TIME SERIES FORECASTING
FINAL PROJECT

TABLE OF CONTENTS

1. Rose Wine Analysis
2. Sparkling Wine Analysis

Project - Rose Wine Analysis

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.

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 Rose wine from 20th century.

Data Dictionary -

Variable Name	Description
YearMonth	Represents the year and month in which the sales were recorded
Rose	Denotes the number of wine units sold

Data Description

1. YearMonth: Data time variable from 1980-01-01 to 1995-07-01
2. Rose: Continuous Data

1 - Define the problem and perform Exploratory Data Analysis

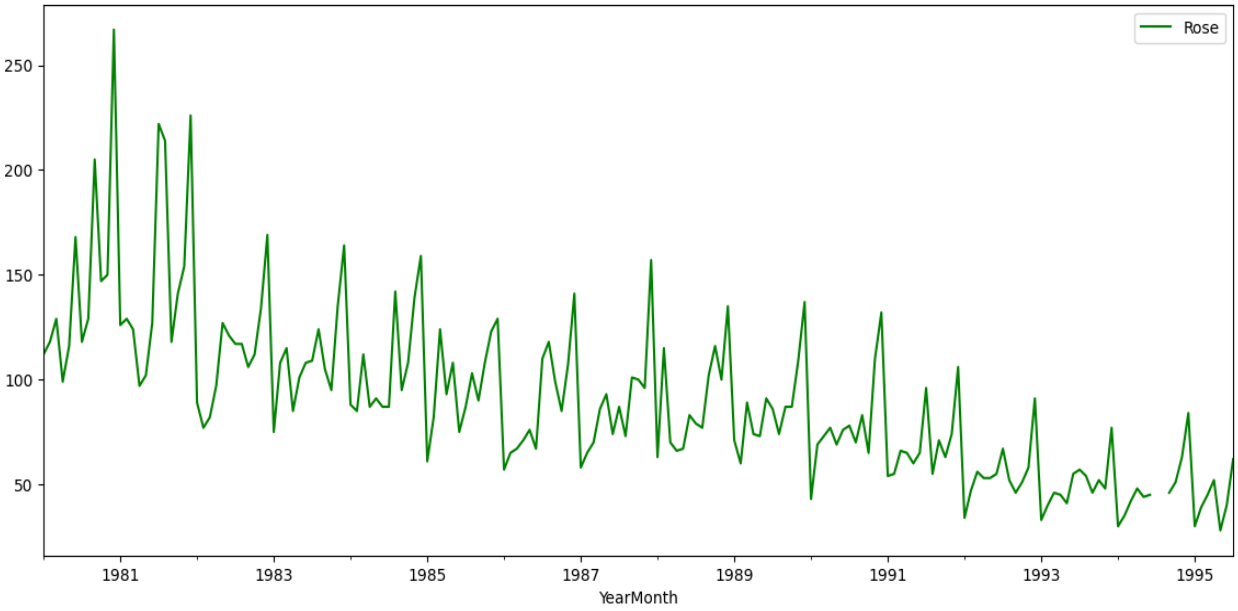
Check the data type and columns details –

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

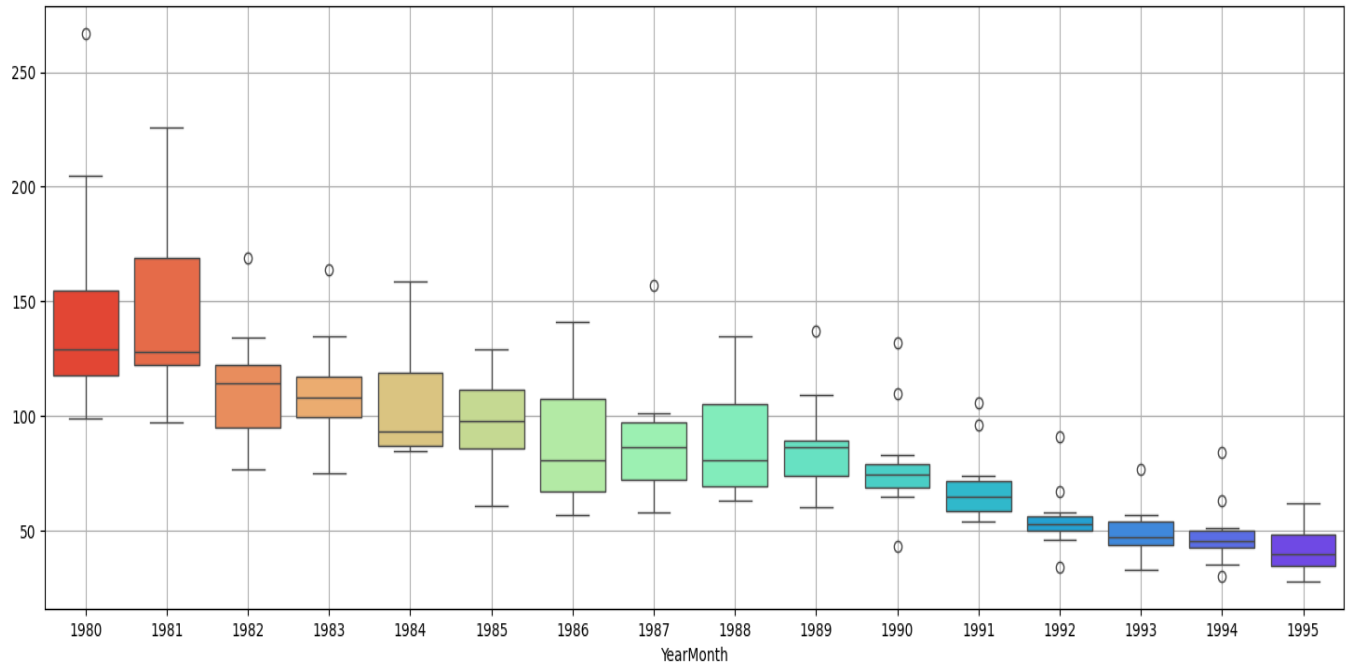
Check the Top 5 and Last 5 Rows details.

```
Top 5 Rows in dataset
              Rose
YearMonth
1980-01-01  112.0
1980-02-01  118.0
1980-03-01  129.0
1980-04-01   99.0
1980-05-01  116.0

Last 5 Rows in dataset
              Rose
YearMonth
1995-03-01  45.0
1995-04-01  52.0
1995-05-01  28.0
1995-06-01  40.0
1995-07-01  62.0
```

Plot the Data - As can be seen from the above figure, there are 2 null values present in the dataset. Since it's a time series we cannot remove it and hence must be imputed



Year Wise Boxplot - To understand the spread of sales across different years and within different months across years.



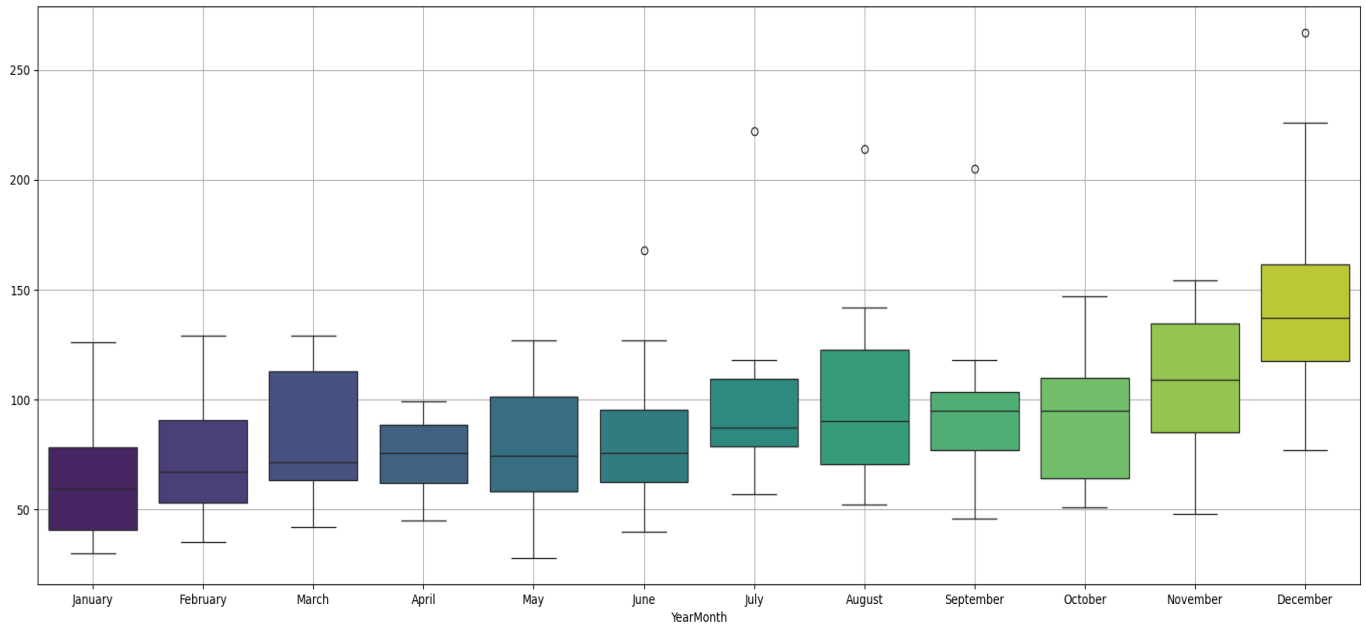
Observation:

We can see from the figure above that sales of rose wine have been declining over time.

• After 1992, the median sales have been at their lowest levels, having peaked in 1980 and 1981.

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

Month-Wise Boxplot -



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.

Convert into Date and Month –

YearMonth	Rose	Month	Year
1980-01-01	112.0	1	1980
1980-02-01	118.0	2	1980
1980-03-01	129.0	3	1980
1980-04-01	99.0	4	1980
1980-05-01	116.0	5	1980

Rename the Column Name –

The below mentioned columns of the data frame have been renamed as shown.

Original Column Name	Renamed Column Name
Rose	Sales

YearMonth	Sales	Month	Year
1980-01-01	112.0	1	1980
1980-02-01	118.0	2	1980
1980-03-01	129.0	3	1980
1980-04-01	99.0	4	1980
1980-05-01	116.0	5	1980

Check the Missing Value – We can observe that there are two 2 values in dataset are missed.

YearMonth	Sales	Month	Year
1994-07-01	NaN	7	1994
1994-08-01	NaN	8	1994

Impute the Missing value - We have imputed the missing values with Mean function.

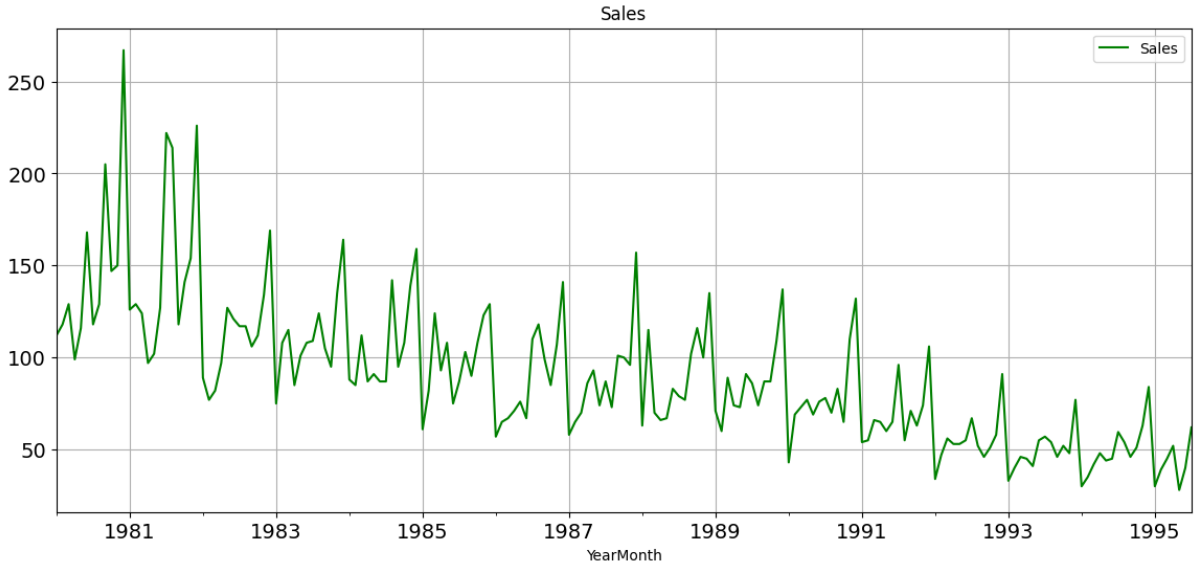
Check the Null value after Treatment -

```
df.isna().sum()
```

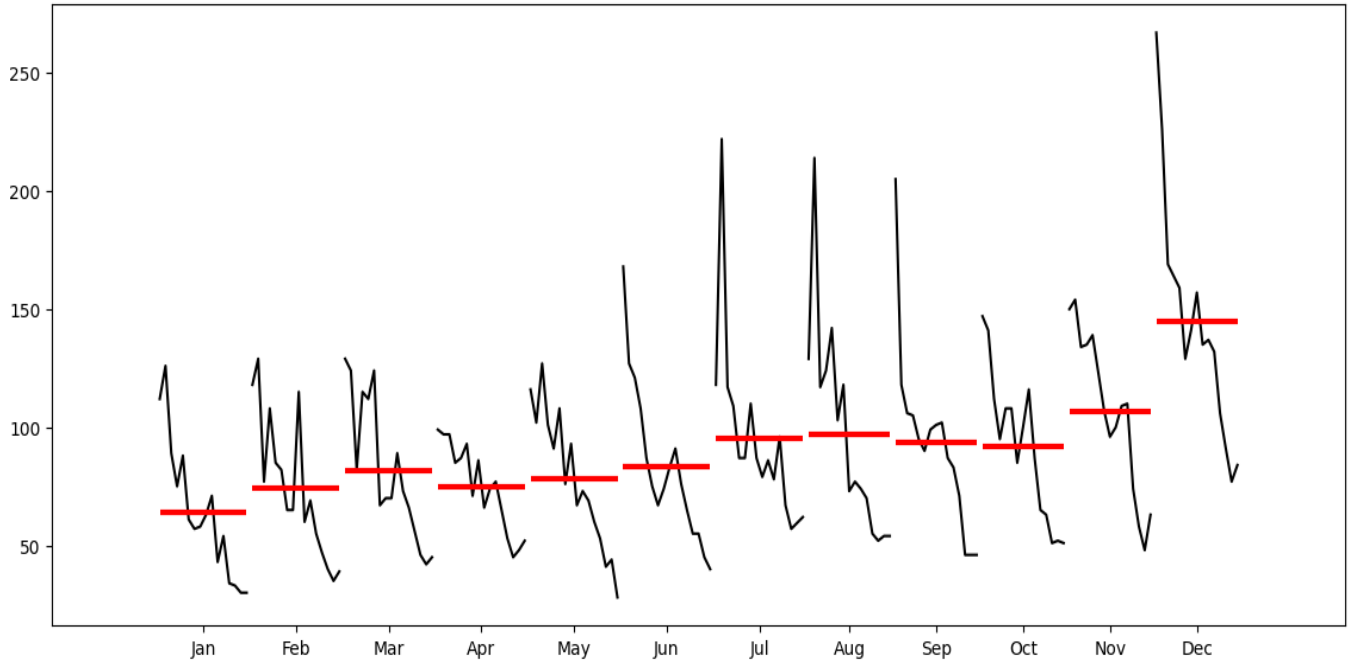
	0
Sales	0
Month	0
Year	0

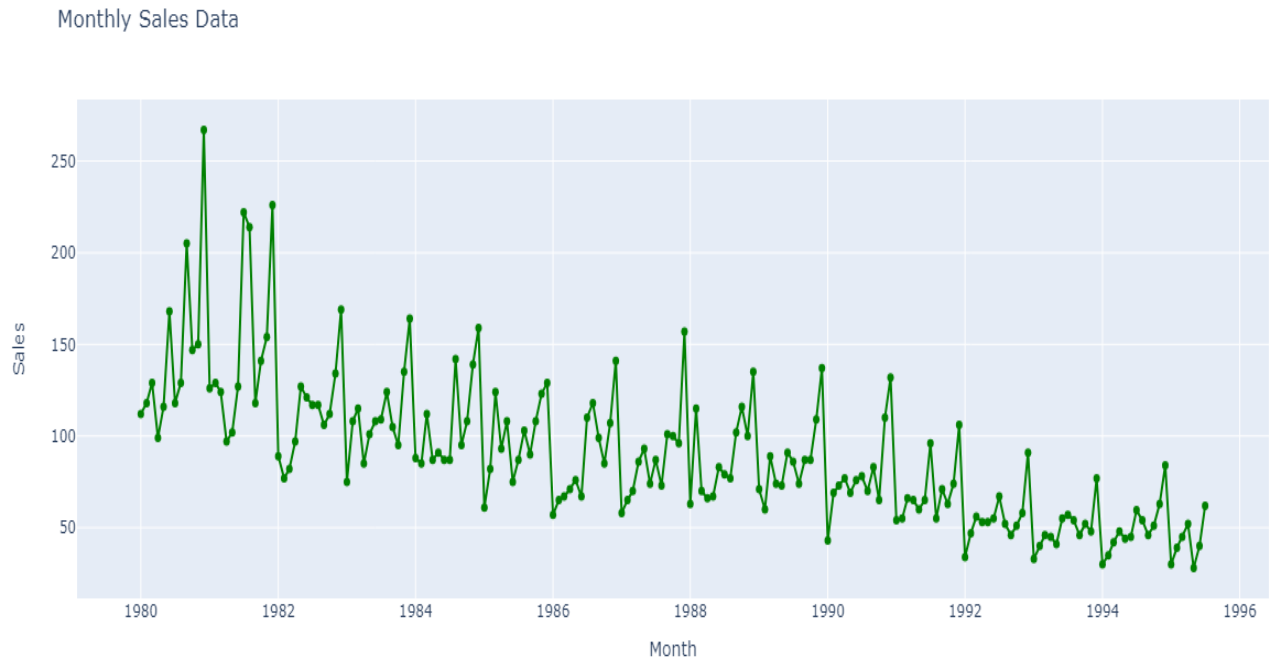
dtype: int64

Plot the Data after Treating –



Plot the Time Series according to different Months for Different Years -



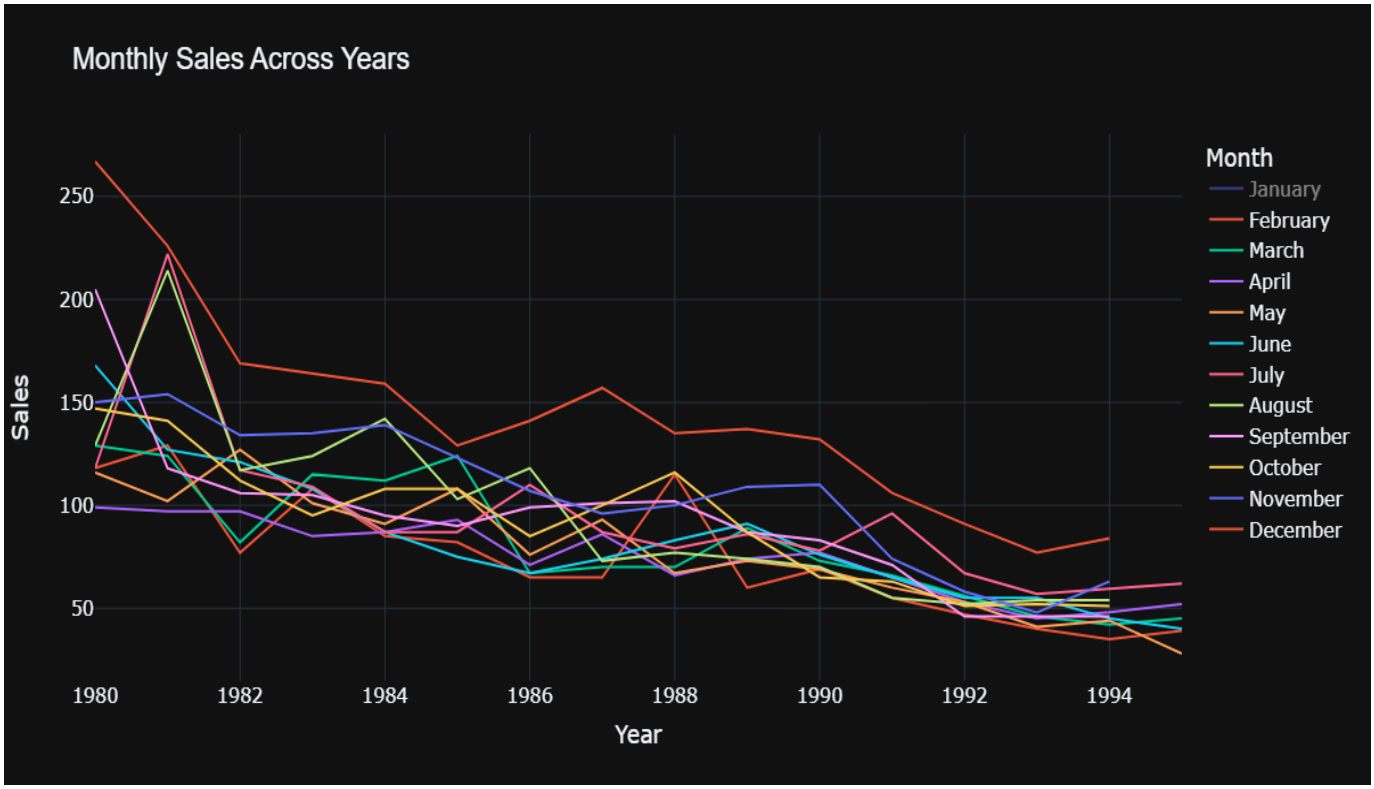


Observation:

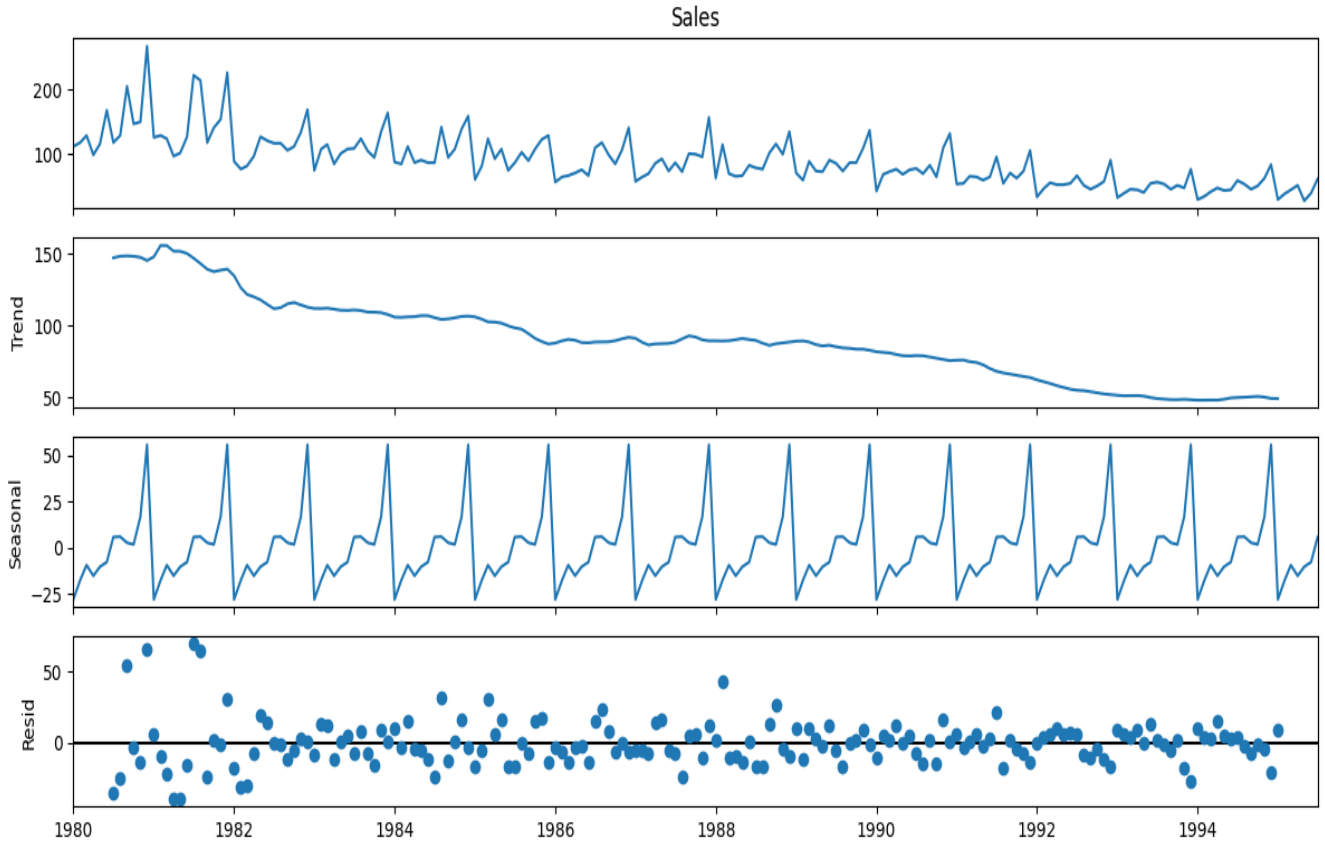
• After 1981, the sales fell drastically. Sales are typically lowest in the first quarter and highest in the fourth quarter.

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

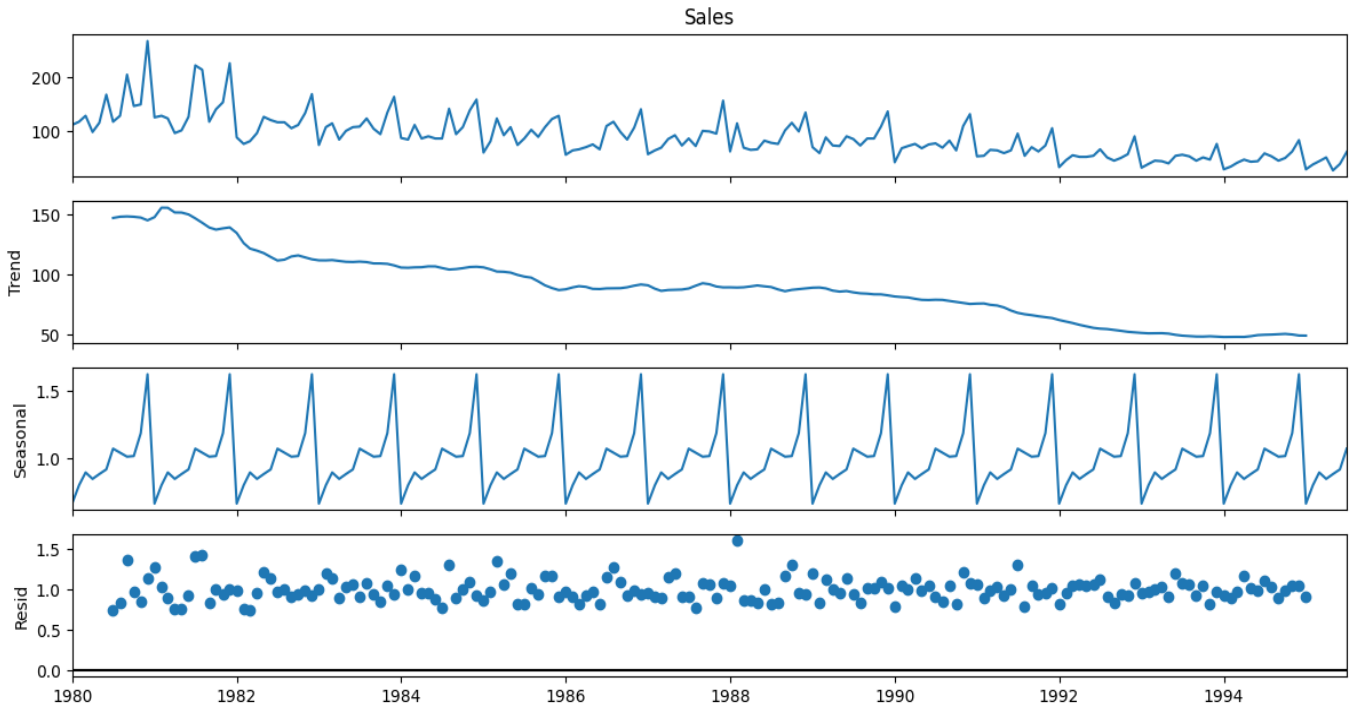
Graph of Monthly Sales across Years –



Perform Decomposition - Additive Decomposition



Perform Decomposition - Multiplicative Decomposition



Observation:

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

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

• Multiplicative Model is selected owing to a more stable residual plot and lower range of residuals.

2 - Data Pre-processing

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.

```
Shape of Training Data (Rows, Columns) is  : (132, 1)
Shape of Test Data (Rows, Columns) is : (55, 1)
```

First few rows of Training Data

YearMonth	Sales
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of Training Data

YearMonth	Sales
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

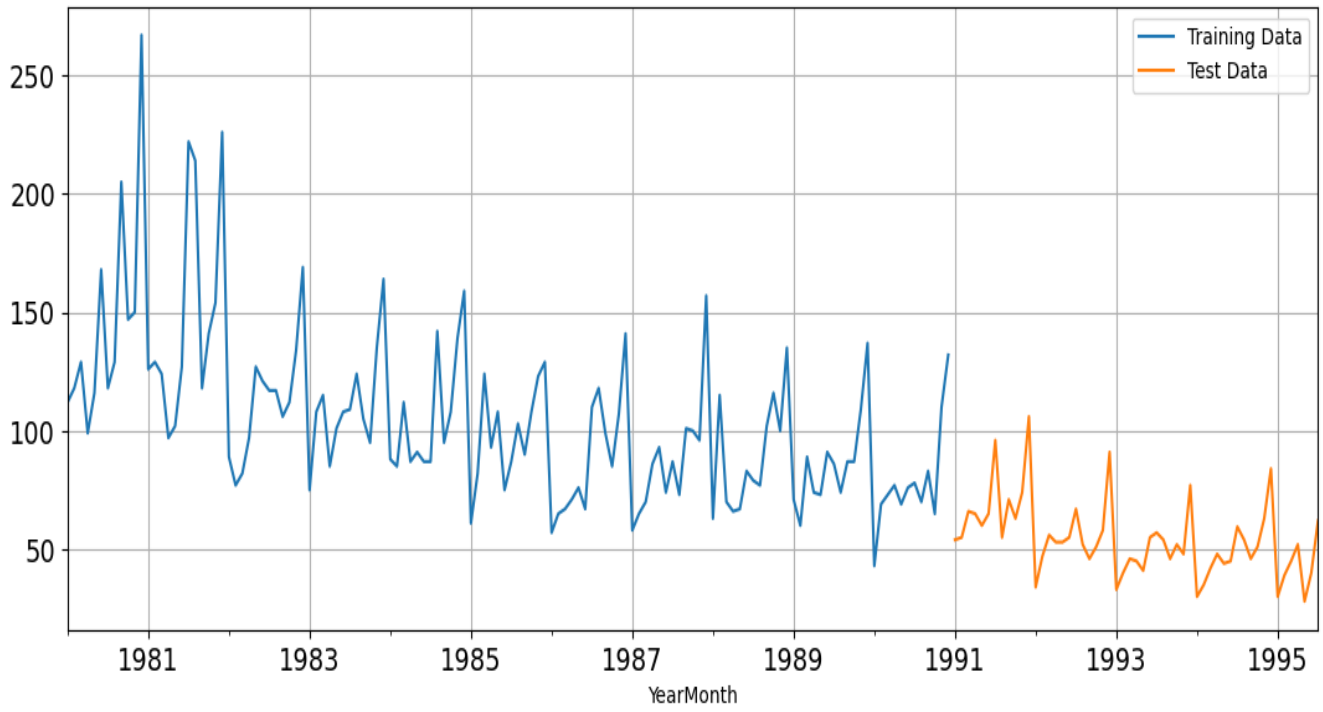
First few rows of Test Data

YearMonth	Sales
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Test Data

YearMonth	Sales
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Plot the Test and Training -



Note: It is difficult to predict the future observations if such an instance has not happened in the past. From our train-test split we are predicting likewise behaviour as compared to the past years.

3 - Model Building and Check the Performance of the Models Built

Build Forecasting Models

- Model 1: Linear Regression
- Model 2: Simple Average
- Model 3: Moving Average (MA)
- Model 4: Simple Exponential Smoothing
- Model 5: Double Exponential Smoothing (Holt's Model)
- Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

Model 1: Linear Regression

For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

Training Time instance

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132]

Test Time instance

[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187]

First few rows of Training Data

YearMonth	Sales	time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Last few rows of Training Data

YearMonth	Sales	time
1990-08-01	70.0	128
1990-09-01	83.0	129
1990-10-01	65.0	130
1990-11-01	110.0	131
1990-12-01	132.0	132

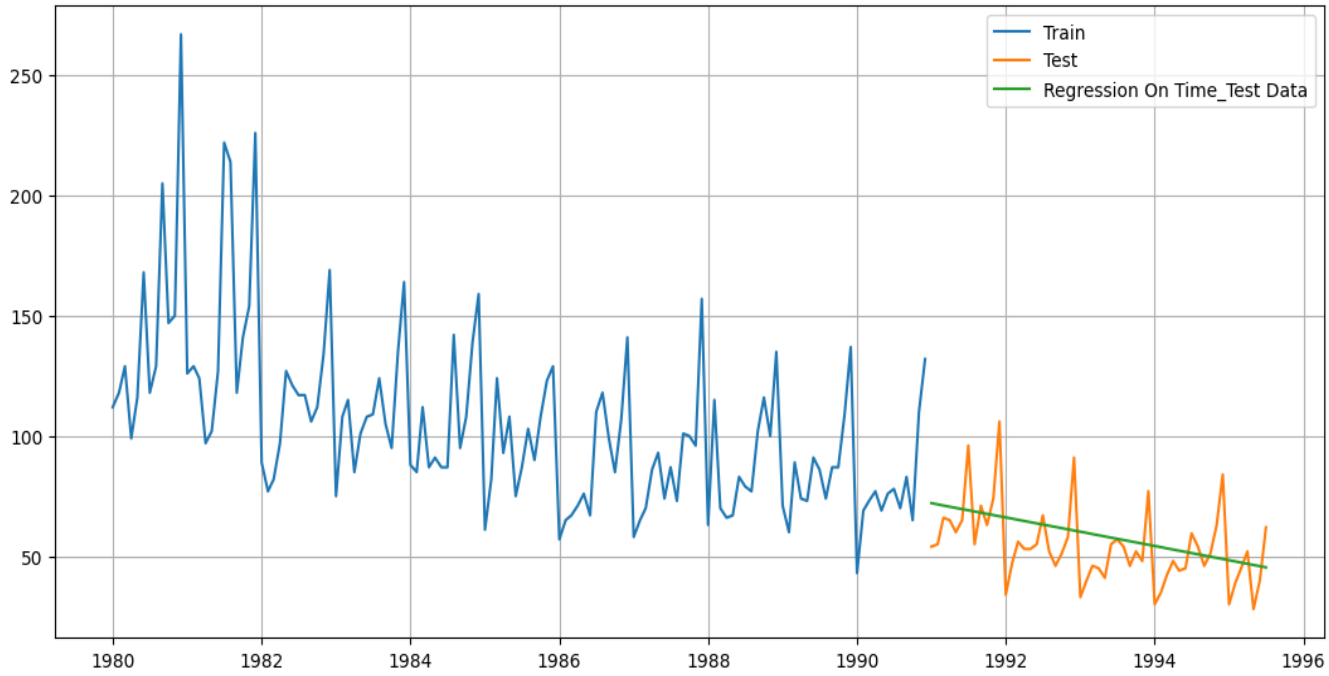
First few rows of Test Data

YearMonth	Sales	time
1991-01-01	54.0	133
1991-02-01	55.0	134
1991-03-01	66.0	135
1991-04-01	65.0	136
1991-05-01	60.0	137

Last few rows of Test Data

YearMonth	Sales	time
1995-03-01	45.0	183
1995-04-01	52.0	184
1995-05-01	28.0	185
1995-06-01	40.0	186
1995-07-01	62.0	187

Now that our training and test data has been modified, let us go ahead use **Linear Regression** to build the model on the training data and test the model on the test data.



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 15.28. The size of the seasonal

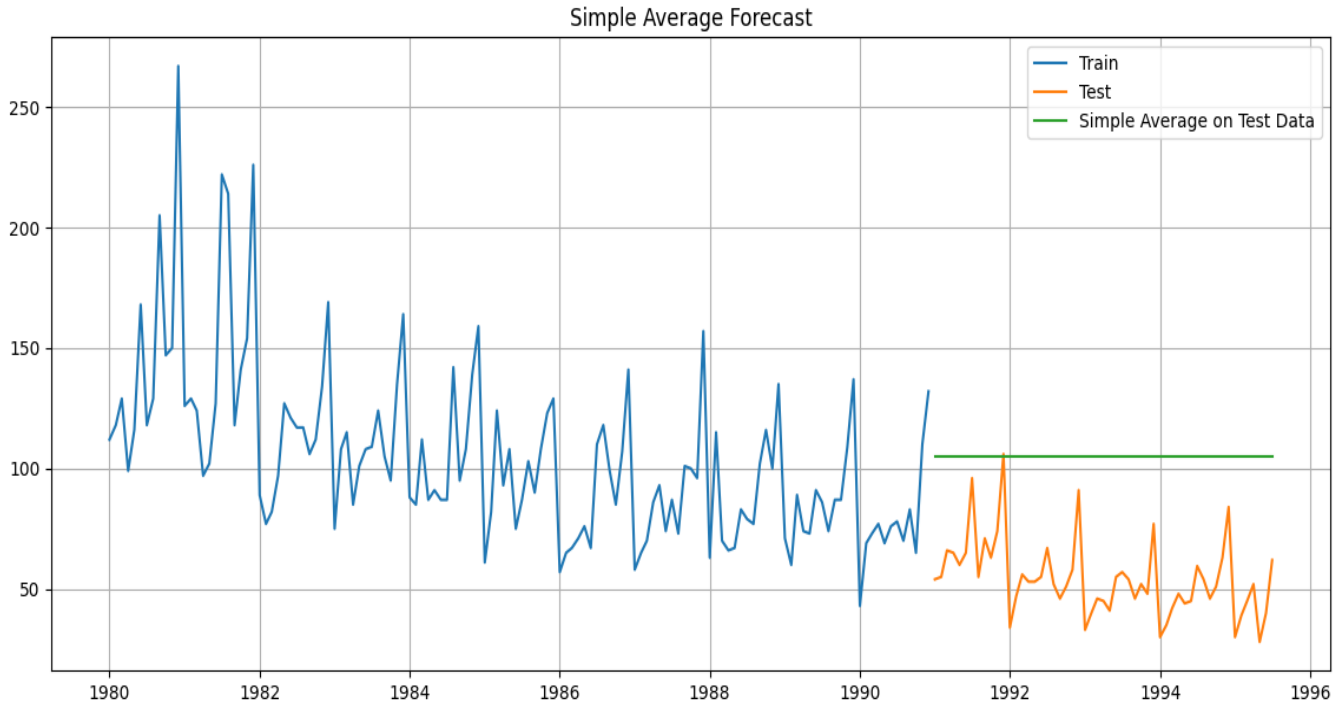
Linear Regression: Model Evaluation

Performance Metric	
Test RMSE	15.28

Model 2: Simple Average

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

YearMonth	Sales	mean_forecast
1991-01-01	54.0	104.939394
1991-02-01	55.0	104.939394
1991-03-01	66.0	104.939394
1991-04-01	65.0	104.939394
1991-05-01	60.0	104.939394



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 **53.0497** which is significantly higher than the regression model but lower than naïve forecast model.

Simple Average: Model Evaluation

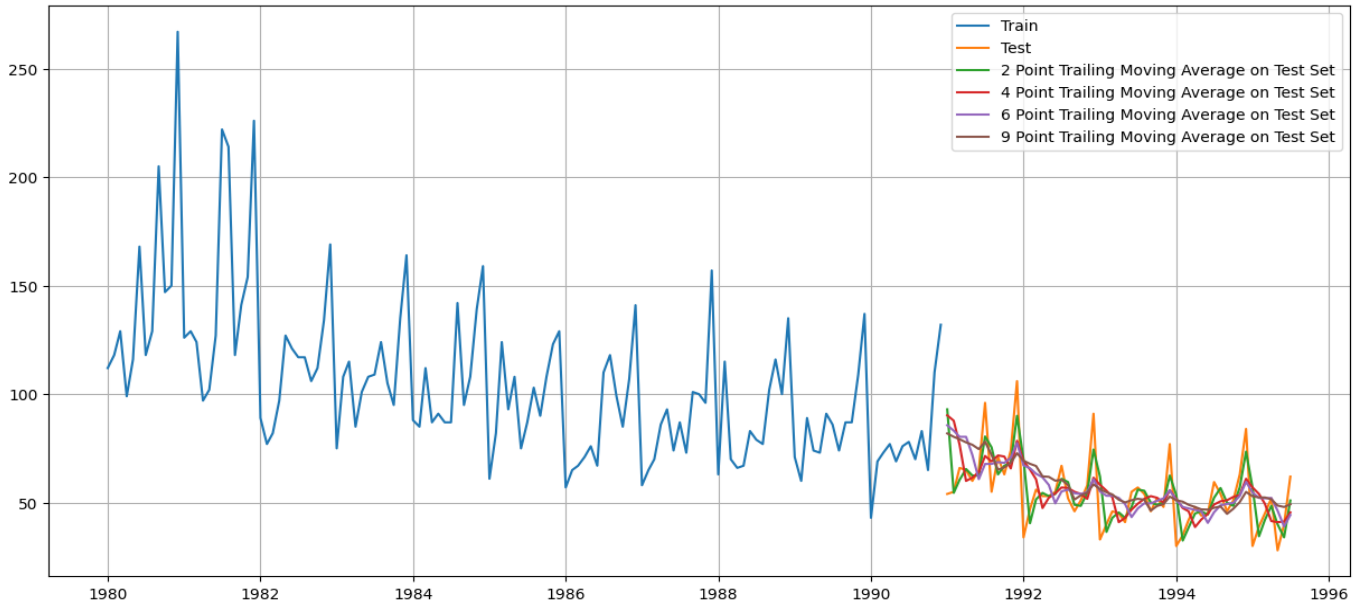
Performance Metric	
Test RMSE	53.0497

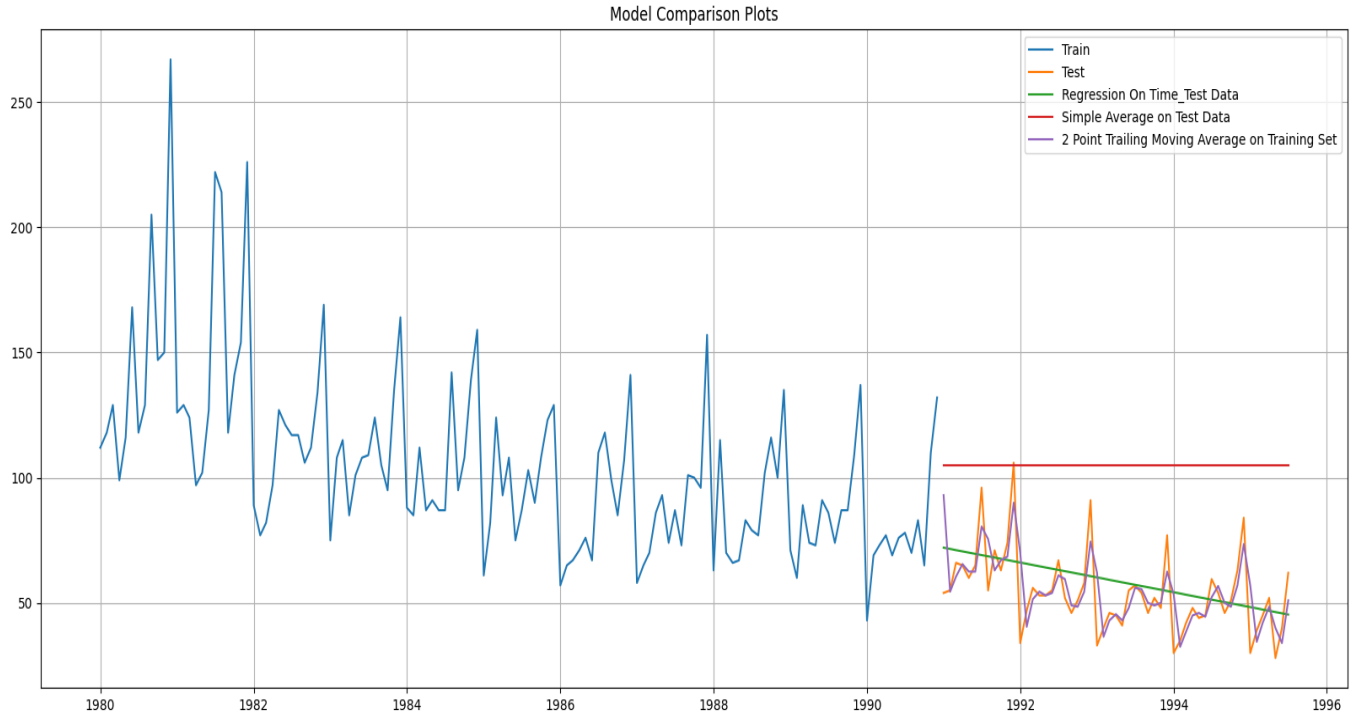
Model 3: Moving Average (MA)

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.

YearMonth	Sales
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

YearMonth	Sales	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.50	NaN	NaN
1980-05-01	116.0	107.5	115.50	NaN	NaN
1980-06-01	168.0	142.0	128.00	123.666667	NaN
1980-07-01	118.0	143.0	125.25	124.666667	NaN
1980-08-01	129.0	123.5	132.75	126.500000	NaN
1980-09-01	205.0	167.0	155.00	139.166667	132.666667
1980-10-01	147.0	176.0	149.75	147.166667	136.555556





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.
- The root means squared error (RMSE) for the 2-point trailing average model is 11.589, which is lowest than all models build so far.

Moving Average: Model Evaluation

Model	Test RMSE
2 Point Trailing Moving Average	11.589082
4 Point Trailing Moving Average	14.506190
6 Point Trailing Moving Average	14.558008
9 Point Trailing Moving Average	14.797139

Model 4: 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.

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

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.

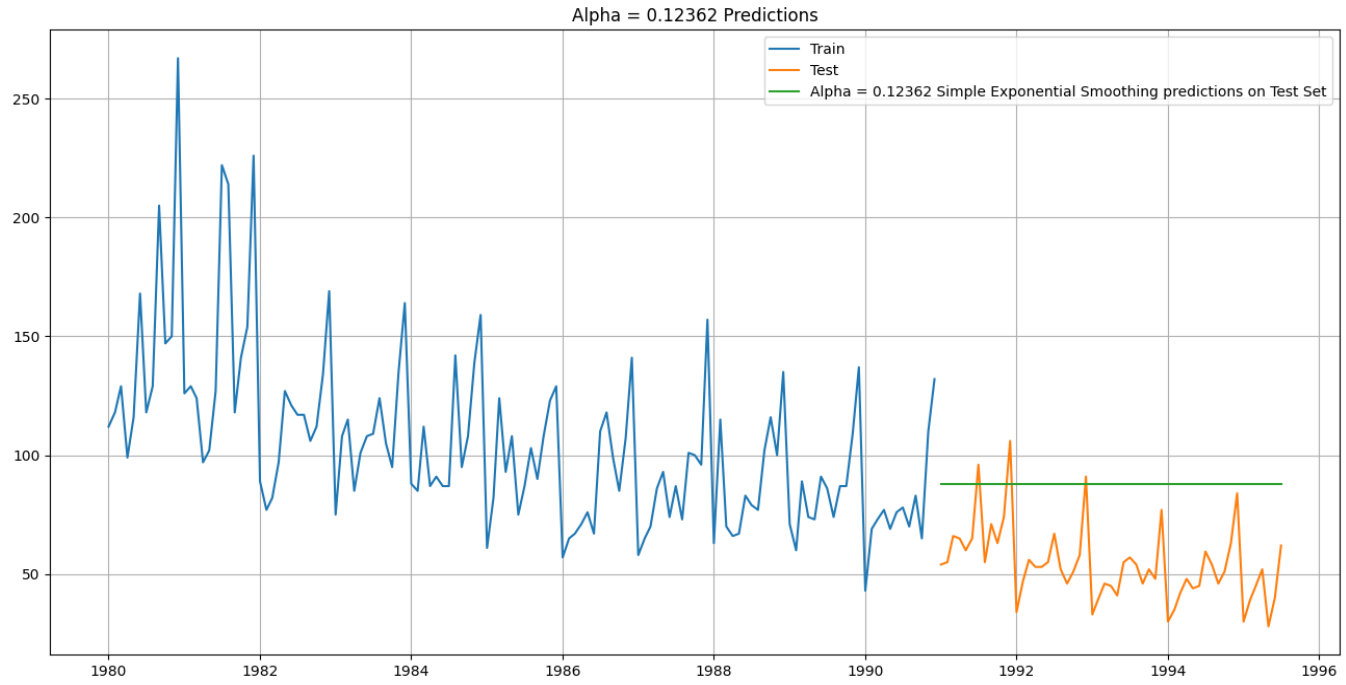
For the selection criteria, the below Simple Exponential Smoothing is built by using optimized parameters.

Simple Exponential Smoothing Model -

```
{'smoothing_level': 0.12362013466760018,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Sample of SES predictions –

YearMonth	Sales	predict
1991-01-01	54.0	87.983765
1991-02-01	55.0	87.983765
1991-03-01	66.0	87.983765
1991-04-01	65.0	87.983765
1991-05-01	60.0	87.983765



For Alpha = 0.12362 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 37.193

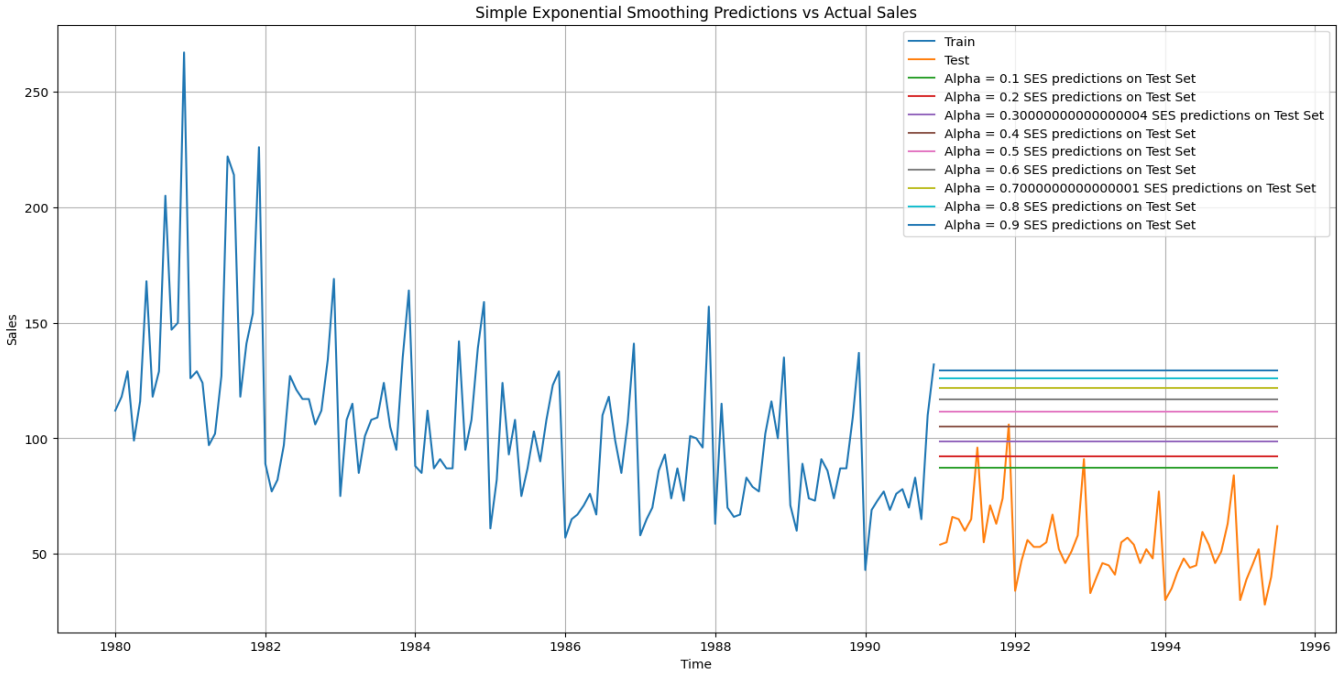
Simple Exponential Smoothing: Model Evaluation

Model	Test RMSE
Simple Exponential Smoothing	37.193

We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

Model Evaluation

	Alpha Values	Train RMSE	Test RMSE
0	0.1	31.815610	36.429535
1	0.2	31.979391	40.957988
2	0.3	32.470164	47.096522
3	0.4	33.035130	53.356493
4	0.5	33.682839	59.229384
5	0.6	34.441171	64.558022
6	0.7	35.323261	69.284383
7	0.8	36.334596	73.359904
8	0.9	37.482782	76.725002



	Test RMSE
RegressionOnTime	15.278158
Simple Average Model	53.049755
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
Alpha = 0.12362, SimpleExponentialSmoothing	37.192623
Alpha=0.1,SimpleExponentialSmoothing	36.429535

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=0.12362 is 37.192 and for Alpha=0.1, RMSE is 36.429.

Simple Exponential Smoothing: Model Evaluation

Model	Test RMSE
SES (Alpha = 0.12362)	37.192
SES (Alpha = 0.1)	36.429

Method 5: 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,

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

The forecast at time t + 1 is given by

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

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

For the selection criteria, the below Double Exponential Smoothing is built by using optimized parameters.

Double Exponential Smoothing Model

```
{'smoothing_level': 0.16213321015010723,
 'smoothing_trend': 0.13152155372234675,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 'initial_trend': 6.0,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Double Exponential Smoothing Model Predictions

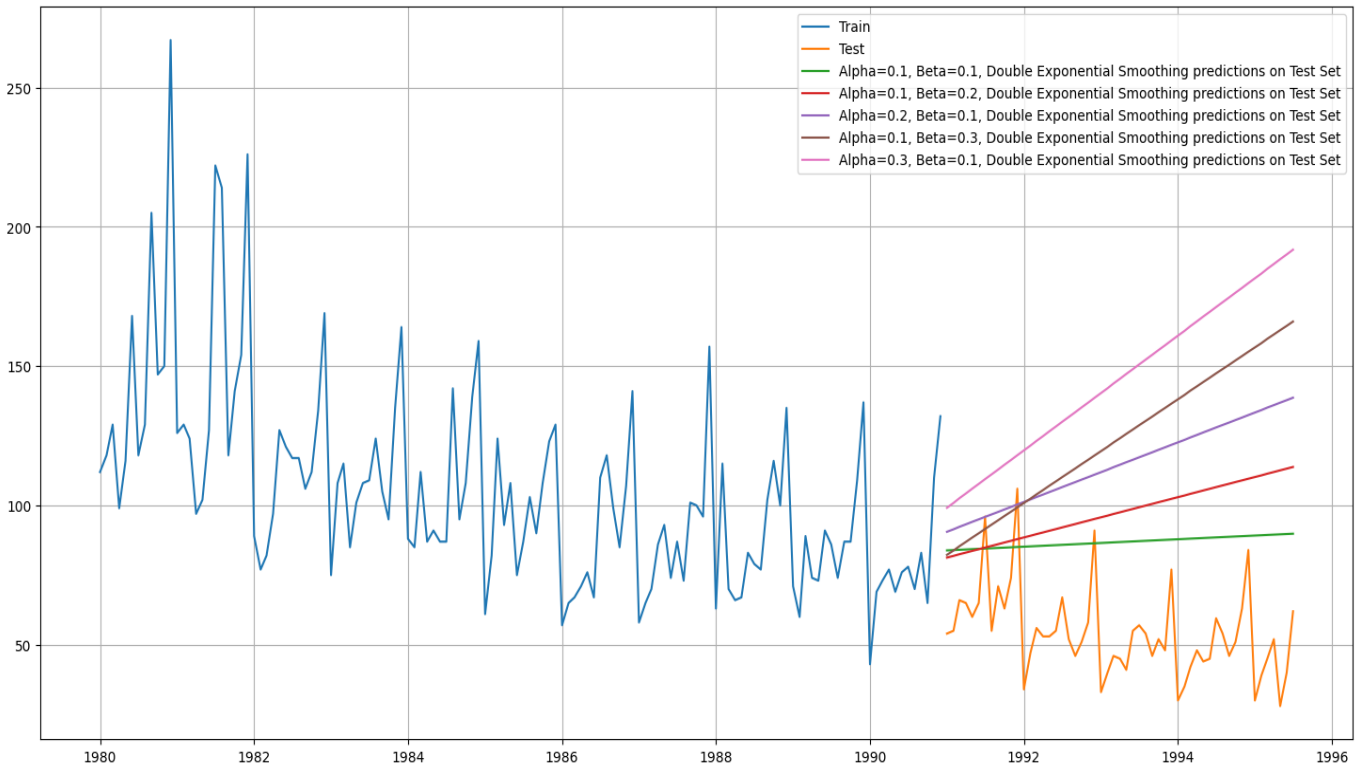
YearMonth	Sales	predict
1991-01-01	54.0	87.249993
1991-02-01	55.0	88.152722
1991-03-01	66.0	89.055451
1991-04-01	65.0	89.958180
1991-05-01	60.0	90.860910

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.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.510010
1	0.1	0.2	33.450729	48.221436
2	0.1	0.3	33.145789	77.649847
3	0.1	0.4	33.262191	99.064536
4	0.1	0.5	33.688415	123.742433
...	...	...	...	...
95	1.0	0.6	51.831610	801.137173
96	1.0	0.7	54.497039	841.349112
97	1.0	0.8	57.365879	853.421959
98	1.0	0.9	60.474309	834.167545
99	1.0	1.0	63.873454	779.536777

100 rows × 4 columns

Plot of DES predictions on Test data



DES List of Sorted RMSE

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.510010
1	0.1	0.2	33.450729	48.221436
10	0.2	0.1	33.097427	65.251675
2	0.1	0.3	33.145789	77.649847
20	0.3	0.1	33.611269	98.152852

Observation:

• The root means squared error (RMSE) for the double exponential smoothing model with Alpha=0.162, Beta= 0.1315 is 36.51 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.

Double Exponential Smoothing: Model Evaluation

Model	Test RMSE
DES (Alpha=0.162, Beta=0.135)	36.51

Method 5: 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, $0 < \alpha < 1$, $0 < \beta < 1$, $0 < \gamma < 1$.

For the selection criteria, the below Triple Exponential Smoothing is built by using optimized parameters

For Alpha =0.08621, Beta = 1.3722, Gamma = 0.4763 Tripple Exponential Smoothing Model (Trend = Additive, Seasonality = Additive) forecast on the Test Data, RMSE is 14.128

For Alpha =0.1542, Beta = 5.3100, Gamma = 0.3713 Tripple Exponential Smoothing Model (Trend = Additive, Seasonality = Multiplicative) forecast on the Test Data, RMSE is 18.683

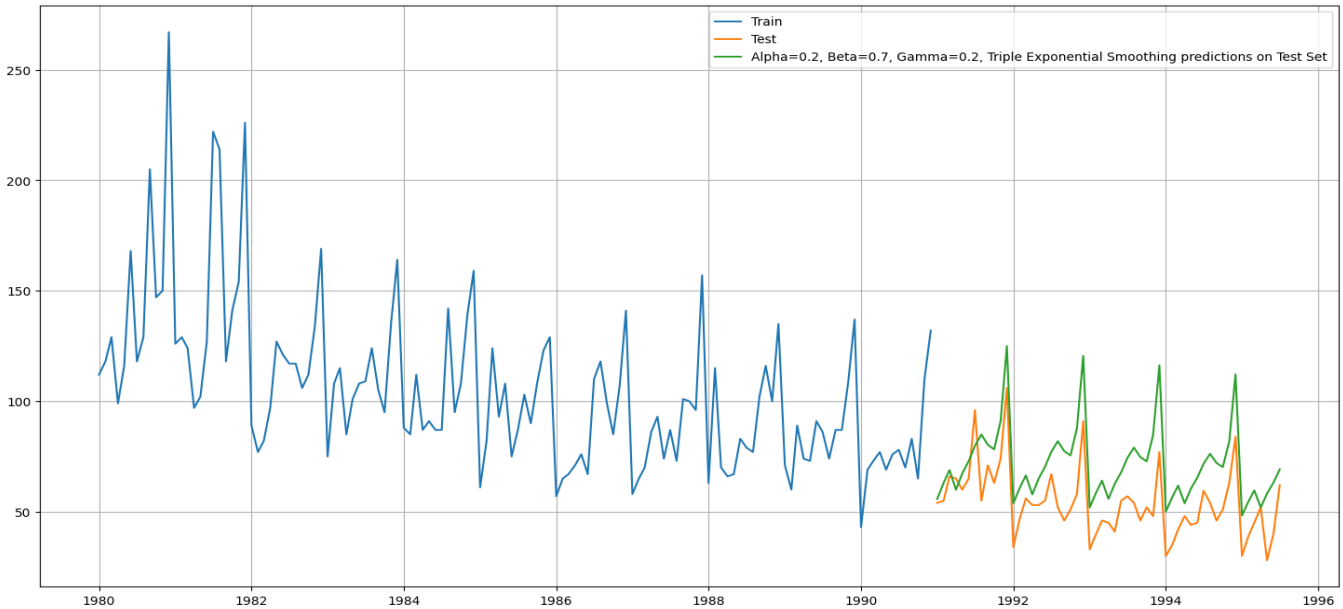
For Alpha =0.1531, Beta = 1.3401, Gamma = 0.3691 Tripple Exponential Smoothing Model (Trend = Multiplicative, Seasonality = Multiplicative) forecast on the Test Data, RMSE is 19.880

For Alpha =0.0831, Beta = 7.8624, Gamma = 0.4910 Tripple Exponential Smoothing Model (Trend = Multiplicative, Seasonality = Additive) forecast on the Test Data, RMSE is 16.254

Check the performance of the models built and Sort in Ascending Order

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
2136	0.2	0.7	0.2	24.042290	8.992350	tm_sm
1010	0.1	0.2	0.1	19.770392	9.221020	ta_sm
1011	0.1	0.2	0.2	20.253487	9.543696	ta_sm
1151	0.2	0.6	0.2	23.129850	9.922552	ta_sm
1012	0.1	0.2	0.3	20.871304	9.952909	ta_sm

Plot of TES forecast



Observation: We got the lowest RMSE when we take Trend – Multiplicative and Seasonality – Multiplicative if Alpha – 0.2, Beta – 0.7 and Gamma – 0.2.

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

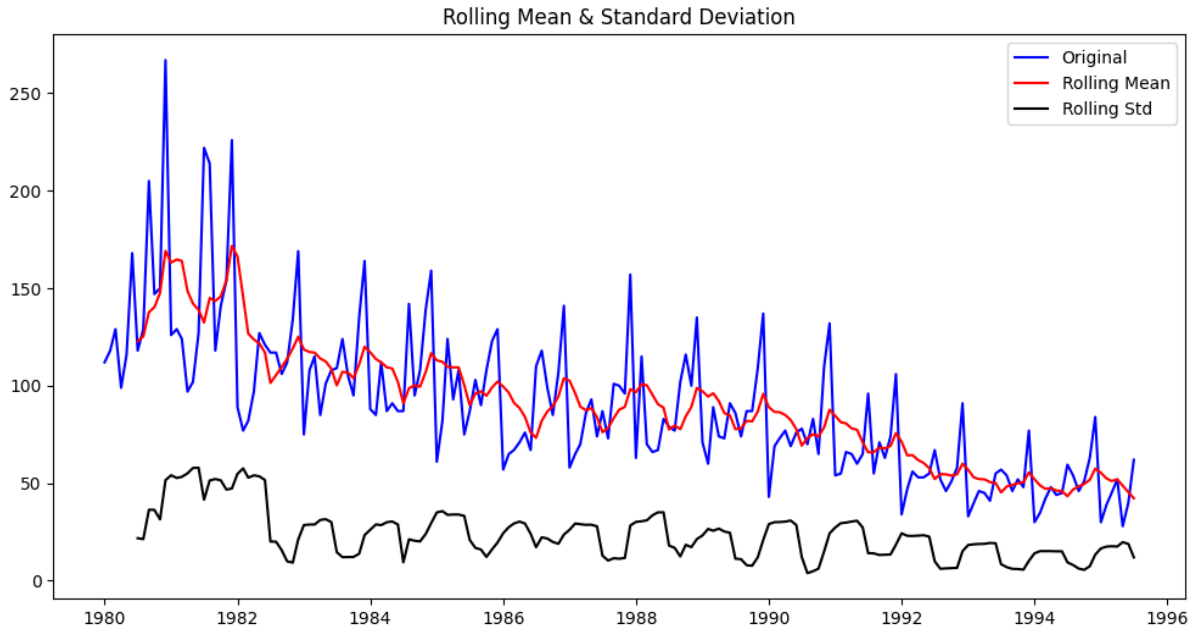
4 - Check for Stationarity

Check for stationarity of the whole Time Series data. The Augmented Dickey-Fuller test is a 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.

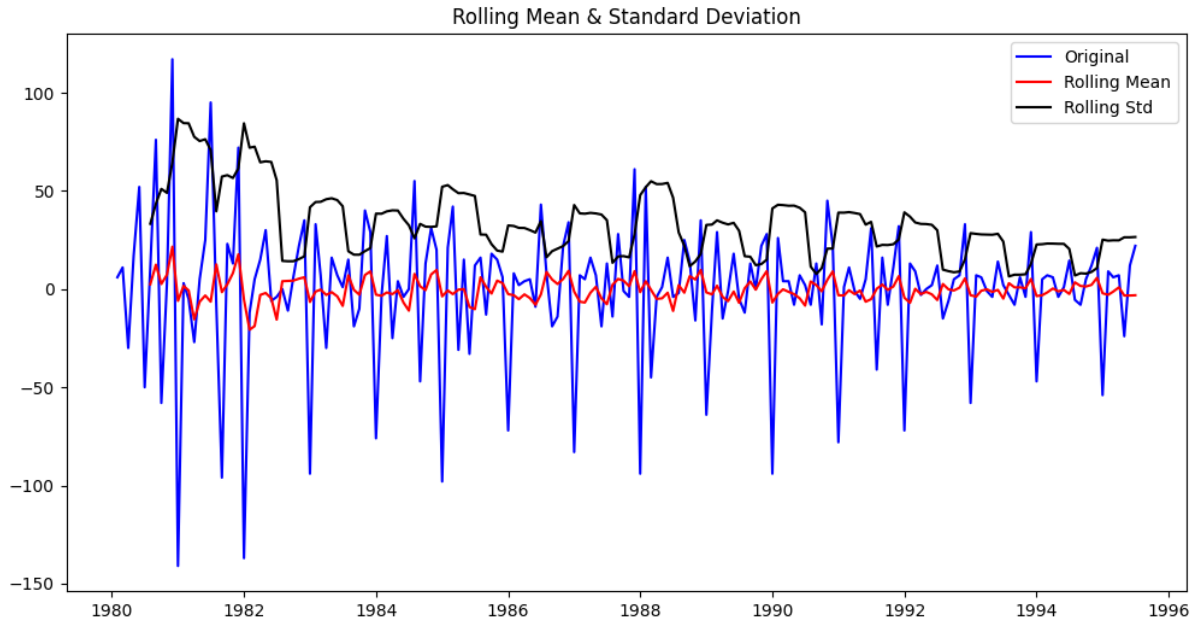
H1 : The Time Series does not have a unit root and is thus stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.892338
p-value                          0.335674
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

We see that at 5% significant level the Time Series is non-stationary.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

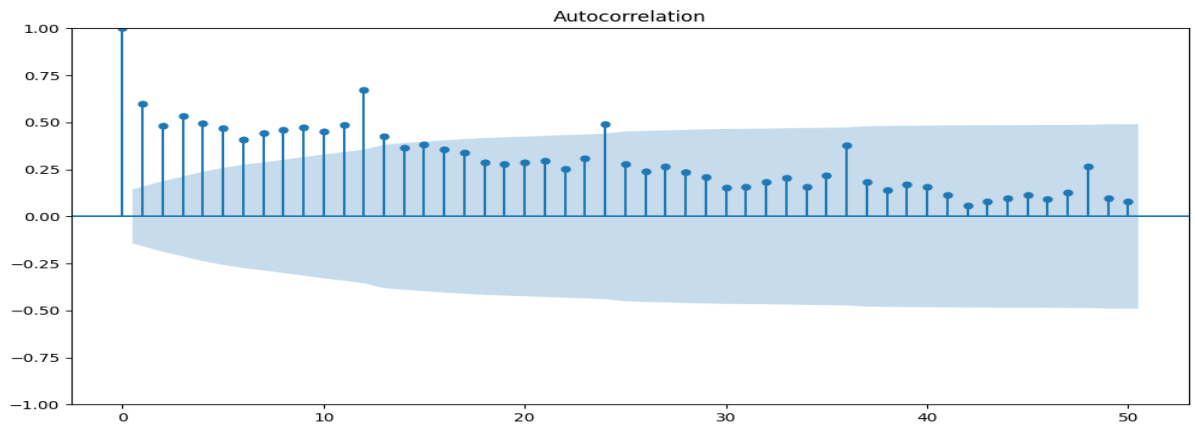


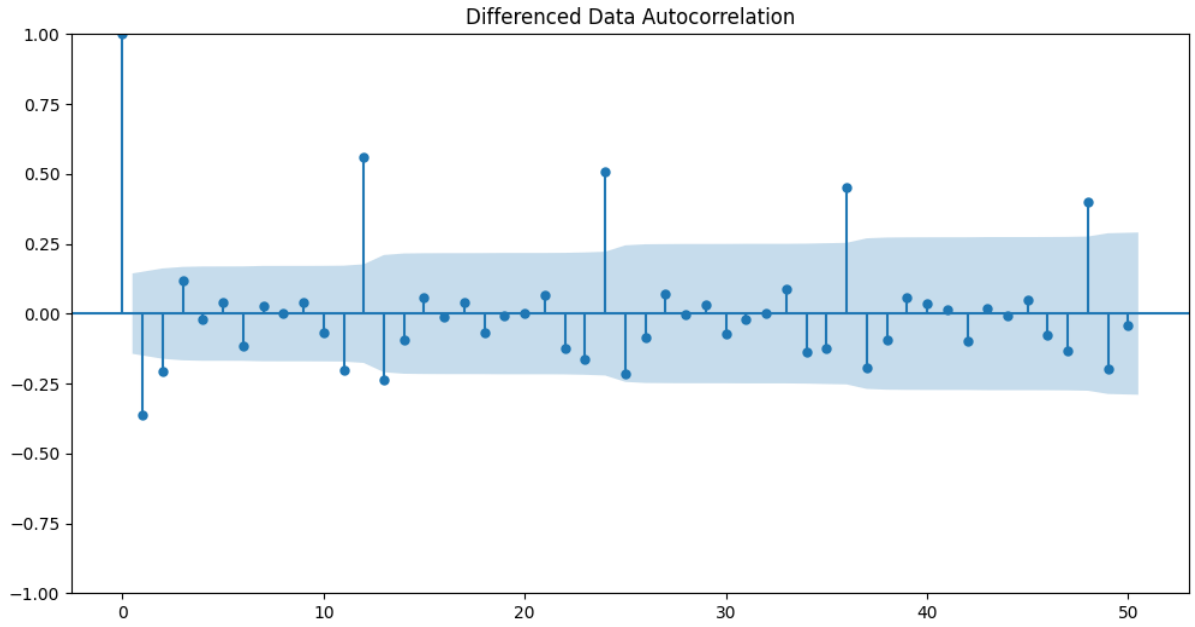

```
Results of Dickey-Fuller Test:
Test Statistic                -8.032729e+00
p-value                        1.938803e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

We observe that the p-value (α) is less than 0.05. Therefore, we reject the null hypothesis. This result suggests that the time series is stationary, meaning its statistical properties such as mean, variance, and autocorrelation remain constant over time.

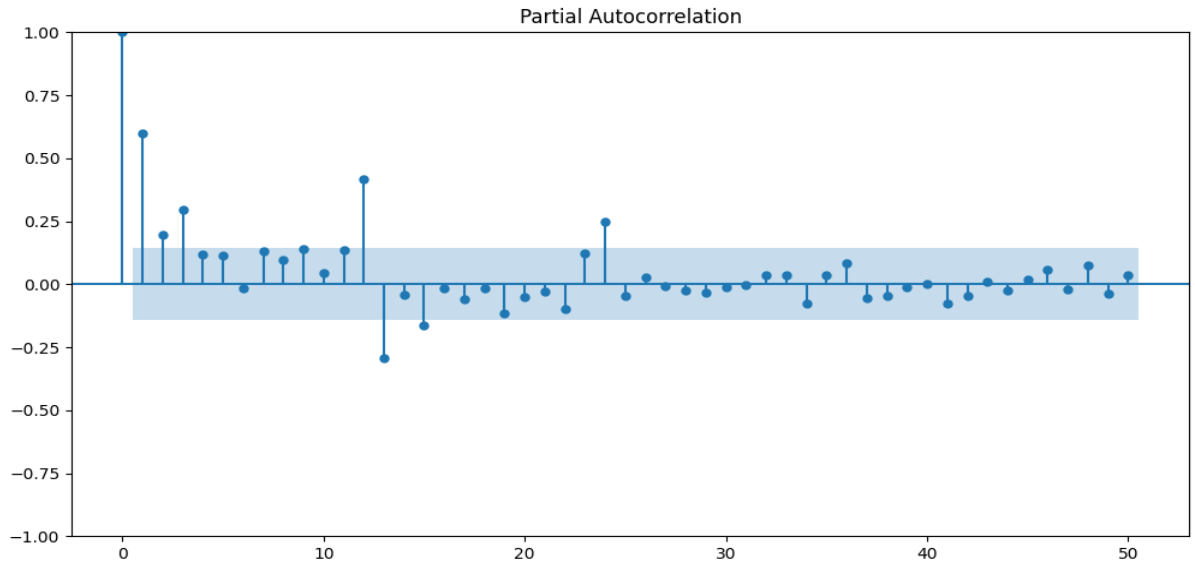
5 - Model Building - Stationary Data

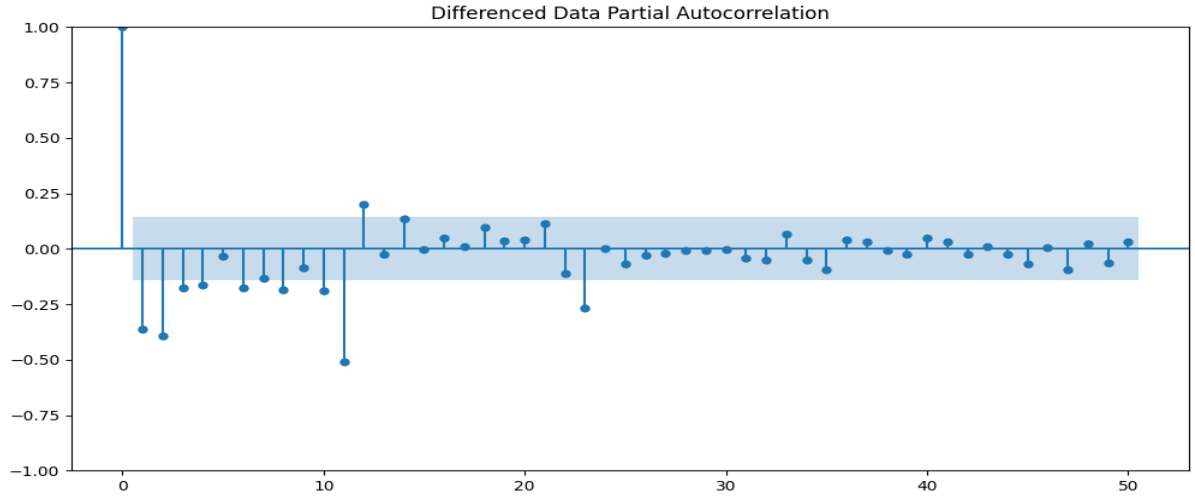
ACF and PACF Plots




Differenced Data Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50

PACF Plot -


Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50


Differenced Data Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50

Split the data into train and test and plot the training and test data.

Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

First few rows of Training Data

YearMonth	Sales
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

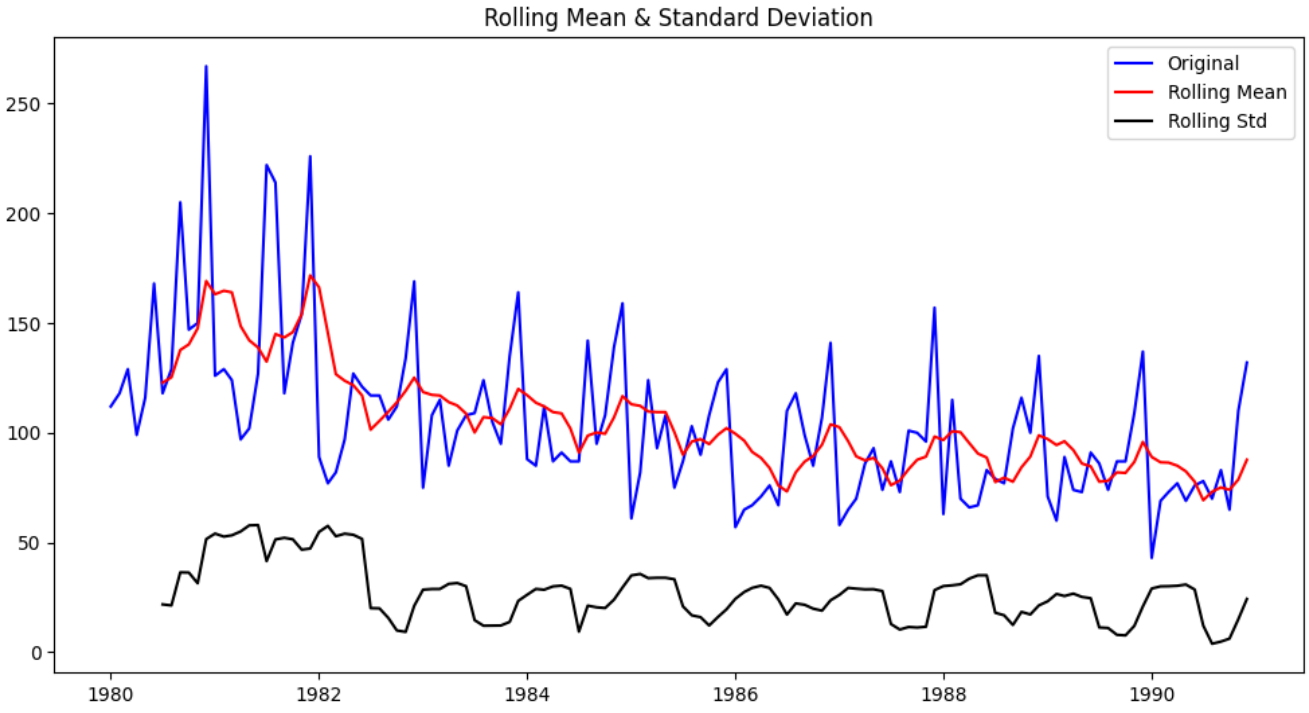
Last few rows of Training Data

First few rows of Test Data

YearMonth	Sales
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Test Data

Check the stationarity of the training data

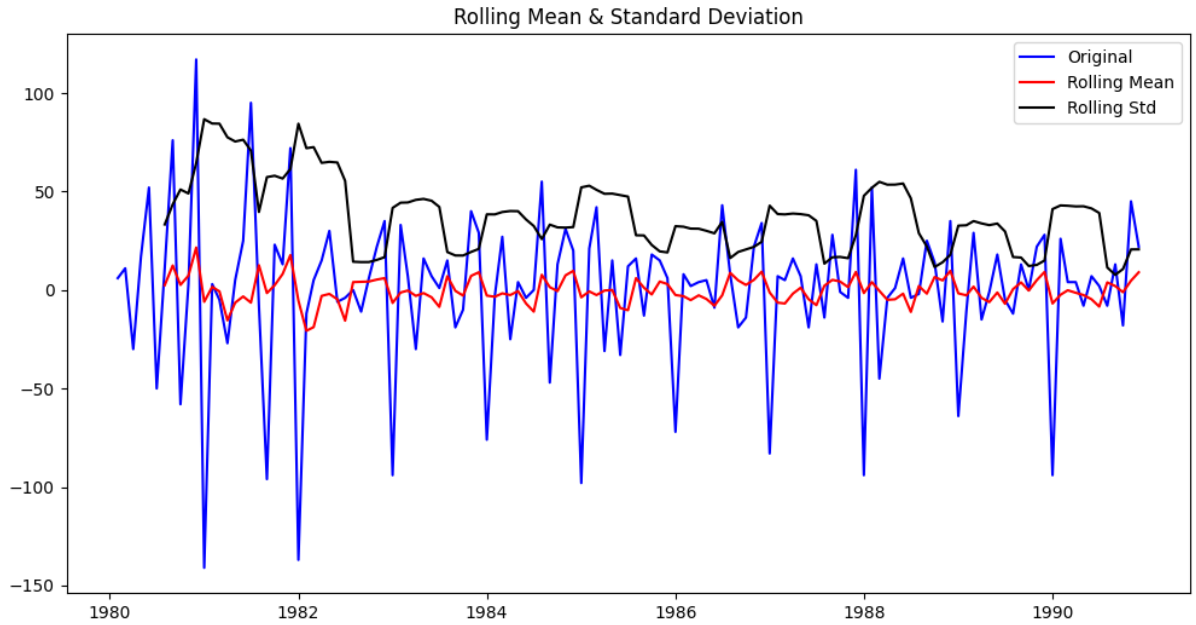



```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

We observe that the p-value (α) is greater than 0.05. Therefore, we fail to reject the null hypothesis. This result suggests that the time series is not stationary, meaning its statistical properties, such as mean, variance, and autocorrelation, change over time.

Check the stationarity of the differenced training data

Results of Dickey-Fuller Test:
Test Statistic -6.592372e+00
p-value 7.061944e-09
#Lags Used 1.200000e+01
Number of Observations Used 1.180000e+02
Critical Value (1%) -3.487022e+00
Critical Value (5%) -2.886363e+00
Critical Value (10%) -2.580009e+00
dtype: float64



We see that after taking a difference of order 1 the series have become stationary at α = 0.05.

Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

Build the Auto ARIMA –

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Creating an empty Data Frame with column names only

param	AIC

Sort the AIC values to find the best parameters

	param	AIC
11	(2, 1, 3)	1274.695127
15	(3, 1, 3)	1278.658004
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.574230
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832

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.

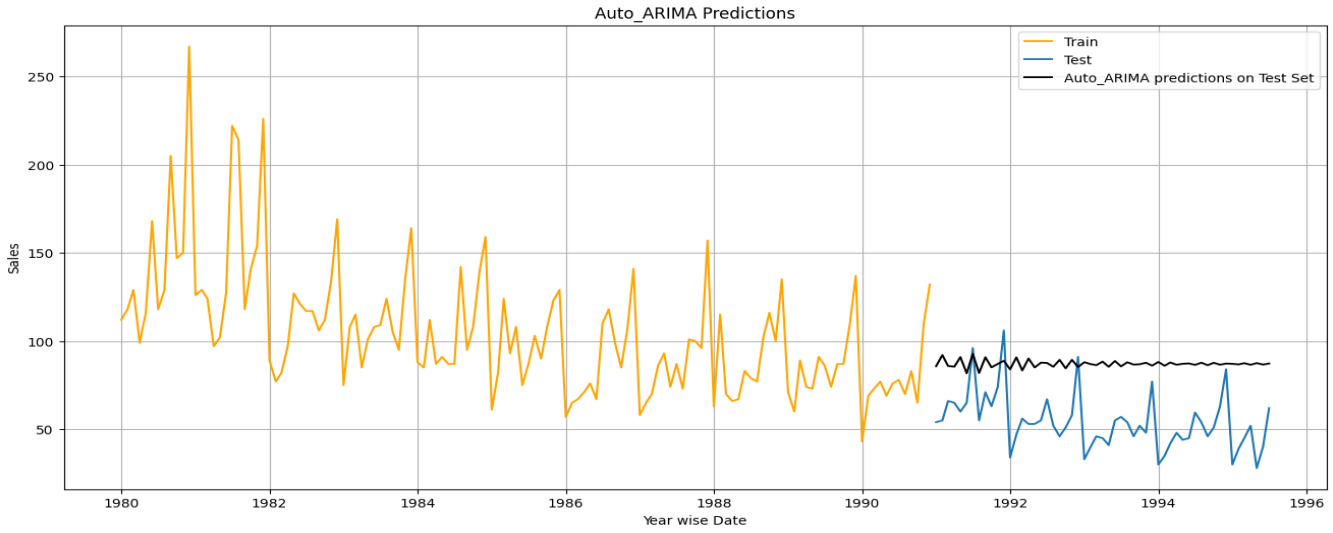
SARIMAX Results

Dep. Variable:	Sales	No. Observations:	132
Model:	ARIMA(3, 1, 3)	Log Likelihood	-632.329
Date:	Sun, 18 Aug 2024	AIC	1278.658
Time:	14:50:12	BIC	1298.784
Sample:	01-01-1980 - 12-01-1990	HQIC	1286.836
Covariance Type:	opg		

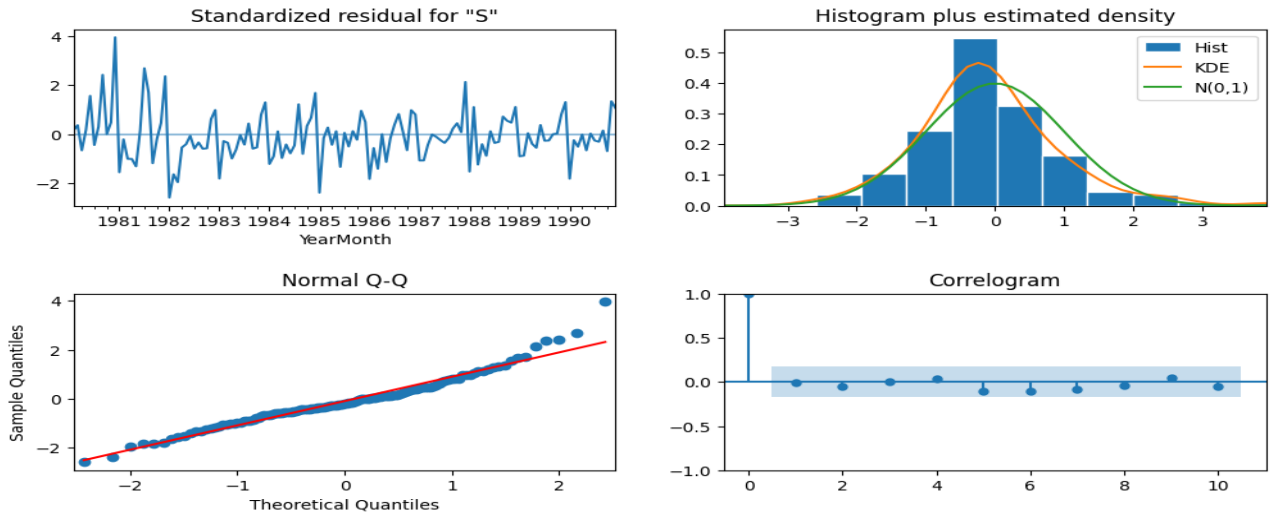
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.5838	0.088	-17.939	0.000	-1.757	-1.411
ar.L2	-0.6387	0.142	-4.497	0.000	-0.917	-0.360
ar.L3	0.1336	0.090	1.492	0.136	-0.042	0.309
ma.L1	0.9452	0.161	5.861	0.000	0.629	1.261
ma.L2	-0.7147	0.106	-6.765	0.000	-0.922	-0.508
ma.L3	-0.9103	0.156	-5.844	0.000	-1.216	-0.605
sigma2	882.9435	139.008	6.352	0.000	610.494	1155.393

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	31.69
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.72
Prob(H) (two-sided):	0.00	Kurtosis:	4.94

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



Diagnostics plot



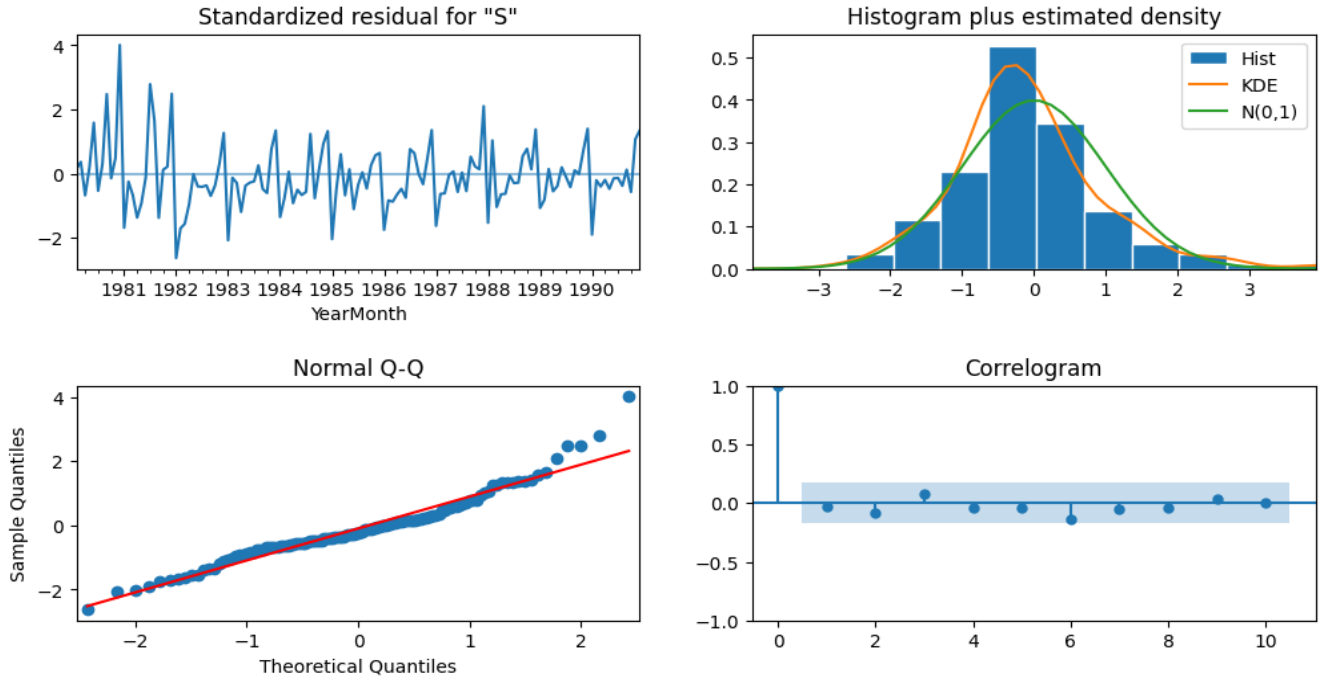
Automated ARIMA: Model Evaluation - For evaluating the model's performance metrics, we look at root means squared error (RMSE)

Model	Test RMSE
ARIMA (p=3, d=1, q=3)	**36.30**

Build the Manual ARIMA -

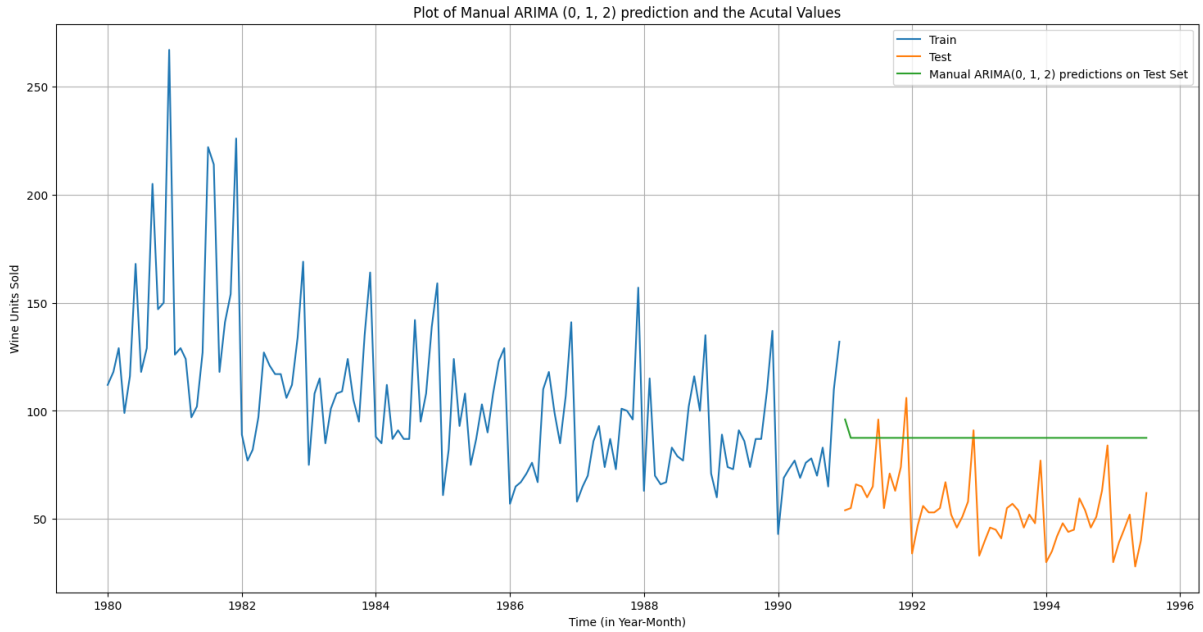
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -636.836
Date:                Sun, 18 Aug 2024   AIC                           1279.672
Time:                        14:50:14   BIC                           1288.297
Sample:                    01-01-1980   HQIC                          1283.176
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.6970      0.072     -9.689      0.000      -0.838      -0.556
ma.L2         -0.2042      0.073     -2.794      0.005      -0.347      -0.061
sigma2       965.8407     88.305     10.938      0.000     792.766    1138.915
===================================================================================
Ljung-Box (L1) (Q):                   0.14   Jarque-Bera (JB):                39.24
Prob(Q):                              0.71   Prob(JB):                         0.00
Heteroskedasticity (H):               0.36   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.13
===================================================================================

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

Create the diagnostic Plot for Manual ARIMA



Manual ARIMA: Model Evaluation

For evaluating the model's performance metrics, we look at root means squared error (RMSE)



Model	Test RMSE
ARIMA (p=0, d=1, q=2)	**36.90**

Build the Auto 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 like regular differencing.

A SARIMA model is characterized by 7 terms: p, d, q, P, Q, D and F

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

```
Examples of some parameter combinations for the Model...
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)
```

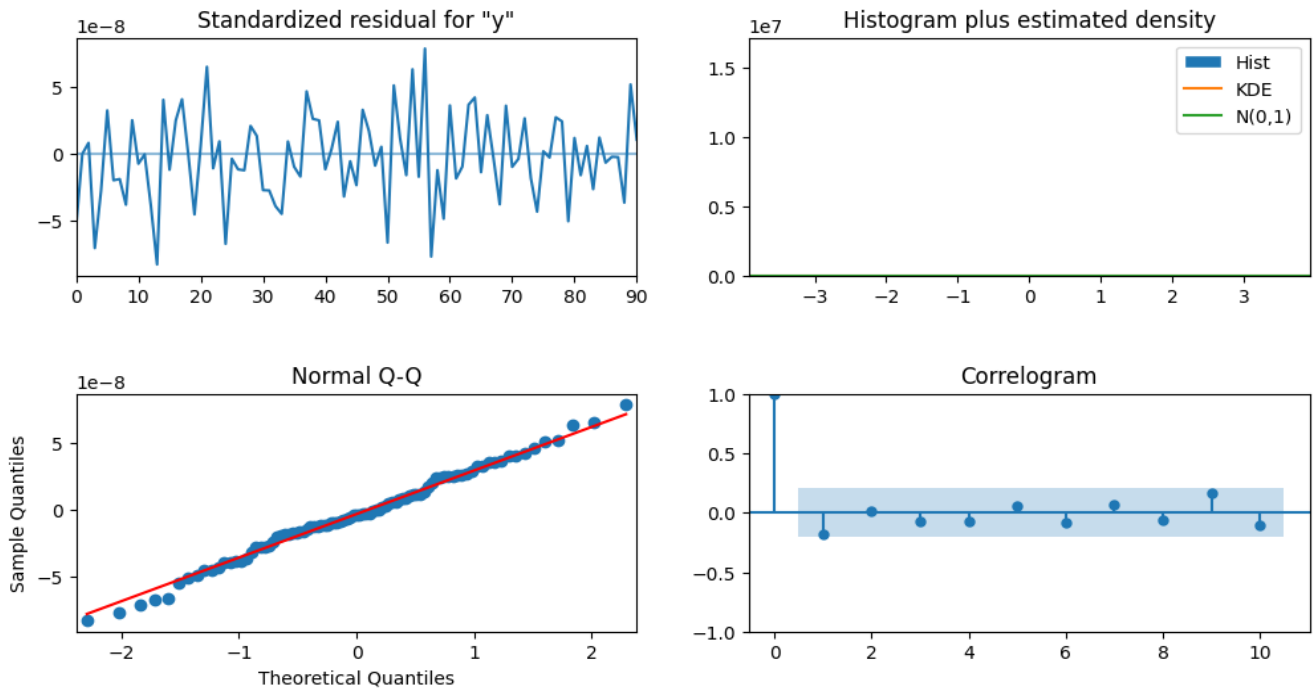
Sort by AIC to find the best parameters

	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400287
238	(3, 1, 2)	(3, 0, 2, 12)	774.894960
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.495330
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018

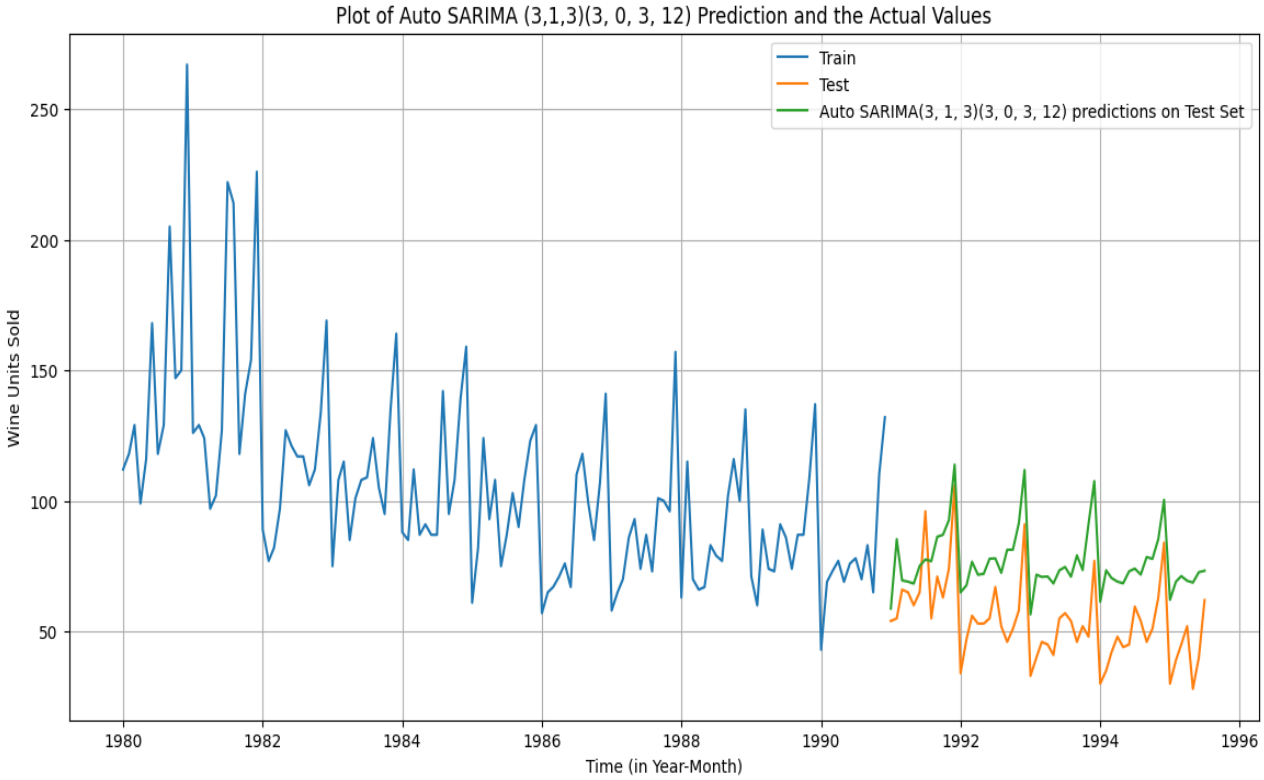
We can see that among all the possible given combinations, the AIC is lowest for the combination (3,1,1) (3,0,2,12). Hence, the model is built with these parameters to determine the RMSE value of test data.

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 0, 3, 12)   Log Likelihood               -1911.627
Date:                            Sun, 18 Aug 2024   AIC                           3849.255
Time:                                    14:57:34   BIC                           3881.896
Sample:                                         0   HQIC                          3862.423
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7770         -0        inf      0.000      -0.777      -0.777
ar.L2         -0.1311         -0        inf      0.000      -0.131      -0.131
ar.L3          0.0122   4.39e-34   2.79e+31      0.000       0.012       0.012
ma.L1          0.1404   3.25e-32   4.32e+30      0.000       0.140       0.140
ma.L2         -0.5865   1.11e-31  -5.31e+30      0.000      -0.586      -0.586
ma.L3         -0.1778   4.44e-32     -4e+30      0.000      -0.178      -0.178
ar.S.L12       0.1794   1.42e-35   1.26e+34      0.000       0.179       0.179
ar.S.L24       0.2567    2.6e-34   9.88e+32      0.000       0.257       0.257
ar.S.L36       0.3231   1.18e-33   2.75e+32      0.000       0.323       0.323
ma.S.L12    1.317e+14   2.91e-33   4.53e+46      0.000    1.32e+14    1.32e+14
ma.S.L24    1.063e+12   1.68e-46   6.34e+57      0.000    1.06e+12    1.06e+12
ma.S.L36   -3.112e+13   1.18e-45  -2.65e+58      0.000   -3.11e+13   -3.11e+13
sigma2      1.619e-11      2e-10      0.081      0.936   -3.76e-10    4.09e-10
==========================================================================================
Ljung-Box (L1) (Q):                   2.92   Jarque-Bera (JB):                 0.11
```

Automated SARIMA – Diagnostics plot



Plot of Auto SARIMA



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) (3,0,3,12) as we see from the above results.

Automated SARIMA: Model Evaluation

Model	Test RMSE
SARIMA (p=3, d=1, q=1) (P=3, D=0, Q=3, F=12)	24.61

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=3, F=12) is 24.61.

Manual SARIMA Model –

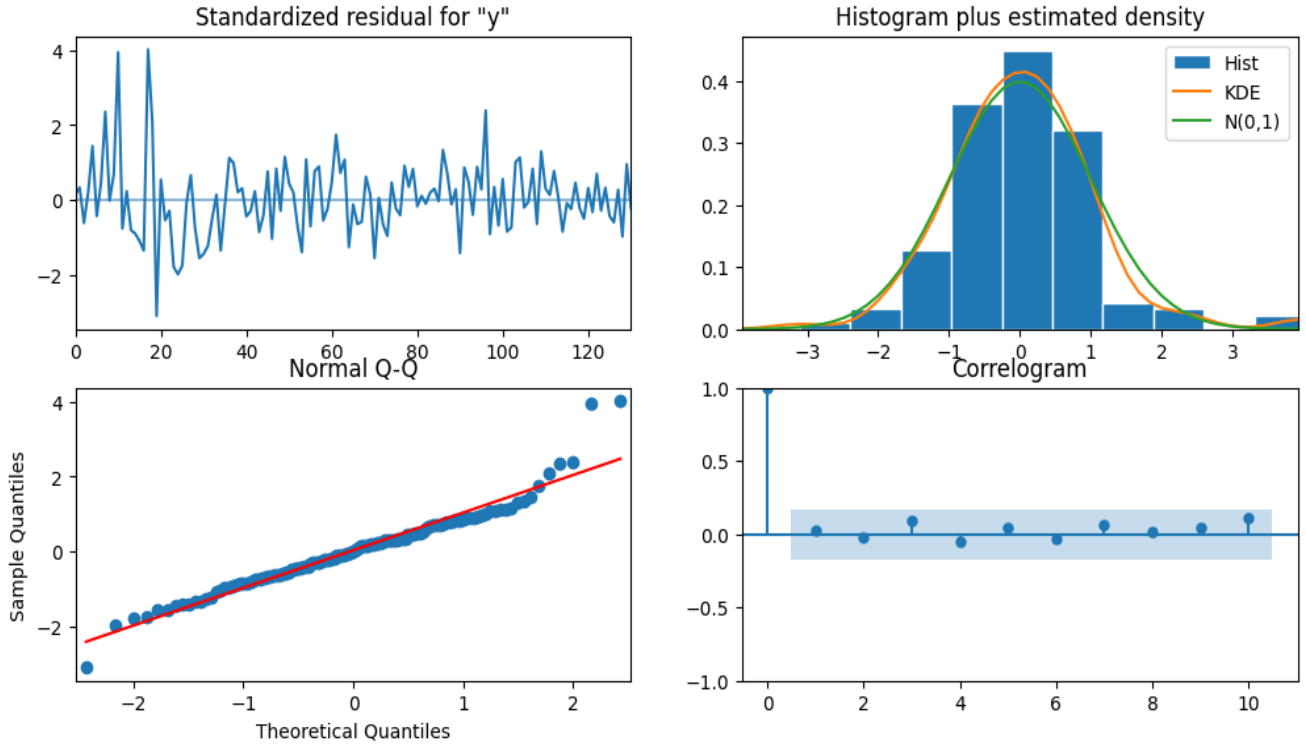
SARIMAX Results

Dep. Variable:		y	No. Observations:			
Model:	SARIMAX(3, 1, 2)x(3, 0, 2, 12)		Log Likelihood			-595.
Date:	Sun, 18 Aug 2024		AIC			1212.
Time:	14:58:03		BIC			1244.
Sample:	0		HQIC			1225.
	- 132					
Covariance Type:	opg					

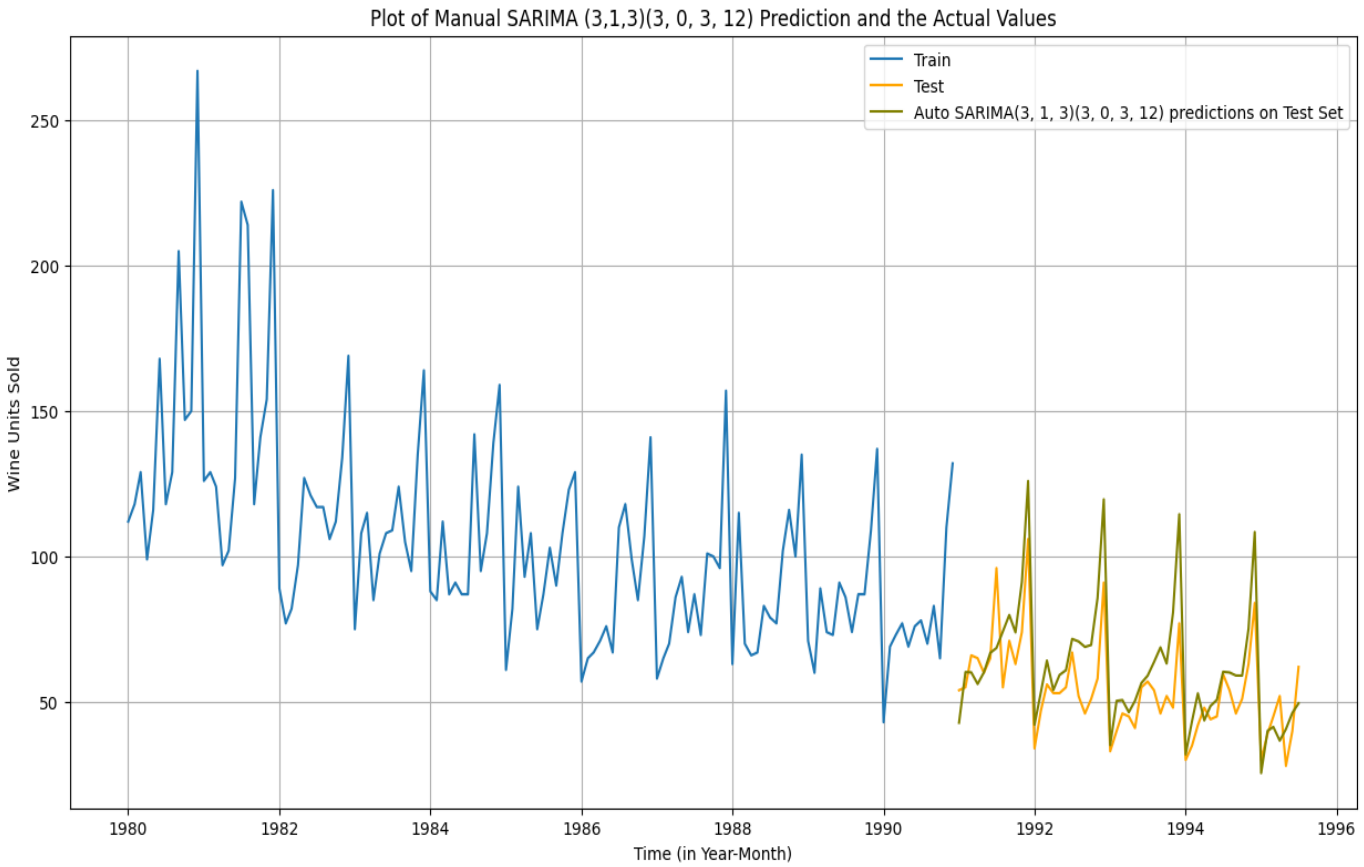
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8496	3.077	-0.276	0.782	-6.881	5.182
ar.L2	-0.0275	0.484	-0.057	0.955	-0.977	0.922
ar.L3	-0.1777	0.554	-0.321	0.749	-1.264	0.909
ma.L1	0.1046	3.038	0.034	0.973	-5.850	6.059
ma.L2	-0.8951	2.690	-0.333	0.739	-6.168	4.378
ar.S.L12	0.0921	0.431	0.214	0.831	-0.752	0.936
ar.S.L24	0.8707	0.504	1.728	0.084	-0.117	1.858
ar.S.L36	0.0206	0.182	0.113	0.910	-0.336	0.377
ma.S.L12	0.1867	1.242	0.150	0.881	-2.248	2.621
ma.S.L24	-0.8517	0.974	-0.875	0.382	-2.760	1.057
sigma2	399.5444	637.205	0.627	0.531	-849.354	1648.443

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	54.92
Prob(Q):	0.74	Prob(JB):	0.00
Heteroskedasticity (H):	0.29	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	5.84

Create the diagnostics plot of Manual SARIMA



Plot of Manual SARIMA



Manual SARIMA: Model Evaluation

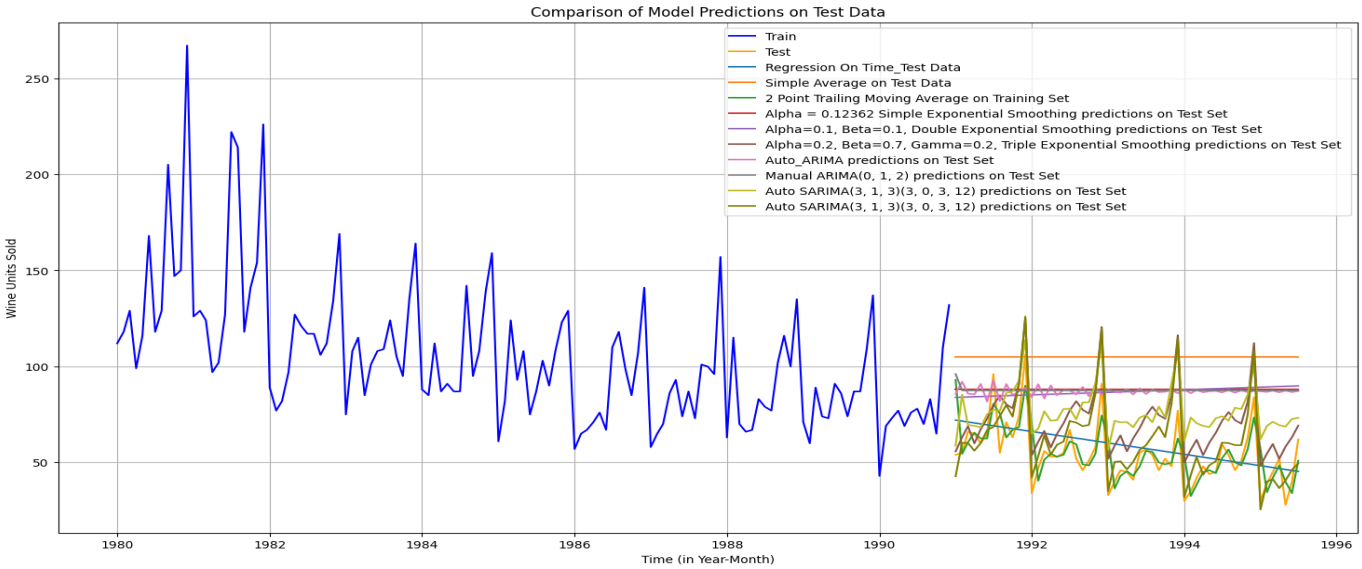
For evaluating the model performance, we look at root means squared error (RMSE)

Model	Test RMSE
ARIMA (p=3, d=1, q=3) (P=3, D=0, Q=3, F=12)	13.9639

6 - Compare the performance of the models

Sort the Data Frame by RMSE in ascending order for ALL created Model

	Test RMSE
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.992350
2pointTrailingMovingAverage	11.589082
Manual_SARIMA(3, 1, 2)(3, 0, 2, 12)	13.963902
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	14.127706
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
RegressionOnTime	15.278158
Auto_SARIMA(3,1,3)(3, 0, 3, 12)	24.615887
Auto ARIMA	36.309428
Alpha=0.1,SimpleExponentialSmoothing	36.429535
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Manual ARIMA(0,1,2)	36.909214
Alpha = 0.12362, SimpleExponentialSmoothing	37.192623
Simple Average Model	53.049755



Analysis Summary

After evaluating various time series forecasting models, the model that exhibited the best performance is the Triple Exponential Smoothing model with parameters Alpha=0.2, Beta=0.7, and Gamma=0.2. This model achieved the lowest Root Mean Square Error (RMSE) of 8.99.

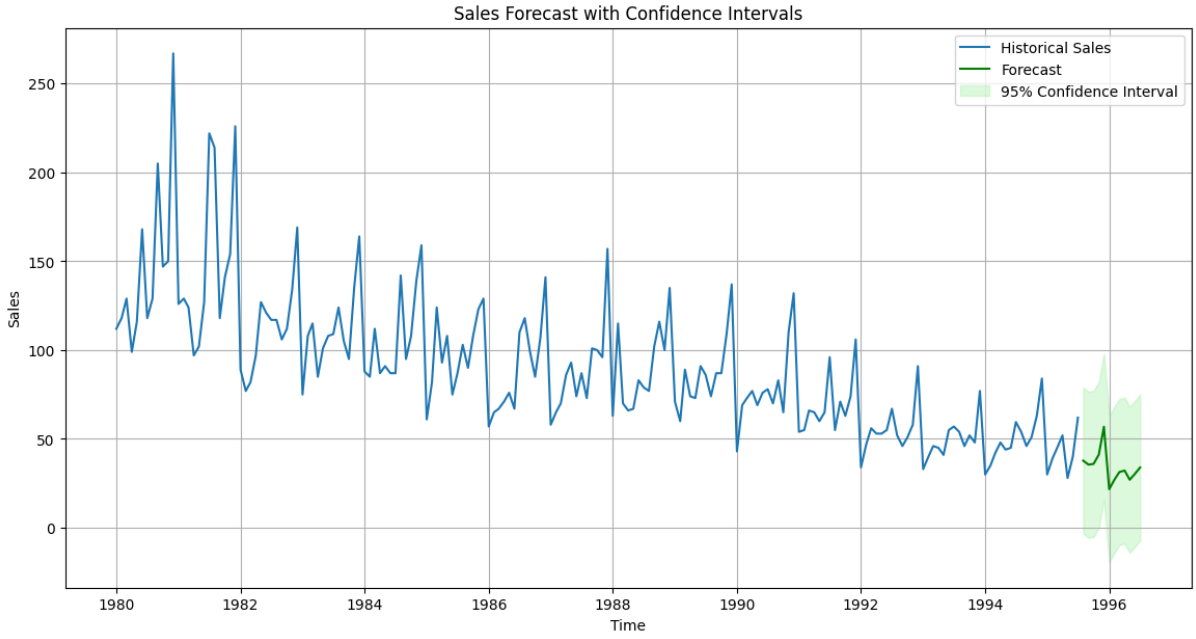
Best Model

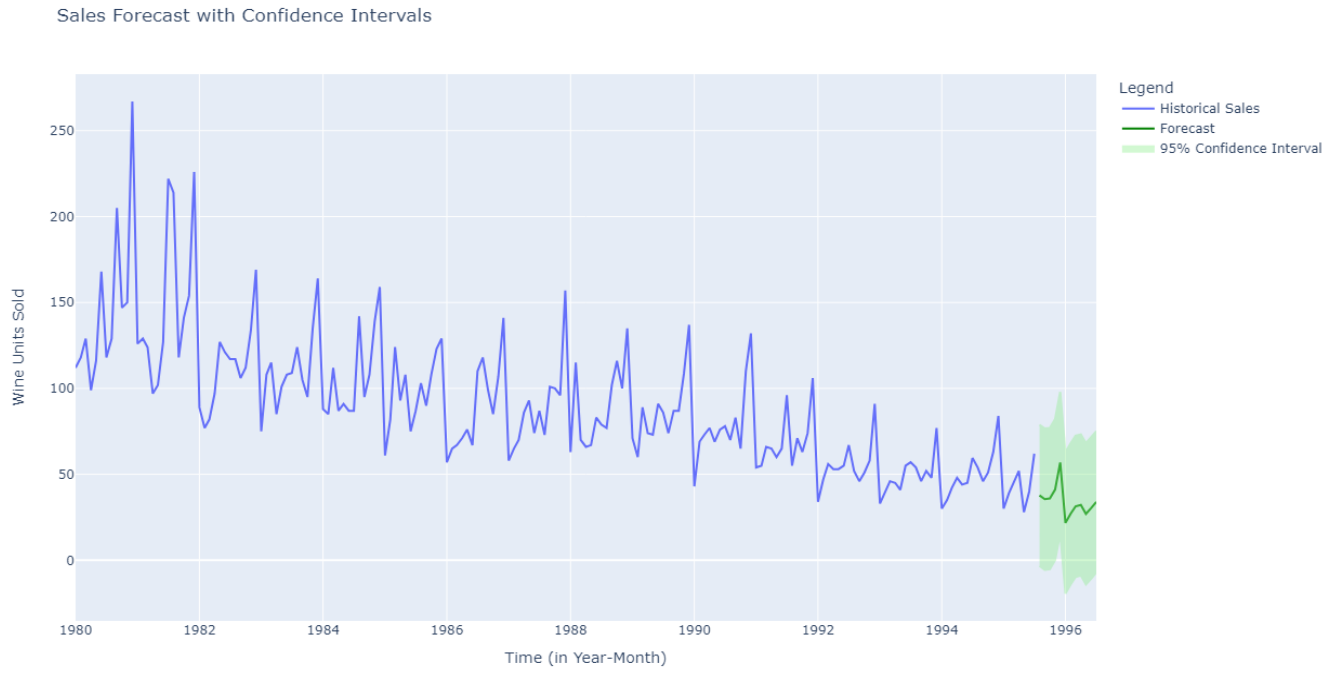
The Triple Exponential Smoothing (Alpha=0.2, Beta=0.7, Gamma=0.2) model demonstrated superior accuracy in predicting the dataset, outperforming other models such as Auto ARIMA, Manual ARIMA, and various moving average techniques. The low RMSE value indicates that this model has the best fit, making it the most reliable choice for forecasting in this context.

Rebuild the best model using the entire data - Make a forecast for the next 12 months

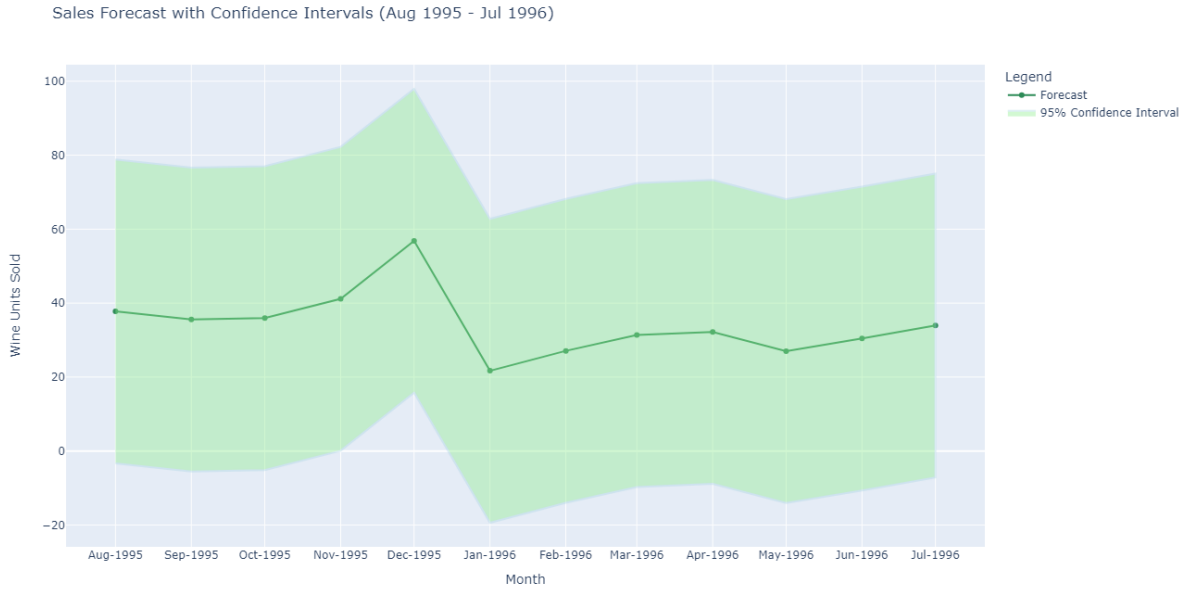
Forecast the next 12 months

```
1995-08-01    37.765009
1995-09-01    35.565255
1995-10-01    35.942844
1995-11-01    41.159755
1995-12-01    56.859748
1996-01-01    21.698798
1996-02-01    27.087773
1996-03-01    31.374379
1996-04-01    32.198521
1996-05-01    27.023700
1996-06-01    30.422915
1996-07-01    33.927989
Freq: MS, dtype: float64
```





TES Optimum Model – Time series plot forecast with confidence intervals



TES Optimum Model – Time series plot forecast for next 12 months with confidence intervals

Forecast: The predicted sales values for the next 12 months.

Confidence Intervals: The range within which the actual values are expected to fall with 95% confidence. The confidence intervals are calculated as the forecasted values ± 1.96 times the standard error of the residuals.

Plot: This shows the historical sales data, the forecasted sales, and the shaded area representing the confidence intervals.

7 - Actionable Insights & Recommendations

Key Findings

- Seasonal Pattern: The sales data exhibits a clear seasonal pattern with higher sales during the holiday season (November-December) and a subsequent decline in January.
- Moderate Growth: Overall, the forecast indicates a moderate growth trend in wine sales throughout the year.
- Forecast Uncertainty: The confidence intervals suggest a degree of uncertainty in the forecast, particularly during the peak and off-peak seasons.

Conclusion Based on the Analysis:

- Sales Performance: The historical data shows volatility in sales, with significant spikes followed by declines. The forecasted trend suggests that future sales are likely to be more stable but at a lower level compared to earlier highs.
- Strategic Insights: ABC Estate Wines may need to investigate the factors behind the earlier peaks in sales to understand what drove higher demand during those periods. Additionally, understanding the causes of the subsequent decline could help in developing strategies to boost future sales.
- Forecast Confidence: The relatively narrow confidence intervals suggest that the forecast is reliable, but the downward trend indicates a need for intervention if the company wishes to see growth.

Recommendations:

- Market Analysis: Further analysis into market conditions, consumer preferences, and competitive factors during the periods of peak sales could provide insights into how to replicate past successes.
- Promotional Strategies: Considering targeted marketing and promotional efforts to re-ignite interest in this wine variety could help to reverse the downward trend in sales.
- New Opportunities: Exploring new markets or introducing innovative products may also provide avenues for growth.

In summary, the forecast suggests stable but subdued sales in the near future, prompting a need for strategic actions to enhance performance and capitalize on potential market opportunities.

Conclusion

By carefully analyzing the sales forecast and identifying key trends, businesses can implement strategic initiatives to optimize operations, increase sales, and mitigate risks. A data-driven approach, coupled with a deep understanding of customer behavior and market dynamics, will be instrumental in achieving sustained growth and profitability.

To further enhance decision-making, consider conducting a more in-depth analysis of:

- Customer behavior: Identify factors influencing purchasing decisions and preferences.

- Competitive landscape: Analyze competitor strategies and market share.

*Economic indicators: Assess the impact of economic factors on sales.

By incorporating these additional insights, businesses can refine their strategies and gain a competitive advantage.

Project - 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.

Data Dictionary

Variable Name	Description
YearMonth	Represents the year and month in which the sales were recorded
Sparkling	Denotes the number of wine units sold

Data Description –

1. YearMonth: Datatime variable from 1980-01 to 1995-07

2. Sparkling: Continuous from 1070 to 7242

1 - Define the problem and perform Exploratory Data Analysis

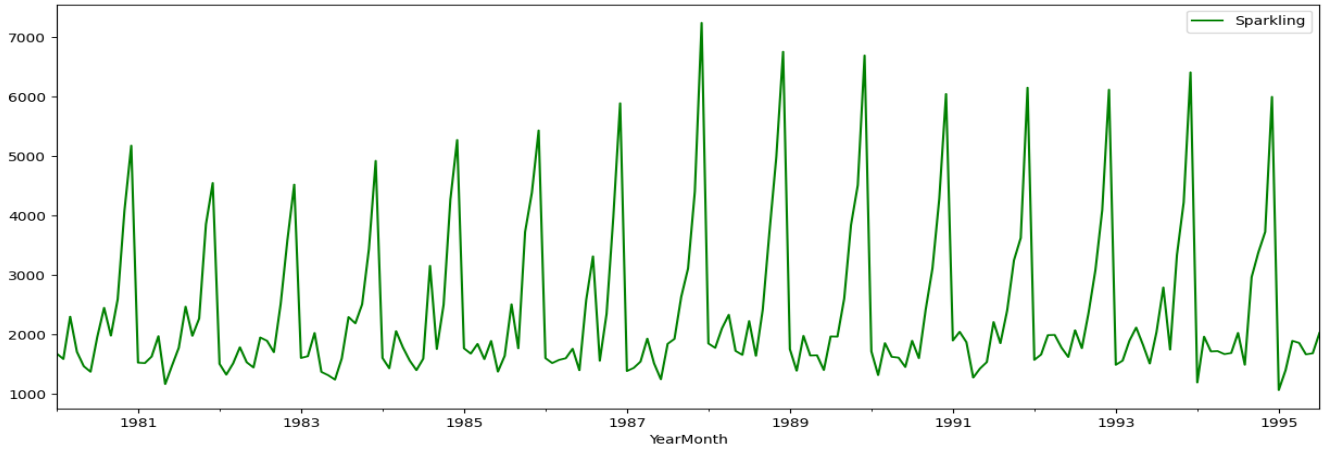
Check the data type and columns details

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Check the Top 5 and Last 5 rows details.

```
Top 5 Rows in dataset
            Sparkling
YearMonth
1980-01-01        1686
1980-02-01        1591
1980-03-01        2304
1980-04-01        1712
1980-05-01        1471

Last 5 Rows in dataset
            Sparkling
YearMonth
1995-03-01        1897
1995-04-01        1862
1995-05-01        1670
1995-06-01        1688
1995-07-01        2031
```

Plot the graph



Trend Analysis

- Upward Trend: The overall direction of the data is upward, indicating an increasing trend in sparkling wine sales over the years.
- Non-Linearity: The upward movement is not perfectly linear. There are periods of faster growth and periods of slower growth, suggesting a non-linear trend.

Seasonality Analysis

- Presence of Seasonality: The data exhibits a clear seasonal pattern with peaks and troughs recurring at regular intervals.
- Magnitude: The magnitude of the peaks and troughs appears to be relatively constant throughout the time series.

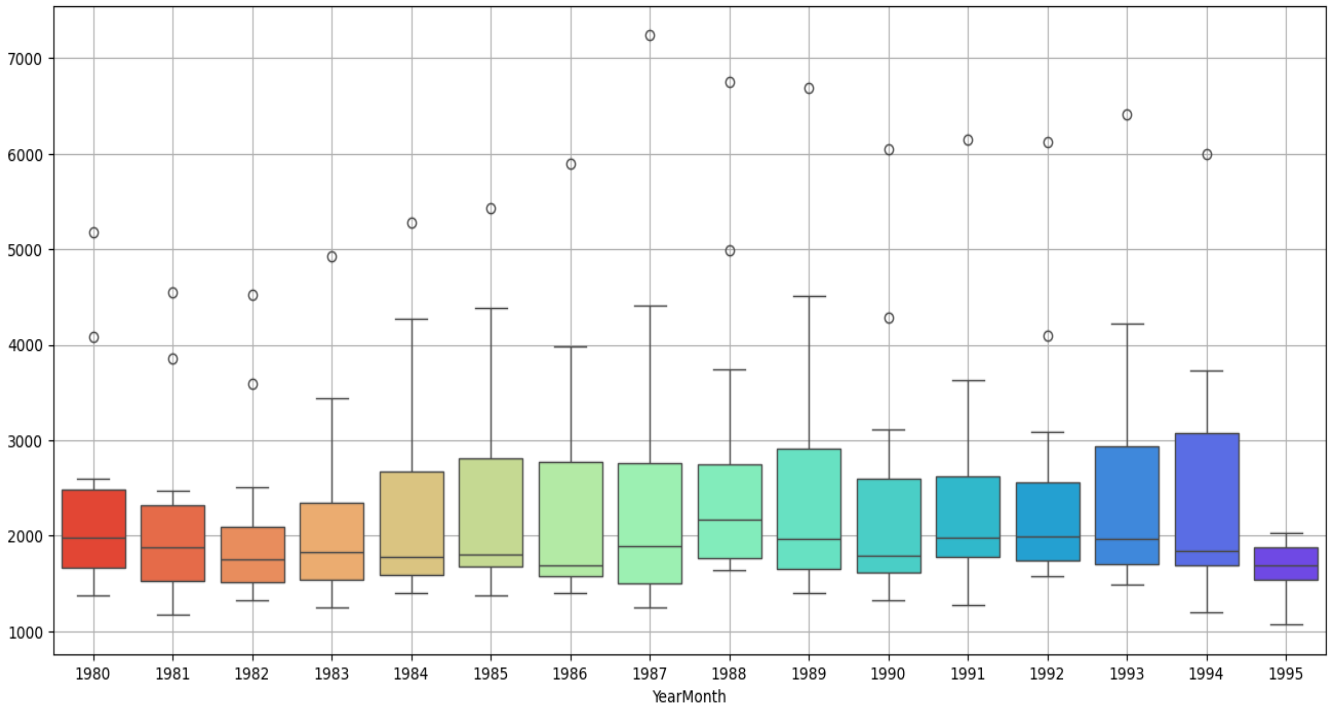
Plot a boxplot to understand the spread of sales across different years and within different months across years.

Convert into Date and Month

YearMonth	Sparkling	Month	Year
1980-01-01	1686	1	1980
1980-02-01	1591	2	1980
1980-03-01	2304	3	1980
1980-04-01	1712	4	1980
1980-05-01	1471	5	1980

Check the Missing Value – We observed that there is no missing value in given dataset

Yearly Boxplot



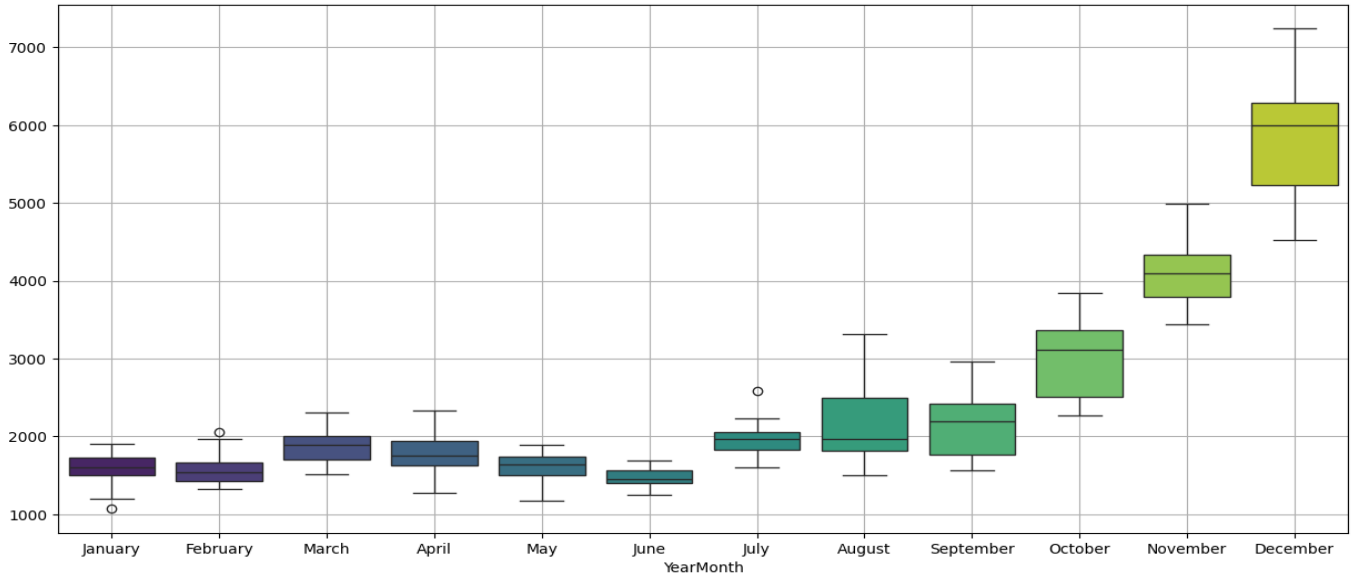
Observation:

• 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 Boxplot –



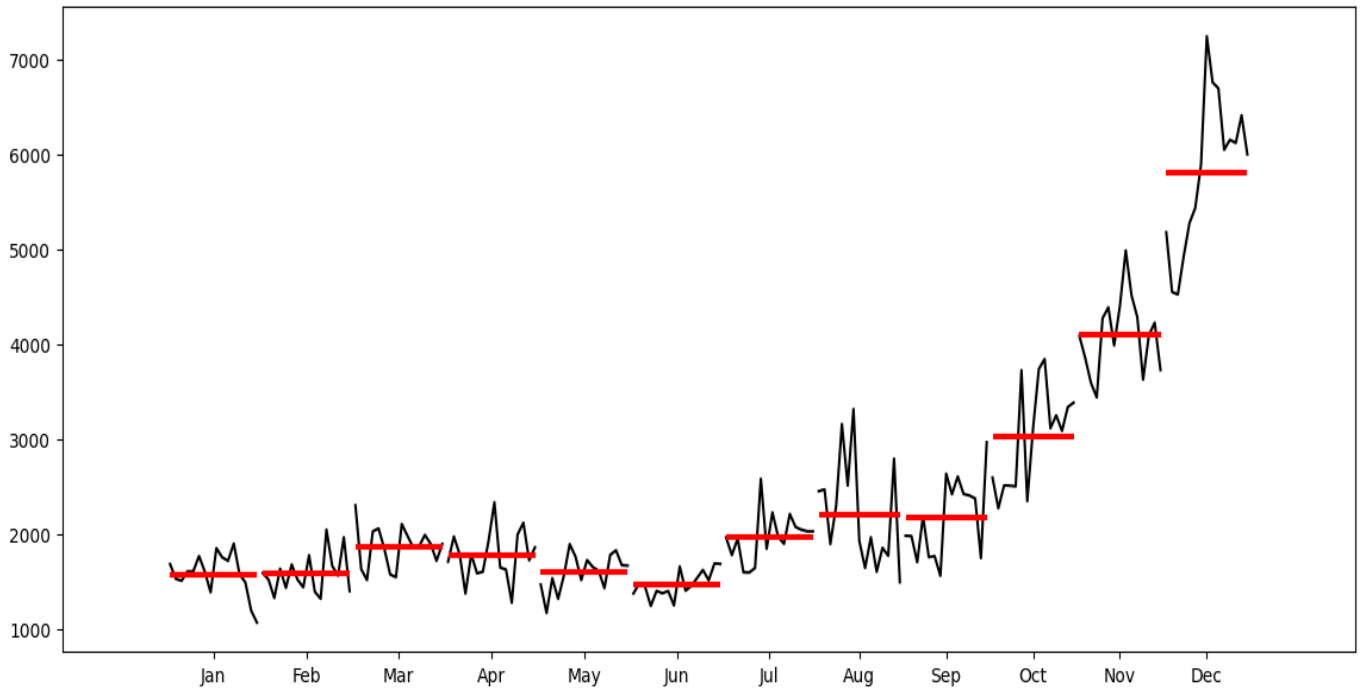
Observation:

• 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.

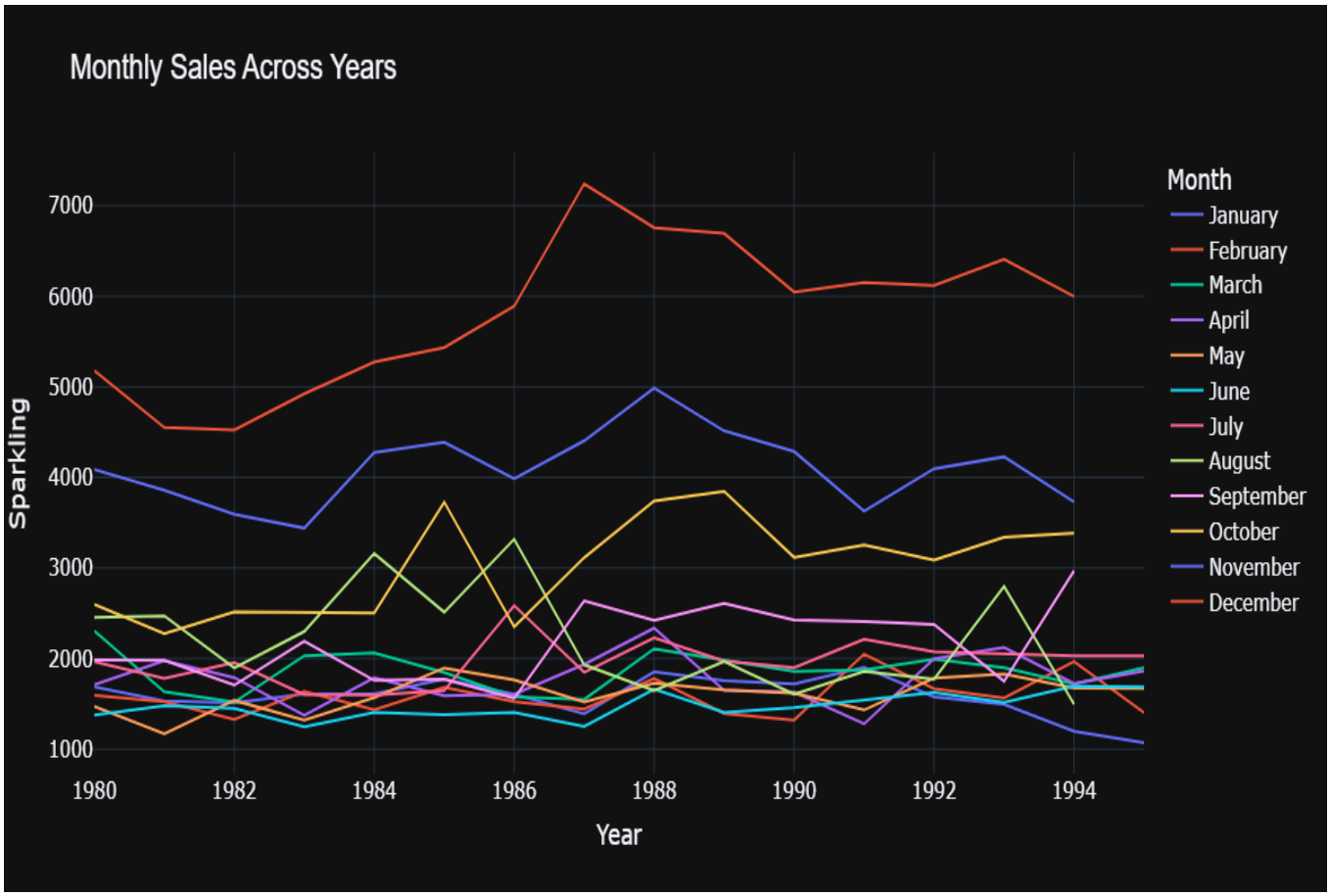
• January has the lowest wine sales while December sees the greatest. The sales modestly grow from January to August and then sharply climb after that.

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

Plot the Time Series according to different Months for Different Years



Graph of Monthly Sales across Years



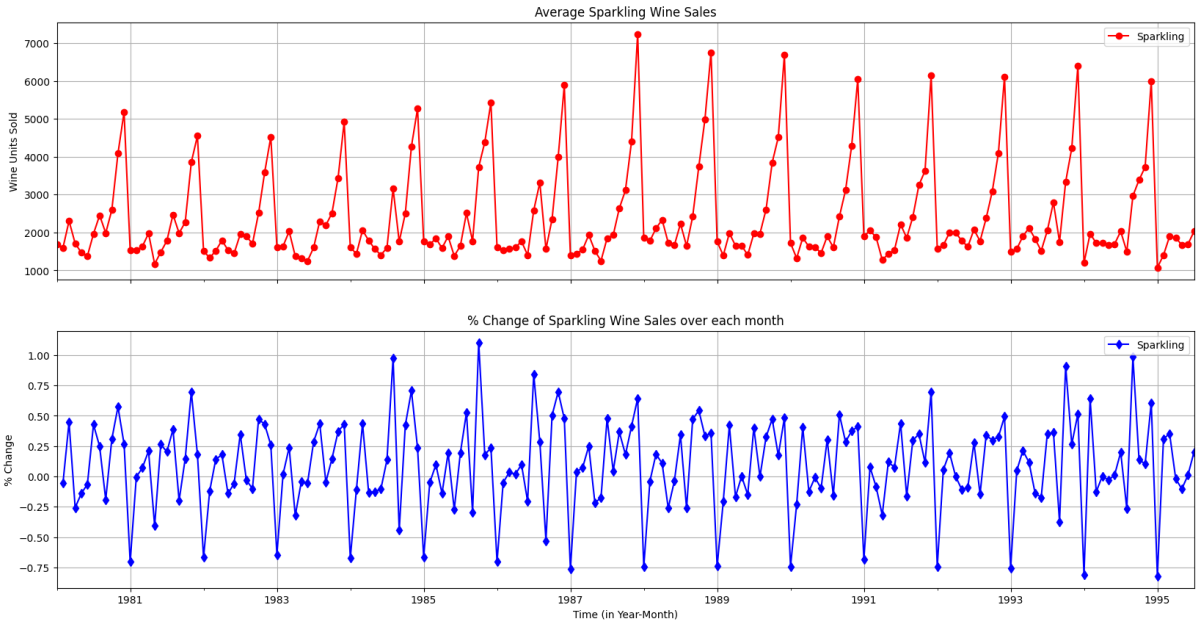
Observation:

• 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.

Plot the average sales per month and the month-on-month percentage change of sales.

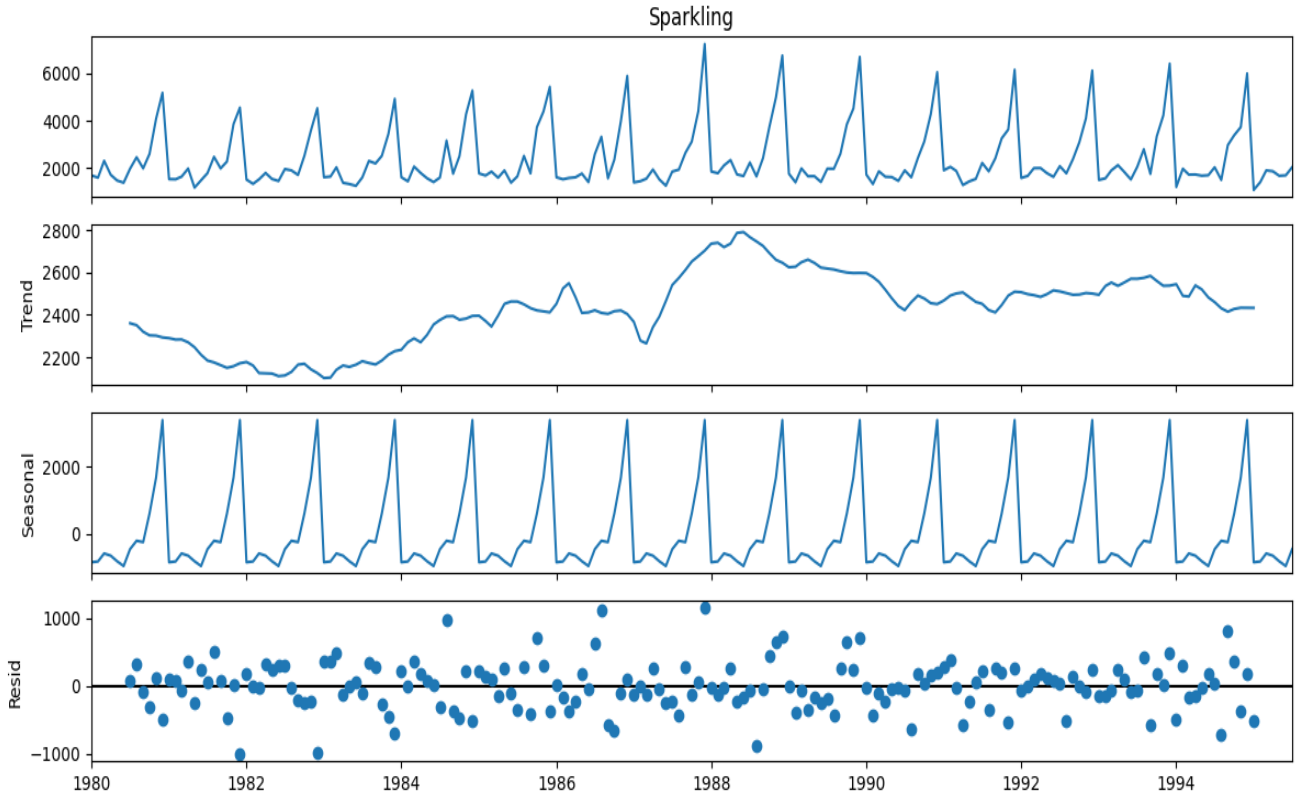


Observation:

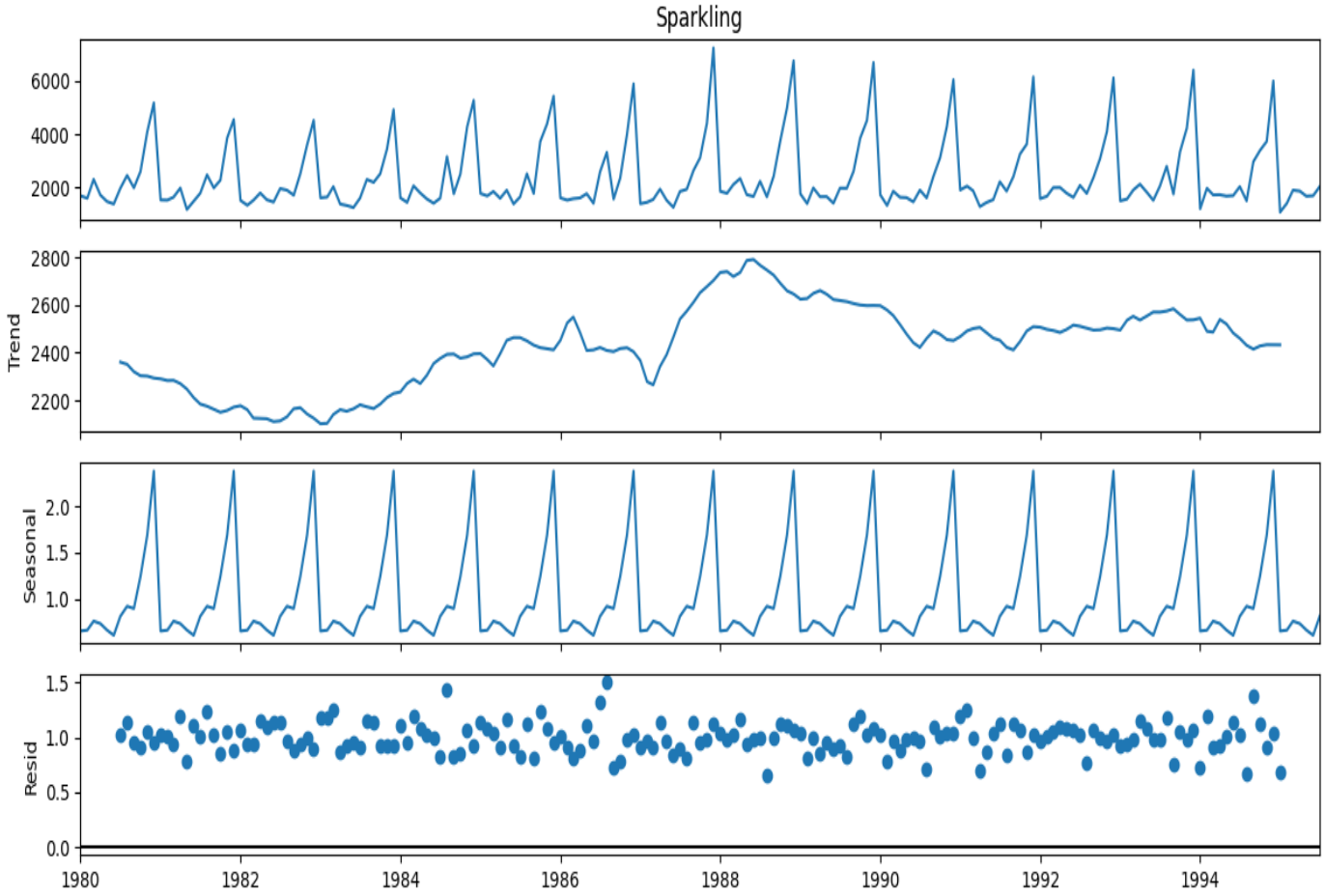
• We can see that there is a no trend but only seasonality from the average sales and % change plots. Additionally, the seasonality in the percentage change appears to be consistent throughout all the years.

Perform Decomposition

Additive Decomposition



Multiplicative Decomposition –



Additive vs. Multiplicative Model

- Multiplicative Model: Given that the magnitude of the seasonal fluctuations appears to be proportional to the level of the time series, a multiplicative model would be more appropriate. This is because the amplitude of the seasonal peaks and troughs increases as the overall level of sales increases.

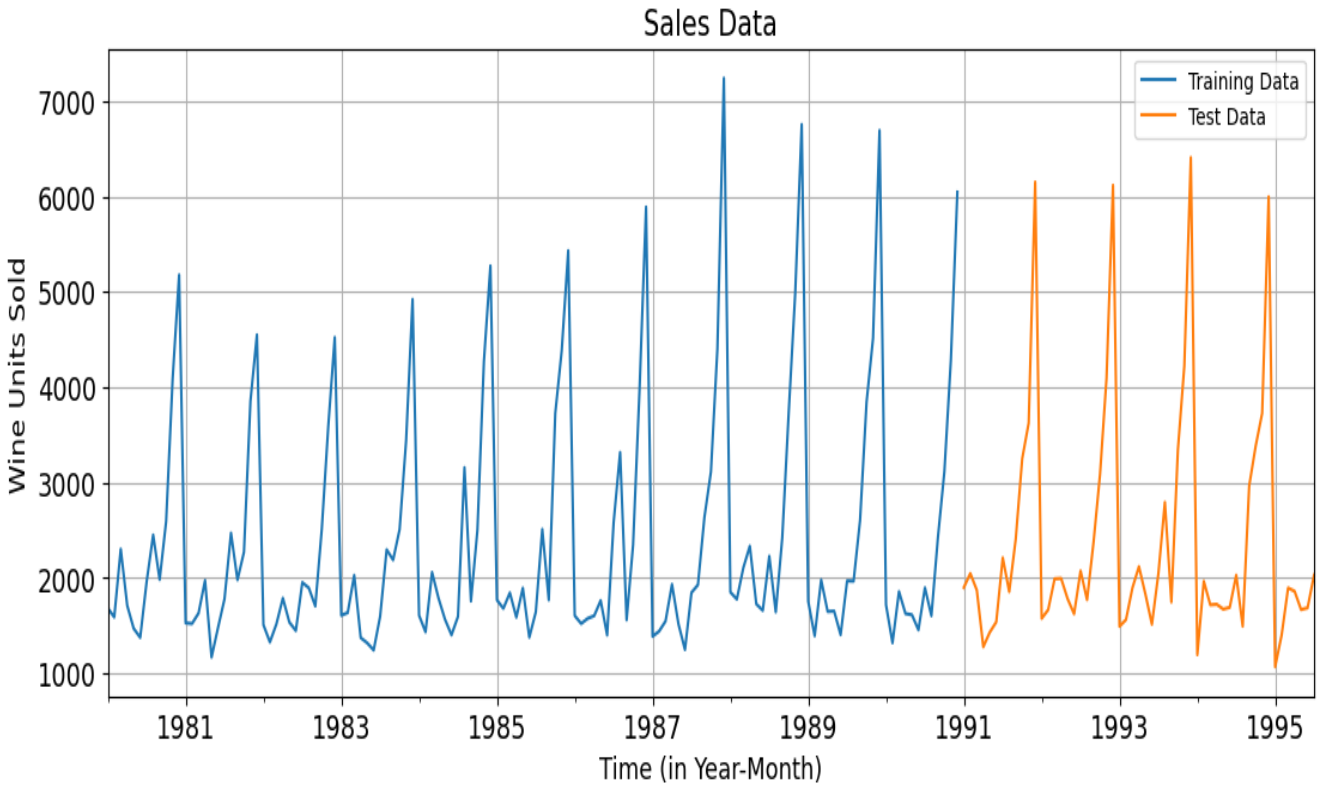
In Summary:

- The time series displays an upward, non-linear trend.
- Seasonality is present with relatively constant magnitude.
- A multiplicative model is suitable due to the changing amplitude of the seasonal fluctuations.

Drop the 'Year' and 'Month' Columns.

2 - Data Pre-processing - Split the Data into Training and Test

Plot the Train and Test -



3 - Model Building And Check the Performance of the Models Built

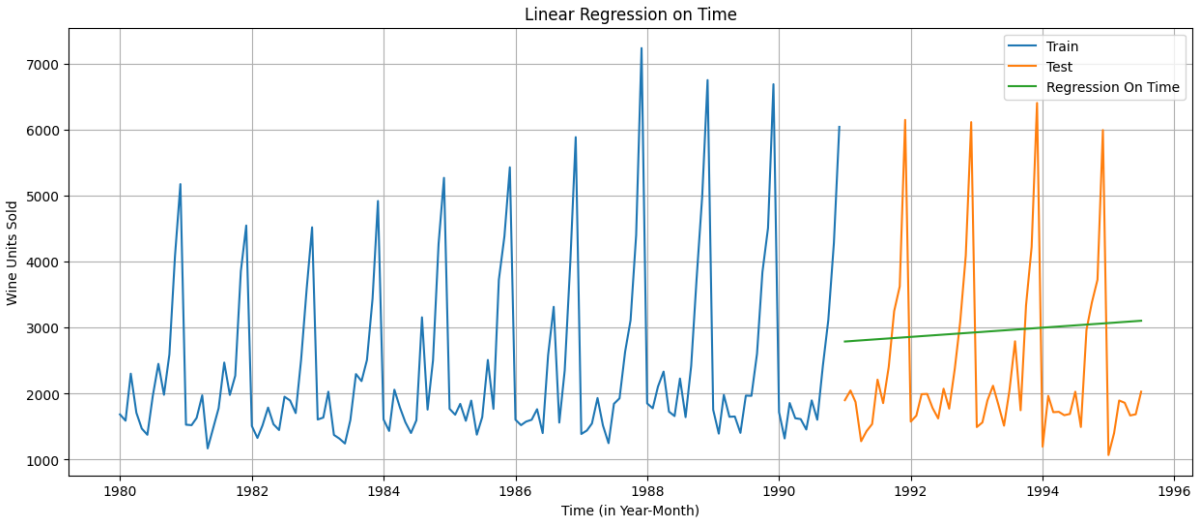
Build Forecasting Models

- Model 1: Linear Regression
- Model 2: Simple Average
- Model 3: Moving Average(MA)
- Model 4: Simple Exponential Smoothing
- Model 5: Double Exponential Smoothing (Holt's Model)
- Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

Model 1: Linear Regression

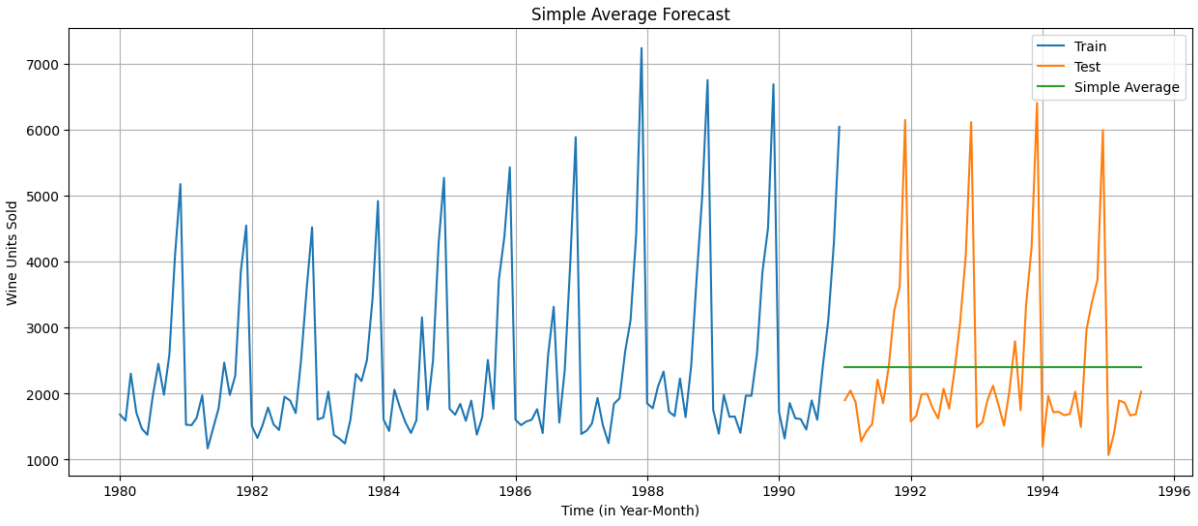
For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

Now that our training and test data has been modified, let us go ahead use **Linear Regression** to build the model on the training data and test the model on the test data.



Defining the Accuracy Metrics – Linear Regression On Time forecast on the Test Data, RMSE is 1389.14

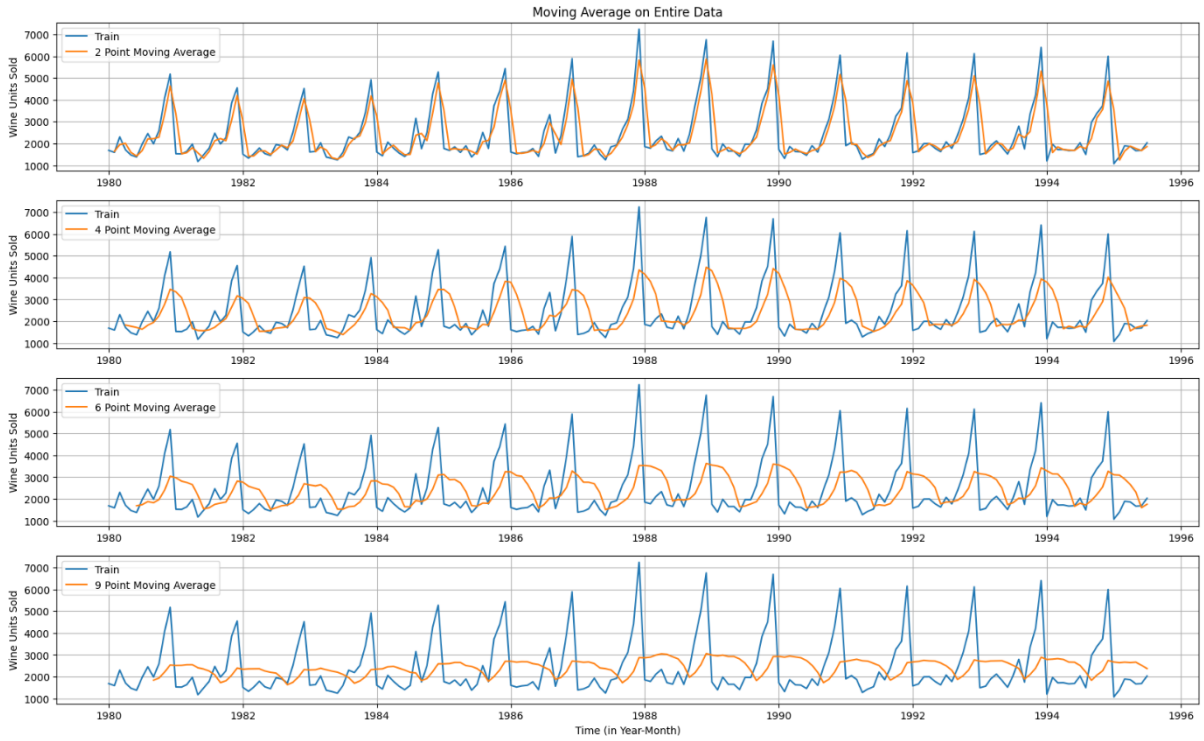
Model 2: Simple Average



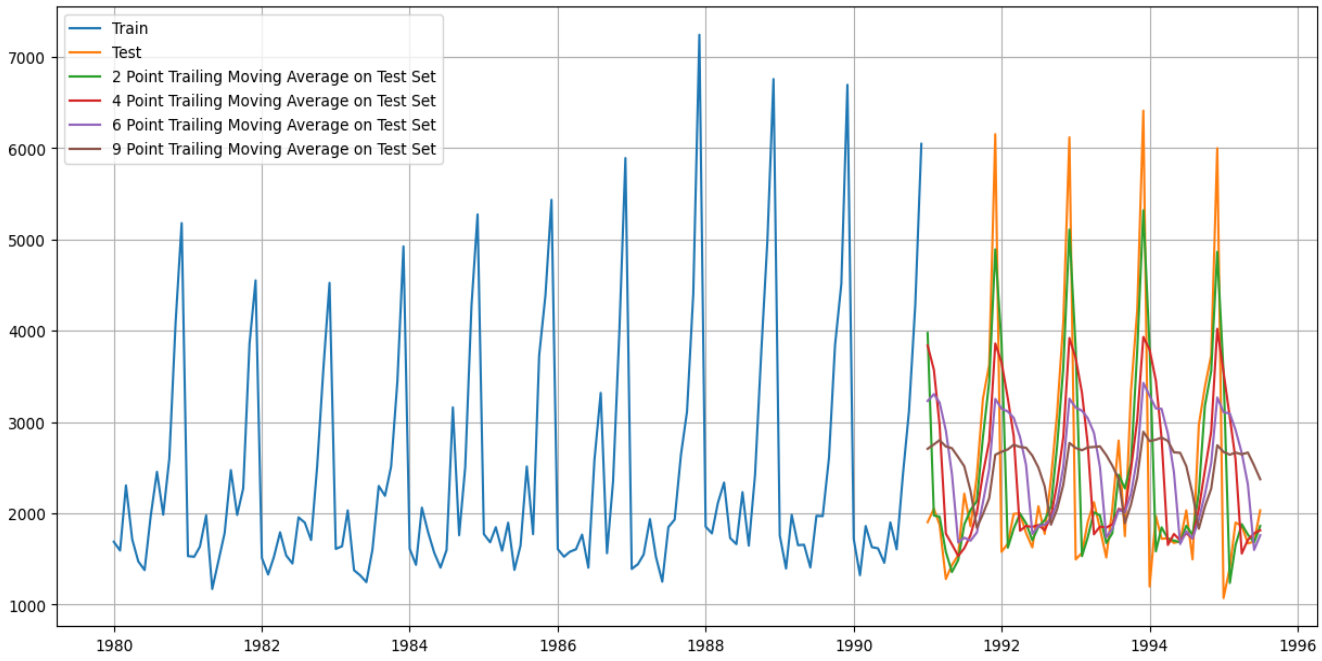
Defining the Accuracy Metrics – Simple Average Model forecast on the Test Data, RMSE is 1275.0818

Model 3: 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.



Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.



Defining the Accuracy Metrics –

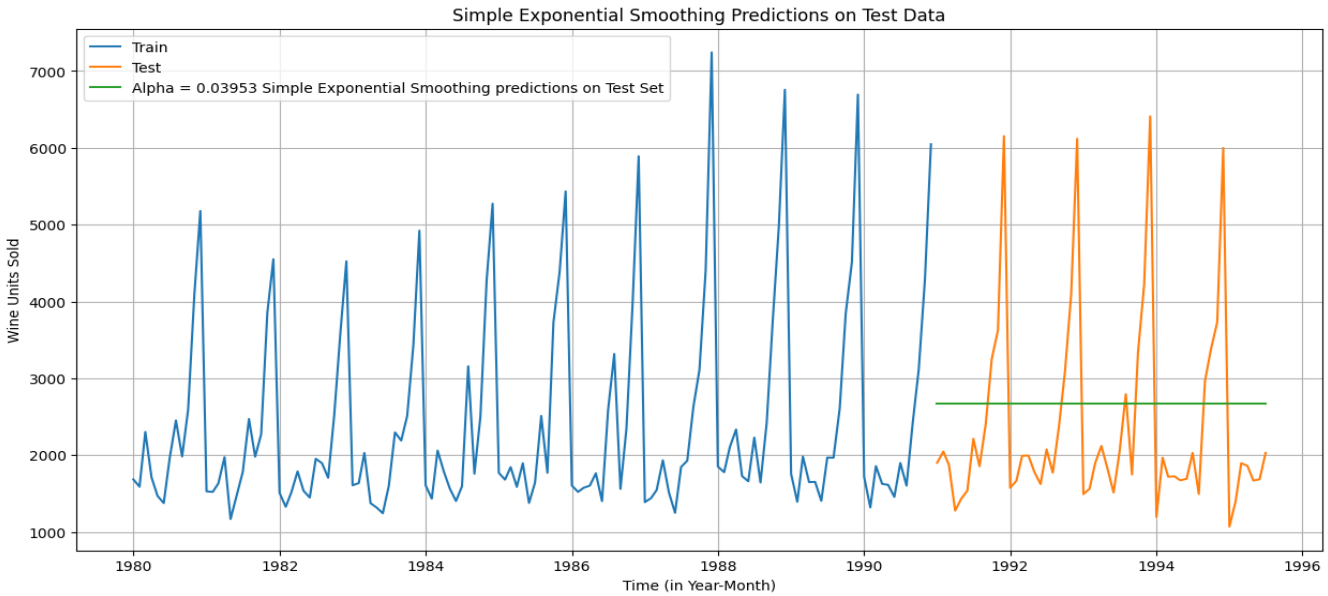
For 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401

For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590

For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927

For 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278

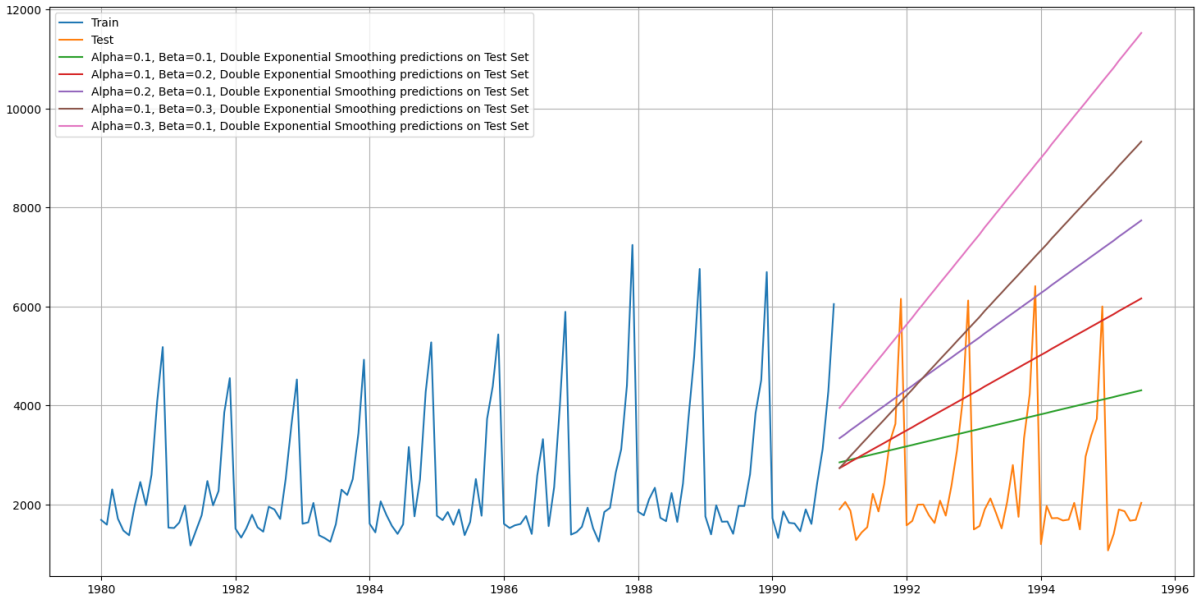
Model 4: Simple Exponential Smoothing



Model Evaluation for Alpha = 0.03953 Simple Exponential Smoothing - For Alpha = 0.03953 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1304.927.

Method 5: Double Exponential Smoothing (Holt's Model)

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

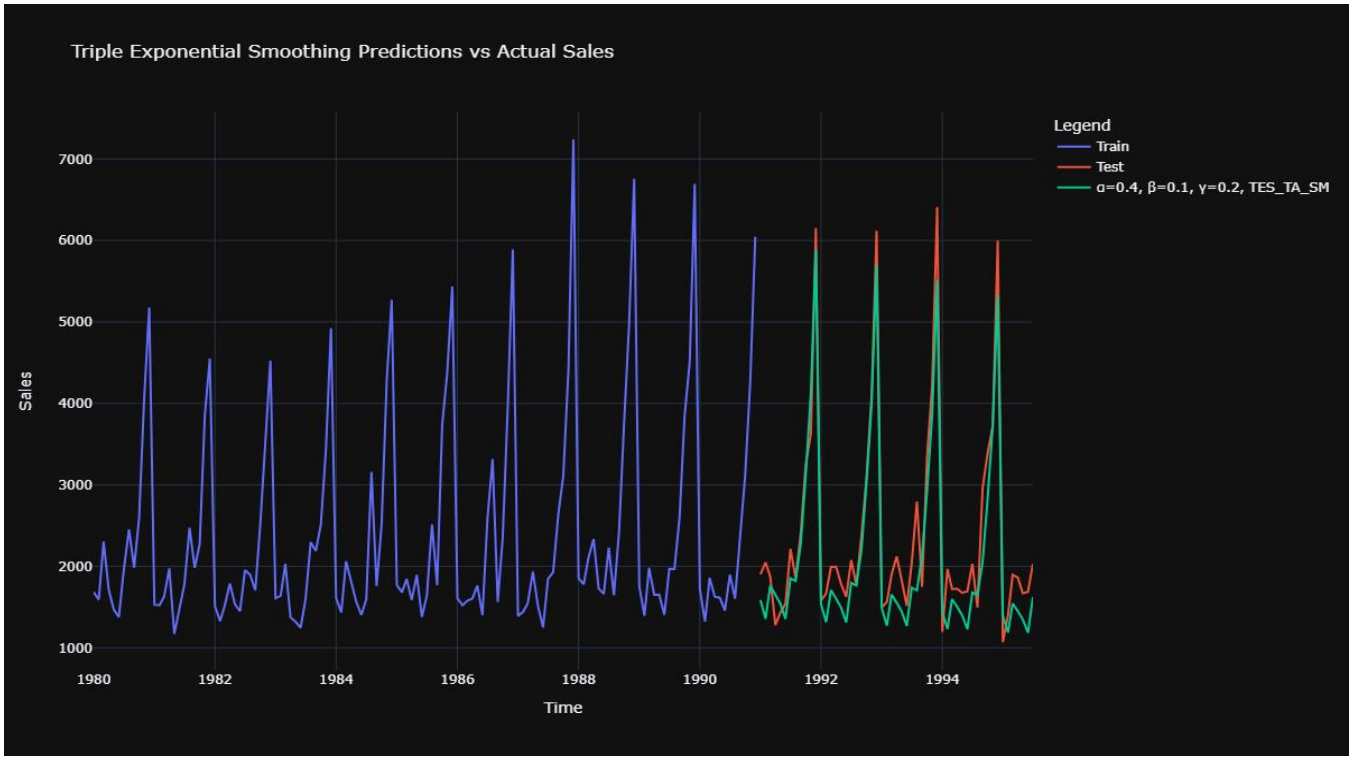


Model Evaluation

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	1382.520870	1778.564670
1	0.1	0.2	1413.598835	2599.439986
10	0.2	0.1	1418.041591	3611.763322
2	0.1	0.3	1445.762015	4293.084674
20	0.3	0.1	1431.169601	5908.185554

Method 5: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters α, β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.



Check the Performance of the models built and Sort in Ascending Order (RMSE)

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
1301	0.4	0.1	0.2	384.467709	317.434302	ta_sm
2245	0.4	0.1	0.3	381.106645	326.579641	tm_sm
1211	0.3	0.2	0.2	388.544148	329.037543	ta_sm
1200	0.3	0.1	0.1	388.220071	337.080969	ta_sm
1110	0.2	0.2	0.1	398.482510	340.186457	ta_sm

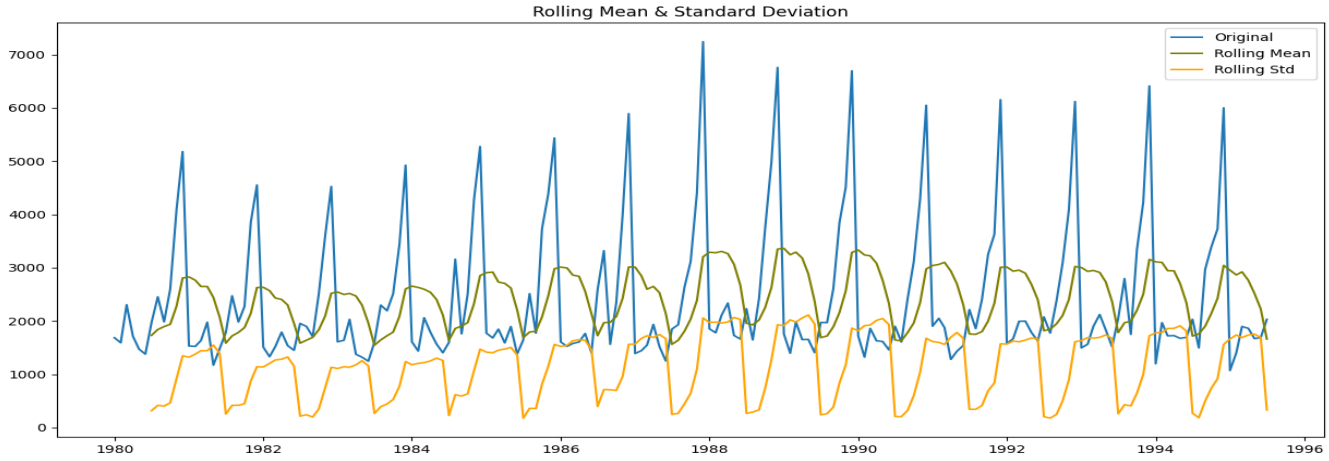
Check for Stationarity

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.

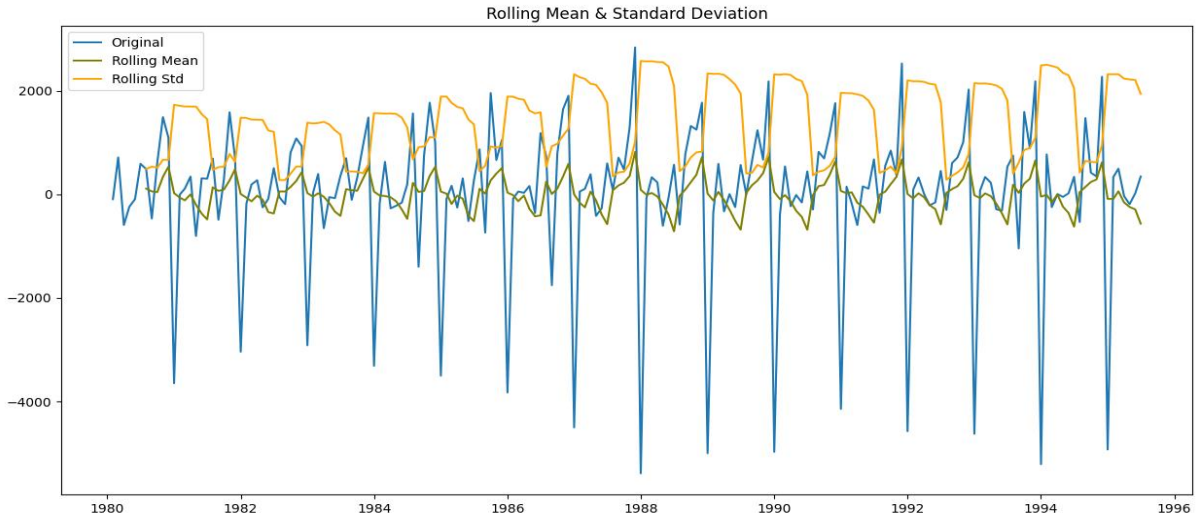


Results of Dickey-Fuller Test:

Test Statistic	-1.360497
p-value	0.601061
#Lags Used	11.000000

We see that at 5% significant level the Time Series is non-stationary.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

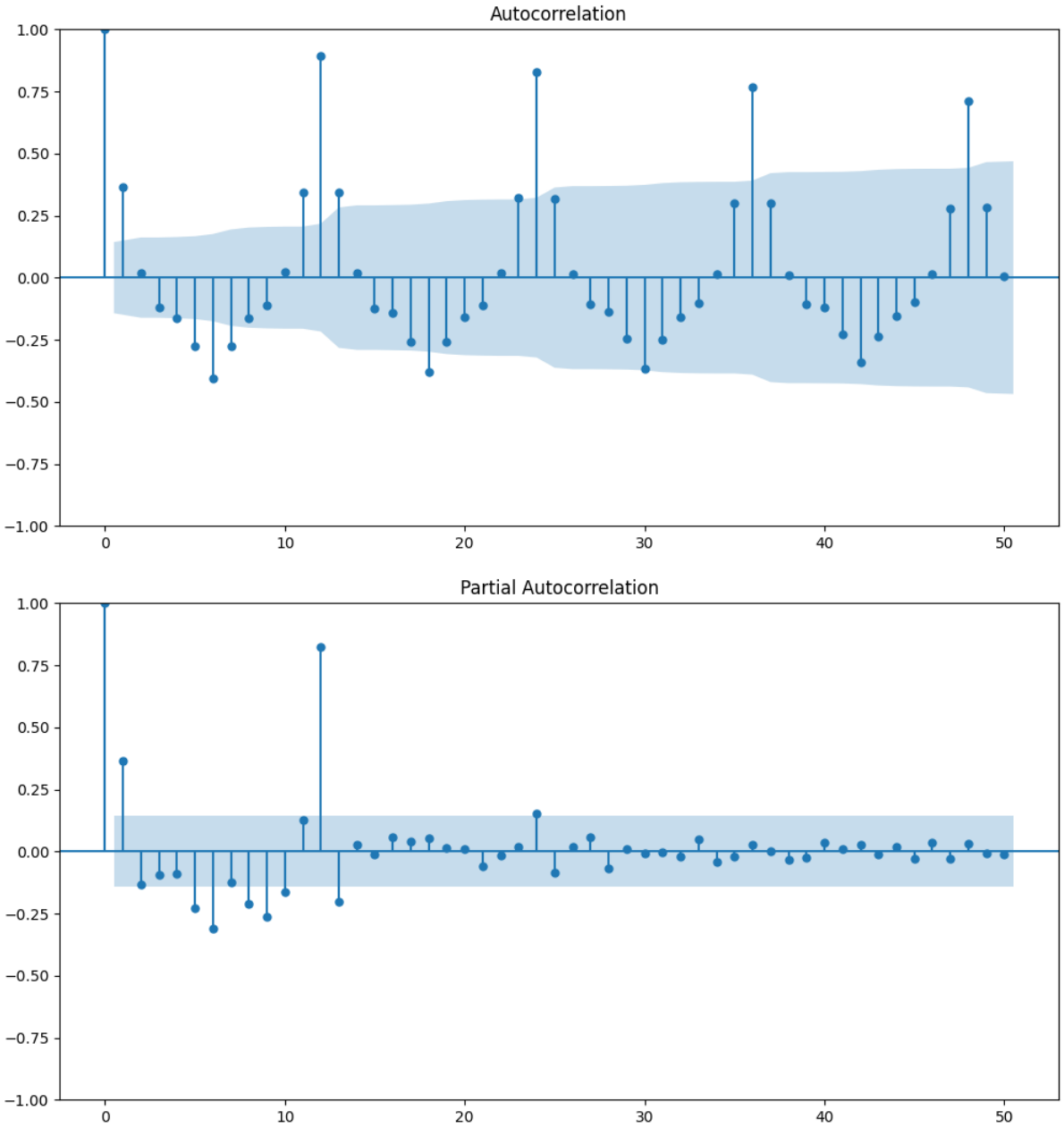


Results of Dickey-Fuller Test:
Test Statistic -45.050301
p-value 0.000000

We observe that the p-value (α) is less than 0.05. Therefore, we reject the null hypothesis. This result suggests that the time series is stationary, meaning its statistical properties such as mean, variance, and autocorrelation remain constant over time.

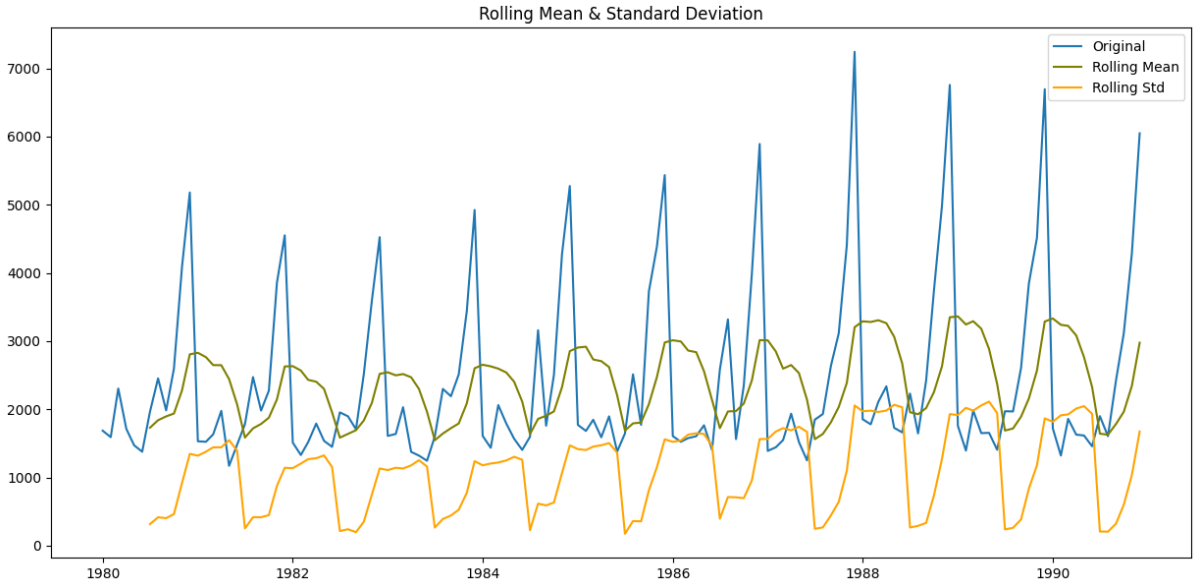
Model Building - Stationary Data

ACF and PACF Plots



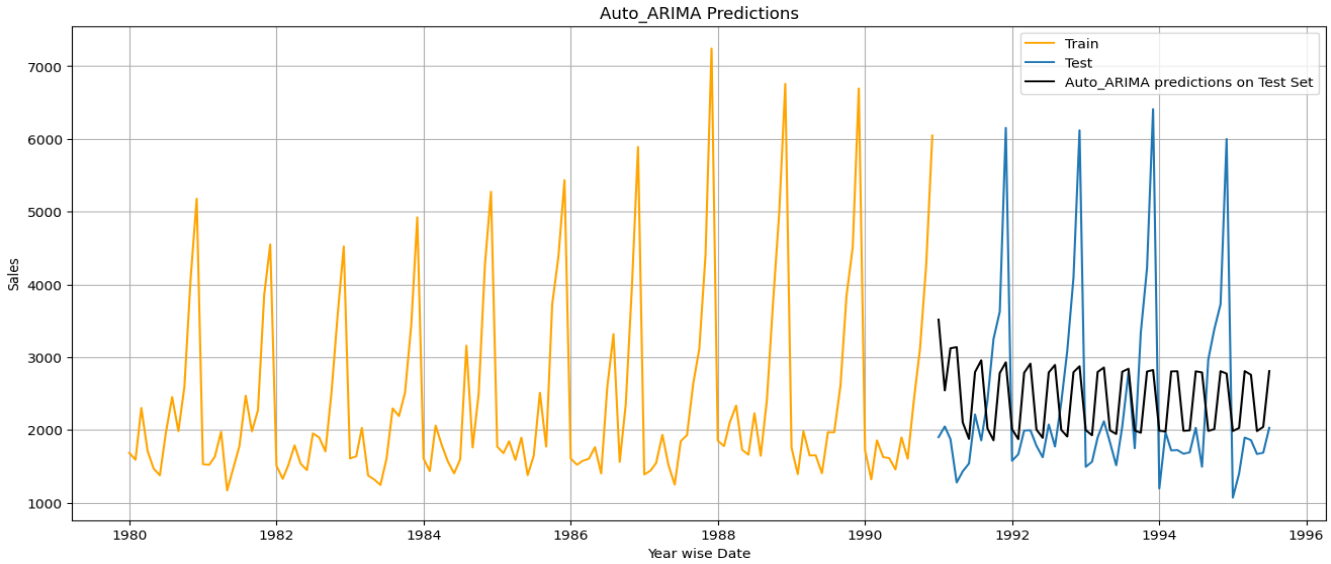
Split the data into train and test and plot the training and test data.
Training Data is till the end of 1990. Test Data is from the beginning of 1991 to the last time stamp provided.

Check the stationarity of the training data



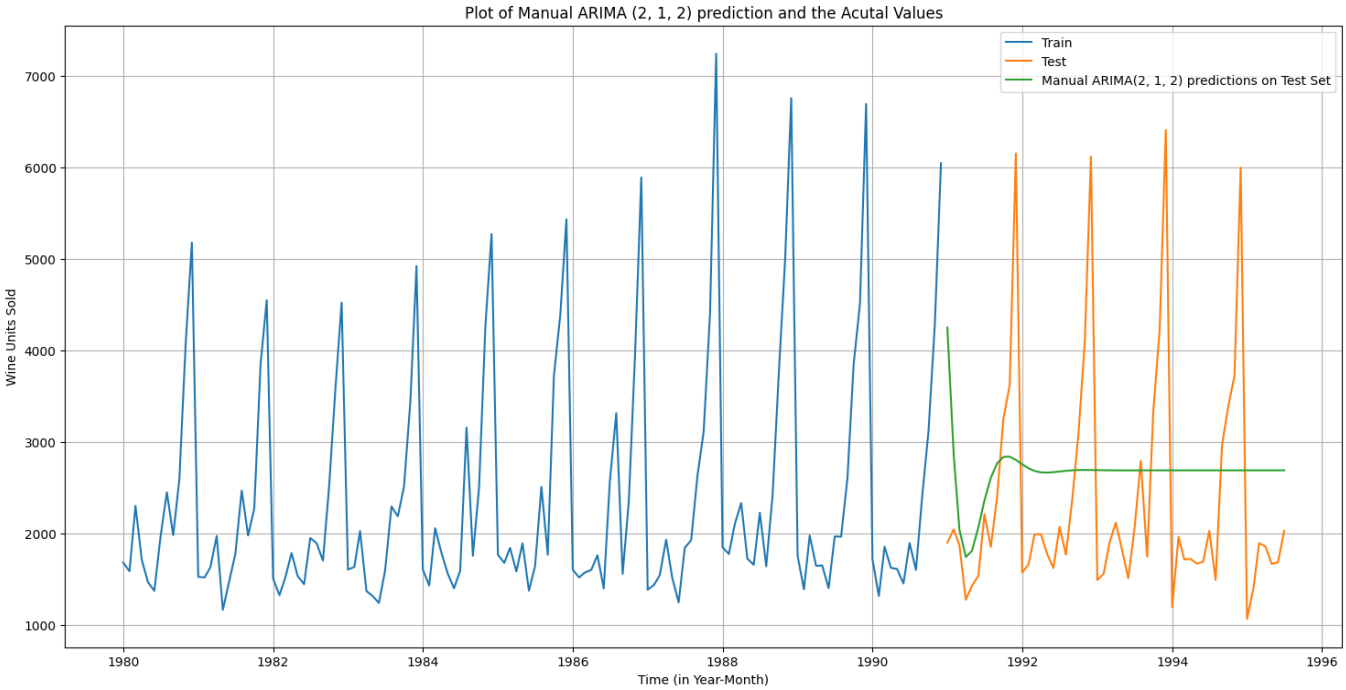
We observe that the p-value (α) is greater than 0.05. Therefore, we fail to reject the null hypothesis. This result suggests that the time series is not stationary, meaning its statistical properties, such as mean, variance, and autocorrelation, change over time.
We see that after taking a difference of order 1 the series have become stationary at α = 0.05.

Build an Automated version of an ARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).
Build the Auto ARIMA



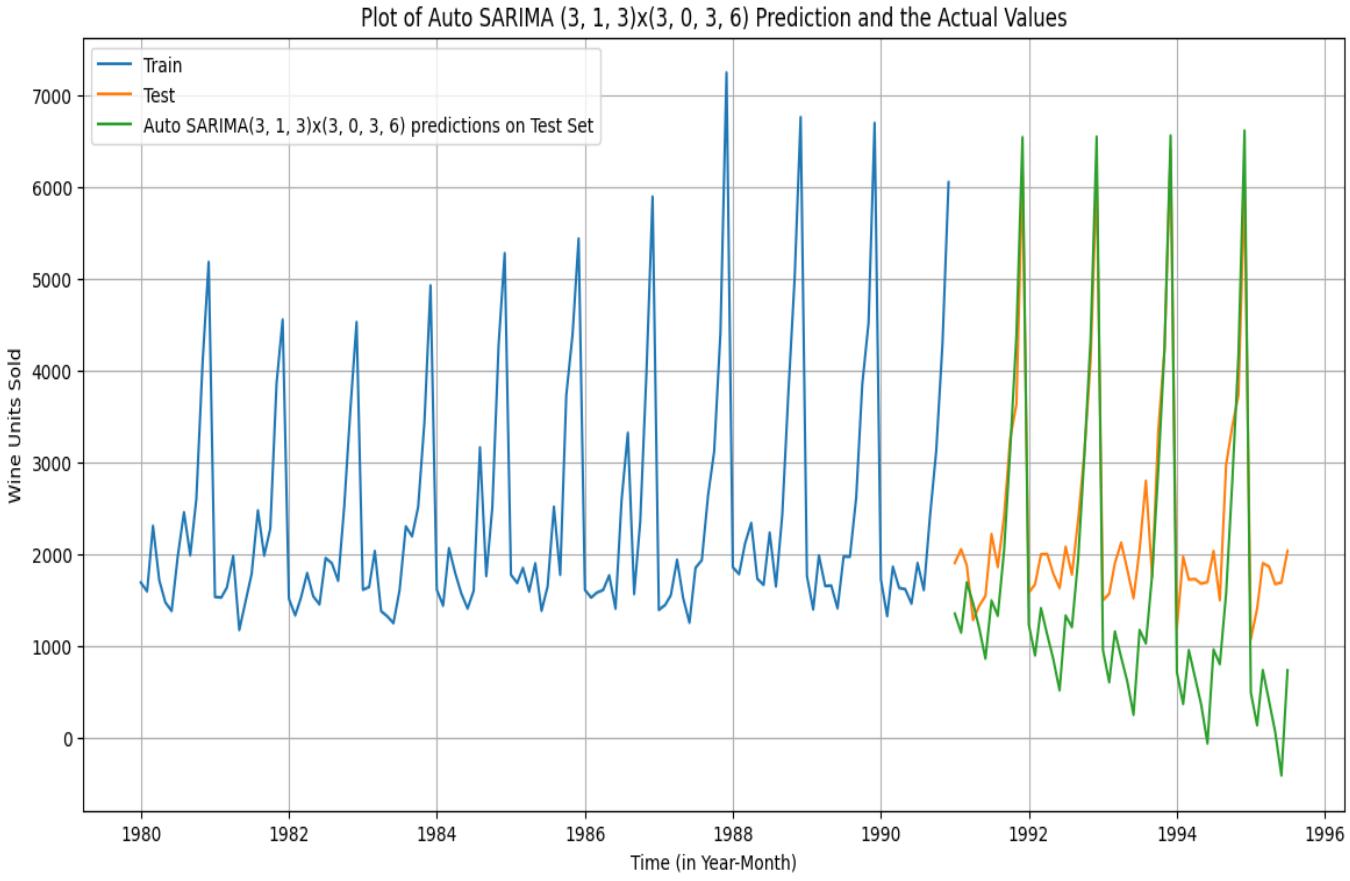
RMSE - 1229.272591696495

Build the Manual ARIMA -



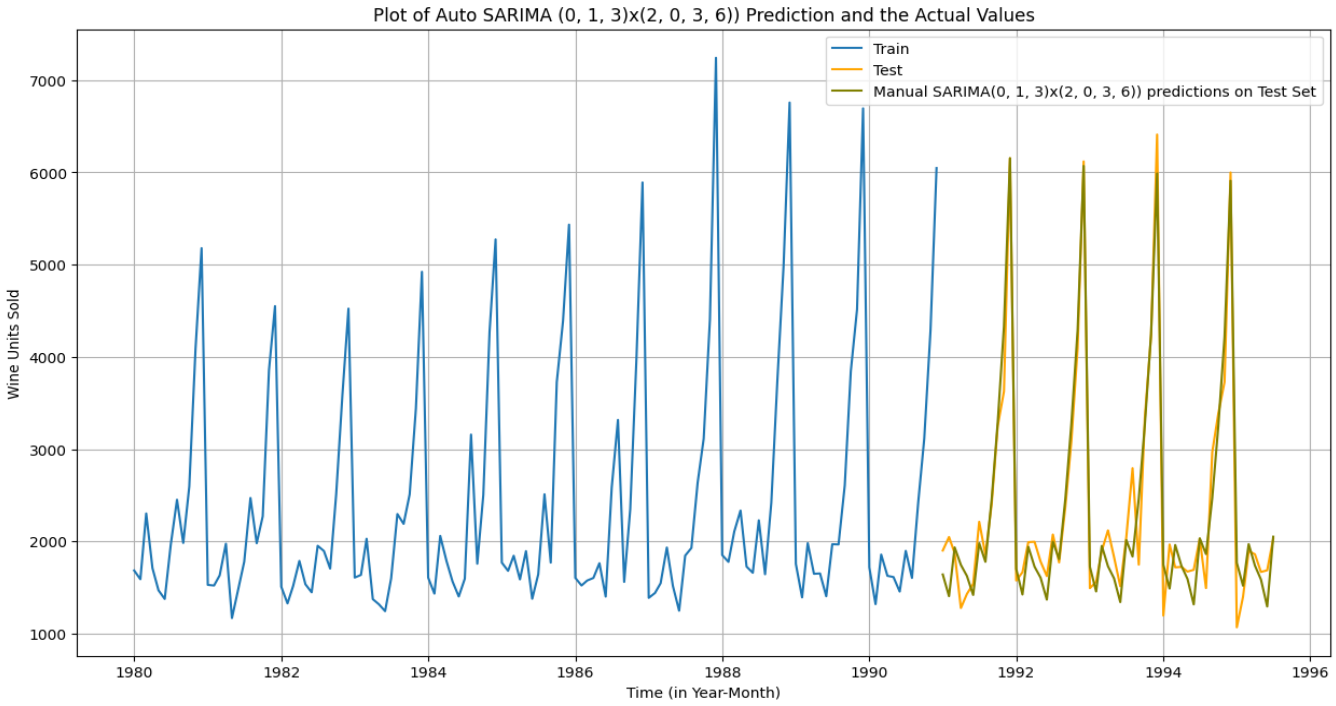
RMSE - 1299.97

Build the Auto SARIMA -



RMSE: 927.2277539348062

Manual SARIMA Model

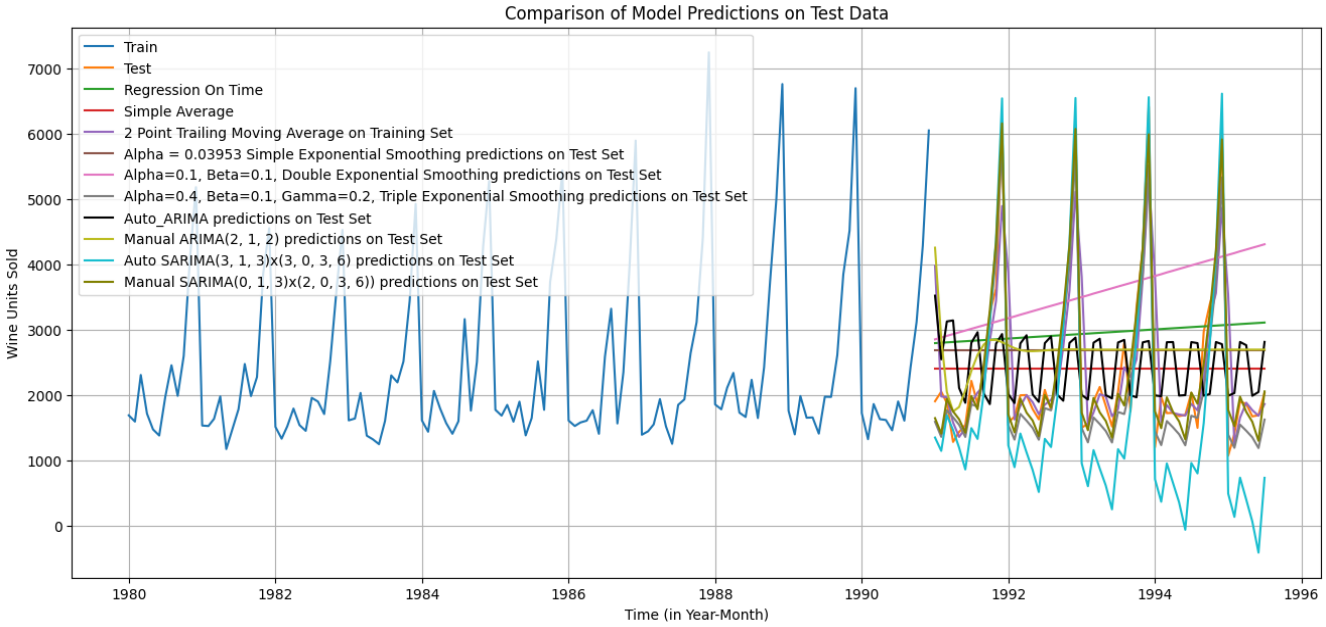


RMSE - 320.21988160804887

6 - Compare the performance of all the models built

Sort the Data Frame by RMSE in ascending order for ALL created Model.

	Test RMSE
α=0.4, β=0.1, γ=0.2,TripleExponentialSmoothing_ta_sm	317.434302
Manual_SARIMA(0, 1, 3)x(2, 0, 3, 6))	320.219882
α = 0.111272, β = 0.012360, γ = 0.460717 Tripple Exponential_auto_fit(Trend = Add, Seasonality = Add)	378.626241
2 Point Trailing Moving Average	813.400684
Auto_SARIMA(3, 1, 3)x(3, 0, 3, 6)	927.227754
4 Point Trailing Moving Average	1156.589694
Auto ARIMA	1229.272592
Simple Average Model	1275.081804
6 Point Trailing Moving Average	1283.927428
Manual ARIMA(2,1,2)	1299.979749
Alpha = 0.03953, Simple Exponential Smoothing	1304.927405
9 Point Trailing Moving Average	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
RegressionOnTime	1389.135175



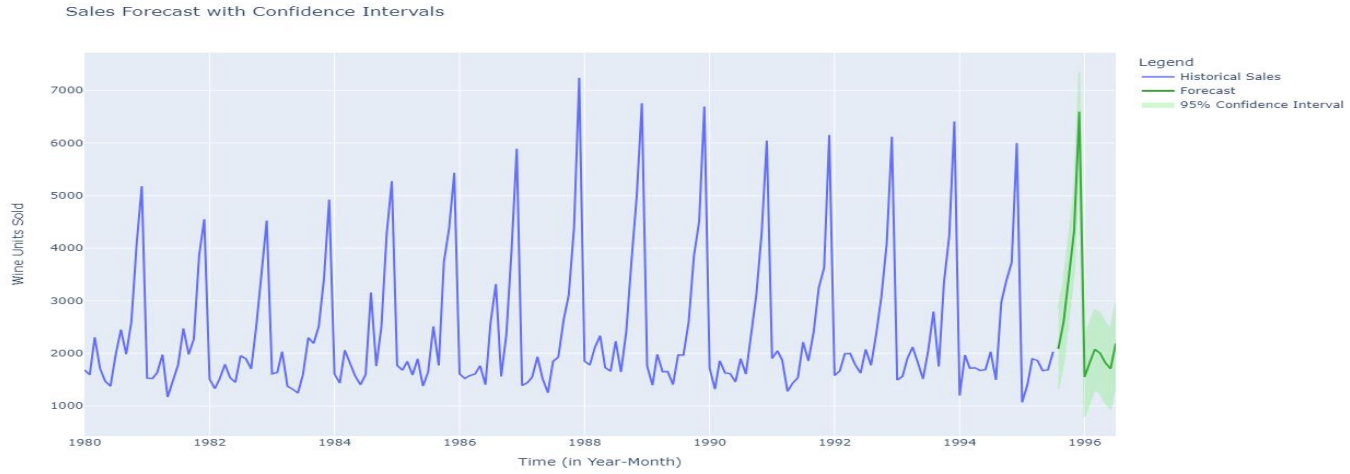
Analysis Summary

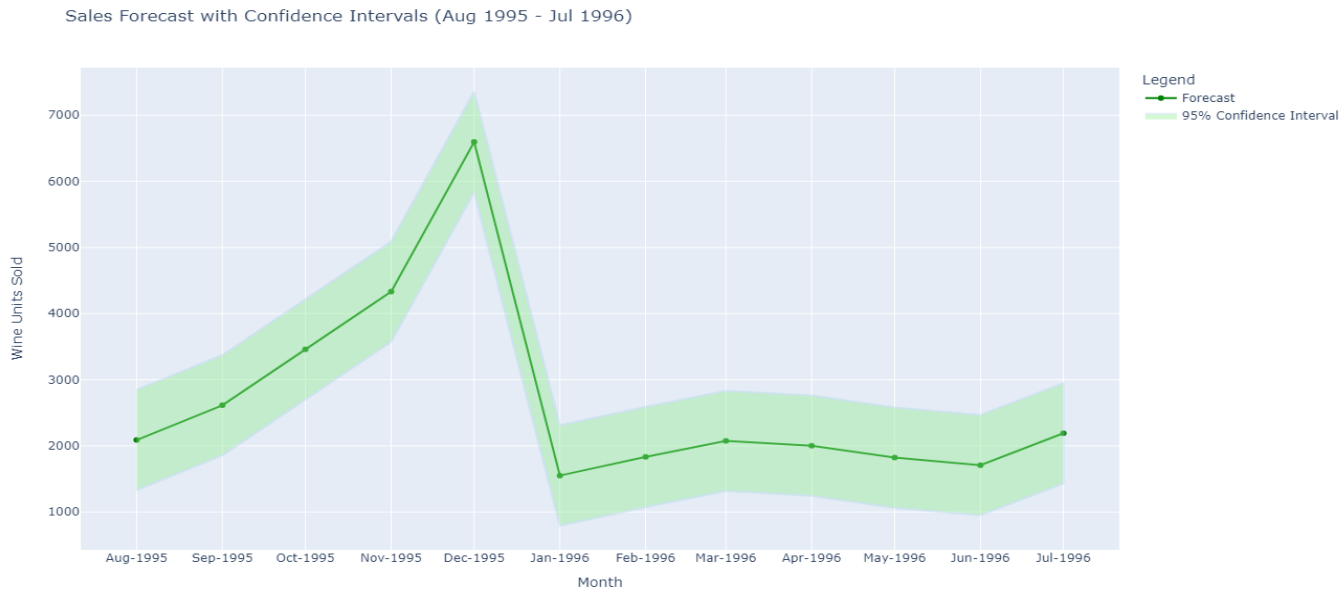
After evaluating various time series forecasting models, the model that exhibited the best performance is the Triple Exponential Smoothing model with parameters Alpha=0.4, Beta=0.1, Gamma=0.2. This model achieved the lowest Root Mean Square Error (RMSE) of 317.434302.

Best Model

The Triple Exponential Smoothing (Alpha=0.4, Beta=0.1, Gamma=0.2) model demonstrated superior accuracy in predicting the dataset, outperforming other models such as Manual ARIMA and various moving average techniques. The low RMSE value indicates that this model has the best fit, making it the most reliable choice for forecasting in this context.

Rebuild the Best Model using the Entire data - Make a Forecast for the Next 12 Months





7 - Actionable Insights & Recommendations

Key Findings

Pronounced Seasonality: Sales exhibit a strong seasonal pattern, with peak demand during the holiday season (November-December) and a subsequent decline in January.

Significant Sales Drop: A substantial decrease in forecasted sales is observed in January, necessitating further investigation.

Forecast Uncertainty: The varying width of confidence intervals indicates differing levels of forecast reliability across different periods.

Actionable Insights

- Leverage Seasonal Trends: Optimize inventory levels, staffing, and marketing efforts to align with seasonal demand fluctuations.
- Deep Dive into January Sales: Conduct a comprehensive analysis to identify the root causes of the January sales decline. Implement targeted strategies to mitigate this trend.
- Risk Management: Employ a risk-based approach by focusing on periods with wider confidence intervals for enhanced planning and resource allocation.

Recommendations

- Enhanced Forecasting: Incorporate additional variables such as economic indicators, competitor activities, and marketing campaign effectiveness into the forecasting model to improve accuracy.
- Customer Segmentation: Implement customer segmentation to tailor marketing efforts and product offerings to specific customer groups.

- Pricing Optimization: Conduct a rigorous pricing analysis to determine optimal price points for different product categories and customer segments.
- Continuous Monitoring: Establish a robust sales performance monitoring system to track key metrics and identify emerging trends.
- Diversification: Explore opportunities to diversify product offerings or target new customer segments to reduce reliance on seasonal peaks.

Conclusion

By carefully analyzing the sales forecast and identifying key trends, businesses can implement strategic initiatives to optimize operations, increase sales, and mitigate risks. A data-driven approach, coupled with a deep understanding of customer behavior and market dynamics, will be instrumental in achieving sustained growth and profitability.