

2022

Time Series Forecasting Business Report



Yogender Singh

Great Learning

10/9/2022

Contents

[Q 1] Read the data as an appropriate Time Series data and plot the data.....	2
[Q 2] Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.....	3
• $y_t = \text{Trend} * \text{Seasonality} * \text{Residual}$	6
[Q 3] Split the data into training and test. The test data should start in 1991.....	6
[Q 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.	8
[Q 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.	19
If the data is found to be non-stationary, take appropriate steps to make it stationary.	19
Check the new data for stationarity and comment. Note: Stationarity should be checked at $\alpha = 0.05$	19
[Q 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	21
[Q 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.....	25
[Q 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.	28
[Q 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.	30
[Q 10] Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.	31
Rose Wine Sales - Comments :	31
Rose Wine Sales - Forecast Models :	32
Rose Wine Sales - Suggestions :	32
.....	32

INTRODUCTION

This report consists of Time Series analysis and forecasting of 1 dataset -

- DATASET 1 - Sales data of Rose Wine

Please find the Jupyter Code Notebook. Analysis code is in Python.

PROBLEM – Rose Wine Sales

For this particular assignment, the data of different types of wine sales in the 20th century is to be analyzed.

Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyze and forecast Wine Sales in the 20th century.

Datasets used - Sales of Rose Wine

SYNOPSIS

1. Total No. Of Rose Data Entries = 187
 No. Of Missing Values in Rose data = 2
 No. Of Duplicate entries in Rose data = 0
2. Both datasets are split in Train : Test at year 1991 - Test data starts at 1991

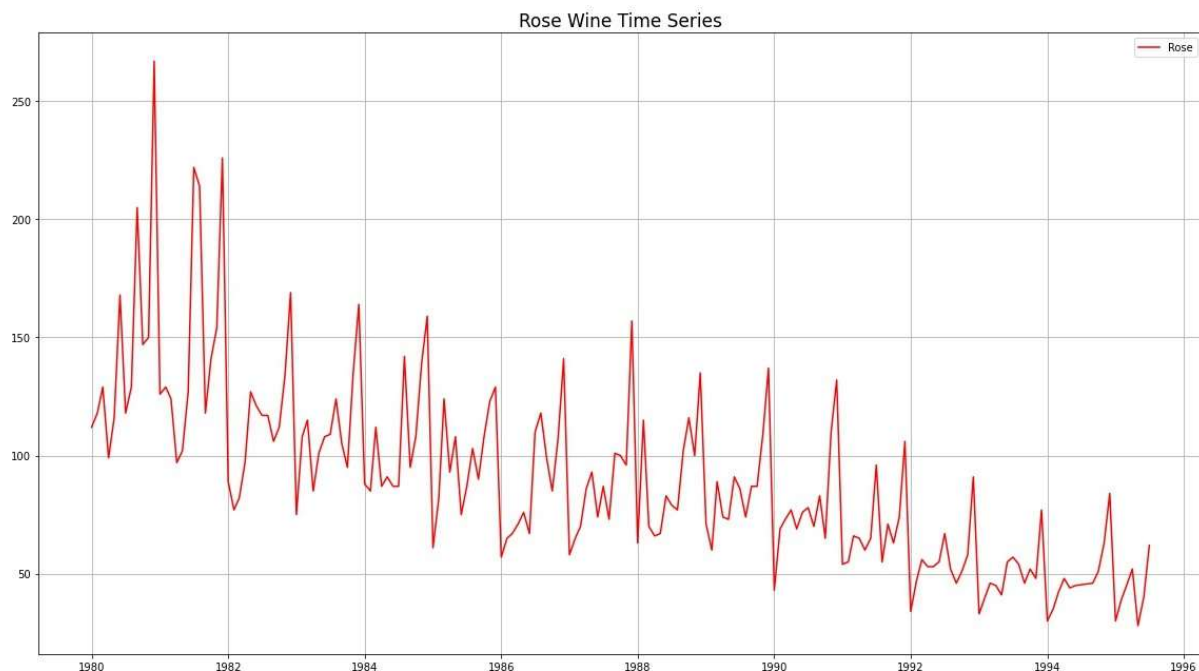
[Q 1] Read the data as an appropriate Time Series data and plot the data.

- Rose Dataset are read and stored as Pandas Data Frames for analysis
- Datasets are read as Time Series data using `parse_dates=True& index_col='YearMonth'`
- First 5 rows of Rose data are given below –

	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0

1980-03-01 129.0
 1980-04-01 99.0
 1980-05-01 116.0

Rose Data plot -



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

◆ Exploratory Data Analysis -

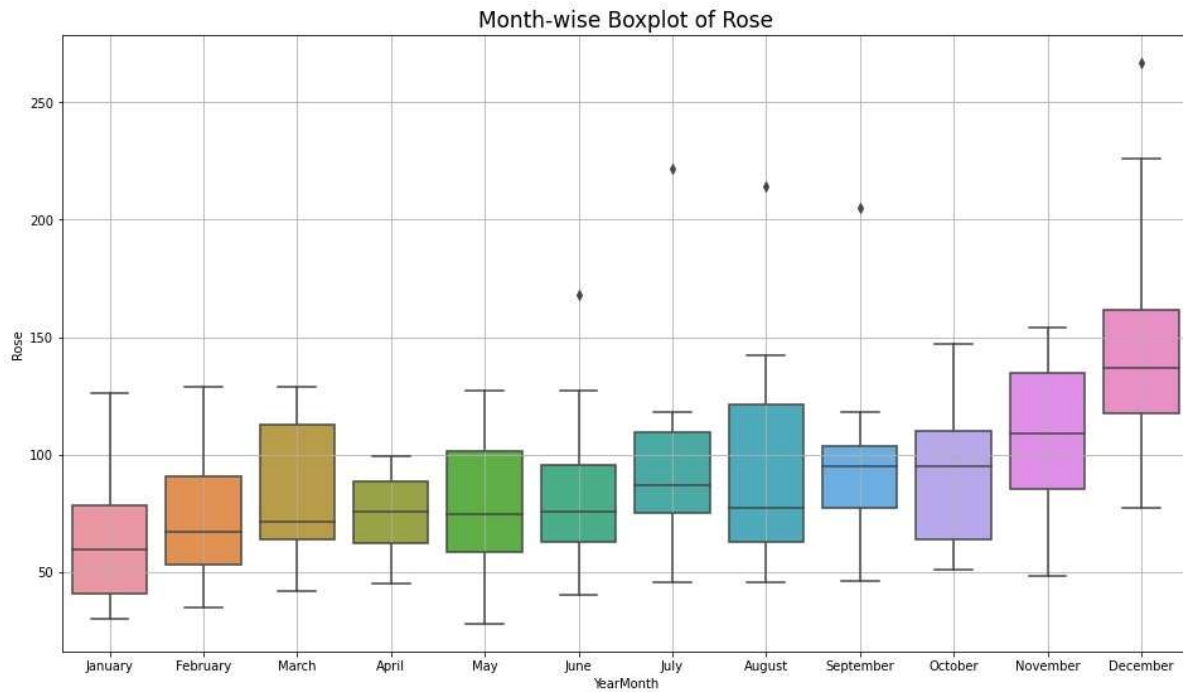
	count	mean	std	min	25%	50%	75%	max
Rose	187.0	89.914	39.238	28.0	62.5	85.0	111.0	267.0

Descriptive Stats of Rose datasets

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

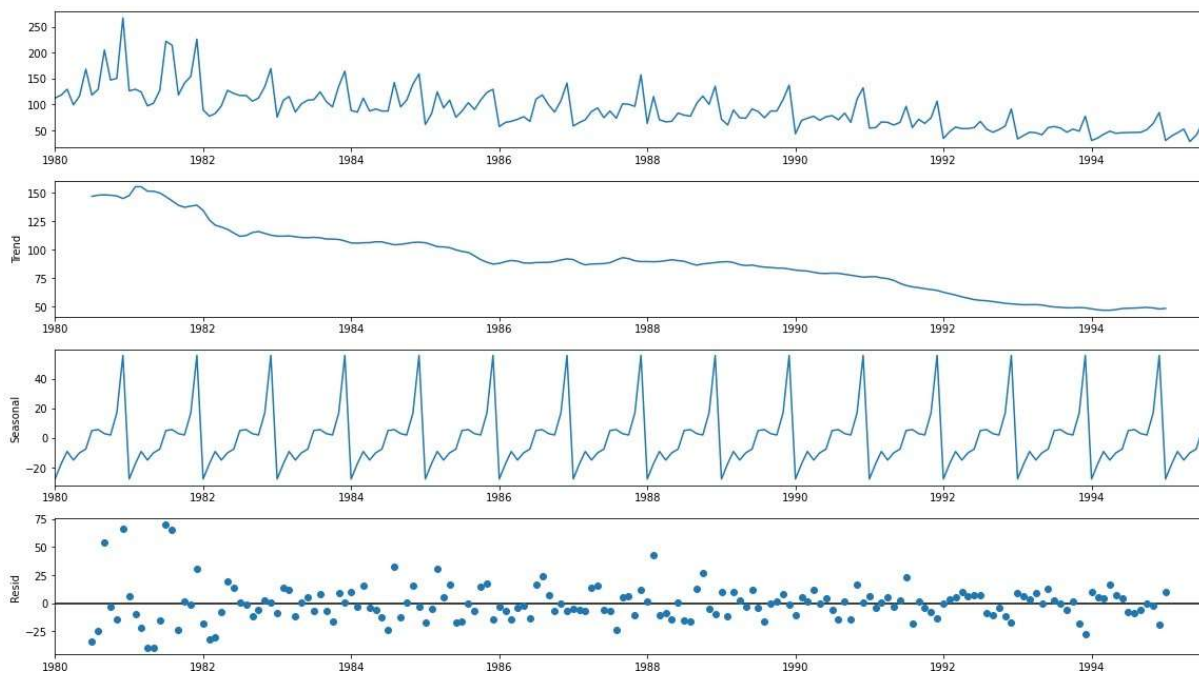
Info - Rose data

- Month-wise Boxplot of Rose -



- Sales of - Rose, show a spike in the last quarter of Oct to Dec
- Spike is much more accentuated in Sparkling sales
- This spike may be due to the Holiday season starting in Oct

- ◆ Additive Decomposition of Rose -



YearMonth	trend
1980-01-01	
1980-02-01	
1980-03-01	
1980-04-01	
1980-05-01	
1980-06-01	
1980-07-01	147.08
1980-08-01	148.13
1980-09-01	148.38
1980-10-01	148.08
1980-11-01	147.42
1980-12-01	145.13

Rose Trend

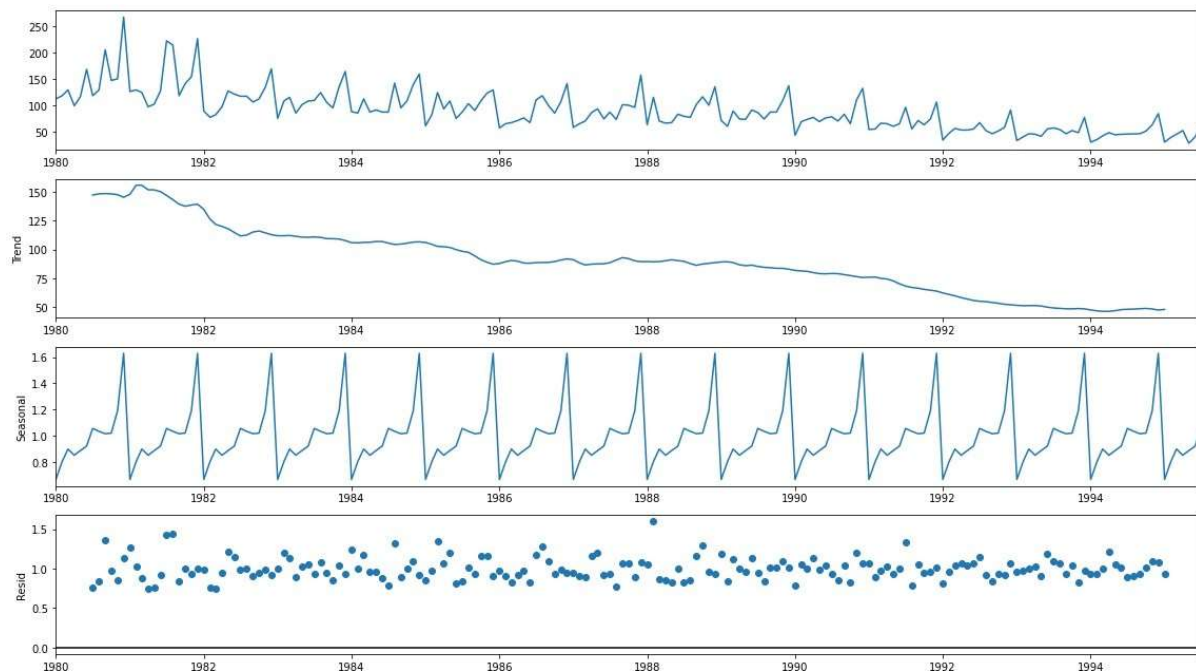
YearMonth	seasonal
1980-01-01	0.670112
1980-02-01	0.806164
1980-03-01	0.901166
1980-04-01	0.854026
1980-05-01	0.889417
1980-06-01	0.923987
1980-07-01	1.058029
1980-08-01	1.035877
1980-09-01	1.017649
1980-10-01	1.022575
1980-11-01	1.192350
1980-12-01	1.628648

Rose Seasonality

YearMonth	resid
1980-01-01	NaN
1980-02-01	NaN
1980-03-01	NaN
1980-04-01	NaN
1980-05-01	NaN
1980-06-01	NaN
1980-07-01	NaN
1980-08-01	0.758
1980-09-01	0.841
1980-10-01	1.358
1980-11-01	0.971
1980-12-01	0.853

Rose Residual

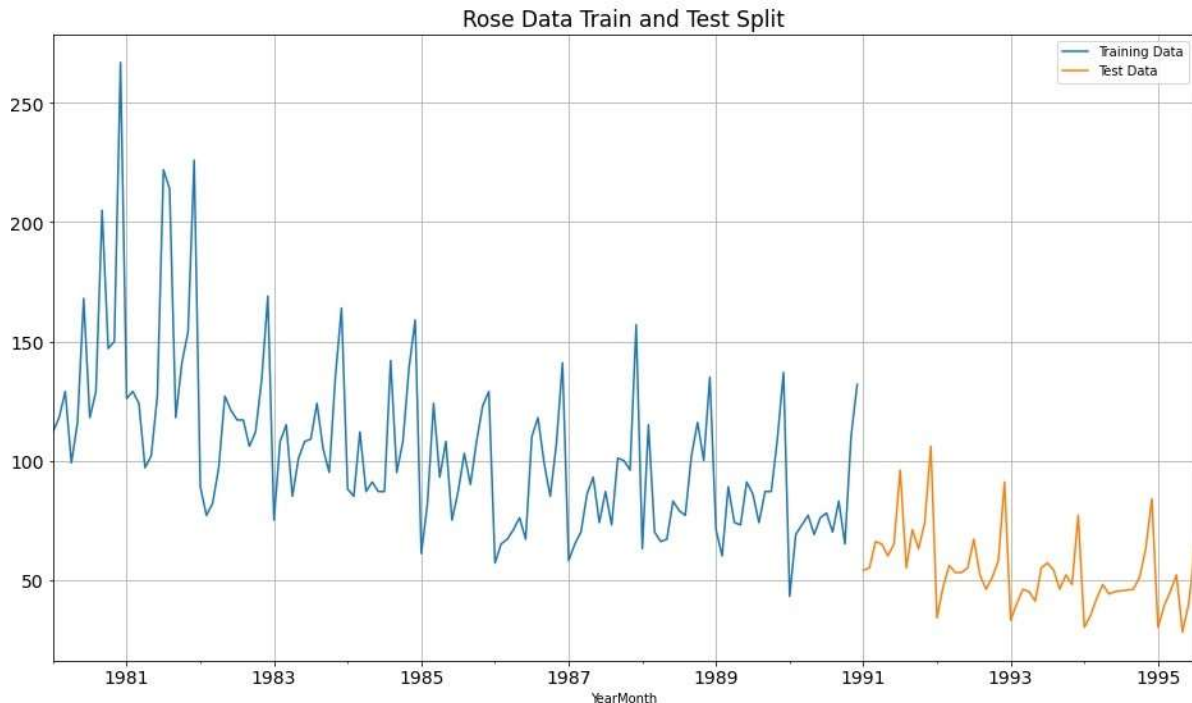
◆ Multiplicative Decomposition of Rose -



- Additive Models -
 - The seasonality is relatively constant over time
 - $y_t = \text{Trend} + \text{Seasonality} + \text{Residual}$
- Multiplicative Models -
 - The seasonality increases or decreases over time. It is proportionate to the trend
 - $y_t = \text{Trend} * \text{Seasonality} * \text{Residual}$
- Here by just observing the Residual patterns of Additive and Multiplicative models of Rose datasets. It seems that -
 - Rose is Multiplicative

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

- Both datasets of Rose are split at the year 1991
- Test datasets start at 1991



-
-
-

- Rose dataset - TRAIN

YearMonth	Rose
1980-01-01	112.00
1980-02-01	118.00
1980-03-01	129.00
1980-04-01	99.00
1980-05-01	116.00

Rose Train - First5 rows

YearMonth	Rose
1990-08-01	70.00
1990-09-01	83.00
1990-10-01	65.00
1990-11-01	110.00
1990-12-01	132.00

Rose Train - Last 5 rows

- Rose dataset - TEST

YearMonth	Rose
1991-01-01	54.00
1991-02-01	55.00
1991-03-01	66.00
1991-04-01	65.00
1991-05-01	60.00

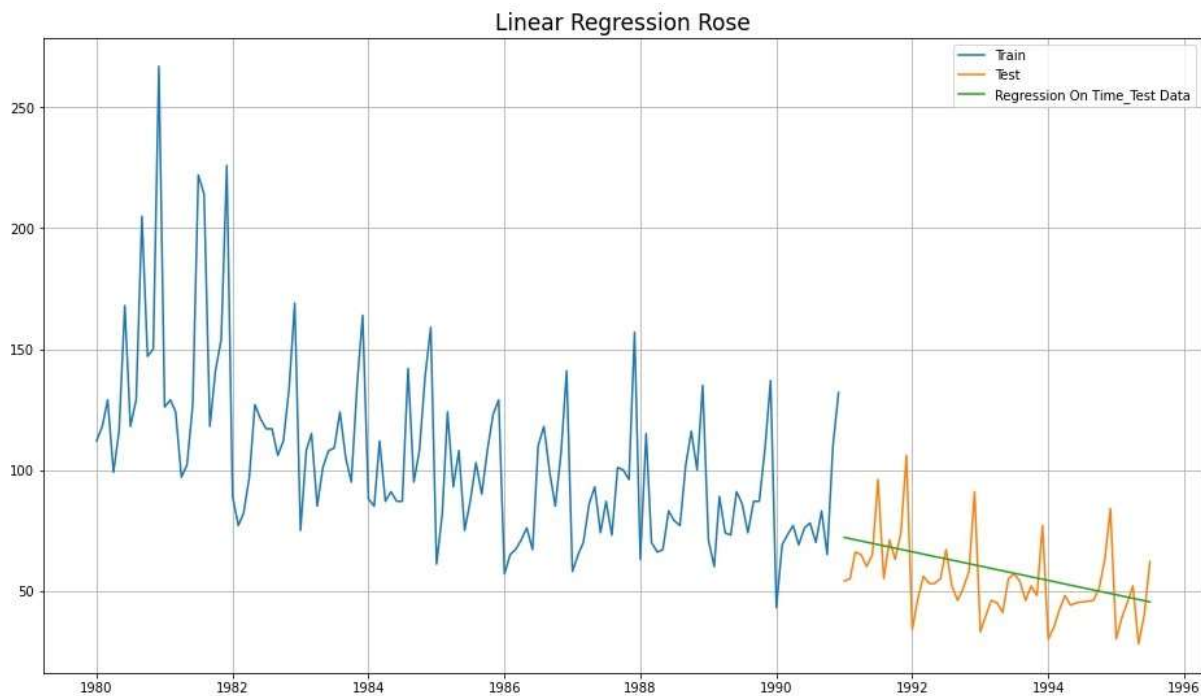
Rose Test - First5 rows

YearMonth	Rose
1995-03-01	45.00
1995-04-01	52.00
1995-05-01	28.00
1995-06-01	40.00
1995-07-01	62.00

Rose Test - Last 5 rows

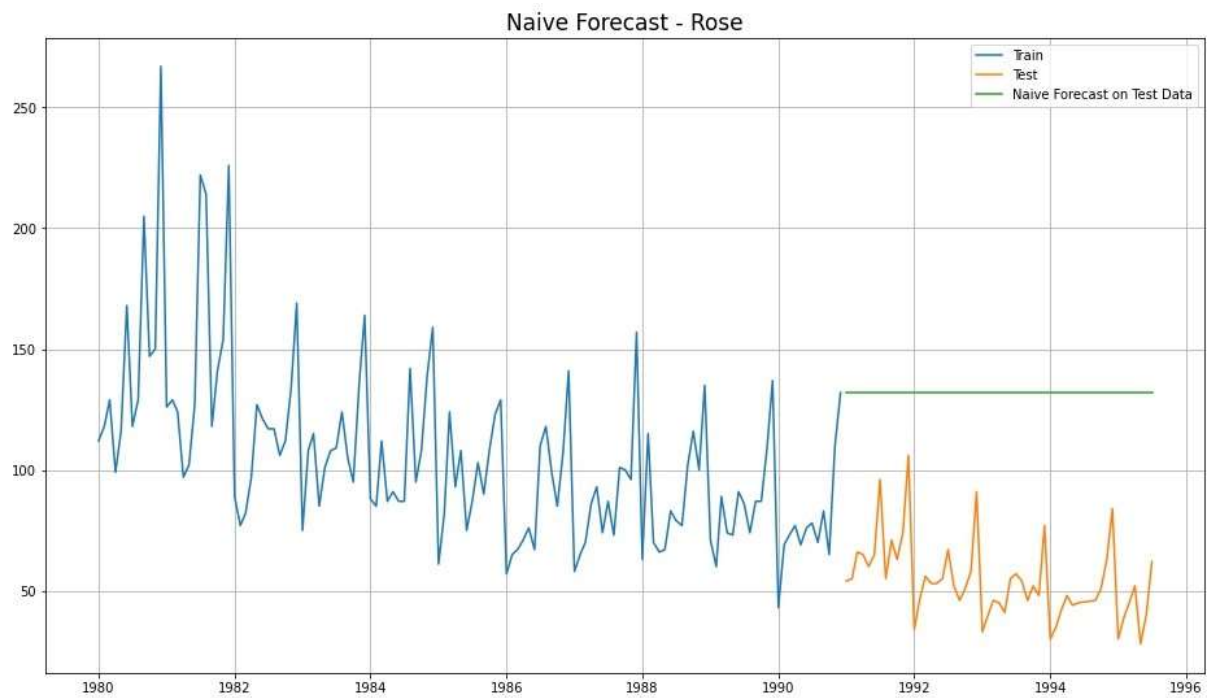
[Q4] Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

✦ **Model 1 - Linear Regression**



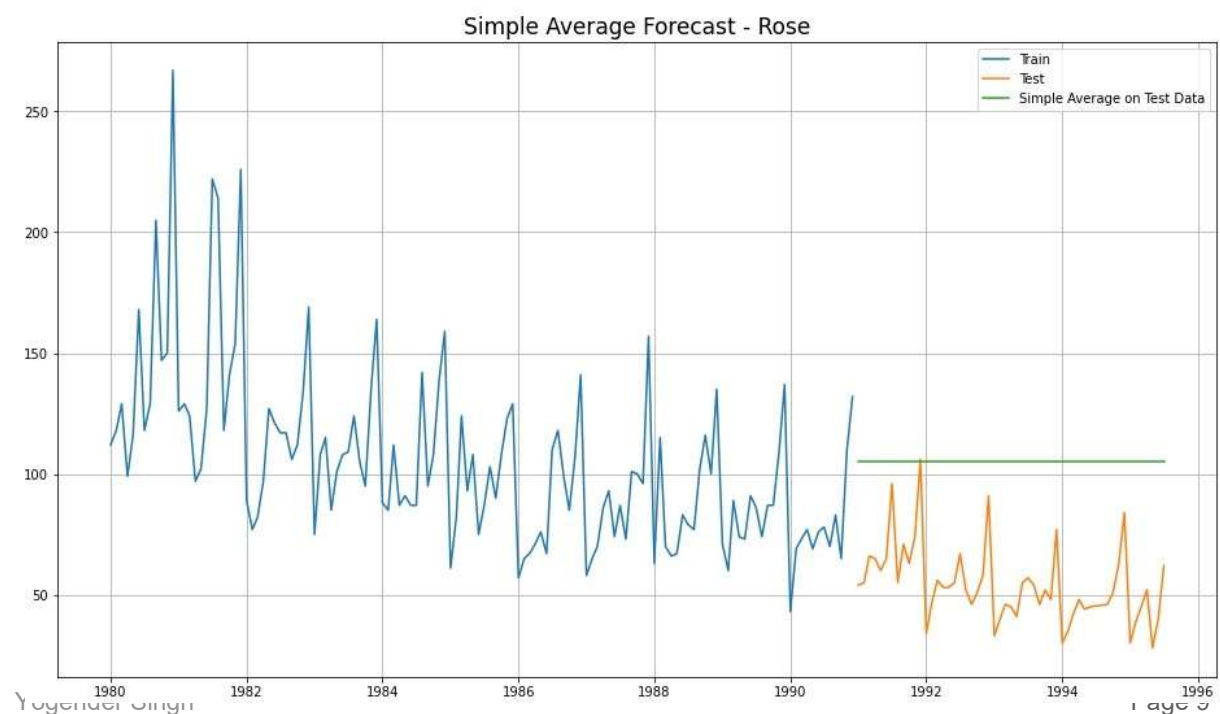
	Test RMSE Rose
RegressionOnTime	15.27

✦ **Model 2 - Naive Bayes**



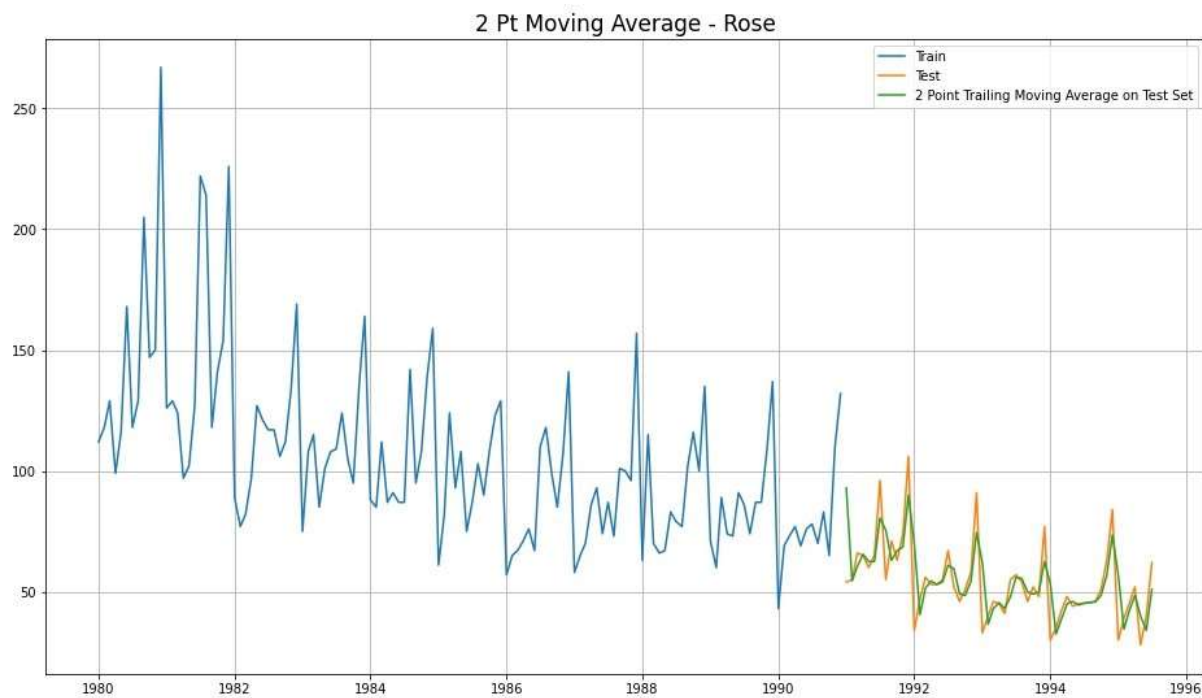
	Test RMSE Rose
RegressionOnTime	15.27
NaiveModel	79.72

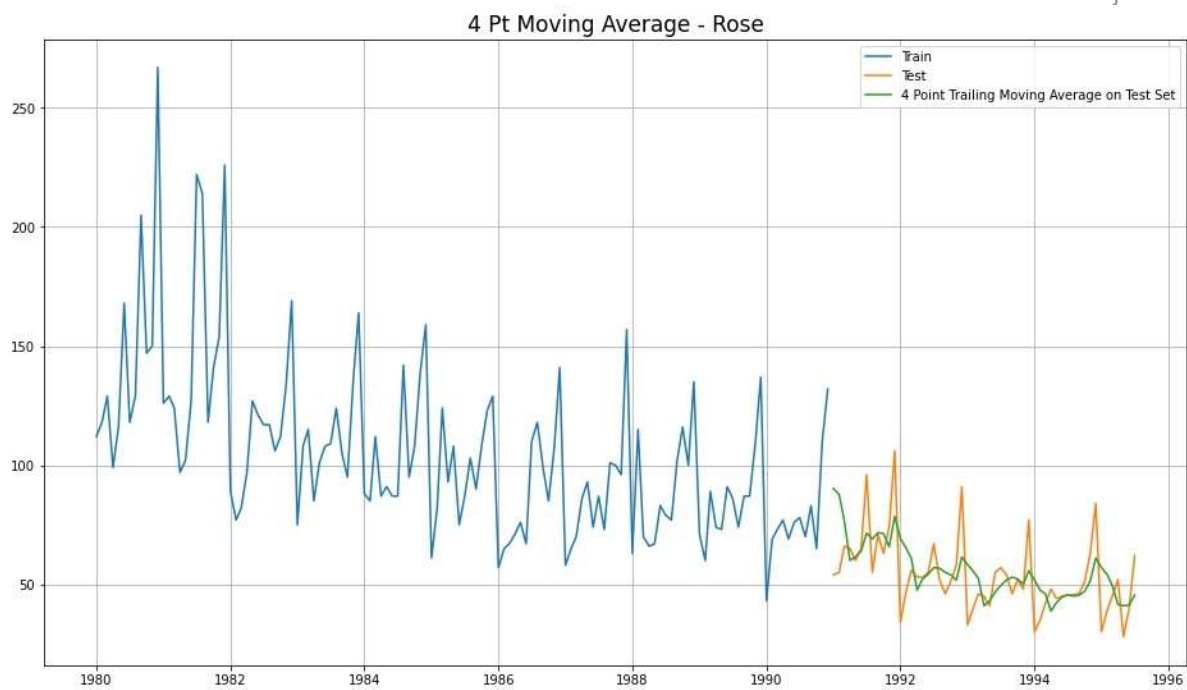
◆ Model 3 - Simple Average

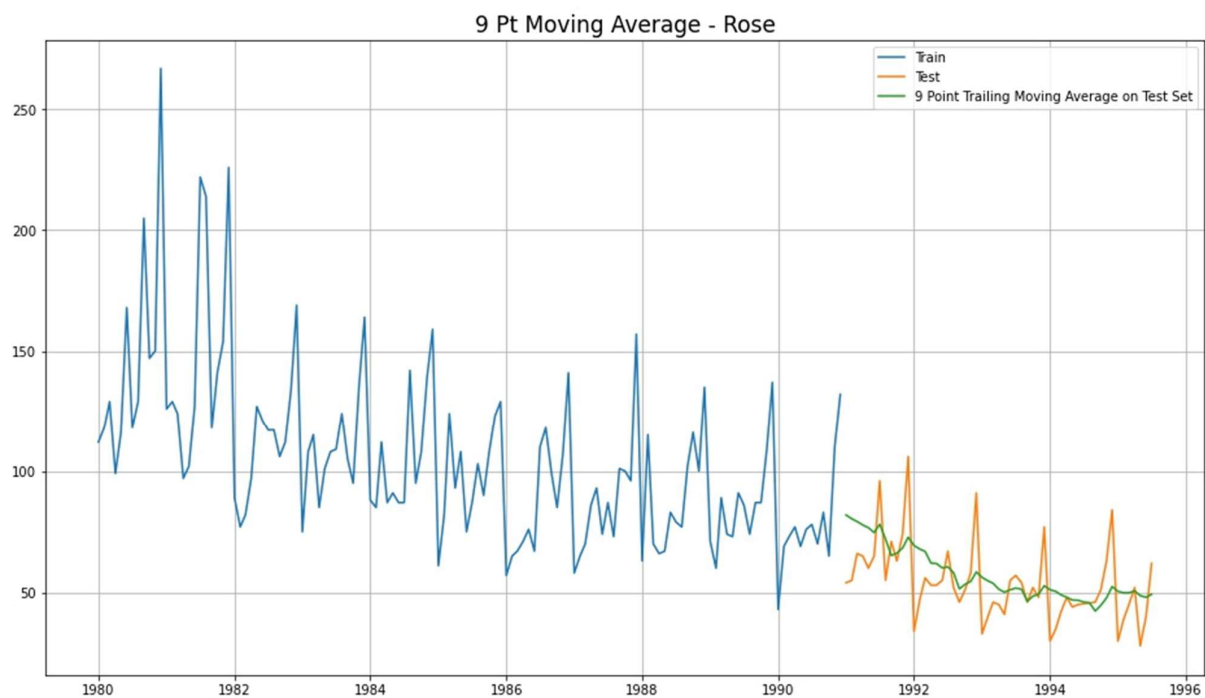
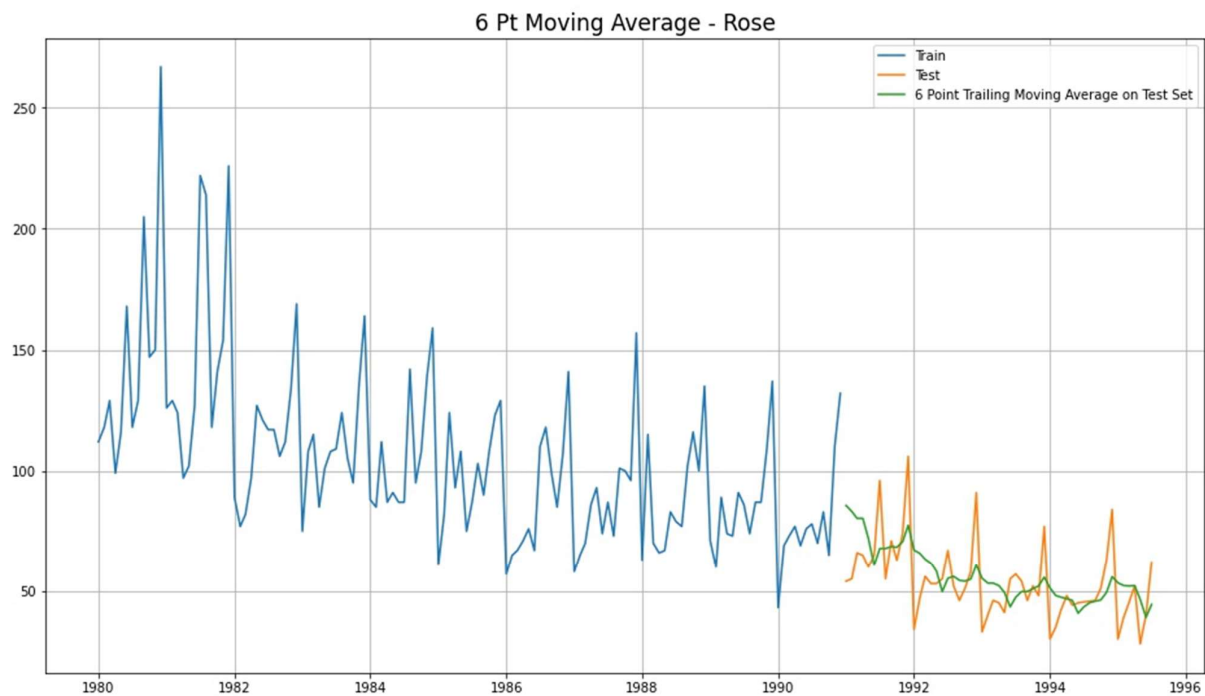


	Test RMSE Rose
RegressionOnTime	15.27
NaiveModel	79.72
SimpleAverageModel	53.46

◆ **Model 4.A - Moving Average (Rose)**

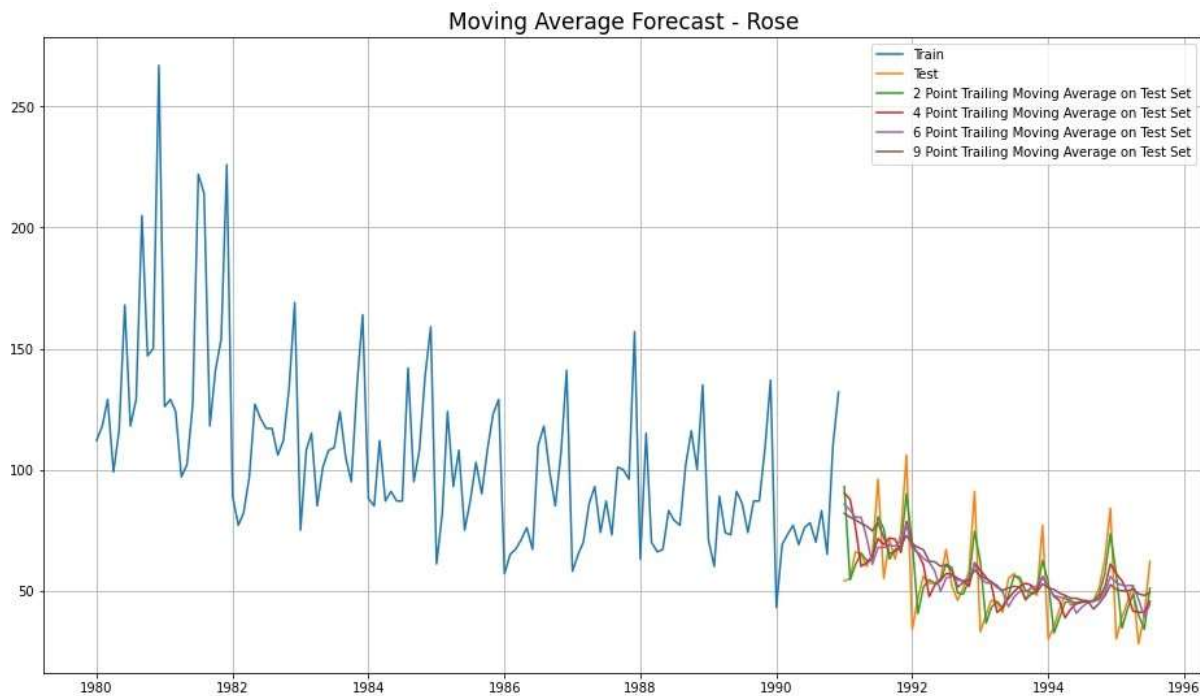






	Test RMSE Rose
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73

◆ Consolidated Moving Average Forecasts (Rose)



◆ NOTE -

- We have built 4 models till now for both Rose and Sparkling Wine datasets
- We fitted various models to the Train split and Tested it on Test split. Accuracy metrics used is Root Mean Squared Error (RMSE) on Test data
- Model 1 - Linear Regression ($y_t = \theta_0 + \theta_1 X_t + \epsilon_t$)
 - We regressed variables 'Rose' and 'Sparkling' against their individual time instances
 - We modified the datasets and tagged individual sales to their time instances
 - TEST RMSE ROSE = 15.27
- Model 2 - Naive Approach ($\hat{y}_{t+1} = y_t$)
 - Naive approach says that prediction for tomorrow is same as today
 - And, prediction for day-after is same as tomorrow
 - So, effectively all future predictions are going to be same as today
 - TEST RMSE ROSE = 79.72

- Model 3 - Simple Average ($\hat{y}_{t+1} = \hat{y}_{t+2} = \dots = \hat{y}_{t+n} = \text{Mean}(y_1, y_2, \dots, y_t)$)
 - All future predictions are the same as the simple average of all data till today
 - TEST RMSE ROSE = 53.46
- Model 4 - Moving Average (MA)
 - We calculate rolling means (Moving averages) over different intervals for the whole train data
 - 2 Pt MA =====> means, we find average of 1st and 2nd to predict 3rd
similarly, average of 2nd and 3rd to predict 4th and so on
 - 4 Pt MA =====> means, we find average of 1st, 2nd, 3rd & 4th to predict 5th
also, average of 2nd, 3rd, 4th & 5th to predict 6th and so on
 - 2 PT MA =====>

TEST RMSE ROSE = 11.53
 - 4 PT MA =====>

TEST RMSE ROSE = 14.45
 - 6 PT MA =====>

TEST RMSE ROSE = 14.57
 - 9 PT MA =====>

TEST RMSE ROSE = 14.73

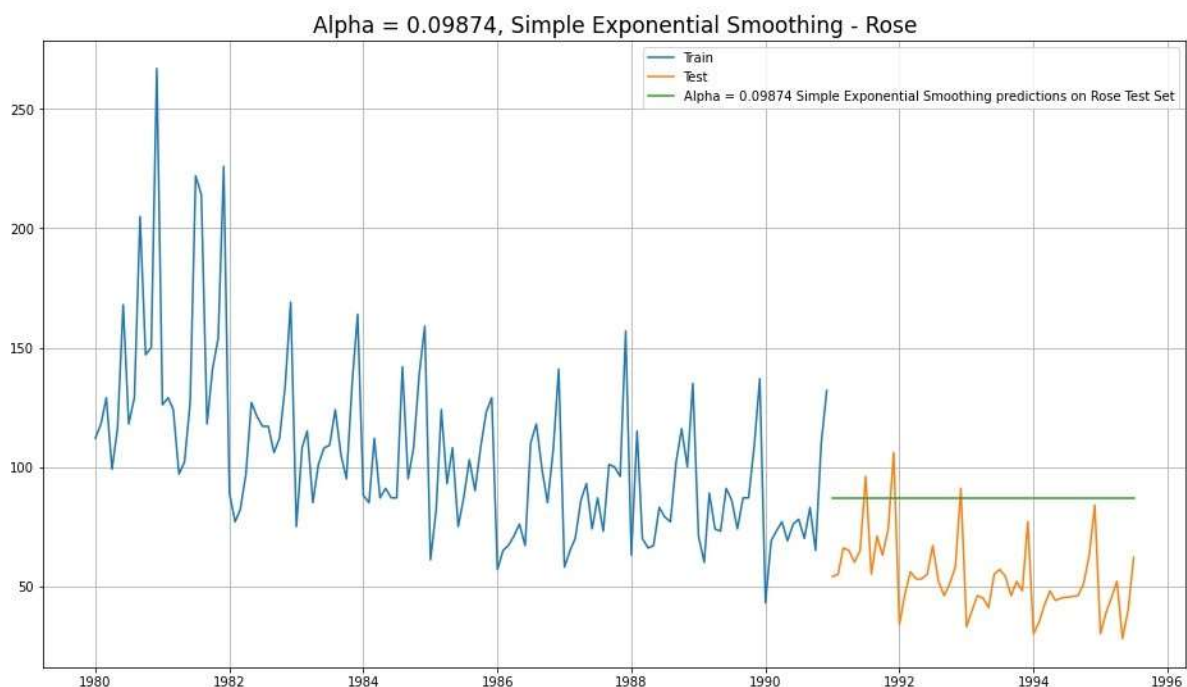
	Test RMSE Rose
RegressionOnTime	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73

Consolidated Scores of Regression, Naive, Simple Average & Moving Average

- **Till now, Best Model which gives lowest RMSE score for Rose is ———>**
2 Pt Moving Average Model

- We'll continue to forecast using Exponential Smoothing Models for datasets of Rose Wine Sales
- Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations
- Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality)
- One or more parameters control how fast the weights decay. The values of the parameters lie between 0 and 1.

◆ **Single Exponential Smoothing with Errors -**

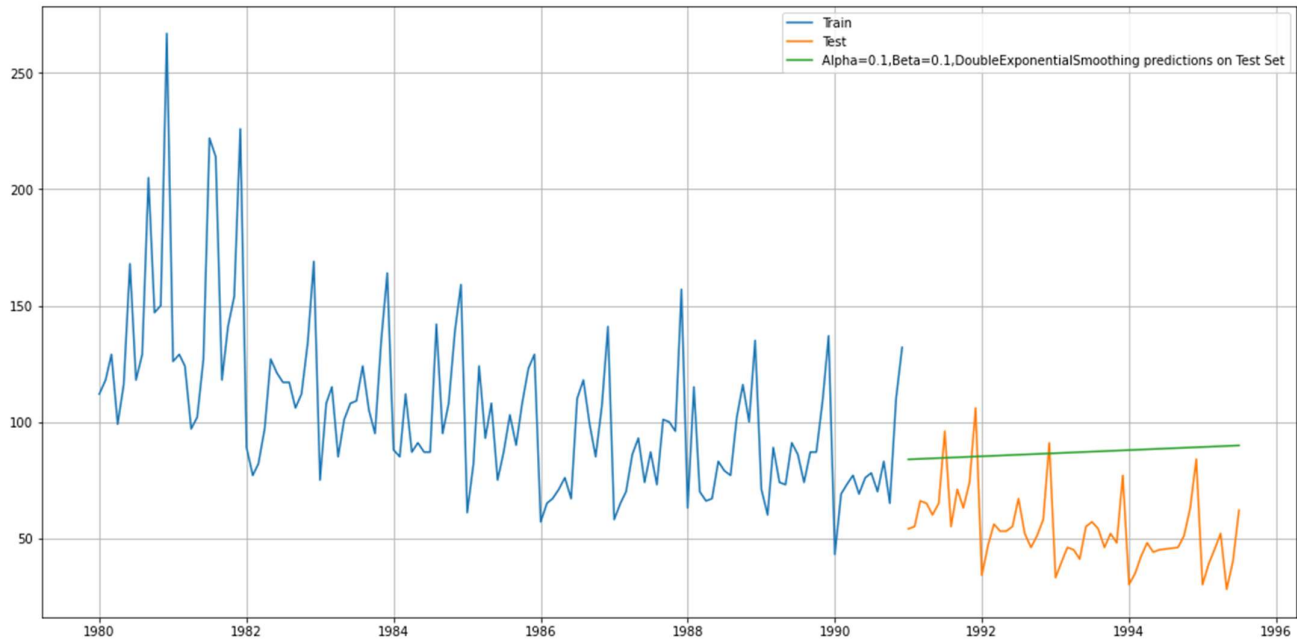


- For Rose - Level Parameter, Alpha = 0.09874

	Test RMSE Rose
RegressionOnTime	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
Simple Exponential Smoothing	36.80

- Best Model till now for Rose—— > 2 Pt Moving Average Model

◆ Double Exponential Smoothing

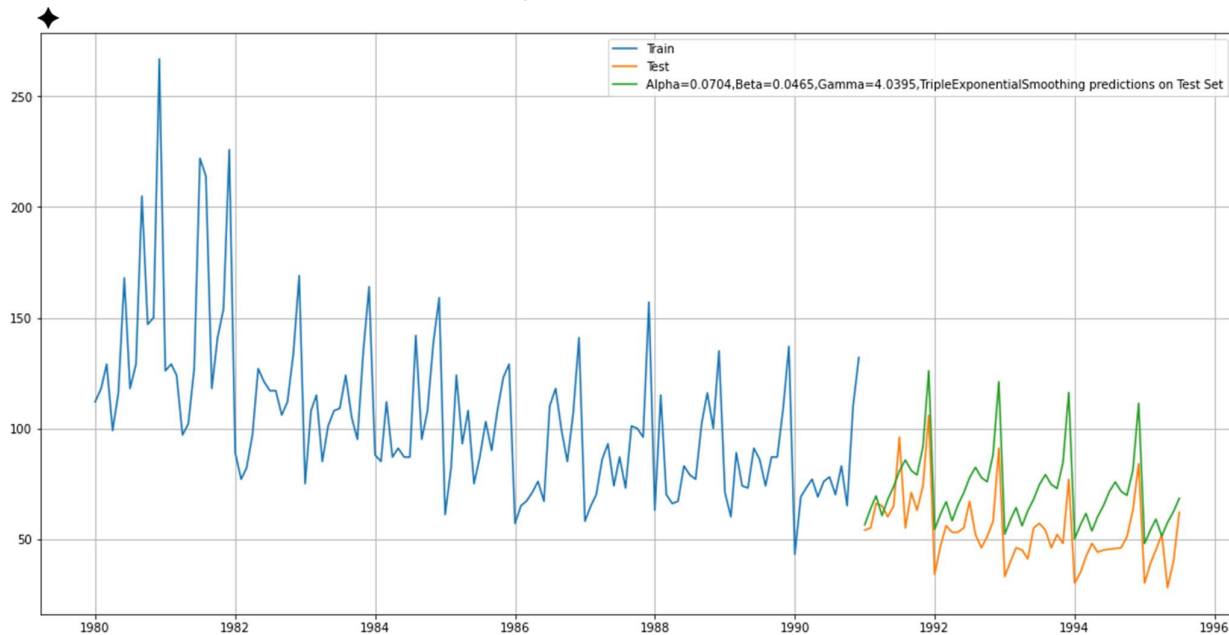


- In Rose - DES has picked up the trend well. DES seems to perform equal to SES here.
- Rose Level parameter, $\text{Alpha} = 0.1$
- Trend parameter, $\text{Beta} = 0.1$

	Test RMSE Rose
RegressionOnTime	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
Simple Exponential Smoothing	36.80
Double Exponential Smoothing	36.82

- Best Model till now for Rose— > 2 Pt Moving Average Model

Triple Exponential Smoothing

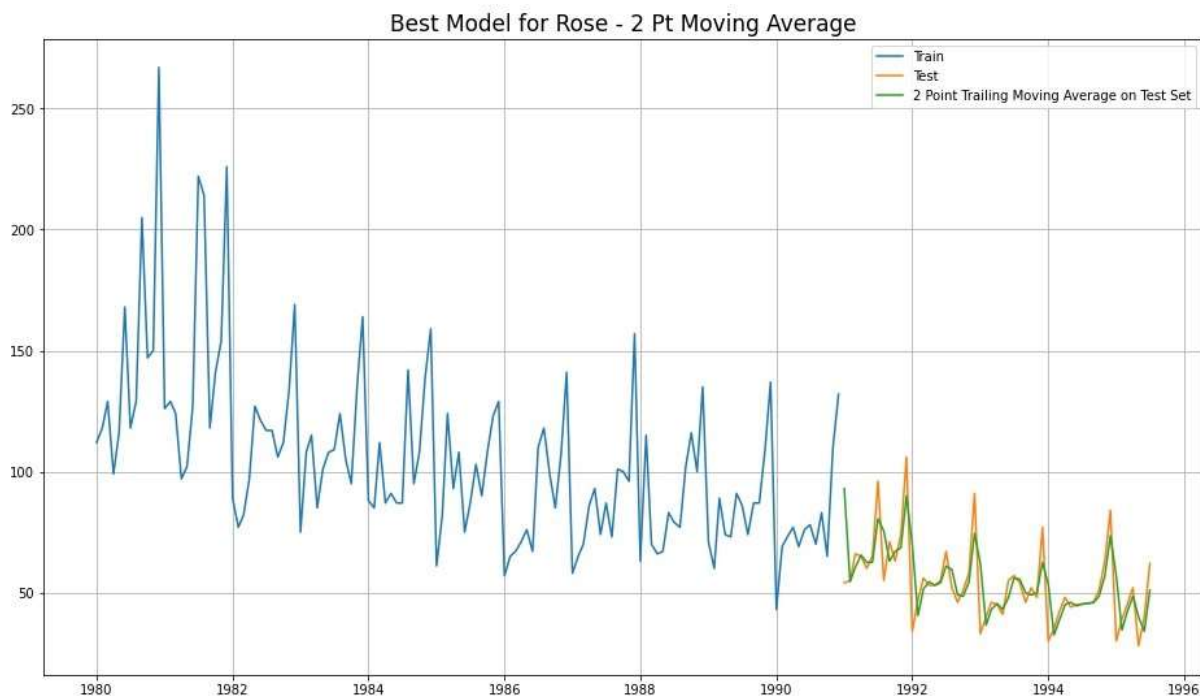


- In Rose - TES has picked up the trend and seasonality very well
- Rose - Level parameter, $\alpha = 0.0704$
Trend parameter, $\beta = 0.0465$
Seasonality parameter, $\gamma = 0.0395$

	Test RMSE Rose
RegressionOnTime	15.27
NaiveModel	79.72
SimpleAverageModel	53.46
2pointTrailingMovingAverage	11.53
4pointTrailingMovingAverage	14.45
6pointTrailingMovingAverage	14.57
9pointTrailingMovingAverage	14.73
Simple Exponential Smoothing	36.80
Double Exponential Smoothing	36.82
Triple Exponential Smoothing (Additive Season)	20.33

- Till now, Best Model for Rose —> 2 Pt Moving Average

◆ **Best Models for Rose -**



[Q 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 $\alpha = 0.05$.

◆ **To Check Stationarity of Data -**

- We use Augmented Dicky - Fuller (ADF) Test to check the Stationarity of Data
- Hypotheses of ADF Test :

H_0 Time Series is not Stationary

H_a Time Series is Stationary

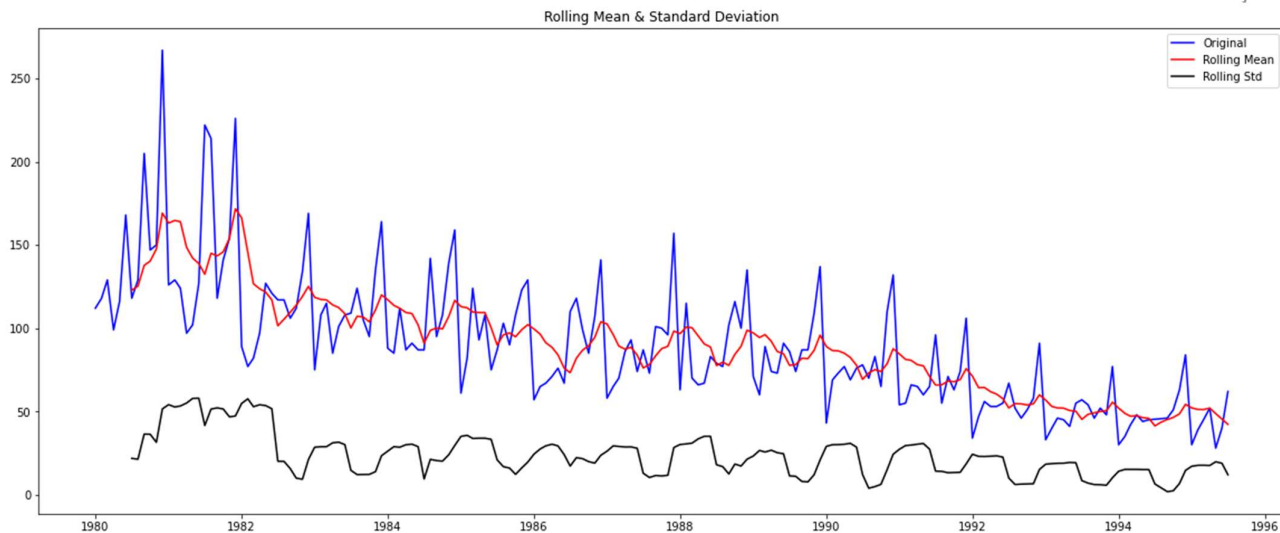
- So for Industry standard (also given for this problem), the Confidence Interval is 95%

- Hence, $\alpha = 0.05$
- So in ADF Test, if $p\text{-value} < \alpha \implies$ We reject the Null Hypothesis and hence conclude that given Time Series is Stationary
- So in ADF Test, if $p\text{-value} > \alpha \implies$ We fail to reject the Null Hypothesis and hence conclude that given Time Series is Not Stationary
- If Time Series is not Stationary then we apply one level of differencing and check for Stationarity again.
- Again, if the Time Series is still not Stationary, we apply one more level of differencing and check for Stationarity again
- Generally, with max 2 levels of differencing, Time Series becomes Stationary
- Once the Time Series is Stationary then we are ready to apply ARIMA / SARIMA models

Stationarity of Rose Wine Dataset -



- Augmented Dicky-Fuller Test was applied to the whole Rose dataset
- We found, $p\text{-value} = 0.3431$
- Here, $p\text{-value} > \alpha = 0.05$
- We fail to reject the Null Hypothesis and hence conclude that Rose Wine Time Series is Not Stationary
- We take 1 level of differencing and check again for Stationarity
- Now, $p\text{-value} = 1.8109e-12$
- Now, $p\text{-value} < \alpha = 0.05$
- Now, we reject the Null Hypothesis and conclude that Rose Time Series is Stationary with a lag of 1



[Q 6] Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE

✦ **ARIMA / SARIMA Models** -

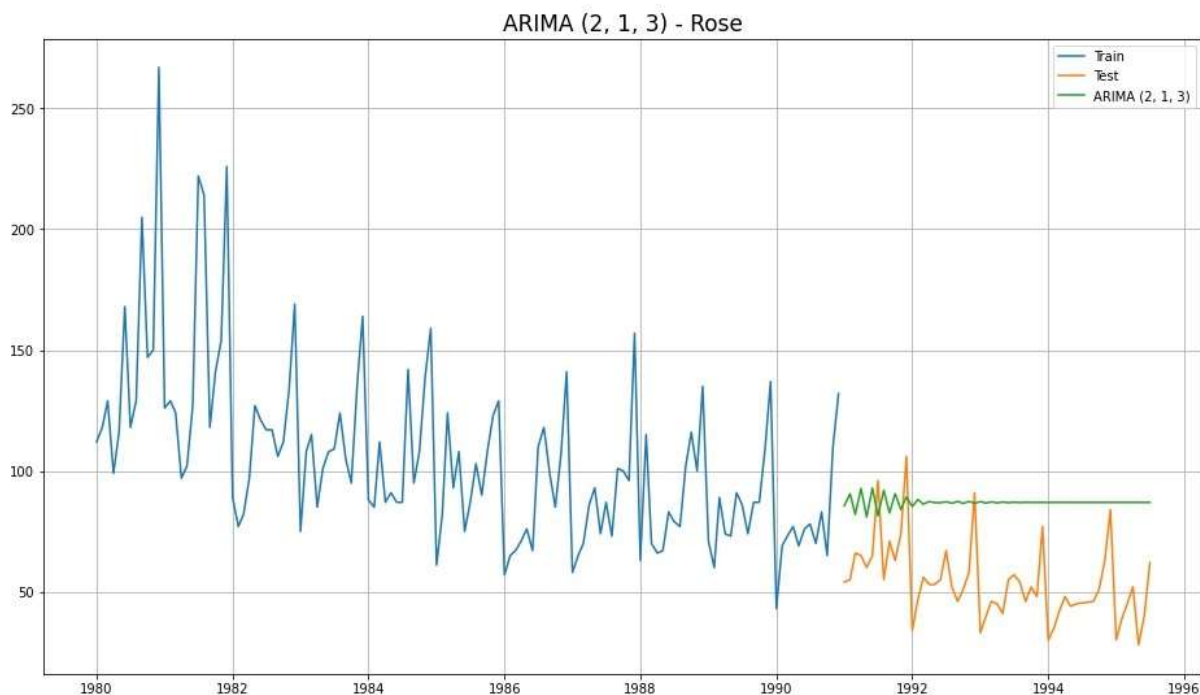
- ARIMA is an acronym for Auto-Regressive Integrated Moving Average
- SARIMA stands for Seasonal ARIMA, when the TS has seasonality
- ARIMA / SARIMA are forecasting models on Stationary Time Series

✦ **ARIMA / SARIMA Modelling on Train Rose Data** -

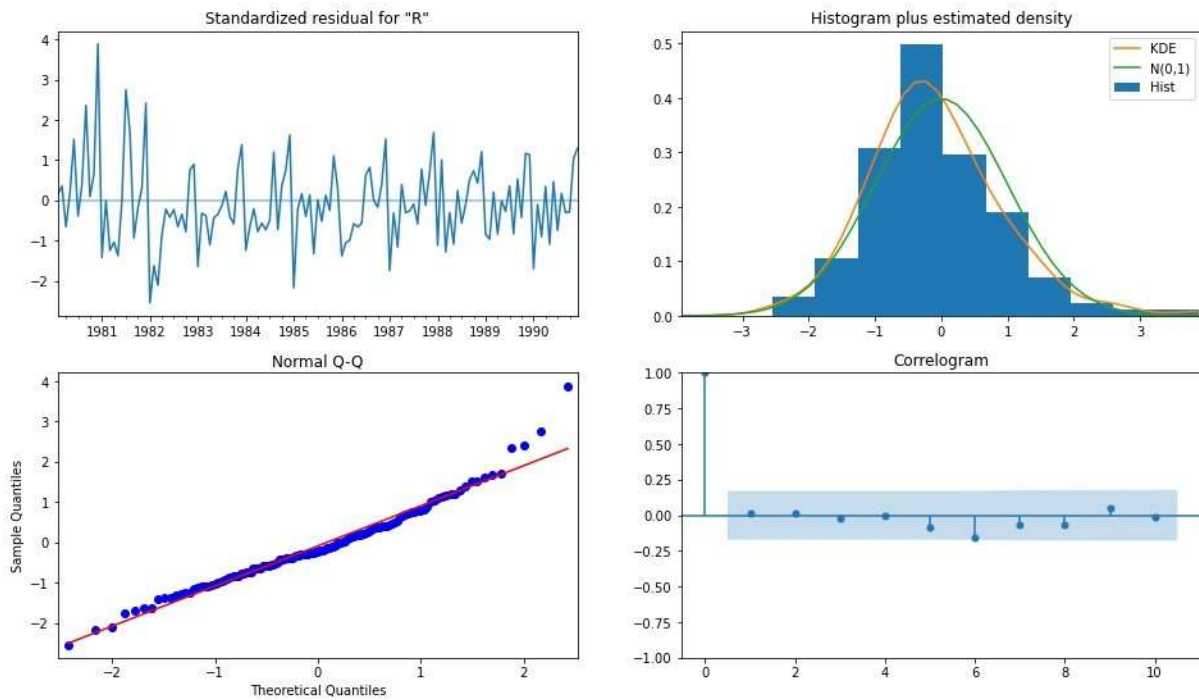
- We check for stationarity of Train Rose & Sparkling data by using Augmented Dicky Fuller Test
- We take a difference of 1 and make both these datasets Stationary
- We apply the following iterations to both these datasets -
 1. ARIMA Automated
 2. SARIMA Automated

1. ARIMA -

- We create a grid of all possible combinations of (p, d, q)
- Range of p = Range of q = 0 to 3, Constant $d = 1$
- Few Examples of the grid -
 - 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)
- We fit ARIMA models to each of these combinations for dataset.
- We choose the combination with the least Akaike Information Criteria (AIC)
- We fit ARIMA to this combination of (p, d, q) to the Train set and forecast on the Test set
- Finally, we check the accuracy of this model by checking RMSE of Test set
- For **Rose**, Best Combination with **Least AIC** is - **$(p, d, q) \rightarrow (2, 1, 3)$**



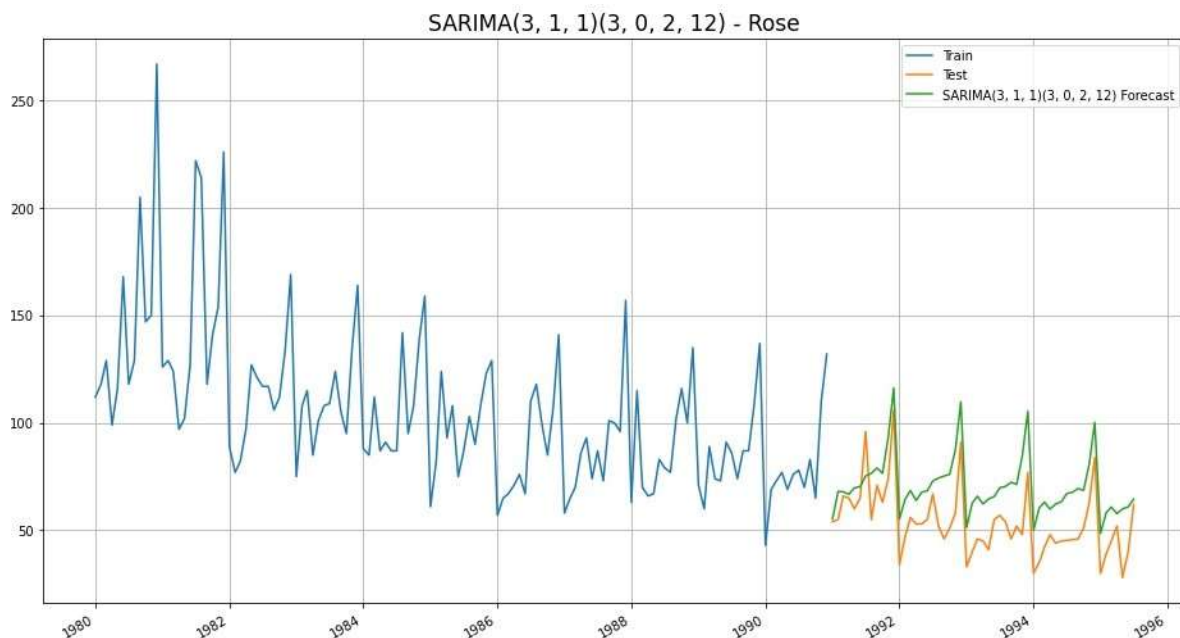
ARIMA (2, 1, 3) Diagnostic Plot - Rose



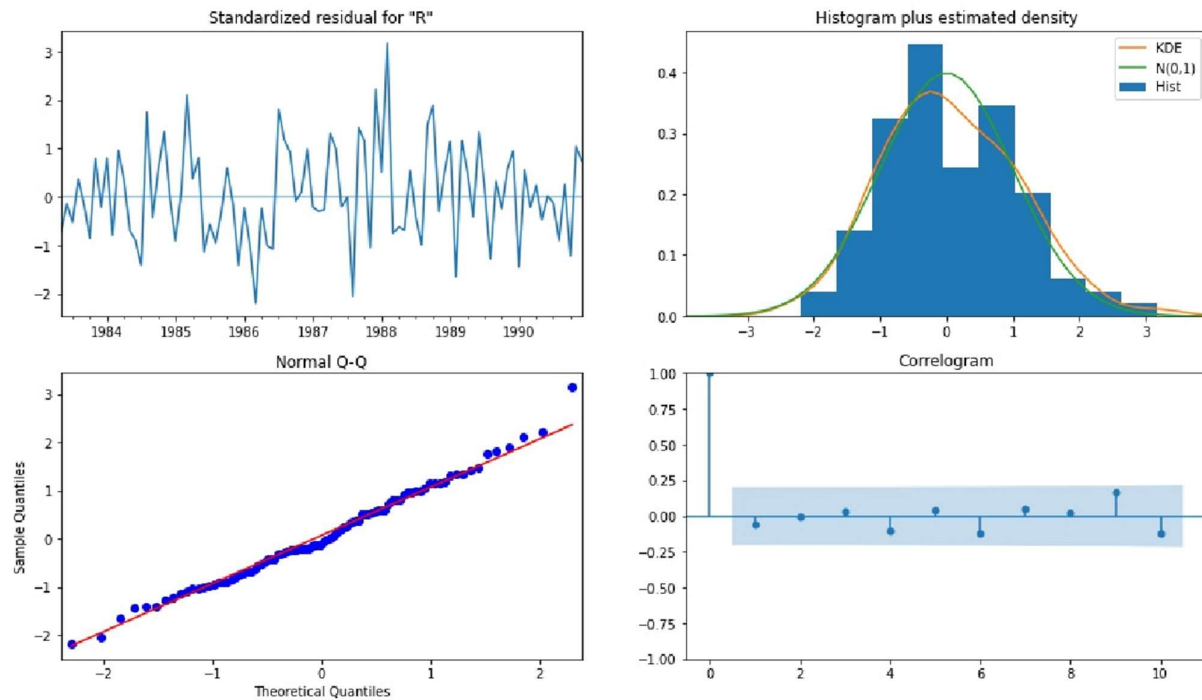
	Test RMSE Rose
ARIMA(2,1,3)	36.81

2. SARIMA Automated-

- We create a grid of all possible combinations of (p, d, q) along with Seasonal (P, D, Q) & Seasonality of 12
- Range of p = Range of q = 0 to 3, Constant d = 1
- Range of Seasonal P = Range of Seasonal Q = 0 to 3, Constant D = 1, Seasonality m = 12
- Few Examples of the grid (p, d, q) (P, D, Q, m) -
 - 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)
- We fit SARIMA models to each of these combinations and select with least AIC
- We fit SARIMA to this best combination of (p, d, q) (P, D, Q, m) to the Train set and forecast on the Test set. Then, we check accuracy using RMSE on Test set
- For **Rose**, Best Combination with **Least AIC** is - **(0, 1, 2) (2, 0, 2, 12)**



SARIMA (0, 1, 2) (2, 0, 2, 12) Diagnostic Plot - ROSE



	Test RMSE Rose
ARIMA(2,1,3)	36.81
SARIMA (0, 1, 2) (2, 0, 2, 12)	26.92

• **Till Now, Best Model for Rose with Least RMSE → SARIMA (0, 1, 2) (2, 0, 2, 12)**

[Q 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.

♦ **Auto-Correlation Function (ACF) -**

Autocorrelation refers to how correlated a time series is with its past values. e.g. y_t with y_{t-1} also y_{t+1} with y_t and so on.

- 'Auto' part of Autocorrelation refers to Correlation of any time instance with its previous time instance in the SAME Time Series
- ACF is the plot used to see the correlation between the points, up to and

including the lag unit.

- ACF indicates the value of 'q' - which is the Moving Average parameter in ARIMA / SARIMA models

✦ **Partial Auto-Correlation Function (PACF)** -

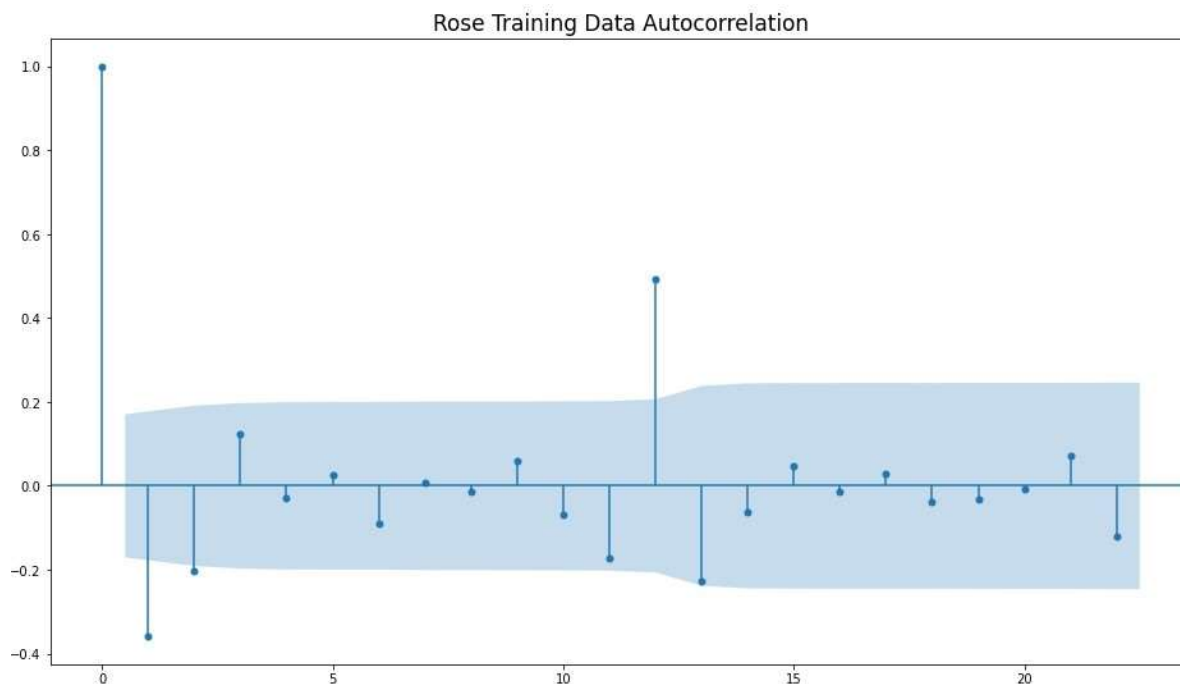
- Partial Autocorrelation refers to how correlated a time series is with its past lag values.
- For example, let lag=k, then Partial Autocorrelation is Correlation of y_t with y_{t-k} , ignoring the effects of all the instances between y_t and y_{t-k}
- PACF is the plot used to see the correlation between the lag points
- PACF indicates the value of 'p' - which is the Auto-Regressive parameter in ARIMA / SARIMA models

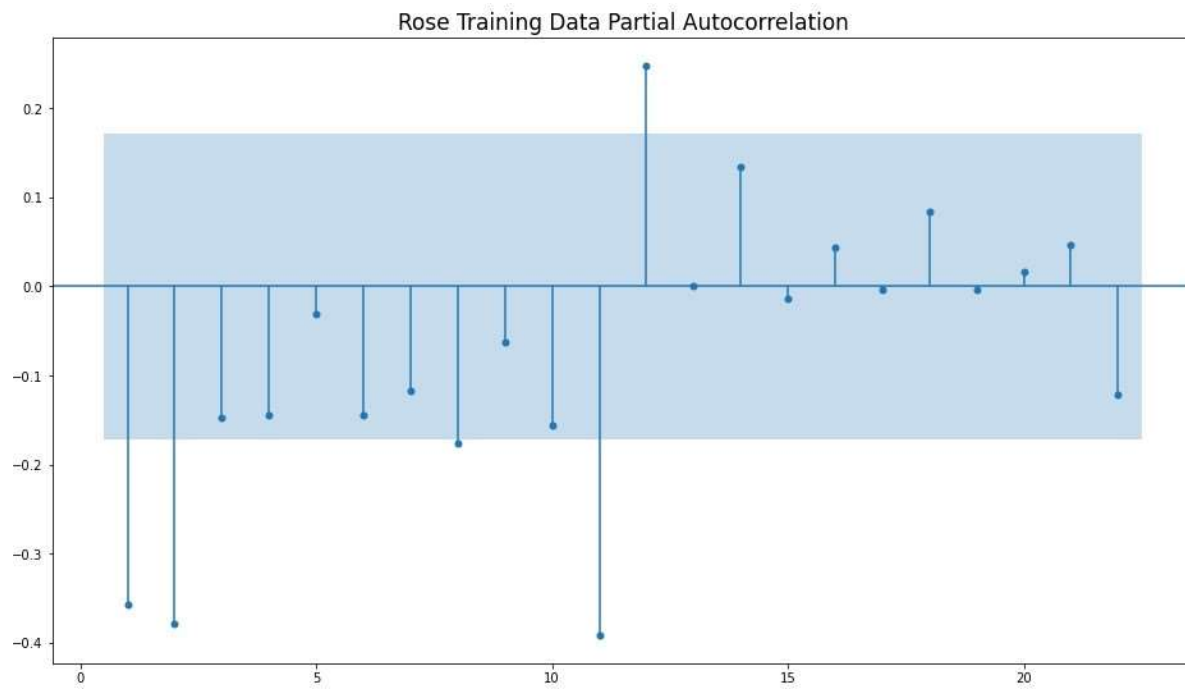
✦ **ACF & PACF of Rose** -

- Observing the cutoffs in ACF and PACF plots for Rose dataset, we get -

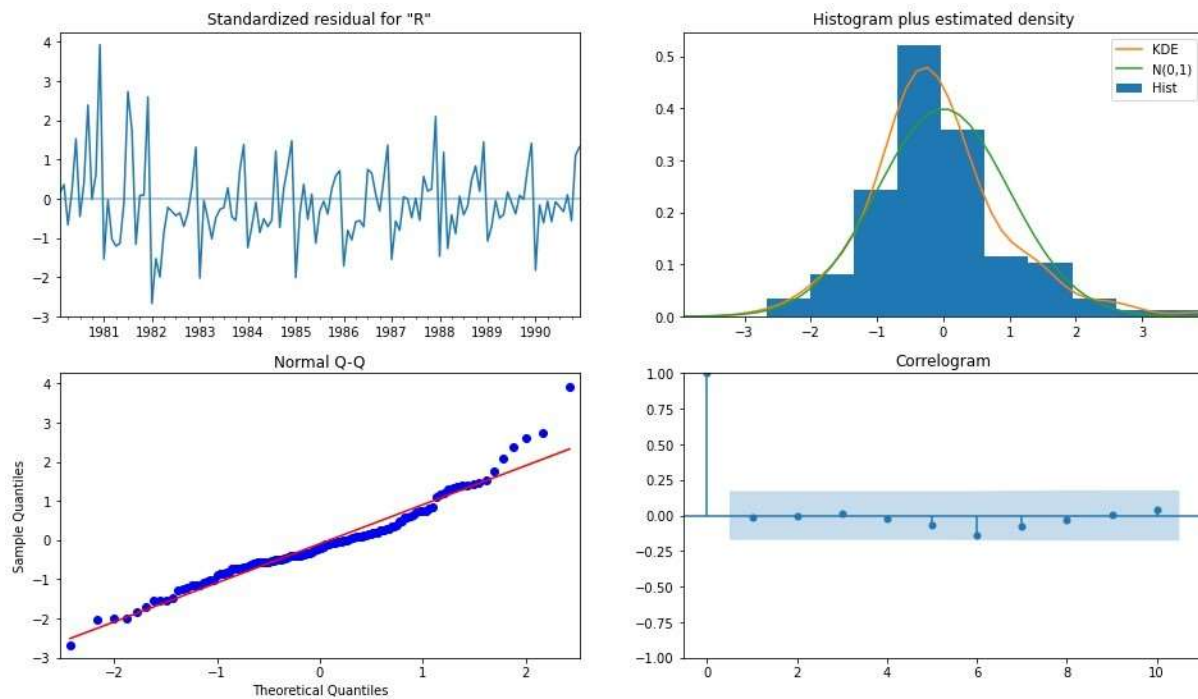
FOR ARIMA → $p = 2$, $q = 2$ and difference $d = 1$

FOR SARIMA → $p = 2$, $q = 2$, $d = 1$ and $P = 2$, $D = 1$, $Q = 2$, Seasonality=12



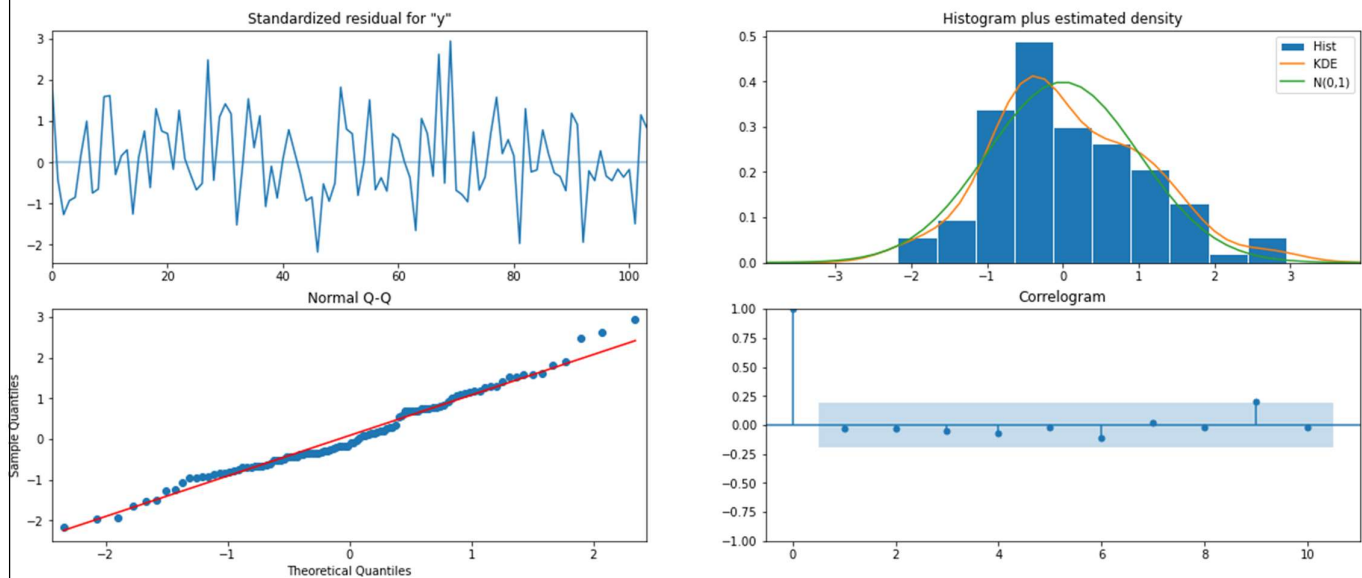


ARIMA (2, 1, 3) Diagnostic Plot - ROSE



	Test RMSE Rose
ARIMA(2, 1, 3)	36.81

3. SARIMA Manual - Rose - (0, 1, 2) (2, 0, 2, 12)



	Test RMSE Rose
ARIMA(2, 1, 3)	36.81
SARIMA (0, 1, 2) (2, 0, 2, 12)	26.92

- In all Manual methods, Best Model for Rose with Least RMSE

—> **SARIMA (0, 1, 2) (2, 0, 2, 12)**

- Seasonal P and Q - it was difficult to gauge the correct values here as the data was not enough and cutoffs were not visible
- Hence, we tried multiple combinations of Seasonal P and Q as given above

[Q 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 testdata.

♦ **All Models built with ROSE (sorted by RMSE) -**

Models with Parameters	Test RMSE
RegressionOnTime	15.269029
NaiveModel	79.718988
SimpleAverageModel	53.460790
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451433
6pointTrailingMovingAverage	14.566399
9pointTrailingMovingAverage	14.727667
Alpha=0.098,SimpleExponentialSmoothing	36.796467
Alpha=0.1,Beta =0.1 DoubleExponentialSmoothing	36.828257
Alpha=0.0704,Beta=0.0465,Gamma=4.0395,TripleExponentialSmoothing	20.333666
Alpha=0.1,Beta=0.2,Gamma=0.1,TripleExponentialSmoothing	9.223633
ARIMA(2,1,3)	36.807423
ARIMA(4,1,2)	37.037862
SARIMA(1,1,2)(0,0,2,11)	36.995556
SARIMA(0,1,2)(2,0,2,12)	26.928591

[Q 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.

• **Best Models as per the Least RMSE on ROSE Test set —>**

- **2 Pt Trailing Moving Average**

- **Triple Exponential Smoothing**

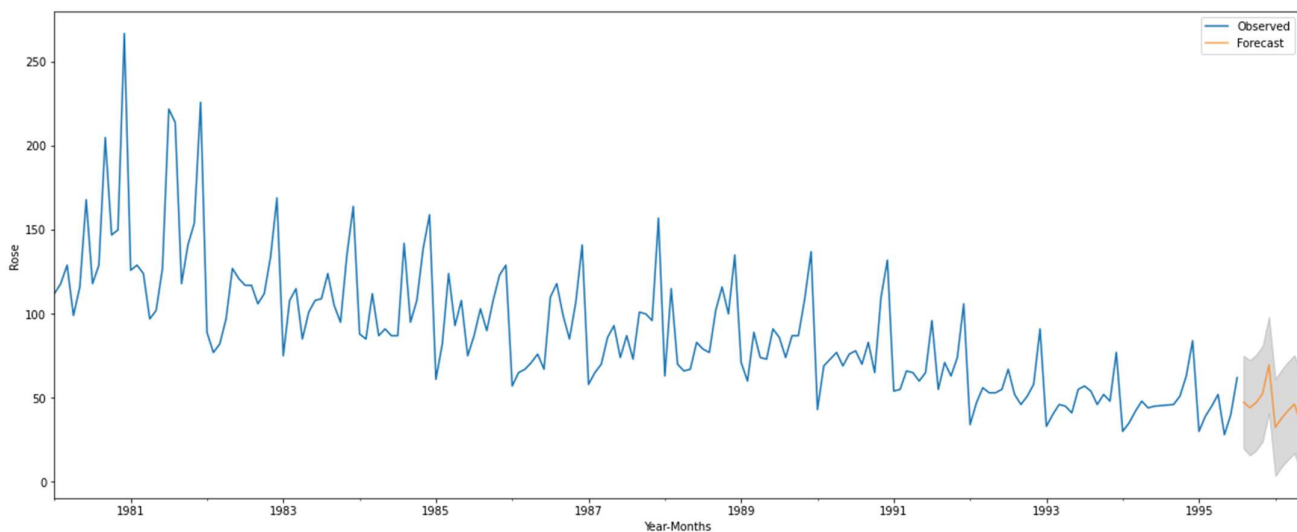
♦ **Rose Forecast Next 10 months -**

- **2 Pt Trailing Moving Average**

• **doesn't seem to be predicting very well**

• **Hence, forecasting on the second best model - Triple Exponential Smoothing
ETS(A, A, A) - Additive Seasonality**

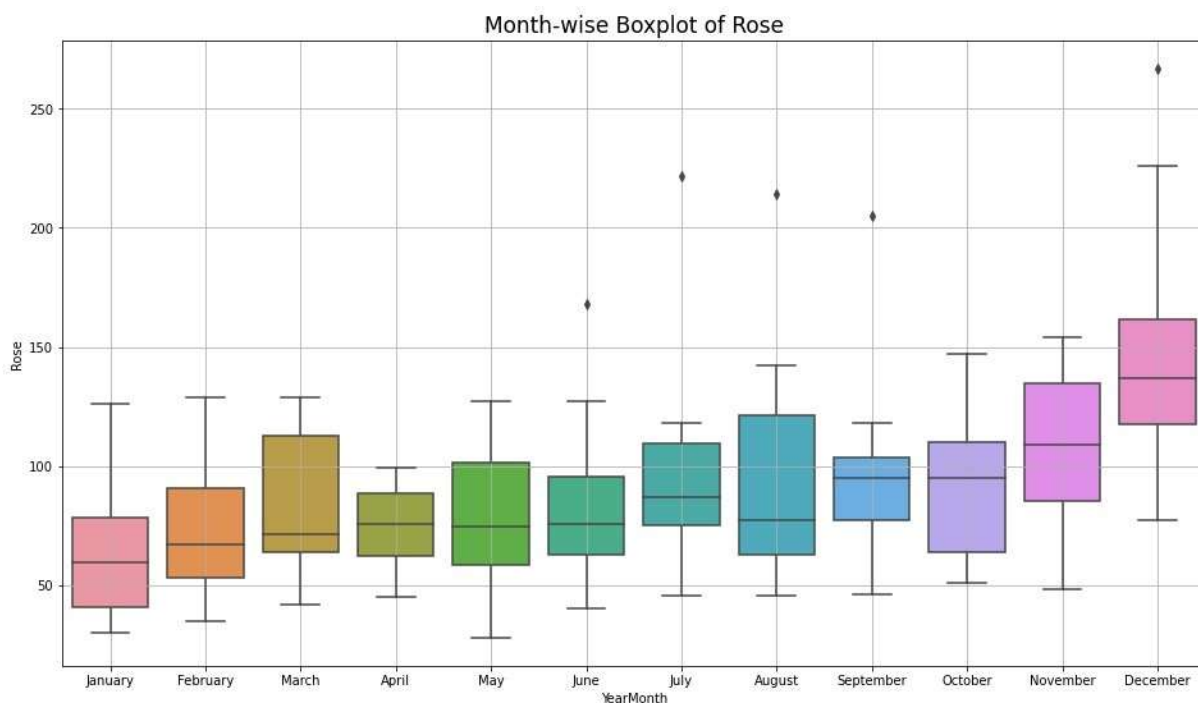
♦ **Rose Forecast Next 10 months - Triple Exponential Smoothing**



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

♦ **Rose Wine Sales - Comments :**

- Rose wine shows a clear trend of declining sales since 1980
 - This shows decline in popularity of this variant of wine
- Also, there is a clear spike in sales seen in the last quarter of every year from Oct to Dec
 - This might be due to the Holiday season in this period
 - Highest peak in sales is seen in Dec every year
- There is also an instant crashing slump in sales in the first quarter of every year from Jan
- This might be due to the after effect or hangover of Holidays.
- Sales slowly pick up only after May-June



◆ Rose Wine Sales - Forecast Models :

- Top 2 best models as per lowest Test RMSE were found to be - 2 Pt Moving Average and Holt-Winters - Additive Seasonality & Trend
- 2 Pt Moving Average model, when used for forecasting do not seem to give good predictions. Forecast values level out after a few iterations
- Holt-Winters seems to give a consistent forecast with respect to the data
- Hence, for **final forecast of Rose Wine Sales - we choose Holt-Winters**

◆ Rose Wine Sales - Suggestions :

- Firstly, Holiday season is around the corner and forecast shows increasing sales and sharp peak in Dec. Hence, Company should stock up
- But Declining sales of Rose Wine over the long period should be investigated with more data crunching
- Company can rebrand its Rose variant along-with a new Wine-master
- Company should take advantage of the oncoming spike from Aug-Oct by introducing aggressive offers and Ad campaigns.
- This will entice first time Wine drinkers and fence sitters (who don't have specific loyalties to any particular brand)
- Still if there is no significant upward trend in sales by this Dec, then Company has 2 options - invest in R&D or think of discontinuing this variant and come up with something completely new.

◆