# TIME SERIES FORECASTING PROJECT

# Table Of Content

# List Of Figures

# List of Tables

# 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.

Data set for the Problem is as below:

### Sparkling Dataset

| YearMonth | Sparkling |
|---|---|
| 1980-01 | 1686 |
| 1980-02 | 1591 |
| 1980-03 | 2304 |
| 1980-04 | 1712 |
| 1980-05 | 1471 |
| ... | ... |
| 1995-03 | 1897 |
| 1995-04 | 1862 |
| 1995-05 | 1670 |
| 1995-06 | 1688 |
| 1995-07 | 2031 |

### Rose Dataset

| YearMonth | Rose |
|---|---|
| 1980-01 | 112.0 |
| 1980-02 | 118.0 |
| 1980-03 | 129.0 |
| 1980-04 | 99.0 |
| 1980-05 | 116.0 |
| ... | ... |
| 1995-03 | 45.0 |
| 1995-04 | 52.0 |
| 1995-05 | 28.0 |
| 1995-06 | 40.0 |
| 1995-07 | 62.0 |

# Time series Forecasting on Sparkling Wine Data

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

### Plot of Sparkling Wine sales

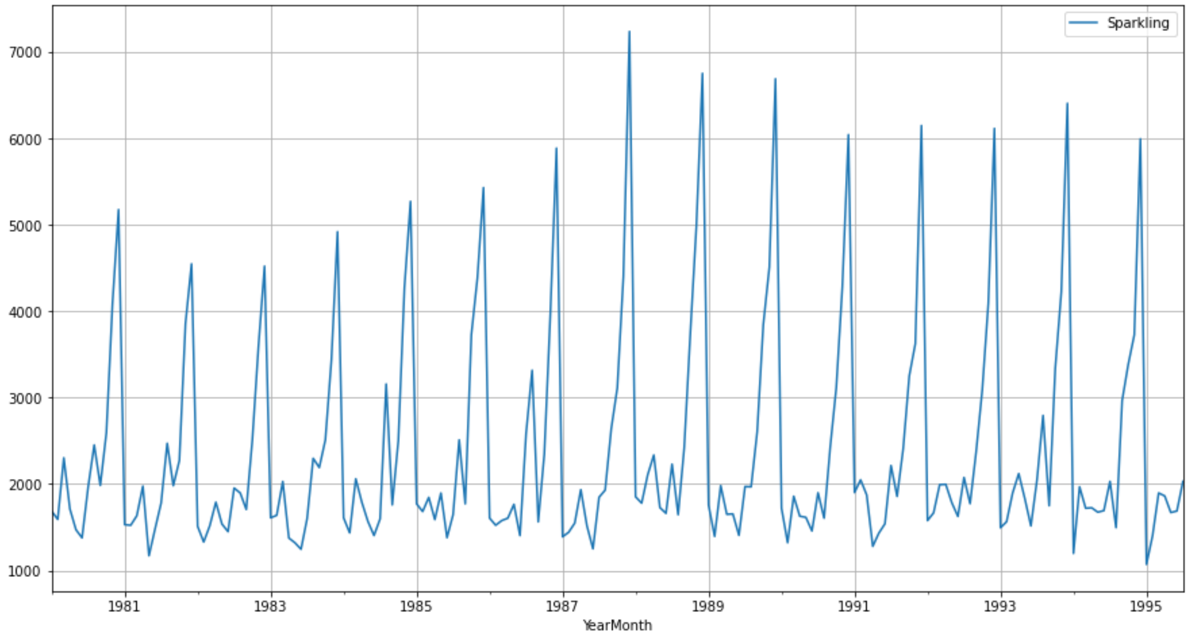



**Fig. 1: Sparkling Wine Sales Data**

- Data values are stored in correct time order and no data is missing.
- Do not observe steady rise in sales every year
- Intra-year stable fluctuations are indicative of seasonal component.

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

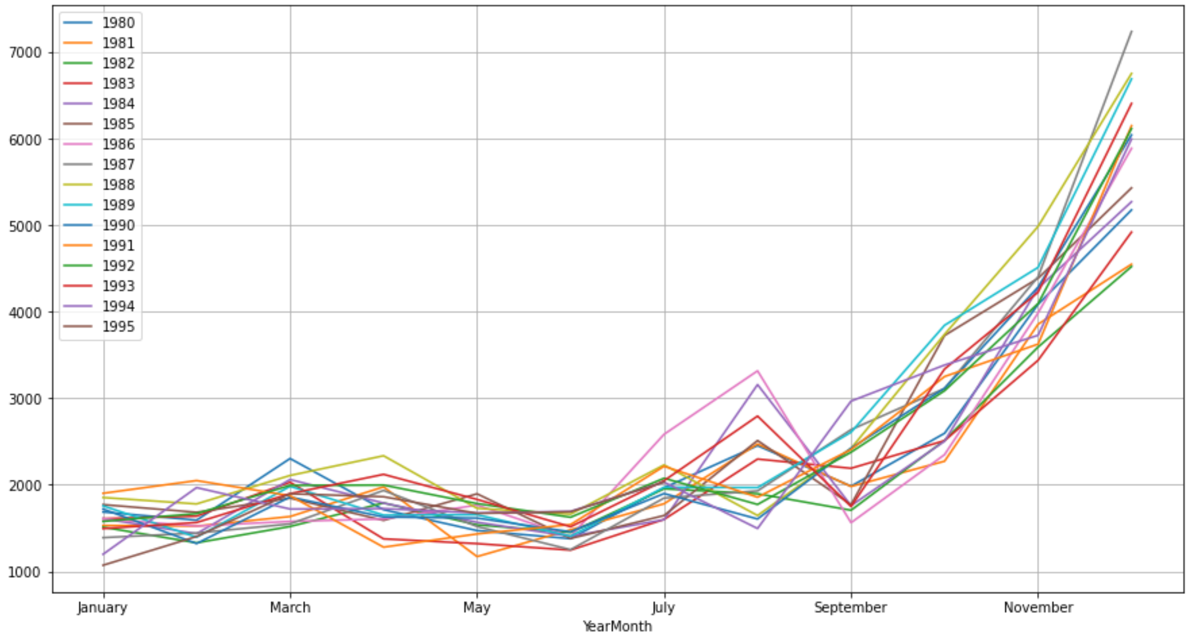



**Fig. 2: Seasonal plot Year Month wise**

- Figure 2 shows that the sales of Sparkling Wine which identifies seasonal fluctuations
- We can observe that the average sales is higher in the month of December as compared to other months

### Decomposition of Sparkling Wine Dataset

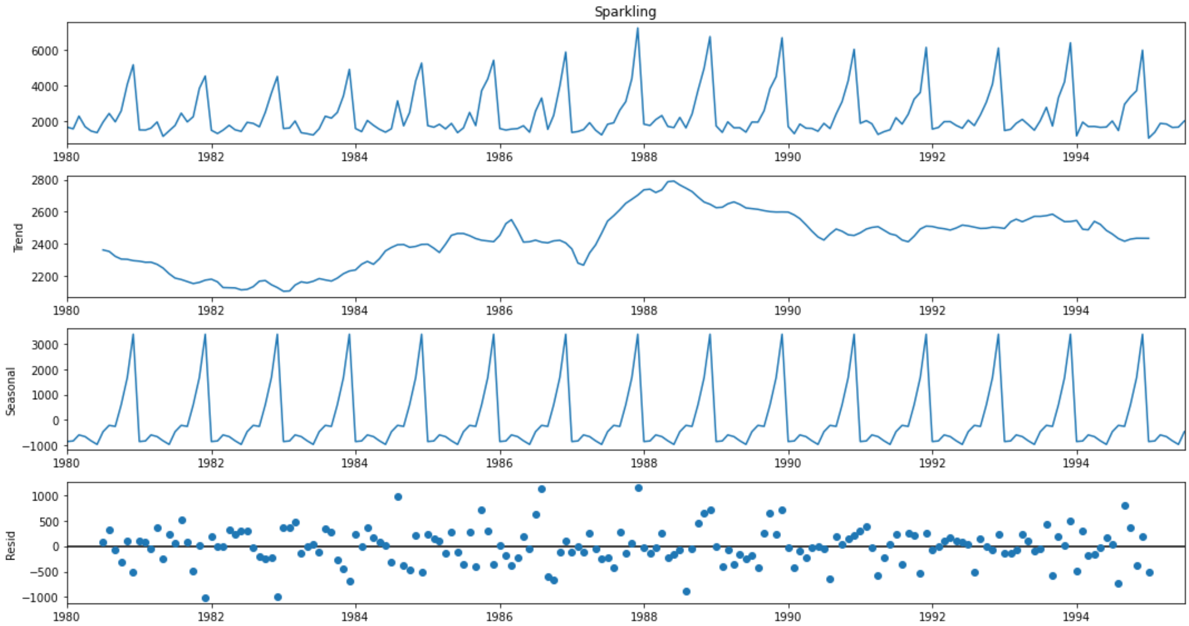



**Fig. 3: Decomposition of Sparkling Wine Dataset**

- From the above Figure we can observe that there is a trend present but not increasing steadily.
- Also we can observe Seasonality in the Data and residuals are random

## 3. Splitting the data into training and test. The test data will start from 1991

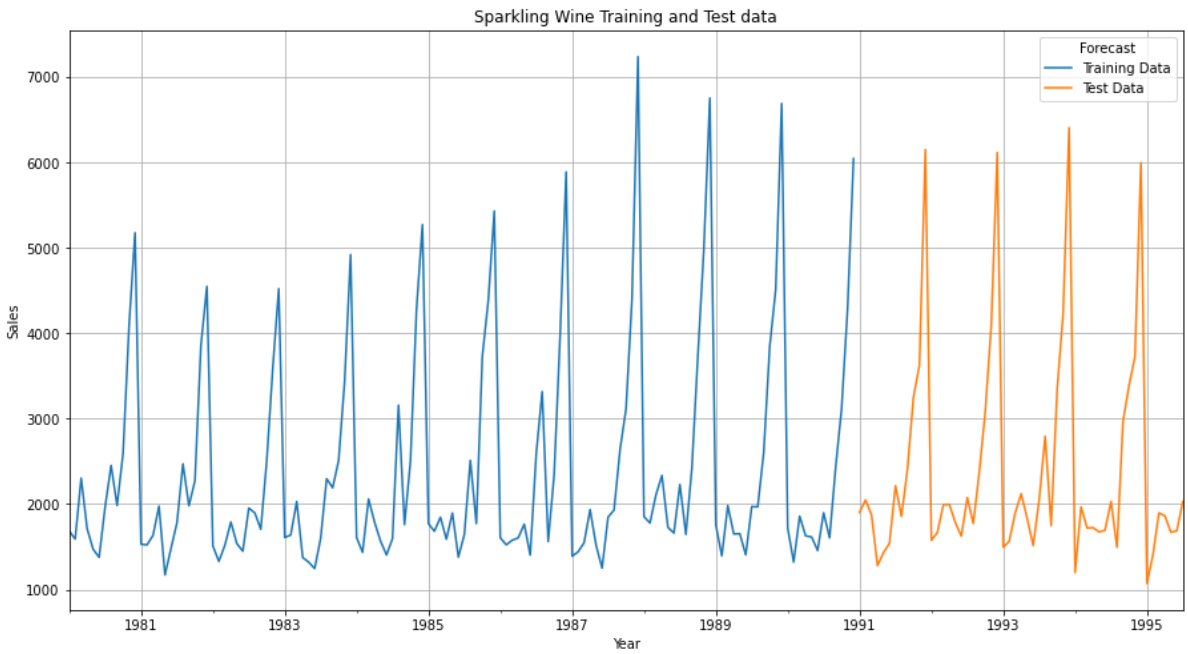



**Fig. 4: Plot of Training and test dataset for Sparkling Wine Sales**

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

## Exponential Smoothing Method for Sparkling Wine Sales Dataset

## SES - ETS(A, N, N) - Simple Exponential Smoothing with additive errors

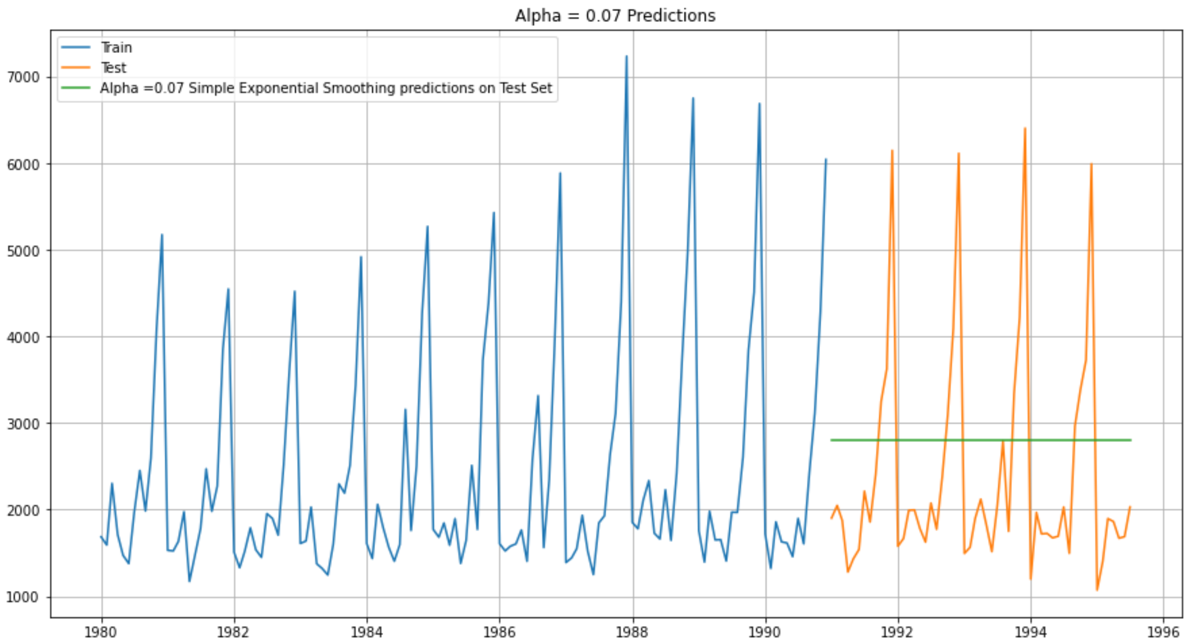



**Fig. 5: Alpha =0.07, Simple Exponential Smoothing predictions on Test Set for Sparkling Wine Sales Dataset**

## Holt - ETS(A, A, N) - Holt's linear method with additive errors

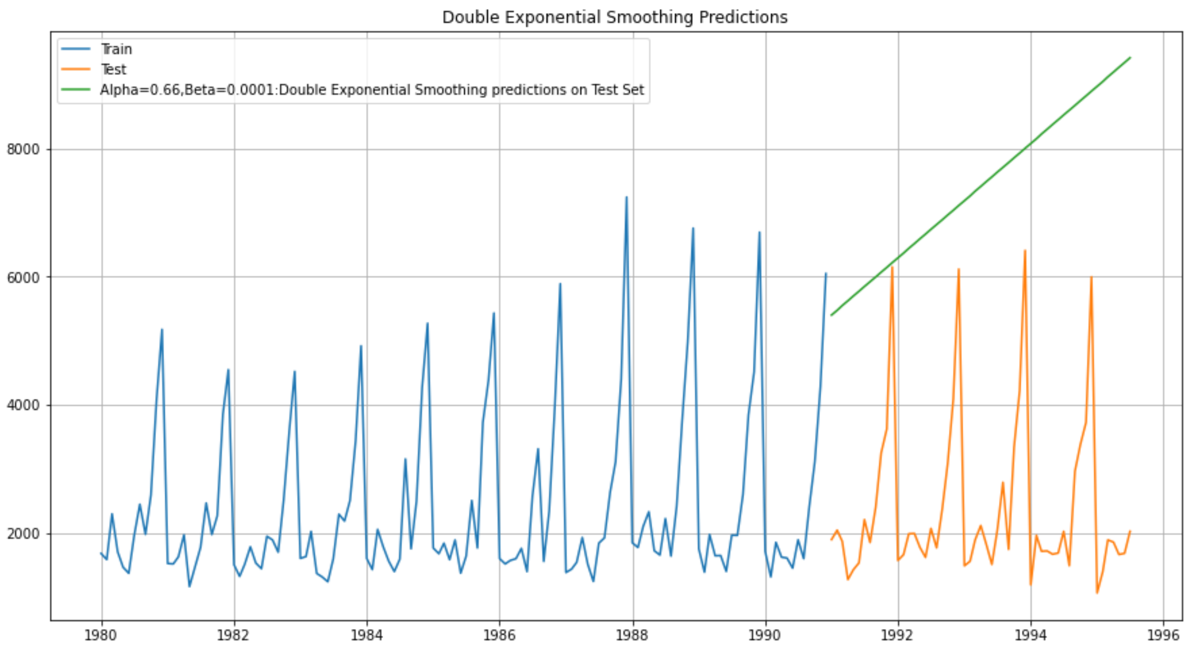



**Fig. 6: Alpha=0.66,Beta=0.0001:Double Exponential Smoothing predictions on Test Set for Sparkling Wine Sales Dataset**

## Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

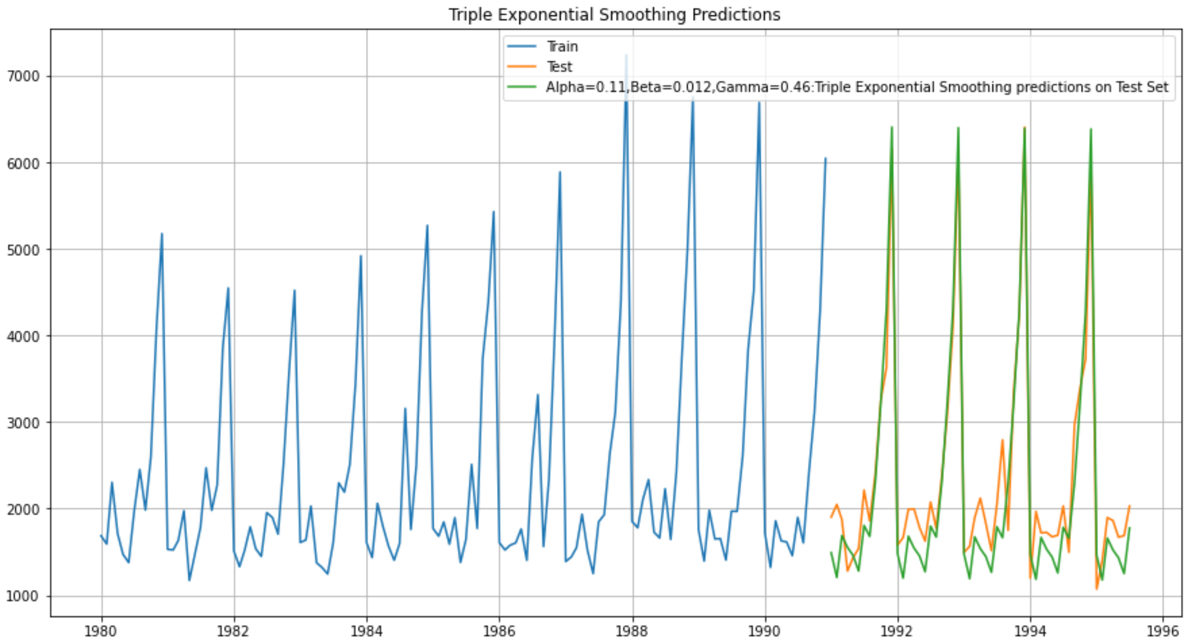



**Fig. 7: Alpha=0.11,Beta=0.012,Gamma=0.46:Triple Exponential Smoothing predictions on Test Set for Sparkling Wine Sales Dataset**

## Holt-Winters - ETS(A, A, M) - Holt Winter's linear method

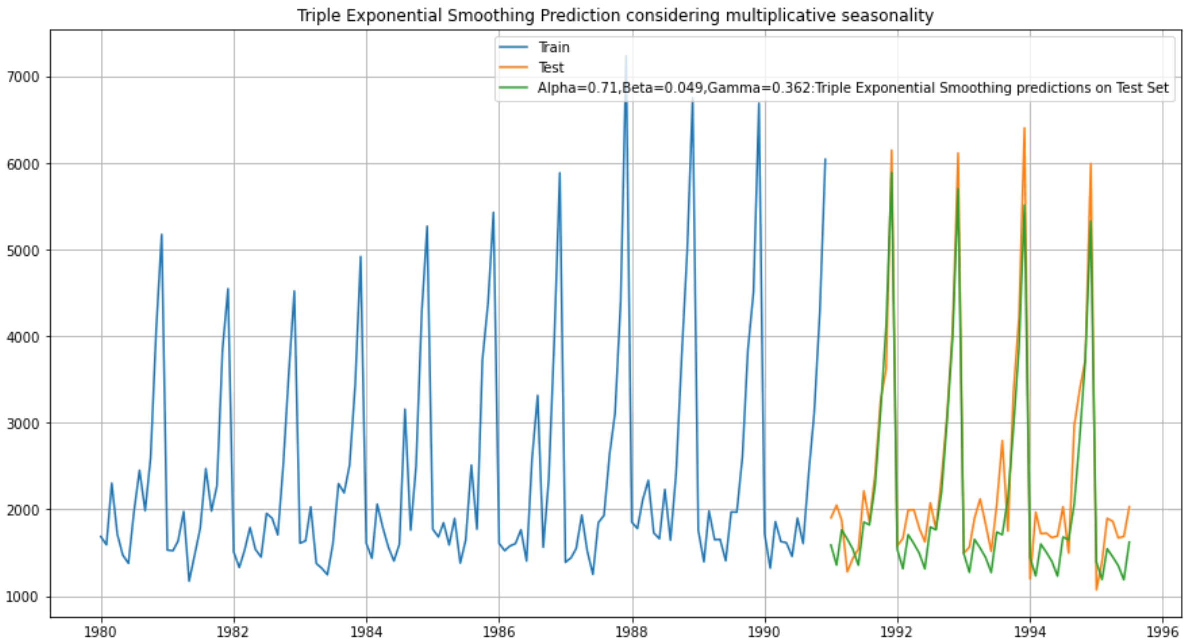



**Fig. 8: Alpha=0.71,Beta=0.049,Gamma=0.362:Triple Exponential Smoothing predictions on Test Set for Sparkling Wine Sales Dataset when considering multiplicative seasonality**

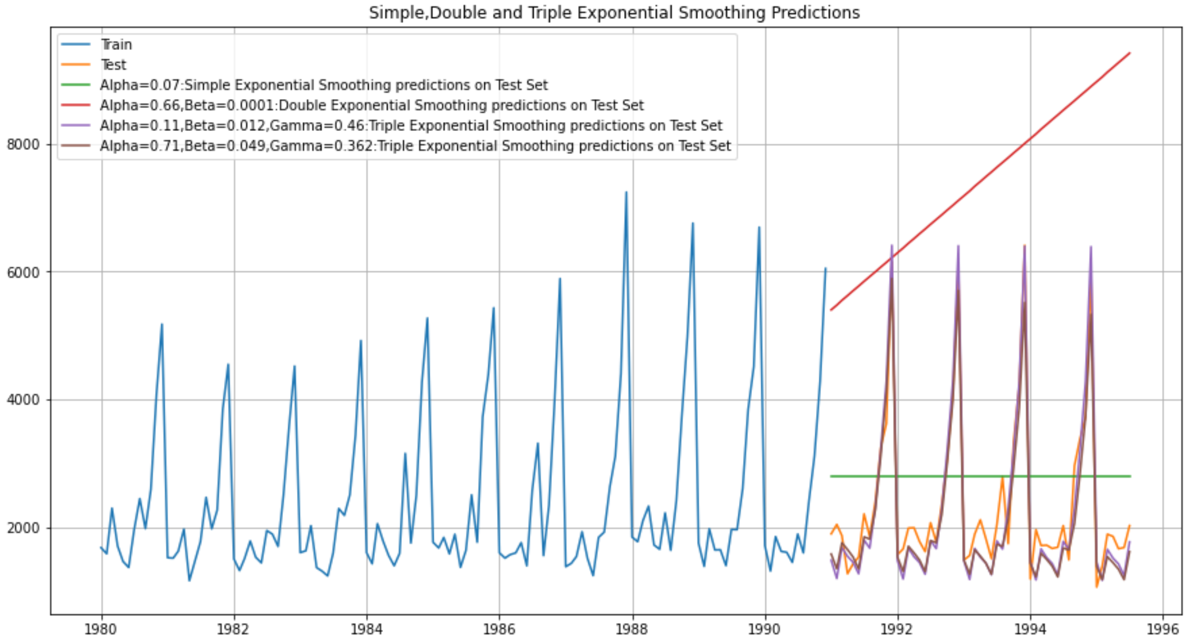



**Fig. 9: Simple,Double and Triple Exponential Smoothing Predictions for Sparkling Wine Sales Dataset**

### Test RMSE for all the Exponential Smoothing used for Sparkling Wine Dataset

| | Test RMSE |
|---|---|
| SES, Alpha=0.07,SES | 1338.008384 |
| Holts, Alpha=.66,Beta=0.0001:DES | 5291.879833 |
| Holt-Winters, Alpha=0.11,Beta=0.012,Gamma=0.46:TES | 378.626008 |
| Holt Winter's linear, Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES | 402.938530 |

## Inference

- Above table represents the RMSE values of all the Exponential Smoothing Methods used on Test Data.
- RMSE of the Holt-Winters Method is less compared to other Exponential Smoothing Methods hence Holt-Winters Method is mostly favorable.
- We see that the multiplicative seasonality model has not done that well when compared to the additive seasonality Triple Exponential Smoothing model.

## Building different models and comparing the accuracy metrics.

### Model 1: Linear Regression

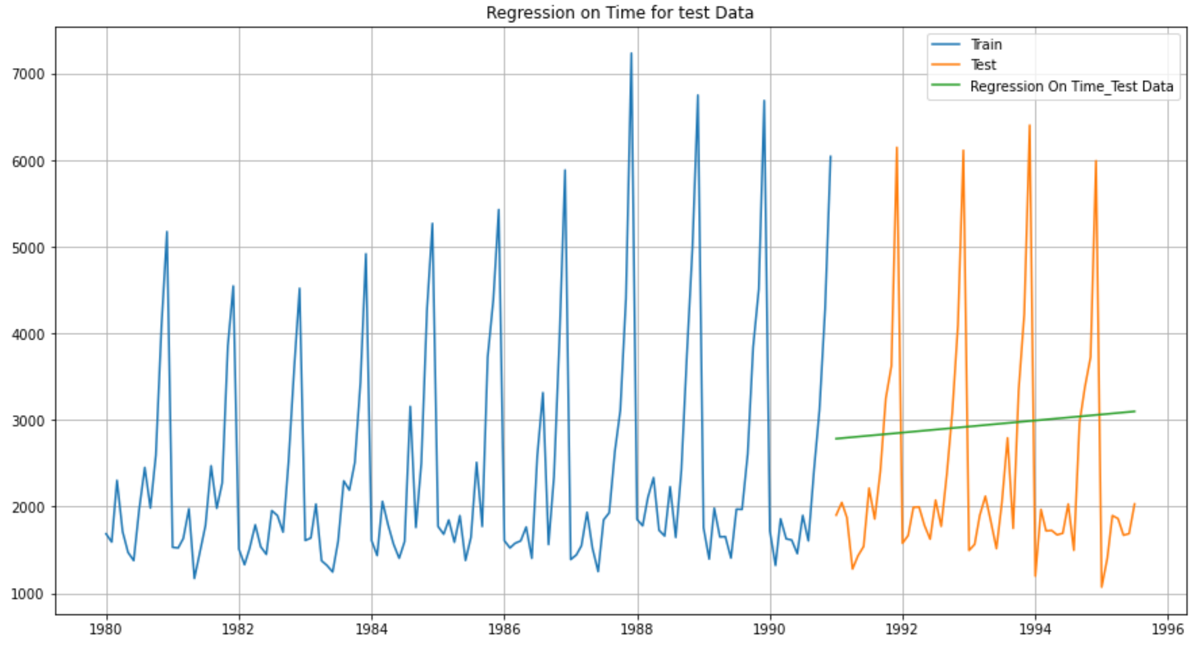



**Fig. 10: Regression on Time Model for Sparkling Wine Sales Dataset**

- For RegressionOnTime forecast on the Test Data, RMSE is 1386.836

## Model 2: Naive Approach: $\hat{y}_{t+1} = y_t$

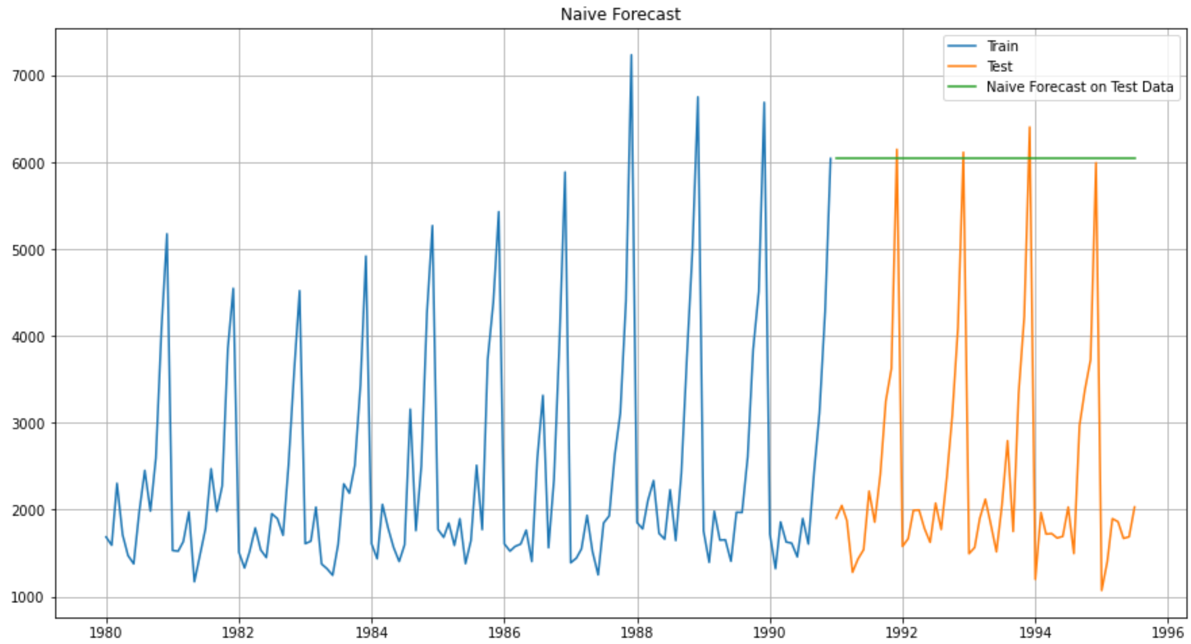



**Fig. 11: Naive Model on test Data for Sparkling Wine Sales Dataset**

- For Naive Model forecast on the Test Data, RMSE is 3864.279

## Method 3: Simple Average

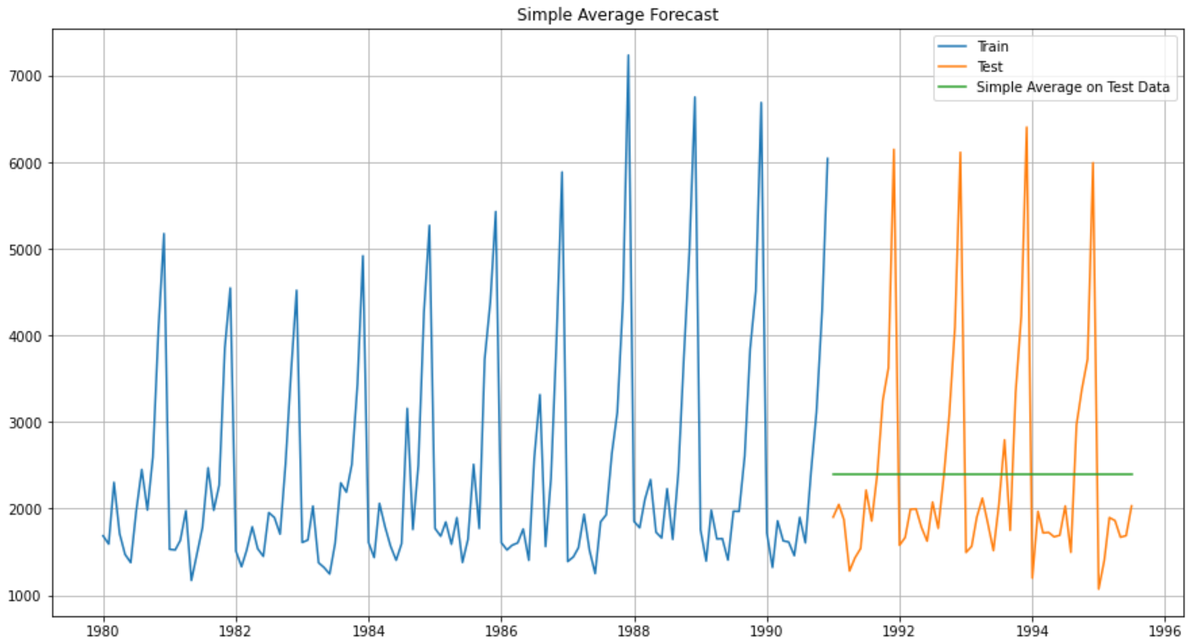



**Fig. 12: Simple Average Model on test data for Sparkling Wine Sales Dataset**

- For Simple Average forecast on the Test Data, RMSE is 1275.082

## Method 4: Moving Average(MA)

- For Moving Average, we consider rolling means (or moving averages) for different intervals.
- The best interval can be determined by the maximum accuracy (or the minimum error) over here.

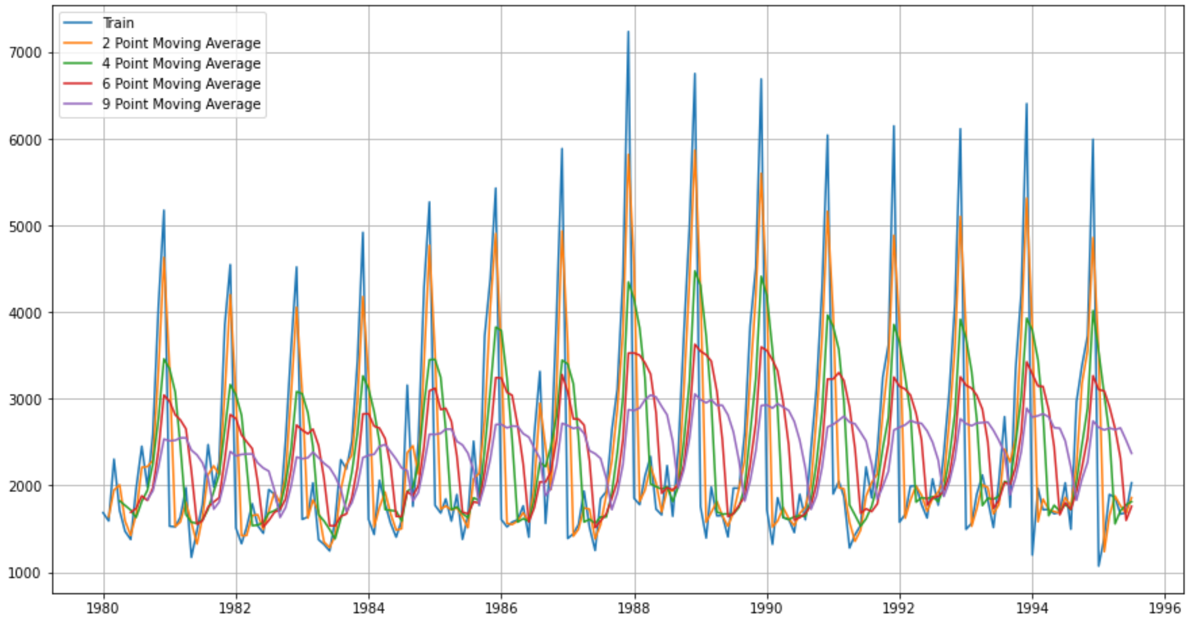



**Fig. 13: Moving Averages at 2,4,6,9 point for Sparkling Wine Sales Dataset**

- We Now Split the Data into Train set and Test Set.
- Test set will start from year 1991

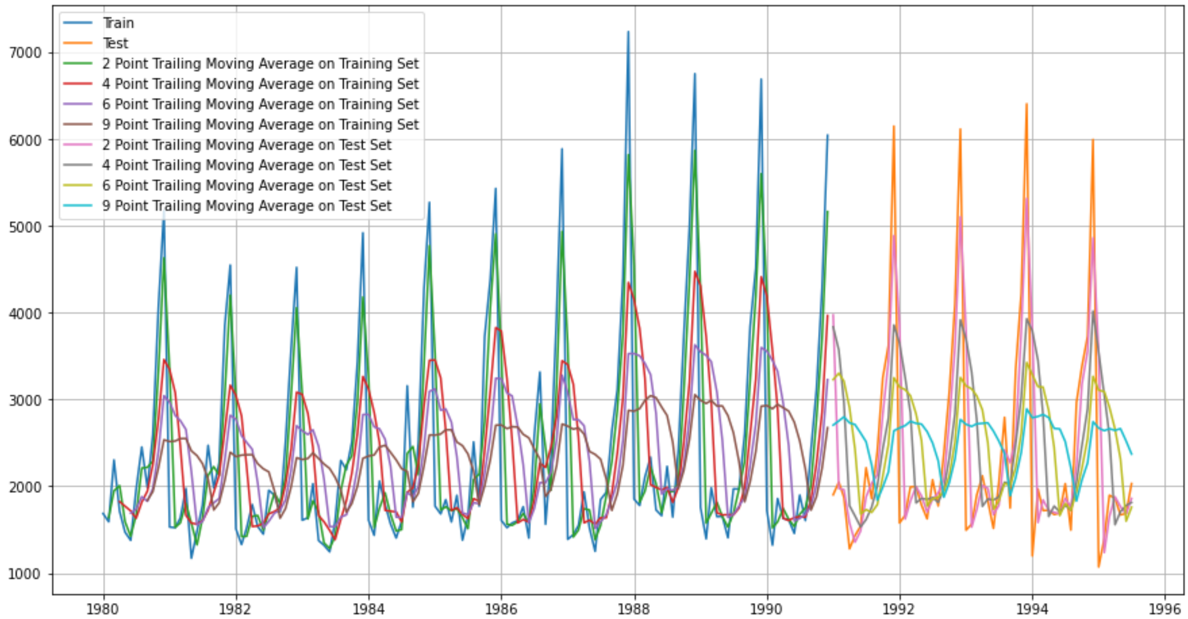



**Fig. 14: Data Split into Train and Test set of Moving Averages at 2,4,6,9 point for Sparkling Wine Sales Dataset**

- For 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401
- For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590
- For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927
- For 9 point Moving Average Model forecast on the Training Data, RMSE is 1346.278

| | Test RMSE |
|---|---|
| Alpha=0.07,SES | 1338.008384 |
| Alpha=.66,Beta=0.0001:DES | 5291.879833 |
| Alpha=0.11,Beta=0.012,Gamma=0.46:TES | 378.626008 |
| Alpha=0.71,Beta=0.049,Gamma=0.362:TES | 402.938530 |
| Linear Regression | 1386.836243 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| 9pointTrailingMovingAverage | 1346.278315 |

## Inference

- We see that Triple Exponential Smoothing method (also called Holt-Winters Method) is the most preferable model as it's RMSE valus is smaller when compared to other Models.

We will further proceed with ARIMA model

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

## found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

## Check for stationarity of the whole Time Series Sparkling Wine Dataset

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:

- $H_0$ : The Time Series has a unit root and is thus non-stationary.
- $H_1$ : 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 $\alpha$ value. $\alpha$ = 0.05.

- DF test statistic is -1.586
- DF test p-value is 0.7980379297590474
- Number of lags used 24

We see that Sparkling Wine Sales is not Stationary, Hence we will take Log transform of the data and apply ADF to check for Stability.

Applied Log Transformation to data and the statistics are as below.

- DF test statistic is -1.871
- DF test p-value is 0.669332209553601
- Number of lags used 24

We can observe that either original nor log-transformed series is stationary. Hence, a stationarization is necessary. Often differencing a non-stationary time series leads to a stationary series.

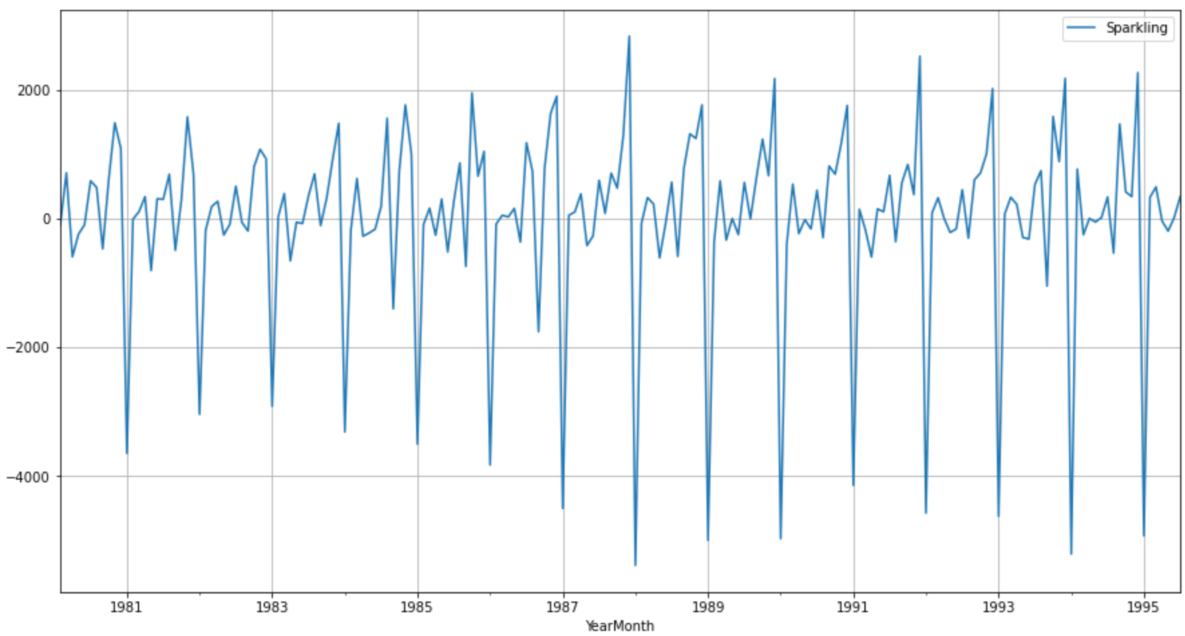



**Fig. 15: Plot of Sparkling Wine Sales Dataset after 1st Order Differencing**

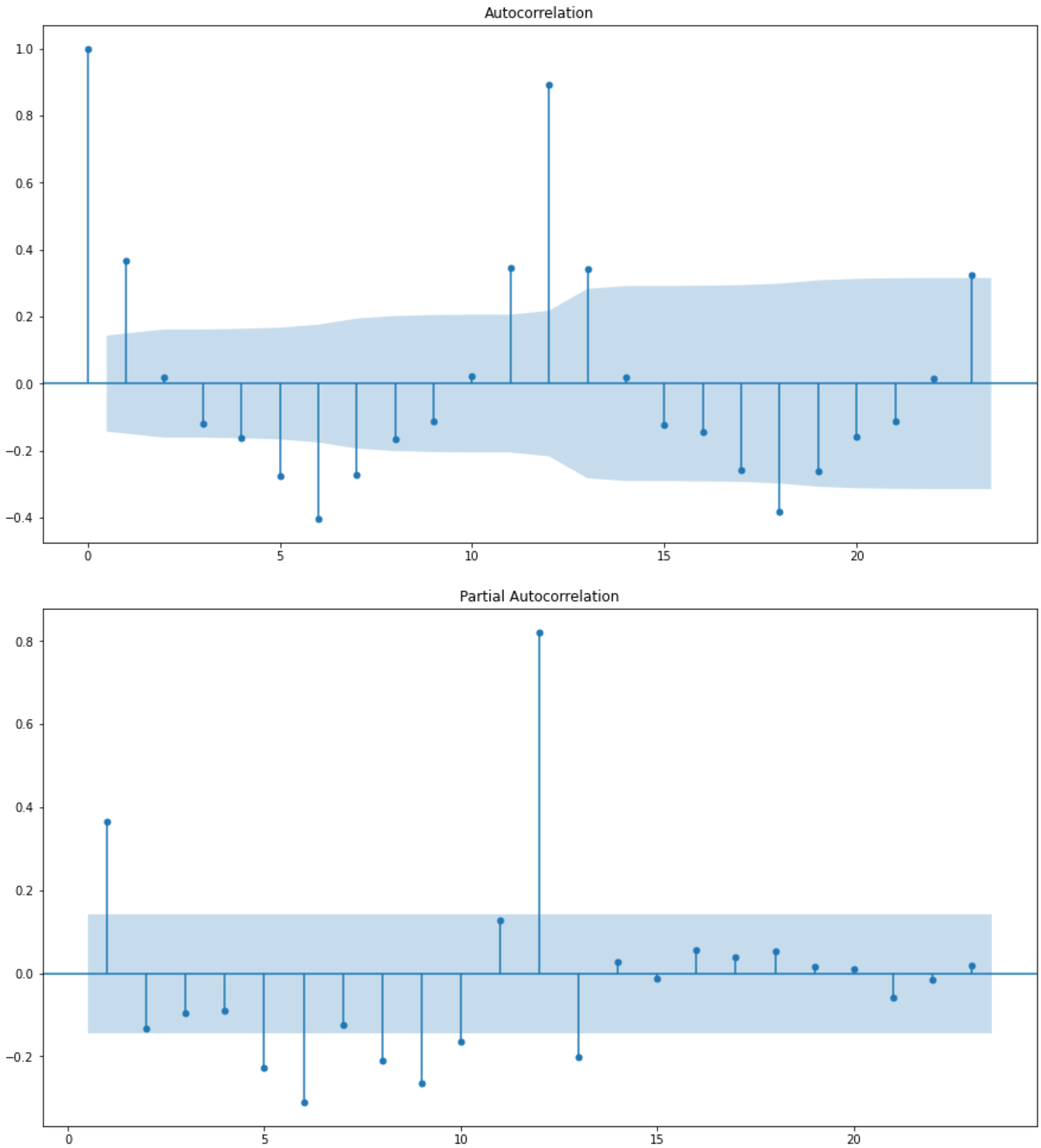



Fig. 16: Plot of PACF and ACF for Sparkling Wine Dataset

## 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.

### 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

- We will Consider p and q to be in the Range of 0 and 3.
- we found data to be stationary at 1st order of Differencing. Hence d = 1
- Hence below table indicates the params that we can use to calculate AIC. Once AIC is calculated, The value with the least AIC score will be considered for ARIMA modeling

**Params**

| Params |
|---|
| Model: (0, 1, 0) |
| 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) |

- After calculating AIC score for each params, the below contains 1st 5 values in ascending order.

| | param | AIC |
|---|---|---|
| 16 | (2, 1, 2) | 2210.626049 |
| 21 | (3, 1, 3) | 2225.661559 |
| 3 | (0, 1, 3) | 2228.672181 |
| 20 | (3, 1, 2) | 2228.927897 |
| 17 | (2, 1, 3) | 2229.358094 |

- From the Above table we can see that AIC value of 2210.62 is the least of all the parmas and respective values of p,d,q are 2,1,2.
- Now Arima Model will be built with Order (2,1,2)

Below are the Observations once the ARIMA model with Order (2,1,2) were built

| | SARIMAX | Results | |
|---|---|---|---|
| Dep. Variable: | Sparkling | No. Observations: | 132 |
| Model: | ARIMA(2, 1, 2) | Log Likelihood | -1101.755 |
| Date: | Wed, 15 Dec 2021 | AIC | 2213.509 |
| Time: | 17:41:34 | BIC | 2227.885 |
| Sample: | 01-01-1980 | HQIC | 2219.351 |
| | - 12-01-1990 | | |

**SARIMAX Results**

| | | | | | | |
|---|---|---|---|---|---|---|
| Covariance Type: | opg | | | | | |
| | **coef** | **std err** | **z** | **P>(z)** | **0.025** | **0.975** |
| ar.L1 | 1.3121 | 0.046 | 28.781 | 0.000 | 1.223 | 1.401 |
| ar.L2 | -0.5593 | 0.072 | -7.741 | 0.000 | -0.701 | -0.418 |
| ma.L1 | -1.9917 | 0.109 | -18.218 | 0.000 | -2.206 | -1.777 |
| ma.L2 | 0.9999 | 0.110 | 9.109 | 0.000 | 0.785 | 1.215 |
| sigma2 | 1.099e+06 | 1.99e-07 | 5.51e+12 | 0.000 | 1.1e+06 | 1.1e+06 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.19 | Jarque-Bera (JB): | 14.46 |
| Prob(Q): | 0.67 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 2.43 | Skew: | 0.61 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 4.08 |

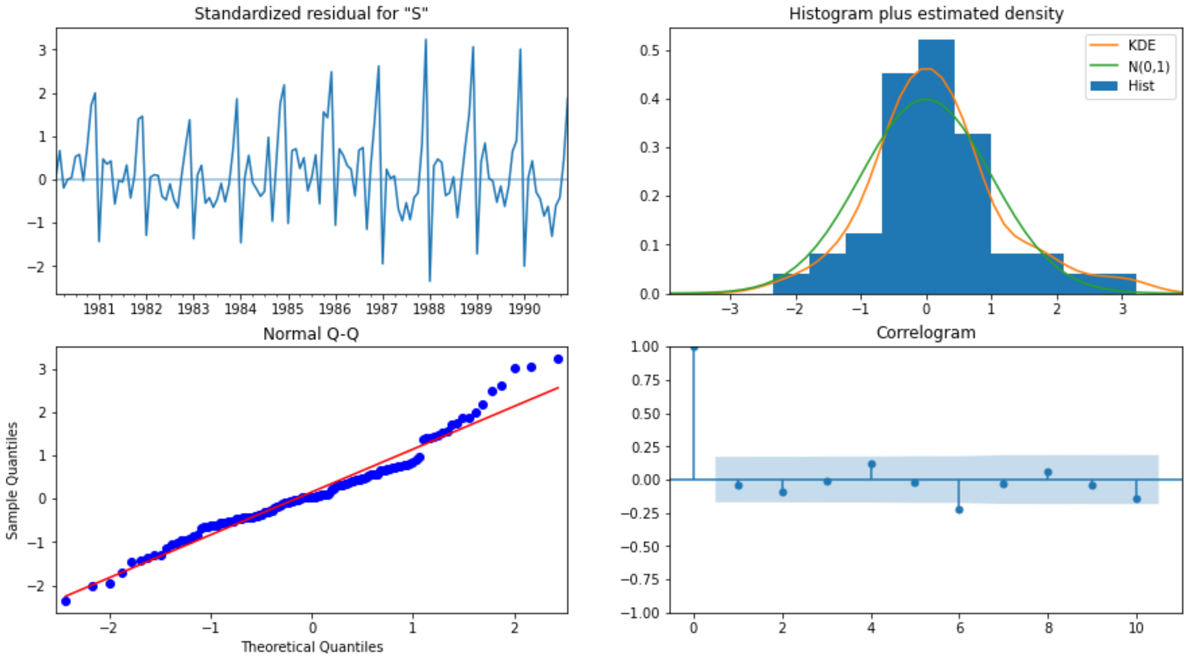



**Fig. 17: Diagnostics Report of Arima Model with Order (2,1,2) for Sparkling Wine Dataset**

- For Auto ARIMA Model forecast on the Training Data, RMSE is 1299.980

## Building a version of the ARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots

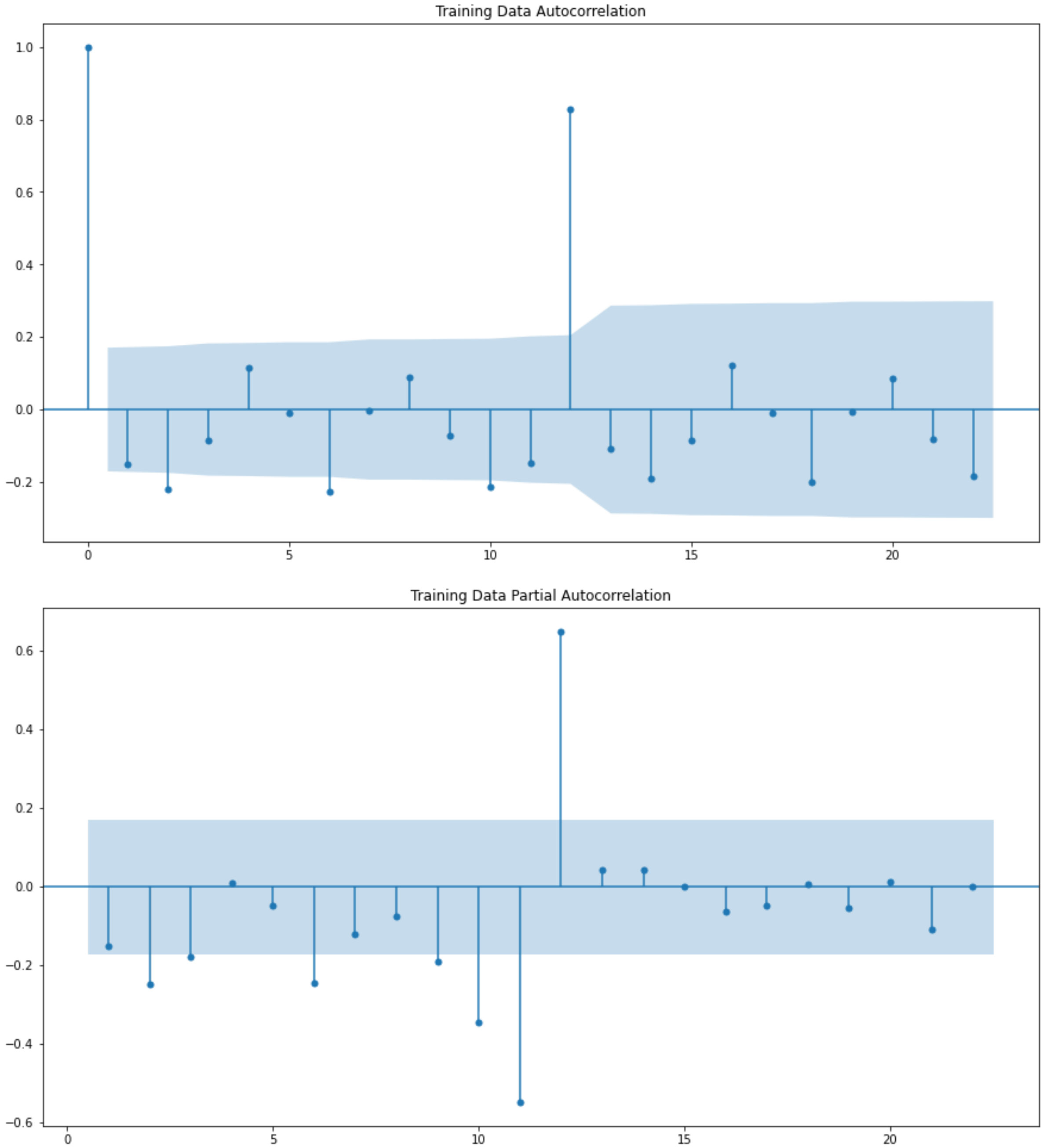



**Fig. 18: PACF and ACF Plot of 1st Order Differentiation on Train data set for Sparkling Wine sales data**

Here, we have taken alpha=0.05.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 2.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 3.

By looking at the above plots, we will take the value of p and q to be 3 and 2 respectively. First order differencing will make the data stationary, hence d = 1.

Once Model is built with Order (3,1,2) the following data is obtained

| | **SARIMAX** | **Results** | |
|---|---|---|---|
| Dep. Variable: | Sparkling | No. Observations: | 132 |
| Model: | ARIMA(3, 1, 2) | Log Likelihood | -1109.476 |

**SARIMAX Results**

| | | | |
|---|---|---|---|
| Date: | Wed, 15 Dec 2021 | AIC | 2230.952 |
| Time: | 18:08:22 | BIC | 2248.204 |
| Sample: | 01-01-1980 | HQIC | 2237.962 |
| | - 12-01-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>(z) | 0.025 | 0.975 |
|---|---|---|---|---|---|---|
| ar.L1 | -0.4155 | 0.043 | -9.746 | 0.000 | -0.499 | -0.332 |
| ar.L2 | 0.3242 | 0.120 | 2.704 | 0.007 | 0.089 | 0.559 |
| ar.L3 | -0.2603 | 0.077 | -3.362 | 0.001 | -0.412 | -0.109 |
| ma.L1 | 0.0218 | 0.134 | 0.163 | 0.871 | -0.241 | 0.284 |
| ma.L2 | -0.9780 | 0.141 | -6.918 | 0.000 | -1.255 | -0.701 |
| sigma2 | 1.327e+06 | 1.94e-07 | 6.86e+12 | 0.000 | 1.33e+06 | 1.33e+06 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.12 | Jarque-Bera (JB): | 3.62 |
| Prob(Q): | 0.73 | Prob(JB): | 0.16 |
| Heteroskedasticity (H): | 2.72 | Skew: | 0.31 |
| Prob(H) (two-sided): | 0.00 | Kurtosis: | 3.52 |

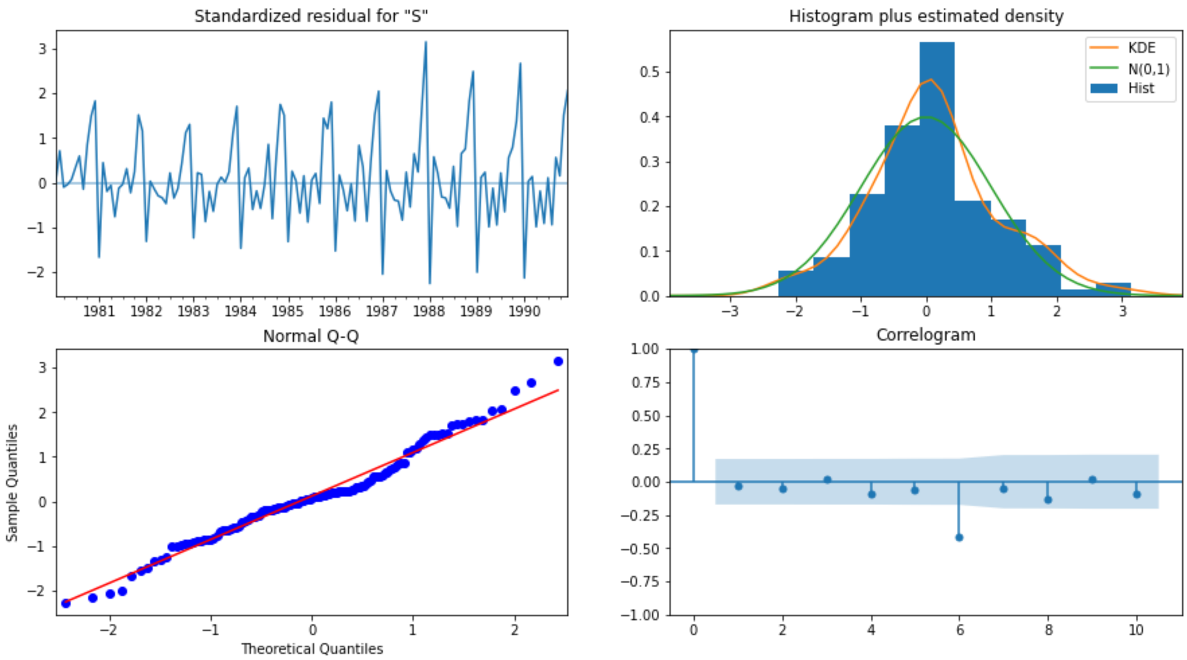



**Fig. 19: Diagnostics Report of Arima Model with Order (3,1,2) for Sparkling Wine Dataset**

- For Manual ARIMA Model forecast on the Training Data, RMSE is 1286.235

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

## SARIMA Model

- We will Consider p, q, P & Q to be in the Range of 0 and 3.
- we found dat to be stationary at 1st order of Differencing. Hence d = 1, D= 0.
- Hence below table indicates the params that we can use to calculate AIC. Once AIC is calculated, The value with the least AIC score will be considered for ARIMA modeling

| Params |
|---|
| 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) |

- After calculating AIC score for each params, the below contains 1st 5 values in ascending order.

| | param | seasonal | AIC |
|---|---|---|---|
| 467 | (3, 1, 1) | (0, 0, 3, 12) | 16.000000 |
| 499 | (3, 1, 3) | (0, 0, 3, 12) | 357.367162 |
| 307 | (0, 1, 3) | (0, 0, 3, 12) | 737.043695 |
| 508 | (3, 1, 3) | (3, 0, 0, 12) | 1387.497014 |
| 476 | (3, 1, 1) | (3, 0, 0, 12) | 1387.788331 |

**SARIMAX Results**

| | | | |
|---|---|---|---|
| Dep. Variable: | Sparkling | No. Observations: | 132 |
| Model: | SARIMAX(3, 1, 3)x(3, 0, [], 12) | Log Likelihood | -683.749 |
| Date: | Wed, 15 Dec 2021 | AIC | 1387.497 |
| Time: | 20:40:32 | BIC | 1412.715 |
| Sample: | 01-01-1980 | HQIC | 1397.675 |
| | - 12-01-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>(z) | 0.025 | 0.975 |
|---|---|---|---|---|---|---|
| ar.L1 | -1.6747 | 0.142 | -11.813 | 0.000 | -1.953 | -1.397 |
| ar.L2 | -0.7437 | 0.258 | -2.885 | 0.004 | -1.249 | -0.238 |
| ar.L3 | -0.0023 | 0.144 | -0.016 | 0.987 | -0.285 | 0.280 |
| ma.L1 | 1.0549 | 0.191 | 5.510 | 0.000 | 0.680 | 1.430 |
| ma.L2 | -0.7784 | 0.172 | -4.522 | 0.000 | -1.116 | -0.441 |
| ma.L3 | -0.9066 | 0.148 | -6.140 | 0.000 | -1.196 | -0.617 |
| ar.S.L12 | 0.5331 | 0.118 | 4.523 | 0.000 | 0.302 | 0.764 |
| ar.S.L24 | 0.2786 | 0.116 | 2.400 | 0.016 | 0.051 | 0.506 |
| ar.S.L36 | 0.2390 | 0.102 | 2.350 | 0.019 | 0.040 | 0.438 |
| sigma2 | 1.527e+05 | 1.94e-06 | 7.89e+10 | 0.000 | 1.53e+05 | 1.53e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.01 | Jarque-Bera (JB): | 4.33 |
| Prob(Q): | 0.93 | Prob(JB): | 0.12 |
| Heteroskedasticity (H): | 1.26 | Skew: | 0.30 |
| Prob(H) (two-sided): | 0.52 | Kurtosis: | 3.88 |

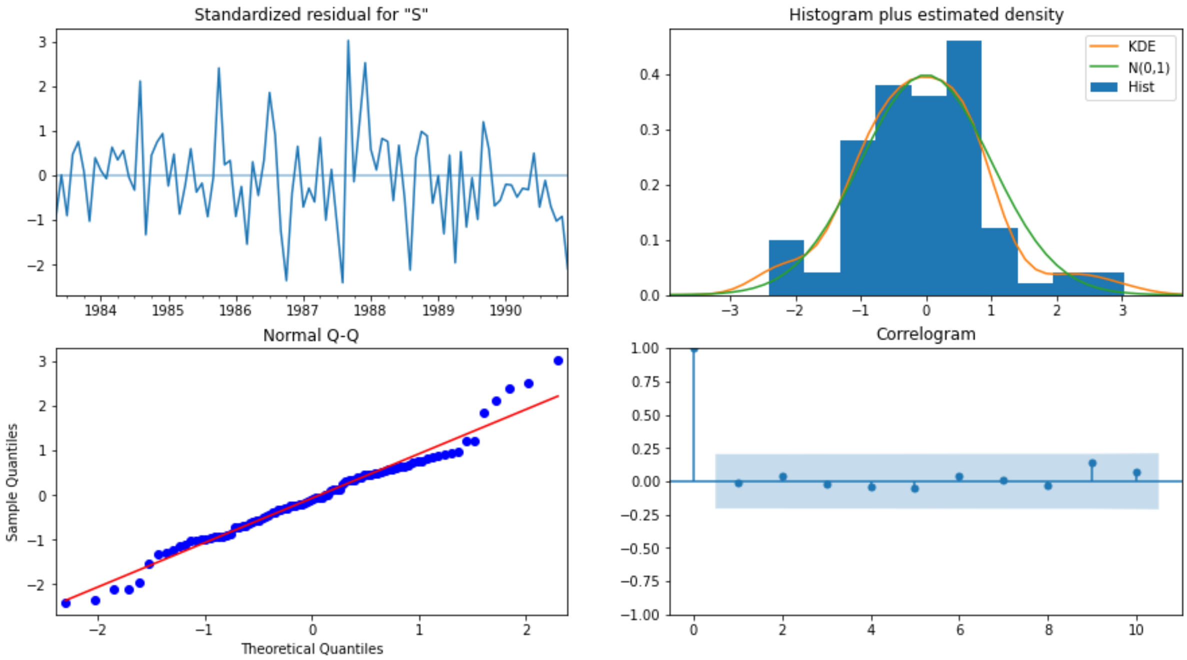



**Fig. 20: Diagnostics Report of SARIMA Model with Order (3,1,3)(3, 0, 0, 12) for Sparkling Wine Dataset**

- For Auto SARIMA Model forecast on the Training Data, RMSE is 611.271

## Building a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 12.

- From Fig 18, Observing the Plot of PACF and ACF of 1st Order difference of Train dataset, we can Conclude that P & Q are 3 and 1 Respectively.
- Building SARIMA Model, following observes are observed.

**SARIMAX Results**

| | | | |
|---|---|---|---|
| Dep. Variable: | Sparkling | No. Observations: | 132 |
| Model: | SARIMAX(3, 1, 2)x(3, 0, [1], 12) | Log Likelihood | -684.301 |
| Date: | Wed, 15 Dec 2021 | AIC | 1388.603 |
| Time: | 20:43:32 | BIC | 1413.821 |
| Sample: | 01-01-1980 | HQIC | 1398.781 |
| | - 12-01-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>(z) | 0.025 | 0.975 |
|---|---|---|---|---|---|---|
| ar.L1 | -0.5430 | 0.416 | -1.306 | 0.192 | -1.358 | 0.272 |
| ar.L2 | -0.0074 | 0.198 | -0.037 | 0.970 | -0.396 | 0.381 |
| ar.L3 | 0.0638 | 0.140 | 0.455 | 0.649 | -0.211 | 0.339 |
| ma.L1 | -0.1995 | 0.404 | -0.494 | 0.622 | -0.992 | 0.593 |
| ma.L2 | -0.6547 | 0.327 | -2.004 | 0.045 | -1.295 | -0.014 |
| ar.S.L12 | 0.7652 | 0.448 | 1.707 | 0.088 | -0.113 | 1.644 |
| ar.S.L24 | 0.1091 | 0.330 | 0.331 | 0.741 | -0.537 | 0.756 |
| ar.S.L36 | 0.1764 | 0.186 | 0.946 | 0.344 | -0.189 | 0.542 |
| ma.S.L12 | -0.2428 | 0.451 | -0.539 | 0.590 | -1.126 | 0.640 |
| sigma2 | 1.663e+05 | 2.63e+04 | 6.326 | 0.000 | 1.15e+05 | 2.18e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.00 | Jarque-Bera (JB): | 9.35 |
| Prob(Q): | 0.96 | Prob(JB): | 0.01 |
| Heteroskedasticity (H): | 1.25 | Skew: | 0.35 |
| Prob(H) (two-sided): | 0.54 | Kurtosis: | 4.40 |

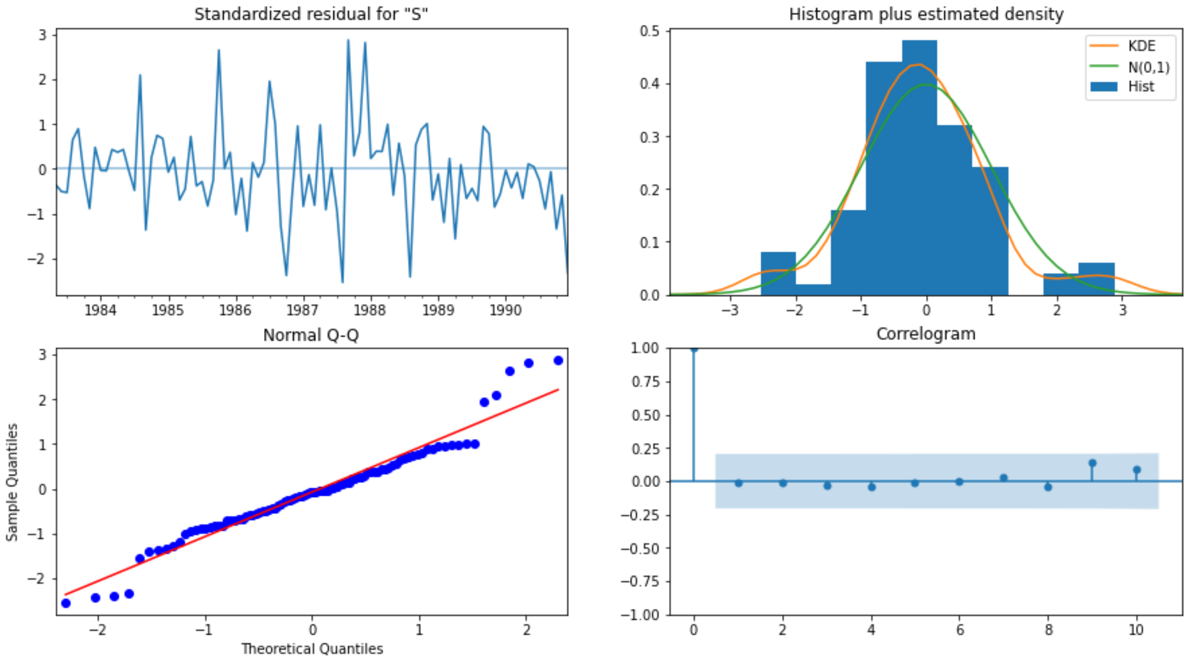


**Fig. 21: Diagnostics Report of SARIMA Model with Order (3,1,2)(3, 0, 1, 12) for Sparkling Wine Dataset**

- For Manual SARIMA Model forecast on the Training Data, RMSE is 579.536

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

Now we Combine all the RMSE values of various Models. Below is the Table containing all the values

| | Test RMSE |
|---|---|
| Alpha=0.07,SES | 1338.008384 |
| Alpha=.66,Beta=0.0001:DES | 5291.879833 |
| Alpha=0.11,Beta=0.012,Gamma=0.46:TES | 378.626008 |
| Alpha=0.71,Beta=0.049,Gamma=0.362:TES | 402.938530 |
| Linear Regression | 1386.836243 |
| NaiveModel | 3864.279352 |
| SimpleAverageModel | 1275.081804 |
| 2pointTrailingMovingAverage | 813.400684 |
| 4pointTrailingMovingAverage | 1156.589694 |
| 6pointTrailingMovingAverage | 1283.927428 |
| 9pointTrailingMovingAverage | 1346.278315 |
| AutoARIMA(2,1,2) | 1299.979640 |
| ManualARIMA(3,1,2) | 1286.234646 |
| AutoSARIMA(3, 1, 3)(3, 0, 0, 12) | 611.271445 |
| ManualSARIMA(3, 1, 2)(3, 0, 1, 12) | 579.536269 |

## Inference

- Most optimum Model is Triple Exponential Smoothing or Holt-Winters Method cause the RMSE of Test Data is Less compared to all other Models.

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

## Forecasting for the next 12 months using Holt-Winters method.

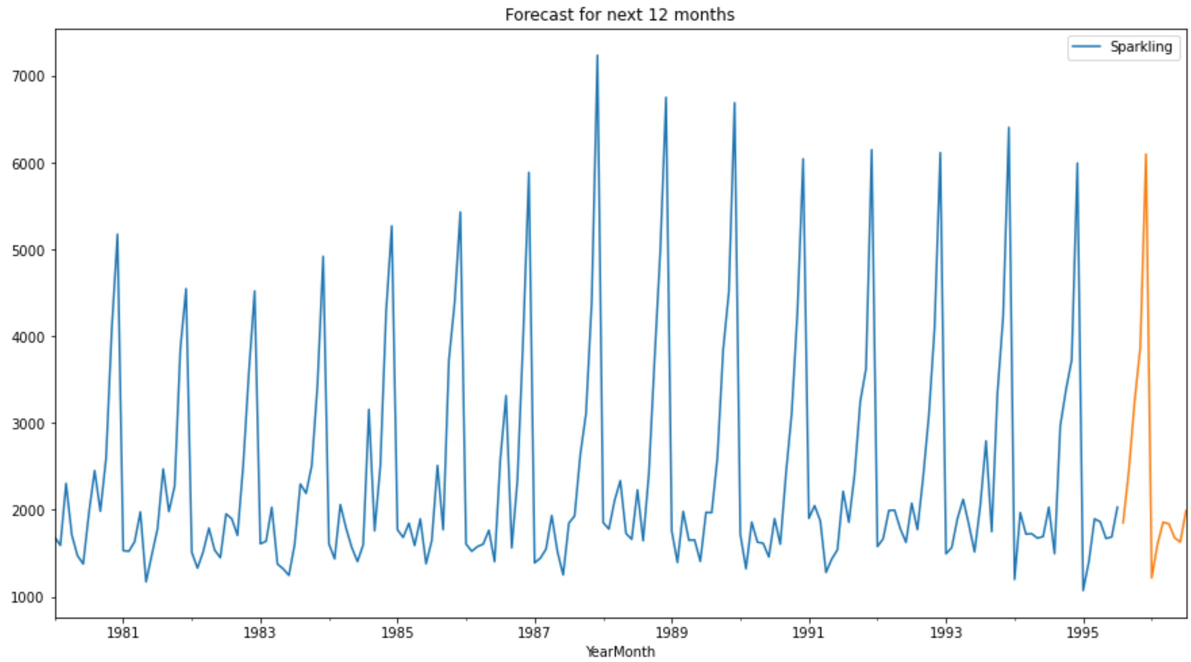



**Fig. 22: Forecasting for the next 12 months using Holt-Winters method for Sparkling Wine Dataset**

- The RMSE Value of the Forecasted Data is RMSE: 368.00814865080315

---

## Time series Forecasting on Rose Wine Data

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

## Plot of Rose Wine sales

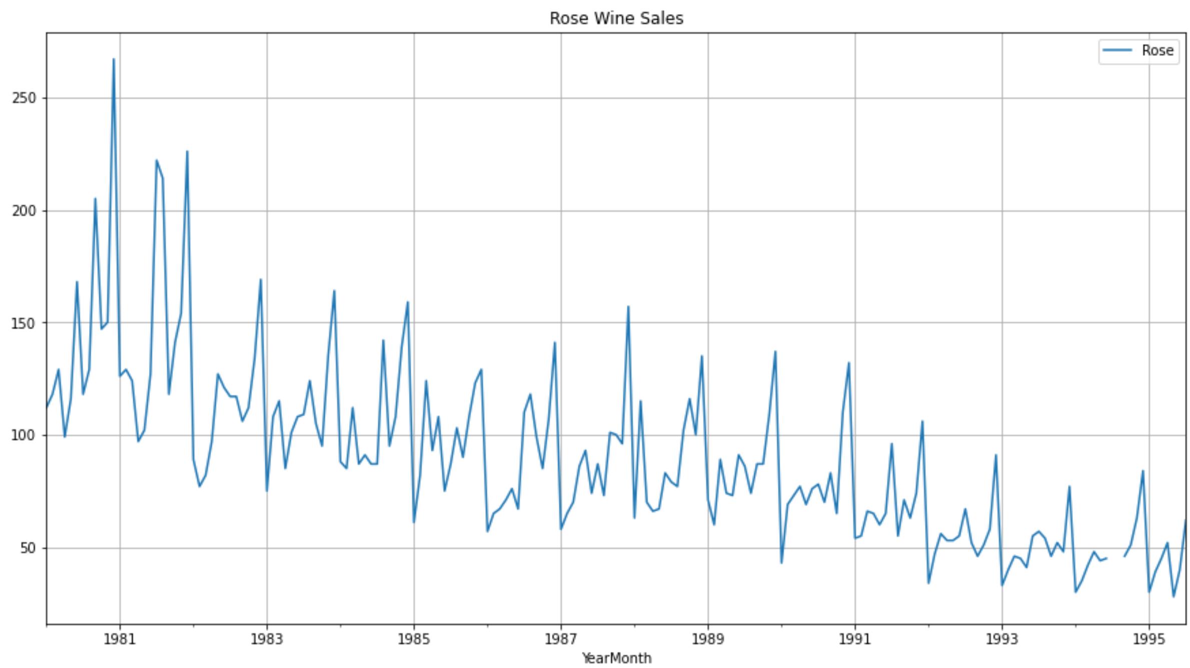



**Fig. 23: Rose Wine Sales Data**

- From the Above Graph we can see that there is a split in Data

- This Indicates that there are missing values. Missing Values are between the year 1994 and 1995
- The Data between 1994 and 1995 is as shown below.

| YearMonth | Rose |
|---|---|
| 1994-01-01 | 30.0 |
| 1994-02-01 | 35.0 |
| 1994-03-01 | 42.0 |
| 1994-04-01 | 48.0 |
| 1994-05-01 | 44.0 |
| 1994-06-01 | 45.0 |
| 1994-07-01 | NaN |
| 1994-08-01 | NaN |
| 1994-09-01 | 46.0 |
| 1994-10-01 | 51.0 |
| 1994-11-01 | 63.0 |
| 1994-12-01 | 84.0 |

- From the Above data we only see 2 missing values
- We can use Interpolation technique to replace the missing values. We we not be removed NaN values as the Data Observation must be contiguous.
- Below table shows how many observations are present in the data.
- We Can also see that Time period is from 1980-01-01 to 1995-07-01.
- Interpolating Missing Values But Only Upto Two Values coz only 2 values are missing

## DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01

Data columns (total 1 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Rose | 185 non-null | float64 |

| 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 |
| ... | ... |
| 1994-05-01 | 44.0 |
| 1994-06-01 | 45.0 |
| 1994-07-01 | 45.33 |

| YearMonth | Rose |
|---|---|
| 1994-08-01 | 45.66 |
| 1994-09-01 | 46.00 |
| ... | ... |
| 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 |

- After Interpolating, the new Data set is as shown above.
- We can see that 2 missing values are now updated.
- Plot of the new Dataset for Rose Wine Sales is as shown Below.

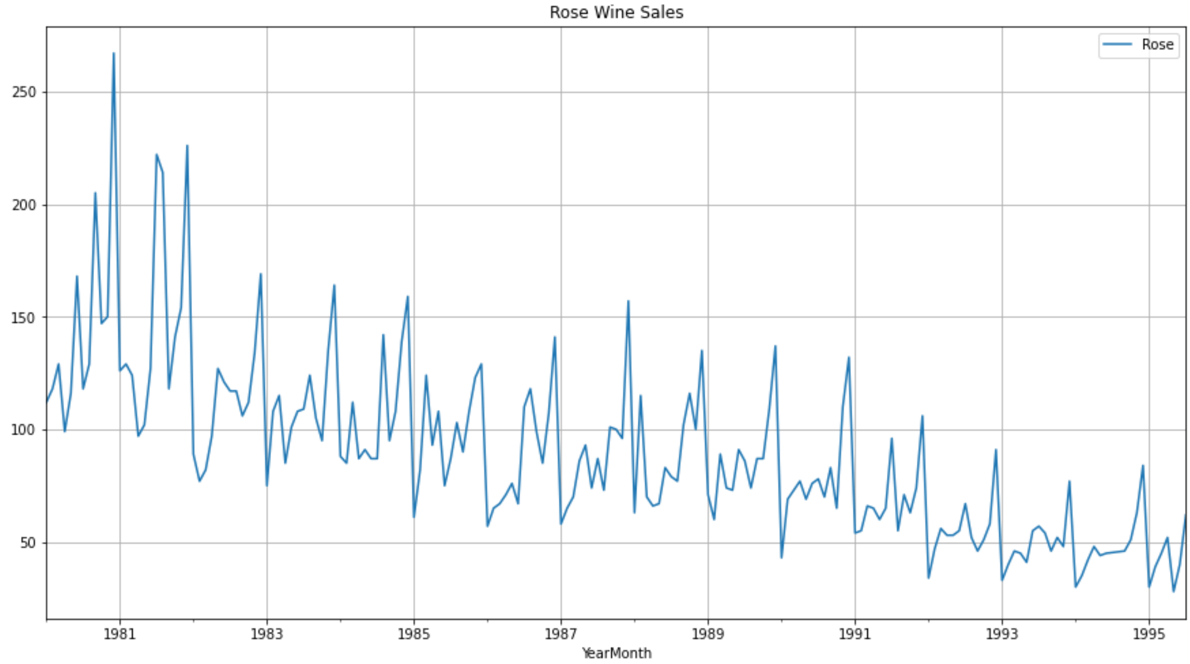



**Fig. 24: Rose Wine Sales Data (Treated Missing Values)**

- Observe steady Decrease in sales every year, indicating downward Trend
- Intra-year stable fluctuations are indicative of seasonal component.

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

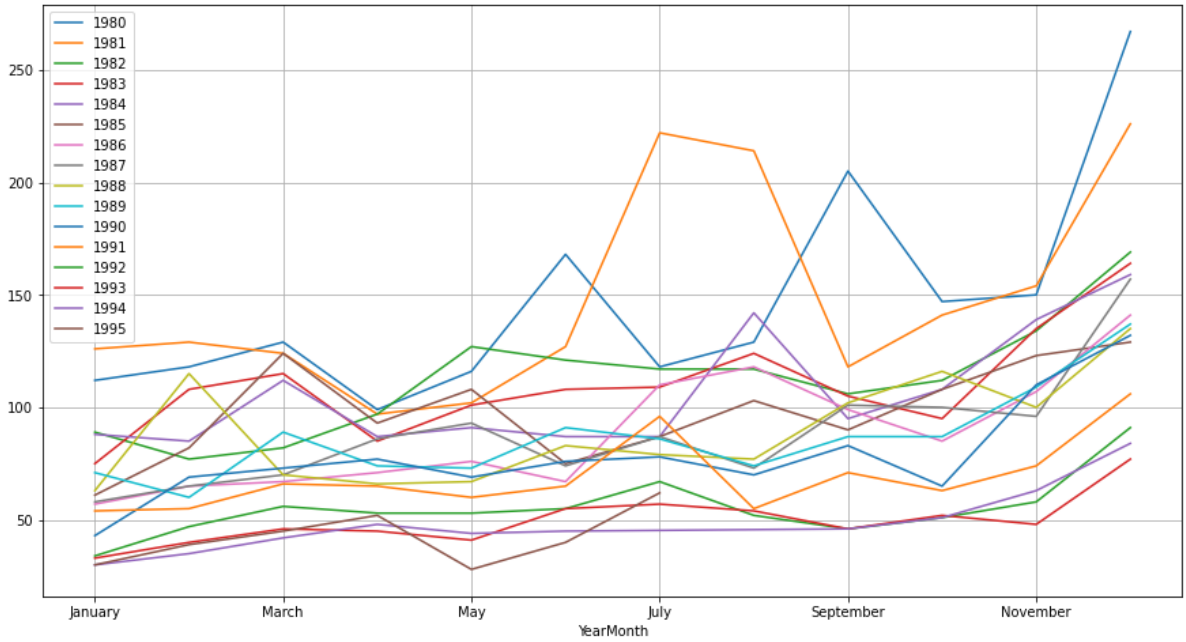



**Fig. 25: Seasonal plot Year Month wise for Rose Wine Sales**

- Figure 25 shows that the sales of Rose Wine which identifies seasonal fluctuations
- We can observe that the average sales is higher in the month of December as compared to other months and also a small rise in month of July-August.

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

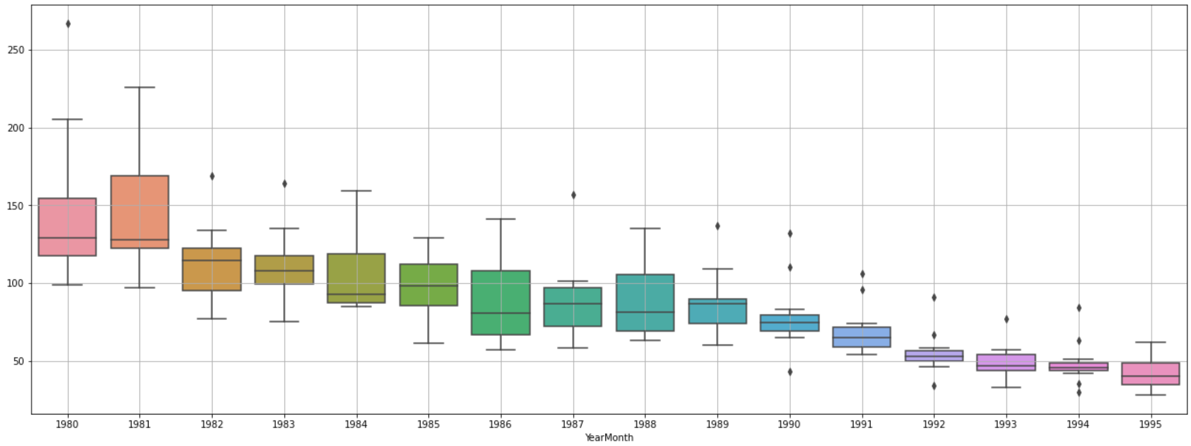



**Fig. 26: Yearly Boxplot**

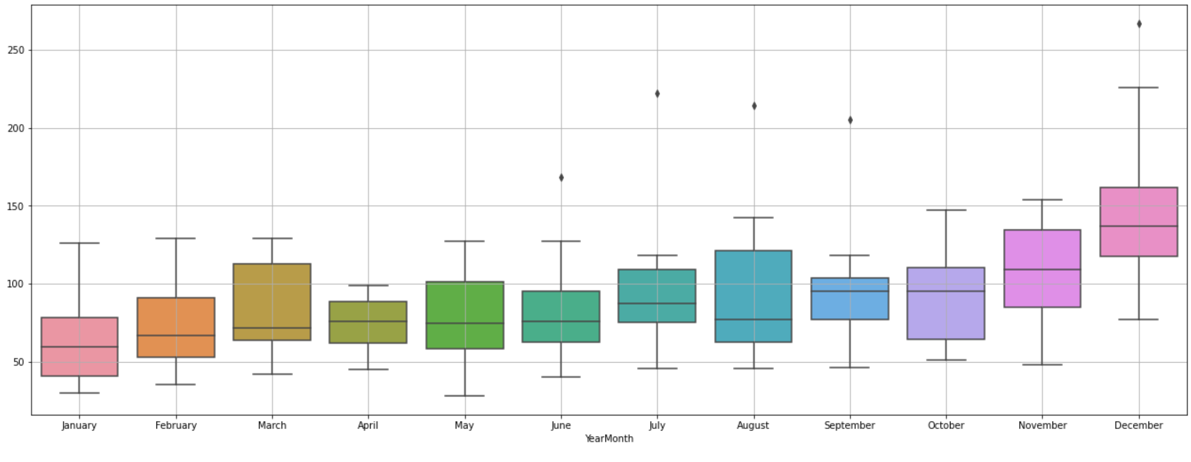



**Fig. 27: Monthly Boxplot**

## Decomposition of Rose Data set

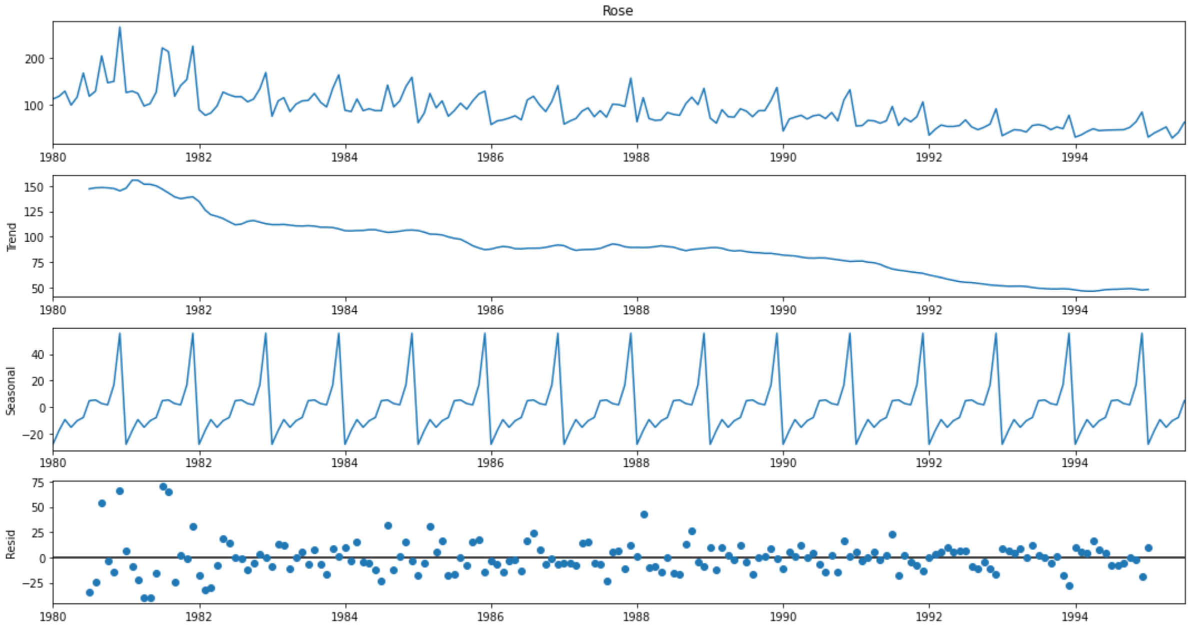



Fig. 28: Decomposition of Rose Wine Dataset

- From the above Figure we can observe that there is a downward trend is present.
- Also we can observe Seasonality in the Data and residuals are random, Additive seasonlity is considered

## 3. Splitting the data into training and test. The test data will start from 1991

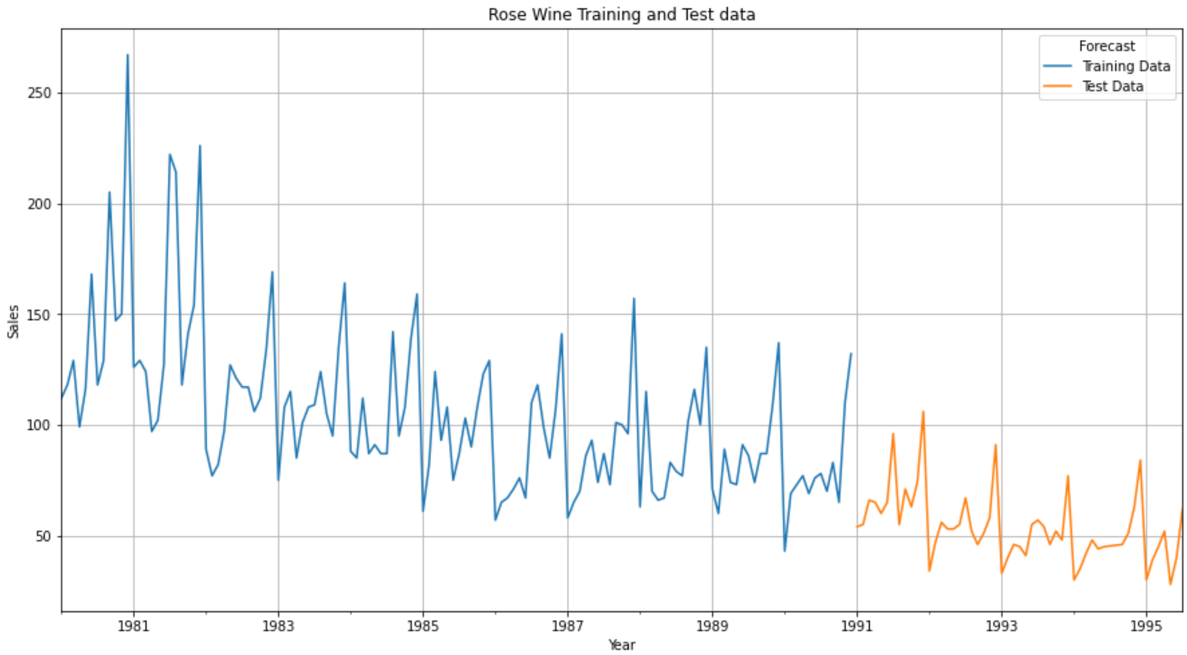



Fig. 29: Plot of Training and test dataset for Rose Wine Sales

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

# Exponential Smoothing Method for Sparkling Wine Sales Dataset

## SES - ETS(A, N, N) - Simple Exponential Smoothing with additive errors

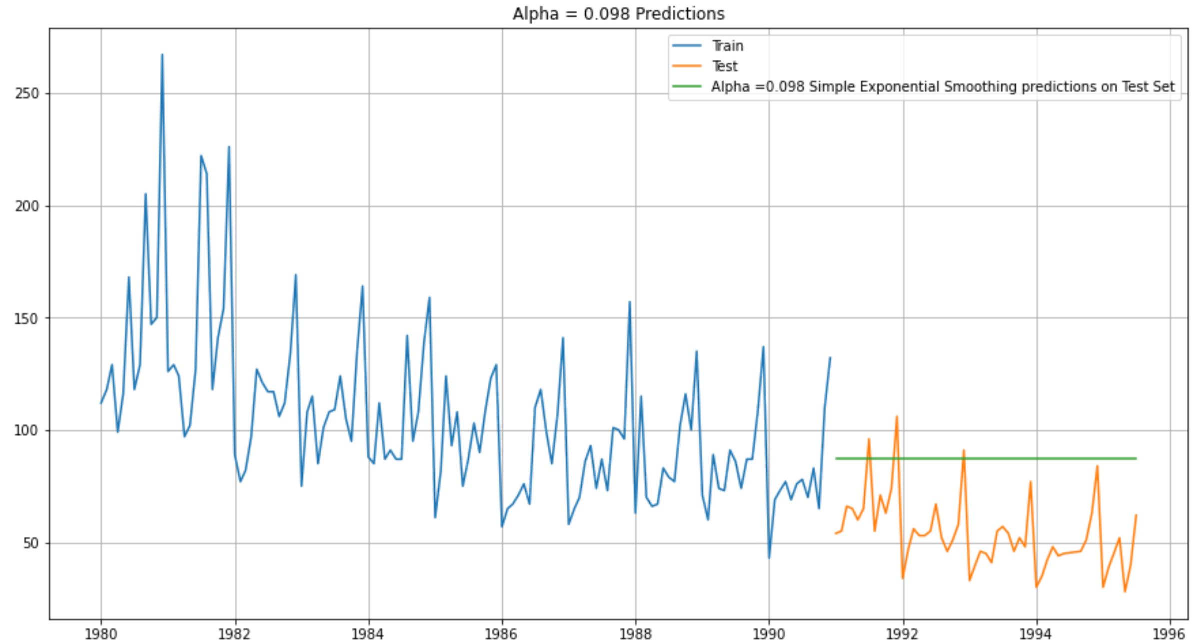



**Fig. 30: Alpha =0.098, Simple Exponential Smoothing predictions on Test Set for Rose Wine Sales Dataset**

- SES RMSE: 36.79624054770398

## Holt - ETS(A, A, N) - Holt's linear method with additive errors

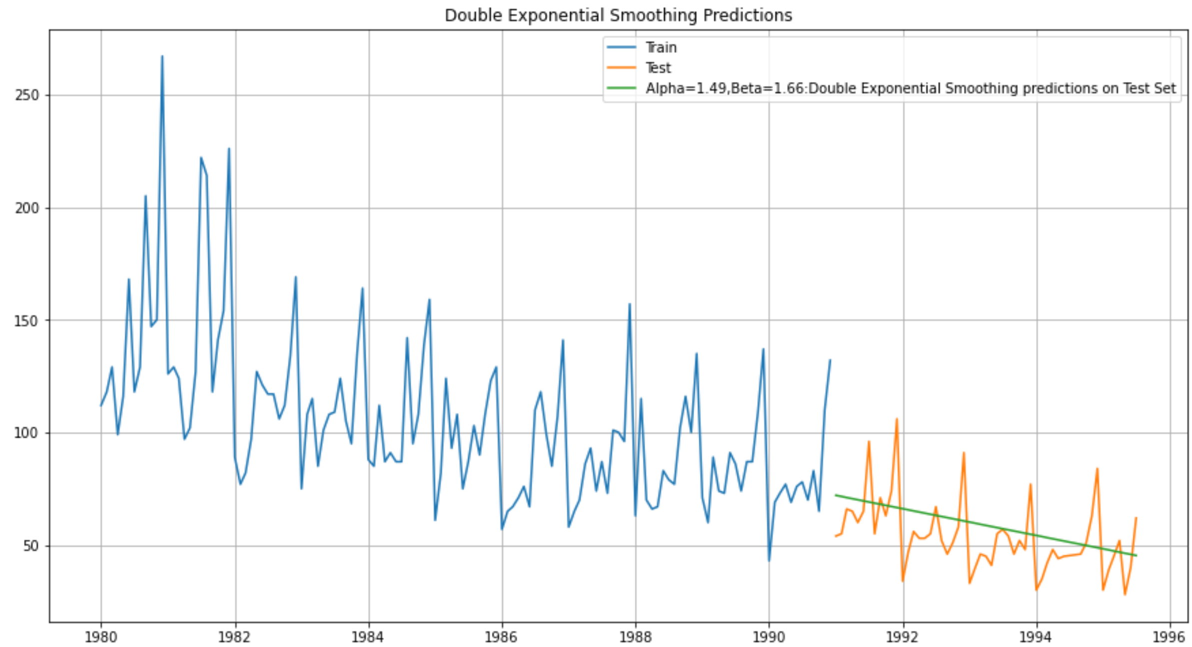



**Fig. 31: Alpha=1.49,Beta=1.66:Double Exponential Smoothing predictions on Test Set for Rose Wine Sales Dataset**

- DES RMSE: 15.268943764436564

## Holt-Winters - ETS(A, A, A) - Holt Winter's linear method with additive errors

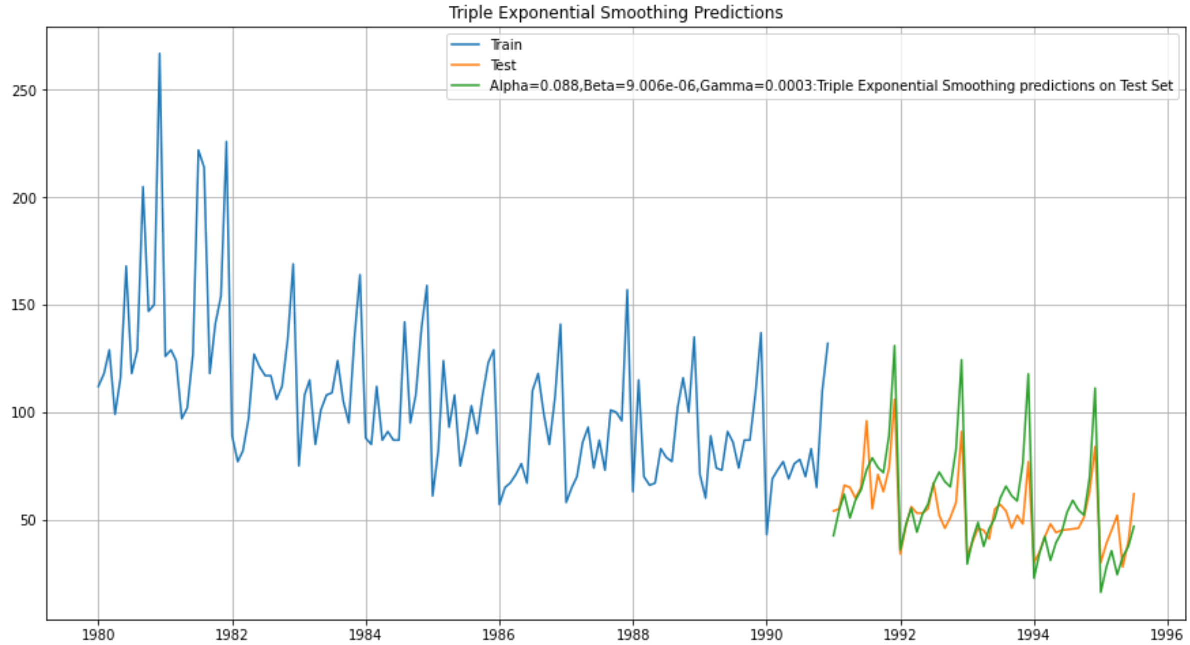



**Fig. 32: Alpha=0.088,Beta=9.006e-06,Gamma=0.0003 Triple Exponential Smoothing predictions on Test Set for Rose Wine Sales Dataset**

- TES RMSE: 14.25561890068539

## Holt-Winters - ETS(A, A, M) - Holt Winter's linear method

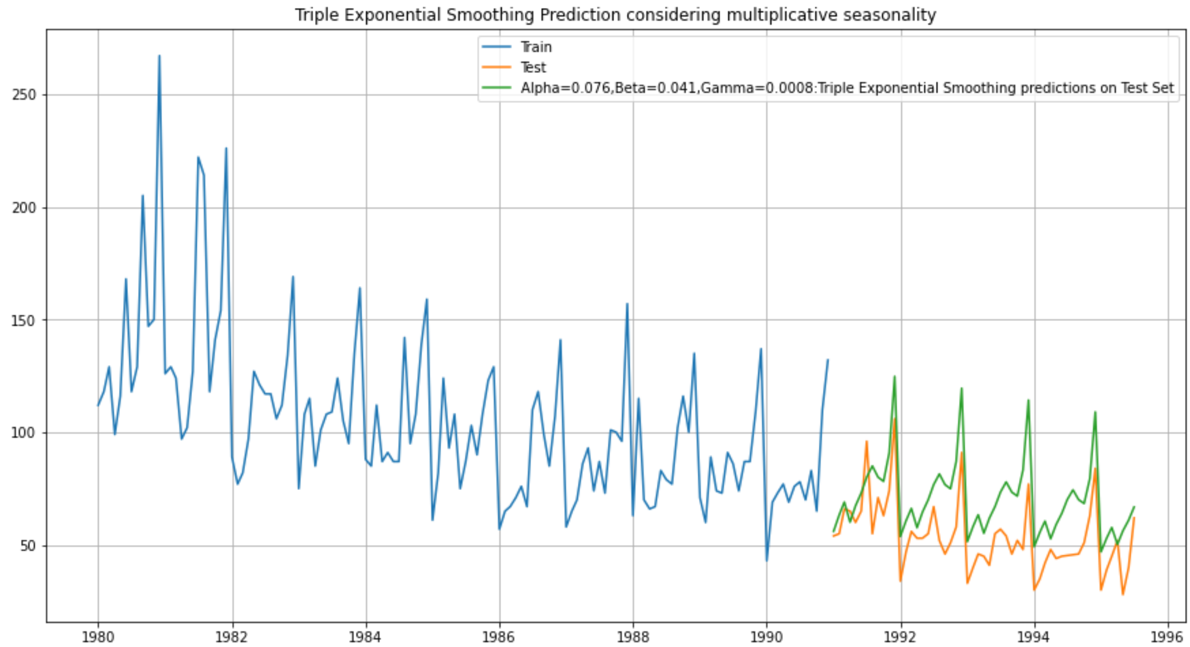



**Fig. 33: Alpha=0.076,Beta=0.041,Gamma=0.0008:Triple Exponential Smoothing predictions on Test Set considering multiplicative seasonality for Rose Wine Sales Dataset**

TES_am RMSE: 19.39714270628386

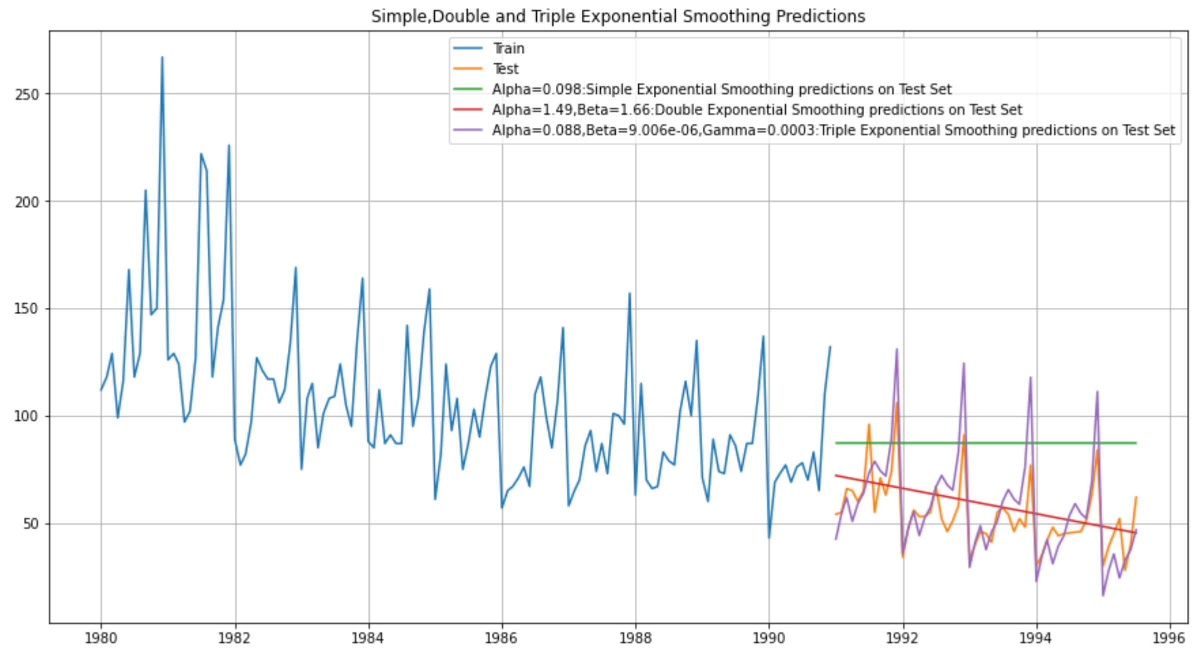



**Fig. 34: Simple,Double and Triple Exponential Smoothing Predictions for Rose Wine Sales Dataset**

### Test RMSE for all the Exponential Smoothing used for Rose Wine Dataset

| | Test RMSE |
|---|---|
| Alpha=0.098,SES | 36.796241 |
| Alpha=1.49,Beta=1.66:DES | 15.268944 |
| Alpha=0.088,Beta=9.006e-06,Gamma=0.0003:TES | 14.255619 |
| Alpha=0.076,Beta=0.041,Gamma=0.0008:TESm | 19.397143 |

## Inference

- Above table represents the RMSE values of all the Exponential Smoothing Methods used on Test Data.
- - RMSE of the Holt-Winters Method is less compared to other Exponential Smoothing Methods hence Holt-Winters Method is mostly favorable.
- We see that the multiplicative seasonality model has not done that well when compared to the additive seasonality Triple Exponential Smoothing model.
- We will further proceed with other methods and compare Accuracy Metrics.

## Building different models and comparing the accuracy metrics.

### Model 1: Linear Regression

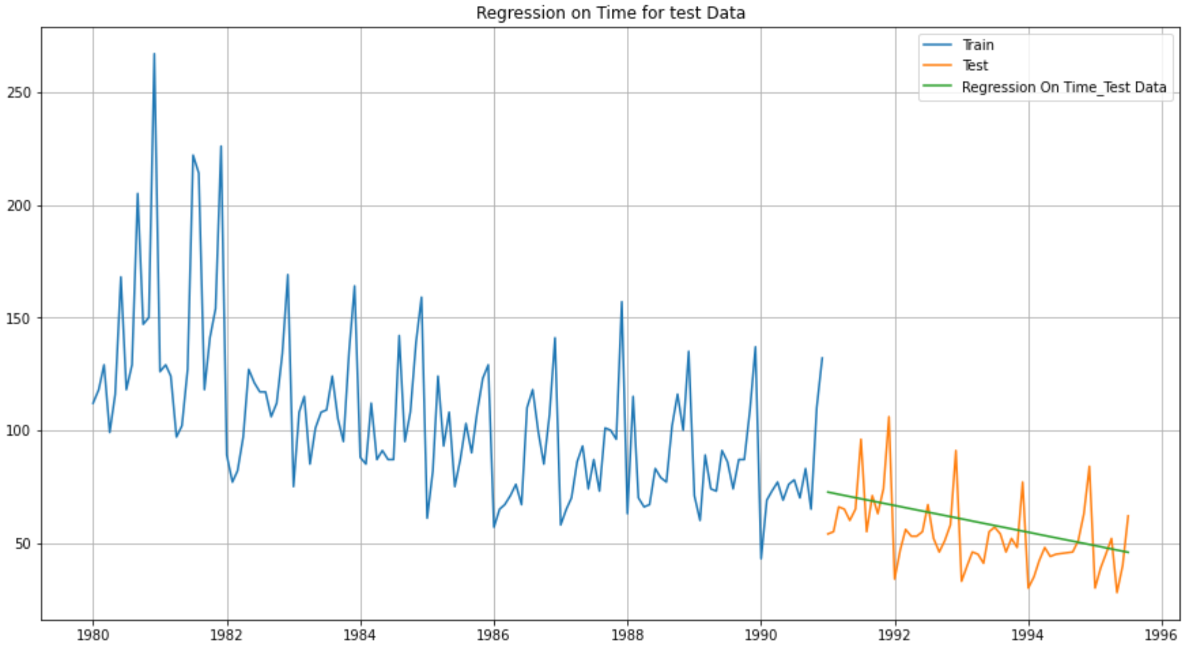



**Fig. 35: Regression on Time Model for Rose Wine Sales Dataset**

For RegressionOnTime forecast on the Test Data, RMSE is 15.433

## Model 2: Naive Approach: $\hat{y}_{t+1} = y_t$

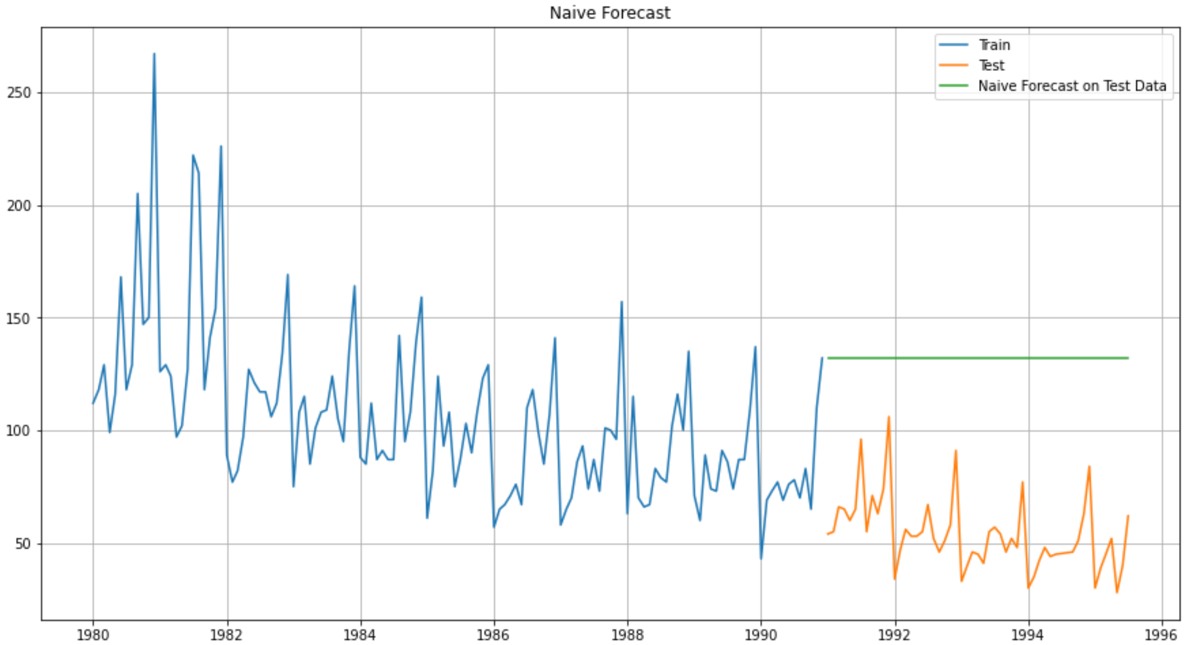



**Fig. 36: Naive Forecast on Test Data for Rose Wine Sales Dataset**

For Naive Model forecast on the Test Data, RMSE is 79.719

## Method 3: Simple Average

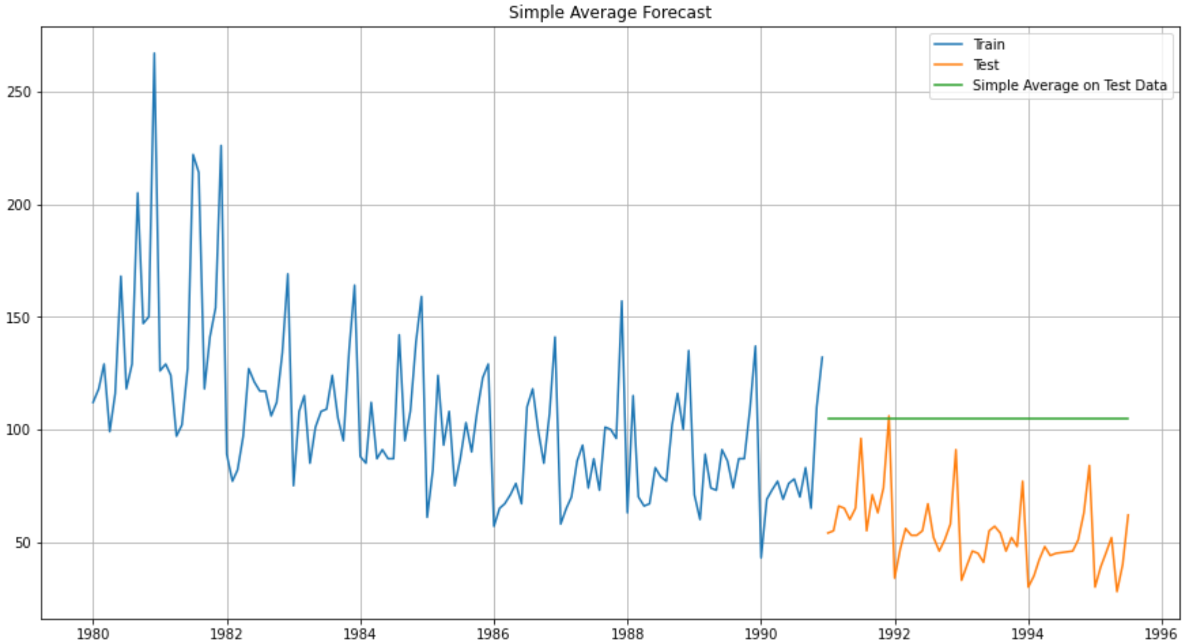



**Fig. 37: Simple Average Model on test data for Rose Wine Sales Dataset**

For Simple Average forecast on the Test Data, RMSE is 53.461

## Method 4: Moving Average(MA)

- For Moving Average, we consider rolling means (or moving averages) for different intervals.
- The best interval can be determined by the maximum accuracy (or the minimum error) over here.

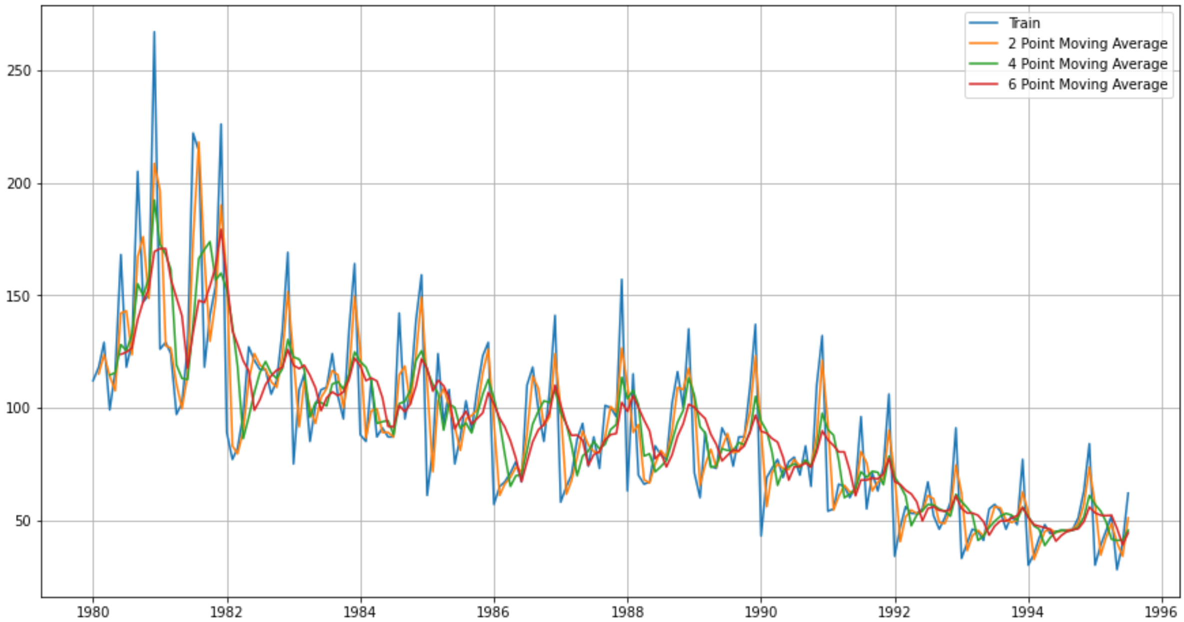



**Fig. 38: Moving Averages at 2,4,6 point for Rose Wine Sales Dataset**

- We Now Split the Data into Train set and Test Set.
- Test set will start from year 1991

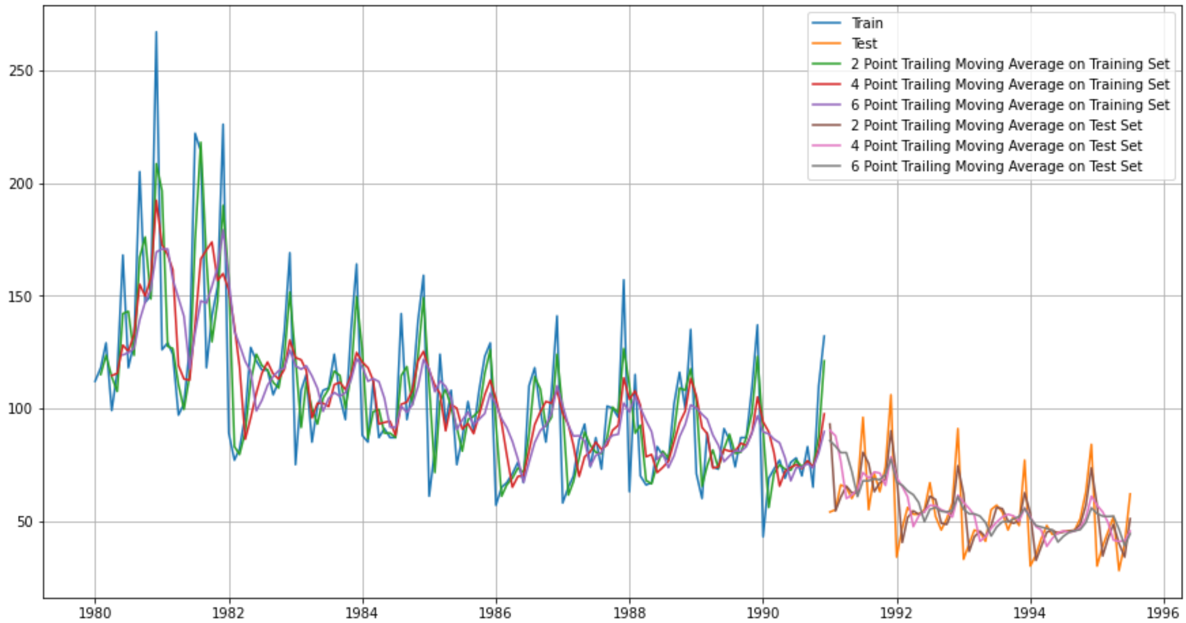



**Fig. 39: Data Split into Train and Test set of Moving Averages at 2,4,6 point for Rose Wine Sales Dataset**

- For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.529
- For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.451
- For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.566

### Test RMSE of all the Models used on Rose Wine Dataset

| | Test RMSE |
|---|---|
| Alpha=0.098,SES | 36.796241 |
| Alpha=1.49,Beta=1.66:DES | 15.268944 |
| Alpha=0.088,Beta=9.006e-06,Gamma=0.0003:TES | 14.255619 |
| Alpha=0.076,Beta=0.041,Gamma=0.0008:TESm | 19.397143 |
| Linear Regression | 15.433446 |
| NaiveModel | 79.718773 |
| SimpleAverageModel | 53.460570 |
| 2pointTrailingMovingAverage | 11.529278 |
| 4pointTrailingMovingAverage | 14.451403 |
| 6pointTrailingMovingAverage | 14.566327 |

## Inference

- We see that Triple Exponential Smoothing method (also called Holt-Winters Method) is the most preferable model as it's RMSE valus is smaller when compared to other Models.
- We will further proceed with SARIMA model.
- We see a seasonality in data hence we will proceed with SARIMA Model instead of ARIMA Model.

## 5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also

## mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05

## Check for stationarity of the whole Time Series Rose Wine Dataset

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:

- $H_0$ : The Time Series has a unit root and is thus non-stationary.
- $H_1$ : 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 $\alpha$ value. $\alpha$ = 0.05.

Using Augmented Dickey-Fuller test on Rose Data set to check Stationary

- DF test statistic is -3.160
- DF test p-value is 0.09272983310684335
- Number of lags used 24

We See that pvalue is > than $\alpha$, Hence Observations are not Stationary.

We Will take Log Transform and check for Stationarity

- DF test statistic is -2.645
- DF test p-value is 0.25968786910763036
- Number of lags used 24

Neither original nor log-transformed series is stationary. Hence, a stationarization is necessary. Often differencing a non-stationary time series leads to a stationary series.

After 1st Order Differencing

- DF test statistic is -6.804
- DF test p-value is 3.894831356782412e-08
- Number of lags used 12

We See that pvalue is < than $\alpha$, Hence Dataset is Stationary after 1st Order Differncing.

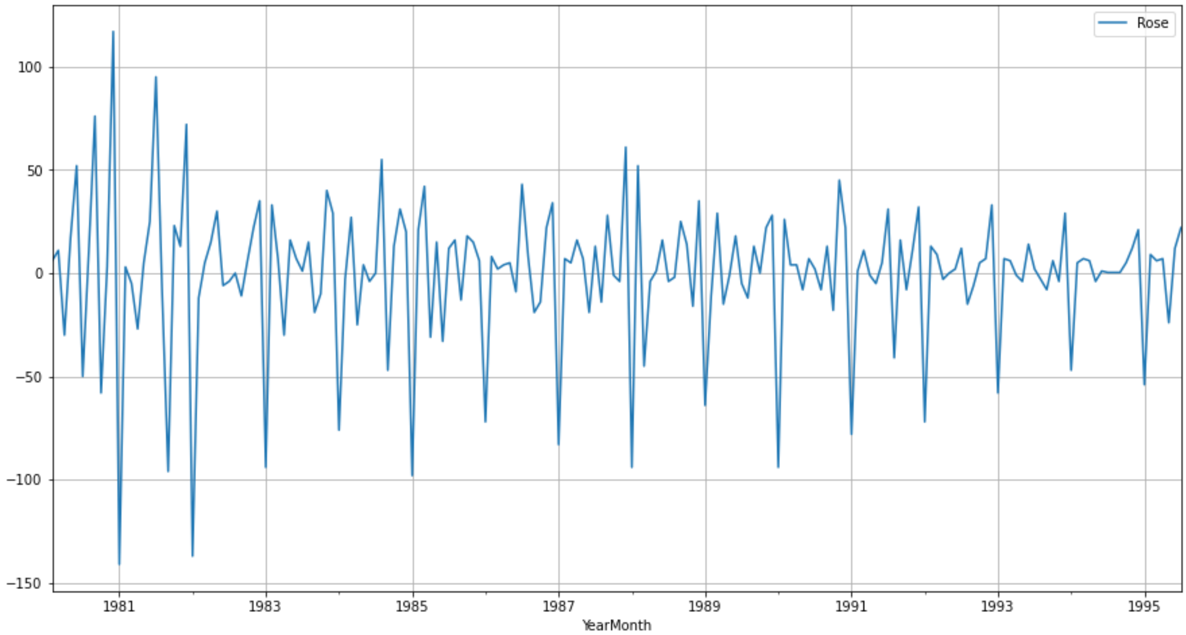
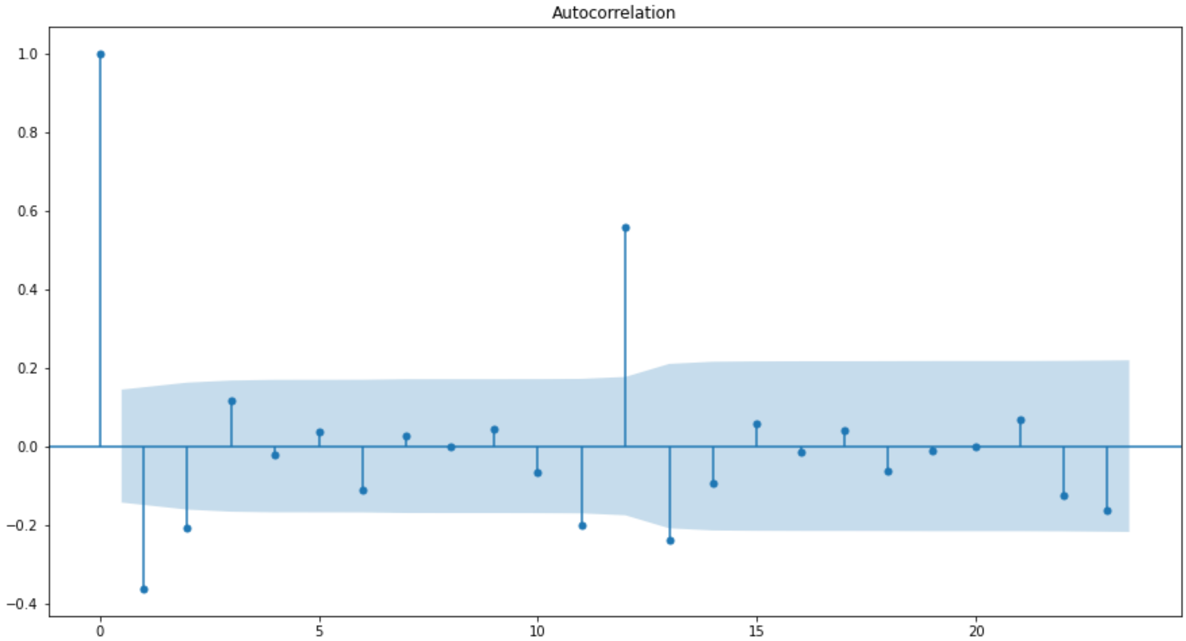



Fig. 40: Rose Wine Dataset Plot after 1st Order Differencing



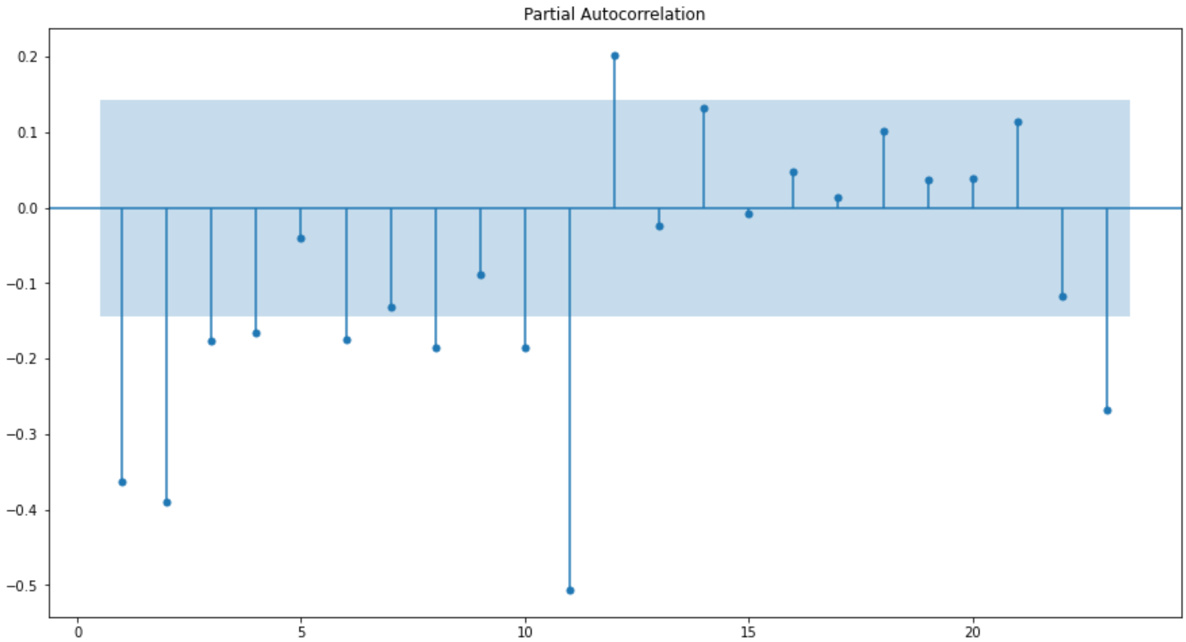



**Fig. 41: Plot of PACF and ACF after 1st Order Differencing for Rose Wine Dataset**

# 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.

## 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.

### SARIMA Model

- We choose SARIMA Model as we see Seasonality and a trend.
- We will Consider p and P to be in the Range of 2 and 5 & q and Q to be either 2 or 3.
- we found data to be stationary at 1st order of Differencing. Hence d = 1
- Hence below table indicates the params that we can use to calculate AIC. Once AIC is calculated, The value with the least AIC score will be considered for ARIMA modeling

| params |
|---|
| Model: (2, 1, 2)(2, 0, 2, 12) |
| Model: (2, 1, 3)(2, 0, 3, 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) |
| Model: (4, 1, 1)(4, 0, 1, 12) |
| Model: (4, 1, 2)(4, 0, 2, 12) |
| Model: (4, 1, 3)(4, 0, 3, 12) |
| Model: (5, 1, 1)(5, 0, 1, 12) |
| Model: (5, 1, 2)(5, 0, 2, 12) |
| Model: (5, 1, 3)(5, 0, 3, 12) |

- After calculating AIC score for each params, the below contains 1st 5 values in ascending order.

| | param | seasonal | AIC |
|---|---|---|---|
| 95 | (4, 1, 2) | (5, 0, 3, 12) | 100.903166 |
| 142 | (5, 1, 3) | (5, 0, 2, 12) | 560.041489 |
| 117 | (5, 1, 1) | (5, 0, 1, 12) | 560.458406 |
| 118 | (5, 1, 1) | (5, 0, 2, 12) | 562.071708 |
| 141 | (5, 1, 3) | (5, 0, 1, 12) | 568.464743 |

- Considering Params (5,1,3) and seasonal Params as (5, 0, 2, 12), we Built a Auto SARIMA Model

**SARIMAX Results**

**SARIMAX Results**

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose | No. Observations: | 132 |
| Model: | SARIMAX(5, 1, 3)x(5, 0, [1, 2], 12) | Log Likelihood | -264.021 |
| Date: | Thu, 16 Dec 2021 | AIC | 560.041 |
| Time: | 13:53:05 | BIC | 595.076 |
| Sample: | 01-01-1980 | HQIC | 573.885 |
| | - 12-01-1990 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P>(z) | 0.025 | 0.975 |
|---|---|---|---|---|---|---|
| ar.L1 | 0.4592 | 0.219 | 2.094 | 0.036 | 0.029 | 0.889 |
| ar.L2 | -0.6799 | 0.188 | -3.620 | 0.000 | -1.048 | -0.312 |
| ar.L3 | -0.1266 | 0.206 | -0.615 | 0.539 | -0.530 | 0.277 |
| ar.L4 | 0.0020 | 0.150 | 0.013 | 0.989 | -0.293 | 0.297 |
| ar.L5 | -0.1851 | 0.152 | -1.217 | 0.223 | -0.483 | 0.113 |
| ma.L1 | -1.5528 | 241.336 | -0.006 | 0.995 | -474.563 | 471.457 |
| ma.L2 | 1.5986 | 881.399 | 0.002 | 0.999 | -1725.911 | 1729.108 |
| ma.L3 | -0.8412 | 570.684 | -0.001 | 0.999 | -1119.361 | 1117.679 |
| ar.S.L12 | 0.1855 | 0.164 | 1.132 | 0.258 | -0.136 | 0.507 |
| ar.S.L24 | 0.3757 | 0.120 | 3.124 | 0.002 | 0.140 | 0.611 |
| ar.S.L36 | 0.0376 | 0.162 | 0.233 | 0.816 | -0.279 | 0.354 |
| ar.S.L48 | 0.0544 | 0.101 | 0.539 | 0.590 | -0.143 | 0.252 |
| ar.S.L60 | 0.2047 | 0.059 | 3.456 | 0.001 | 0.089 | 0.321 |
| ma.S.L12 | 0.2406 | 0.394 | 0.610 | 0.542 | -0.532 | 1.013 |
| ma.S.L24 | -0.3897 | 0.304 | -1.280 | 0.200 | -0.986 | 0.207 |
| sigma2 | 142.9076 | 9.69e+04 | 0.001 | 0.999 | -1.9e+05 | 1.9e+05 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.09 | Jarque-Bera (JB): | 0.04 |
| Prob(Q): | 0.76 | Prob(JB): | 0.98 |
| Heteroskedasticity (H): | 1.79 | Skew: | 0.02 |
| Prob(H) (two-sided): | 0.58 | Kurtosis: | 2.89 |

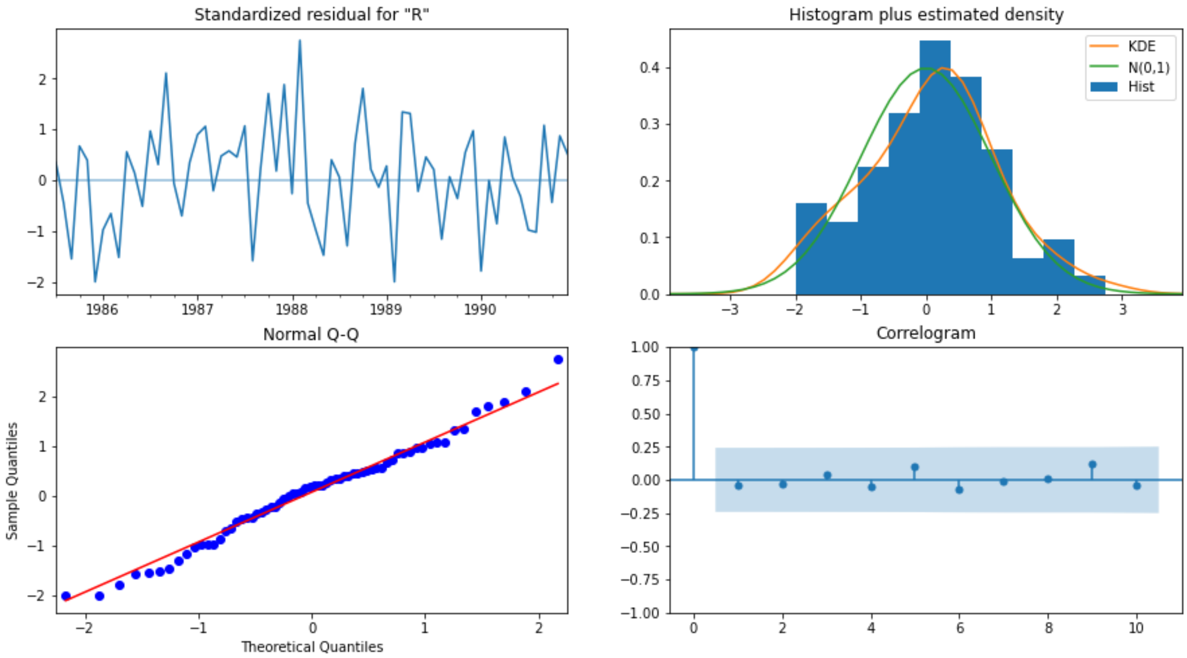



**Fig. 42: Diagnostics Report of SARIMA Model with Order (5, 1, 3) (5, 0, 2, 12) for Rose Wine Dataset**

For Auto SARIMA Model forecast on the Training Data, RMSE is 18.609

## Building a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 12.

- From Fig 18, Observing the Plot of PACF and ACF of 1st Order difference of Train dataset, we can Conclude that P & Q are 3 and 1 Respectively.

- Building SARIMA Model, following observes are observed.

Here, we have taken $\alpha$=0.05.

- We are going to take the seasonal period as 12
- We are taking the p value to be 5 and the q value to be 3 as the parameters.
- Since !st Order Differenencing makes the data stationary, we take d=1

- The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 4.

- The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 2.

Considering Params (5,1,3) and seasonal Params as (4, 0, 2, 12), we Built a Manual SARIMA Model

**SARIMAX Results**

| | | | |
|---|---|---|---|
| Dep. Variable: | Rose | No. Observations: | 132 |
| Model: | SARIMAX(5, 1, 3)x(4, 0, [1, 2], 12) | Log Likelihood | -321.063 |
| Date: | Thu, 16 Dec 2021 | AIC | 672.127 |
| Time: | 14:02:41 | BIC | 707.477 |
| Sample: | 01-01-1980 | HQIC | 686.278 |

**SARIMAX Results**

| | |
|---|---|
| | - 12-01-1990 |
| Covariance Type: | opg |

| | coef | std err | z | P>(z) | 0.025 | 0.975 | |
|---|---|---|---|---|---|---|---|
| ar.L1 | -1.0920 | | 0.162 | -6.742 | 0.000 | -1.409 | -0.775 |
| ar.L2 | -0.9855 | 0.207 | -4.754 | 0.000 | -1.392 | -0.579 | |
| ar.L3 | -0.1857 | 0.294 | -0.632 | 0.527 | -0.762 | 0.390 | |
| ar.L4 | -0.2367 | 0.251 | -0.944 | 0.345 | -0.728 | 0.255 | |
| ar.L5 | -0.1912 | 0.169 | -1.131 | 0.258 | -0.523 | 0.140 | |
| ma.L1 | 0.2160 | 13.485 | 0.016 | 0.987 | -26.215 | 26.647 | |
| ma.L2 | -0.0300 | 11.673 | -0.003 | 0.998 | -22.908 | 22.848 | |
| ma.L3 | -0.9122 | 21.949 | -0.042 | 0.967 | -43.931 | 42.107 | |
| ar.S.L12 | 0.3525 | 0.423 | 0.834 | 0.405 | -0.476 | 1.181 | |
| ar.S.L24 | 0.4346 | 0.347 | 1.253 | 0.210 | -0.245 | 1.114 | |
| ar.S.L36 | 0.0946 | 0.088 | 1.078 | 0.281 | -0.077 | 0.267 | |
| ar.S.L48 | 0.0346 | 0.087 | 0.399 | 0.690 | -0.136 | 0.205 | |
| ma.S.L12 | -0.1780 | 0.512 | -0.348 | 0.728 | -1.181 | 0.825 | |
| ma.S.L24 | -0.3867 | 0.356 | -1.086 | 0.277 | -1.084 | 0.311 | |
| sigma2 | 190.1454 | 4559.932 | 0.042 | 0.967 | -8747.157 | 9127.448 | |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 0.24 | Jarque-Bera (JB): | 14.23 |
| Prob(Q): | 0.63 | Prob(JB): | 0.00 |
| Heteroskedasticity (H): | 0.46 | Skew: | 0.85 |
| Prob(H) (two-sided): | 0.05 | Kurtosis: | 4.23 |

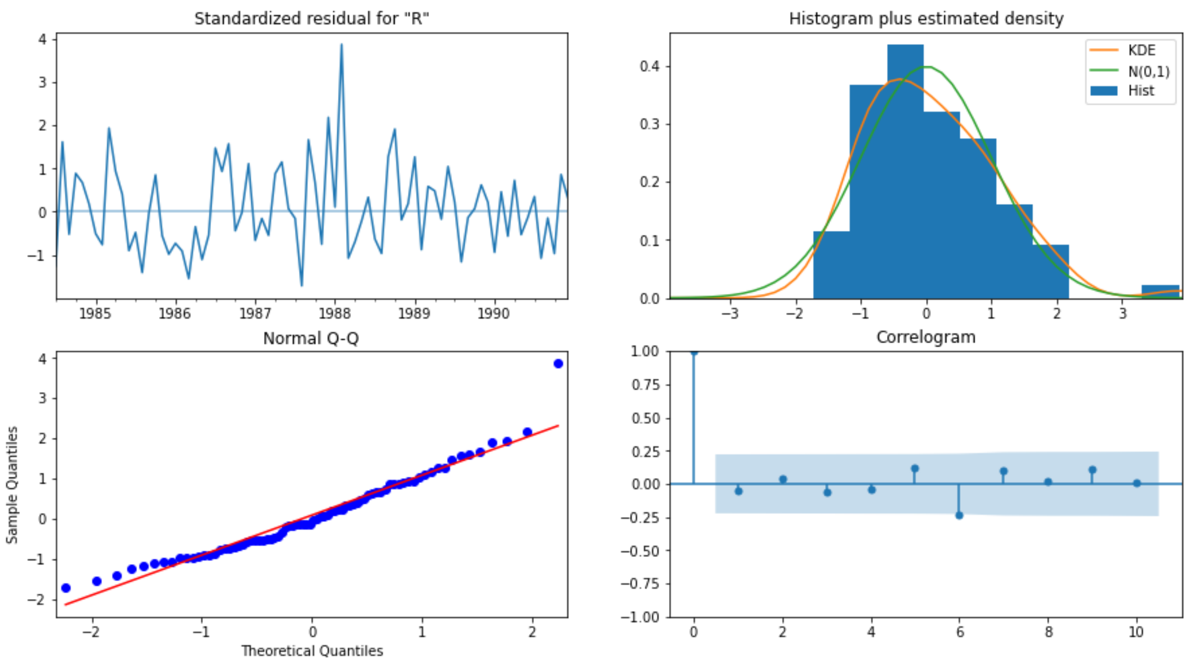


**Fig. 43: Diagnostics Report of SARIMA Model with Order (5,1,3)(4, 0, 2, 12) for Sparkling Wine Dataset**

For Manual SARIMA Model forecast on the Training Data, RMSE is 19.015

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

Now we Combine all the RMSE values of various Models. Below is the Table containing all the values

| | Test RMSE |
|---|---|
| Alpha=0.098,SES | 36.796241 |
| Alpha=1.49,Beta=1.66:DES | 15.268944 |
| Alpha=0.088,Beta=9.006e-06,Gamma=0.0003:TES | 14.255619 |
| Alpha=0.076,Beta=0.041,Gamma=0.0008:TESm | 19.397143 |
| Linear Regression | 15.433446 |
| NaiveModel | 79.718773 |
| SimpleAverageModel | 53.460570 |
| 2pointTrailingMovingAverage | 11.529278 |
| 4pointTrailingMovingAverage | 14.451403 |
| 6pointTrailingMovingAverage | 14.566327 |
| AutoSARIMA(5, 1, 3)(5, 0, 2, 12) | 18.608984 |
| AutoSARIMA(5, 1, 3)(4, 0, 2, 12) | 19.014706 |

### Inference

- Most optimum Model is 2pointTrailingMovingAverage because the RMSE of Test Data is Less compared to all other Models.
- We Can use the 2 Point Moving Average Model and forecast for next 12 months and results will be fairly accuarate.

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

## Forecasting for the next 12 months using Holt-Winters method.

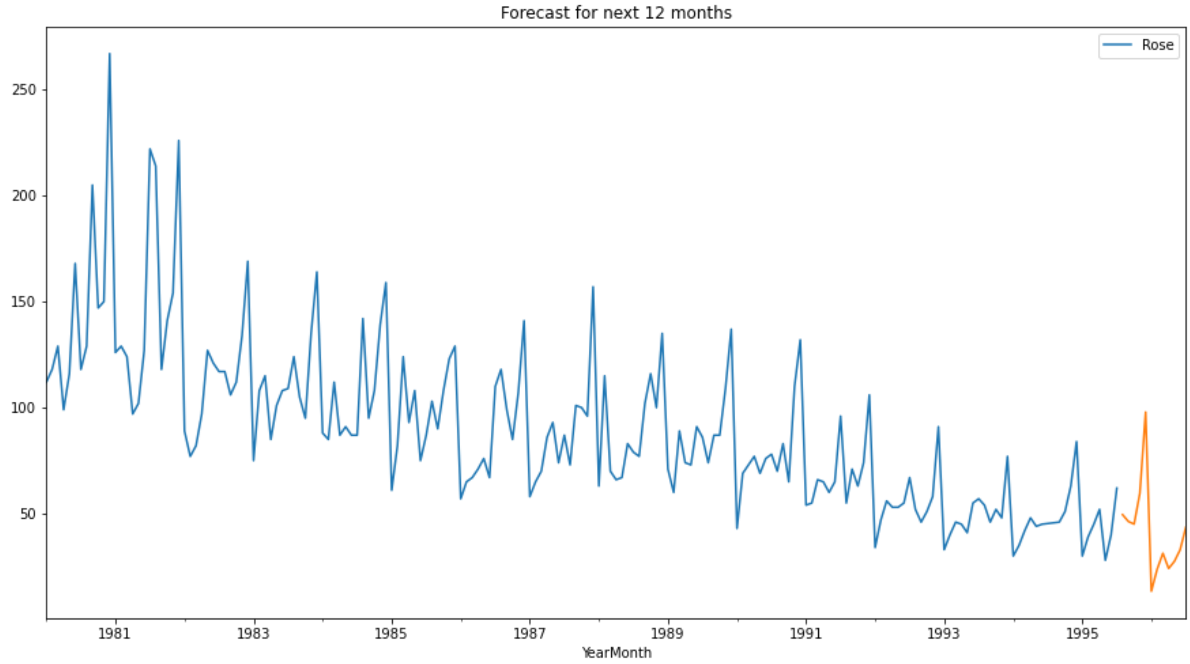



**Fig. 44: Forecast for the next 12 months for Rose Wine Sales**

RMSE: 17.664104202163752

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

- We Found that Holt-Winters Method (TES) produces the lowest RMSE values when compared to other models
- Using TES, we have forecasted for next 12 months. The Respective graphs are shown above
- RMSE Values are quit close to the training model values, and fairly accurate which tells us that model is a good model.
- For Sparkling Wine, Sales per year doesn't degrade hence manufacturing This type of Wine will not bring losses in the near foreseeable future.
- We see sudden rise in sales in the month of November and December. As both months are holiday season, Production of Sparkling Wine should be fairly high when compared to other months.
- Company needs to make sure that this particular Wine is made available at major sector of retails stores in order for the Consumer to readily purchase. Thus increasing in sales and profit.
- For Rose Wine, We see a downward trend in the Market, and not a preferred Choice of Wine.
- Company should either promote this brand of Wine and provide offers to Consumers. Company do this for next few years and see if any change in sales. If the trend remains the same, then it would be effective to stop the production of sure wine considering the production cost and concentrate on other products.

**End Of Report**