Final_Project_Richard_Antony_PSTAT 174

Richard Antony

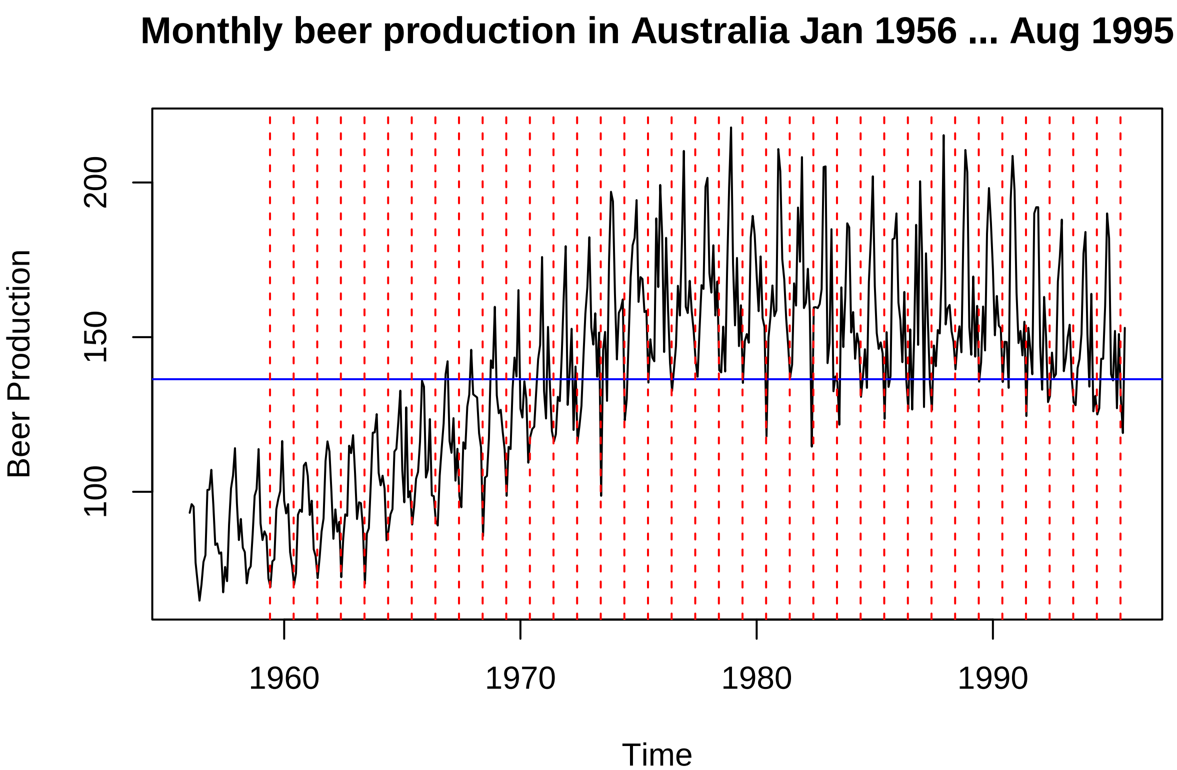
2022-12-10

##Abstract, Introduction, Main Body

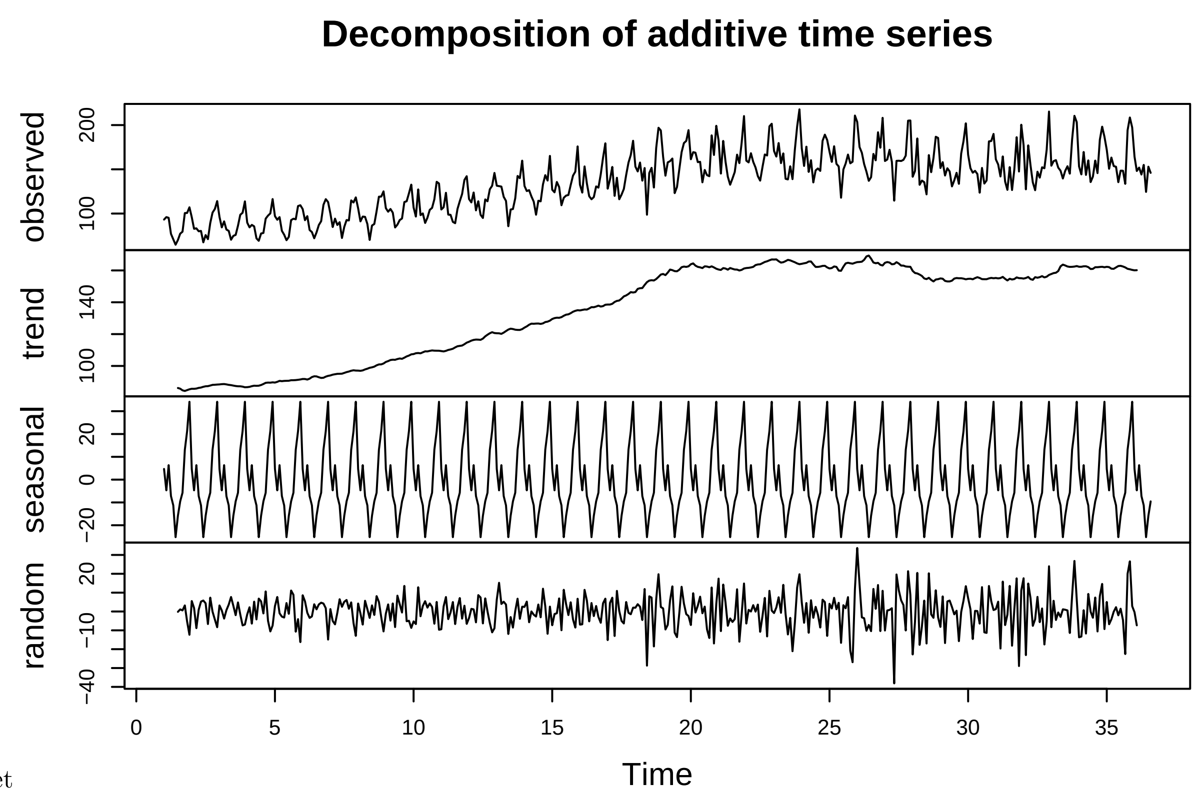
In this final project I will go over to try to forecast the data of Monthly beer production in Australia Jan 1956 – Aug 1995. I chose this data set because beer production really interest me. I also love to drink beer on the weekends so I wanted to know more about the production of beer. This data comes from the tsdl library. To achieve this, I split the data into two parts to a training set and a test set. Then I will perform time series analysis on the train set from looking at the acf pacf, differencing, transforming , looking at the AICc to choose the best model and Diagnostic checking of the model and forecast.

First I take 90% of the the first observation from the dataset to be my training set and last 10% observations to be my test set. I then transform my training set in hope that it can assume normality. I transform it with log and box cox. After interpreting the result from a histogram viewpoint, the box cox transformation did better. After transforming the data with box cox I then proceed to see the data PACF and ACF for model assumputions. The model I assume at first is SARIMA$(0,1,0) \times (1,1,0)_6$. I then compare the AICc of the model with a pure AR(1) and pure MA(1) model to see if my model really did better than a pure model. Turns out it did. After assuming the model to be SARIMA$(0,1,0) \times (1,1,0)_6$, I then proceed to do diagnostic check to improve and check asummptions for the model. From diagnostic checking and seeing the residuals PACF and ACF I ended with the model SARIMA$(8,1,1) \times (2,1,2)_6$. This model resiudals pass the Box pierce and Box ljung test but failed the Mcloid test.

After doing the diagnostic check , I did forecasting with the model to predict the test set (last 10% of the original data). It turn out pretty good as the prdict the observation quite close .



By looking at data plot, we see that it has a seasonal componet of quarterly, but this needs to be explore further down the project.



#Look on the train set

From the Decomposition of the Beer Data, we see that there is trend going up wards linearlly to approximiately 1970 and after that it stays

From the Decomposition of the Beer Data, we see that there is some seasonailty and from the plot itself, I see that there is a seasonal every 3 months (Quarterly). But this needs to be explore further. Differencing could eliminate these.

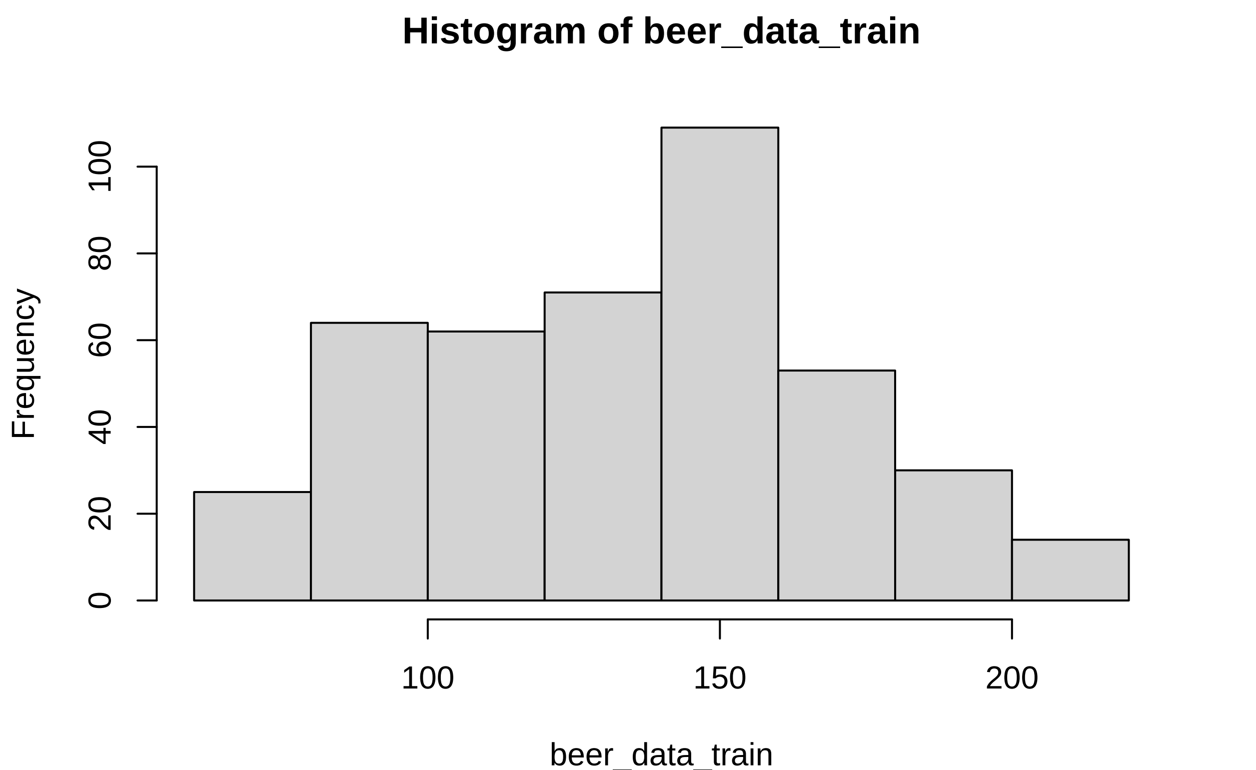
Check if the train data has constant variance and constant mean

```
## $stat
## [1] 230.177
##
## $sum
## [1] 252.8813
```

```
##          Part     Mean Variance
## 1  First Half 109.3481 622.3758
## 2 Second Half 160.5758 461.9746
```

From here, we saw that the beer data does not have constant mean and constant variance. I split into approximiately half of the data to see each of its mean and variance. I saw that the mean and variance in each splitted range is not relatively similar. To further prove if the data does not have constant variance, I use Automatic Variance Ratio Test in which the statistical value should be small if there is constant variance. But in here, the Automatic Variance Ratio shows a very large number.

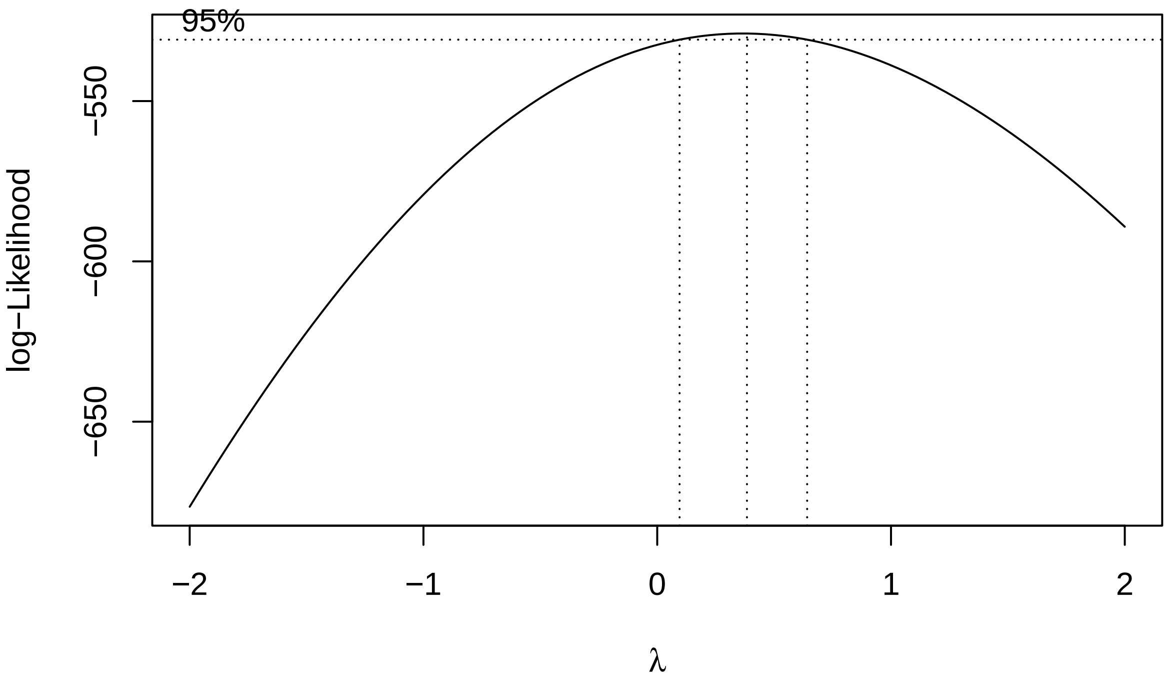
```
hist(beer_data_train)
```




```
#shapiro.test(beer_data_train)
```

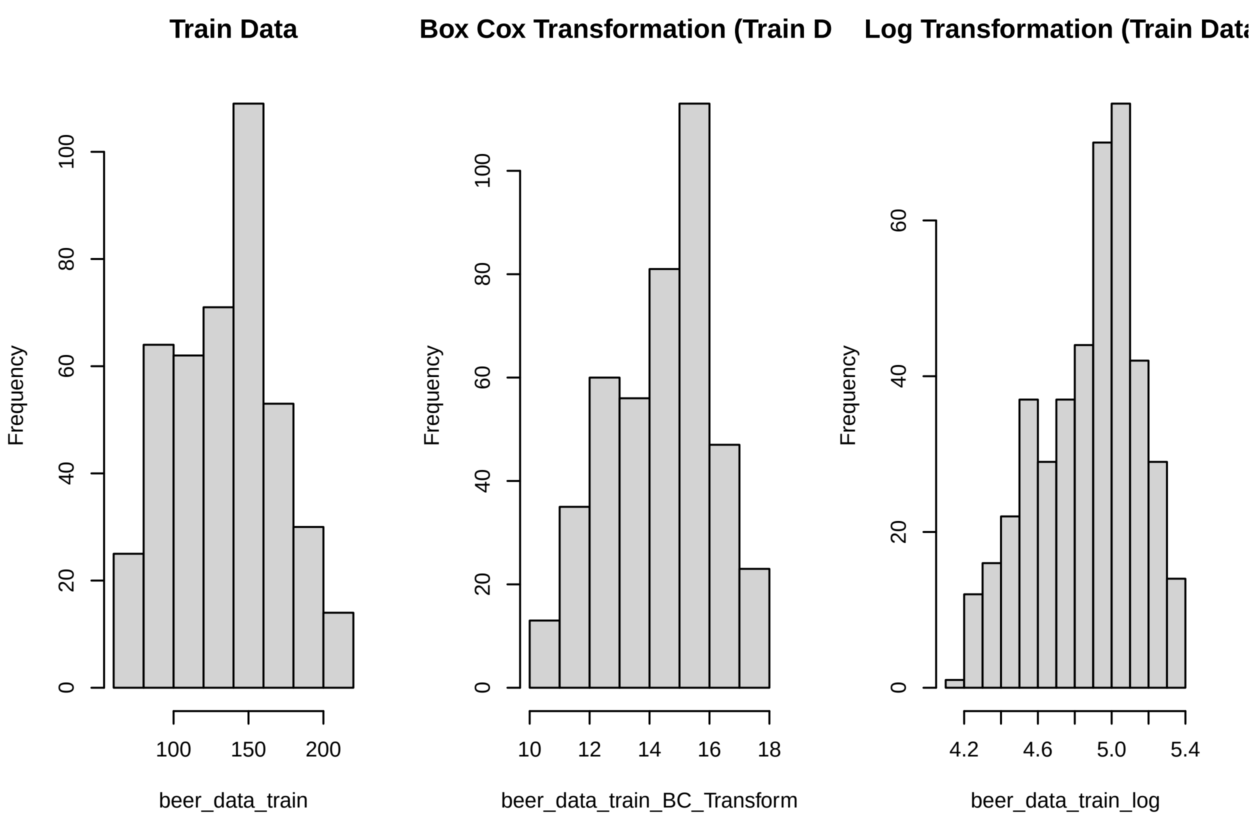
From the Histogram we see that the data is fairly symmetrical (Gaussian). But this needs to explore futher down the Project.

Transformation box cox and log



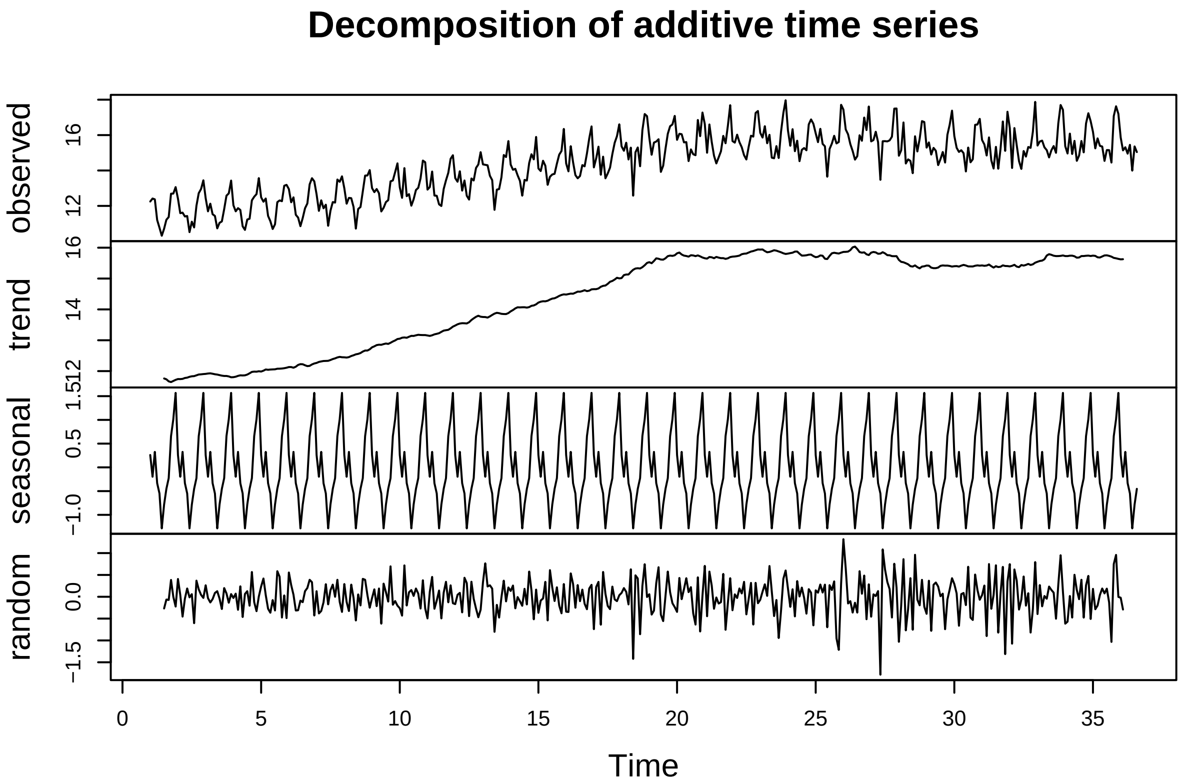
For the Box cox, I found out that $\lambda = 0.3838384$. So I use power transfromation

#Compare Histogram

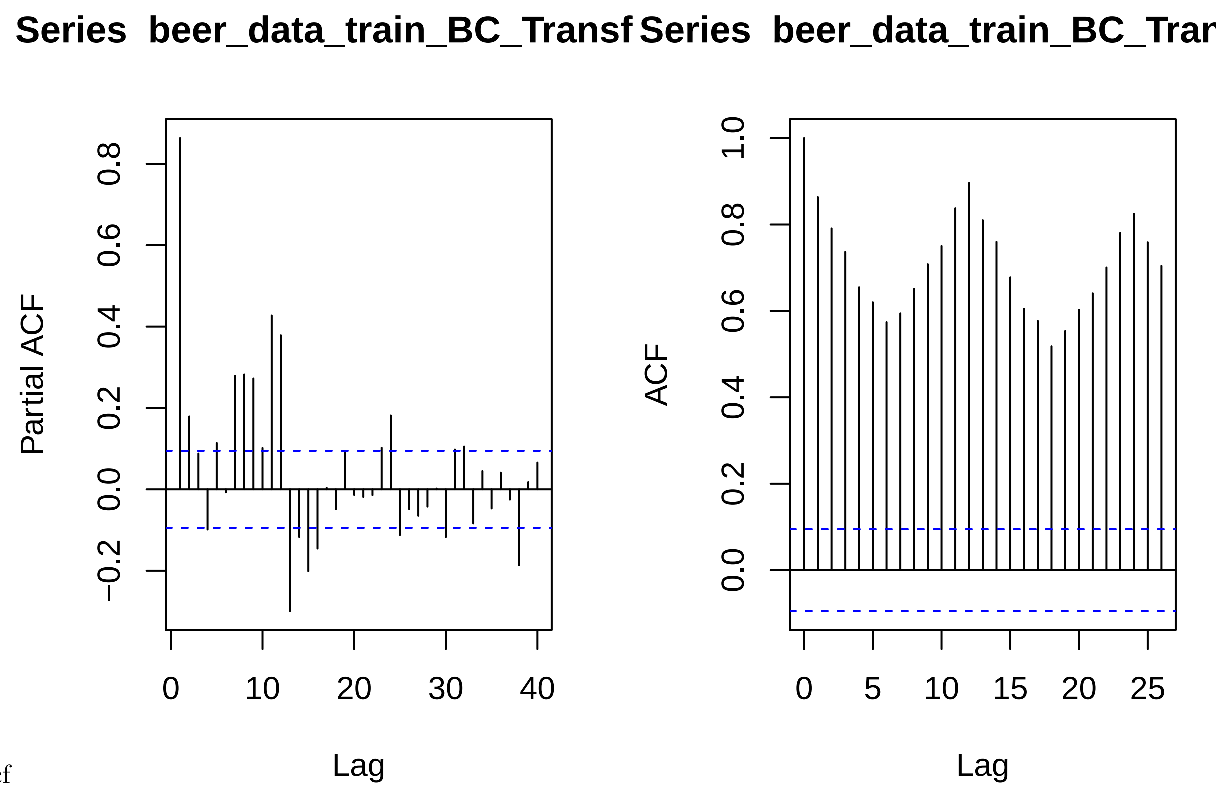


From comparing the Histogram, I choose the training data that is transformed by Box-Cox as it looks the most symetrical.

##Differencing to elimanite seasonality for the box cox transformation train data



We see that after doing the Box-Cox Transformation towards the training data, there is still seasonanility and trend. Trend seems increasing linearly to 1975 and stays relatively flats onwards. So we need to difference it. Its a pieceswise function.

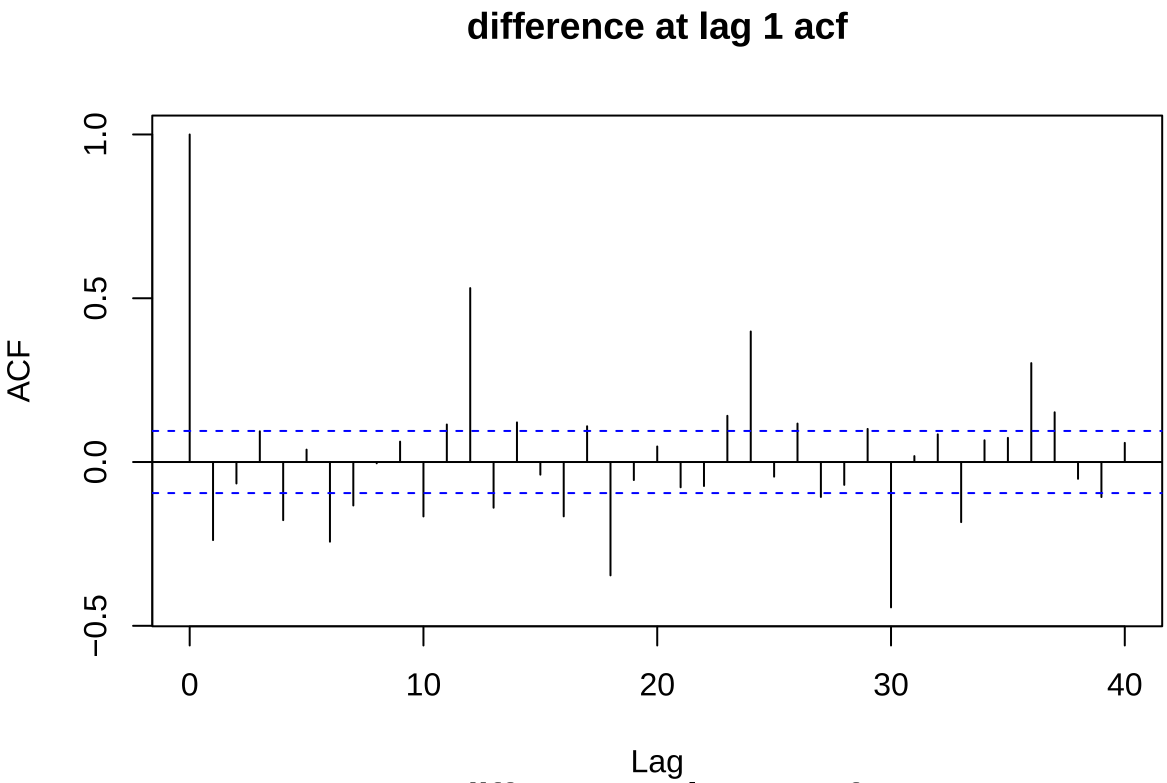


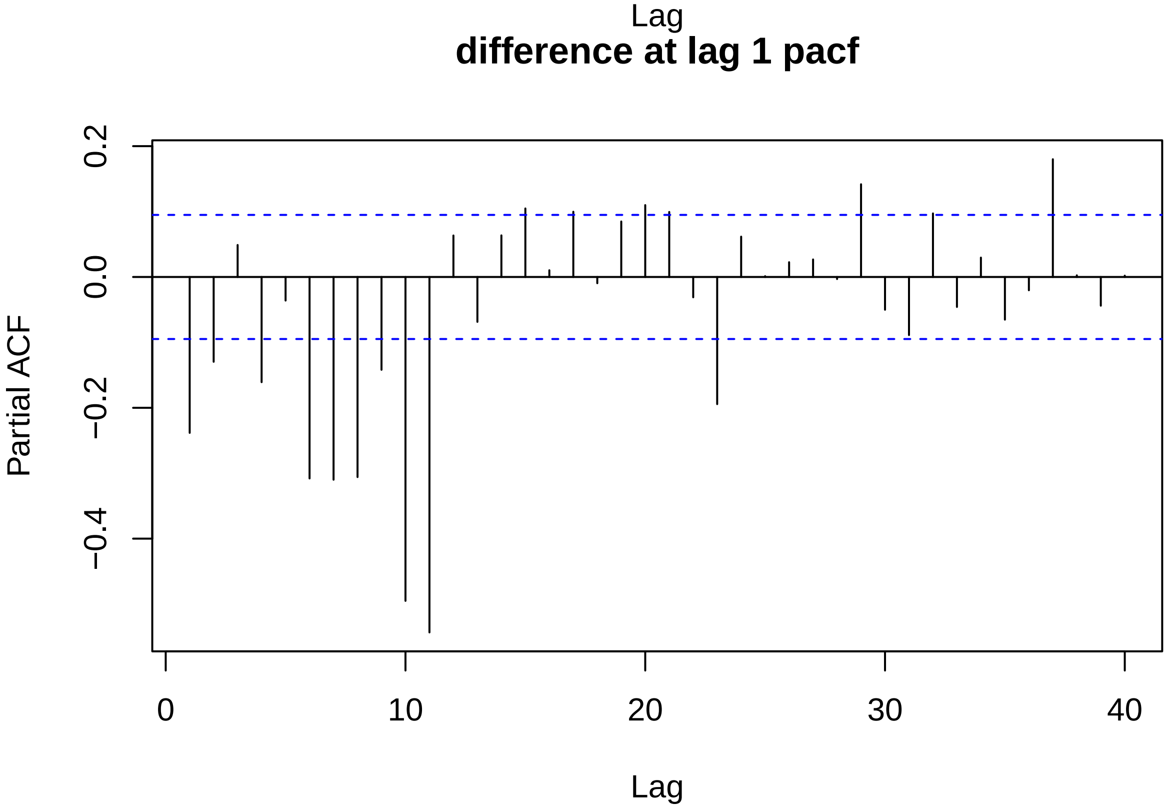
#Check acf and pacf

In this part I am comparing the PACF and ACF for both the train and the transformed train data. All the ACF lags are significant meaning that there is a trend on going which we need to difference at lag 1. I also see that from the PACF there is a seasonly of semi-annual with lag 6 , lag 12 being significant. So it needs to be difference at 6 to remove seasonality.

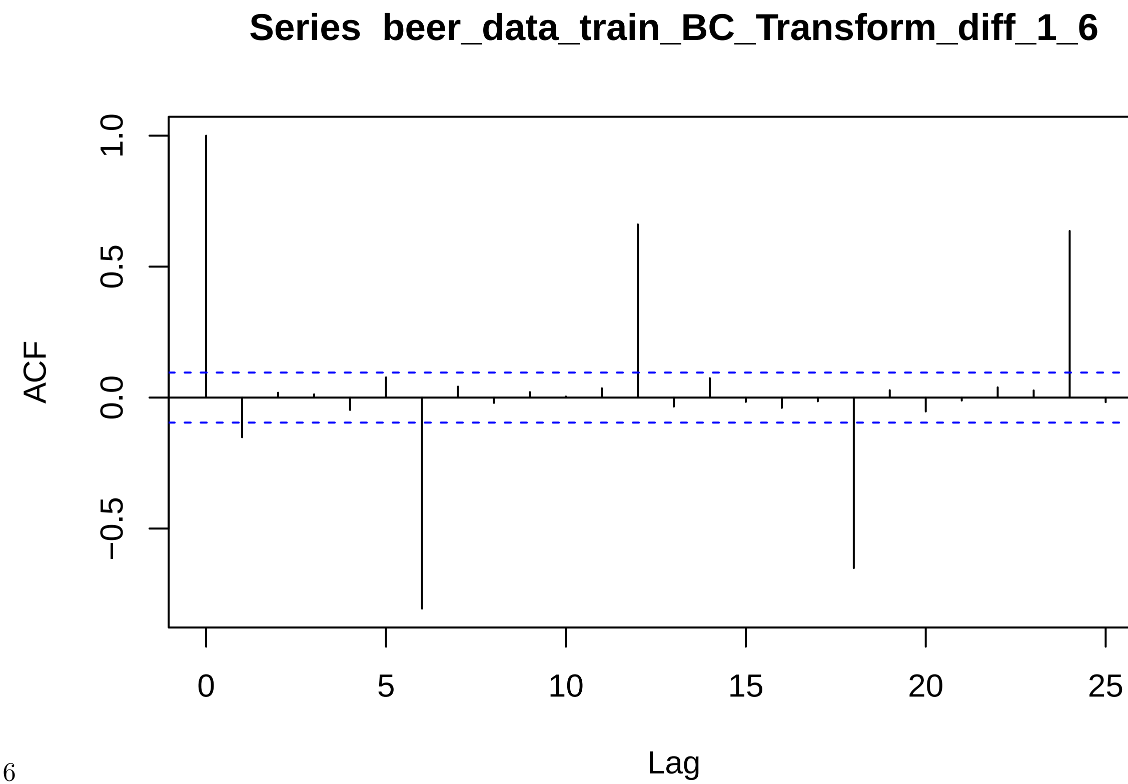
#Differencing and seeing their ACF & PACF

diff lag 1

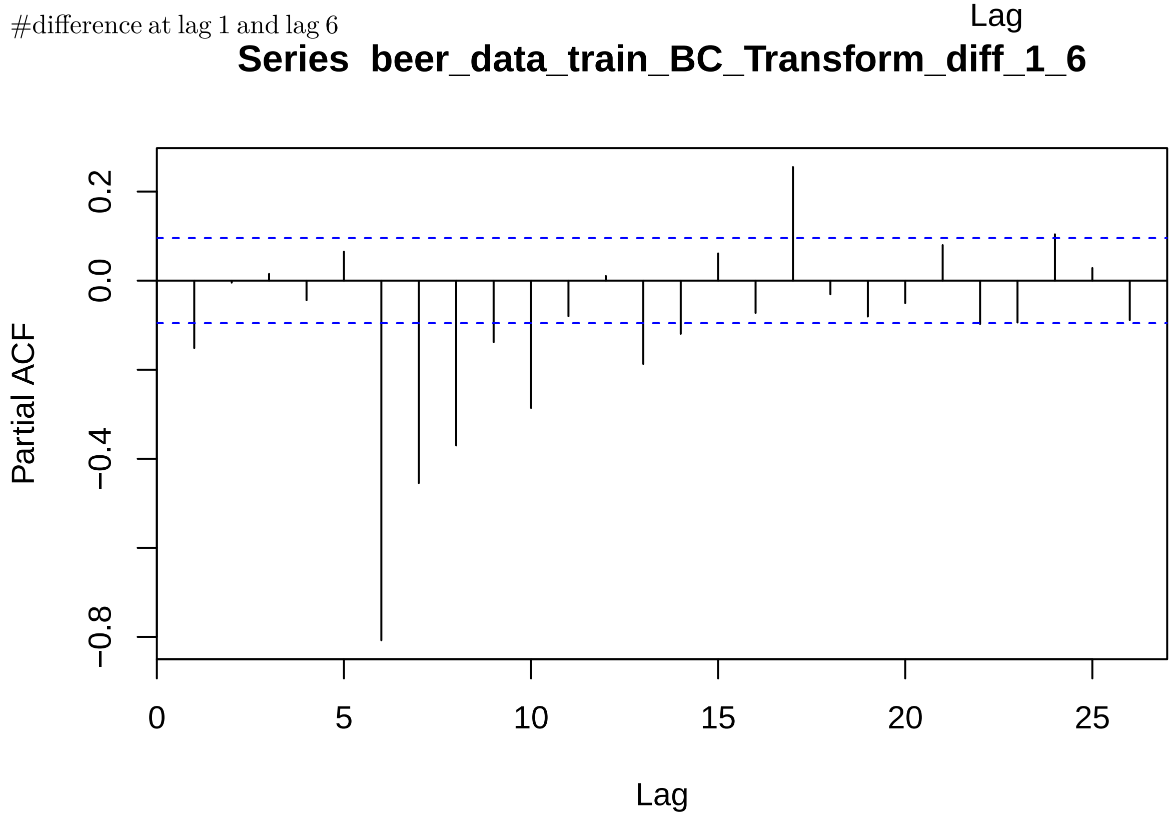


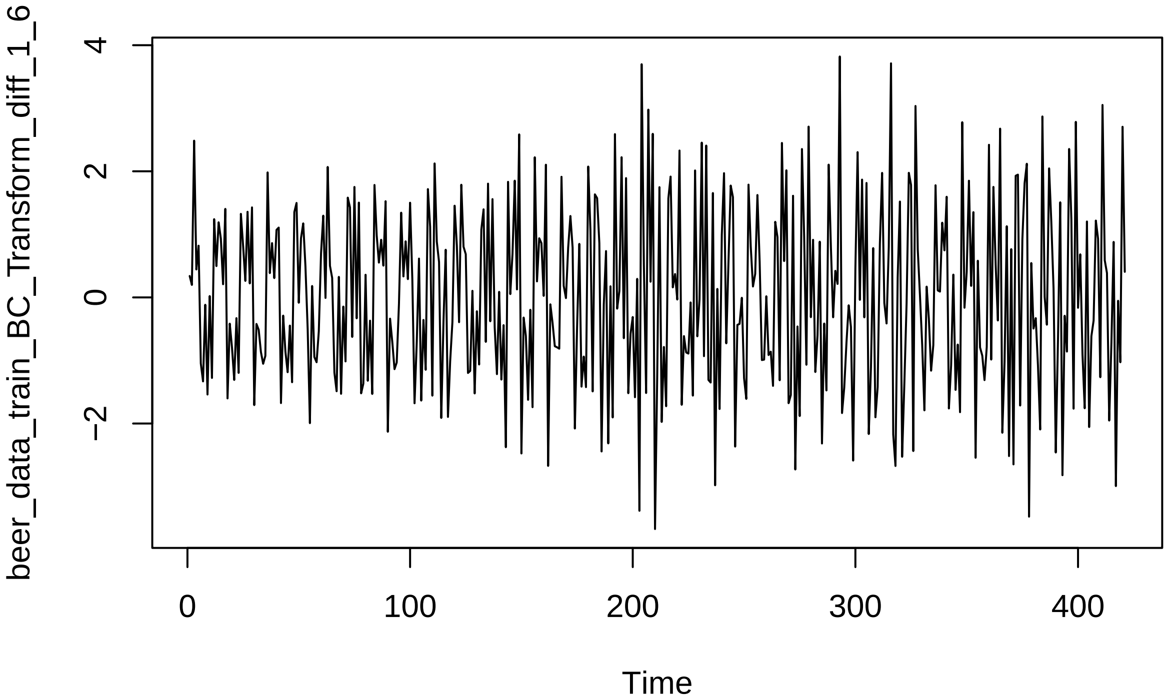


We see that when differencing at lag 1, the data is detrended. But there is also component of seasonality every 6 lags from the ACF. So we will be differecing at lag 6 to reduce that semi annual seasonal component.



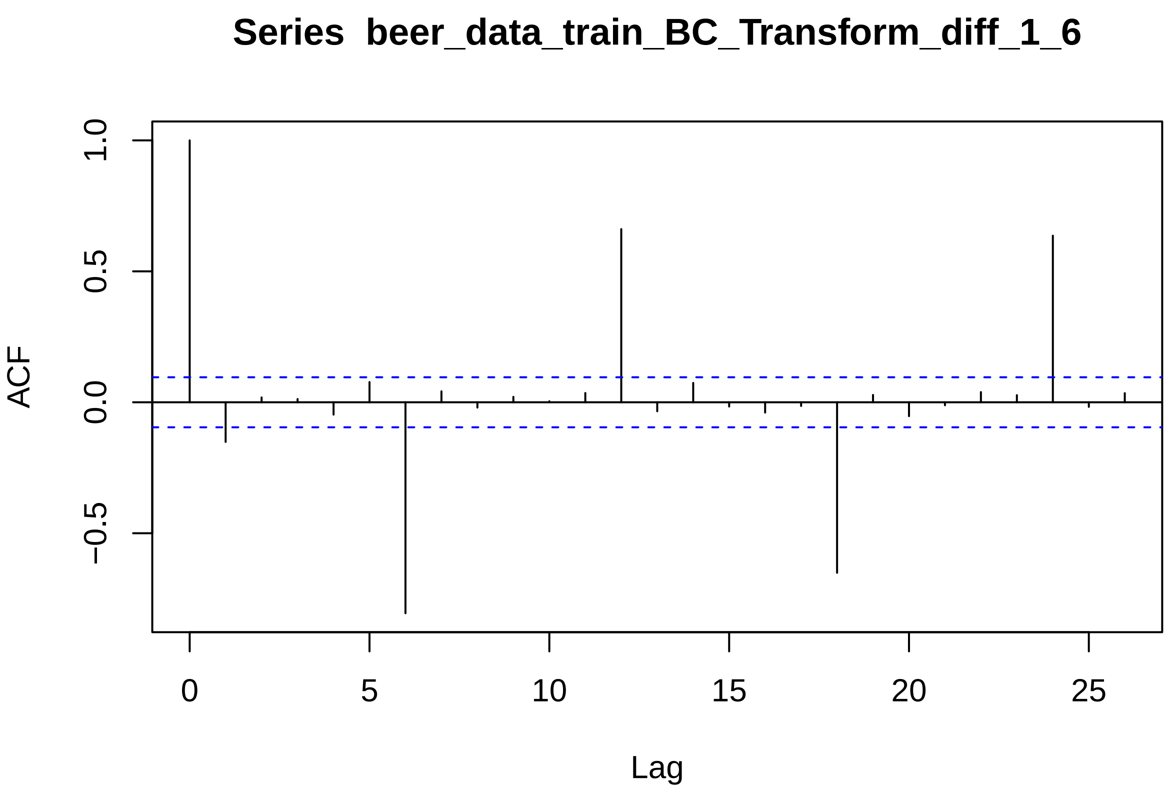
#difference at lag 1 and lag 6

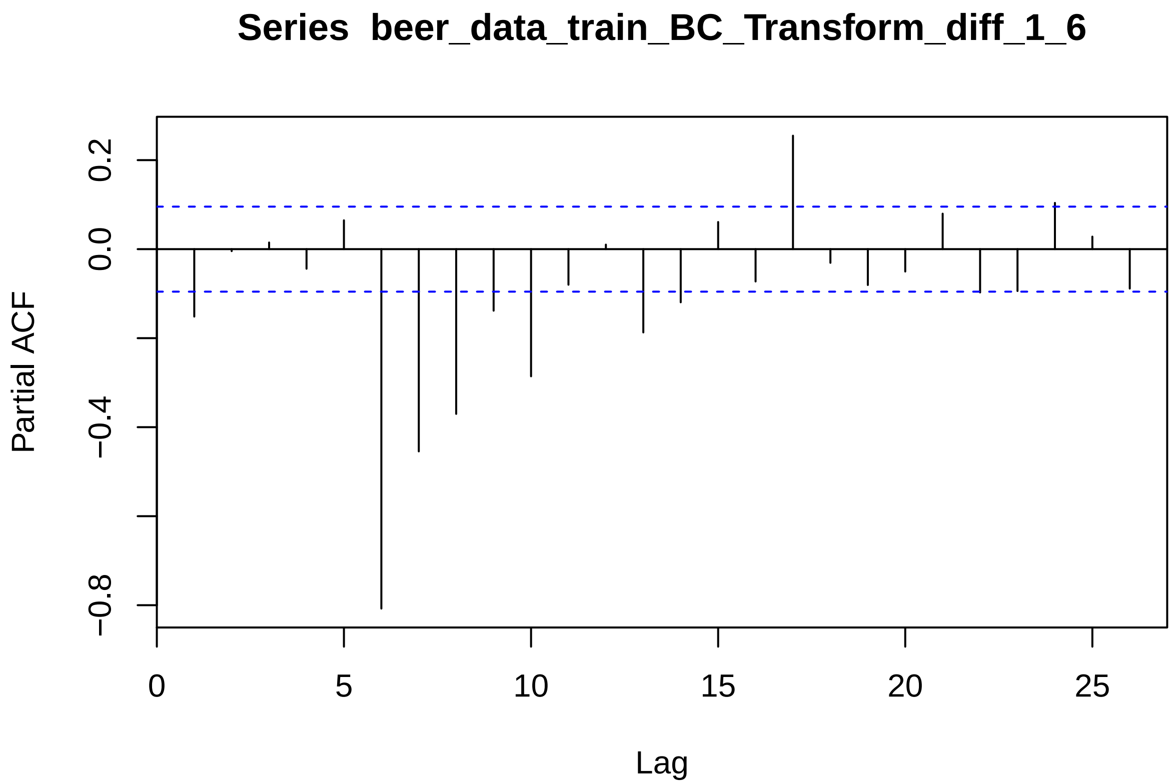




Every 6 lags theres a peak , typical SMA. P=1 , this looks like a seasonal AR(1) since its geomatrically decreasing and signs alternate in the acf. Typical AR(1) feature but since spike at lag 6 , so its seasoanal AR(1). From the PACF we see that lag 6 has a very significant negative spike, this indicates there are seasonal AR(1). P=1

#Model Assumptions





From the ACF we see that lag 1 has a significant spike which might suggest MA(1). So p=1. From the PACF we see that lag 1, lag 2, lag 3 has significant spike but its decreasing. So this can be either AR(1) or AR(3). So q=1 or q=3. From both PACF and ACF, I did not see a seasonal part after differencing it at lag 12.So D=1 and s=12.

From all of these we can assume a couple of models: $\mathrm{SARIMA}(0,1,0)\times(1,1,0)_6$

#Comparing AICc

```
# Pure MA(1)
fit_ma1 <- arima(beer_data_train_BC_Transform, order = c(0, 0, 1))
fit_ma1
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(0, 0, 1))
##
## Coefficients:
##          ma1  intercept
##       0.7437    14.3704
## s.e.  0.0276     0.1010
##
## sigma^2 estimated as 1.439:  log likelihood = -685.64,  aic = 1377.27
```

```
AICc(fit_ma1)#1387.953
```

```
## [1] 1377.328
```

```
#Pure AR(1)
fit_ar1 <- arima(beer_data_train_BC_Transform, order = c(1, 0, 0))
fit_ar1
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.8646    14.3542
## s.e.  0.0241     0.3058
##
## sigma^2 estimated as 0.7553:  log likelihood = -547.95,  aic = 1101.9
```

```
AICc(fit_ar1)#1113.916
```

```
## [1] 1101.955
```

```
#SARIMA(0,1,0) X (1,1,0)
fit_sarima010_110 <- arima(beer_data_train_BC_Transform,order = c(0, 1, 0),seasonal = list(order=c(1,1,(
fit_sarima010_110
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(0, 1, 0), seasonal = list(order = c(1,
##     1, 0), period = 6), method = "ML")
##
## Coefficients:
##          sar1
##       -0.8194
## s.e.   0.0275
##
## sigma^2 estimated as 0.6697:  log likelihood = -516.31,  aic = 1036.62
```

```
AICc(fit_sarima010_110 ) #1041.356
```

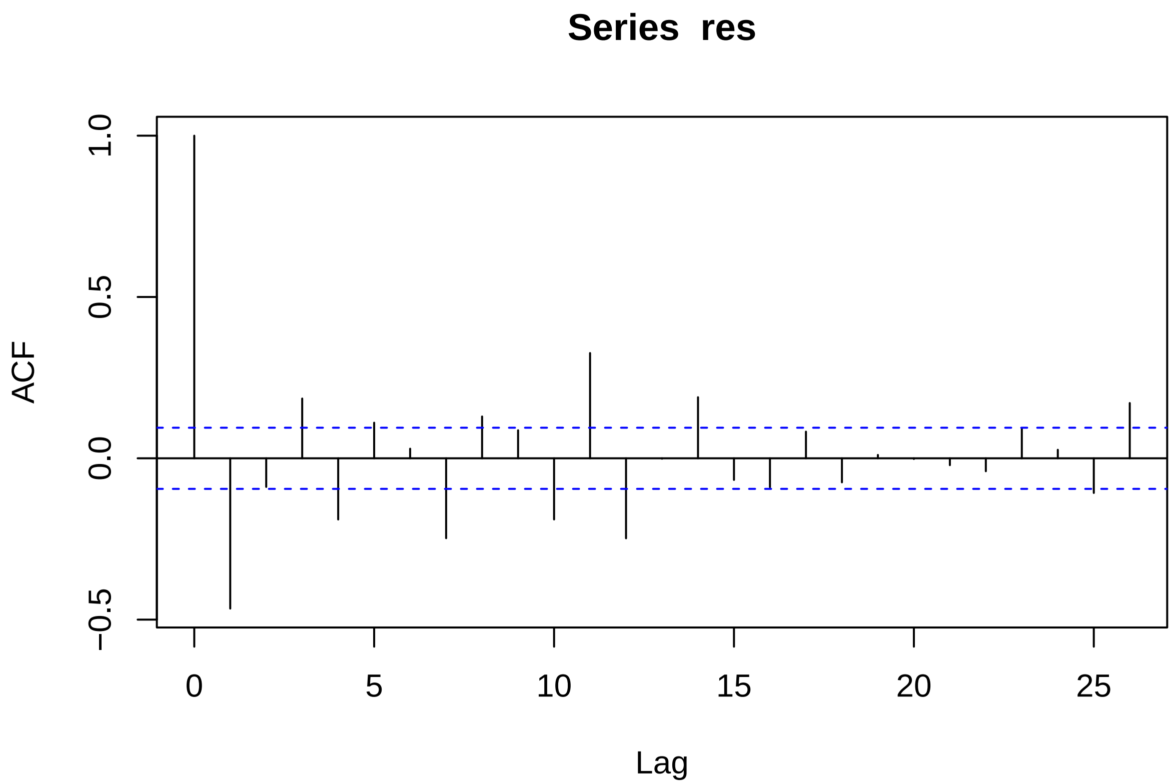
```
## [1] 1036.651
```

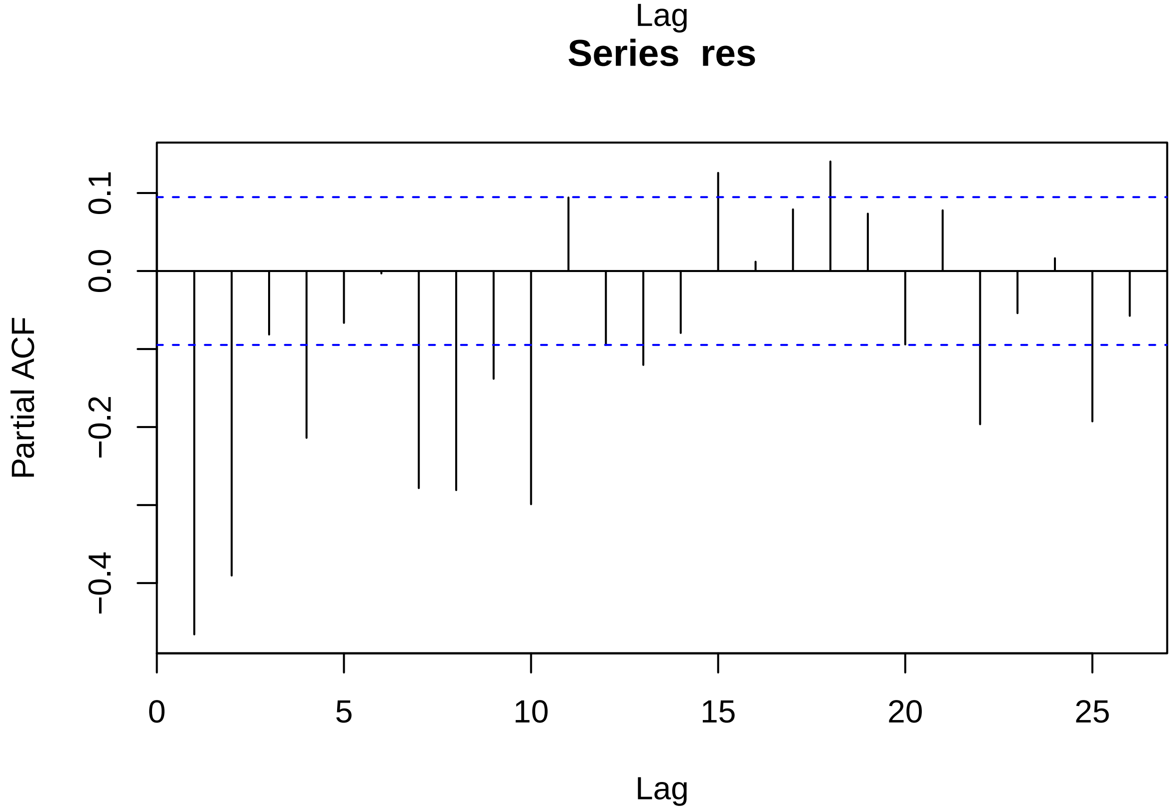
```
Models<-c("MA1","AR1","SARIMA(0,1,0) X (1,1,0), s=6")

AICc=c(AICc(fit_ma1),AICc(fit_ar1),AICc(fit_sarima010_110 ))

data.frame(Models,AICc)
```

```
##                         Models     AICc
## 1                          MA1 1377.328
## 2                          AR1 1101.955
## 3 SARIMA(0,1,0) X (1,1,0), s=6 1036.651
```

From comparing the AICc the SARIMA(0,1,0) X (1,1,0), s=6 did the best with the least AICc.

#Diagnostic checking for residuals





We see tha tthe Residual plot still have significant spikes in both the ACF and PACF which is not allowed. So we have to update the model gain. From the ACF we see that lag 1 is significant which might indicate q=1. From the PACF we see that lag 1, lag 2 are significinat for the non-seasonal part. So p=2 .

```r
#SARIMA(0,1,0) X (1,1,0)
fit_sarima010_110 <- arima(beer_data_train_BC_Transform,order = c(0, 1, 0),seasonal = list(order=c(1,1,
fit_sarima010_110
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(0, 1, 0), seasonal = list(order = c(1,
##     1, 0), period = 6), method = "ML")
##
## Coefficients:
##          sar1
##       -0.8194
## s.e.   0.0275
##
## sigma^2 estimated as 0.6697:  log likelihood = -516.31,  aic = 1036.62
```

```r
AICc(fit_sarima010_110 ) # 1036.651
```

```
## [1] 1036.651
```

```r
#SARIMA(2,1,1) X (1,1,0)
fit_sarima211_110 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,1,
fit_sarima211_110
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(2, 1, 1), seasonal = list(order = c(1,
##     1, 0), period = 6), method = "ML")
##
## Coefficients:
##           ar1      ar2      ma1     sar1
##       -0.1727  -0.1636  -0.9546  -0.9597
## s.e.   0.0524   0.0514   0.0190   0.0130
##
## sigma^2 estimated as 0.3035:  log likelihood = -356.1,  aic = 722.2
```

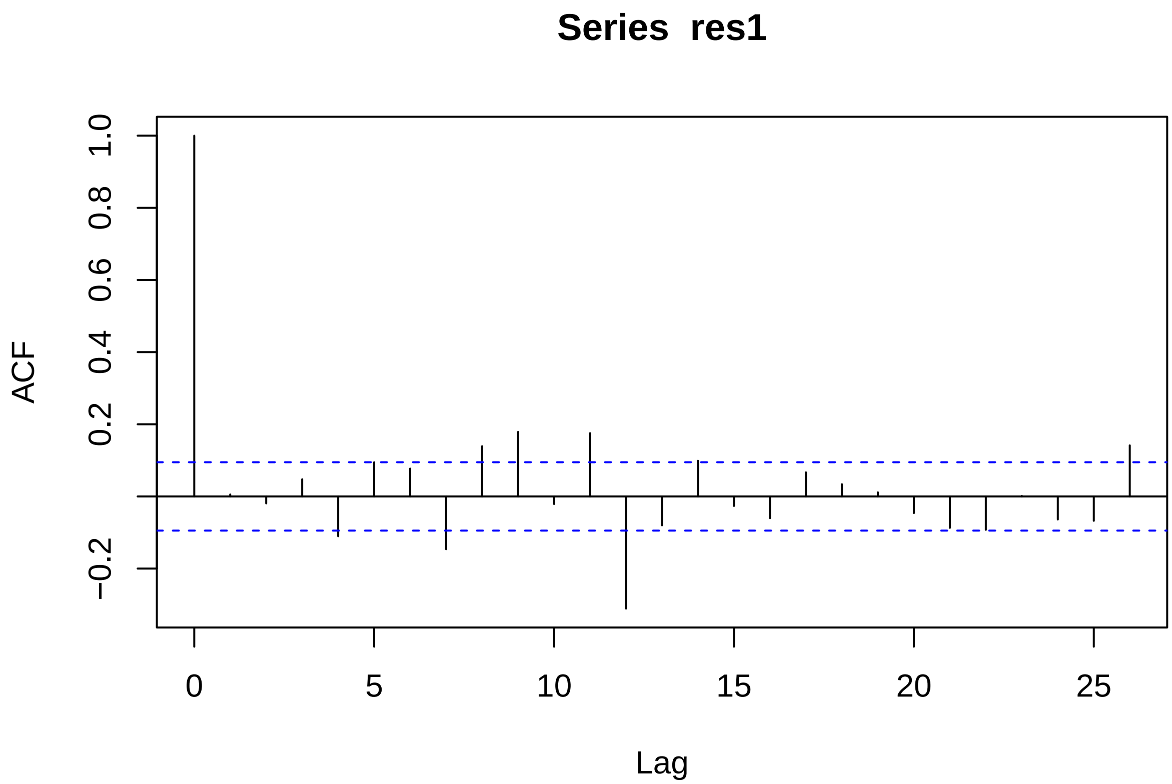
```r
AICc(fit_sarima211_110  ) #722.3407
```

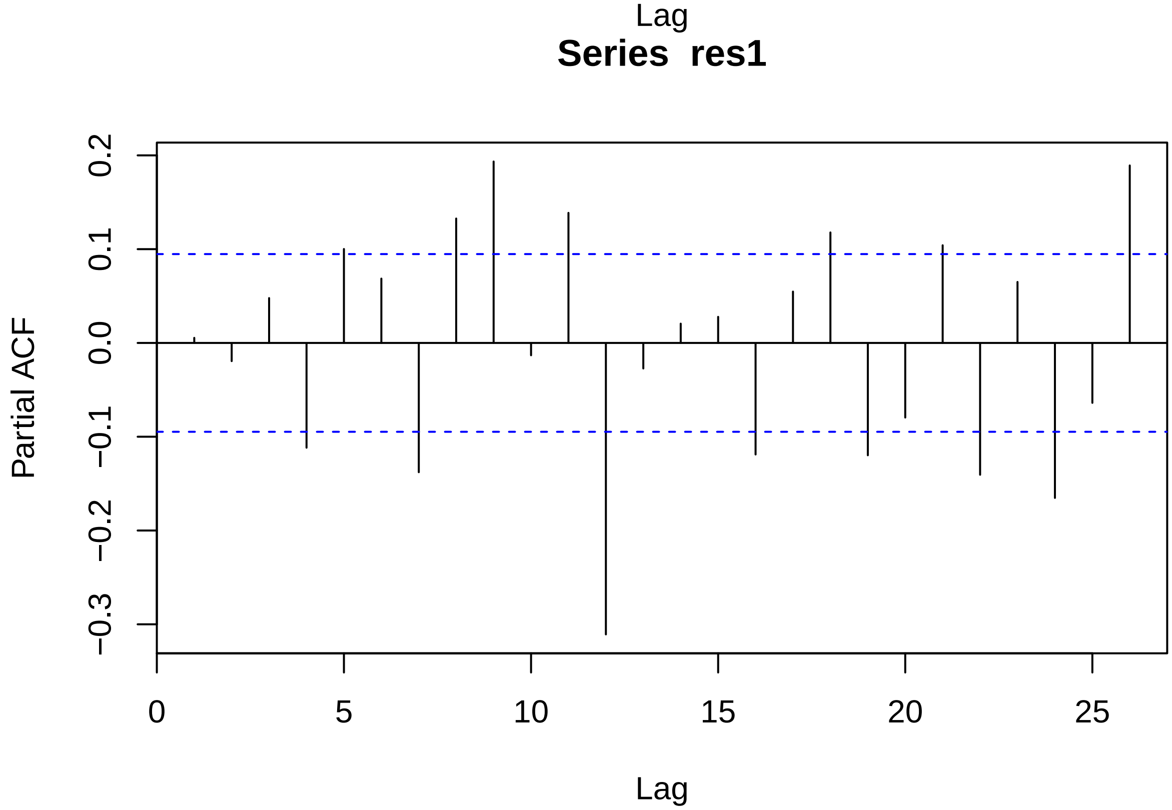
```
## [1] 722.3407
```

```r
Models=c('sarima010_110','sarima211_110 ')
AICc=c(AICc(fit_sarima010_110 ),AICc(fit_sarima211_110  ))

data.frame(Models,AICc)
```

```
##           Models      AICc
## 1  sarima010_110 1036.6506
## 2 sarima211_110   722.3407
```

We see that the updated model has a lower AICc . So we will proceed with that and check again its resiudal.





From the acf and pacf plot , we see that the residuals still have significant spikes.In the ACF lag 12 , in the PACF lag 7 , lag 9 , lag 22, lag 24 , lag 26. Most of the significant lag seems to be coming from a seasonal component. So I tried a variaety of models to try fix these. Assumptions can be Q=2 (caught in the residual 1 acf plot theres a single spike in lag 12 and in the residual plot there is a single spike too in lag 12. And its seasonal.), or try to to differencing the seasonal component twice again which makes D =2.

```r
#SARIMA(2,1,1) X (1,1,0)
fit_sarima211_110 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,1,
fit_sarima211_110
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(2, 1, 1), seasonal = list(order = c(1,
##     1, 0), period = 6), method = "ML")
##
## Coefficients:
##           ar1      ar2      ma1     sar1
##       -0.1727  -0.1636  -0.9546  -0.9597
## s.e.   0.0524   0.0514   0.0190   0.0130
##
## sigma^2 estimated as 0.3035:  log likelihood = -356.1,  aic = 722.2
```

```r
AICc(fit_sarima211_110  ) #722.3407
```

```
## [1] 722.3407
```

```r
#SARIMA(2,1,1) X (1,2,0)
fit_sarima211_120 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,2,
fit_sarima211_120
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(2, 1, 1), seasonal = list(order = c(1,
##     2, 0), period = 6), method = "ML")
##
## Coefficients:
##           ar1      ar2      ma1     sar1
##       -0.1360  -0.1774  -1.0000  -0.9806
## s.e.   0.0492   0.0486   0.0066   0.0078
##
## sigma^2 estimated as 0.5368:  log likelihood = -473.56,  aic = 957.13
```

```r
AICc(fit_sarima211_120) #957.2718
```

```
## [1] 957.2718
```

```r
#SARIMA(2,1,1) X (1,1,2)
fit_sarima211_112 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,1,
fit_sarima211_112
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(2, 1, 1), seasonal = list(order = c(1,
##     1, 2), period = 6), method = "ML")
##
## Coefficients:
##           ar1      ar2      ma1     sar1    sma1     sma2
##       -0.2385  -0.2111  -0.8527  -0.9999  0.0728  -0.8670
## s.e.   0.0528   0.0518   0.0285   0.0002  0.0420   0.0431
##
## sigma^2 estimated as 0.1994:  log likelihood = -277.71,  aic = 569.41
```

```
AICc(fit_sarima211_112) #572.2513
```

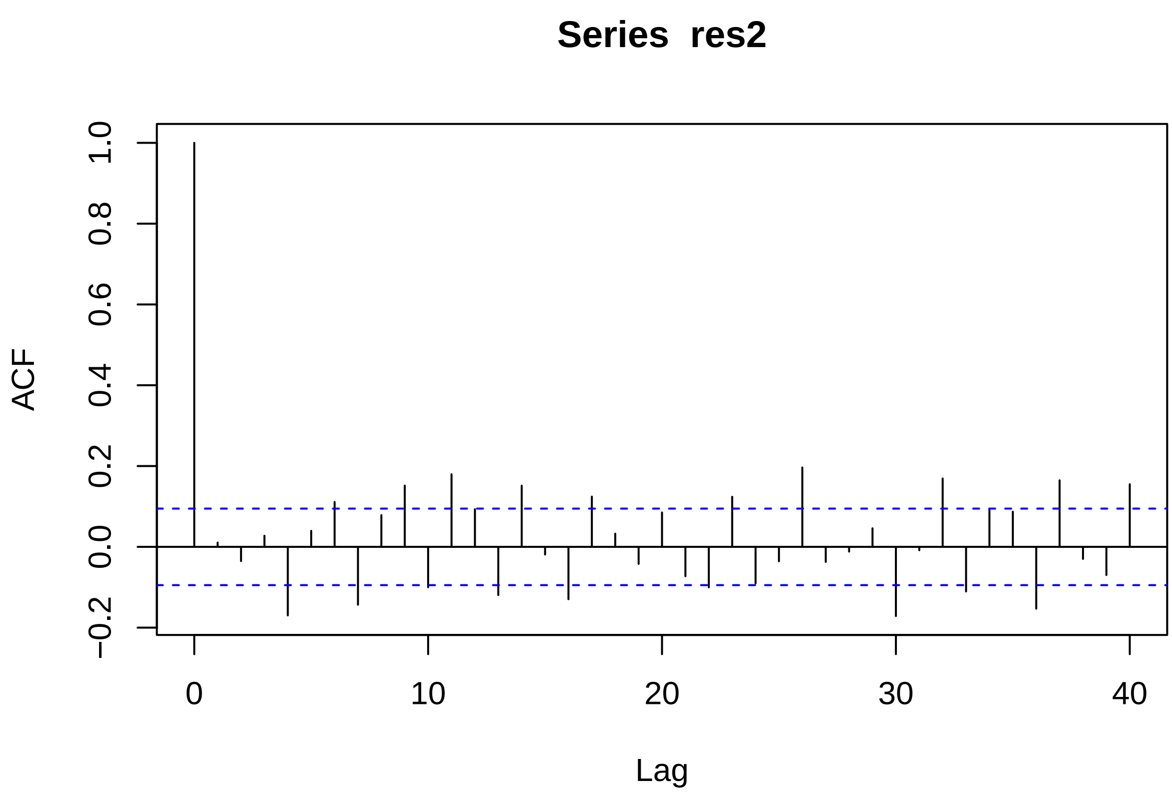
```
## [1] 569.6817
```

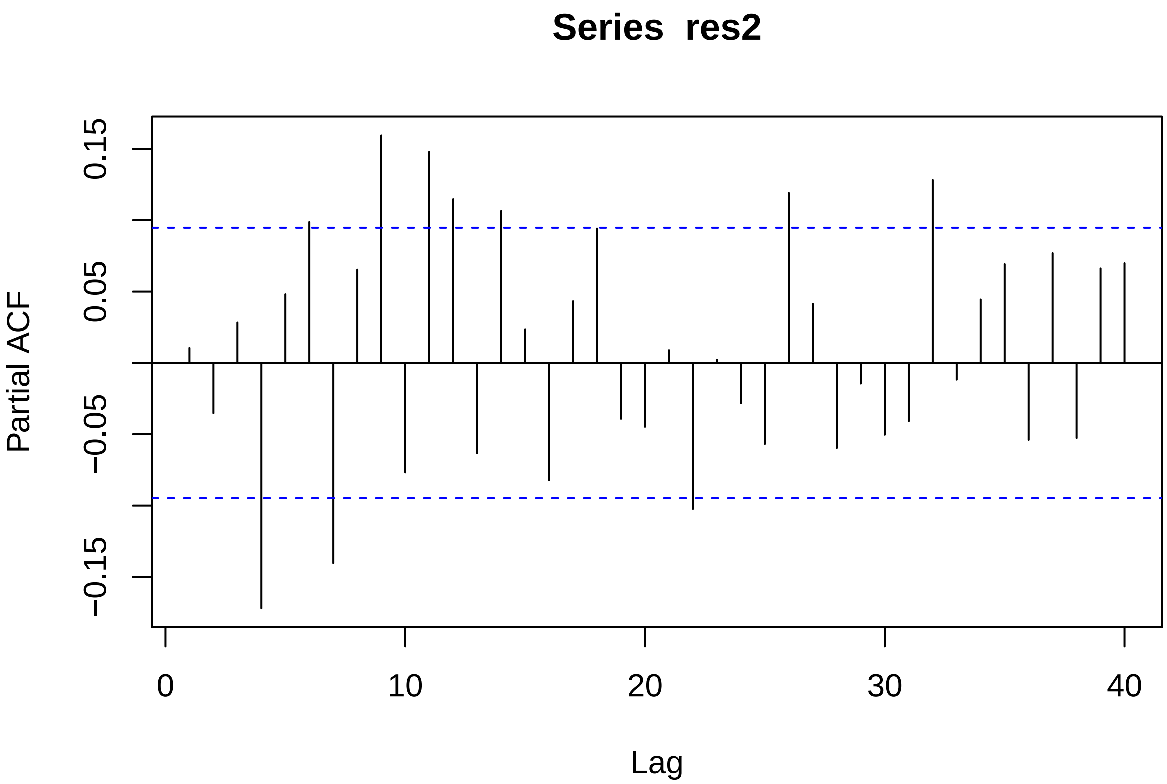
```
Models= c('SARIMA(2,1,1) X (1,1,0)','SARIMA(2,1,1) X (1,2,0)s=6','SARIMA(2,1,1) X (1,1,2)s=6')
AICc=(c(AICc(fit_sarima211_110  ),AICc(fit_sarima211_120),AICc(fit_sarima211_112)))
data.frame(Models,AICc)
```

```
##                        Models     AICc
## 1    SARIMA(2,1,1) X (1,1,0) 722.3407
## 2 SARIMA(2,1,1) X (1,2,0)s=6 957.2718
## 3 SARIMA(2,1,1) X (1,1,2)s=6 569.6817
```

Now we see that after updating the model from seeing the pacf and acf of the resiudal plot the model SARIMA(2,1,1) X (1,1,2)s=6 works best as there is the lowest AICc.

#resiudal diagnostic: Model assumptions and room for improvement.





At the ACF of the residuals, I saw that 2 lags (lag 6 ,lag 9, lag 11 and lag 13) have significant spike. So I tried p=4

```
#SARIMA(4,1,1) X (1,1,2)
fit_sarima411_112 <- arima(beer_data_train_BC_Transform,order = c(4, 1, 1),seasonal = list(order=c(1,1,
fit_sarima411_112
```

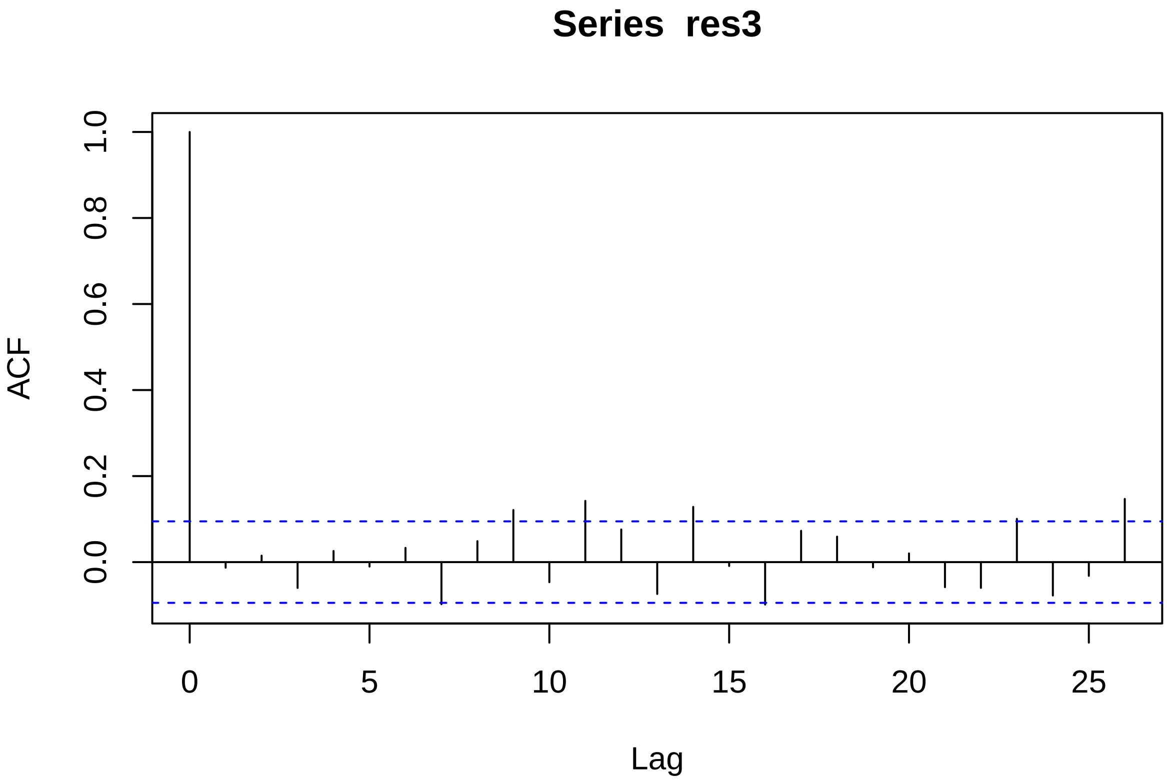
```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(4, 1, 1), seasonal = list(order = c(1,
##     1, 2), period = 6), method = "ML")
##
## Coefficients:
##          ar1      ar2      ar3      ar4      ma1     sar1    sma1     sma2
##       -0.252  -0.2869  -0.0175  -0.2363  -0.7998  -0.9998  0.0366  -0.8726
## s.e.   0.059   0.0631   0.0613   0.0535   0.0423   0.0002  0.0376   0.0369
##
## sigma^2 estimated as 0.1899:  log likelihood = -266.9,  aic = 551.81
```

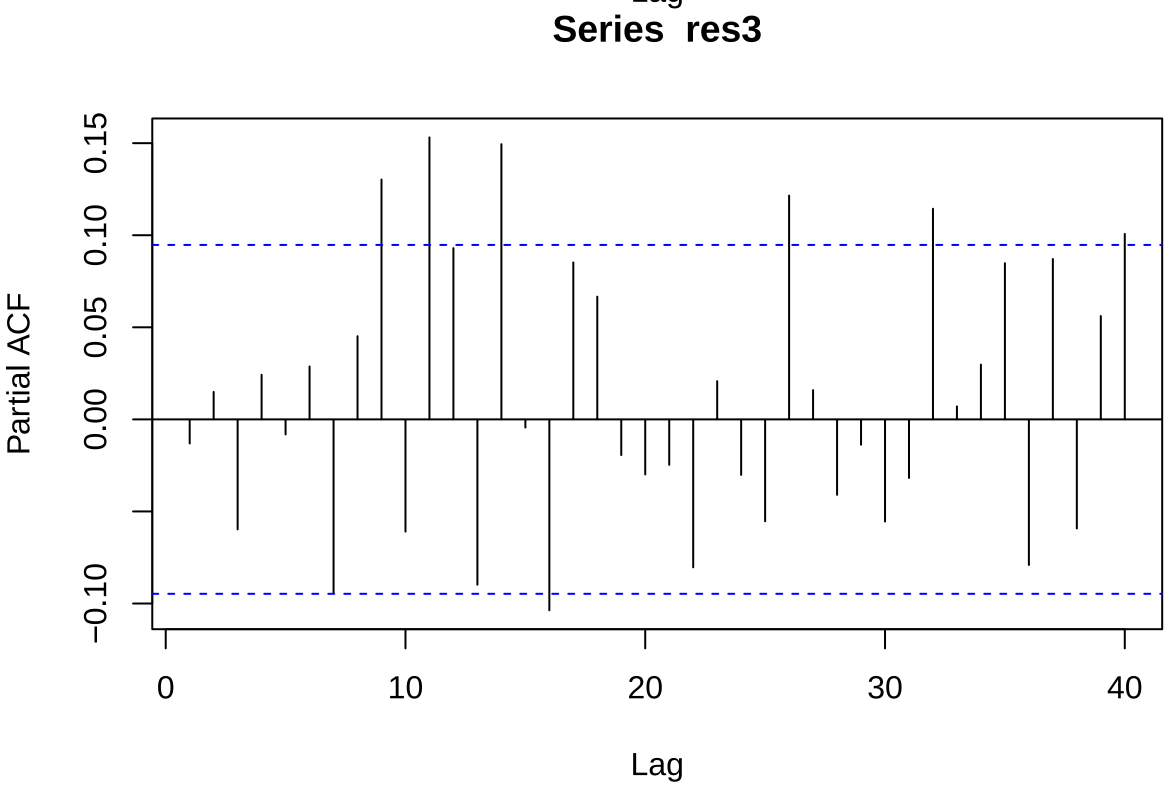
```
AICc(fit_sarima411_112 ) #552.2447
```

```
## [1] 552.2447
```

It improves the model slightly by having a lower AICc. So we will go with the model SARIMA(4,1,1) X (1,1,2). Lets check its resiudal again.

Diagnostic Checking





Now , there is no significant spikes for the residuals in the ACF. However there is still significant spikes in the PACF. There are 3 lags that is significant (lag 9,lag 11 and lag 13 and lag 26). So I decided that p=4.

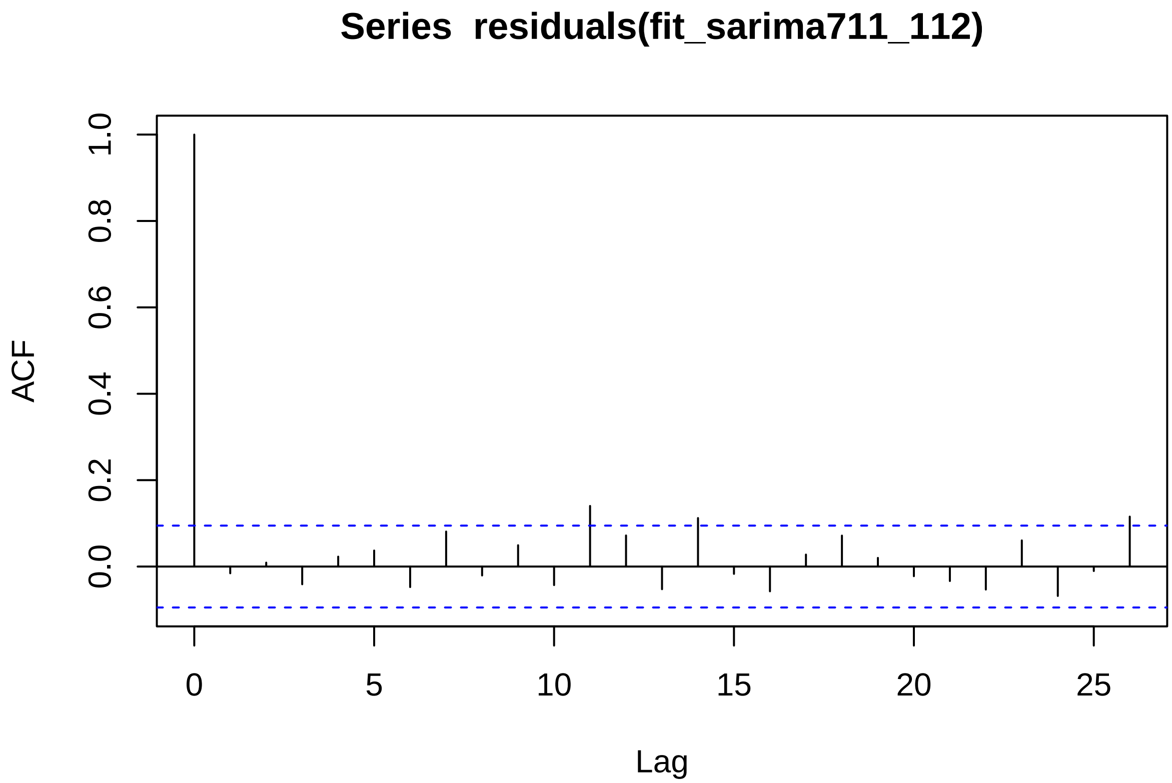
```r
#SARIMA(7,1,1) X (1,1,2)
fit_sarima711_112 <- arima(beer_data_train_BC_Transform,order = c(7, 1, 1),seasonal = list(order=c(1,1,2
fit_sarima711_112
```

```
##
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(7, 1, 1), seasonal = list(order = c(1,
##     1, 2), period = 6), method = "ML")
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ar6      ar7      ma1
##       -0.3189  -0.3483  -0.1535  -0.2895  -0.1301  -0.0014  -0.2096  -0.7186
## s.e.   0.0792   0.0839   0.0876   0.0735   0.0737   0.0644   0.0519   0.0689
##          sar1    sma1     sma2
##       -0.9997  0.0236  -0.8658
## s.e.   0.0003  0.0403   0.0356
##
## sigma^2 estimated as 0.1818:  log likelihood = -256.8,  aic = 537.61
```

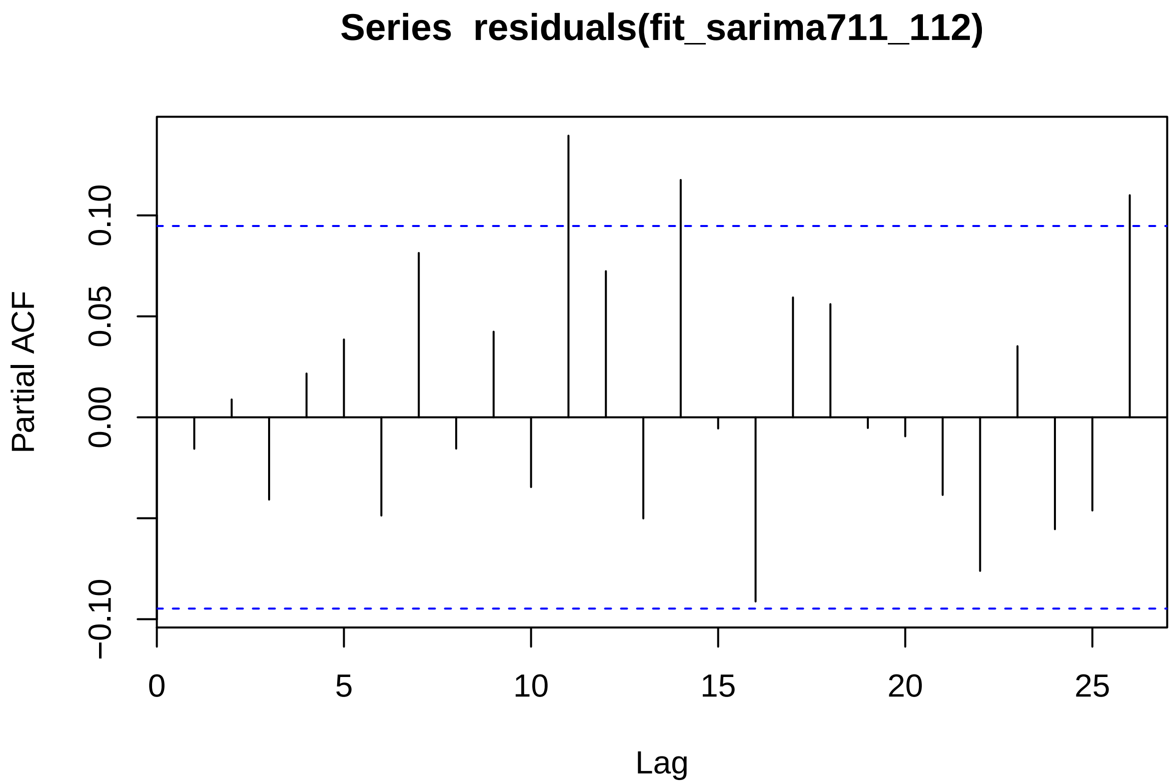
```r
AICc(fit_sarima711_112 ) #538.3702
```

```
## [1] 538.3702
```

```r
acf(residuals(fit_sarima711_112))
```




```r
pacf(residuals(fit_sarima711_112))
```




```
Box.test(residuals(fit_sarima711_112) ^ 2, lag = 20, type = c('Ljung-Box'), fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit_sarima711_112)^2
## X-squared = 57.931, df = 20, p-value = 1.481e-05
```

```
Box.test(residuals(fit_sarima711_112), lag = 20, type = c('Ljung-Box'), fitdf = 5)
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit_sarima711_112)
## X-squared = 30.087, df = 15, p-value = 0.01161
```

```
Box.test(residuals(fit_sarima711_112), lag = 20, type = c("Box-Pierce"), fitdf = 5)
```

```
##
##  Box-Pierce test
##
## data:  residuals(fit_sarima711_112)
## X-squared = 29.134, df = 15, p-value = 0.01546
```

THe model SARIMA(7,1,1) X (1,1,2) failed all the diagnostic checks. So at this point with I bruteforce my model to try to pass the diagnostic test.

```
#SARIMA(8,1,1) X (2,1,2)
fit_sarima811_212 <- arima(beer_data_train_BC_Transform,order = c(8, 1, 1),seasonal = list(order=c(2,1,2
fit_sarima811_212
```

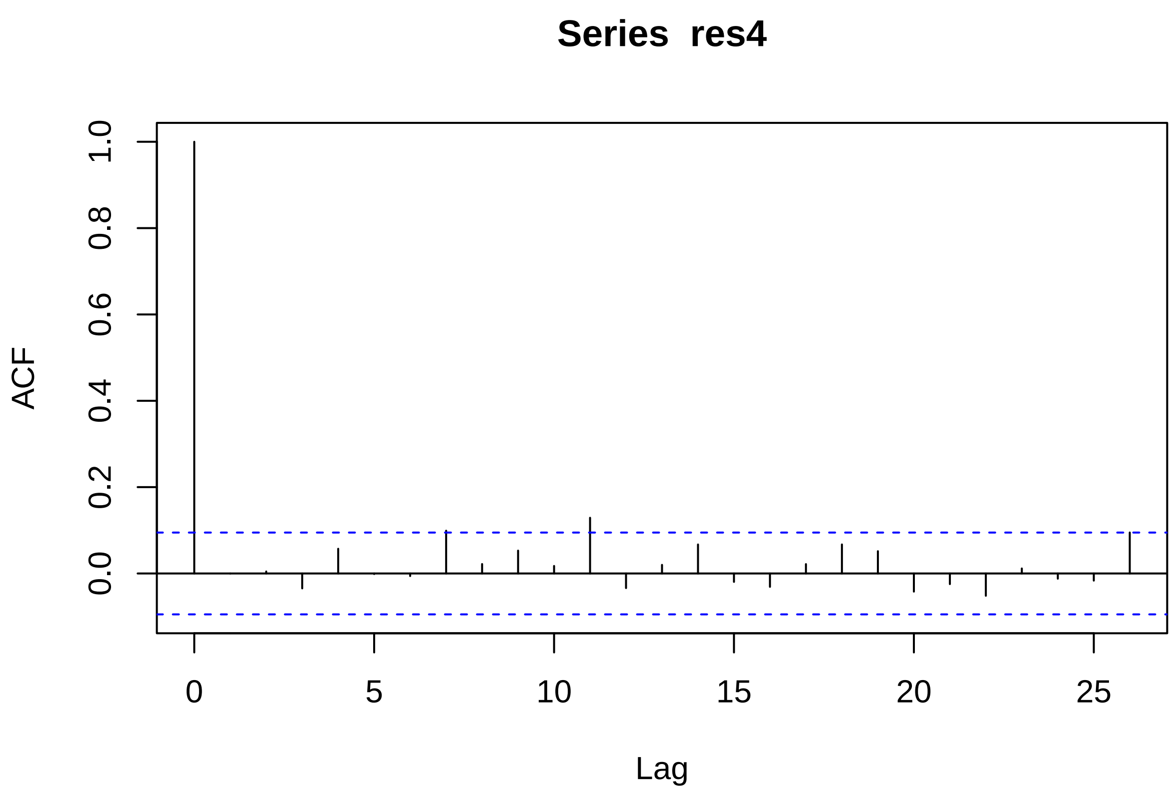
```
##
```

```
## Call:
## arima(x = beer_data_train_BC_Transform, order = c(8, 1, 1), seasonal = list(order = c(2,
##     1, 2), period = 6), method = "ML")
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5     ar6      ar7     ar8
##       -0.5023  -0.5355  -0.3817  -0.4822  -0.2771  0.1386  -0.1858  0.0111
## s.e.      NaN      NaN   0.0069      NaN   0.0042  0.0085      NaN     NaN
##           ma1     sar1     sar2    sma1     sma2
##       -0.5298  -1.3755  -0.3755  0.0671  -0.9071
## s.e.   0.0410      NaN      NaN  0.0249   0.0256
##
## sigma^2 estimated as 0.1743:  log likelihood = -250.89,  aic = 529.78
```

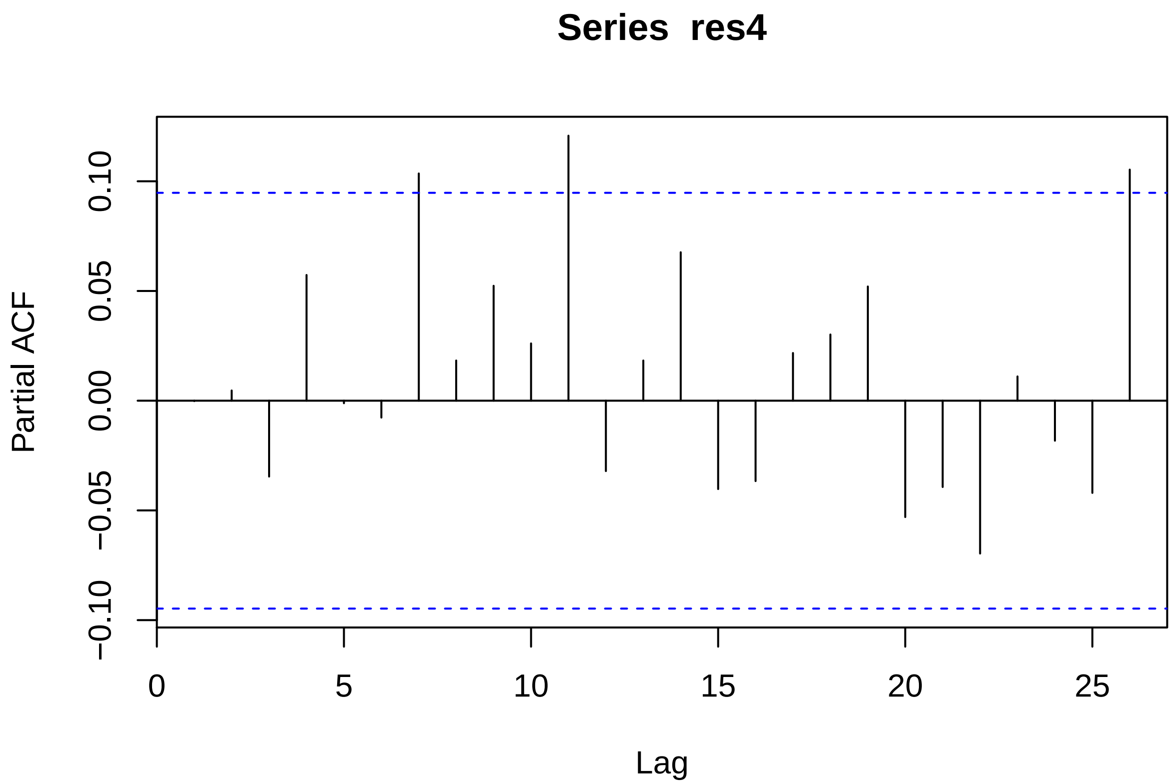
```
AICc(fit_sarima811_212 ) #530.813
```

```
## [1] 530.813
```

```
res4 <-residuals(fit_sarima811_212)
acf(res4 )
```




```
pacf(res4 )
```




```
sqrt(length(res4 ))
```

```
## [1] 20.68816
```

```
#Box pierce
Box.test(res4 , lag = 20, type = c("Box-Pierce"), fitdf = 5)#pass
```

```
##
##  Box-Pierce test
##
## data:  res4
## X-squared = 21.999, df = 15, p-value = 0.1078
```

```
#Box ljung test
Box.test(res4 , lag = 20, type = c('Ljung-Box'), fitdf = 5) # Pass
```

```
##
##  Box-Ljung test
##
## data:  res4
## X-squared = 22.703, df = 15, p-value = 0.09066
```

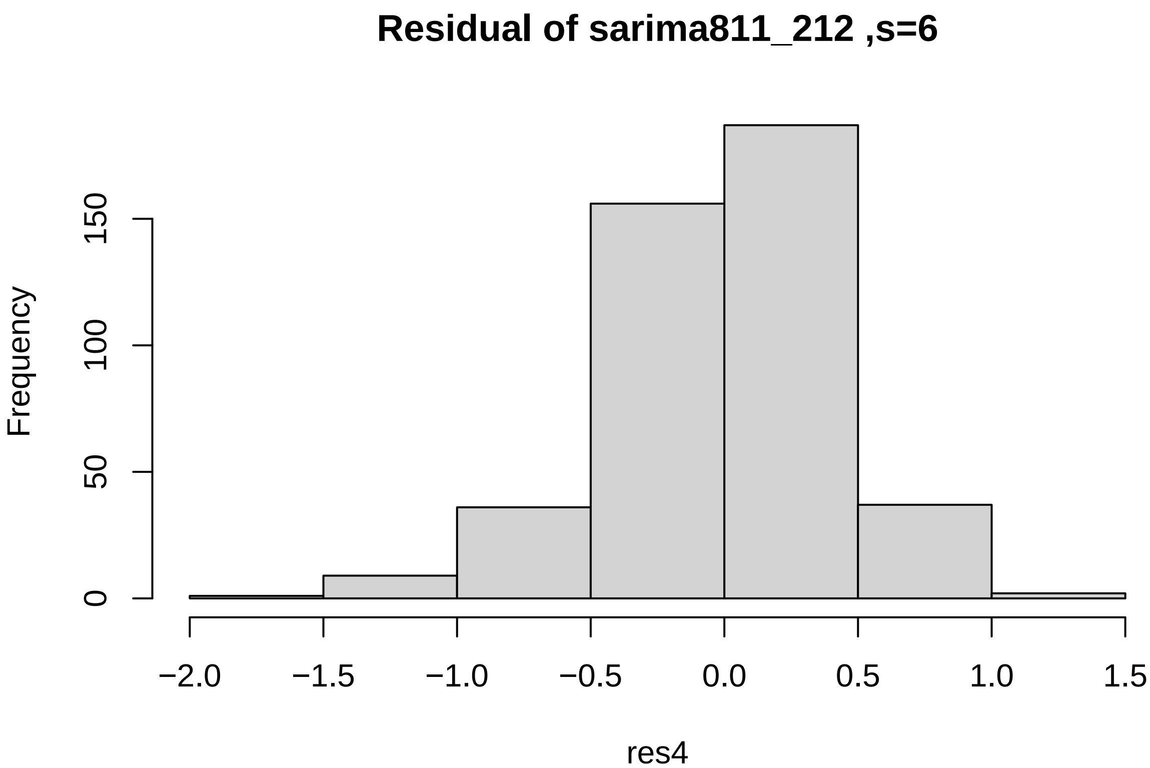
```
# Mc Loid test
Box.test(res4 ^2 , lag = 20, type = c('Ljung-Box'), fitdf = 0) # fail
```

```
##
##  Box-Ljung test
##
## data:  res4^2
## X-squared = 57.649, df = 20, p-value = 1.635e-05
```

```r
shapiro.test(res4 ) # shapiro fail
```

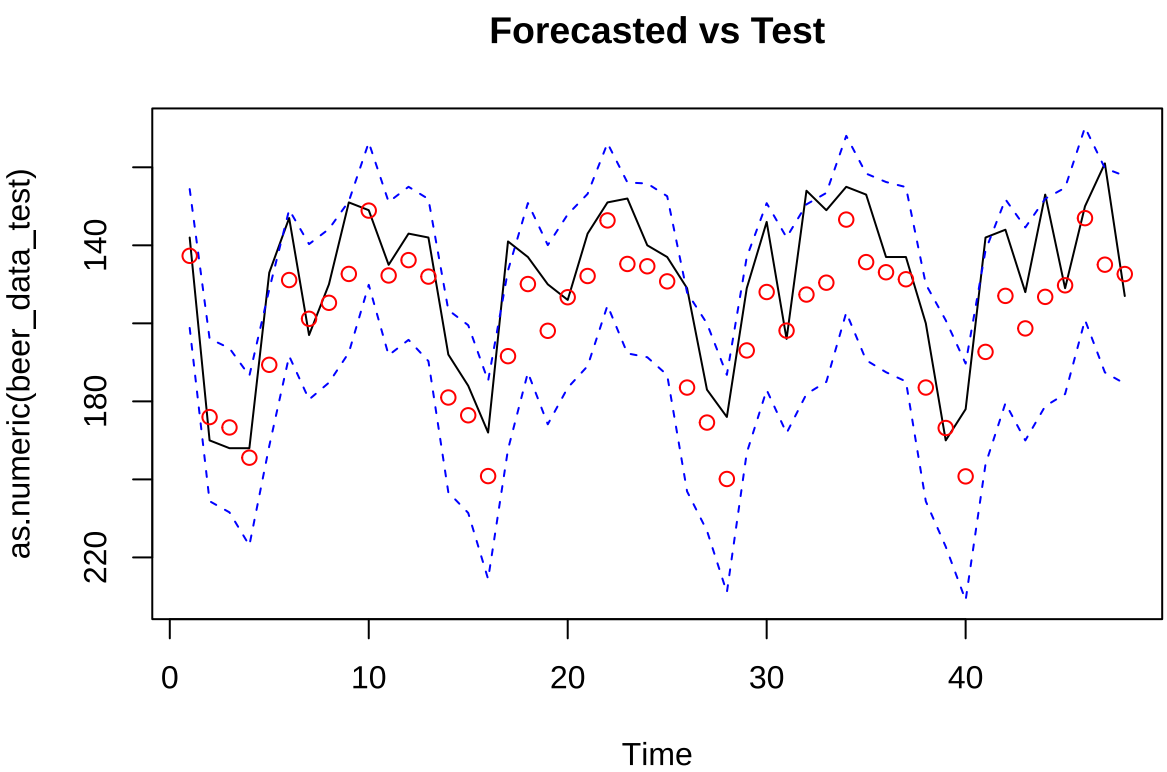
```
##
##  Shapiro-Wilk normality test
##
## data:  res4
## W = 0.9813, p-value = 2.499e-05
```

```r
hist(res4,main = "Residual of sarima811_212 ,s=6")
```



From the diagnotic checks, my model pass the Box pierce and Box ljung test but failed the Mcloid test. Since its passed the box pierce test, we conclude that our residuals are iid. Since We pass the Box Ljung test, we conclude that our resiudals are having nonlinear relationships. We failed the McLoid linear test thus our residuals have non linear dependence. For both the PACF and ACF there is no longer significant spikes so residulas can assume white noise. So I will be forecasting with these model.

#Model Forecasting



The prediction turns out pretty well as it most of the predicted points are close to the real observations. In the end I am happy with my results. Out of the 48 predictions only 10 predicted observations are similar to the real observations.

#Conclusion

I am happy on the project but I could improve more on it. My model failed the Mcloid test and I think it have an impact on the forecasting. It is a fun project to make but a difficult one. From this project I learned alot from making models to forecast from scratch by looking at the PACF, ACF , residuals , AICc and diagnostic check.

#Apendix

```
#lib
library(MuMIn)
library(ggplot2)
library(dplyr)
library(tsdl)
library(astsa)
library(MASS)
#install.packages('MuMIn')
#library(MuMIn)
library(ggplot2)
#install.packages('ggfrtify')
#library(ggfortify)
library(forecast)
#install.packages('quant')
#install.packages('tsereis')
#######
#install.packages("FinTS")
library(FinTS) # Arch Test
#install.packages('rugarch')
```

```
library(rugarch)# Garch Models
library(tseries) # Unit root test
library(zoo)
#install.packages('dynlm')
library('dynlm') # use labes in models
#install.packages('vars')
library(vars) # Use VAR
#install.packages('nlwaldTest') # Testing non-linear wald test or can use Mcloid
#library(nlwaldTest)
#install.packages('lmtest')
library(lmtest)# BP test
library(broom) # table presentations
library(car) #Robust standard error
library(sandwich)
library(knitr)
library(forecast)
library(ggplot2)
#install.packages('pdftech') # import financial data
#library(pdftech)
#install.packages('tsbox')
library(tsbox)
#install.packages('vrtest')
#install.packages('tsdl')
library(tsdl)
library(vrtest)
#install.packages('devtools')
devtools::install_github("FinYang/tsdl")

#meta_tsdl$source[[98]]
#meta_tsdl$description[[98]] = Monthly beer production in Australia Jan 1956 - Aug 1995
#meta_tsdl$frequency[[98]]



beer_data<-ts(tsdl[[98]],start = c(1956,1),end = c(1995,8),frequency = 12)

#beer_data_train <-ts(tsdl[[98]],start = c(1956,1),end = c(1991,11),frequency = 12)

#beer_data_test<-ts(tsdl[[98]],start = c(1991,12),end = c(1995,8),frequency = 12)

beer_data_train <-beer_data[1:428]
beer_data_test<-beer_data[429:476]

#data.frame(beer_data[476]) %>%  dim()
# 476 entries
#476*0.9 = 428 , train =[1:428]


length(beer_data_test)
#test = [429:476]

#data.frame(beer)
```

```
plot(beer_data,ylab="Beer Production",main="Monthly beer production in Australia Jan 1956 - Aug 1995")
abline(h=mean(beer_data),col="Blue")
#abline(v= ts(c(1959.4,1960.4,1961.4,1962.4)),lty=2,col="red")
abline(v=ts(seq(1959.4,1995.4,by=1)),lty=2,col="red") # quarterly seasonal

ts(beer_data_train,frequency = 12) %>% decompose() %>%  plot()

beer_data_train %>%  Auto.VR()
#data.frame(beer_data_train)

#mean(beer_data_train[c(1:214)]) #109.3481
#mean(beer_data_train[c(214:428)]) #160.5758

#var(beer_data_train[c(1:214)]) #622.3758
#var(beer_data_train[c(214:428)]) #461.9746

Part<-c("First Half","Second Half")
Mean<-c(mean(beer_data_train[c(1:214)]),mean(beer_data_train[c(214:428)]))
Variance<- c(var(beer_data_train[c(1:214)]) ,var(beer_data_train[c(214:428)]) )

data.frame(Part,Mean,Variance)

hist(beer_data_train)
#shapiro.test(beer_data_train)

par(mfrow=c(1,3))

hist(beer_data_train, main = "Train Data")
hist(beer_data_train_BC_Transform,main="Box Cox Transformation (Train Data)")
hist(beer_data_train_log,main="Log Transformation (Train Data)")

par(mfrow=c(1,1))
ts(beer_data_train_BC_Transform,frequency = 12) %>% decompose() %>% plot()

# Calculating the confidence interval 95%
#1.96 * 1/sqrt(n)
n_train<-as.numeric(length(beer_data_train))
margin_error<- 1.96 * 1/sqrt(n_train)

#Beer Data Train PACF and ACF

par(mfrow=c(1,2))
#PACF train
#Beer_Data_Train_pacf<-pacf(beer_data_train)
#plot(Beer_Data_Train_pacf$lag*12,Beer_Data_Train_pacf$acf,type="h",ylim=c(-0.3,1),xlab="Lag",ylab="Par
#abline(h=c(-margin_error,0,margin_error),col=c("blue","black","blue"),lty=c(2,1,2))

#ACF train
#Beer_data_Train_acf<-acf(beer_data_train)
#plot(Beer_data_Train_acf$lag*12,Beer_data_Train_acf$acf,type="h",ylim=c(-0.3,1),xlab="Lag",ylab="ACF",
#abline(h=c(-margin_error,0,margin_error),col=c("blue","black","blue"),lty=c(2,1,2))

#beer_data_train_BC_Transform PACF and ACF
```

```
#PACF train box cox
beer_data_train_BC_Transform_pacf<-pacf(beer_data_train_BC_Transform,lag.max = 40)
#plot(beer_data_train_BC_Transform_pacf$lag*12,beer_data_train_BC_Transform_pacf$acf,type="h",ylim=c(-0
#abline(h=c(-margin_error,0,margin_error),col=c("blue","black","blue"),lty=c(2,1,2))

#ACF train box cox
beer_data_train_BC_Transform_acf<-acf(beer_data_train_BC_Transform)
#plot(Beer_data_Train_acf$lag*12,Beer_data_Train_acf$acf,type="h",ylim=c(-0.3,1),xlab="Lag",ylab="ACF",
#abline(h=c(-margin_error,0,margin_error),col=c("blue","black","blue"),lty=c(2,1,2))

beer_data_train_BC_Transform_diff_1<-diff(beer_data_train_BC_Transform,1) # Difference at lag 1

acf(beer_data_train_BC_Transform_diff_1,lag.max = 40,main="difference at lag 1 acf") # difference at la

pacf(beer_data_train_BC_Transform_diff_1,lag.max = 40,main="difference at lag 1 pacf")# difference at l

#plot(beer_data_train_BC_Transform_diff_1,main="difference at lag 1")
beer_data_train_BC_Transform_diff_1_6 <-diff(beer_data_train_BC_Transform_diff_1,6) # difference at lag
acf(beer_data_train_BC_Transform_diff_1_6) # Every 6 lags theres a peak , typical SMA. P=1 , this looks

pacf(beer_data_train_BC_Transform_diff_1_6)# From the PACF we see that lag 6 has a very significant neg

plot.ts(beer_data_train_BC_Transform_diff_1_6)

beer_data_train_BC_Transform_diff_1_6 <-diff(beer_data_train_BC_Transform_diff_1,6) # difference at lag
acf(beer_data_train_BC_Transform_diff_1_6) # Every 6 lags theres a peak , typical SMA. P=1 , this looks

pacf(beer_data_train_BC_Transform_diff_1_6)# From the PACF we see that lag 6 has a very significant neg

# Pure MA(1)
fit_ma1 <- arima(beer_data_train_BC_Transform, order = c(0, 0, 1))
fit_ma1
AICc(fit_ma1)#1387.953

#Pure AR(1)
fit_ar1 <- arima(beer_data_train_BC_Transform, order = c(1, 0, 0))
fit_ar1
AICc(fit_ar1)#1113.916

#SARIMA(0,1,0) X (1,1,0)
fit_sarima010_110 <- arima(beer_data_train_BC_Transform,order = c(0, 1, 0),seasonal = list(order=c(1,1,0
fit_sarima010_110
AICc(fit_sarima010_110 ) #1041.356

Models<-c("MA1","AR1","SARIMA(0,1,0) X (1,1,0), s=6")

AICc=c(AICc(fit_ma1),AICc(fit_ar1),AICc(fit_sarima010_110 ))

data.frame(Models,AICc)

#SARIMA(0,1,0) X (1,1,0)
fit_sarima010_110 <- arima(beer_data_train_BC_Transform,order = c(0, 1, 0),seasonal = list(order=c(1,1,0
fit_sarima010_110
```

```r
#Non seaosonal

res<-residuals(fit_sarima010_110)

acf(res) # q=1 , lag 1 very significant
pacf(res) # PACF shows that p =2, lag 1 & lag 2 are significant


#SARIMA(0,1,0) X (1,1,0)
fit_sarima010_110 <- arima(beer_data_train_BC_Transform,order = c(0, 1, 0),seasonal = list(order=c(1,1,0
fit_sarima010_110
AICc(fit_sarima010_110 ) # 1036.651

#SARIMA(2,1,1) X (1,1,0)
fit_sarima211_110 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,1,0
fit_sarima211_110
AICc(fit_sarima211_110  ) #722.3407

Models=c('sarima010_110','sarima211_110 ')
AICc=c(AICc(fit_sarima010_110 ),AICc(fit_sarima211_110  ))

data.frame(Models,AICc)

res1<-residuals(fit_sarima211_110 )

acf(res1) # we see spike at lag 12, do another differencing
pacf(res1)

#SARIMA(2,1,1) X (1,1,0)
fit_sarima211_110 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,1,0
fit_sarima211_110
AICc(fit_sarima211_110  ) #722.3407

#SARIMA(2,1,1) X (1,2,0)
fit_sarima211_120 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,2,0
fit_sarima211_120
AICc(fit_sarima211_120) #957.2718

#SARIMA(2,1,1) X (1,1,2)
fit_sarima211_112 <- arima(beer_data_train_BC_Transform,order = c(2, 1, 1),seasonal = list(order=c(1,1,2
fit_sarima211_112
AICc(fit_sarima211_112) #572.2513

Models= c('SARIMA(2,1,1) X (1,1,0)','SARIMA(2,1,1) X (1,2,0)s=6','SARIMA(2,1,1) X (1,1,2)s=6')
AICc=(c(AICc(fit_sarima211_110  ),AICc(fit_sarima211_120),AICc(fit_sarima211_112)))
data.frame(Models,AICc)

res2<-residuals(fit_sarima211_112 )
acf(res2,lag.max=40)
pacf(res2,lag.max = 40) # From the PACF we could try quarterly seasonility because there is a significa

res2<-residuals(fit_sarima211_112 )
acf(res2,lag.max=40)
```

```r
pacf(res2,lag.max = 40) # From the PACF we could try quarterly seasonility because there is a significa

#SARIMA(4,1,1) X (1,1,2)
fit_sarima411_112 <- arima(beer_data_train_BC_Transform,order = c(4, 1, 1),seasonal = list(order=c(1,1,
fit_sarima411_112
AICc(fit_sarima411_112 ) #552.2447

res3<-residuals(fit_sarima411_112)
acf(res3)
pacf(res3,lag.max = 40) # P =2
#shapiro.test(res3) # The trend is piece-wice linear but our model assume the data is linear so I think

#SARIMA(7,1,1) X (1,1,2)
fit_sarima711_112 <- arima(beer_data_train_BC_Transform,order = c(7, 1, 1),seasonal = list(order=c(1,1,
fit_sarima711_112
AICc(fit_sarima711_112 ) #538.3702

acf(residuals(fit_sarima711_112))
pacf(residuals(fit_sarima711_112))


Box.test(residuals(fit_sarima711_112) ^ 2, lag = 20, type = c('Ljung-Box'), fitdf = 0)

Box.test(residuals(fit_sarima711_112), lag = 20, type = c('Ljung-Box'), fitdf = 5)

Box.test(residuals(fit_sarima711_112), lag = 20, type = c("Box-Pierce"), fitdf = 5)

#length(res4)
#476-429

beer_data_train <-beer_data[1:428]
beer_data_test<-beer_data[429:476]

pred_transform<-predict(fit_sarima811_212 ,n.ahead = 48)
#pred_transform

pred<-(pred_transform$pred*lambda + 1) ^ (1/lambda)

up_trans<-pred_transform$pred +2*pred_transform$se
low_trans<-pred_transform$pred -2 *pred_transform$se

up<- (up_trans*lambda +1) ^ (1/lambda)
low<- (low_trans*lambda +1) ^(1/lambda)



ts.plot(as.numeric(beer_data_test),main="Forecasted vs Test",ylim=c(max(up),min(low)))
lines(1:48,up,lty="dashed" ,col='blue')
lines(1:48, low, lty="dashed",col='blue')
points(1:48,pred,col='red')
```