Time Series Forecasting-Rose Wine

25/02/2024

BALAJI S

PGP-DSBA

Module 8 - Time series

Problem:

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

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

YearMonth	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

YearMonth	Rose
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

Fig1.Heads & Tails of the Rose Dataset

- *There are 187 rows and 1 column*

Plot:

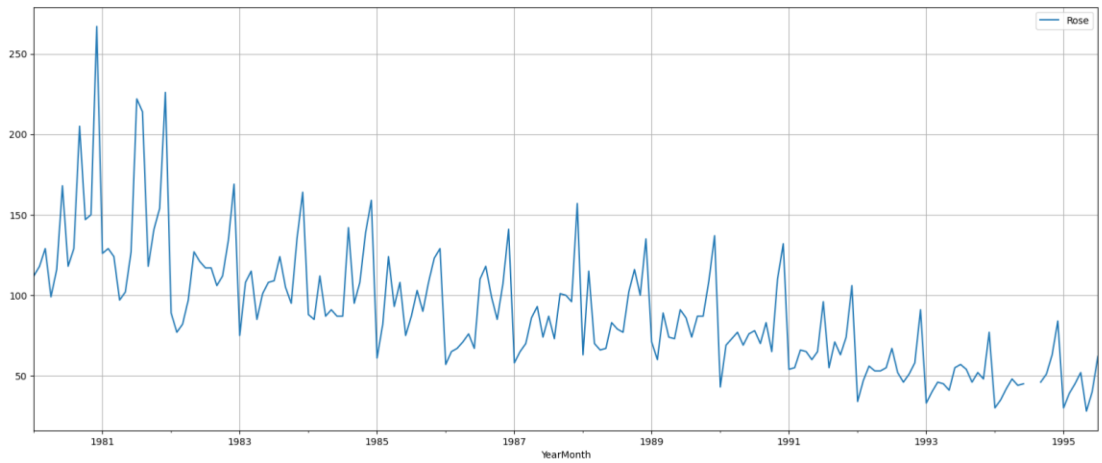


Fig 2: plot of the dataset

- Following the dataset ***ingestion*** process, we proceeded to refine its structure for enhanced analytical depth. We achieved this by segregating the dataset into distinct month and year categories, based on the information extracted from the 'YearMonth' column.

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

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

- Additionally, we renamed the 'Rose' column to 'Sales', a decision made to foster a clearer understanding and streamlined analysis of the dataset.

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

Data Type

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sales   185 non-null    float64
 1   Year    187 non-null    int64
 2   Month   187 non-null    int64
dtypes: float64(1), int64(2)
memory usage: 5.8 KB
```

Statistical Summary:

	count	mean	std	min	25%	50%	75%	max
Sales	185.0	90.0	39.0	28.0	63.0	86.0	112.0	267.0
Year	187.0	1987.0	5.0	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.0	3.0	1.0	3.0	6.0	9.0	12.0

Null values:

- There are two null values in the data set one is from August and September in the sales column found in the year 1994

```
Sales     2
Year      0
Month     0
dtype: int64
```

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

- "After exploring different methods to handle missing data, we employed the following approach:

- Mean Imputation - Before & After:
  Filling in missing values is critical for accurate analysis. Rather than using the mean from the 7th month across all years, we opted for a more precise approach. We calculated the mean of the 7th month using data from the same month in the preceding and following years. The same approach was utilized for imputing missing values in the 8th month.

```
Year      0
Month     0
Sales     0
dtype: int64
```

- This technique was chosen to preserve data accuracy and prepare it for further analysis.

Boxplot of the dataset:

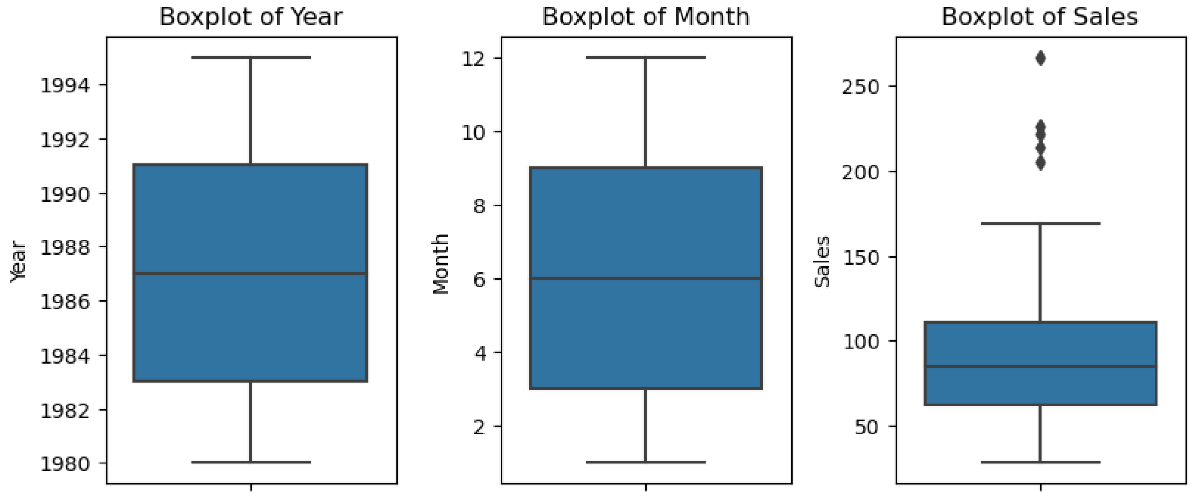
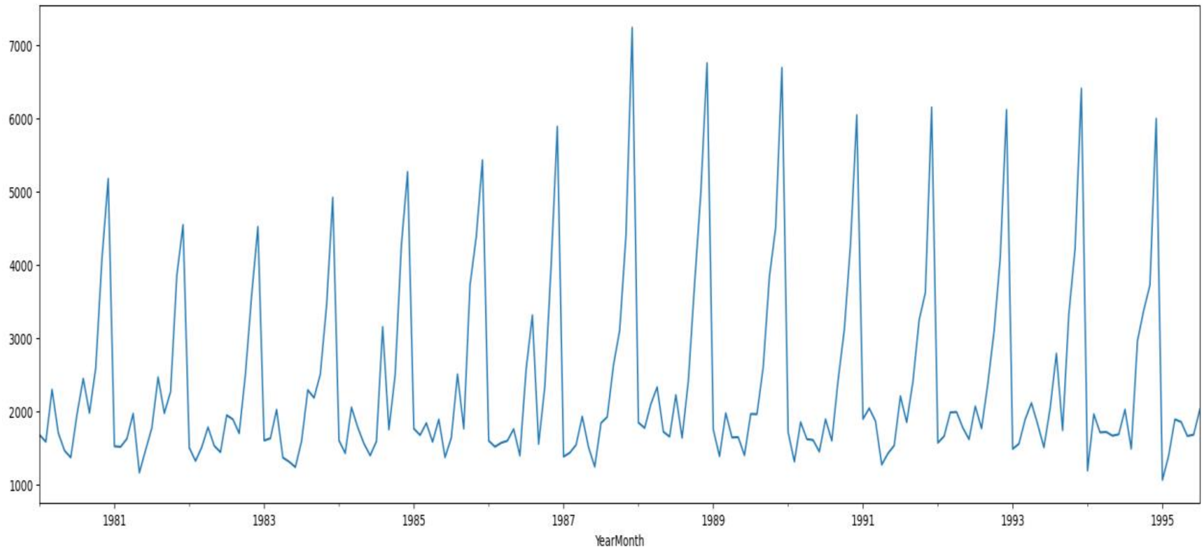


Fig 3: Boxplot

- The box plot reveals the presence of outliers in the 'Sales' data.
- Although these outliers could be addressed, we have decided against them as they have minimal impact on the time series model.
- Instead, our focus is on other aspects that significantly influence the model's performance.

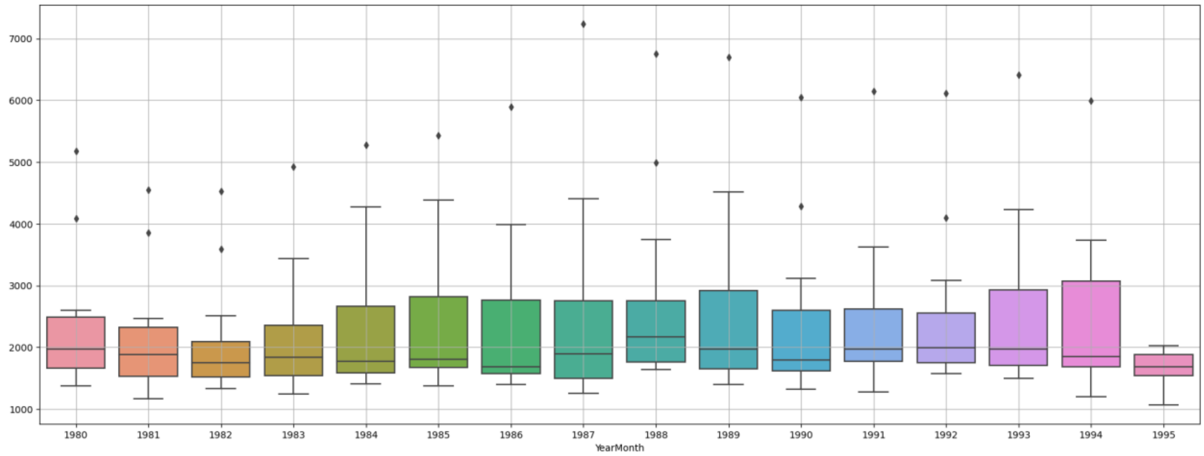
Line plot of sales:



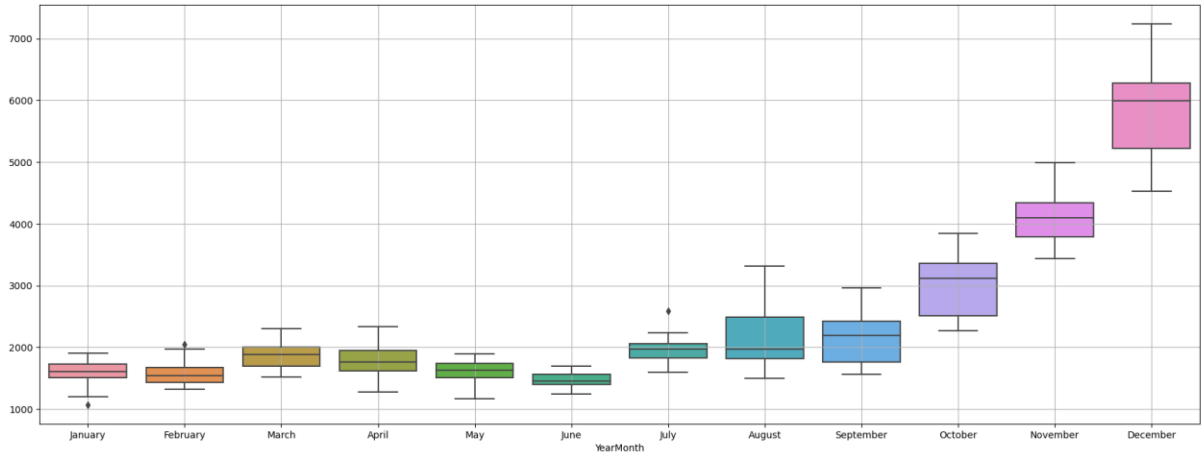
- *Seeing the line plot shows that there was a peak in 1997 to 1998*

Boxplot

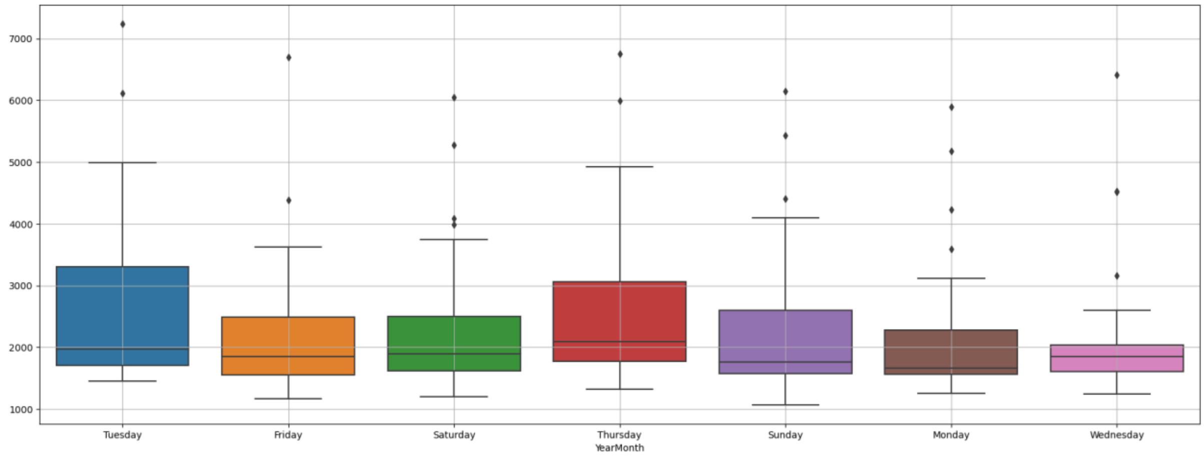
- ***yearly***



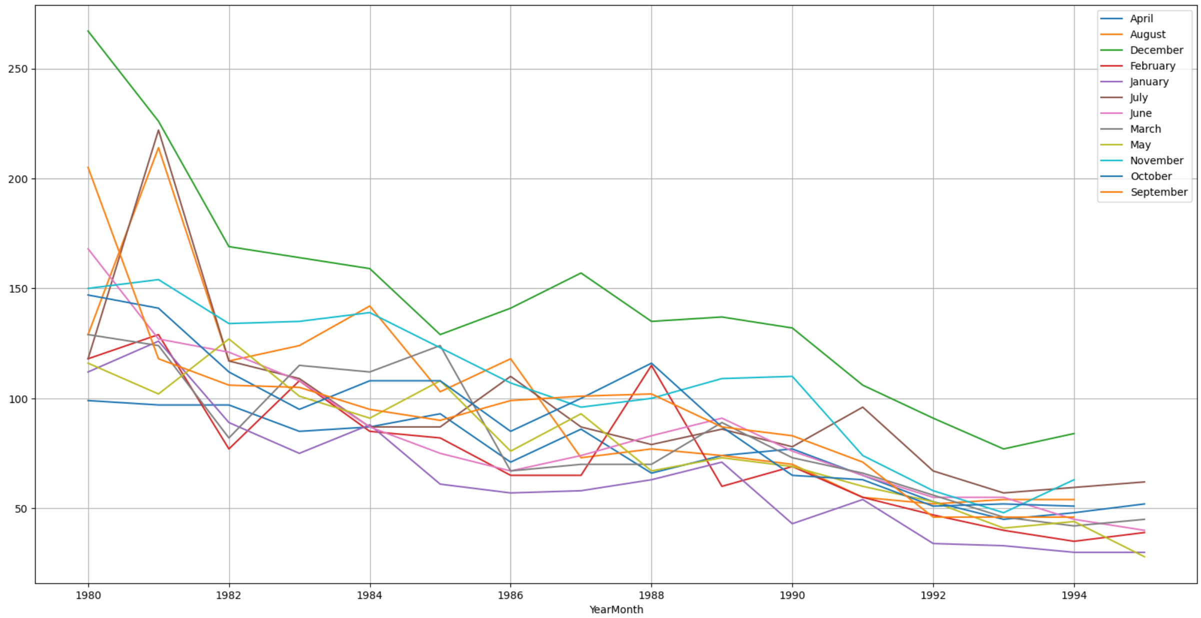
Monthly



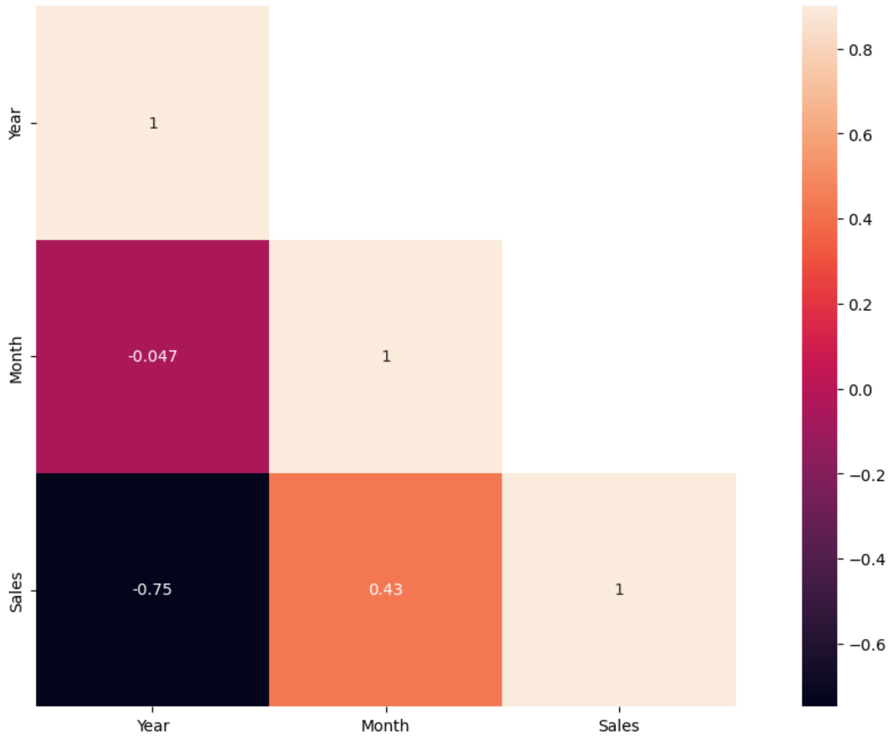
Weekly



Graph of Monthly sales across years



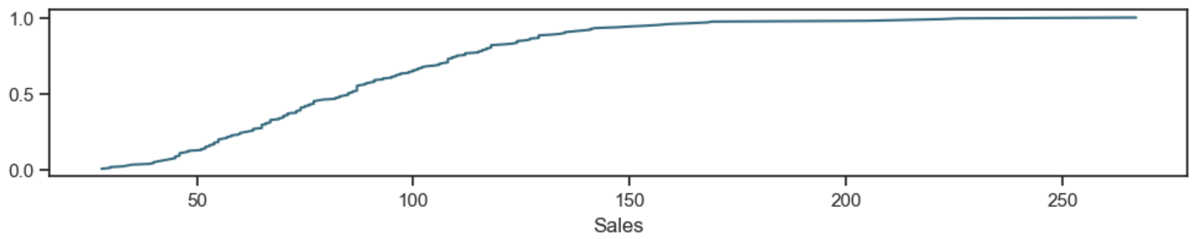
CORRELATION:



This heat map shows that there was little correlation between Sales and the Years data,
there was significantly more correlation between the month and Sales columns.

Clearly indicating a seasonal pattern in our Sales data. Certain months have higher sales, while certain months have fewer

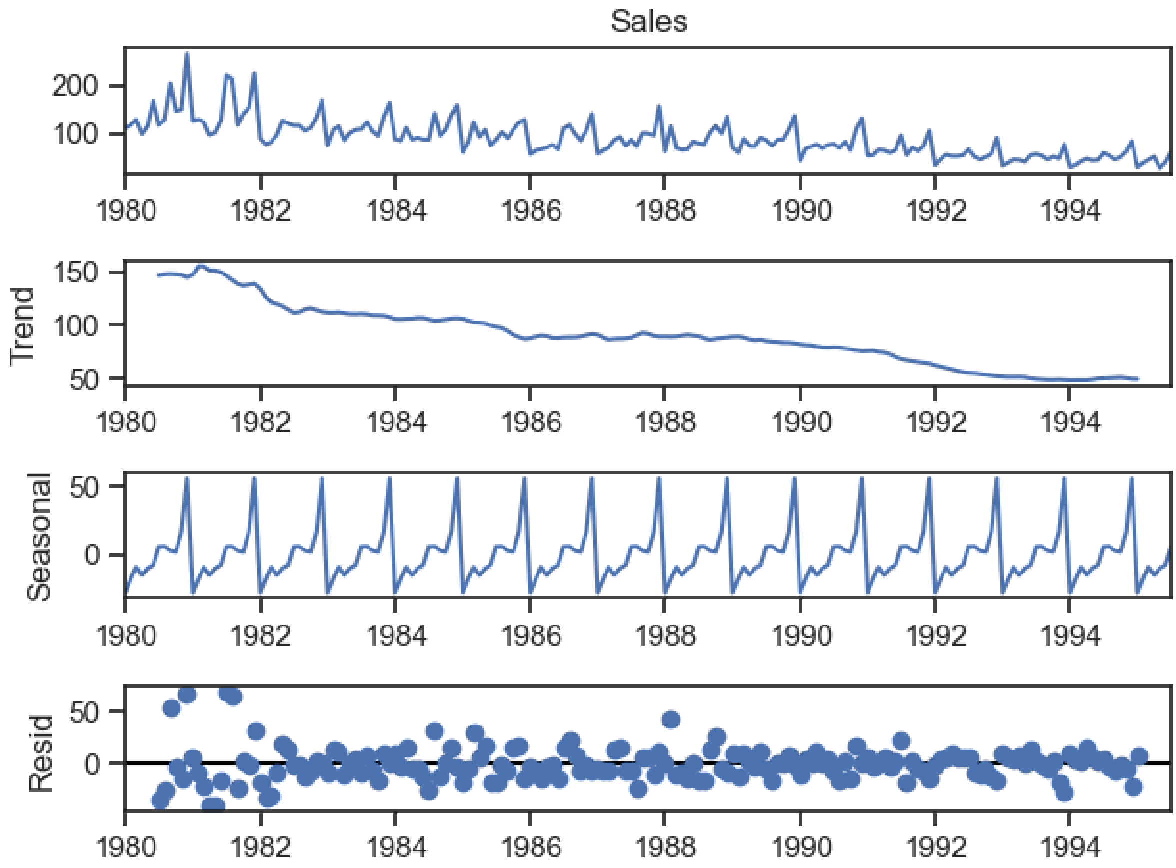
Plot ECDF: Empirical Cumulative Distribution Function.



This plot shows:

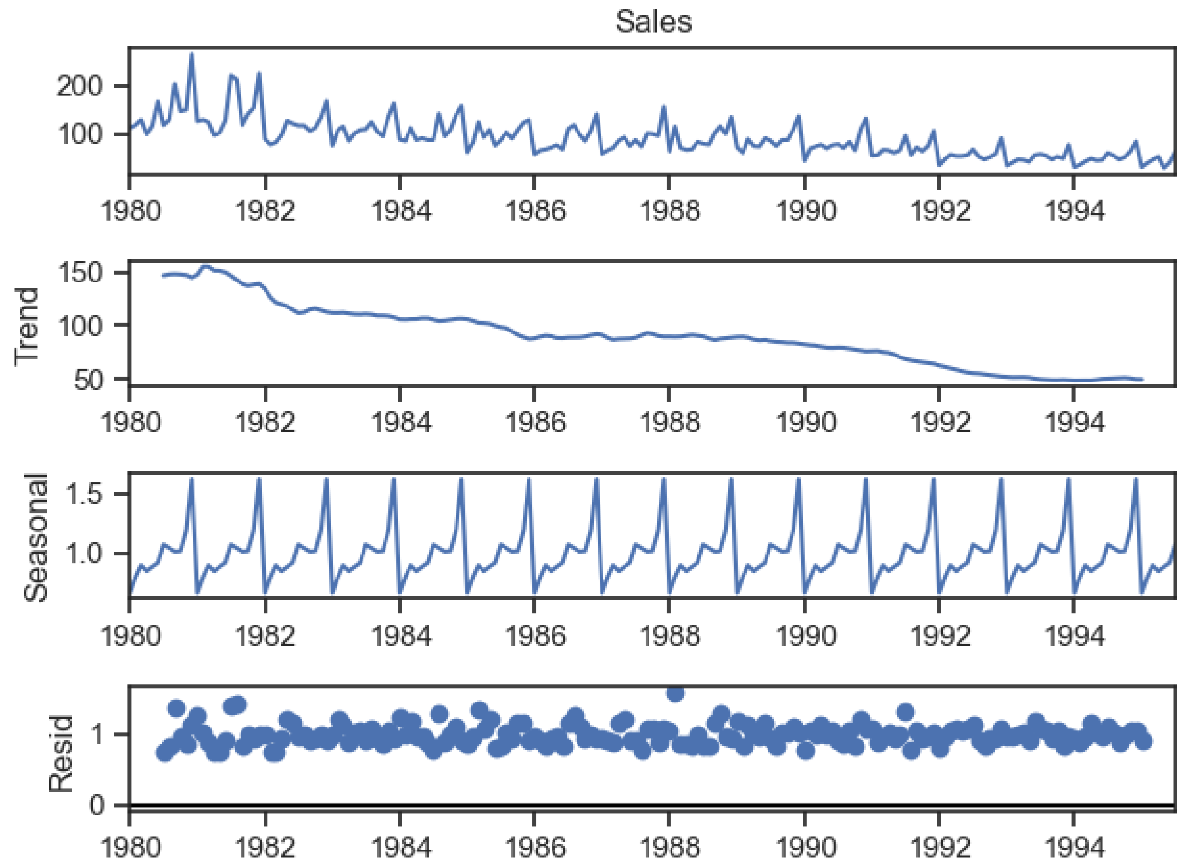
- 50% of sales has been less than 100
- Highest value is 250
- Approx 90% of sales has been less than 150

Decomposition -addictive



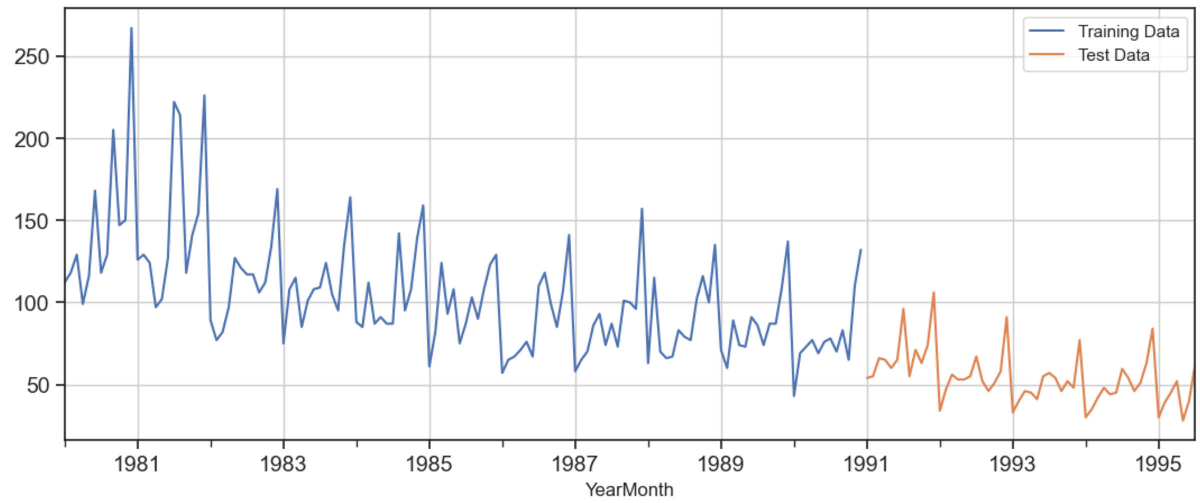
The plots show:
● Peak year 1981
● It also shows that the trend has declined over the year after 1981
● Residue is spread and is not in a straight line.
● Both trend and seasonality are present.

Decomposition -multiplicative



The plots show:
● Peak year 1981
● It also shows that the trend has declined over the year after 1981.
● Residue is spread and is in approximately a straight line.
● Both trend and seasonality are present.
● Reside is 0 to 1, while for additive is 0 to 50.
● So multiplicative model is selected owing to a more stable residual plot and lower range of residuals.

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



Data split from 1980-1990 is training data, then 1991 to 1995 is training data.

Rows and Columns:

- train dataset has 132 rows and 3 columns.
- test dataset has 55 and 3 columns.

```
Rows of dataset:
First few rows of Training Data
            Year  Month  Sales
YearMonth
1980-01-01  1980      1  112.0
1980-02-01  1980      2  118.0
1980-03-01  1980      3  129.0
1980-04-01  1980      4   99.0
1980-05-01  1980      5  116.0

Last few rows of Training Data
            Year  Month  Sales
YearMonth
1990-08-01  1990      8   70.0
1990-09-01  1990      9   83.0
1990-10-01  1990     10   65.0
1990-11-01  1990     11  110.0
1990-12-01  1990     12  132.0
```

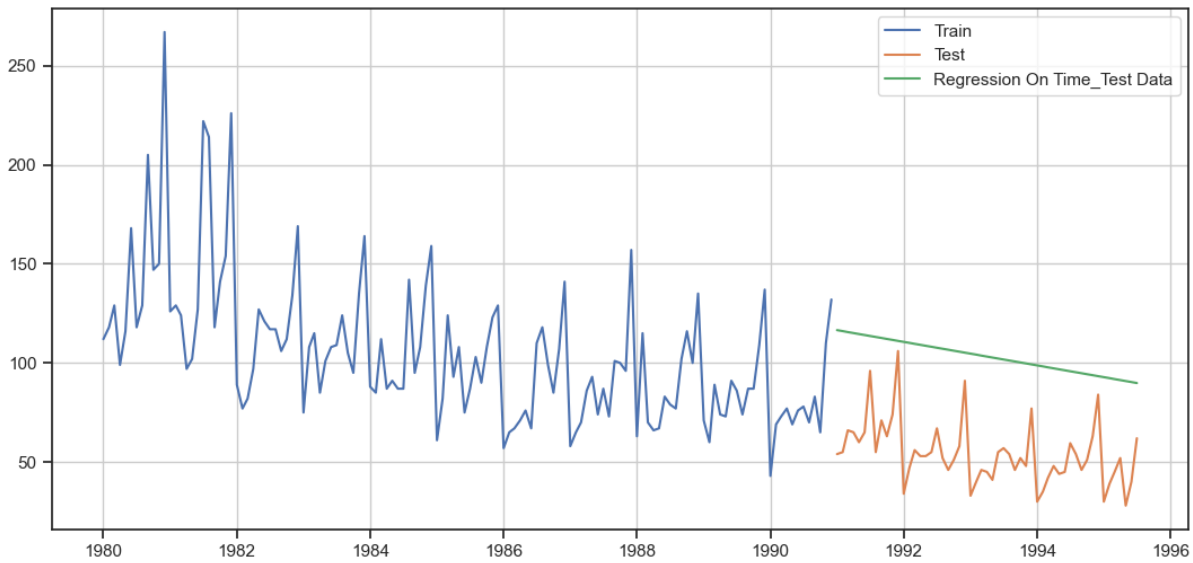
```
First few rows of Test Data
            Year  Month  Sales
YearMonth
1991-01-01  1991      1   54.0
1991-02-01  1991      2   55.0
1991-03-01  1991      3   66.0
1991-04-01  1991      4   65.0
1991-05-01  1991      5   60.0

Last few rows of Test Data
            Year  Month  Sales
YearMonth
1995-03-01  1995      3   45.0
1995-04-01  1995      4   52.0
1995-05-01  1995      5   28.0
1995-06-01  1995      6   40.0
1995-07-01  1995      7   62.0
```

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.

- Model 1:Linear Regression
- Model 2: Naive Approach
- Model 3: Simple Average
- Model 4: Moving Average(MA)
- Model 5: Simple Exponential Smoothing
- Model 6: Double Exponential Smoothing (Holt's Model)
- Model 7: Triple Exponential Smoothing (Holt - Winter's Model)

LINEAR REGRESSION

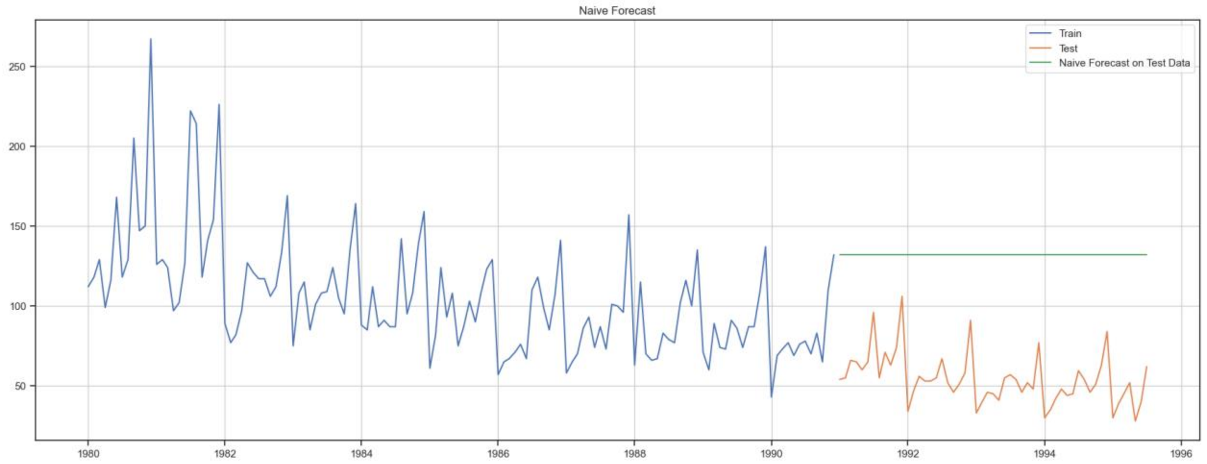


The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

	Test RMSE
Linear Regression	51.080941

NAÏVE APPROACH

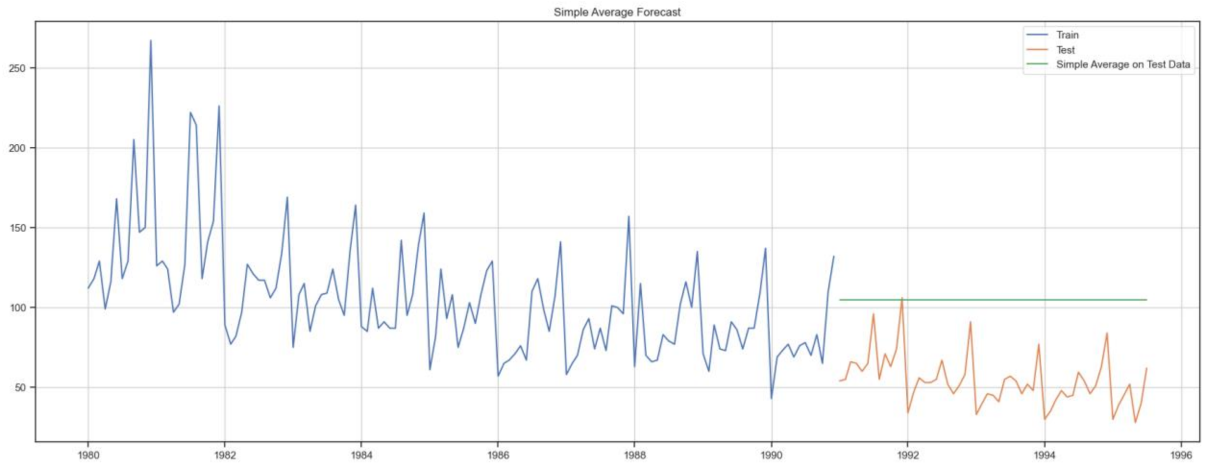


The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

The model was evaluated using the RMSE metric. Below is the

RMSE calculated for this model: **Naive Model** 79.304391

SIMPLE AVERAGE:

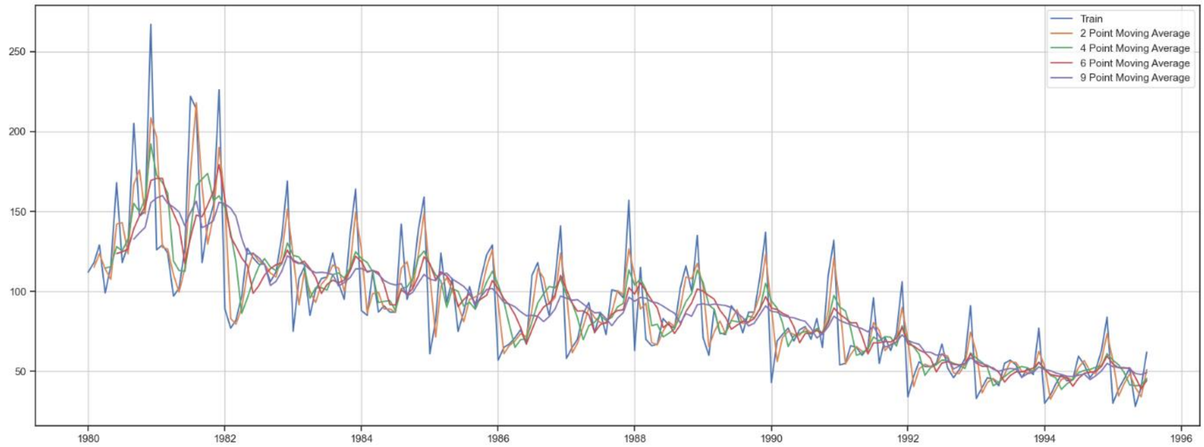


The green line indicates the predictions made by the model, while the orange values are the actual test values. It is clear the predicted values are very far off from the actual values

Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

Simple Average Model	53.049755

MOVING AVERAGE:

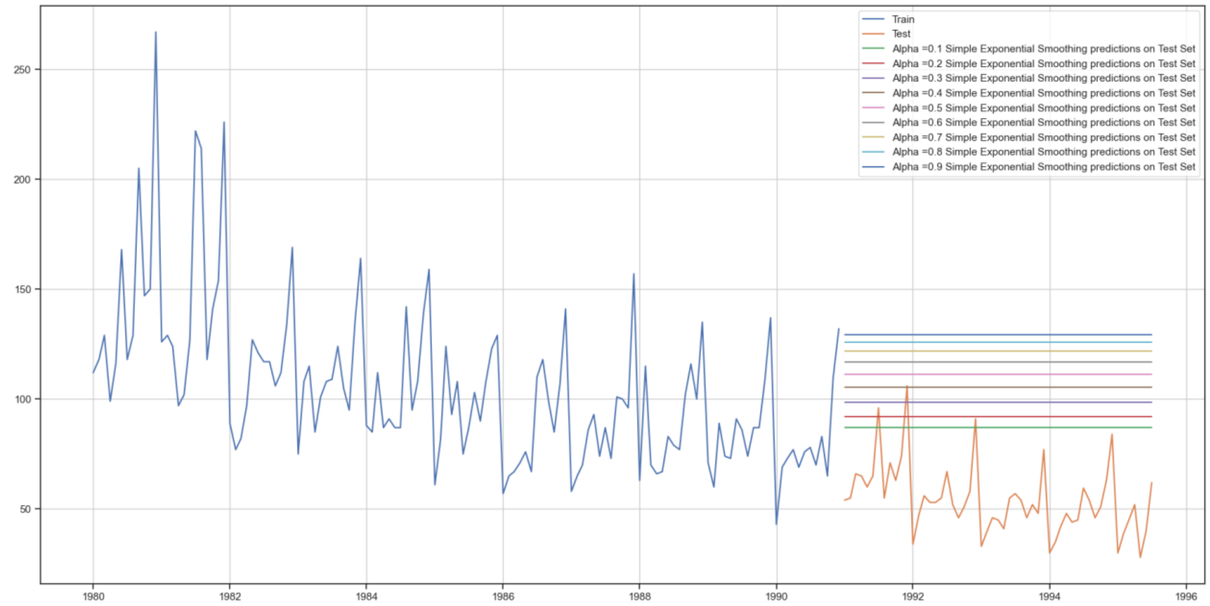


Model was evaluated using the RMSE metric. Below is the RMSE calculated for this model:

2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139

- We created multiple moving average models with rolling windows varying from 2 to 9.
- Rolling average is a better method than simple average as it takes into account only the previous n values to make the prediction, where n is the rolling window defined.
- This takes into account the recent trends and is in general more accurate.
- The higher the rolling window, the smoother it will be its curve, since more values are being taken into account

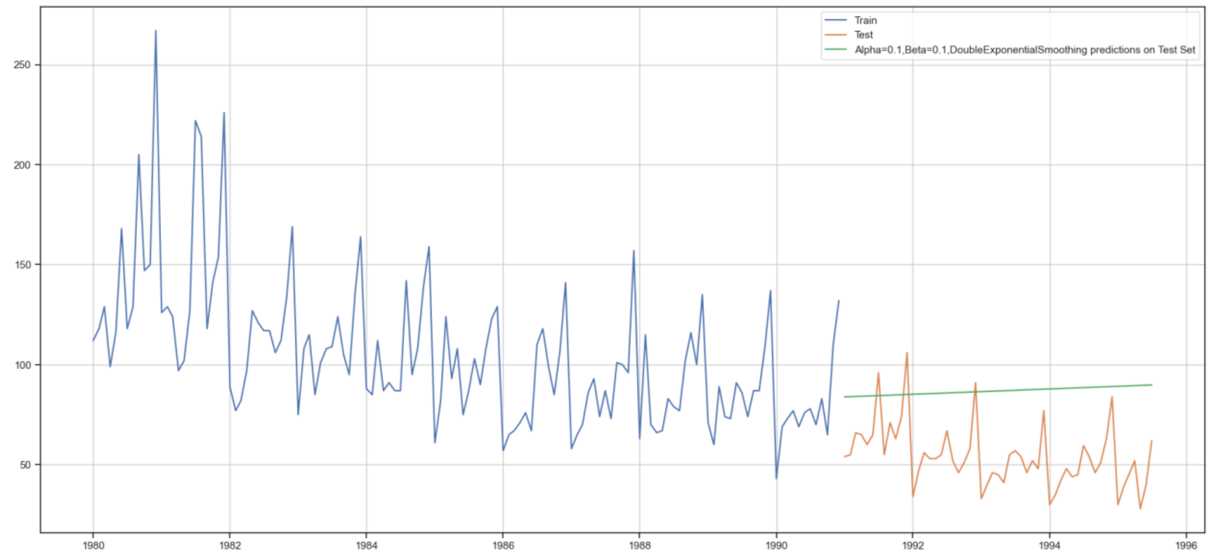
SIMPLE EXPONENTIAL SMOOTHING:



The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha=0.1,SimpleExponentialSmoothing	36.429535

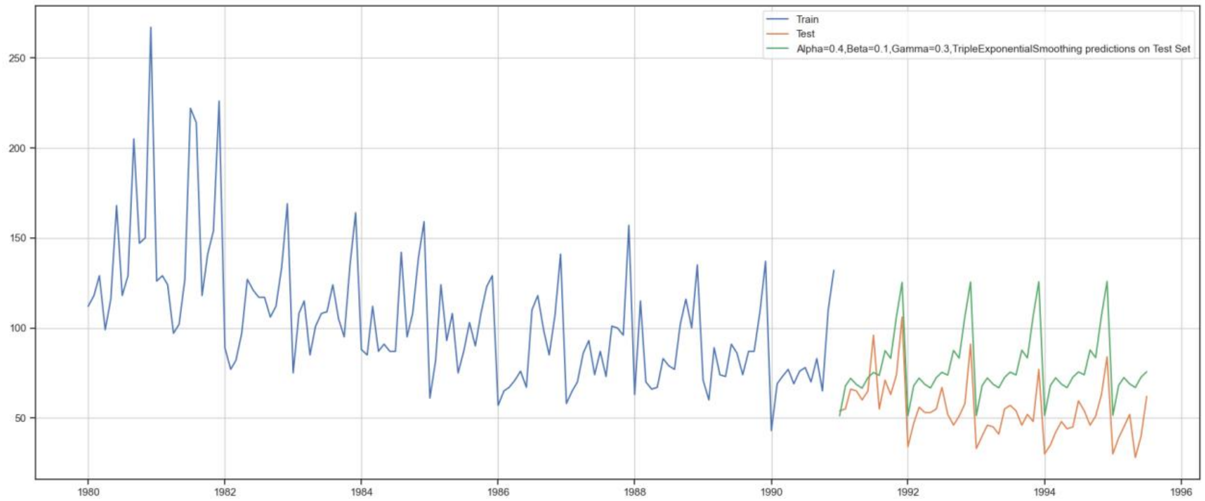
Double Exponential smoothing(Holt's model)



The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010

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



Output for best alpha, beta, and gamma values is shown by the green color line in the
above plot. The best model had both multiplicative trend as well as seasonality. So far this is the best model

The model was evaluated using the RMSE metric. Below is the RMSE calculated for this model.

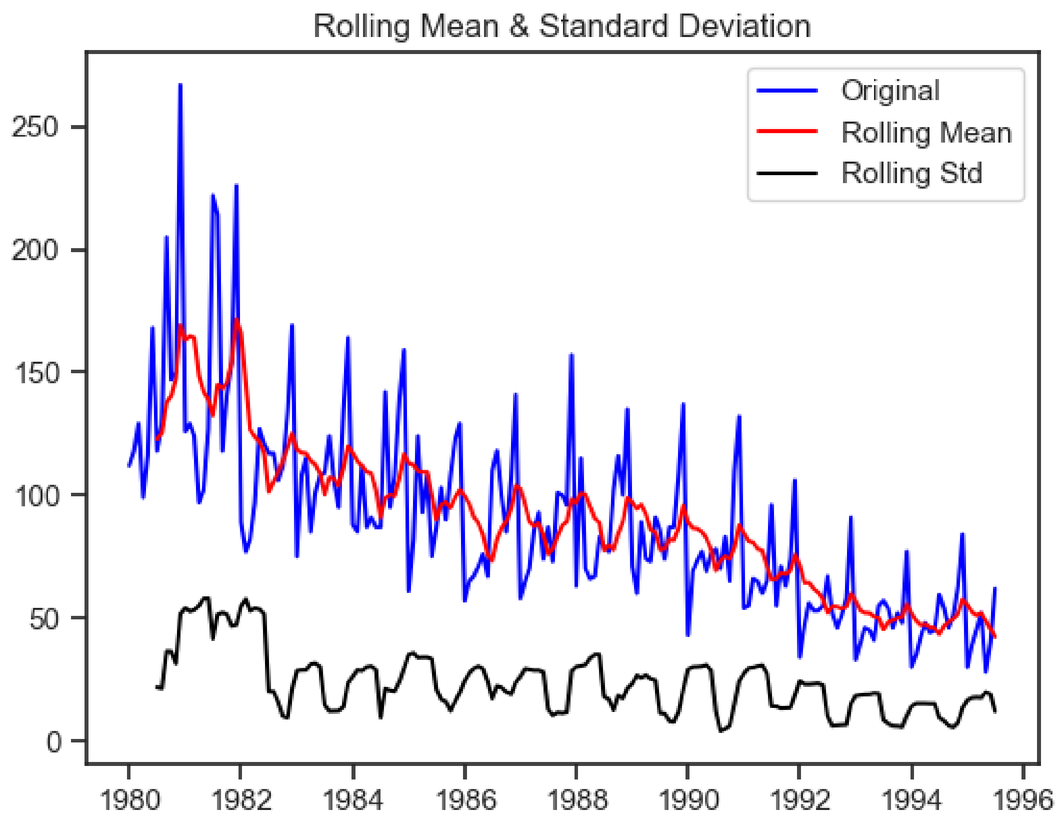
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.992350

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

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

● H0 : The Time Series has a unit root and is thus non-stationary.

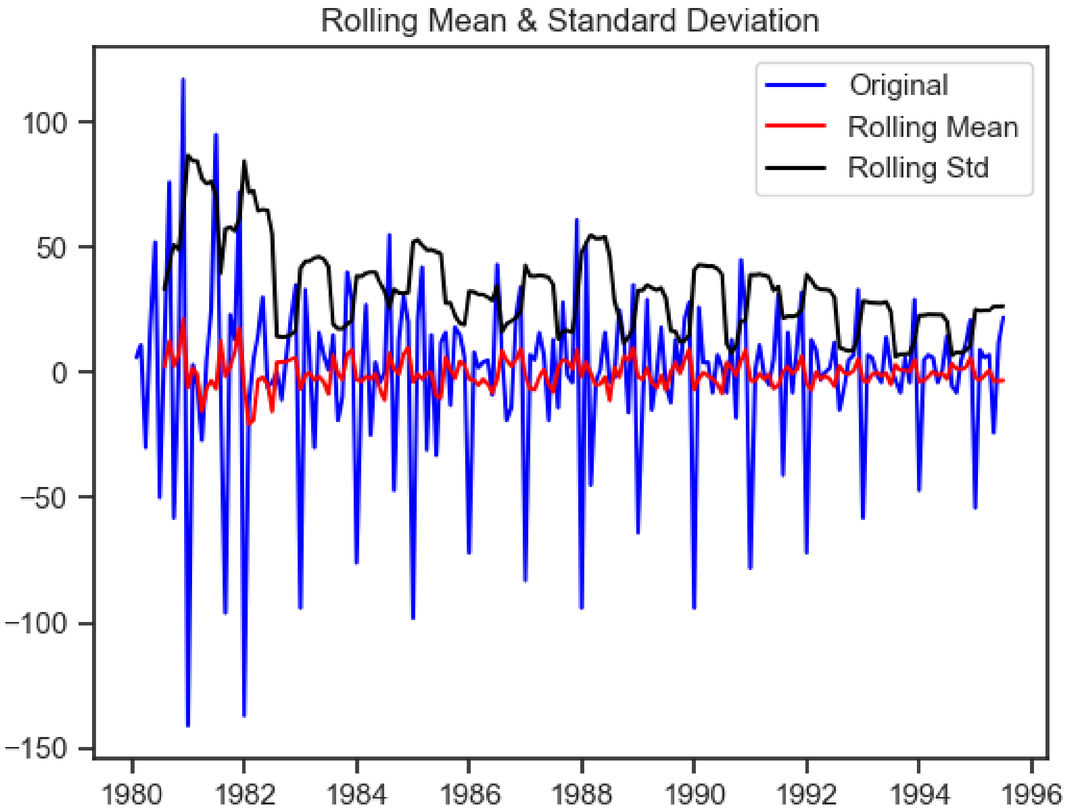
● H1 : The Time Series does not have a unit root and is thus stationary.
We would want the series to be stationary for building ARIMA models and thus we would want the
p-value of this test to be less than the α value.
We see that at 5% significant level the Time Series is non-stationary.




```
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 failed to reject the null hypothesis, which implies the Series is not stationary.

- To try and make the series stationary we used the differencing approach.
- We used the .diff() function on the existing series without any argument, implying the default diff value of 1, and also dropped the NaN values, since differencing of order 1 would generate the
- First value as NaN which needs to be dropped




```
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
```

- The null hypothesis that the series is not stationary at difference = 1 was rejected, which implied that the series has indeed become stationary after we performed the differencing.
- We could now proceed ahead with ARIMA/ SARIMA models since we had made the series stationary

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.

AUTO - ARIMA model

We employed a for loop for determining the optimum values of p,d,q, where p is the order of the AR (Auto-Regressive) part of the model, while q is the order of the MA (Moving Average) part of the model. d is the differencing that is required to make the series stationary. p,q values in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for d since we had already determined d to be 1 while checking for stationarity using the ADF test.

Some parameter combinations for the Model

```
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)
```

Akaike information criterion (AIC) value was evaluated for each of these models and the model with the least AIC value was selected

	param	AIC
11	(2, 1, 3)	1274.695692
15	(3, 1, 3)	1278.654372
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
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

The summary report for the ARIMA model with values (p=2,d=1,q=3).

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 25 Feb 2024   AIC                           1274.696
Time:                        14:54:12   BIC                           1291.947
Sample:                    01-01-1980   HQIC                          1281.706
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6778      0.084    -20.027      0.000      -1.842      -1.514
ar.L2         -0.7286      0.084     -8.698      0.000      -0.893      -0.564
ma.L1          1.0444      0.602      1.734      0.083      -0.136       2.225
ma.L2         -0.7722      0.130     -5.924      0.000      -1.028      -0.517
ma.L3         -0.9046      0.546     -1.658      0.097      -1.974       0.165
sigma2       859.1641    505.896      1.698      0.089    -132.374    1850.703
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.47
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================
```

RMSE values are as below: 36.415310298964606

AUTO- SARIMA Model

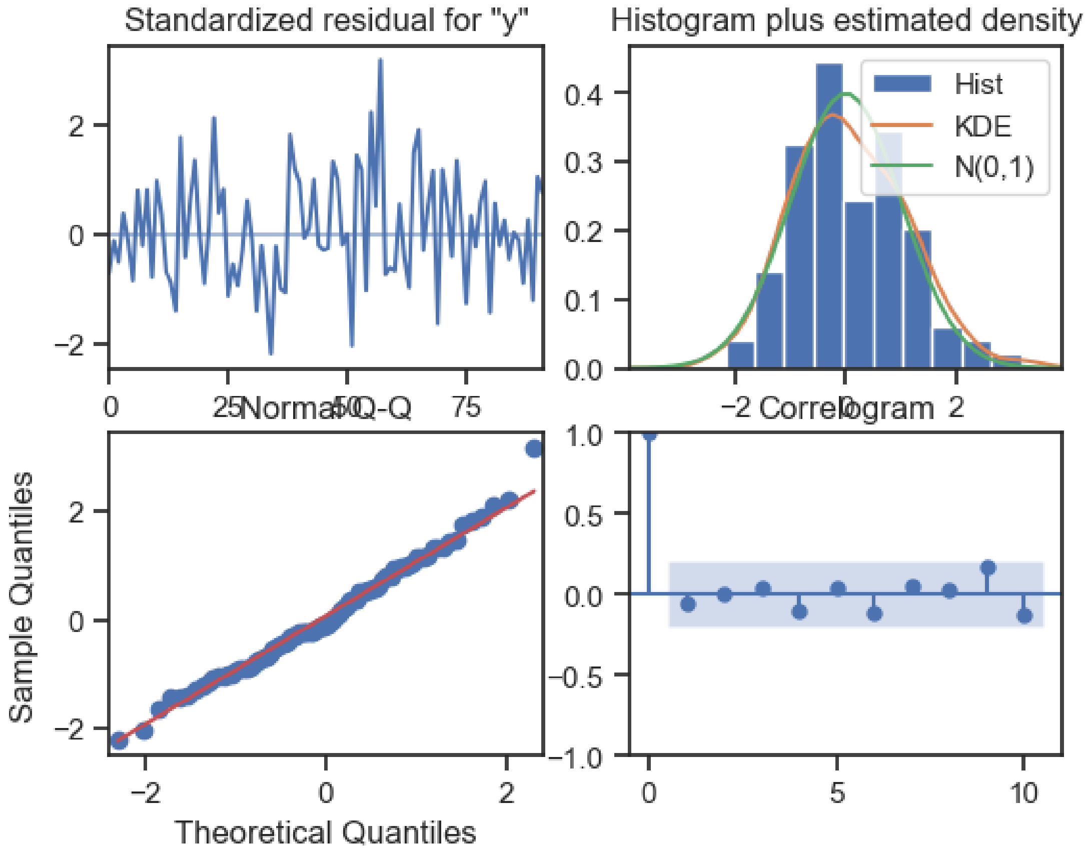
A similar for loop like AUTO_ARIMA with the below values was employed, resulting in the models shown below.

p = q = range(0, 4) d= range(0,2) D = range(0,2) pdq = list(itertools.product(p, d, q))
model_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, D, q))]

```
Examples of some parameter combinations for 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)
```

```
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)   Log Likelihood                -377.200
Date:                            Sun, 25 Feb 2024   AIC                            774.400
Time:                                    14:59:30   BIC                            799.618
Sample:                                         0   HQIC                           784.578
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0464      0.126      0.367      0.714      -0.202       0.294
ar.L2         -0.0060      0.120     -0.050      0.960      -0.241       0.229
ar.L3         -0.1808      0.098     -1.838      0.066      -0.374       0.012
ma.L1         -0.9370      0.067    -13.905      0.000      -1.069      -0.805
ar.S.L12       0.7639      0.165      4.640      0.000       0.441       1.087
ar.S.L24       0.0840      0.159      0.527      0.598      -0.229       0.397
ar.S.L36       0.0727      0.095      0.764      0.445      -0.114       0.259
ma.S.L12      -0.4969      0.250     -1.988      0.047      -0.987      -0.007
ma.S.L24      -0.2191      0.210     -1.044      0.296      -0.630       0.192
sigma2       192.1509     39.627      4.849      0.000     114.484     269.818
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                 1.64
Prob(Q):                              0.58   Prob(JB):                         0.44
Heteroskedasticity (H):               1.11   Skew:                             0.33
Prob(H) (two-sided):                  0.77   Kurtosis:                         3.03
-----------------------------------------------------------------------------------
```

We also plotted the graphs for the residual to determine if any further information can be extracted or if all the usable information has already been extracted. Below are the plots
for the best auto SARIMA model.

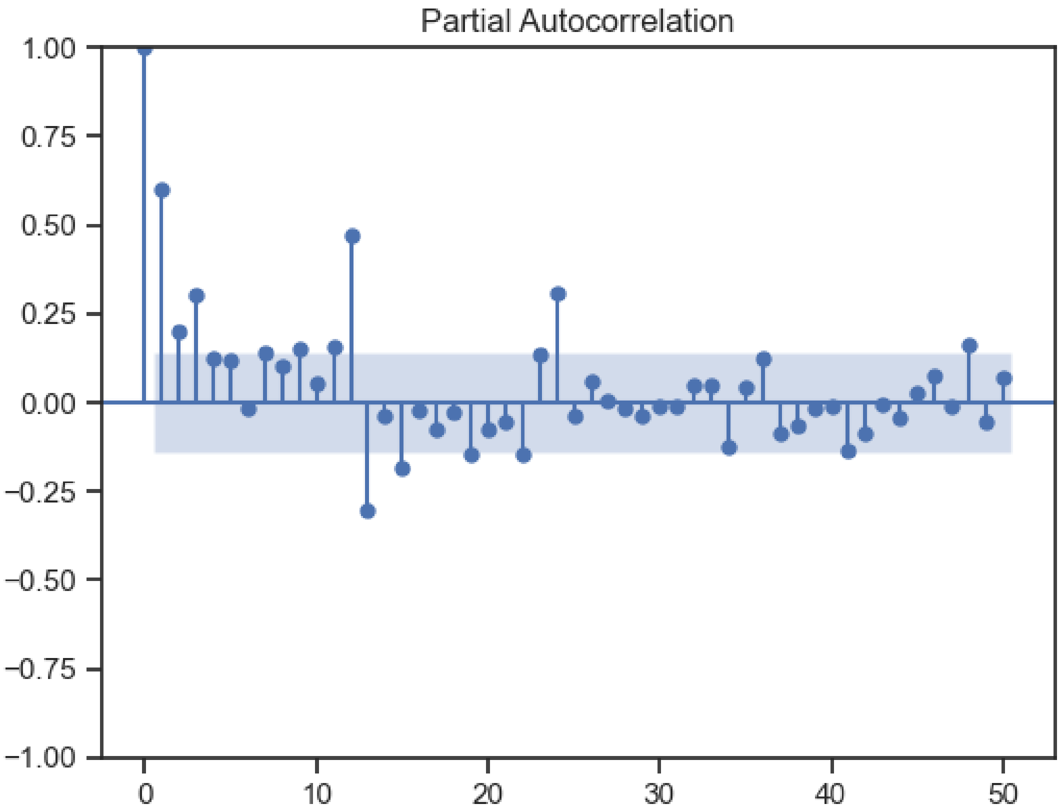


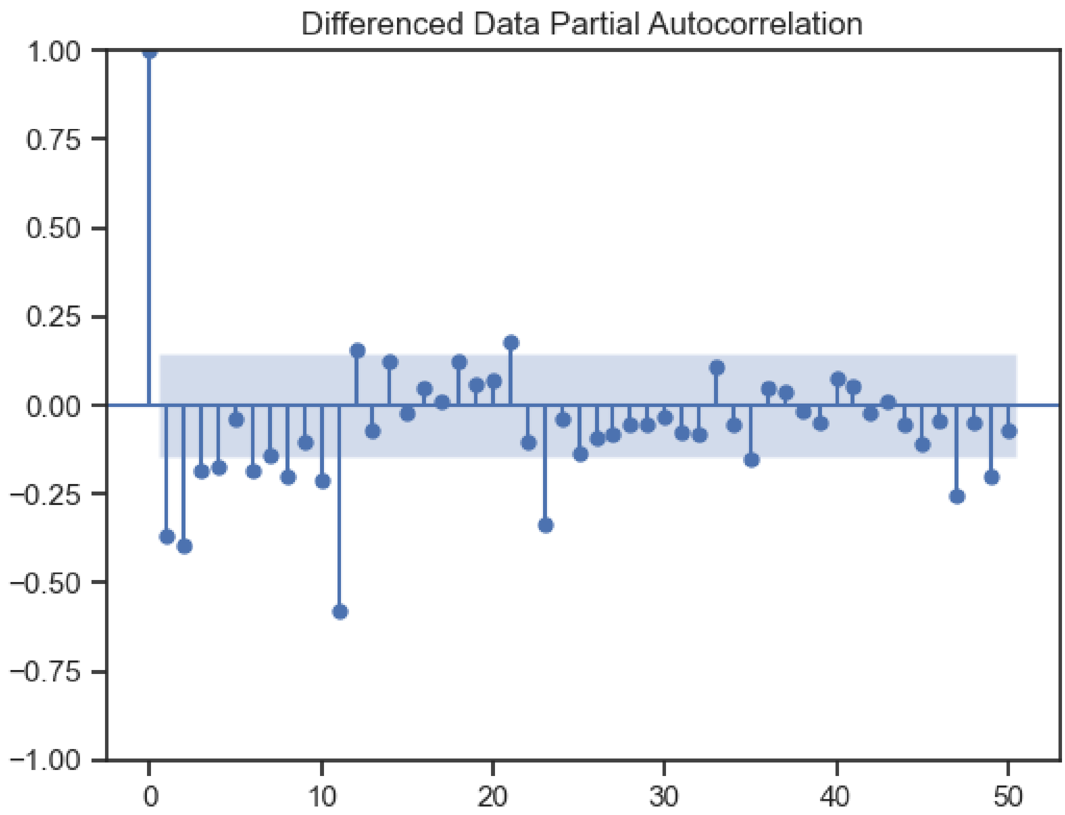
RSME of Model: 18.53560834178629

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

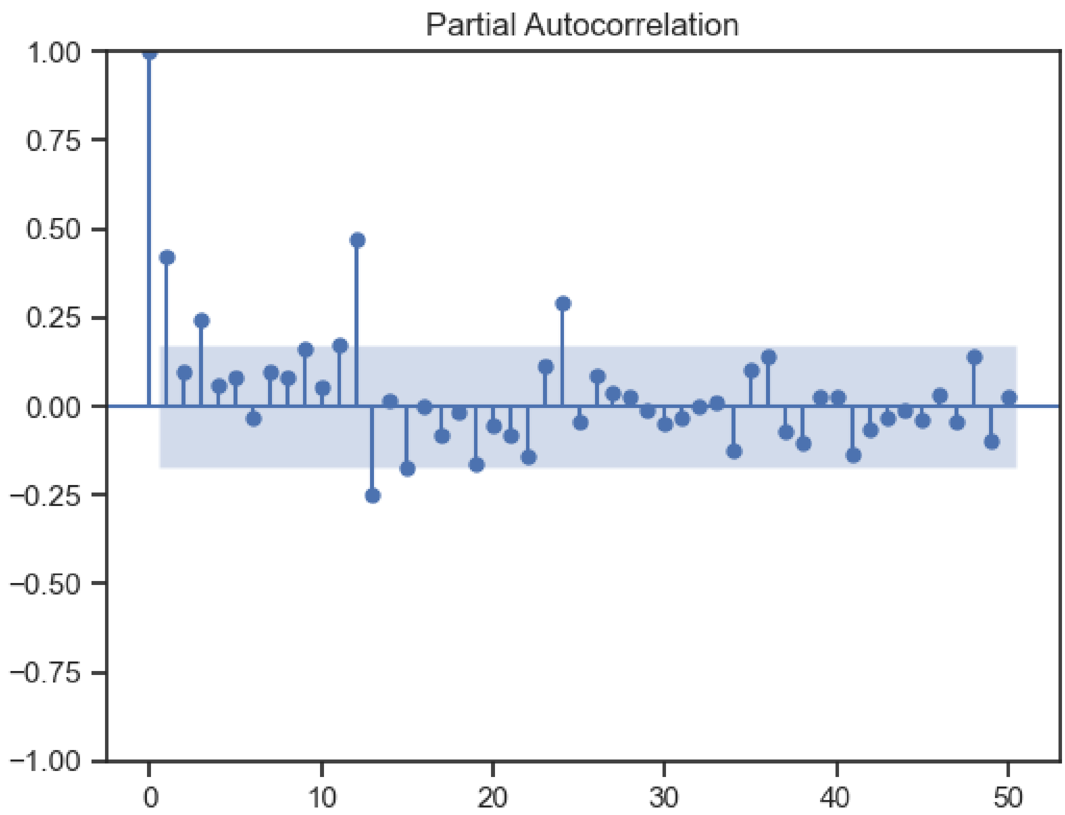
Manual- ARIMA Model

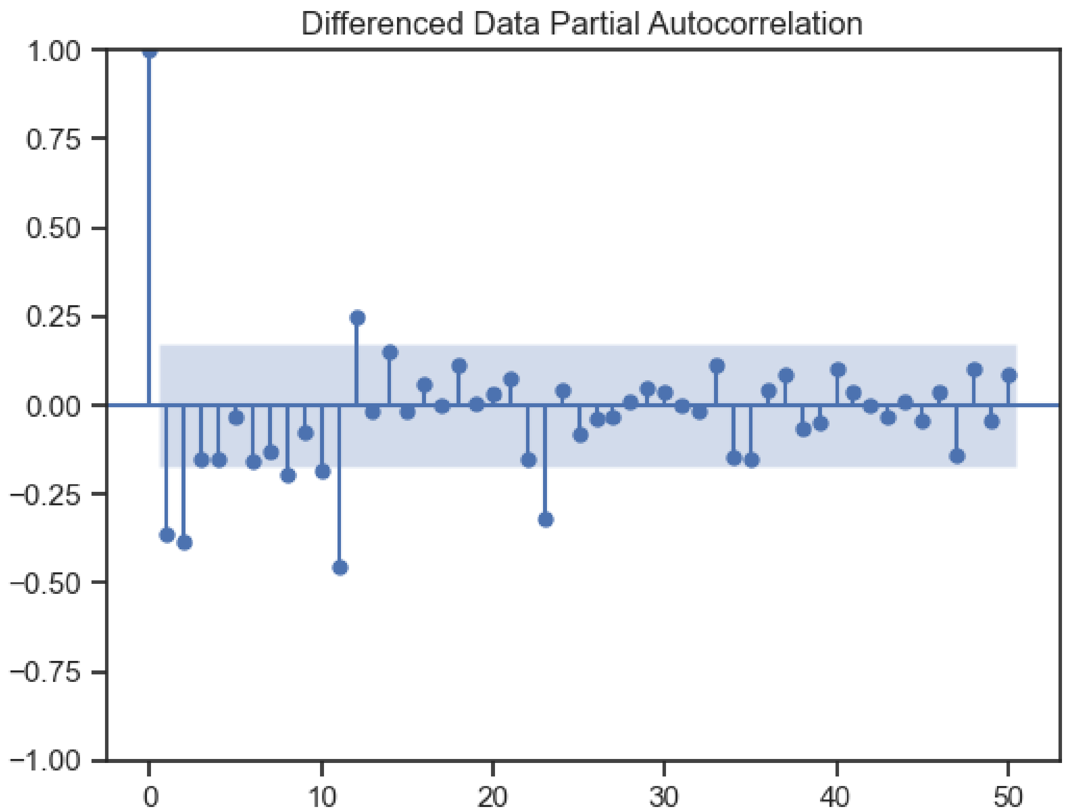
PACF the ACF plot on data :





PACF and ACF plot of train date:



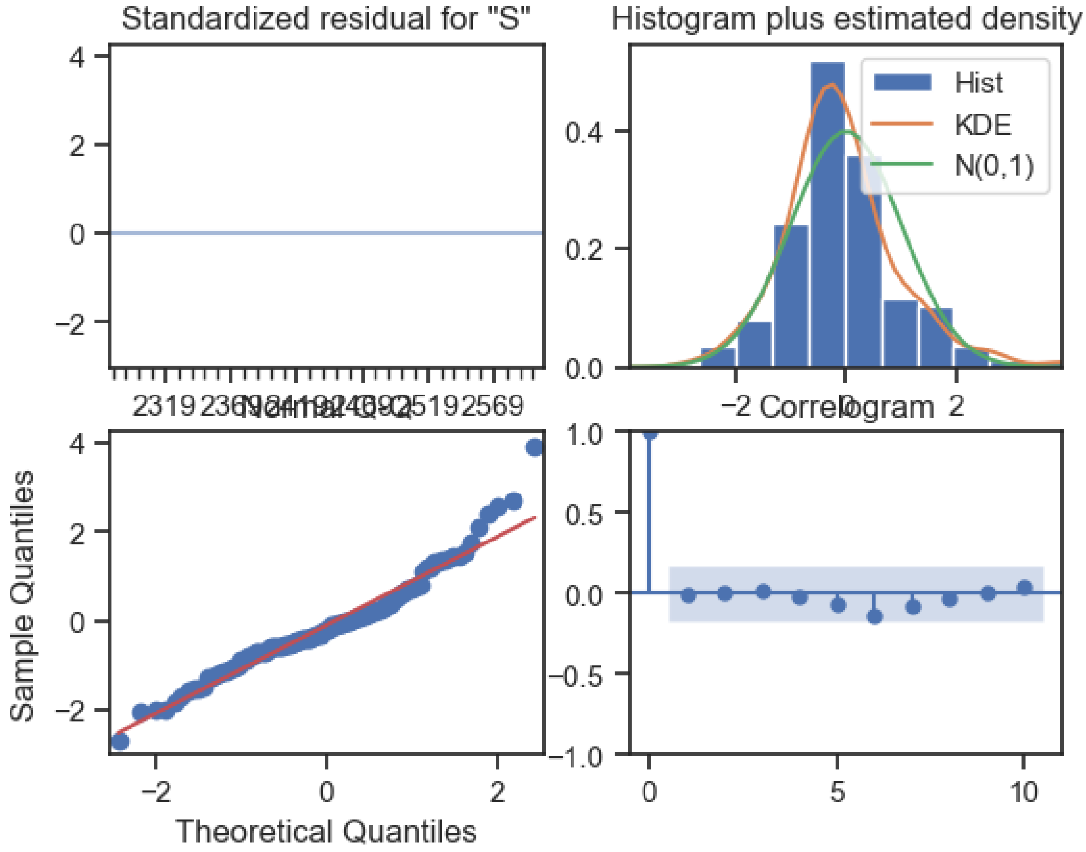


Hence the values selected for manual ARIMA:- p=2, d=1, q=2

summary from this manual ARIMA model:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Sun, 25 Feb 2024   AIC                           1281.871
Time:                        14:59:33   BIC                           1296.247
Sample:                    01-01-1980   HQIC                          1287.712
                         - 12-01-1990
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
===================================================================================
```

manual arima model plots:



Model Evaluation: RSME: `RMSE: 36.473224886646065`

Manual SARIMA Model

Looking at the ACF and PACF plots for training data, we can clearly see significant spikes at lags 12,24,36,48 etc, indicating a seasonality of 12. The parameters used for manual

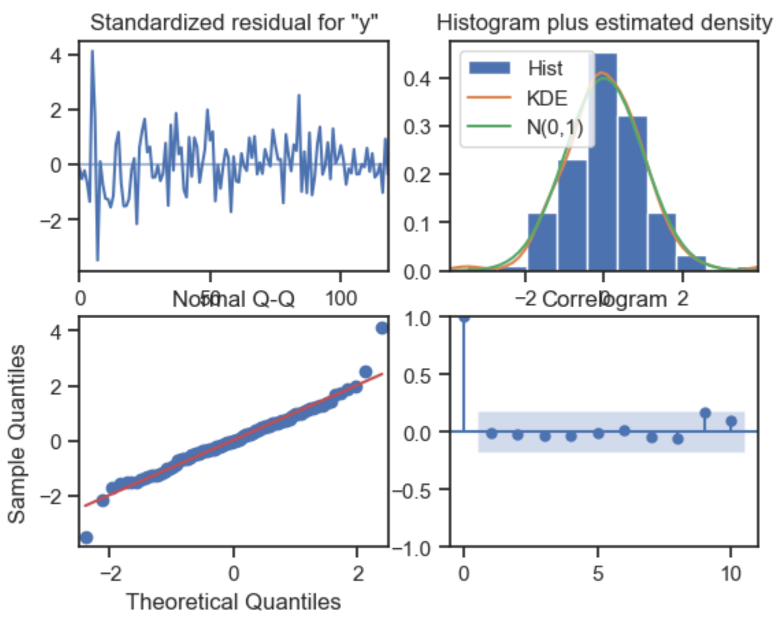
SARIMA model are as below.

SARIMAX(2, 1, 2)x(2, 1, 2, 12)

Below is the summary of the manual SARIMA model

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 1, 2, 12)   Log Likelihood                -538.016
Date:                            Sun, 25 Feb 2024   AIC                           1094.031
Time:                                    14:59:36   BIC                           1119.044
Sample:                                         0   HQIC                          1104.188
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5491      0.228     -2.408      0.016      -0.996      -0.102
ar.L2         -0.0744      0.099     -0.753      0.451      -0.268       0.119
ma.L1         -0.1703      0.216     -0.787      0.431      -0.594       0.254
ma.L2         -0.6694      0.228     -2.937      0.003      -1.116      -0.223
ar.S.L12      -1.0135      0.524     -1.935      0.053      -2.040       0.013
ar.S.L24      -0.1003      0.175     -0.572      0.567      -0.444       0.243
ma.S.L12       0.2906     20.998      0.014      0.989     -40.864      41.445
ma.S.L24      -0.7076     14.965     -0.047      0.962     -30.038      28.623
sigma2       430.5088   8838.340      0.049      0.961   -1.69e+04    1.78e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                27.15
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               0.33   Skew:                             0.26
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.28
```

manula sarima plots:



Model Evaluation: RSME 14.974952993897551

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

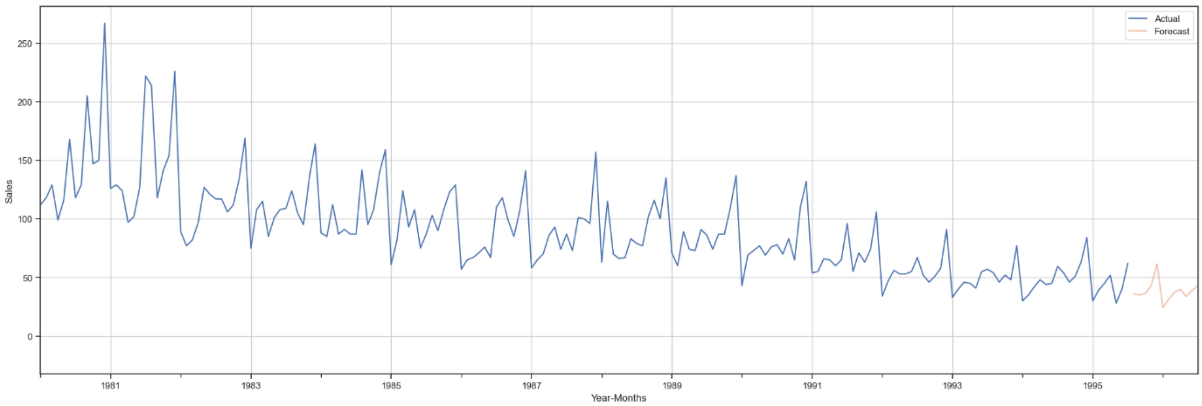
	Test RMSE
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.992350
2pointTrailingMovingAverage	11.589082
4pointTrailingMovingAverage	14.506190
6pointTrailingMovingAverage	14.558008
9pointTrailingMovingAverage	14.797139
(2,1,2)(2,1,2,12),Manual_SARIMA	14.974953
(3,1,1),(3,0,2,12),Auto_SARIMA	18.535608
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	36.397793
Auto_ARIMA	36.415310
Alpha=0.1,SimpleExponentialSmoothing	36.429535
ARIMA(3,1,3)	36.473225
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.510010
Linear Regression	51.080941
Simple Average Model	53.049755
Naive Model	79.304391

We can see that the triple exponential smoothing model with ***alpha 0.1, beta 0.7, and gamma 0.2*** *is the best as it he the <u>lowest</u> RSME score.*

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

	Sales_Predictions
1995-08-01	36.096841
1995-09-01	34.999961
1995-10-01	36.289937
1995-11-01	43.126839
1995-12-01	61.593978
1996-01-01	24.293852
1996-02-01	31.406019
1996-03-01	37.545514
1996-04-01	39.735393
1996-05-01	33.753457
1996-06-01	38.868148
1996-07-01	43.093112

The sales prediction on the graph along with the confidence intervals. PFB the graph.



Predictions, 1 year into the future are shown in orange color, while the confidence interval has been shown in grey color.

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

- Rose wine sales have been consistently declining for more than a decade and this trend is expected to continue in the future, according to our most reliable predictions.
- Wine sales fluctuate seasonally, with peaks during festive periods and drops during winter, notably in January.
- The company should focus on campaigns to boost sales during lean periods, particularly between April and June.
- Campaigns during peak periods, like festivals, may not yield significant gains since sales are already high then.
- Campaigns during peak winter (January) are discouraged since weather conditions deter people from buying wine.

It's vital to assess the reasons behind the decline in Rose wine popularity and possibly alter production and marketing strategies to regain market share.