Time Series S05

Group 3

6/23/2020

### Import the necessary libraries

library(fpp2)

## Warning: package 'fpp2' was built under R version 3.6.3

## Loading required package: ggplot2

## Loading required package: forecast

## Warning: package 'forecast' was built under R version 3.6.3

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

## Loading required package: fma

## Warning: package 'fma' was built under R version 3.6.3

## Loading required package: expsmooth

## Warning: package 'expsmooth' was built under R version 3.6.3

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(imputeTS)

## Warning: package 'imputeTS' was built under R version 3.6.3

### Load the dataset

data\_project <- readxl::read\_excel("Data Set for Class.xls")  
head(data\_project)

## # A tibble: 6 x 7  
## SeriesInd group Var01 Var02 Var03 Var05 Var07  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 40669 S03 30.6 123432400 30.3 30.5 30.6  
## 2 40669 S02 10.3 60855800 10.0 10.2 10.3  
## 3 40669 S01 26.6 10369300 25.9 26.2 26.0  
## 4 40669 S06 27.5 39335700 26.8 27.0 27.3  
## 5 40669 S05 69.3 27809100 68.2 68.7 69.2  
## 6 40669 S04 17.2 16587400 16.9 16.9 17.1

### Subset the dataset

S05 <- subset(data\_project, group == 'S05', select = c(SeriesInd, Var02, Var03))  
summary(S05)

## SeriesInd Var02 Var03   
## Min. :40669 Min. : 4156600 Min. : 55.94   
## 1st Qu.:41304 1st Qu.: 11201200 1st Qu.: 77.21   
## Median :41946 Median : 14578800 Median : 84.87   
## Mean :41945 Mean : 16791350 Mean : 82.97   
## 3rd Qu.:42586 3rd Qu.: 20021200 3rd Qu.: 89.74   
## Max. :43221 Max. :118023500 Max. :103.95   
## NA's :141 NA's :145

#### Get the subsets Var02 and Var03

var02 <- S05 %>% filter(SeriesInd <= 43021) %>% select(Var02)

var03 <- S05 %>% filter(SeriesInd <= 43021) %>% select(Var03)

Explore the subsets Var02

summary(var02)

## Var02   
## Min. : 4156600   
## 1st Qu.: 11201200   
## Median : 14578800   
## Mean : 16791350   
## 3rd Qu.: 20021200   
## Max. :118023500   
## NA's :1

Var02 has 1 missing value

Explore the subsets Var03

summary(var03)

## Var03   
## Min. : 55.94   
## 1st Qu.: 77.21   
## Median : 84.87   
## Mean : 82.97   
## 3rd Qu.: 89.74   
## Max. :103.95   
## NA's :5

Var03 has 5 missing values Median and mean are in the same order. #### converse Var02 and Var03 to time series

var02 <- ts(var02)  
str(var02)

## Time-Series [1:1622, 1] from 1 to 1622: 27809100 30174700 35044700 27192100 24891800 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : NULL  
## ..$ : chr "Var02"

var03 <- ts(var03)  
str(var03)

## Time-Series [1:1622, 1] from 1 to 1622: 68.2 68.8 69.3 69.4 69.2 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : NULL  
## ..$ : chr "Var03"

#### Imputing missing values

var02 <- na\_interpolation(var02)  
summary(var02)

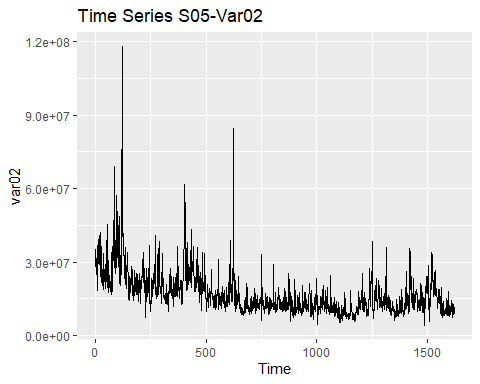
## Var02   
## Min. : 4156600   
## 1st Qu.: 11206550   
## Median : 14575700   
## Mean : 16788872   
## 3rd Qu.: 20014325   
## Max. :118023500

var03 <- na\_interpolation(var03)  
summary(var03)

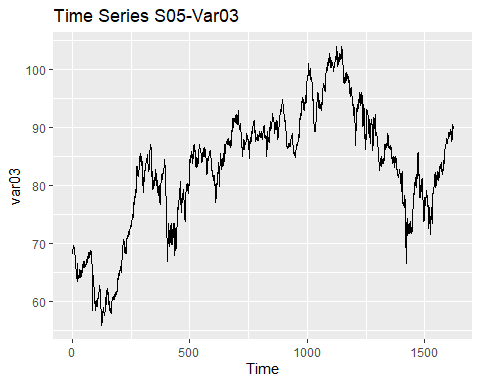
## Var03   
## Min. : 55.94   
## 1st Qu.: 77.25   
## Median : 84.88   
## Mean : 82.97   
## 3rd Qu.: 89.71   
## Max. :103.95

### Visualization

autoplot(var02) + ggtitle("Time Series S05-Var02")

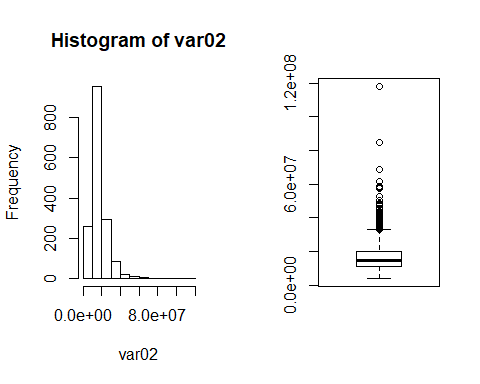
 Trend non seasonal time series

autoplot(var03) + ggtitle("Time Series S05-Var03")

 Trend non seasonal time serie. It can also be cyclic time series

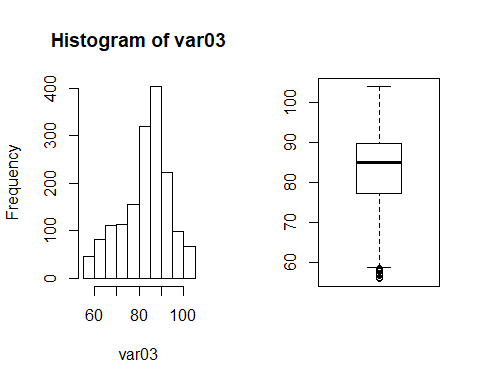
#### The distribution of data

par(mfrow= c(1,2))  
hist(var02)  
boxplot(var02)



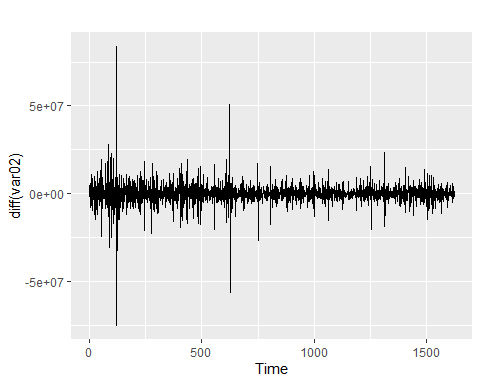
Va02 is right skewed. It need to be centralised.

par(mfrow= c(1,2))  
hist(var03)  
boxplot(var03)

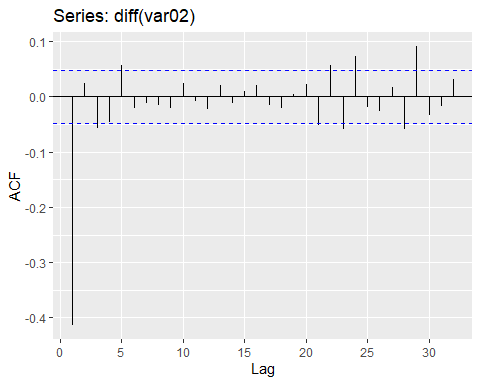
 Var03 is nearly normal distributed and has outliers at the left.

ACF of Var02 difference

par(mfrow=c(1,2))  
autoplot(diff(var02))

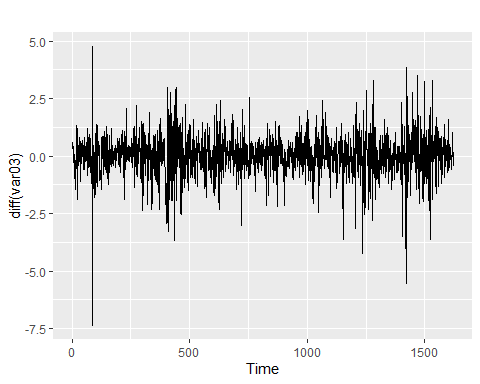


ggAcf(diff(var02))

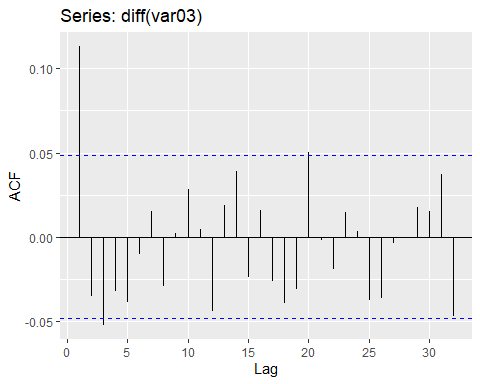
 There is one order correlation

ACF of Var03 difference

par(mfrow=c(1,2))  
autoplot(diff(var03))



ggAcf(diff(var03))

 The ACF shows one order correlation

#### Subset the train set

train.var02 <- window(var02, end = as.integer(length(var02)\*0.8))

train.var03 <- window(var03, end = as.integer(length(var03)\*0.8))

#### Look for lambda transformation

lambda2 <- BoxCox.lambda(var02)

lambda3 <- BoxCox.lambda(var03)

#### Apply the models to the train sets

forecast horizon here the length of test set

h <- length(var02)- as.integer(length(var03)\*0.8)

Get the arima model for Var02

fvar02 <- train.var02 %>% auto.arima(lambda = lambda2, stepwise = FALSE) %>% forecast(h = h)

Get the arima model for Var03

fvar03 <- train.var03 %>% auto.arima(lambda = lambda2, stepwise = FALSE) %>% forecast(h = h)

#### Naive model for Var02

naive.var02 <- naive(train.var02, h=h)

#### Naive model for Var03

naive.var03 <- naive(train.var03, h=h)

#### Accuracy compare to the naive model

RMSE accuracy of Var02 arima model

accuracy(fvar02, var02)["Test set", "RMSE"]

## [1] 5213174

RMSE accuracy of Var02 naive model

accuracy(naive.var02, var02)["Test set", "RMSE"]

## [1] 5289689

Comparing the RMSEs, the arima model for Var02 is better than the naive model.

RMSE accuracy for Var03 arima model

accuracy(fvar03, var03)["Test set", "RMSE"]

## [1] 8.585339

RMSE accuracy for Var03 naive model

accuracy(naive.var03, var03)["Test set", "RMSE"]

## [1] 8.564479

comparing with the RMSE, the naive method for Var03 is better than the arima of the same data. But the residuals of the naive method are not white noise. There still information to get from the naive residuals. We chose arima method instead.

Ljung-Box test of the naive residuals

Box.test(residuals(naive.var03))

##   
## Box-Pierce test  
##   
## data: residuals(naive.var03)  
## X-squared = 13.752, df = 1, p-value = 0.0002086

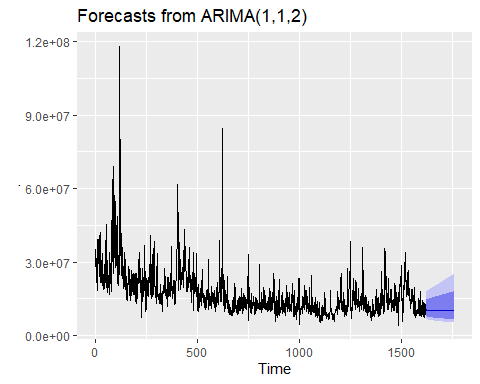
### Forecast the time series

farima.var02 <- var02 %>% auto.arima(lambda = lambda2, stepwise = FALSE) %>% forecast(h = 140)

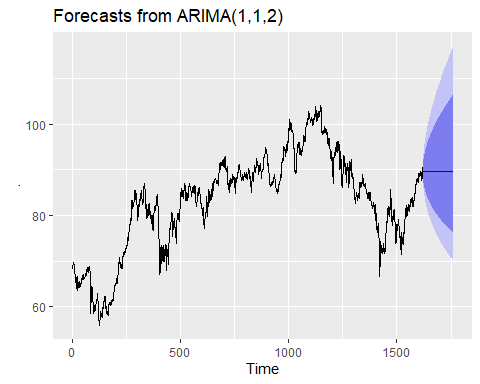
farima.var03 <- var03 %>% auto.arima(lambda = lambda2, stepwise = FALSE) %>% forecast(h = 140)

##### Var02 forecast

autoplot(farima.var02)

 ##### Var03 forecast

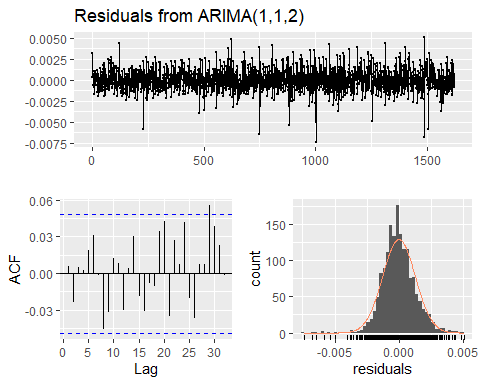
autoplot(farima.var03)



#### Check the residuals if the model is valid

##### Var02 residuals

checkresiduals(farima.var02)

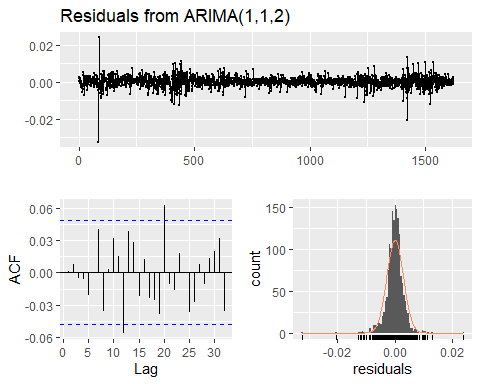


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,1,2)  
## Q\* = 8.5874, df = 7, p-value = 0.2836  
##   
## Model df: 3. Total lags used: 10

with p-value greater than 0.05, there is convaincing evidence that residuals for Var02 are white noise. On ACF, the residuals are uncorrelated. The histogram shows that the residuals are normal distributed.

##### Var03 residuals

checkresiduals(farima.var03)



##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,1,2)  
## Q\* = 7.4704, df = 7, p-value = 0.3816  
##   
## Model df: 3. Total lags used: 10

with p-value greater than 0.05, there is convaincing evidence that residuals for Var03 are white noise. On ACF, the residuals are uncorrelated. The histogram shows that the residuals are normal distributed.