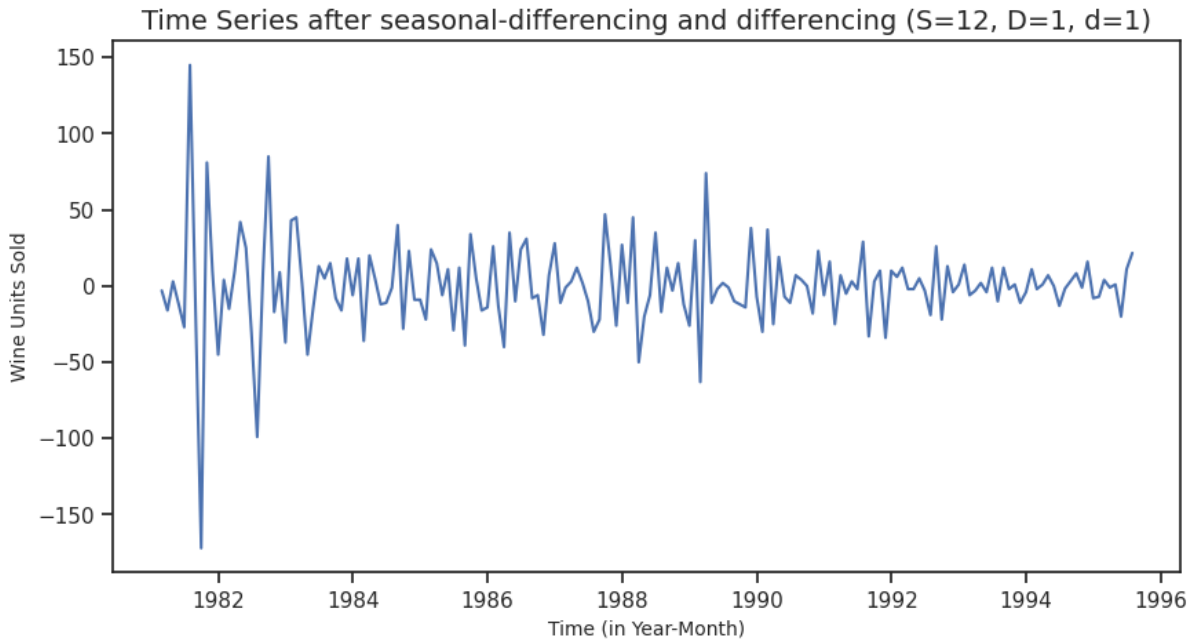
PACF plot



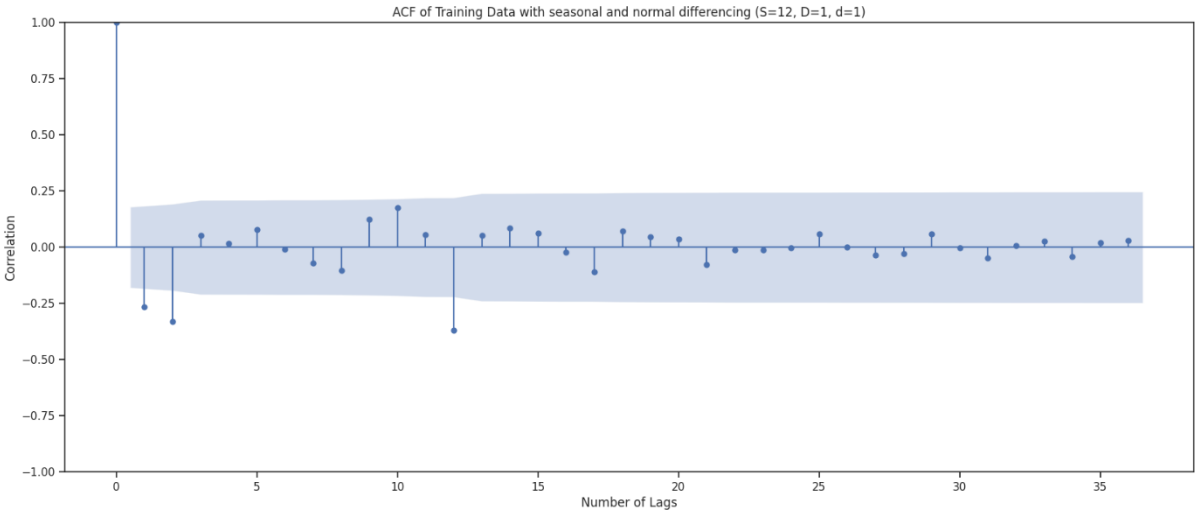
We see that our ACF plot at the seasonal interval (12) does not taper off quickly. So, we go ahead and take a seasonal differencing of the original series. Before that let us look at the original series.

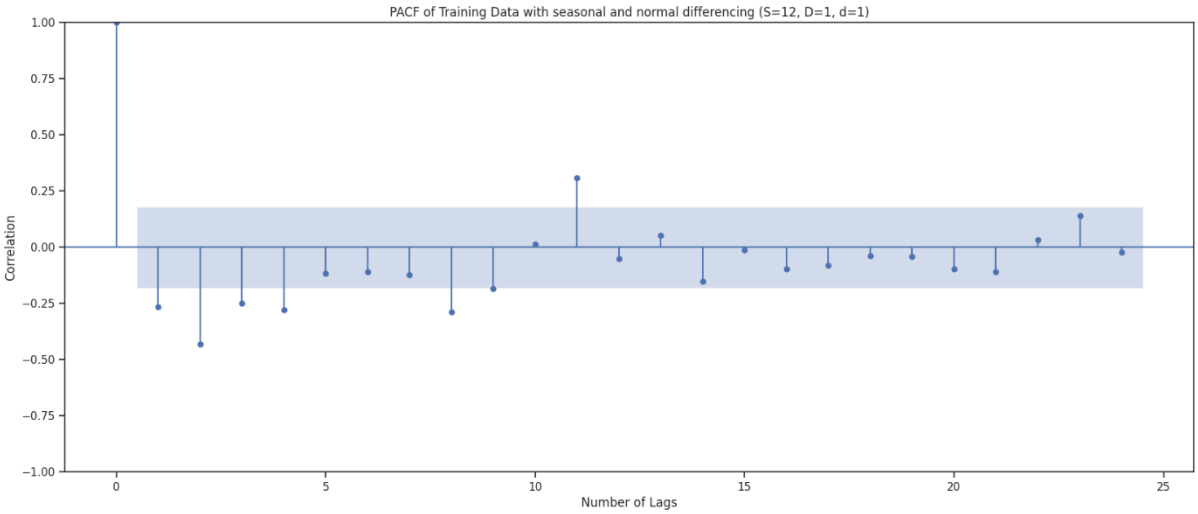

Sales of Rose wine
Rose_Wine_Sales
Units Sold
250
200
150
100
50
1981
1983
1985
1987
1989
1991
1993
1995
Time (in Year-Month)


Time Series after seasonal-differencing (S=12, D=1)
Wine Units Sold
100
50
0
−50
−100
1982
1984
1986
1988
1990
1992
1994
1996
Time (in Year-Month)



Let us look at the ACF and the PACF plots once more with seasonal and normal differencing on train data





- Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till lag 4 is significant before cut-off, so AR term ‘p = 4’ is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’.
- From ACF plot, lag 1 and 2 are significant before it cuts off, so lets keep MA term ‘q = 2’ and at seasonal lag of 12, a significant lag is apparent and no seaonal lags are apparent at lags 24, 36 or afterwards, so lets keep ‘Q = 1'.
- The final selected terms for SARIMA model is (4, 1, 2)x(0, 1, 1, 12), as inferred from the ACF and PACF plots.

SARIMAX Results

==========

Dep. Variable: Rose_Wine_Sales No. Observations: 132

Model: SARIMAX(4, 1, 2)x(0, 1, [1], 12) Log Likelihood -446.102

Date: Sun, 04 May 2025 AIC 908.203

Time: 16:08:33 BIC 929.358

Sample: 01-31-1980 HQIC 916.774

- 12-31-1990

Covariance Type: opg

==========

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8046	0.119	-6.778	0.000	-1.037	-0.572
ar.L2	0.0387	0.140	0.276	0.783	-0.237	0.314
ar.L3	-0.2310	0.147	-1.568	0.117	-0.520	0.058
ar.L4	-0.1875	0.108	-1.741	0.082	-0.398	0.024
ma.L1	0.1434	1054.999	0.000	1.000	-2067.617	2067.904
ma.L2	-0.8566	903.712	-0.001	0.999	-1772.099	1770.386
ma.S.L12	-0.5406	0.085	-6.385	0.000	-0.707	-0.375
sigma2	296.7693	3.13e+05	0.001	0.999	-6.13e+05	6.14e+05

==========

Ljung-Box (L1) (Q): 0.01 Jarque-Bera (JB): 0.03

Prob(Q): 0.94 Prob(JB): 0.98

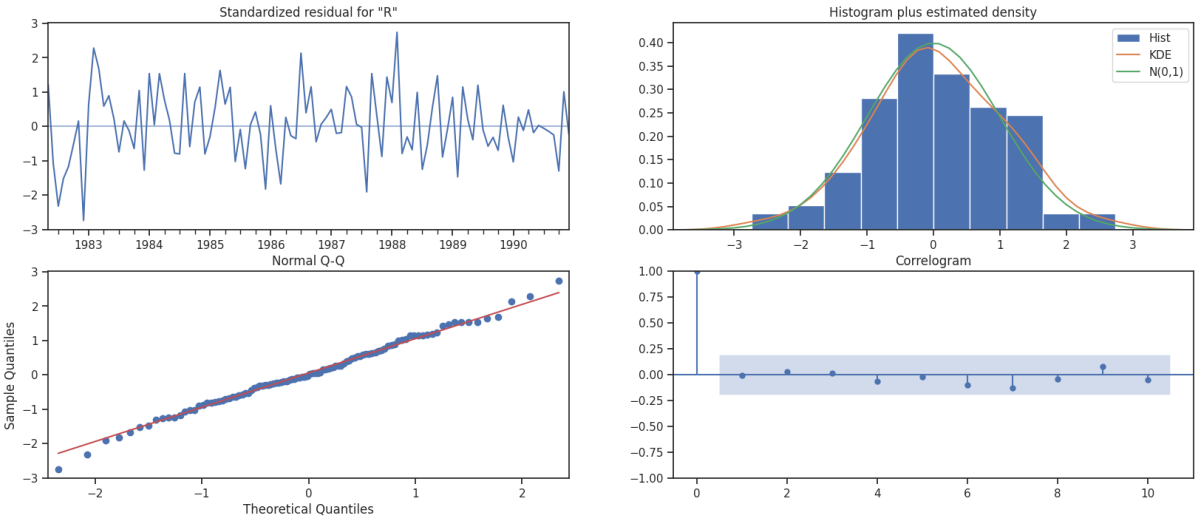
Heteroskedasticity (H): 0.55 Skew: -0.02

Prob(H) (two-sided): 0.08 Kurtosis: 3.07

==========

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).



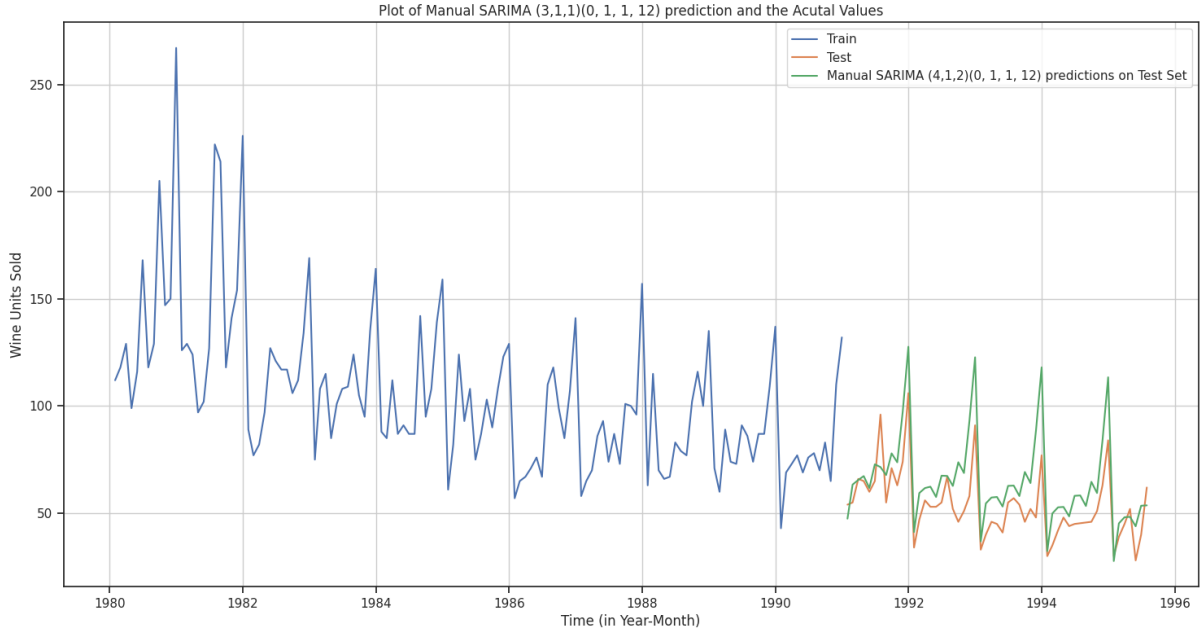
Predict on the Test Set using this model and evaluate the model.

	predicted_mean
1991-01-31	47.471764
1991-02-28	63.350018
1991-03-31	65.513285
1991-04-30	67.317182
1991-05-31	61.756416
1991-06-30	72.816609
1991-07-31	71.513832
1991-08-31	67.808638
1991-09-30	77.918588
1991-10-31	73.748671
1991-11-30	97.305484

	predicted_mean
1991-12-31	127.634754
1992-01-31	41.226569
1992-02-29	59.413092
1992-03-31	61.742040
1992-04-30	62.374894
1992-05-31	57.582093
1992-06-30	67.508022
1992-07-31	67.387086
1992-08-31	62.729133
1992-09-30	73.769516
1992-10-31	68.753681
1992-11-30	93.025714
1992-12-31	122.710417
1993-01-31	36.869550
1993-02-28	54.567994
1993-03-31	57.326451
1993-04-30	57.584591
1993-05-31	53.116324
1993-06-30	62.758893

	predicted_mean
1993-07-31	62.884565
1993-08-31	58.012490
1993-09-30	69.239330
1993-10-31	64.061327
1993-11-30	88.474298
1993-12-31	118.036386
1994-01-31	32.302144
1994-02-28	49.907891
1994-03-31	52.746966
1994-04-30	52.935004
1994-05-31	48.527690
1994-06-30	58.117254
1994-07-31	58.289016
1994-08-31	53.376864
1994-09-30	64.638554
1994-10-31	59.430248
1994-11-30	83.869568
1994-12-31	113.408745
1995-01-31	27.694425

	predicted_mean
1995-02-28	45.282848
1995-03-31	48.136987
1995-04-30	48.311927
1995-05-31	43.916002
1995-06-30	53.495663
1995-07-31	53.676037



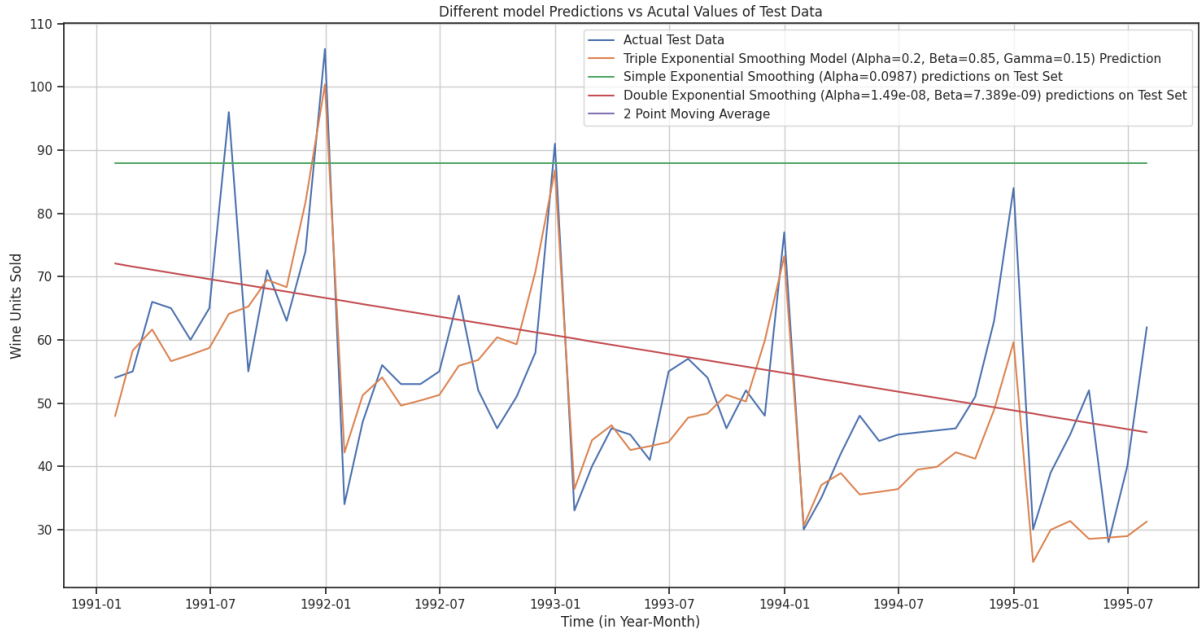
index	Test RMSE	MAPE
Auto ARIMA (2,1,3)	36.8169146419244	75.84685520971229
Auto SARIMA (3,1,1)(3,0,2,12)	18.8818485219433	75.84685520971229
Manual ARIMA(2,1,2)	36.870991041964494	75.84685520971229
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	15.907182350411736	23.71261553502886

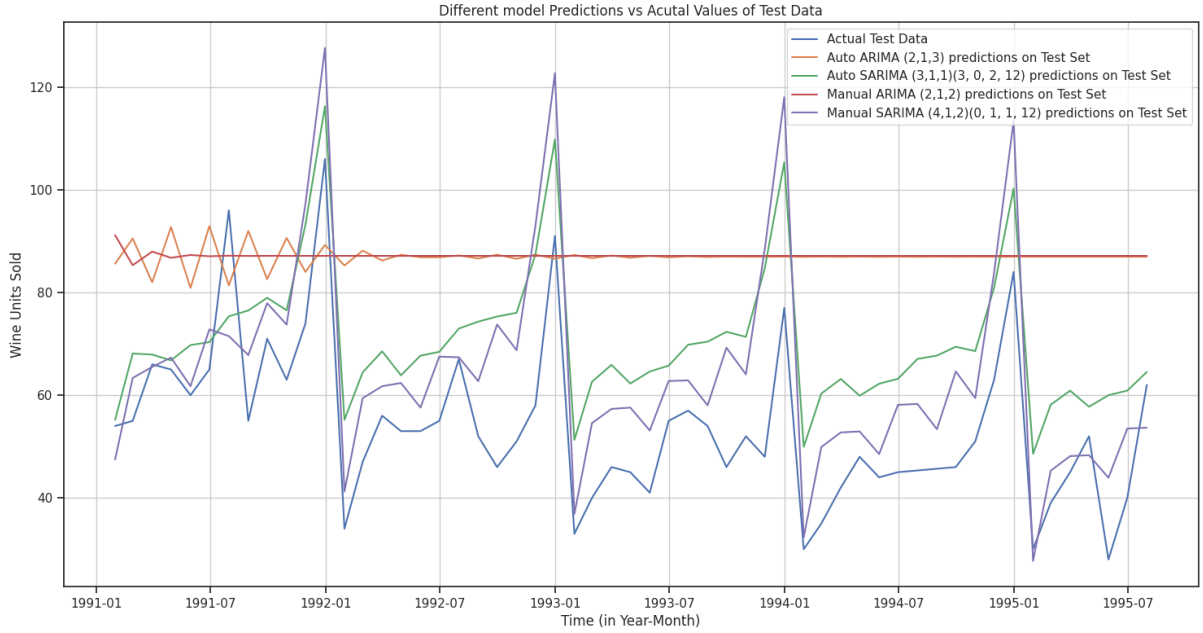
8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

index	Test RMSE	MAPE
Linear Regression	15.268887473349798	NaN
Naive Model	79.71857596912689	NaN
Simple Average	53.460367314208234	NaN
2 point TMA	11.529277632500342	NaN
4 point TMA	14.451376090500547	NaN
6 point TMA	14.566261685653924	NaN
9 point TMA	14.727595576680468	NaN
Alpha=0.0987,SimpleExponentialSmoothing	37.59200650745667	NaN
Alpha=1.49e-08, Beta=7.389e-09, Double Exponential Smoothing	15.270900637349518	NaN
Alpha=0.064,Beta=0.053,Gamma=0.0,Triple Exponential Smoothing	19.112865532946618	NaN
Alpha=0.2,Beta=0.85,Gamma=0.15,Triple Exponential Smoothing	10.27987620009132	NaN
Auto ARIMA (2,1,3)	36.8169146419244	75.84685520971229
Auto SARIMA (3,1,1)(3,0,2,12)	18.8818485219433	75.84685520971229
Manual ARIMA(2,1,2)	36.870991041964494	75.84685520971229
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	15.907182350411736	23.712615535028866

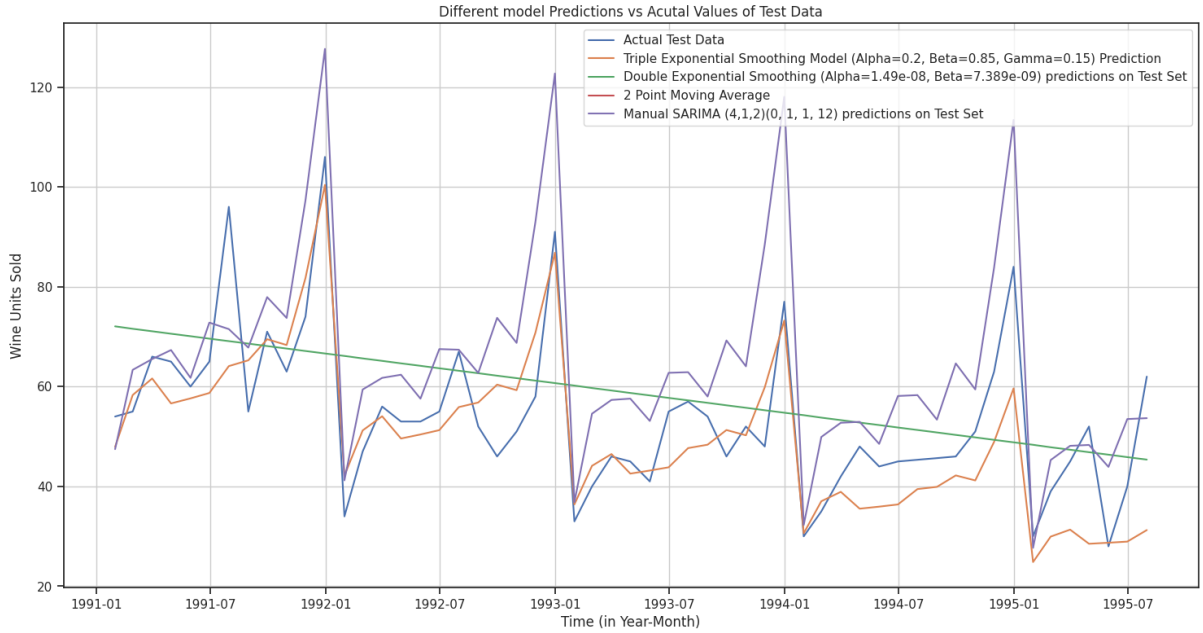
9) Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

we observed the Triple Exponential Smoothing model is the optimum model for the given data set as it has the lowest RMSE value. However, as we know SARIMA models tend to perform better with seasonal time series, we are also considering SARIMA model for the forecast.

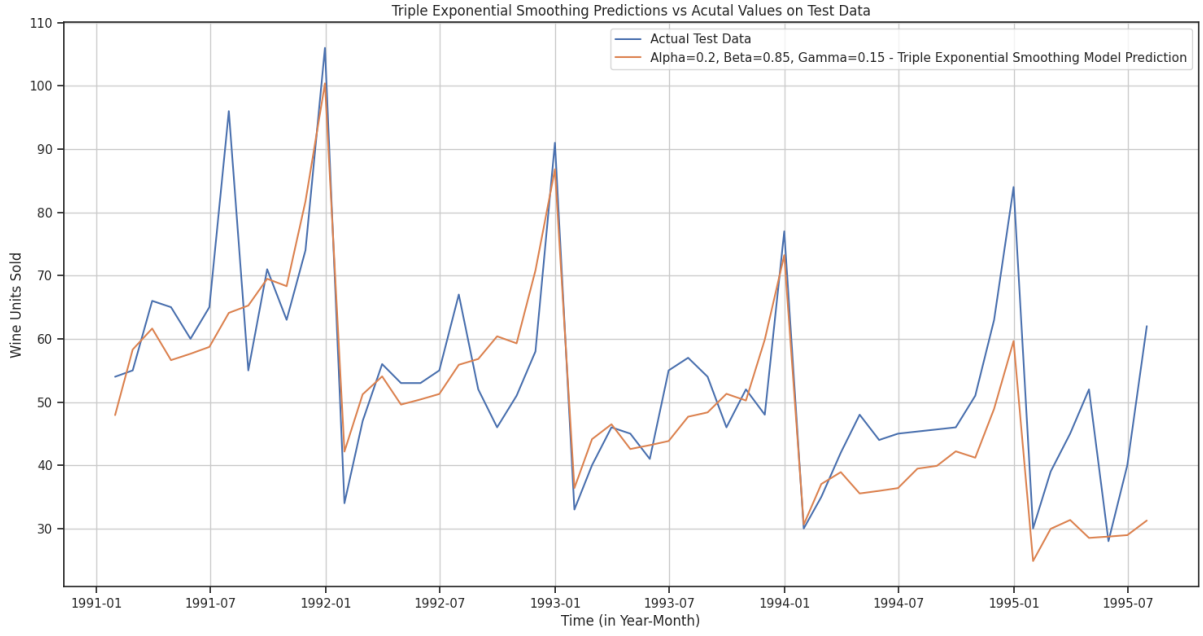




Plotting on both the Training data

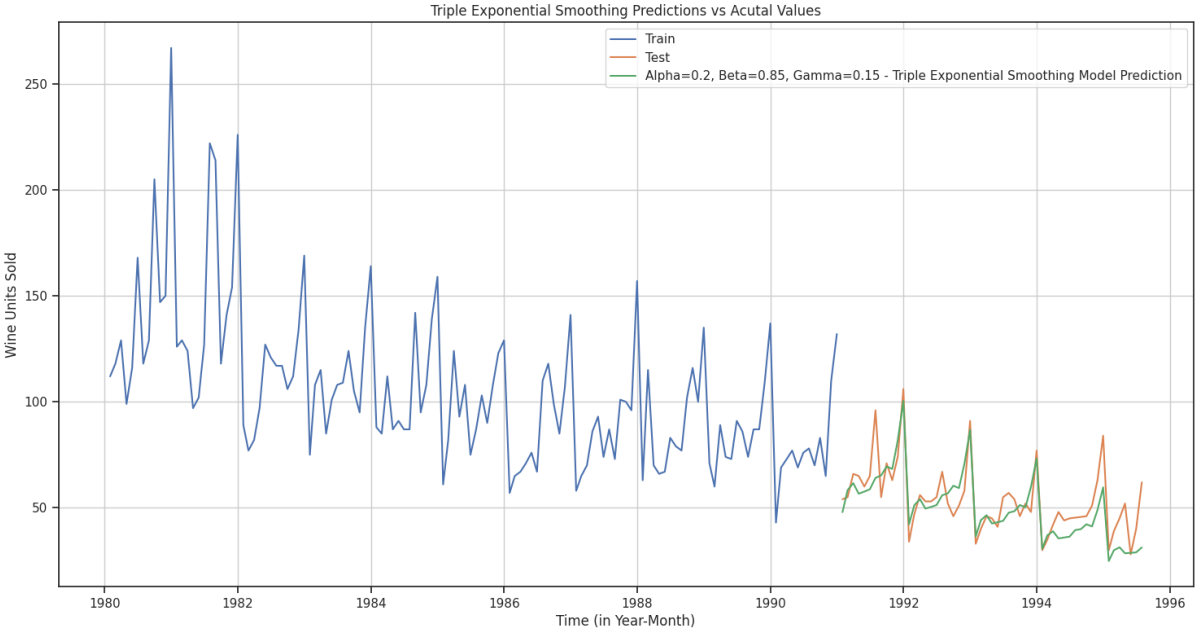


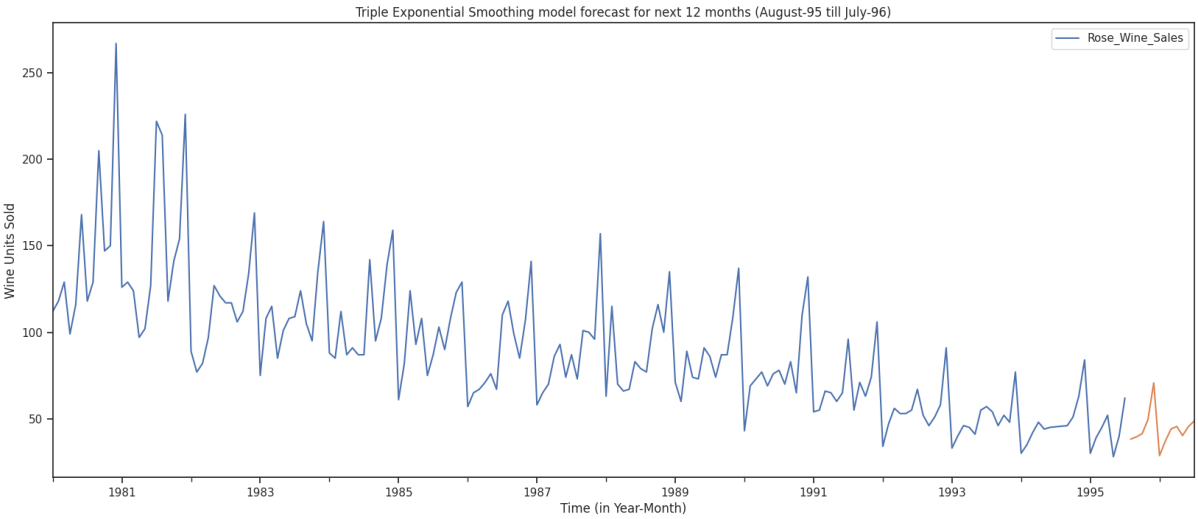
Optimum Model - Triple Exponential Smoothing Model (Alpha=0.2, Beta=0.85, Gamma=0.15)



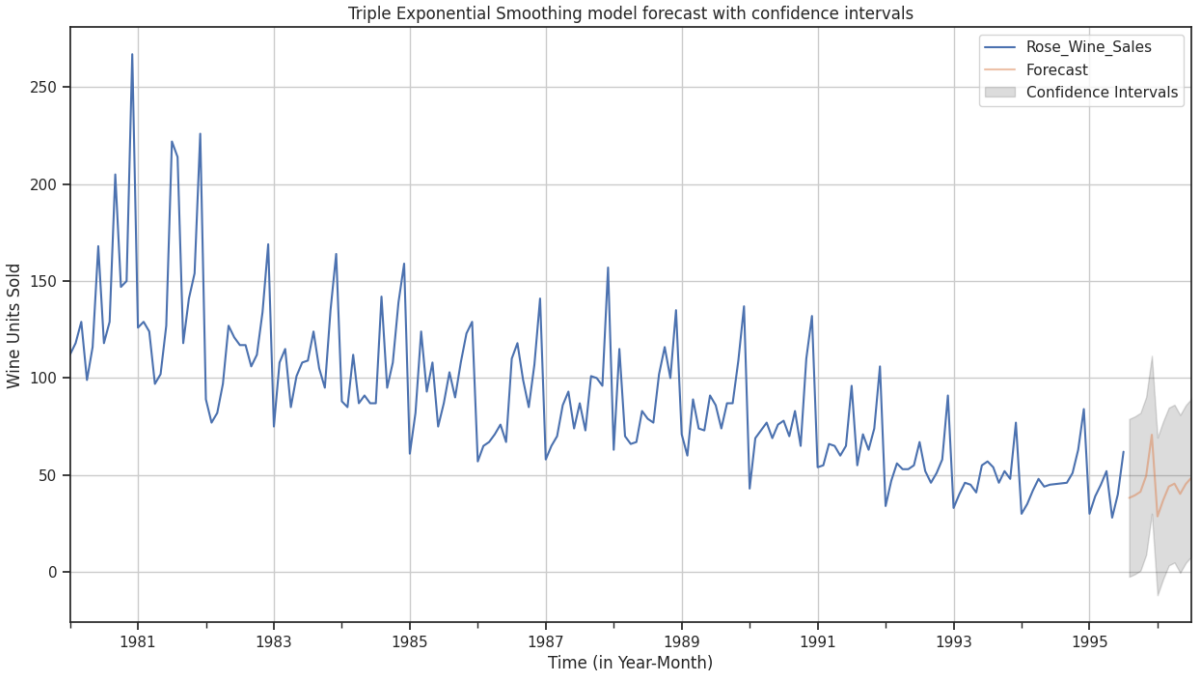
Optimum Model 1:

Triple Exponential Smoothing Model (Alpha=0.2, Beta=0.85, Gamma=0.15)

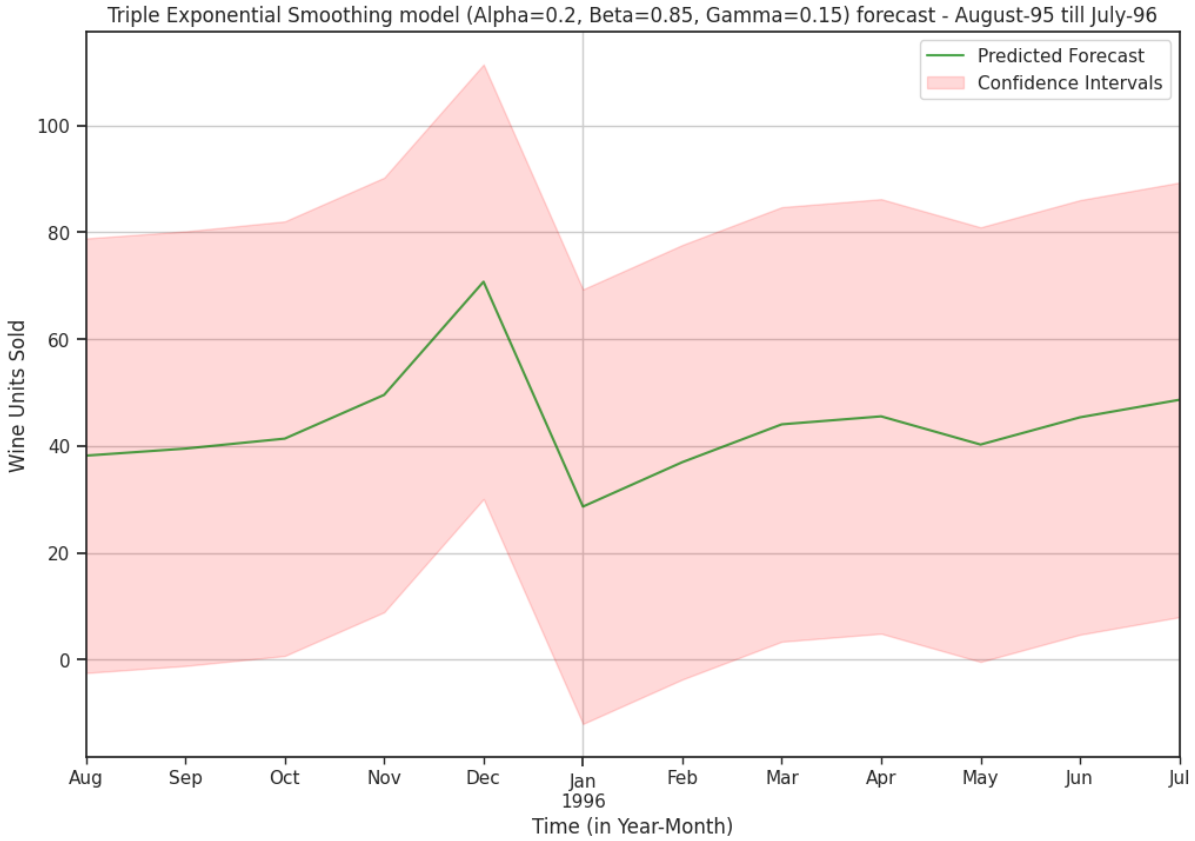




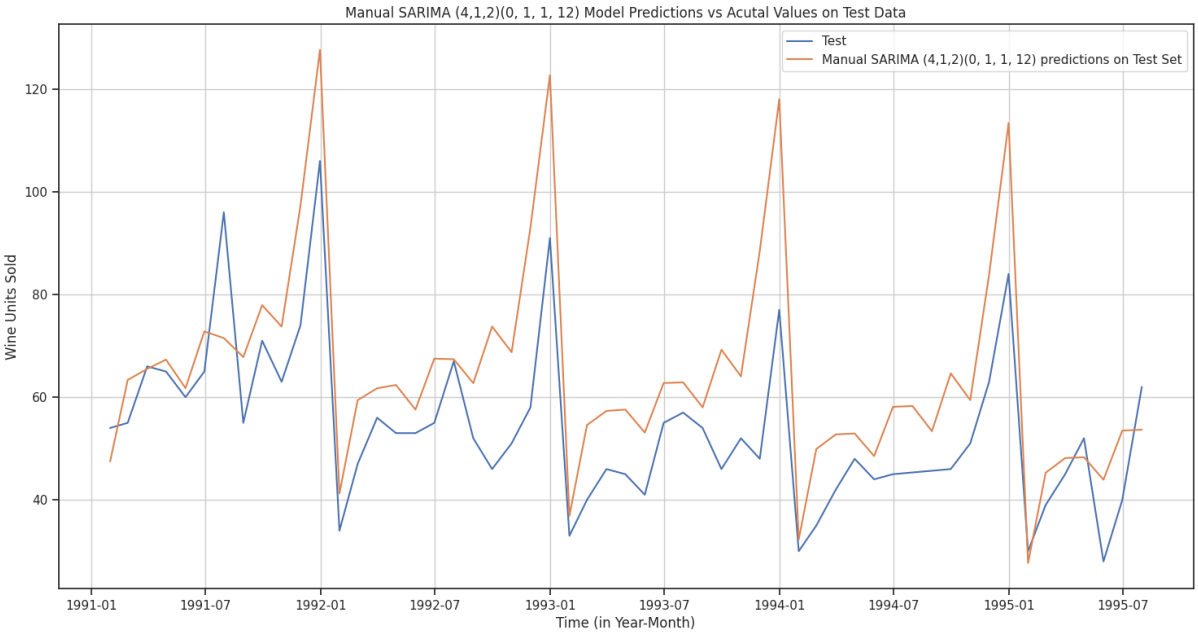
index	lower_ci	prediction	upper_ci
1995-08-31 00:00:00	-2.4896464666249045	38.19283765590618	78.87532177843727
1995-09-30 00:00:00	-1.1745442150162972	39.50793990751479	80.19042403004588
1995-10-31 00:00:00	0.6937525010717636	41.37623662360285	82.05872074613393
1995-11-30 00:00:00	8.885705699608273	49.56818982213936	90.25067394467044
1995-12-31 00:00:00	30.07369774375038	70.75618186628147	111.43866598881255
1996-01-31 00:00:00	-12.031866834957373	28.650617287573713	69.3331014101048
1996-02-29 00:00:00	-3.7116891877222997	36.97079493480879	77.65327905733987
1996-03-31 00:00:00	3.36925849444831	44.051742616979396	84.73422673951049
1996-04-30 00:00:00	4.8631004568024565	45.54558457933354	86.22806870186463
1996-05-31 00:00:00	-0.41669782306095016	40.265786299470136	80.94827042200123
1996-06-30 00:00:00	4.6823778239876575	45.364861946518744	86.04734606904984
1996-07-31 00:00:00	7.964693381497312	48.6471775040284	89.32966162655948

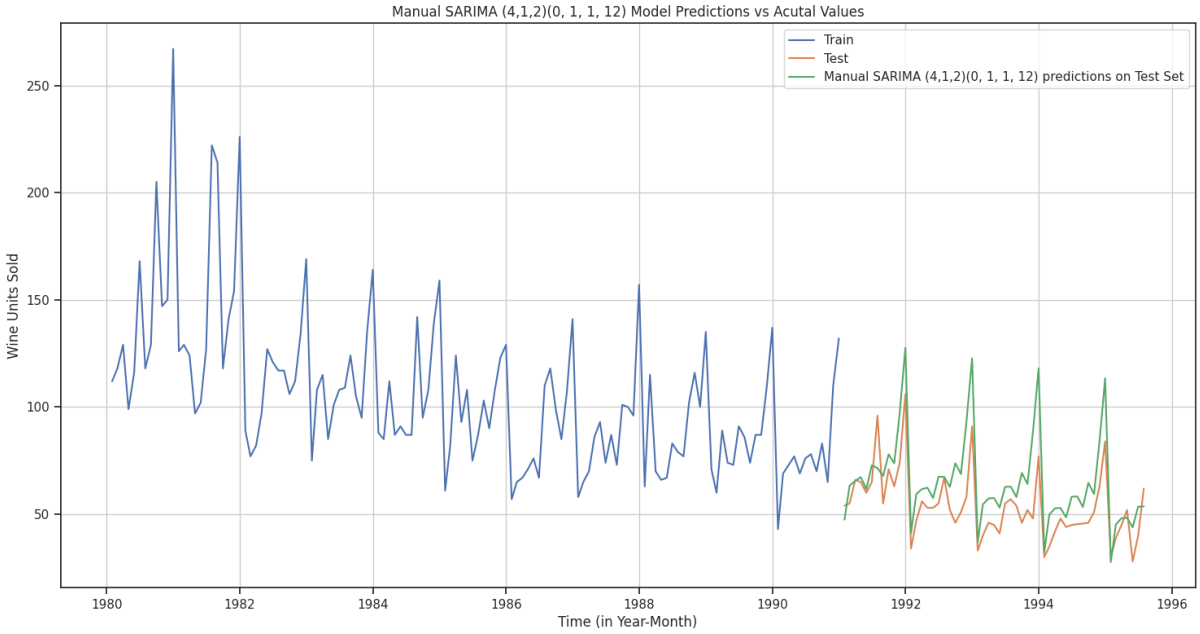


TES Optimum Model – Time series plot forecast with confidence intervals



Optimum Model 2: Manual SARIMA Model (4, 1, 2) (0, 1, 1, 12)





SARIMAX Results

==========

Dep. Variable: Rose_Wine_Sales No. Observations: 187

Model: SARIMAX(4, 1, 2)x(0, 1, [1], 12) Log Likelihood -658.935

Date: Sun, 04 May 2025 AIC 1333.870

Time: 16:08:41 BIC 1358.421

Sample: 01-31-1980 HQIC 1343.840

- 07-31-1995

Covariance Type: opg

==========

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8239	0.083	-9.939	0.000	-0.986	-0.661
ar.L2	0.0471	0.107	0.439	0.661	-0.163	0.258
ar.L3	-0.2146	0.110	-1.954	0.051	-0.430	0.001
ar.L4	-0.1693	0.078	-2.182	0.029	-0.321	-0.017
ma.L1	0.1564	81.834	0.002	0.998	-160.235	160.547
ma.L2	-0.8436	69.031	-0.012	0.990	-136.143	134.456
ma.S.L12	-0.5418	0.061	-8.899	0.000	-0.661	-0.423

sigma2 225.0273 1.84e+04 0.012 0.990 -3.59e+04 3.63e+04

==========
=

Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 3.12

Prob(Q): 0.88 Prob(JB): 0.21

Heteroskedasticity (H): 0.24 Skew: 0.04

Prob(H) (two-sided): 0.00 Kurtosis: 3.68

==========
=

Warnings:

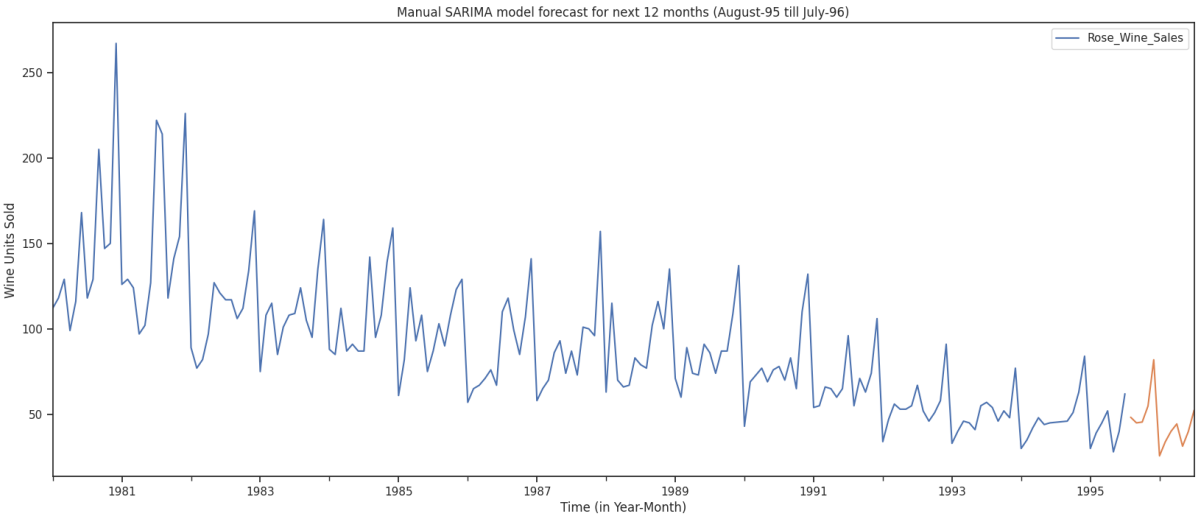
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

index	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31 00:00:00	48.27253248634685	15.053102427094847	18.768993873648473	77.77607109904523
1995-09-30 00:00:00	44.98571142539495	15.830702741595937	13.958104201907421	76.01331864888249
1995-10-31 00:00:00	45.474487743412176	15.889167551181327	14.33229159877429	76.61668388805006
1995-11-30 00:00:00	54.80669576090882	15.89916314770612	23.644908607078346	85.9684829147393
1995-12-31 00:00:00	81.9042219783203	15.913863548702016	50.71362256797957	113.09482138866102

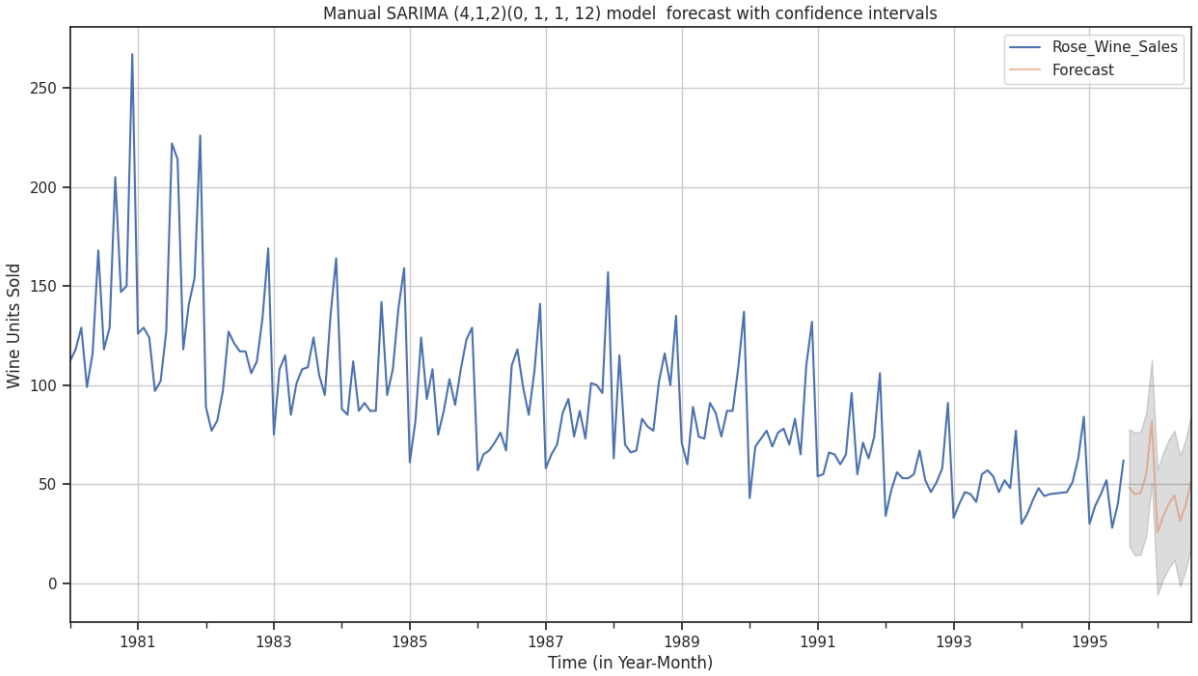
index	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31 00:00:00	48.27253248634685	15.053102427094847	18.768993873648473	77.77607109904523
1995-09-30 00:00:00	44.98571142539495	15.830702741595937	13.958104201907421	76.01331864888249
1995-10-31 00:00:00	45.474487743412176	15.889167551181327	14.33229159877429	76.61668388805006
1995-11-30 00:00:00	54.80669576090882	15.89916314770612	23.644908607078346	85.9684829147393
1995-12-31 00:00:00	81.9042219783203	15.913863548702016	50.71362256797957	113.09482138866102
1996-01-31 00:00:00	25.670657016958035	16.19962672438213	-6.0800279258235115	57.42134195973958
1996-02-29 00:00:00	33.89450887919349	16.307730507317096	1.9319444152668765	65.8570733431201
1996-03-31 00:00:00	40.04754404328722	16.589672238679416	7.53238394015159	72.56270414642285
1996-04-30 00:00:00	44.382924130047435	16.65767400775607	11.734483008636552	77.03136525145831
1996-05-31 00:00:00	31.338504960826413	16.871987078059078	-1.7299820597945654	64.40699198144739
1996-06-30 00:00:00	39.915062049499234	16.946055207673822	6.701404162431118	73.12871993656735
1996-07-31 00:00:00	52.37497998758772	17.155986794802033	18.749863750530977	86.00009622464447

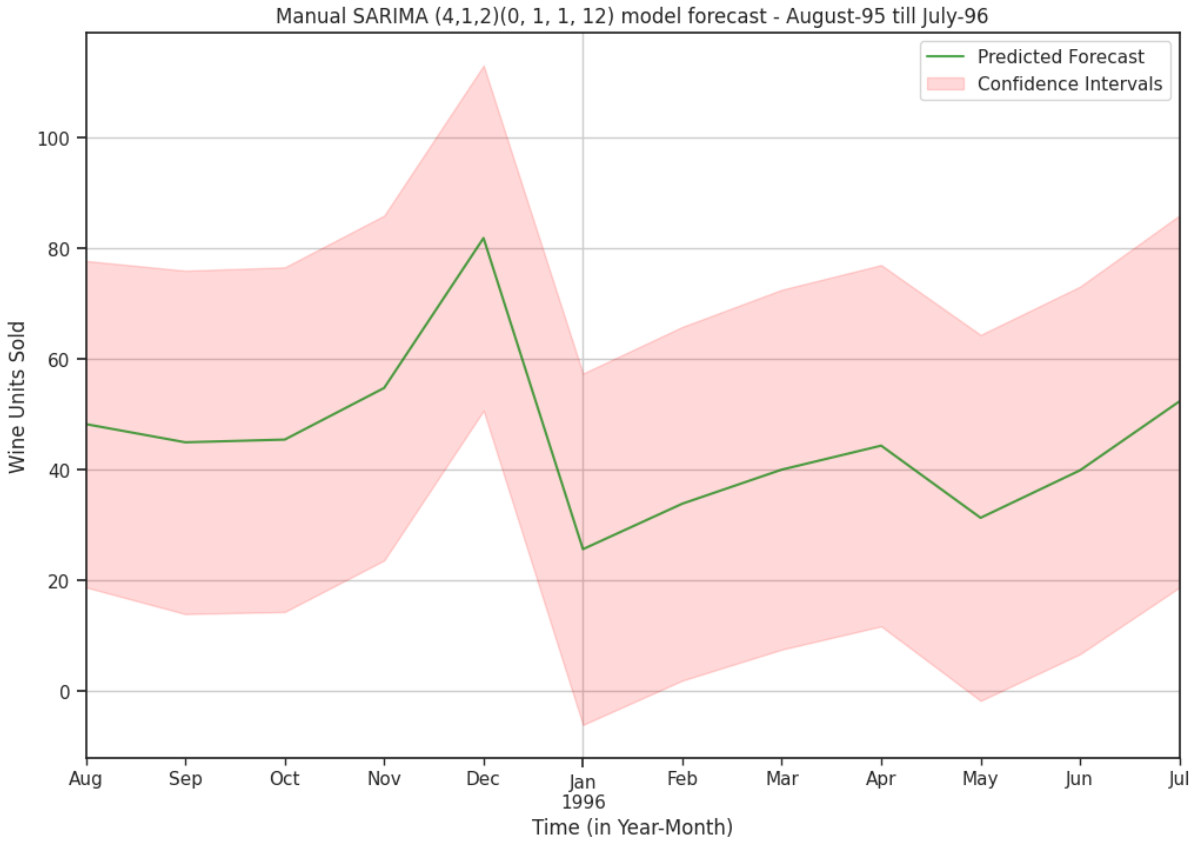
	predicted_mean
1995-08-31	48.272532
1995-09-30	44.985711
1995-10-31	45.474488
1995-11-30	54.806696
1995-12-31	81.904222
1996-01-31	25.670657
1996-02-29	33.894509

	predicted_mean
1996-03-31	40.047544
1996-04-30	44.382924
1996-05-31	31.338505
1996-06-30	39.915062
1996-07-31	52.374980



Manual SARIMA Optimum Model – Time series plot forecast with confidence intervals




Manual SARIMA (4,1,2)(0, 1, 1, 12) model forecast - August-95 till July-96
Predicted Forecast
Confidence Intervals
Wine Units Sold
100
80
60
40
20
0
Aug
Sep
Oct
Nov
Dec
Jan
1996
Feb
Mar
Apr
May
Jun
Jul
Time (in Year-Month)

10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

We needed to construct an optimum model to forecast the rose wine sales for the next 12 months. The model information, insights and recommendations are as follows.

Model Insights:

• The time series in consideration exhibits a declining trend and stable seasonality. When comparing the various models, we can see that Triple Exponential Smoothing and SARIMA models frequently deliver the greatest results. This is due to the fact that these models are excellent at predicting time series that demonstrate trend and seasonality. Apart from these Double Exponential Smoothing and Moving Average Models also tend to perform moderately good.

• We examine the root mean squared value of the forecast model to assess its performance (RMSE). The model with the lowest RMSE value and characteristics that match the test data is regarded as being a superior model.

• We observed that Triple Exponential Smoothing model had the lowest RMSE and the characteristics that most closely fit test data. As a result, its regarded as the best model for forecasting and can thus be used by the company for forecast analysis.

Historical Insights:

• The rose wine sales have declined throughout time. Rose wine sales peaked in 1980 & 1981 and fell to their present low position in 1995 (as we have data for only first 7 months).

• The monthly sales trajectory appears to be exactly the opposite of the yearly plot, with a progressive increase towards the end of each year. January has the lowest wine sales, while December has the highest. From January to August, sales increase gradually, and then they quickly increase after that.

• The average monthly sales of Rose wine are 90 bottles. More than 50% of the sold units of rose wine fall between 62 and 111. 28 units were sold as the lowest and 267 units as the most. Only 20% of monthly sales that were recorded were for more than 120 units.

• Around 70 to 75 percent of the units sold are fewer than 100, and 90% of the units sold are less than 150. Only 15% of sales involved more than 50 items. Therefore, it is clear that the bulk of sales were in the range of 50 to 100 units.

Forecast Insights:

• Based on the forecast made by the Triple Exponential Smoothing model previously presented, the following insights are offered.

• The forecast calls for average sale of 44 units, down by 45 units from the historical average of 89 units. Thus, we might observe an alarming decrease in average sales by 50%.

• The prediction for minimum sales volume of 28 units equals the minimum sales volume in the past. Consequently, a no percentage change could be seen in minimum quantity sold.

• The projection estimates a maximum sales volume of 70 units, which is 197 units fewer than the largest sales volume recorded in the past, which was 267 units. Consequently, a 73% decrease in maximum sales is visible.

• In comparison to the historical standard deviation of 62 recorded in the past, the forecast's standard deviation is 10 units, or 52 units lower. It's gone down by 83%. This is not anticipated because historical data tends to have less volatility than future data.

• We can see from the prediction that the months of October, November, and December have increased sales. December is often when the sales are at their highest. There is a startling decline in sales in January following December. The months after January appear to witness a gradual improvement in sales until October, when it jumps sharply.

Recommendations:

• Records show that the months of September, October, November, and December account for 40% of the total sales forecast. Many festivities take place in these months, and many people travel during this time. One of the most premium types of wine used during festive and event celebrations is rose wine.

• Wine sales often climb in the final two months of the year as people hurry to buy holiday beverages. For forthcoming occasions like Thanksgiving, Christmas, and New Year's, people typically stock up. The majority of individuals also buy in bulk for holiday gatherings and gift-giving. • Many individuals choose wine as their go-to gift when it comes to occasions like parties and gift-giving. Sales of Rose wine rise just before the winter holidays as more collectors purchase these wines as presents or look for vintages to serve at holiday gatherings.

• This blush wine works nicely with nearly anything, including spicy dishes, sushi, salads, grilled meats, roasts, and rich sauces. It is well renowned for its outdoor-friendly drinking style.

• The festival seasons may vary depending on where you are geographically, however the most of the celebrations take place in the last four months.

▪ In these months, promotional offers might be implemented to lower costs and significantly boost revenue.

▪ To increase sales, we must take advantage of all holiday events and set prices appropriately.

▪ Many individuals order in bulk to prepare for upcoming festivities, which may result in a high shipping expenditure. Businesses may provide significant discounts or free shipping beyond a certain threshold at these times.

▪ Giving customers gifts to improve their user experience is one of the greatest marketing strategies to deploy. In order to attract more consumers and increase sales, the company might provide free gifts on orders with significant sales.

▪ To target various client demographics, the proper marketing campaigns must be run

▪ Numerous ecommerce campaigns and competitions may be performed to broaden the product's audience and enhance sales.

• The period from January to June is one of the key challenges for Rose wine sales.

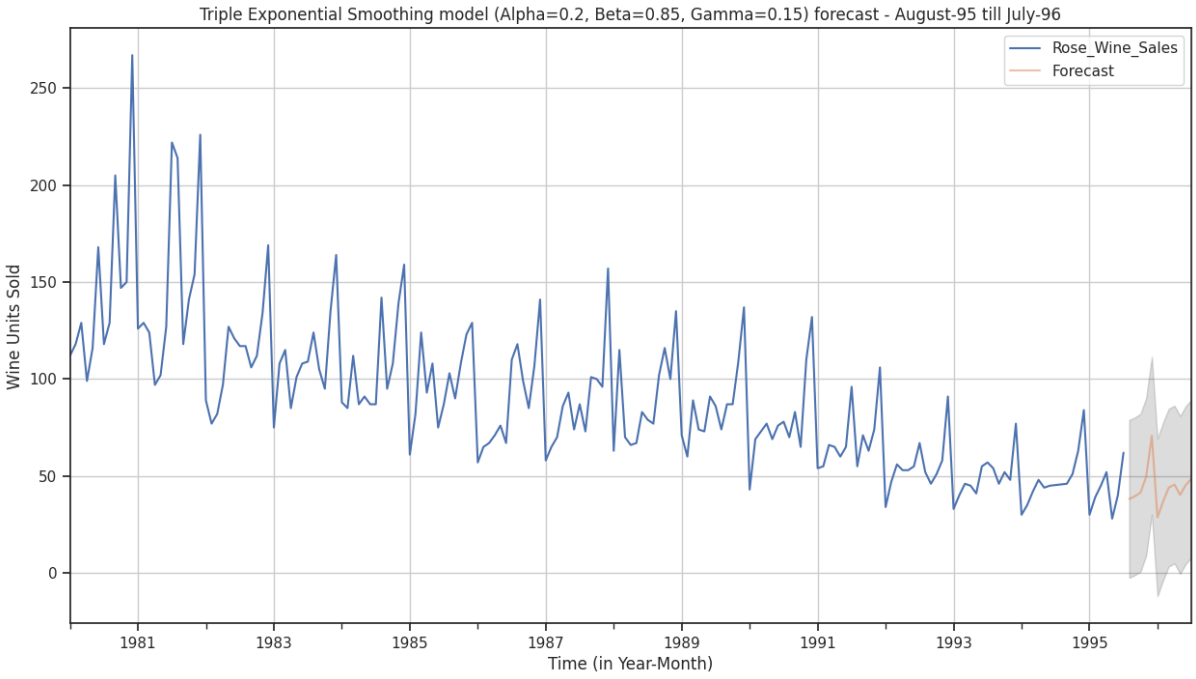
▪ To identify the elements affecting sales, in-depth market research must be conducted.

▪ Due to the fact that rose wines are premium category of wine, a marketfriendly version of the existing product might be introduced by the company, helping to make up for the drop in sales. Long-term, this may bring in additional clients.

▪ The company can rebrand its product to instill a fresh perspective towards the product and break the declining sales trend.

• There are other key elements that might be driving the sales, despite the present model's ability to closely track the historical sales trend.

▪ The forecast might be improved by doing in-depth market research on the factors that influence sales and incorporating that information into the model for projection



Sparkling Wine Analysis

Executive Summary

Data on wine sales from the 20th century are available from ABC Estate Wines, a wine producing firm, and should be examined. With the provided information, an estimate of wine sales in the 20th century must be forecasted.

Introduction

The purpose of this report is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of sales of Sparkling wine from 20th century.

DATASET

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

	YearMonth	Sparkling
182	1995-03	1897
183	1995-04	1862
184	1995-05	1670
185	1995-06	1688
186	1995-07	2031

Dataset has 2 columns which captures the Year and Month of recorded data and the number of units sold on corresponding Year-Month respectively.

1. Read the data as an appropriate Time Series data and plot the data.

Let us check the types of variables in the data frame and check for missing values in the dataset

	YearMonth	Sparkling	Time_Stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

The dataset has 2 variables and 187 rows in total. The "YearMonth" column can be deleted after creating a suitable time stamp column because it is not necessary for our modelling. The column Sparkling is of float type. Additionally, we can observe from the data above that Sparkling column has no missing values.

Time Stamp created from 'YearMonth' column

DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',

'1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',

'1980-09-30', '1980-10-31',

...

'1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',

'1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',

'1995-06-30', '1995-07-31'],

dtype='datetime64[ns]', length=187, freq='ME')

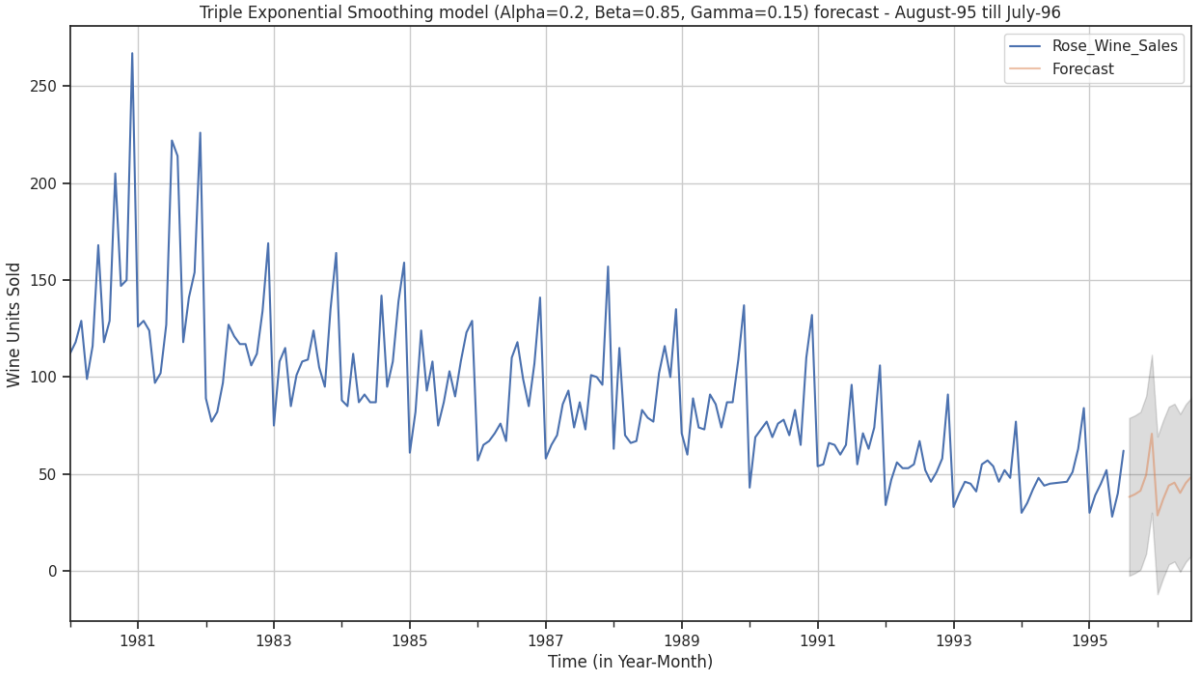
Renaming the column

	Sparkling_Wine_Sales
Time_Stamp	
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Time_Stamp column has been set as index of the dataset and column Sparkling has been renamed as Sparkling_Wine_Sales.

Checking null values in the dataset

Null values in the dataset As can be seen from the above figure, there are no null values present in the dataset

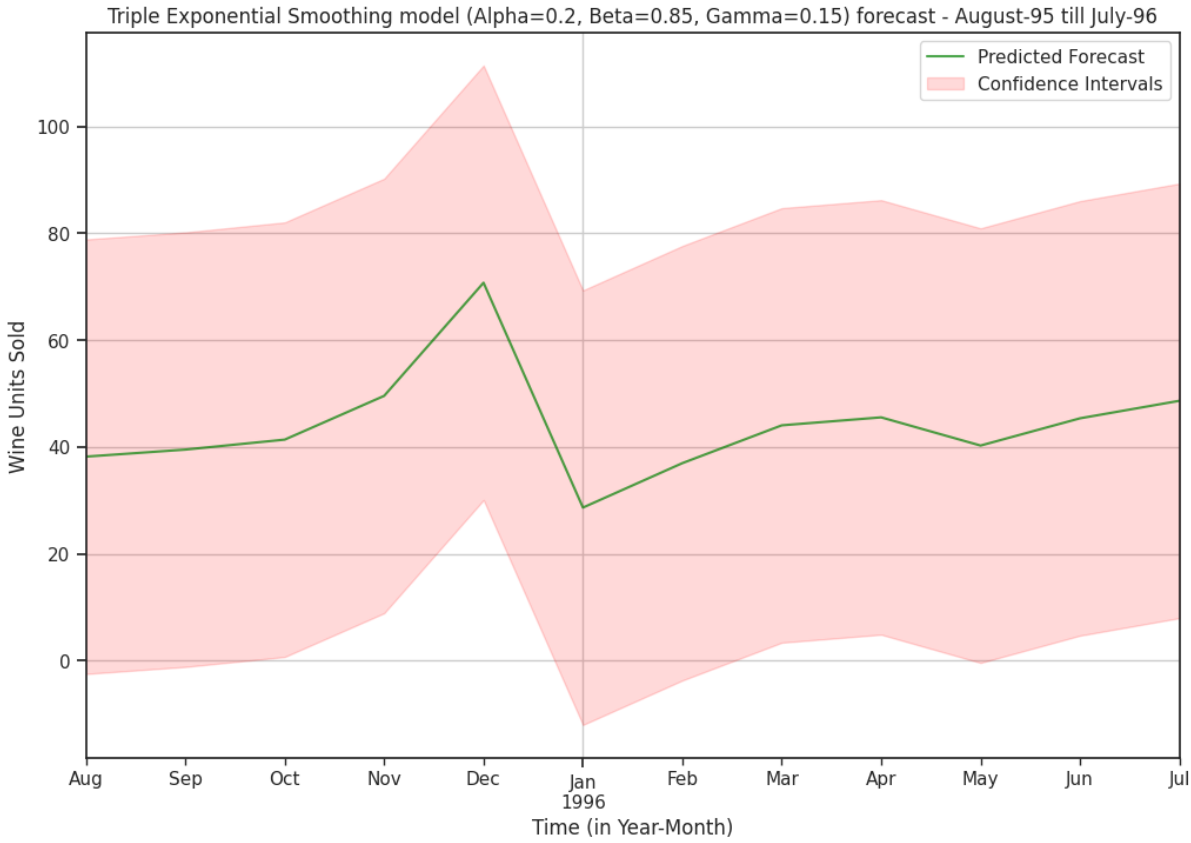


Observation:

• The data set provided contains sales information from January 1980 to July 1995.

• We can see from the plot that there has been a constant pattern of sales with seasonality. Over the years, the sales have consistent. The data also exhibits seasonality, as may be shown.

• There are no missing values which must be imputed

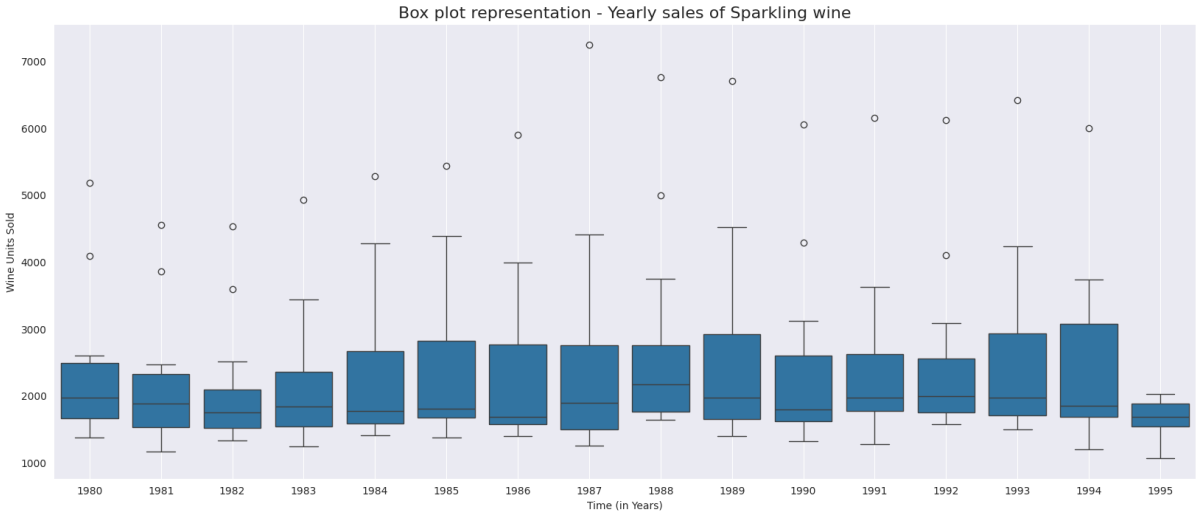


Observation:

- 2402 bottles of sparkling wine are typically sold each month.
- Between 1605 and 2549 units make up more than 50% of the sold sparkling wine units.
- The lowest unit sold is 1070 units, while the highest unit sold is 7242 units.
- Only 25% of monthly sales that recorded are more than 2549 units.

2) Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Exploratory Analysis



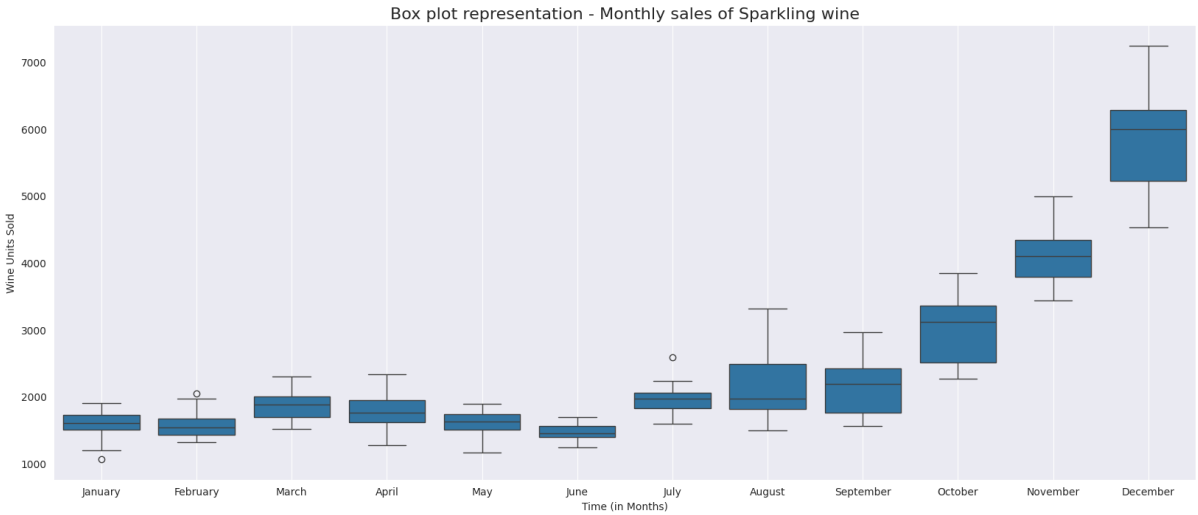
Observation:

• We can see from the figure above that sales of sparkling wine have remained constant over the years.

• The median sales of sparkling wine reached their peak in 1988 and their current low point in 1995.

• Additionally, we can see that there are outliers in the box plots.

Monthly Plot



1. December month records the highest average sales
2. The sales seems to usually pick in the last 4 months

Observation:

• We can see from the figure above that sales of sparkling wine have remained constant over the years.

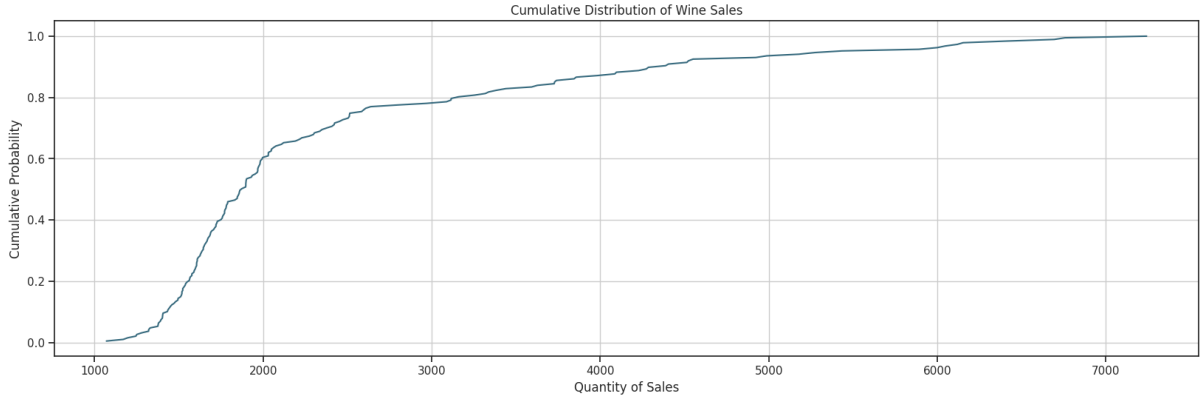
• The median sales of sparkling wine reached their peak in 1988 and their current low point in 1995.
• Additionally, we can see that there are outliers in the box plots.

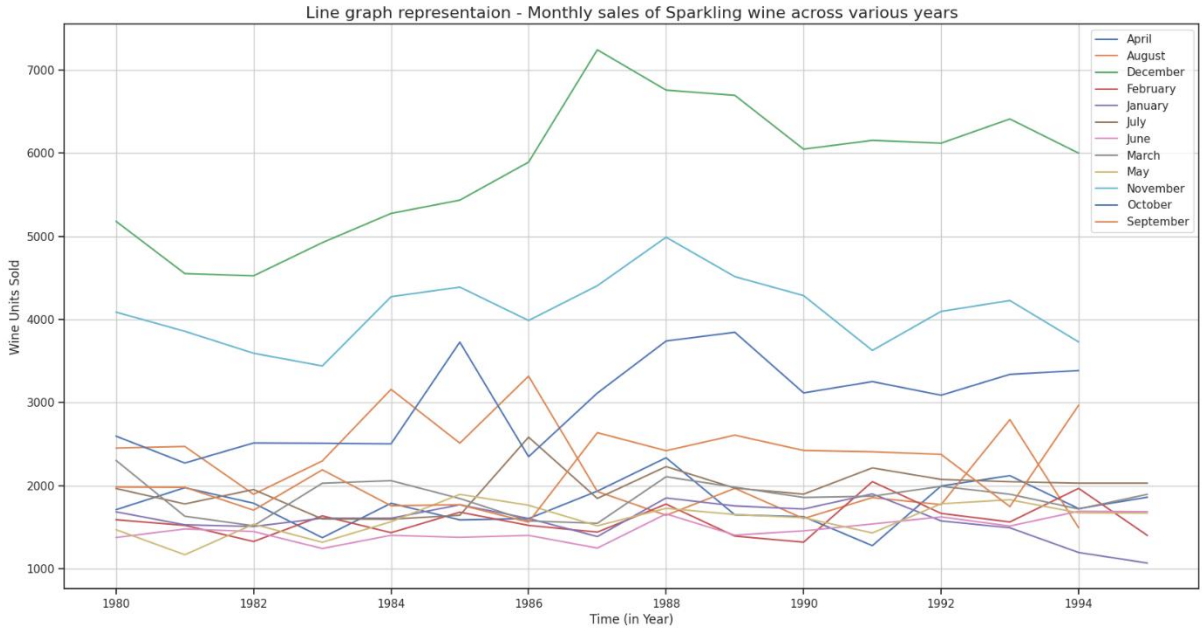
• The sales trajectory appears to be precisely the reverse of that seen in the yearly plot, seeing a gradual increase towards the end of each year.

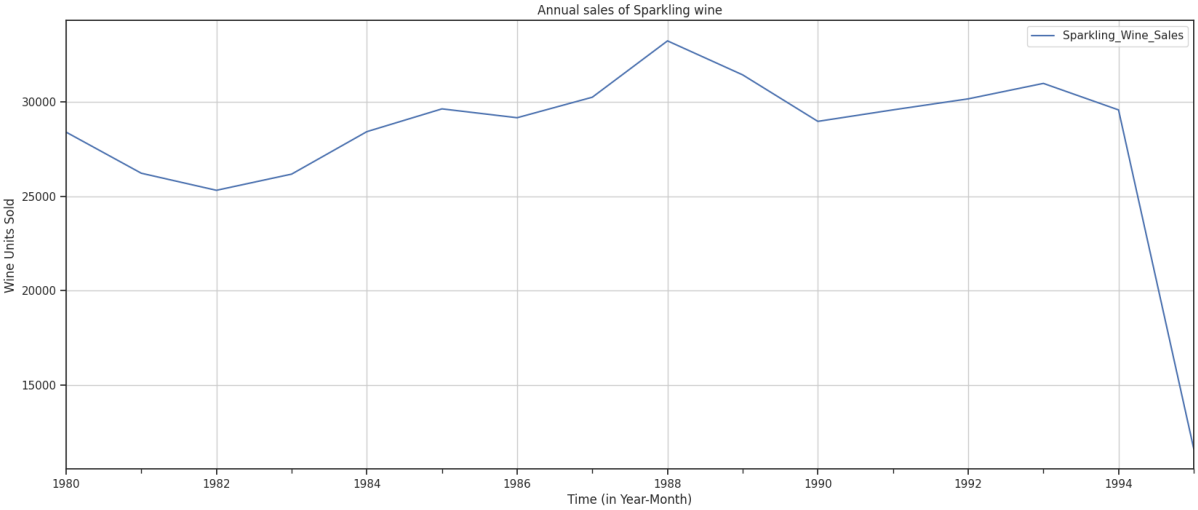
• January has the lowest wine sales while December sees the greatest. The sales modestly grow from January to August and then sharply climb after that.

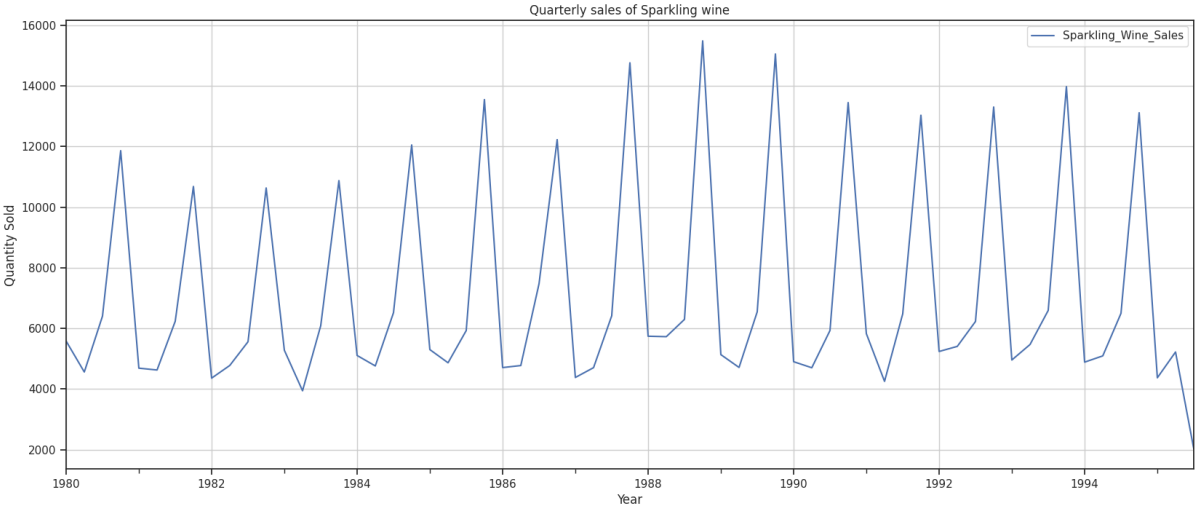
• Additionally, we can see that there are few outliers in the box plots.

Plot the empirical cumulative Distribution function

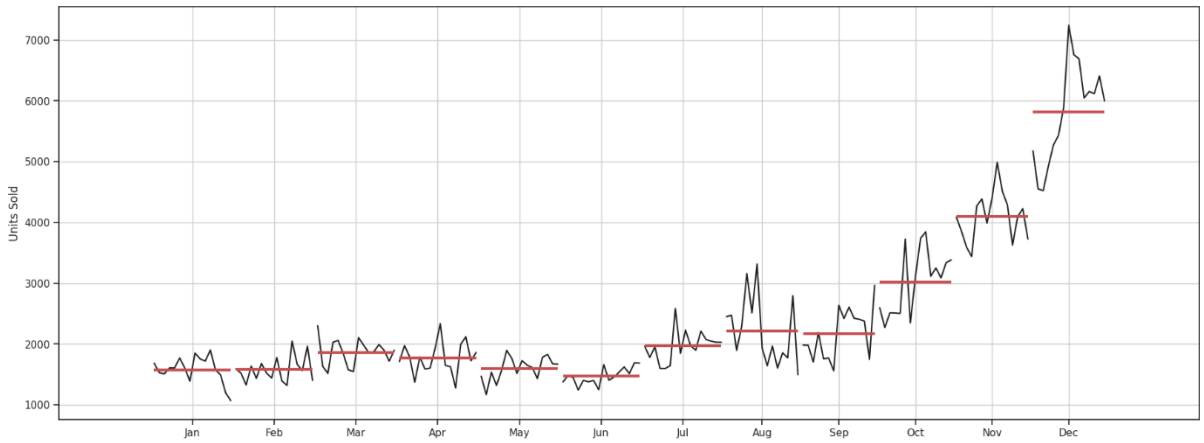





Annual sales of Sparkling wine
Sparkling_Wine_Sales
Wine Units Sold
30000
25000
20000
15000
1980
1982
1984
1986
1988
1990
1992
1994
Time (in Year-Month)


Quarterly sales of Sparkling wine
Sparkling_Wine_Sales
Quantity Sold
16000
14000
12000
10000
8000
6000
4000
2000
1980
1982
1984
1986
1988
1990
1992
1994
Year

Month plot of times series



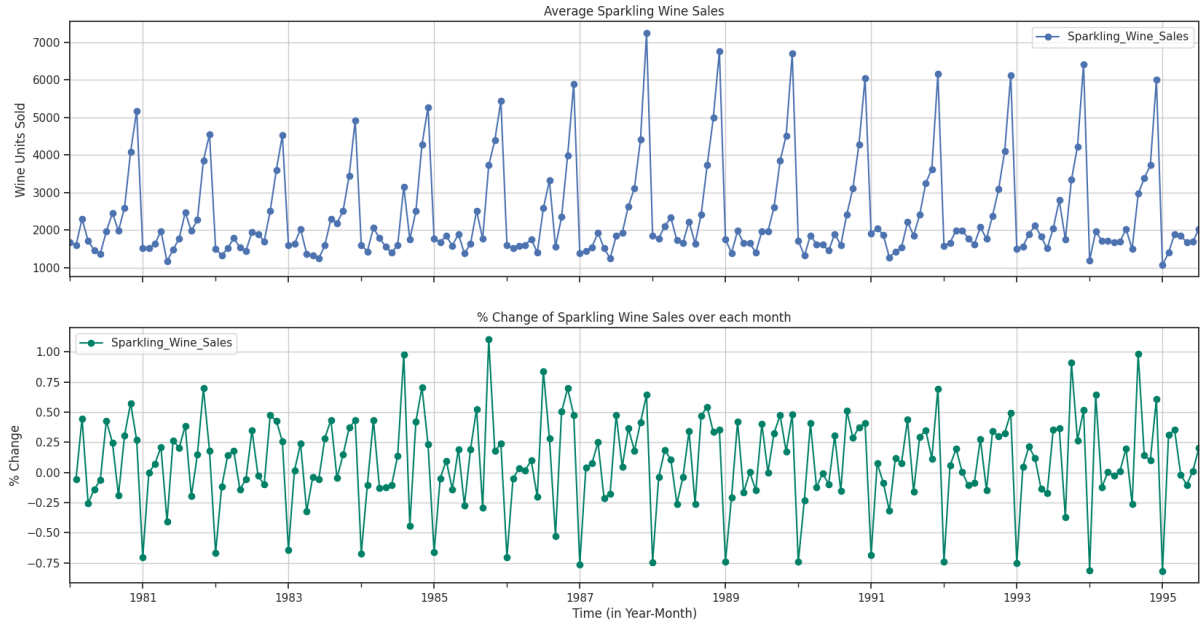
1. The sales are incresingly high in Q3 and Q4

Observation:

• Over the years, sales have stayed steady. The sales climbed gradually starting in 1982 until 1988, then decreased until 1990, then slightly increased again until 1994.

• Every year, December has the highest sales, followed by November and October. The first 2 months January and February have the lowest median sales.

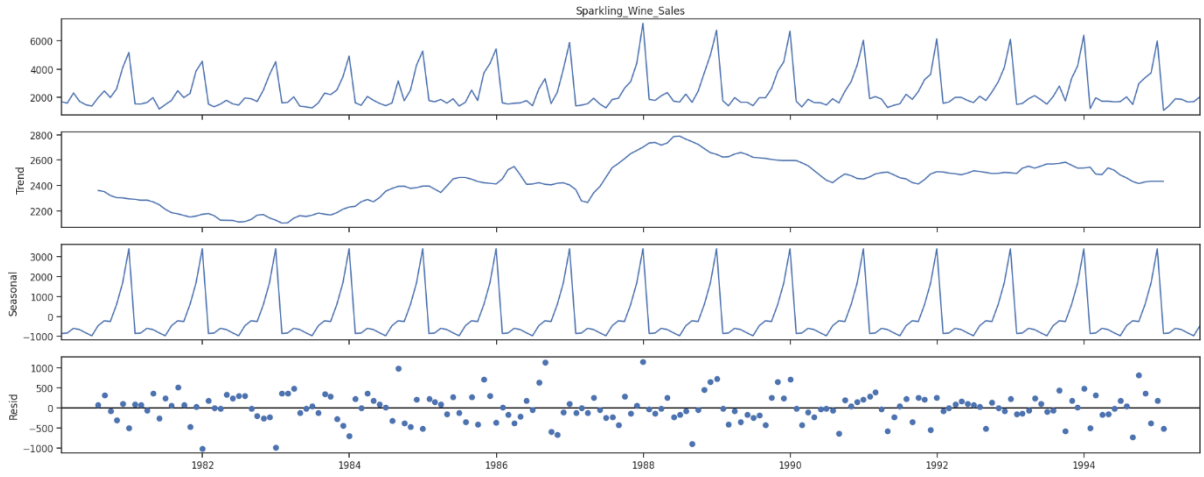
• From the cumulative distribution graph, we can observe that around 60 to 70 percent of the units sold are fewer than 2500, and 80% of the units sold are less than 4000. Only 20% of sales involved more than 3000 items. Therefore, it is clear that the bulk of sales were in the range of 1000 to 3000 units.



1. Average Sparkling wine sales graph shows us a no trend however yearly seasonality present in it
2. % change graph shows us the seasonality of the change in sales to be constant throghout the lifetime of sales

Decomposition of the Time Series

Additive decomposition



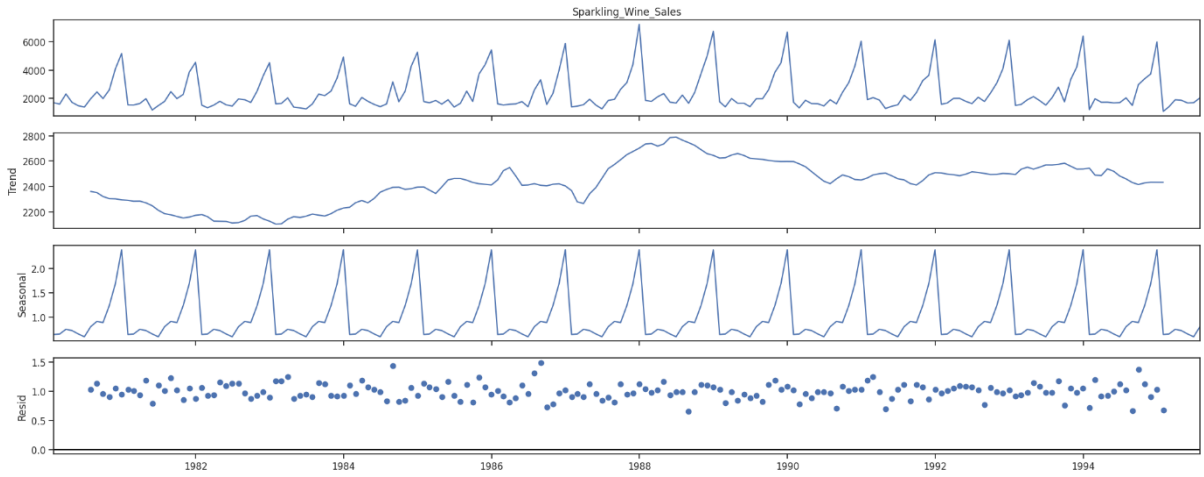
```
Trend
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: trend, dtype: float64

Seasonality
 Time_Stamp
1980-01-31    -854.260599
1980-02-29    -830.350678
1980-03-31    -592.356630
1980-04-30    -658.490559
1980-05-31    -824.416154
1980-06-30    -967.434011
1980-07-31    -465.502265
1980-08-31    -214.332821
1980-09-30    -254.677265
1980-10-31     599.769957
1980-11-30    1675.067179
1980-12-31    3386.983846
Name: seasonal, dtype: float64
```

Code cell output actions

Web search | Co

Multiplicative Decomposition



```
Trend
 Time_Stamp
1980-01-31            NaN
1980-02-29            NaN
1980-03-31            NaN
1980-04-30            NaN
1980-05-31            NaN
1980-06-30            NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Name: trend, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
1980-06-30    0.603468
1980-07-31    0.809164
1980-08-31    0.918822
1980-09-30    0.894367
1980-10-31    1.241789
1980-11-30    1.690158
1980-12-31    2.384776
Name: seasonal, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31         NaN
1980-02-29         NaN
1980-03-31         NaN
1980-04-30         NaN
1980-05-31         NaN
1980-06-30         NaN
1980-07-31    1.029230
1980-08-31    1.135407
1980-09-30    0.955954
1980-10-31    0.907513
1980-11-30    1.050423
1980-12-31    0.946770
Name: resid, dtype: float64
```

Observation:

• The residual patterns after additive decomposition of the time series appear to represent the seasonal element and exhibit substantial variation.

• In the multiplicative decomposition of the time series, it has been observed that the seasonal fluctuation of residuals is under control.

• The size of the seasonal variations doesn't change on comparison, but the residuals are tightly controlled by the multiplicative decomposition. In addition to this, the residuals are not independent of seasonality thus we may assume that it is multiplicative.

3. Split the data into training and test. The test data should start in 1991.

Train and test data are separated from the provided dataset. Sales data up to 1991 is included in the training data, while data from 1991 through 1995 is used for testing.

Number of observations in Train data : (132, 1)
Number of observations in Test data : (55, 1)
Total Observations : 187

First few rows of Training Data

Time_Stamp	Sparkling_Wine_Sales
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471
1980-06-30	1377
1980-07-31	1966
1980-08-31	2453
1980-09-30	1984
1980-10-31	2596

Last few rows of Training Data

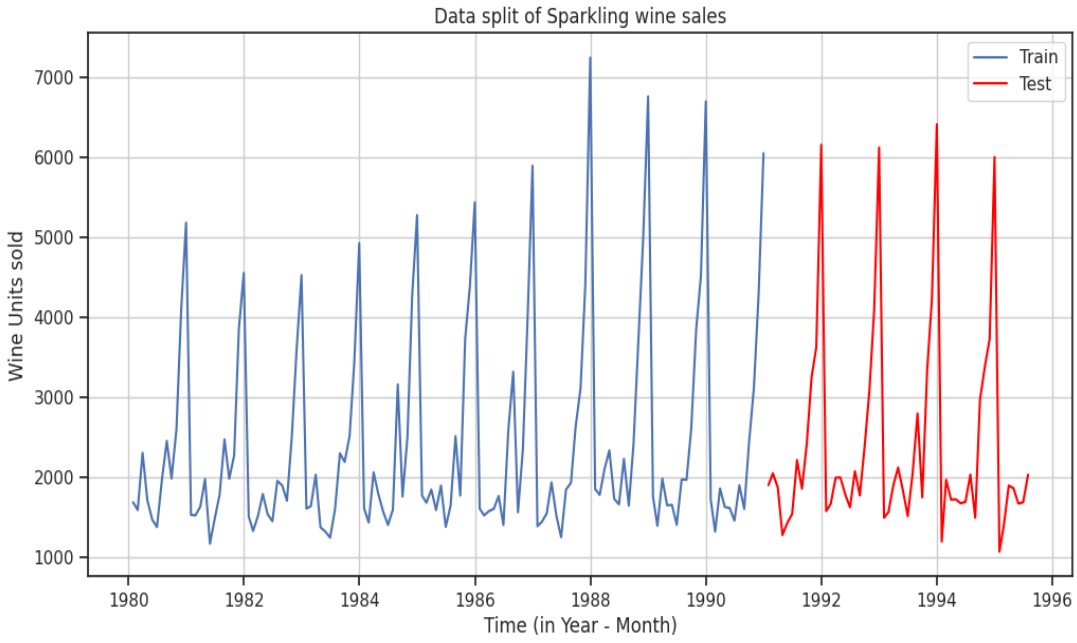
Time_Stamp	Sparkling_Wine_Sales
1990-03-31	1859
1990-04-30	1628
1990-05-31	1615
1990-06-30	1457
1990-07-31	1899
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

First few rows of Test Data

Time_Stamp	Sparkling_Wine_Sales
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432
1991-06-30	1540
1991-07-31	2214
1991-08-31	1857
1991-09-30	2408
1991-10-31	3252

Last few rows of Test Data

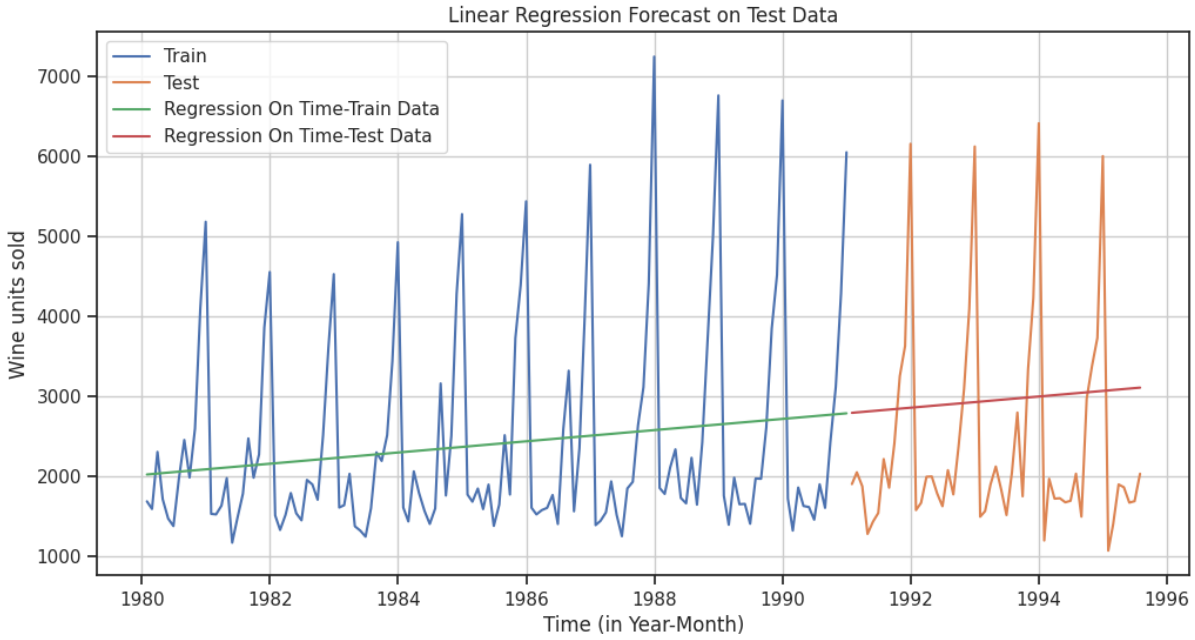
Time_Stamp	Sparkling_Wine_Sales
1994-10-31	3385
1994-11-30	3729
1994-12-31	5999
1995-01-31	1070
1995-02-28	1402
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031



Line Plot – Splitting of time series into Train & Test data

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression



Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• The train and test data trends have been caught by the linear regression model however, it is unable to account for seasonality

• The root means squared error (RMSE) for the linear regression model is 1389.135. The size of the seasonal

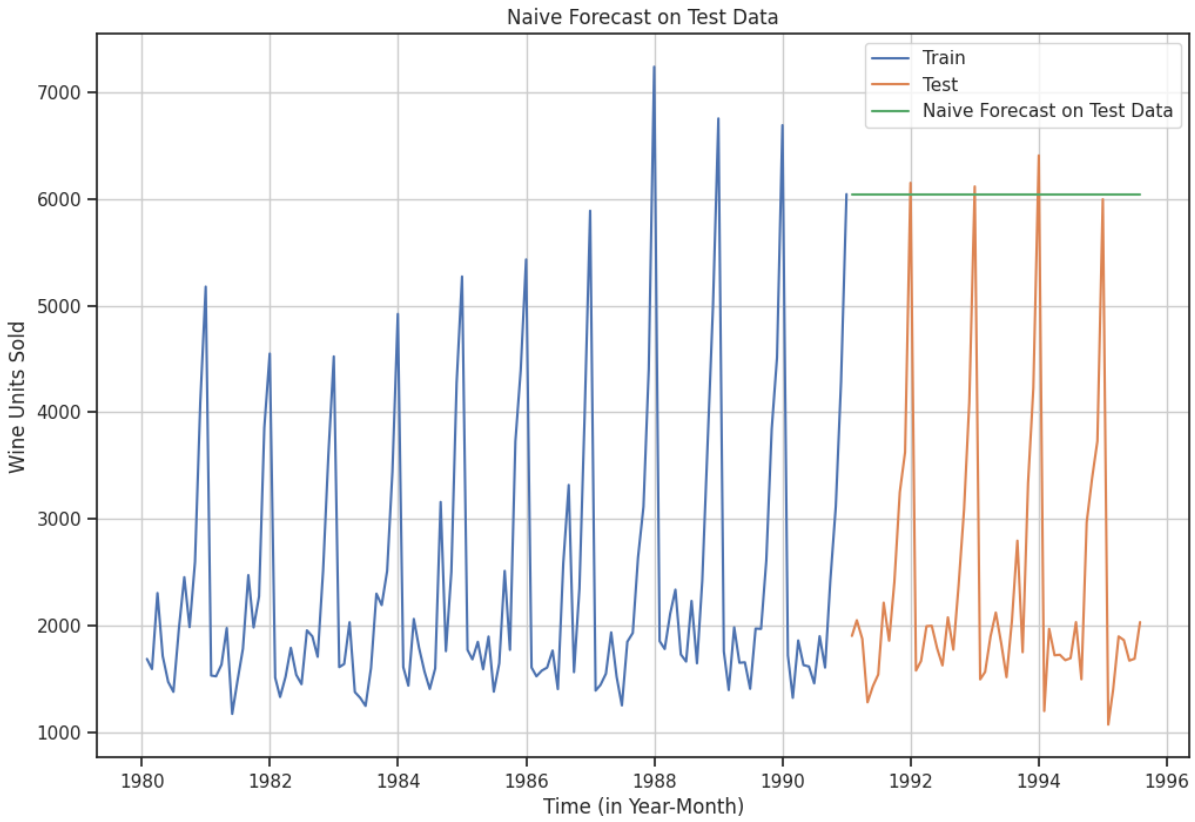
Linear Regression: Model Evaluation

Performance Metric

Test RMSE	1389.135175

Model 2 – Naïve Forecast

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

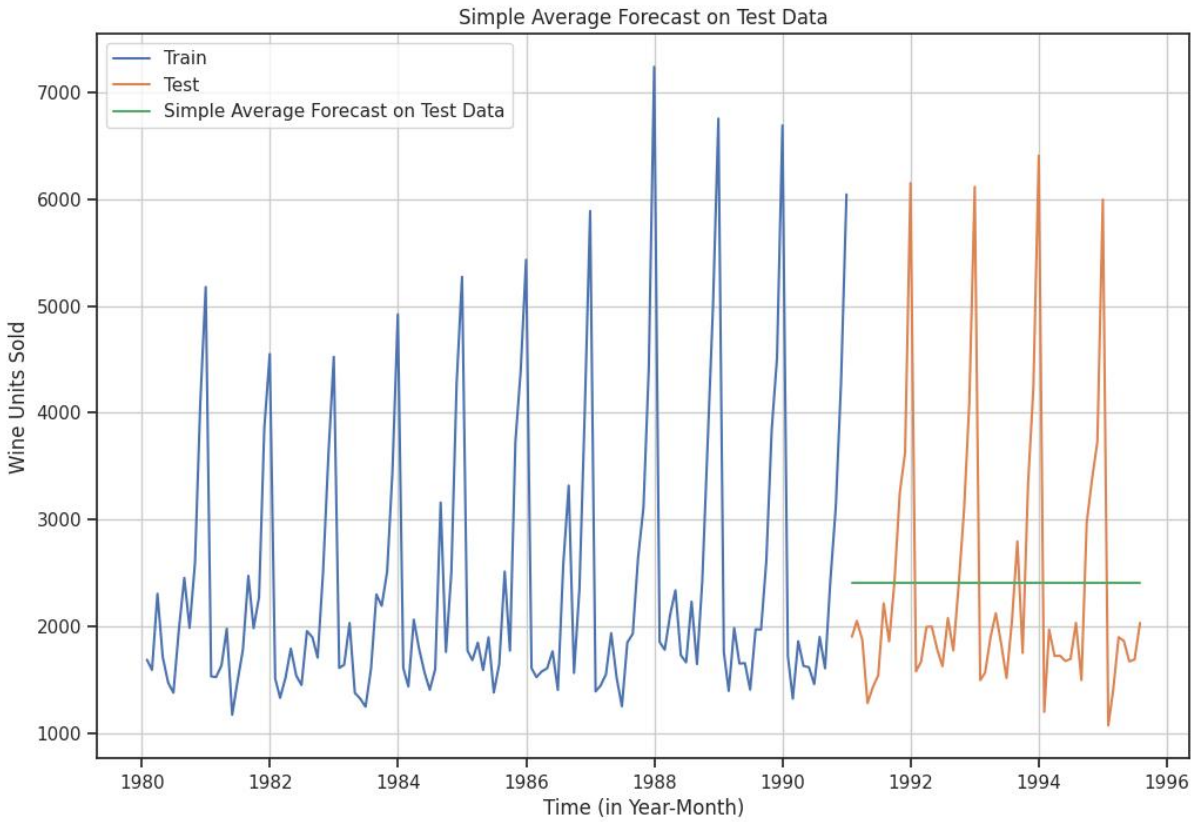


Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• The seasonality and trend of the time series data cannot be captured by the naive forecast model.
• The root means squared error (RMSE) for the naïve forecast model is 3864.279 which is significantly higher than the regression model.

Model 3 – Simple Average



Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• The seasonality and trend of the time series data cannot be captured by the simple average model.
• The root means squared error (RMSE) for the simple average model is 1275.081 which is significantly lower than the naïve forecast model and slightly lower than Linear regression model.

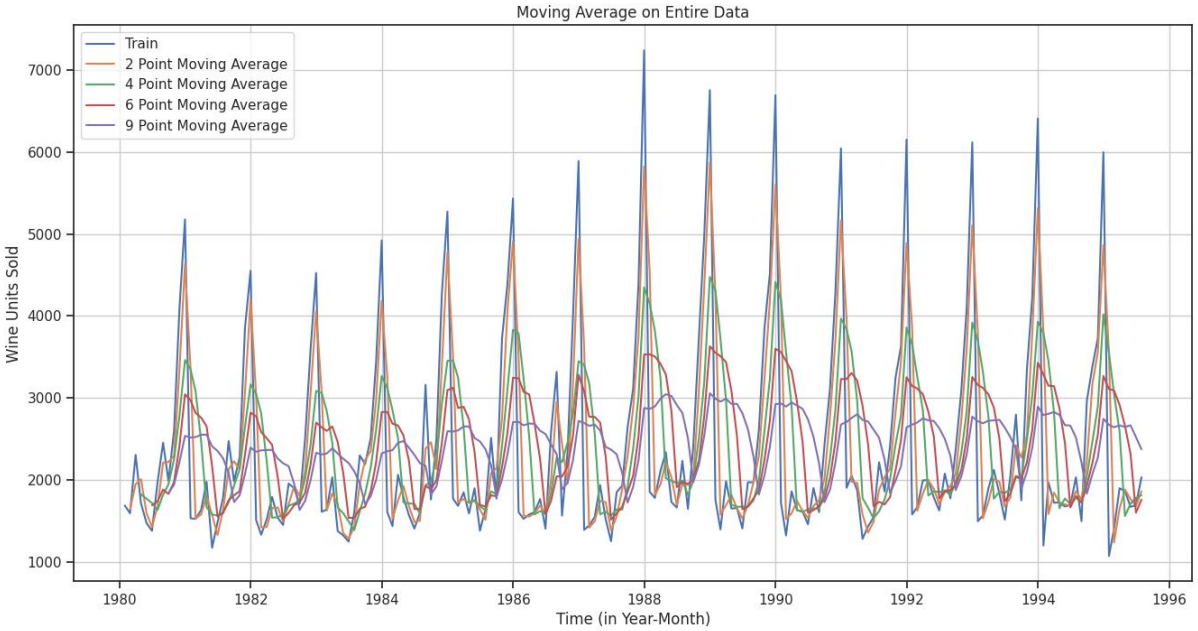
Simple Average: Model Evaluation

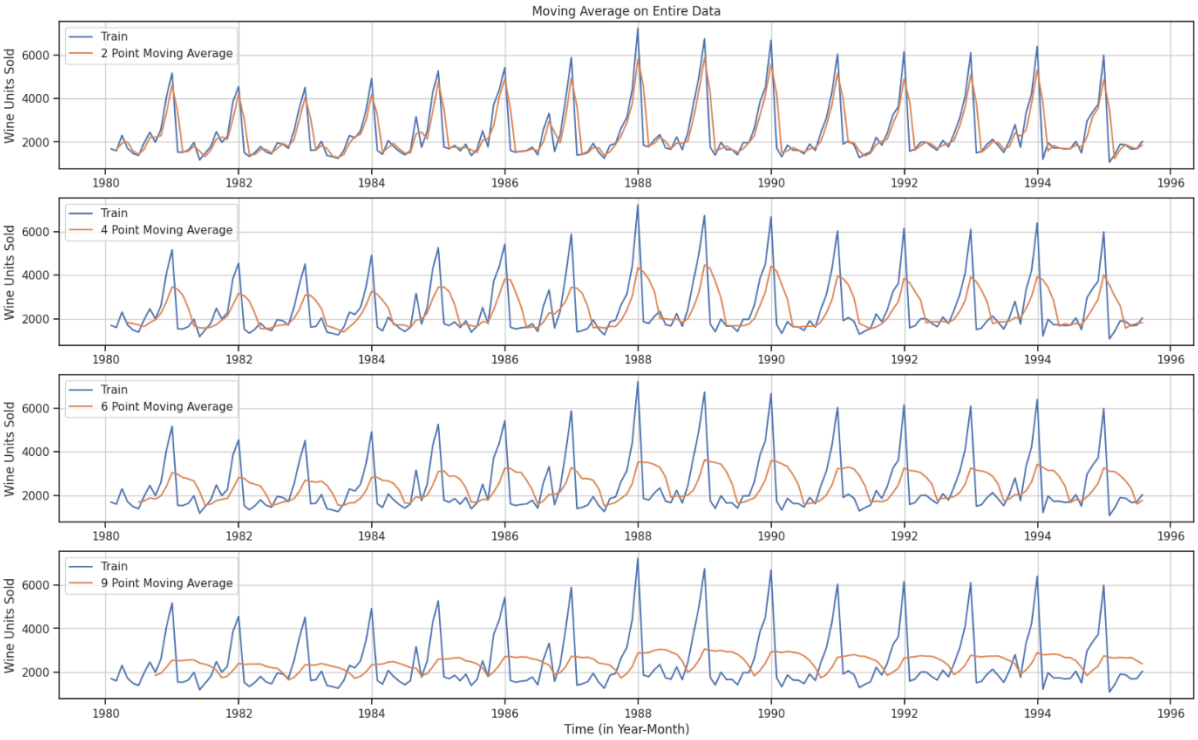
Performance Metric

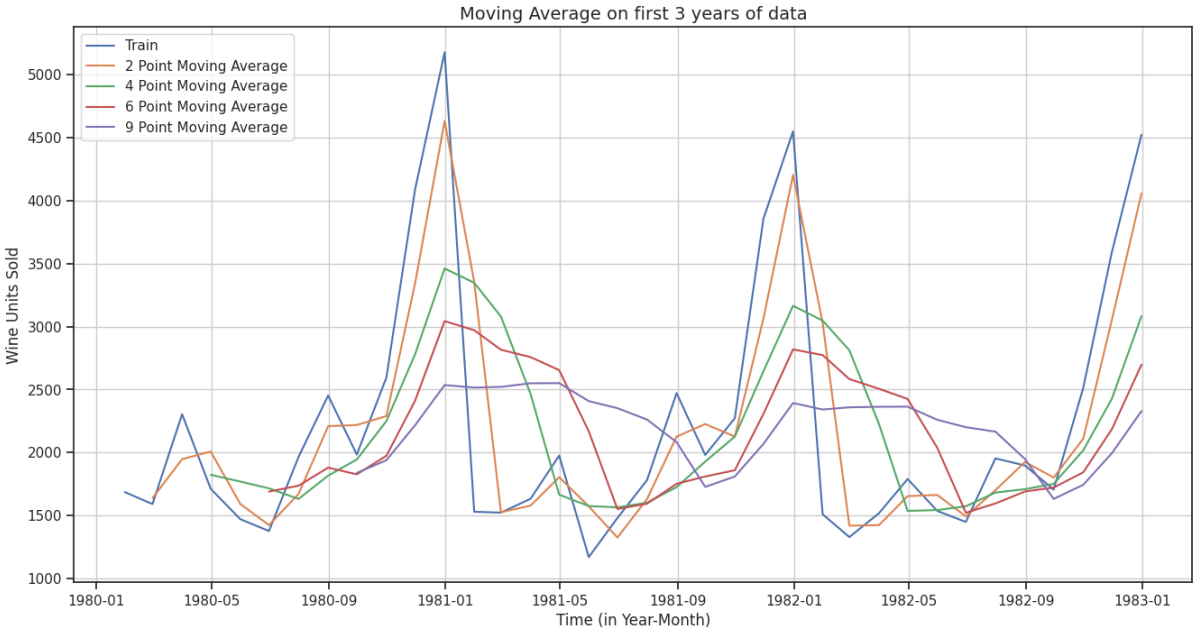
Test RMSE	1275.081804

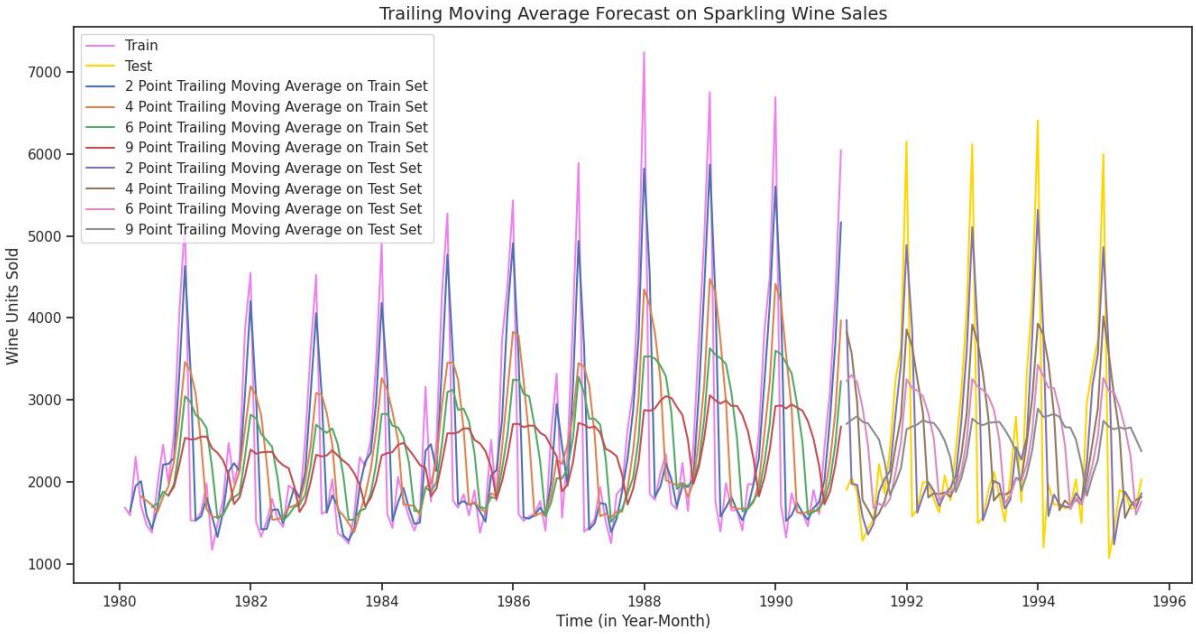
Model 4 – Moving Average (MA)

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.









Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• The seasonality and trend of the time series data may both be predicted using moving average models.

• We can see how the data smooth out as the number of observation points taken increases. The 2-point TMA has characteristics that are more similar to test results than the 9-point TMA.

• The root means squared error (RMSE) for the 2-point trailing average model is 813.4, which is lowest than all models build so far.

Model Evaluation- Moving Average

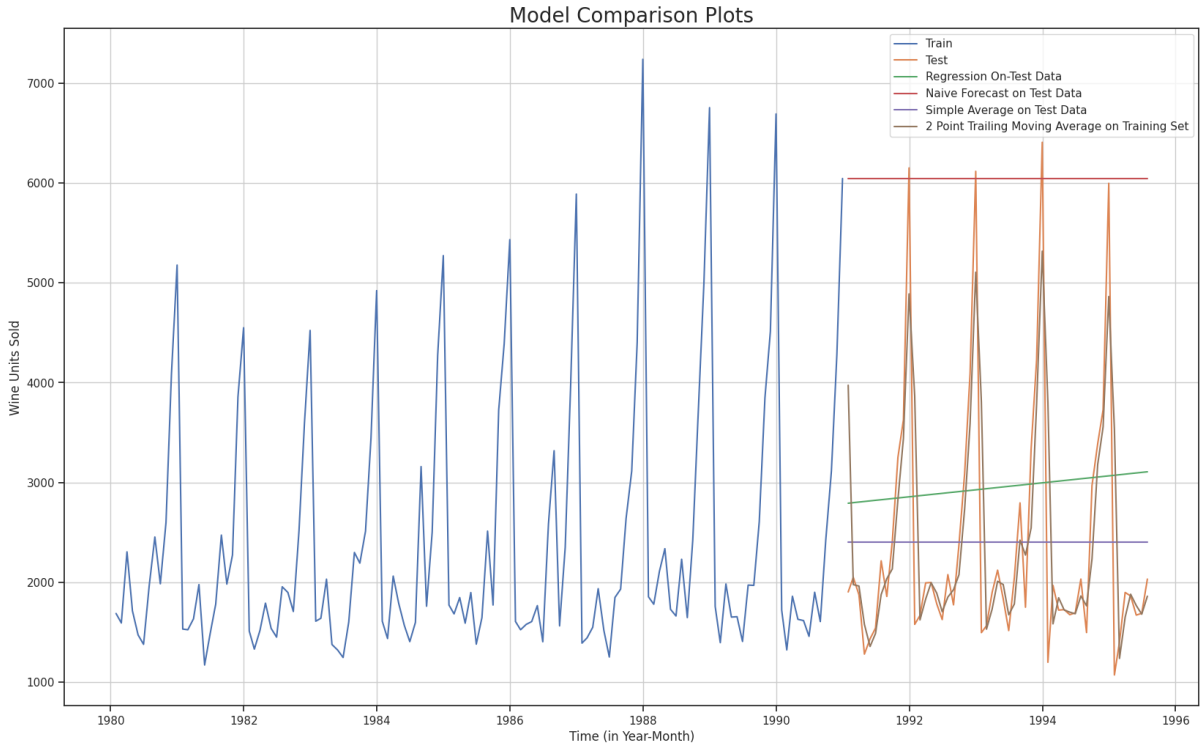
For 2 point Moving Average Model forecast on the Training Data, RMSE is 661620.673

For 4 point Moving Average Model forecast on the Training Data, RMSE is 1337699.720

For 6 point Moving Average Model forecast on the Training Data, RMSE is 1648469.640

For 9 point Moving Average Model forecast on the Training Data, RMSE is 1812465.30

Let's compare the visualization of each model's predictions that we have constructed so far before investigating exponential smoothing methods



Comparison of different models on test data (Regression, Naïve, Simple and Moving Average)

Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• We can see from the graph above that simple average and naive forecast models fail to adequately describe the characteristics of the test data.

• The trend portion of the series has been caught using linear regression, however the seasonality has been missed

• Both trend and seasonality may be accounted for using moving average models

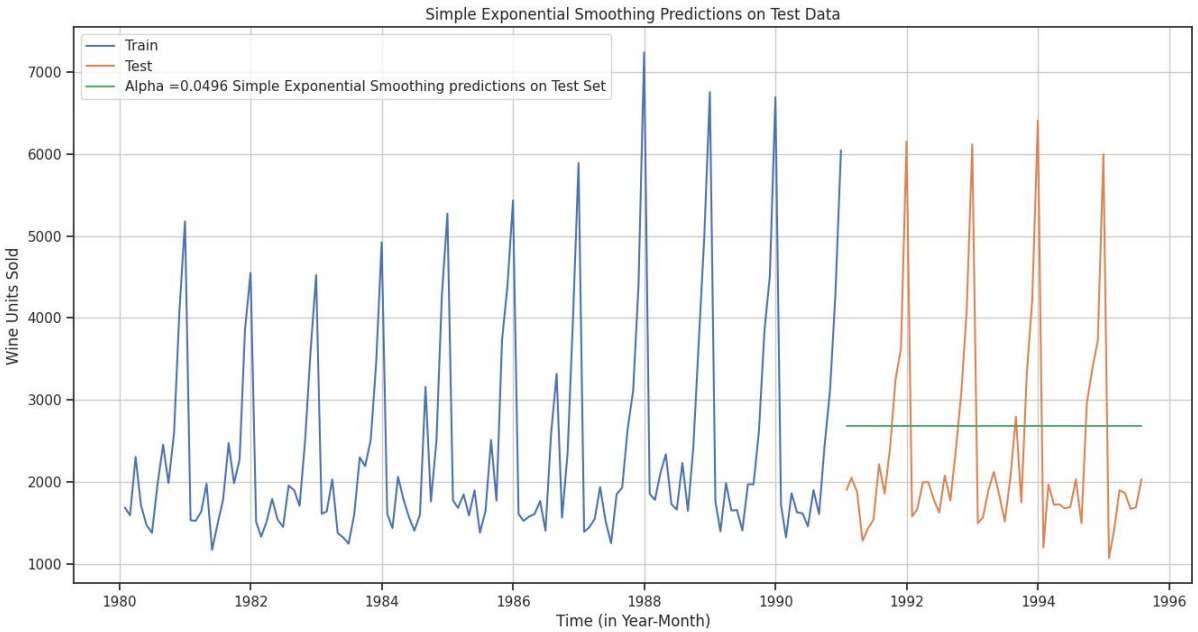
Model 5: Simple Exponential Smoothing

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES). This method is suitable for forecasting data with no clear trend or seasonal pattern.

In Single ES, the forecast at time (t + 1) is given by Winters,1960

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing. For the selection criteria, the below Simple Exponential Smoothing is built by using optimized parameters.

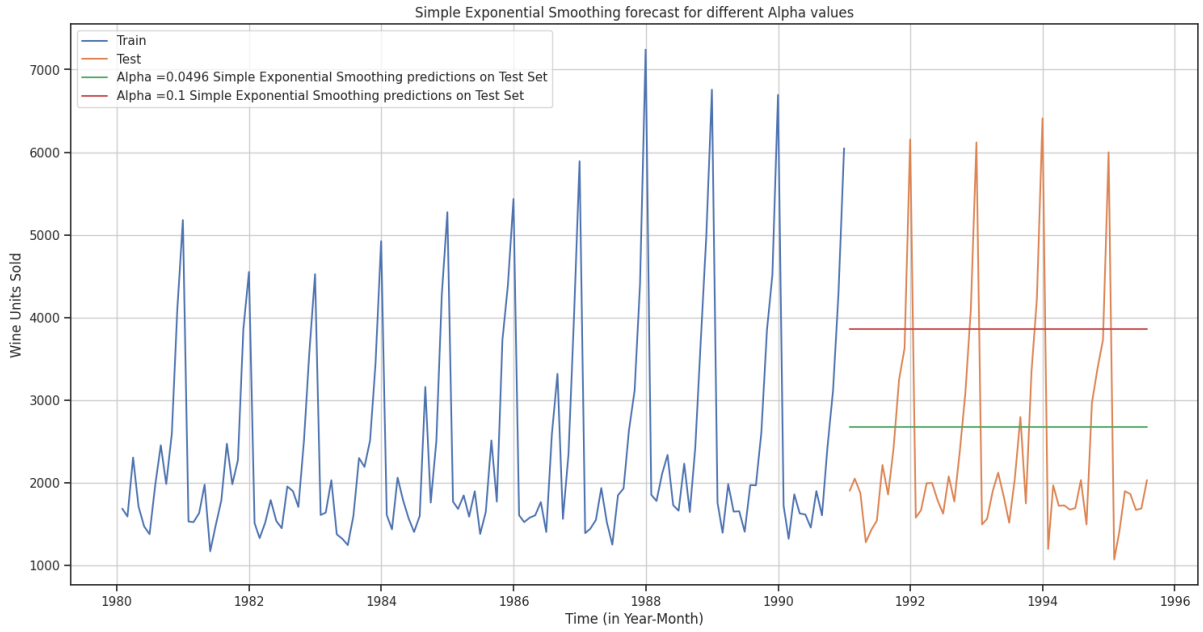


The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set.

The range of alpha value is from 0.1 to 0.95 and the respective RMSE for train and test data are calculated for analyzing the performance metrics.

Model Evaluation for α = 0.0496 : Simple Exponential Smoothing

index	Test RMSE
Linear Regression	1389.135174897992
Naive Model	14932654.909090908
Simple Average	1625833.6061179985
2 point TMA	661620.6727272727
4 point TMA	1337699.7204545455
6 point TMA	1648469.64040404
9 point TMA	1812465.3025813692
Alpha =0.0496,SimpleExponentialSmoothing	1702835.5310907199



Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• When there is neither a trend nor a seasonal component to the time series, simple exponential smoothing is typically used. It is due to this reason, it unable to capture the characteristics of the time series data.

• The root means squared error (RMSE) for the simple exponential smoothing model with Alpha =0.0496 is 1316.135and for Alpha=0.1, RMSE is 1375.393. • The Simple Exponential Smoothing with alpha=0.0496 is taken as the best model among two as it has the lowest test RMSE.

Simple Exponential Smoothing: Model Evaluation

Model	Test RMSE
SES (Alpha = 0.0496)	1316.135411
SES (Alpha = 0.1)	1375.393398

Method 6: Double Exponential Smoothing (Holt's Model)

This model is an extension of SES known as Double Exponential model which estimates two smoothing parameters. Applicable when data has Trend but no seasonality. Two separate components are considered: Level and Trend. Level is the local mean. One smoothing parameter α corresponds to the level series. A second smoothing parameter β corresponds to the trend series.

Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short-term average value or level and the other for capturing the trend.

Intercept or Level equation, Lt is given by: $L_t = \alpha Y_t + (1-\alpha)F_t$ Trend equation is given by $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

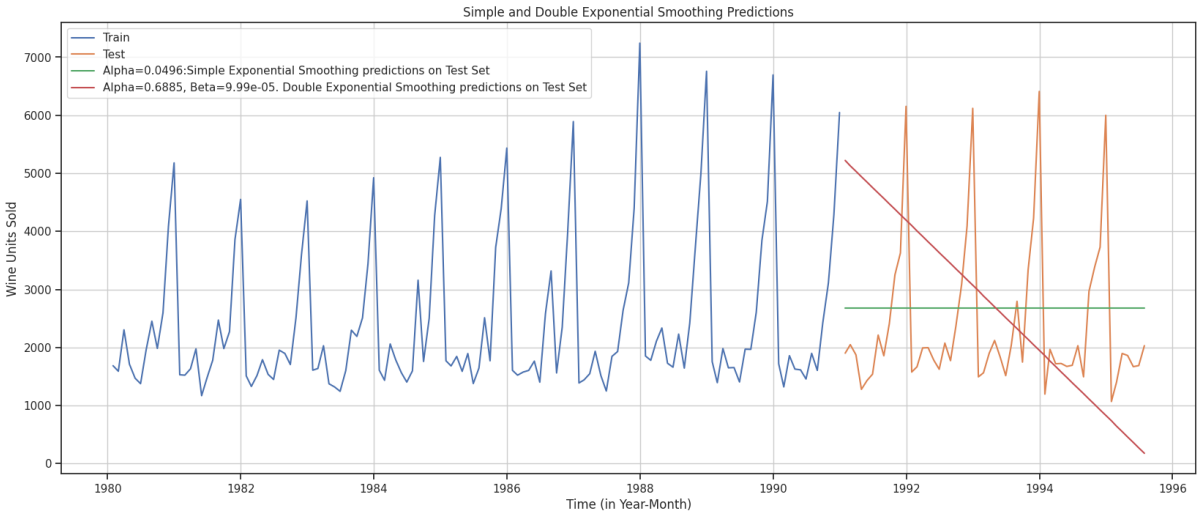
Here, αα and ββ are the smoothing constants for level and trend, respectively,

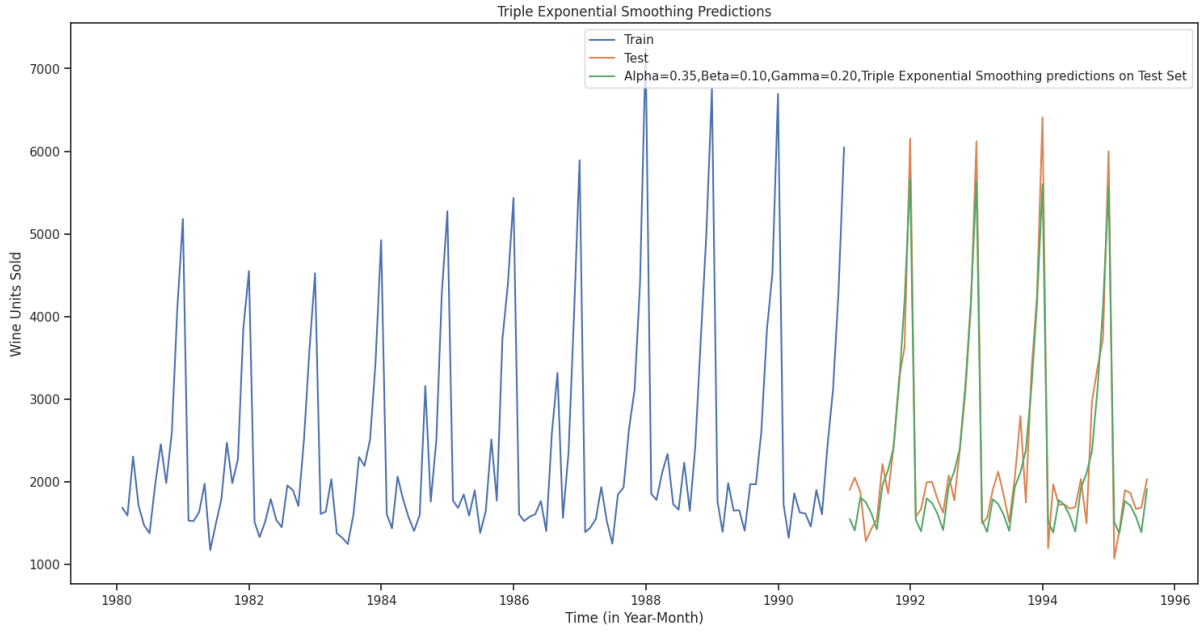
$0 < 1$ and $0 < \beta < 1$.

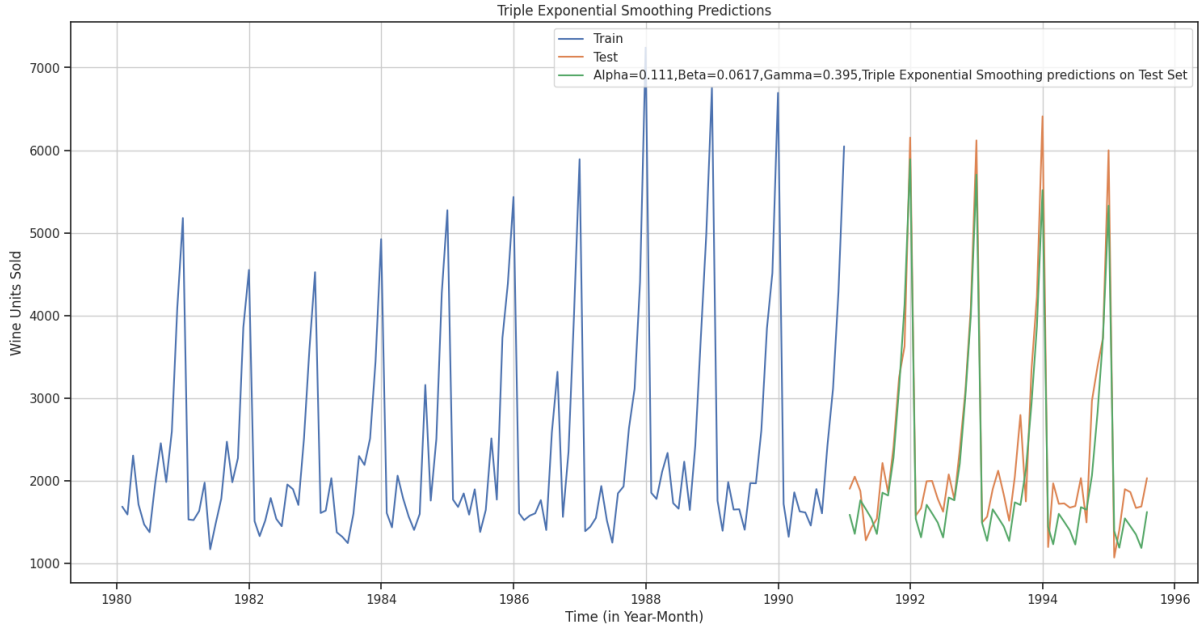
The forecast at time t + 1 is given by

$F_{t+1} = L_t + T_t$

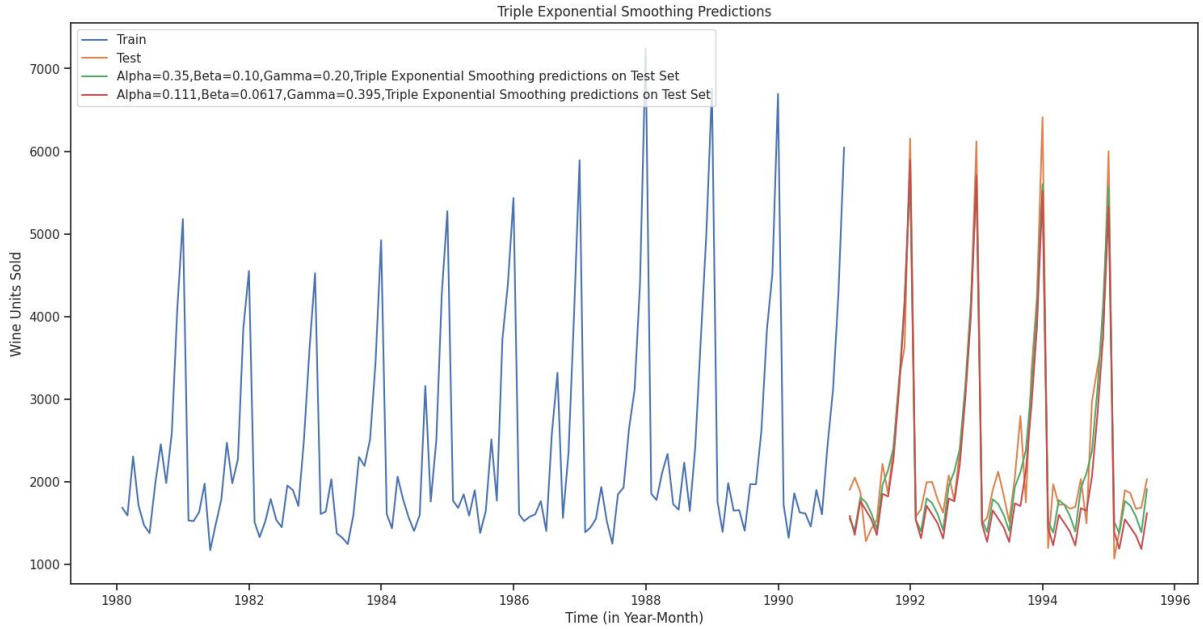
$F_{t+n} = L_t + nT_t$



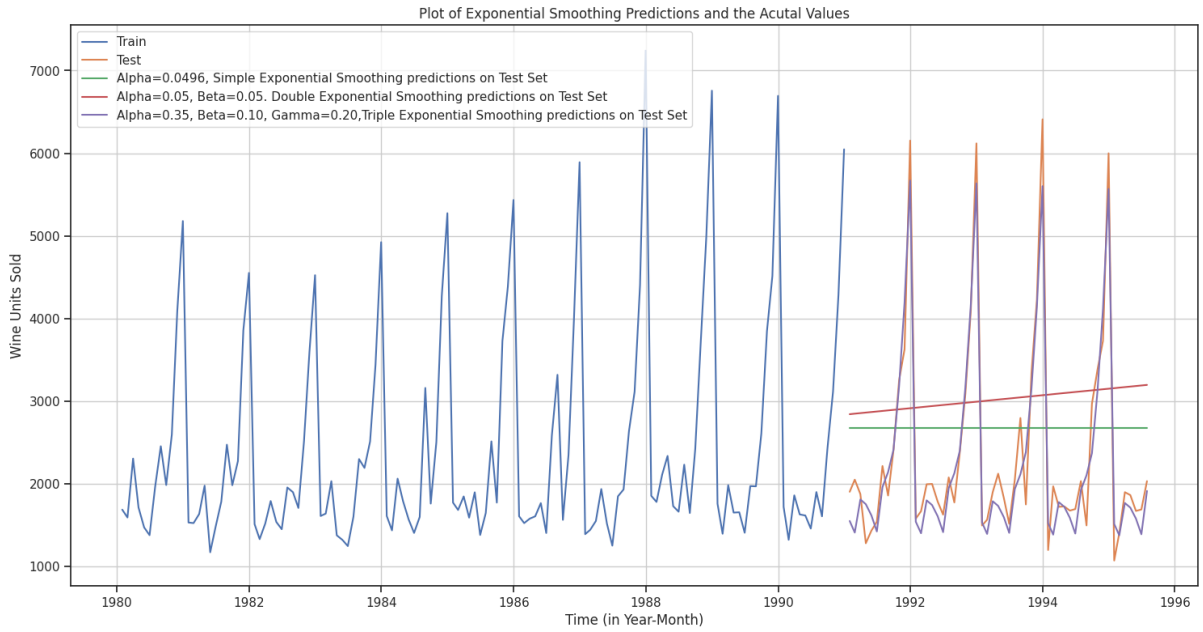




The more recent observation is given more weight the higher the alpha value. That implies that the recent events will repeat again. A loop with different alpha values is run to understand which particular value works best for alpha on the test set. The range of alpha value is from 0.05 to 1.0 and the respective RMSE for train and test data are calculated for analyzing the performance metrics



We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters α = 0.35, β = 0.10 and γ = 0.20.



Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• When there is simply trend and no seasonality in the time series data, the double exponential smoothing model performs well. It is due to this reason it is only able to capture the trend characteristics of the data and seasonality is not accounted for.

• The root means squared error (RMSE) for the double exponential smoothing model with Alpha=0.6885, Beta=9.99e-05 is 2007.238and for Alpha=0.05, Beta=0.05 (Auto tuned model), RMSE is 1418.407.

• The Double Exponential Smoothing with Alpha=0.05, Beta=0.05 is taken as the best model among two as it has the lowest test RMSE.

• Additionally, it should be highlighted that compared to the simple exponential smoothing model, the double exponential smoothing model has slightly higher RMSE.

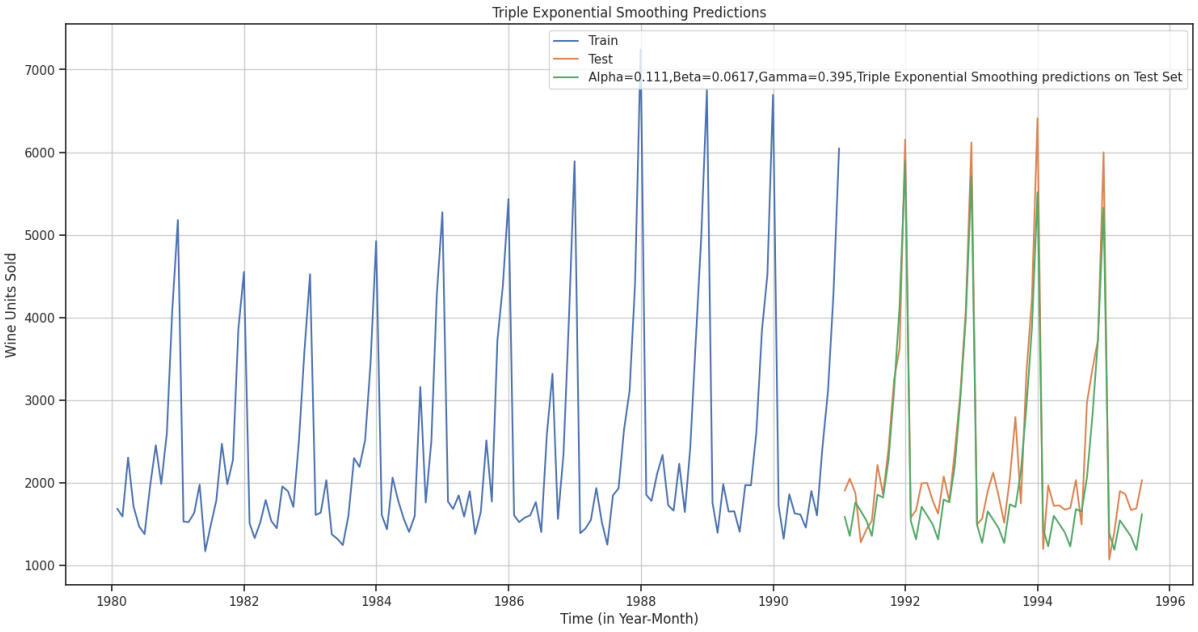
Double Exponential Smoothing: Model Evaluation

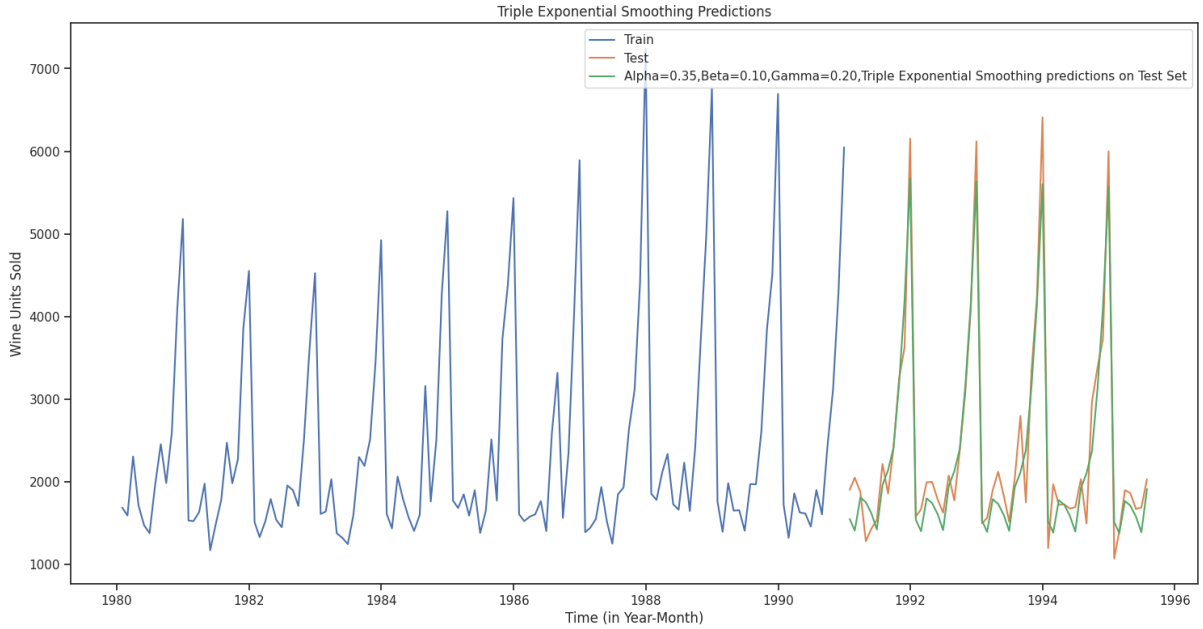
Model	Test RMSE
DES (Alpha=0.6885, Beta=9.99e-05)	2007.238526
DES (Alpha=0.05, Beta=0.05)	1418.407668

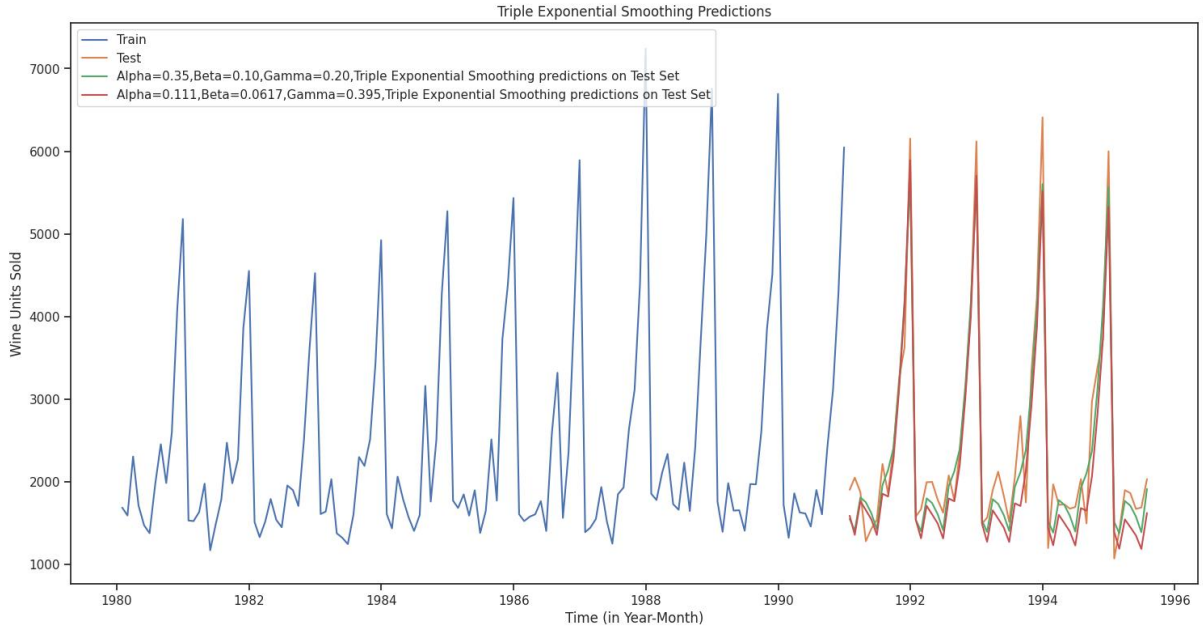
Model 7 – Triple Exponential Smoothing (Holt-Winter's Model)

This model is an extension of DES known as Triple Exponential Smoothing model which estimates three smoothing parameters. Applicable when data has both Trend and seasonality. Three separate components are considered: Level, Trend and Seasonality

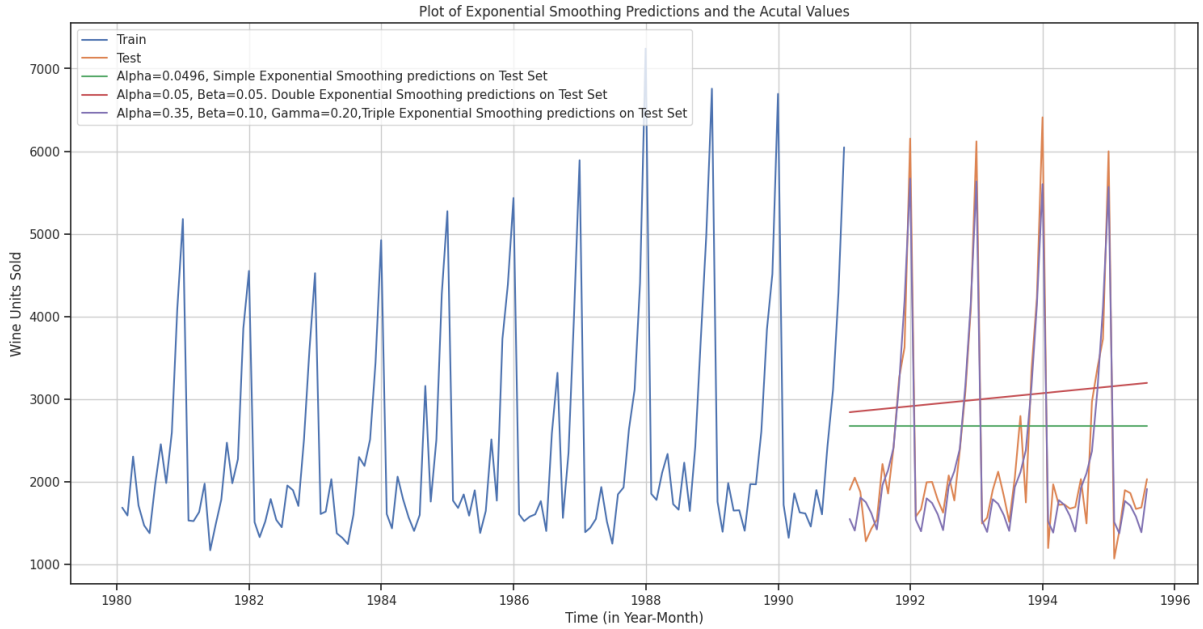
One smoothing parameter α corresponds to the level series. A second smoothing parameter β corresponds to the trend series. A third smoothing parameter γ corresponds to the seasonality series where, 0 < α < β < γ







We see that the best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters α = 0.35, β = 0.10 and γ = 0.20.



Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• When there is both trend and seasonality in the time series data, the triple exponential model works well. It is due to this reason it able to capture both the trend and seasonal characteristics and nearly match the actual test data plot.

• The root means squared error (RMSE) for the double exponential smoothing model with Alpha=0.111, Beta=0.0617, Gamma=0.395 is 469.659 and for Alpha=0.35, Beta=0.10, Gamma=0.20 (Auto tuned model), RMSE is 319.498.

• The Triple Exponential Smoothing with Alpha=0.35, Beta=0.10, Gamma=0.20 is taken as the best model among two as it has the lowest test RMSE.

• Additionally, it should be highlighted that compared to the double exponential smoothing model, the triple exponential smoothing model has almost reduced the RMSE value by 75%.

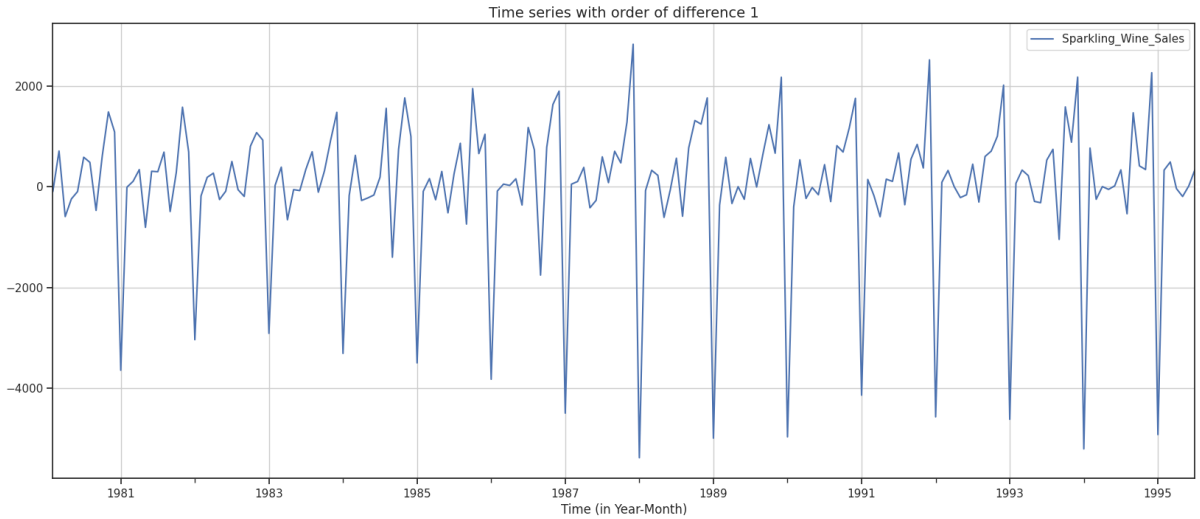
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Check for stationarity of the whole Time Series data.

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.
The hypothesis in a simple form for the ADF test is:

- H0 : The Time Series has a unit root and is thus non-stationary.
- H1 : The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

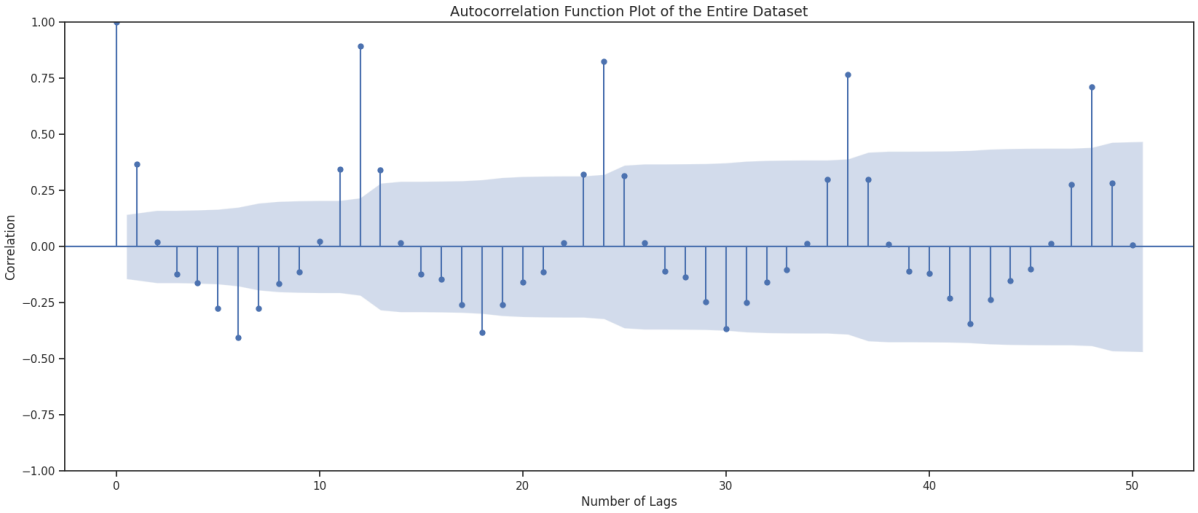


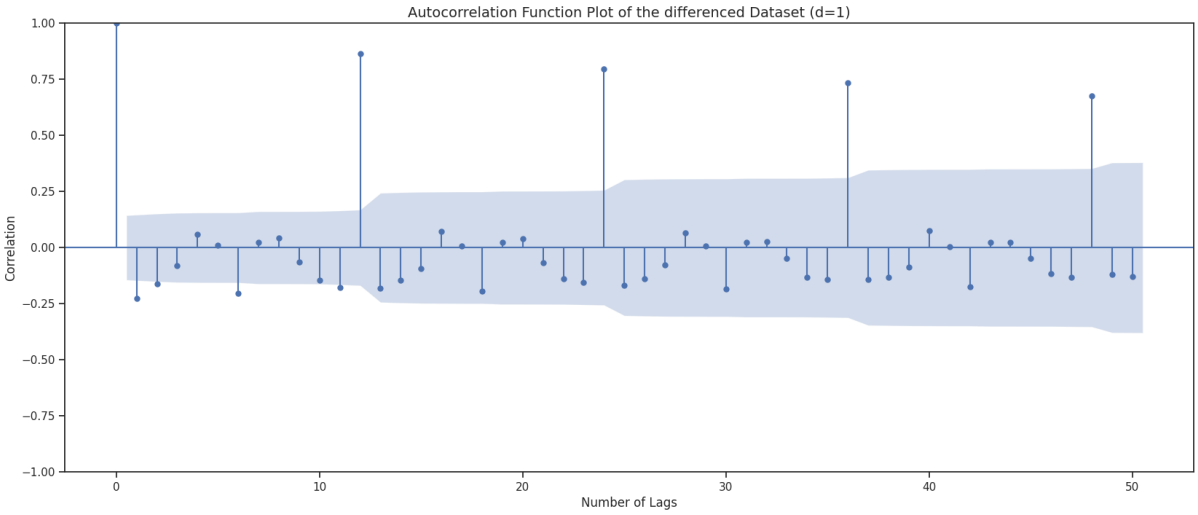
Inference:

We see that at 5% significant level the Time Series is non-stationary as p-value is 0.705 which is more than alpha value (0.05), therefore we fail to reject the null hypothesis. Let us take one level of differencing to see whether the series becomes stationary.

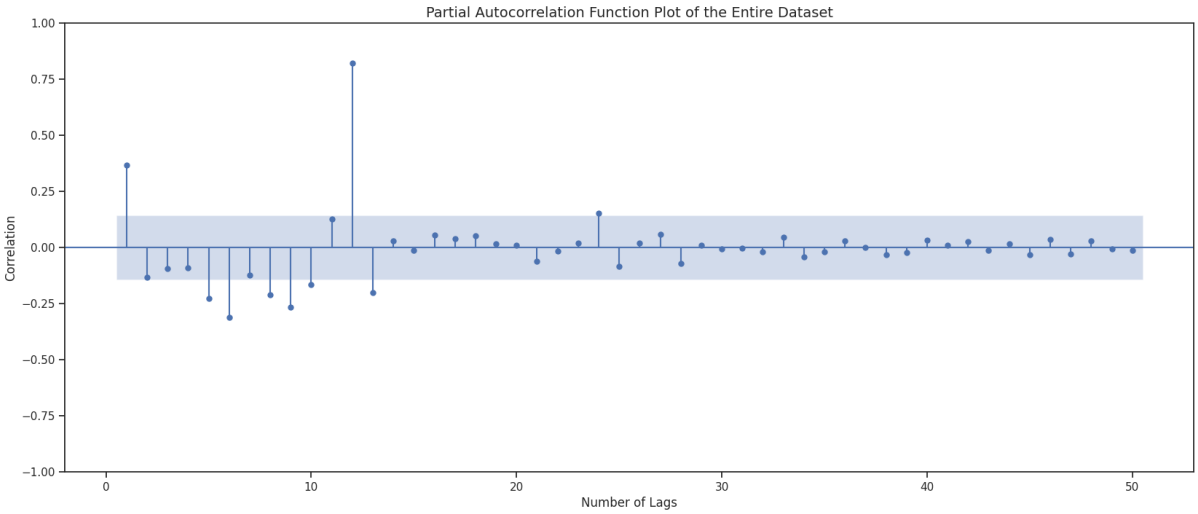
Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.

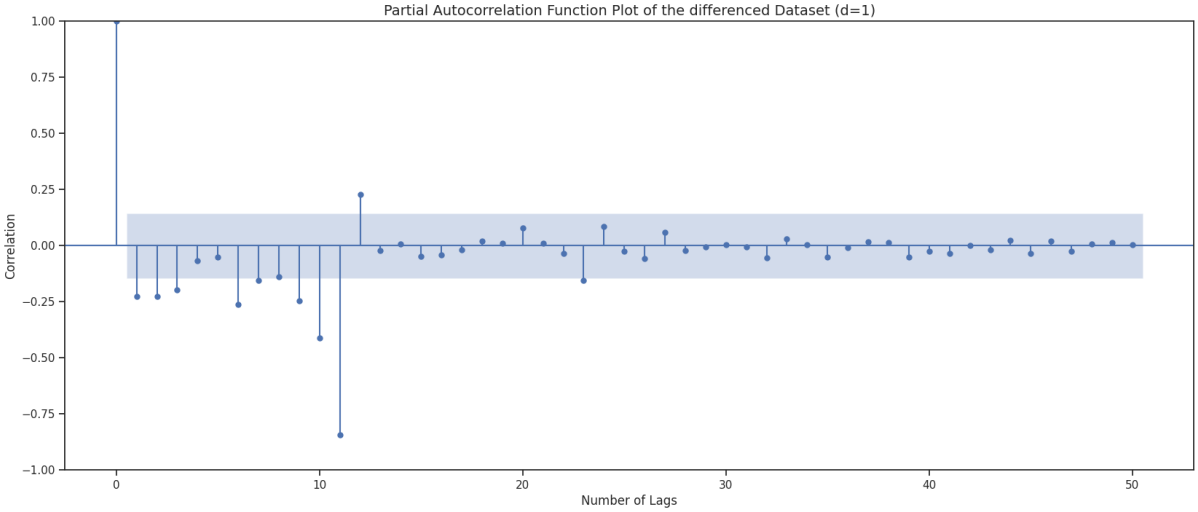
ACF plot

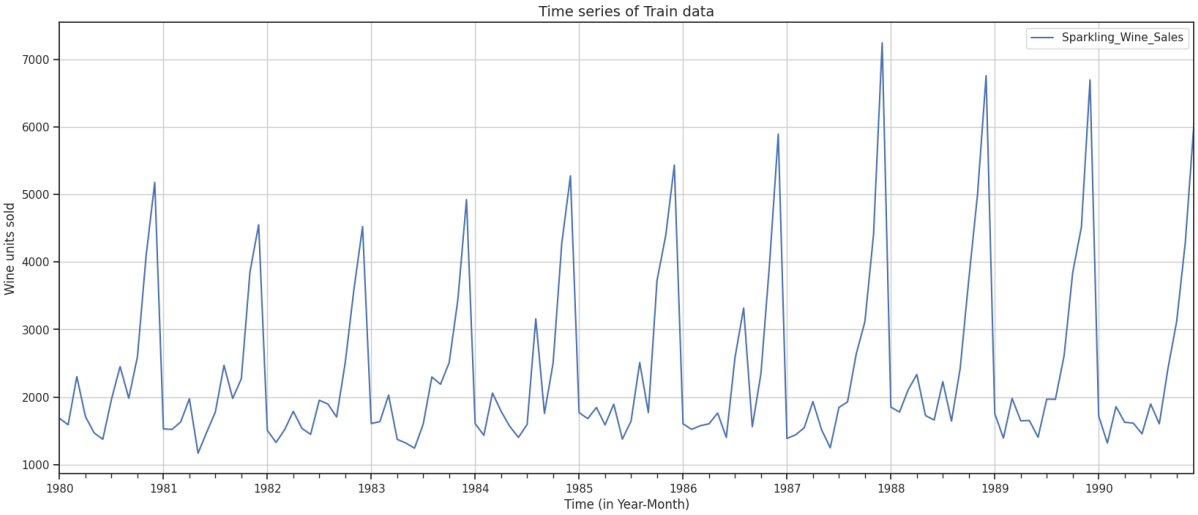


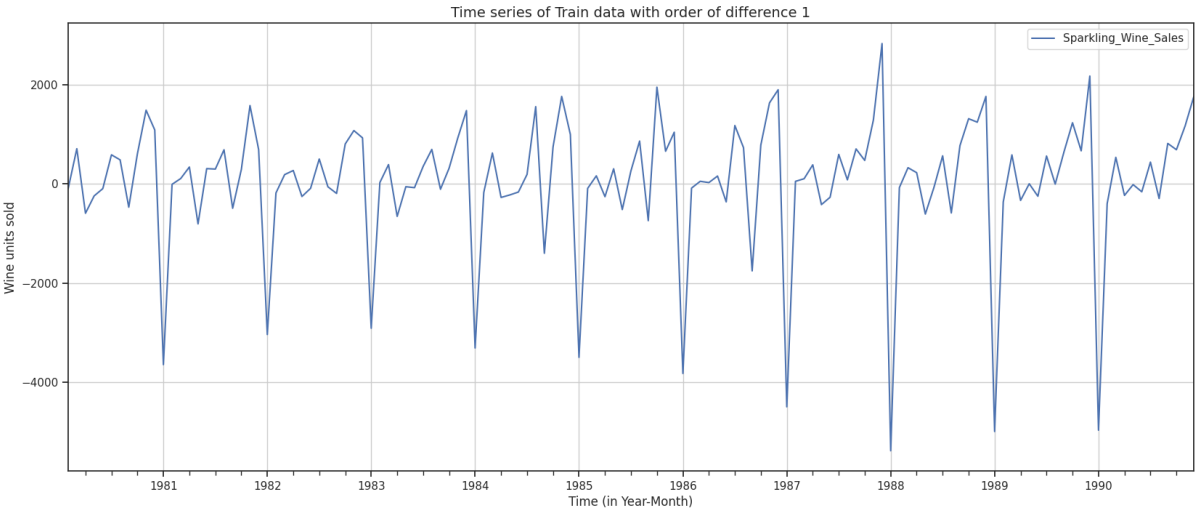


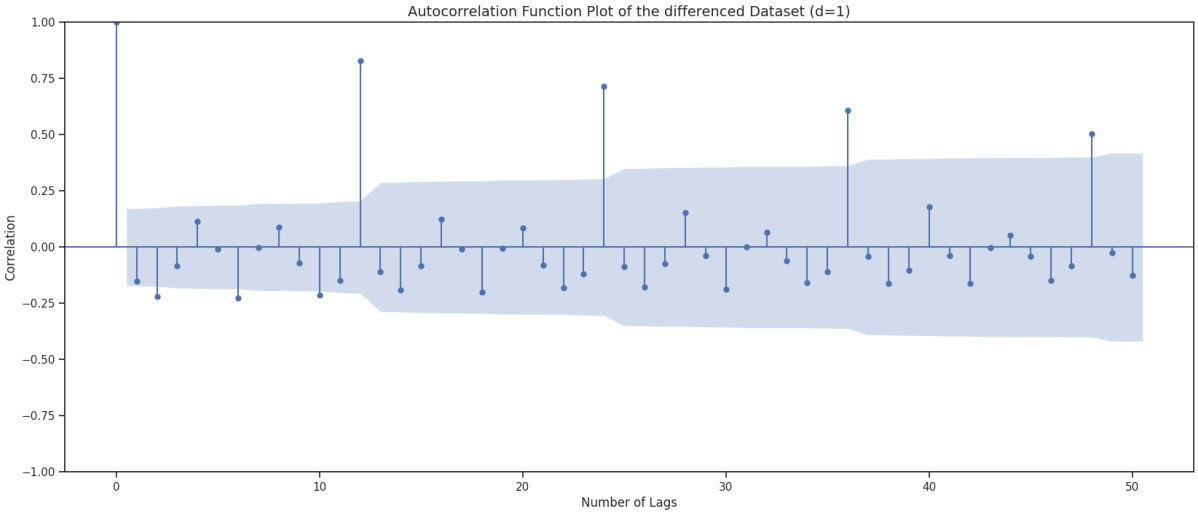
PACF plot

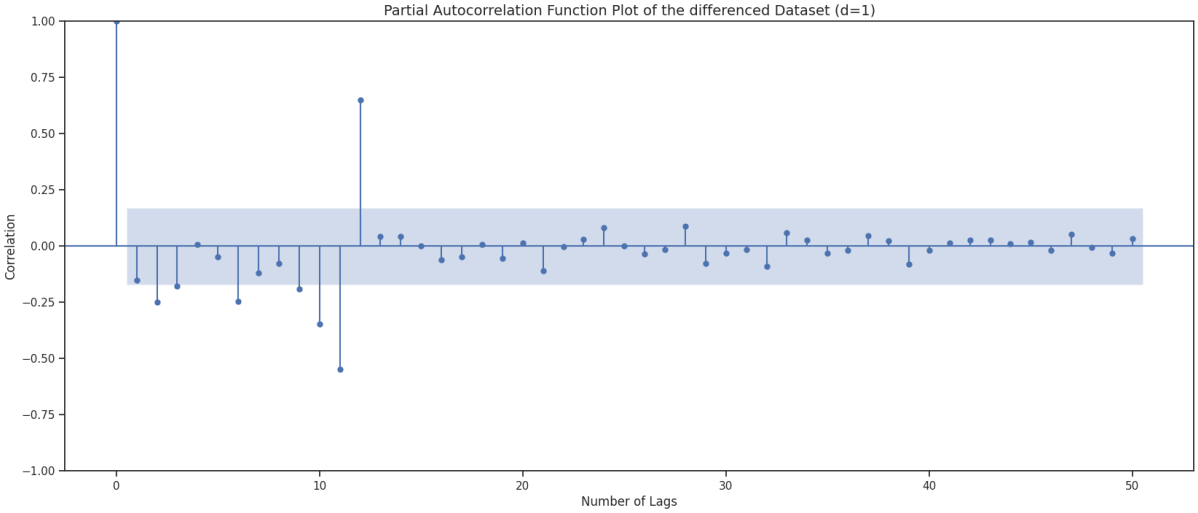

Partial Autocorrelation Function Plot of the Entire Dataset
Correlation
Number of Lags


Partial Autocorrelation Function Plot of the differenced Dataset (d=1)
Correlation
Number of Lags


Time series of Train data
Sparkling_Wine_Sales
Wine units sold
Time (in Year-Month)


Time series of Train data with order of difference 1
Sparkling_Wine_Sales
Wine units sold
2000
0
−2000
−4000
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
Time (in Year-Month)


Autocorrelation Function Plot of the differenced Dataset (d=1)
Correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
Number of Lags


Partial Autocorrelation Function Plot of the differenced Dataset (d=1)
Correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
Number of Lags

6. . Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Model 8 – Auto-Regressive Integrated Moving Average (ARIMA1)

Auto-regression means regression of a variable on itself. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR. ARIMA models may be used to represent any "non-seasonal" time series that has patterns and isn't just random noise.

An ARIMA model is characterized by 3 terms: p, d, q where, p is the order of the Auto Regressive (AR) term q is the order of the Moving Average (MA) term d is the number of differencing required to make the time series stationary

ARIMA Model

Examples of the parameter combinations for the Model

Model: (0, 1, 0)

Model: (0, 1, 1)

Model: (0, 1, 2)

Model: (0, 1, 3)

Model: (0, 1, 4)

Model: (1, 1, 0)

Model: (1, 1, 1)

Model: (1, 1, 2)

Model: (1, 1, 3)

Model: (1, 1, 4)

Model: (2, 1, 0)

Model: (2, 1, 1)

Model: (2, 1, 2)

Model: (2, 1, 3)

Model: (2, 1, 4)

Model: (3, 1, 0)

Model: (3, 1, 1)

Model: (3, 1, 2)

Model: (3, 1, 3)

Model: (3, 1, 4)

Model: (4, 1, 0)

Model: (4, 1, 1)

Model: (4, 1, 2)

Model: (4, 1, 3)

Model: (4, 1, 4)

SARIMAX Results

==========

Dep. Variable: Sparkling_Wine_Sales No. Observations: 132

Model: ARIMA(4, 1, 4) Log Likelihood -1097.960

Date: Sun, 04 May 2025 AIC 2213.920

Time: 17:56:01 BIC 2239.797

Sample: 01-31-1980 HQIC 2224.435

- 12-31-1990

Covariance Type: opg

==========

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4654	0.117	-3.976	0.000	-0.695	-0.236
ar.L2	-0.4749	0.072	-6.637	0.000	-0.615	-0.335
ar.L3	-0.4571	0.095	-4.811	0.000	-0.643	-0.271
ar.L4	0.5246	0.069	7.614	0.000	0.390	0.660
ma.L1	0.0152	1.601	0.010	0.992	-3.123	3.154
ma.L2	0.0258	3.067	0.008	0.993	-5.985	6.036
ma.L3	-0.0710	1.427	-0.050	0.960	-2.868	2.726

ma.L4 -0.9700 0.155 -6.261 0.000 -1.274 -0.666

sigma2 9.083e+05 6.54e-06 1.39e+11 0.000 9.08e+05 9.08e+05

==========
=

Ljung-Box (L1) (Q): 0.04 Jarque-Bera (JB): 1.20

Prob(Q): 0.83 Prob(JB): 0.55

Heteroskedasticity (H): 2.84 Skew: 0.22

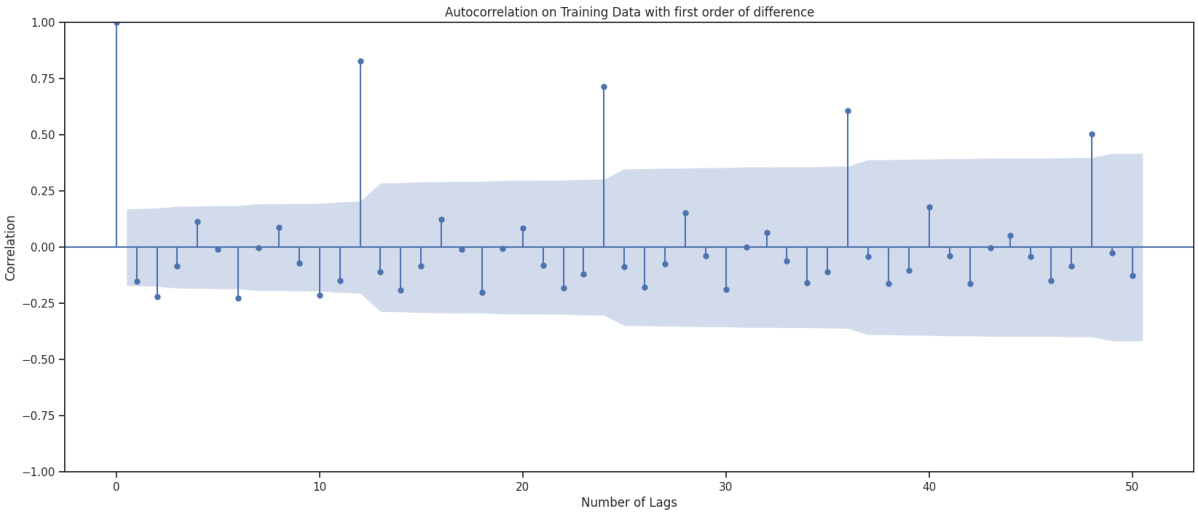
Prob(H) (two-sided): 0.00 Kurtosis: 3.14

Model 9 – Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

SARIMA models or also known as Seasonal ARIMA is an extension of ARIMA for a time series data with defined seasonality. SARIMA models use seasonal differencing which is similar to regular differencing.

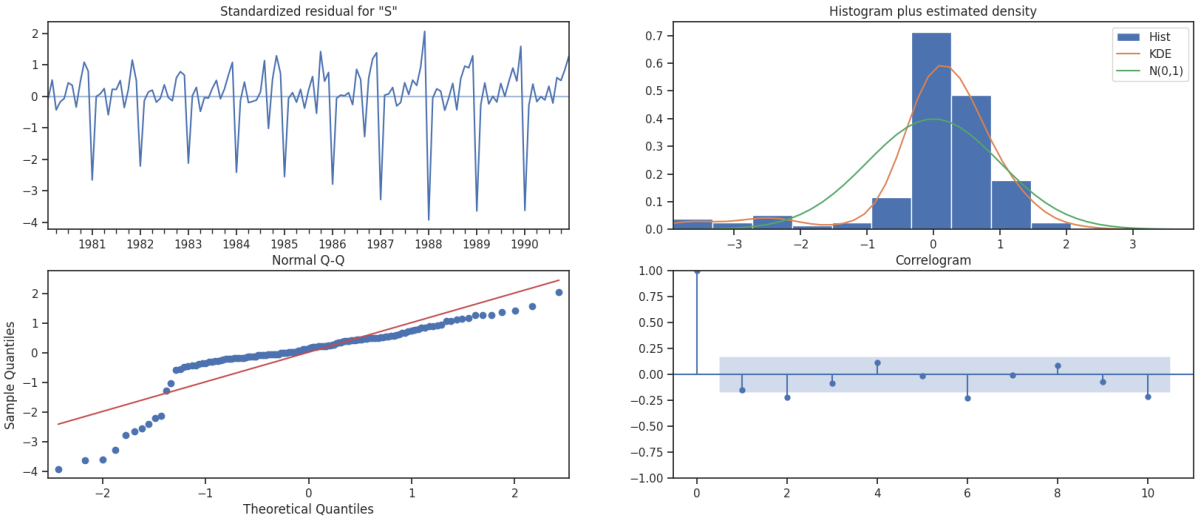
A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F where, p is the order of the Auto Regressive (AR) term q is the order of the Moving Average (MA) term d is the number of differencing required to make the time series stationary P is the order of the Seasonal Auto Regressive (AR) term Q is the order of the Seasonal Moving Average (MA) term D is the number of seasonal differencing required to make the time series stationary F is the seasonal frequency of the time series

We must examine the PACF and ACF plots, respectively, at delays that are the multiple of "F" in order to determine the "P" and "Q" values, and determine where these cut-off values are (for appropriate confidence interval bands). By examining the lowest AIC values, we can also estimate "p," "q," "P," and "Q" for the SARIMA models. By examining the ACF plots, one may calculate the seasonal parameter 'F'. The existence of seasonality should be shown by a spike in the ACF plot at multiples of "F."



From the above ACF plot we can observe that at every 12th lag is significant indicating the presence of seasonality. Hence for our model building we will consider the term F=12.

Automated SARIMA – Diagnostics plot



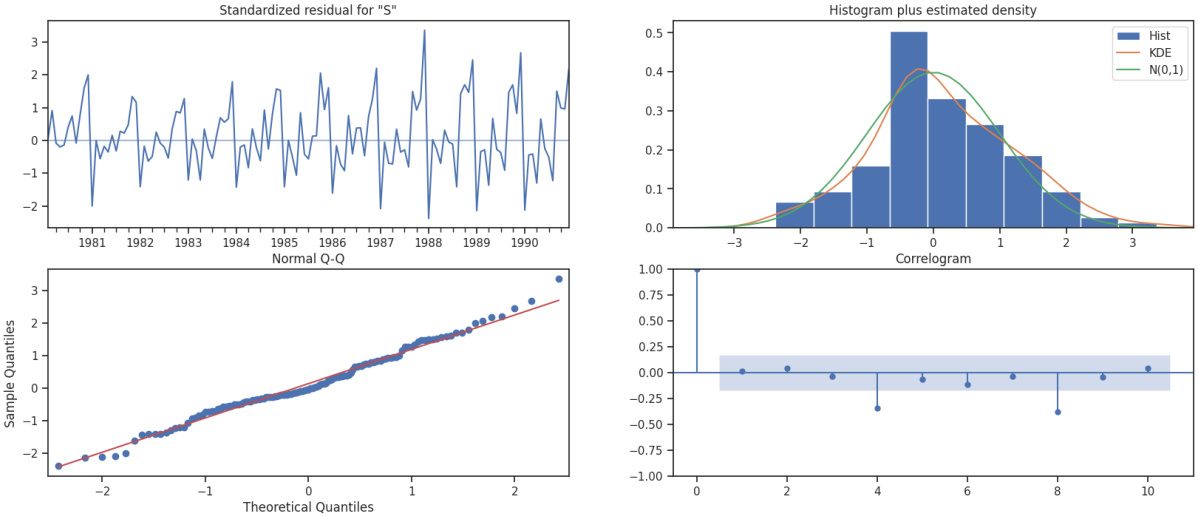
Observation:

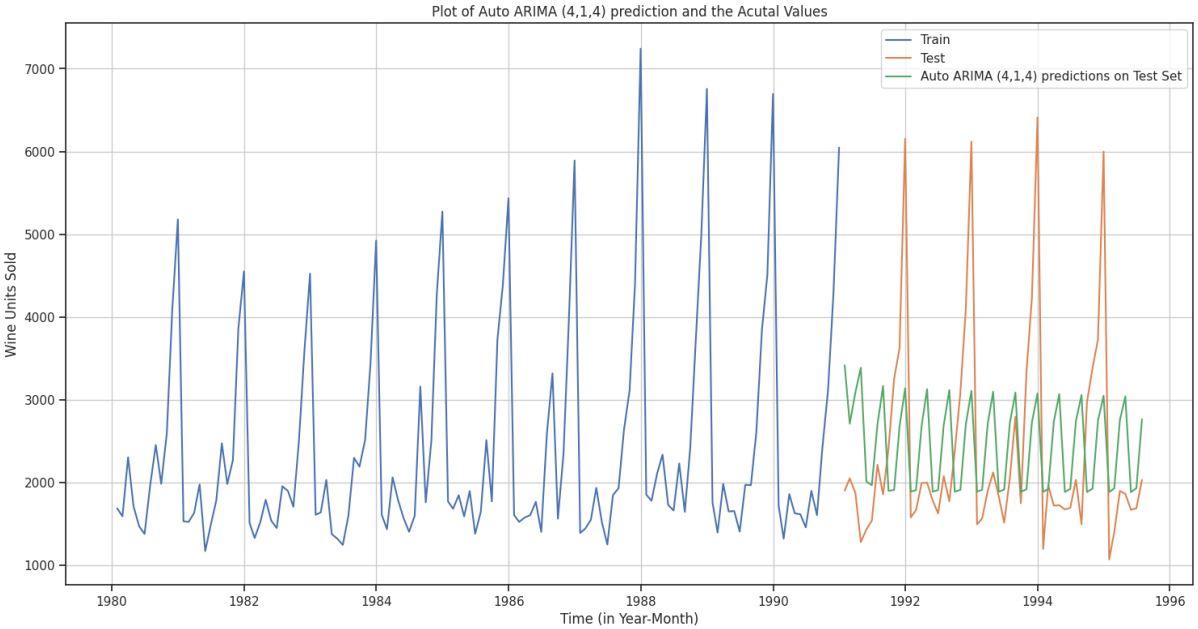
• The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. The AIC is lowest for the combination (3,1,2) (3,0,1,12) as we see from the above results.

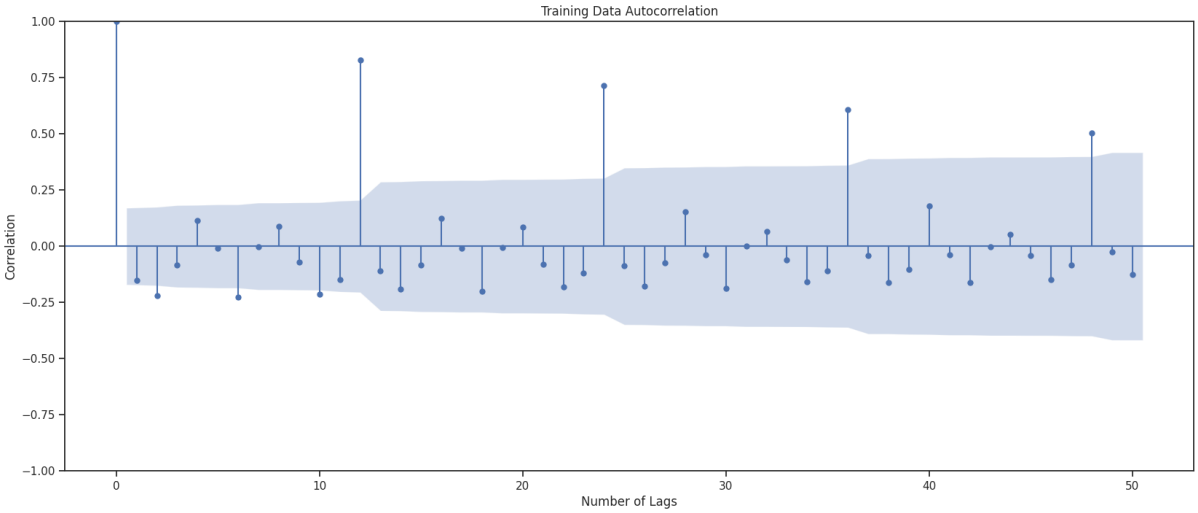
• From the Standardized residual plot above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.

• The histogram plus estimated density plot suggests a slightly uniform distribution with mean zero and slightly skewed to the right.

• In Normal Q-Q plot, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution • The correlogram plot of residuals shows that the residuals are not auto correlated.







Observation:

• The optimal parameters are decided based on the lowest Akaike Information Criteria (AIC) values. The AIC is lowest for the combination (4,1,4) as we see from the above results.

• From the Standardized residual plot above, we can notice that the residuals seem to fluctuate around the mean of zero and have uniform variance.

• The histogram plus estimated density plot suggests a slightly uniform distribution with mean zero and slightly skewed to the right.

• In Normal Q-Q plot, all the dots fall more or less in line with the red line. Few deviations are present implying minor skewed distribution.

• The correlogram plot of residuals shows that the residuals are not auto correlated.

Automated ARIMA: Model Evaluation

For evaluating the model's performance metrics, we look at root means squared error (RMSE) & mean absolute percentage error (MAPE)

Model	Test RMSE	Test MAPE
ARIMA (p=4, d=1, q=4)	1212.918	40.214

Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• ARIMA models performs well on non-seasonal time series. It is due to this reason it is unable to capture the entire characteristics of the test data.

• The root means squared error (RMSE) of test data for the ARIMA model with (p=4, d=1, q=4) is 1212.918.

• Not surprisingly, the RMSE of the aforementioned ARIMA model is lower than the majority of previously constructed models but significantly higher than triple exponential smoothing model.

Examples of the parameter combinations for the Model are

Model: (0, 1, 0)(0, 0, 0, 12)

Model: (0, 1, 1)(0, 0, 1, 12)

Model: (0, 1, 2)(0, 0, 2, 12)

Model: (0, 1, 3)(0, 0, 3, 12)

Model: (1, 1, 0)(1, 0, 0, 12)

Model: (1, 1, 1)(1, 0, 1, 12)

Model: (1, 1, 2)(1, 0, 2, 12)

Model: (1, 1, 3)(1, 0, 3, 12)

Model: (2, 1, 0)(2, 0, 0, 12)

Model: (2, 1, 1)(2, 0, 1, 12)

Model: (2, 1, 2)(2, 0, 2, 12)

Model: (2, 1, 3)(2, 0, 3, 12)

Model: (3, 1, 0)(3, 0, 0, 12)

Model: (3, 1, 1)(3, 0, 1, 12)

Model: (3, 1, 2)(3, 0, 2, 12)

Model: (3, 1, 3)(3, 0, 3, 12)

SARIMAX Results

==========

Dep. Variable: Sparkling_Wine_Sales No. Observations: 132

Model: SARIMAX(3, 1, 2)x(3, 0, [1], 12) Log Likelihood -684.301

Date: Sun, 04 May 2025 AIC 1388.603

Time: 18:10:31 BIC 1413.820

Sample: 01-31-1980 HQIC 1398.781

- 12-31-1990

Covariance Type: opg

==========

coef std err z P>|z| [0.025 0.975]

---

ar.L1	-0.5433	0.416	-1.307	0.191	-1.358	0.272
ar.L2	-0.0076	0.198	-0.038	0.970	-0.396	0.381
ar.L3	0.0636	0.140	0.453	0.651	-0.212	0.339
ma.L1	-0.1992	0.404	-0.493	0.622	-0.992	0.593
ma.L2	-0.6548	0.327	-2.005	0.045	-1.295	-0.015
ar.S.L12	0.7651	0.448	1.706	0.088	-0.114	1.644
ar.S.L24	0.1092	0.330	0.331	0.741	-0.537	0.756
ar.S.L36	0.1764	0.186	0.946	0.344	-0.189	0.542
ma.S.L12	-0.2427	0.451	-0.539	0.590	-1.126	0.640
sigma2	1.663e+05	2.63e+04	6.326	0.000	1.15e+05	2.18e+05

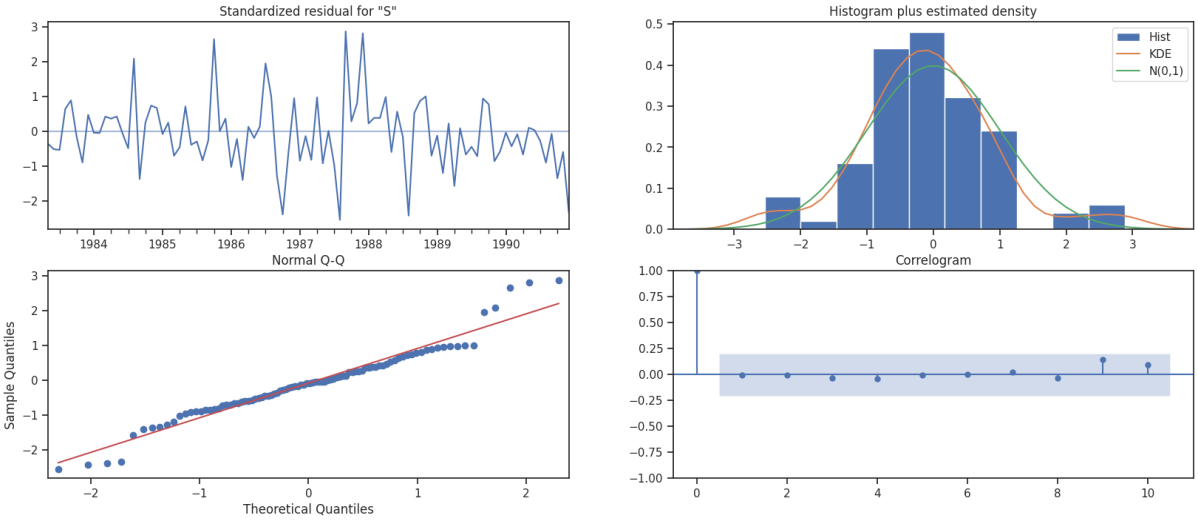
==========
=

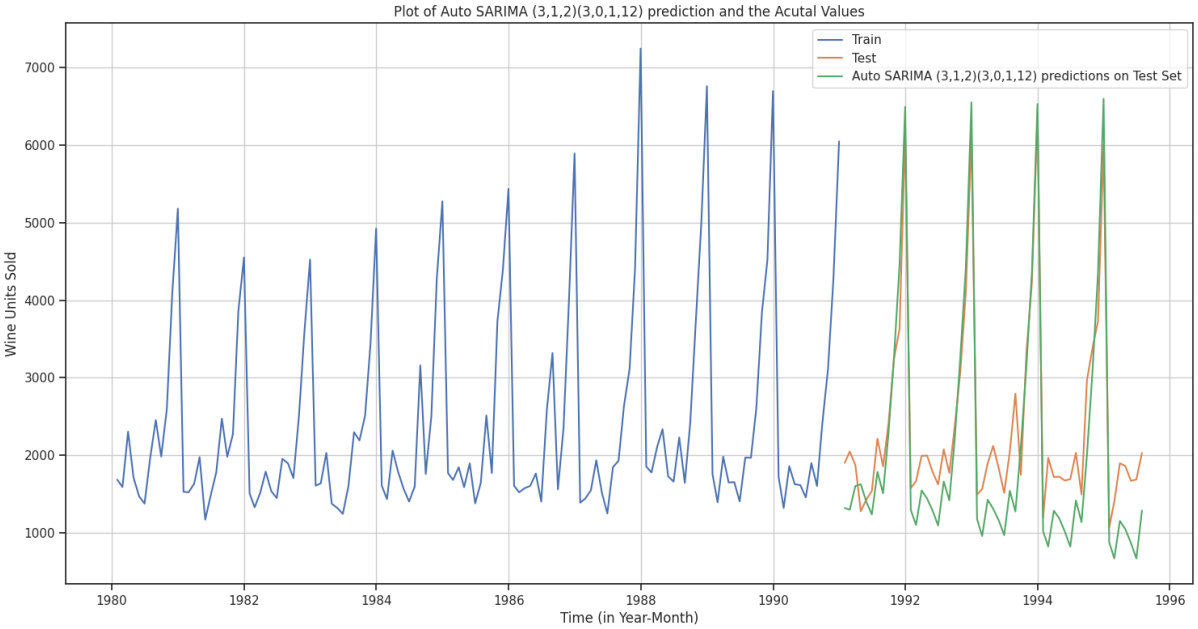
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	9.36
Prob(Q):	0.96	Prob(JB):	0.01
Heteroskedasticity (H):	1.25	Skew:	0.35
Prob(H) (two-sided):	0.54	Kurtosis:	4.40

==========
=

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).





Observation:

• We can see from the graphs above that the time series has a marginal upward trend and seasonality

• SARIMA model performs well on seasonal time series. It is due to this reason it is able to capture the entire characteristics of the test data.

• The root means squared error (RMSE) of test data for the SARIMA model with (p=3, d=1, q=2) (P=3, D=0, Q=1, F=12) is 579.925.

• Additionally, it should be highlighted that compared to the ARIMA model, the SARIMA model has almost more than halved the RMSE value.

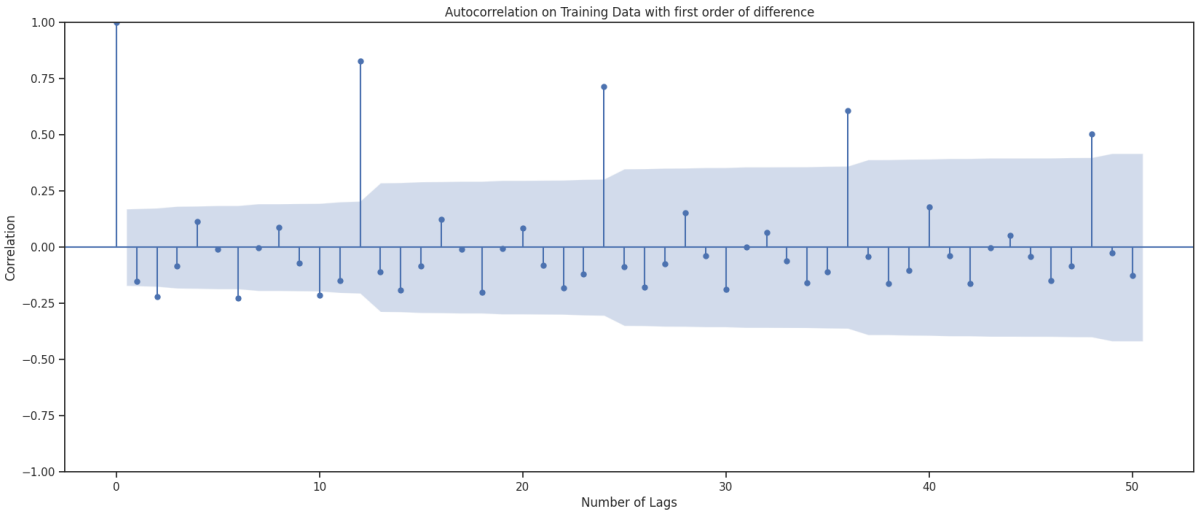
6. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

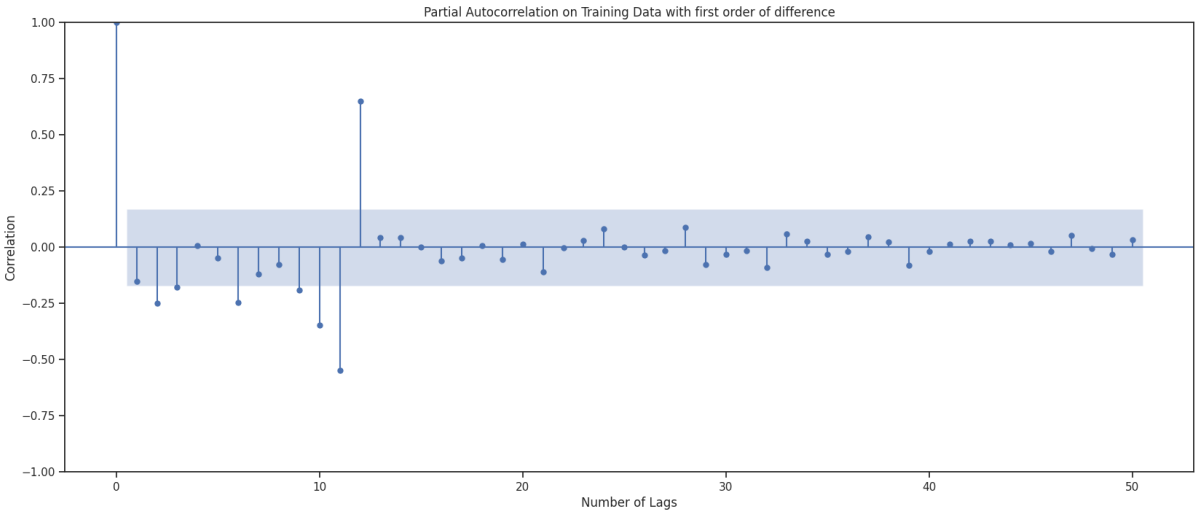
Model 10 – Auto-Regressive Integrated Moving Average (ARIMA) – Manual

An ARIMA model is characterized by 3 terms: p, d, q where, p is the order of the Auto Regressive (AR) term q is the order of the Moving Average (MA) term d is the number of differencing required to make the time series stationary

Indicating which previous series values are most beneficial in forecasting future values, autocorrelation and partial autocorrelation are measures of relationship between present and past series values. You may identify the sequence of processes in an ARIMA model using this information. The parameters p & q can be determined by looking at the PACF & ACF plots respectively. Autocorrelation function (ACF) - At lag k, this is the correlation between series values that are k intervals apart. Partial autocorrelation function (PACF) - At lag k, this is the correlation between series values that are k intervals apart, accounting for the values of the intervals between. In an ACF & PACF plots, each bar represents the size and direction of the connection. Bars that cross the red line are statistically significant.

Manual ARIMA Model



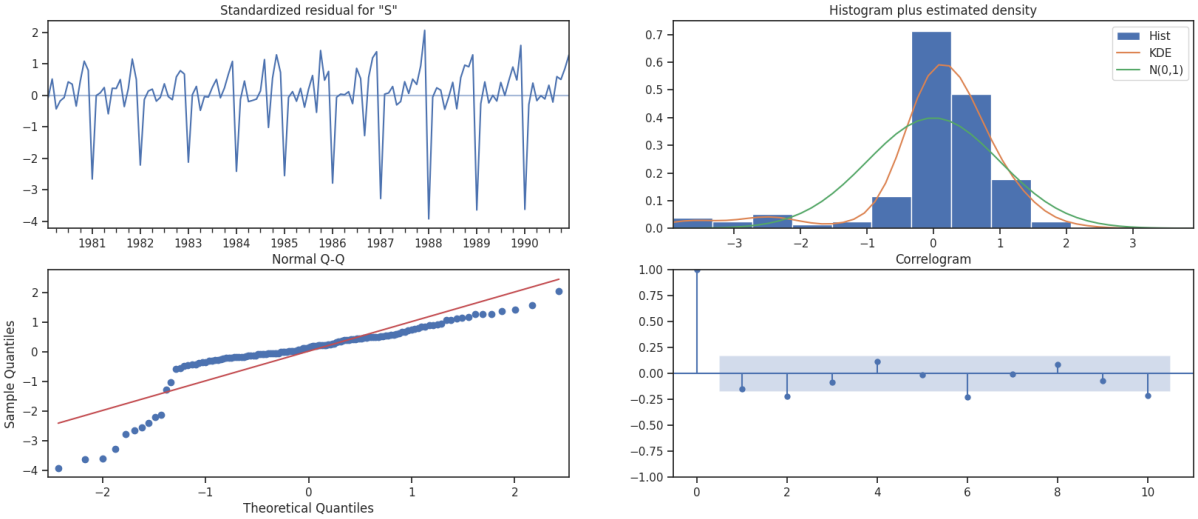


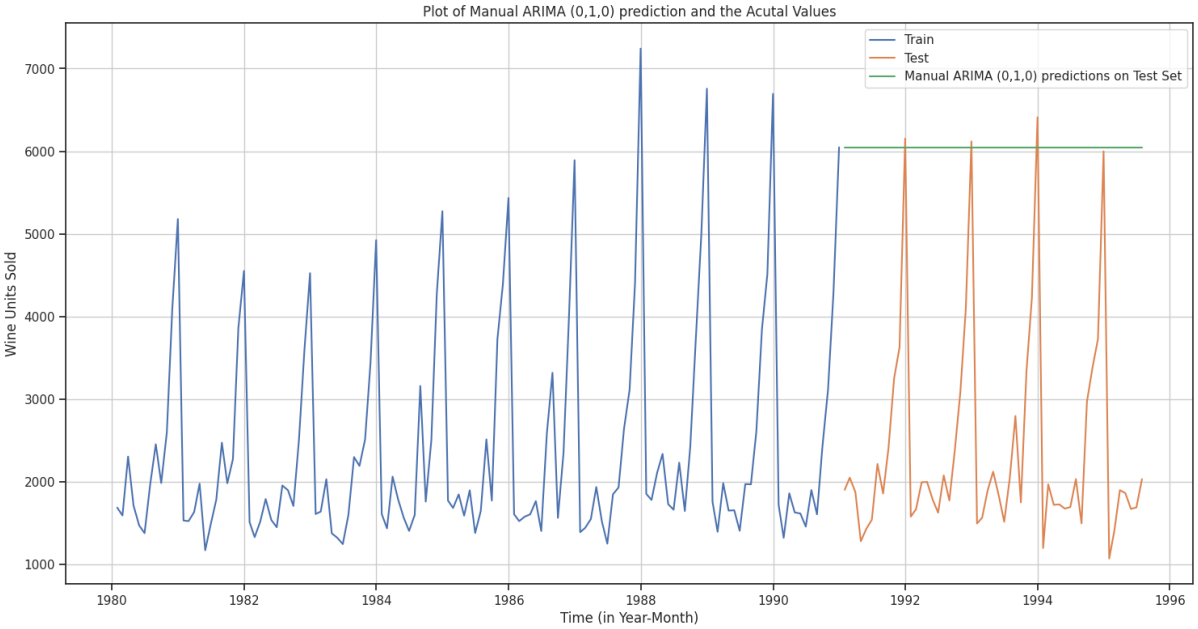
Here, we have taken alpha=0.05.

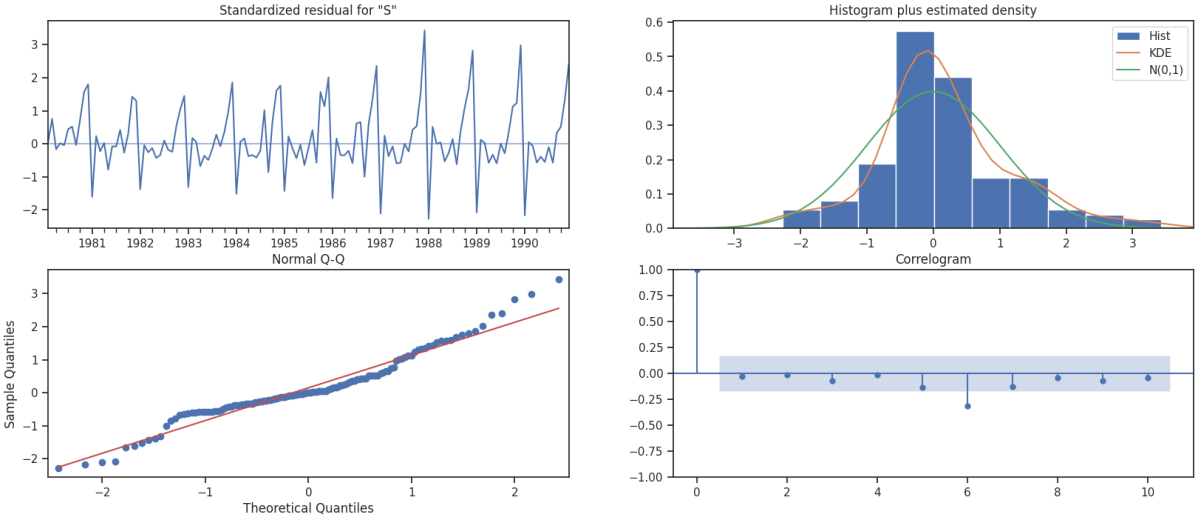
- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag after which the PACF plot cuts-off to 0.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off to 0.

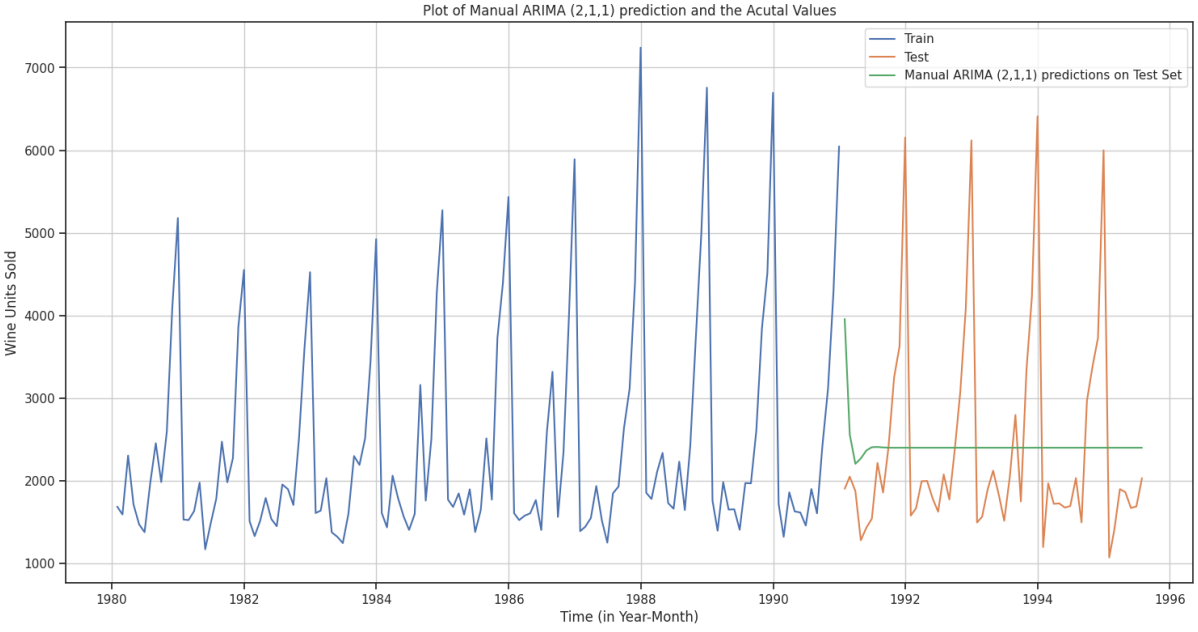
By looking at the above plots, we can see that first lag cuts off in both plots and hence we start from lags after lag 0. Therefore we have taken the value of p and q to be 2 and 1 respectively.

We would also build a model with p=0,q=0




Plot of Manual ARIMA (0,1,0) prediction and the Acutal Values
Train
Test
Manual ARIMA (0,1,0) predictions on Test Set
Wine Units Sold
Time (in Year-Month)


Standardized residual for "S"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
Correlogram


Plot of Manual ARIMA (2,1,1) prediction and the Acutal Values
Train
Test
Manual ARIMA (2,1,1) predictions on Test Set
Wine Units Sold
Time (in Year-Month)

Observation:

• The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag after which the PACF plot cuts-off below the confidence interval.

• The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag after which the ACF plot cuts-off below the confidence interval.

• We can observe from the above plots that after lag 1, we have few significant lags and hence we would also build another model by taking value of p=2 and q=1 respectively

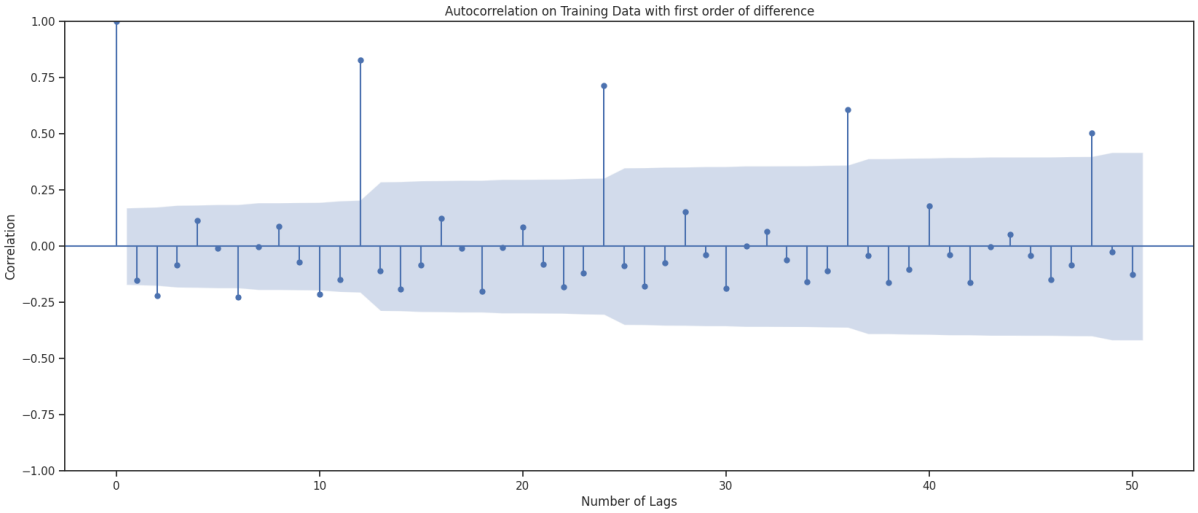
Manual SARIMA Model

Let us look at the ACF and the PACF plots once more.

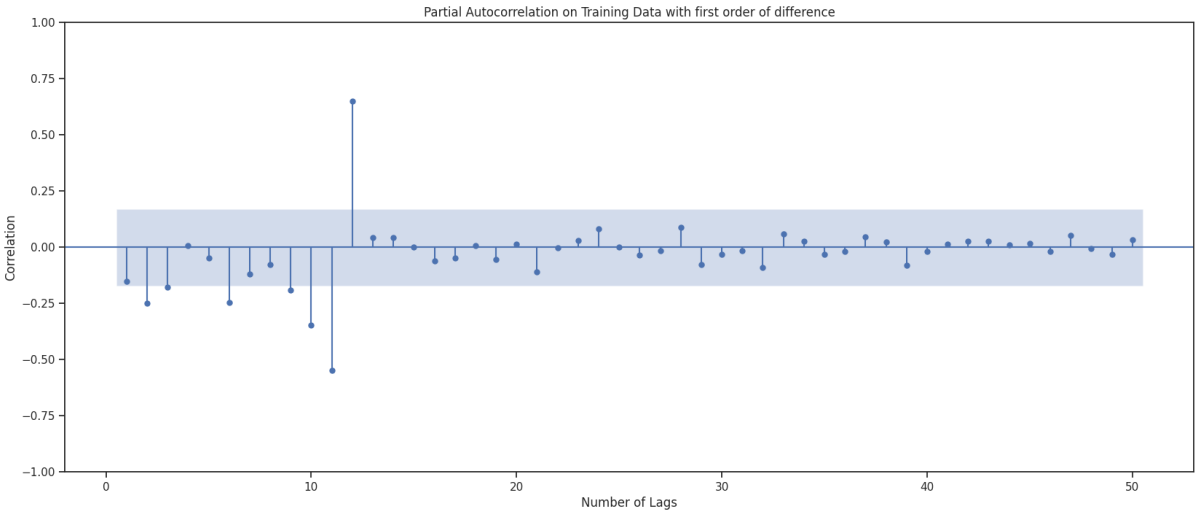
---

keyboard_arrow_down

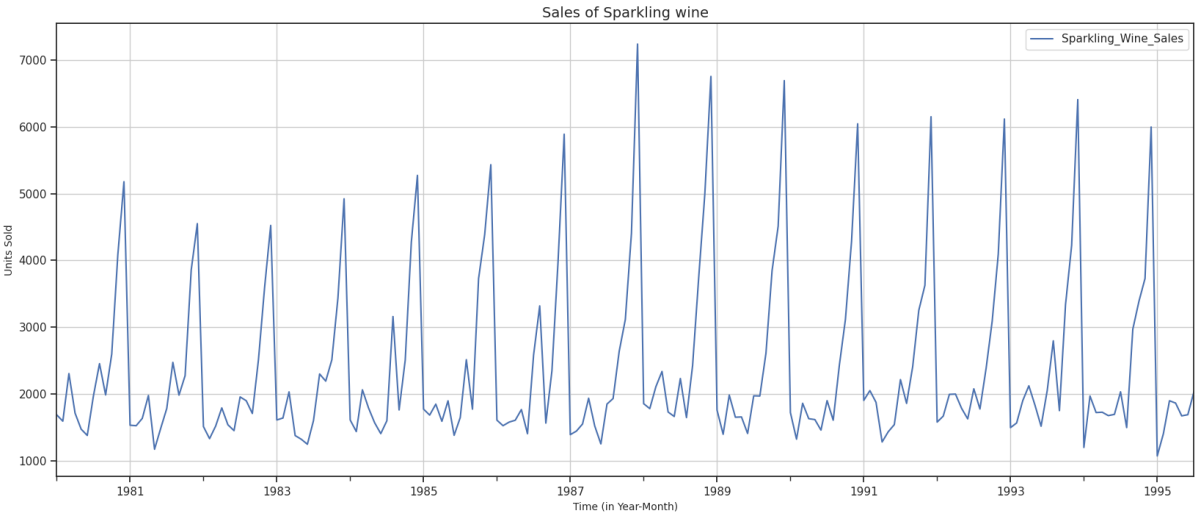
ACF plot

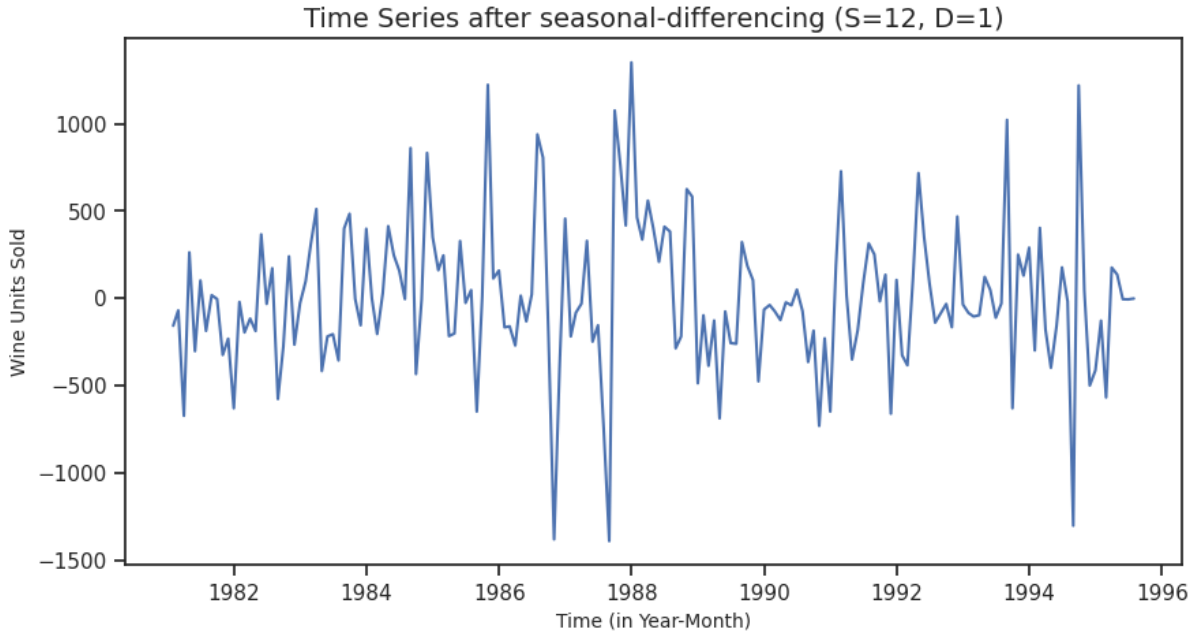


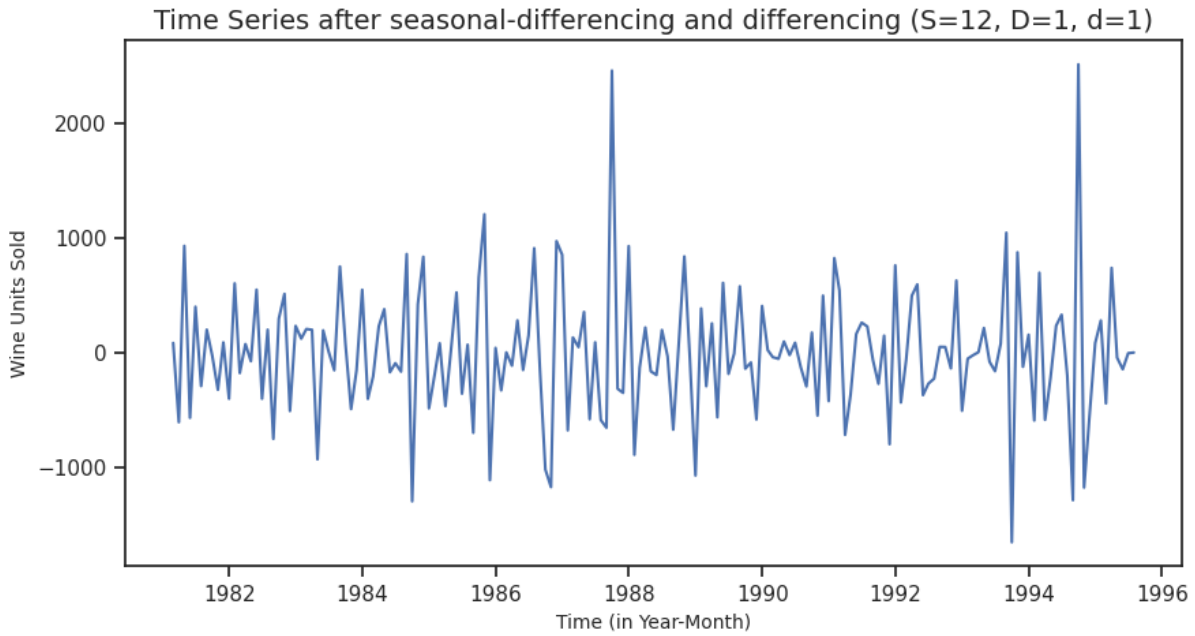
PACF plot

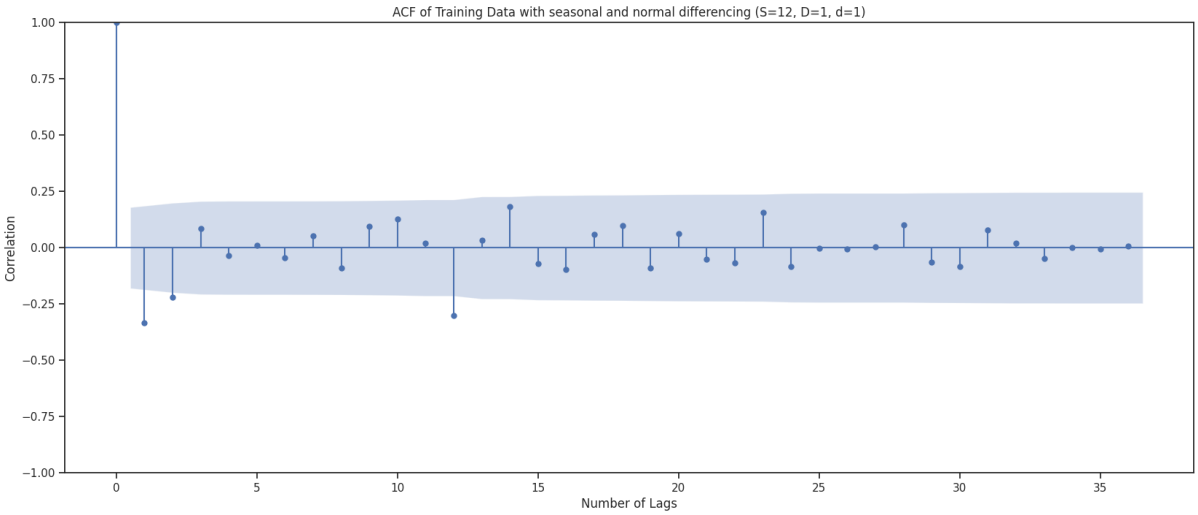


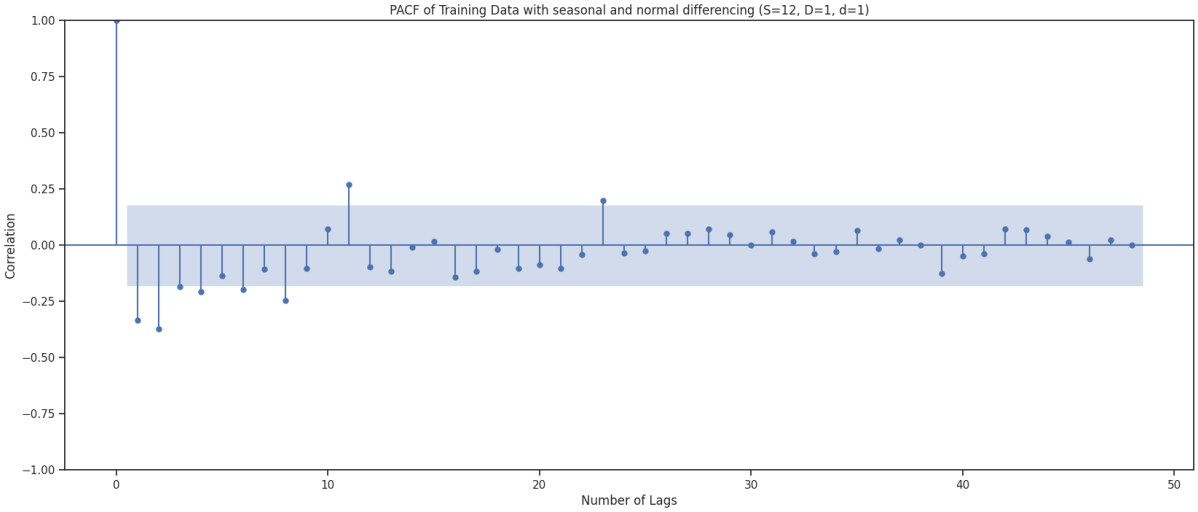
Observation: • From the PACF plot it can be seen in early lags that till lag 4 is significant before cutoff, so AR term ‘p = 4’ is chosen. From the multiples of seasonal lags, after first seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’. • From ACF plot, it can be seen in early lags, lag 1 and 2 are significant before it cuts off, so let’s keep MA term ‘q = 2’ and at seasonal lag of 12, a significant lag is apparent and no seasonal lags are apparent at lags 24, 36 or afterwards, so let’s keep ‘Q = 1'. • The final selected terms for SARIMA model are (4, 1, 2) (0, 1, 1, 12), as inferred from the ACF and PACF plots.

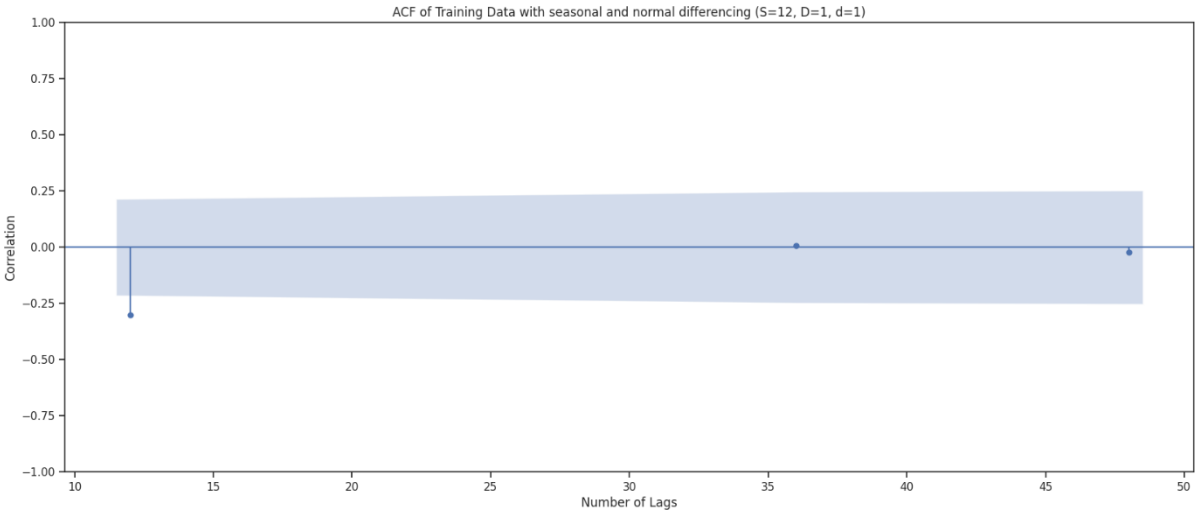

Sales of Sparkling wine
Sparkling_Wine_Sales
Units Sold
7000
6000
5000
4000
3000
2000
1000
1981
1983
1985
1987
1989
1991
1993
1995
Time (in Year-Month)

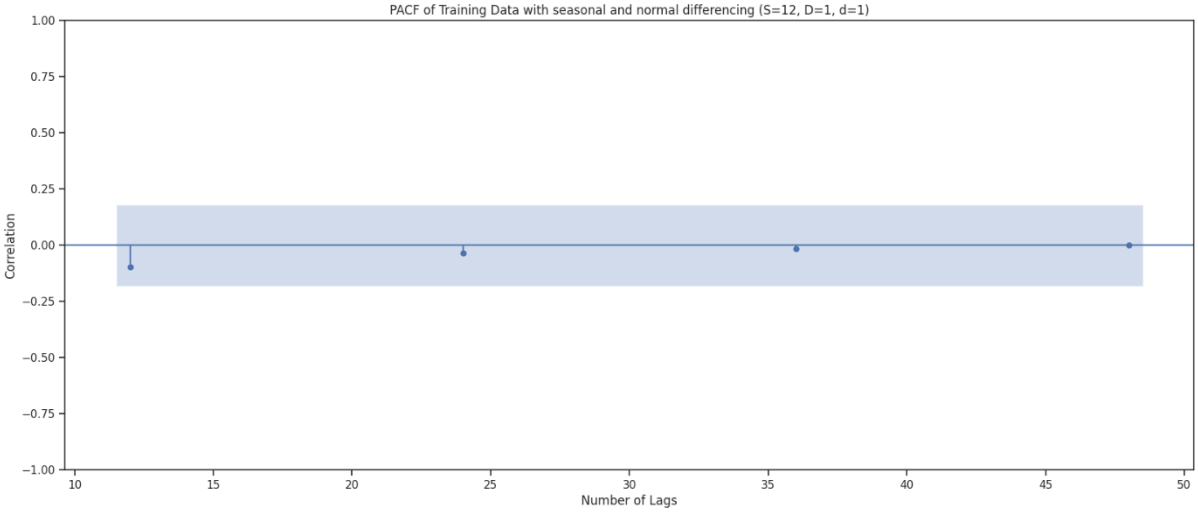

Time Series after seasonal-differencing (S=12, D=1)
Wine Units Sold
1000
500
0
−500
−1000
−1500
1982
1984
1986
1988
1990
1992
1994
1996
Time (in Year-Month)


Time Series after seasonal-differencing and differencing (S=12, D=1, d=1)
Wine Units Sold
2000
1000
0
−1000
1982
1984
1986
1988
1990
1992
1994
1996
Time (in Year-Month)

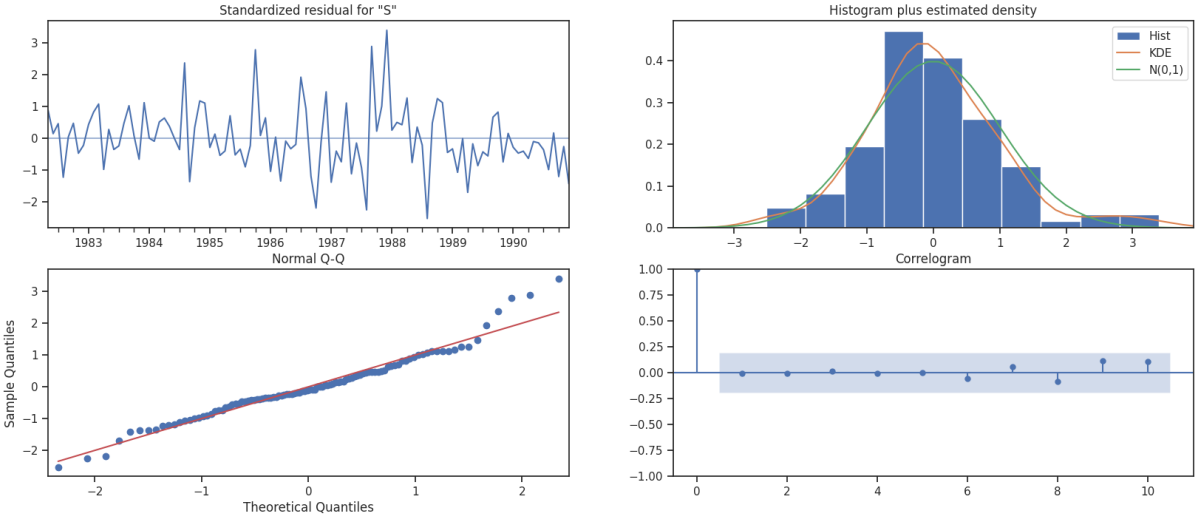

ACF of Training Data with seasonal and normal differencing (S=12, D=1, d=1)
Correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30
35
Number of Lags

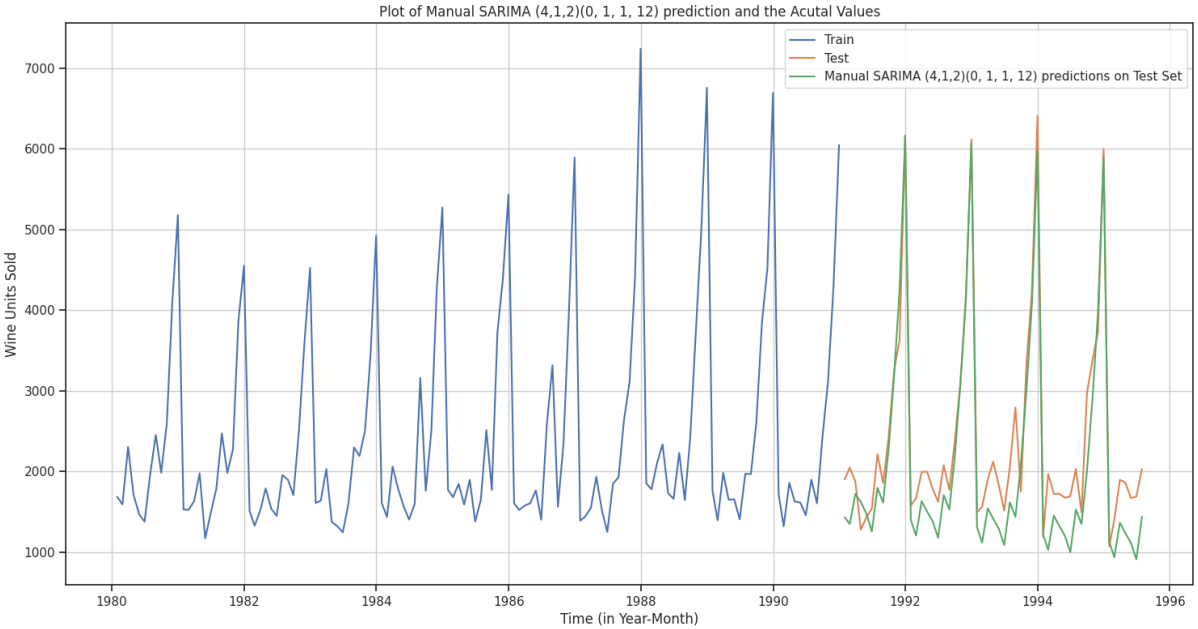

PACF of Training Data with seasonal and normal differencing (S=12, D=1, d=1)
Correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
Number of Lags





- Here we have taken alpha = 0.05 and seasonal period as 12.
- From the PACF plot it can be seen that till lag 4 is significant before cut-off, so AR term ‘p = 4’ is chosen. At seasonal lag of 12, it cuts off, so keep seasonal AR ‘P = 0’.
- From ACF plot, lag 1 and 2 are significant before it cuts off, so lets keep MA term ‘q = 2’ and at seasonal lag of 12, a significant lag is apparent and no seaonal lags are apparent at lags 24, 36 or afterwards, so lets keep ‘Q = 1'.
- The final selected terms for SARIMA model is (4, 1, 2)x(0, 1, 1, 12), as inferred from the ACF and PACF plots.

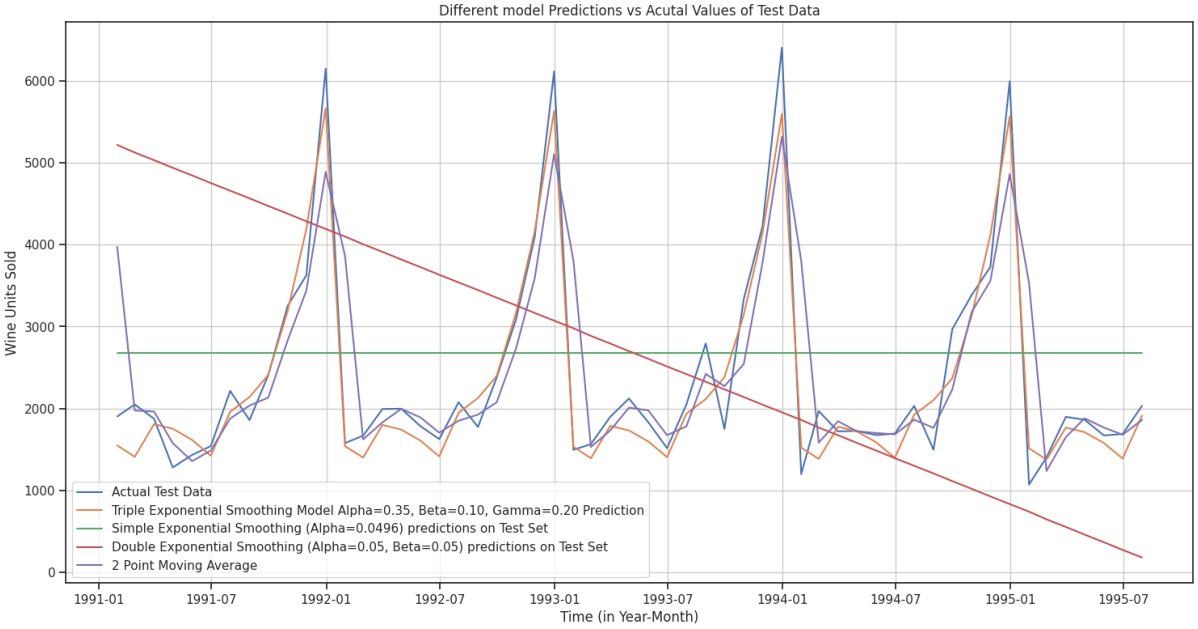

Standardized residual for "S"
Histogram plus estimated density
Hist
KDE
N(0,1)
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
Correlogram

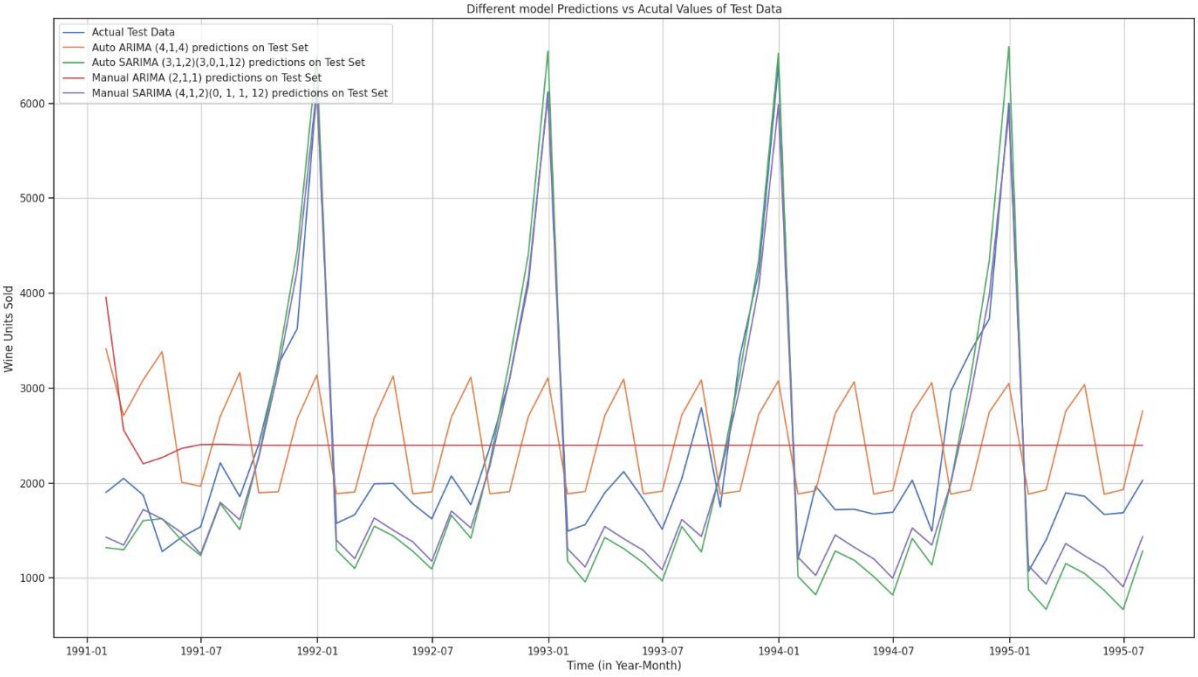

Plot of Manual SARIMA (4,1,2)(0, 1, 1, 12) prediction and the Acutal Values
Train
Test
Manual SARIMA (4,1,2)(0, 1, 1, 12) predictions on Test Set
Wine Units Sold
Time (in Year-Month)

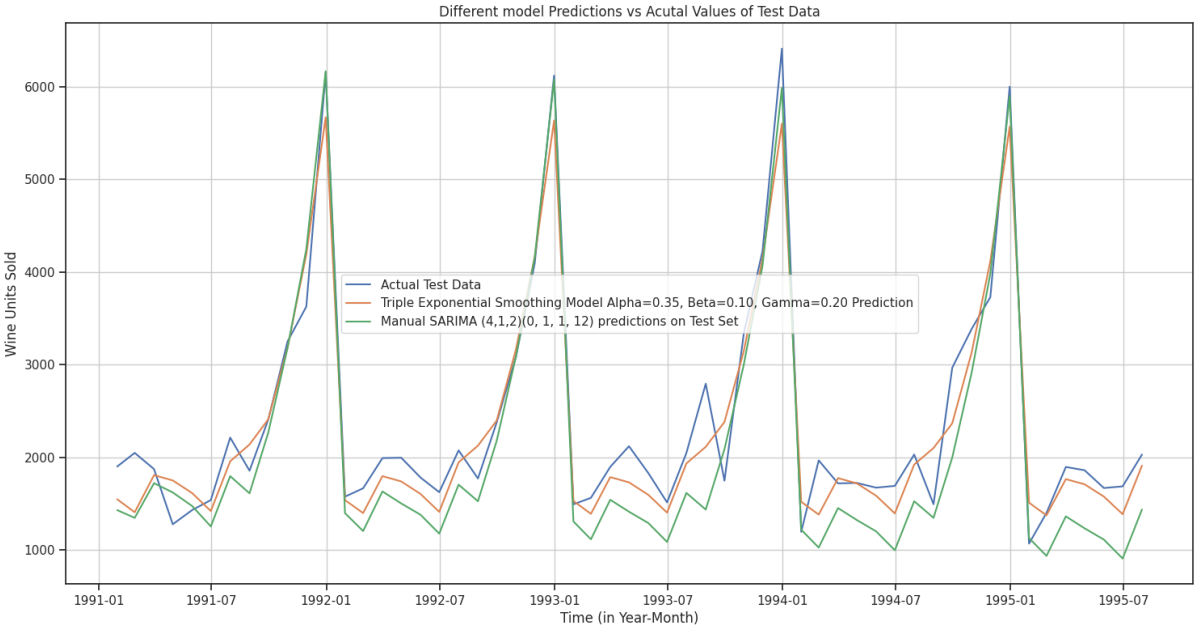
8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

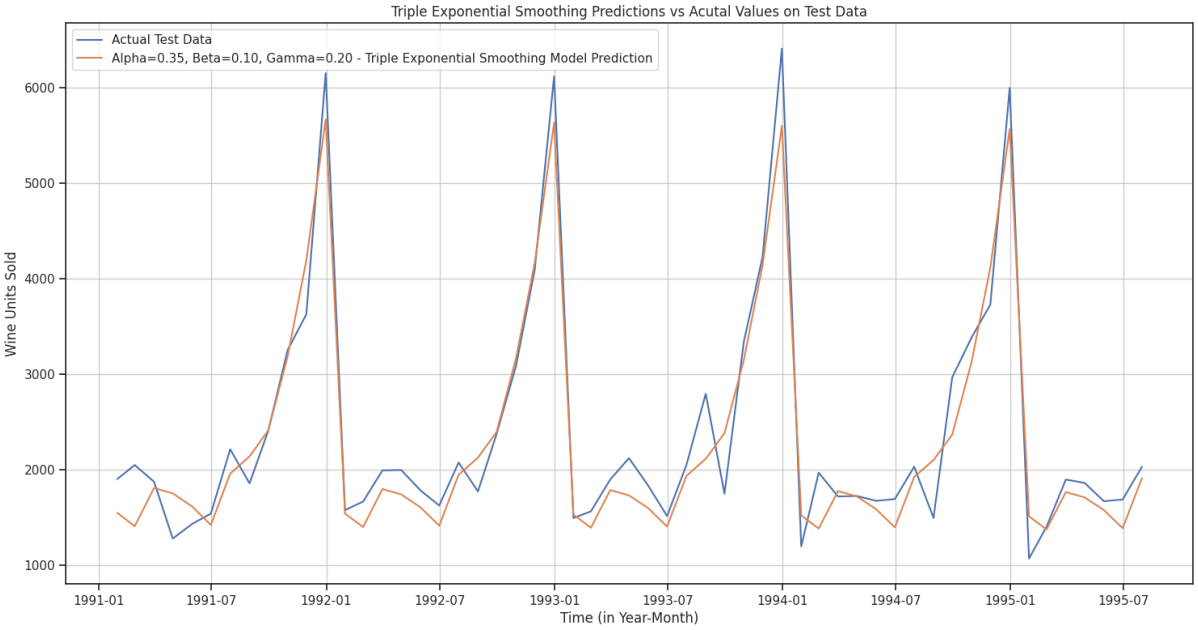
index	Test RMSE	MAPE
Linear Regression	1389.135174897992	NaN
Alpha=0.35,Beta=0.10,Gamma=0.20,Triple Exponential Smoothing	109585.97480942453	NaN
Alpha=0.111,Beta=0.0617,Gamma=0.395,Triple Exponential Smoothing	162978.71835291386	NaN
Manual SARIMA (4, 1, 2)(0, 1, 1, 12)	219658.91381149445	19.32492405578
Auto SARIMA (3,1,2)(3,0,1,12)	336424.589379628	25.057898047024
2 point TMA	661620.6727272727	NaN
4 point TMA	1337699.7204545455	NaN
Auto ARIMA (4,1,4)	1472744.4442627183	39.756151559436
Simple Average	1625833.6061179985	NaN
6 point TMA	1648469.64040404	NaN
Manual ARIMA (2,1,1)	1691876.1243400597	40.225668183948
Alpha =0.0496,SimpleExponentialSmoothing	1702835.5310907199	NaN
9 point TMA	1812465.3025813692	NaN
Alpha=0.05, Beta=0.05, Double Exponential Smoothing	2011880.3132099581	NaN
Alpha=0.6885, Beta=9.99e-05, Double Exponential Smoothing	4029006.499289429	NaN
Naive Model	14932654.909090908	NaN

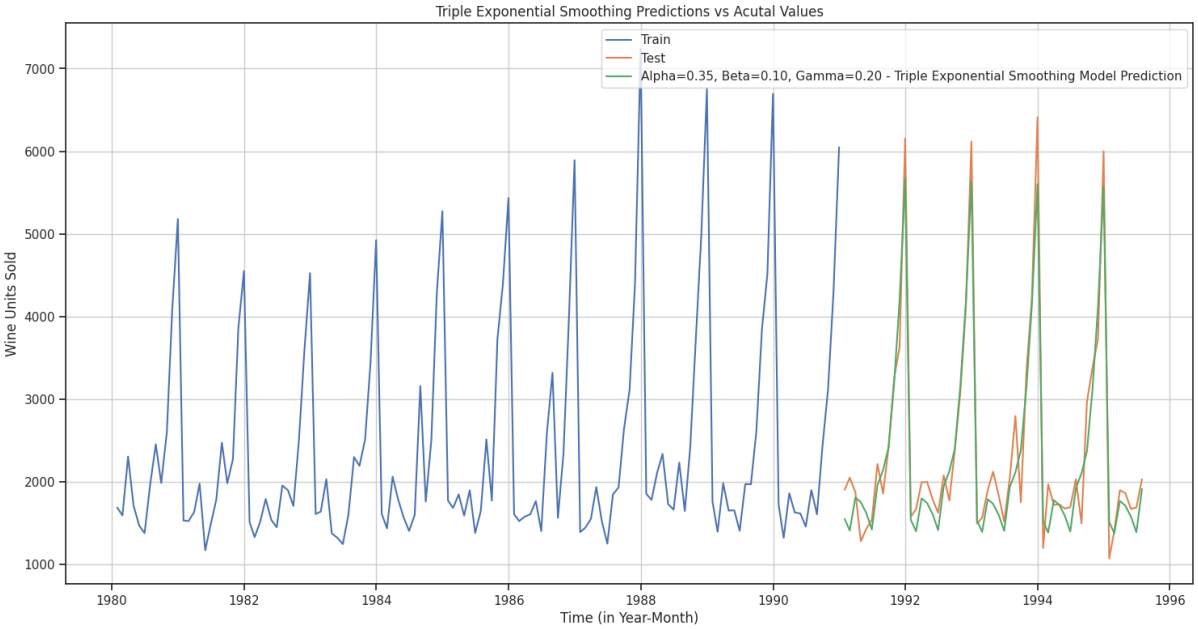
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

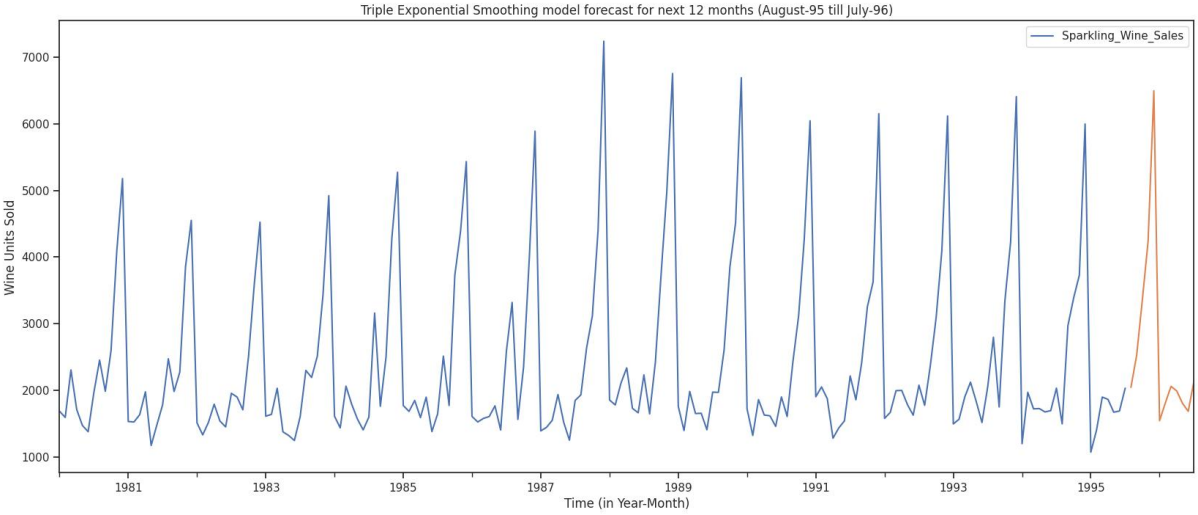




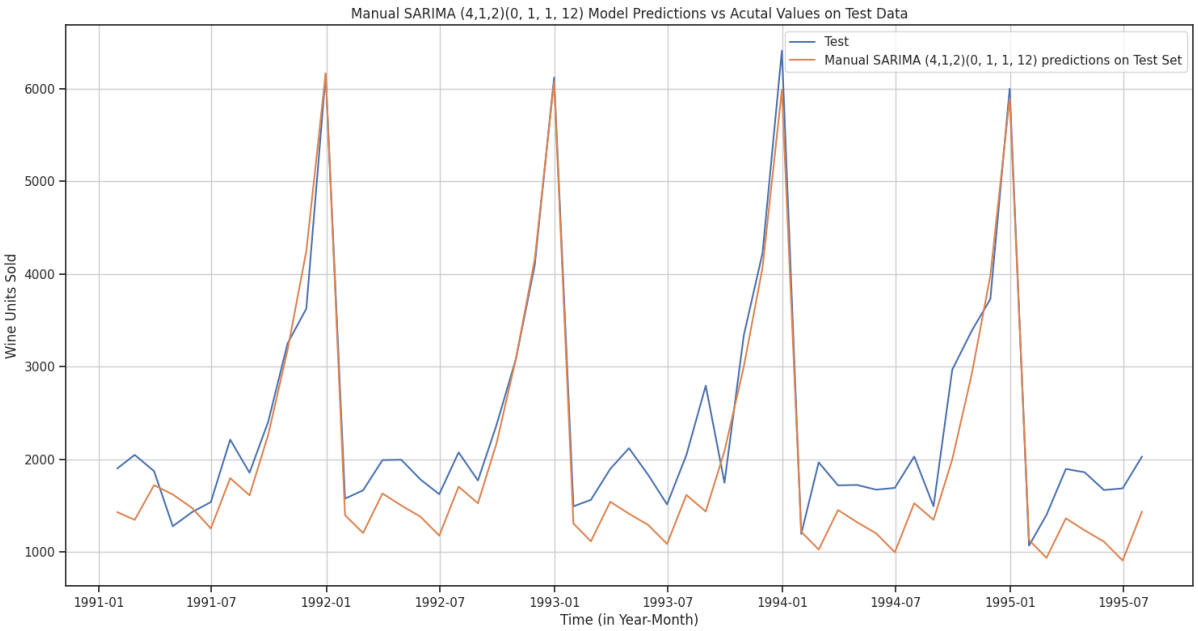

Different model Predictions vs Acutal Values of Test Data
Actual Test Data
Triple Exponential Smoothing Model Alpha=0.35, Beta=0.10, Gamma=0.20 Prediction
Manual SARIMA (4,1,2)(0, 1, 1, 12) predictions on Test Set
Wine Units Sold
Time (in Year-Month)
1000
2000
3000
4000
5000
6000
1991-01
1991-07
1992-01
1992-07
1993-01
1993-07
1994-01
1994-07
1995-01
1995-07


Triple Exponential Smoothing Predictions vs Acutal Values on Test Data
Actual Test Data
Alpha=0.35, Beta=0.10, Gamma=0.20 - Triple Exponential Smoothing Model Prediction
Wine Units Sold
Time (in Year-Month)
1000
2000
3000
4000
5000
6000
1991-01
1991-07
1992-01
1992-07
1993-01
1993-07
1994-01
1994-07
1995-01
1995-07


Triple Exponential Smoothing Predictions vs Acutal Values
Train
Test
Alpha=0.35, Beta=0.10, Gamma=0.20 - Triple Exponential Smoothing Model Prediction
Wine Units Sold
Time (in Year-Month)


Triple Exponential Smoothing model forecast for next 12 months (August-95 till July-96)
Sparkling_Wine_Sales
Wine Units Sold
Time (in Year-Month)

Optimum Model- Manual SARIMA Model (4, 1, 2)(0, 1, 1, 12)



10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

We needed to construct an optimum model to forecast the sparkling wine sales for the next 12 months. The model information, insights and recommendations are as follows.

Model Insights:

• The time series in consideration exhibits a little rising trend and stable seasonality. When comparing the various models, we can see that Triple Exponential Smoothing and SARIMA models frequently deliver the greatest results. This is due to the fact that these models are excellent at predicting time series that demonstrate trend and seasonality. • We examine the root mean squared value of the forecast model to assess its performance (RMSE). The model with the lowest RMSE value and characteristics that match the test data is regarded as being a superior model. • We observed that SARIMA and the Triple Exponential Smoothing model had the lowest RMSE and the characteristics that most closely fit test data. As a result, they are regarded as the best models for forecasting. Historical Insights: • The sparkling wine sales have remained stable throughout time. Sparkling wine sales peaked in 1988 and fell to their present low position in 1995 (as we have data for only first 7 months). • The monthly sales trajectory appears to be exactly the opposite of the yearly plot, with a progressive increase towards the end of each year.

January has the lowest wine sales, while December has the highest. From January to August, sales increase gradually, and then they quickly increase after that. • The average monthly sales of sparkling wine are 2402 bottles. More than 50% of the sold units of sparkling wine fall between 1605 and 2549. 1070 units were sold as the lowest and 7242 units as the most. Only 25% of monthly sales that were recorded were for more than 2549 units. • Around 60 to 70 percent of the units sold are fewer than 2500, and 80% of the units sold are less than 4000. Only 20% of sales involved more than 3000 items. Therefore, it is clear that the bulk of sales were in the range of 1000 to 3000 units.

Forecast Insights:

• Based on the forecast made by the Triple Exponential Smoothing model previously presented, the following insights are offered. • The forecast calls for average sale of 2639 units, up 237 units from the historical average of 2402 units. Thus, we might observe an increase in average sales of 10%. • The prediction is for a minimum sales volume of 1540 units, which is 470 units more than the minimum sales volume of 1070 units in the past. Consequently, a 43% increase in minimum sales is seen. • The projection estimates a maximum sales volume of 6487 units, which is 755 units fewer than the largest sales volume recorded in the past, which was 7242 units. Consequently, a 10% decrease in maximum sales is visible.

• In comparison to the historical standard deviation of 1295 recorded in the past, the forecast's standard deviation is 1439 units, or 144 units higher. It's gone up by 11%. This is also anticipated because historical data tends to have less volatility than future data. • We can see from the prediction that the months of October, November, and December have increased sales. December is often when the sales are at their highest. There is a startling decline in sales in January following December. The months after January appear to witness a gradual improvement in sales until October, when it jumps sharply.

Recommendations:

• Records show that the months of September, October, November, and December account for 50% of the total sales forecast. Many festivities take place in these months, and many people travel during this time. One of the most popular types of wine used during festive and event celebrations is sparkling wine. • Wine sales often climb in the final two months of the year as people hurry to buy holiday beverages. For forthcoming occasions like Thanksgiving, Christmas, and New Year's, people typically stock up. The majority of individuals also buy in bulk for holiday gatherings and gift-giving. • Many individuals choose wine as their go-to gift when it comes to occasions like parties and gift-giving. Sales of sparkling wine rise just before the winter holidays as more collectors purchase these wines as presents or look for vintages to serve at holiday gatherings. • The festival seasons may vary depending on where you are geographically, however the most of the celebrations take place in the last four months. ♣ In these months, promotional offers might be implemented to lower costs and significantly boost revenue. ♣ To increase sales, we must take advantage of all holiday events and set prices appropriately. ♣ Many individuals order in bulk to prepare for upcoming festivities, which may result in a high shipping expenditure. Businesses may provide significant discounts or free shipping beyond a certain threshold at these times. ♣ Giving customers gifts to improve their user experience is one of the greatest marketing strategies to deploy. In order to attract more consumers and increase sales, the company might provide free gifts on orders with significant sales. ♣ To target various client demographics, the proper marketing campaigns must be run ♣ Numerous ecommerce campaigns and competitions may be performed to broaden the product's audience and enhance sales. • The period from January to June is one of the key challenges for sparkling wine sales. ♣ To identify the elements affecting sales, in-depth market research must be conducted. ♣ Due to the fact

that sparkling wines are typically used while celebrating, a market-friendly version of the existing product might be introduced by the company, helping to make up for the drop in sales. Long-term, this may bring in additional clients. • There are other key elements that might be driving the sales, despite the present model's ability to closely track the historical sales trend. ♣ The forecast might be improved by doing in-depth market research on the factors that influence sales and incorporating that information into the model for projection.

THANK YOU