|  |
| --- |
|  |
| CP G1 Group Assignment |
| Time Series Forecasting – Souvenir Sales Forecasting |

Amit Sharma

Amitesh Bajpai

Sahil Sachdeva

Kushal Maheshwari

Neeraj Singh

**PROBLEM STATEMENT**

Sales of souvenir data have been provided in the ***fancy.txt*** file.

Part A)

Using the Holts-winters methods and model the data and predict for the next 5 years.

Your submission should contain the complete modelling steps with explanations.

Include pictures and R-code where applicable.

Part B)

Using the ARIMA method model the data and predict for the next 5 years. Your

submissions should contain the complete modelling steps with explanations. Include

pictures and R-code where applicable.

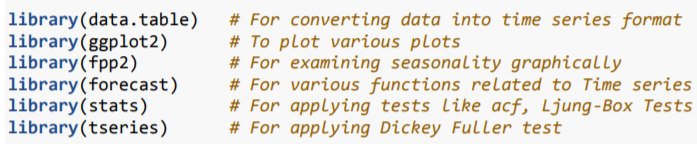
Info on Data:

contains monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, for

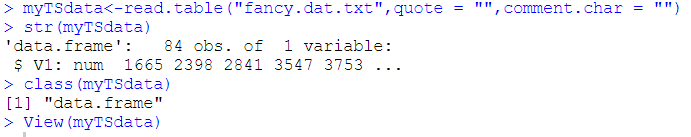
January 1987-December 1993

**INSTALL/LOAD REQUIRED LIBRARIES**

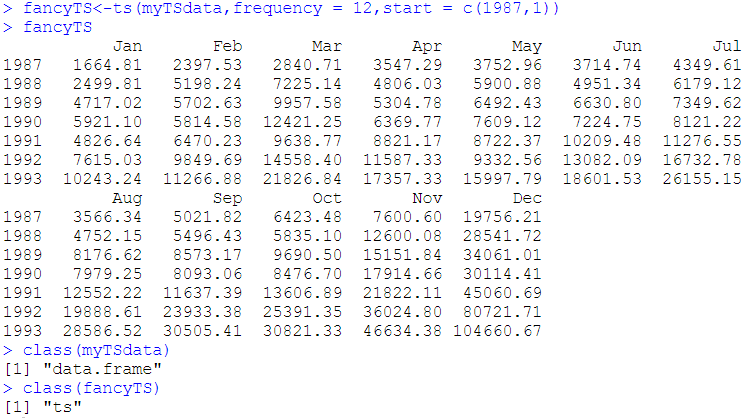
We install and below libraries in R studio



**LOADING THE DATA AND CHECKING THE SALES SERIES PATTERN**

There are 84 observations in the dataset. This dataset contains monthly sales for a souvenir shop at beach resort town in Queensland, Australia, from January 1987 to December 1993. The above data is not of Time Series class as we stored the data in a data-frame.

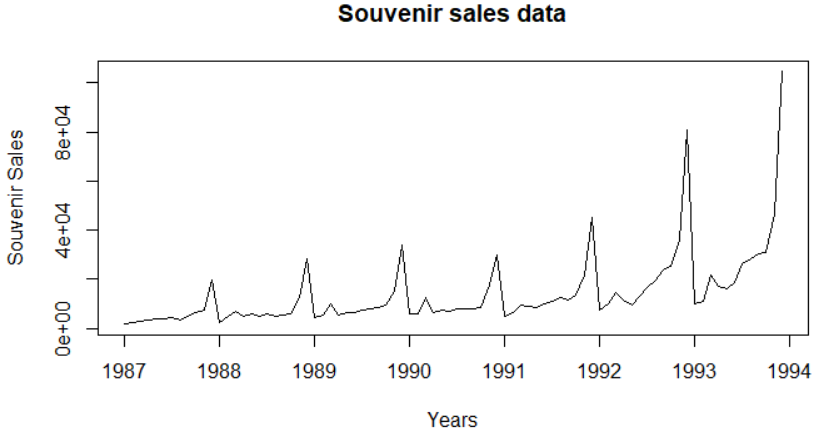
We need to convert the data in the data frame into “***ts***” class. This must be done, so that we can use the data for forecasting based on the time series data. We will use the below command to convert the data set into a time series class and then we will cross validate the class of the dataset.



By looking at the sales data we can clearly say that the periodicity is monthly. In other words, Time series data has a monthly periodicity as each observation represents monthly sales for a souvenir shop.

Next, we plot the time series data.





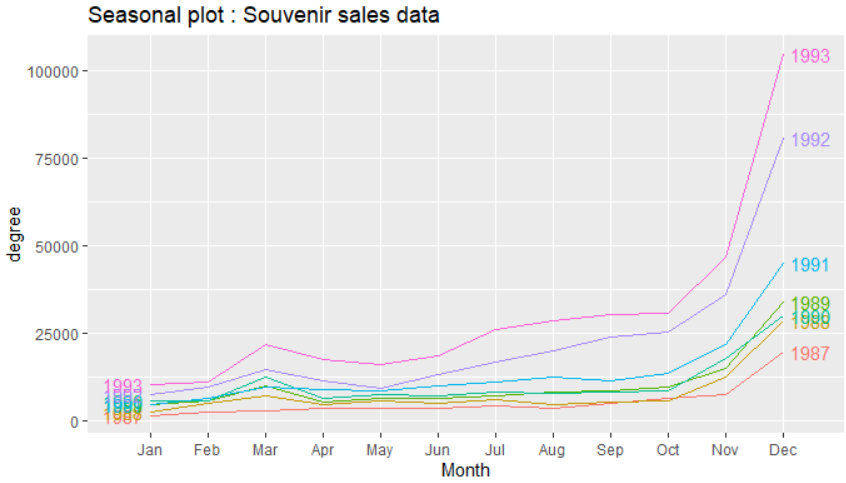
Following are the observations from above figure.

1. Data values are stored in correct time order and no data is missing.
2. The sales are increasing in numbers, implying presence of trend component.
3. Intra-year stable fluctuations are indicative of seasonal component. As trend increases, fluctuations are also increasing. This is indicative of multiplicative seasonality.

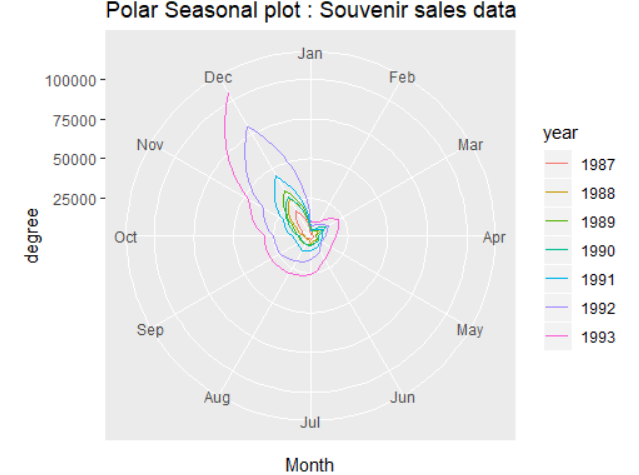
**VISUALIZING THE DATA**

We use three plots to understand seasonality fluctuations better.

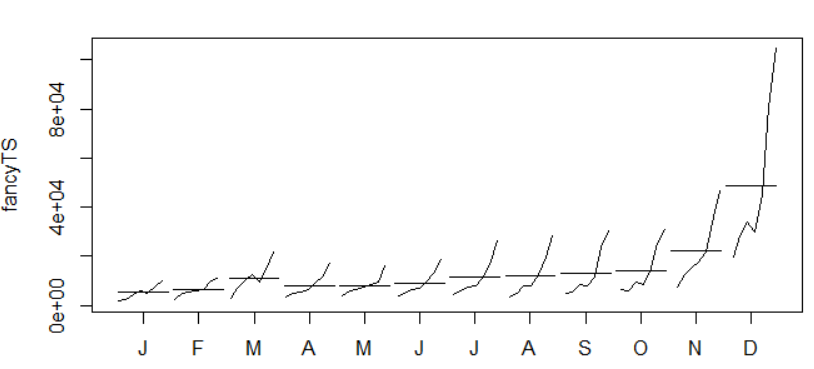












First two graphs show that the sales of tractors are increasing every year in number. Last graph emphasizes that, the vertical lines represent monthly sales and the horizontal lines represent average sales of the given month. Here, it can be observed that average sales are higher towards the year end i.e. November and December as compered other months.

In all these above plots the increasing lines that represent sales have seasonal fluctuations along with a trend. Thus, we can say that it is multiplicative seasonality.

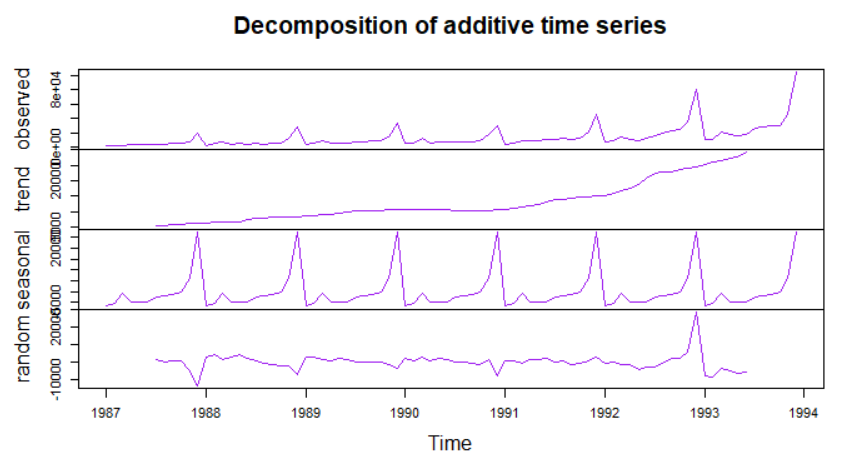
**IDENTIFY COMPONENTS OF SOUVENIR SALES DATA**

Now decomposition method is applied to identify and separate out the three components (i.e. trend, seasonality and noise components) from the given series to observe their independent properties.

We use decompose function in R to decompose the time series by assuming the additive decomposition first.

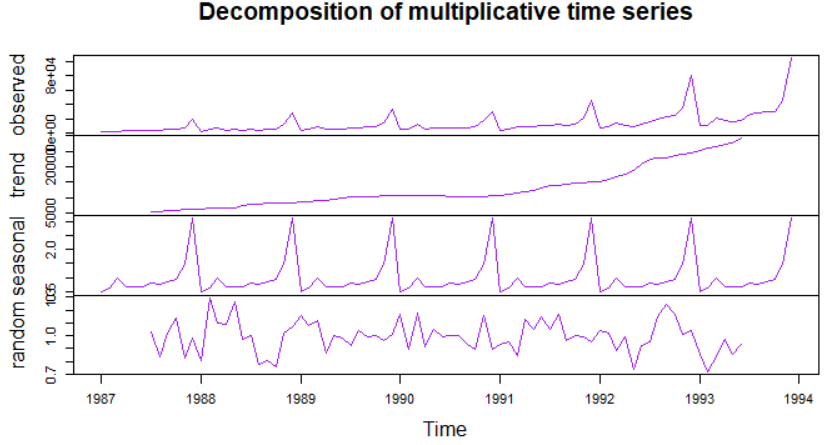




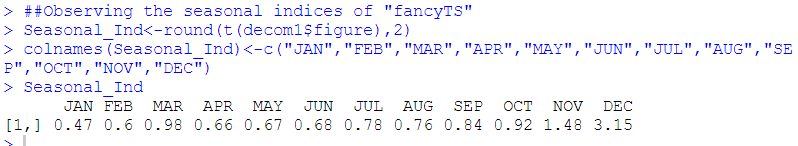


Let’s now see how the components behave when we consider multiplicative decomposition.





The only difference we can see is in the random component of the time series compared to additive decomposition. Above figure indicates that trend is increasing linearly.



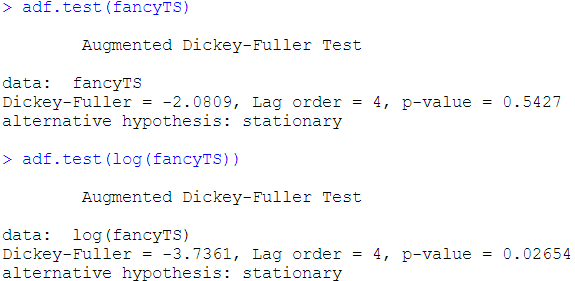
Since this is monthly data, there are 12 seasonal indices. Sum of the monthly indices is 12. In December tractor sales is the highest among all months in the same year, as borne by the highest value of the seasonal component whereas in January (lowest seasonality) tractor sales is the lowest.

**TIME SERIES STATIONARY CHECK**

A process is said to be stationary if its mean and variance are constant over a period and, the correlation between the two time periods depends only on the distance or lag between the two periods. A formal stationarity test needs to be applied to the time series under consideration. We use Augmented Dickey-Fuller Test. This test will determine whether time series data follows stationary process.

H0: Time series is non-stationary

H1: Time series is stationary

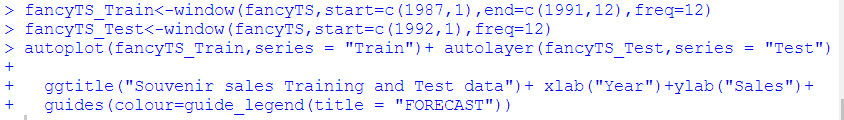


From the above it is clear Souvenir sales data is not stationary. Log transformed data is also checked for stationarity. We conclude that data is non-stationary, and we will convert it into a stationary series later before applying ARIMA model.

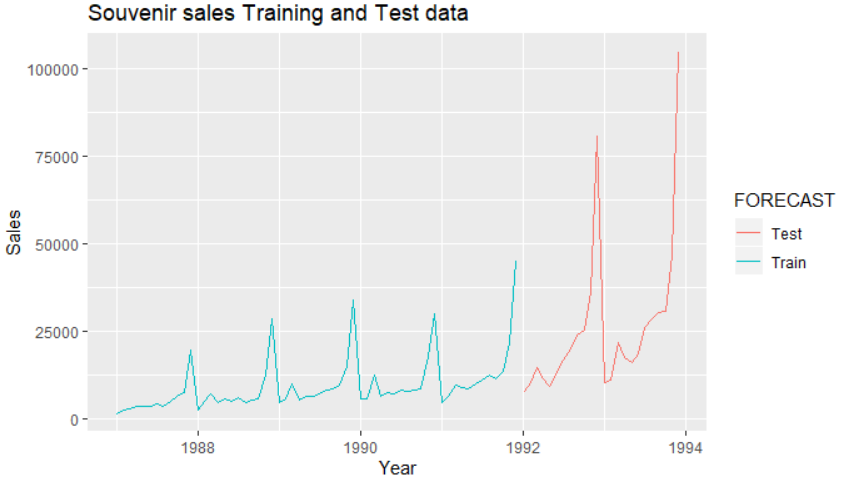
**SPLITTING DATA INTO TEST/TRAIN**

Before a forecast method is proposed, the method needs to be validated. For that purpose, data must be split into two sets i.e. training and testing. Training data helps in identifying and fitting right model(s) and test data is used to validate the same.

In case of time series data, the test data is the most recent part of the series so that the ordering in the data is preserved. For Souvenir sales series, the first 5 years of data is used for training purpose and last 2 years of data is for testing purpose.

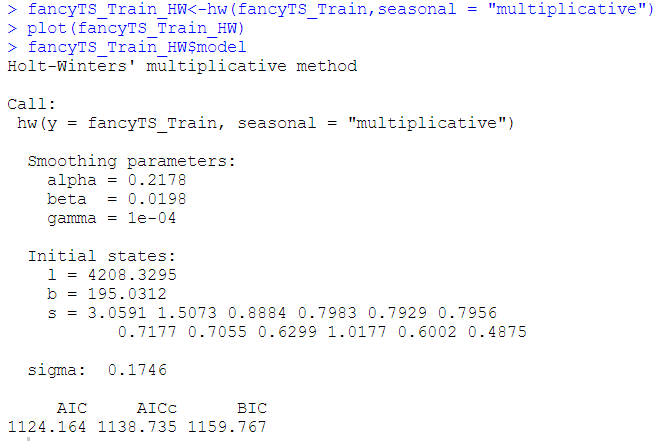


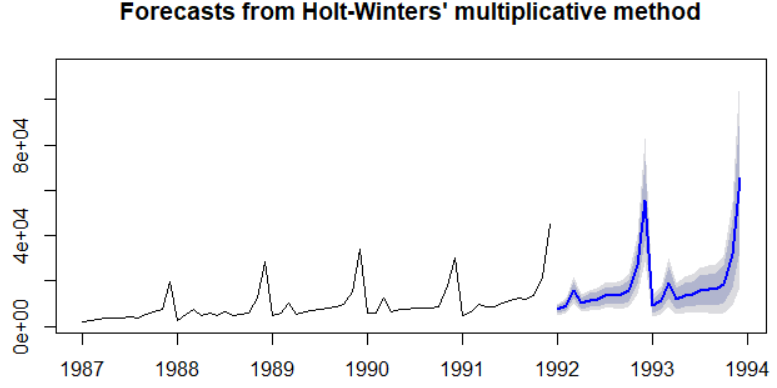




**HOLT-WINTERS’ MODEL CREATION**

We forecast Souvenir sales series using Holt-Winters’ method since it contains both trend and seasonality.

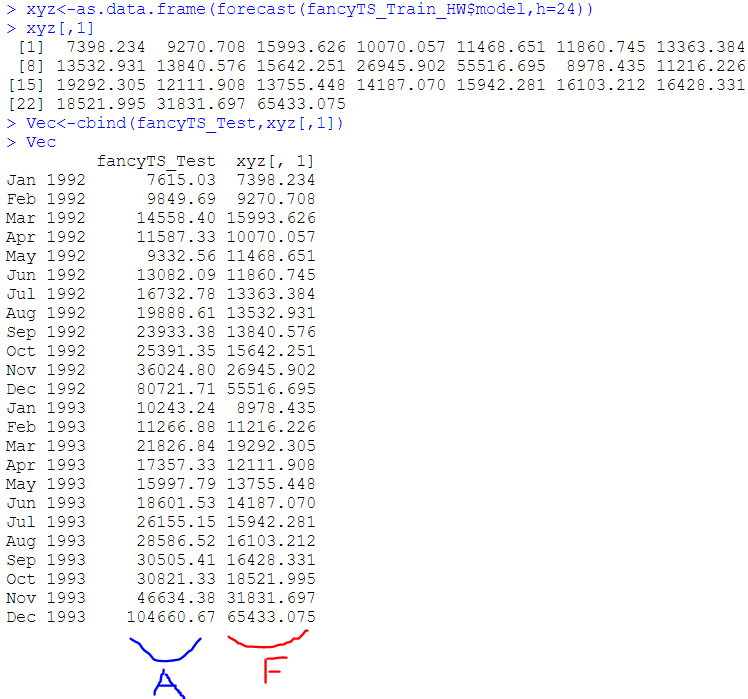




A user may also choose values of **𝛼, 𝛽 𝑎𝑛𝑑 𝛾** and can observe the differences in the model.

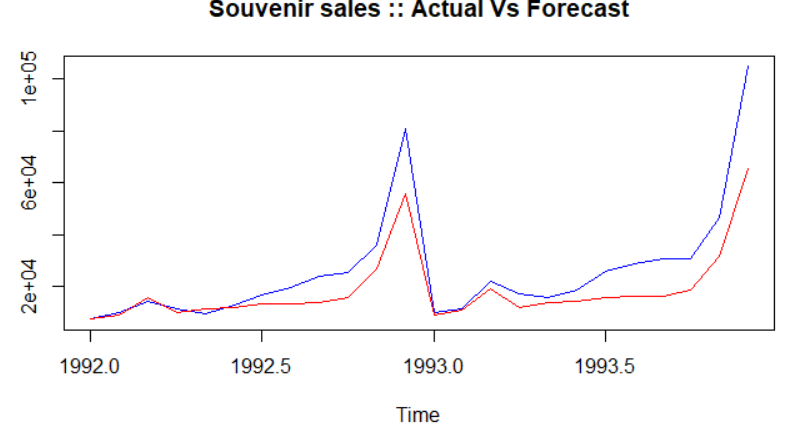
*Predicting the values for test dataset using HW model*

Using the above built model, we forecast values for test dataset and then compare it with actual sales to determine how good our model is. We forecast for 2 years (24 months) i.e. 1992-1993.



The blue marked sales are the actual ones and red ones are the forecasted ones. They have been placed side by side just to have an idea. Also, below plot with closely show the relationship between actual vs forecasted sales.





*Validation using MAPE (Mean Absolute Percentage Error)*

MAPE is computed as the average of the absolute difference between the forecasted and actual values, expressed as a percentage of the actual values. We look at how large the miss was relative to the size of the actual value. Lesser the MAPE better is the model.

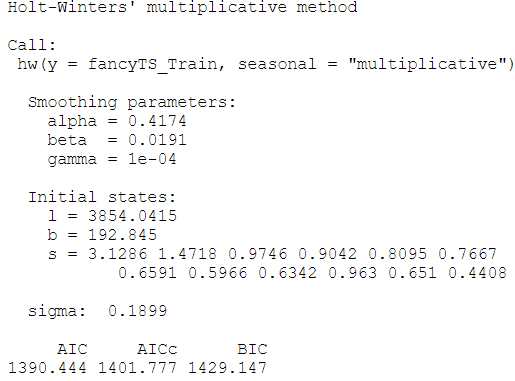
We calculate MAPE for the HW model to ascertain how accurate our model is. The accuracy comes around 76% which is reasonably OK.

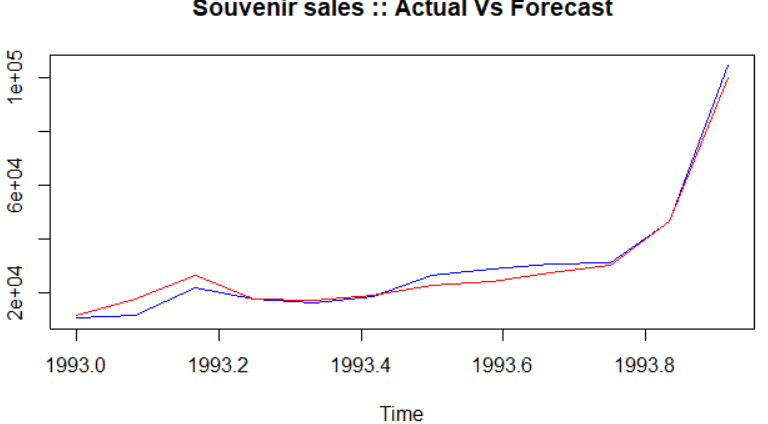


Here, we took first 5 years data as train and last 2 years as test, which makes the Train to Test ratio around 70:30. Now we will look to partition the data into other ratios to look for an improved accuracy. Finally, we consider the least MAPE value model to be the accepted one. We tabulate the accuracy for all considered cases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No. | Train:Test Ratio | Train (Yrs.) | Test (Yrs.) | MAPE |
| 1 | 70:30 | 5 | 2 | 24.31% |
| 2 | 57:43 | 4 | 3 | 67.6% |
| 3 | 86:14 | 6 | 1 | 12.04% |

We can see that model 3 has maximum accuracy, close to 90% which is perfect. We consider HW model 3 as final one for prediction purposes. Below is the final model and its actual vs forecast comparison.



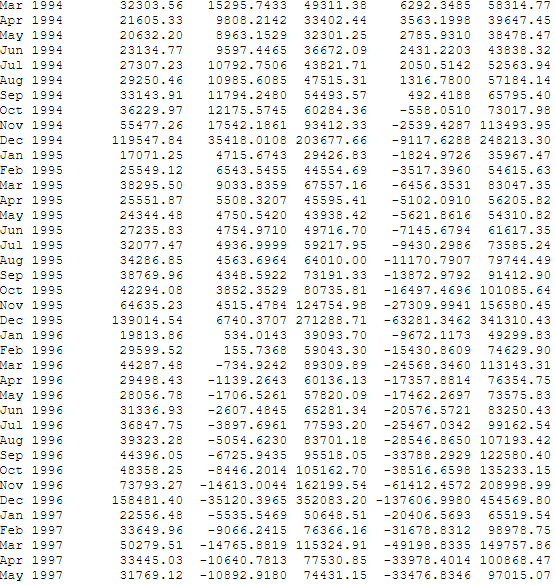


*FINAL PREDICTIONS for next 5 years using HW model*

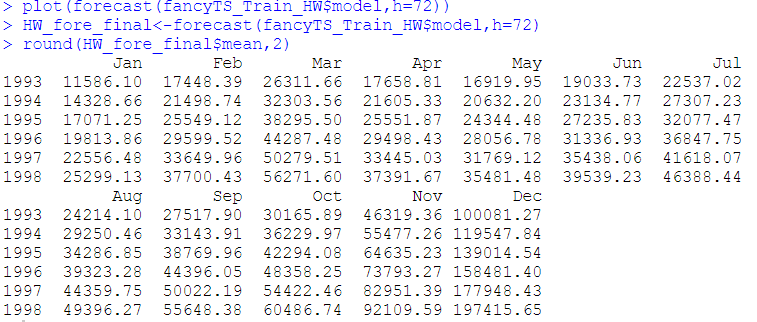
Once a model is chosen and validated, forecasts into the future to be determined. Souvenir sales to be forecasted for 5 years: 1994 – 1998. We predict the sales using chosen HW model 3. We have taken h value as 72 as one extra year is from the test dataset.



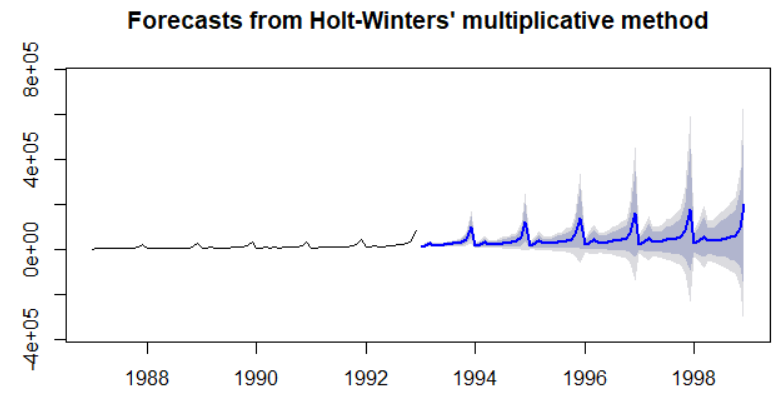




We also take a mean of the forecasted values given by the model to have a deeper understanding.

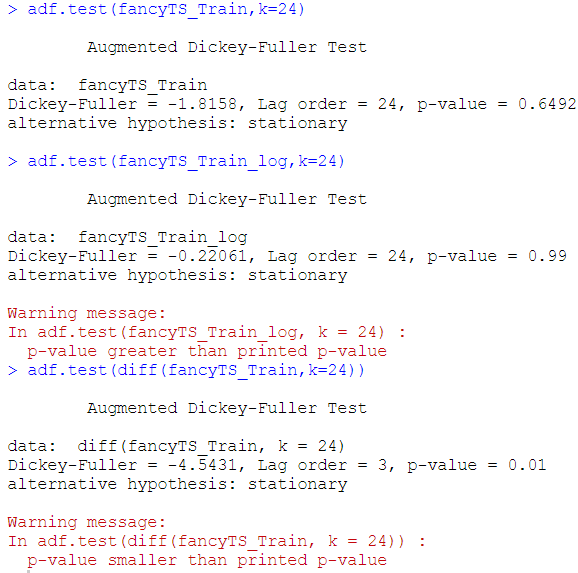


Plot of 5 years of prediction is given below.



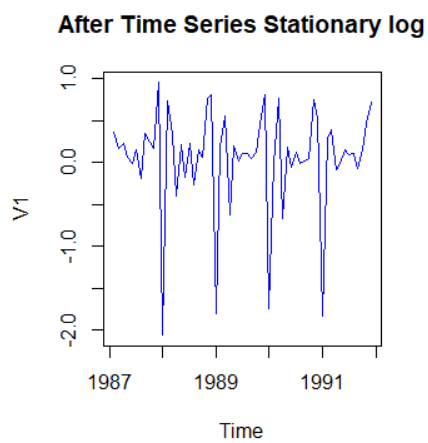
**ARIMA MODEL CREATION**

Since ARIMA model requires a stationary series, we had conducted a formal stationary test earlier and found out that time series under consideration is not stationary. Therefore, we need to convert it into a stationary series. Often differencing a time series lead to a stationary series.



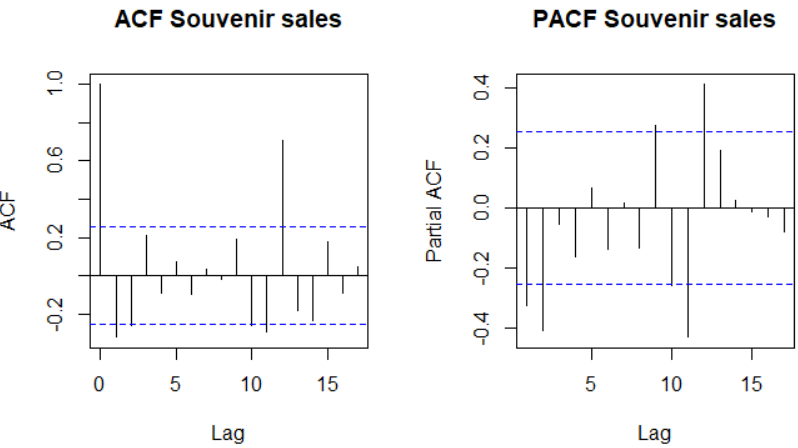
We produce a stationary time series by taking a diff of log Souvenir time series and get a good stationary plot.



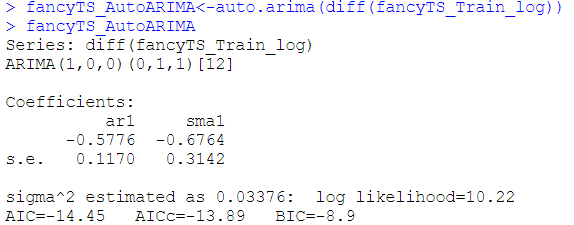


Next, we run ACF and PACF plots. ACF measures strength of dependency of current observations on past observations. PACF provides the correlation value between current and 𝑘 - lagged series by removing the influence of all other series that exist in between. ACF and PACF used together to identify the order of the ARMA. Seasonal ACF and PACF examines correlations for seasonal data.



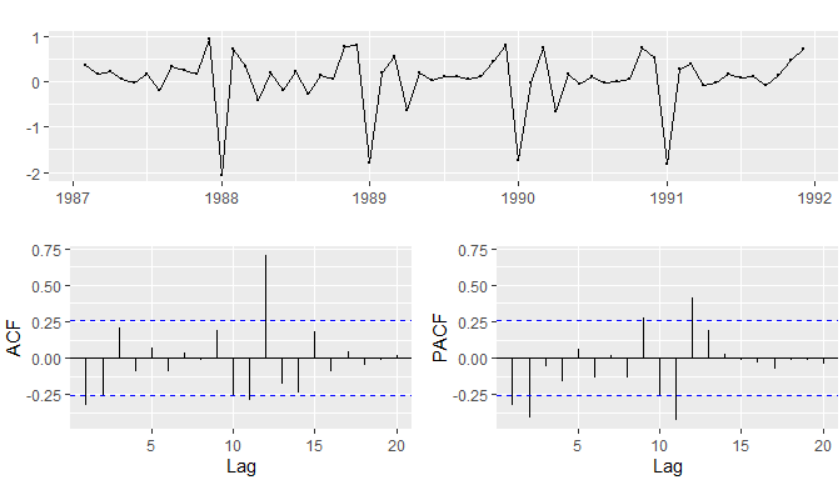


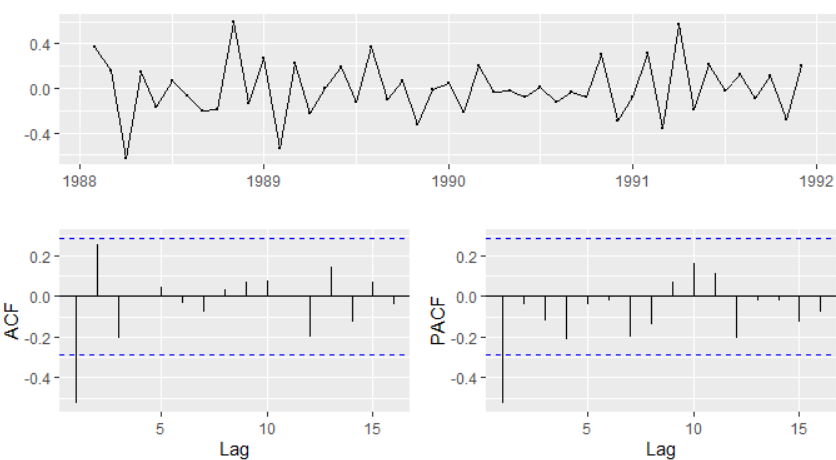
Above figure indicates the presence of seasonality in using ACF and PACF plots of Souvenir sales data as the past values are significantly correlated. And, it is also clear that at every multiple of 12 the ACF swings higher than neighbouring values. We run Auto ARIMA.



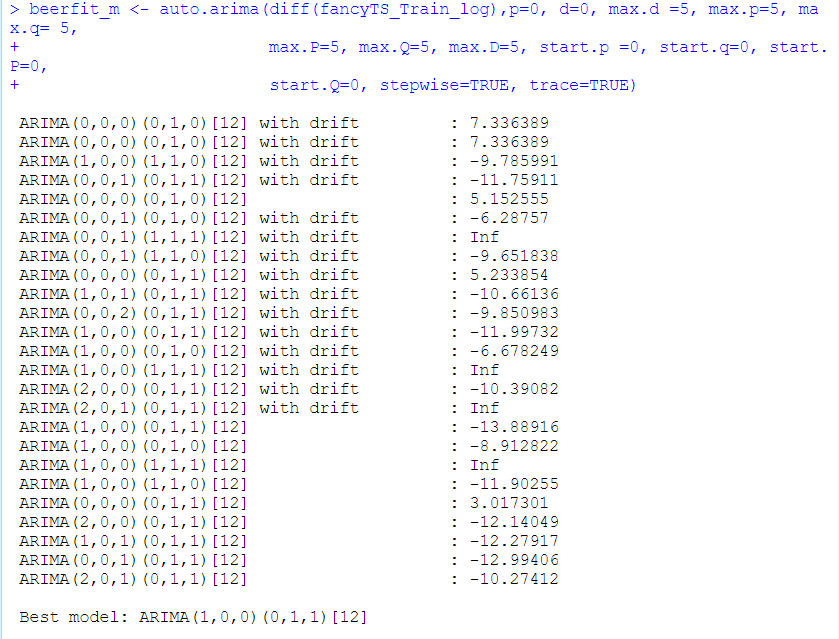
According to the result, ***ARIMA (1,0,0) (0,1,1) [12]*** is the indicated model for Souvenir Sales. Alternatively, one might investigate other suitable model(s) for a time series using ACF and PACF for the differenced series as seen below.



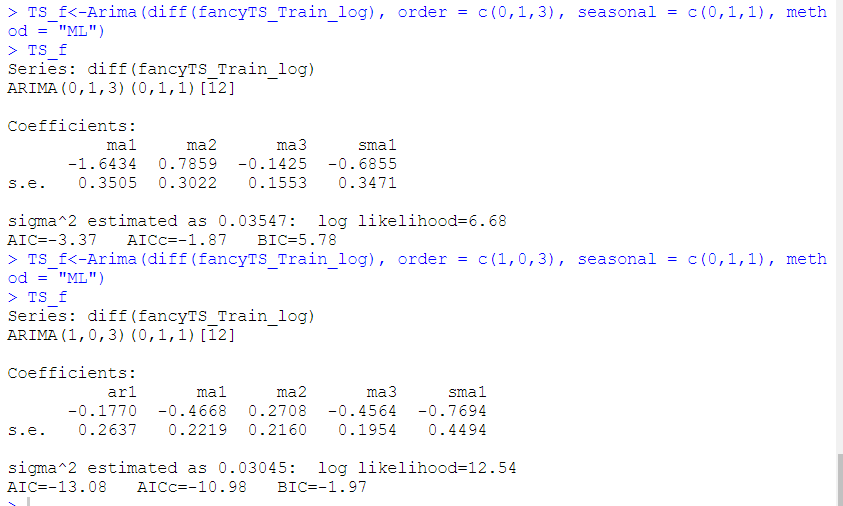




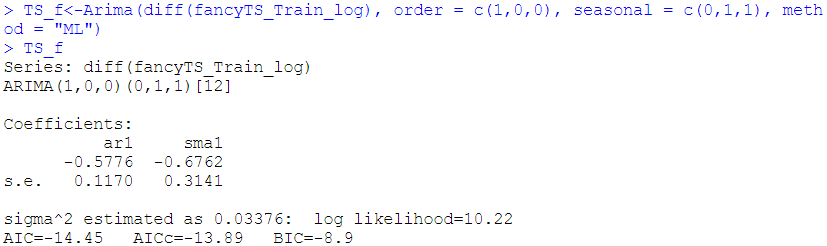
We will use auto.arima which will run iteratively to find out optimal values of p, d and q. Best model will be selected based on AIC value and our optimal model is Best model: ARIMA(1,0,0)(0,1,1)[12]



A user may choose different values of **𝑝, 𝑑, 𝑞 (𝑜𝑟 𝑃, 𝐷 𝑎𝑛𝑑 𝑄)** and compare AIC, BIC values and select the ‘best’ model accordingly. But none of the combinations give lower AIC and BIC values.



We run ARIMA with finalized values.



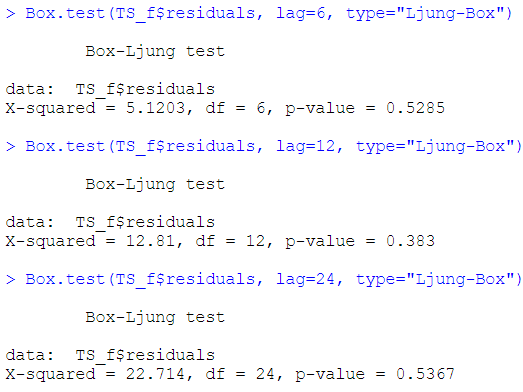
Before ARIMA model is assumed to be reasonable for a series, it is important to check whether the residuals are following white noise or not. Towards that goal Box-Ljung test is applied.

Box- Ljung test: This checks whether the residuals of time series data are stationary or not.

H0: Residuals are stationary

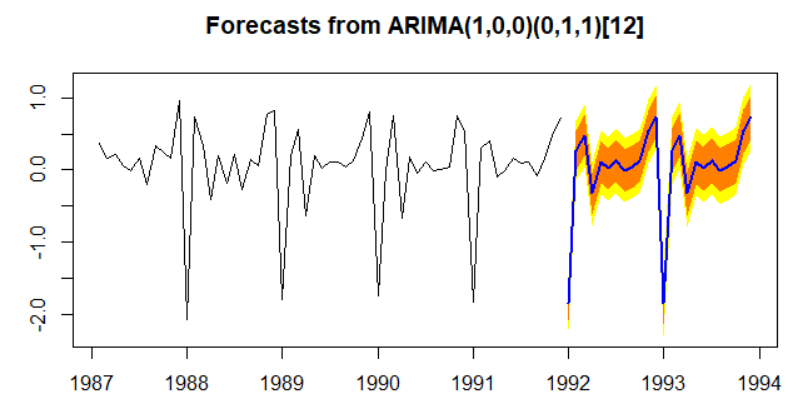
H1: Residuals are not stationary

To check the residual of the series Box-Ljung test has been applied at different lags (k=6, 12 and 24). None of the p-values are significant implying that the series is stationary.



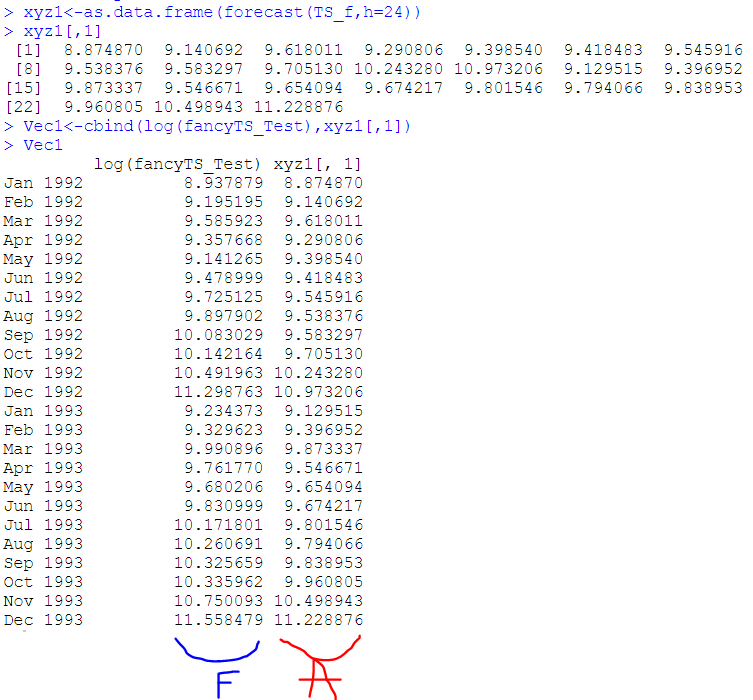
We forecast sample sales to get a first preview of the fitted model.



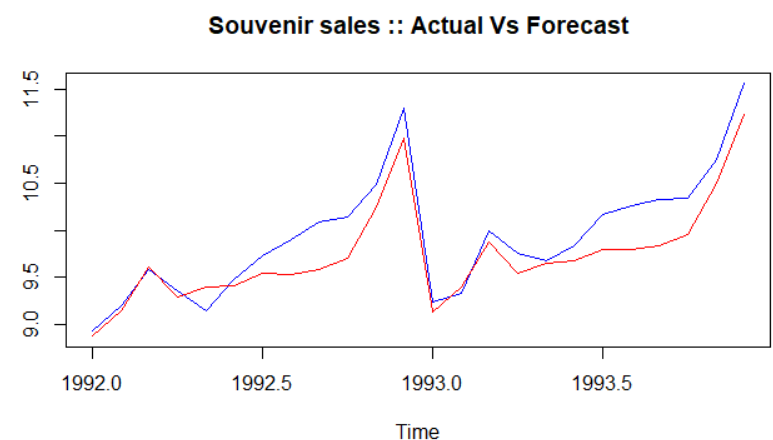


*Predicting the values for test dataset using ARIMA model*

Using the above built model, we forecast values for test dataset and then compare it with actual sales to determine how good our model is. We forecast for 2 years (24 months) i.e. 1992-1993.



The blue marked sales are the actual ones and red ones are the forecasted ones. They have been placed side by side just to have an idea. Also, below plot with closely show the relationship between actual vs forecasted sales.



*Validation using MAPE (Mean Absolute Percentage Error)*

MAPE is computed as the average of the absolute difference between the forecasted and actual values, expressed as a percentage of the actual values. We look at how large the miss was relative to the size of the actual value. Lesser the MAPE better is the model.

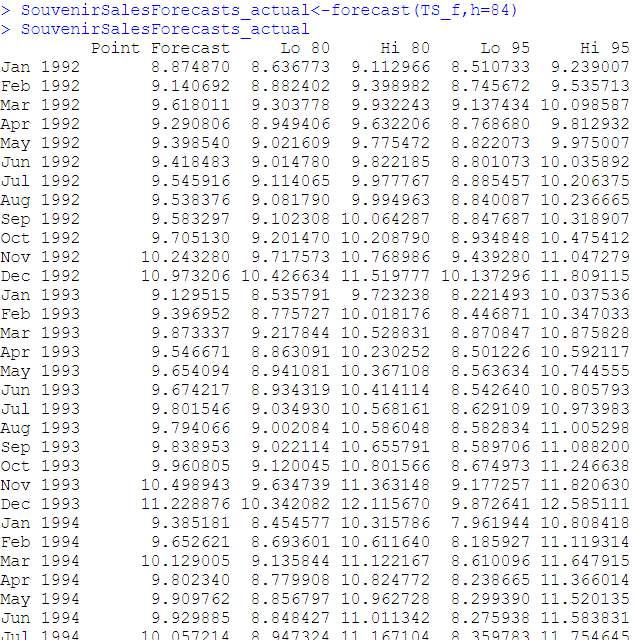
We calculate MAPE for the ARIMA model to ascertain how accurate our model is. The accuracy comes around 2% which is kind of perfect.



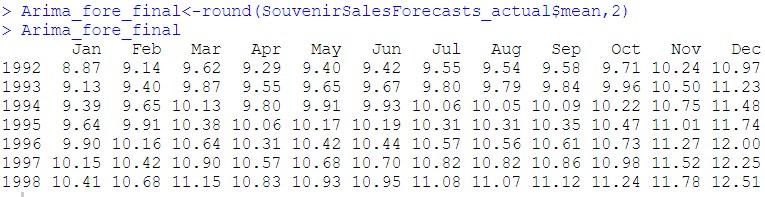
Here, we took first 5 years data as train and last 2 years as test, which makes the Train to Test ratio around 70:30. We consider this ARIMA model as final one for prediction purposes.

*FINAL PREDICTIONS for next 5 years using ARIMA model*

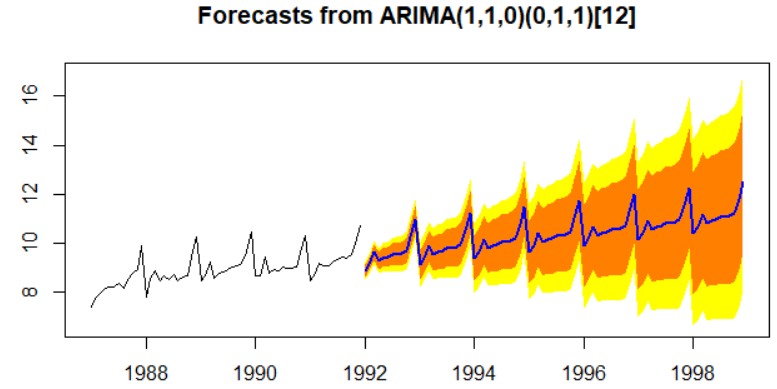
Once a model is chosen and validated, forecasts into the future to be determined. Souvenir sales to be forecasted for 5 years: 1994 – 1998. We predict the sales using chosen ARIMA model. We have taken h value as 84 as two extra years is from the test dataset.



We also take a mean of the forecasted values given by the model to have a deeper understanding.



Plot of 5 years of prediction is given below.



R code is attached.

