Project: Time Series Forecasting (Sparkling wine)

Name- Manisha Rout PGP-DSBA Online

Date: 17.07.2022

Table of Content

|  |  |  |
| --- | --- | --- |
|  | Q/A | Page no |
| 1 | Read the data as an appropriate Time Series data and plot the data. | 4-5 |
| 2 | Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. | 5-12 |
| 3 | Split the data into training and test. The test data should start in 1991. | 12-13 |
| 4 | Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE. | 13-25 |
| 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 | 25-28 |
| 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. | 29-34 |
| 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 | 34-38 |
| 8 | Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data. | 38-39 |
| 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. | 39-43 |
| 10 | Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. | 45-49 |

List of Figures

1. Time Series forecasting Plot
2. Decomposition Plot
3. Box Plot
4. Line Plot
5. ACF and PACF plot
6. Diagnostic Plot

List of Tables

1. Summary tables for SARIMA and ARIMA
2. Tables containing RMSE scores

Problem 1.1

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

* The data has been Time stamped from the year 1980 to 1995, It’s a Year-Month dataset with corresponding sales of sparkling wine
* Data had been collected from January 1980 to July 1995. Around 15.5yrs data had been collected
* It has 187 rows and 1 column
* ‘Sparkling’ variable of data is of integer type
* There are no missing values present in the data
* No anomalies are present in the data. Overall, the data seems good for Model building

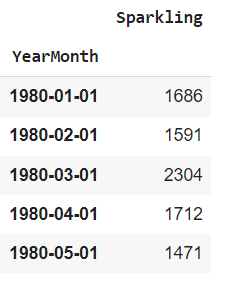


Figure : 1st five instances of Data

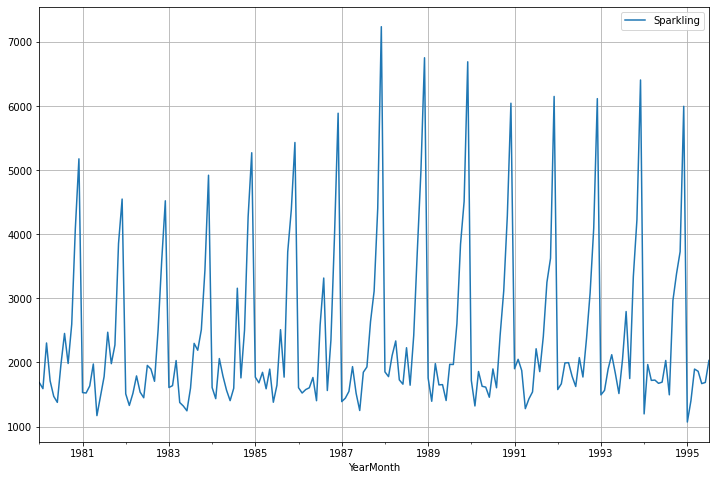
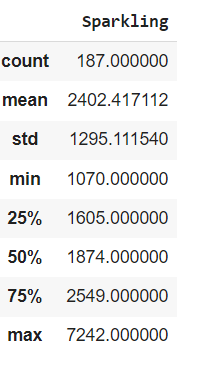


Figure : Initial Time Series graph

* From the initial Time series plot It can be inferred that; Trend might be absent but a Seasonal effect seems to be present
* Sales in the year 1988 seem to be maximum

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

* From the descriptive statistics, It can be seen that the average number of sales is approx. 2402, the Maximum number of sales is 7242 and the minimum number of sales is 1070



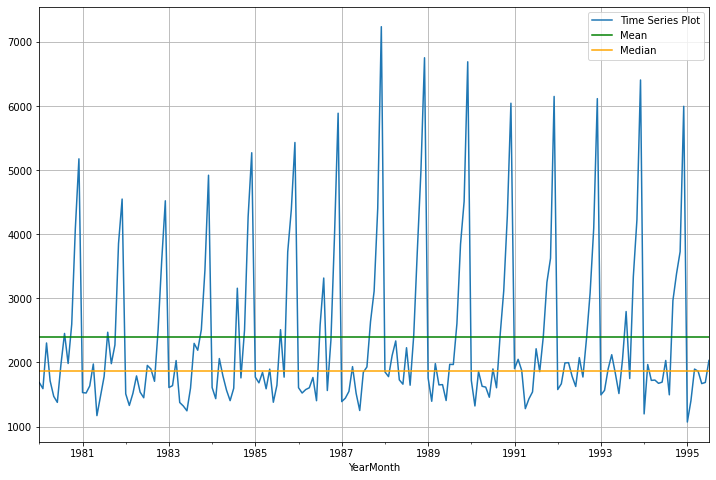


Figure : Time Series Plot Mean & Median of Sales of Sparkling wine

* The data doesn’t have any missing values.

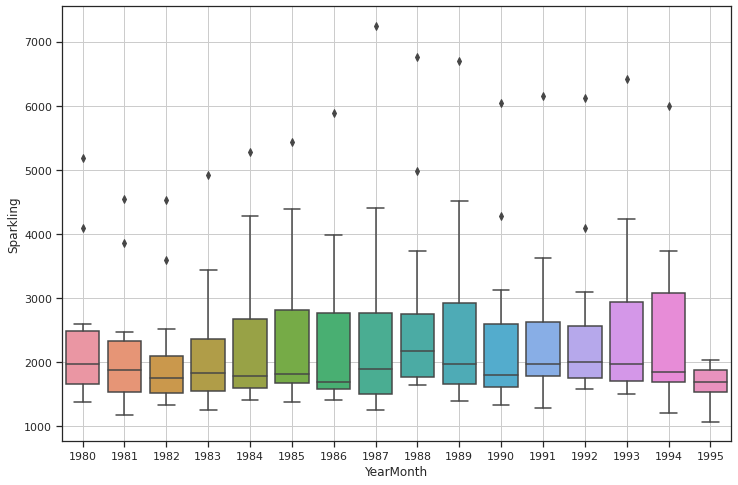


Figure : Box Plot per Year wise

* From the box plot It can be inferred that, Across the year, there is not much noticeable difference, though there are a few high ranges in the middle year. Also, the last year is showing fewer sales due to the fact that the data is recorded for 7 out of 12 months in the year 1995

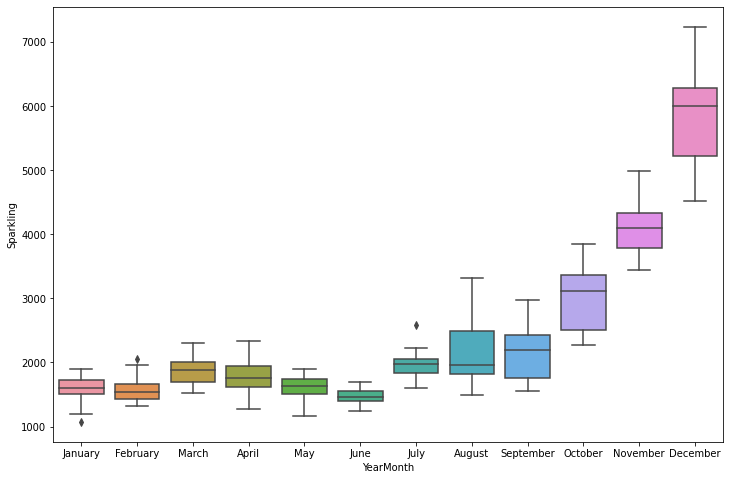


Figure : Boxplot per month wise

* It can be inferred from the boxplot per month wise that, the maximum sale of sparkling wine is in the month of November and December and the minimum in the month of the. Sales have been increasing during the winter months, the dramatic sales lift can be attributed to the months being holiday months (such as Christmas, New Year). Variance and the mean values for the winter months are also much higher than any of the other months. Considering the months from the year 1980 to 1995 July, it can be inferred that seasonality is present in the data. Outliers present in the month of July

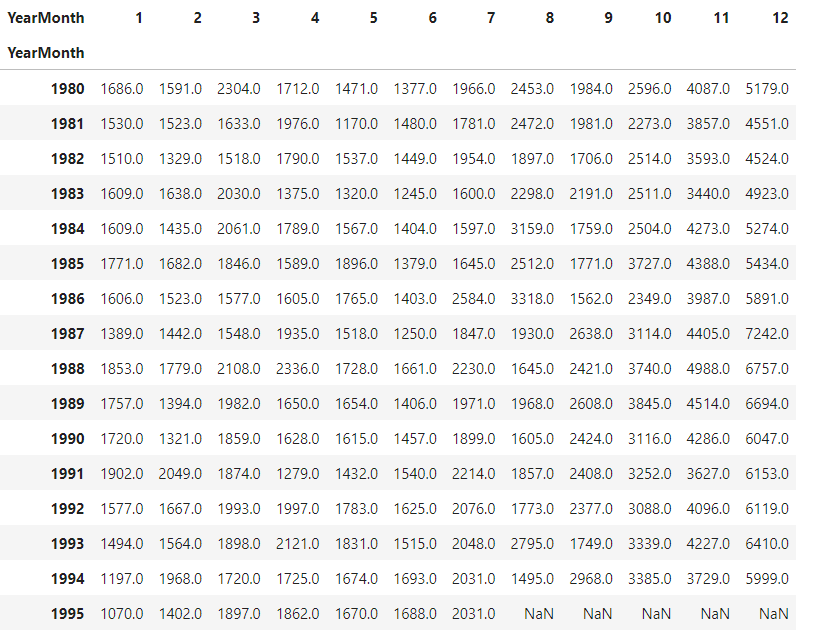


Figure : Years-Month Sales table

* Monthly Sales across years also shows seasonality

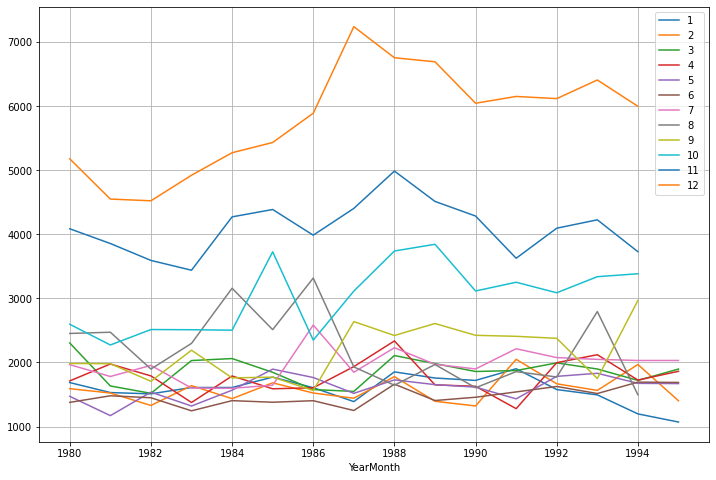
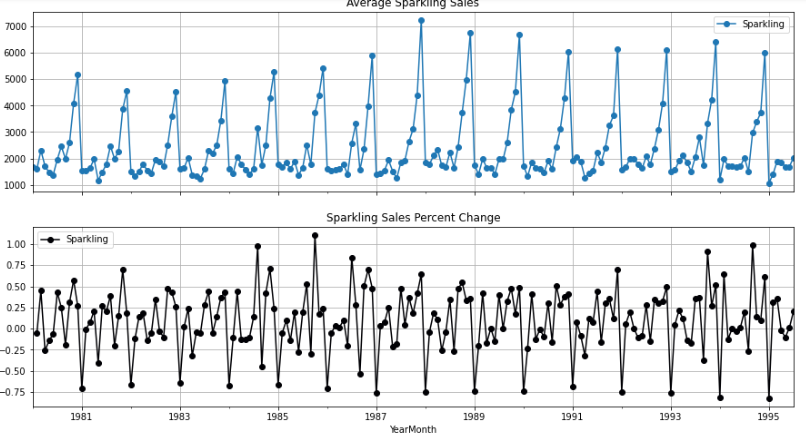


Figure : year- Month Plot

* Month Plot for each month Distribution also shows that December has the highest sales across years



* Average ' Sparkling Sales' and the Percentage change of 'Sparkling Sales' with respect to the time shows lot of seasonal variation every year.
* Additive Decomposition

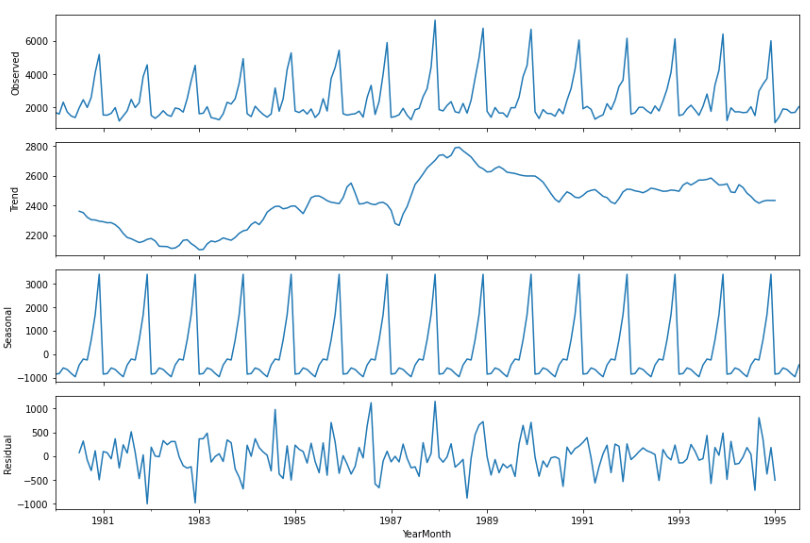


Figure : Additive Decomposition

* From the additive decomposition, it can be inferred that seasonality is present in the data, no distinct trend is present in the data; However, it seems like there is an increase in the sales during the 1st 7yrs and then it moves almost to stationary to downward position. Residual shows patterns.
* Multiplicative Decomposition

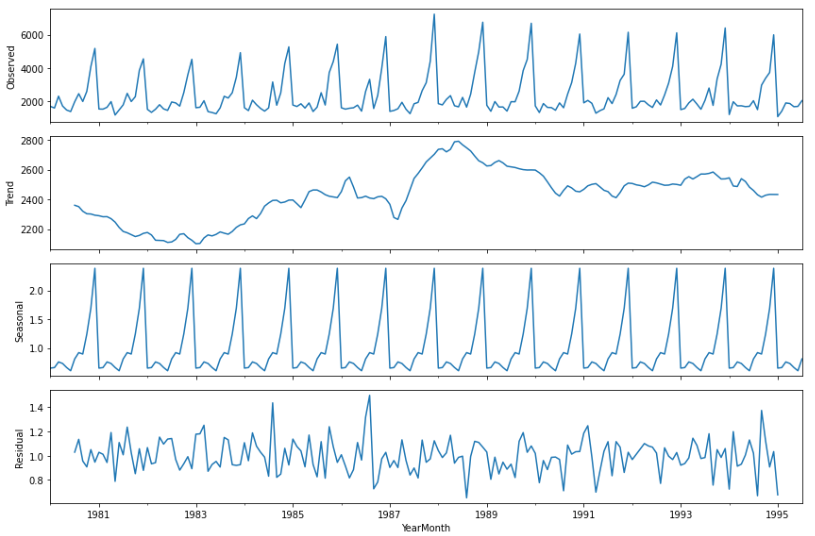


Figure : Multiplicative decomposition

* From the multiplicative decomposition, it can be inferred that seasonality present in the data, no distinct trend presents in the data; However, it seems like there is an increase in the sales during the 1st 7yrs and then it moves almost to stationary to a downward position. Residual shows patterns.
* Hence, either of the methods is suitable for analysis

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

* For the time series analysis, the data should be in continuous form. Any randomization in the train-test split will not be helpful in model building. Hence, the train-test splits require a specific number/ year/ month for model building.
* It is also important to mention that, In time series analysis It is difficult to predict future observations if such an instance has not happened in the past. Train- Test split will predict likewise behavior as compared to the past year
* Two new dataframe have been created with a train set containing sales information from 1980 to 1991 and test data containing sales information from 1991 onwards.
* Shape of the train set = (132, 1) and shape of the test set= (55, 1)

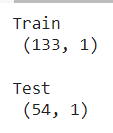


Figure : train-test split

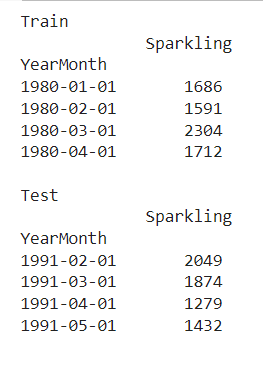


Figure : train and test data frame

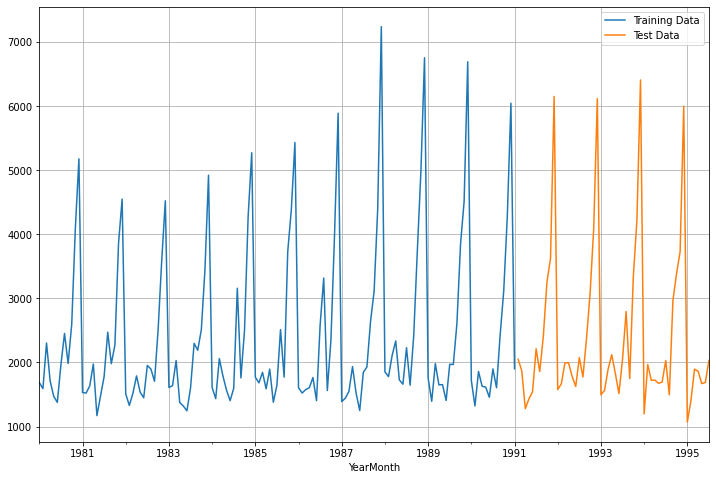


Figure : Joint Plot showing Train test Split

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

**Linear Regression**

* For this particular linear regression, we have regressed the 'Sales' variable against the order of the occurrence. The training data has been modified before fitting it into a linear regression
* The numerical time instance order for both the training set and test has been generated
* A new column time (containing the numeric instances) has been added to the training as well as test data
* Linear Regression imported from Sklearn packages and fitted to training data to build the model
* Performance of the test data has been checked by using the Root mean square error value.
* RMSE value for the Linear Regression model is 1381.320, which is a very high value. As we know smaller the RMSE value, the better the model. Hence, this is not a very decent model for prediction

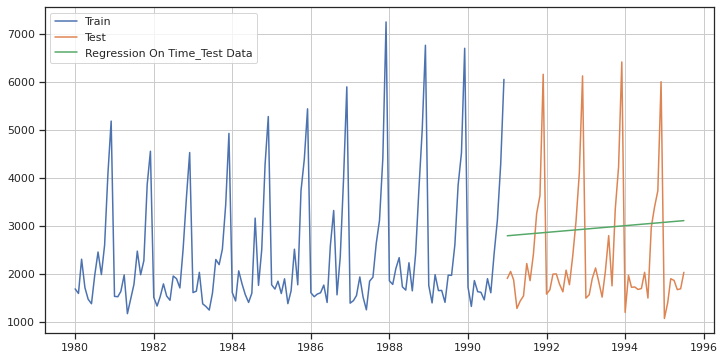


Figure : Linear Regression Model

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a flat line with slight steep

**Naive Model**

* For this particular approach, the prediction for tomorrow is the same as today and the prediction for the day after tomorrow is tomorrow since the prediction of tomorrow is the same as today. So, the prediction for the day after tomorrow is also today.
* Numerical time instance order for both the training set and test is not required for this approach
* Performance of the test data has been checked by using the Root mean square error value.
* RMSE value for the Naïve Approach is 1381.177, which is a very high value. As we know smaller the RMSE value, the better the model. Hence, this is not a very decent model for prediction

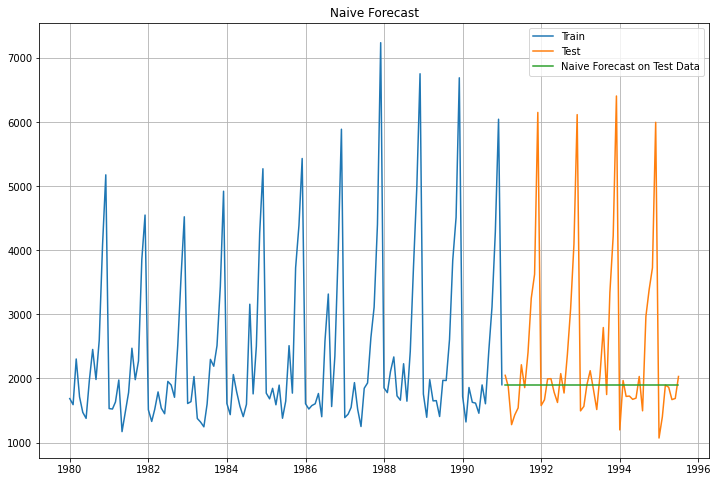


Figure : Naive Approach Model

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a horizontal flat line

**Simple Average**

* For this particular approach, the average of the training data used for the forecast
* Numerical time instance order for both the training set and test is not required for this approach
* A new column time (containing the simple average of training data) has been added to the test data
* Performance of the test data has been checked by using the Root mean square error value.
* RMSE value for the Naïve Approach is 1285.039, which is a very high value. As we know smaller the RMSE value, the better the model. Hence, this is not a very decent model for prediction

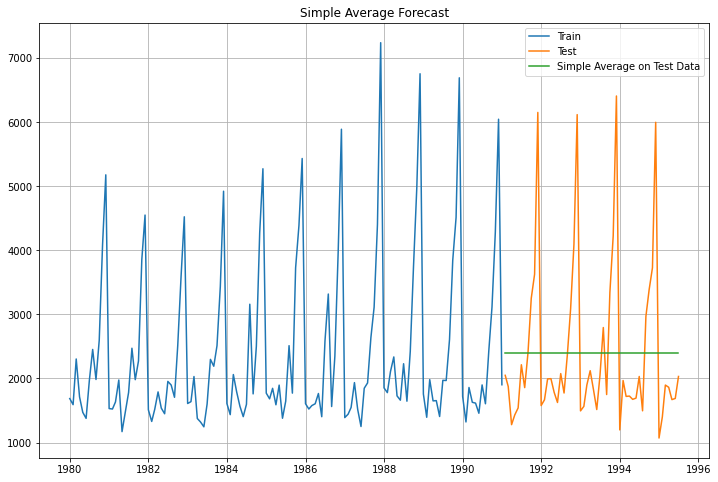


Figure : Simple Average method

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a horizontal flat line

**Moving Average**

* For this particular approach, rolling means of different intervals have been calculated
* The best interval can be determined by the maximum accuracy
* For this particular approach data has not been split initially and the model has been built on original data
* 4 new columns (containing the rolling mean) have been added to the original data

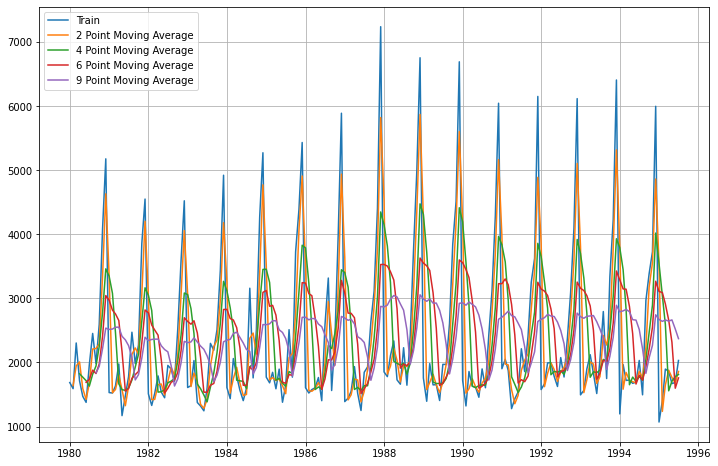
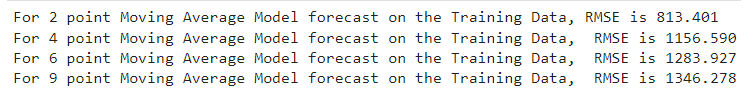


Figure : Moving average on entire data

* Since, there are multiple colors line present in the data, it’s difficult to evaluate which moving average did a better job
* Hence, the data has been split into training- test data, with training data containing information about sales from 1980 to 1991 and test data from 1991 onwards
* RMSE scores for each moving average model have been calculated



Based on the RMSE scores 2-point rolling has done a better job compared to all other models with an RMSE score of 813.401, which is significantly lower than all other scores. As we know smaller the RMSE score, the better the model. It can be inferred that Rolling 2 model is a decent model and can be used for prediction

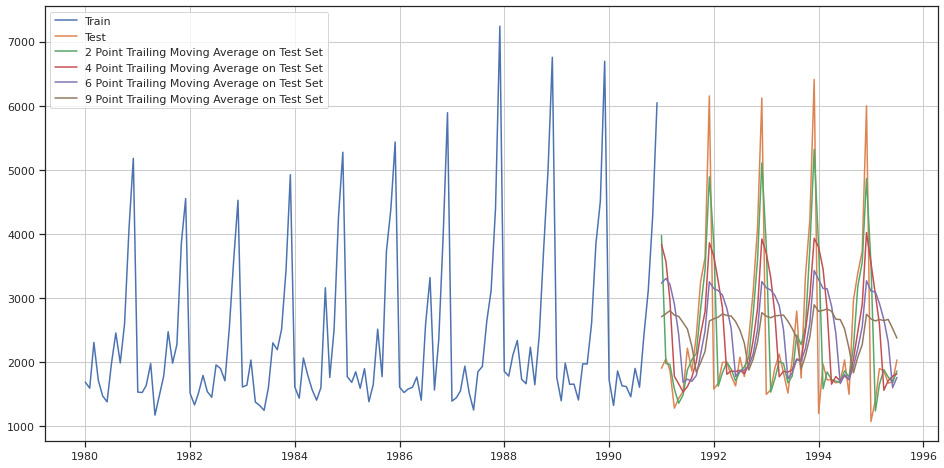


Figure : Moving average model with train test split

* It is also seen from the time series plot that it (green line) does look like it has done very well forecast against the original data compared to other lines

**Comparison of Time Series Plot (Linear Regression, Naïve Approach, Simple Average, Moving Average model)**

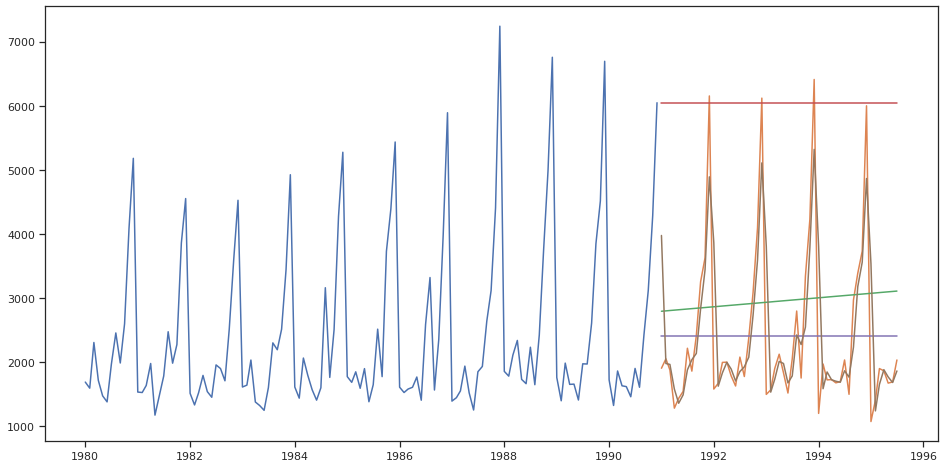
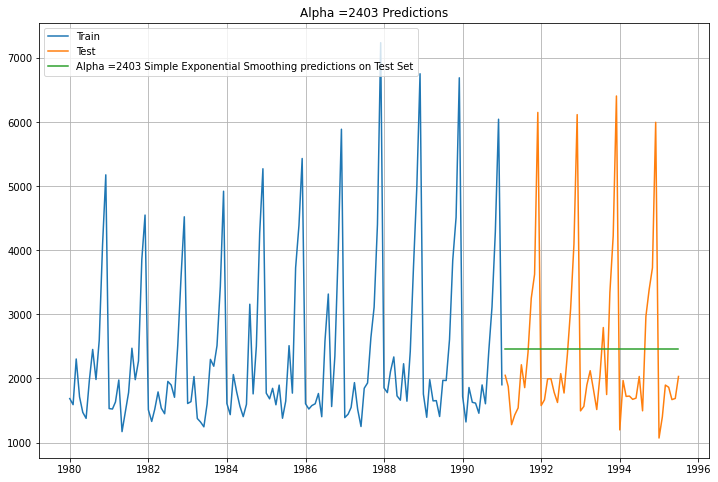
****

Figure : Comparison model

**Simple Exponential Model**

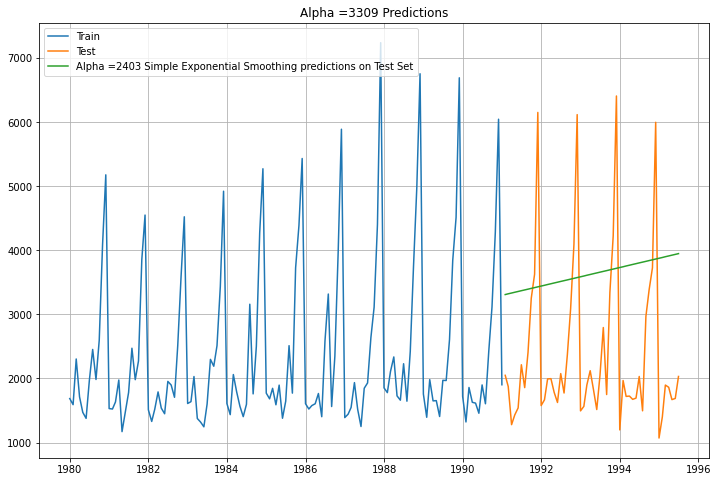
* For this particular approach, it is assumed that the data set has no trends or no seasonality. However, It has only a level also known as alpha present in the data
* Initialization method used as estimated and optimization set as true, the smoothing level or alpha obtained, which is very less 0.01, which means the previous period demand is not very accurate to forecast for the next period
* It is a flat forecast, hence forecasting it on test data gave a constant value
* RMSE value for the Simple exponential smoothing is 1285.771, which is a very high value. As we know smaller the RMSE value, the better the model. Hence, this is not a very decent model for prediction



* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a horizontal flat line

**Double Exponential Model**

* For this particular approach, it is assumed that the data set has trends and levels but seasonality.
* Initialization method used as estimated and optimization set as true, the smoothing level or alpha obtained, which is very less 0.53, which means the previous period demand is a good parameter to forecast for the next period. The beta value is 0.00 means no trend present in the data
* RMSE value for the Simple exponential smoothing is 1780.013, which is a very high value. As we know smaller the RMSE value, the better the model. Hence, this is not a very decent model for prediction



* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a flat steep line.

**Triple Exponential Model**

* For this particular approach, it is assumed that the data set has trends, levels, seasonality
* Initialization method used as estimated and optimization set as true, the smoothing level or alpha obtained, which is very less 0.147, which means the previous period demand is a good parameter to forecast for the next period. The beta value is 2.11 e-30and the gamma value is 0.368
* RMSE value for the Simple exponential smoothing is 427.963, which is a small value. As we know smaller the RMSE value, the better the model. Hence, this is a very decent model for prediction

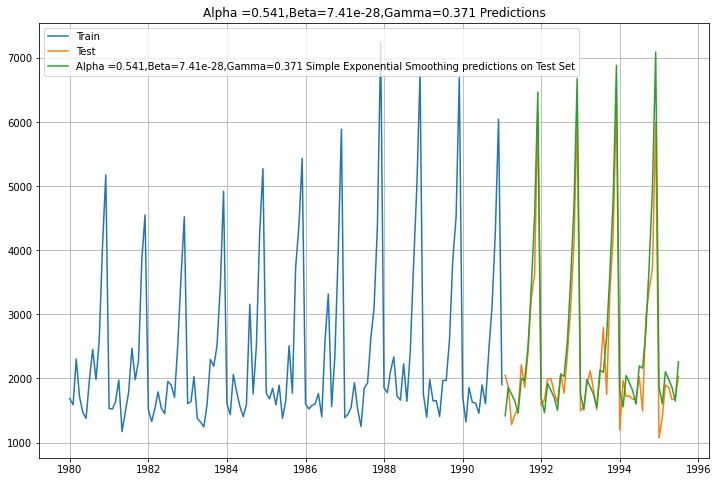


Figure : Triple Exponential smoothing

* It is also seen from the time series plot that it (green line) does look like it has done very well forecast against the original data compared to other lines

**Model Performance by putting the alpha value manually in simple exponential smoothing**

* Range of alpha value has been given from 0.3 to 1 with a difference of 0.1. RMSE value for alpha = 0.3 is minimum compared to all other alpha values
* However, the RMSE value for alpha = 0.3 is itself a large value is 1289.773. Hence, this is not a very decent model for prediction

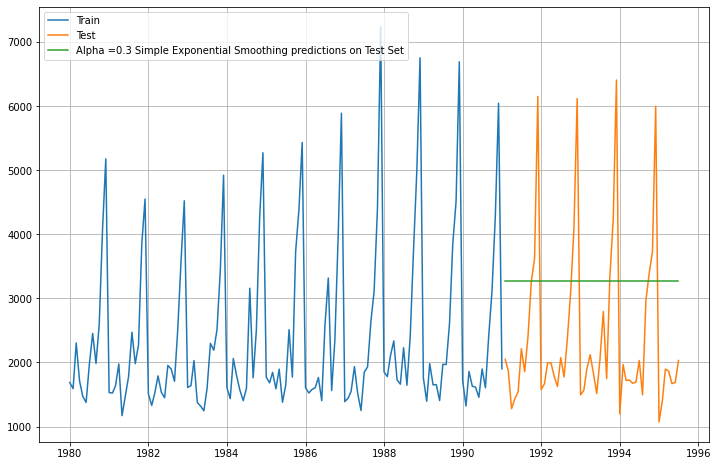
****

Figure : Simple exponential smoothing with alpha value 0.3

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a horizontal flat line

**Model Performance by putting the alpha and beta value manually in double exponential smoothing**

* Range of alpha value and beta value have been given from 0.3 to 1 with a difference of 0.1. RMSE value for alpha = 0.6 and beta=0.4 is minimum compared to all other alpha and beta values
* However, the RMSE value for alpha =0.6 and beta= 0.4 is itself a large value is 1656.603 Hence, this is not a very decent model for prediction

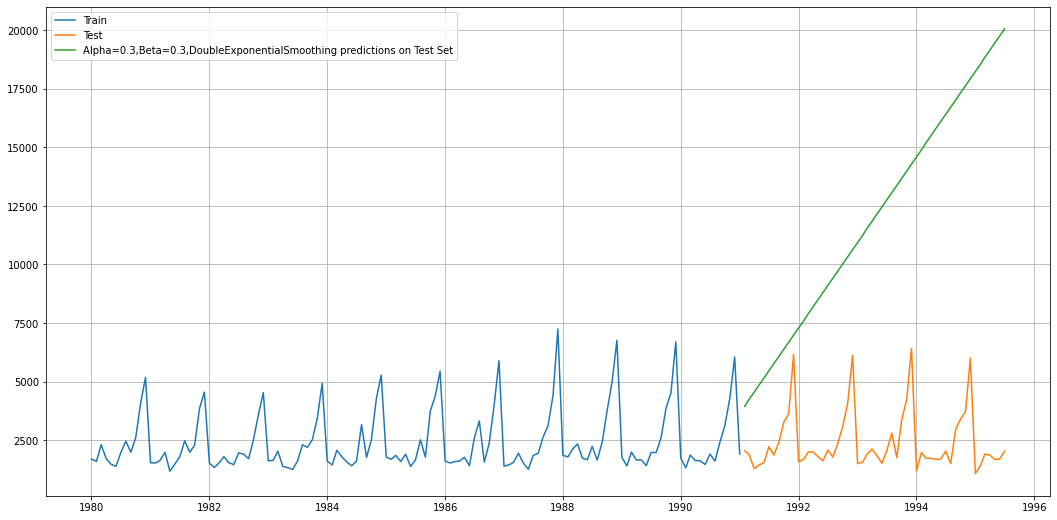
****

Figure : DES with manual alpha and beta value

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data, because it’s a horizontal flat line

**Model Performance by putting the alpha, beta, and gamma values manually in triple exponential smoothing**

* Range of alpha value, beta value, and gamma value has been given from 0.3 to 1 with a difference of 0.1. RMSE value for alpha = 0.8, beta=0.9 and gamma=0.3 is minimum compared to all other alpha, beta, and gamma values
* However, the RMSE value for alpha =0.8, beta= 0.9, and is gamma= 0.3 is a small value 4.323879e+02. Hence, this is a very decent model for prediction

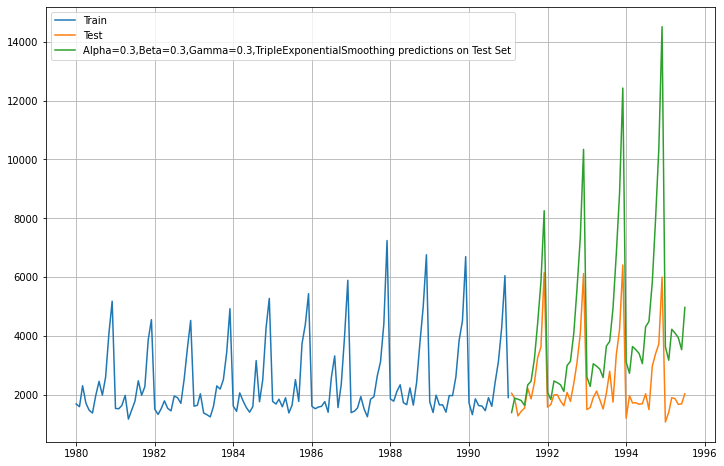
****

Figure : Triple exponential smoothing

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data.
* Brute force triple exponential smoothing gives the with alpha value= 0.3, beta value= 0.4 and gamma value= 0.3 gives lower rmse value. 458.683

**Final Table of RMSE score of all the models**

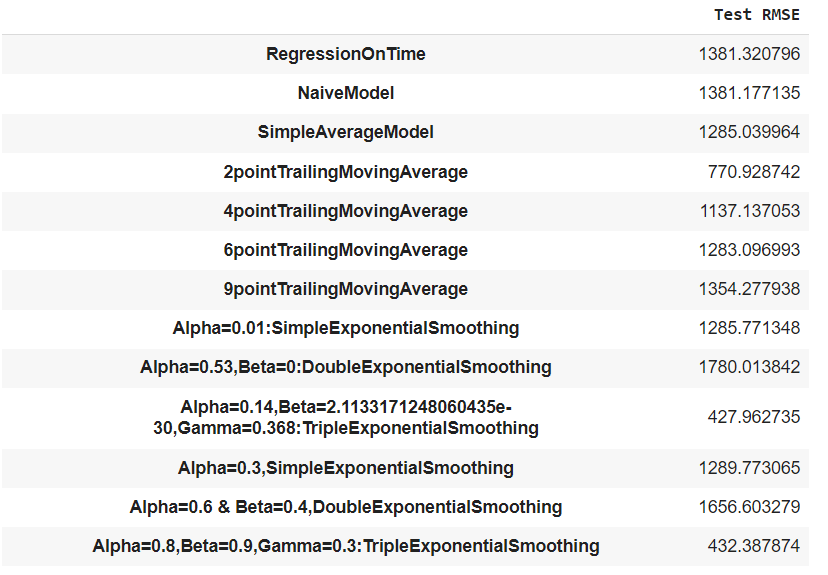
****

Figure : RMSE score of different models

**Conclusion**- From the above model we can safely assume that Triple Exponential smoothing with Alpha=0.14, Beta=2.1133171248060435e-30, Gamma=0.368 gives the minimum RMSE value which is 427.96. Lower the RMSE, the better the prediction/ forecasting

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

* It is a prerequisite for the ARIMA model is to have a stationary time series. From the Decomposition plot, it can be seen that the time series has seasonality but no distinct trend.
* To check the stationarity, we have to conduct a statistical test known as the Augmented Dickey-Fuller test, which is a unit root test that determines whether there is a unit root and subsequently whether the series is non-stationary
* The hypothesis for the ADF test is as follows:
* Null hypothesis- H0: Time series is not stationary
* Alternative hypothesis- H1: Time series is stationary

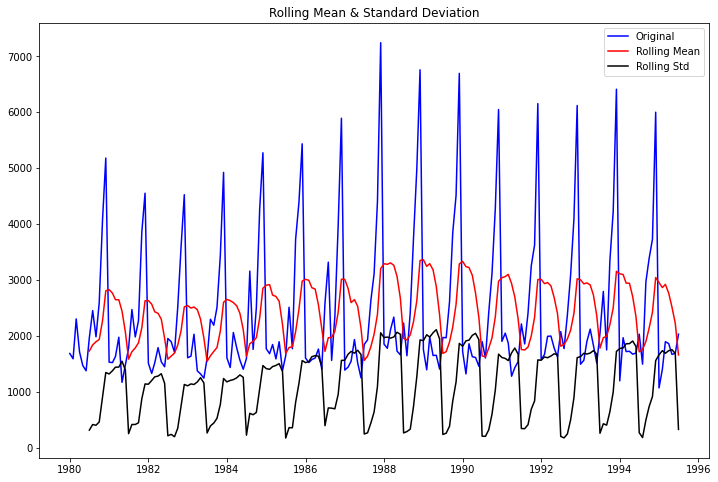


Figure : Stationarity check for ADF

* After performing the ADF test, the p-value obtained= 0.601 which is greater than 0.05, we failed to reject the null hypothesis. Hence, our data is not stationary. The test shows that the test statistic is greater than the 1% value. This shows that we cannot reject the null hypothesis and the series is non-stationary. The plot shows that the rolling means, as well as the standard deviation, is not constant, this gives us a hint of non-stationarity. Let us confirm it by using the Dicky Fuller Test.
* After taking a difference of order 1 and checking whether the Time series is stationary or not. The differencing method used to make the non-stationary time series to stationarity

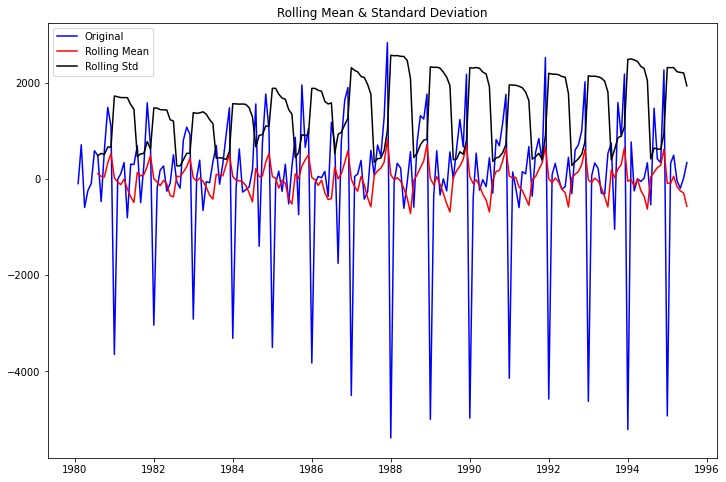
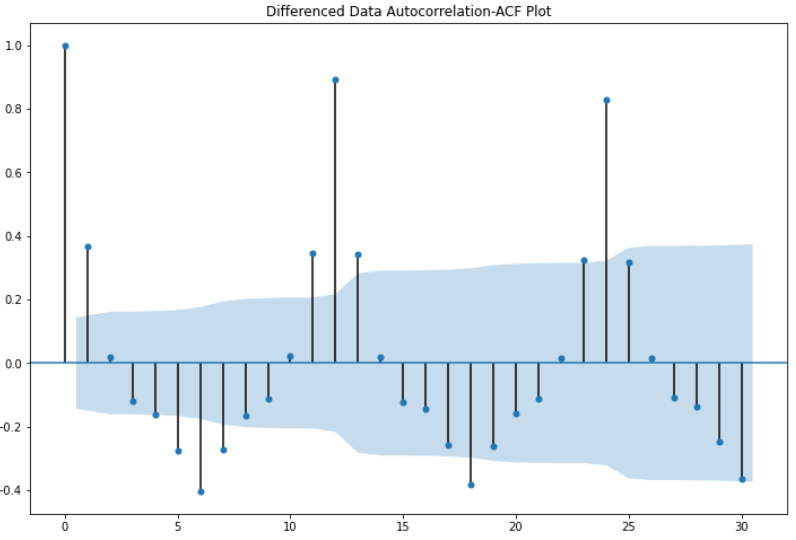
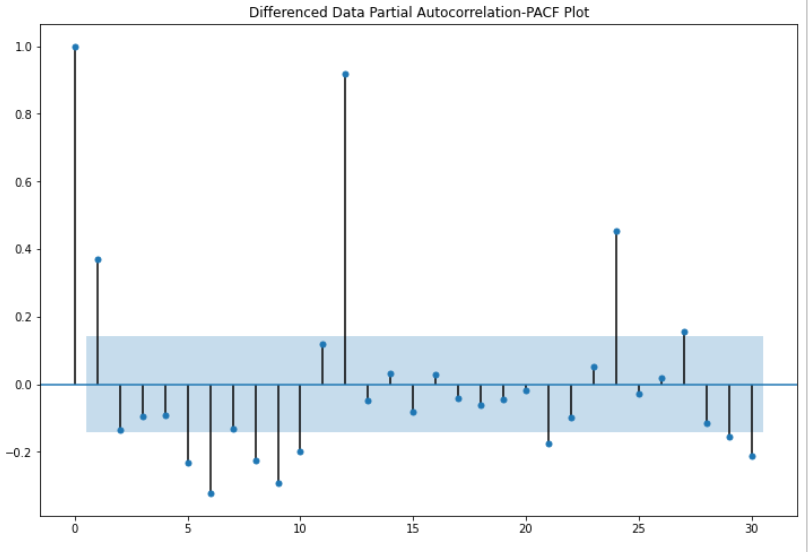


Figure : Differencing method for stationarity check

* After performing the ADF test, the p-value obtained= 0.00 which is less than 0.05, we are accepting the alternative hypothesis. Hence, our data is stationary after differencing. The test shows that the test statistic is lesser than the 1% value.
* Stationarity has been removed from the time series, ACF, and PACF plots allows us to decide the parameters of our ARIMA model.





1. 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**.**

**Auto-ARIMA model using AIC**

* ARIMA model- Autoregressive integrated moving average model where the p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and. q is the number of lagged forecast errors in the prediction equation.
* For the ARIMA model building, the range of p and q have been taken as 0 to 4 and for d range is 1 to 2. From these parameters, different combinations of the models have been created.
* Akaike Information has been obtained from these models. It basically checks the goodness of fit, and the simplicity/parsimony, of the model into a single statistic.
* When comparing two models, the one with the lower AIC is generally “better”
* For the value of pdq= (2,1,2) gives the best AIC value, which is equal to 2231.214 compared to all other combinations.

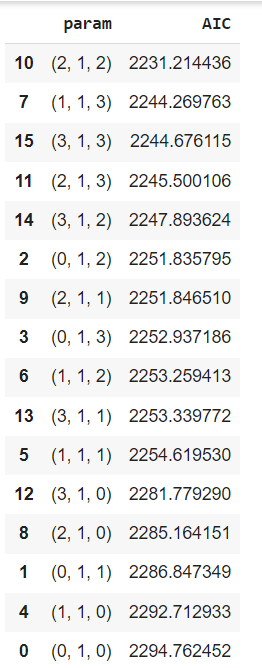


Figure : Best Param for ARIMA based on AIC

* Further the ARIMA model built on the training data set with the pdq value (212).

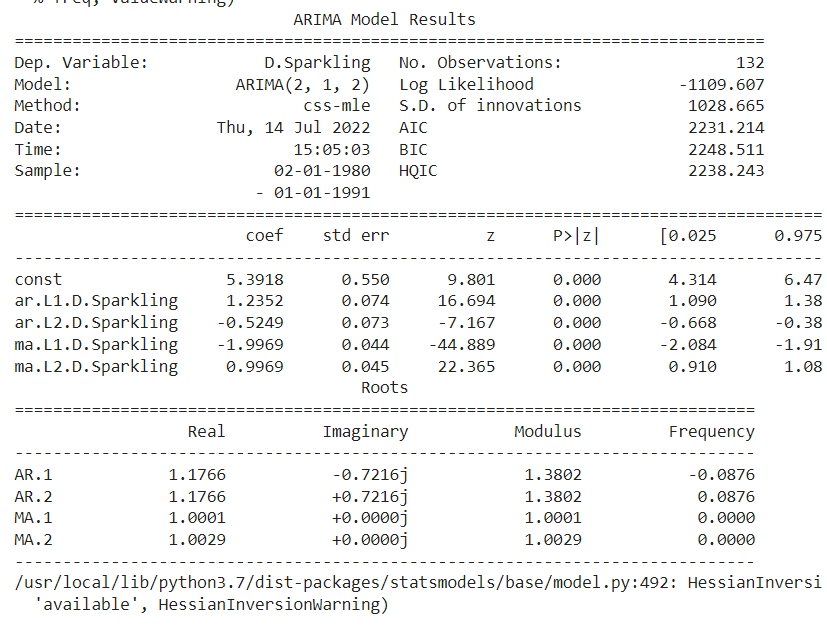


Figure : Auto- ARIMA (2,1,2) summary

* From the summary, it can be inferred that the model has taken 2 order of AR and 2 orders of MA. It is also can be inferred that all the orders are significant based on the p-values, which are less than 0.05 for all the orders.
* To check the performance RMSE test has been performed and for auto-ARIMA (2,1,2) model RMSE score is 1450.816, which is a very high value. It is not a very decent model for forecasting the time series

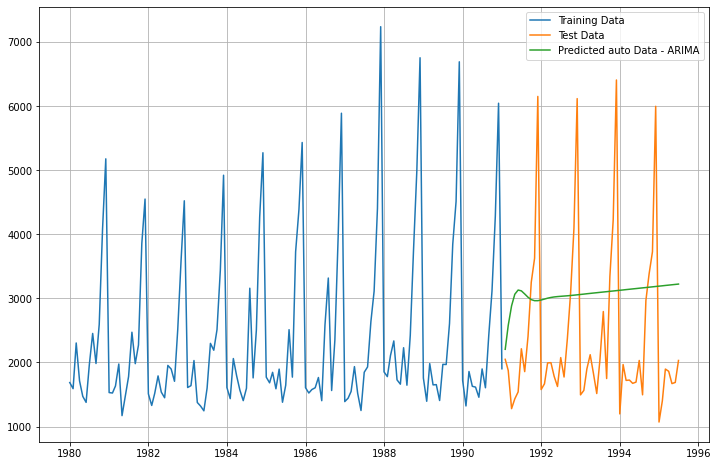


Figure : Auto- ARIMA model (2,1,2)

* It is also seen from the time series plot that it (green line) doesn’t look like it has done very well forecast against the original data

**Auto-SARIMA model using AIC**

* SARIMA model- Seasonal Autoregressive integrated moving average model where the p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and. q is the number of lagged forecast errors in the prediction equation.
* There are four seasonal elements that are not part of ARIMA that must be configured; they are: P-Seasonal autoregressive order, D-Seasonal difference order, Q-Seasonal moving average order, m-The number of time steps for a single seasonal period.
* For the SARIMA model building, the range of p and q have been taken as 0 to 3 and for d range is 1 to 2, D range is 0 to 1 and seasonality is 12. From these parameters, different combinations of the models have been created.
* Akaike Information has been obtained from these models. It basically checks the goodness of fit, and the simplicity/parsimony, of the model into a single statistic.

When comparing two models, the one with the lower AIC is generally “better”.

* For the value of pdq= (1,1,2) and seasonal (1,0,2,12) gives the best AIC value, which is equal to 1571.213 compared to all other combinations

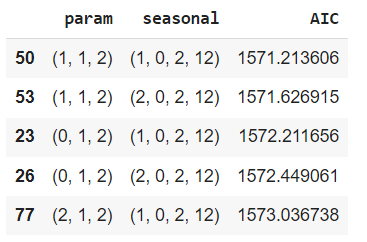
****

Figure : Best param and seasonal for Auto-SARIMA model

* Further the SARIMA model built on the training data set with the pdq value (1,1,2) and seasonal (1,0,2,12).

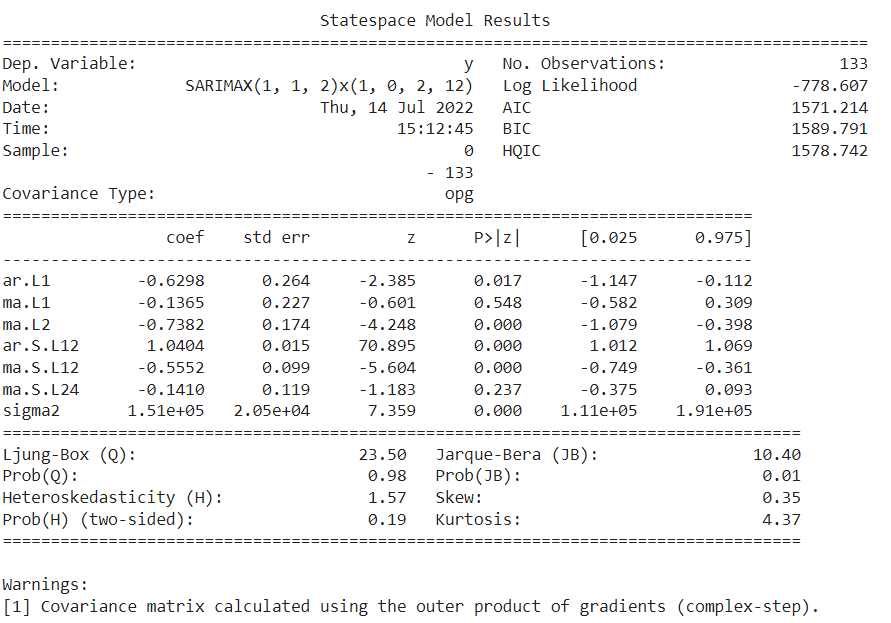
****

Figure : Summary Of SARIMA model

* From the summary, it can be inferred that the model has taken 1 order of AR and 2 orders of MA, 1 orders of seasonal AR and 2 orders of seasonal MA. It is also can be inferred that all the orders are significant based on the p-values except for the MA L1 and MA.S.L24.

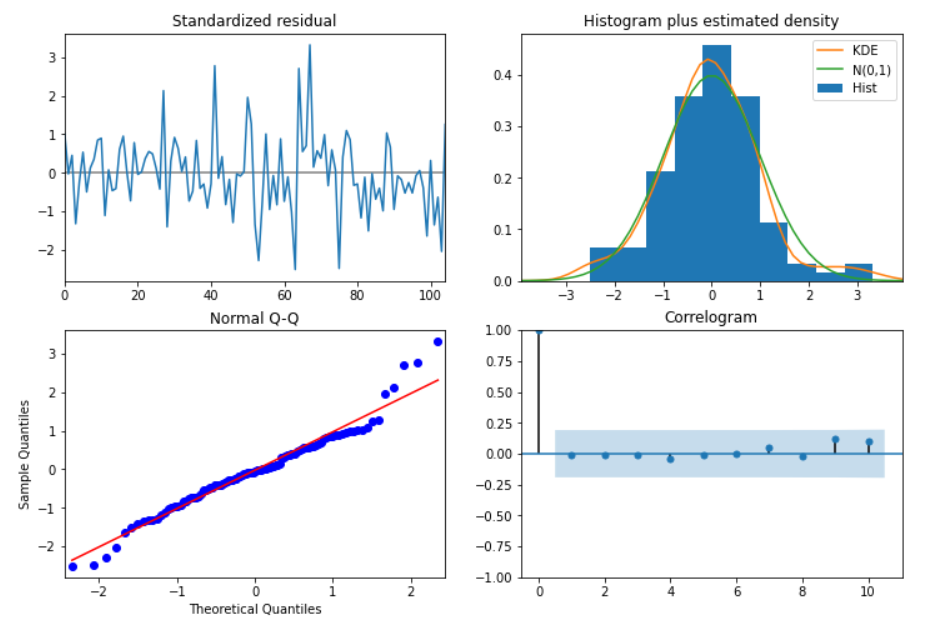


Figure : Diagnostic plot of the SARIMA model

* To check the performance RMSE test has been performed and for auto-ARIMA (2,1,2) model RMSE score is 455.756, which is a less. It is a decent model can be used for forecasting the time series

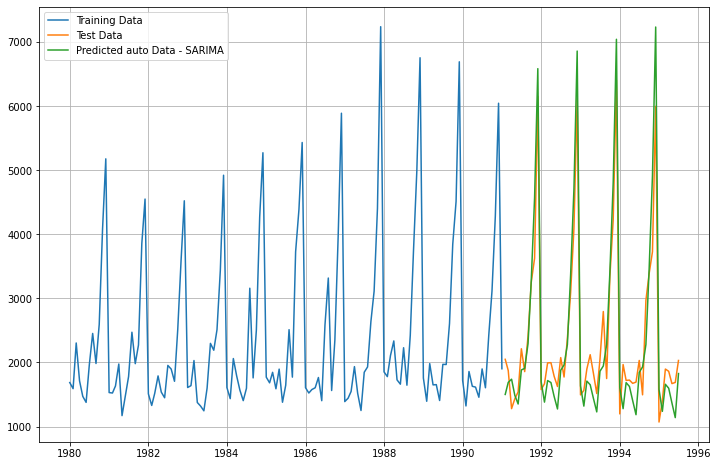
****

Figure : Auto- SARIMA model

* It is also seen from the time series plot that it (green line) has done very well forecast against the original data

**Conclusion**

* Based on the RMSE score SARIMA (1,1,2)(1,0,2,12) is less compared to Auto- ARIMA (2,1,2). Hence, the former is a better model.

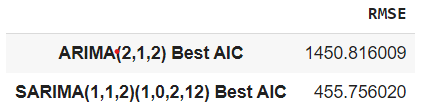
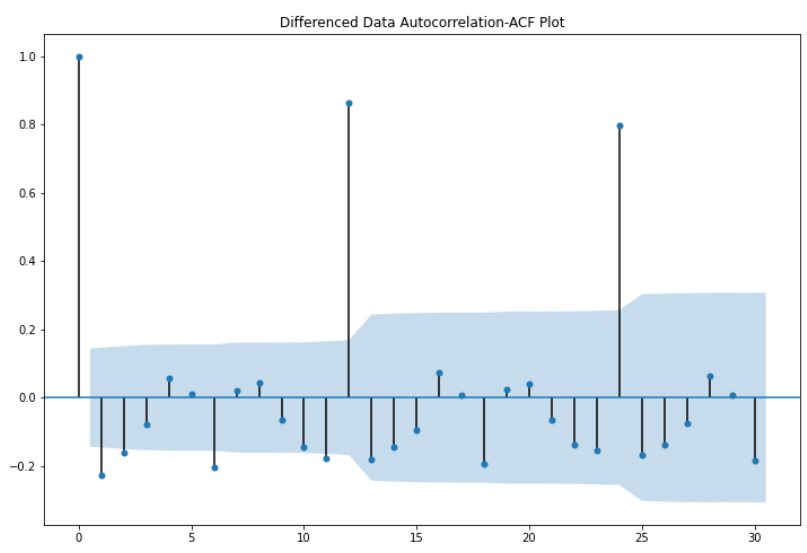
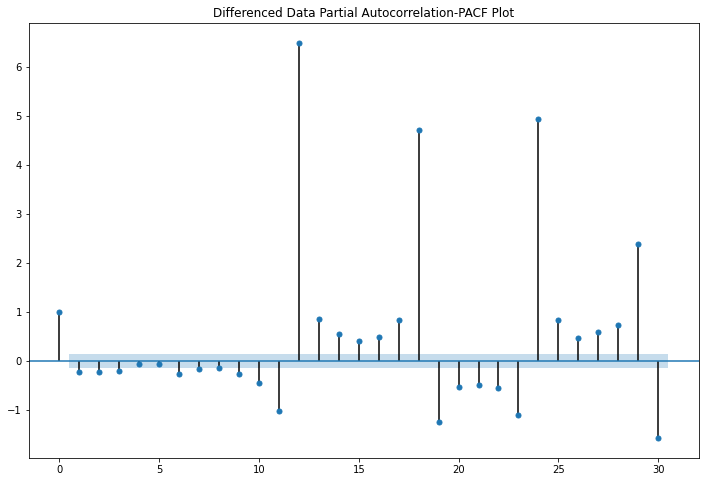
****

Figure : RMSE Score for ARIMA and SARIMA

1. 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

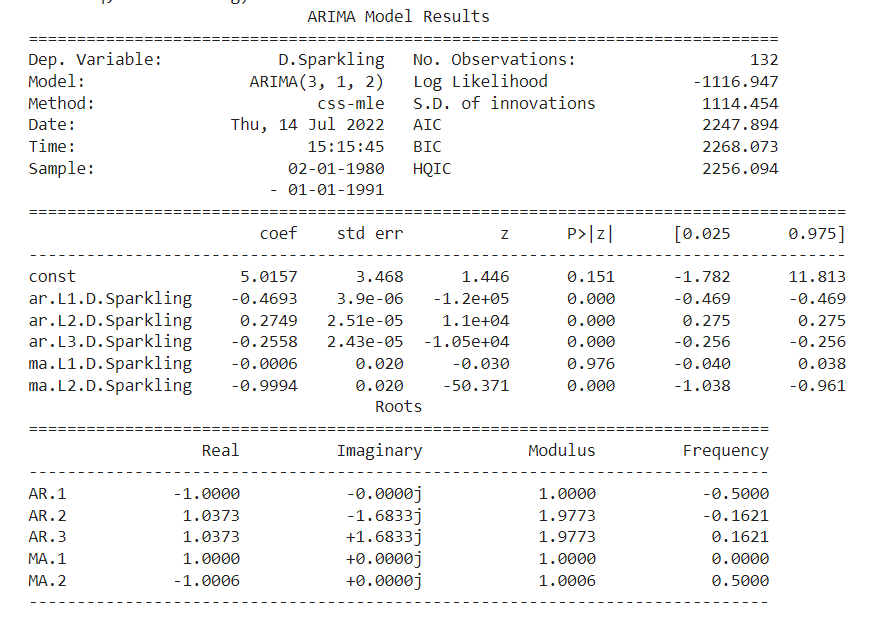
* ACF, and PACF plots allow us to decide the parameters of our ARIMA model.





* Seasonality after certain 12 lags that is every 12 months is visible in plots
* p =3(PACF plot)3 lag are falling outside the significance blue band ranged=1(order of differencing is 1 as original series was non-stationary but became stationary after differencing), and q=2(ACF plot)-two lags are falling above the significance blue band range. Seasonality after 12 lags is seen in plots. ACF & PACF plots are done using 95% confidence interval bands.

**ARIMA model by using p,d,q values from ACF and PACF plots**

****

* From the summary, it can be inferred that the model has taken 3 orders of AR and 2 orders of MA, 1. It is also can be inferred that all the orders are significant based on the p-values except for the MA.
* RMSE score for this ARIMA- (3,1,2) model is 1341.272, which is a very high value. Hence, It is not a decent model for forecasting

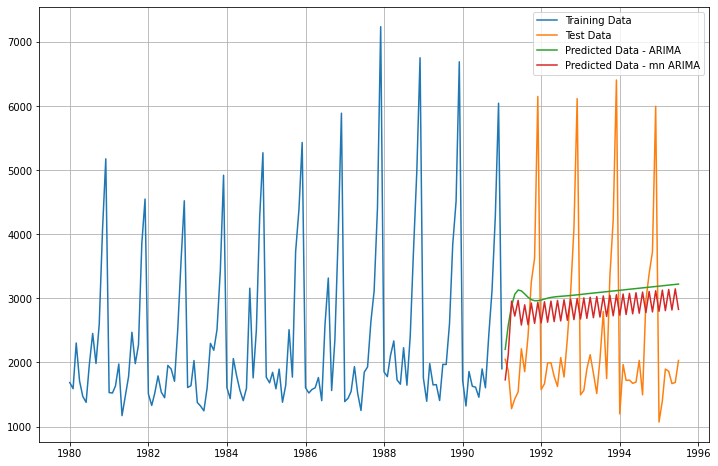


Figure : ARIMA with ACF and PACf- (3, 1, 12)

* It is also seen from the time series plot that it (red line) has not done very well forecast against the original data

**SARIMA model by using p,d,q values from ACF and PACF plots**

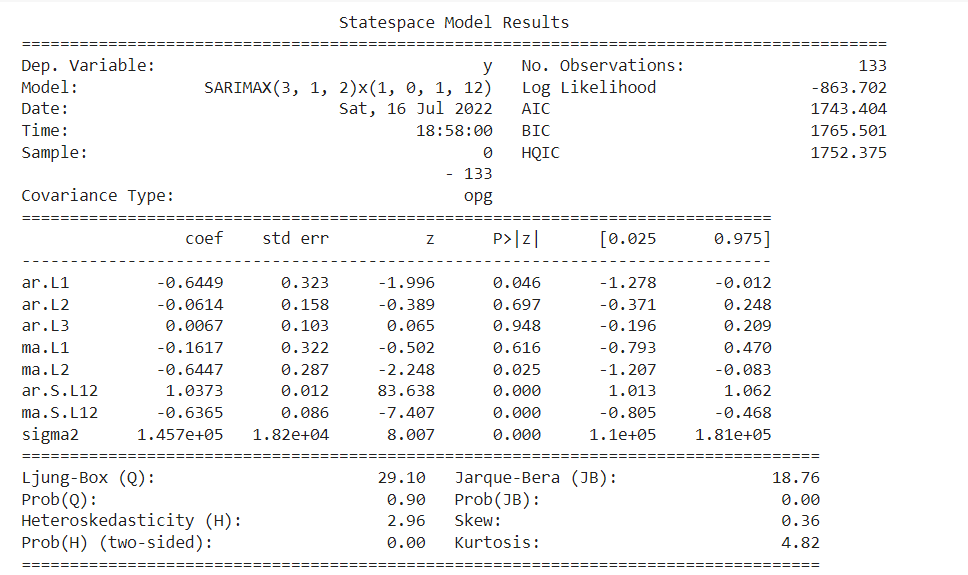
****

Figure : SARIMA model ACF and PACF

* From the summary, it can be inferred that the model has taken 3 orders of AR and 2 orders of MA, 1 order of seasonal AR and 1 order of seasonal MA. It is also can be inferred that AR L1, Seasonal AR and Seasonal MA orders are significant based on the p-value
* RMSE score for this ARIMA- (3,1,2)(1,0,1,12) model is 1341.272, which is a very high value. Hence, It is not a decent model for forecasting

**Conclusion**

* Based on the RMSE score both model performed similarly. We cannot take either of the models for final consideration

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

|  |  |
| --- | --- |
| MODELS | TEST RMSE |
| Regression On Time | 1381.320 |
| Naïve Model | 1381.177 |
| Simple Average model | 1285.039 |
| 2 point Trailing Moving Average | 770.928 |
| 4 point Trailing Moving Average | 1137.137 |
| 6 point Trailing Moving Average | 1283.096 |
| 9 point Trailing Moving Average | 1354.277 |
| Alpha=0.01: Simple Exponential Smoothing | 1285.771 |
| Alpha=0.53,Beta=0: Double Exponential Smoothing | 1780.013 |
| Alpha=0.14,Beta=2.1133171248060435e-30,Gamma=0.368:Triple Exponential Smoothing | 427.962 |
| Alpha=0.3, Simple Exponential Smoothing | 1289.773 |
| Alpha=0.6 & Beta=0.4, Double Exponential Smoothing | 1656.603 |
| Alpha=0.8,Beta=0.9,Gamma=0.3:Triple Exponential Smoothing | 432.387 |
| ARIMA (2,1,2) Best AIC | 1450.816 |
| SARIMA (1,1,2)(1,0,2,12) Best AIC | 455.756 |
| ARIMA (3,1,2) ACF & PACF | 1341.272 |
| SARIMA (3,1,2) (1,0,1,12) ACF and PACF | 1341.272 |
| SARIMA (1,1,2) (1,0,2,12) on 12 months | 539.981 |
| Alpha=0.14,Beta=2.1133171248060435e-30,Gamma=0.368:Triple Exponential Smoothing on 12 moths | 458.683 |

Table : Different Models RMSE score

**DATA FRAME of RMSE scores**

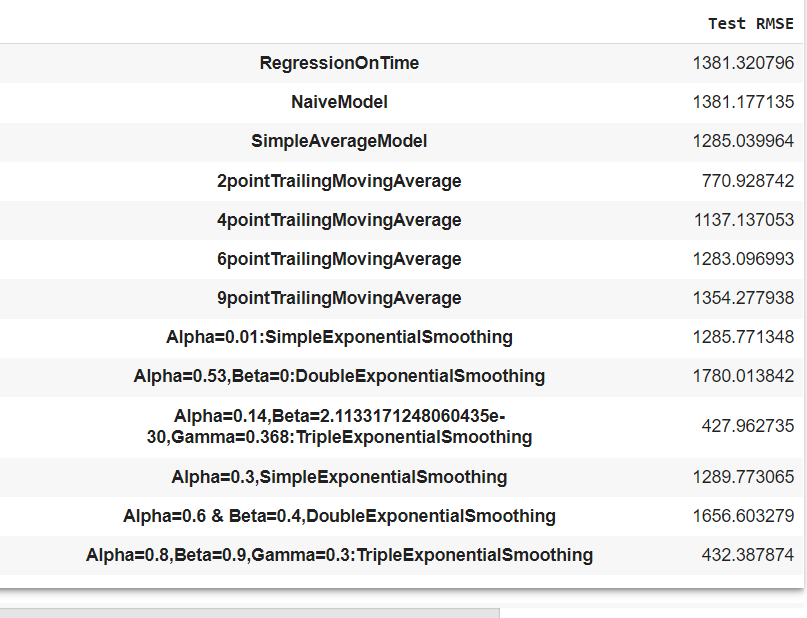




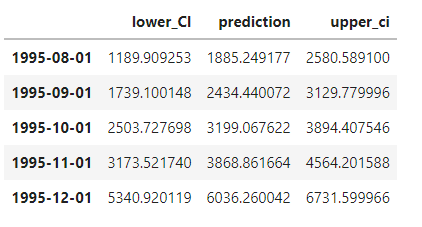
Table : Data frame containing models RMSE scores

Conclusion: Triple Exponential smoothing with Alpha=0.14, Beta=2.1133171248060435e-30, Gamma=0.368 gives the minimum RMSE score (427.962) among non- parametric models. And among the parametric models SARIMA (1,1,2) (1,0,1,12) gives the minimum RMSE score 455.756.

Further, for final model selection, we had taken both the final model of the parametric and non-parametric model.

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

* Based on the RMSE score for different models, Triple Exponential smoothing with Alpha=0.14, Beta=2.1133171248060435e-30, Gamma=0.368 gives the minimum RMSE score (427.962) among non-parametric models. And among the parametric models SARIMA (1,1,2) (1,0,1,12) gives the minimum RMSE score 455.756
* Two final model has been instantiated for 12 months of prediction on whole data with an assumption that while calculating the confidence bands is that the standard deviation of the forecast distribution is almost equal to the residual standard deviation
* For**, Non- Parametric**- **Triple Exponential smoothing** with Alpha=0.14, Beta=2.1133171248060435e-30, Gamma=0.368, multiplier taken for this= 1.96



RMSE score for this is 458.683

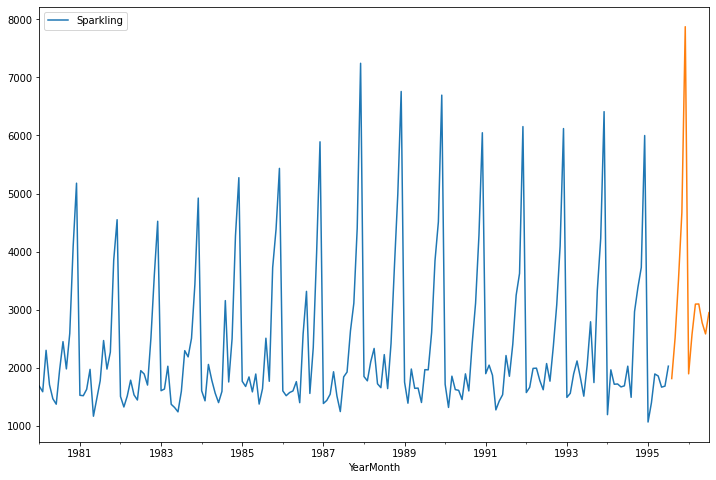


Figure : Prediction Plot for 12 months using Triple Exponential smoothing

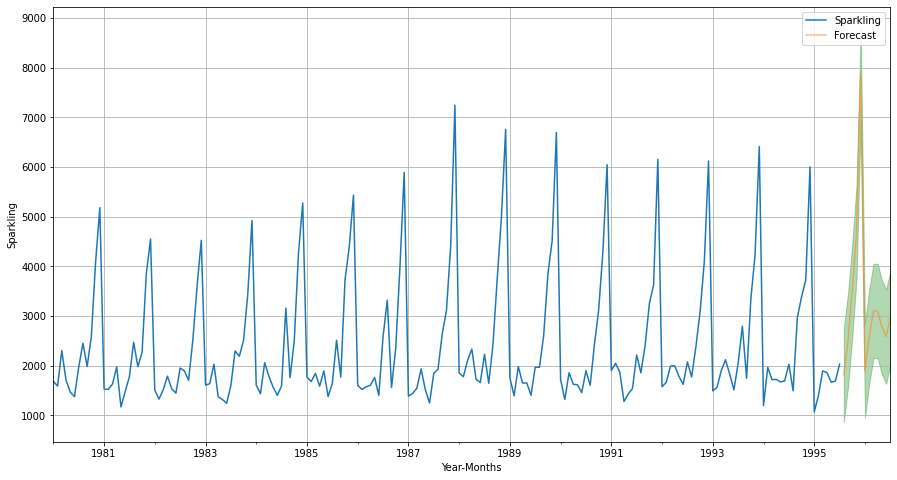
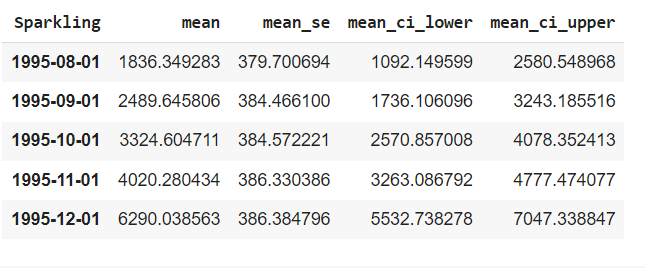


Figure : Prediction Plot for 12 months with CI using Triple Exponential smoothing

* It can be inferred from the prediction plot as well as prediction plot with confidence interval that, the forecast has done good job at forecasting based on the previous information. RMSE score is less (427.962), hence; this model can be used as a final model for forecasting.
* For**, Parametric- SARIMA (1,1,2) (1,0,1,12)** with p=1, d=1, q=2 and P= 1, D=0, Q=2, M= 12
  + - SARIMA model requires the time series to be stationary. The original data is not stationary. Since the requirement is to predict the original time series, the Same parameters are used for training data along with a level of differencing added to it, which allowed to maintain the stationarity of the data; necessary for the SARIMA model.



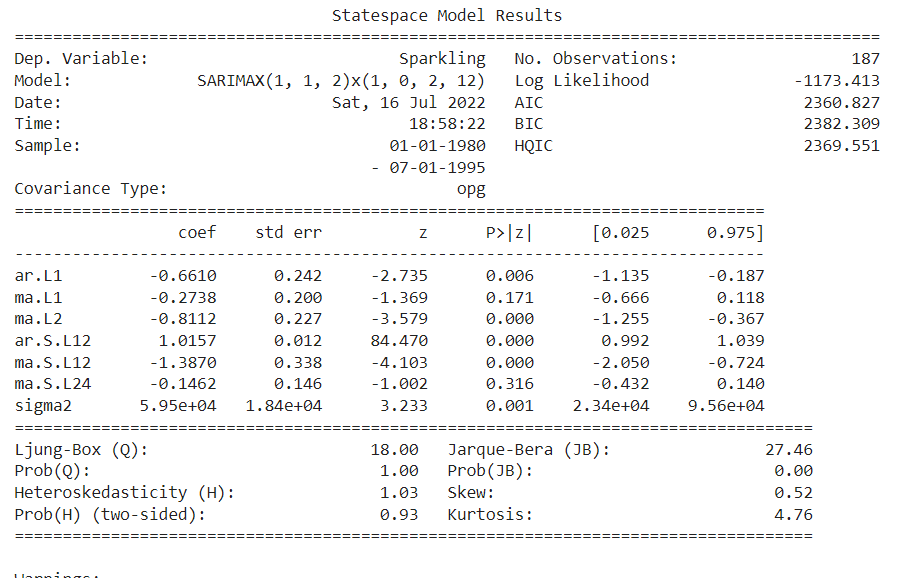


Figure : Summary of Final SARIMA model

* + - From the summary, it can be inferred that the model has taken 1 order of AR and 2 orders of MA, 1 orders of seasonal AR and 2 orders of seasonal MA. It is also can be inferred that all the orders are significant based on the p-values except for the MA L1 and MA.S.L24.

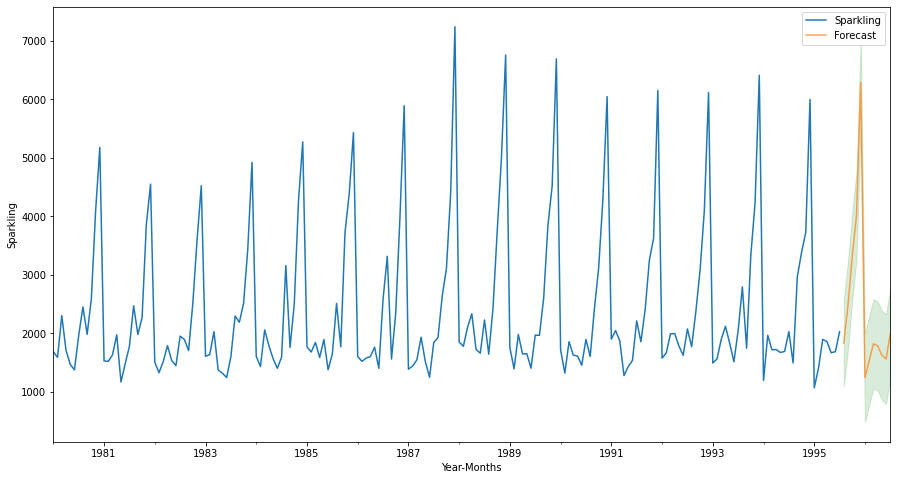
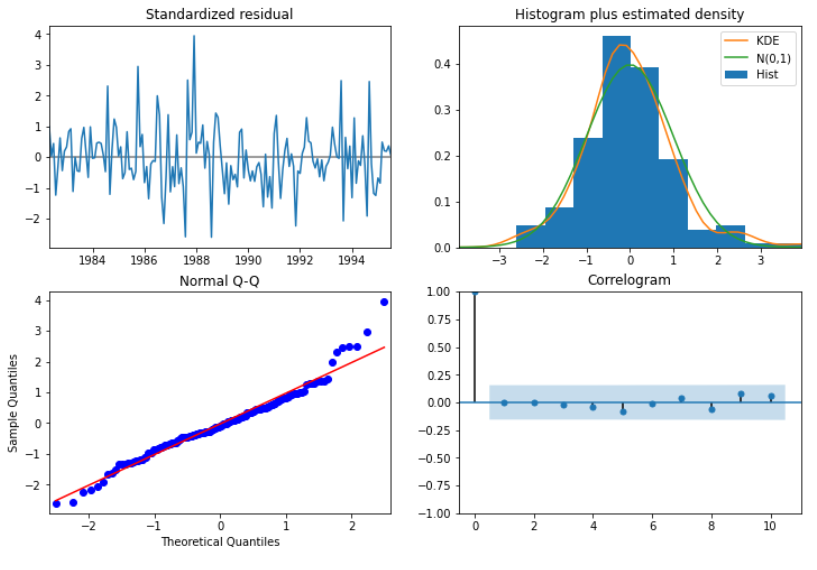


Figure : Prediction Plot for 12 months with CI using SARIMA (1,1,2) (1,0.1,12)

* + - It can be inferred from the prediction plot as well as prediction plot with confidence interval that, the forecast has done good job at forecasting based on the previous information. RMSE score is less (455.756), hence; this model can be used as a final model for forecasting.

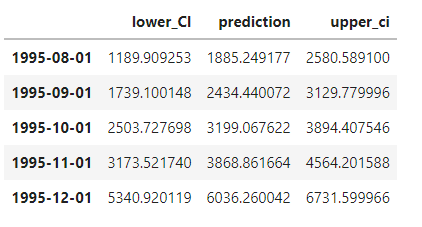


* + - diagnostic suggests that the model residuals are normally distributed, the residuals of this model are uncorrelated and normally distributed with zero mean. The KDE plot of the residuals on top right is almost similar with the normal distribution.
    - The qq-plot on the bottom lefts shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0,1).
    - The residuals over time (top left plot) don’t display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i,.e. correlogram ) plot on the bottom right, which shows that the time series residuals have a low correlation with lagged versions of itself
    - There are no spikes outside the insignificant zone for both ACF and PACF plots leading us to conclude that residuals are random with no information in them and our model produces a satisfactory fit that could help us understand our time series data and forecast future values

**Conclusion**: It can be inferred that Both the Triple Exponential Smoothing final model and SARIMA final model have performed well and can be used as the final 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

* Based on the RMSE score for different models, Triple Exponential smoothing with Alpha=0.14, Beta=2.1133171248060435e-30, Gamma=0.368 gives the minimum RMSE score (427.962) among non-parametric models. And among the parametric models SARIMA (1,1,2) (1,0,1,12) gives the minimum RMSE score 455.756
* Two final model has been instantiated for 12 months of prediction on whole data with an assumption that while calculating the confidence bands is that the standard deviation of the forecast distribution is almost equal to the residual standard deviation
* For**, Non- Parametric**- **Triple Exponential smoothing** with Alpha=0.14, Beta=2.1133171248060435e-30, Gamma=0.368, multiplier taken for this= 1.96



RMSE score for this is 458.683

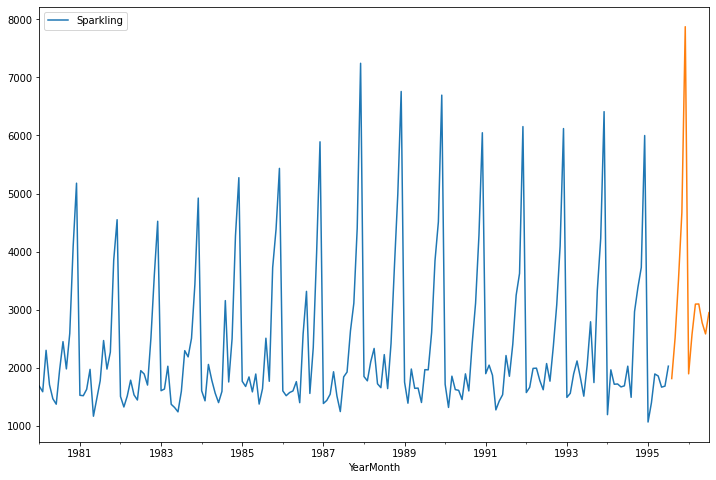


Figure : Prediction Plot for 12 months using Triple Exponential smoothing

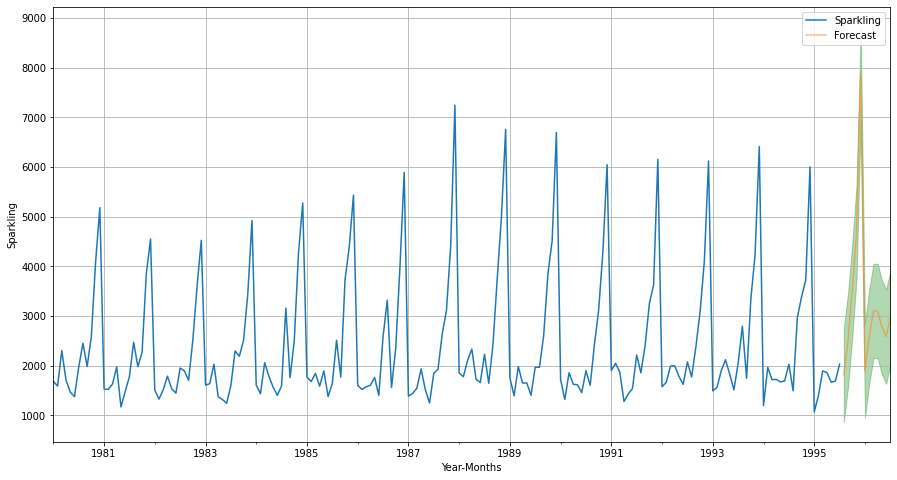
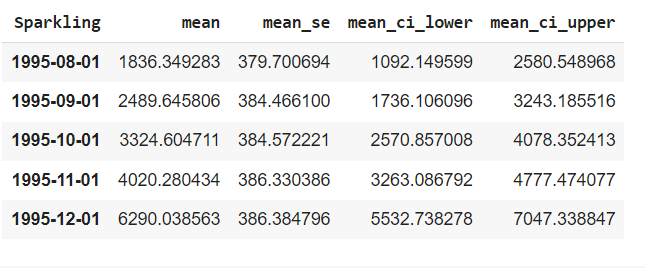


Figure : Prediction Plot for 12 months with CI using Triple Exponential smoothing

* It can be inferred from the prediction plot as well as prediction plot with confidence interval that, the forecast has done good job at forecasting based on the previous information. RMSE score is less (427.962), hence; this model can be used as a final model for forecasting.
* For**, Parametric- SARIMA (1,1,2) (1,0,1,12)** with p=1, d=1, q=2 and P= 1, D=0, Q=2, M= 12
  + - SARIMA model requires the time series to be stationary. The original data is not stationary. Since the requirement is to predict the original time series, the Same parameters are used for training data along with a level of differencing added to it, which allowed to maintain the stationarity of the data; necessary for the SARIMA model.



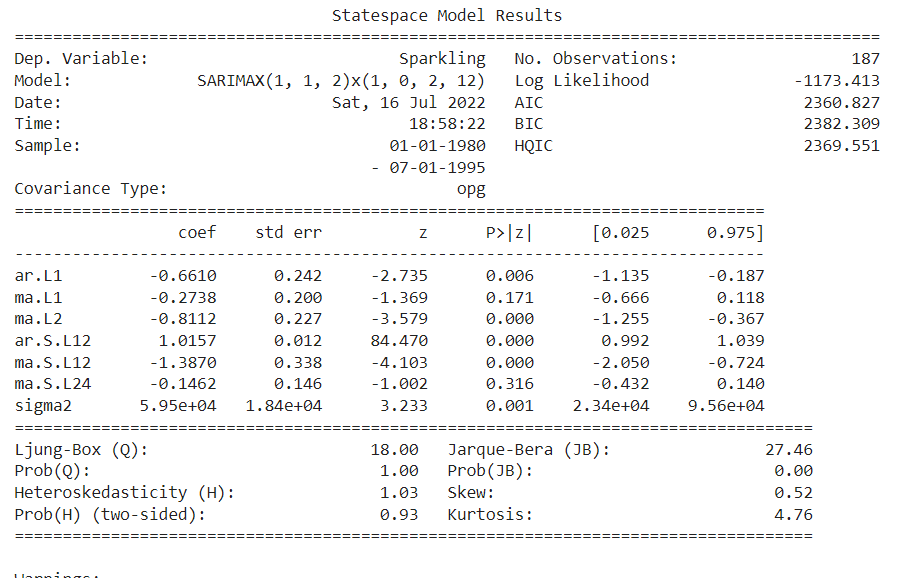


Figure : Summary of Final SARIMA model

* + - From the summary, it can be inferred that the model has taken 1 order of AR and 2 orders of MA, 1 orders of seasonal AR and 2 orders of seasonal MA. It is also can be inferred that all the orders are significant based on the p-values except for the MA L1 and MA.S.L24.

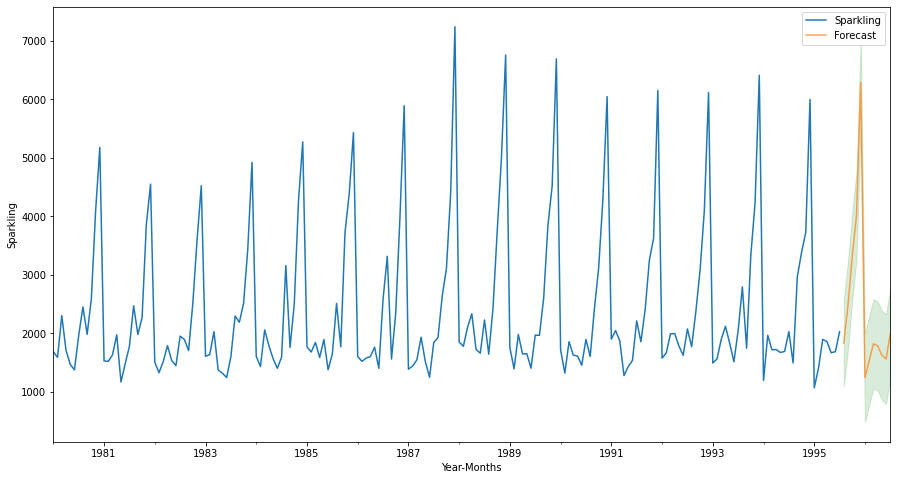
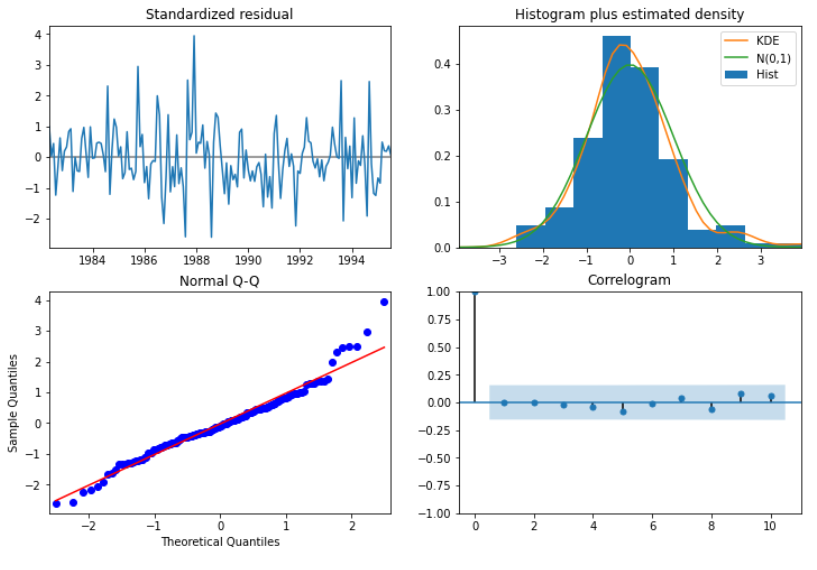


Figure : Prediction Plot for 12 months with CI using SARIMA (1,1,2) (1,0.1,12)

* + - It can be inferred from the prediction plot as well as prediction plot with confidence interval that, the forecast has done good job at forecasting based on the previous information. RMSE score is less (455.756), hence; this model can be used as a final model for forecasting.



* + - diagnostic suggests that the model residuals are normally distributed, the residuals of this model are uncorrelated and normally distributed with zero mean. The KDE plot of the residuals on top right is almost similar with the normal distribution.
    - The qq-plot on the bottom lefts shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0,1).
    - The residuals over time (top left plot) don’t display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i,.e. correlogram ) plot on the bottom right, which shows that the time series residuals have a low correlation with lagged versions of itself
    - There are no spikes outside the insignificant zone for both ACF and PACF plots leading us to conclude that residuals are random with no information in them and our model produces a satisfactory fit that could help us understand our time series data and forecast future values

**Conclusion**: It can be inferred that Both the Triple Exponential Smoothing final model and SARIMA final model have performed well and can be used as the final model.

**FINDINGS-**

* Each model shows increase in seasonality sparkling wine sales and no distinct trend present in the data.
* It is also seen that number of sparkling wine sales increases during the winter months due to the holiday seasons. Sparkling wine sales tend to be seasonal

**Measures-**

* The growth of fruit wines has had an impact on declining sales of wine, flavored cider peaks also play a crucial role in wine production
* sparkling wine is smaller and more niche than other types of wine. To increase the number of sales, it should look at appealing to older consumers more effectively by communicating premium messages, emphasizing the production process, mentioning the uniqueness of wine aged in whiskey barrels, the wine aged in amphoras under the sea, wine with chilly, wine with hemp, etc. Different wine bottles and unique labels will do the job as well
* People generally buy sparkling wine during celebration, giving sales during exclusive customers birthday, anniversary etc will help in sales increase
* Increase in tourism may help in the increase in sales.
* Building brand recognition is extremely important in order to attract younger generations.
* Apps are also a great way for any winery to introduce their wines to the younger people