Business Report on Time Series Analysis of sales of Rose Wine

Objective:- To forecast the approximate sales number for 12 months into the future basis the past data provided for 187 months for Rose wine.

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

Solution: - We have started the time series analysis of the given dataset by importing the usual libraries with an additional library for the decomposition of the time series data.

- The data talks about the sales number for a particular given month in a year.
- We checked the data types of the columns of the dataset and found that "YearMonth" column is of Object data type and column "Rose" is of float data type. Hereby we need to instruct python that we are reading a time series data.

```
YearMonth     object
Rose         float64
dtype: object
```

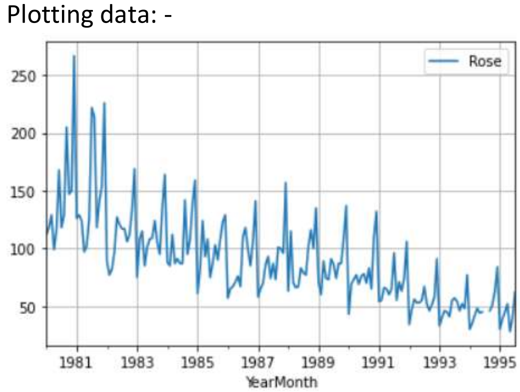
- We need to parse the data to make python understand that we are working in time series data and now we can observe that the "YearMonth" column has been identified as datetime data type. (Note: ns stands for nano seconds)

```
YearMonth    datetime64[ns]
Rose                float64
dtype: object
```

- It is also recommended that for all time series analysis we should mostly put the time series reference column as the index. It makes it easy while slicing and dicing the data. Which can be achieved by passing an additional function known as "index_col". And now we see that the "YearMonth" variable as set as index now.

	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

- Checking the shape of data and we can find that there are 187 observations and 1 target variable.
- Plotting data: -



- From the above plot we can visually conclude that the data have presence of both trend and seasonality.
- Checking for null values we can find 2 missing data in the dataset.
- **The missing values needs treatment we just cannot delete as it will create a hole / gap in a continuous time series data.**
- There are various ways of treating/imputing missing values time series analysis and one of the ways is imputation via **Interpolation**. We are using this method as we understand that the data have both trend and seasonality as shown in graph/plot above.
- Describing data before and after missing value imputation and we can observe the changes in statistical data once the missing values are imputed.

Describing data before missing value inputaion

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

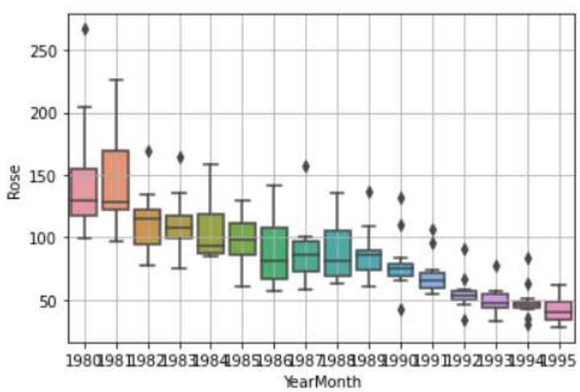
Describing data after missing value imputaion

	Rose
count	187.000000
mean	89.927087
std	39.224153
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

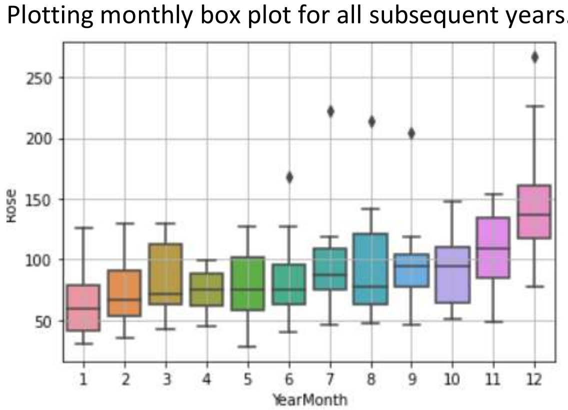
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Solution: - Performing exploratory data analysis: -

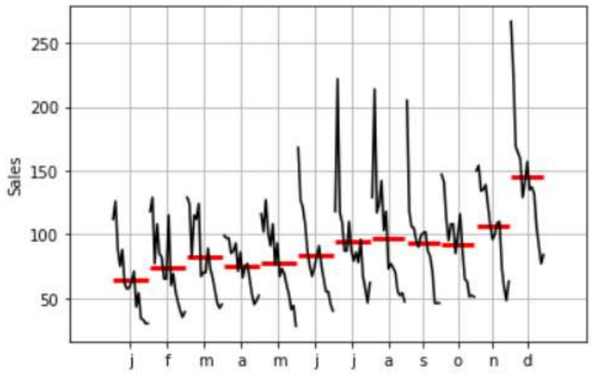
- Plotting yearly box plot to check on the sales distribution and trends across years and it also gives a glimpse of trend and outliers in the dataset.



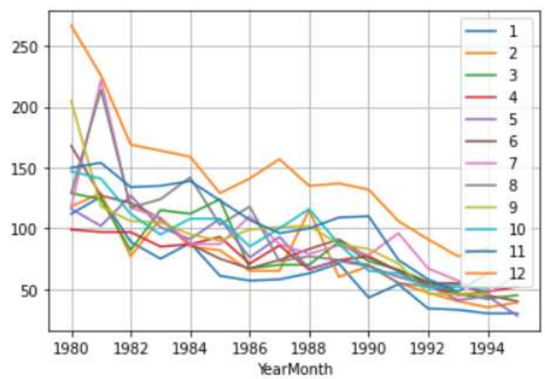
- Plotting monthly box plot for all subsequent years.



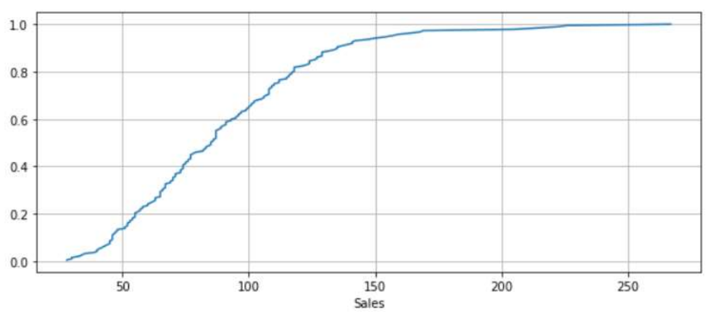
- Plot a time series month plot to understand the spread of accidents across different years and within different months across years.



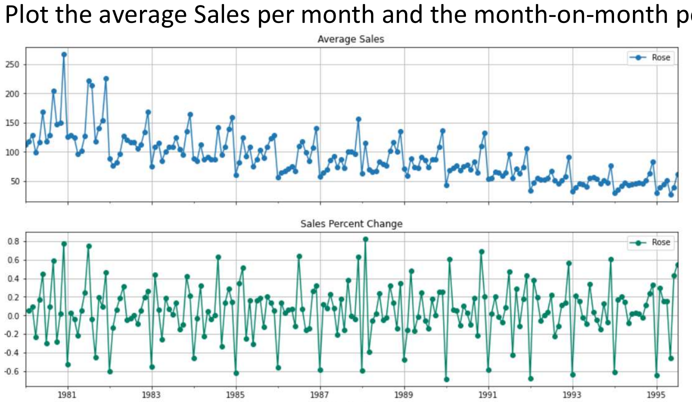
- Plot a graph of monthly Sales across years.



- We can notice that the sales are at decent numbers at starting of every year however it shows decreasing pattern at the end of the year.
- Plotting the Empirical Cumulative Distribution



- Plot the average Sales per month and the month-on-month percentage change of Sales.



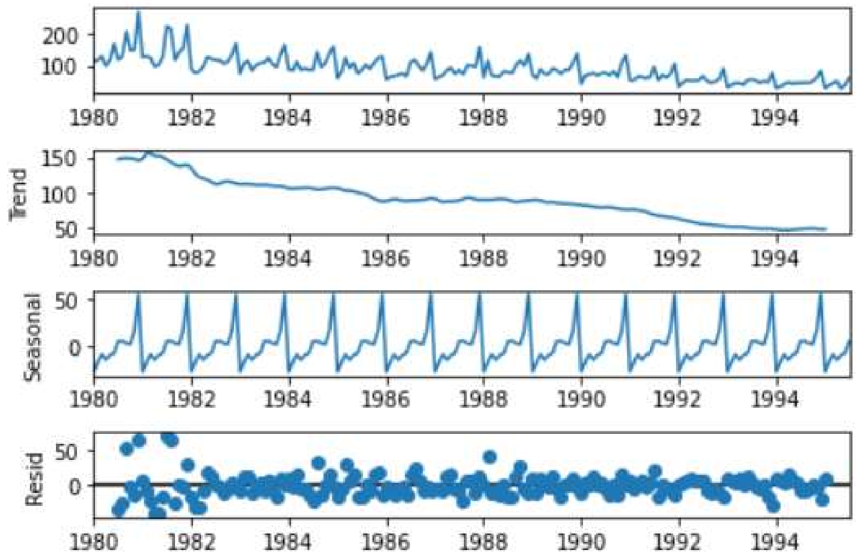
<u>Performing Decomposition of Data: -</u>

From the above plot we can see that we have a time series data which is not a constant time series. It has a decreasing trend, so the slope of the trend is negative.

It also seems to have a repetitive nature which is a repeatable pattern every 2 years. Which is known as seasonality. But it does not seem to be a constant seasonality. The peaks are repeated but peaks are decreasing as we move along the years. And we can a certain it even better when we decompose this time series data.

1. **<u>Additive Decomposition</u>**

- Below is the plot of the decomposed time series using additive method



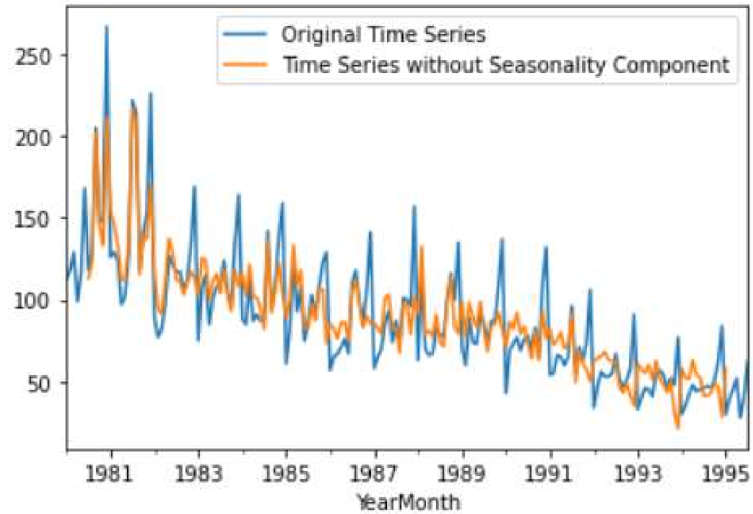
- The above shows the plot of original data and then the plot of trend. It has a decreasing trend. It has a seasonality component.
- We also have some residual/ error component available in data and it seem to be showing some pattern.
- Inspecting the trend, seasonal and residual elements of data: -

```
Trend
 YearMonth
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    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64
```

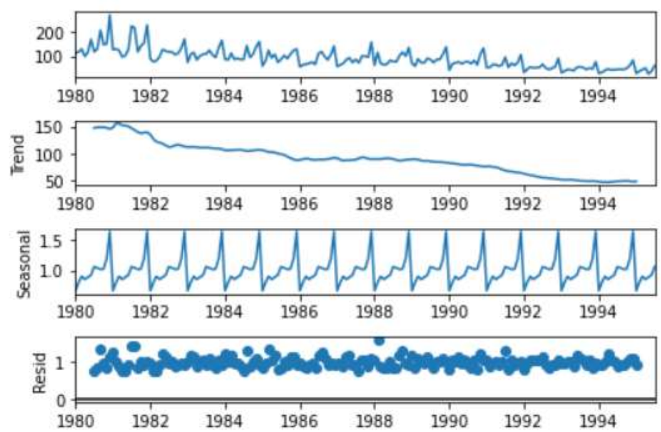
```
Seasonality
 YearMonth
1980-01-01    -27.921780
1980-02-01    -17.445103
1980-03-01     -9.299901
1980-04-01    -15.112401
1980-05-01    -10.210615
1980-06-01     -7.692758
1980-07-01      4.938434
1980-08-01      5.589575
1980-09-01      2.761554
1980-10-01      1.858776
1980-11-01     16.833776
1980-12-01     55.700443
Name: seasonal, dtype: float64
```

```
Residual
 YearMonth
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   -34.021767
1980-08-01   -24.714575
1980-09-01    53.863446
1980-10-01    -2.942109
1980-11-01   -14.250443
1980-12-01    66.174557
Name: resid, dtype: float64
```

- Plotting graph with and without seasonality



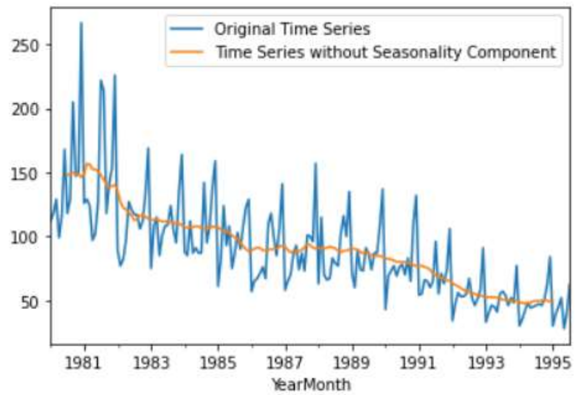
2. **Multiplicative Decomposition**
   - Below is the plot of the decomposed time series using multiplicative method



- We can notice that between the additive and multiplicative plot the scales have changed.
- We can see the error component is also flat and near to "1".
- So, we can conclude that for the given dataset multiplicative decomposition works well.
- Inspecting the trend, seasonality and residual components of data

```
Trend                              Seasonality                        Residual
 YearMonth                          YearMonth                          YearMonth
1980-01-01           NaN           1980-01-01     0.669946           1980-01-01         NaN
1980-02-01           NaN           1980-02-01     0.806019           1980-02-01         NaN
1980-03-01           NaN           1980-03-01     0.900899           1980-03-01         NaN
1980-04-01           NaN           1980-04-01     0.853719           1980-04-01         NaN
1980-05-01           NaN           1980-05-01     0.889143           1980-05-01         NaN
1980-06-01           NaN           1980-06-01     0.923718           1980-06-01         NaN
1980-07-01    147.083333           1980-07-01     1.058920           1980-07-01    0.757627
1980-08-01    148.125000           1980-08-01     1.037754           1980-08-01    0.839203
1980-09-01    148.375000           1980-09-01     1.017402           1980-09-01    1.358003
1980-10-01    148.083333           1980-10-01     1.022303           1980-10-01    0.971028
1980-11-01    147.416667           1980-11-01     1.192007           1980-11-01    0.853623
1980-12-01    145.125000           1980-12-01     1.628173           1980-12-01    1.129974
Name: trend, dtype: float64        Name: seasonal, dtype: float64     Name: resid, dtype: float64
```

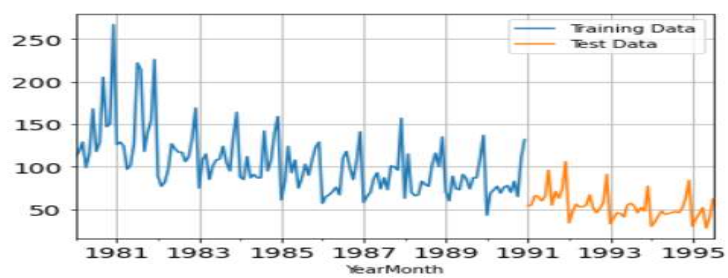
➢ Plotting graph with and without seasonality



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

Solution: - splitting data into test and train dataset in such a way that the train data has all the observations before 1991 and in test data observations start from 1991.After splitting the data and checking the shape we find that the train data have 132 observations and test dataset have 55 observations.

Below is the plot of train and test dataset: -



We understand that the train-test split cannot be done randomly as we are dealing with continuous data and time series has to be in continuous manner.

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.

Solution: - Solution: - Building various exponential smoothing models: -

To build various exponential smoothing model we have to start with importing ExponentialSmoothing library from statsmodels and also mean_sqaured_error library to compare the various models built which in-turn helps in selecting the best optimal model. The model with least Mean squared error will be considered the best optimal model and that model denoted less error components.

Here, we are about to build 4 models, 1st will be only with level (alpha), 2nd with level and trend component (alpha and beta), 3rd with Level, trend and additive seasonality component and 4th with level, trend and multiplicative seasonality component.

❖ **<u>Building Simple Exponential smoothing model: -</u>**

This method involves only the "Level" component of data (alpha) with no trend and no seasonality. which mean it is best suitable for data with no clear trend and seasonality. The value of alpha lied between 0 and 1. This model uses only single exponential component so also called as "Single Exponential Smoothing".

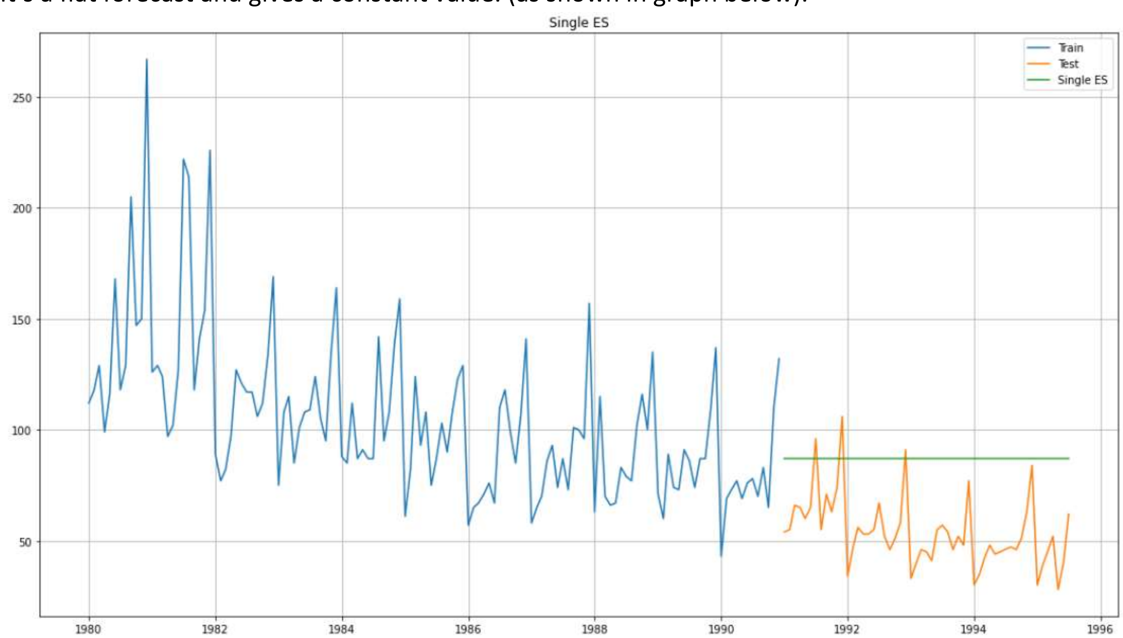
Here we are building model on train data and predicting on testing data and to check the accuracy of the built model we are using RMSE as the parameter.

<u>Below are the parameter details for the model: -</u>

```
{'smoothing_level': 0.09874933517484011,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38703609891138,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

We can see that the value is close to "0" (zero), which means that the previous time series data not that that accurately related to the forecast for the next period.

It's a flat forecast and gives a constant value. (as shown in graph below).



BY looking at the graph we can conclude that this is not the right model as the forecast is flat and doesn't acknowledge the element of trend and seasonality in the data.

<u>Below is the RMSE value of this model: -</u>

```
SES RMSE: 36.74838945471327
SES RMSE (calculated using statsmodels): 36.74838945471326
```

- **Building Double Exponential smoothing model: -**

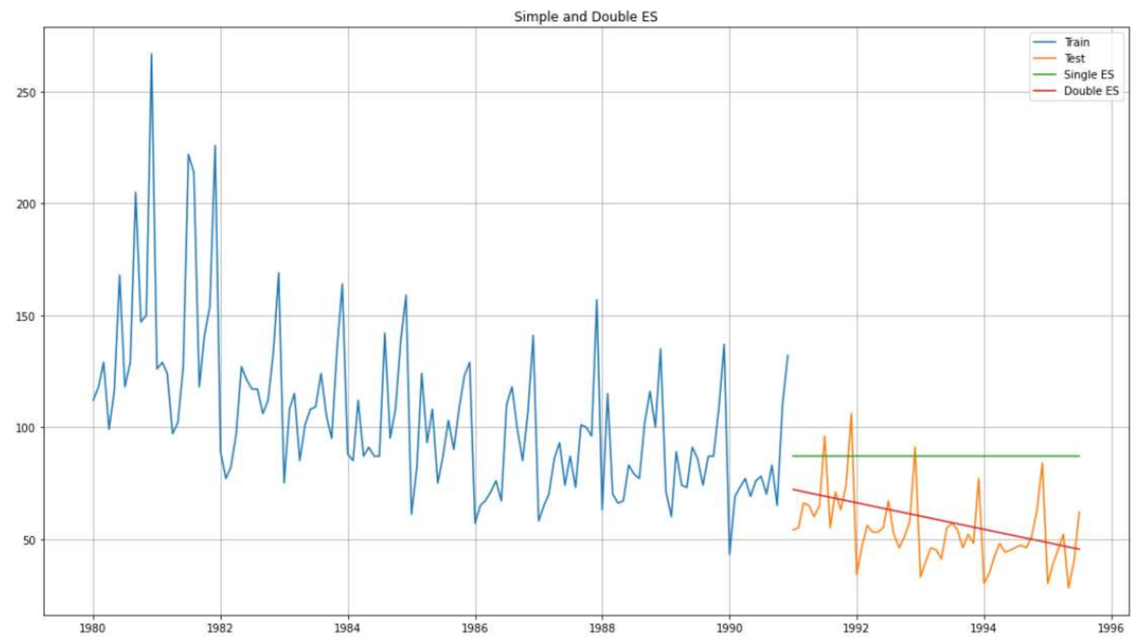
This method involves the "Level" component of data (alpha) as well as the "Trend" component (beta) with no seasonality. which mean it is best suitable for data with clear trend but no clear seasonality. This is also called as "Holts Linear method".

Below are the parameter details for the model: -

```
==Holt model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 1.9086427682180844e-08, 'smoothing_trend': 7.302464353829351e-09, 'smoothing_seasonal': nan, 'damping_tren
d': nan, 'initial_level': 137.81629861505857, 'initial_trend': -0.4943753249082896, 'initial_seasons': array([], dtype=float6
4), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

Below is the forecast plot using double exponential method: -



BY looking at the graph we can see that this time the forecast values are not flat this time but looks linear in nature.

Below is the RMSE value of this model: -

```
DES RMSE: 15.255861145392286
```

- **Building Triple Exponential smoothing model with additive seasonality: -**

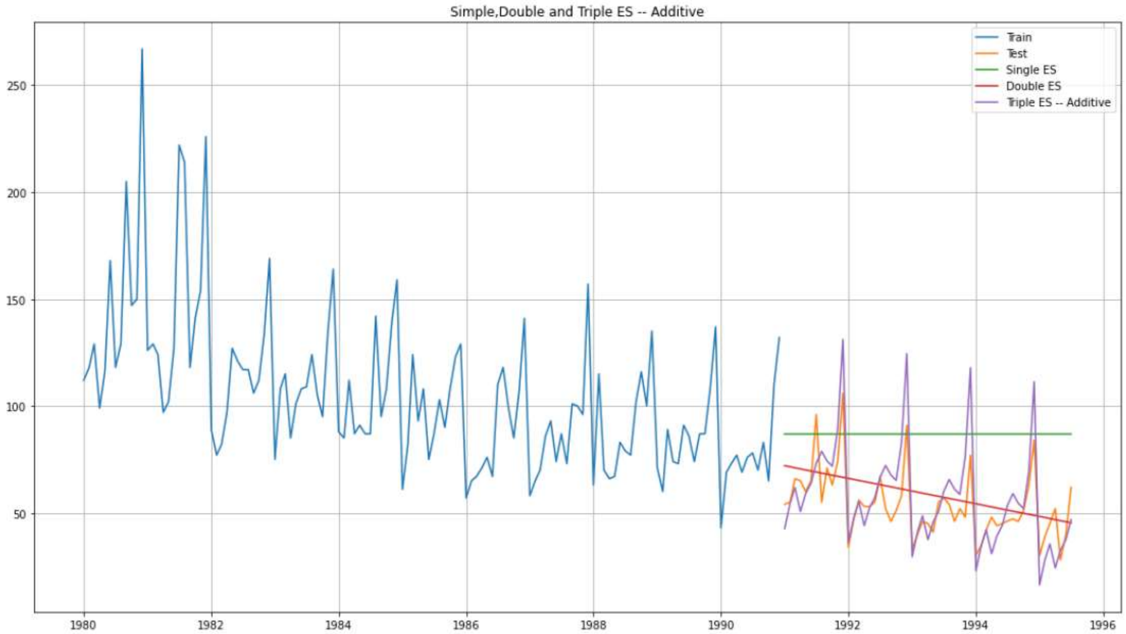
This method involves the "Level" component of data (alpha) as well as the "Trend" component (beta) along with "seasonality" component (Gamma) with additive nature. which mean it is best suitable for data with clear trend and clear seasonality. This is also called as "Holt winter's Linear method".

Below are the parameter details for the model: -

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.08830330642635406, 'smoothing_trend': 6.730635331927582e-05, 'smoothing_seasonal': 0.004455138229351625,
'damping_trend': nan, 'initial_level': 146.88752868155674, 'initial_trend': -0.5492163940406024, 'initial_seasons': array([-31.
12207537, -18.81171138, -10.86052241, -21.52235816,
       -12.68359535,  -7.17529564,   2.7456236 ,   8.84900094,
         4.85724354,   2.9520333 ,  21.05004912,  63.29916317]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

Below is the forecast plot using triple exponential method: -



BY looking at the graph we can see that the forecast values are getting better as it acknowledge both trend and seasonality component of data.

Below is the RMSE value of this model: -

```
TES RMSE: 14.23282652785053
```

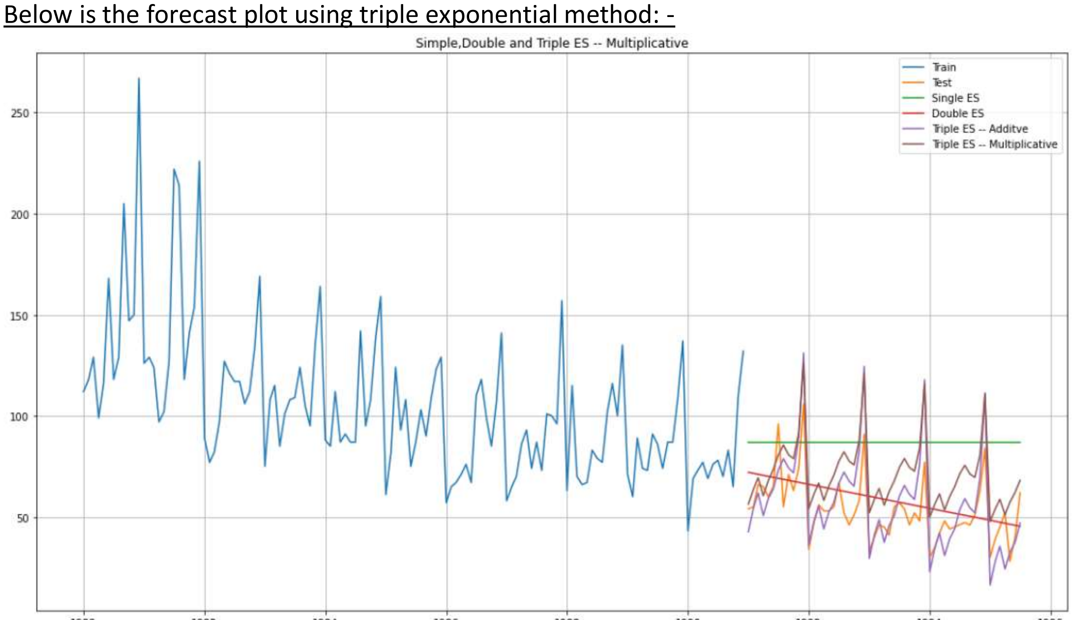
Building Triple Exponential smoothing model with multiplicative seasonality: -

Below are the parameter details for the model: -

```
==Holt Winters model Exponential Smoothing Estimated Parameters ==

{'smoothing_level': 0.07132109562890512, 'smoothing_trend': 0.04553831096563722, 'smoothing_seasonal': 8.356711212063695e-07,
'damping_trend': nan, 'initial_level': 134.25655591779326, 'initial_trend': -0.8038265942903572, 'initial_seasons': array([0.83
746068, 0.94985307, 1.03812083, 0.90732186, 1.02043162,
       1.11131741, 1.22228039, 1.30104211, 1.23132915, 1.20610008,
       1.40577823, 1.93832412]), 'use_boxcox': False, 'lamda': None, 'remove_bias': False}
```

Below is the forecast plot using triple exponential method: -



Below is the RMSE value of this model: -

```
TES_am RMSE: 20.13155549909118
```

BELOW IS THE CONSOLIDATE RMSE VALUES OF ALL THE EXPONENTIAL MODEL BUILT: -

	Test RMSE
Single ES	36.748389
Double ES	15.255861
Triple ES -- Additive	14.232827
Triple ES -- Multiplicative	20.131555

Inferences: - From the above all the exponential models built we can conclude that the **triple exponential model with additive seasonality** out performs all the other model based on respective RMSE scores.

Building Linear Regression Model: -

In this particular linear regression, we are going to regress the “Rose” variable against the order of the occurrence. For this we need to modify our training data before fitting into linear regression model.

We see that the total observation in data is 187 out of which 132 are in training and 55 are in testing dataset.

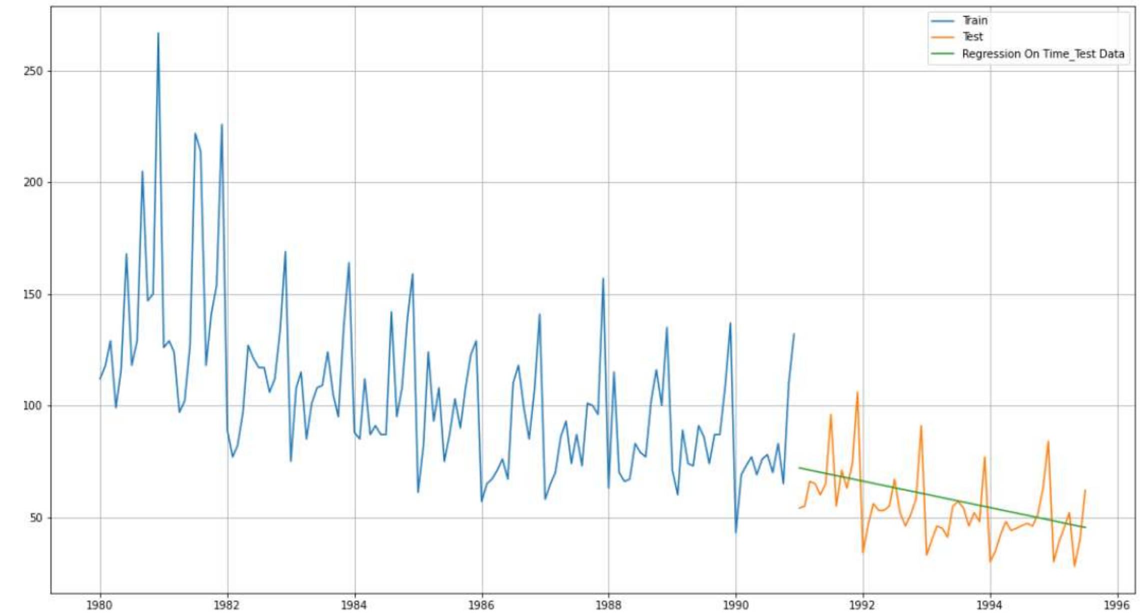
As linear regression requires at least one “X” variable to get the outcome of our interest but in our data, we have the time element. So, we will take the time as our independent variable and the sales number as the dependent variable. However, we have set the time variable as index. Now, we can create a dummy variable in sequence which will represent time variable as X but not as an index.

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

We can see above that as we define the train data it contains 132 observations and test data contains observations from 133 till 187 and this is the independent variable.

Now we need to add the above created dummy variable to the original dataset and to do that we have taken a copy of the original dataset so set we do not mess up with the original data.

Now we will build the linear regression model in a usual way and below is the graphical representation of the forecast.

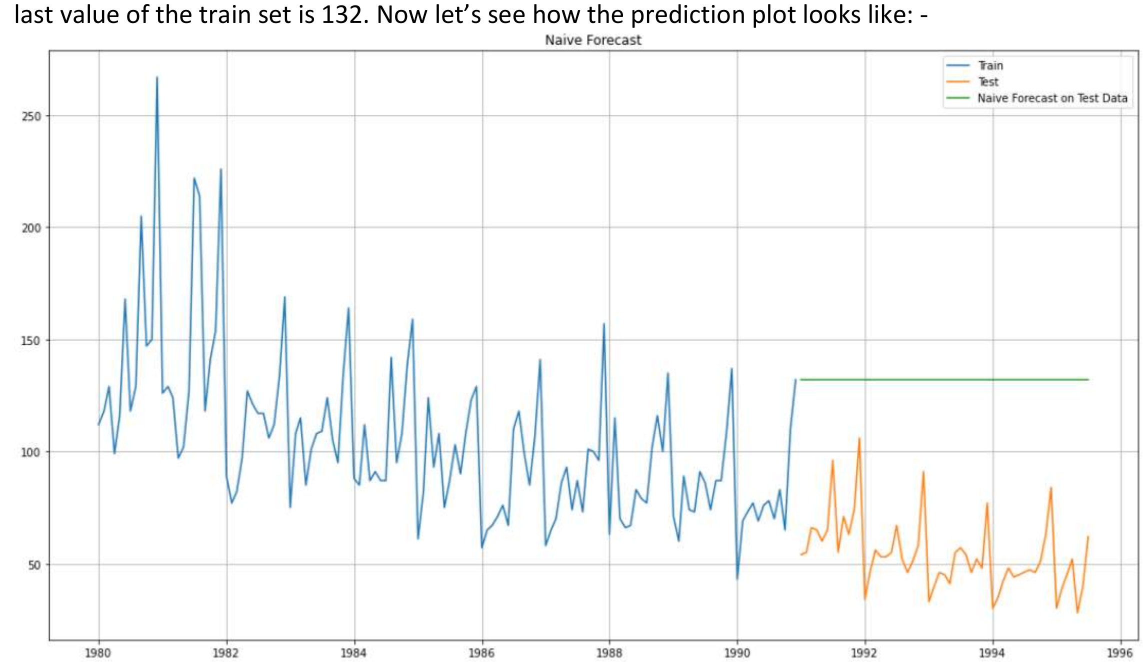


The green line shown above is the linear regression model.
Below is the RMSE score of linear regression model: -

```
For RegressionOnTime forecast on the Test Data,  RMSE is 15.255
```

Building Naive Model: -

The naïve model works on the basis of the last value of the data and the same value repeats for rest of the data. Which mean that the forecast is going to be flat in nature. In this case the last value of the train set is 132. Now let's see how the prediction plot looks like: -

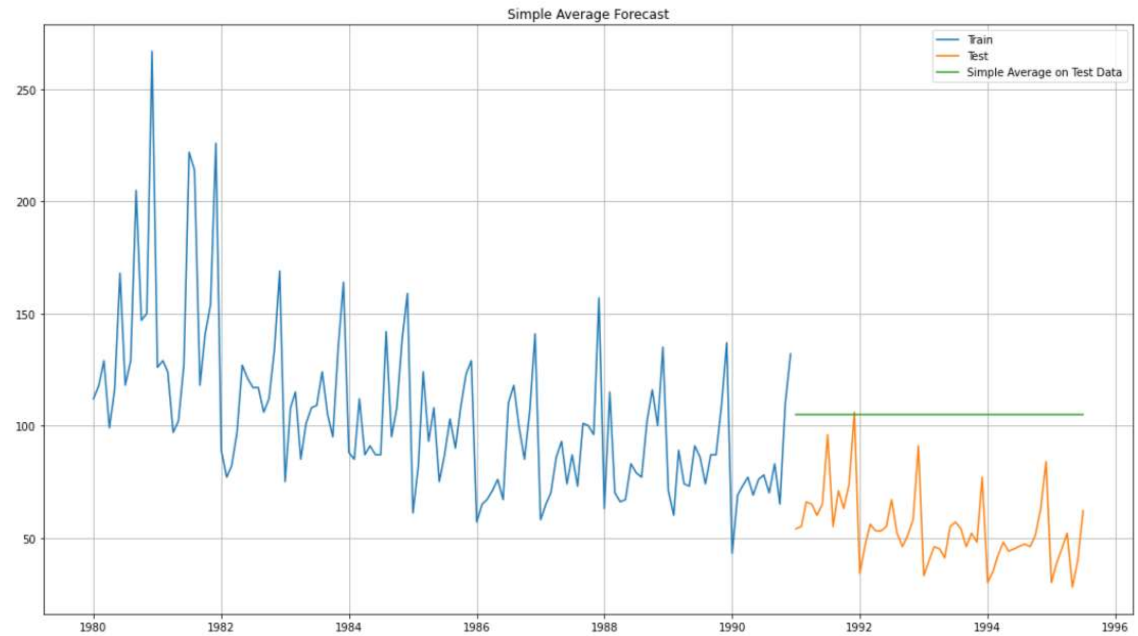


Lets check the model evaluation using RMSE: -

```
For RegressionOnTime forecast on the Test Data,  RMSE is 79.672
```

Building Simple Average Model: -

In this model the average of the train data becomes forecast for the test data and we will get a flat forecast. Below is the graphical representation of the forecast using this model: -



Below is the model evaluation value using RMSE: -

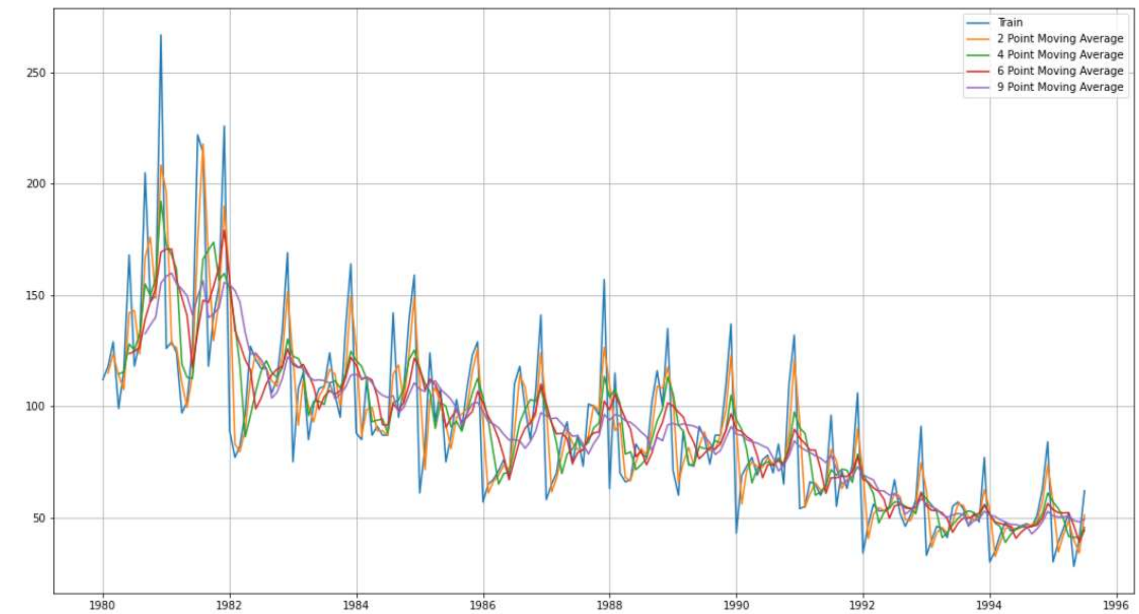
```
For Simple Average forecast on the Test Data,  RMSE is 53.413
```

Building Moving Average Model: -

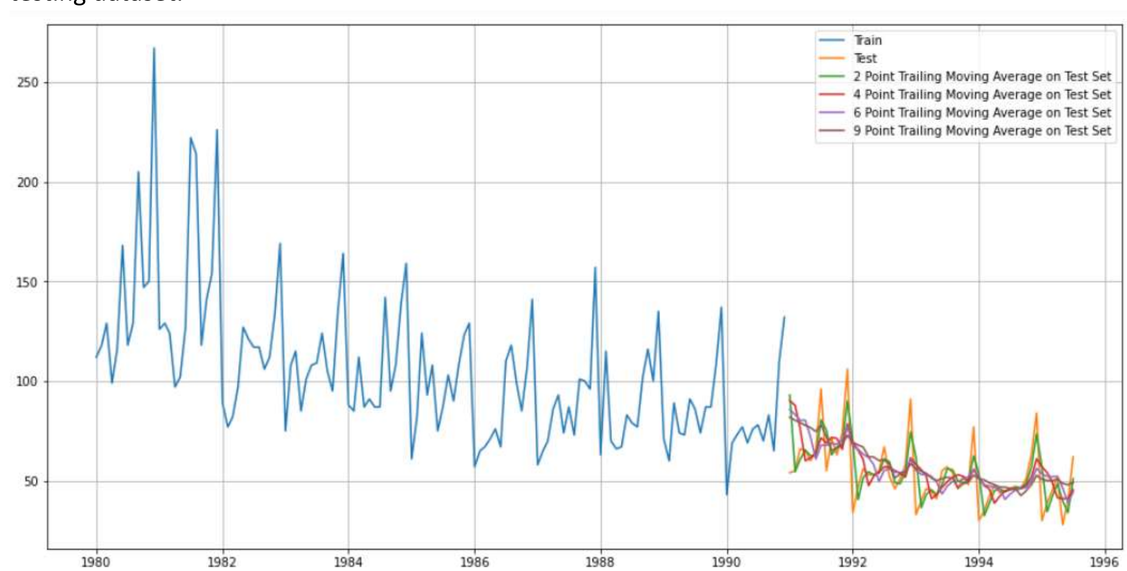
This model creates a cascading window which helps in calculating a rolling mean for different intervals. In this model we cannot use the train or test dataset as the components changes in moving average so we need to consider the complete data for this model. And once the model is built, we can then divide them into training and testing data.

Here are building model with different window size of 2, 4, 6 and 9.

Below pot shows the graphical representation of model built on the whole data.



Now let's divide the data into train and test dataset. Below is the visualization of forecast onto testing dataset.

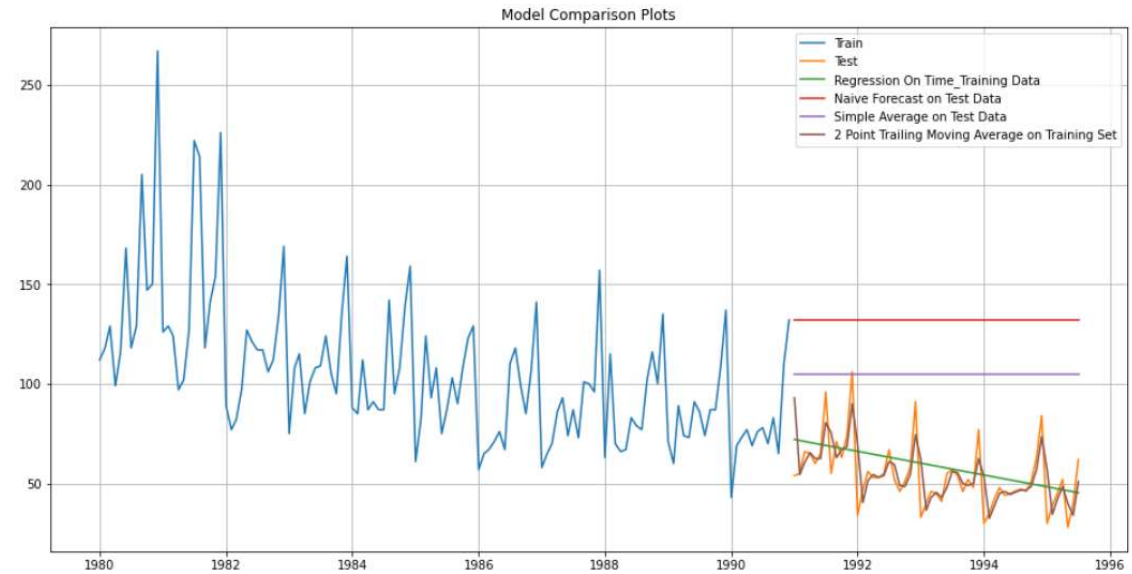


Below are the RMSE scores of the model built with different rolling window.

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.530
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.444
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.555
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.722
```

From above we can conclude that the moving average with rolling window "2" gives the best forecast with least RMSE value.

Now let's visualize the forecast with moving average and rolling window as "2".



Below is the consolidated RMSE values of all the model built above.

	Test RMSE
RegressionOnTime	15.255492
NaiveModel	79.672475
SimpleAverageModel	53.413298
2pointTrailingMovingAverage	11.529985
4pointTrailingMovingAverage	14.444375
6pointTrailingMovingAverage	14.554986
9pointTrailingMovingAverage	14.721520

From above we can conclude that the model built with moving average with rolling window "2" out performs all other models in terms of accuracy with minimal error among other models.

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.

Solution: -

Stationarity is a process of making a time series data stationary through out the time. That means the data won't be having upward or downward trend. The stationarity of data can be tested using dickey fuller test alongside the required level of significance. If we find that the data is non-stationary, we can have various level of differencing done within the data and stationarity can be achieved. Overall, this process will eliminate the trend part.

We could see visually that there are some elements of trend present in the data. Now lest perform a statistical test to confirm if the series is stationary or not.

The null hypothesis of this test says "the time series is not stationary" and alternate hypothesis would be "time series is stationary".

Ho == Time series is not stationary
Ha == Time series is stationary

We see ta 5% significance as instructed.

Below is the outcome of the test performed: -

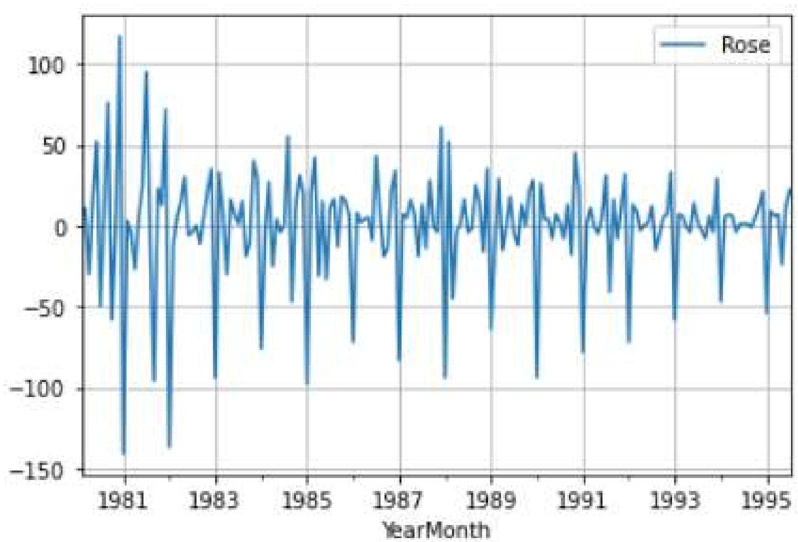
```
DF test statistic is -2.240
DF test p-value is 0.4675494470630276
Number of lags used 13
```

We see that the p value is more than 0.05 and we fail to reject the null hypothesis and conclude that the given time series is not stationary.

In order to make the time series stationary we start with taking a 1st order difference and test again is stationarity is achieved or not.
While performing 1st order differencing we understand that the 1st value will be a NULL value or missing data. Hence, we are dropping the missing value.
Below is the outcome of dickey-fuller test performed after 1st level differencing.

```
DF test statistic is -8.164
DF test p-value is 2.9904329878167767e-11
Number of lags used 12
```

We notice that this time the P-value is very small and less than 0.05% of significance level. Therefore, the stationarity in data is nor achieved.
Now let's visually inspect the differenced time series.



We can see that there is no trend element present in the data and its stationary now.

YearMonth	Rose
1980-02-01	6.0
1980-03-01	11.0
1980-04-01	-30.0
1980-05-01	17.0
1980-06-01	52.0
...	...
1995-03-01	6.0
1995-04-01	7.0
1995-05-01	-24.0
1995-06-01	12.0
1995-07-01	22.0

This is how the difference data looks like and same can used further for analysis and forecasting. The differenced data should be used for building ARIMA or SARIMA models.

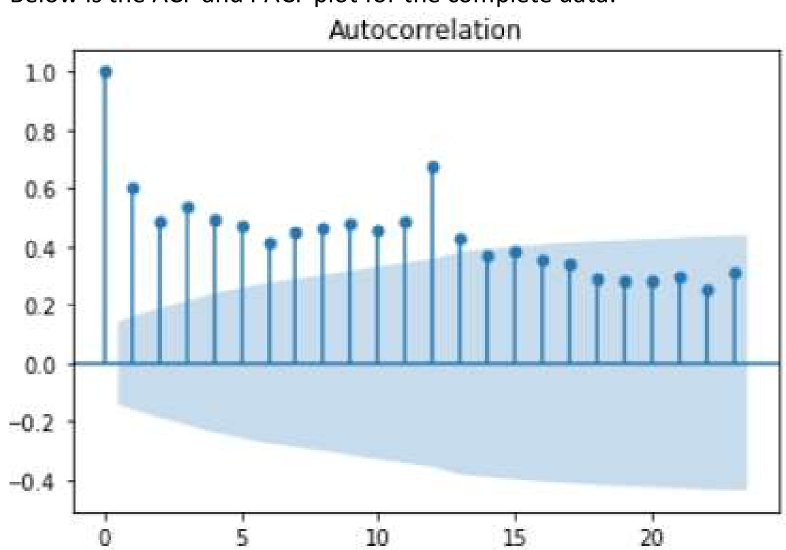
6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

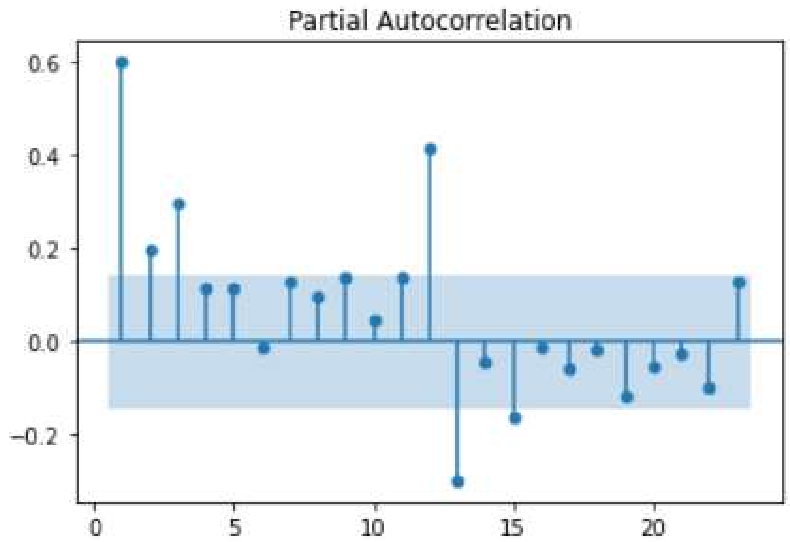
<u>Solution: -</u>

While building ARIMA/SARIMA model's it is recommended at we try and figure out the order of ARIMA (p, d and q) and that can be done by plotting ACF and PACF plot.

ACF does account for the intermediary series where as in PACF the influence on intermediary series is completely discounted.

Below is the ACF and PACF plot for the complete data: -





Now let's discuss a little about the above plots and see how to read and get the cut off points from these plots.

- The ACF plot gives the value of the "q" term, which is the moving average.
- The PACF plot gives us the value of the "p" term.
- The term "d" is the level of differencing considered while performing stationarity in data.
- There are two possibilities of these plots, 1st we may have a cut-off or we may not have a cut-off.
- Lag "0" (zero) is never counted because a series will always have 100% correlation with itself.
- Now let's count the number of significant lags. The significant lags are the once whose tip points are outside the shaded region. Shaded region is our 95% confidence band.
- The cut-off starts when the tip of lag lies within the shaded region. From above ACF plot we can see that the cut-off will be "14" as 15th lag is within the shaded region. Value of q = 14.
- From above we know the value of d = 1
- There may be chances at we will not get a proper cut off till the higher order lag. We see a cut off at much later stage. By market practice we account up to 12 lags and if we did not get any cut-off, we conder the value as "0" (zero). Therefore, now out value of q changes to "0".
- From the above PACF plot we can see that value of p will be "3".
- So manually we got all the values which is **p = 3, d = 1 and q = 0**.

Earlier we have divided data into train and test split. Let check if we have stationarity in training data set using dickey-fuller test and below are the results: -

```
rose test statistic is -1.686
rose test p-value is 0.7569093051047057
Number of lags used 13
```

From above output we can conclude that the data is not stationary as P-value is greater than 0.05% level of significance and we fail to reject the null hypothesis of stationarity. we need to make it stationary before building the model. We will perform stationary using 1s order differencing and check for stationarity. Below is the outcome: -

```
rose test statistic is -6.804
rose test p-value is 3.894831356782385e-08
Number of lags used 12
```

Now we see a change in P-value and its very much less than level of significance of 0.05% and our train data is now stationary and ready for model building.

Building Automated ARIMA MODEL: -

In this we will try with various combinations of "p", "d" and "q" terms. which is more like a grid search approach and will pick the best combination based on AIC value. "d" value is fixed at "1" as we see that the data attains stationarity with 1st order differencing and this grid approach we will try with values of "p" and "q" ranging from 1 to 3.

Below is the combination generated over which we will try and fit the ARIMA model: -

```
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)
```

Now let's fit ARIMA model to all the above combination and get their respective AIV value to choose the best. Please note that the model with least AIC Value will be considered as the best. Below is the top 5 AIC model sorted in ascending order: -

	param	AIC
11	(2, 1, 3)	1274.695319
15	(3, 1, 3)	1278.654399
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

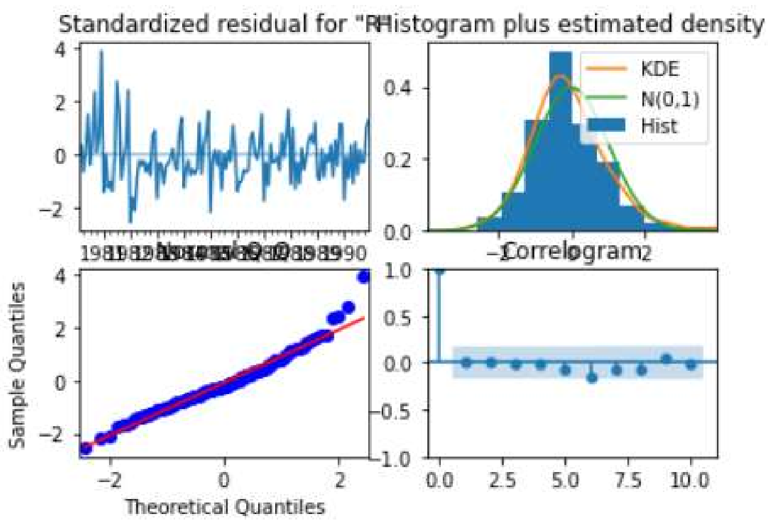
From above we can conclude that the model with p, d and q values of 2,1 and 3 respectively gives the least AIC value.

Now let's fit this ARIMA model into train set, below is the output of ARIMA model for train set:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Wed, 21 Jul 2021   AIC                           1274.695
Time:                        00:12:38   BIC                           1291.947
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6780      0.084    -20.029      0.000      -1.842      -1.514
ar.L2         -0.7287      0.084     -8.697      0.000      -0.893      -0.565
ma.L1          1.0447      0.616      1.695      0.090      -0.163       2.253
ma.L2         -0.7716      0.132     -5.856      0.000      -1.030      -0.513
ma.L3         -0.9044      0.558     -1.620      0.105      -1.999       0.190
sigma2       858.9120    517.873      1.659      0.097    -156.100    1873.924
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.43
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).
```

Below is the diagnostic graph for the same: -



Forecasting on the test data set using the same parameter as train dataset and below is the RMSE score for the same: -

```
RMSE: 36.76869087421142
MAPE: 75.67053334070707
```

Building Automated SARIMA Model: -

While building SARIMA model there are extra components attached along with the above components of p, d and q. The additional components are below: -

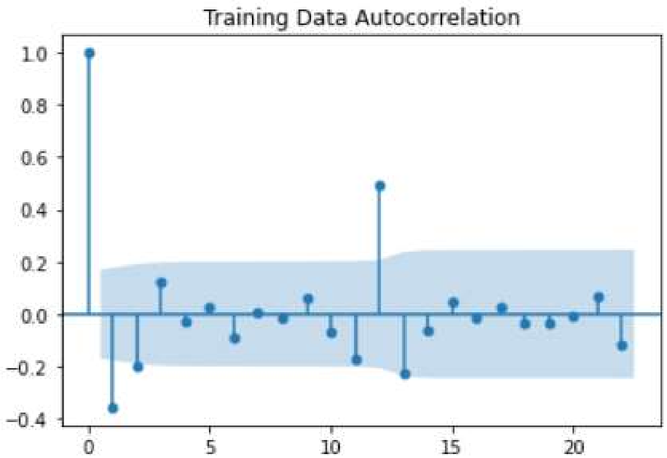
P = AR component with seasonality

D = Differencing for data with seasonality

Q = MA component with seasonality

S = number of lags between the two seasonality components.

to build a model we need to set a mark for the “S” component and same can be decided using ACF plot of training dataset. Below is the ACF plot: -



From above ACF plot we can’t really see anything for S value. As we see significant lag after 9th lag. So, we considering that as the value of S.

To build an automated SARIMA model we will use a grid search approach with range values of p, q, P, Q and S. The values of d = 1 and D = 0 as obtained from level of differencing while performing stationarity.

Below are the combinations obtained from all the above values of parameters: -

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 9)
Model: (0, 1, 2)(0, 0, 2, 9)
Model: (0, 1, 3)(0, 0, 3, 9)
Model: (1, 1, 0)(1, 0, 0, 9)
Model: (1, 1, 1)(1, 0, 1, 9)
Model: (1, 1, 2)(1, 0, 2, 9)
Model: (1, 1, 3)(1, 0, 3, 9)
Model: (2, 1, 0)(2, 0, 0, 9)
Model: (2, 1, 1)(2, 0, 1, 9)
Model: (2, 1, 2)(2, 0, 2, 9)
Model: (2, 1, 3)(2, 0, 3, 9)
Model: (3, 1, 0)(3, 0, 0, 9)
Model: (3, 1, 1)(3, 0, 1, 9)
Model: (3, 1, 2)(3, 0, 2, 9)
Model: (3, 1, 3)(3, 0, 3, 9)
```

Let’s fit these combinations into SARIMAX and obtain their AIC values. The combination with least AIC value will be considered as the best and can be used for prediction on training dataset.

Below are the AIC scores obtained for top 5 combinations: -

	param	seasonal	AIC
255	(3, 1, 3)	(3, 0, 3, 9)	914.734579
127	(1, 1, 3)	(3, 0, 3, 9)	917.317491
191	(2, 1, 3)	(3, 0, 3, 9)	919.151413
119	(1, 1, 3)	(1, 0, 3, 9)	926.359588
63	(0, 1, 3)	(3, 0, 3, 9)	926.727658

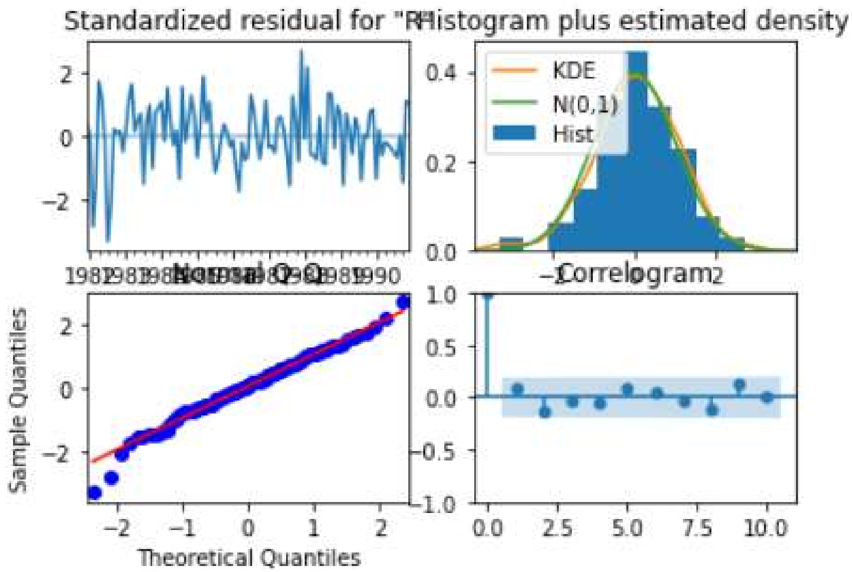
Form above we can conclude that the parameter values with p = 3, d = 1, q = 3, P = 3, D = 0, Q = 3 and S = 9 outperforms will less AIC scores.

Now we can build a SARIMA model after getting the right values of the parameters and below is the result of building the model: -

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -464.872
Date:                             Wed, 21 Jul 2021   AIC                            951.744
Time:                                     19:47:50   BIC                            981.349
Sample:                                 01-01-1980   HQIC                           963.750
                                      - 12-01-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5027      0.083     -6.082      0.000      -0.665      -0.341
ar.L2         -0.6628      0.084     -7.918      0.000      -0.827      -0.499
ma.L1         -0.3714    578.654     -0.001      0.999   -1134.513    1133.771
ma.L2          0.2033    363.737      0.001      1.000    -712.708     713.114
ma.L3         -0.8320    481.376     -0.002      0.999    -944.313     942.649
ar.S.L6       -0.0838      0.049     -1.720      0.085      -0.179       0.012
ar.S.L12       0.8099      0.052     15.466      0.000       0.707       0.913
ma.S.L6        0.1702      0.248      0.686      0.493      -0.316       0.656
ma.S.L12      -0.5646      0.199     -2.834      0.005      -0.955      -0.174
ma.S.L18       0.1710      0.143      1.198      0.231      -0.109       0.451
sigma2       260.7811   1.51e+05      0.002      0.999   -2.96e+05    2.96e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):                 4.77
Prob(Q):                              0.40   Prob(JB):                         0.09
Heteroskedasticity (H):               0.54   Skew:                            -0.36
Prob(H) (two-sided):                  0.06   Kurtosis:                         3.73
===================================================================================

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

Below is the diagnostic plot for the same: -



With the same model build above we can now perform forecast in test dataset and obtain its RMSE value: -

```
RMSE: 27.068898978940556
MAPE: 55.07586918293619
```

The top 5 forecasts are shown below with 95% confidence interval: -

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-01	66.900092	16.350226	34.854239	98.945946
1991-02-01	65.988158	16.481446	33.685118	98.291198
1991-03-01	74.438688	16.587371	41.928039	106.949337
1991-04-01	76.040407	16.709956	43.289494	108.791320
1991-05-01	78.415084	16.710569	45.662970	111.167198

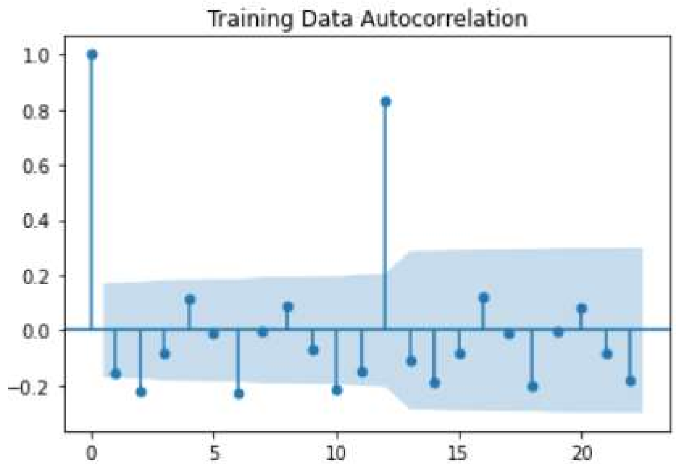
7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

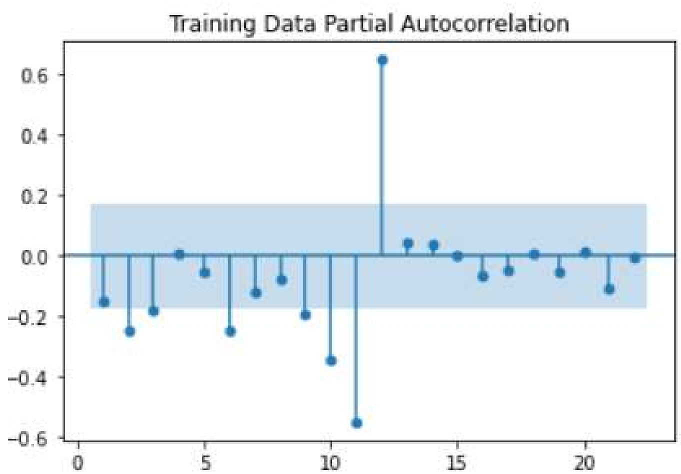
Solution: -

Building ARIMA model basis cut-off obtained from ACF and PACF plot: -

Earlier we have discussed about how to read the ACF and PACF plot for the overall data. Now we will plot the same ACF and PACF plot for training data and determine the cut-offs for p, d and q.

Below are the ACF and PACF plots for training data: -



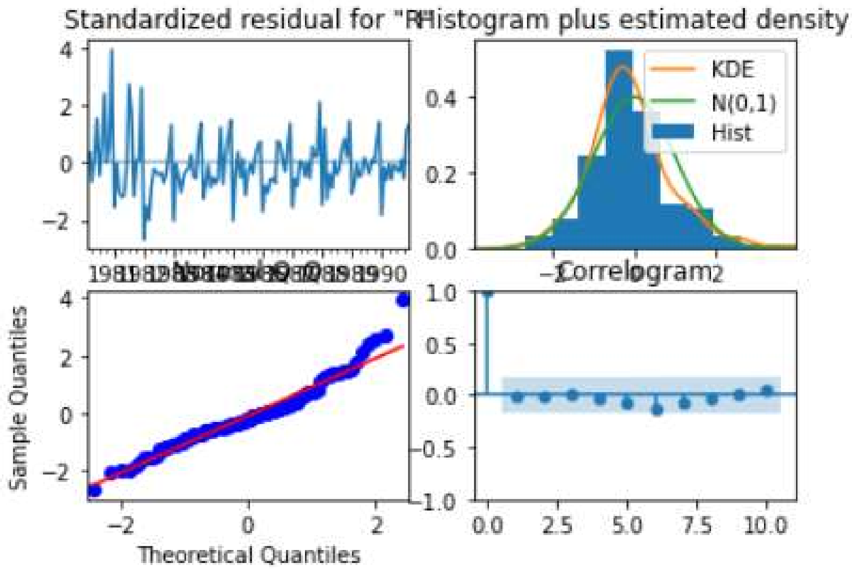


From plots as shown above indicates that the value of p = 2 and q = 2 and we know the value of d = 1 using which we have made the training data stationary. Now let’s fit the training data with these values and below is the output: -

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -635.935
Date:                Wed, 21 Jul 2021   AIC                           1281.871
Time:                        20:51:46   BIC                           1296.247
Sample:                    01-01-1980   HQIC                          1287.712
                         - 12-01-1990
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
===================================================================================

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

Below is the diagnostic plot of the model built: -

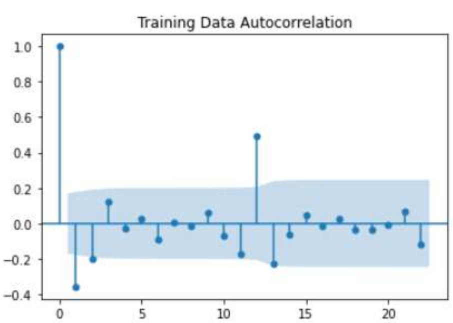


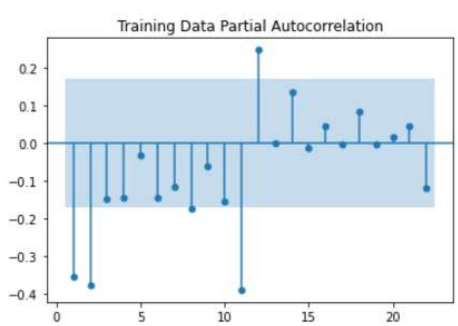
Now as the model is built, we can use the same for prediction on to testing dataset and check on its RMSE value. Below is the RMSE value obtained out of this model: -

```
RMSE: 36.82342004988253
MAPE: 75.88057965112233
```

<u>Building SARIMA model basis cut-off obtained from ACF and PACF plot: -</u>

To build a SARIMA model we need to obtain the value of "P" and "Q" in a seasonal way from the below plots: -





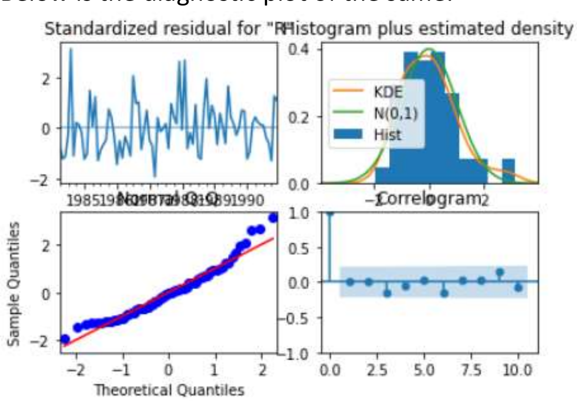
- From above manually built ARIMA model we know that the values of p, d and q as 2,1 and 2 respectively.
- From the above ACF plot we don't see any significant lag at a larger scale and we can conclude the value of Q as 0.
- From the above PACF plot we can see that every 6th lag is significant can we can conclude the value of P as 6.
- Value of D will be "0" as we are not performing and differencing and value of S will be 8 as shown above.

  Now will fit the above obtained parameter from plots to build SARIMA model on train dataset and below is the output: -

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                132
Model:             SARIMAX(2, 1, 2)x(6, 0, [], 8)   Log Likelihood                -343.066
Date:                              Wed, 21 Jul 2021   AIC                            708.131
Time:                                      21:02:00   BIC                            734.470
Sample:                                  01-01-1980   HQIC                           718.699
                                       - 12-01-1990
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7223      0.516     -1.400      0.162      -1.733       0.289
ar.L2         -0.0067      0.164     -0.041      0.967      -0.327       0.314
ma.L1         -0.2189    481.659     -0.000      1.000    -944.253     943.815
ma.L2         -0.7811    376.253     -0.002      0.998    -738.224     736.662
ar.S.L8        0.0250      0.128      0.195      0.846      -0.227       0.277
ar.S.L16      -0.1611      0.130     -1.239      0.215      -0.416       0.094
ar.S.L24       0.4639      0.111      4.187      0.000       0.247       0.681
ar.S.L32      -0.1037      0.084     -1.233      0.218      -0.269       0.061
ar.S.L40       0.1689      0.088      1.918      0.055      -0.004       0.341
ar.S.L48       0.1620      0.078      2.064      0.039       0.008       0.316
sigma2       266.9613   1.29e+05      0.002      0.998   -2.52e+05    2.52e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 9.81
Prob(Q):                              0.93   Prob(JB):                         0.01
Heteroskedasticity (H):               0.58   Skew:                             0.79
Prob(H) (two-sided):                  0.16   Kurtosis:                         3.65
===================================================================================

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

  Below is the diagnostic plot of the same: -



We can perform prediction on test dataset using the model built above and check on its RMSE score: -

```
RMSE: 25.947909131383774
MAPE: 51.98614705755057
```

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.

Solution: -

	Test RMSE
Single ES	36.748389
Double ES	15.255861
Triple ES -- Additive	14.232827
Triple ES -- Multiplicative	20.131555
RegressionOnTime	15.255492
NaiveModel	79.672475
SimpleAverageModel	53.413298
2pointTrailingMovingAverage	11.529985
4pointTrailingMovingAverage	14.444375
6pointTrailingMovingAverage	14.554986
9pointTrailingMovingAverage	14.721520
ARIMA(2,1,3) - Auto	36.768691
SARIMA(2,1,3)(2,0,3,6) - Auto	27.068899
ARIMA(2,1,2)- Manual	36.823420
SARIMA(2,1,2)(6,0,0,8)- Manual	25.947909

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.

Solution: -

From above data-frame we can conclude that the best optimized model has the values of 2, 1, 2, 6, 0, 0, 8 for p, d, q, P, D, Q and S respectively as it gives the least RMSE score among all the models. We can use the same values of parameters for the complete dataset and output is shown below: -

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 Rose   No. Observations:                187
Model:             SARIMAX(2, 1, 2)x(6, 0, [], 8)   Log Likelihood                -556.419
Date:                              Wed, 21 Jul 2021   AIC                           1134.837
Time:                                      21:16:02   BIC                           1166.877
Sample:                                  01-01-1980   HQIC                          1147.857
                                       - 07-01-1995
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7499      0.347     -2.160      0.031      -1.430      -0.069
ar.L2         -0.0074      0.112     -0.066      0.947      -0.227       0.212
ma.L1         -0.0973      0.357     -0.273      0.785      -0.797       0.602
ma.L2         -0.7027      0.337     -2.083      0.037      -1.364      -0.041
ar.S.L8        0.1123      0.080      1.397      0.162      -0.045       0.270
ar.S.L16      -0.1155      0.087     -1.331      0.183      -0.286       0.055
ar.S.L24       0.5186      0.074      6.969      0.000       0.373       0.664
ar.S.L32      -0.1005      0.063     -1.600      0.110      -0.224       0.023
ar.S.L40       0.1419      0.059      2.400      0.016       0.026       0.258
ar.S.L48       0.1600      0.051      3.112      0.002       0.059       0.261
sigma2       207.9052     21.827      9.525      0.000     165.125     250.685
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                21.97
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               0.21   Skew:                             0.76
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.25
===================================================================================

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

Predicting for 12 months into the future: -

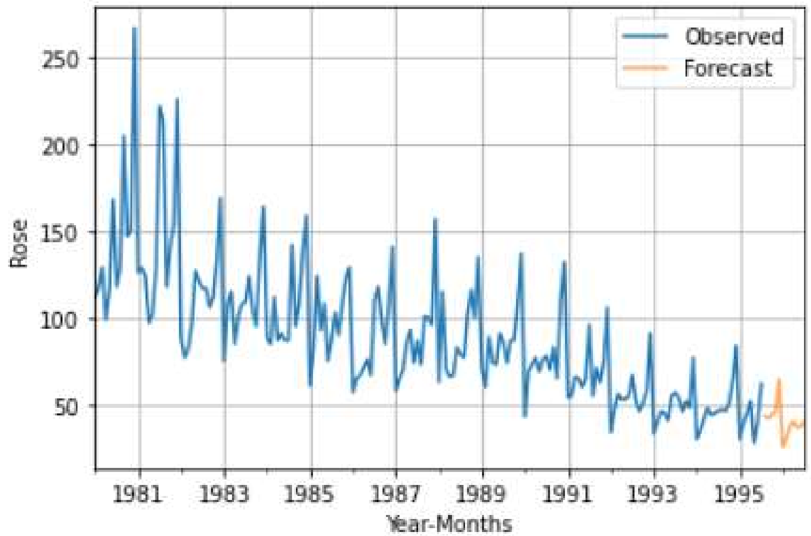
The forecast for 12 months into the future is as below: -

Rose	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	43.290622	14.418919	15.030060	71.551185
1995-09-01	41.956414	14.586460	13.367477	70.545351
1995-10-01	44.305924	14.629753	15.632136	72.979713
1995-11-01	45.599687	14.769179	16.652628	74.546746
1995-12-01	64.416160	14.831457	35.347038	93.485282
1996-01-01	25.504298	14.946533	-3.790368	54.798964
1996-02-01	30.262000	15.020230	0.822890	59.701109
1996-03-01	37.197958	15.122548	7.558308	66.837607
1996-04-01	40.295210	15.452243	10.009370	70.581051
1996-05-01	36.876788	15.575006	6.350337	67.403238
1996-06-01	37.333747	15.669807	6.621490	68.046003
1996-07-01	40.553838	15.784939	9.615925	71.491750

The RMSE score of this forecast is below: -

```
RMSE of the Full Model 30.69046635681222
```

The graphical representation of forecast is as below: -



We do understand that modelling is an iterative process and trying various other combinations of values may lead to better RMSE score and better model.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Solution: -

Solution: -

From all the above model built we have concluded the SARIMA model with p,d,q,P,D,Q and S values as 2,2,2,6,0,0 and 8 respectively as it gives less RMSE score when compared with other models which indicates that this model loses the least data ana captures the maximum for modelling.

Also, triple exponential model works well will minimally difference in RMSE score and can be used for future forecasting.

Various Steps performed in this project: -

- Plotting data and understand about trend and seasonality about the data
- Multiplicative and additive decomposition of data.
- Splitting data into training and test dataset.
- Performing exponential smoothing on train dataset and forecasting on test dataset and evaluating the model using RMSE scores.
- Building Linear regression, Naïve base, simple average model and moving average model on training dataset and forecasting on test dataset, preforming model evaluation using RMSE score.
- Checking on stationarity using dickey-fuller test and performing stationarity using level of differencing.
- Building automated ARIMA and SARIMA model over the stationary data, using grid search approach and selecting the best using AIC score. The model with least AIC score considered

the best. Forecasting on test data using the model with best p,d and q values with least AIC score and perform model evaluation using RMSE score.

- Manually building ARIMA and SARIMA model but deciding on p,d and q values basis cut offs from ACF and PACF plot. Fitting these cut-offs on training dataset and forecasting on test dataset and evaluating performance using RMSE score.
- From all the model built above selecting the best model with least RMSE score and which accounts best for the trend and seasonality component in the data.
- Building the most optimal model on the complete dataset and forecasting into the future for 12 months and performing model evaluation using RMSE score.
- Finally plotting the data provided along with the forecast.

Insight from data: -

From the above exploratory data analysis we can conclude that the data definitely have a trend and seasonality component, we can see a negative trend. Which implies that the sales are decreasing year on year.

Business Recommendations: -

- Conducting a customer survey to check on if the quality is depreciated.
- Training sales staff with latest trend and them creating USP about this product accordingly.
- Digital marketing campaign to increase in visibility.
- Various offers and discount’s during festive season.
- In store tasting counters.
- Various offers for store dealers to promote this product.

------------------------END--------------------------