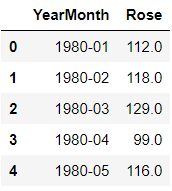
**Problem summary:**

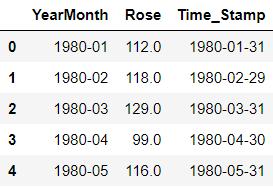
To forecast the Rose wine sales in the 20th Century using the time series data at our disposal. We are given the monthly sales figures of the Rose wine sales from January 1980 to July 1995.

**EDA:**

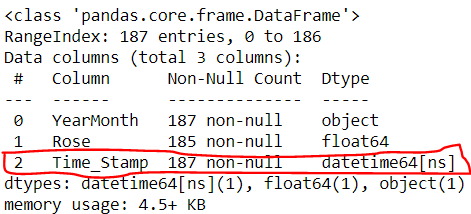
After importing the necessary libraries, we read the data and import the data into a dataframe as follows:



The data has been read as a normal data and so we use the datetime library to read the data as a time series data using the start and end dates and the frequency as monthly.

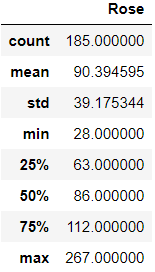


We then use the info command on the data to check if the time series data has been read and identified appropriately:



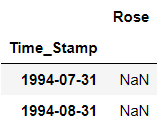
It can be seen that the required column is read correctly. We then remove the column: YearMonth from our dataframe.

We then use the describe function on the data to look at the description of the Rose wine sales



The mean and the median are fairly close to each other. The distance between the 25th percentile and the 50th percentile is almost the same as the distance between the 50th percentile and the 75th percentile. We see that the data is fairly normally distributed with extreme values on the upper end of the spectrum. We will have to closely analyze these extreme values as they are of Rose wine sales for particular months.

Moving on, we check for the null values, na values, non-numerical object values in the data and find that there are two null values present in the data. They are present for the following dates in the data:

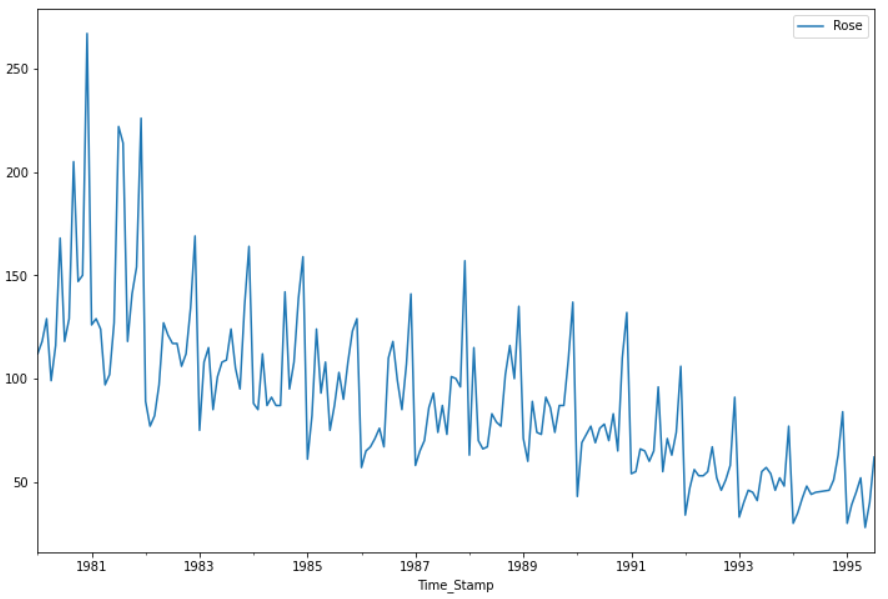


These two values are replaced with the interpolated values as a method of linear interpolation. The absence of any non-numerical object values in the data can be confirmed by the data type that is present. Since there are no object data types in our data this can be confirmed.

The shape of the data looks like this:

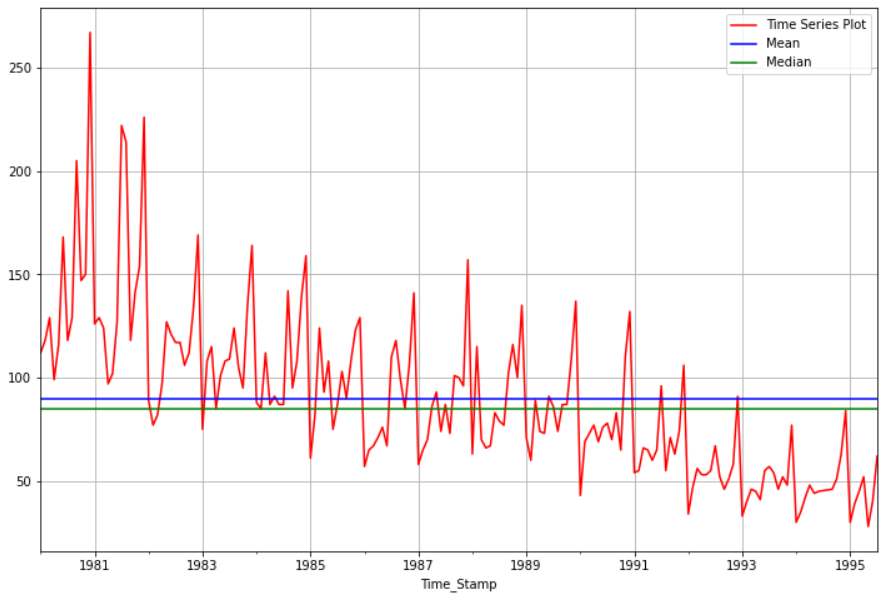


**Plotting Time series data:**



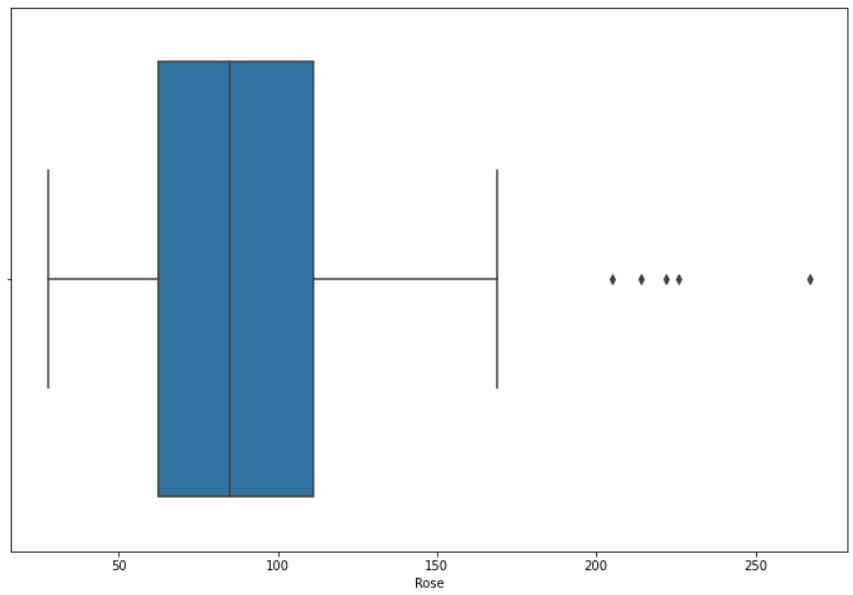
From the above plot it can be seen that the data has both trend and seasonality as well.

**Plotting Time series data along with Mean and Median:**



As stated earlier, the mean and median are close to each other and seem to be at the middle point of the data.

**Boxplot of Time series data:**



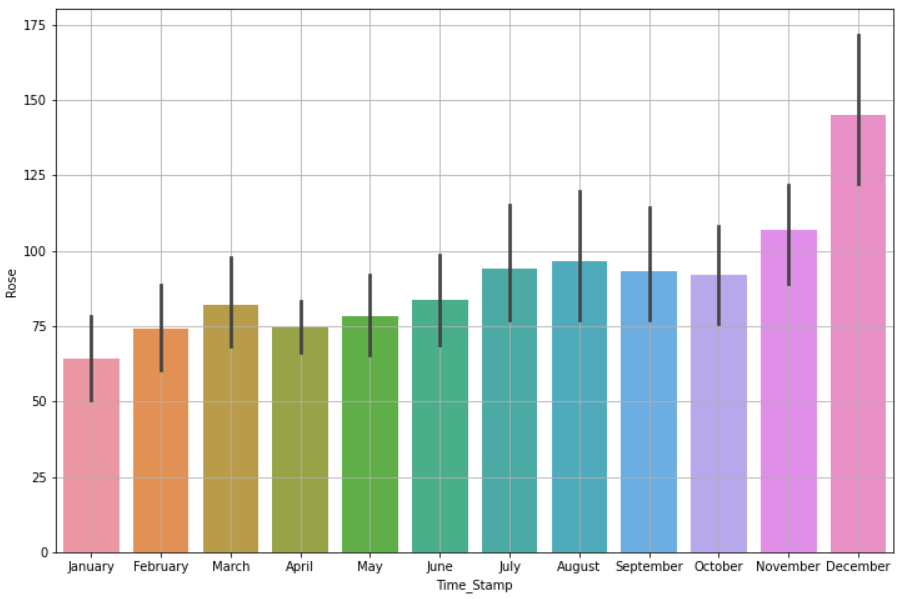
The plot above shows us that there are outliers present in the data beyond the upper whisker of the boxplot.

### Quarterly comparison of shipments using barplot:

### 

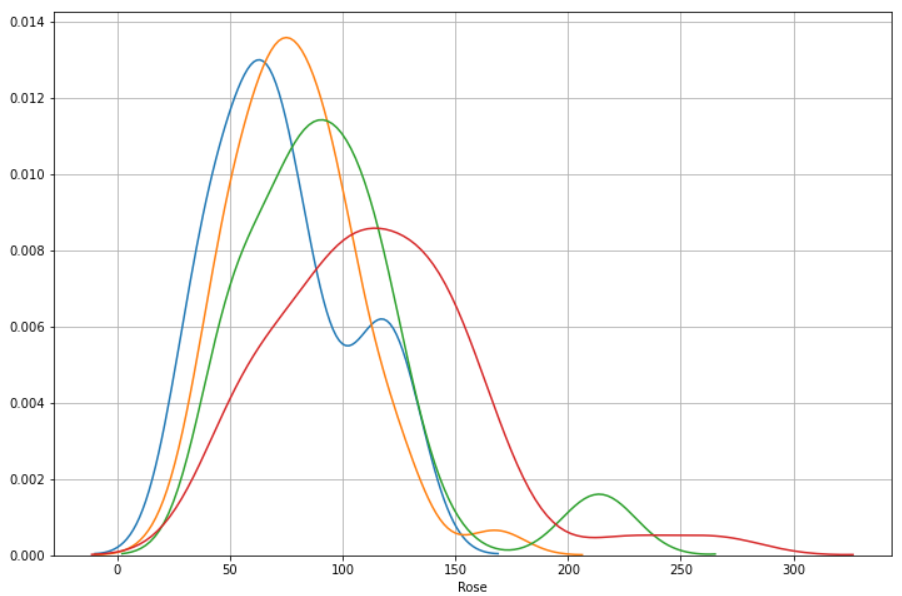
The above plot shows us that the sales were highest in the 4th quarter of each year. The sales figures are lowest at the start of each year and increases gradually. The sales in the third quarter is 35% more than that in the first quarter. This must be due to the presence of festival seasons in the 4th quarter of a year such as Christmas and New Year.

**Monthly comparison of shipments using barplot:**



The Rose wine sales is the highest in the month of December and lowest in the month of January. The sales in March of every year is more than that of sales in Jan / Feb / April / May.

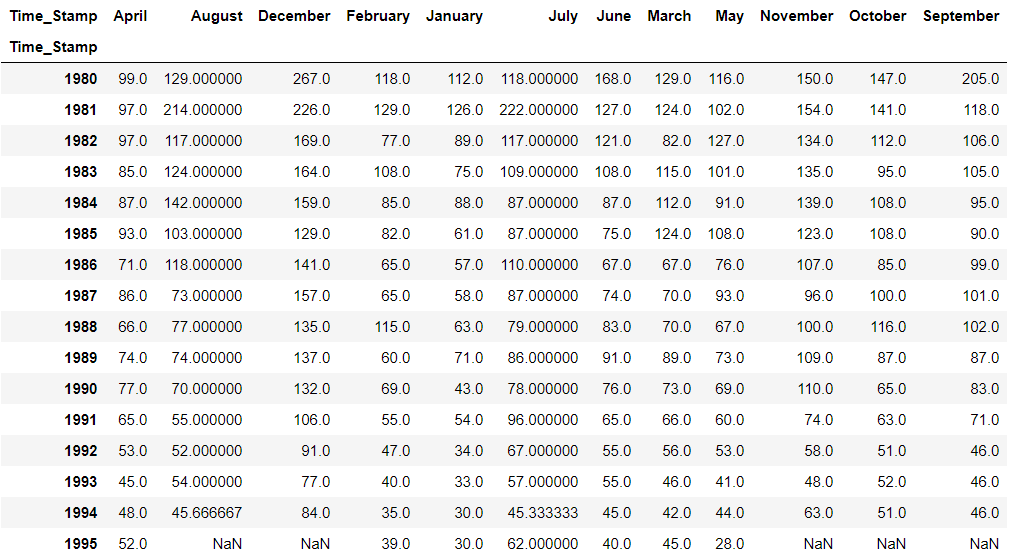
**Distribution plot for quarterly comparison of Rose wine sales:**



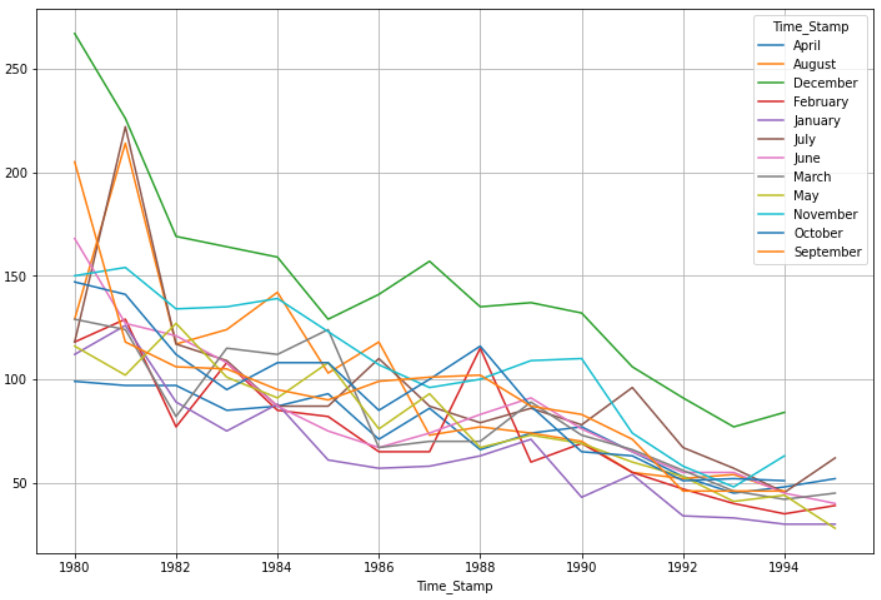
The quarterly distribution plot for the sales is shown above. The red line is the sales for the 4th quarter of years. Blue line corresponds to the first quarter, the orange line to the second quarter and the green line to the third quarter. It is to observe that all the quarters except for the first quarter follow a normal distribution. The first quarter – Blue line, has two noticeable peaks else it would have also been termed for a normal distribution.

**Monthly sales across years:**

We create a data frame to observe the monthly sales across each year and then plot the same.

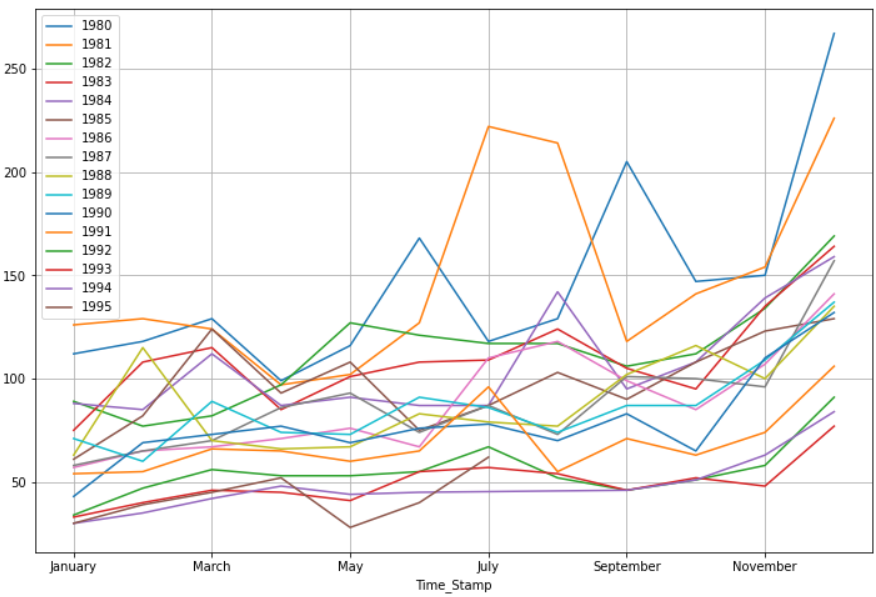


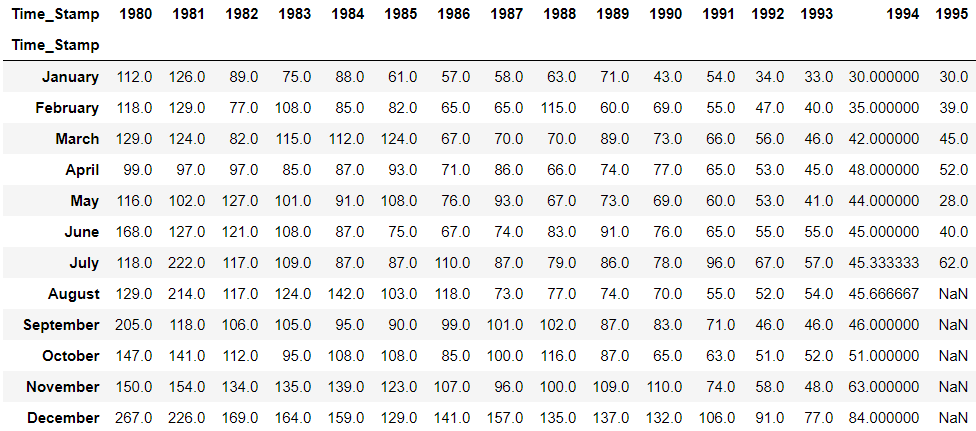
It is to note that ‘Time\_Stamp’ is the time index that we have assigned to the data frame.



* It is to observe that December has the highest sales as observed before as well.
* In the year of 1981 – August, there was a spike in the Rose wine sales that had touched the 225-230 sales units.
* A similar trend was observed in July of the same year that contributes to around 230 sales units of that year. This calls for a closer look into what exactly had happened during those two months of that year that resulted in the sales units soaring up this high.
* Another such a noticeable spike was observed between the years 1987 and 1989 February where the sales units were close to 120. Even in the year 1983 February, the sales numbers were more than 110 units.
* In the year 1982, the sales numbers reduced in the months August and February.
* It is to note that the median sales numbers for the Rose wine sales are 86. So, if there are sales observed more than 100 during any month apart from November and December, then that calls for a closer introspection. It is also advised to look into the fall of the sales numbers.
* The month of March across the years 1983 to 1985 has observed the sales of 110 – 125 units consistently.

**Yearly sales across months:**





* From June till September for the year 1981 the sales numbers have increased tremendously.
* The year 1980 has also seen extraordinary sales across all the months but more so in June and Sep.
* The period 1982 May – Sep has seen sales more than 110 units.
* The sales units between the years 1991 – 1995 has seen an overall low level. The year 1991 has seen a sudden increase in sales in July with the numbers close to 100 units.
* It is to note that there is an overall decline in the Rose wine sales numbers over the years.

### Empirical Cumulative Distribution Function Plot:

### 

### Close to 15 % of the sales is less than 50 sales units.

### Close to 35% of the sales happens beyond 100 sales units.

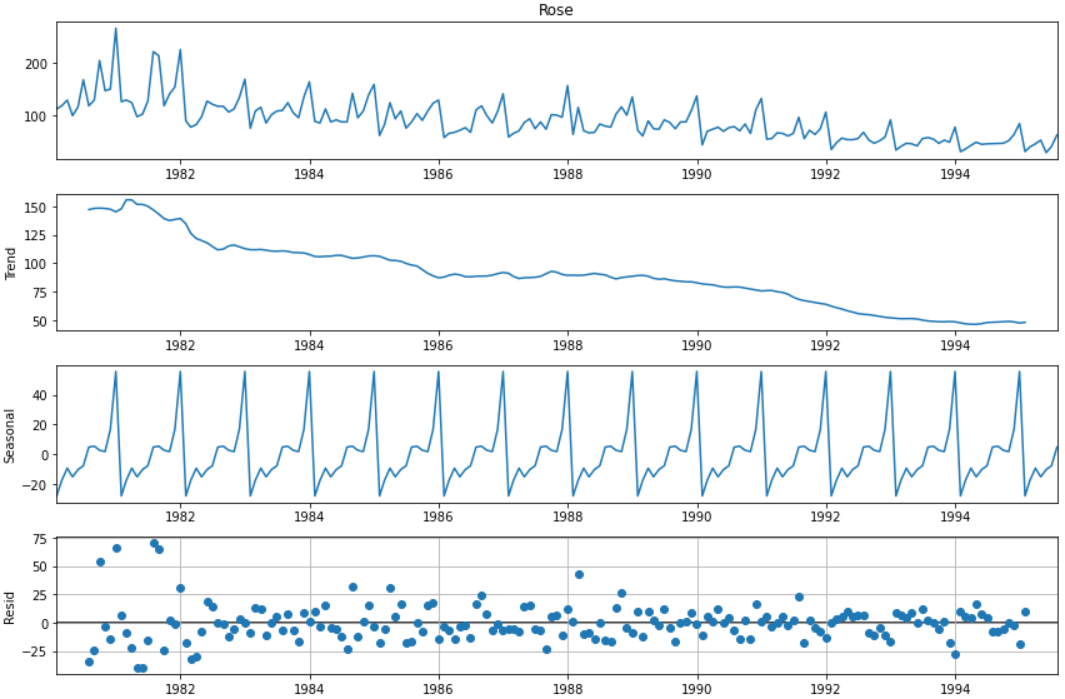
### Average sales and percentage change:

### 

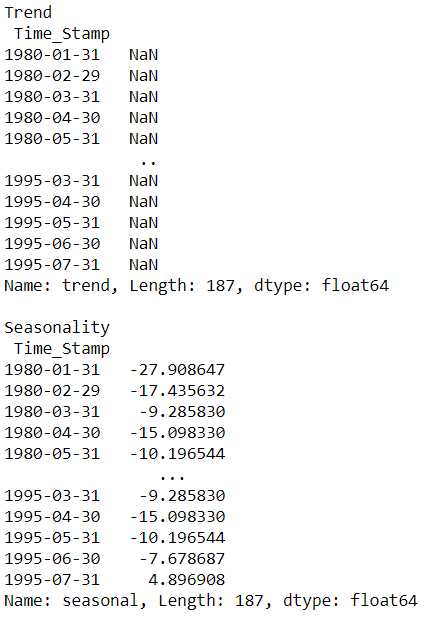
The average customers have been declining with the progress in the years. The percentage change in the sales units can be tallied in accordance with the previous observations thus mentioned.

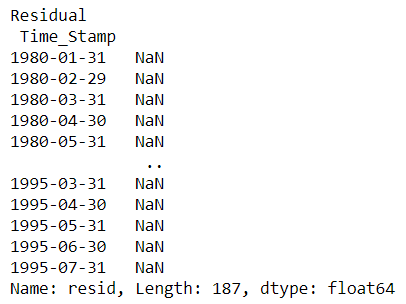
**Decomposition of Time Series:**

Let us decompose the time series in an additive way:

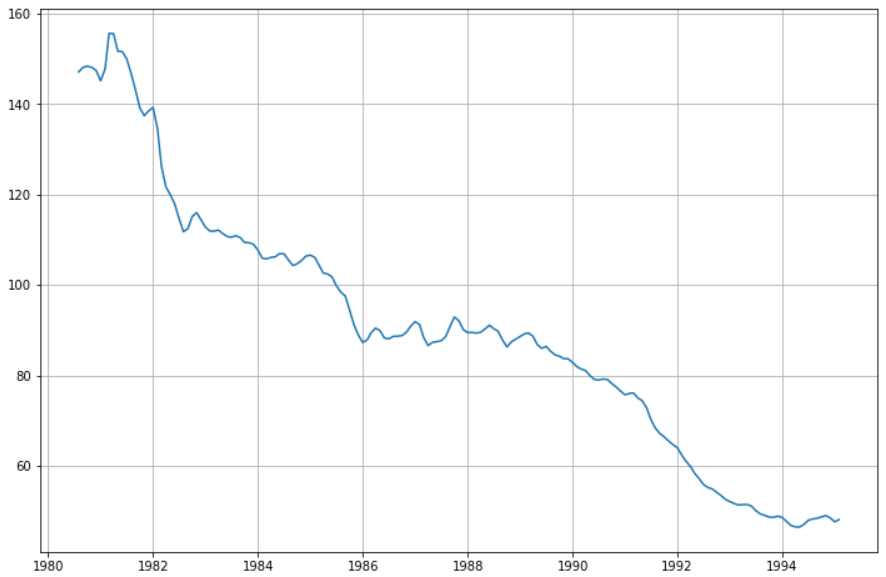


There is a decreasing trend, a constant seasonality and a high residual. The residual is high in the initial years until 1983. In 1983 the residual is centered around 0. Similarly beyond 1990 the residual is centered around 0. In the year 1985 the residual is high.

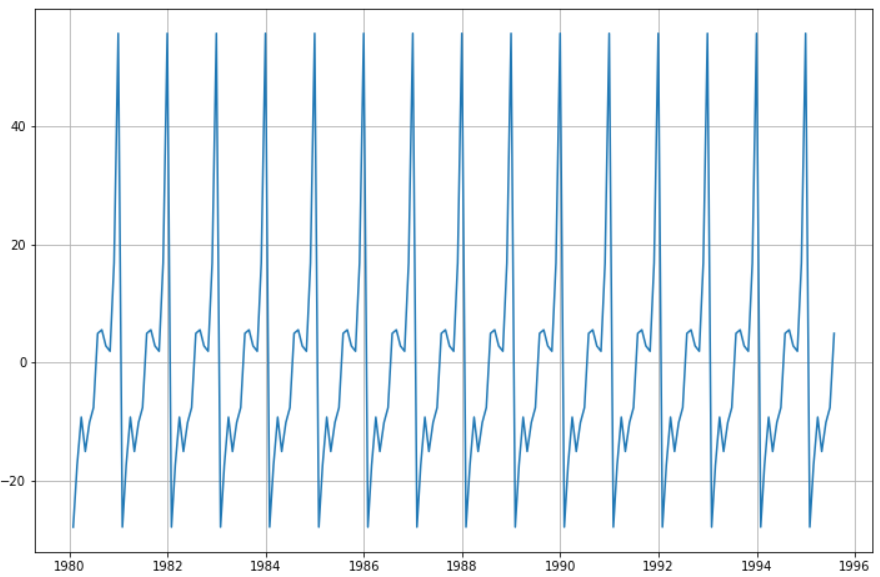




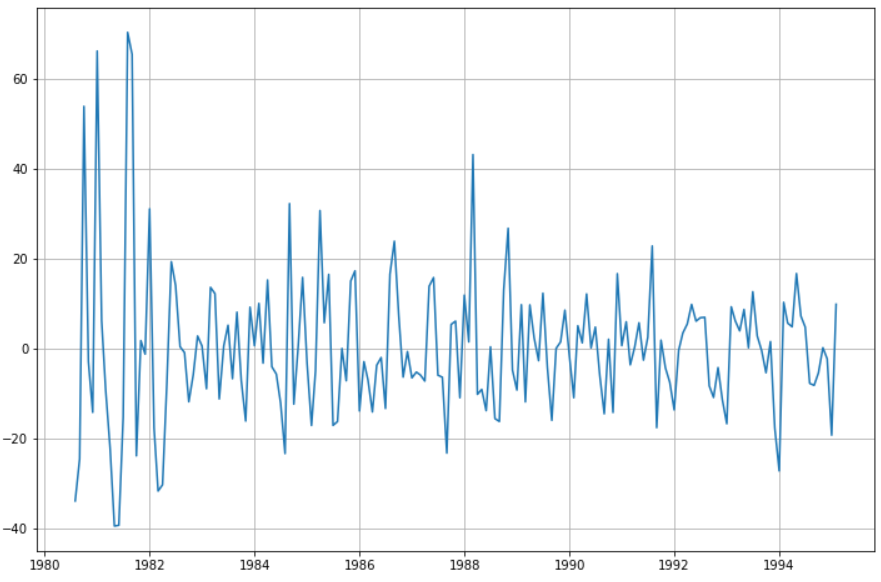
**Trend of the additive series:**



**Seasonality of the additive series:**



**Residual of the additive series:**

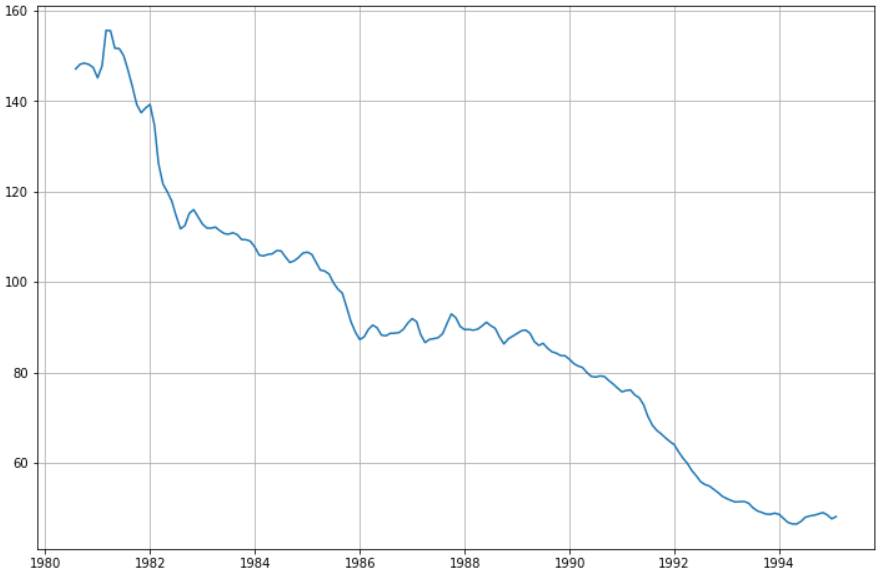


### Multiplicative decomposition:

### 

### In this the residual is centred around 1. Magnifying the trend, seasonal and residual components.

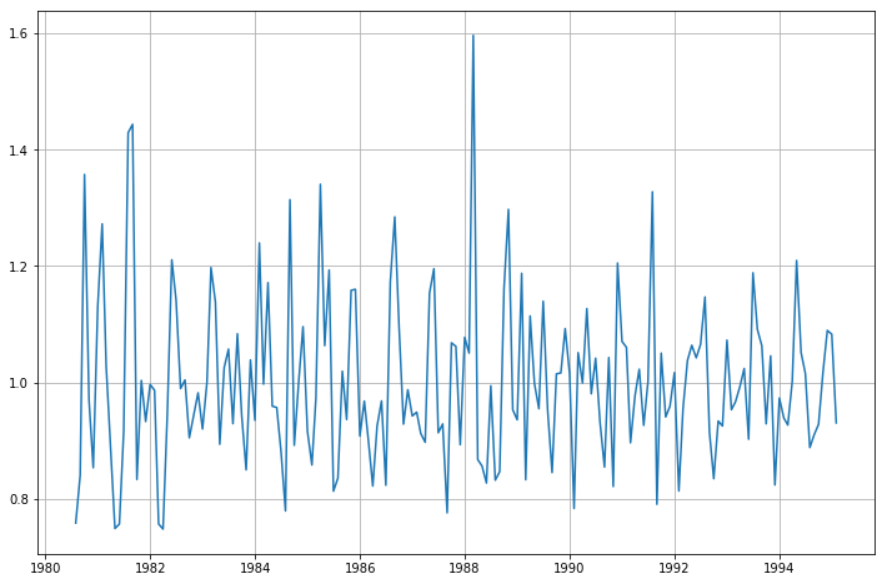
### Trend:



**Seasonality:**



**Residual:**

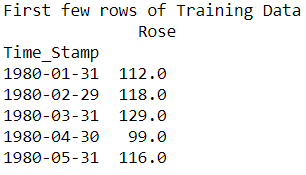


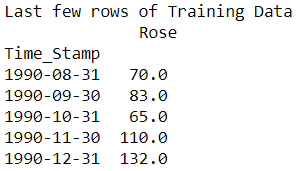
**Splitting the data into train and test:**

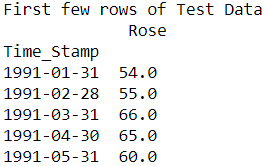
We have split the data into train and test such that the values before the year 1991 is considered as the train set and the values from the year 1991 is considered as the test set.

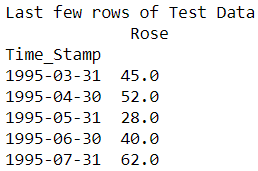
The shape of train and test sets look like this:

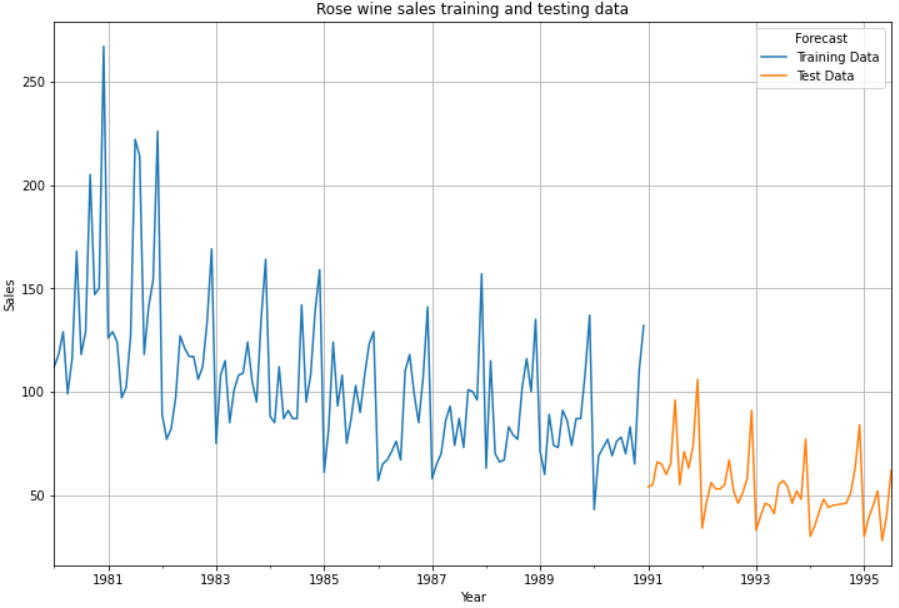




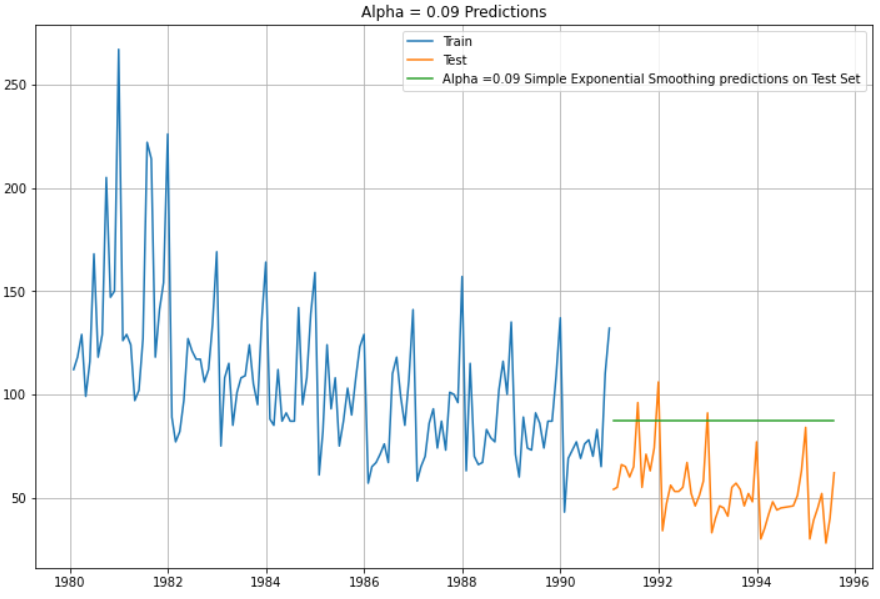








**Simple Exponential Smoothing model:**

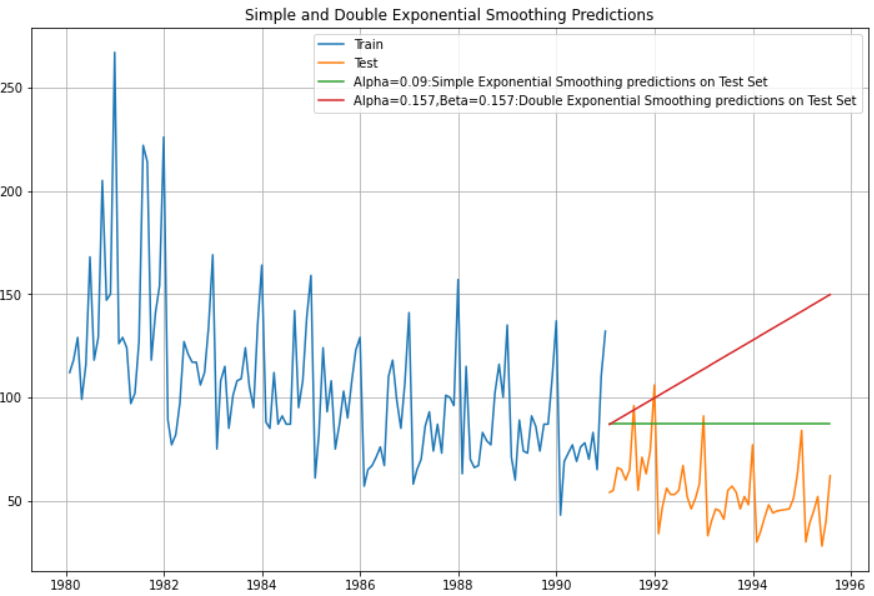


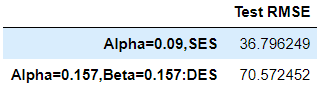
We see that the SES model gives us a straight line prediction. We check the RMSE value on the test data:



**Double Exponential Smoothing - Holt's linear method with additive errors**

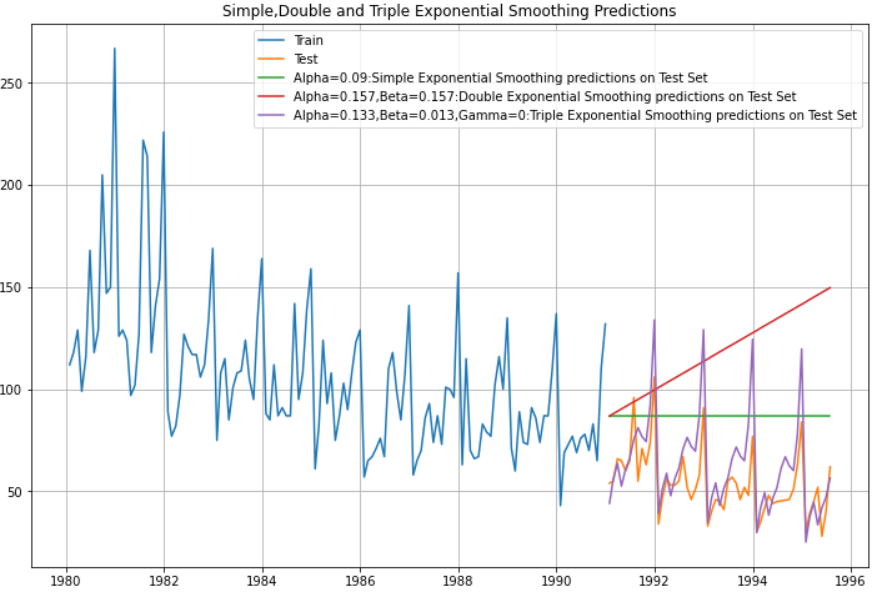
Apart from the alpha this model also considers the trend Beta in the model.



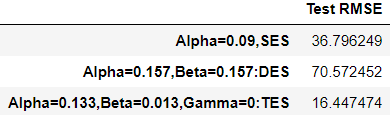


**Triple Exponential Smoothing model or Holt Winter's linear method:**

### ETS(A, A, A) : With additive trend and seasonality:



We see that the Triple Exponential Smoothing is picking up the seasonal component as well.



### ETS(A, A, M) model : Additive trend and multiplicative seasonality

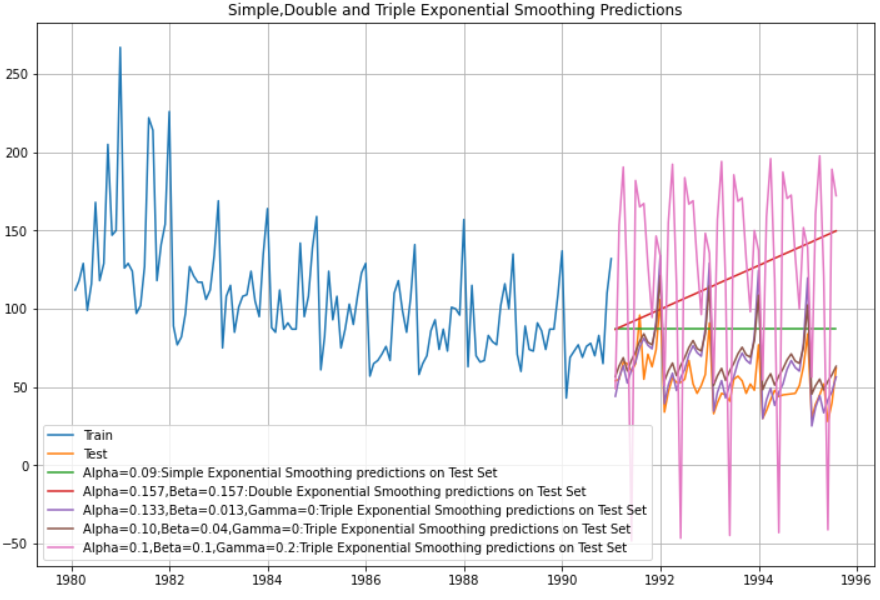
### 

### 

### Triple Exponential Smoothing model by taking the best alpha, beta and gamma [all in the range of 0.1 to 1 taking an interval of 0.1]

### We use the brute force in the TES model and also search the best value of the RMSE for different values of Alpha, Beta and Gamma in the given range. We get the best values as alpha = 0.1, Beta = 0.3 and Gamma = 0.2

### 



It can be seen that although this model has given us the least RMSE values so far, this model does not do well on the test data.

### Linear Regression model:

### Here is the brief snapshot of the train and test data with the respective time instances.

### 

### We see that this model has given us a lower RMSE value than the previous model.

### 

### Again, this model fails miserably on the prediction of the test data.

### Naive approach model:

### 

### The Naïve model gives us a Test RMSE value of 79.718 that is higher than any of the previous models so far.

### 

### This model as well, fails on the visual forecast of the test data.

### Simple Average model:

### 

### 

### The Simple Average model fails on the Test RMSE and on the visual forecast of the test data.

### Moving Average(MA) models:

### Although the moving average model was not asked for in the project questions, we have performed the model on this data to see its efficacy. We take the trailing means of 2, 4, 6 and 9.

### The head of the data looks like this:

### 

### We plot the moving average model variations on the entire data at first:

### 

### The 2 point trailing moving average looks promising based on the visual appeal.

### We split the data into train and test and then run the different moving average variations on the train data and evaluate the test data:

### 

### As expected, the 2 point trailing moving average model has performed very well on the test data.

### 

### Also, on the RMSE values of the test data, the 2 point trailing moving average model has performed exceptionally well.

### Checking for Stationarity of the whole time series data:

### 

### The test for stationarity is performed using the Dickey Fuller’s test:

### 

### 

### 

### 

### Automated version of an SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) and for a seasonality as 6:

### The loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary. The seasonal differencing 'D' will be between 0 and 1 to check if it is needed or not.

### The best combinations are listed as follows with the lowest AIC – Akaike Information Criteria

### 

### We run the SARIMAX model with the train data, order as (2,1,3) , seasonal order as (2,0,3,6), enforce stationarity as false as the stationary series is already being considered in the parameters entered in the order, enforce invertibility as False as the effect on the y variable alone is required and not the inverse of it on the X variable (time variable). We choose the maximum iterations as 1000. The SARIMAX results are as follows:

### 

### Based on the p values below 0.05, we can see that the customer count Autoregressive model of order 1 and 2 are both important. The seasonal autoregressive of second order is shown to be important. The AIC is 951.744.

### Diagnostic plot:

### 

### Prediction on the test set with mean values and lower confidence intervals and upper confidence intervals:

### 

### 

### The test RMSE for this model is 27.12.

### ACF and PACF on the time series data:

### 

### 

### 

### 

### Autocorrelation and Partial Autocorrelation plot with a higher lag count:

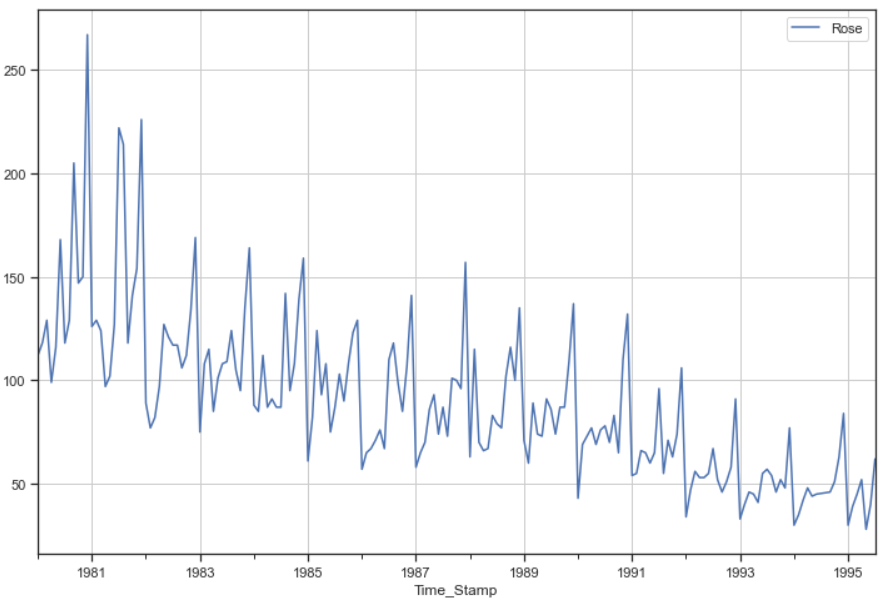
### 

### 

By looking at the above plots, we can say the following:

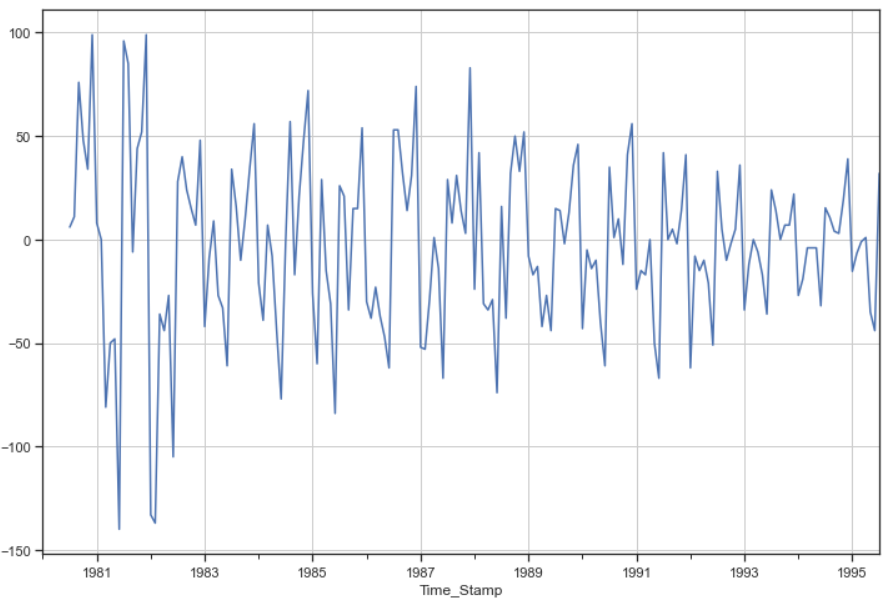
1. The ACF plot cuts off at lag 2. i.e. The Moving-Average parameter in an ARIMA model, q = 2
2. The PACF plot cuts off at lag 4. i.e. The Auto-Regressive parameter in an ARIMA mode, p = 4
3. We can also see there is a seasonality present and can be considered as either 6 or 12.

Let us plot the original data and the differenced series:



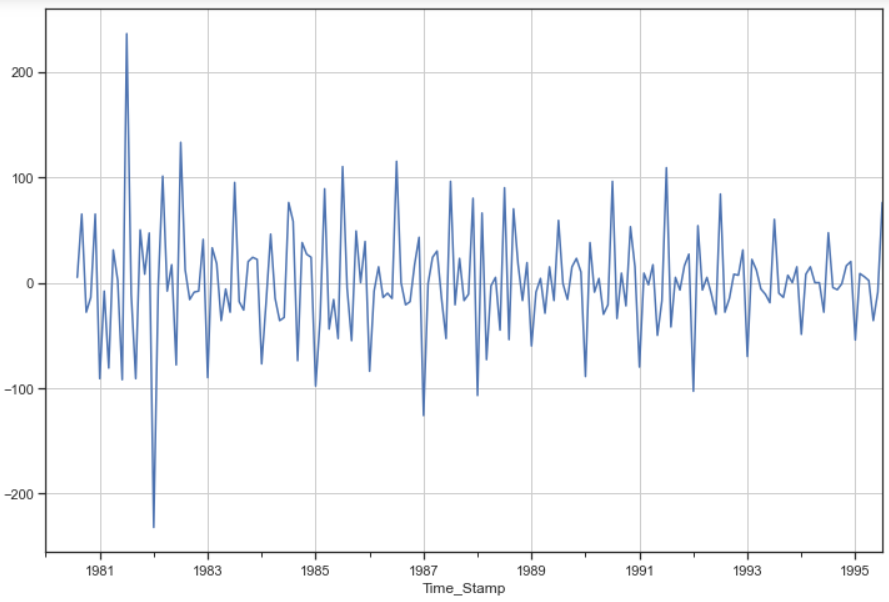
In the above plot, we can see there is both trend and seasonality.

Let us take a seasonal differencing and look at the series:

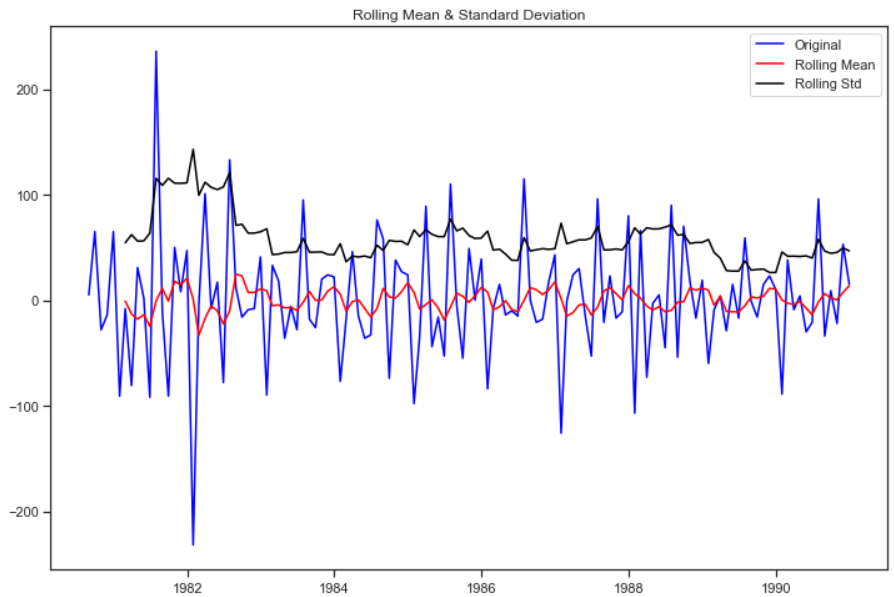


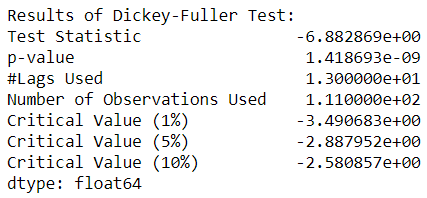
We can see there is a slight trend present in the data.

So, we take a differencing of first order on the seasonally differenced series.

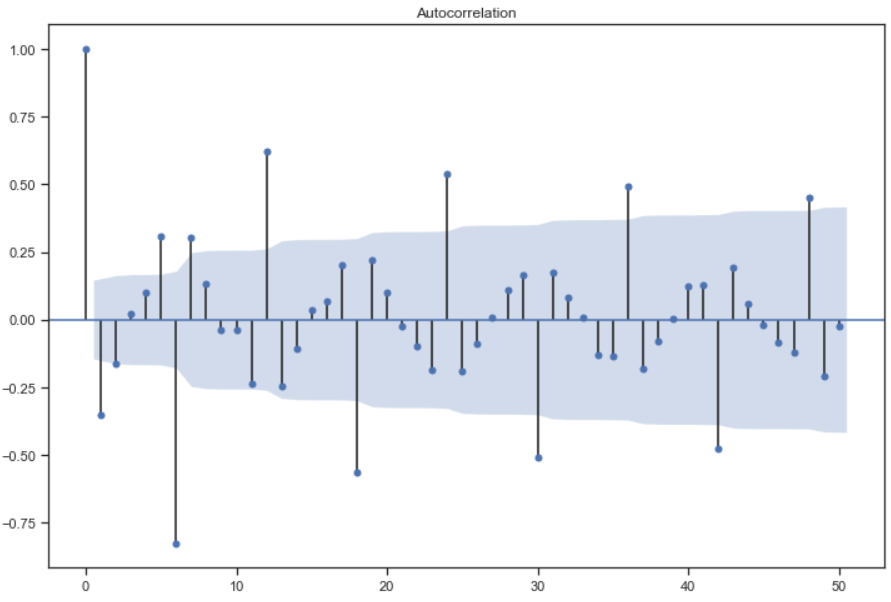


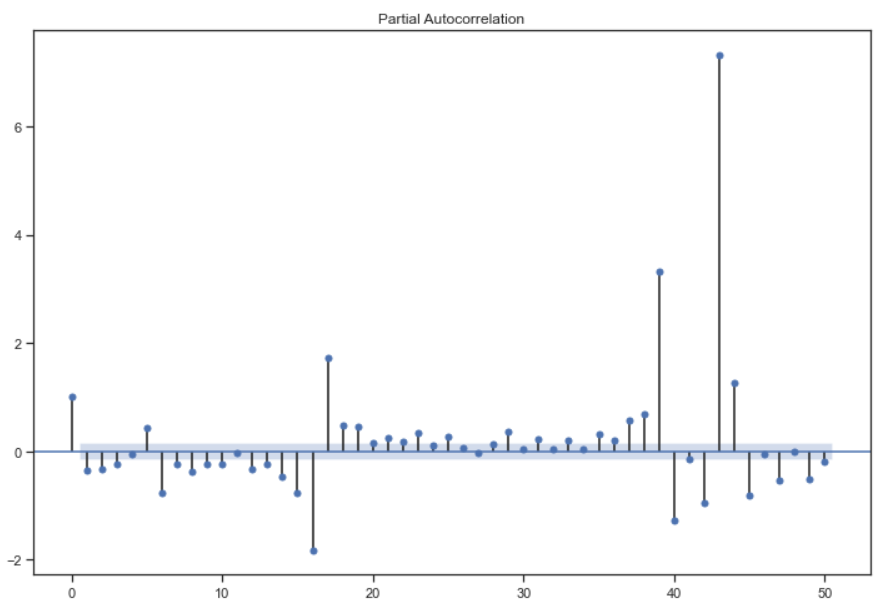
Checking for the stationarity of the above series:





Checking the new ACF and PACF plots for the new modified time series:

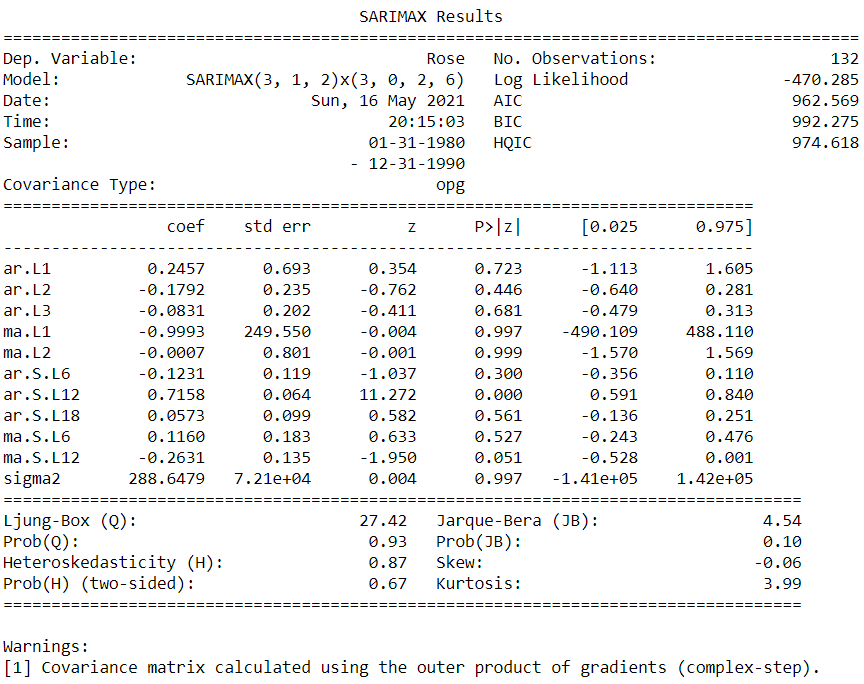




Here, P = 3 and Q is hard to interpreset as the lags in the ACF plot for seasonal differences do not cut off.

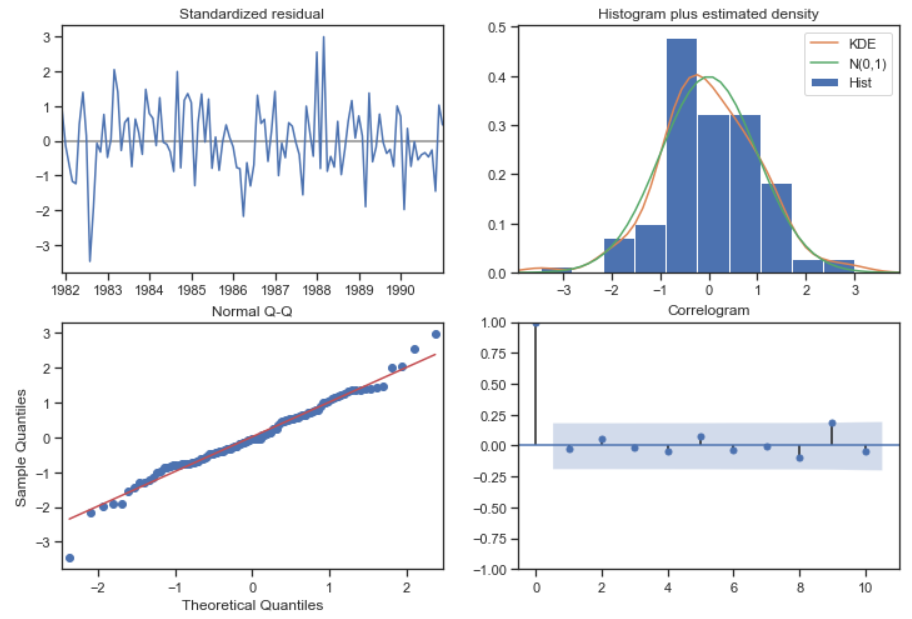
So, we will take a range of P and Q values and the q = 2 and p = 3

So running the SARIMAX model as (3,1,2)(3,0,2,6)



The AIC values have actually increased as compared to the previous iteration. This time the Auto regressive 2 intervals of seasonal count and the moving average 2 intervals of seasonal count have shown to be important based on the above p values.

Plot diagnostics:



### Predict on the Test Set using this model and evaluate the model:

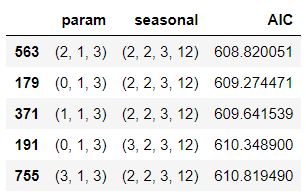
### 

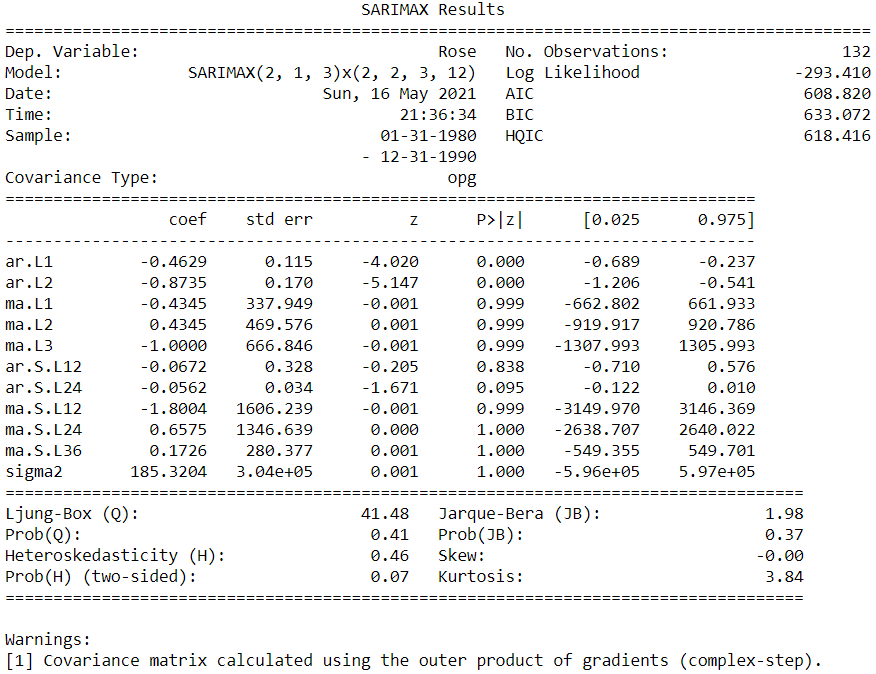
### Automated version of an SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC) and for a seasonality as 12:

### The following loop helps us in getting a combination of different parameters of p and q in the range of 0, 1,2 and 3. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary. The seasonal differencing 'D' will be between 0, 1 and 2.

### 

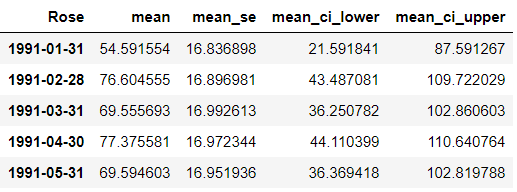
We then list the best combinations of the parameters based on the lowest AIC:





Predict on the Test Set using this model and evaluate the model:

The head of the mean predicted values, the upper confidence intervals and the lower confidence intervals are as follows:





### ACF and PACF plot of the time series data with the seasonality as 12:

### Seasonally differenced time series:

### 

### We can see there is a slight trend present in the data.

### So, we take a differencing of first order on the seasonally differenced series.

### 

### Checking for the stationarity of the above series:

### 

### 

### Checking the new ACF and PACF plots for the new modified time series:

### 

### 

### We run the SARIMAX model with the following parameters: Here, P = 2 and Q = 1 , p = 4 and q = 2.

### We have seasonally differenced twice, so D = 2.

### 

### The AIC is 712.992 in this combination which is higher than that of the previous model. In this model, the moving average one period count and two intervals of moving average period count are shown to be important. The Auto regressive seasonal one period count and two intervals of period count are shown to be important. The moving average seasonal count is also important. The sigma2 is also shown to be important in this case which is the estimate of the variance of the error term. Also, the Jarque-Bera (JB) test for the normality is shown, the hypothesis of the test is as follows:

### JB(PValue>0.05)= Accept Ho (Normal Distribution)

### JB(PValue<0.05)= Reject Ho (Non-Normal Distribution)

### In this case the p value obtained from the above model is 0.22 which is corresponding to the Normal distribution.

### Plot diagnostics:

### 

### Standardized residual: Residual mean is not so constant. It should be around 0. But as we had observed before the residual component in this data is pretty high so this is okay.

### Histogram plus estimated density: Residuals have an almost perfectly normal distribution and so this is okay.

### Normal Q-Q: All the data points should fall in the line of the Q-Q. They should be linear. The above plot satisfies the beforementioned conditions and hence looks okay.

### Correlogram: The residuals themselves should not have any autocorrelation. They should not have any significant lags. As observed, they do not.

### If any of this is violated, then we cannot claim that the point forecast: upper confidence level and lower confidence level is true 95% of the time.

### Predict on the Test Set using this model and evaluate the model

### 

### 

### Building the most optimum model on the full data:

### 

### The AIC values have shot up to 1303 when forecasting for 12 months into the future. The only two components that are important in this iteration are autoregressive 3 intervals of period count, moving average one interval of period count, autoregressive seasonal with one, two and three intervals of period count. Moving average seasonal count is also important and the sigma2 important in this iteration.

### Plot diagnostics:

### 

### Evaluate the model on the whole and predict 12 months into the future:

### Following is the head of the forecasted values with the mean values, mean standard error, mean lower CI and mean upper CI values:

### 

### 

### The forecast of the model for 12 months into the future along with the confidence intervals are shown in the following plot as follows:

### 

### Table listing of all the models with their respective RMSE values on the test data:

### 

**Observations and suggestions:**

* The probability of the Jarque-Bera test is 0.01 thus indicating that the distribution is normal.
* The standardized residual plot has a high error variation. The histogram plus estimated density shows a perfectly normal distribution. The normal Q-Q plot is okay as all the data points are lying in the line of the Q-Q. They are linear. The correlogram shows us that the residuals do not have any autocorrelation. They do not have any significant lags.
* The only two components that are important in this iteration are autoregressive seasonal with one and two intervals of period count. Even the sigma2 is not important in this iteration.
* Prob(H) (Two-sided) Heteroskedasticity is 0 and hence there is heteroskedasticity present in the time series. This means that the residuals are not distributed with equal variance. The null hypothesis of this test shows that there is no heteroskedasticity and the alternate hypothesis shows us that there is. At 95% confidence level, we can say that there is heteroskedasticity present. This comes as a bit of a surprise as the sigma2 value is shown to be non-important and thus does indicate that the variance of the error term is not significant. This means that the model thus built is unable to explain some pattern in the response ‘y’ variable which in this case is our Rose wine sales. ARCH modelling is the method proposed to reduce Heteroskedasticity. Reference: Datascienceplus, towardssciencedirect.
* The inferences from the EDA have been mentioned previously in the appropriate sections and are to be taken a closer look at to identify any reason for the spike in sales. If possible, the same methods can be adopted with a view to increase the sales.
* This calls for careful investigation on the periods where the spike in the Rose wine sales was observed. The spike could have been the result of any one or many of the following factors, which is to be considered in no particular order: The composition of the Rose wine during that period, the offer price with respect to the economic conditions prevailing during that period, the promotional strategies during that period, the location of the stores during that period and the sales in each of those stores, the medium of advertising and the ambassador for endorsements, government policies during that period, the management team during that period, etc.
* The Rose wine needs to be offered at a discounted price especially during the last quarter of every year to increase sales even more than what they already are. Another way to look at this would be to increase the Rose wine price during the fourth quarter of every year as there is surely an increase in sales during this period, thus the income would increase.
* Some important variables to be included in such an analysis are: Location of the store, sales with respect to each location / region / store. The economic condition of the people in that location as the low economic folks tend to drink a low-priced wine as opposed to the high economic ones.