Business Report

Time Series Forecasting

Sparkling Wine Sales

Ayush Sharma

Table of Contents

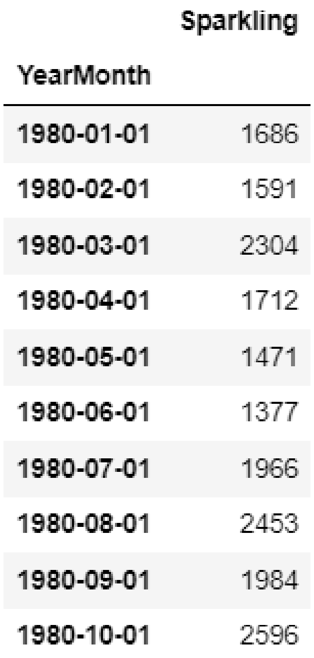
A. Read the data as an appropriate Time Series data and plot the data.	2-3
B. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.	4-6
C. Split the data into training and test. The test data should start in 1991.	7
D. Build all the exponential smoothing 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.	8-9
E. 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.	10
F. 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.	11-14
G. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.	15
H. 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.	16
I. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.	17

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

Ans: Following inferences can be made about the dataset:

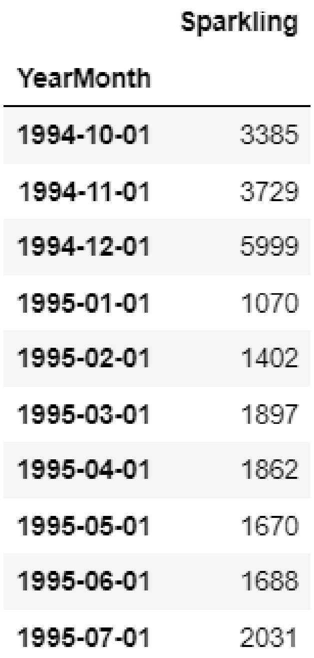
- The Sparkling Wine dataset consists of 187 values with no null values
- The data has been recorded from the year 1980 till the year 1995 capturing the data of 15 years

Data head

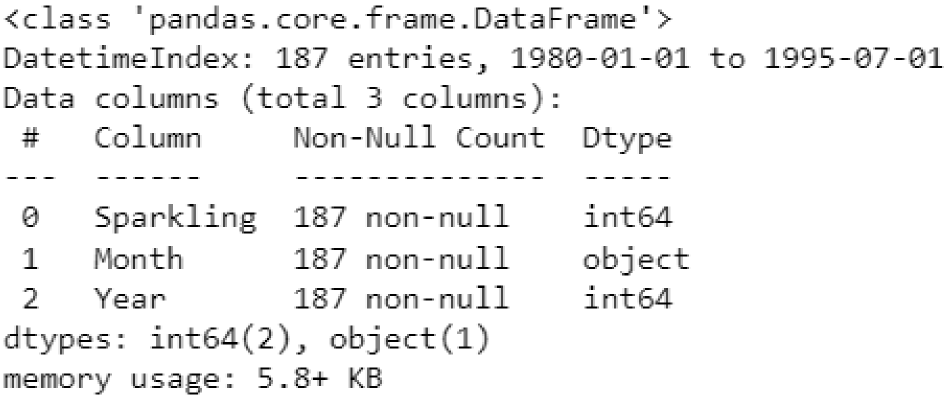


	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471
1980-06-01	1377
1980-07-01	1966
1980-08-01	2453
1980-09-01	1984
1980-10-01	2596

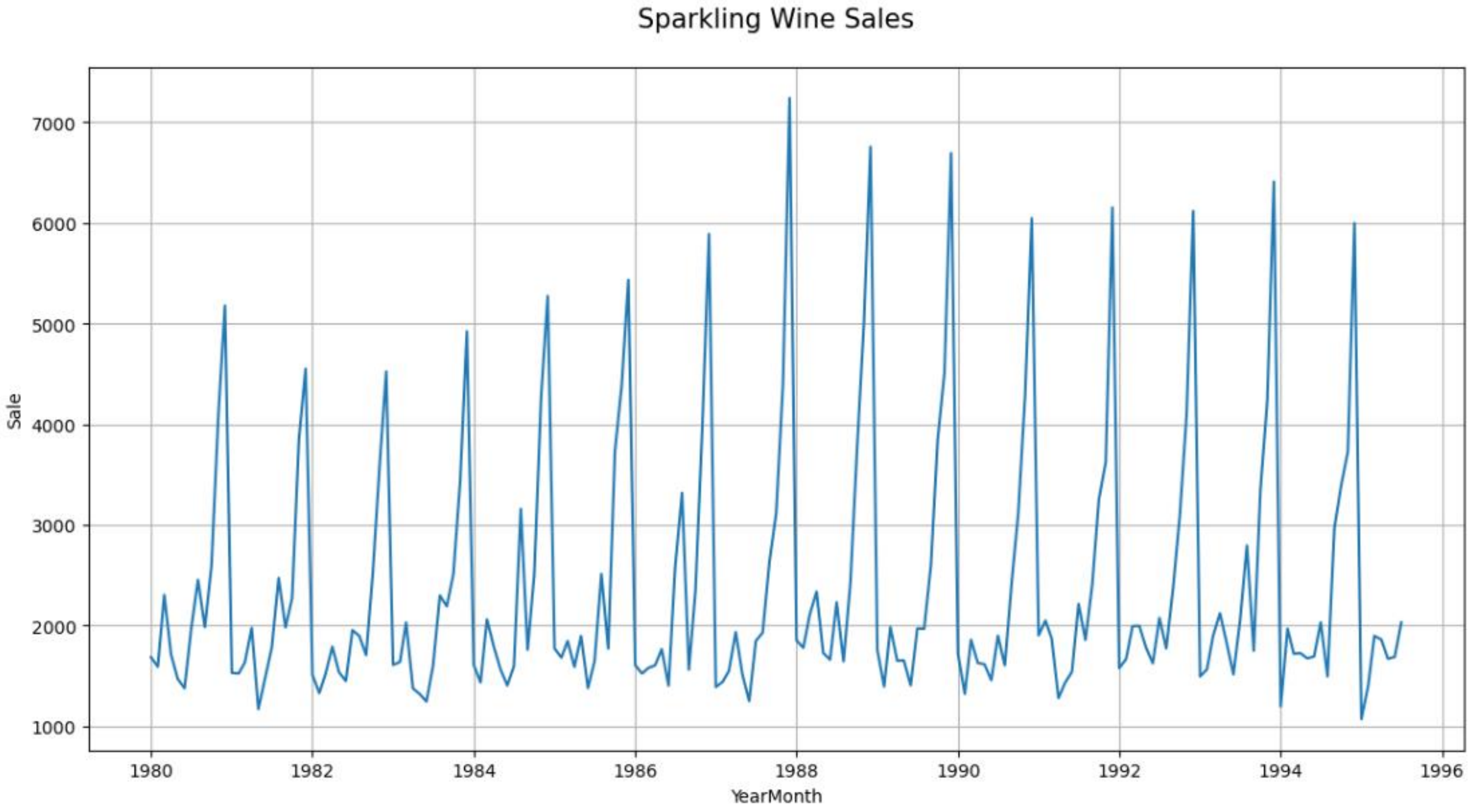
Data tail



	Sparkling
YearMonth	
1994-10-01	3385
1994-11-01	3729
1994-12-01	5999
1995-01-01	1070
1995-02-01	1402
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031



```
<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   Sparkling  187 non-null    int64 
 1   Month      187 non-null    object
 2   Year       187 non-null    int64 
dtypes: int64(2), object(1)
memory usage: 5.8+ KB
```

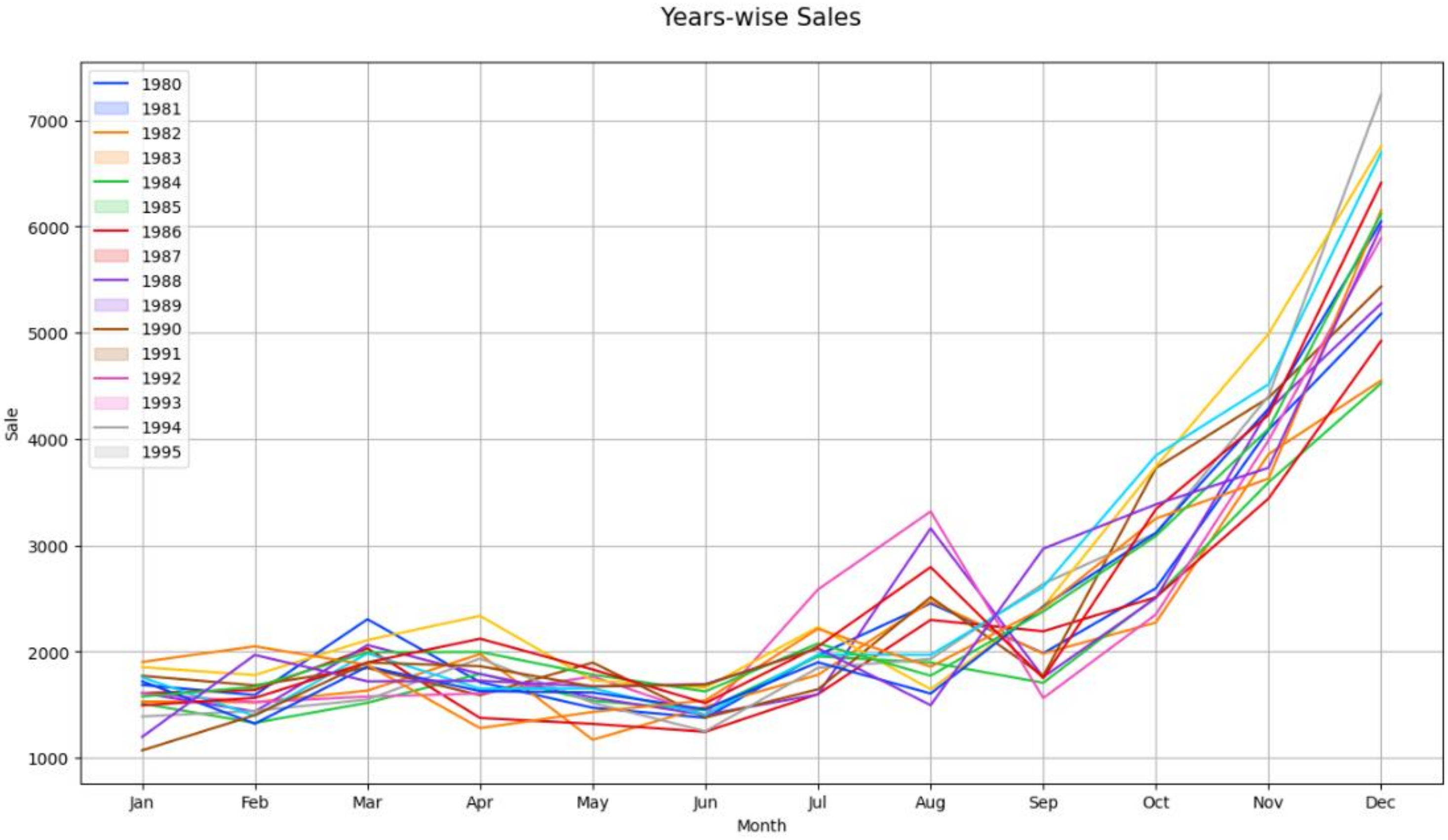


- It can be visually observed from the plot that the Sparkling Wine Sales trend steadily increases up until the year 1988, where the sale was the highest followed by a steady decline up until the year 1995.
- 12-month seasonality is evidently prominent in the dataset
- Variance seems to be non-uniform thanks to the different peaks and valleys in the sales values

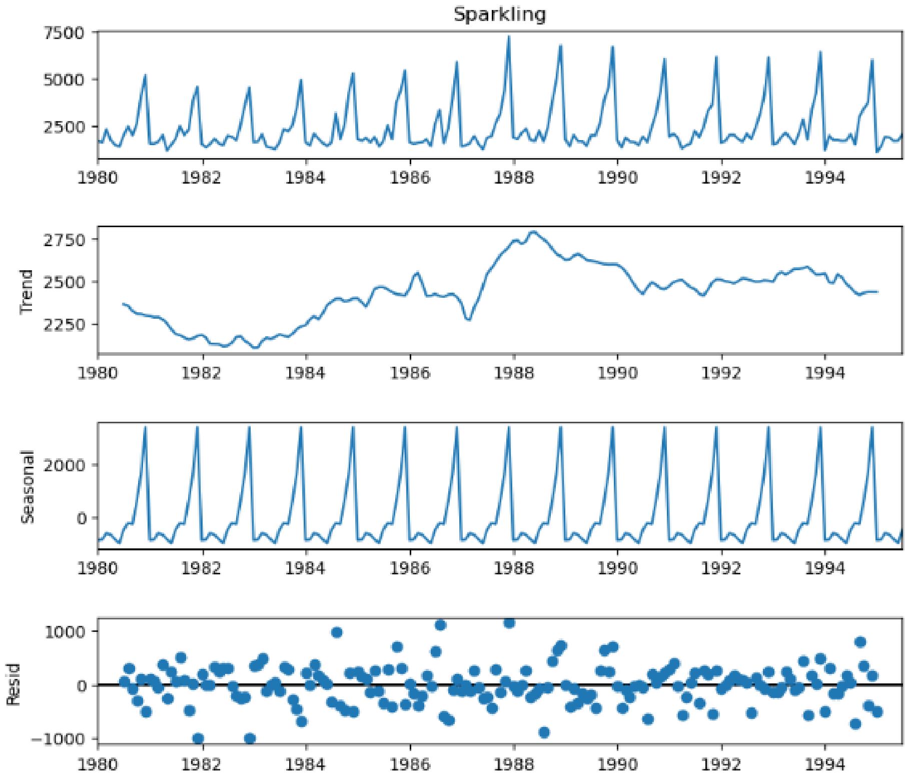
B. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Ans: Following inferences can be observed from the Exploratory Data Analysis and the time series decomposition:

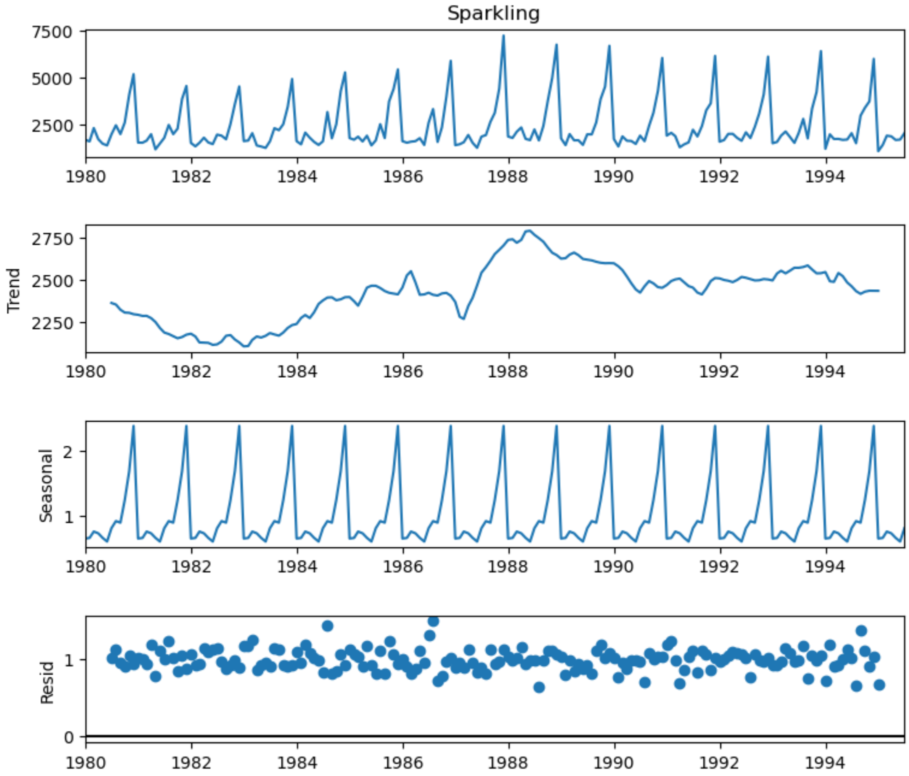
- Multiplicative decomposition seems to appropriate due to the scattered and non-uniform distribution of the residual values.
- Looking at the yearly and monthly sales plots, it can be clearly noted that certain months have much higher wine sales in comparison to the rest
- Up until the month of August, the sale is within the limits of 1000 to 3000 units
- From September onwards, the sale shoots exponentially and the month of December aces the wine sales throughout the years with an upper limit of approximately 7000 units

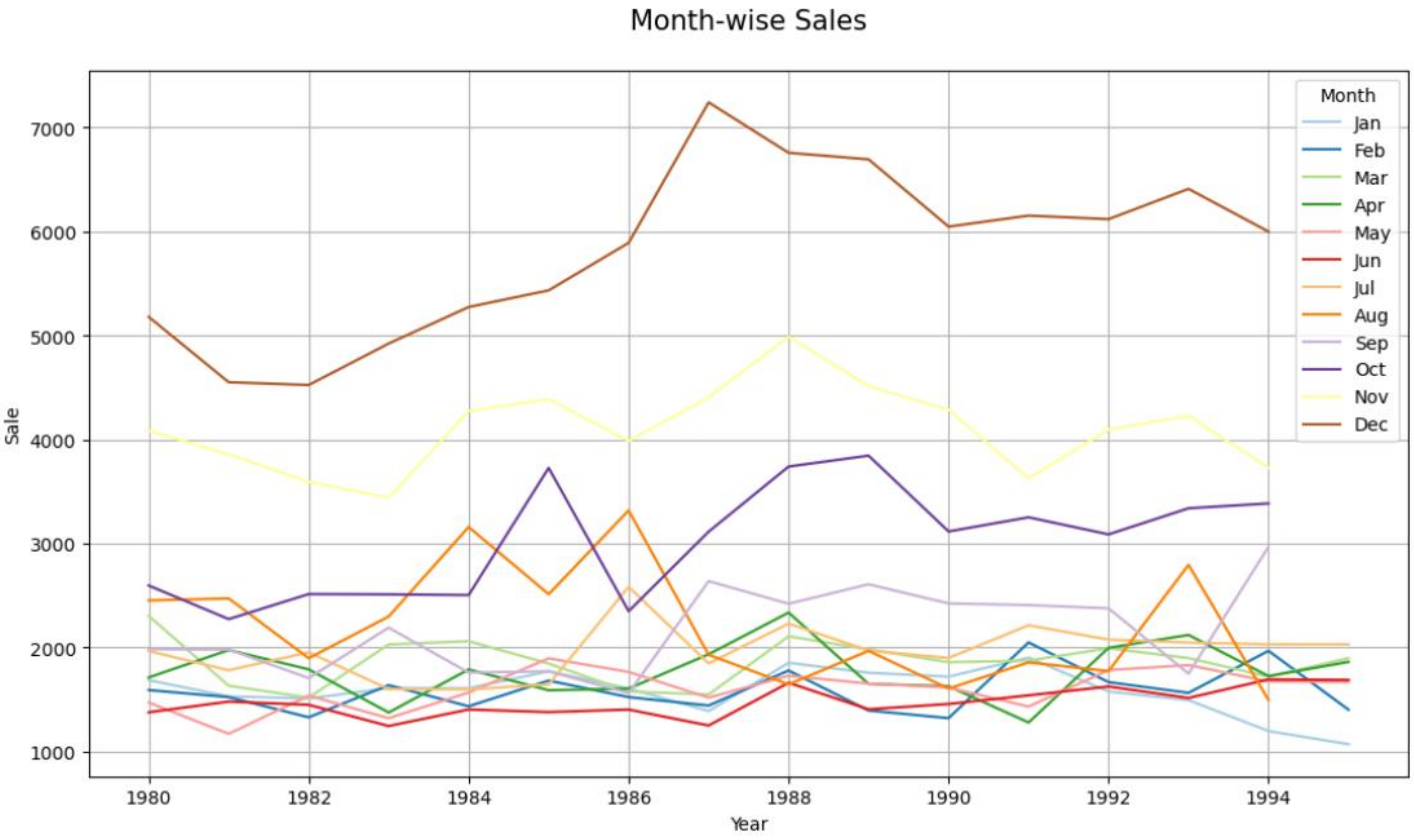


Decomposed Sparkling Wine Series (Additive)



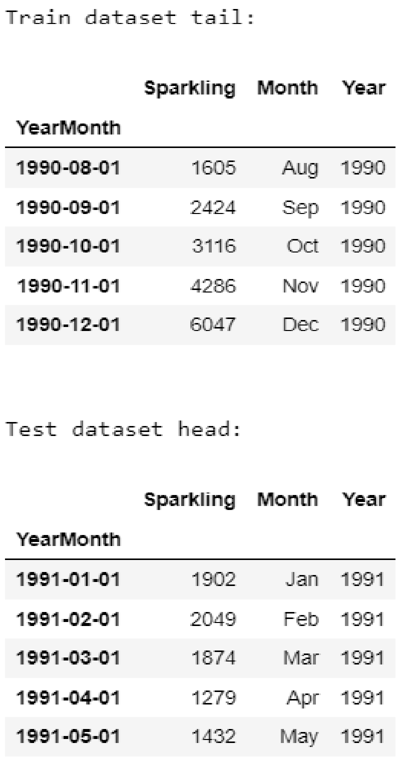
Decomposed Sparkling Wine Series (Multiplicative)




Month-wise Sales
Sale
Year
1000
2000
3000
4000
5000
6000
7000
1980
1982
1984
1986
1988
1990
1992
1994
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

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

Ans:

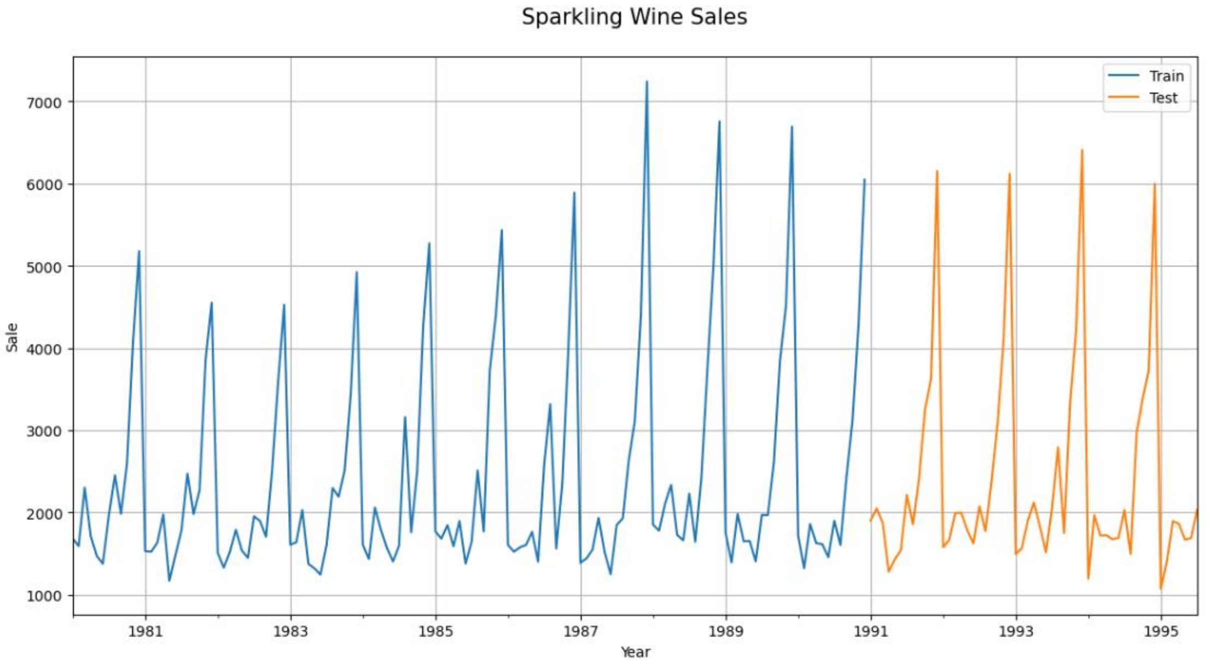


```
Train dataset tail:
```

	Sparkling	Month	Year
YearMonth			
1990-08-01	1605	Aug	1990
1990-09-01	2424	Sep	1990
1990-10-01	3116	Oct	1990
1990-11-01	4286	Nov	1990
1990-12-01	6047	Dec	1990

```
Test dataset head:
```

	Sparkling	Month	Year
YearMonth			
1991-01-01	1902	Jan	1991
1991-02-01	2049	Feb	1991
1991-03-01	1874	Mar	1991
1991-04-01	1279	Apr	1991
1991-05-01	1432	May	1991



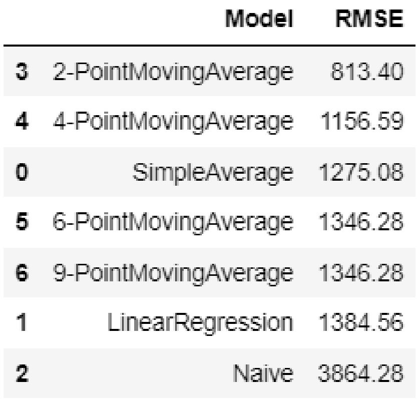
D. **Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance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

Ans: Following are the inferences drawn from the Liner Regression, Naïve, Simple and Moving averages model built on the dataset:

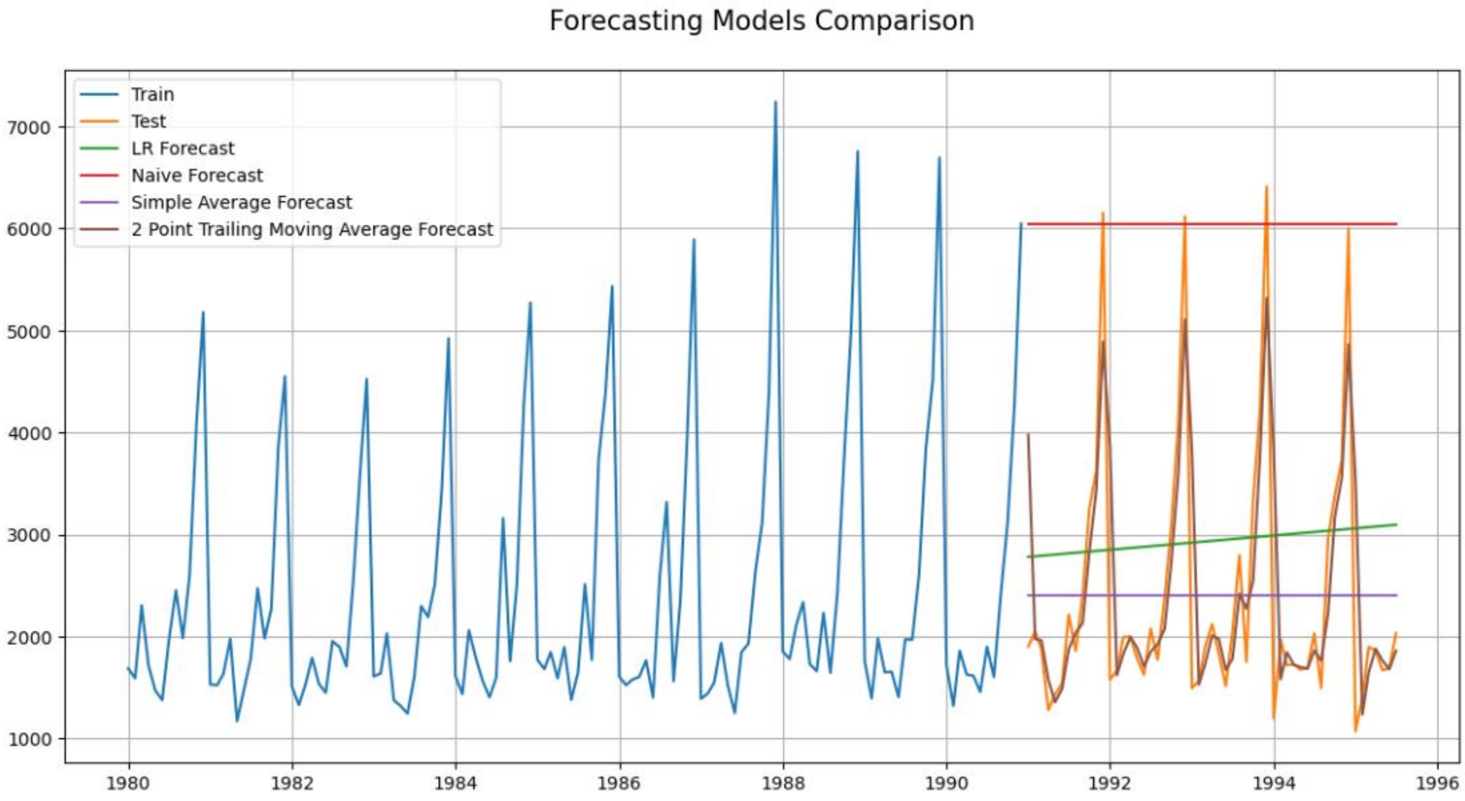
J. The 2-point moving average model has outperformed all the other models with an RMSE score of about 813.

K. It can also be visually observed in the Forecasting Models plot that the 2-point moving average model is fitting the test dataset much better in comparison to the other models.

L. The Naïve Forecast seems to be the most underperforming model which can also be visually observed in the Forecasting Models Comparison plot.

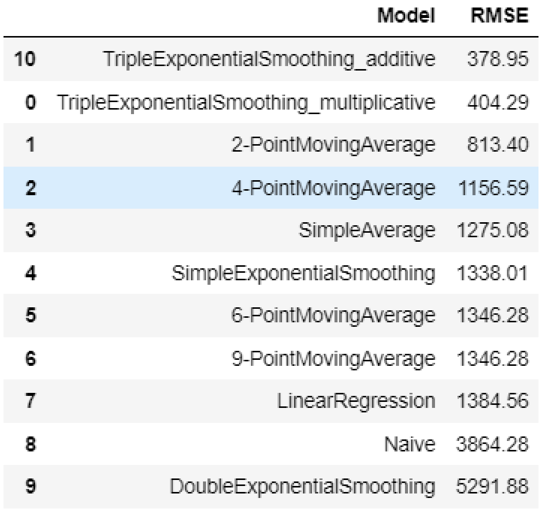


	Model	RMSE
3	2-PointMovingAverage	813.40
4	4-PointMovingAverage	1156.59
0	SimpleAverage	1275.08
5	6-PointMovingAverage	1346.28
6	9-PointMovingAverage	1346.28
1	LinearRegression	1384.56
2	Naive	3864.28



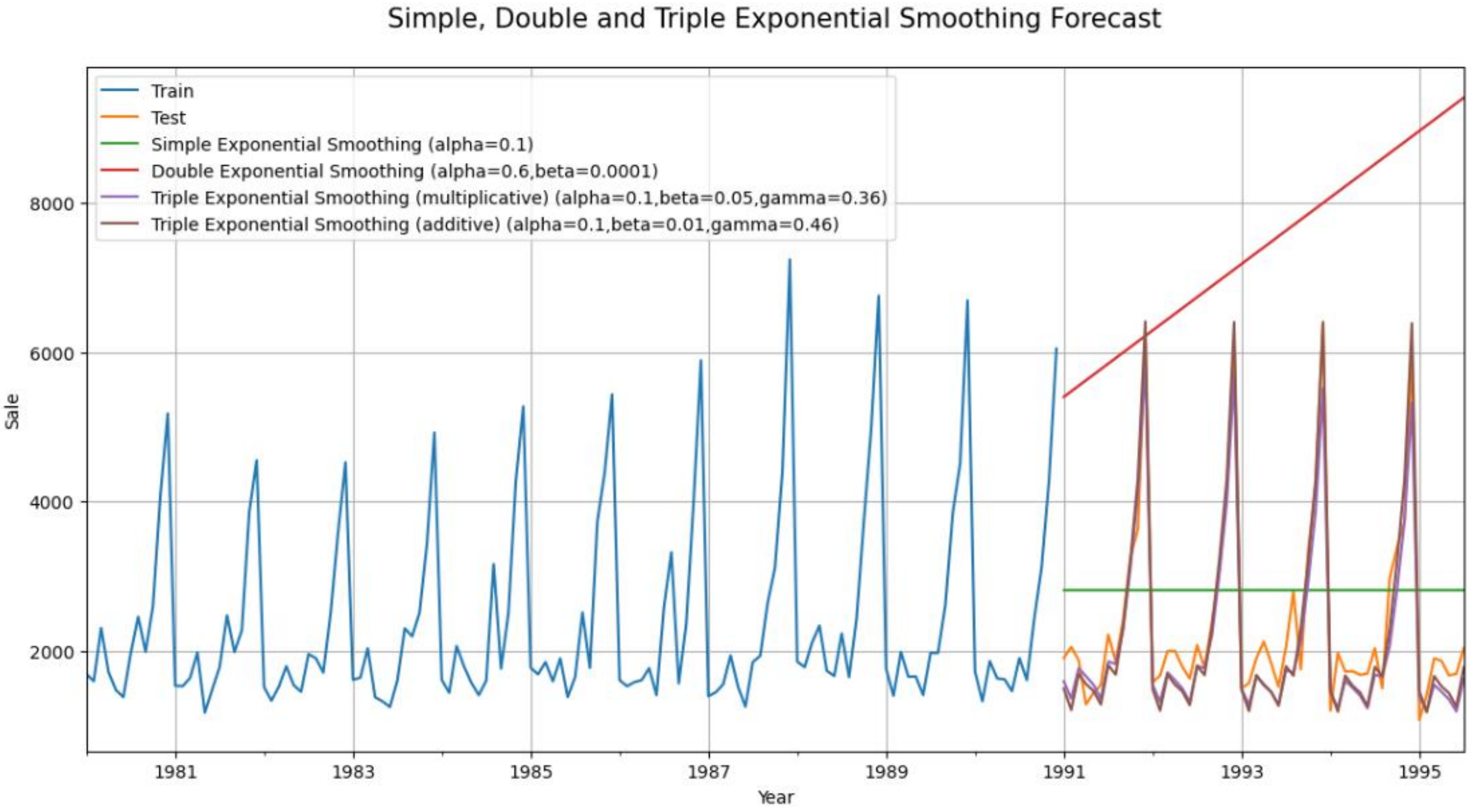
Following inferences can be drawn from the Exponential Smoothing Models:

- The Triple Exponential Smoothing model with an additive seasonality has outperformed all the other models with the lowest RMSE score of 378
- This holds true as this method of exponential smoothing effectively captures the level, trend and seasonality of the time series.



	Model	RMSE
10	TripleExponentialSmoothing_additive	378.95
0	TripleExponentialSmoothing_multiplicative	404.29
1	2-PointMovingAverage	813.40
2	4-PointMovingAverage	1156.59
3	SimpleAverage	1275.08
4	SimpleExponentialSmoothing	1338.01
5	6-PointMovingAverage	1346.28
6	9-PointMovingAverage	1346.28
7	LinearRegression	1384.56
8	Naive	3864.28
9	DoubleExponentialSmoothing	5291.88

- The Triple Exponential Smoothing model with a multiplicative seasonality follows after this model with an RMSE score of 404
- The Double Exponential Smoothing model seems to be the most underperforming model which is evident as the model only captures the trend and level of the time series and doesn't account for the seasonality present in the model.
- It can be observed visually from the Forecasting Comparison plot that the Simple and Double Exponential Smoothing Forecasting models are nowhere close to the Triple Exponential Smoothing models.



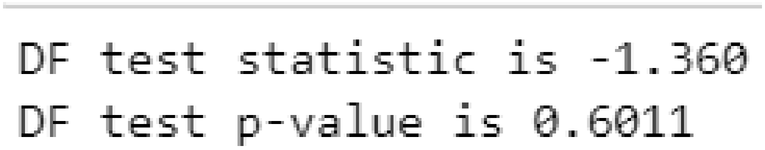
E. **Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and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

Ans: The Stationarity Test is done using the Dickey-Fuller Test which assumes the null-hypothesis (H_0)of the time series to be non-stationary and the alternate hypothesis (H_a) to be stationary. This means that if the p-value (α) < 0.05, then the null hypothesis is rejected meaning that the time series is stationary else it is accepted meaning that it is non-stationary.

It has been observed that the p-value when conducting the Dickey-Fuller test on the dataset is greater than the alpha value of 0.05 which means that we cannot reject the null hypothesis. This means that the time series is non-stationary. We can proceed by differencing the series to see if this results into making it stationary.

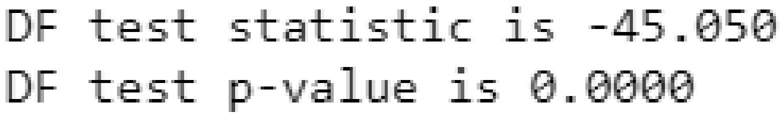
When differencing the series, the p-value becomes less than the alpha value of 0.05 which means that we can reject the null hypothesis. This means that the time series when differenced at (d=1) becomes stationary.

p-value before differencing (>0.05)



```
DF test statistic is -1.360
DF test p-value is 0.6011
```

p-value before differencing (>0.05)

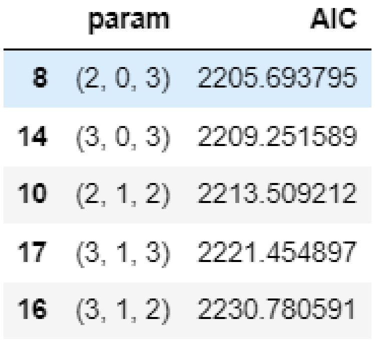


```
DF test statistic is -45.050
DF test p-value is 0.0000
```

F. **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.**

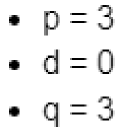
Ans: ARIMA Model (p,d,q)

The ARIMA model accounts for the Auto-Regressive component (p), Moving-Average component(q) and the Integrated component (d). The model similar to the Double Exponential Smoothing model, misses to incorporate the stationarity component which leads to the model acquiring a low RMSE score of 1001. This can also be visualised graphically when looking at the ARIMA Forecast plot wherein the model fails to superimpose the test data successfully.



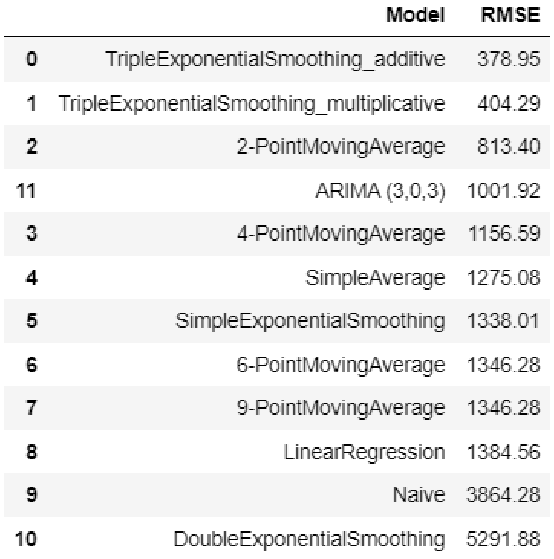
	param	AIC
8	(2, 0, 3)	2205.693795
14	(3, 0, 3)	2209.251589
10	(2, 1, 2)	2213.509212
17	(3, 1, 3)	2221.454897
16	(3, 1, 2)	2230.780591

Building ARIMA model with best parameters p,d,q for the following values:

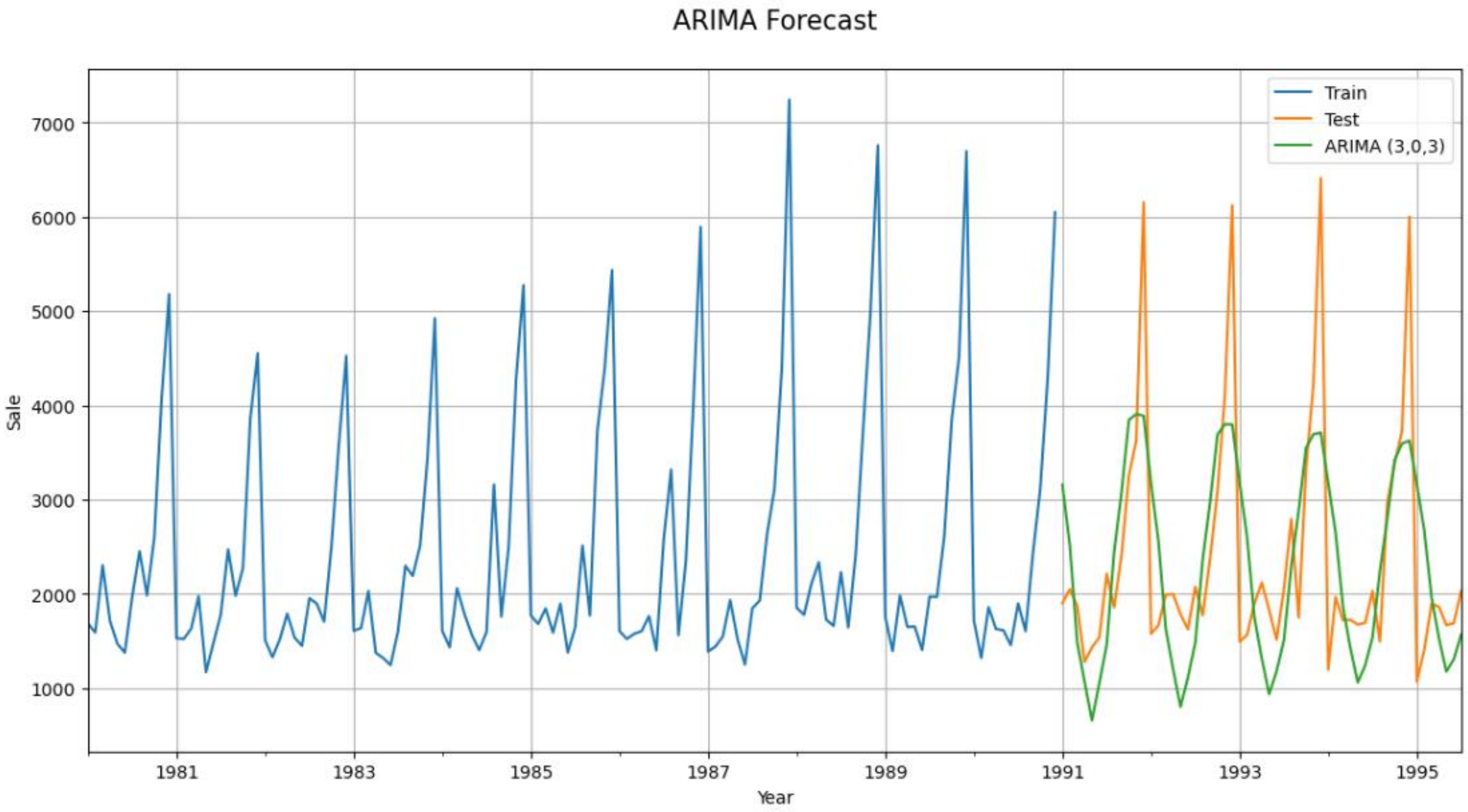


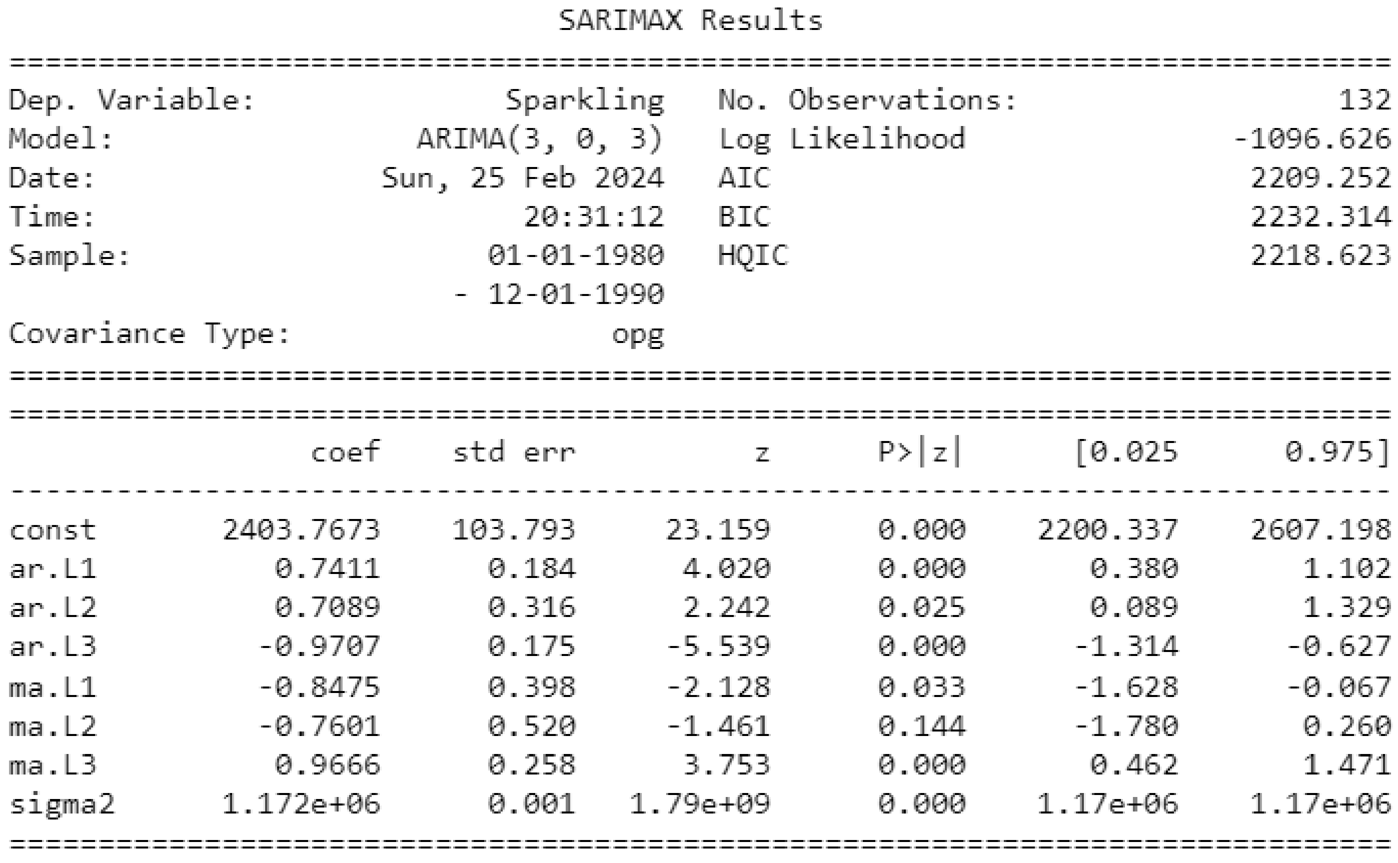
- p = 3
- d = 0
- q = 3

The model parameters (p,d,q) are chosen by determining the lowest Akaike Information Criterion (AIC) values. The model however does well in comparison to the Simple and Double Exponential Smoothing models however is still underperforming when compared to the 2-point average model.



	Model	RMSE
0	TripleExponentialSmoothing_additive	378.95
1	TripleExponentialSmoothing_multiplicative	404.29
2	2-PointMovingAverage	813.40
11	ARIMA (3,0,3)	1001.92
3	4-PointMovingAverage	1156.59
4	SimpleAverage	1275.08
5	SimpleExponentialSmoothing	1338.01
6	6-PointMovingAverage	1346.28
7	9-PointMovingAverage	1346.28
8	LinearRegression	1384.56
9	Naive	3864.28
10	DoubleExponentialSmoothing	5291.88





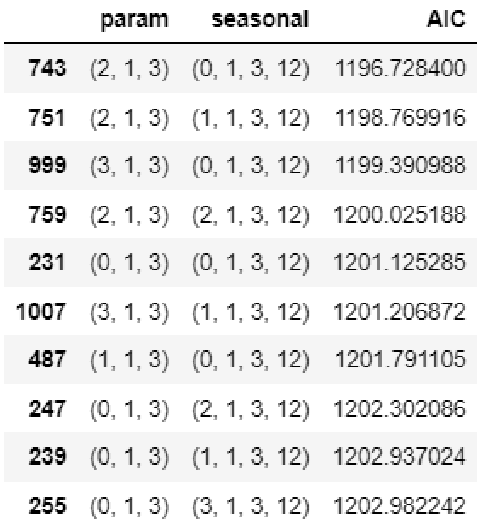
```
                               SARIMAX Results
==============================================================================
Dep. Variable:               Sparkling   No. Observations:                  132
Model:                  ARIMA(3, 0, 3)   Log Likelihood               -1096.626
Date:                 Sun, 25 Feb 2024   AIC                           2209.252
Time:                         20:31:12   BIC                           2232.314
Sample:                     01-01-1980   HQIC                          2218.623
                          - 12-01-1990
Covariance Type:                   opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2403.7673    103.793     23.159      0.000    2200.337    2607.198
ar.L1          0.7411      0.184      4.020      0.000       0.380       1.102
ar.L2          0.7089      0.316      2.242      0.025       0.089       1.329
ar.L3         -0.9707      0.175     -5.539      0.000      -1.314      -0.627
ma.L1         -0.8475      0.398     -2.128      0.033      -1.628      -0.067
ma.L2         -0.7601      0.520     -1.461      0.144      -1.780       0.260
ma.L3          0.9666      0.258      3.753      0.000       0.462       1.471
sigma2      1.172e+06      0.001   1.79e+09      0.000    1.17e+06    1.17e+06
==============================================================================
```

SARIMA Model (p,d,q,P,D,Q,F)

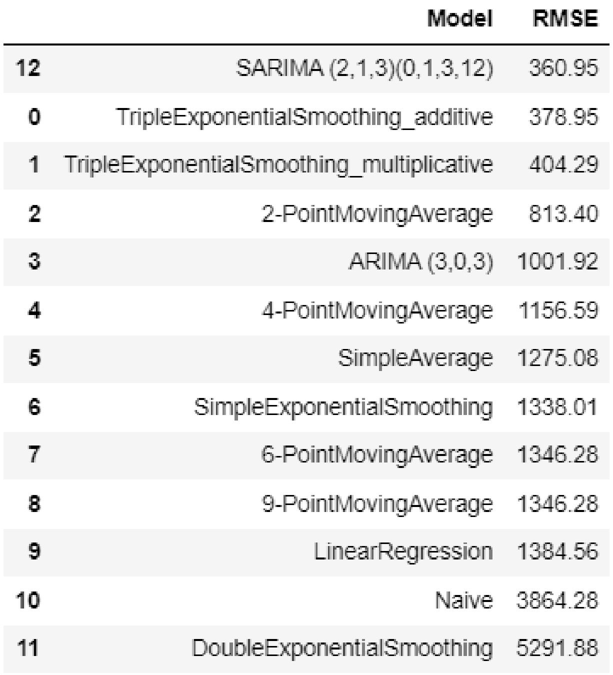
The SARIMA model accounts for the Auto-Regressive component (p), Moving-Average component(q), the Integrated component (d) along with the Seasonality components (P,D,Q,F). The model hence is able to incorporate seasonality parameter which results in better forecasting.

The model naturally performs well with the lowest RMSE score of 360. The model's effective forecasting can also be visualised graphically when looking at the SARIMA Forecast plot wherein the model is effectively able to superimpose the test data.

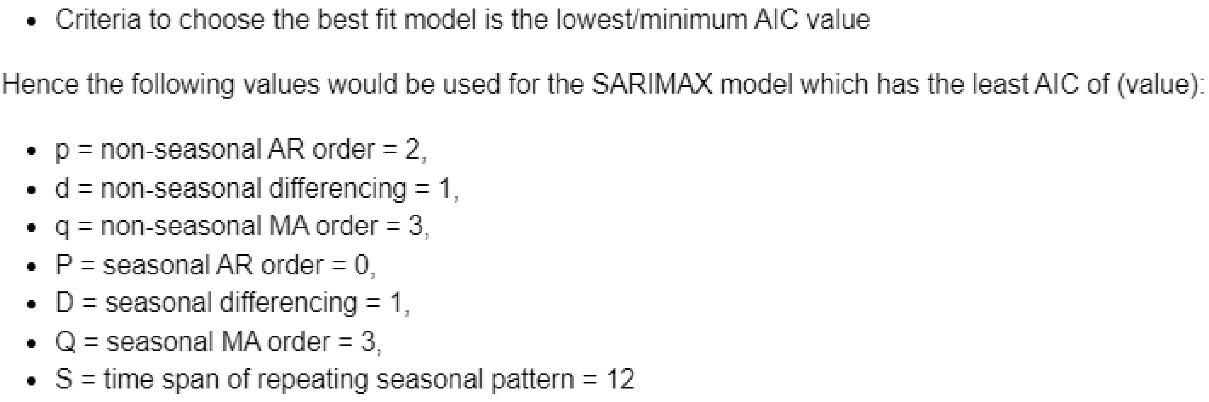
Similar to the ARIMA model, the model parameters are determined by calculating the lowest AIC values.



	param	seasonal	AIC
743	(2, 1, 3)	(0, 1, 3, 12)	1196.728400
751	(2, 1, 3)	(1, 1, 3, 12)	1198.769916
999	(3, 1, 3)	(0, 1, 3, 12)	1199.390988
759	(2, 1, 3)	(2, 1, 3, 12)	1200.025188
231	(0, 1, 3)	(0, 1, 3, 12)	1201.125285
1007	(3, 1, 3)	(1, 1, 3, 12)	1201.206872
487	(1, 1, 3)	(0, 1, 3, 12)	1201.791105
247	(0, 1, 3)	(2, 1, 3, 12)	1202.302086
239	(0, 1, 3)	(1, 1, 3, 12)	1202.937024
255	(0, 1, 3)	(3, 1, 3, 12)	1202.982242



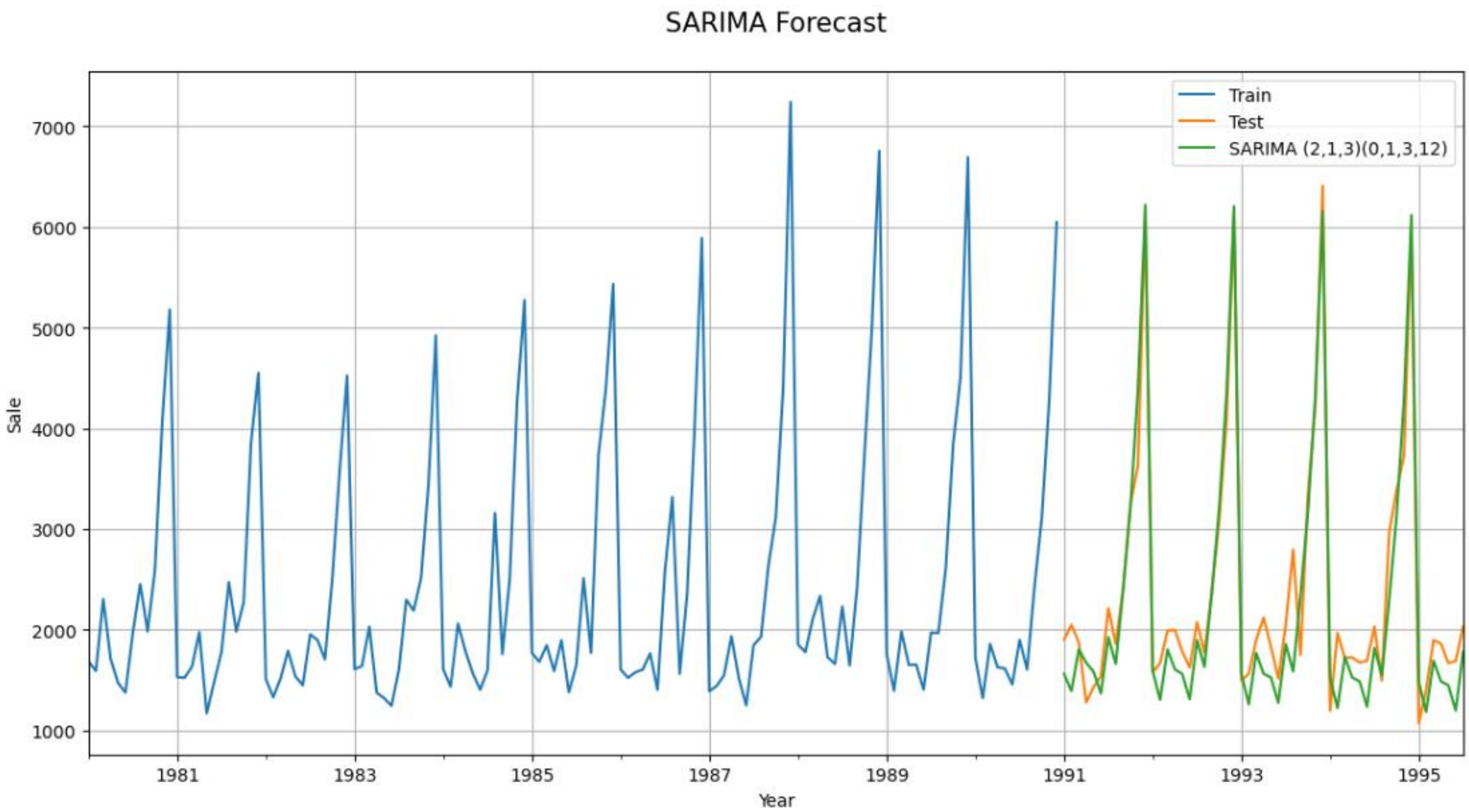
	Model	RMSE
12	SARIMA (2,1,3)(0,1,3,12)	360.95
0	TripleExponentialSmoothing_additive	378.95
1	TripleExponentialSmoothing_multiplicative	404.29
2	2-PointMovingAverage	813.40
3	ARIMA (3,0,3)	1001.92
4	4-PointMovingAverage	1156.59
5	SimpleAverage	1275.08
6	SimpleExponentialSmoothing	1338.01
7	6-PointMovingAverage	1346.28
8	9-PointMovingAverage	1346.28
9	LinearRegression	1384.56
10	Naive	3864.28
11	DoubleExponentialSmoothing	5291.88

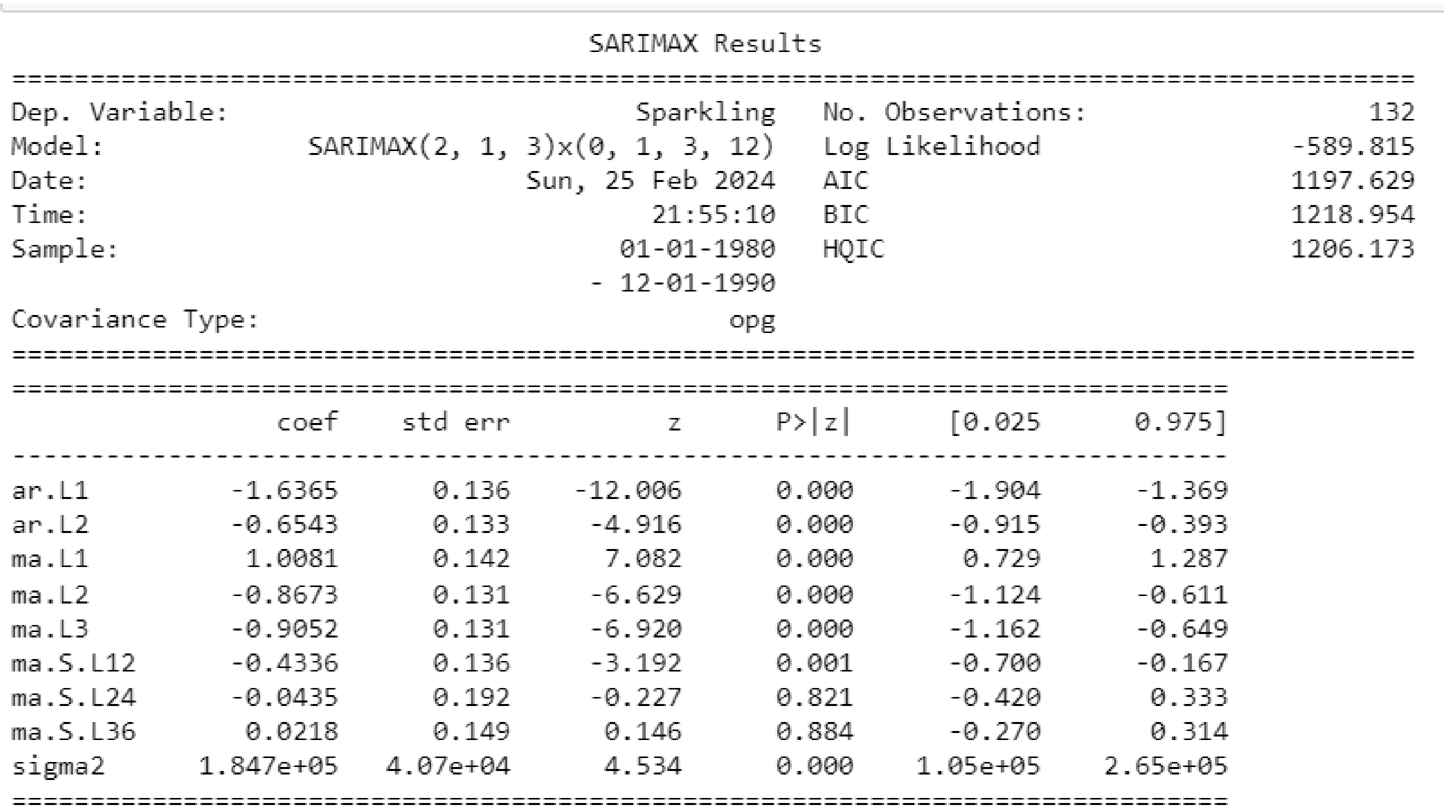


- Criteria to choose the best fit model is the lowest/minimum AIC value

Hence the following values would be used for the SARIMAX model which has the least AIC of (value):

- p = non-seasonal AR order = 2,
- d = non-seasonal differencing = 1,
- q = non-seasonal MA order = 3,
- P = seasonal AR order = 0,
- D = seasonal differencing = 1,
- Q = seasonal MA order = 3,
- S = time span of repeating seasonal pattern = 12



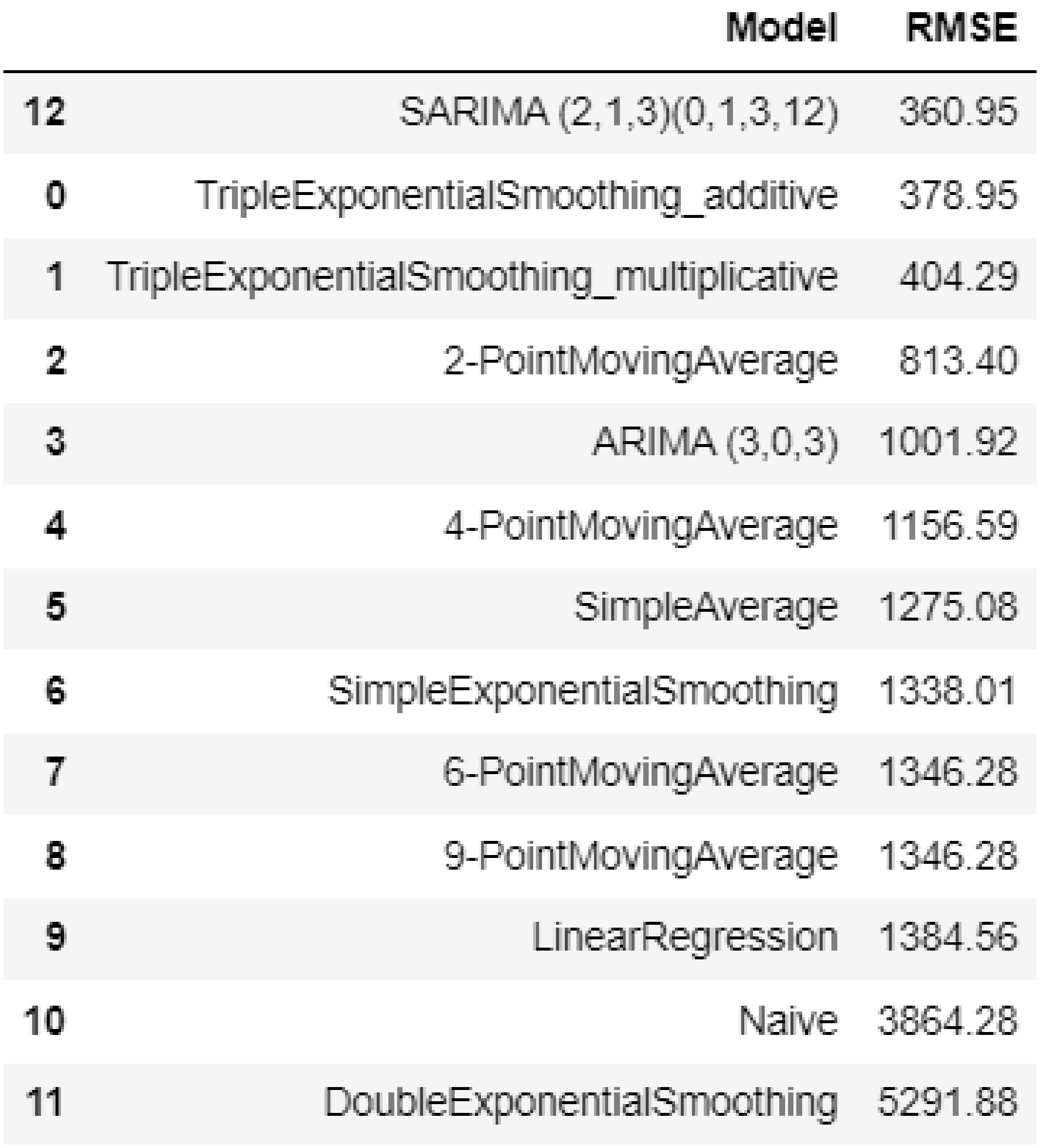


```
                                    SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                132
Model:             SARIMAX(2, 1, 3)x(0, 1, 3, 12)   Log Likelihood                -589.815
Date:                              Sun, 25 Feb 2024   AIC                           1197.629
Time:                                      21:55:10   BIC                           1218.954
Sample:                                  01-01-1980   HQIC                          1206.173
                                       - 12-01-1990
Covariance Type:                                opg
==========================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6365      0.136    -12.006      0.000      -1.904      -1.369
ar.L2         -0.6543      0.133     -4.916      0.000      -0.915      -0.393
ma.L1          1.0081      0.142      7.082      0.000       0.729       1.287
ma.L2         -0.8673      0.131     -6.629      0.000      -1.124      -0.611
ma.L3         -0.9052      0.131     -6.920      0.000      -1.162      -0.649
ma.S.L12      -0.4336      0.136     -3.192      0.001      -0.700      -0.167
ma.S.L24      -0.0435      0.192     -0.227      0.821      -0.420       0.333
ma.S.L36       0.0218      0.149      0.146      0.884      -0.270       0.314
sigma2      1.847e+05   4.07e+04      4.534      0.000    1.05e+05    2.65e+05
==============================================================================
```

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

Ans: Following are the top-performing models along with their respective RMSE scores:

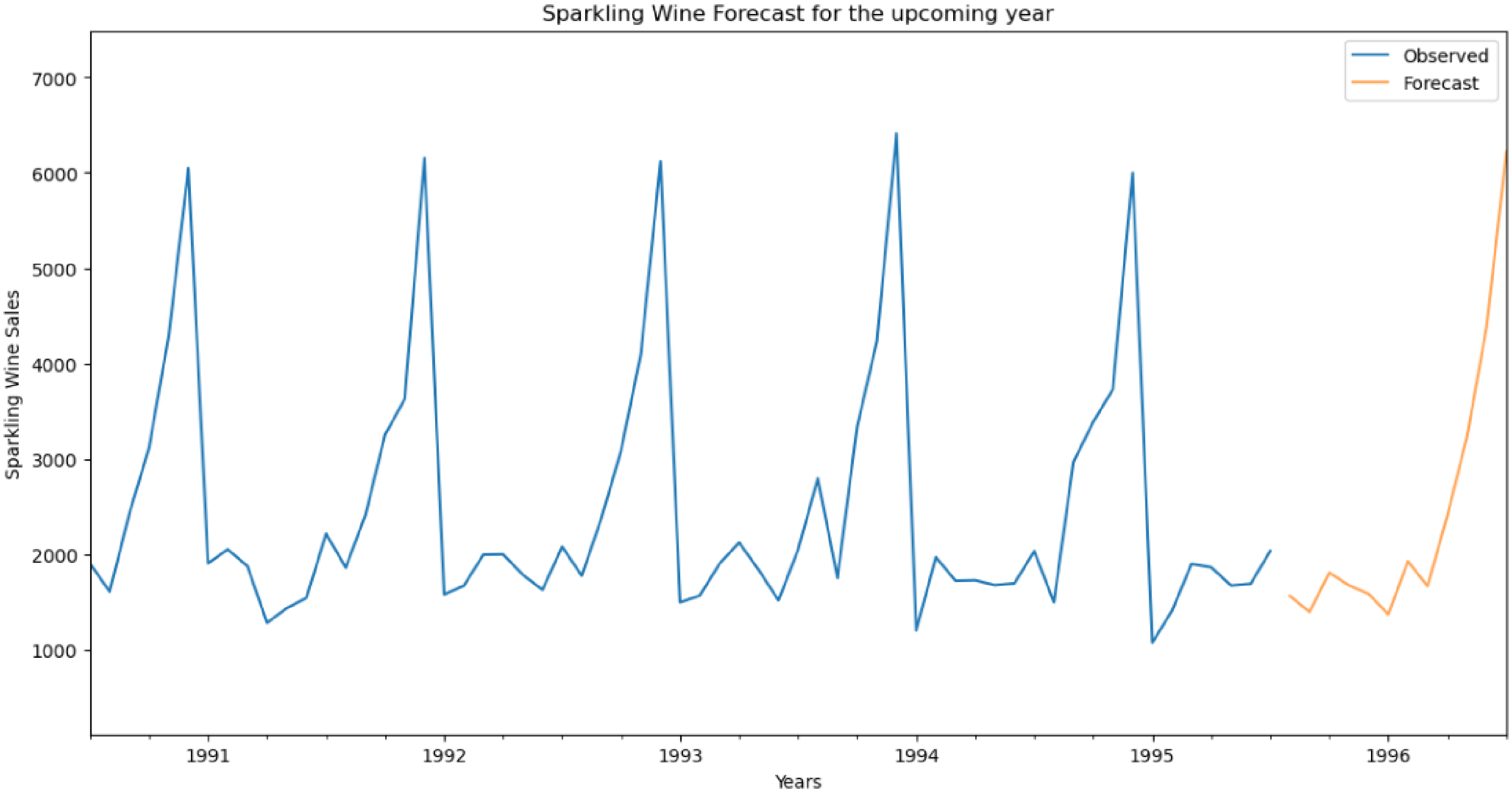
M. SARIMA (RMSE = 360)
N. Triple Exponential Smoothing (Additive)(RMSE = 378)
O. Triple Exponential Smoothing (Multiplicative) (RMSE = 404)



	Model	RMSE
12	SARIMA (2,1,3)(0,1,3,12)	360.95
0	TripleExponentialSmoothing_additive	378.95
1	TripleExponentialSmoothing_multiplicative	404.29
2	2-PointMovingAverage	813.40
3	ARIMA (3,0,3)	1001.92
4	4-PointMovingAverage	1156.59
5	SimpleAverage	1275.08
6	SimpleExponentialSmoothing	1338.01
7	6-PointMovingAverage	1346.28
8	9-PointMovingAverage	1346.28
9	LinearRegression	1384.56
10	Naive	3864.28
11	DoubleExponentialSmoothing	5291.88

H. 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.

Ans: The SARIMA model was chosen as the most optimum model for forecasting. Following is the 12-month forecast represented graphically along with the appropriate confidence intervals/bands:



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

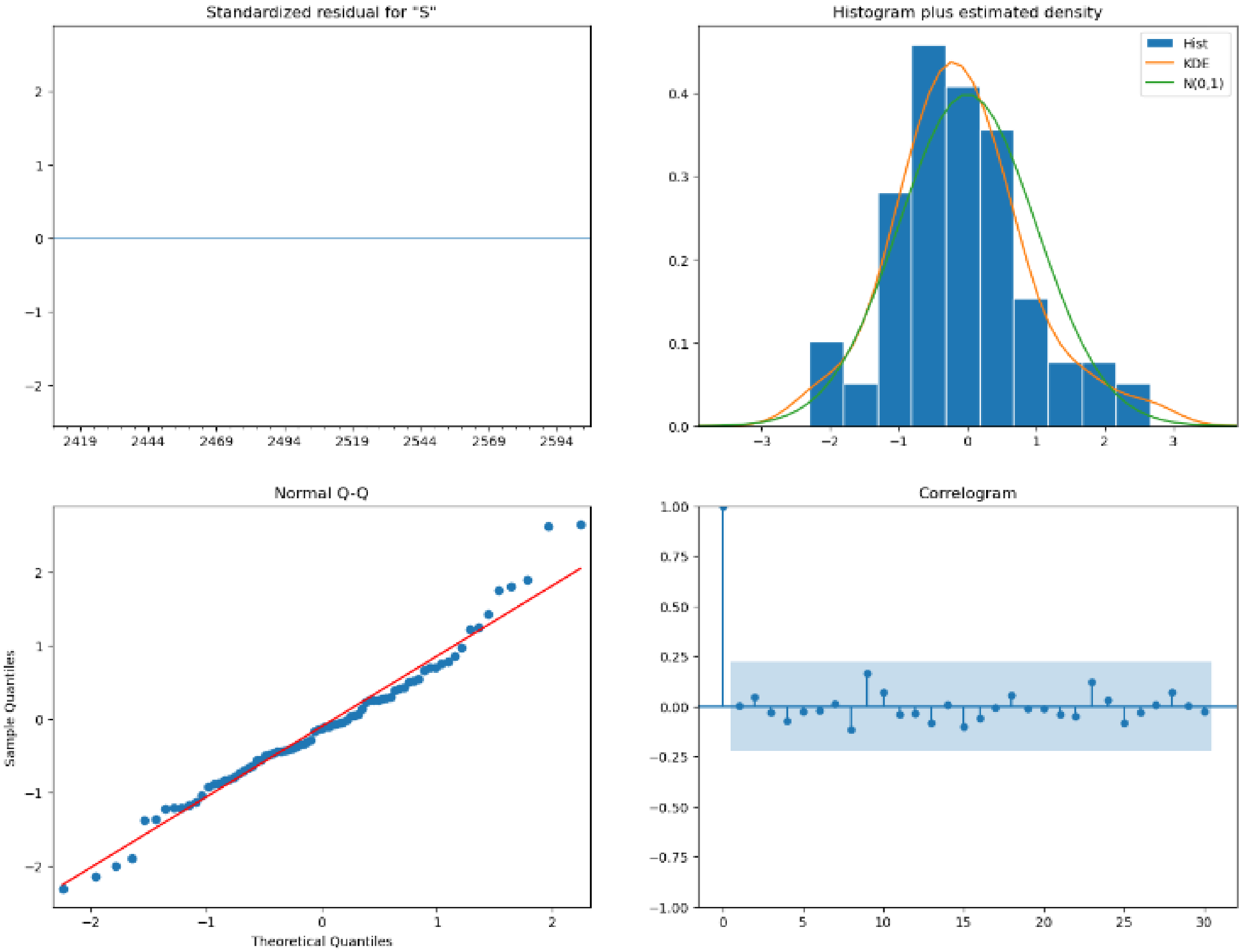
Ans: Model Diagnostics/Analysis

Following inferences can be drawn from the SARIMA Forecasting model:

P. <u>Standardized residuals plot</u> : No pattern is visible in the residuals in the Standardized Residual Plot indicating that the model is working efficiently.

Q. <u>Histogram plus estimated density plot</u> : The residual distribution is shows slight variation between the normal distribution which indicates that are model's performance is good enough. Vast deviation from this might indicate a poor performing model.

R. <u>Normal Q-Q plot</u> : The Q-Q plot also shows a normal residual distribution as they are distributed along the line


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