

Time Series Forecasting  
Group Assignment – Group6

**Submitted by:**

[Alex James](https://olympus.greatlearning.in/groups/13713/users/7830)

[Krishna Teja Rekapalli](https://olympus.greatlearning.in/groups/13713/users/7838)

Naveen Kumar Sambangi

[Venkateswara Rao Kunche](https://olympus.greatlearning.in/groups/13713/users/7863)

**Problem Statement:**

The attached data shows monthly demand of two different types of consumable items in a certain store from January 2002 to September 2017. The ultimate objective of this exercise is to predict sales for the period October 2017 to December 2018.

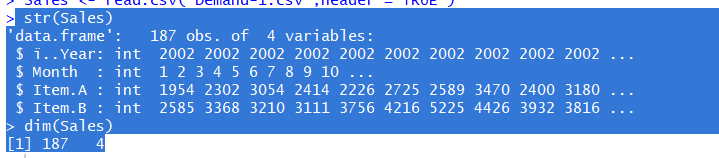
1. Read the data as time series objects in R. Plot the data. What are the major features you notice in the series? How do the two series differ?
2. Before a formal extraction of time series components is done, can you check for seasonal changes in the data for the two series separately? Particularly whether there are more variability in a season compared to the others, whether seasonal variations are changing across years etc. Compare the behavior of the two series.
3. Decompose each series to extract trend and seasonality, if there are any. Which seasonality is more appropriate – additive or multiplicative? Explain the seasonal indices. In which month(s) do you see higher sales and which month(s) you see lower sales? Any difference in the nature of demand of the two items?
4. Can you extract the residuals for the two decomposition exercises and check if they form a stationary series? Do a formal test for stationarity writing down the null and alternative hypothesis. What is your conclusion in each case?
5. Before the final forecast is undertaken one would like to compare a few models. Use the last 21 months as hold-out sample fit a suitable exponential smoothing model to the rest of the data and calculate MAPE. What are the values of α, β and γ? What role do they play in the modeling? For the same hold-out period compare forecast by decomposition and compute MAPE. Which model gives smaller MAPE? Give a comparison for the two demands.
6. Use the ‘best’ model obtained from above to forecast demand for the period Oct 2017 to December 2018 for both items. Provide forecasted values as well as their upper and lower confidence limits. If you are the store manager what decisions would you make after looking at the demand of the two items over years?

**Introduction:** The demand (Demand-1.csv) dataset contains the monthly demand of two different types of consumable items in a certain store from January 2002 to September 2017 and it contains the continuous monthly data without any missing value. The objective is to predict sales for the period October 2017 to December 2018 by using time series forecasting.

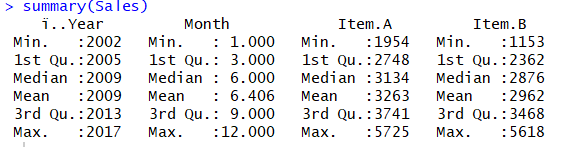
**Steps to resolve the problem:**

**1)Loading the Dataset:**  Firstly, we load the dataset into R and identify the number of variables/columns and records/rows.

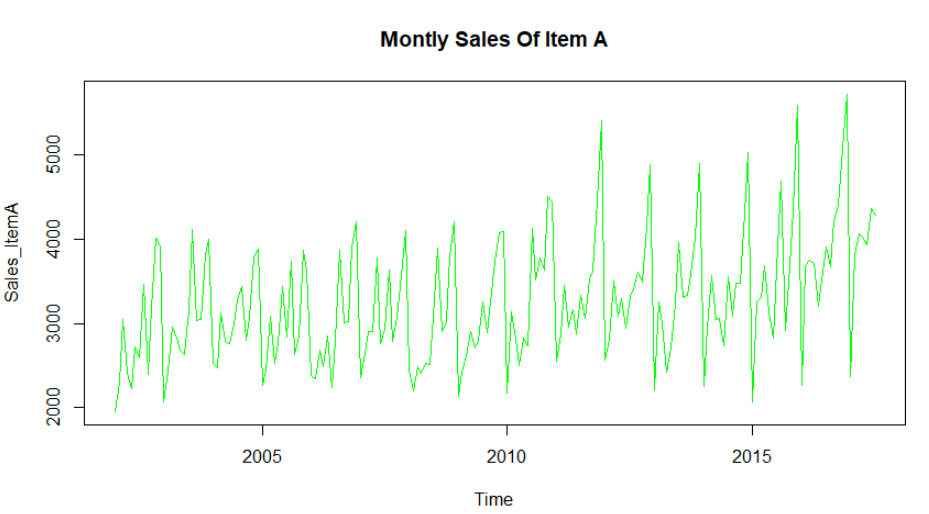
|  |
| --- |
| setwd("E:/r direct/TimeSeries/Assignment") Sales <- read.csv("Demand-1.csv",header = TRUE ) str(Sales) dim(Sales) |
|  |

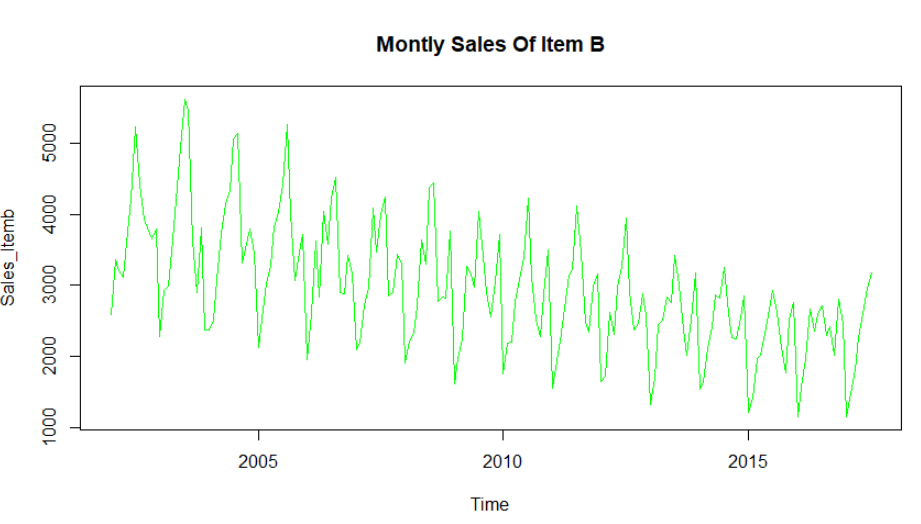


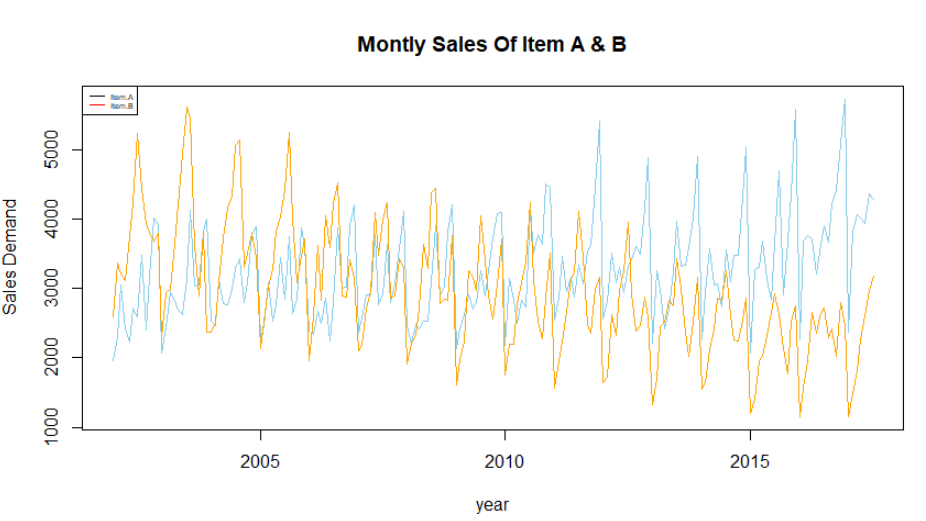
The dataset contains the 187 observations with variables of year, month, Item A and Item B details.



**2)Converting the data into Time Series Components:** Plotting the monthly sales of item A & Item B.







From the above plots, we can identify that Item A has an increasing demand and Item B has fall in demand and also, there is some seasonality and trend in demands.

**3)Decomposition of Time series Dataset:** It is a [statistical](https://en.wikipedia.org/wiki/Statistical) task that deconstructs a [time series](https://en.wikipedia.org/wiki/Time_series) into several components, each representing one of the underlying categories of patterns and it’s decomposed into below components.

* [**Trend component**](https://en.wikipedia.org/wiki/Trend_estimation) at time t, which reflects for the long-term progression of the series. A trend exists when there is a persistent increasing or decreasing direction in the data. The trend component does not have to be linear.
* **Cyclical component** at time t, which reflects repeated but non-periodic fluctuations. The duration of these fluctuations is usually of at least two years.
* **Seasonal component** at time t, reflecting [seasonality](https://en.wikipedia.org/wiki/Seasonality) (seasonal variation). A seasonal pattern exists when a time series is influenced by seasonal factors. Seasonality occurs over a fixed and known period (e.g., the quarter of the year, the month, or day of the week).
* **Irregular component** (or "noise") at time t, which describes random, irregular influences. It represents the residuals or remainder of the time series after the other components have been removed.

In order to decompose the timeseries components we need to choose the additive or multiplicative model.

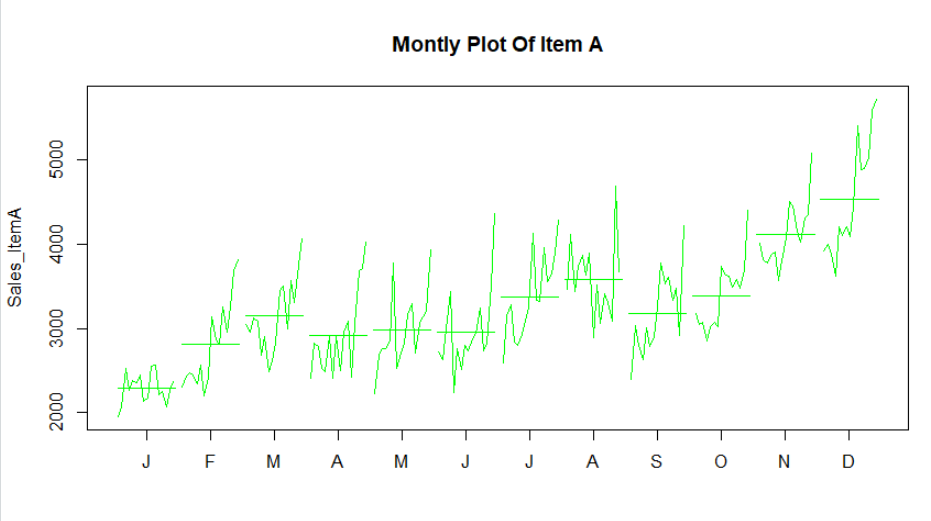
**Additive model:** It is useful when the seasonal variation is relatively constant over time.

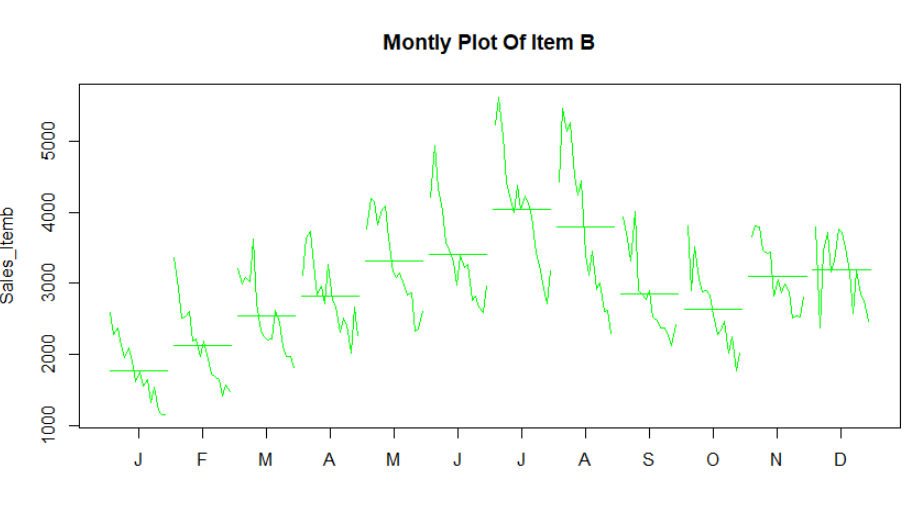
Additive:  Trend + Seasonal + Irregular +Cyclical component

**Multiplicative model:** It is useful when the seasonal variation increases over time.

Multiplicative: Trend \* Seasonal \* Irregular \*Cyclical component

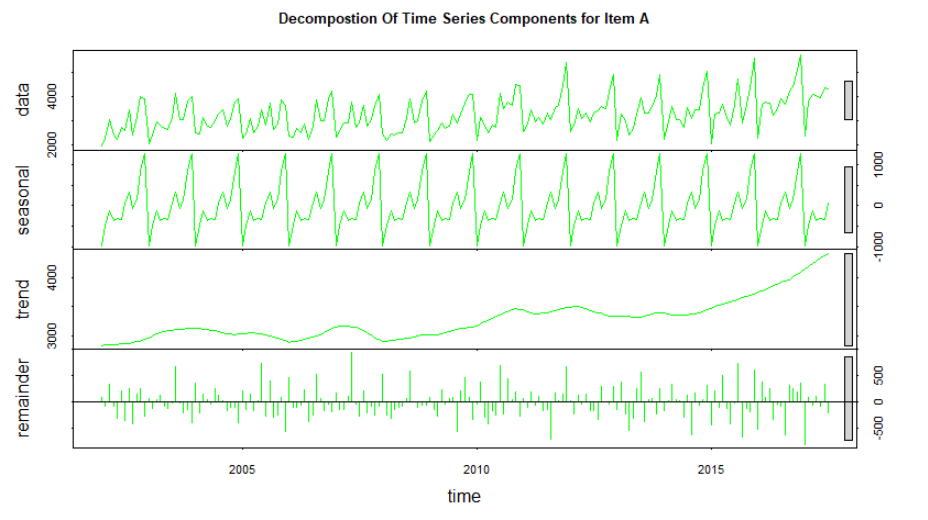
Hence, we are plotting the monthly plot for each of the items to identify the seasonal variation is constant or increasing over time.

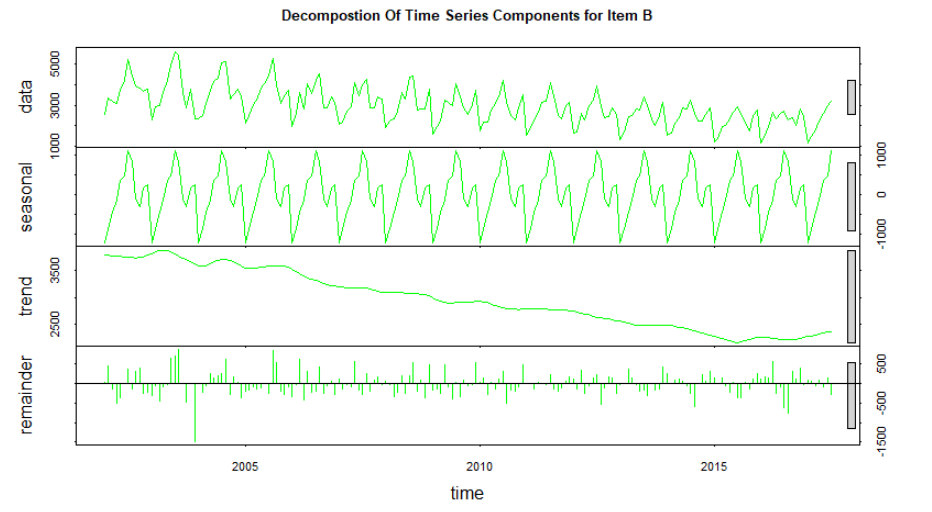




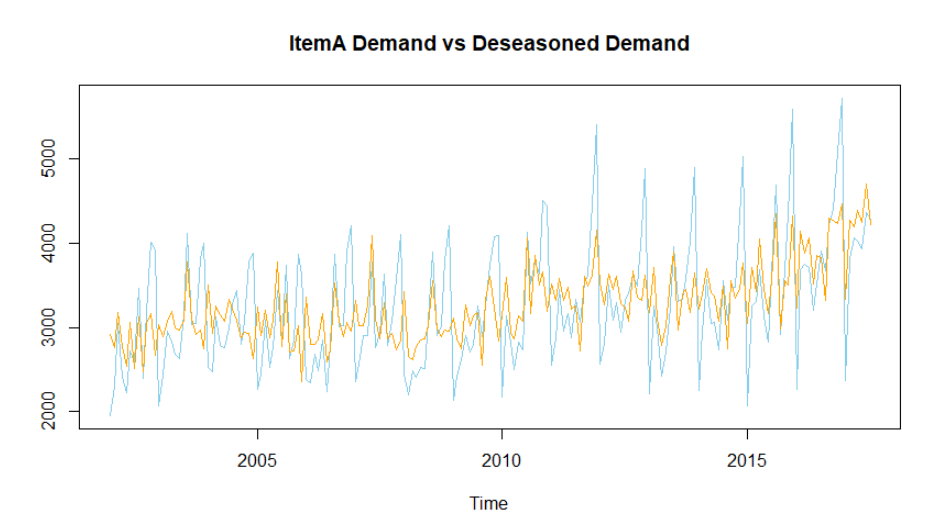
From the above plots identified that there is not much(constant) seasonal variation for both the items. Hence, we are using the **additive model** for decomposition of time series. For Item-A **Dec** month is highest demand and **Jan** is the lowest demand and Item-B **Jul** month is the highest and **Jan** is lowest demand.

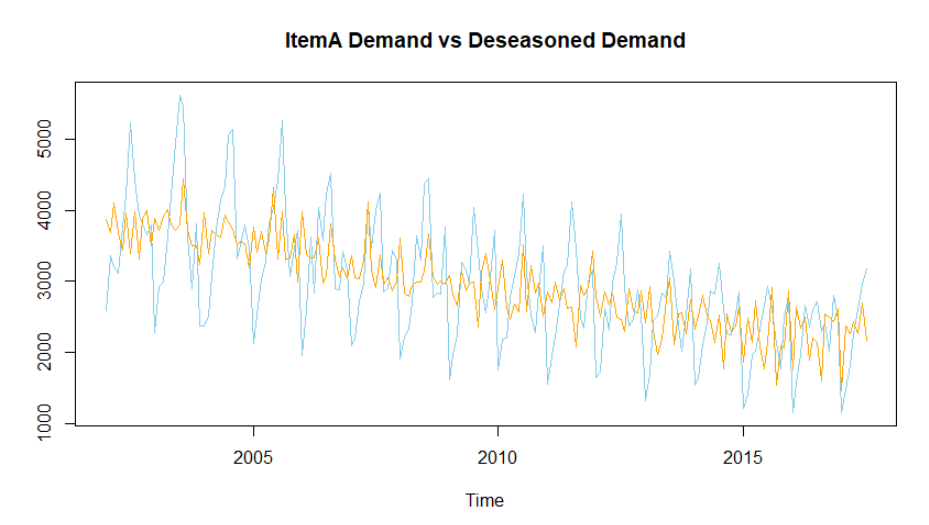
**Decomposition Using STL:**





From the above time series components, it has been identified that Item A has increasing in demand and for the Item B has the decreasing demand over the time. We might like to do is remove just the seasonal effect and leave any trend and the random ups and downs back in the data. The resulting series gives us De-seasonalized data which may give us a clearer picture of what is happening.

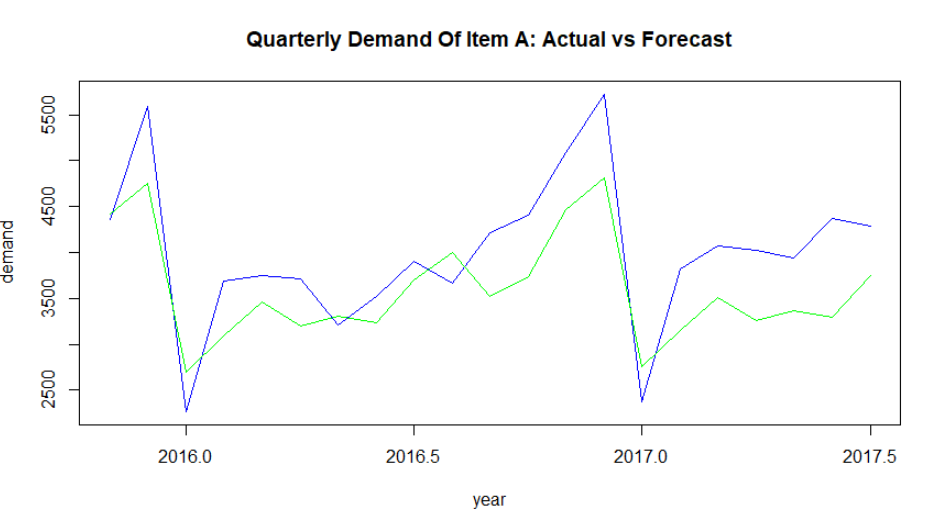




The skyblue color is the actual demand and the orange color is the de-seasonalized demand and we can see the increase trend in demand for Item A and decrease trend for Item B.

**4)Model Building:** For model building we need to convert the data set into dev and holdout same as per the requirement and then decompose it to timeseries components like trend, seasonality and irregular parts.

**Random walk with drift model:** The random walk model predicts that the value at time "t" will equal the last period's value plus a constant, or drift (α), and a white noise term (εt), It also does not revert to a long-run mean and has variance dependent on time.



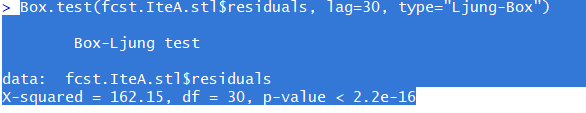
**MAPE for Item A:** The mean absolute percentage error function for the forecast and the output as mentioned below.



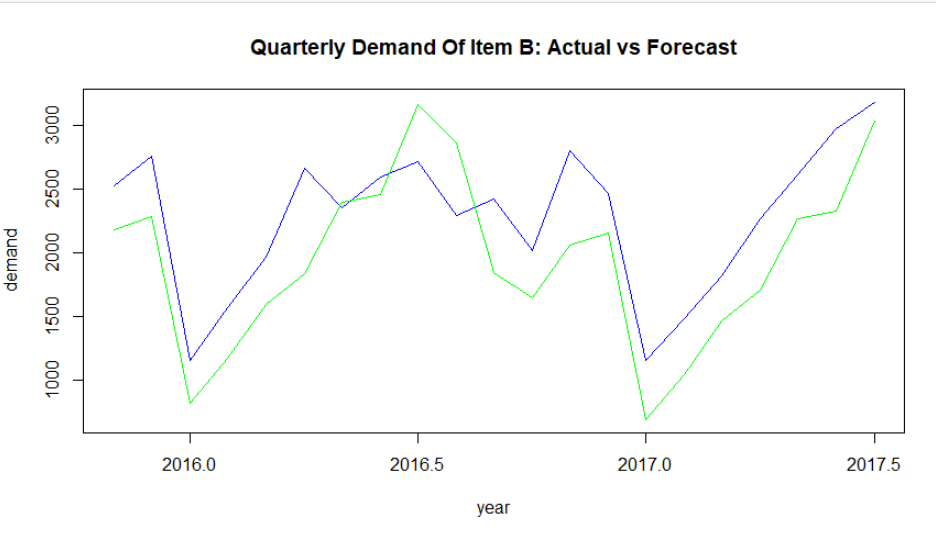
Box-Ljung Test:To check the random/residual components are independent or not.

H0: residuals are independent

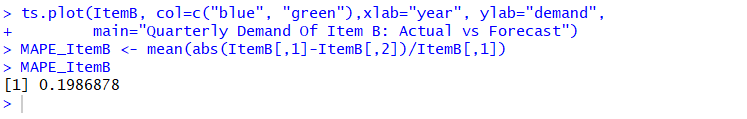
HA: residuals are not independent.



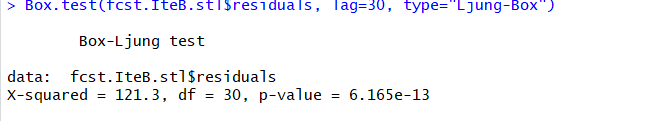
The p-value is less than the 0.05 hence rejecting the **Null hypothesis** which means the residuals are not independent.



**MAPE for Item B:** The mean absolute percentage error function for the forecast of Item B and the output as mentioned below.



Box-Ljung Test:To check the random/residual are independent or not.

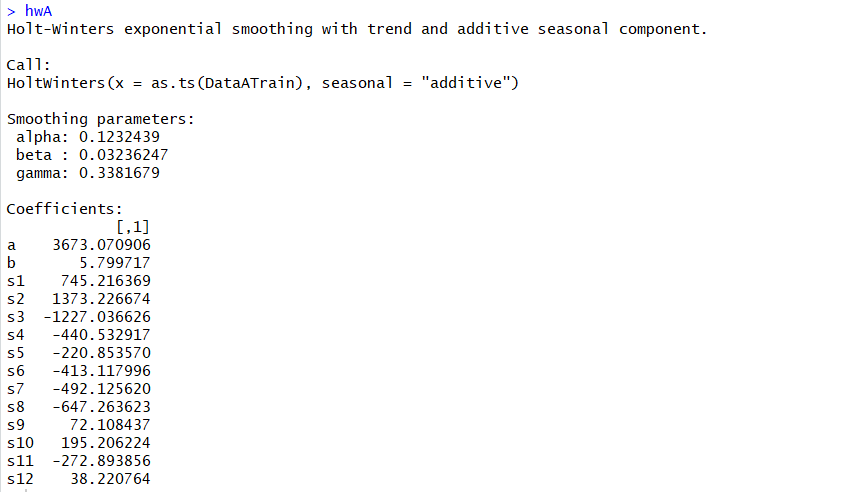


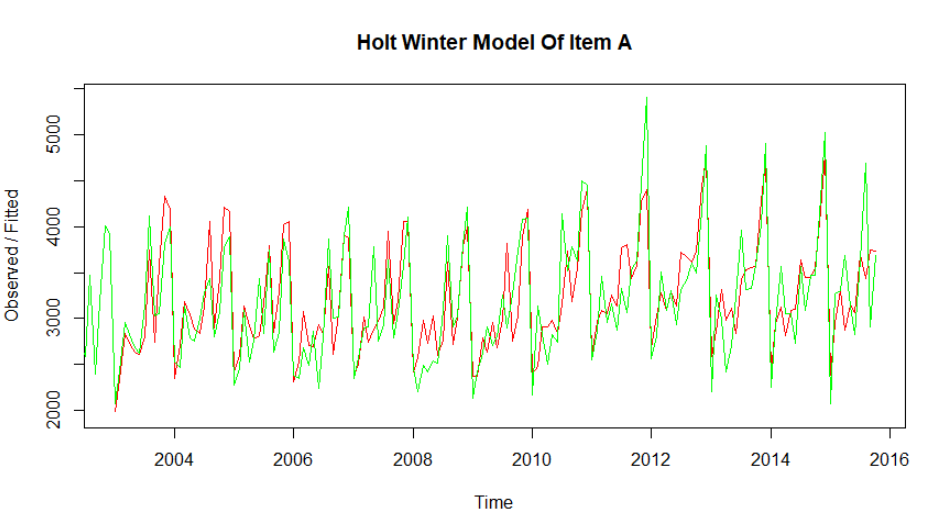
The p-value is less than the 0.05 hence rejecting the **Null hypothesis** which means the residuals are not independent.

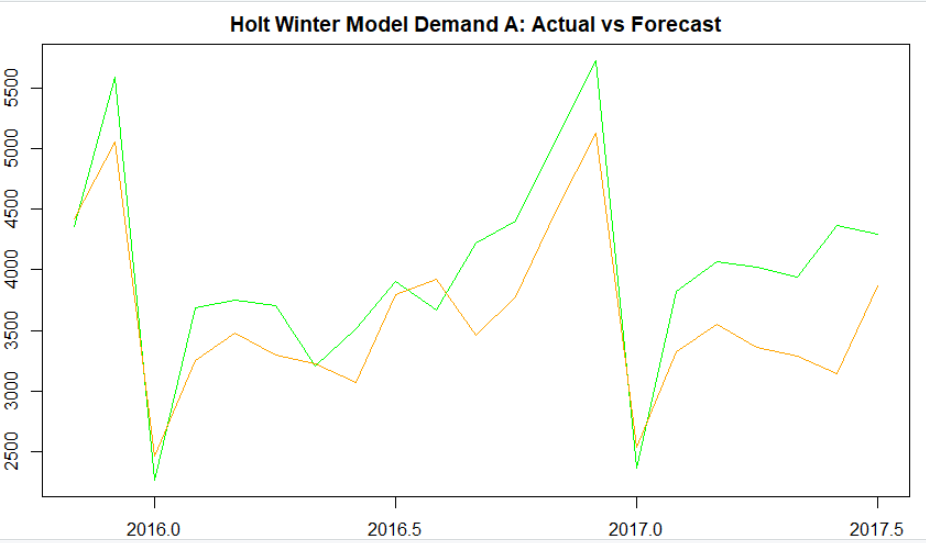
The MAPE results from random walk with drift model we observed that **13.7%** and **19.8%** less accuracy in model.

**Holt Winter Model:** It is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict “typical” values for the present and future.

**Item A:**





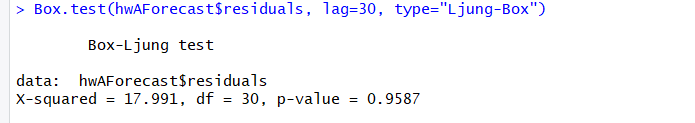


Box-Ljung Test:To check the random/residual components are independent or not.

H0: residuals are independent

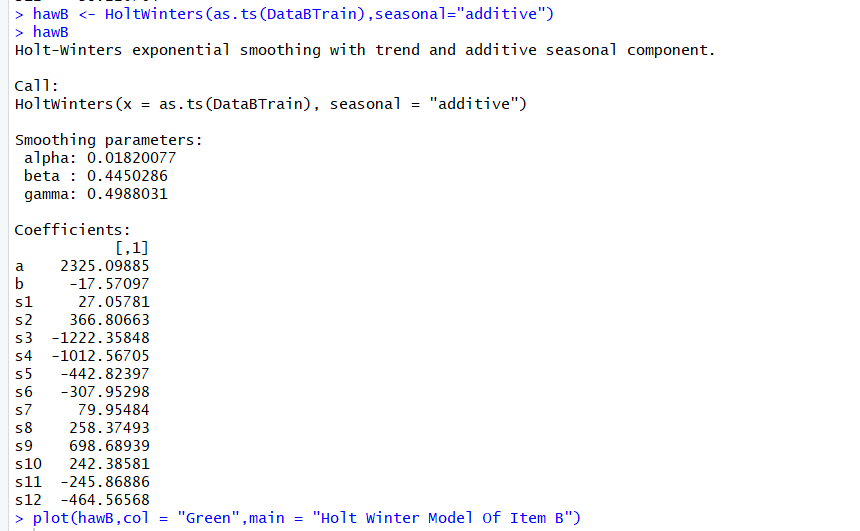
HA: residuals are not independent.

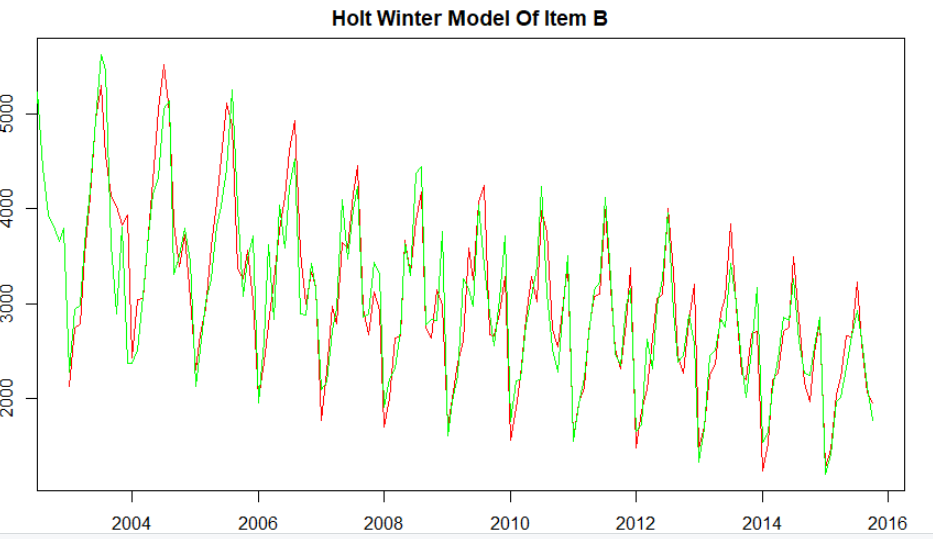
**Item A:**

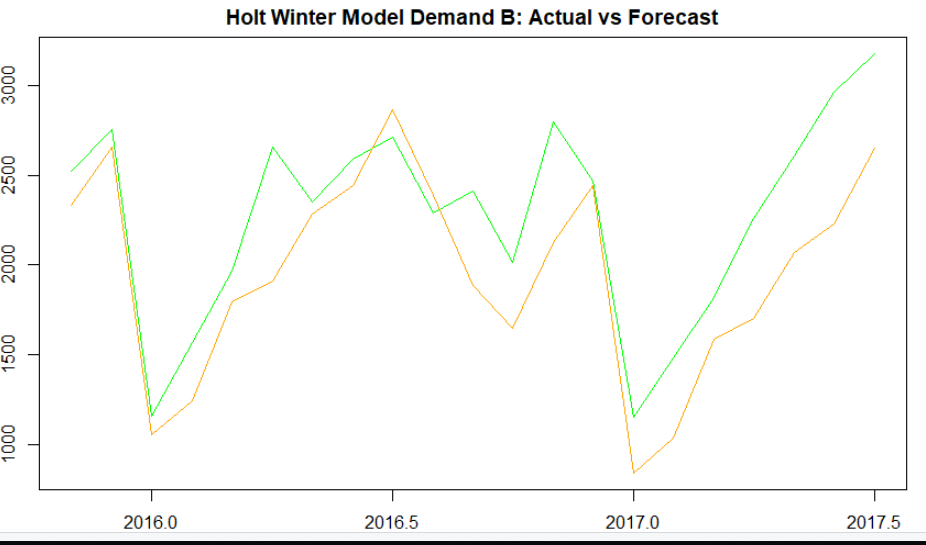


The p-value is greater than the 0.05 hence accepting the **Null hypothesis** which means the residuals are independent.

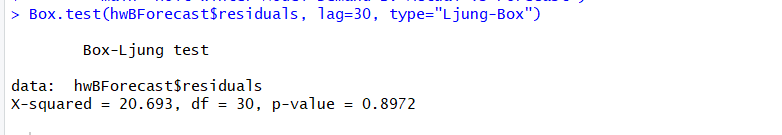
**Item B:**







**Item B:**



The p-value is greater than the 0.05 hence accepting the **Null hypothesis** which means the residuals are independent.

The MAPE results from holt’s winter model we observed that **12.6%** and **19.3%** less accuracy in model.

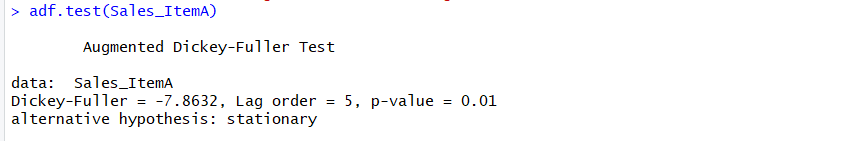
**ARIMA Model:**

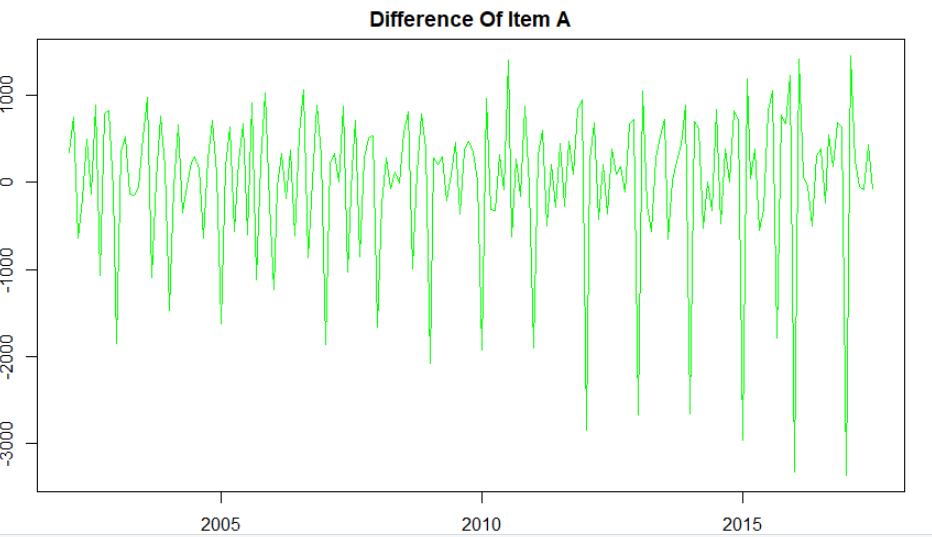
**Step 1:** To apply ARIMA model need to check the stationary of time series, If Stationarity condition is violated, the first step is to stationarize the series.

**Dickey-Fuller test:** Tocheck and confirmatory evidence that your time series is stationary or non-stationary.

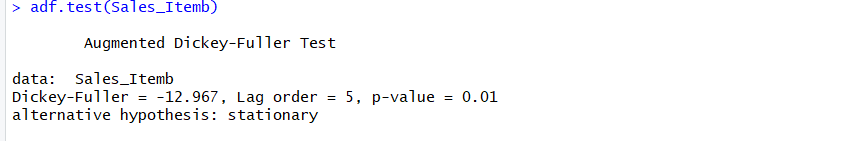
**Null Hypothesis (H0):** If accepted, Time series is non-stationary. It has some time dependent structure.  
**Alternate Hypothesis (H1):** The null hypothesis is rejected; it suggests the time series is stationary. It does not have time-dependent structure.

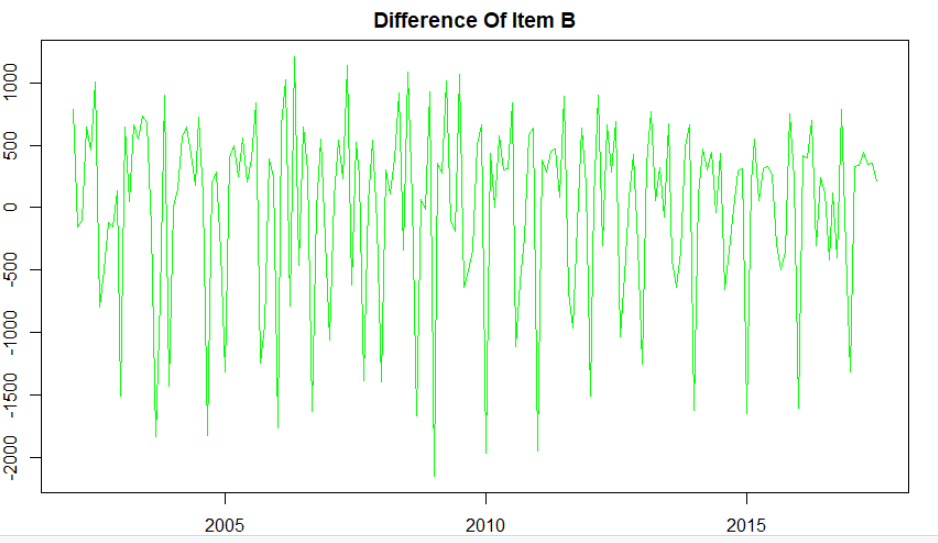
**Item A:** P-value is less than the 0.05 hence rejecting the null hypothesis which means it’s stationary





**Item B:** P-value is less than the 0.05 hence rejecting the null hypothesis which means it’s stationary



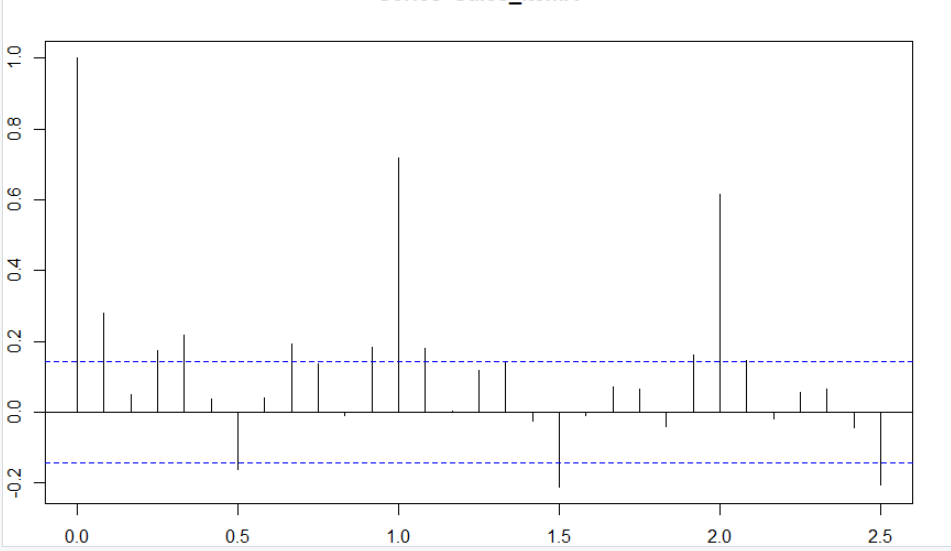


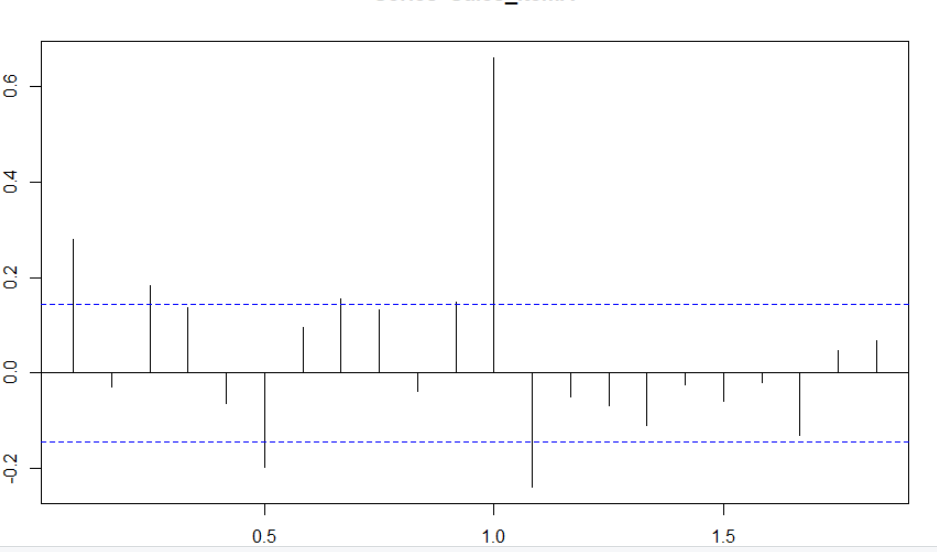
From the above test we have identified that both items(A&B) have the p-value of less than the 0.05 hence both are stationary time series.

**Step 2:** **Explore Autocorrelations and Partial Autocorrelations:**

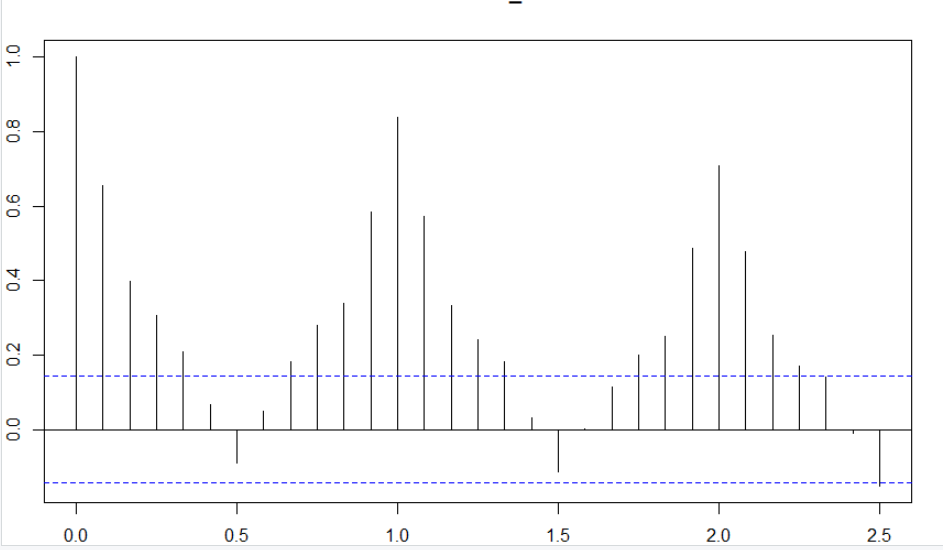
The function ACF and PACF used to check the auto correlation and partial auto correlations.

**Item A:**

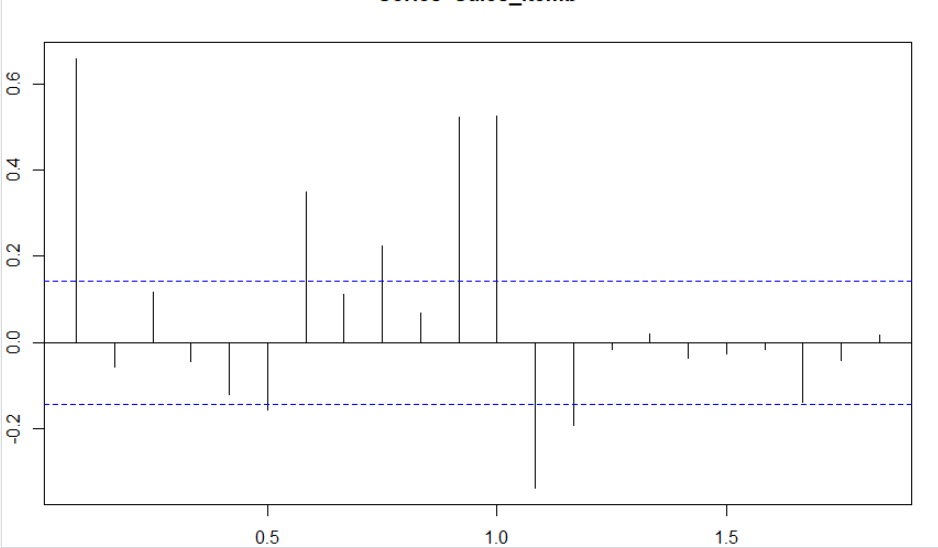
**ACF: With difference of 30 points.**

**PACF**

**Item B:**

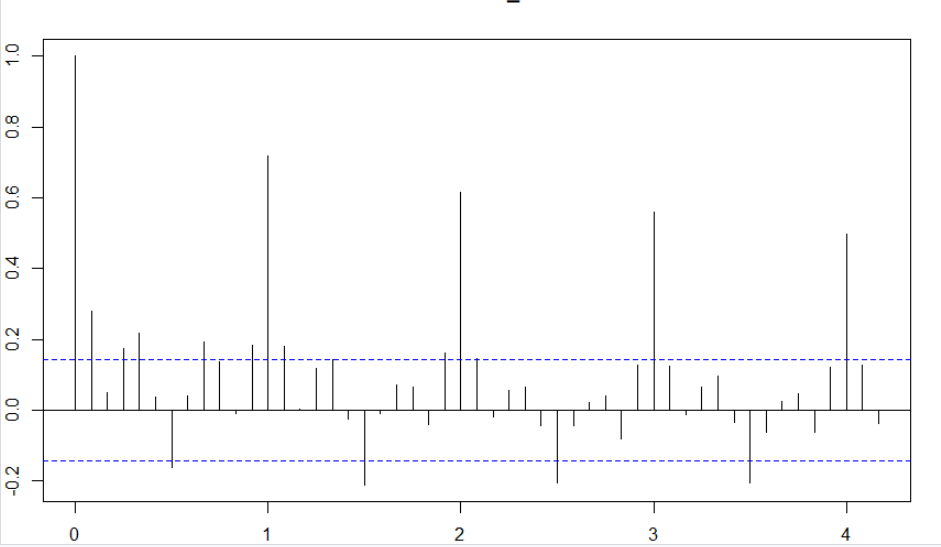
**ACF:** **With difference of 30 points** 

**PCF:**

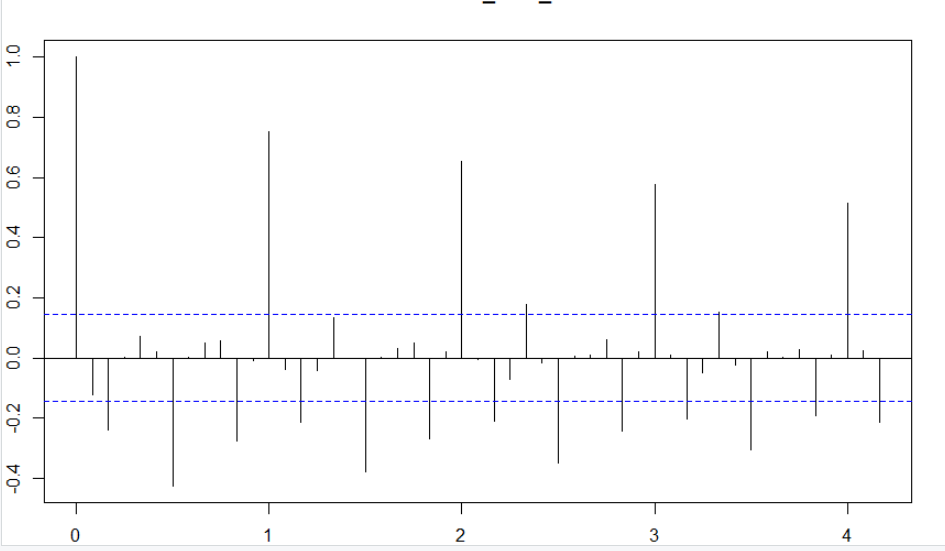


Now checking with difference of 50points.

**Item A:**



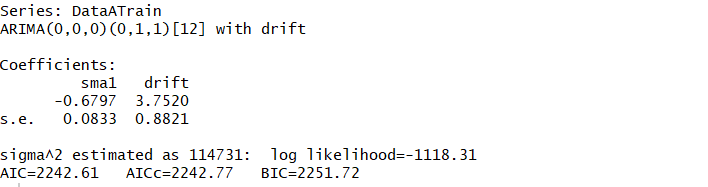
**Item B:**

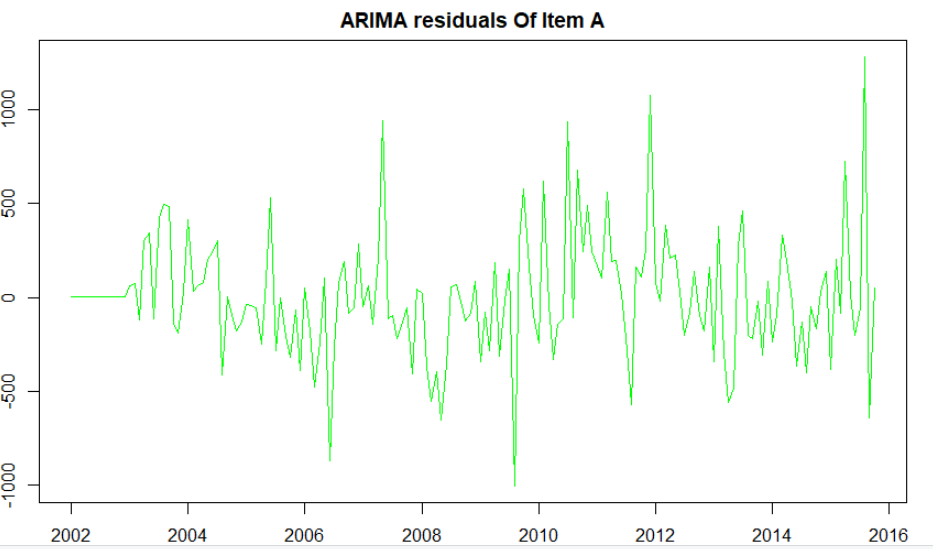


The above ACF and PACF we have found out that the positive and negative values mean this is due to data is stationary.

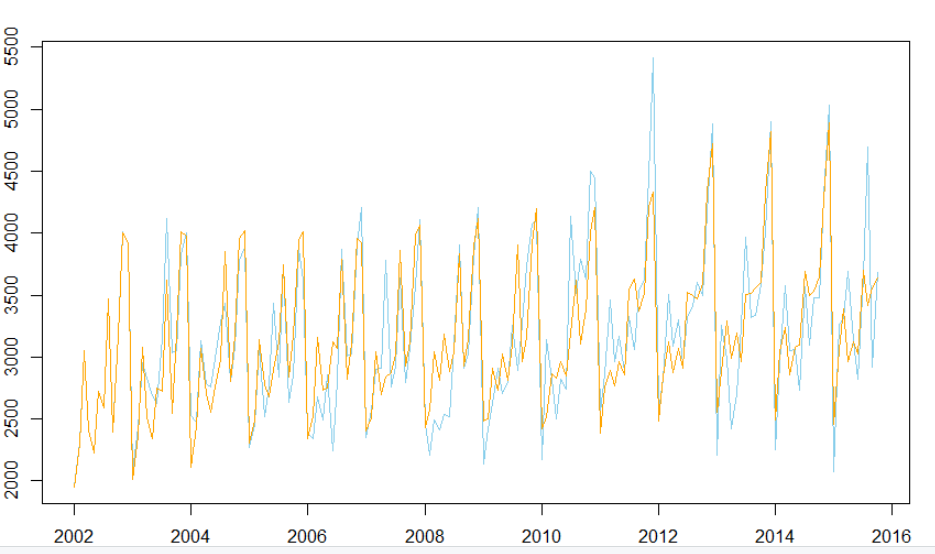
**Step 3: Building ARIMA Model**

**Item A:**



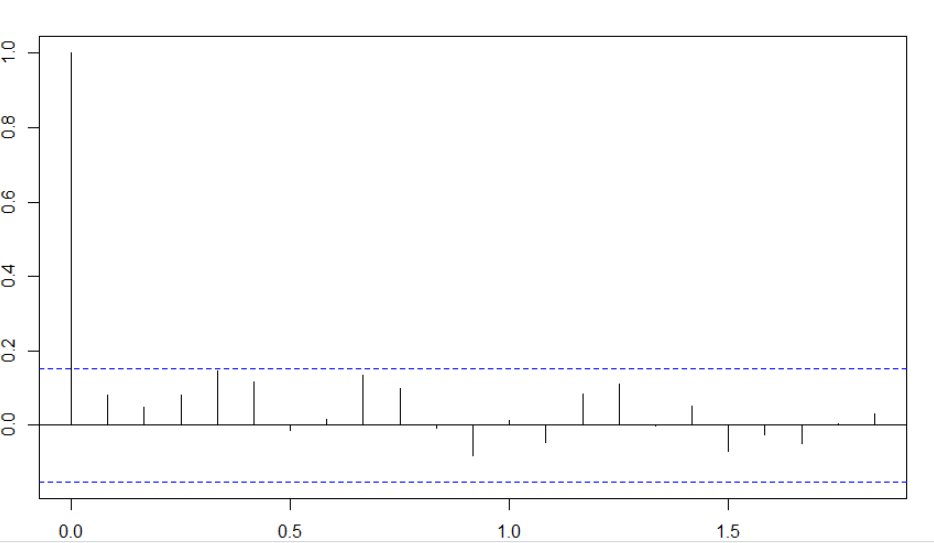
**Plotting the residuals:** 

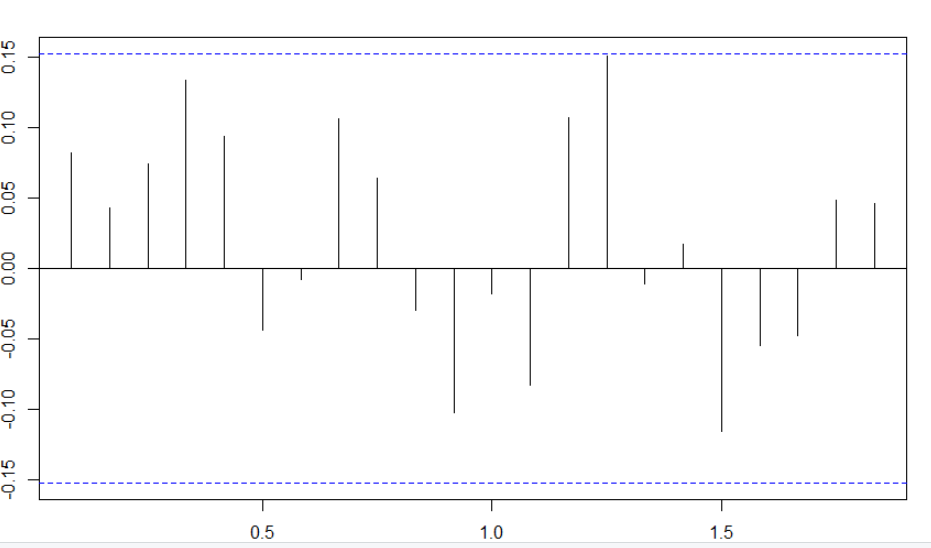
**Item A: Demand of Actual values Vs forecast values:**



**The MAPE value 7.33% from the ARIMA model which is less than the above two models.**

**ACF:**

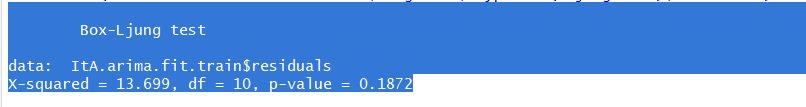


**PACF:** 

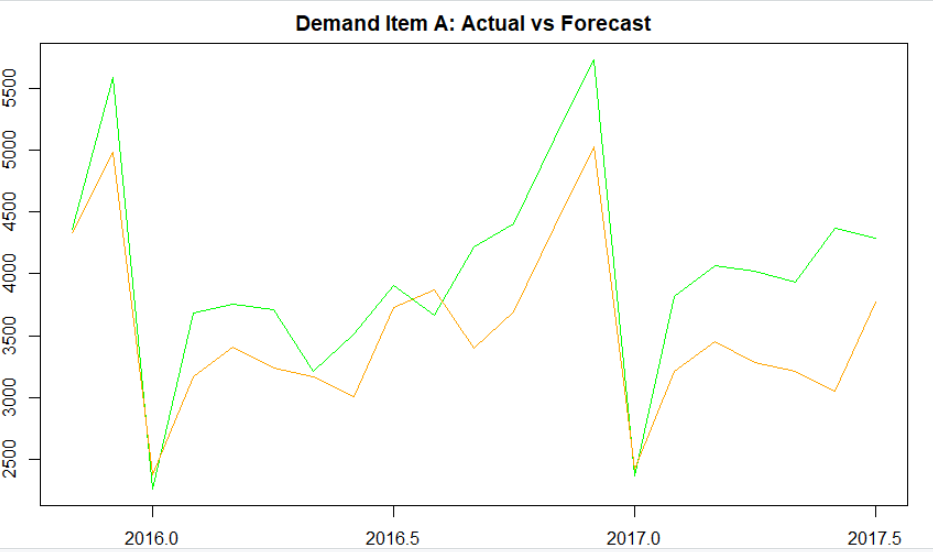
Box-Ljung Test:To check the random/residual components are independent or not.

H0: residuals are independent

HA: residuals are not independent.

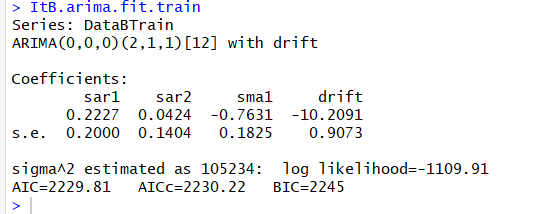


The p-value is greater than the 0.05 hence accepting the **Null hypothesis** which means the residuals are independent.

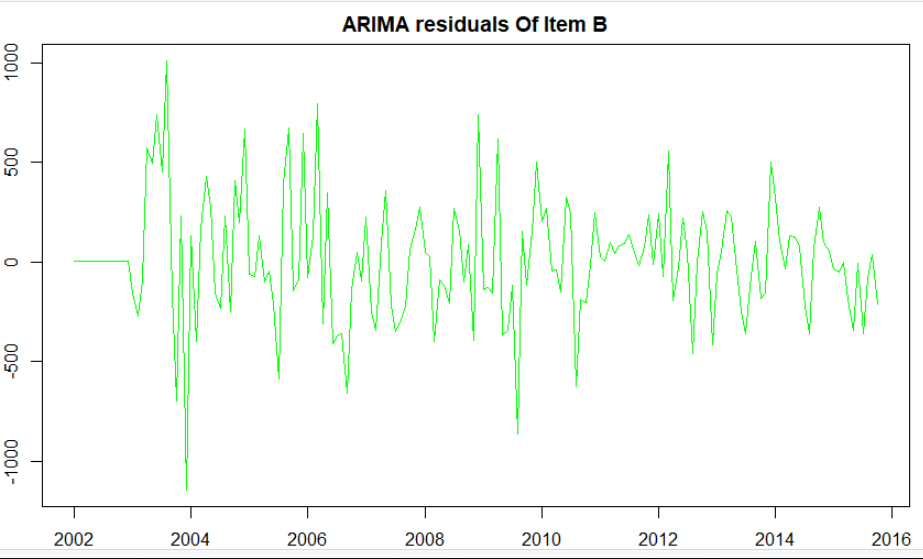


From the plotted data, we can see the forecasted value follows almost the same as actual value, there are point of interaction at Dec 2015, Jul 2016, Dec 2016, Jan 2017.

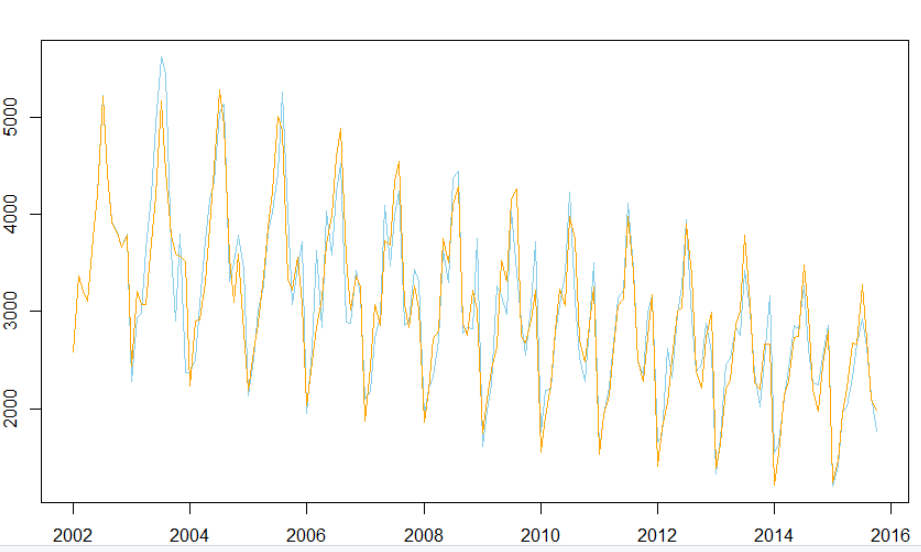
**Item B:**



**Plotting the residuals:**

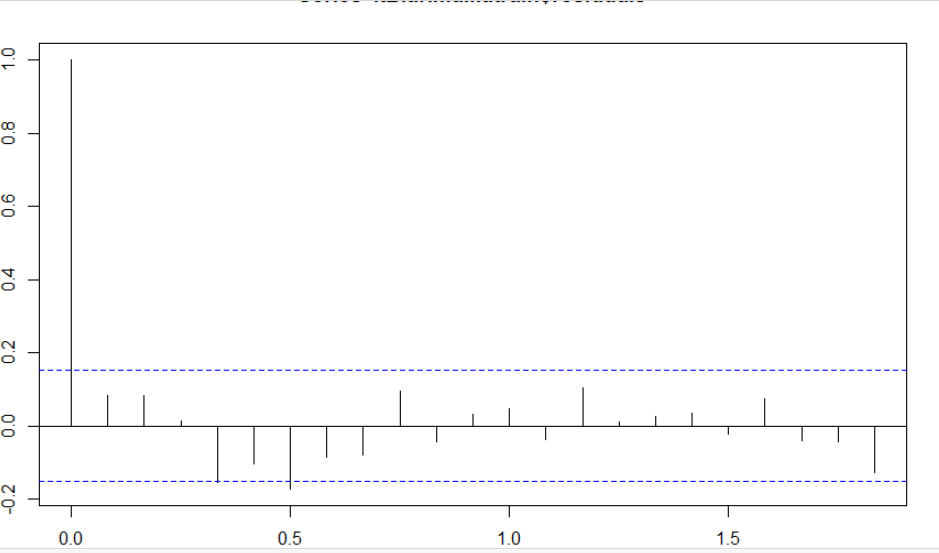


**Item B: Demand of Actual values Vs forecast values:**

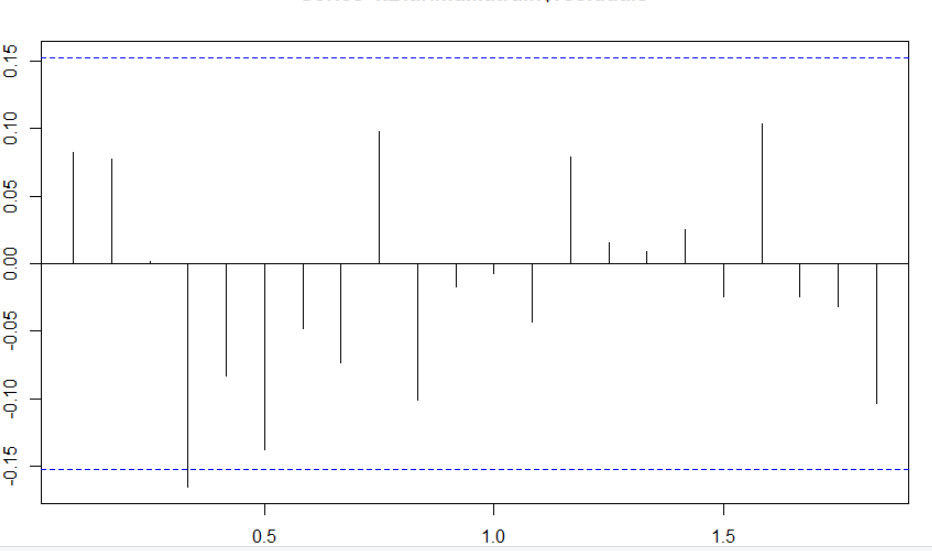


**The MAPE value 7.34% for the item B from the ARIMA model which is less than the above two models.**

**ACF Item B:**



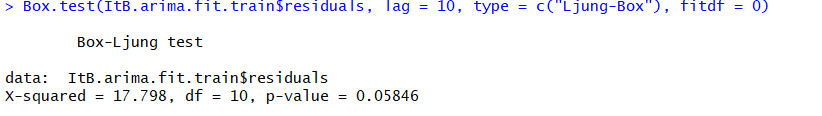
**PACF Item B:**



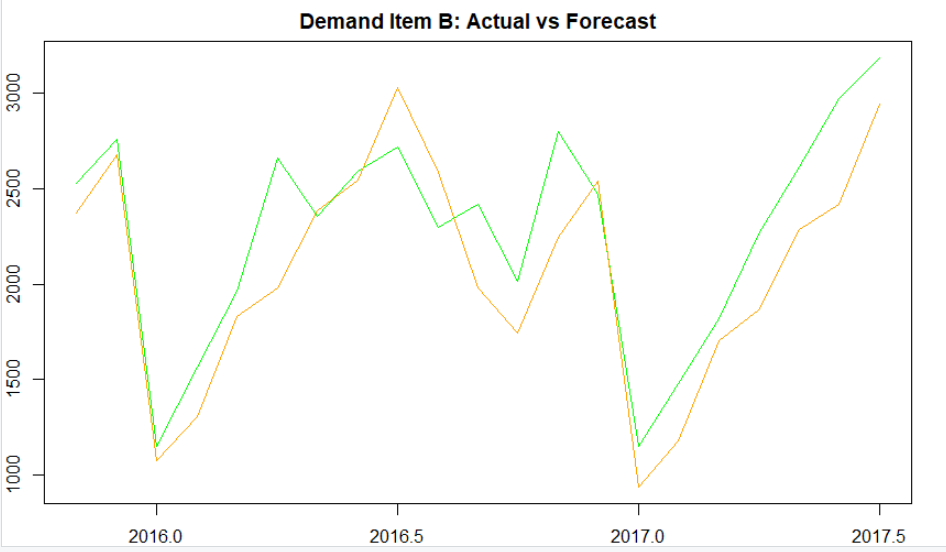
Box-Ljung Test:To check the random/residual components are independent or not.

H0: residuals are independent

HA: residuals are not independent.



The p-value is greater than the 0.05 hence accepting the **Null hypothesis** which means the residuals are independent.



From the plotted data, we can see the forecasted value follows almost the same as actual value, there are point of interaction at Apr 2015, Dec 2015, May 2016, July 2016, Nov 2016.

**5)Conclusion:**

For given problem, we observed the trend and seasonality in the data.  
and also observed that the Item A has increasing trend, but for Item B the trend is declining.  
for both items there are few months with high variation in seasonality.

We have used the “Additive” seasonality and performed the below three models.

* Random Walk with Drift
* Holt Winters
* ARIMA model.

**MAPE, Box-Ljung and α, β and γ values:**

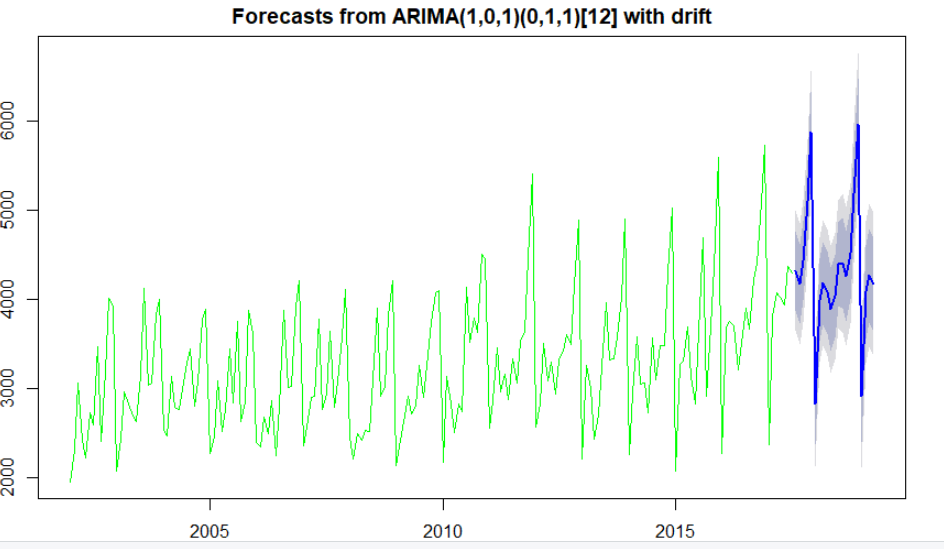
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Item A** | | **Item B** | |
| **MAPE** | **P-Value** | **MAPE** | **P-Value** |
| **Random Walk with Drift** | 13.1 | 2.2e^-16 | 19.8 | 6.16e^-13 |
| **Holt Winters** | 12.6 | 0.95 | 19.3 | 0.89 |
| **ARIMA** | 7.33 | 0.18 | 7.34 | 0.058 |

From the above table of MAPE values it has been observed that ARIMA model provided the lowest values and hence we selected for forecasting. The **α, β and γ values from the holt winter test are mentioned blow.**

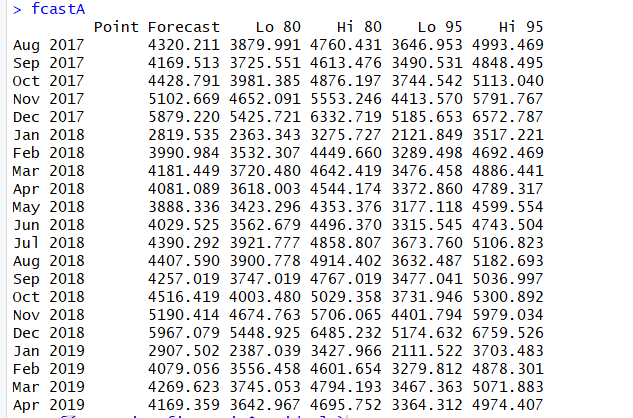
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Alpha** | **Beta** | **Gama** |
| **Item A** | 0.12 | 0.03 | 0.33 |
| **Item B** | 0.018 | 0.44 | 0.49 |

Forecasting the values for Item A and Item B as mentioned below from the period of Aug 2017 to Apr 2019.

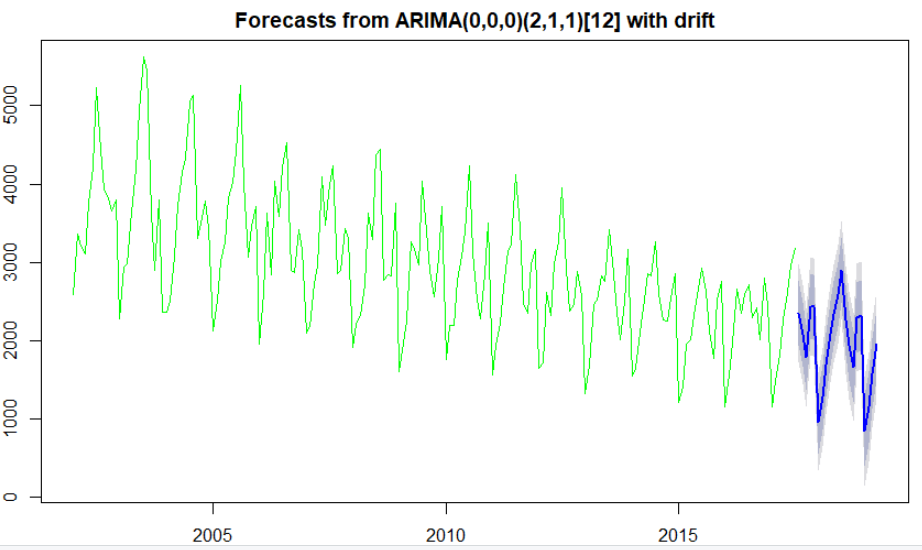
**Demand for Item: A forecasting Values:**



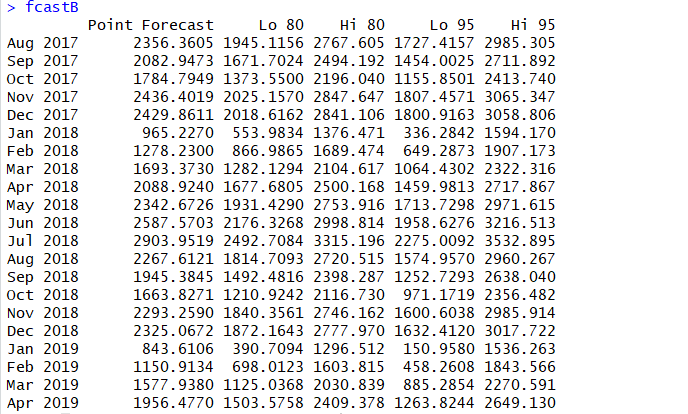
**Forecasted Values:**



**Demand for Item: B forecasting Values:**



**Forecasting Values:**



As a sales manner the demand for the future of Item-A is more hence we will look for the more availability of Item A then the Item B. We will also perform the market basket analysis on item B and look for any correlation with any other then we will club those two items as there is no demand of item B.

===============================**Appendicies**=====================================

library(forecast)

library(lubridate)

library(fpp2)

library(tseries)

library(MLmetrics)

setwd("E:/r direct/TimeSeries/Assignment")

Sales <- read.csv("Demand-1.csv",header = TRUE )

str(Sales)

dim(Sales)

summary(Sales)

Sales\_ItemA <- ts(Sales[,3], start=c(2002,1),end=c(2017,7),frequency=12)

plot(Sales\_ItemA)

plot(Sales\_ItemA, col = "Green", main = "Montly Sales Of Item A")

Sales\_Itemb <- ts(Sales[,4], start=c(2002,1),end=c(2017,7),frequency=12)

plot(Sales\_Itemb, col = "Green", main = "Montly Sales Of Item B")

ts.plot(Sales\_ItemA, Sales\_Itemb, gpars = list(col = c("Skyblue", "Orange")),

xlab="year", ylab="Sales Demand",main = "Montly Sales Of Item A & B")

legend("topleft", colnames(Sales[3:4]), col=1:ncol(Sales), lty=1.9, cex=.45)

monthplot(Sales\_ItemA,col='Green', main = "Montly Plot Of Item A")

monthplot(Sales\_Itemb,col='Green', main = "Montly Plot Of Item B")

Itema\_Sea<-stl(Sales\_ItemA, s.window='p')

plot(Itema\_Sea,col = "Green",main = "Decompostion Of Time Series Components for Item A")

Itemb\_Sea<-stl(Sales\_Itemb, s.window='p')

plot(Itemb\_Sea,col = "Green",main = "Decompostion Of Time Series Components for Item B")

Deseason\_ItemA <- (Itema\_Sea$time.series[,2]+Itema\_Sea$time.series[,3])

ts.plot(Sales\_ItemA, Deseason\_ItemA, col=c("Skyblue", "Orange"), main="ItemA Demand vs Deseasoned Demand")

Deseason\_ItemB <- (Itemb\_Sea$time.series[,2]+Itema\_Sea$time.series[,3])

ts.plot(Sales\_Itemb, Deseason\_ItemB, col=c("Skyblue", "Orange"), main="ItemA Demand vs Deseasoned Demand")

DataATrain <- window(Sales\_ItemA, start=c(2002,1), end=c(2015,10), frequency=12)

DataATest <- window(Sales\_ItemA, start=c(2015,11), frequency=12)

DataBTrain <- window(Sales\_Itemb, start=c(2002,1), end=c(2015,10), frequency=12)

DataBTest <- window(Sales\_Itemb, start=c(2015,11), frequency=12)

ItemATrn <- stl(DataATrain, s.window="p")

ItemBTrn <- stl(DataBTrain, s.window="p")

fcst.IteA.stl <- forecast(ItemATrn, method="rwdrift", h=21)

fcst.IteB.stl <- forecast(ItemBTrn, method="rwdrift", h=21)

ItemA<- cbind(DataATest,fcst.IteA.stl$mean)

ItemB<- cbind(DataBTest,fcst.IteB.stl$mean)

ts.plot(ItemA, col=c("blue", "green"),xlab="year", ylab="demand",

main="Quarterly Demand Of Item A: Actual vs Forecast")

MAPE\_ItemA <- mean(abs(ItemA[,1]-ItemA[,2])/ItemA[,1])

MAPE\_ItemA

ts.plot(ItemB, col=c("blue", "green"),xlab="year", ylab="demand",

main="Quarterly Demand Of Item B: Actual vs Forecast")

MAPE\_ItemB <- mean(abs(ItemB[,1]-ItemB[,2])/ItemB[,1])

MAPE\_ItemB

Box.test(fcst.IteA.stl$residuals, lag=30, type="Ljung-Box")

Box.test(fcst.IteB.stl$residuals, lag=30, type="Ljung-Box")

hawA <- HoltWinters(as.ts(DataATrain),seasonal="additive")

hawA

plot(hawA,col = "Green",main = "Holt Winter Model Of Item A")

hwAForecast <- forecast(hawA, h=21)

Item1 <- cbind(DataATest,hwAForecast)

par(mfrow=c(1,1), mar=c(2, 2, 2, 2), mgp=c(3, 1, 0), las=0)

ts.plot(Item1[,1],Item1[,2], col=c("Green","orange"),xlab="year", ylab="demand",

main="Holt Winter Model Demand A: Actual vs Forecast")

Box.test(hwAForecast$residuals, lag=30, type="Ljung-Box")

MAPE(Item1[,1],Item1[,2])

hawB <- HoltWinters(as.ts(DataBTrain),seasonal="additive")

hawB

plot(hawB,col = "Green",main = "Holt Winter Model Of Item B")

hwBForecast <- forecast(hawB, h=21)

Item2 <- cbind(DataBTest,hwBForecast)

ts.plot(Item2[,1],Item2[,2], col=c("Green","orange"),xlab="year", ylab="demand",

main="Holt Winter Model Demand B: Actual vs Forecast")

Box.test(hwBForecast$residuals, lag=30, type="Ljung-Box")

MAPE(Item2[,1],Item2[,2])

adf.test(Sales\_ItemA)

diff\_dem\_ItA <- diff(Sales\_ItemA)

plot(diff\_dem\_ItA,col = "Green",main = "Difference Of Item A")

adf.test(diff(Sales\_ItemA))

adf.test(Sales\_Itemb)

diff\_dem\_ItB <- diff(Sales\_Itemb)

plot(diff\_dem\_ItB,col = "Green",main = "Difference Of Item B")

adf.test(diff(Sales\_Itemb))

acf(Sales\_ItemA,lag=30)

pacf(Sales\_ItemA)

acf(diff\_dem\_ItA,lag=30)

pacf(diff\_dem\_ItA)

acf(Sales\_ItemA,lag=50)

acf(Sales\_Itemb,lag=30)

pacf(Sales\_Itemb)

acf(diff\_dem\_ItB,lag=30)

pacf(diff\_dem\_ItB)

acf(diff\_dem\_ItB,lag=50)

ItA.arima.fit.train <- auto.arima(DataATrain, seasonal=TRUE)

ItA.arima.fit.train

plot(ItA.arima.fit.train$residuals,col = "Green",main = "ARIMA residuals Of Item A")

plot(ItA.arima.fit.train$x,col="Skyblue")

lines(ItA.arima.fit.train$fitted,col="Orange",main="Demand A: Actual vs Forecast")

MAPE(ItA.arima.fit.train$fitted,ItA.arima.fit.train$x)

acf(ItA.arima.fit.train$residuals)

pacf(ItA.arima.fit.train$residuals)

Box.test(ItA.arima.fit.train$residuals, lag = 10, type = c("Ljung-Box"), fitdf = 0)

ArimafcastA <- forecast(ItA.arima.fit.train, h=21)

ItemA2 <- cbind(DataATest,ArimafcastA)

par(mfrow=c(1,1), mar=c(2, 2, 2, 2), mgp=c(3, 1, 0), las=0)

ts.plot(ItemA2[,1],ItemA2[,2], col=c("Green","Orange"),xlab="year",

ylab="demand", main="Demand Item A: Actual vs Forecast")

ItB.arima.fit.train <- auto.arima(DataBTrain, seasonal=TRUE)

ItB.arima.fit.train

plot(ItB.arima.fit.train$residuals,col = "Green",main = "ARIMA residuals Of Item B")

plot(ItB.arima.fit.train$x,col="Skyblue")

lines(ItB.arima.fit.train$fitted,col="Orange",main="Demand Of Item B: Actual vs Forecast")

MAPE(ItB.arima.fit.train$fitted,ItB.arima.fit.train$x)

acf(ItB.arima.fit.train$residuals)

pacf(ItB.arima.fit.train$residuals)

Box.test(ItB.arima.fit.train$residuals, lag = 10, type = c("Ljung-Box"), fitdf = 0)

ArimafcastB <- forecast(ItB.arima.fit.train, h=21)

ItemB2 <- cbind(DataBTest,ArimafcastB)

par(mfrow=c(1,1), mar=c(2, 2, 2, 2), mgp=c(3, 1, 0), las=0)

ts.plot(ItemB2[,1],ItemB2[,2], col=c("Green","Orange"),xlab="year",

ylab="demand", main="Demand Item B: Actual vs Forecast")

ItA.arima.fit <- auto.arima(Sales\_ItemA, seasonal=TRUE)

fcastA <- forecast(ItA.arima.fit, h=21)

plot(fcastA,col = "Green")

ItB.arima.fit <- auto.arima(Sales\_Itemb, seasonal=TRUE)

fcastB <- forecast(ItB.arima.fit, h=21)

plot(fcastB,col = "Green")