Time Series Forecasting

Table of contents

[**Executive summary** 4](#_Toc101126010)

[**Introduction** 5](#_Toc101126011)

[**Problem 1:** 6](#_Toc101126012)

[**Data Dictionary and description** 6](#_Toc101126013)

[**Sample of Dataset1** 6](#_Toc101126014)

[**1.** **Read the data as an appropriate Time Series data and plot the data.** 6](#_Toc101126015)

[**2.** **Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.** 7](#_Toc101126016)

[**3.** **Split the data into training and test. The test data should start in 1991.** 12](#_Toc101126017)

[**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 th training data and check the performance on the test data using RMSE.** 13](#_Toc101126018)

[**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.** 23](#_Toc101126019)

[**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.** 24](#_Toc101126020)

[**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.** 28](#_Toc101126021)

[**8.** **Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.** 31](#_Toc101126022)

[**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.** 32](#_Toc101126023)

[**10.** **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.** 33](#_Toc101126024)

List of Figures

[Figure 1.Boxplot for yearly sales- dataset1(Sparkling) 8](file:///D:\Jupyter%20Codes\M7-%20Time%20series%20forcasting\Time%20series%20-%20Assignment\Time%20Series%20-%20Project-Ankita%20Juneja-Sparkling.docx#_Toc101126025)

[Figure 2. Boxplot for monthly sales- dataset1(Sparkling) 9](#_Toc101126026)

[Figure 3.. Month plot for dataset 1- Sparkling 9](#_Toc101126027)

[Figure 4.Month sales across years for Dataset 1- Sparkling 10](file:///D:\Jupyter%20Codes\M7-%20Time%20series%20forcasting\Time%20series%20-%20Assignment\Time%20Series%20-%20Project-Ankita%20Juneja-Sparkling.docx#_Toc101126028)

[Figure 5.Month sales across years for Dataset 2- Rose 10](file:///D:\Jupyter%20Codes\M7-%20Time%20series%20forcasting\Time%20series%20-%20Assignment\Time%20Series%20-%20Project-Ankita%20Juneja-Sparkling.docx#_Toc101126029)

[Figure 6. Mean sales for all years- Sparkling 10](#_Toc101126030)

[Figure 7. decomposition of dataset1- Sparkling 11](#_Toc101126031)

[Figure 8. Time series without seasonality for dataset 1- Sparkling 12](#_Toc101126032)

[Figure 9. First five entries of train and test data respectively 14](#_Toc101126033)

[Figure 10.Plot for prediction using Linear regression 15](#_Toc101126034)

[Figure 11. plot for model prediction using simple average 17](#_Toc101126035)

[Figure 12. prediction on data for 4 point moving average and 6 point moving average. 18](#_Toc101126036)

[Figure 13.Plot for model prediction using Moving average(4point MA and 6 point MA) 19](#_Toc101126037)

[Figure 14. Model prediction using DES 21](#_Toc101126038)

[Figure 15. Plot for TES(Holt Winter's method) 22](#_Toc101126039)

[Figure 17.Predictions using ARIMA 25](#_Toc101126040)

[Figure 18. Lowest AIC parameters for SARIMA 27](#_Toc101126041)

[Figure 19. Predictions using SARIMA(1,1,2)(0,1,2,12) 27](#_Toc101126042)

[Figure 20. ACF plot on differenced series 29](#_Toc101126043)

[Figure 21.PACF plot on differenced series 29](#_Toc101126044)

[Figure 22. RMSE for ARIMA from parameters using ACF and PACF 30](#_Toc101126045)

[Figure 23. Model diagnostics 34](#_Toc101126046)

**List of Table**

[Table 1.Sample of Dataset1 6](#_Toc101126047)

[Table 2. Datatype and missing value for dataset 1 – Sparkling 6](#_Toc101126048)

[Table 3.Datatype and missing value for dataset 2 – Rose 6](#_Toc101126049)

[Table 4. Time Series for data set 1(Sparkling) 7](#_Toc101126050)

[Table 5.plot for dataset set 1 – Sparkling 7](#_Toc101126051)

[Table 6. Descriptive analysis of dataset 1- Sparkling 8](#_Toc101126052)

[Table 7. first five and last five entries of train data of dataset 1- Sparkling 13](#_Toc101126053)

[Table 8. first five and last five entries of test data of dataset 1- Sparkling 13](#_Toc101126054)

[Table 9. first five entries for train and test data after Linear regression 14](#_Toc101126055)

[Table 10.Initial five entries of dataset after naïve prediction 15](#_Toc101126056)

[Table 11. prediction using Naive model 16](#_Toc101126057)

[Table 12. first five entries using simple average 17](#_Toc101126058)

[Table 13. RMSE Simple Exponential Smoothing 20](#_Toc101126059)

[Table 14. RMSE Double Exponential Smoothing 21](#_Toc101126060)

[Table 15. RMSE for TES 23](#_Toc101126061)

[Table 16. Lowest AIC Parameters 25](#_Toc101126062)

[Table 17. RMSE for ARIMA(2,1,2) 26](#_Toc101126063)

[Table 18. RMSE for SARIMA 28](#_Toc101126064)

[Table 19. RMSE for all models 31](#_Toc101126065)

[Table 20. Full model summary using SARIMA(3,1,2)x(0,1,0,12) 32](#_Toc101126066)

[Table 21. Forcasted values for initial 5 months 32](#_Toc101126067)

[Table 22. Prediction plot using SARIMA 33](#_Toc101126068)

[Table 23. comparison of RMSE for all models 34](#_Toc101126069)

[Table 24. Prediction plot 35](#_Toc101126070)

**Executive summary**

The report covers the methods used to analyze Time Series data and draw inferences from two different datasets.

Time series forecasting is applied to extract information from historical series and is used to predict future behaviour of the same series based on past pattern

In this project we will deep dive in Time series forecasting of Sales of two different types of Wines

Through the use of models we will forecast sales based and recommend sales in 20th century for ABC Estate Wines.

**Introduction**

The purpose of this exercise is to analyse data and predict through use of Time series forcasting techniques

Objective of Time series forecasting is to extract information from historical series and is used to predict future behaviour of the same series based on past pattern.

The regular time intervals can be daily,weekly, monthly, quarterly , annual , quinquennial, that is in every 5 years, or decennial .

There are various forces that may affect the observations in a time series. The three important components are:

I. Trend

II. Seasonal component

III. Irregular component

Response variable is the variable of interest that one wants to model or predict using one variable whose values are known. This variable is a continuous variable

In this exercise we predict the Sales of wine through use of time series forecasting models .

**Problem 1:**

**Data Dictionary and description**

**Dataset 1 : Sparkling**

1. YearMonth: This variable coveys Year and month of sales of Wine
2. Sparkling: The variable represent the sales of wine named Sparkling

**Sample of Dataset1**

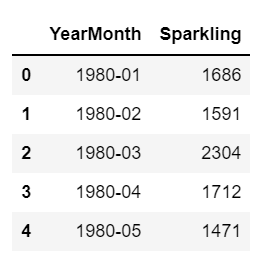


Table 1.Sample of Dataset1

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

Lets us first check the date type and any missing values for Data set 1- Sparkling

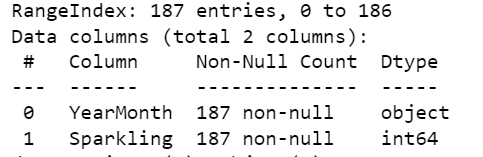


Table 2. Datatype and missing value for dataset 1 – Sparkling

As we can see from above that there table that data type of YearMonth is ‘object’ hence this needs to be converted to Time Series format for appropriate analysis. Also we can see that there is no missing values in data set.

Check for Data type and any missing values for Data set 2- Rose

Table 3.Datatype and missing value for dataset 2 – Rose

Time Series for Dataset Sparkling:

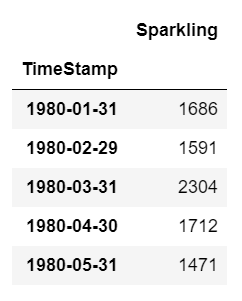


Table 4. Time Series for data set 1(Sparkling)

)

The plot for data sets are as follows:

Plot for dataset 1- Sparkling

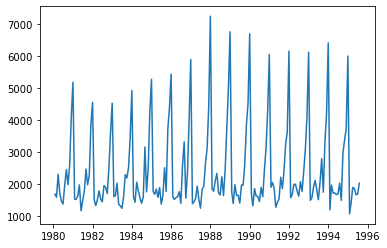


Table 5.plot for dataset set 1 – Sparkling

As we can see from above plot that there is a presence of seasonality every year. However, trend seems irregular i.e. sales are neither continuously increasing, nor it is continuously decreasing. We will get more insights on this once we decompose the series in later part of report.

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

**Let us perform descriptive analysis**

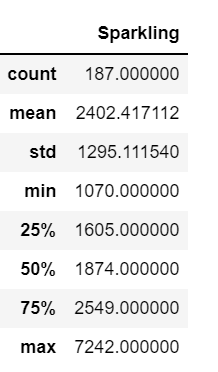


Table 6. Descriptive analysis of dataset 1- Sparkling

As seen from above table that mean is greater than median i.e. the series is right skewed

Exploratory data analysis

Lets us now plot boxplots :

Yearly Boxplots

**Dataset 1- Sparkling**

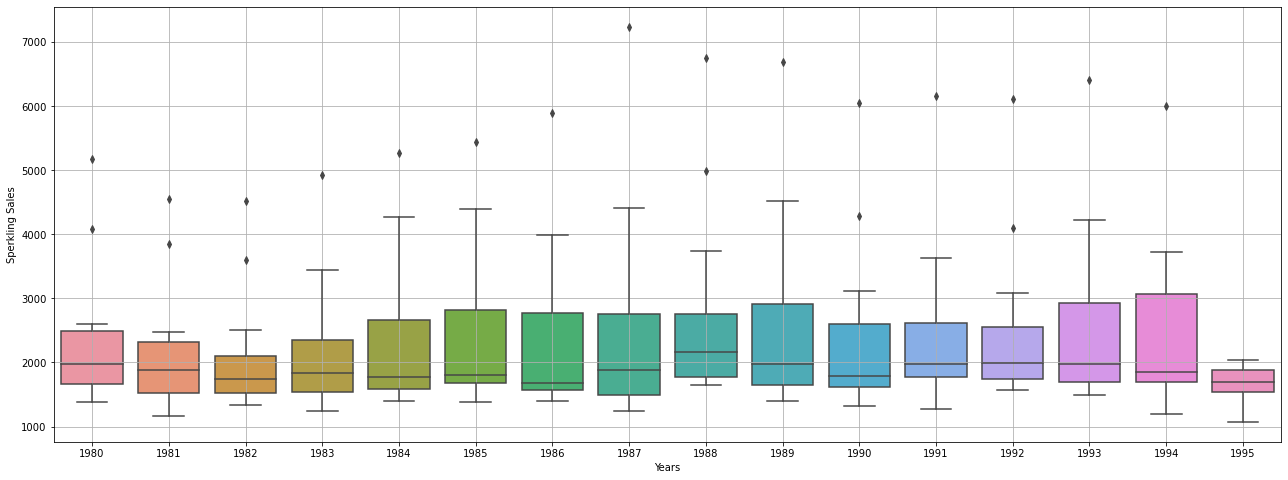


Figure 1.Boxplot for yearly sales- dataset1(Sparkling)

As we got to know from the Time Series plot, the boxplots over here also indicates a measure of trend being present. Also, we see that the sales has outliers in most of the years.

Monthly Boxplots

**Dataset 1 - Sparkling**

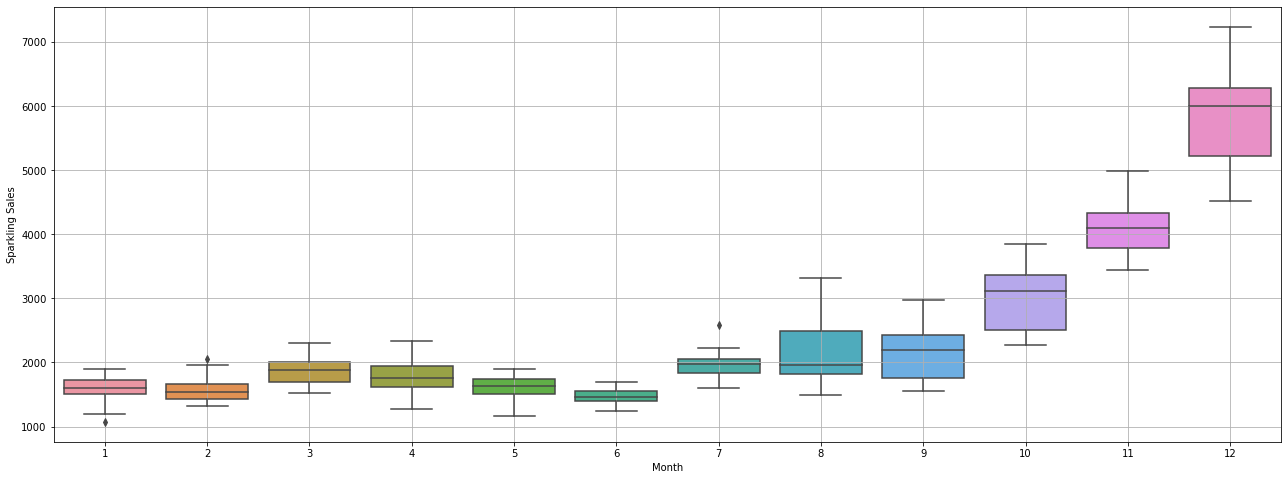


Figure 2. Boxplot for monthly sales- dataset1(Sparkling)

As can be seen from figure, sales of sparkling wine increases towards the end of the year and is maximum in December. At the same time we see that there are no Outliers in Monthly boxplots except for the month of July.

Month Plot for dataset 1- Sparkling

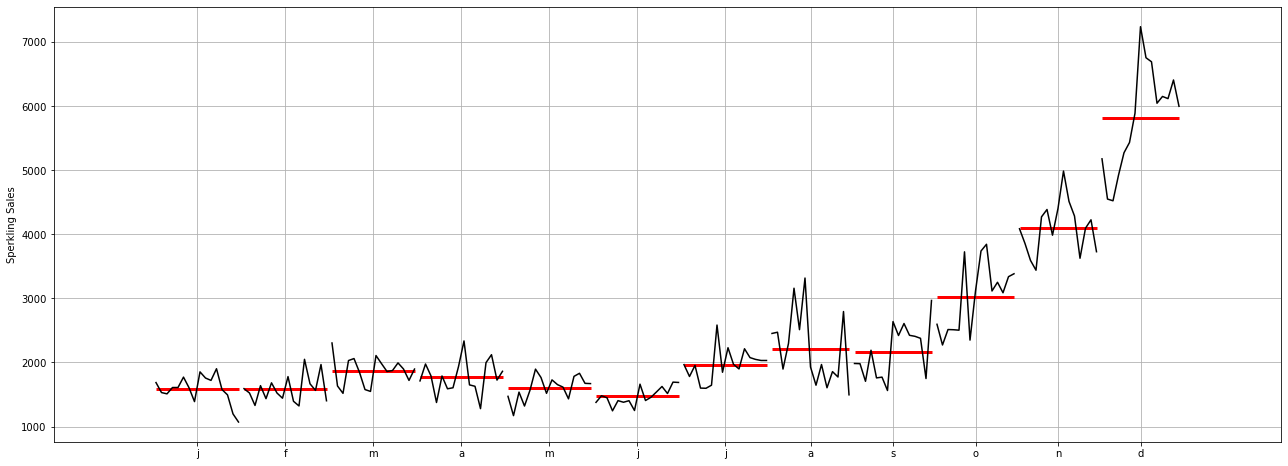


Figure 3.. Month plot for dataset 1- Sparkling

As seen from above figure we can conclude that average monthly sales goes up from sept onwards and sharply increases till dec

Let us plot the monthly sales across years for both data set.

Plot for dataset 1 - Sparkling

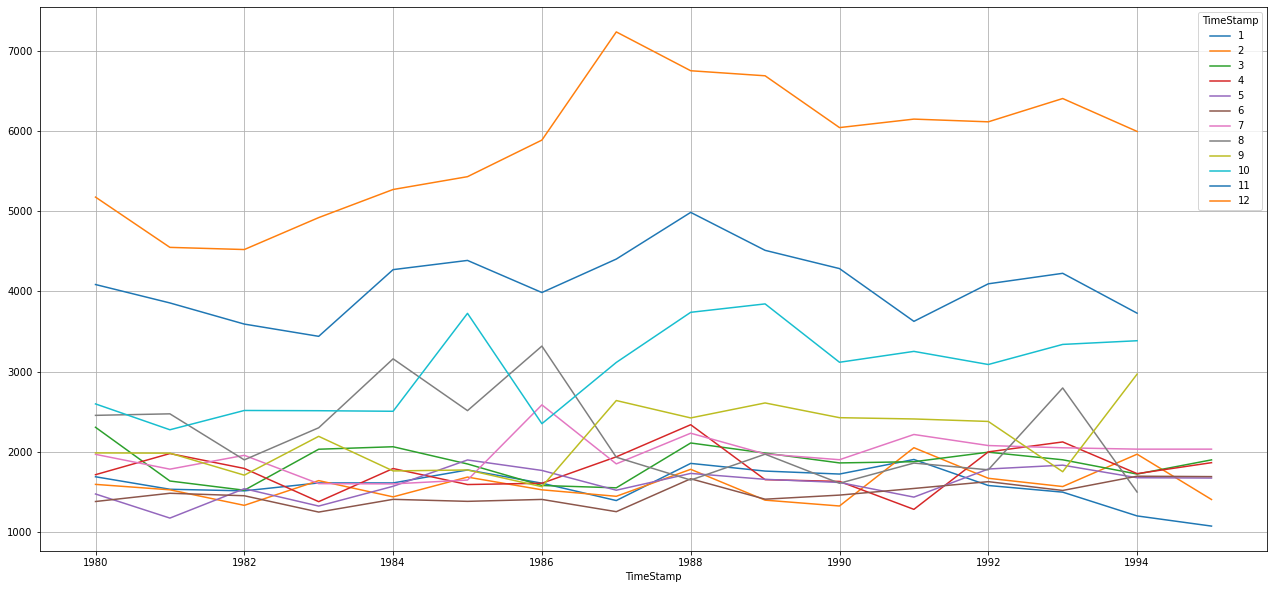
As seen from above figure for Sparkling dataset we can see that months Nov and Dec months stand out in sales. These months have comparatively higher sales than other months. Also we notice that from year 1986 onwards Oct month also has increased sales. We also see that there is some sesonality in data and also some irregular component which effects sales.

Figure 4.Month sales across years for Dataset 1- Sparkling

Figure 5.Month sales across years for Dataset 2- Rose

On find the mean sales for all years we can see below mentioned trend for Sparkling

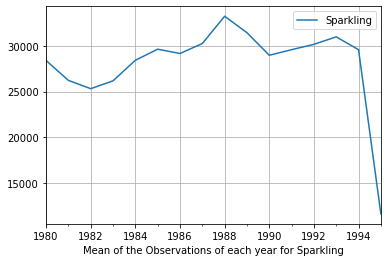


Figure 6. Mean sales for all years- Sparkling

Decomposition

There are various forces that may affect the observations in a time series. The three important components are:

* **Trend** (Long term movement ***Tt***): When the series increases (or decreases) over the entire length of time.
* **Seasonal component** (St**):**: When a series is observed with more frequently than a year (quarterly or monthly for example), the series is subject to rhythmic fluctuations which are stable and repeatable each year.
* **Irregular component** (Random movements- ***It***): These fluctuations are purely random, erratic, unforeseen, and unpredictable. This is the random component of time series.

Trend and seasonal components are part of systematic components of time series.

Decomposition of a time series leads to identification and extraction of the individual components.

**Decomposition of dataset 1- Sparkling into trend, seasonality and residual.**

Primary objective of decomposition is to study the components of the time series.

Based on the EDA done above we can see that this is additive decomposition as seasonal variation is constant over time.

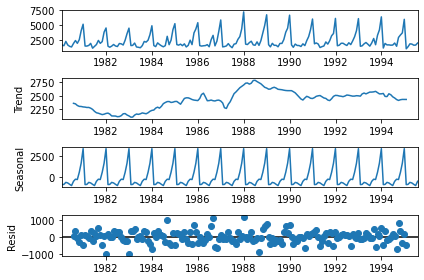


Figure 7. decomposition of dataset1- Sparkling

As we can see from above figure, seasonality is constant over years. Lets plot another plot where seasonality is removed, so that we can understand trend and residual better.

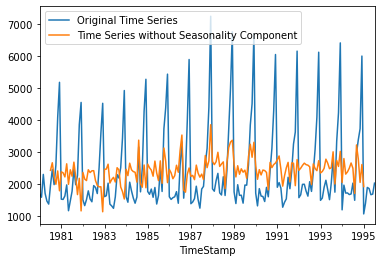


Figure 8. Time series without seasonality for dataset 1- Sparkling

We can infer that data without seasonality follows a trend which is neither increasing nor decreasing.

Inferences for Dataset 1-Sparkling:

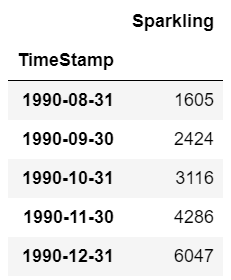
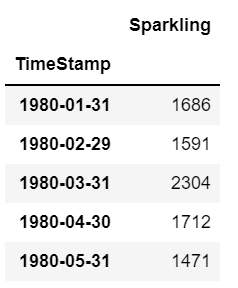
* The Sales of Sparkling wine does not seem to follow specific trend which is neither increasing nor decreasing
* Dataset follows seasonality which is almost constant over the years
* We also see that sales in last 4 months sharply increases with Dec having highest sales
* On removal of seasonality data follows trend with some randomness

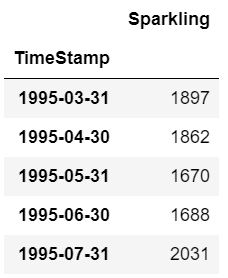
1. **Split the data into training and test. The test data should start in 1991.**

For any model creation we need to split the data in Training data set and Testing dataset. For Time series forecasting approach is to use most recent years data as test data and rest data as train data.

In our data set we have time series starting from year 1980 uptill 1995 . Hence we will use data from 1980 to 1990 as train data and data from 1991 uptill 1995 as test data.

For data set for train data is as follows

Table 7. first five and last five entries of train data of dataset 1- Sparkling

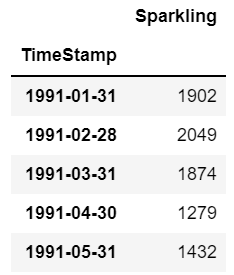
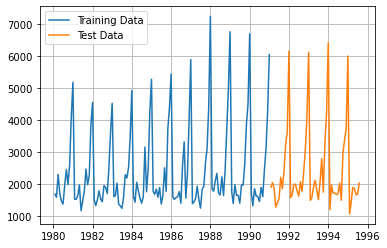


Table 8. first five and last five entries of test data of dataset 1- Sparkling

Plot for 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 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.**

Linear regression Model :

#### For this particular linear regression, we are going to regress the 'Sparkling' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

Data after modification for Linear regression is as follows

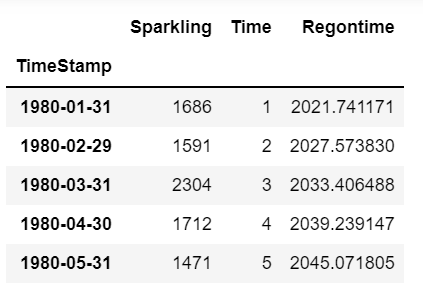
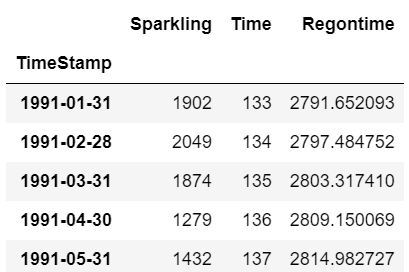
 

Figure 9. First five entries of train and test data respectively

After forecasting using Linear regression model following is the output

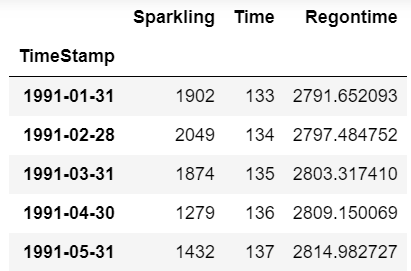
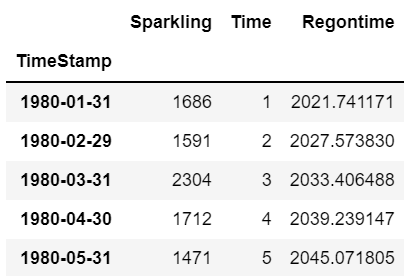


Table 9. first five entries for train and test data after Linear regression

Model Evaluation:

Now lets check the Evaluate model by calculating RMSE



As can be seen that test RMSE is quite high for Linear Regression model. Hence we will explore other models to find the lowest RMSE.

**Plot for prediction using Linearregression**

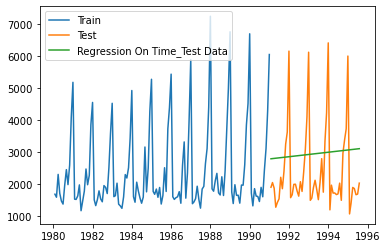


Figure 10.Plot for prediction using Linear regression

**Naïve Forcast Models :**

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

Naïve approach

𝑦̂ 𝑡+1 = 𝑦

In This we checked the last five entries of dataset and will use the last entry of dataset as prediction for all entries of testset

Initial five entries of dataset after naïve prediction are as follows

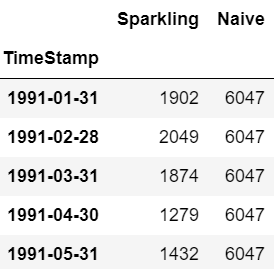
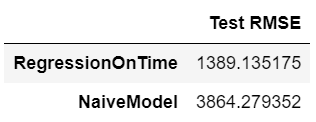


Table 10.Initial five entries of dataset after naïve prediction

Model Evaluation:

Now lets check the Evaluate model by calculating RMSE



The model prediction for naïve model is worst than Linear regression.

**Plot for Naïve Forecast prediction**

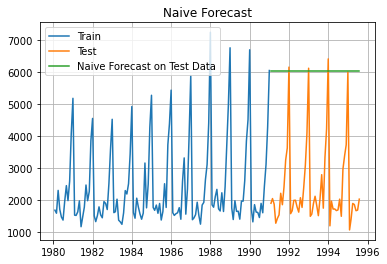


Table 11. prediction using Naive model

As we can see from plot that Prediction of model is far from actual test data values

**Simple Average method**

For this particular simple average method, we will forecast by using the average of the training values.

Initial five entries after forecast using this method is as follows

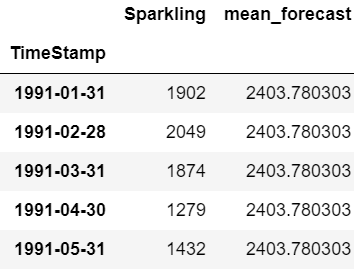
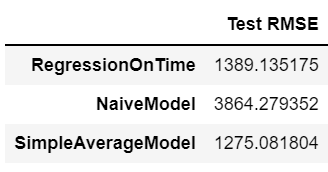


Table 12. first five entries using simple average

Model Evaluation:

Now lets check the Evaluate model by calculating RMSE



As we can see Simple average method has performed better than Naïve model and Linear regression model.

**Plot for model prediction using Simple average**

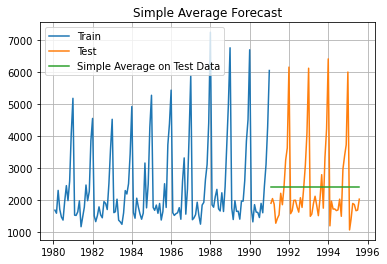


Figure 11. plot for model prediction using simple average

As we can see from above figure that predictions are very far from actual observations. This plot does not seem to be suitable.

**Moving Average(MA)**

A moving average process, or the moving average model, says that the current value is linearly dependent on the current and past error terms

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

For Moving Average, we are going to average over the entire data*.*

For this problem we will use 4 point moving average and 6 point moving average

The sample data and prediction after using rolling means or moving average on data is as follows:

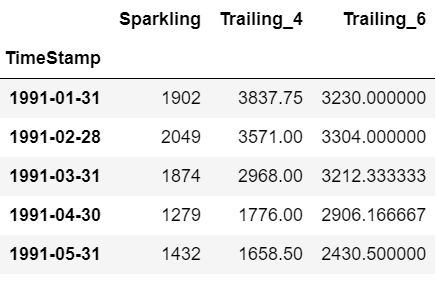
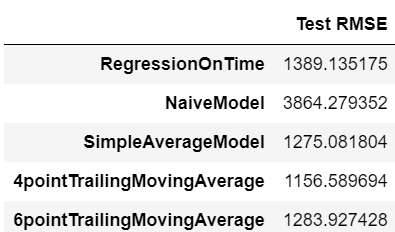


Figure 12. prediction on data for 4 point moving average and 6 point moving average.

Model Evaluation:

Now lets check the Evaluate model by calculating RMSE



We can conclude from above table that 4 point moving average model is giving best results amongst all the models used so far. 4 point MA model has lowest RMSE as per above table

**Plot for model prediction using Moving average(4point MA and 6 point MA)**

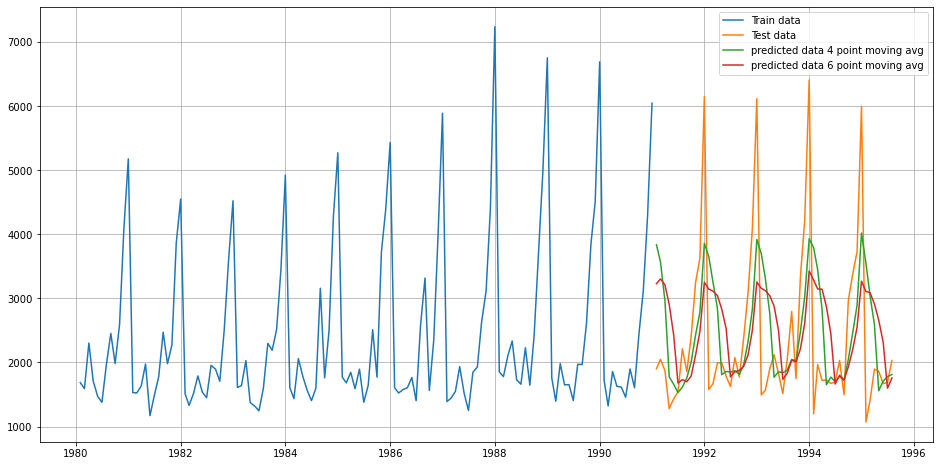
****

Figure 13.Plot for model prediction using Moving average(4point MA and 6 point MA)

**Simple smoothing method(SES)**

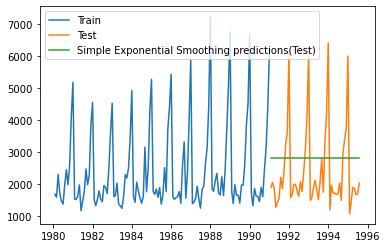
SES or one-parameter exponential smoothing is applicable to time series which do not contain either of trend or seasonality. Forecast by SES is given by:

Ŷt+1 = 𝛼𝑌𝑡+𝛼(1−𝛼)𝑌𝑡−1+𝛼(1−𝛼)2𝑌𝑡−1+⋯, 0<𝛼<1

where, 𝛼 is the smoothing parameter for the level. In reality such a series is hard to find. This is a one-step-ahead forecast where all the forecast values are identical.

After running Simple exponential model we get smoothing\_level(𝛼)= 0.0703

The prediction plot for SES is as below



Model Evaluation:

Now lets check the Evaluate model by calculating RMSE

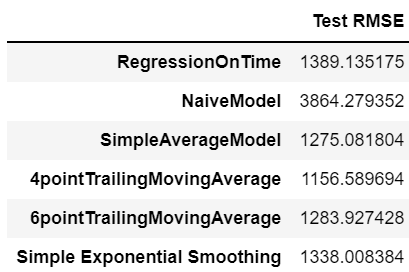


Table 13. RMSE Simple Exponential Smoothing

We can conclude from above table that SES has performed better than Liner regression but worst than any other model.

**Double Exponential Smoothing(Holt’s method)**

This method is an extension of SES method, proposed by Holt in 1957. This method is applicable where trend is present in the data but no seasonality.

The forecast values are given as:

Forecast equation :

Ŷ𝑡+1=𝑙𝑡+𝑏𝑡

Level Equation :

𝑙𝑡=𝛼𝑌𝑡+𝛼(1−𝛼)𝑌𝑡−1, 0 < 𝛼 < 1

Trend Equation :

𝑏𝑡=𝛽(𝑙𝑡−𝑙𝑡−1)+(1−𝛽)𝑏𝑡−1, 0 < 𝛽 < 1

where, 𝑙𝑡 is the estimate of level and 𝑏𝑡 is the trend estimate. 𝛼 is the smoothing parameter for the level and 𝛽 is the smoothing parameter for trend.

Post use of DES model model parameters for smoothing\_level(𝛼) is 0.666,and smoothing\_trend( 𝛽) is 0.0001

**Plot for model prediction using Double exponential smoothing**

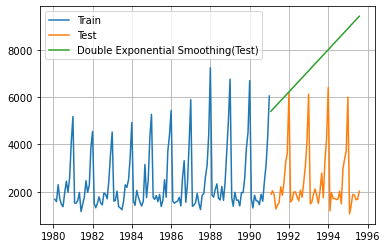


Figure 14. Model prediction using DES

Model Evaluation:

Now lets check the Evaluate model by calculating RMSE

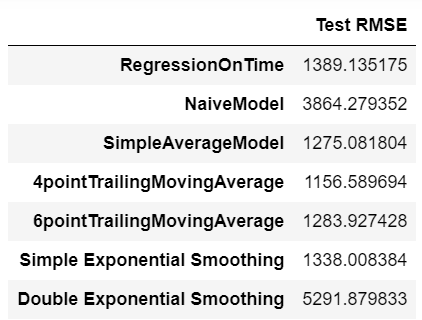


Table 14. RMSE Double Exponential Smoothing

We can conclude from above table that DES has performed worst for the dataset.

**Plot for model prediction using Double exponential smoothing**

**Triple Exponential Smoothing(Holt-Winter’s method )**

This is an extension of Holt’s method when seasonality is found in the data.

Forecast equation:

Y𝑡+1=𝑙𝑡+𝑏𝑡+𝑠𝑡−𝑚(𝑘+1)

Level Equation:

𝑙𝑡=𝛼(𝑌𝑡−𝑠𝑡−𝑚)+𝛼(1−𝛼)𝑌𝑡−1, 0 < 𝛼 < 1

Trend Equation:

𝑏𝑡=𝛽(𝑙𝑡−𝑙𝑡−1)+(1−𝛽)𝑏𝑡−1, 0 < 𝛽 < 1

Seasonal Equation :

𝛾(𝑌𝑡−𝑙𝑡−1−𝑏𝑡−1)+(1−𝛾)𝑠𝑡−𝑚, 0 < 𝛾 < 1

This is also known as three parameters exponential or triple exponential because of the three smoothing parameters 𝛼, 𝛽 and 𝛾. This is a general method and a true multi-step ahead forecast.

Post use of TES model model parameters for smoothing\_level(𝛼) is 0.1113, smoothing\_trend( 𝛽) is 0.0124 and smoothing\_seasonal(𝛾): 0.4607

**Plot for Triple exponential smoothing predictions**

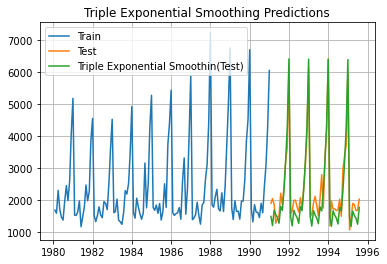


Figure 15. Plot for TES(Holt Winter's method)

Model Evaluation:

Now lets check the Evaluate model by calculating RMSE

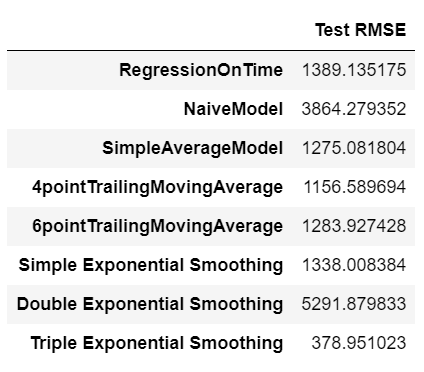


Table 15. RMSE for TES

As we can see from above table. Triple exponential smoothing has performed best so far. It has lowest RMSE than any other model.

1. **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.**

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

The hypothesis in a simple form for the ADF test is:

* 𝐻0: The Time Series has a unit root and is thus non-stationary.
* 𝐻1: The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the 𝛼α value.

After performing adfuller test on series we get following result

DF test Statistics -2.062

P-value for series is : 0.5674110388593686

No. of lags : 119

P-value is higher than 5% significant level and hence the **Time Series is non-stationary**

In order to solve the non-stationary data problem we use differencing of data which makes it stationary series

Let us take one level of differencing to see whether the series becomes stationary.

After performing adfuller test on differenced series we get following result

DF test Statistics -7.968

P-value for series is : 8.479210655514366e-11

No. of lags : 119

As P value now is below 5% of significance level, we can reject null hypothesis and conclude series is stationary.

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 Regressive Integrated Moving Average (ARIMA) models are applied on time series data when the current value is assumed to be correlated to past values and past prediction errors. Therefore, these models are used in defining current value as a linear combination of past values and past prediction errors.

𝐴𝑅𝐼𝑀𝐴(𝑝,𝑑,𝑞) Model: ARIMA is defined by 3 parameters

𝑝 : No of autoregressive terms

𝑑 : No of differencing to stationarize the series

𝑞: No of moving average terms

As from the stationary test we confirmed that series is non stationary and using one level differencing we get a stationary series . hence d=1

**Autoregressive Process(p)**

When the current value of variable can be expressed as a linear function of its past values then, it is known as an auto-regression process

An autoregressive process of order p is a sequence of a random variable 𝒀𝒕 defined by the rule:

𝒀𝒕=𝜷𝟏𝒀𝒕−𝟏+𝜷𝟐𝒀𝒕−𝟐+𝜷𝟑𝒀𝒕−𝟑+⋯+𝜷𝒑𝒀𝒕−𝒑+𝜺𝒕

**Moving Average(q)**

When the current value of the series is a function of past forecast errors this model is known as a moving average model

A moving average process of order q i.e. MA(q), is a sequence Yt defined by the rule:

𝒀𝒕=𝝁+𝜺𝒕+𝜶𝟏𝜺𝒕−𝟏+𝜶𝟐𝜺𝒕−𝟐+⋯+𝜶𝒒𝜺𝒕−𝒒

We will do the grid search to find out the best paramaters i.e. the paramters with lowest AIC .

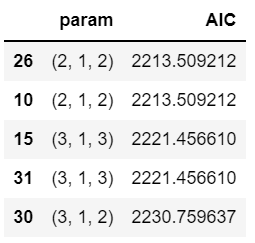


Table 16. Lowest AIC Parameters

As we can see from above table value of p =2 and q=2 .

We already have differenced series of 1 level hence d=1

After using ARIMA model we get following plot for

**PLOT for ARIMA predictions**

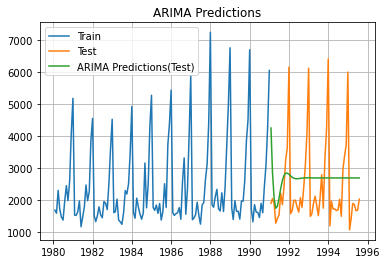


Figure 17.Predictions using ARIMA

As we can see from graph that initially predictions were considerably close to actual observations. However towards the end the predictions are approximately following straight line.

**Model Evaluation:**

Let us evaluate model by quantifying the error RMSE

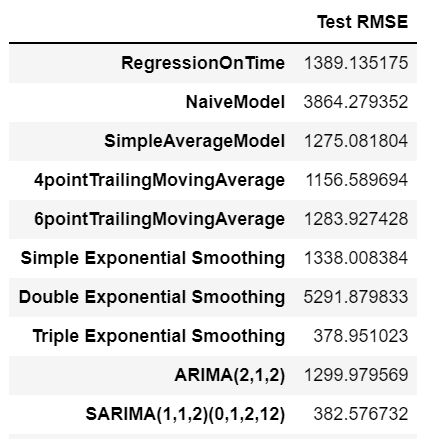


Table 17. RMSE for ARIMA(2,1,2)

As we can see from above table that ARIMA model also has high RMSE. In fact ARIMA has performed worst than 4 point MA, 6 point MA and Triple exponential smoothing

Now Lets include Seasonality in the data an build SARIMA model

**SARIMA(p,d,q)(P, D, Q)[m]**

Seasonal ARIMA models are more complex models with seasonal adjustments. These models are used when time series data has significant seasonality. The most general form of seasonal ARIMA is 𝐴𝑅𝐼𝑀𝐴(𝑝,𝑑,𝑞)∗𝐴𝑅𝐼𝑀𝐴(𝑃,𝐷,𝑄)[m], where P, D, Q are defined as seasonal AR component, seasonal difference and seasonal MA component respectively. And, ‘𝑚’ represents the frequency (time interval) at which the data is observed.

In this case, we have a monthly series will have 𝑚 = 12.

We will again use the grid search to find the best parameters i.e. lowest AIC.

The following parameters have lowest AIC values:



Figure 18. Lowest AIC parameters for SARIMA

That means parameters for ARIMA(p,d,q) is p=1, d=1 and q=2

Paramaters for ARIMA(P,D, Q) is P=0, D=1 and Q=2

Now lets build model and predict on test data. Following plot shows the predictions using SARIMA model

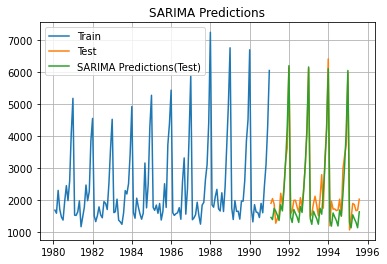


Figure 19. Predictions using SARIMA(1,1,2)(0,1,2,12)

As we can see that predictions are very close to actual observations. SARIMA has perfomed better than other models considering the plot above. Lets quantify the model evaluation.

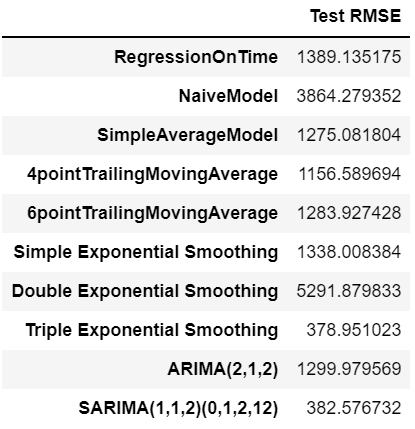


Table 18. RMSE for SARIMA

As we can see from above table that SARIMA has performed considerably well on series. RMSE for SARIMA is one of the lowest except Triple exponential smoothing.

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.**

**Autocorrelation Function (ACF):**

Autocorrelation of order k is the correlation between

𝑌𝑡 and 𝑌𝑡+𝑘 for all values of 𝑘=0,1,…, −1≤𝐴𝐶𝐹≤1 and 𝐴𝐶𝐹(0)=1. ACF measures strength of dependency of current observations on past observations.

**Partial Autocorrelation Function (PACF):**

PACF of order 𝑘 is the autocorrelation between 𝑌𝑡 and 𝑌𝑡+𝑘 adjusting for all the intervening periods i.e. it provides the correlation value between current and 𝑘 - lagged series by removing the influence of all other observations that exist in between.

ACF and PACF used together to identify the order of the ARMA. Seasonal ACF and PACF examines correlations for seasonal data

Let us plot the ACF and PACF on train data. As we already know that series in non stationary from dickey fuller test. Considering single level of differencing make series stationary hence we will use differential series for acf an PACG plots.

**ACF plot for train data:**

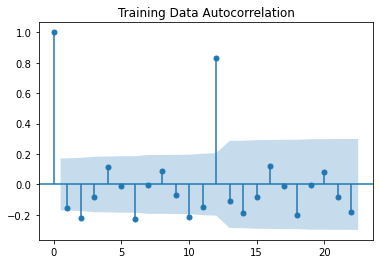
****

Figure 20. ACF plot on differenced series

**The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.**

**PACF plot for train data**

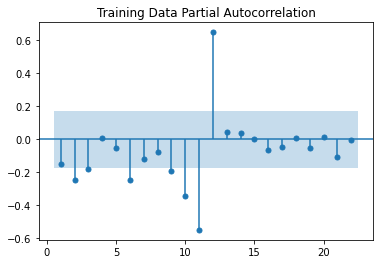
****

Figure 21.PACF plot on differenced series

**The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 3**

By looking at the above plots, we will take the value of p and q to be 3 and 2 respectively.

After using manual Arima we get below summary for ARIMA

THe AIC for Manual ARIMA model is 2230.759636959836

The AIC from this method is quite high as compared to other methods

Model Evaluation:

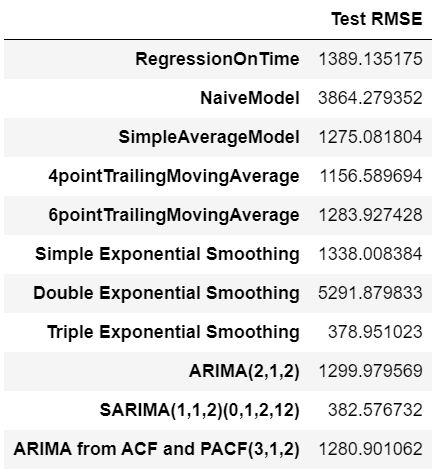


Figure 22. RMSE for ARIMA from parameters using ACF and PACF

As seen from above table selecting p, d, q manually by using ACF and PACF plots have given better ARIMA results than by grid search.

Now Lets decide on P, D, Q considering seasonality

Here, we have taken alpha=0.05.

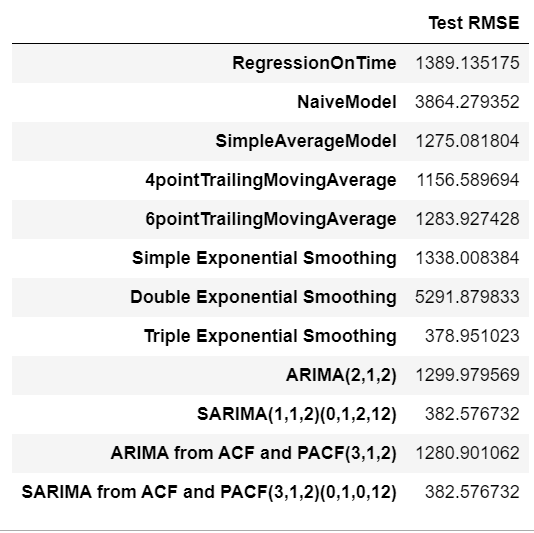
We are taking the p value to be 3 and the q value also to be 2 and d =1 as the parameters same as the ARIMA model.

* The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0. As there is no significant pattern observed.
* The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0. As there is no significant pattern observed.

Hence our model P=0,D=1 and Q=0.

We will build model based on this

The AIC using ACF and PCF plots with parameters as (3,1,2)X(0,1,0,12) is 1382.3477



As we can see from above RMSE from SARIMA using ACF and PACF plots have given same result as SARIMA using grid search

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

The Table for all models is as below:

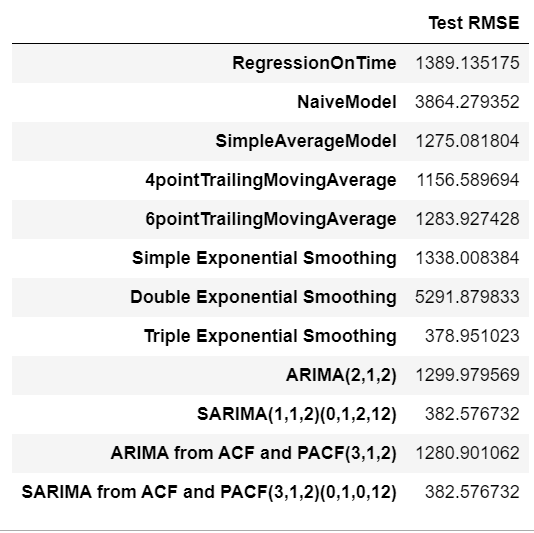


Table 19. RMSE for all models

As seen from table above Triple exponential smoothing and SARIMA models has performed best as the RMSE for these are lowest. As SARIMA model takes care of seasonality in better way we will use **SARIMA model for our predictions**

1. **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.**

As per the conclusion in previous table we will use SARIMA model for model prediction

The model thus built using SARIMA has following summary with AIC as 2562.53

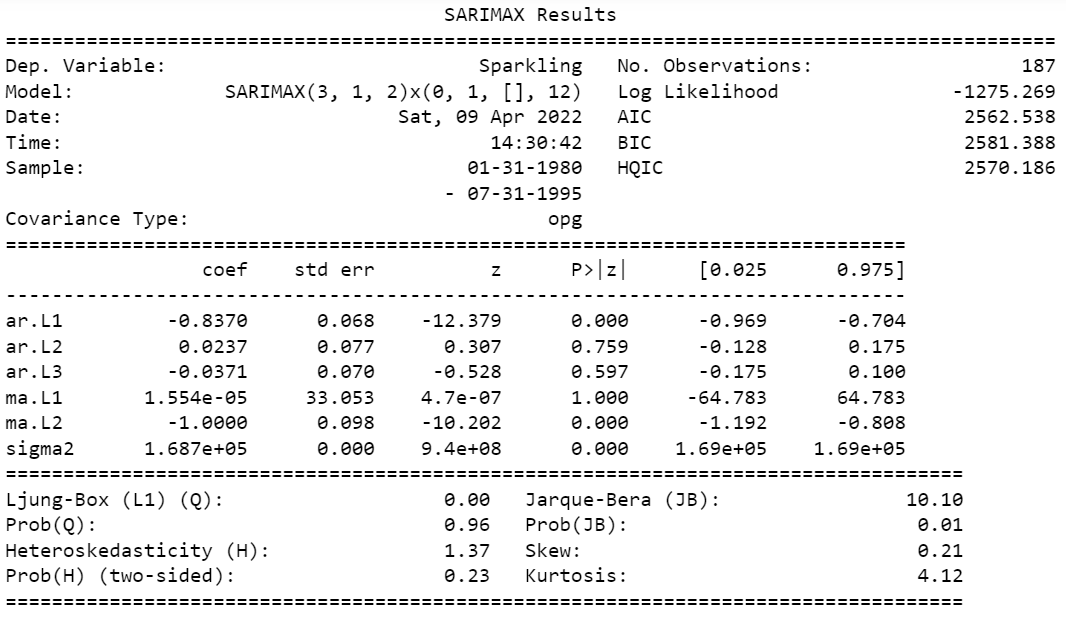


Table 20. Full model summary using SARIMA(3,1,2)x(0,1,0,12)

Below are the first five entries of forecasted values. The table below also shows mean lower value at confidence interval of .05 and mean upper value at confidence interval of .05.

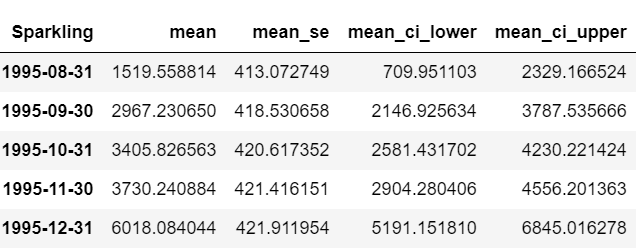


Table 21. Forcasted values for initial 5 months

Lets evaluate the model by calculating RMSE on full model

**RMSE of the Full Model is 575.948**

This is calculated by fittedvalues of model.

**Plot for predictions using SARIMA(3,1,2)(0,1,0,12)**

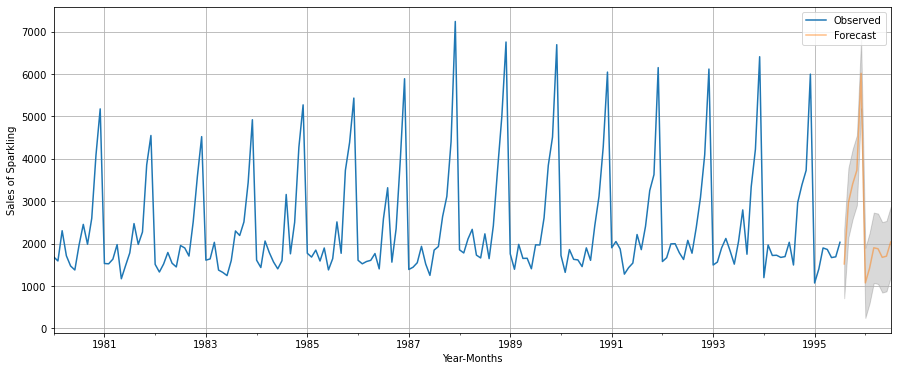


Table . Prediction plot using SARIMA

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

As we had concluded that model used for forecasting of the series is SARIMA model with parameters as (3,1,2)x(0,1,0,12)

We find from the below table that bot Holt Winter’s model and SARIMA model has performed better than any other model and RMSE for both are approximately same. However SARIMA being amore sophisticated method, we have narrowed down to use SARIMA model for predictions.

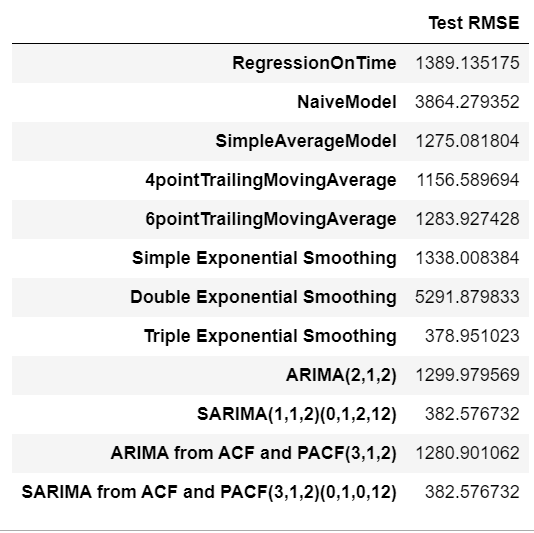


Table 23. comparison of RMSE for all models

Diagnostics of the model thus built

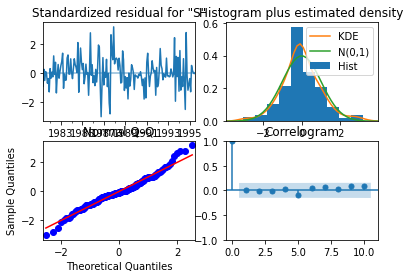


Figure 23. Model diagnostics

Also from the prediction plot we can see that the predictions are following trend and sesonality both.

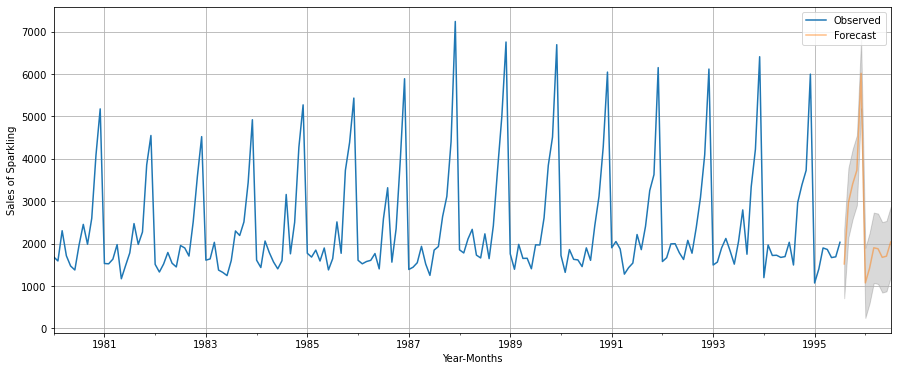
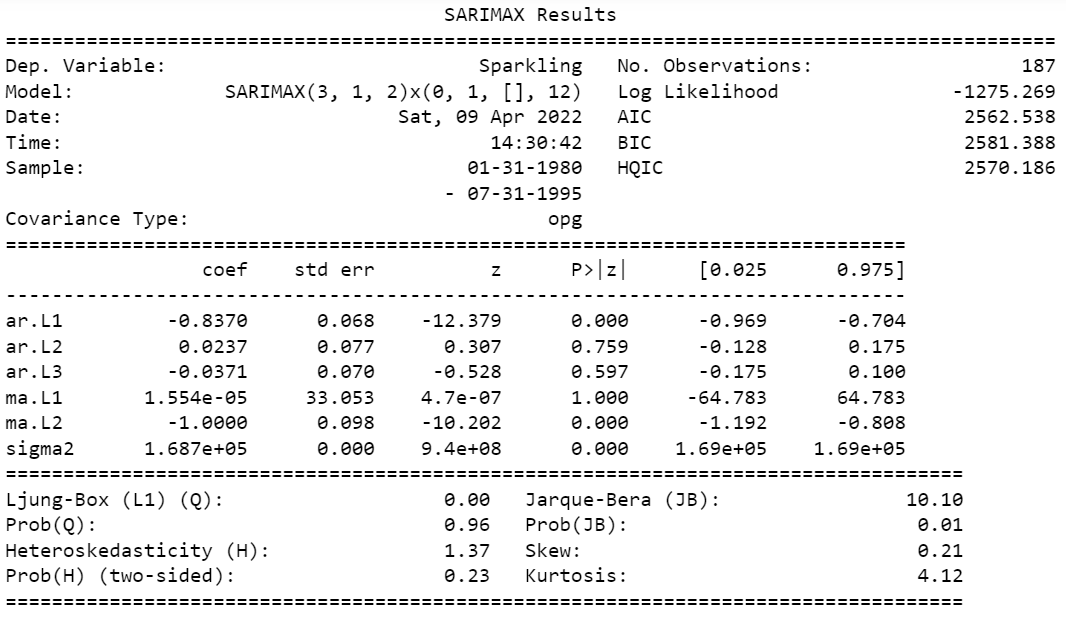


Table 24. Prediction plot

We also calculated the RMSE of predicted values using fitted values function of model and found RMSE of the Full Model is 575.948 which seems considerably low with respect to other models .



Inferences from model predictions:

* From the model summary above we can see that coefficient for auto regression AR2 and AR3 term is close to 0 and nd the P-Value in ‘P>|z|’ column is highly insignificant.It should ideally be less than 0.05 for the respective X to be significant. We can drop these terms and rebuilt the model again.
* We also see that coefficient of the MA1 term is close to zero and the P-Value in ‘P>|z|’. This means this is also insignificant,
* We also see from diagnostics plots. The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals that should be used in computing forecasts.
* The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.
* Confidence interval of .05 is centered around for the mean with lower CI and upper CI following same trend as mean.

Business Implication

* From the plot of model prediction we see that the predicted values follows trend and seasonality. We see increased sales of sparkling wine from September onwards with its peak in Dec(highest sales). This peak as predicted can reach upto means sales of 6000 with higher confidence interval reaching to 7000. Hence business should be prepared to stock up the volume starting from sept and upto Dec. This means that next 4 months will see steep increase in sales and wine inventories and stock should go up considerablly from sept upto Dec.
* We also see from plot that after Dec the sales are very less and company can save their warehousing and stocking expenses after Dec by maintaining low stock volume.
* We also see the values of p and q from ACF and PACF as 3 and 2 respectively which means that the current value is dependent on previous values of order 3 and hence our forecast value will also be function of previous 3 values. Also the value of q=2 suggests that current value is dependent on previous forecast errors of order 2. This means that future sales can be predicted using these values. This understanding can be used for all future forecasts and hence company may choose to perform forecasting again based on actual data of every month.
* We also see from diagnostics plot that the residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.
* For overall prediction there seems to be considerable seasonality in data. However,trend does not seem to show considerable decline or incline effect which ascertains the fact that future sales are predominantly dependent on previous years/months sales.