Time Series Forecasting

PROJECT - 7

REPORT – Part 2

Piyush Kumar Singh
PGP – DSBA Online
May-21 Batch

Date: 16/01/2022

Table of Contents

- List of Figures 3
- List of Tables 4
- Dataset 1 – Rose.csv 5

1.1) Read the data as an appropriate Time Series data and plot the data 5

1.2) Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition 7

1.3) Split the data into training and test. The test data should start in 1991 12

1.4) 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 14

1.5) 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. 16

1.6) 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 18

1.7) 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. 21

1.8) Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data 25

1.9) 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 26

1.10) Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales 28

List of Figures

Fig.1 – Line plot of rose Time Series along with mean and median of our Time Series 6
Fig.2 – Yearly boxplot of rose sales 7
Fig.3 – Monthly boxplot of rose sales 8
Fig.4 – Monthly plot showing mean and variation of units sold 8
Fig.5 – Line chart of sales across years 9
Fig.6 – Empirical cumulative distribution graph 10
Fig.7 – Average sales and sales percentage change across years 10
Fig.8 – Additive model decomposition 11
Fig.9 – Multiplicative model decomposition 11
Fig.10 – Original Time Series Vs Time Series with decomposed component 12
Fig.11 – Train and Test set line plot 13
Fig.12 – Prediction graph of various models w.r.t to test set 16
Fig.13 – Time Series of difference of order 1 17
Fig.14 – Diagnostic plot of ARIMA(2,1,3) 19
Fig.15 – Diagnostic plot of SARIMA(3,1,1)(3,0,2,12) 20
Fig.16 – ACF and PACF plot of differenced time series 21
Fig.17 – Diagnostic plot of ARIMA(0,1,0) 22
Fig.18 – Diagnostic plot of SARIMA(0,1,0)(0,0,0,12) 24
Fig.19 – Line plot of Time Series Vs fitted values by model vs future prediction for 12 months 26
Fig.20 – Line plot of actual data Vs predicated data with 95% confidence interval 27
Fig.21 – Line plot of 12 months future prediction with 95% confidence interval 27

List of Tables

Table 1. Time Series dataset info 6
Table 2. Summary of Time Series Rose Dataset 7
Table 3. Monthly sales across the years 9
Table 4. ARIMA Automated Summary results 18
Table 5. SARIMA Automated Summary results 20
Table 6. ARIMA Manual Summary results 22
Table 7. SARIMA Manual Summary results 23
Table 8. Model comparison table using test RMSE 25
Table 9. Sales Prediction values for 12 months in future 26

Dataset: Rose.csv

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.

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

- Monthly sales of 'rose' wine from period Jan – 1980 to July – 1995 is provided in the Rose.csv file.

- The given data files is read and date range is inserted and the YearMonth column is to date-range and set as index to create a time series data having one column of 'Rose' showing sales value.

YearMonth	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Head

YearMonth	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Tail

- Numbers of records in the dataset is 187.
- There are 2 null values in our time series.

YearMonth	Rose
1994-07-31	NaN
1994-08-31	NaN

- I am using interpolate method 'pad' to replace the missing values.
- After replacement we get the following values in our time series for year 1994.

YearMonth	Rose
1994-01-31	30.0
1994-02-28	35.0
1994-03-31	42.0
1994-04-30	48.0
1994-05-31	44.0
1994-06-30	45.0
1994-07-31	45.0
1994-08-31	45.0
1994-09-30	46.0
1994-10-31	51.0
1994-11-30	63.0
1994-12-31	84.0

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Rose    187 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Table.1

- Info of dataset after replacement of missing values. As we can see there are no missing values now in our dataset.

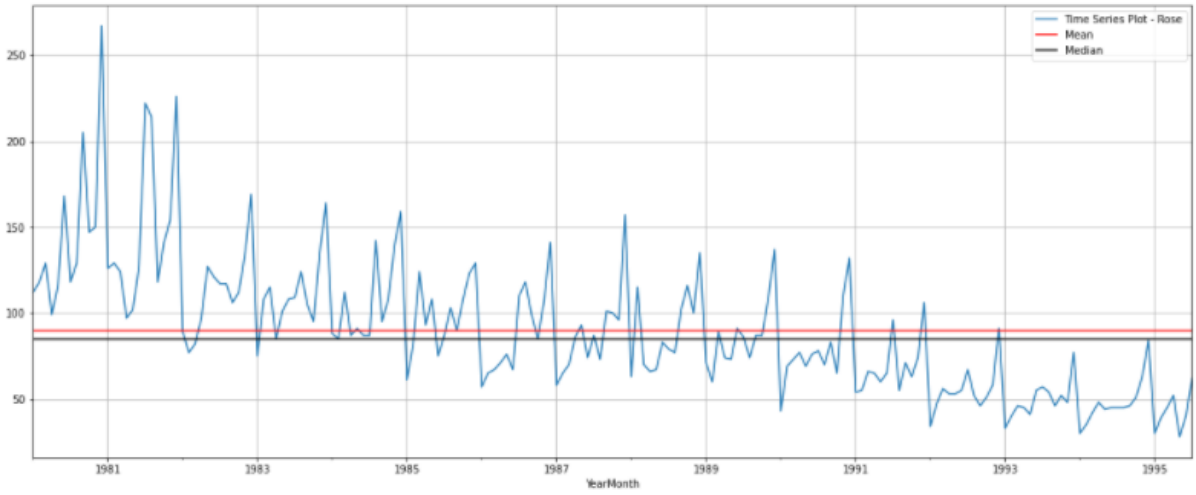

Fig.1

- Here we have plotted our time series to analyse it behaviour visually.
- Sales of Rose wine is showing a downward trend with sales decreasing over the given time period.
- Rose time series has significant seasonality.

1.2 Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition Read the data as an appropriate Time Series data and plot the data.

	Rose
count	187.000
mean	89.909
std	39.244
min	28.000
25%	62.500
50%	85.000
75%	111.000
max	267.000

Table.2

- The descriptive summary of the data shows that on an average 89 units of rose were sold each month on the given time period. 50 % of the sales have more than 85 unit's sales every month.
- Maximum sale in a month was of 267 units and minimum sale in a month was 28 units.

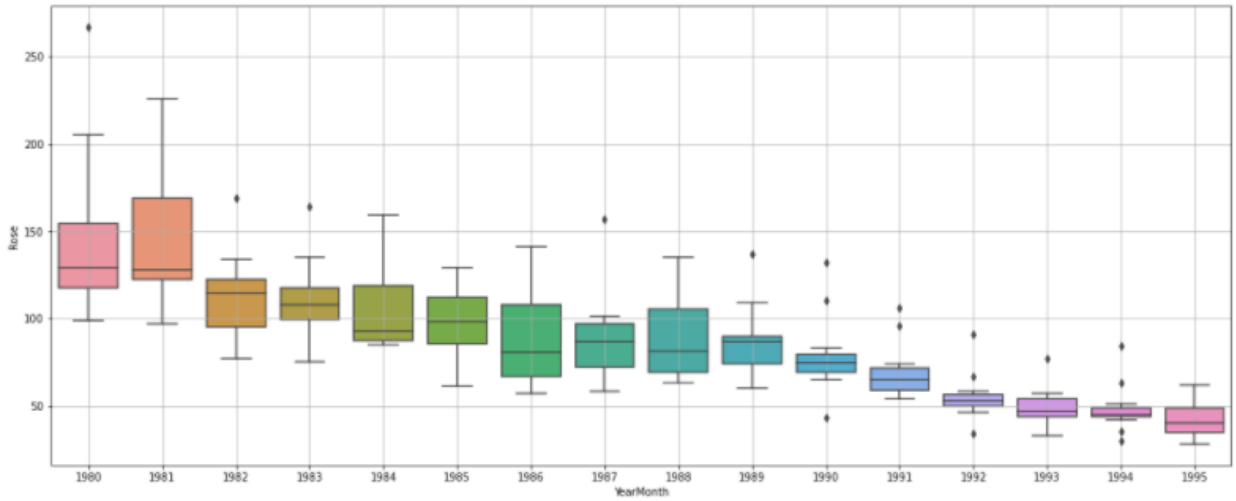

Fig.2

- The yearly boxplot shows that the average sales of rose wine is showing a downward trend over the years.
- The outliers in the yearly-boxplot most probably represents seasonal sales during the seasonal months.

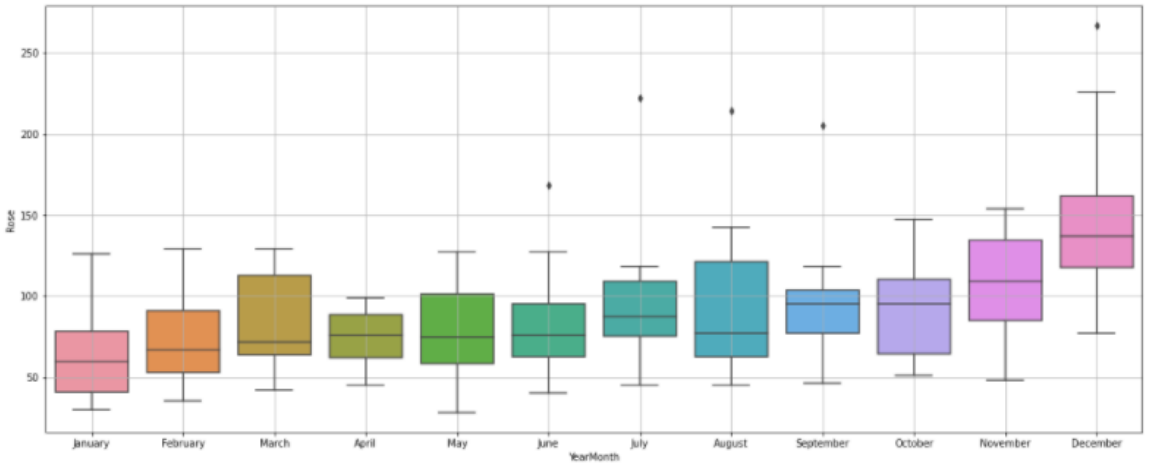

Fig.3

- The monthly boxplot show sales within different months spread across various years.
- The monthly boxplot shows clear seasonality during the months of November and December. Though sales drops drastically in month of January, it slowly picks up during the course of a year.
- The highest such numbers are being recorded in the month of December across various years.

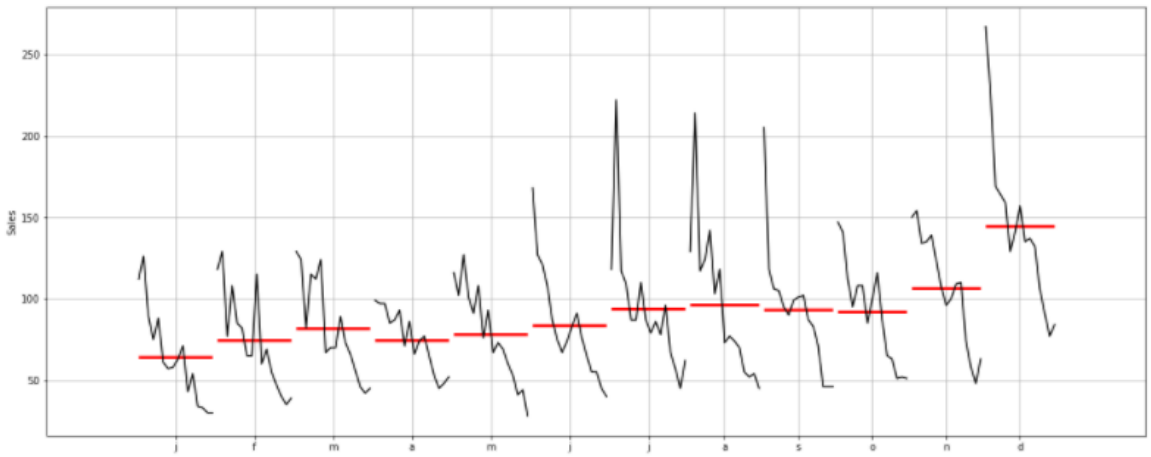

Fig.4

- This monthly plot shows mean and variation of units sold each month across various years.
- Sales in the seasonal months show higher variation.
- The red line is the mean sales value of each month across various years.
- Sales from Jan - Oct has been very slowly increasing over a particular the year.
- Dec month shows highest mean value representing highest sales amongst all the months.

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	45.0	45.0	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

Table.3

- Table of monthly sales across all the years.
- December 1980 has highest number of rose wine units sold 267 among all the years.
- Least sale in December month was in 1994 having only 84 units sold.

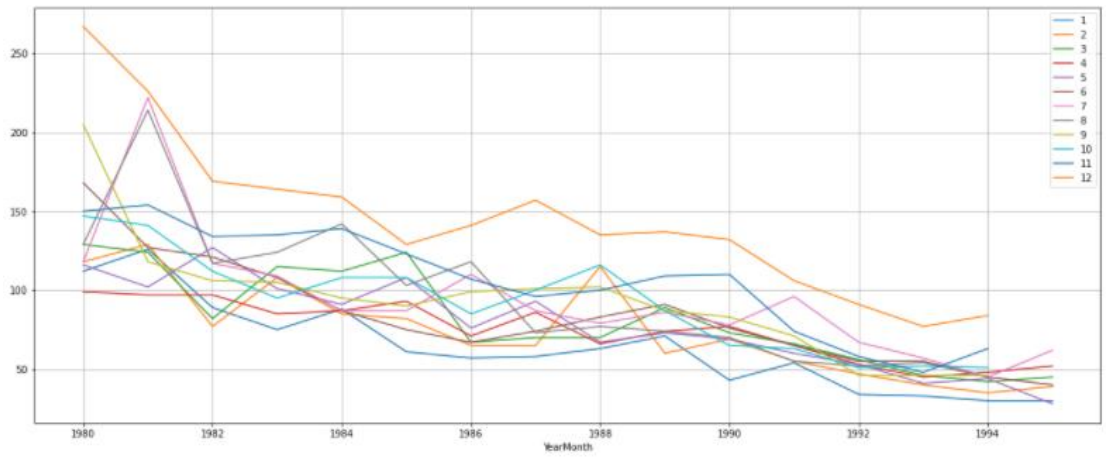

Fig.5

- Line chart of sales across various years. We can see December month has the highest number of sales of rose wine.
- Highest units sold in December was in 1980.
- Charts shows a decreasing trend in sale of the rose wine.

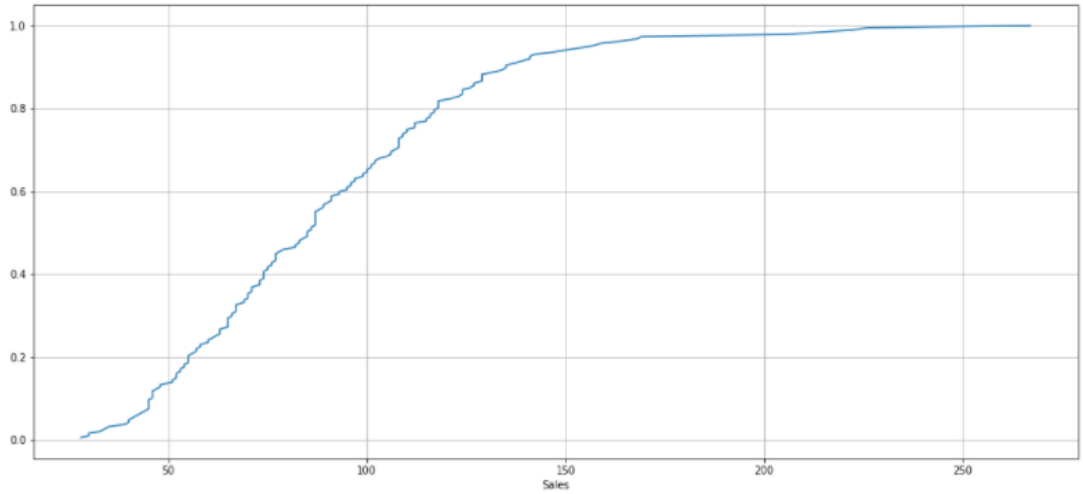

Fig.6

- This is an empirical cumulative distribution graph. This graph tells us what percentage of data points refer to what number of sales.
- 80% of the month have at least 120 units' sales of rose wine.

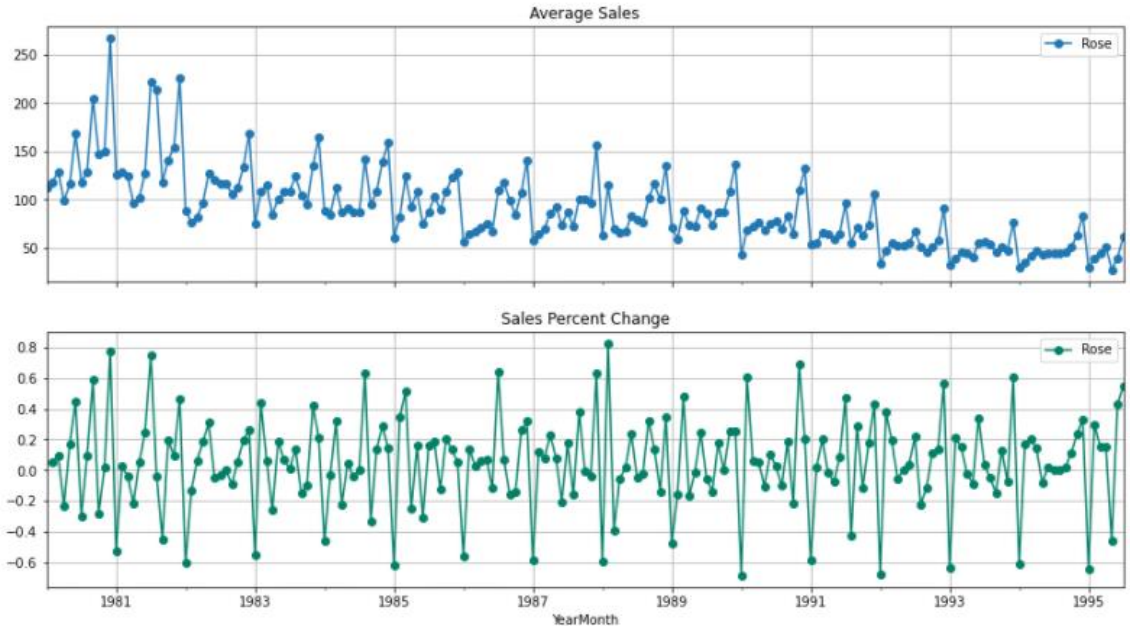

Fig.7

- The above two graphs tells us the Average 'Sales' and the Percentage change of 'Sales' with respect to the time.
- Average sales shows decreasing trend over the years with sales reaching below 50 units in the most recent years.

Time Series Decomposition:

- Additive Model

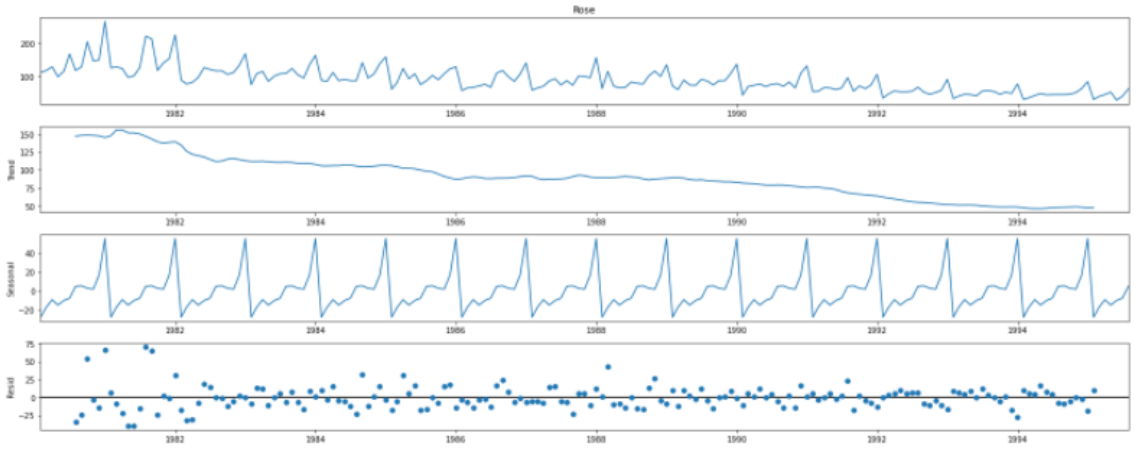

Fig.8

- Multiplicative Model

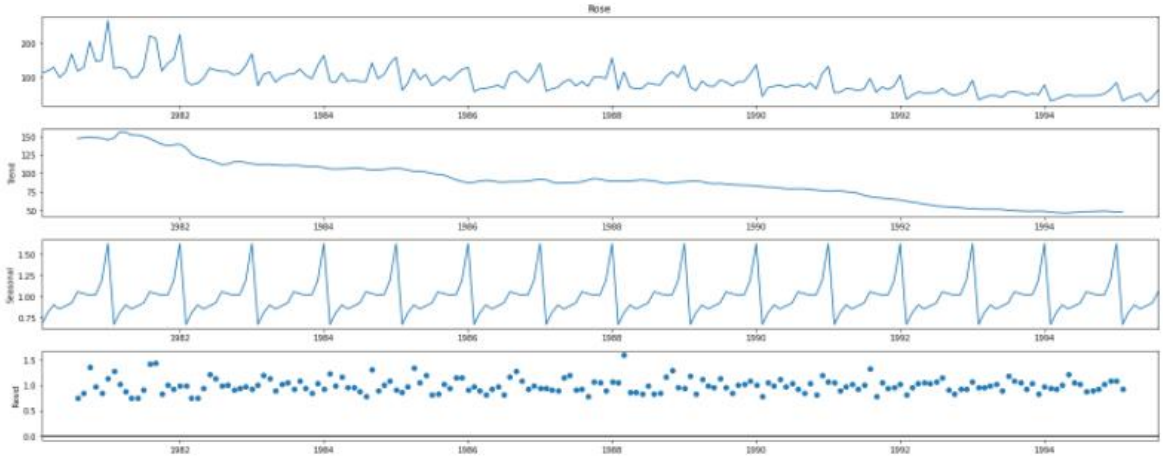

Fig.9

- As we can see from decomposition plot shows visible seasonality and a downward trend.
- The residual shows high variability across the period of time series, which is more or less consistent in both additive and multiplicative decompositions.
    - Residual mean of additive decomposition is : -0.08
    - Residual mean of multiplicative decomposition is : 0.99
- P-value of shapiro test for both models is also not significant.
- As the seasonality peaks are consistently reducing in altitude along with the trend, our time series can be treated as multiplicative in model building.

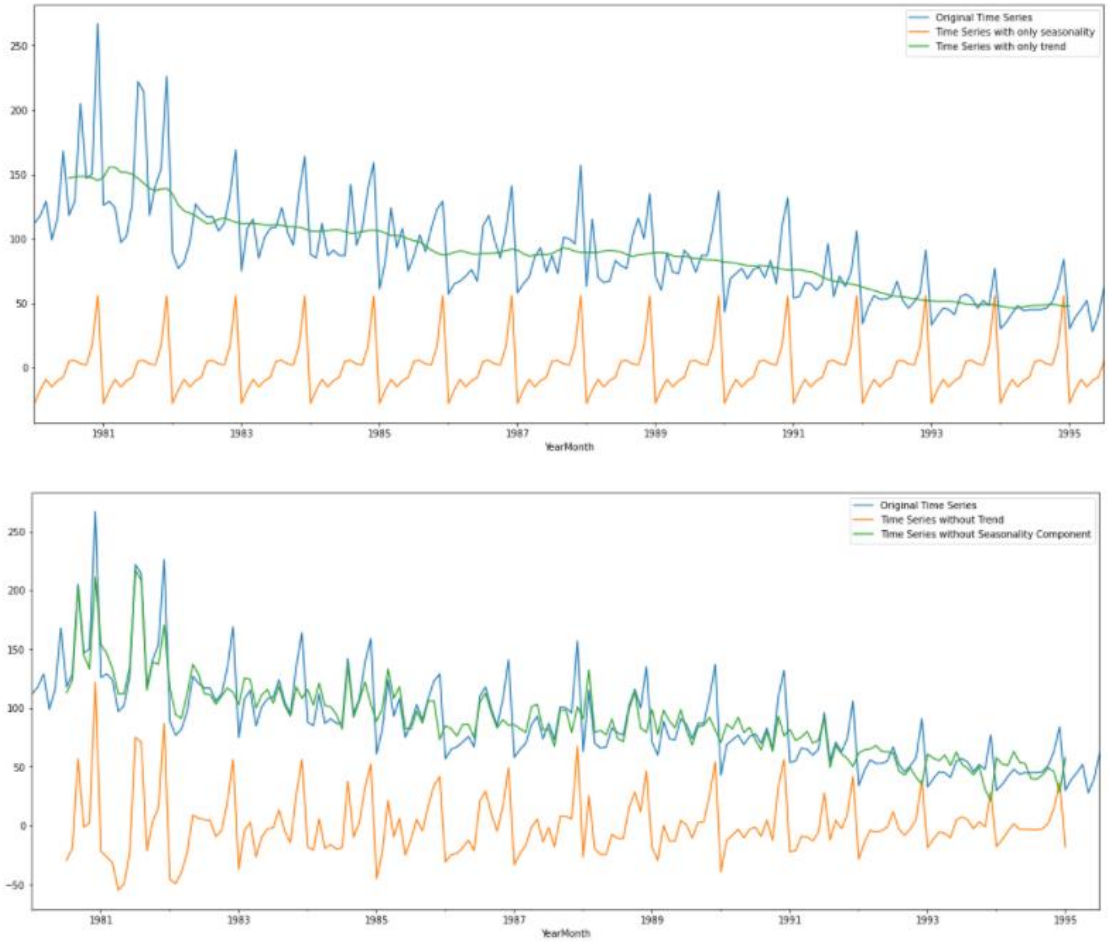

Fig.10

- The above graphs shows the original time series, Seasonality plot and trend plot.
- Second graph shows time series without trend and time series without seasonality.

1.3 Split the data into training and test. The test data should start in 1991. Read the data as an appropriate Time Series data and plot the data.

- The time series is split into train and test set with test set starting from year 1991.
- Train Set

YearMonth	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

Train Head

YearMonth	Rose
1990-08-31	70.0
1990-09-30	83.0
1990-10-31	65.0
1990-11-30	110.0
1990-12-31	132.0

Train Tail

- Test Set

YearMonth	Rose
1991-01-31	54.0
1991-02-28	55.0
1991-03-31	66.0
1991-04-30	65.0
1991-05-31	60.0

Test Head

YearMonth	Rose
1995-03-31	45.0
1995-04-30	52.0
1995-05-31	28.0
1995-06-30	40.0
1995-07-31	62.0

Test Tail

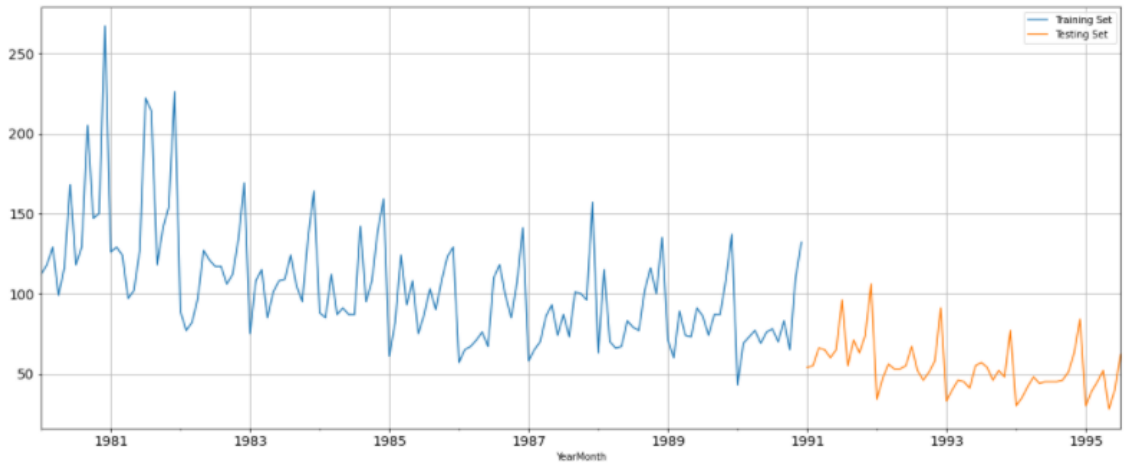

Fig.11

- Plot of train and test set for the rose time series.

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

Linear Regression on Time:

- To regress the sale of rose wine, numerical time instance needs to be added to train and test set respectively for regression algorithm to work on our time series.
- Model is trained on training set and RMSE evaluated on test set.
    - Test RMSE – 15.275

Naïve Forecasting:

- In naïve model, the value at the end of train set is applied to all the test set records. Prediction is made using these values with actual values in test set.
    - Test RMSE – 79.738
- Model is very poor fitted and does not capture neither trend nor seasonality present in the dataset.

Simple Average Forecasting:

- In simple average model, the forecast is done using the mean of the target variable of training set of the time series.
    - Test RMSE – 53.480
- Model is very poor fitted and does not capture neither trend nor seasonality present in the dataset.

Trailing Moving Average:

- For this model we will calculate the rolling means for different time intervals. The best interval can be determined by the least RMSE value.
- Trailing moving average models are built for 2, 4, 6 and 9 points moving averages.
    - Test RMSE - 2 Trailing Average – 11.529
    - Test RMSE - 4 Trailing Average – 14.455
    - Test RMSE - 6 Trailing Average – 14.572
    - Test RMSE - 9 Trailing Average – 14.731
- The best trailing interval for moving average is 2.

Simple Exponential Smoothing-Auto fit:

- Simple exponential smoothing is applied if the time series has neither trend nor seasonality, which is not the case for our dataset.
- For SES auto fit we are just passing the train set to the base model and evaluating on test set. We get best value for alpha – 0.098 from auto fit SES model.
    - Test RMSE – 36.816

Simple Exponential Smoothing Manual:

- Here we are manually finding the best value for alpha between 0 and 1.
- For different value of alpha, we are fitting it to the SES model and we are evaluating the test RMSE value. The best test RMSE value found is for 0.1 alpha.
    - Test RMSE – 36.848

Double Exponential Smoothing-Manual:

- Double exponential smoothing model is applied when the data has trend, but no seasonality, which is not the case for our dataset. Our dataset has downward trend and seasonality both.
- For DES model we trying to find the best value for both alpha and beta respectively. The best combination of both alpha and beta are chosen based on the test RMSE value. We get the best test RMSE value at alpha = 0.1 and beta = 0.1.
    - Test RMSE – 36.900

Triple Exponential Smoothing-Auto fit:

- Triple exponential smoothing model is applied when data has both trend and seasonality. Our wine time series has both trend and seasonality present.
- For auto fit TES model we are just passing the train set to the base model and evaluating on the test set.
- We get the best value for alpha = 0.06, beta=0.05, gamma=0.0 from the auto fit model.
    - Test RMSE – 21.045

Triple Exponential Smoothing Manual:

- For TES manual model we try to find the best value of alpha, beta and gamma from range between 0.1 -1 for each respectively. Trying to find the best combination of value and testing on test set to find the least RMSE value.
- We see that we get the least value for alpha=0.1, beta=0.2 and gamma=0.2.
    - Test RMSE – 9.647

Model Comparison:

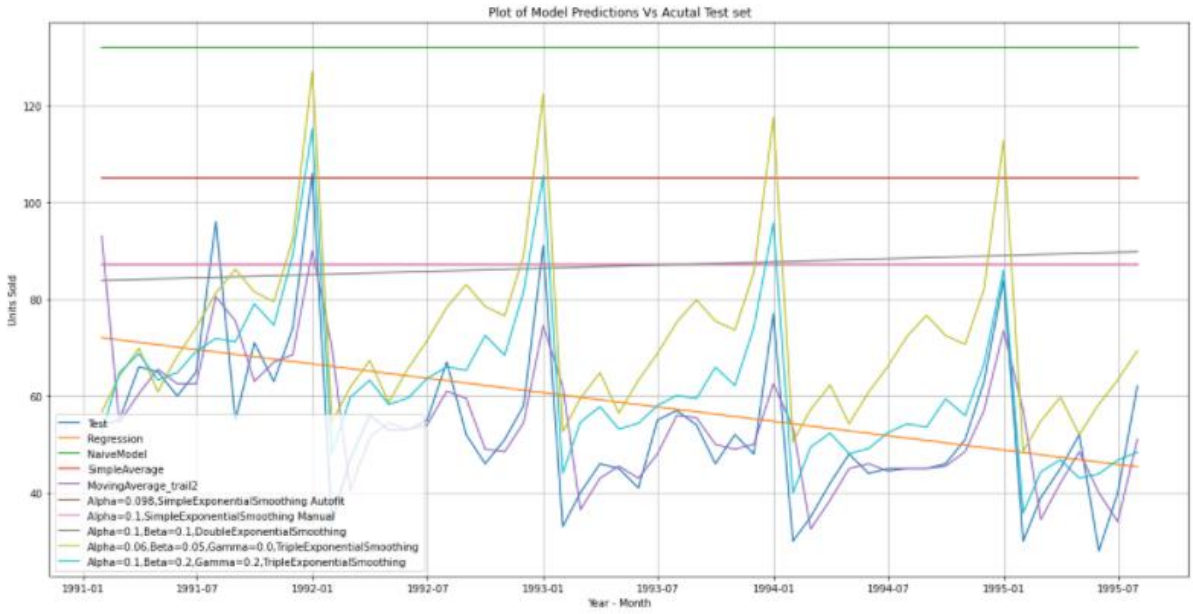

Fig.12

- We have plotted the test data against predictions from all the models.
- Out of all the models we can see that the best model is TES manual model. Test RMSE value for the TES manual model is also least among all the models built so far.

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

- Augmented Dickey Fuller test is the statistical test to check the stationarity of a time series. The test determine the presence of unit root in the series to understand if the series is stationary or not.

- For checking the stationary of the series we have the following hypothesis as below:
  H0: The series has a unit root, hence it is not stationary.
  H1: The series has no unit root, hence it is stationary.

- Now to test our hypothesis we perform the Dickey Fuller Test on the whole data.
  Output:
  DF test statistic is -2.241
  DF test p-value is 0.46694206026101637
  Number of lags used 13

  Since p-value is significant (greater than 0.05) we fail to reject the null hypothesis. So data is not stationary.

- Taking differencing of order 1 on the whole series and again testing for stationarity.
  Output:
  DF test statistic is -8.161
  DF test p-value is 3.028272263687806e-11
  Number of lags used 12

  Since the p-value is low less than alpha=0.05 we reject the null hypothesis. So for differencing of order 1 we get a stationary time series.

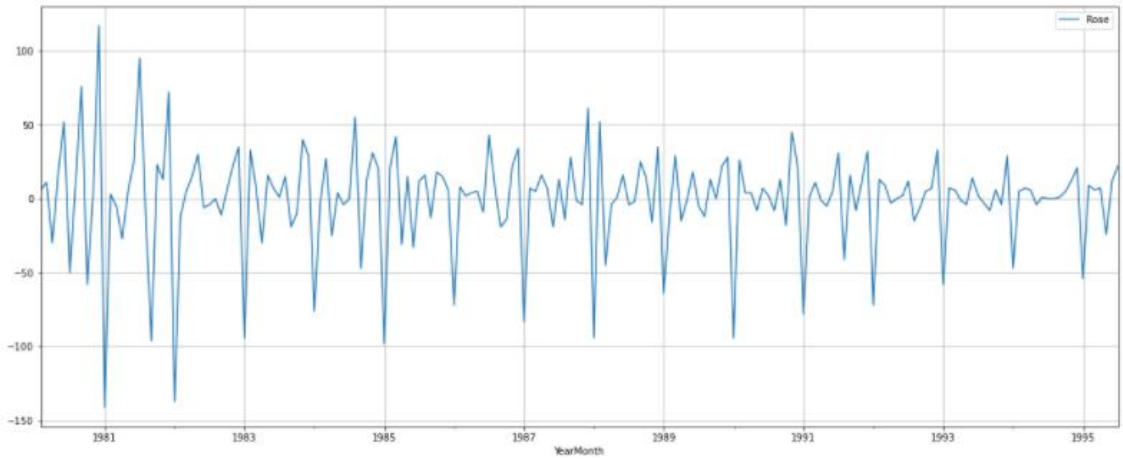

Fig.13

- We can see that the data has trend and seasonality component present in it but the series is now stationary around 0.
- We also check for stationarity of the train set and get the value of d=1.

1.6 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:

- We build an automated ARIMA model on the dataset taking order of differencing as 1 and p and q value from range 0-4. Finding all combination of values of (p,d,q) and fitting in the ARIMA model to find the least AIC score.
- We get least AIC value for ARIMA(2,1,3).
    - ❖ AIC value for ARIMA(2,1,3) – 1274.694

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.347
Date:	Thu, 30 Dec 2021	AIC	1274.695
Time:	12:13:45	BIC	1291.946
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.705
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6781	0.084	-20.035	0.000	-1.842	-1.514
ar.L2	-0.7289	0.084	-8.703	0.000	-0.893	-0.565
ma.L1	1.0450	0.685	1.527	0.127	-0.297	2.387
ma.L2	-0.7716	0.137	-5.636	0.000	-1.040	-0.503
ma.L3	-0.9046	0.622	-1.455	0.146	-2.123	0.314
sigma2	858.3595	576.845	1.488	0.137	-272.237	1988.956

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.45
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table.4

- From the model summary it can be inferred that all MA lag 1 and MA lag 2 terms are not significant, as their values are above 0.05.

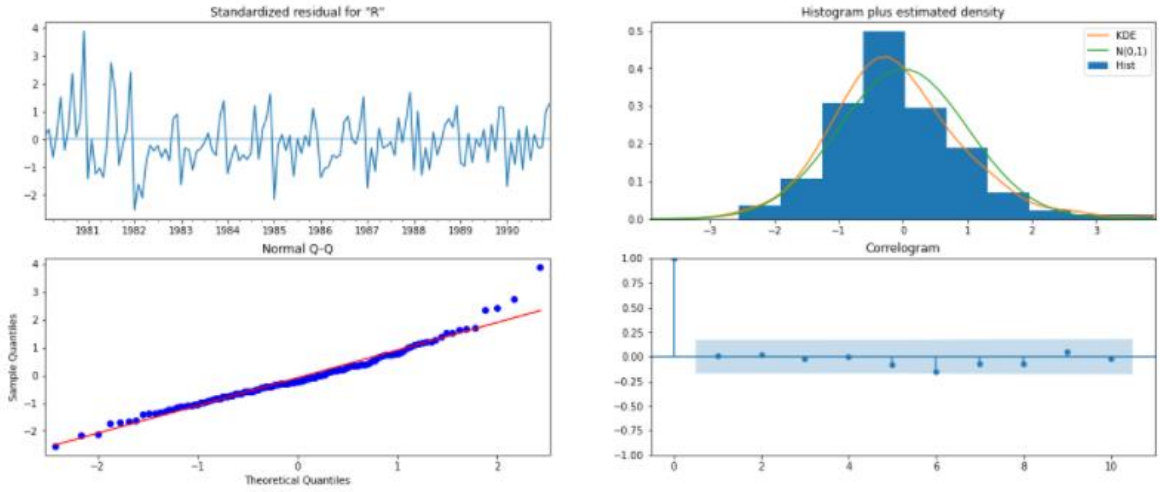

Fig.14 Diagnostics plot – ARIMA(2,1,3)

- The diagnostics plot of model was derived and the standardized residuals are found to follow a mean of zero, and the histogram shows the residuals follow a normal distribution.
- Test RMSE for ARIMA(2,1,3) – 36.838

SARIMA:

- We build an automated SARIMA model on the dataset taking order of differencing as 1 and P,p,Q,q value from range 0-4. We have taking seasonality value as 12, as data is of monthly period in the entire series and D=0. Finding all combination of values of (p,d,q)(P,D,Q,S) and fitting in the SARIMA model to find the least AIC score.
- The value of (p,d,q)(P,D,Q,S) for which SARIMA model is stable is (3,1,1)(3,0,2,12).
- We get the least AIC values for the above values and evaluate the model with these values on test set.
    - AIC value for SARIMA (3,1,1)(3,0,2,12) – 774.400

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)	Log Likelihood	-377.200
Date:	Fri, 14 Jan 2022	AIC	774.400
Time:	15:56:29	BIC	799.618
Sample:	0	HQIC	784.578
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0464	0.126	0.367	0.714	-0.202	0.294
ar.L2	-0.0060	0.120	-0.050	0.960	-0.241	0.229
ar.L3	-0.1808	0.098	-1.837	0.066	-0.374	0.012
ma.L1	-0.9370	0.067	-13.904	0.000	-1.069	-0.805
ar.S.L12	0.7639	0.165	4.639	0.000	0.441	1.087
ar.S.L24	0.0840	0.159	0.527	0.598	-0.229	0.397
ar.S.L36	0.0727	0.095	0.764	0.445	-0.114	0.259
ma.S.L12	-0.4968	0.250	-1.988	0.047	-0.987	-0.007
ma.S.L24	-0.2191	0.210	-1.044	0.296	-0.630	0.192
sigma2	192.1613	39.630	4.849	0.000	114.487	269.835

Ljung-Box (L1) (Q):	0.30	Jarque-Bera (JB):	1.64
Prob(Q):	0.58	Prob(JB):	0.44
Heteroskedasticity (H):	1.11	Skew:	0.33
Prob(H) (two-sided):	0.77	Kurtosis:	3.03

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table.5

- From the model summary it can be inferred that most AR and MA terms are not significant.

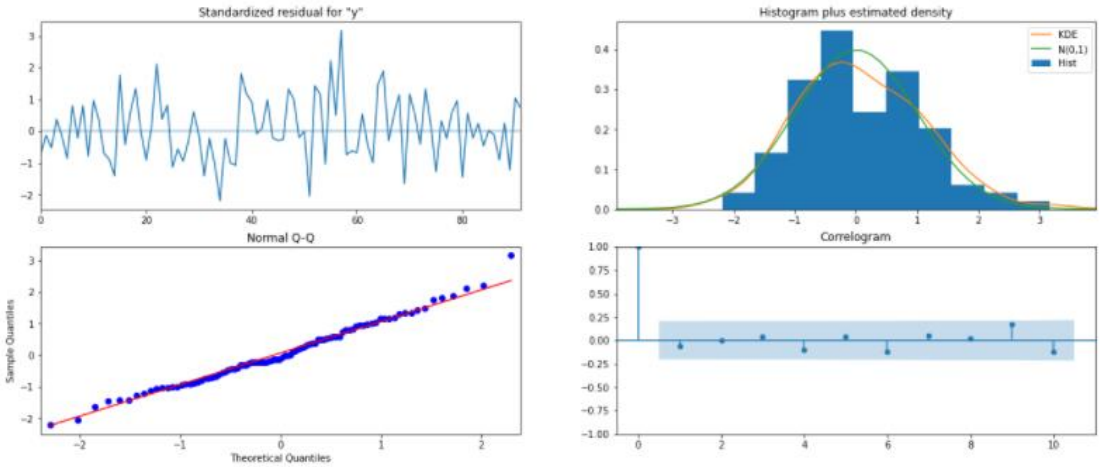

Fig.15 Diagnostics plot – SARIMA(3,1,1)(3,0,2,12)

- Inference from model diagnostics confirms that the model residuals are normally distributed.
- **Standardised residual** - Do not display any obvious seasonality.
- **Histogram plus estimated density** –The KDE plot of the residuals is similar with the normal distribution, hence the model residuals are normally distributed.
- **Normal Q-Q plot** - There is an ordered distribution of residuals(blue dots) following the linear trend of the samples taken from the standard normal distribution with N(0,1)
- **Correlogram** - The time series residuals have low correlation with lagged versions of itself.
- Test RMSE for SARIMA(3,1,1)(3,0,2,12) – 18.903

1.7 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.

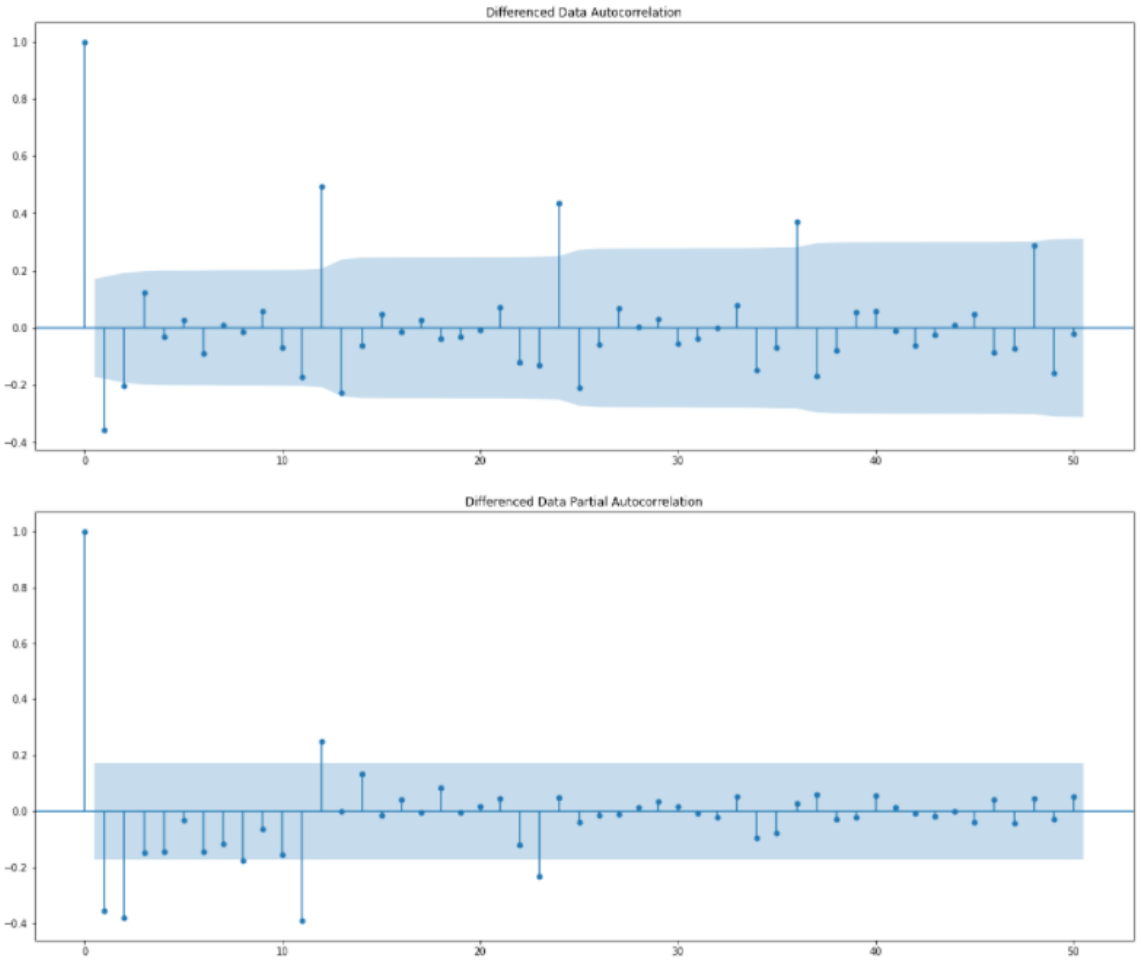

Fig.16

- While reading the PACF and ACF plot:
  We always look at the positive side.
  We exclude 1 lag as it represents the series itself.
  Look at consecutive bars that exceed the threshold to find the values from graph.

ARIMA Model Using PACF and ACF:

- Best parameters are selected by looking at the ACF and the PACF plot
- 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 0.
  The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
  By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 0.
  Difference order is of 1.

- ARIMA Model using p, q = 0 and d=1.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(0, 1, 0)	Log Likelihood	-665.577
Date:	Sat, 15 Jan 2022	AIC	1333.155
Time:	20:12:11	BIC	1336.030
Sample:	01-31-1980 - 12-31-1990	HQIC	1334.323
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1515.6738	122.418	12.381	0.000	1275.740	1755.608

Ljung-Box (L1) (Q):	17.11	Jarque-Bera (JB):	59.55
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.38	Skew:	-0.95
Prob(H) (two-sided):	0.00	Kurtosis:	5.70

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table.6

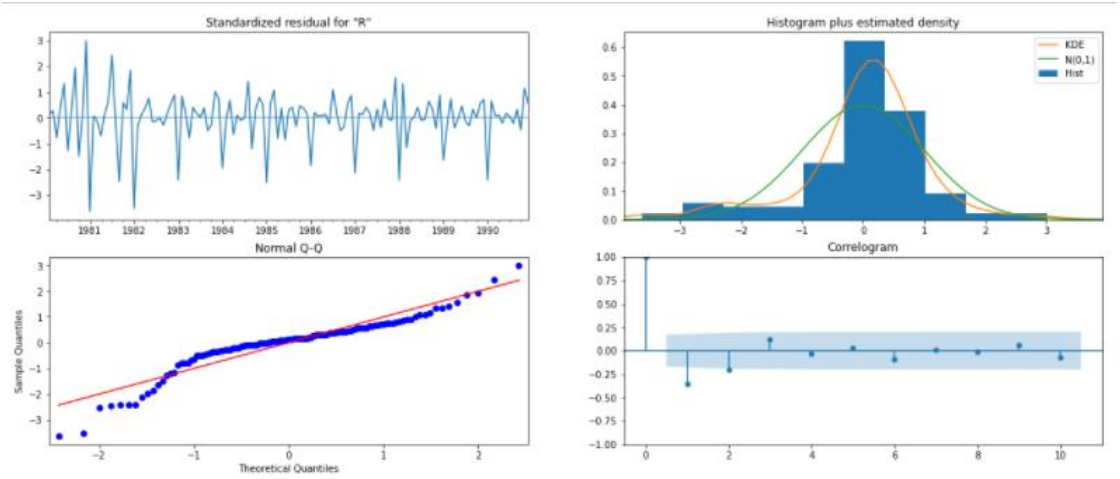

Fig.17

- Since we know that our time series has both trend and seasonality components ARIMA model thus built using p and q = 0 is not correct at all.
- Test RMSE for ARIMA(0,1,0) – 79.738
- As we can see from the test RMSE also the model is performing very poorly.

SARIMA Model Using PACF and ACF:

- Here, we have taken alpha=0.05.

  We are going to take the seasonal period as 12. We will keep the p = 0, q = 0 and d=1 parameters same as the ARIMA model.
  The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.
  The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0.
  We have checked the ACF and the PACF plots only at multiples of 12 (since 12 is the seasonal period).

- SARIMA Model using p,d,q = 0,1,0 and P,D,Q = 0,0,0,12.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 0)	Log Likelihood	-660.983
Date:	Sat, 15 Jan 2022	AIC	1323.966
Time:	20:15:08	BIC	1326.833
Sample:	0	HQIC	1325.131
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1527.0558	124.361	12.279	0.000	1283.314	1770.798

Ljung-Box (L1) (Q):	17.02	Jarque-Bera (JB):	57.49
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.38	Skew:	-0.95
Prob(H) (two-sided):	0.00	Kurtosis:	5.65

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table.7

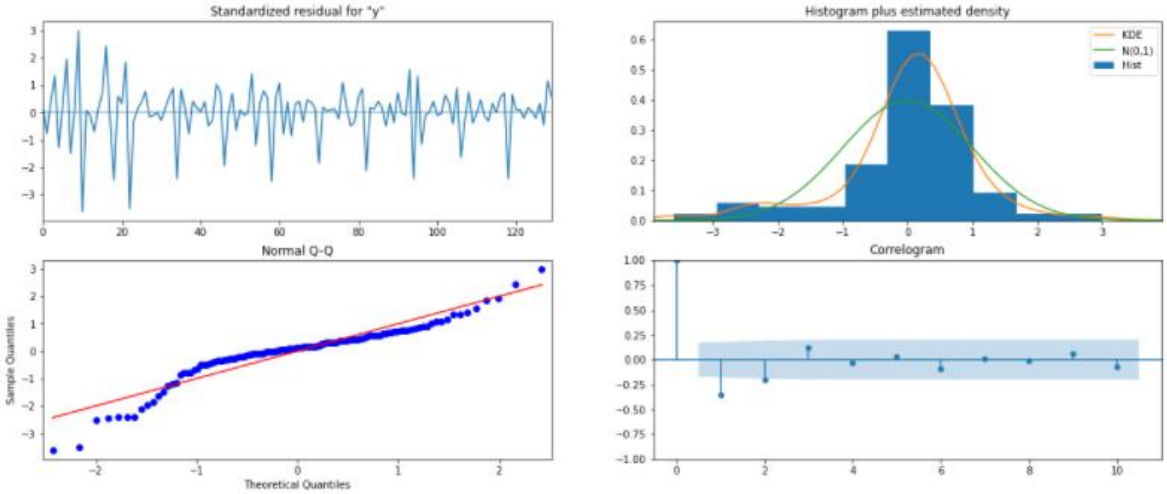

Fig.18

- Again as the value of p,q and P,Q has been taken from the ACF and PACF plots our model is not generalising well and performing very poorly for the SARIMA model as well.
- Test RMSE for SARIMA(0,1,0)(0,0,0,12) – 79.738

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

	Test RMSE
Alpha=0.1,Beta=0.2,Gamma=0.2,TripleExponentialSmoothing	9.647756
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.455221
6pointTrailingMovingAverage	14.572009
9pointTrailingMovingAverage	14.731209
RegressionOnTime	15.275732
pdq=(3,1,1),PDQS=(3,1,1,12),SARIMA Automated	18.903315
Alpha=0.06,Beta=0.05,Gamma=0.0,TripleExponentialSmoothing-Autofit	21.045505
Alpha=0.098,SimpleExponentialSmoothing-Autofit	36.816889
Alpha=2,Beta=1,Gamma=3,ARIMA Automated	36.838008
Alpha=0.1,SimpleExponentialSmoothing	36.848694
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	36.900871
Simple Average	53.480857
Naive Model	79.738550
Alpha=0,Beta=1,Gamma=0,ARIMA Manual	79.738550
pdq=(0,1,0),PDQS(0,0,0,12),SARIMA Manual	79.738550

Table.8

- The above table provides test RMSE value for all the models that we have performed so far.
- We have arranged the table in ascending order of Test RMSE values.
- Model with lowest Test RMSE value will help in predicting the future sales of rose wine much better.
- The best model as per the table is Triple Exponential Smoothing manual model. Having the least RMSE value of 9.64.

1.9 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.

- Now we built our best model that is Triple Exponential Smoothing on the complete dataset using the best parameters values we found earlier.
- Parameter values used for Triple Exponential Smoothing are:
  Alpha: 0.1
  Beta: 0.2
  Gamma: 0.2
  Optimized: True
  Use_brute: True
- Now using the above model we predict for 12 month sales figure of 'rose' wine in the future.

```
1995-08-31    47.427154
1995-09-30    48.269176
1995-10-31    50.273106
1995-11-30    58.463029
1995-12-31    82.124581
1996-01-31    31.701252
1996-02-29    39.442812
1996-03-31    45.376744
1996-04-30    46.826433
1996-05-31    40.738726
1996-06-30    47.021406
1996-07-31    53.990737
Freq: M, dtype: float64
```

Table.9

- Plotting our time series against the values found out by our model for the complete data and our prediction for 12 months in the future.
- We can visually see that our model is able to mimic the time series pattern quite reasonably well.

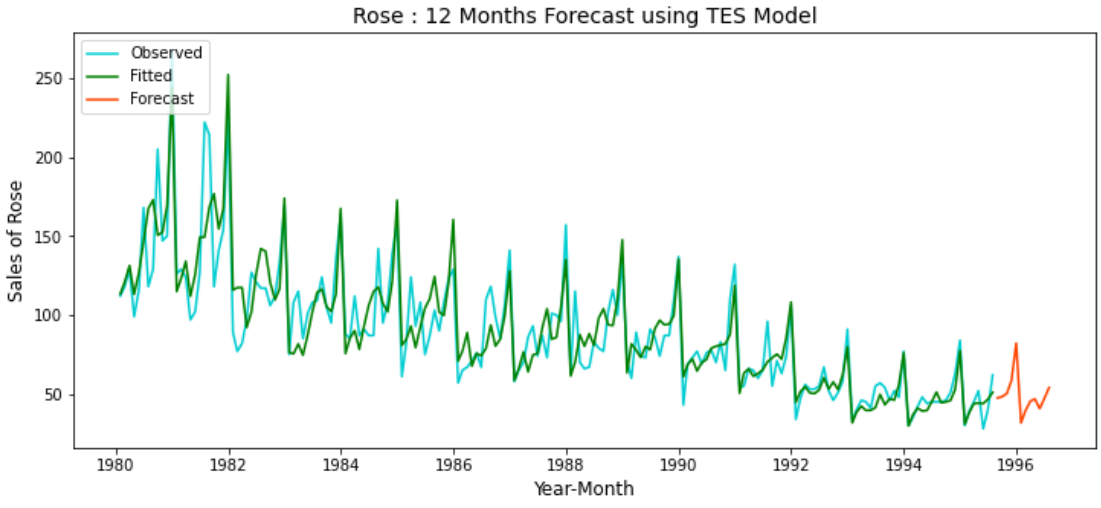

Fig.19

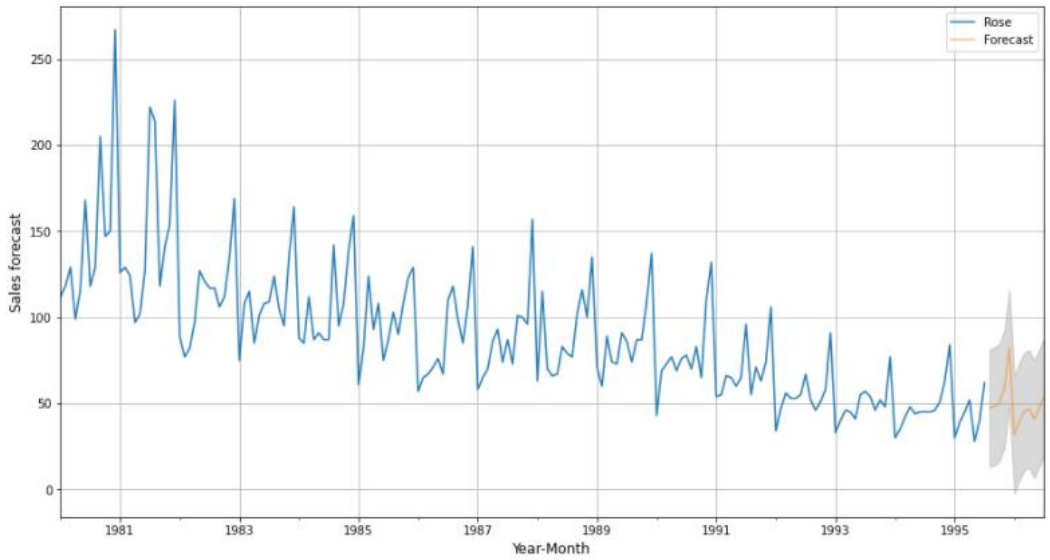

Fig.20

- Plot original data and of our prediction for 12 month sale in the future against confidence interval of 95%.

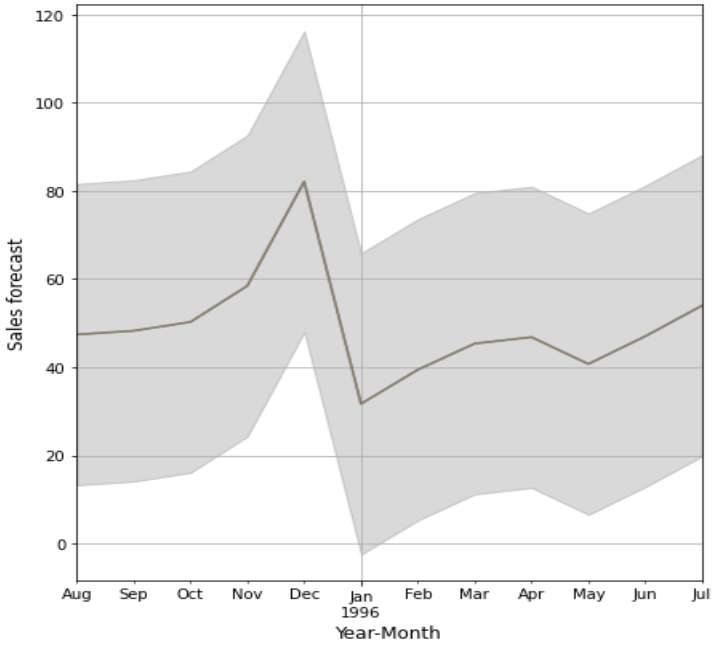

Fig.21

- Plot of our forecast along with the confidence interval band of 95%.

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

- The model forecast sale of 592 units of rose wine in 12 months in the future with average sale of 49 units per month.
- Highest sales is predicted to occur in Dec-1995 with unit's sales hitting maximum of 82 units.
- Rose sales shows a decrease in trend compared to the previous years.
- November and December month shows the highest Sales across the years from 1980-1995.
- The models are built considering the Trend and Seasonality into account and we see from the output plot that the future prediction is in line with the trend and seasonality in the previous years.
- The Sales of Rose wine is seasonal, hence the company cannot have the same stock through the year. The predictions would help here to plan the Stock need basis the forecasted sales.
- The company should use the prediction results and capitalize on the high demand seasons and ensure to source and supply the high demand.
- The company should use the prediction results to plan the low demand seasons to stock as per the demand.
- The forecast also indicates that year-on-year the sales of rose wine is declining. The wine company should investigate the low demand of rose wine and must adopt innovative marketing skills to stop the sales from decreasing further.

THE END