PROJECT
TIME SERIES FORECAST
ROSE WINE

TABLE OF CONTENT

List of Figure and Table

Problem 2: Rose Wine

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

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

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Figure 1: Dataset

	count	mean	std	min	25%	50%	75%	max
Rose	187.0	90.042781	39.114366	28.0	62.5	85.0	111.0	267.0

Figure 2: Description of data

```
<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    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Figure 4: Info of data set

YearMonth	Rose
1994-07-01	NaN
1994-08-01	NaN

Figure 3 : Null value

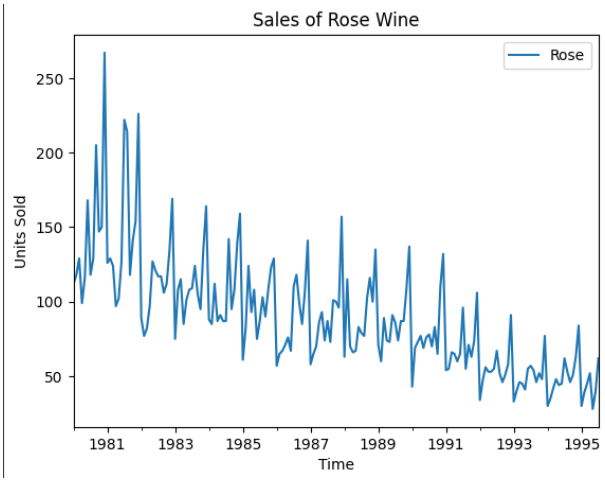


Figure 5 :Time series plot

Inference

- The data type of first column YearMonth is datetime64 and rose is int64, first column is converted to index column
- There are two columns with 187 rows
- Rose column represents sales of rose wine
- 2 null value in the data, it is missing month value is taken the average of consecutive year I.e July 94 value is the average of July 93 & July 95
- From the description of the data, min value is 28, max value is 267.

- From plot of time series observed that data has seasonality and trend, and it is decreasing

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

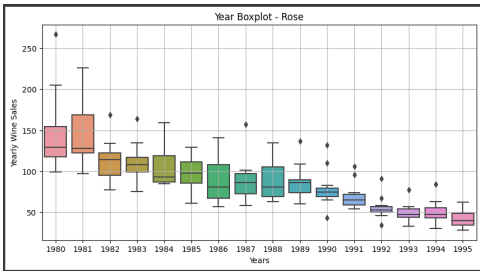


Figure 6: Boxplot of year sales

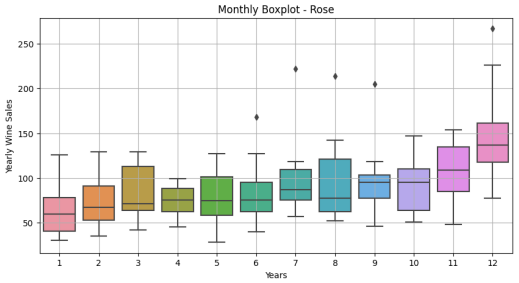


Figure 7: Boxplot of monthly sales

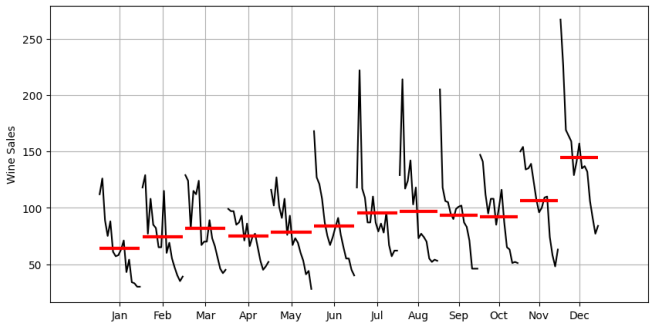


Figure7: Monthly time series

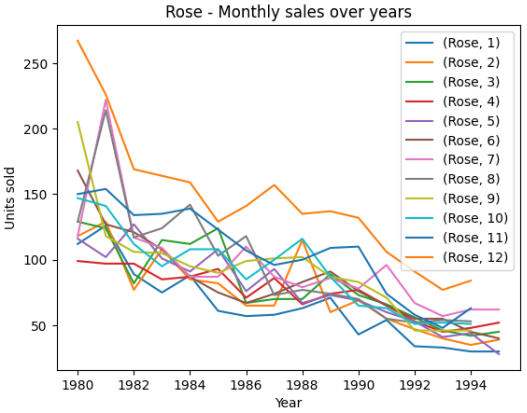


Figure8: Monthly sales over year

	Rose											
YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	62.0	53.0	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Figure9: Monthly sales over year table

Inference

- From fig 6 , Boxplot of year sales indicates the presence of outlier and presence of trend (downwards)
- From Figure 7 Boxplot of monthly sales increasing gradually from Jan to Dec
- Figure7: Monthly time series, the sales of wine increasing gradually from Jan to Dec

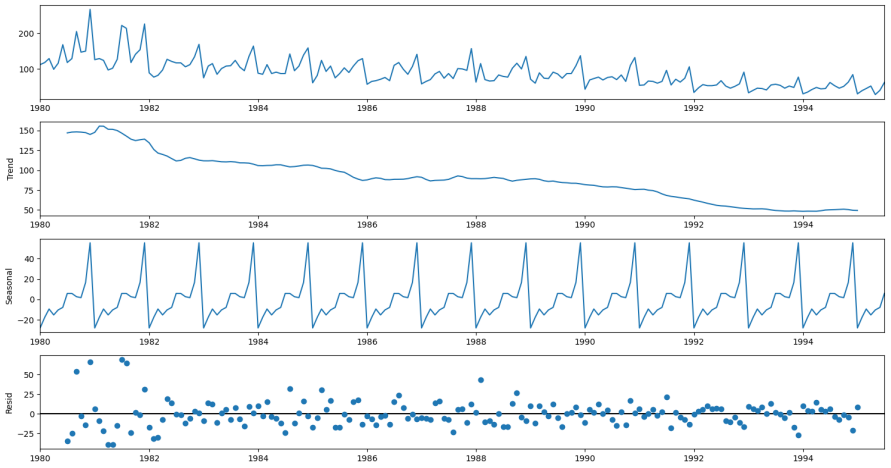


Figure10: Additive decomposition

- Decreasing trend is observed
- Yearly Seasonality is present is data.
- Residual has no pattern

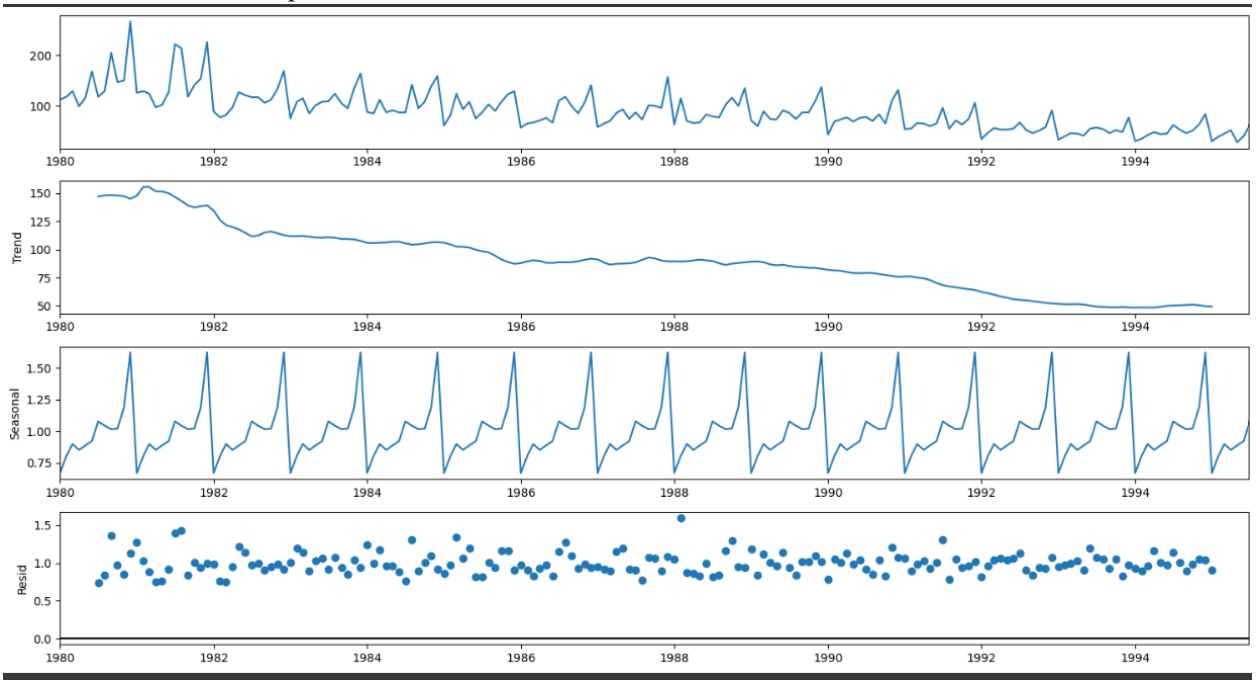


Figure11: Multiplication decomposition

- Decreasing trend is observed
- Yearly Seasonality is present is data. Sales reaches maximum in december every year
- Residual has no pattern

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

The data is split to training and test data, and test data split starts at 1991
Train has 132 records and 55 records..

First few rows of Training Data

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Last few rows of Training Data

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

Fig 12: Head of train and test data

First few rows of Test Data

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Last few rows of Test Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Fig 13: Tail of train and test data

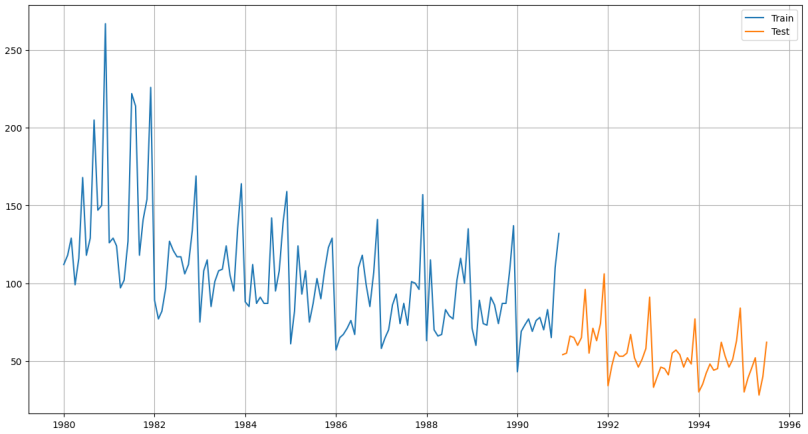


Fig 14: Plot of train and test data

4. Build all the 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.

Model 1: Linear Regression

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

First Few Rows of Train Data

YearMonth	Rose	time
1990-08-01	70.0	128
1990-09-01	83.0	129
1990-10-01	65.0	130
1990-11-01	110.0	131
1990-12-01	132.0	132

Last Few Rows of Train Data

Fig 15: train and test data

YearMonth	Rose	forecast_lr
1991-01-01	54.0	72.063
1991-02-01	55.0	71.569
1991-03-01	66.0	71.075
1991-04-01	65.0	70.580
1991-05-01	60.0	70.086

Fig 16: Sample forecast data

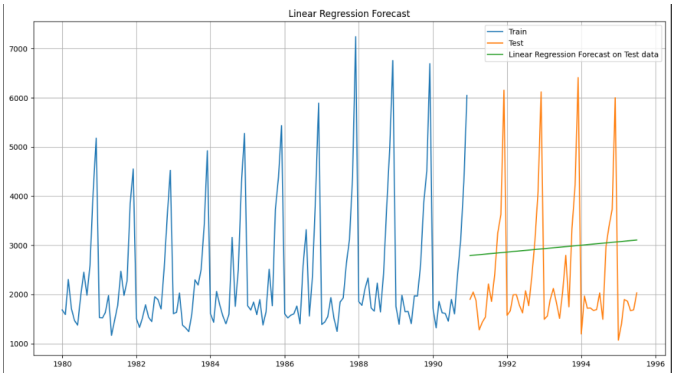


Fig 17: Plot of forecasted sales in LD Model

- The data is split accordingly and numerical time instance is generated
- Head and tail is shown in fig 15
- Fig 16 is the forecasted value

- Figure 17 us the Plot of forecasted sales in LD Model , green line indicated LD which is not matching with actual value
- RMSE calculated is 15.3

Model 2: Naive Approach

YearMonth	Rose	forecast_naive
1991-01-01	54.0	132.0
1991-02-01	55.0	132.0
1991-03-01	66.0	132.0
1991-04-01	65.0	132.0
1991-05-01	60.0	132.0

Fig 18: Forecasted value table

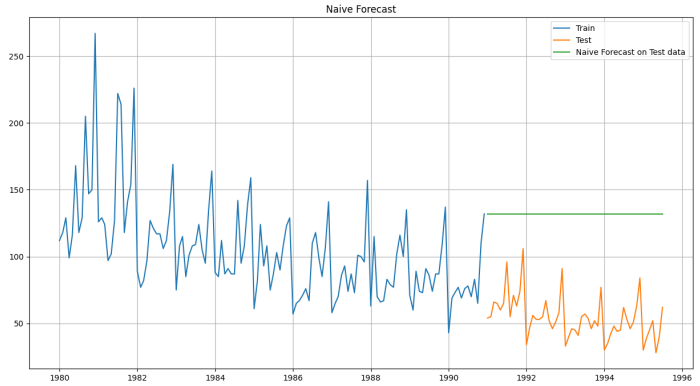


Figure 19: Plot of forecasted sales in Naive model

- The forecast value is same throughout
- Forecasted value is a straight line and not matching with actual value
- RMSE is 79.28

Model 3: Simple average approach

YearMonth	Rose	forecast_sa
1991-01-01	54.0	104.939
1991-02-01	55.0	104.939
1991-03-01	66.0	104.939
1991-04-01	65.0	104.939
1991-05-01	60.0	104.939

Fig 20: Forecasted value table

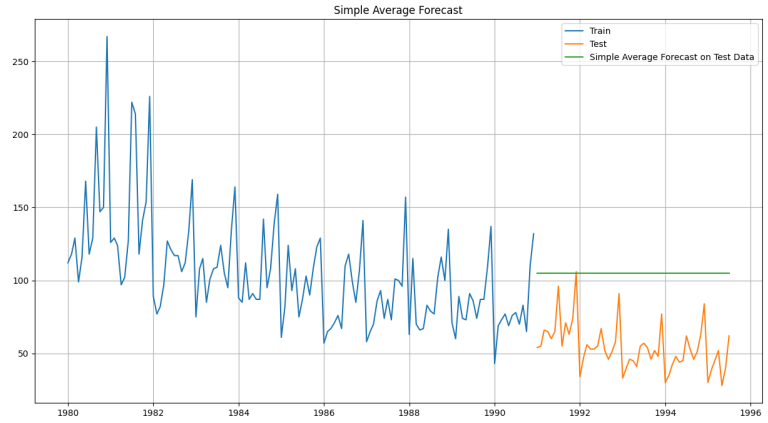


Figure 21: Plot of forecasted sales in Simple average model

- The forecast value is same throughout
- Forecasted value is a straight line and not matching with actual value
- RMSE is 53.02

Model 4: Simple Exponential Smoothing

YearMonth	Rose	forecast_ses_optimized
1991-01-01	54.0	87.105
1991-02-01	55.0	87.105
1991-03-01	66.0	87.105
1991-04-01	65.0	87.105
1991-05-01	60.0	87.105

Fig 22: Forecasted value table

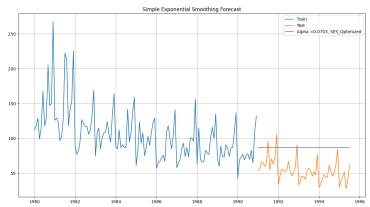


Figure 23: Plot of forecasted sales in SES

- The model doesnt consider trend and seasonality
- The forecasted value is same throughout

- Forecasted value is a straight line and not matching with actual value
- RMSE is 36.3

Optimizing Alpha based on Test RMSE

- Different values are used to find the lowest RMSE. From it best alpha is 0.1
- Forecasted value is a straight line and not matching with actual value
- RMSE is 36.4

YearMonth	Rose	forecast_ses_optimized	forecast_ses
1991-01-01	54.0	87.105	87.14
1991-02-01	55.0	87.105	87.14
1991-03-01	66.0	87.105	87.14
1991-04-01	65.0	87.105	87.14
1991-05-01	60.0	87.105	87.14

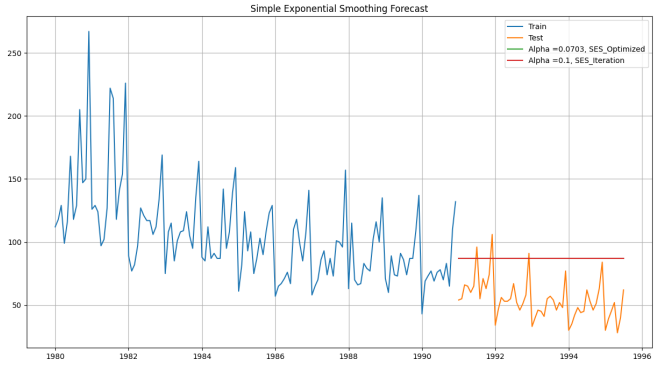


Fig 24: Forecasted value table Figure 25: Plot of forecasted sales in SES optimized alpha

Model5: Double Exponential Smoothing

YearMonth	Rose	forecast_des_optimized
1991-01-01	54.0	72.069
1991-02-01	55.0	71.575
1991-03-01	66.0	71.080
1991-04-01	65.0	70.586
1991-05-01	60.0	70.092

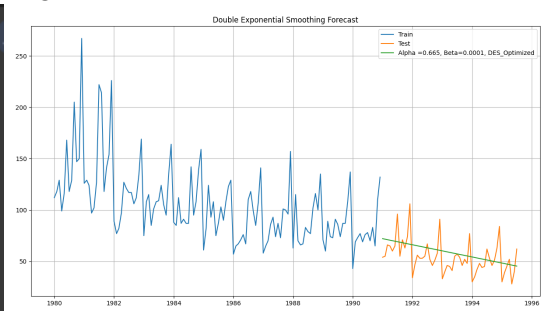


Fig 26: Forecasted value table Figure 27: Plot of forecasted sales in DES

YearMonth	Rose	forecast_des_optimized	forecast_des
1991-01-01	54.0	72.069	83.851
1991-02-01	55.0	71.575	83.961
1991-03-01	66.0	71.080	84.070
1991-04-01	65.0	70.586	84.180
1991-05-01	60.0	70.092	84.290

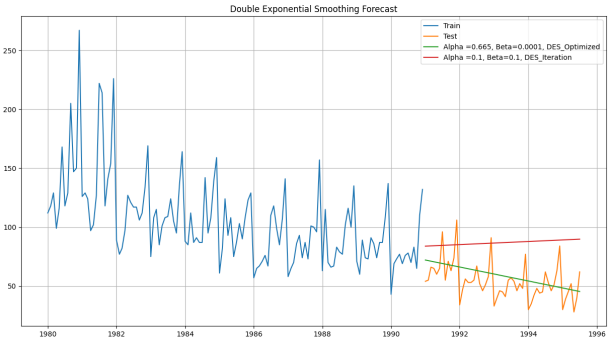
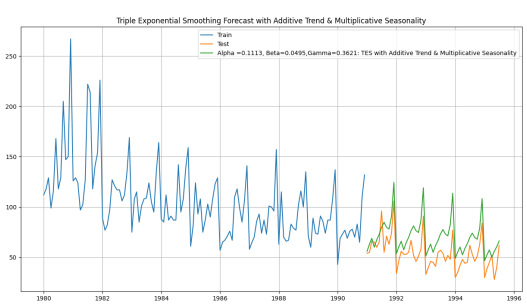


Fig 28: Forecasted value table Figure 29: Plot of forecasted sales in DES optimized alpha, beta

- This model has trend but no seasonality the rmse without optimisation is 15.3
- After optimizing alpha and beta between 0 to 1. The best model is optimised, RMSE : 36.4

Fig 30: Triple Exponential Smoothing with Additive trend & Additive seasonality
Fig 31: Triple Exponential Smoothing with Additive trend & Multiplication seasonality



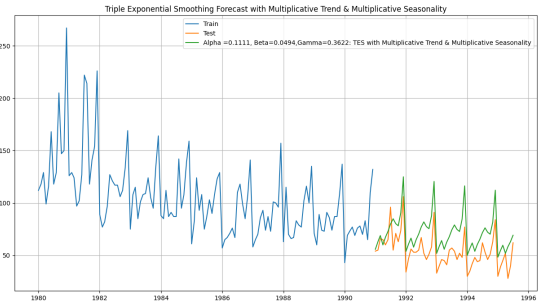
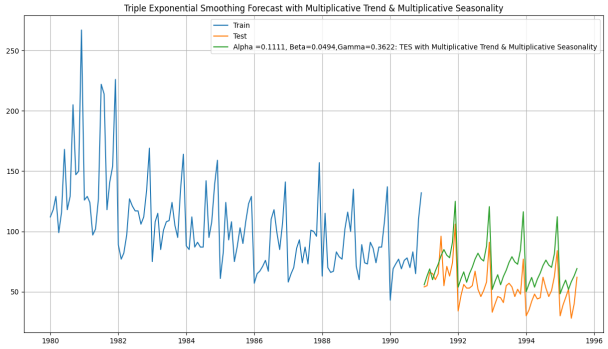
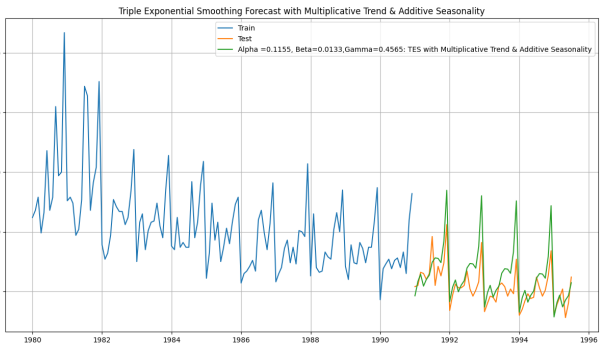


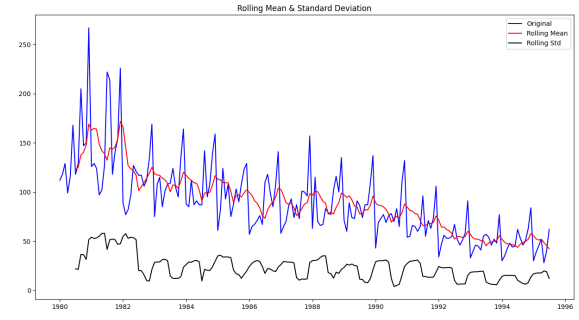
Fig 32: Triple Exponential Smoothing with Multiplication trend & Additive seasonality
Fig 33: Triple Exponential Smoothing with Multiplication trend & Multiplication seasonality





- RMSE of Triple Exponential Smoothing with Additive trend & Additive seasonality .. 23.0
- RMSE of Triple Exponential Smoothing with Additive trend & Multiplication seasonality..1.6
- RMSE of Triple Exponential Smoothing with Multiplication trend & Additive seasonality..19.8
- RMSE of Triple Exponential Smoothing with Multiplication trend & Multiplication seasonality..27.2

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.



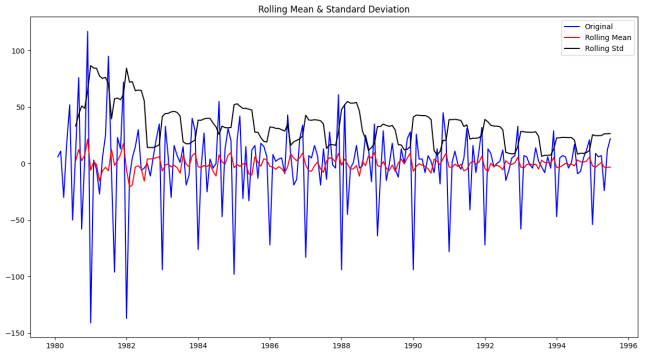


Fig 34 Dickey Fuller test on whole time series Fig 35 Dickey Fuller test on differenced whole time series

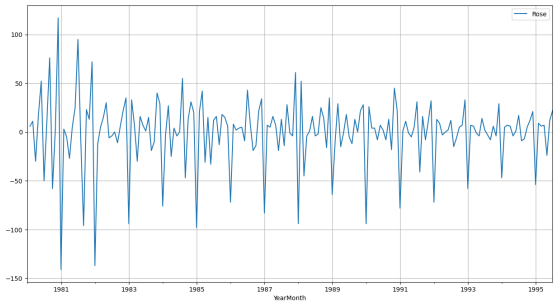


Fig 36:Plot of differenced whole time series

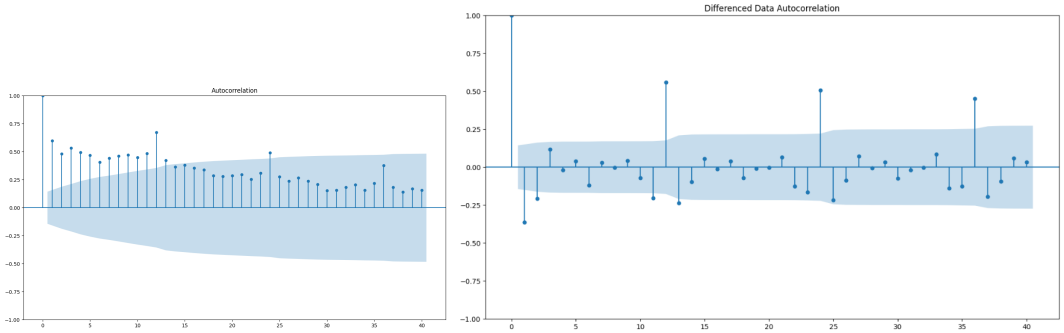


Fig 37:Autocorrelation plot of whole time series Fig 38:Autocorrelation plot of differenced whole time series

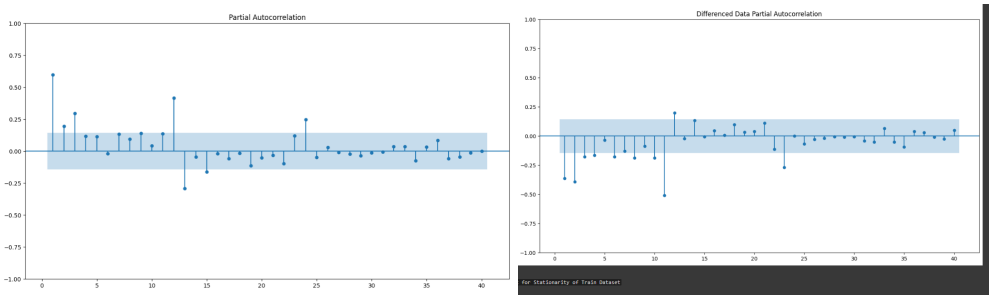


Fig 39:Partial Autocorrelation plot of whole time series Fig 40: Partial Autocorrelation plot of differenced whole time series

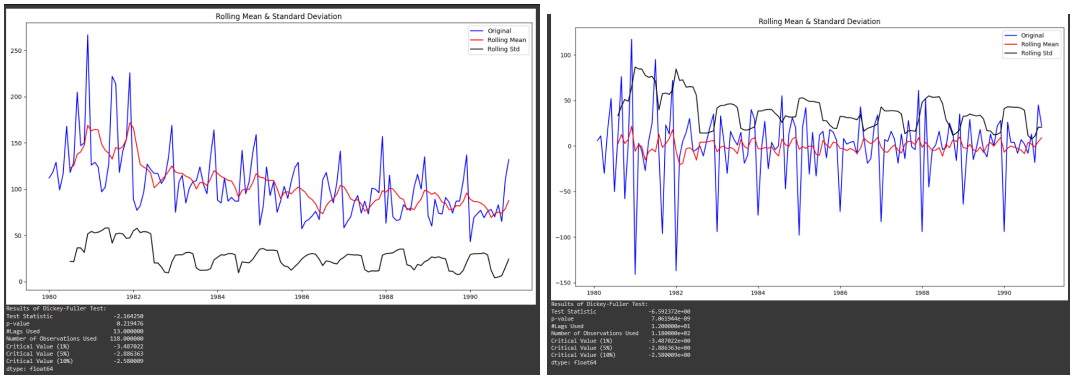


Fig 41: Stationarity of Train Dataset Fig 42: Stationarity of differencing Train Dataset

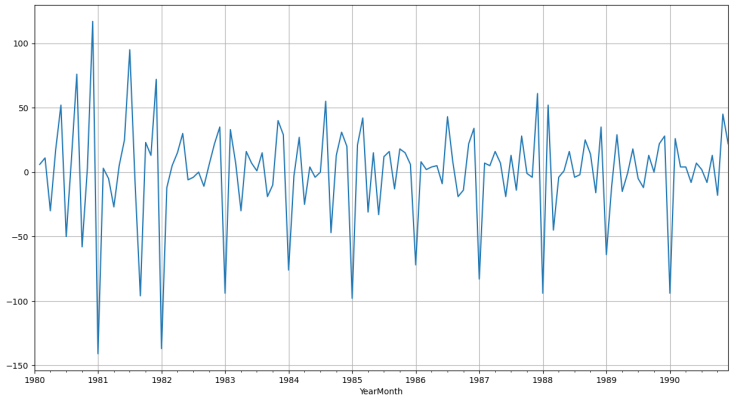
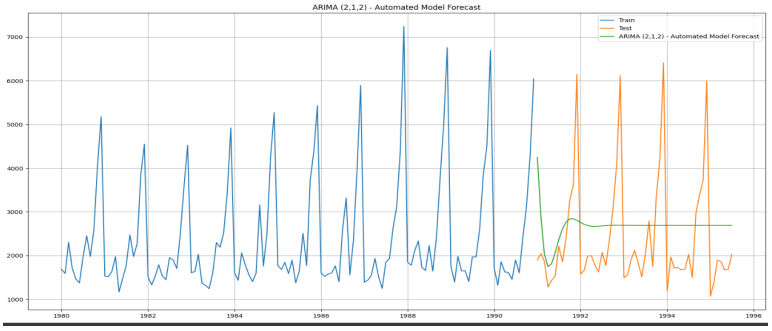


Fig 43: Differenced Time series

Inteference

- ADF is a test to find if the model is stationary or not
- Ho states model is non stationary, H1 states model is stationary
- If alpha is <=0.05 then null value can be disproved
- p is 0.219 training dataset is non stationary. Take differencing on training set
- p is 0.0 training dataset is stationary at one level differenced training 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.

The arima model is created with p =0 to 3, q = 0 to 3, d =1

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-635.935
Date:	Sun, 23 Apr 2023	AIC	1281.871
Time:	11:44:46	BIC	1296.247
Sample:	01-01-1980 - 12-01-1990	HQIC	1287.712

Covariance Type: opg

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

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

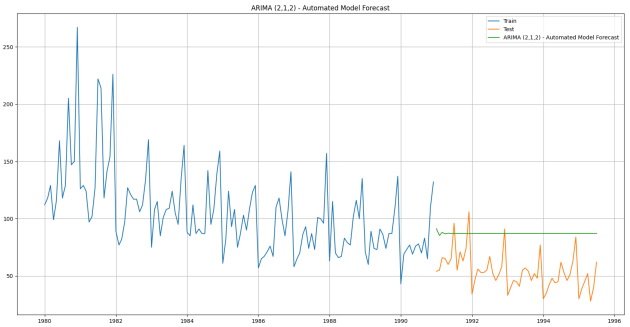
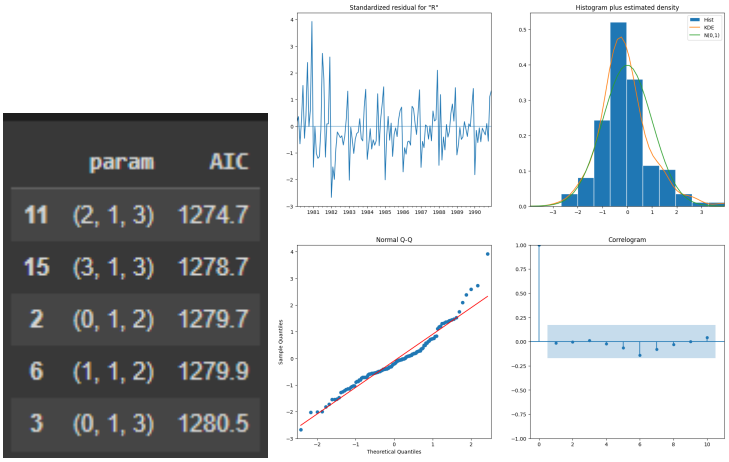


Fig 44: summary automated arima model Fig 45 Plot of forecasted sales automated arima model



	param	AIC
11	(2, 1, 3)	1274.7
15	(3, 1, 3)	1278.7
2	(0, 1, 2)	1279.7
6	(1, 1, 2)	1279.9
3	(0, 1, 3)	1280.5

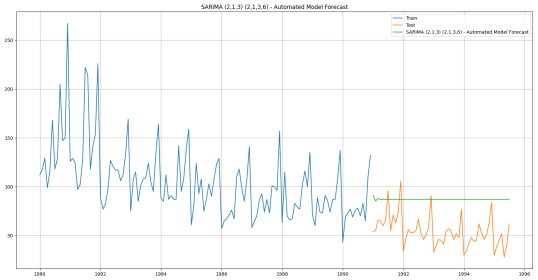
Fig 46:diagnostic automated arima model Fig 47:Order based on AIC

- From this observed actual not matching with forecast, it is thus not a good model for forecasting.
- RMSE is 36.45
- Based on the Lowest AIC the best parameter is (2,1,3)

SARIMA Model

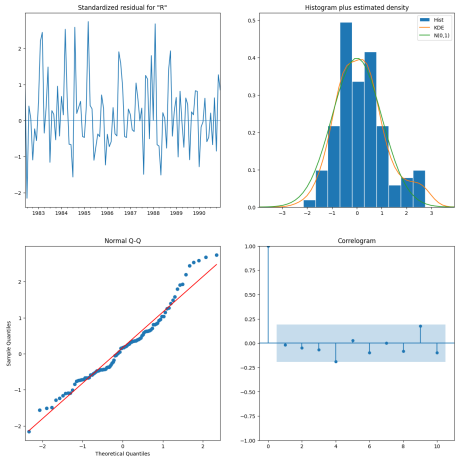
Following value p between 0 to 3, q between 0 to 3, d between 0 to 1

- From this observed actual not matching with forecast, it is thus not a good model for forecasting.
- RMSE is 16.67
- Based on the Lowest AIC the best parameter is (2, 1, 3, 6) AIC = 889.2



	param	param_seasonal	AIC
375	(2, 1, 3)	(2, 1, 3, 6)	889.2
503	(3, 1, 3)	(2, 1, 3, 6)	891.1
511	(3, 1, 3)	(3, 1, 3, 6)	893.1
367	(2, 1, 3)	(1, 1, 3, 6)	894.8
127	(0, 1, 3)	(3, 1, 3, 6)	894.9

Fig 48: Plot of forecasted sales automated sarima model Fig 51::Order based on AIC



Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 1, 3, 6)	Log Likelihood	-433.595
Date:	Sun, 23 Apr 2023	AIC	889.189
Time:	11:53:14	BIC	918.172
Sample:	01-01-1980	HQIC	900.928
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5746	0.023	25.059	0.000	0.530	0.620
ar.L2	-0.9163	0.021	-43.475	0.000	-0.958	-0.875
ma.L1	-1.4571	155.859	-0.009	0.993	-306.936	304.022
ma.L2	1.5182	636.951	0.002	0.998	-1246.884	1249.920
ma.L3	-0.8408	425.368	-0.002	0.998	-834.547	832.866
ar.S.L6	-0.4345	0.106	-4.088	0.000	-0.643	-0.226
ar.S.L12	0.4837	0.102	4.741	0.000	0.284	0.684
ma.S.L6	-1.6622	79.404	-0.021	0.983	-157.292	153.968
ma.S.L12	-1.0720	211.531	-0.005	0.996	-415.666	413.521
ma.S.L18	1.5908	126.462	0.013	0.990	-246.270	249.452
sigma2	68.5368	3.56e+04	0.002	0.998	-6.98e+04	6.99e+04

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	5.81
Prob(Q):	0.86	Prob(JB):	0.05
Heteroskedasticity (H):	0.45	Skew:	0.56
Prob(H) (two-sided):	0.02	Kurtosis:	3.29

Fig 49:diagnostic automated sarima model Fig 50: summary automated sarima model

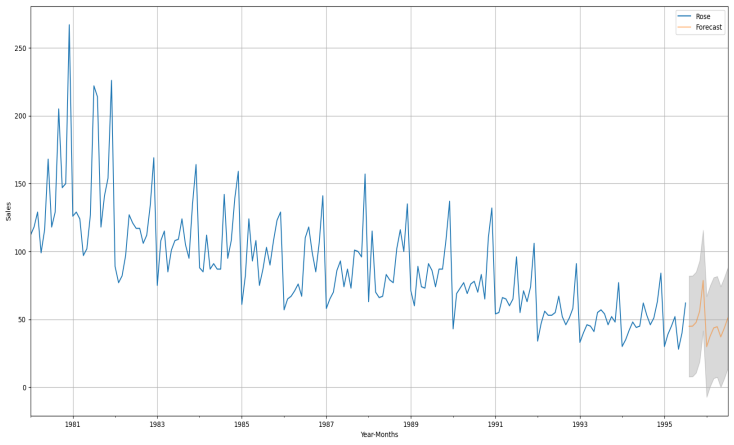
7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

	Test RMSE
Triple Exponential Smoothing Forecast with Multiplicative Trend & Multiplicative Seasonality Alpha =0.2, Beta=0.7,Gamma=0.2	9.053
Triple Exponential Smoothing Forecast with Additive Trend & Multiplicative Seasonality smoothing_level=0.1,smoothing_trend=0.2,smoothing_seasonal=0.1	9.266
Triple Exponential Smoothing Forecast with Additive Trend & Additive Seasonality	14.158
Linear Regression	15.303
Double Exponential Smoothing Forecast	15.305
Triple Exponential Smoothing Forecast with Multiplicative Trend & Additive Seasonality Alpha=0.1, Beta=0.8,Gamma=0.2	15.640
forecast_SARIMA_auto	16.673
Triple Exponential Smoothing Forecast with Additive Trend & Multiplicative Seasonality	18.681
Triple Exponential Smoothing Forecast with Multiplicative Trend & Multiplicative Seasonality	19.876
Triple Exponential Smoothing Forecast with Multiplicative Trend & Multiplicative Seasonality	19.876
Triple Exponential Smoothing Forecast with additive trend & additive seasonality Alpha =0.1, Beta=0.4,Gamma=0.3	23.029
Triple Exponential Smoothing Forecast with Multiplicative Trend & Multiplicative Seasonality Alpha =0.1111, Beta=0.0494,Gamma=0.3622	27.222
Simple Exponential Smoothing Forecast	36.382
Simple Exponential Smoothing Forecast with alpha =0.1	36.413
Double Exponential Smoothing Forecast Alpha =0.665, Beta=0.0001	36.449
forecast_ARIMA_auto	36.457
Simple Average	53.030
Naive Forecast	79.282

Fig 51: summary of all models

From the above the table, the best model is the triple exponential smoothing with additive trend and multiplicative seasonality with the parameters with alpha = 0.4, beta =0.1 and gamma = 0.2

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



	lower_CI	Forecast	upper_ci
1995-08-01	7.7	44.8	82.0
1995-09-01	7.8	45.0	82.1
1995-10-01	10.6	47.7	84.9
1995-11-01	18.8	55.9	93.1
1995-12-01	41.4	78.6	115.7
1996-01-01	-7.3	29.9	67.0
1996-02-01	0.6	37.7	74.8
1996-03-01	6.5	43.7	80.8
1996-04-01	7.4	44.6	81.7
1996-05-01	-0.1	37.1	74.2
1996-06-01	6.3	43.5	80.6
1996-07-01	13.7	50.9	88.0

Fig 52: Forecast with 95% confidence to 12 month in future fig Fig 53: Forecast with 95% confidence to 12 month in future values

Fig 53: Forecast with 95% confidence to 12 month in future table

Best model is the triple exponential smoothing with additive trend and multiplicative seasonality with the parameters with smoothing_level=0.2,smoothing_trend=0.7,smoothing_seasonal=0.2

- RMSE is 9.05
- The value is forecasted to 12 month interval with this model

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

- Best model is the triple exponential smoothing with additive trend and multiplicative seasonality with the parameters with smoothing_level=0.2,smoothing_trend=0.7,smoothing_seasonal
- RMSE is 9.05
- Total wine sales follows a similar downwards pattern overall
- Sales of wine is highest during the month of december
- Sales of wine is expected to have few outlier
- Sales increase gradually from aug to dec

Measures to increase sales:

- During peak season maximum stock should be available for more sales
- Find out the reason why sales is less in other months
- Since the trend is downwards action should be taken to improve some aspects to improve sales