ROSÉ WINE TIME SERIES ANALYSIS REPORT

DIVYA
BALASUBRAMANYAM

9/3/23

Time series Analysis

Table of Contents

List of Figures

List of Tables

INTRODUCTION

The data of Rose wine sales in the 20th century is to be analysed. As an analyst in the ABC Estate Wines, we will analyse and forecast Wine Sales in the 20th century. The purpose of this whole exercise is to analyse the wine sales, do the exploratory data analysis build Models using the dataset to forecast Wine Sales for the next 12 months.

Problem1: Rosé Wine Sales

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

Below is the 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

	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

Figure 1: Top 5 and bottom 5 records of Rosé Dataset

Shape of the Dataset:

There are 187 records with 2 columns in the dataset.

Information on the Dataset :

```
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    object
 1   Rose       185 non-null    float64
dtypes: float64(1), object(1)
```

Table 1: Data Information of Rosé Wine Dataset

There are 2 columns YearMonth and Rose in the dataset . YearMonth is an object type and Rose is integer type column that indicates the sales numbers of the wine . **There are 2 null values in the dataset .**

Since we are doing time series analysis , we will convert the column YearMonth to datetime column and make it an DatetimeIndex

```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
```

Table 2: Data Information after transformation of YearMonth Column

Time_Stamp	Rose
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

Table 3: Top 5 records of Rosé Dataset after transformation of YearMonth Column

Plotting the Time series data of Rosé Wine sales

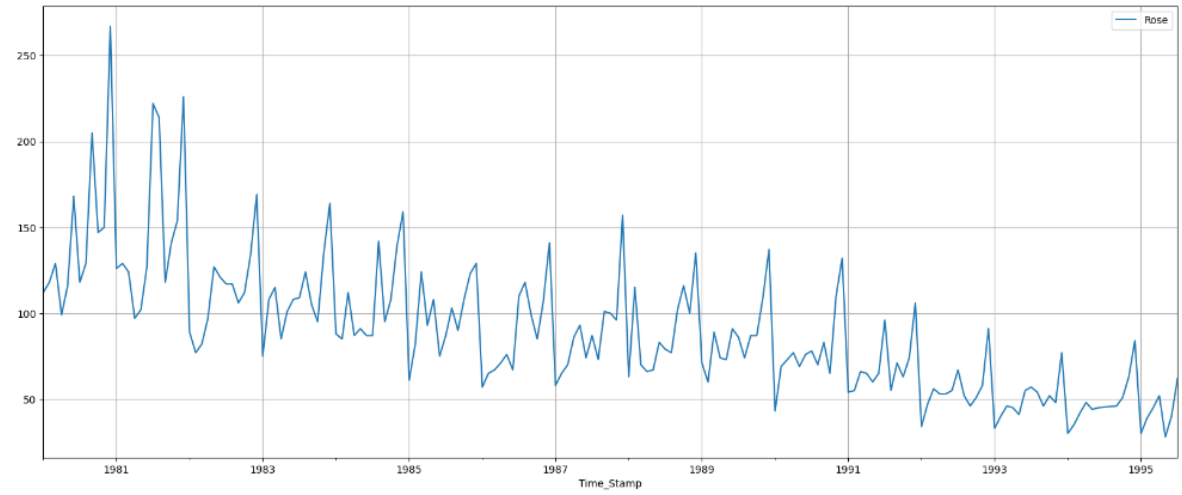


Figure 2: Time series plot of Rosé Wine Data

As we can see from the time series plot above , there is a trend in the sales and possibly some seasonality in the sales data

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

Before we perform EDA we have to impute the 2 missing values , because time series has to be continuous and there shouldn't be any missing values . The 2 records 1994-07 and 1994-08 have missing data, so can impute it using interpolate function.

After Imputing the values there are no null values in the dataset now, let's see the data info again

```
1994-07-01    45.333333
1994-08-01    45.666667
```

Table 4: Data Information after imputing Null values

The 2 values are now imputed .

```
Non-Null Count  Dtype
--------------  -----
187 non-null    float64
dtypes: float64(1)
```

Univariate Analysis

As there is only one numerical variable 'Rose' , let's see data description and boxplot of this variable

count	mean	std	min	25%	50%	75%	max
187	89.914439	39.238325	28	62.500000	85	111	267

Table 5: Data description of the dataset

Boxplot of the 'Rose' variable

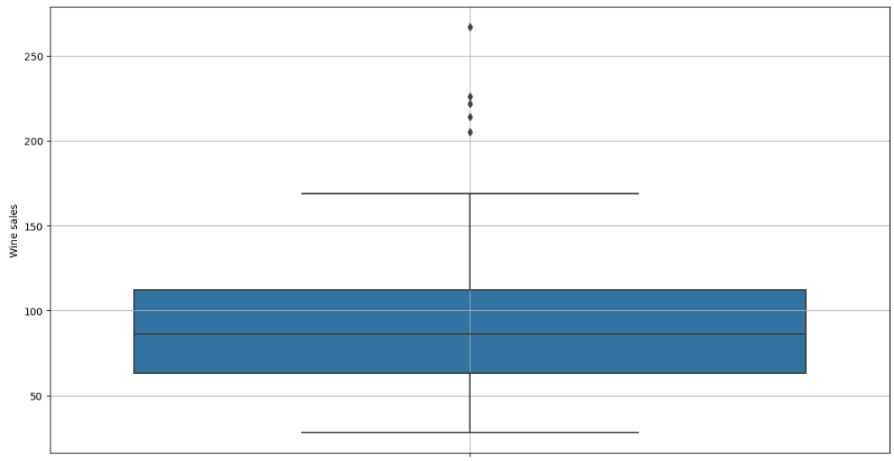


Figure 3:Boxplot of the Sales Variables

Boxplot of sales indicates it's a right skewed distribution

Mean of the data is 89.92 and Median is 32.23..

Minimum sales recorded for a month is 28.

Maximum sales recorded for a month is 267.

25 % of sales is below 62.5

50 % of sales is below 85

70 % of sales is below 111

Standard deviation of the data is 39.238325

There are outliers in the sales data.

Yearly Plot of wine sales

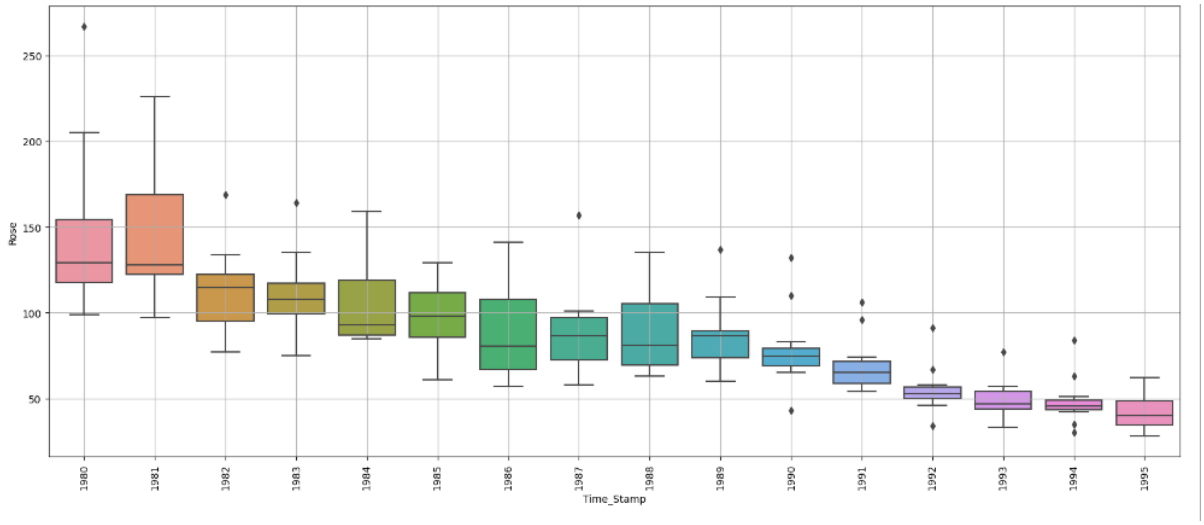


Figure 4: Yearly plot of Sales

As we can see from the yearly plot year 1981 had the highest average sales of 148.33 and year 1995 had the lowest average sales of 42.29.

Monthly Plot of wine sales

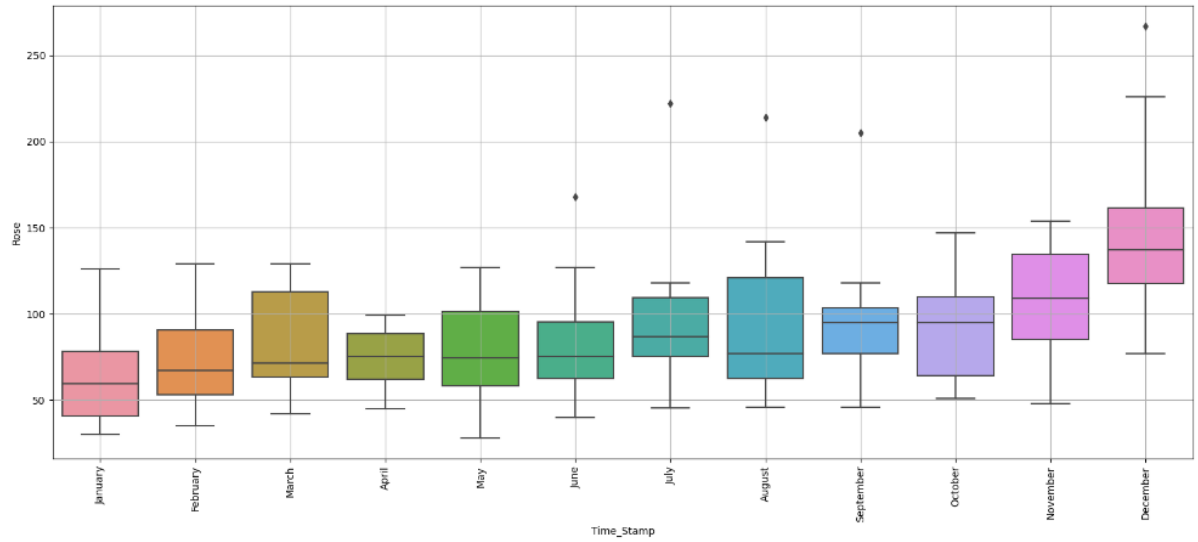


Figure 5: Monthly plot of Sales

As we can see from the Monthly plot December month had the highest average sales year on year and January had the lowest average sales

Month plot of sales

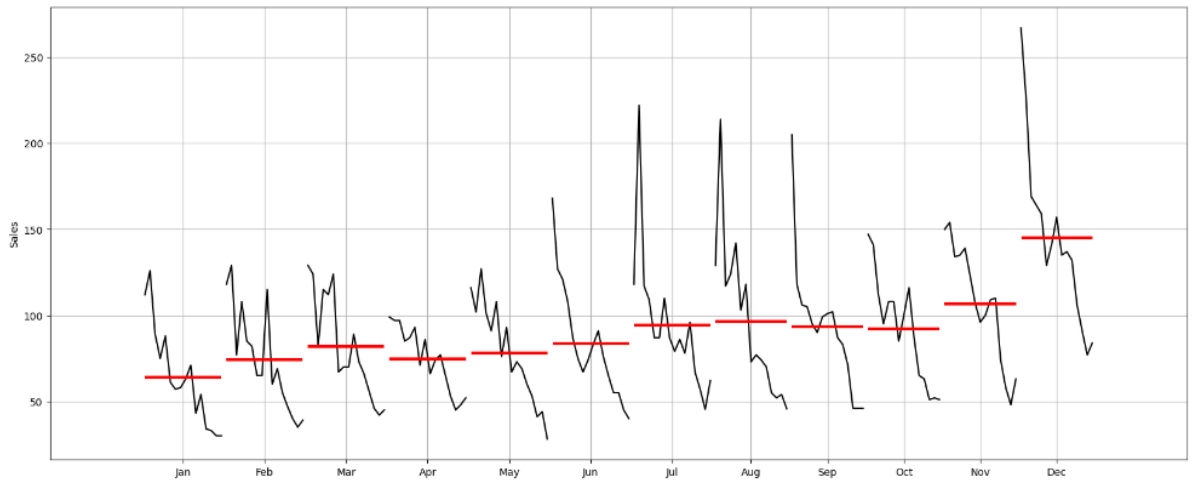


Figure 6:Month Plot of sales

This plot shows us the behaviour of the Time Series ('Wine Sales' in this case) across various months. The red line is the median value. December has the highest median value and January the least

Bivariate Analysis

Let's see a how the yearly sales have been across all the months

Time_Stamp	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Time_Stamp																
January	112.0	126.0	89.0	75.0	88.0	61.0	57.0	58.0	63.0	71.0	43.0	54.0	34.0	33.0	30.000000	30.0
February	118.0	129.0	77.0	108.0	85.0	82.0	65.0	65.0	115.0	60.0	69.0	55.0	47.0	40.0	35.000000	39.0
March	129.0	124.0	82.0	115.0	112.0	124.0	67.0	70.0	70.0	89.0	73.0	66.0	56.0	46.0	42.000000	45.0
April	99.0	97.0	97.0	85.0	87.0	93.0	71.0	86.0	66.0	74.0	77.0	65.0	53.0	45.0	48.000000	52.0
May	116.0	102.0	127.0	101.0	91.0	108.0	76.0	93.0	67.0	73.0	69.0	60.0	53.0	41.0	44.000000	28.0
June	168.0	127.0	121.0	108.0	87.0	75.0	67.0	74.0	83.0	91.0	76.0	65.0	55.0	55.0	45.000000	40.0
July	118.0	222.0	117.0	109.0	87.0	87.0	110.0	87.0	79.0	86.0	78.0	96.0	67.0	57.0	45.333333	62.0
August	129.0	214.0	117.0	124.0	142.0	103.0	118.0	73.0	77.0	74.0	70.0	55.0	52.0	54.0	45.666667	NaN
September	205.0	118.0	106.0	105.0	95.0	90.0	99.0	101.0	102.0	87.0	83.0	71.0	46.0	46.0	46.000000	NaN
October	147.0	141.0	112.0	95.0	108.0	108.0	85.0	100.0	116.0	87.0	65.0	63.0	51.0	52.0	51.000000	NaN
November	150.0	154.0	134.0	135.0	139.0	123.0	107.0	96.0	100.0	109.0	110.0	74.0	58.0	48.0	63.000000	NaN
December	267.0	226.0	169.0	164.0	159.0	129.0	141.0	157.0	135.0	137.0	132.0	106.0	91.0	77.0	84.000000	NaN

Table 6: Tabular column of yearly sales across all months

As we can see from the tabular column above December month clocks highest sales across all . Also, we should note that that sales data is unavailable for months of August, September, October , November, December for the year 1995

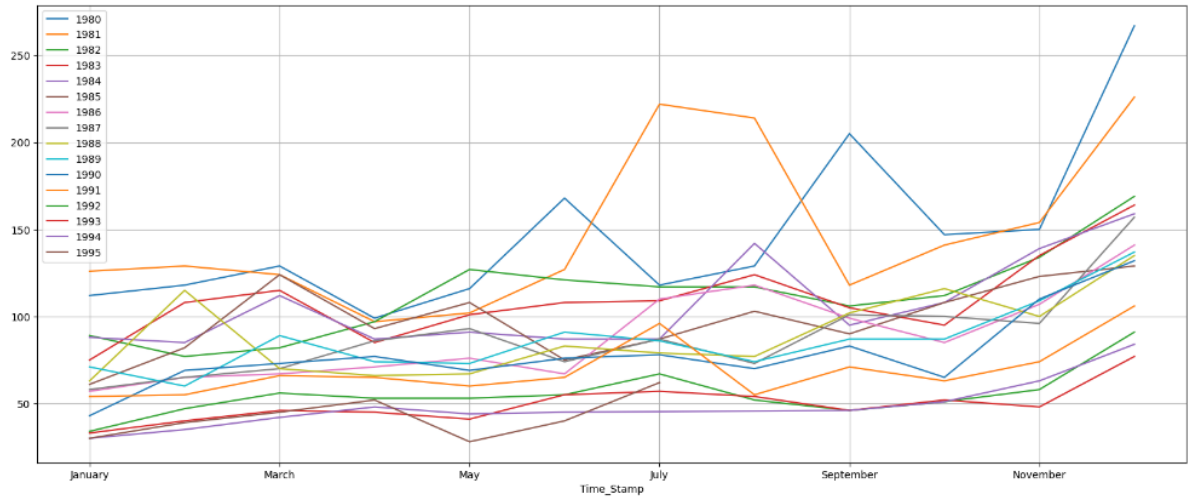


Figure 7: Months Vs Sales across all years

After plotting the sales vs month across the years, we can note that December had the maximum sales across all years

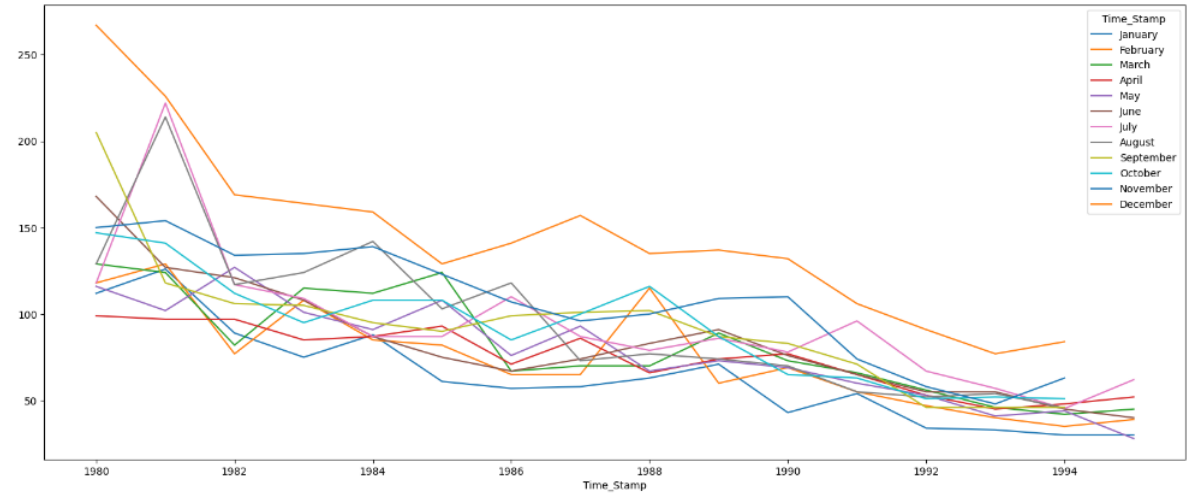


Figure 8: Years Vs Sales across all months

After plotting the Sales vs Years across the months , we can see that sales have been declining year on year considering the fact that data is unavailable for months of August, September, October , November, December for the year 1995.

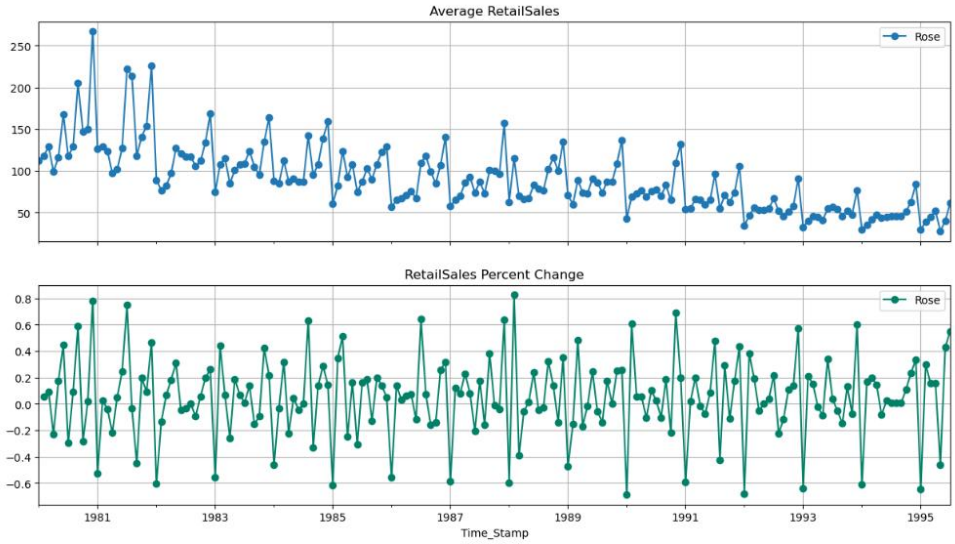


Figure 9 Average sales across years and percentage change is sales .

From the Average sales across years and percentage change sales plot , we can see that the mean is changing across the years and variance is not changing

Decomposition of the Time series data

We decompose the time series

- To understand revenue generation without the quarterly effects
- De-seasonalize the series
- Estimate and adjust by seasonality
- Compare the long-term movement of the series (Trend) vis-a-vis short-term movement (seasonality) to understand which has the higher influence

Decomposition Model can be Additive or Multiplicative

Additive model: Observation = Trend + Seasonality + Error

$$Yt = Tt + St + It$$

Multiplicative model: Observation = Trend * Seasonality * Error

$$Yt = Tt * St * It$$

Yt: time series value (actual data) at period t.

St: seasonal component (index) at period t.

Tt: trend cycle component at period t.

It: irregular (remainder) component at period

Let's decompose the data and check the trend, seasonality and the irregular/residual/error component.

Additive Decomposition

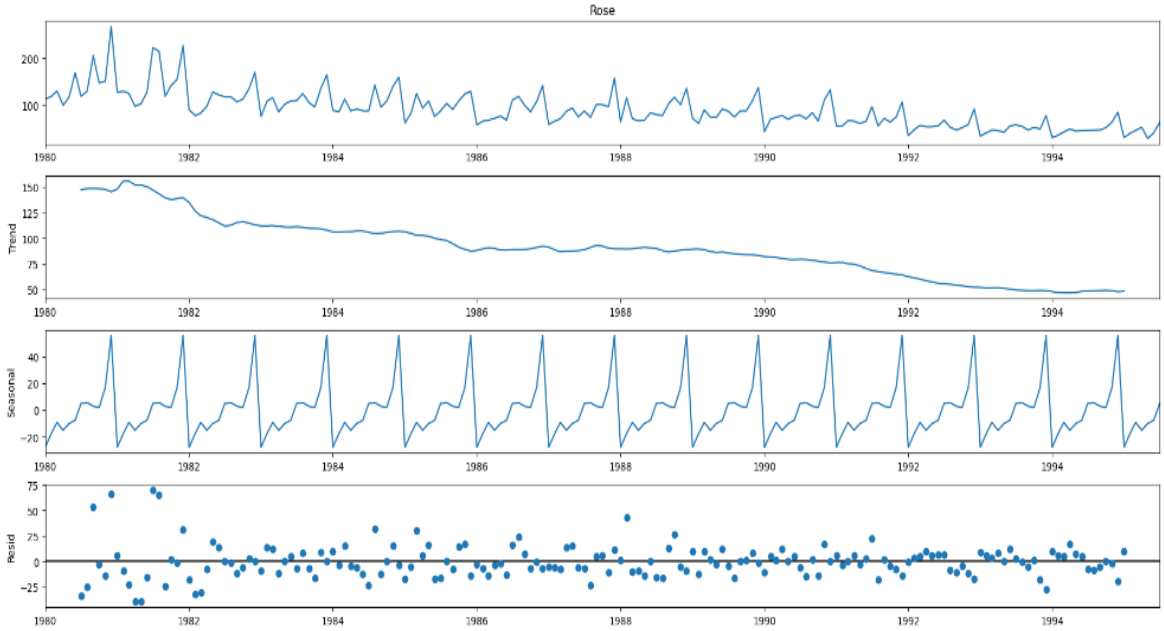


Figure 10: Additive decomposition of Time Series Data

Above plot is break-up of time series plot into Trend, seasonality and Resdiuals using Addtitve model

Trend		Seasonality		Residual	
Time_Stamp		Time_Stamp		Time_Stamp	
1980-01-01	NaN	1980-01-01	-27.91	1980-01-01	NaN
1980-02-01	NaN	1980-02-01	-17.44	1980-02-01	NaN
1980-03-01	NaN	1980-03-01	-9.29	1980-03-01	NaN
1980-04-01	NaN	1980-04-01	-15.10	1980-04-01	NaN
1980-05-01	NaN	1980-05-01	-10.20	1980-05-01	NaN
1980-06-01	NaN	1980-06-01	-7.68	1980-06-01	NaN
1980-07-01	147.08	1980-07-01	4.90	1980-07-01	-33.98
1980-08-01	148.12	1980-08-01	5.50	1980-08-01	-24.62
1980-09-01	148.38	1980-09-01	2.77	1980-09-01	53.85
1980-10-01	148.08	1980-10-01	1.87	1980-10-01	-2.96

Above numbers is break-up of sales into Trend, seasonality and Resdiuals using Addtitve model. Time series value at period t can be obtained by adding the Trend, Seasonality and Residual Data

As we can see from the Model there is a visible trend in the data and there is a seasonality component . The residual component doesn't seem to have a pattern

Multiplicative Decomposition

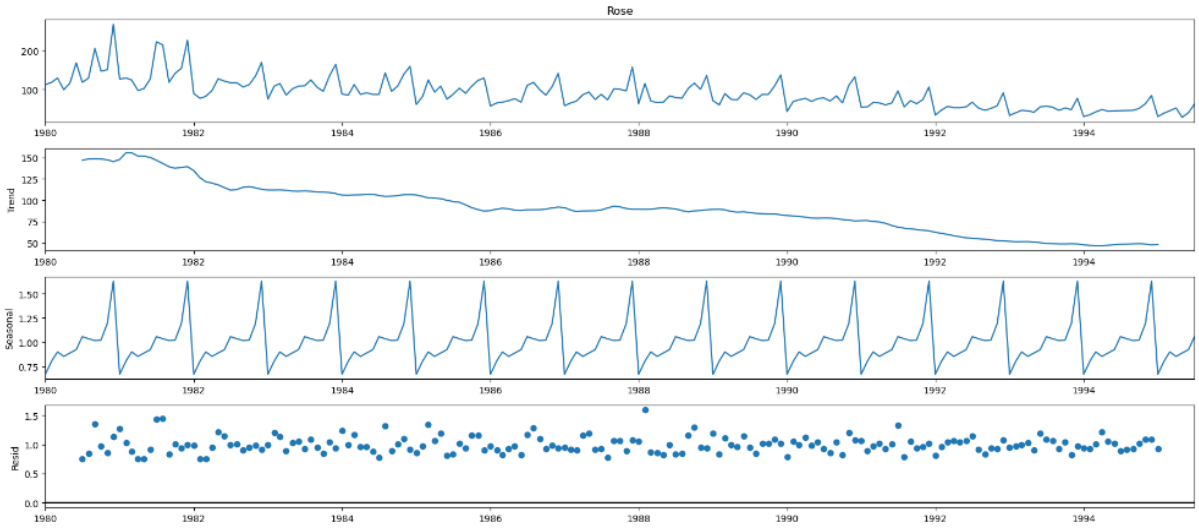


Figure 11: Multiplicative decomposition of Time Series Data

Above plot is break-up of data into Trend, seasonality and Resdiuals using Multiplcative model

Trend		Seasonality		Residual	
Time_Stamp		Time_Stamp		Time_Stamp	
1980-01-01	NaN	1980-01-01	0.67	1980-01-01	NaN
1980-02-01	NaN	1980-02-01	0.81	1980-02-01	NaN
1980-03-01	NaN	1980-03-01	0.90	1980-03-01	NaN
1980-04-01	NaN	1980-04-01	0.85	1980-04-01	NaN
1980-05-01	NaN	1980-05-01	0.89	1980-05-01	NaN
1980-06-01	NaN	1980-06-01	0.92	1980-06-01	NaN
1980-07-01	147.08	1980-07-01	1.06	1980-07-01	0.76
1980-08-01	148.12	1980-08-01	1.04	1980-08-01	0.84
1980-09-01	148.38	1980-09-01	1.02	1980-09-01	1.36
1980-10-01	148.08	1980-10-01	1.02	1980-10-01	0.97

Above data is break-up of data into Trend, seasonality and Resdiuals using Multiplcative model. Time series value at period t can be obtained by multiplying the Trend, Seasonality and Residual Data .

As we can see from the Model there is a visible trend in the data and there is a seasonality component . The residual component doesn't seem to have a pattern

Additive Model seems to be giving a better prediction of Time series value at period t.

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

The dataset is split into training and testing data with testing data starting from Year 1991.

After the train and test split of 187 records, there are 132 records in the training dataset and

55 records in the testing dataset .

Training Data is used to train (develop) the model. Training Data is used to identify a few working models. The forecasts for training data are called fitted values. Each of the models is tested against the observed values of the series for hold-out period.

The model is selected to be the best were observed and forecasted values are the closest.

Predictive power of a model is estimated by comparing its forecasting performance on a Test Data

Let's see a sample of Training and Testing Dataset

Training dataset is ending at 1990 December Let's see a sample of Training and Testing Dataset

```
First few rows of Training Data
 Time_Stamp
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
Freq: MS, Name: Rose, dtype: float64

Last few rows of Training Data
 Time_Stamp
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
Freq: MS, Name: Rose, dtype: float64
```

Table 7: Training Dataset

```
First few rows of Test Data
 Time_Stamp
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
Freq: MS, Name: Rose, dtype: float64

Last few rows of Test Data
 Time_Stamp
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
Freq: MS, Name: Rose, dtype: float64
```

Table 8: Testing Dataset

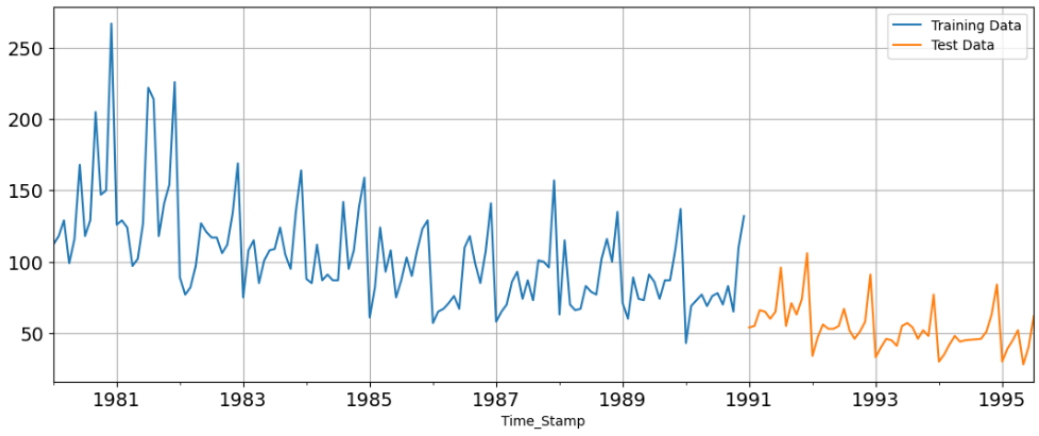


Figure 12: Time series Data split into Train and Test Data

As we can see from the plot above , the training data is marked in blue and testing data is marked in orange starts from 1991 and goes on till the end of the timeseries dataset

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

Exponential Smoothing Models

Exponential Smoothing Models take weighted averages of past observations **,w**eights decay as observations gets older**.** One or more parameters control how fast the weights decay **.**These parameters have values between 0 and 1.

Simple Exponential Smoothing (SES)

SES model is used If the time series neither has a pronounced trend nor seasonality:

Performance of the smoothing parameter α controls performance of the method. If α is closer to 1, forecasts follow the actual observations more closely. If α is closer to 0, forecasts are farther from the actual observations and the line is smooth .

The SES model gives the following parameters

	name	param	optimized
smoothing_level	alpha	0.098749	True
initial_level	l.0	134.387202	True

Table 9:SES Parameters

The smoothing level α is 0.0987 which is close to 0 , hence forecasts are farther from the actual observations and the line is smooth .

The following values are predictions on the top 5 records in the test dataset

1991-01-01	87.104983
1991-02-01	87.104983
1991-03-01	87.104983
1991-04-01	87.104983
1991-05-01	87.104983

Table 10: Test Data predictions using SES Model

As we can see from the above table , the predicted value is same for all the data points

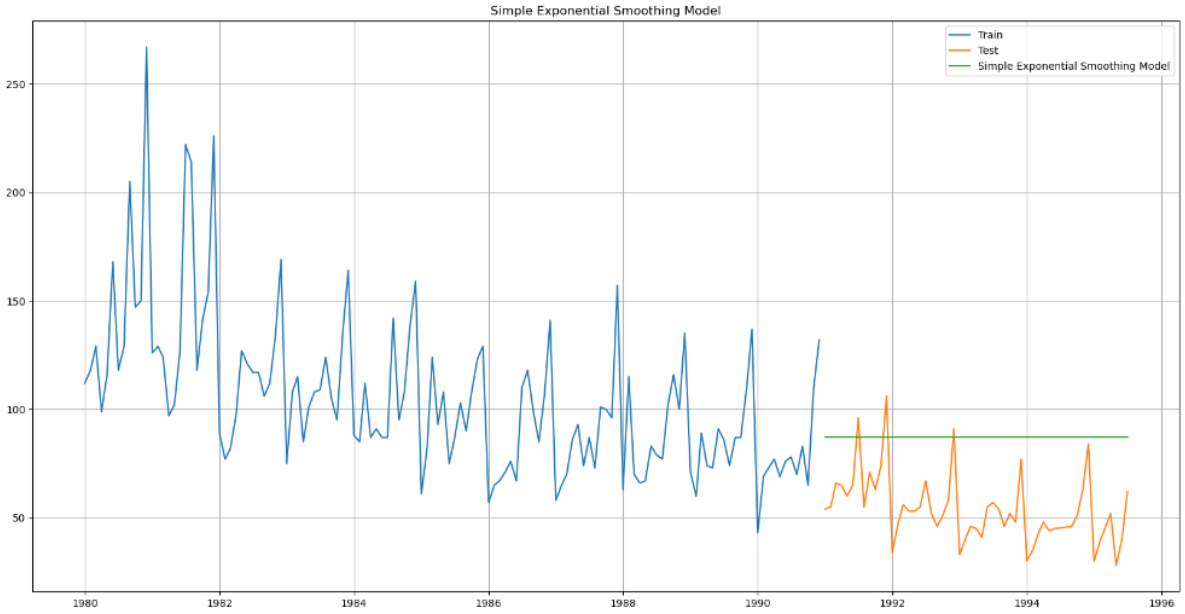


Figure 13: Time series plot with Testing and Training Data. The green line is the SES prediction data values

As we can see from the plot above , the SES Model represented by the green line is not good at predicting the test data values . It's a smooth line with a constant value

Let's evaluate the model using RMSE

Test RMSE score for SES Model is 36.79

	Test RMSE
Alpha=0.098749: SimpleExponentialSmoothing	36.796227

Table 11: RMSE score for SES Model

Simple Exponential Smoothing with Iteration (SES)

Using Iterative Method to find the best values for smoothing parameter α and we get following values for Test RMSE for different α values

	Alpha Values	Test RMSE
6	0.07	36.435772
7	0.08	36.462965
5	0.06	36.580469
8	0.09	36.604118
9	0.10	36.828033
...	...	...
94	0.95	78.532696
95	0.96	78.786884
96	0.97	79.032686
97	0.98	79.270003
98	0.99	79.498734

Table 12: SES Parameters after iteration

With α value 0.07 , so we can build SES model with α 0.07

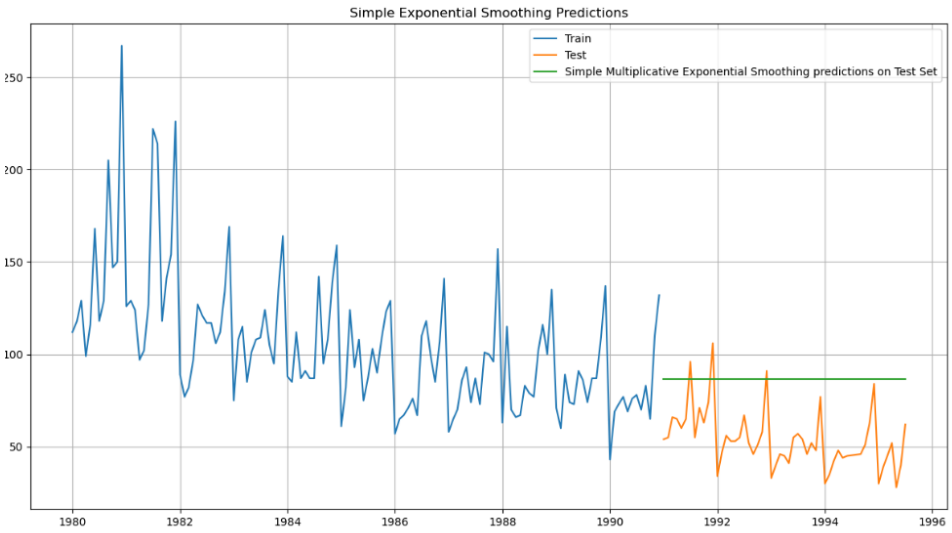


Figure 14: Time series plot with Testing and Training Data.

The green line is the SES prediction data values

As we can see from the above table , the predicted value is same for all the data points

Let's evaluate the model using RMSE

Test RMSE score for SES Model is 36.79

RMSE values is 36.43

Double Exponential Smoothing (DES)

- DES is applicable when data has Trend but no seasonality .It's an extension of SES
- 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
- Also known as Holt mode

The DES model gives the following parameters :

	name	param	optimized
smoothing_level	alpha	0.017550	True
smoothing_trend	beta	0.000032	True
initial_level	l.0	138.820815	True
initial_trend	b.0	-0.492580	True

Table 13: DES Parameters

The smoothing level α is 0.017550 which is not very significant and β is 0.000032 .

Following table shows the predictions on testing dataset

1991-01-01	73.259732
1991-02-01	72.767150
1991-03-01	72.274569
1991-04-01	71.781987
1991-05-01	71.289405

Table 14: Test Data predictions using DES Model

As we can see from the above table , the predicted value is not the same for all the data points as in SES model

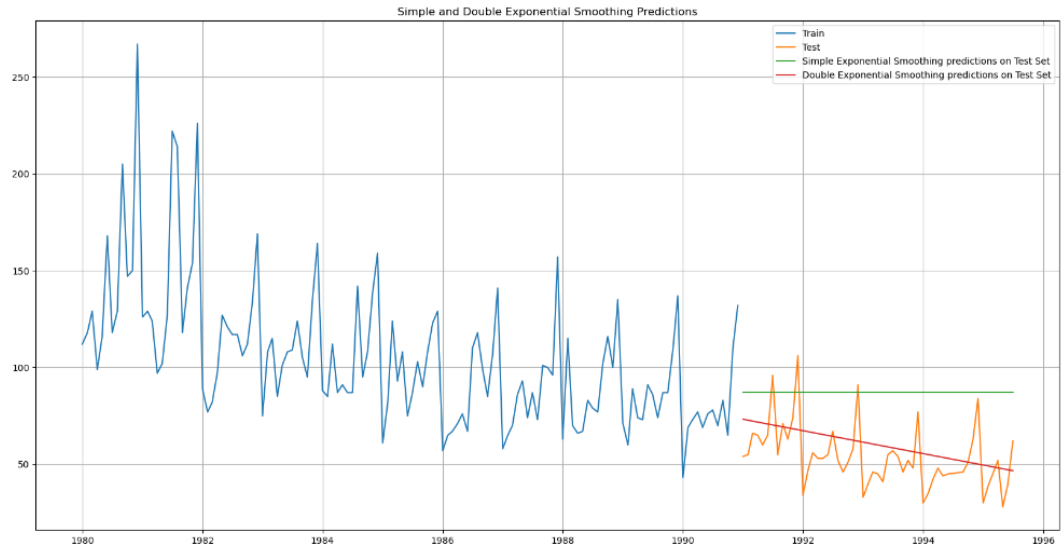


Figure 15: Time series plot with Training, Testing, SES and DES Model

As we can see from the plot above , the DES Model represented by the red line is also not good at predicting the test data values .

Let's evaluate the model using RMSE

The Test RMSE score for DES model is 15.707

	Test RMSE
Alpha=0.098749:SimpleExponentialSmoothing	36.796227
Alpha=0.017550,Beta=0.000032:DoubleExponentialSmoothing	15.707052

Table 15:RMSE score for SES and DES Models

Double Exponential Smoothing with Iteration (DES)

Using Iterative Method to find the best values for smoothing parameter α and we get following values for

Test RMSE for different α and β values

	Alpha Values	Beta Values	Test RMSE
343	0.04	0.47	14.560058
222	0.03	0.25	14.683478
262	0.03	0.65	14.714918
300	0.04	0.04	14.895847
342	0.04	0.46	14.907935
...	...	...	...
6731	0.68	0.99	1118.328808
7127	0.72	0.99	1118.464278
6830	0.69	0.99	1119.095888
7028	0.71	0.99	1119.179428
6929	0.70	0.99	1119.384014

Table 16: DES Parameters after iteration

With α value 0.04 and β values 0.47 we can build DES model

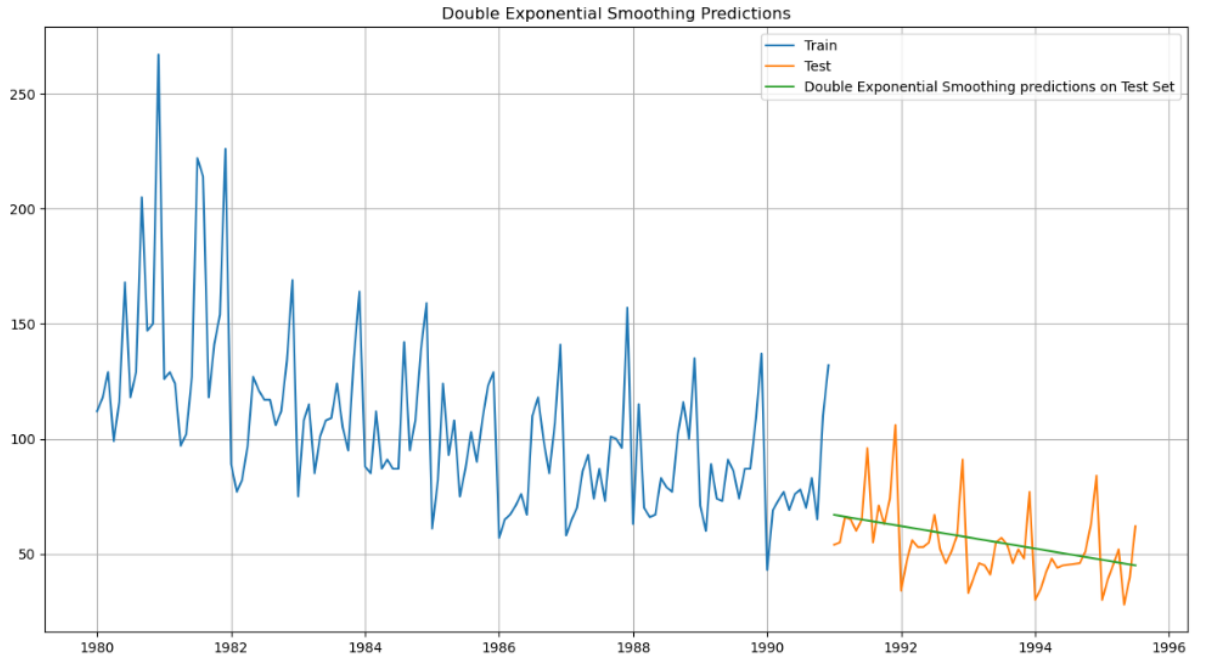


Figure 16: Time series plot with Testing and Training Data and DES prediction data values

As we can see from the plot above , the DES Model represented by the red line is also not good at predicting the test data values .

Let's evaluate the model using RMSE

The Test RMSE score for DES model is 14.56

Triple Exponential Smoothing (TES) or Holt-Winters' Model

- TES is applicable when data has Level, Trend and Seasonality, It's an extension of DES
- Three separate components are considered: Level , Trend and Seasonality
- Because Seasonality can be additive or multiplicative, TES model can be additive or multiplicative
- Simultaneously smooths the level, trend and seasonality

Three separate smoothing parameters

α: Smooths level; $0 < \alpha < 1$

β: Smooths trend; $0 < \beta < 1$

γ : Smooths seasonality; $0 < \gamma < 1$

TES Additive Model

The TES Additive model gives the following parameters

	name	param	optimized
smoothing_level	alpha	0.089541	True
smoothing_trend	beta	0.000240	True
smoothing_seasonal	gamma	0.003467	True
initial_level	l.0	146.557016	True
initial_trend	b.0	-0.547197	True
initial_seasons.0	s.0	-31.174785	True
initial_seasons.1	s.1	-18.748399	True
initial_seasons.2	s.2	-10.769618	True
initial_seasons.3	s.3	-21.367410	True
initial_seasons.4	s.4	-12.637755	True
initial_seasons.5	s.5	-7.274303	True
initial_seasons.6	s.6	2.612798	True
initial_seasons.7	s.7	8.696036	True
initial_seasons.8	s.8	4.793811	True
initial_seasons.9	s.9	2.961101	True
initial_seasons.10	s.10	21.057388	True
initial_seasons.11	s.11	63.182799	True

Table 17:TES Parameters

The smoothing level α is 0.089541 , Smoothing trend β is 0.000240 and smoothing seasonal γ is 0.003467.

Following table shows the predictions on testing dataset

1991-01-01	42.684928
1991-02-01	54.564005
1991-03-01	61.995209
1991-04-01	50.852018
1991-05-01	59.034271

Table 18: Test Data predictions using TES Additive Model

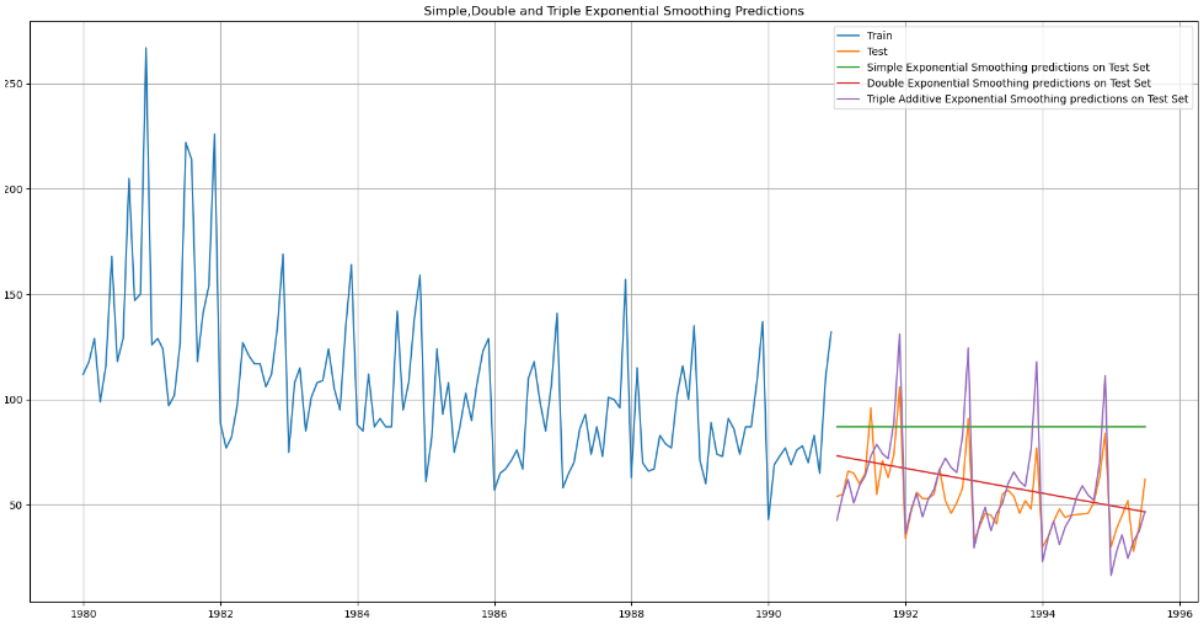


Figure 17: Time series plot with Training, Testing, SES,DES and TES Additive Model

As we can see from the plot above , the TES Additive Model represented by the purple line is good at predicting the test data values as it is following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for TES Additive model is 14.24

TES Multiplicative Model

The TES Multiplicative model gives the following parameters

	name	param	optimized
smoothing_level	alpha	0.071511	True
smoothing_trend	beta	0.045292	True
smoothing_seasonal	gamma	0.000072	True
initial_level	l.0	130.408391	True
initial_trend	b.0	-0.779857	True
initial_seasons.0	s.0	0.862190	True
initial_seasons.1	s.1	0.977675	True
initial_seasons.2	s.2	1.068773	True
initial_seasons.3	s.3	0.934039	True
initial_seasons.4	s.4	1.050625	True
initial_seasons.5	s.5	1.144110	True
initial_seasons.6	s.6	1.258369	True
initial_seasons.7	s.7	1.339378	True
initial_seasons.8	s.8	1.267788	True
initial_seasons.9	s.9	1.241313	True
initial_seasons.10	s.10	1.447246	True
initial_seasons.11	s.11	1.995537	True

Table 19: TES multiplicative model Parameters

The smoothing level α is 0.071511 , Smoothing trend β is 0.0452 and smoothing seasonal γ is 0.000072.

Following table shows the predictions on testing dataset

1991-01-01	56.321655
1991-02-01	63.664690
1991-03-01	69.374024
1991-04-01	60.435528
1991-05-01	67.758341

Table 20: Test Data predictions using TES Multiplicative Model

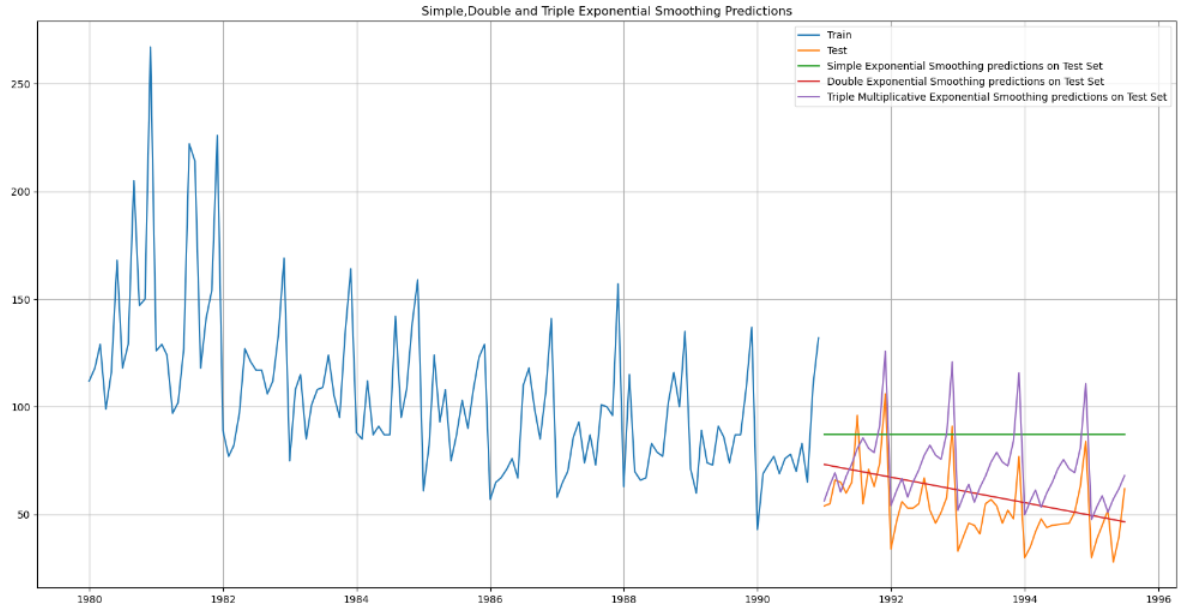


Figure 18: TES Multiplicative Model

As we can see from the plot above , the TES Multiplicative Model represented by the purple line is good at predicting the test data values as it is following the test data variations .

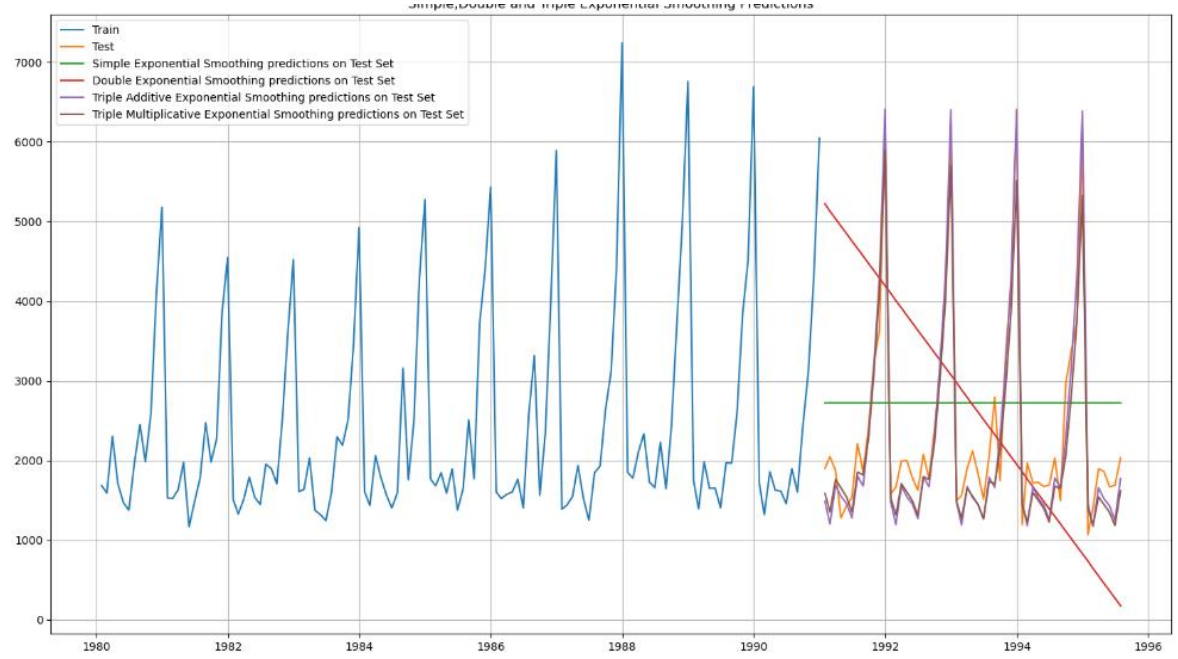


Figure 19: Time series plot with Training, Testing, SES,DES and TES Additive and Multiplicative Model

As we can see from the plot above , the TES Multiplicative Model represented by the brown line is good at predicting the test data values as it is following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for TES Multiplicative model is 20.15

	Test RMSE
Alpha=0.098749:SimpleExponentialSmoothing	36.796227
Alpha=0.017550,Beta=0.000032:DoubleExponentialSmoothing	15.707052
Alpha=0.071511,Beta=0.045292,Gamma=0.000072:TripleExponentialSmoothingMultiplicative	20.156763
Alpha=0.089541,Beta=0.000240,Gamma=0.003467:TripleExponentialSmoothingAdditive	14.249661
Alpha==0.07:SimpleExponentialSmoothing	36.435772
Alpha==0.04,,Beta=0.47:DoubleExponentialSmoothing	14.560058

Table 21: Test RMSE for various Exponential smoothing Models

Linear Regression Model(LR)

For this particular linear regression, we are going to regress the 'Rose' 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, 3
4, 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, 18
3, 184, 185, 186, 187]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

Let's see the first and last few rows of training and test data

```
First few rows of Training Data
```

Time_Stamp	Rose	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
```

Time_Stamp	Rose	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

Table 22: Sample of Training Data for LR model

```
First few rows of Test Data
```

Time_Stamp	Rose	time
1991-01-01	54.0	131
1991-02-01	55.0	132
1991-03-01	66.0	133
1991-04-01	65.0	134
1991-05-01	60.0	135

```
Last few rows of Test Data
```

Time_Stamp	Rose	time
1995-03-01	45.0	181
1995-04-01	52.0	182
1995-05-01	28.0	183
1995-06-01	40.0	184
1995-07-01	62.0	185

Table 23: Sample of Testing Data for LR model

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 and plot the time series data with predictions using LR Model alongside Training and Testing Data .

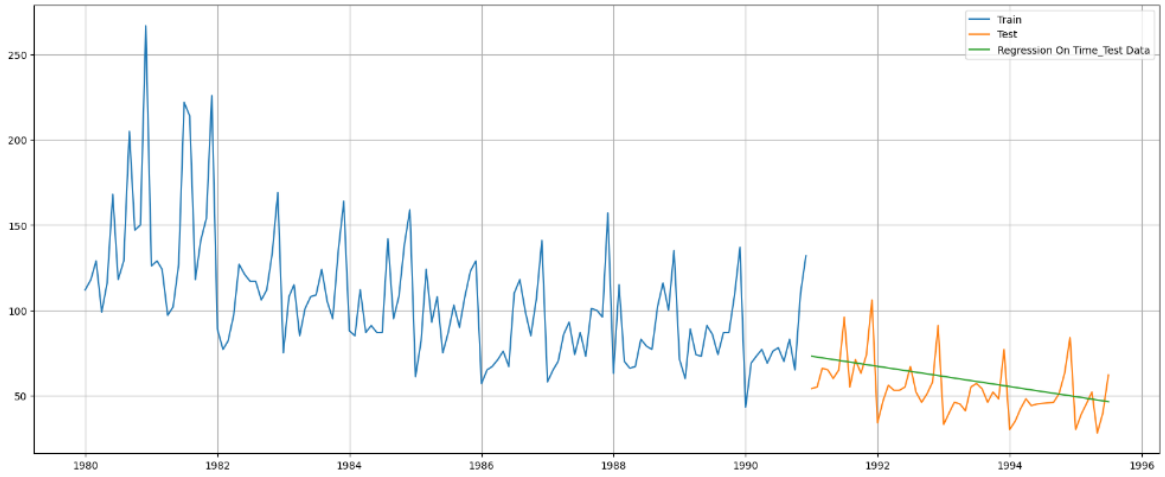


Figure 20: Time series plot with Training, Testing and Linear Regression Model

As we can see from the plot above , the Linear regression Model represented by the green line is not good at predicting the test data values as it is not following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for Linear Regression model is 15.269

Naïve Model : y^t+1=yt+1

For this particular Naïve model, we say that the prediction for next month is the same as current month and the prediction for month after next is same as prediction for next month and since the prediction of next month is same as current month, therefore the prediction for month after next is also same current month.

Since Naïve model prediction for next month is the same current month, the prediction for first record in test data is same as last record of training data , so let's see the last record of training data and the predictions on test data

	Rose
Time_Stamp	
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

Table 24: Last 5 records of training data

Using Naïve Approach to build the model on the training data and test the model on the test data we get the following predictions. As we can see last record of training data has value 132, therefore the test data will have the same value for all records

```
Time_Stamp
1991-01-01    132.0
1991-02-01    132.0
1991-03-01    132.0
1991-04-01    132.0
1991-05-01    132.0
```

Table 25: First 5 records of predicted test data

Let's plot the time series data with predictions using Naïve Model alongside Training and Testing Data

.

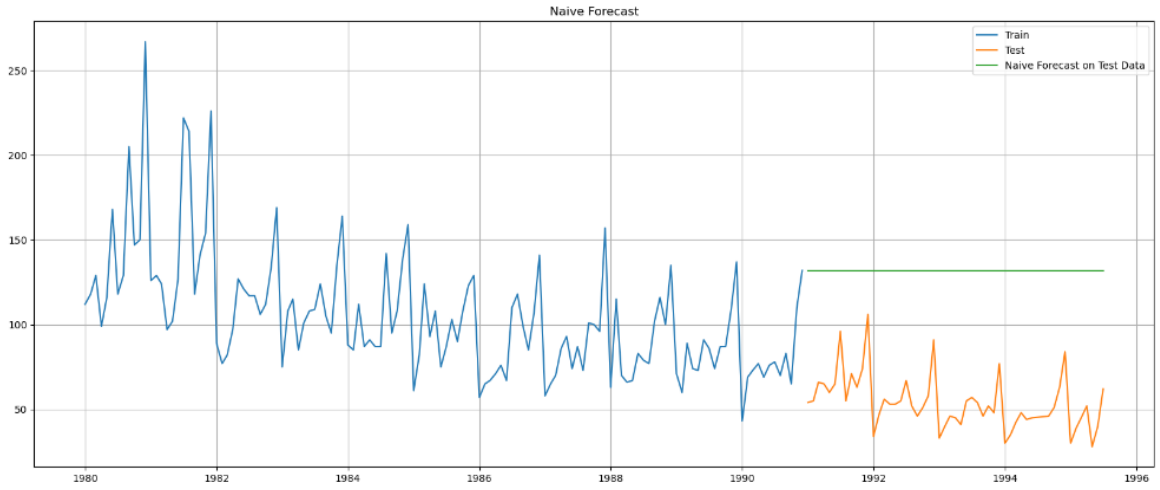


Figure 21: Time series plot with Training, Testing and Naïve Model

As we can see from the plot above , the Linear regression Model represented by the green line is not good at predicting the test data values as it is not following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for Naive model is 79.719

Simple Average Model

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

Time_Stamp	Rose	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

Table 26:Mean Forecast for simple average model against the actual values of test data

Let's plot the time series data with predictions using Simple Average Model alongside Training and Testing Data

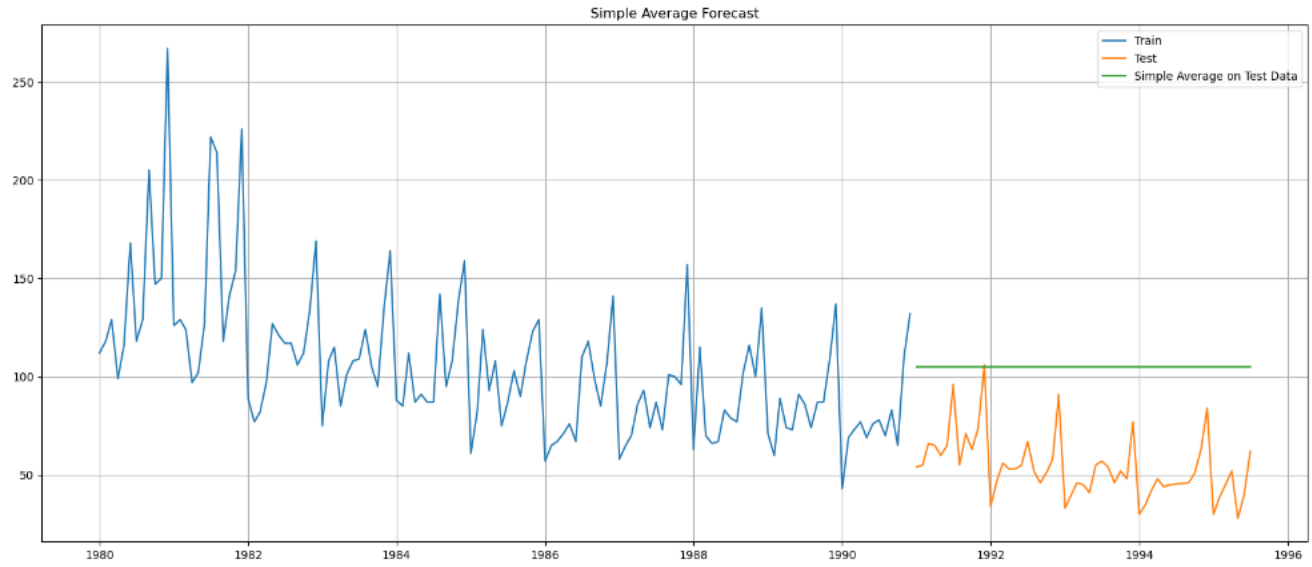


Figure 22: Time series plot with Training, Testing and Simple Average Model

As we can see from the plot above , the Simple Average Model represented by the green line is not good at predicting the test data values as it is not following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for Simple Average model is 53.461

Moving Average Model(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. For Moving Average, we are going to average over the entire data.

Let's compute moving averages with intervals 2,4,6,9 on the entire dataset:

First 10 records of the dataset with moving average at different intervals

Time_Stamp	Rose	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

Figure 23: 2,4,6,9 point Moving Average

Plotting the Moving average vis-a-vis training data

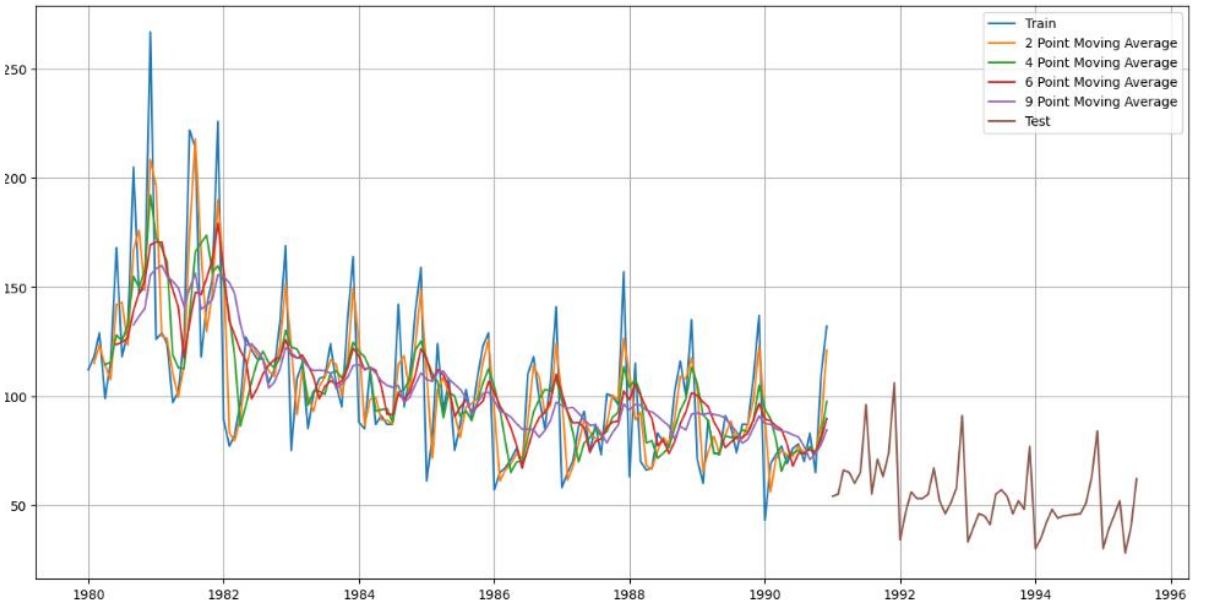


Figure 24: Time series plot with Training Dataset and Moving Average Model with different intervals on training dataset

Let's apply the each of the moving average model on the Testing dataset

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.

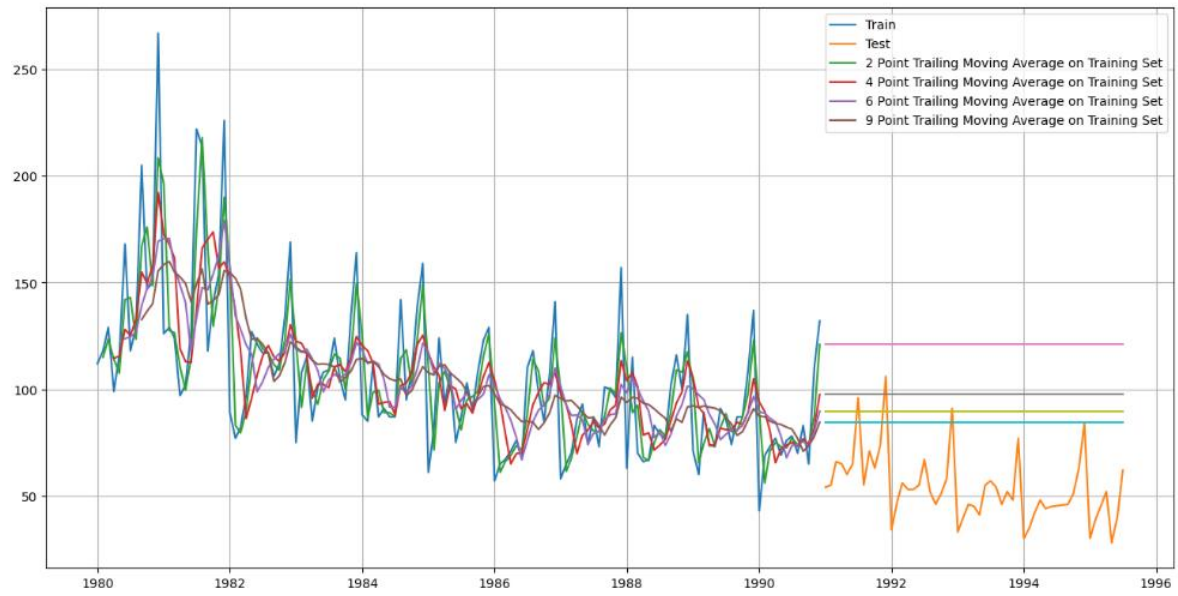


Figure 25: Time series plot with Training and Testing Dataset and Moving Average Model with different intervals

As we can see from the plot above , the Moving Average Model for different intervals is plotted vis- a-vis testing dataset . 2 point MA model is represented by green line, 4 point MA model is represented by red line, 6 point MA model is represented by purple line and 9 point MA model is represented by brown line. 2 point MA model seems to be performing best on the testing dataset

Let's evaluate the model using RMSE for all the MA models

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 68.970
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 46.404
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 39.129
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 34.407
```

Among Moving Average models RMSE score of 2 point Moving Average Model has the leas RMSE score , hence that is the best model among all other MA models

RMSE score of all the models

	Test RMSE
Alpha=0.098749:SimpleExponentialSmoothing	36.796227
Alpha=0.017550,Beta=0.000032:DoubleExponentialSmoothing	15.707052
Alpha=0.071511,Beta=0.045292,Gamma=0.000072:TripleExponentialSmoothingMultiplicative	20.156763
Alpha=0.089541,Beta=0.000240,Gamma=0.003467:TripleExponentialSmoothingAdditive	14.249661
Alpha==0.07:SimpleExponentialSmoothing	36.435772
Alpha==0.04,,Beta=0.47:DoubleExponentialSmoothing	14.560058
RegressionOnTime	15.268955
Simple Average	53.460570
2pointTrailingMovingAverage	68.970159
4pointTrailingMovingAverage	46.403626
6pointTrailingMovingAverage	39.129497
9pointTrailingMovingAverage	34.406988

Table 27: RMSE score of all models

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.

A Time Series is considered to be stationary when statistical properties such as the variance and (auto) correlation are constant over time.

Stationary Time Series allows us to think of the statistical properties of the time series as not changing in time, which enables us to build appropriate statistical models for forecasting based on past data.

Stationarity means that the autocorrelation of lag 'k' depends on k, but not on time t.

Let Xt denote the time series at time t.

Autocorrelation of lag k is the correlation between Xt and $X(t-k)$

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

Null Hypothesis H_0: Time Series is non-stationary.

Alternate Hypothesis H_a: Time Series is stationary.

So Ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected .

Dicky Fuller Test on the complete dataset to check stationarity

DF test statistic is -2.240

DF test p-value is 0.4671

Observations:

- As the p value is larger than 0.05, we fail to reject the null hypotheses that Time Series is non-stationary.
- The Training data is non-stationary at 95% confidence level.

Differencing 'd' to make time series stationary

Differencing 'd' is done on a non-stationary time series data one or more times to convert it into stationary.

(d=1) 1st order differencing is done where the difference between the current and previous (1 lag before) series is taken and then checked for stationarity using the ADF(Augmented Dicky Fueller) test. If differenced time series is stationary, we proceed with AR modelling.

Else we do (d=2) 2nd order differencing, and this process repeats till we get a stationary time series

1st order differencing equation is : y_t=y_t-y_{t-1}

2nd order differencing equation is : y_t=(y_t- y_{t-1}) − (y_{t-1}- y_{t-2}) and so on…

The variance of a time series may also not be the same over time. To remove this kind of non-stationarity, we can transform the data. If the variance is increasing over time, then a log transformation can stabilize the variance.

Let's apply differencing of order 1 on the dataset and check for stationarity

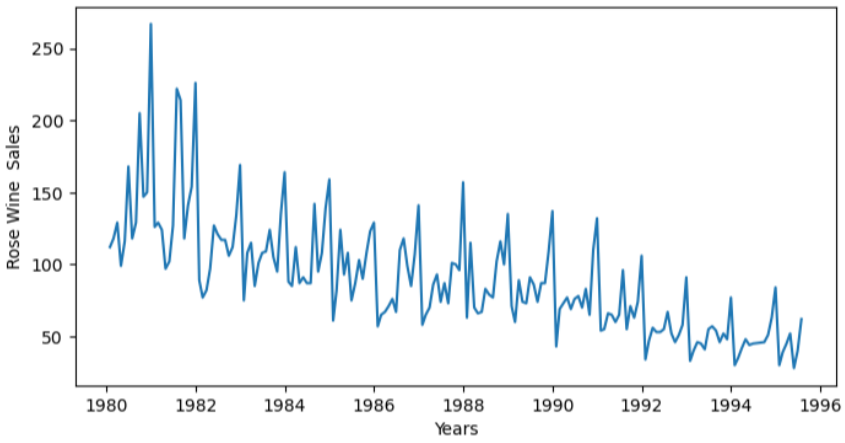


Figure 26:Original Time series

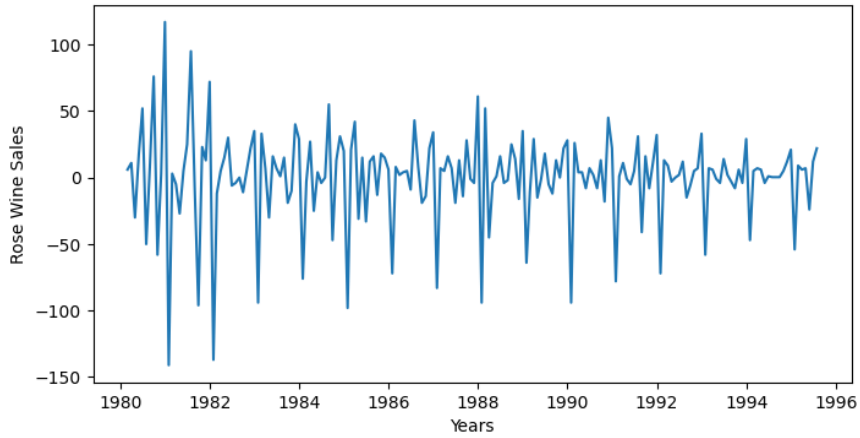


Figure 27:Time series after differencing d=1

Dicky Fuller Test on the differenced(d=1) dataset to check stationarity

DF test statistic is -8.162

DF test p-value is 3.015976115827911e-11

Observations:

- As the p value is less than 0.05, we reject the null hypotheses that Time Series is non-stationary.
- The data is stationary at 95% confidence level.

Dicky Fuller Test on the Training dataset to check stationarity

DF test statistic is -1.686

DF test p-value is 0.7569093051047061

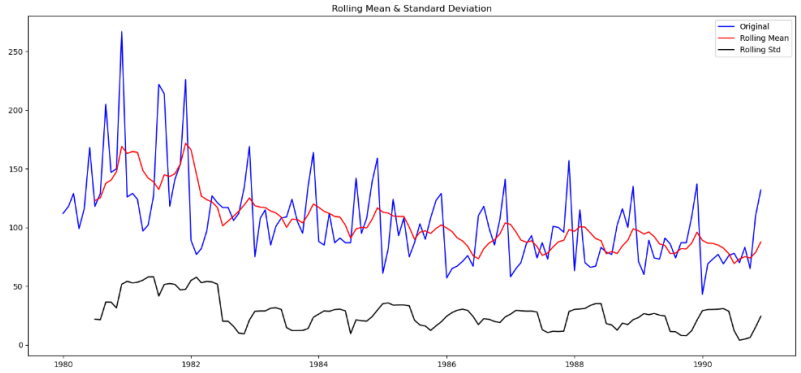


Figure 28: Training Data Time series

Observations:

- As the p value is larger than 0.05, we fail to reject the null hypotheses that Time Series is non-stationary.
- The Training data is non-stationary at 95% confidence level.

Dicky Fuller Test on the differenced(d=1) Training dataset to check stationarity

DF test statistic is -6.804

DF test p-value is 3.894831356782305e-08

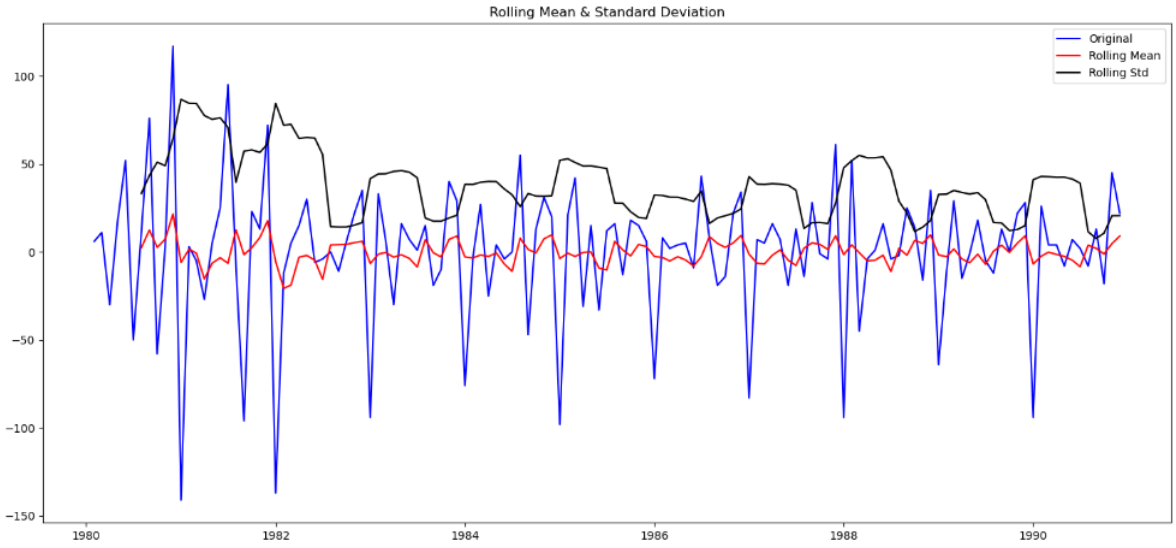


Figure 29: Training Data Time series after differencing

Observations:

- As the p value is less than 0.05, we reject the null hypotheses that Time Series is non-stationary.
- The Training data is stationary at 95% confidence level.

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.

ARIMA model

ARIMA:- Auto Regressive Integrated Moving Average is a way of modelling time series data for forecasting or predicting future data points.

Improving AR Models by making Time Series stationary through Moving Average Forecasts

ARIMA models consist of 3 components:-

AR model: The data is modelled based on past observations.

Integrated component: Whether the data needs to be differenced/transformed.

MA model: Previous forecast errors are incorporated into the model.

ARIMA Model building to estimate best 'p' , 'd' , 'q' parameters (Lowest AIC Approach)

We split the dataset into Training and Testing set , with recent observations in the testing dataset. Training Data is used to train (develop) the ARIMA model. Training Data is used to identify a few

working models with different values of p,d,q. Estimate p,d,q by looking at the lowest AIC for the models built on training data

The model built parameters is then used on the training data to forecast the test data and calculate model evaluation parameters like RMSE .

After the best model is selected , model is checked using diagnostics on the whole data and forecast for the desired future time points using this model .

The table below shows AIC scores for different values of p,d,q listed in ascending order of AIC scores

	param	AIC
13	(2, 1, 3)	1274.696103
23	(4, 1, 3)	1278.451407
18	(3, 1, 3)	1278.654108
14	(2, 1, 4)	1278.768962
9	(1, 1, 4)	1279.605263
2	(0, 1, 2)	1279.671529
7	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
6	(1, 1, 1)	1280.574230
11	(2, 1, 1)	1281.507862
4	(0, 1, 4)	1281.676698
12	(2, 1, 2)	1281.870722
8	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
24	(4, 1, 4)	1282.360356
16	(3, 1, 1)	1282.419278

Table 28: AIC scores in ARIMA model

AS we can see Param 2,1,3 has the lowest AIC score 1274.696. Let's build the model using the param 2,1,3

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.348
Date:	Fri, 01 Sep 2023	AIC	1274.696
Time:	23:15:13	BIC	1291.947
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.706
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6773	0.084	-19.962	0.000	-1.842	-1.513
ar.L2	-0.7281	0.084	-8.661	0.000	-0.893	-0.563
ma.L1	1.0460	0.684	1.530	0.126	-0.294	2.386
ma.L2	-0.7698	0.138	-5.565	0.000	-1.041	-0.499
ma.L3	-0.9037	0.620	-1.458	0.145	-2.119	0.311
sigma2	859.7747	576.982	1.490	0.136	-271.090	1990.640

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.16
Prob(Q):	0.89	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.70
Prob(H) (two-sided):	0.00	Kurtosis:	4.56

Figure 30: Summary of ARIMA model

We can see from the summary above that ar.L1, ar.L2,ma.L2 are significant variables in building the model equation .

Let's see the diagnostics plot

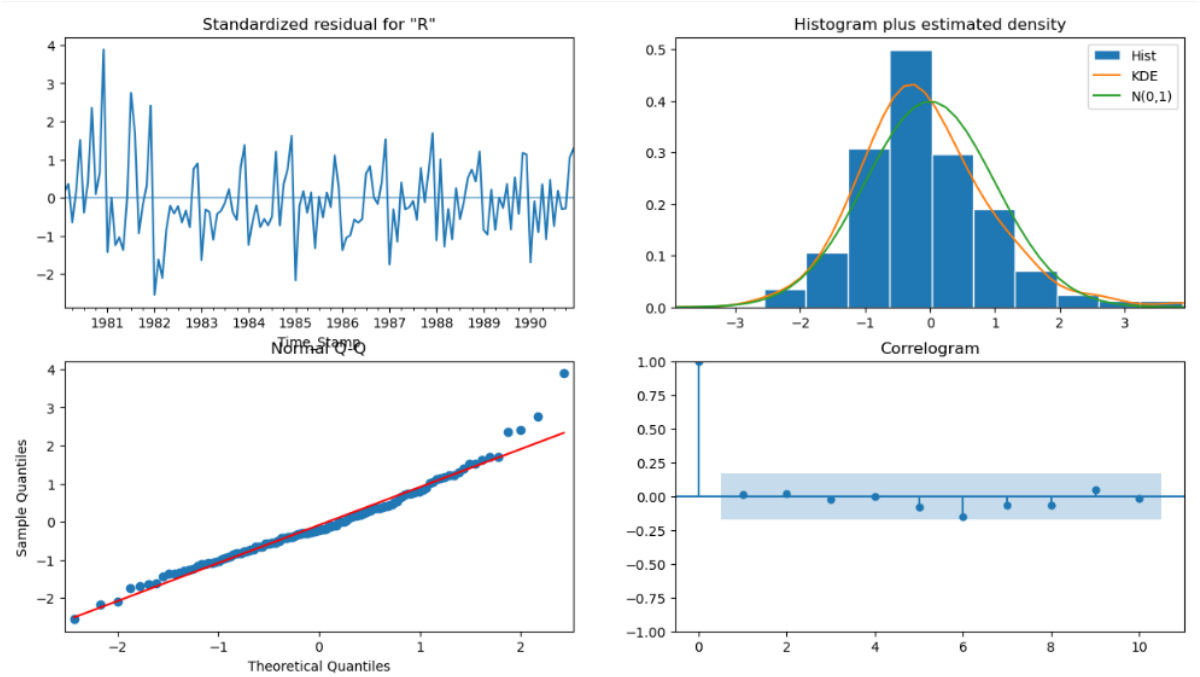


Figure 31: Diagnostics plot of ARIMA model

The residual plot looks close to normal and Q-Q plot look ok

Now we can predict on the Test Set using this model and evaluate the model.

1991-01-01	85.624392
1991-02-01	90.570708
1991-03-01	81.982766
1991-04-01	92.786210
1991-05-01	80.917993
1991-06-01	92.959156
1991-07-01	81.403124
1991-08-01	92.019526
1991-09-01	82.625971
1991-10-01	90.652535
1991-11-01	84.028536
1991-12-01	89.295245
1992-01-01	85.283987
1992-02-01	88.177650
1992-03-01	86.244502
1992-04-01	87.380241
1992-05-01	86.882695
1992-06-01	86.890351
1992-07-01	87.239751
1992-08-01	86.648126
1992-09-01	87.386082
1992-10-01	86.579037
1992-11-01	87.395428
1992-12-01	86.613662
1993-01-01	87.330547
1993-02-01	86.697277
1993-03-01	87.237534
1993-04-01	86.792412
1993-05-01	87.145682
1993-06-01	86.877213
1993-07-01	87.070319
1993-08-01	86.941881

Table 29:Predictions on Test set

Let's evaluate the model using RMSE

The Test RMSE score for ARIMA model is 36.83

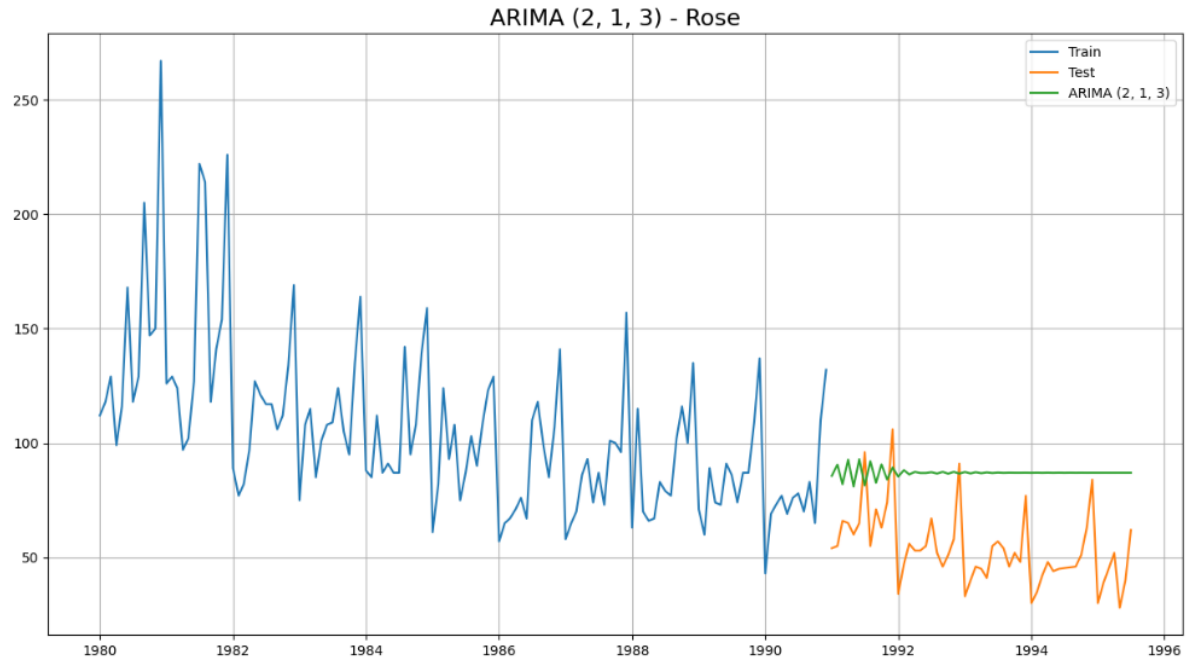


Figure 32: plot of ARIMA(2,1,3) model

ARIMA model on the training data for which the best parameters are selected by looking at the ACF and the PACF

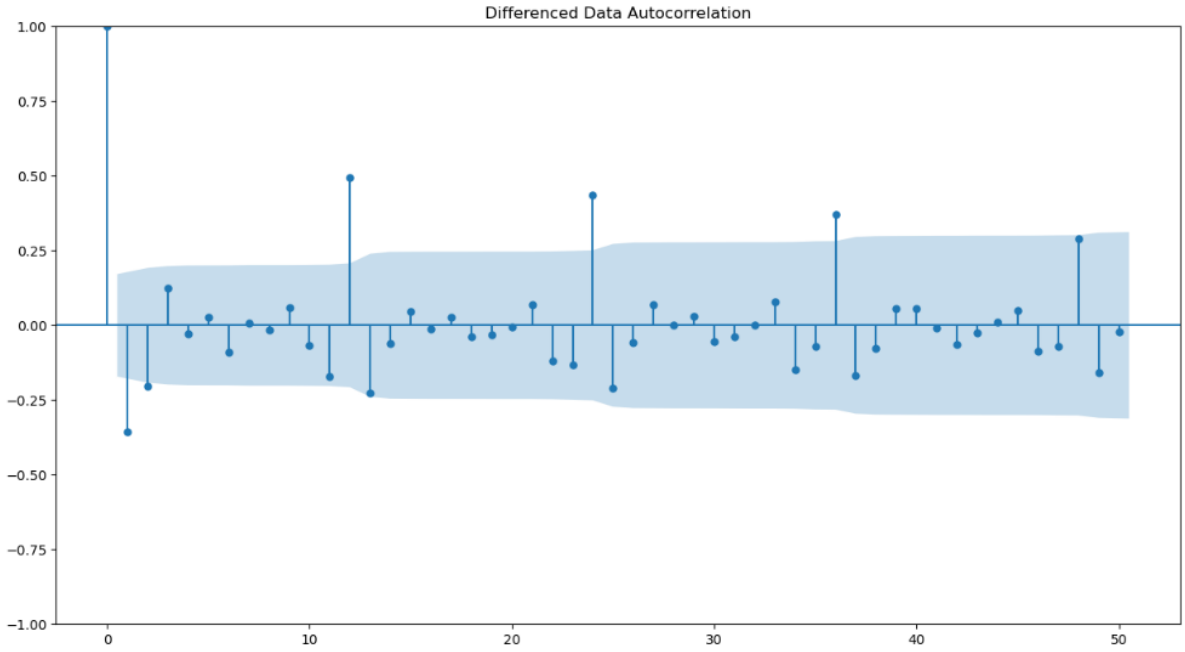


Figure 33: ACF of Training dataset

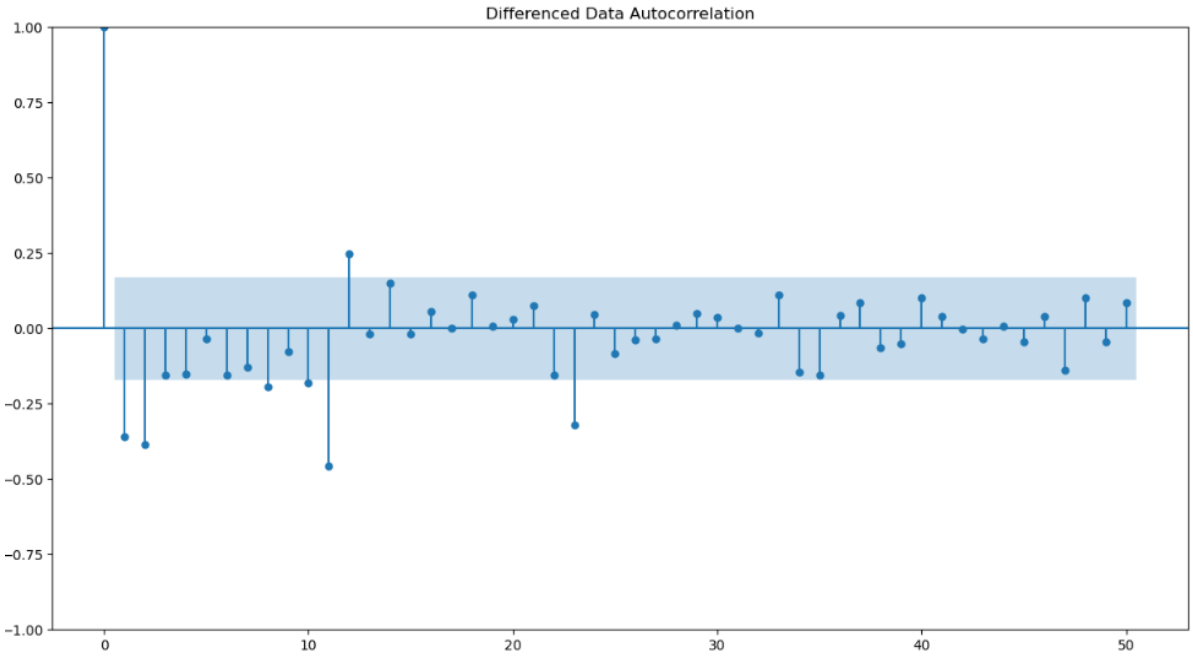


Figure 34: PACF plot of Training Dataset

There are significant peaks in the ACF and PACF plots, so let's take p as 2 and q as 2 . Other significant peaks are only at 12, 24, 36 in ACF plots because of seasonal effect

Manual ARIMA(2, 1, 2)

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Fri, 01 Sep 2023	AIC	1281.871
Time:	23:22:18	BIC	1296.247
Sample:	01-01-1980 - 12-01-1990	HQIC	1287.712
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4540	0.469	-0.969	0.333	-1.372	0.464
ar.L2	0.0001	0.170	0.001	0.999	-0.334	0.334
ma.L1	-0.2541	0.459	-0.554	0.580	-1.154	0.646
ma.L2	-0.5984	0.430	-1.390	0.164	-1.442	0.245
sigma2	952.1601	91.424	10.415	0.000	772.973	1131.347

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	34.16
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.79
Prob(H) (two-sided):	0.00	Kurtosis:	4.94

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

Table 30: Summary of ARIMA(2, 1, 2) model

. As we can see there are no significant components in the model as the probability of the components is > than 0.05

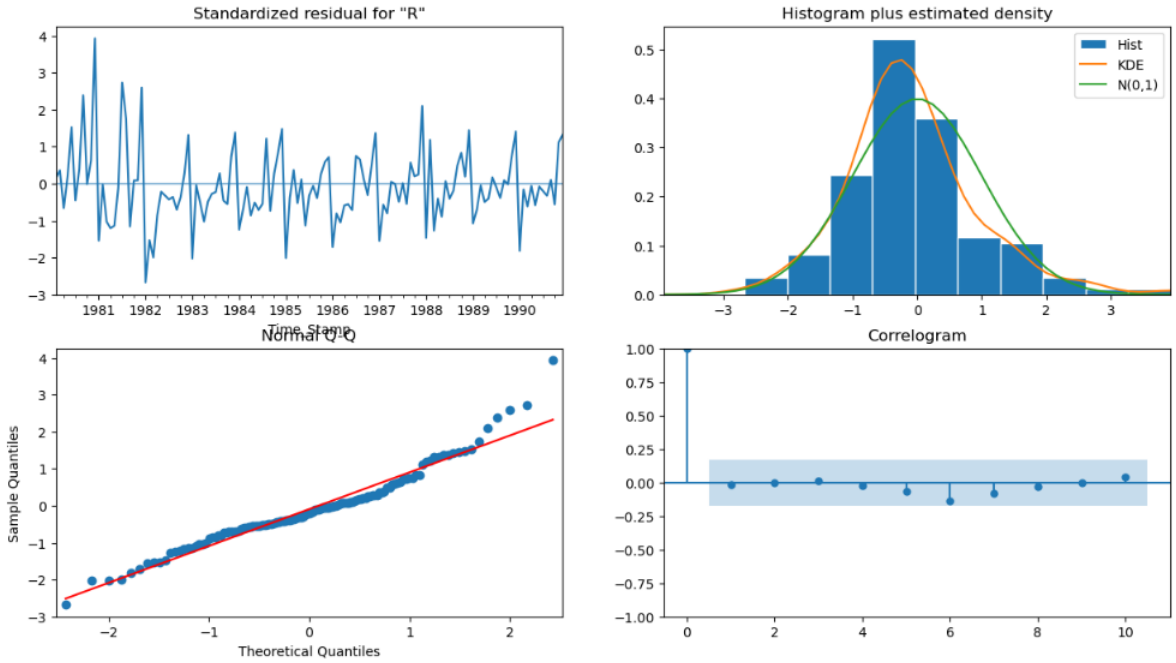


Figure 35: Diagnostics plot of Auto ARIMA(2, 1, 2)

Predict on the Test Set using these models and evaluate the model.

RMSE score for the Manual ARIMA models(2,1,2) is 36.87

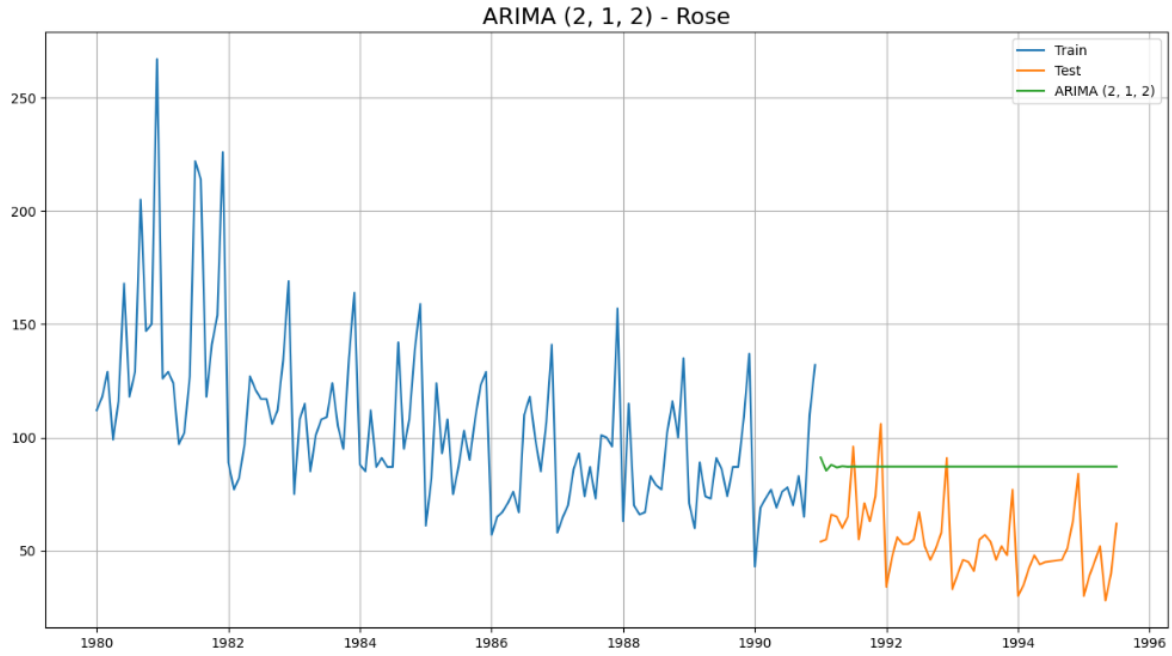


Figure 36: plot of ARIMA(2,1,2) model

SARIMA model

Although ARIMA method can handle data with a trend, it does not support time series with a seasonal component. An extension to ARIMA that supports the direct modelling of the seasonal component of the series is called SARIMA. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.

Trend Elements

There are three trend elements that require configuration.

They are the same as the ARIMA model; specifically:

p: Trend autoregression order.

d: Trend difference order.

q: Trend moving average order.

Seasonal Elements

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

P: Seasonal autoregressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

F: The number of time steps for a single seasonal period.

Together, the notation for a SARIMA model is specified as:

The value for the parameters (p,d,q) and (P, D, Q) can be decided by comparing different values for each and taking the lowest AIC value for the model build. The value for F can be consolidated by ACF plot

Training Data is used to identify a few working models with different values of 'p' , 'd' , 'q' and 'P','D','Q' . Estimate p,d,q and P,D,Q by looking at the lowest AIC for the models built on training data

The model built parameters is then used on the training data to forecast the test data and calculate model evaluation parameters like RMSE .

After the best model is selected , model is checked using diagnostics on the whole data and forecast for the desired future time points using this model .

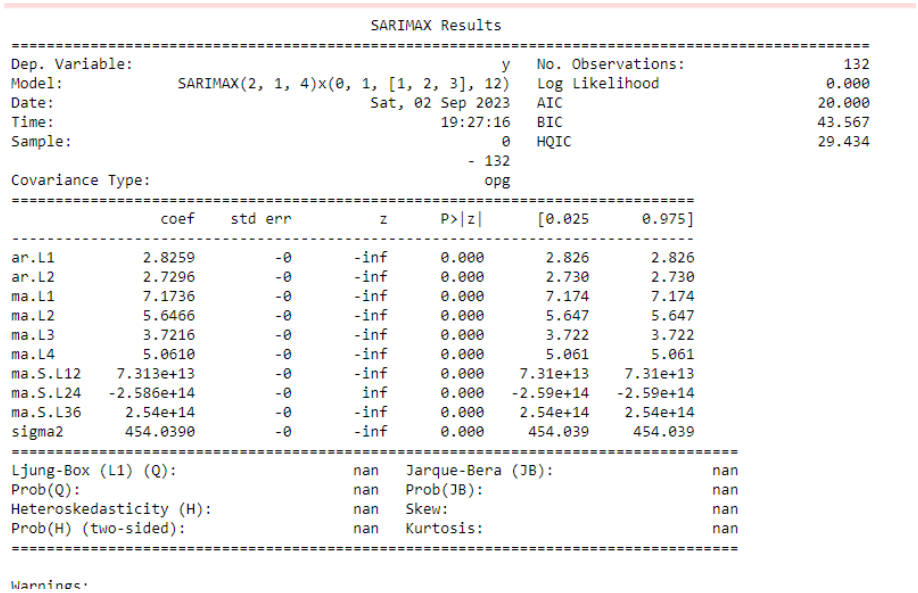
The following are the some of the AIC parameters

	param	seasonal	AIC
708	(2, 1, 4)	(0, 1, 3, 12	20
823	(3, 1, 1)	(2, 0, 3, 12	20
953	(3, 1, 4)	(0, 0, 3, 12	22
1073	(4, 1, 1)	(2, 0, 3, 12	22
1118	(4, 1, 2)	(1, 1, 3, 12	22
723	(2, 1, 4)	(2, 0, 3, 12	24
1173	(4, 1, 3)	(2, 0, 3, 12	26
948	(3, 1, 3)	(4, 1, 3, 12	28
993	(3, 1, 4)	(4, 0, 3, 12	30
673	(2, 1, 3)	(2, 0, 3, 12	194.4436
213	(0, 1, 4)	(1, 0, 3, 12	267.1612
223	(0, 1, 4)	(2, 0, 3, 12	323.3252
173	(0, 1, 3)	(2, 0, 3, 12	433.0358
908	(3, 1, 3)	(0, 1, 3, 12	522.0019
928	(3, 1, 3)	(2, 1, 3, 12	552.3098
1229	(4, 1, 4)	(2, 1, 4, 12	556.7764
729	(2, 1, 4)	(2, 1, 4, 12	558.8795
1249	(4, 1, 4)	(4, 1, 4, 12	559.198
979	(3, 1, 4)	(2, 1, 4, 12	560.6822
739	(2, 1, 4)	(3, 1, 4, 12	560.8791

Table 31: AIC scores of SARIMA model

The table above shows AIC scores for different values of p,d,q and P,D,Q listed in ascending order of AIC scores

Let's build model with different AIC scores and see the diagnostics plot to determine the best model



SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 4)x(0, 1, [1, 2, 3], 12)	Log Likelihood	0.000
Date:	Sat, 02 Sep 2023	AIC	20.000
Time:	19:27:16	BIC	43.567
Sample:	0	HQIC	29.434
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	2.8259	-0	-inf	0.000	2.826	2.826
ar.L2	2.7296	-0	-inf	0.000	2.730	2.730
ma.L1	7.1736	-0	-inf	0.000	7.174	7.174
ma.L2	5.6466	-0	-inf	0.000	5.647	5.647
ma.L3	3.7216	-0	-inf	0.000	3.722	3.722
ma.L4	5.0610	-0	-inf	0.000	5.061	5.061
ma.S.L12	7.313e+13	-0	-inf	0.000	7.31e+13	7.31e+13
ma.S.L24	-2.586e+14	-0	inf	0.000	-2.59e+14	-2.59e+14
ma.S.L36	2.54e+14	-0	-inf	0.000	2.54e+14	2.54e+14
sigma2	454.0390	-0	-inf	0.000	454.039	454.039

Ljung-Box (L1) (Q):	nan	Jarque-Bera (JB):	nan
Prob(Q):	nan	Prob(JB):	nan
Heteroskedasticity (H):	nan	Skew:	nan
Prob(H) (two-sided):	nan	Kurtosis:	nan

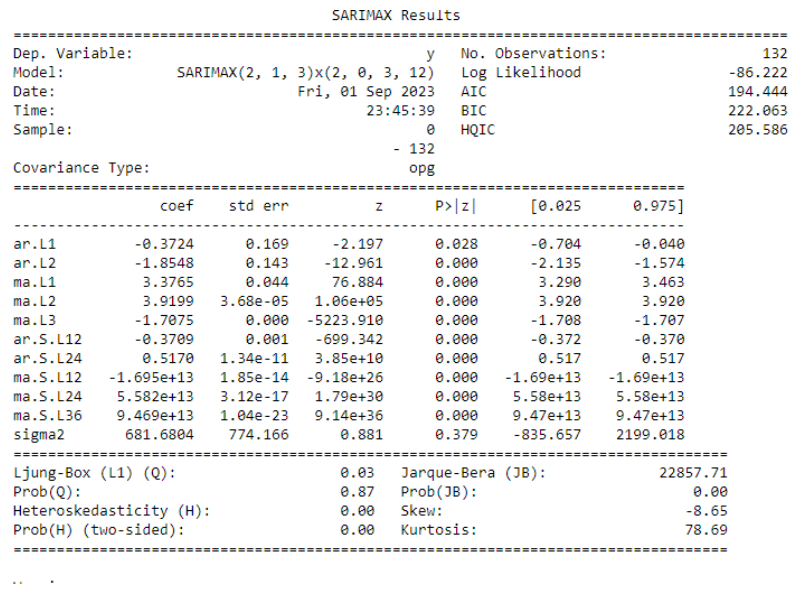
Warnings:

Figure 37: Result Summary of Auto SARIMA(2, 1, 4)(0, 1, 3, 12) Model

This is a model with least AIC score 20

As we can see the model is not suitable for building because as we can see most of the values in the residual summary are Not available , so we cannot diagnose using this model.

So, we can try building model with higher AIC scores



SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 0, 3, 12)	Log Likelihood	-86.222
Date:	Fri, 01 Sep 2023	AIC	194.444
Time:	23:45:39	BIC	222.063
Sample:	0	HQIC	205.586
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3724	0.169	-2.197	0.028	-0.704	-0.040
ar.L2	-1.8548	0.143	-12.961	0.000	-2.135	-1.574
ma.L1	3.3765	0.044	76.884	0.000	3.290	3.463
ma.L2	3.9199	3.68e-05	1.06e+05	0.000	3.920	3.920
ma.L3	-1.7075	0.000	-5223.910	0.000	-1.708	-1.707
ar.S.L12	-0.3709	0.001	-699.342	0.000	-0.372	-0.370
ar.S.L24	0.5170	1.34e-11	3.85e+10	0.000	0.517	0.517
ma.S.L12	-1.695e+13	1.85e-14	-9.18e+26	0.000	-1.69e+13	-1.69e+13
ma.S.L24	5.582e+13	3.12e-17	1.79e+30	0.000	5.58e+13	5.58e+13
ma.S.L36	9.469e+13	1.04e-23	9.14e+36	0.000	9.47e+13	9.47e+13
sigma2	681.6804	774.166	0.881	0.379	-835.657	2199.018

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	22857.71
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	0.00	Skew:	-8.65
Prob(H) (two-sided):	0.00	Kurtosis:	78.69

Figure 38:Diagnostic plot of Auto SARIMA(2, 1, 3)(2,0,3,12) Model

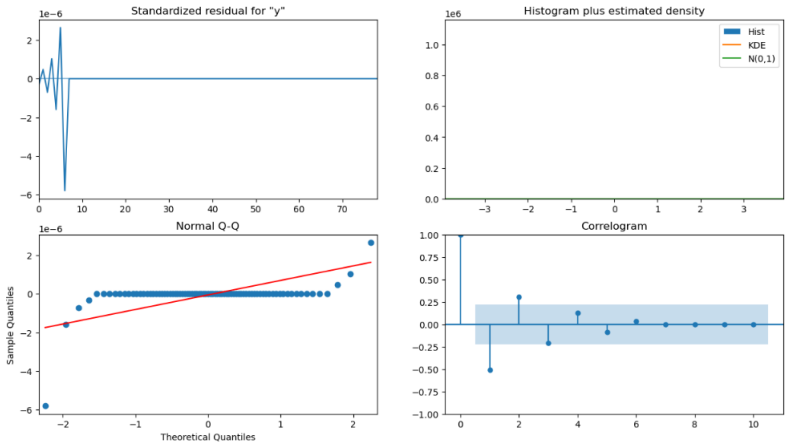


Figure 39:Residual Plot of SARIMA(2, 1, 3)(2,0,3,12) Model

By looking the SARIMA model summary, we can see all the parameters are significant since p-values is less than 0.05. But the Q-Q plot the residuals generated are not following a normal distribution

Let's build a model with higher AIC scores

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(4, 1, 4)x(2, 1, 4, 12)	Log Likelihood	-263.390
Date:	Sat, 02 Sep 2023	AIC	556.780
Time:	00:17:23	BIC	589.625
Sample:	0	HQIC	569.759
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7293	0.221	-3.305	0.001	-1.162	-0.297
ar.L2	-0.7916	0.217	-3.648	0.000	-1.217	-0.366
ar.L3	-0.8832	0.217	-4.079	0.000	-1.308	-0.459
ar.L4	-0.0585	0.188	-0.311	0.756	-0.428	0.311
ma.L1	-0.2377	4.958	-0.048	0.962	-9.955	9.480
ma.L2	0.0877	17.105	0.005	0.996	-33.437	33.613
ma.L3	0.4641	7.594	0.061	0.951	-14.419	15.348
ma.L4	-0.8579	15.271	-0.056	0.955	-30.789	29.073
ar.S.L12	-1.1705	0.101	-11.575	0.000	-1.369	-0.972
ar.S.L24	-0.4887	0.094	-5.195	0.000	-0.673	-0.304
ma.S.L12	0.6070	8.375	0.072	0.942	-15.807	17.021
ma.S.L24	-0.3101	2.488	-0.125	0.901	-5.186	4.565
ma.S.L36	-0.6190	8.392	-0.074	0.941	-17.068	15.830
ma.S.L48	0.2030	2.418	0.084	0.933	-4.536	4.942
sigma2	114.7602	2414.647	0.048	0.962	-4617.861	4847.381

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	4.73
Prob(Q):	1.00	Prob(JB):	0.09
Heteroskedasticity (H):	0.66	Skew:	0.31
Prob(H) (two-sided):	0.34	Kurtosis:	4.16

Figure 40: Result Summary of Auto SARIMA(4, 1, 4)(2, 1,4 12)

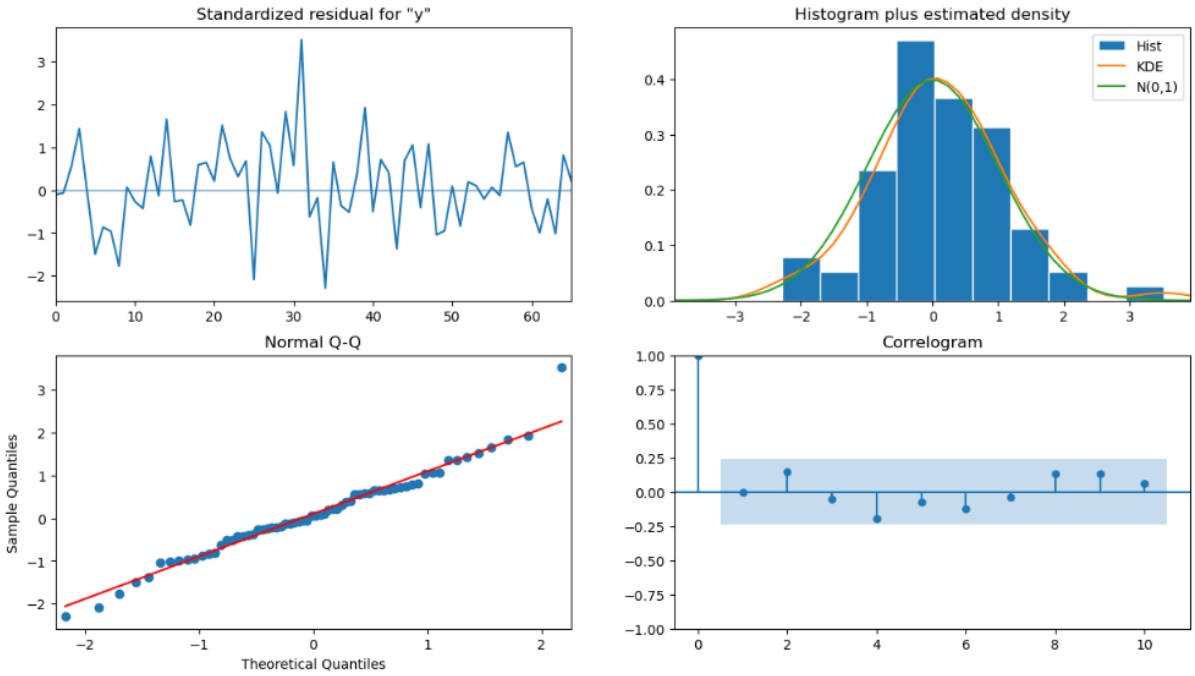


Figure 41: Diagnostics plot of Auto SARIMA(4, 1, 4)(2, 1,4 12)

By looking the SARIMA model summary, we can see that the coefficients for components and p-values of the components like – maL1,maL2,maL3,maL4 and ma.S.L12, ma.S.L24, ma.S.L36, ma.S.L48 of the SARIMA model are more than 0.05 so these are insignificant variables in prediction whereas the coefficients for components and p-values of the components like – ar.L1, ar.L2 , ar.S.L12, ar.S.L24of the SARIMA model are less than 0.05, hence significant in prediction. Q_Q Plot is following a normal distribution

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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	46.051770	12.143426	22.251093	69.852447
1	56.409142	12.029596	32.831567	79.986717
2	73.143321	12.025141	49.574478	96.712163
3	74.492167	12.672338	49.654840	99.329493
4	69.701260	12.735149	44.740826	94.661694

Table 32: Predictions on the test set with Auto SARIMA(4, 1, 4)(2, 1,4 12)

RMSE score for the Auto SARIMA(4, 1, 4)(2, 1,4 12)models is 18.34

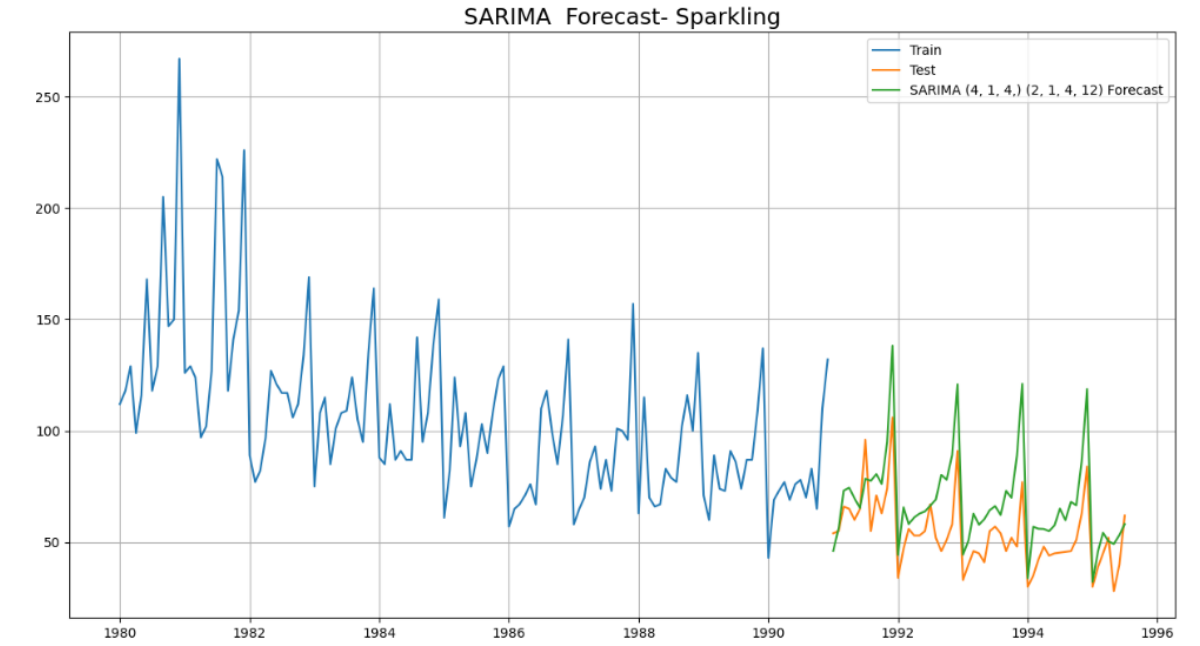


Figure 42: Plot of the SARIMA model vis-à-vis Training and Testing Graphs

As we can see plot of the SARIMA model **(4, 1, 4)(2, 1,4 12)** vis-à-vis Testing plot is a good model

SARIMA model on the training data for which the best parameters are selected by looking at the ACF and the PACF

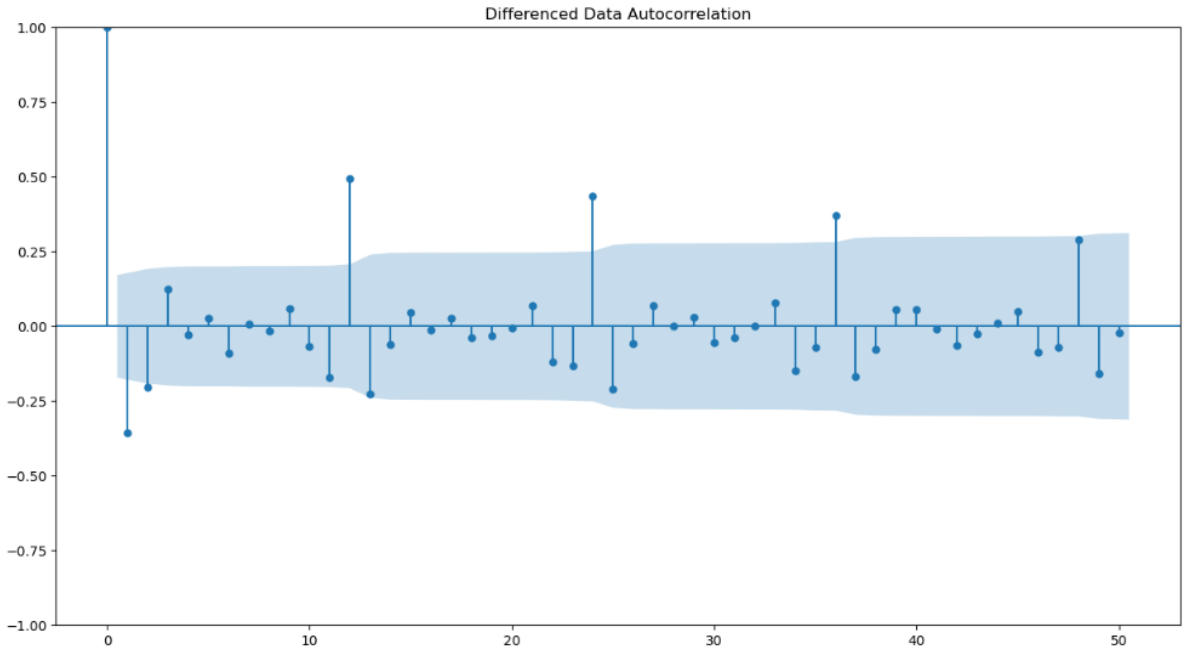


Figure 43: ACF of Training dataset

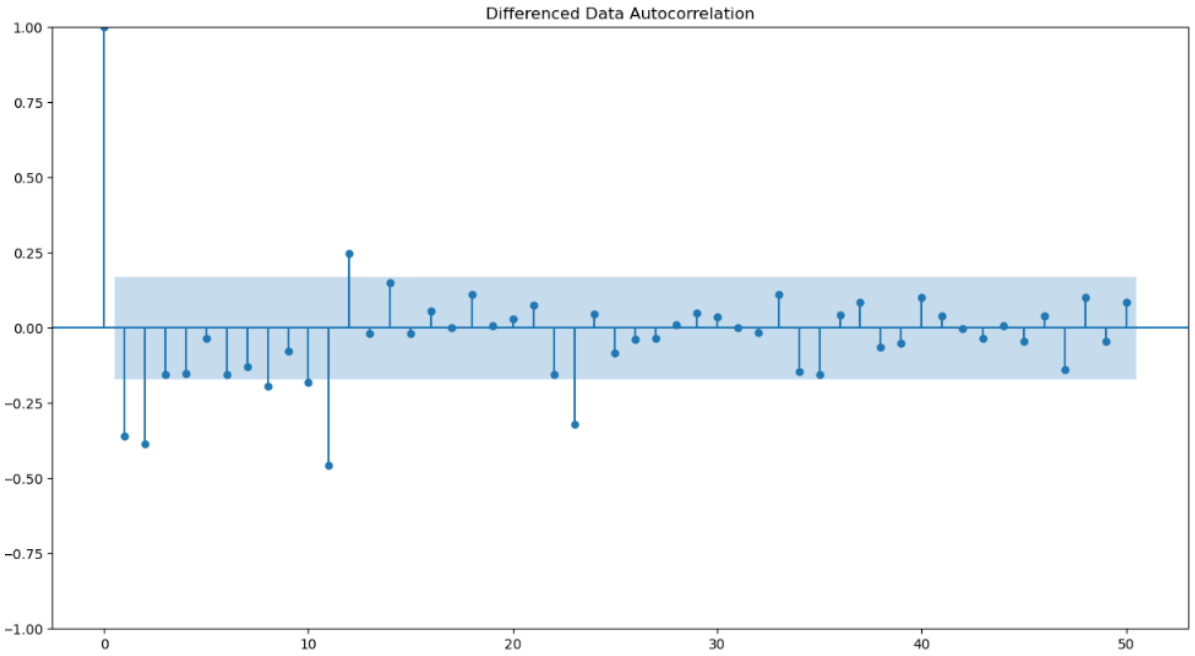


Figure 44: PACF plot of Training Dataset

Let's take p and q same as ARIMA plot which both as 2

To determine P and Q let's take values as 1 and 3 for one model and 1 and 0 for another model

Manual SARIMA(2, 1, 2)(1, 1, 3, 12).

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(1, 1, [1, 2, 3], 12)	Log Likelihood	-2505.649
Date:	Sat, 02 Sep 2023	AIC	5029.299
Time:	19:48:08	BIC	5050.737
Sample:	0 - 132	HQIC	5037.894
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.1152	2242.544	-5.14e-05	1.000	-4395.420	4395.190
ar.L2	-0.1491	1634.210	-9.12e-05	1.000	-3203.142	3202.844
ma.L1	-0.4493	42.396	-0.011	0.992	-83.543	82.645
ma.L2	-0.3120	40.430	-0.008	0.994	-79.554	78.930
ar.S.L12	-0.2841	2.08e+04	-1.36e-05	1.000	-4.08e+04	4.08e+04
ma.S.L12	3.59e+13	1.53e-05	2.35e+18	0.000	3.59e+13	3.59e+13
ma.S.L24	-2.635e+14	2.08e-06	-1.27e+20	0.000	-2.63e+14	-2.63e+14
ma.S.L36	2.097e+14	2.56e-07	8.19e+20	0.000	2.1e+14	2.1e+14
sigma2	569.4616	146.713	3.881	0.000	281.910	857.014

Ljung-Box (L1) (Q):	3.04	Jarque-Bera (JB):	11661.73
Prob(Q):	0.08	Prob(JB):	0.00
Heteroskedasticity (H):	1778827.82	Skew:	-7.27
Prob(H) (two-sided):	0.00	Kurtosis:	60.34

Figure 45: Result Summary of Auto SARIMA(2, 1, 2)(1, 1, 3, 12).

As we can from the summary parameters ma.L1 and ma.L2 and ar,S.L12 are insignificant because the P values is greater than 0.05.

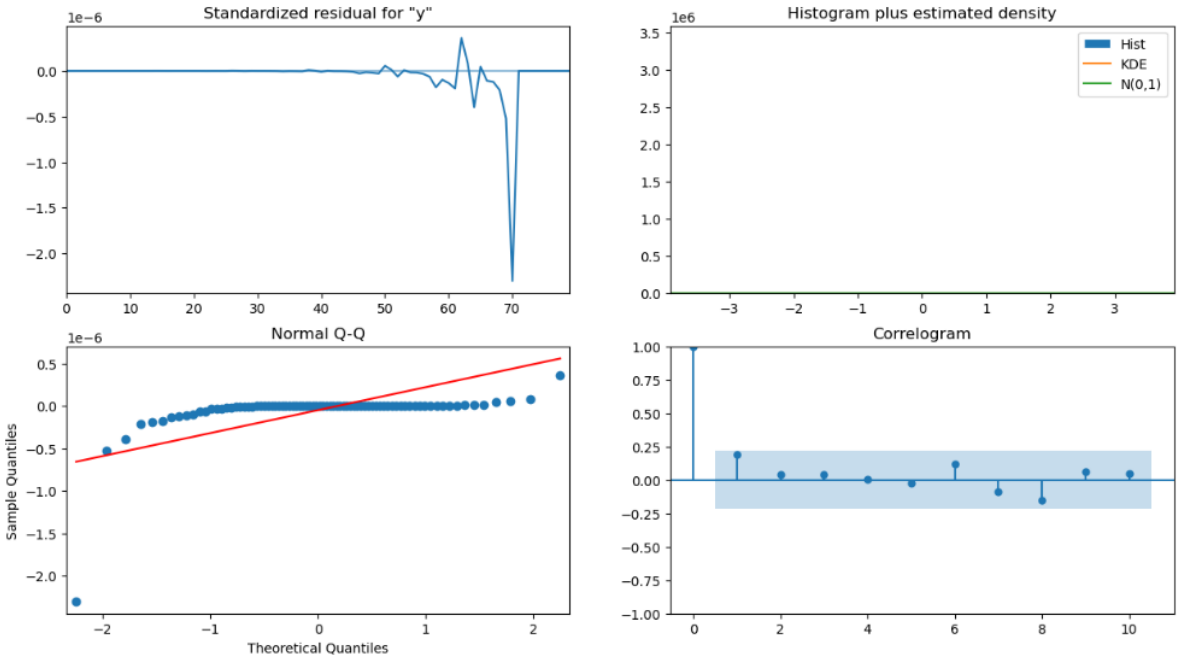


Figure 46: Diagnostics plot of Auto SARIMA(2, 1, 2)(1, 1, 3, 12).

Q_Q plot is not normal so this is not a good model

Manual SARIMA(2, 1, 2)(1, 1, 0, 12).

For P and Q, we go by AR(1) model

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(1, 1, [], 12)	Log Likelihood	-456.685
Date:	Fri, 01 Sep 2023	AIC	925.371
Time:	23:40:44	BIC	941.294
Sample:	0	HQIC	931.823
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0748	0.146	7.365	0.000	0.789	1.361
ar.L2	-0.3376	0.113	-2.986	0.003	-0.559	-0.116
ma.L1	-1.7907	0.123	-14.610	0.000	-2.031	-1.550
ma.L2	0.8432	0.109	7.764	0.000	0.630	1.056
ar.S.L12	-0.4085	0.058	-6.991	0.000	-0.523	-0.294
sigma2	347.3434	51.920	6.690	0.000	245.581	449.105

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	0.02
Prob(Q):	0.89	Prob(JB):	0.99
Heteroskedasticity (H):	0.64	Skew:	0.03
Prob(H) (two-sided):	0.20	Kurtosis:	3.03

Figure 47: Result Summary of Auto SARIMA(2, 1, 2)(1, 1, 0, 12).

As we can from the summary parameters ar.L2 is insignificant because the P values is greater than 0.05. All other parameters are significant in model building .

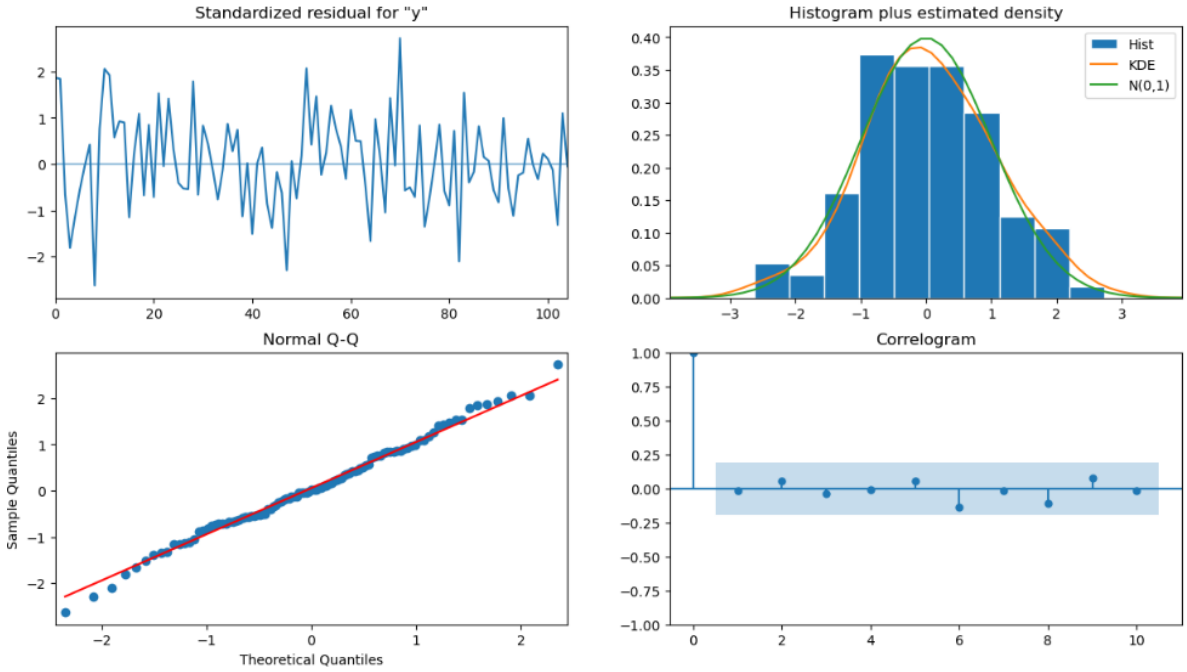


Figure 48: Diagnostics plot Auto SARIMA(2, 1, 2)(1, 1, 0, 12).

Q-Q plot is almost normal with few outliers

Predict on the Test Set using these models and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	46.660836	18.637150	10.132693	83.188979
1	55.895842	19.374674	17.922179	93.869505
2	69.551352	19.378366	31.570453	107.532250
3	65.745795	19.382518	27.756758	103.734831
4	60.747634	19.387091	22.749634	98.745634

Table 33: Predictions on the test set with SARIMA(0, 1, 0)(2, 1,4, 12) Model

RMSE score for the Manual SARIMA(2, 1, 2)(1, 1,0, 12) models is 13.35

Let's plot the best Manual SARIMA models **(2, 1, 2)(1, 1,0, 12)**

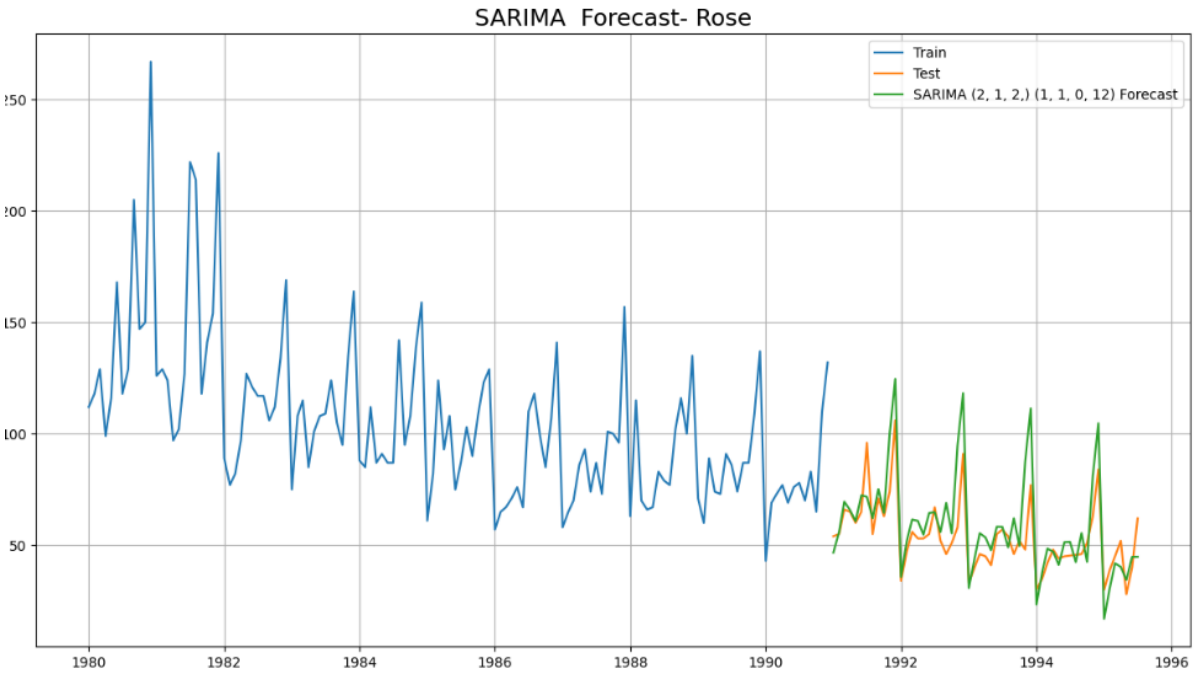


Figure 49: Plot of the SARIMA model vis-à-vis Training and Testing Graphs

As we can see plot of the SARIMA model **(2, 1, 2)(1, 1,0, 12)** vis-à-vis Testing plot is a good model

7. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

After building various time series forecasting models like Linear Regression , Navie Forecast ,Simple Average , Moving Average and exponential smoothing models like (Simple , Double , Triple Exponential) and ARIMA / SARIMA models on the Sparkling Wine Sales dataset , here are the models and their RMSE on the test data

	Test RMSE
Alpha=0.098749:SimpleExponentialSmoothing	36.796227
Alpha=0.017550,Beta=0.000032:DoubleExponentialSmoothing	15.707052
Alpha=0.071511,Beta=0.045292,Gamma=0.000072:TripleExponentialSmoothingMultiplicative	20.156763
Alpha=0.089541,Beta=0.000240,Gamma=0.003467:TripleExponentialSmoothingAdditive	14.249661
Alpha==0.07:SimpleExponentialSmoothing	36.435772
Alpha==0.04,,Beta=0.47:DoubleExponentialSmoothing	14.560058
RegressionOnTime	15.268955
Simple Average	53.460570
2pointTrailingMovingAverage	68.970159
4pointTrailingMovingAverage	46.403626
6pointTrailingMovingAverage	39.129497
9pointTrailingMovingAverage	34.406988
ARIMA(2,1,3)	36.837680
ARIMA(2,1,2)	36.871197
SARIMA(4,1,4)(2,1,4,12)	18.346267
SARIMA(2,1,2)(1,1,0,12)	13.352731

Figure 50:RMSE values of all models built

As we can see from the above table, SARIMA(2,1,2)(1,1,0,12) is having least RMSE on test data, hence this is the best models

8.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.

SARIMA (2,1,2)((1,1,0,12) model seems to performing the best among all the models .

Let's build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

SARIMAX Results

Dep. Variable:		Rose	No. Observations:		187	
Model:	SARIMAX(2, 1, 2)x(1, 1, [], 12)		Log Likelihood		-671.732	
Date:	Sat, 02 Sep 2023		AIC		1355.464	
Time:	00:27:54		BIC		1373.915	
Sample:	01-01-1980		HQIC		1362.956	
	- 07-01-1995					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0760	0.102	10.531	0.000	0.876	1.276
ar.L2	-0.3208	0.082	-3.914	0.000	-0.482	-0.160
ma.L1	-1.8015	0.082	-22.098	0.000	-1.961	-1.642
ma.L2	0.8541	0.072	11.869	0.000	0.713	0.995
ar.S.L12	-0.4079	0.042	-9.663	0.000	-0.491	-0.325
sigma2	257.7201	26.536	9.712	0.000	205.711	309.729

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	3.80
Prob(Q):	0.84	Prob(JB):	0.15
Heteroskedasticity (H):	0.21	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.75

Figure 51:Result Summary of Auto SARIMA(2 1, 2)(1, 1,0, 12)

As we can all the variables are significant because p-values are less than 0.05 for all the variables

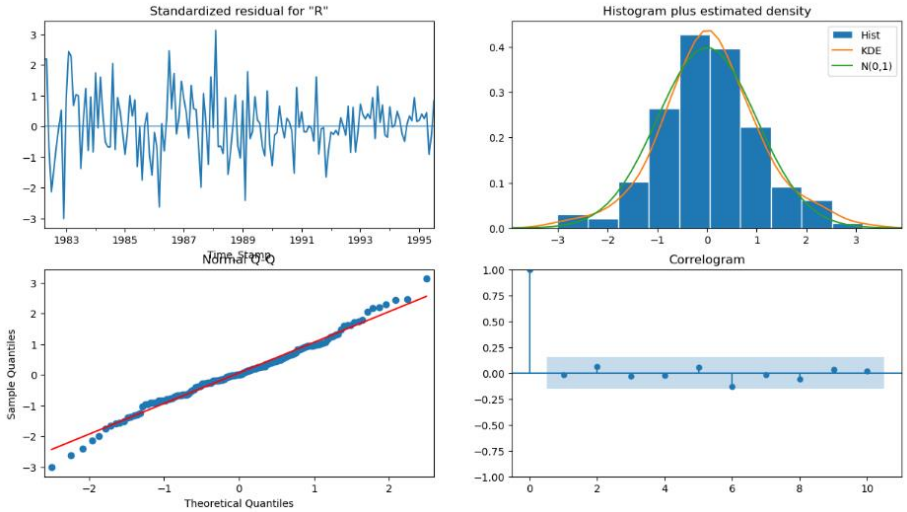


Figure 52: Diagnostics plot Auto SARIMA(2, 1, 2)(1, 1,0, 12)

Predict on the Test Set using these models and evaluate the model

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	54.041541	16.053664	22.576938	85.506145
1995-09-01	48.177628	16.647165	15.549783	80.805472
1995-10-01	52.798750	16.652786	20.159889	85.437611
1995-11-01	58.324183	16.653115	25.684677	90.963689
1995-12-01	82.894768	16.663626	50.234660	115.554875

Table 34: Predictions on the Entire data set with SARIMA(2, 1, 2)(1, 1,0, 12) Model

RMSE score for the SARIMA(2 1, 2)(1, 1,0, 12) model on full data is 38.271

Let's plot the future prediction of 12 months alongside original time series

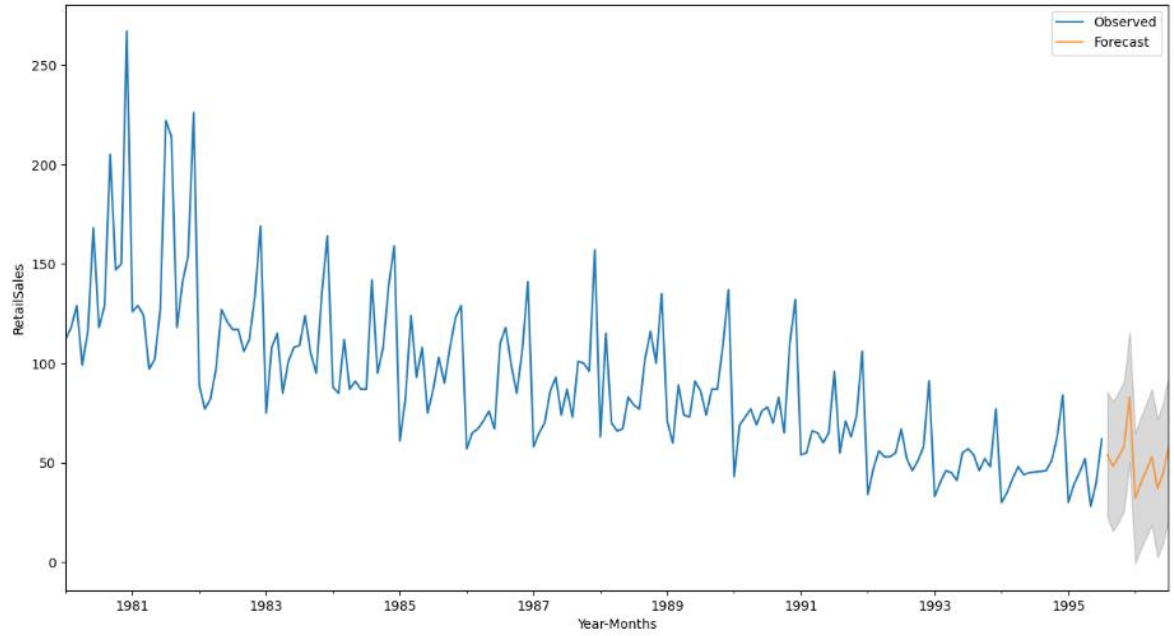


Figure 53: Prediction for the next 12 months with confidence intervals

The orange line traces the next 12 months forecast and as we can see the forecast indicates there the peak sales in December and steep decrease in month of January across all years.

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

Comments on Final Model

After building various time series forecasting models like Linear Regression , Naive Forecast ,Simple Average , Moving Average and exponential smoothing models like (Simple , Double , Triple Exponential) and ARIMA / SARIMA models on the Sparkling Wine Sales dataset and on comparing their RMSE on

the test data we deduce that the Test RMSE of SARIMA (2 1, 2)(1, 1,0, 12)) is least among all the models with different parameters.

So, we take Manual SARIMA(2 1, 2)(1, 1,0, 12)) model to build best fit mode on complete data taking into considerations of order and seasonality. and predict 12 months into the future with appropriate confidence intervals/bands. The forecast also indicates there the peak sales in December, and a drop in sales in the following months as the original dataset.

RMSE of the Full Model is 38.271

From the Plot of the forecast on full data along with the 95% confidence interval we infer that we that forecast also follows the same pattern as original Rose wine sales series follows.

Findings based on EDA / Data Visualization and Time Series Forecasting Models

- The data is from year 1980 to 1995 .
- The highest sales are recorded in the month of December across all years. Across years, from Sep sales started increasing but it ends in the month of December.
- The lowest sales are recorded in the month of January across all years.
- Sales is decreasing year on year as indicated by downward trend in the sales plot.
- Average sales are around is 89.92
- Minimum sales recorded for a month is 28.
- Maximum sales recorded for a month is 267.
- There are outliers in the sales data
- From the Plot of the forecast on full data along with the confidence band we infer that with 95% of the confidence level we found that forecast also follows the same pattern as original sparkling wine sales series follows.

Measures for Future Sales

- Sales is highest in December and lowest in January, it could be due to holiday and tourist season coming to an end. But there could be other factors due to which sales is dropping which is not available as part of the dataset .
- Various factors like wine quality, wine supply and demand, pricing, availability of better alternatives, not enough marketing of the product, shelf life could be reasons for drop in sales over the years. Company should to take measures to address these factors
- The ABC Estate Wines company should develop marketing strategies to promote Rose Wine Sales . During the months wine sales is low , company can run various offers during this period to boost their wine sales to attract more customers.
- Wine pairings are a great way to introduce customers to new choices. Many associate Rose Wine with dessert, so suggesting it with savoury dish might leave customers pleasantly surprised, so the company can let customers sample the Rose wine and have culinary experiences to promote Rose wine
- Give a variety of Rosé bottle sizes to offer to a customer. Having different bottle sizes is a great way to appeal to large groups and couples. You'll have an easier time selling bottles to groups if

they only have to buy 2 or 3, and couples can commit to smaller half bottles. Marketing studies state that if you make your product more accessible to your customers, they're more likely to buy. Serving different-sized bottles does exactly that

- Tying up with restaurants and hotels that serve alcohol to run offers on Rose wine for customers to try and suggest wine pairings, so that customers take a liking towards the Wine
- The staff should be approachable and knowledgeable enough to make informed recommendations and conversations about the wine. The staff members who are well informed will be more likely to confidently make a recommendation or upsell the wine to the customer. Teach them to sell the story and not just the wine and dramatically increases wine sales.
- Offering a box of miniature-wines which allows clients to purchase several of your wines as tasting samples. This allows customers to taste wine before committing to large purchases.
- Online marketplaces and Platforms, like Google Products, can make your wine more visible when someone googles a specific type of wine or similar products. Its free marketing and potential sales, to anyone interested in the exact product that you offer.
- Offering an efficient Wine Delivery-Service that delivers quickly and efficiently by allowing customers the opportunity to purchase wine online.
- Paid Ads are a way of targeting a particular group of people .Facebook ads are relatively affordable considering how targeted the advertising can be.
- Host Tasting Events and Offer Wine Subscription Boxes could be another way to boost sales

----------End of Report ----------