ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 3 - Due date 02/01/24

Student Name

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp24.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

Please keep this R code chunk options for the report. It is easier for us to grade when we can see code and output together. And the tidy.opts will make sure that line breaks on your code chunks are automatically added for better visualization.

When you have completed the assignment, **Knit** the text and code into a single PDF file. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A03_Sp24.Rmd"). Submit this pdf using Sakai.

Questions

Consider the same data you used for A2 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumpti The data comes from the US Energy Information and Administration and corresponds to the January 2022 **Monthly** Energy Review. You will work only with the following columns: Total Renewable Energy Production and Hydroelectric Power Consumption. Create a data frame structure with these two series only.

R packages needed for this assignment:"forecast","tseries", and "cowplot". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

```
#Importing data set - using readxl package
energy_data <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
                          skip = 12,
                          sheet="Monthly Data",
                          col_names=FALSE) #startRow is equivalent to skip on read.table
```

```
## New names:
## * ‘‘ -> ‘...1‘
## * ‘‘ -> ‘...2‘
## * ‘‘ -> ‘...3‘
## * ‘‘ -> ‘...4‘
## * ‘‘ -> ‘...5‘
## * ‘‘ -> ‘...6‘
## * ‘‘ -> ‘...7‘
## * ‘‘ -> ‘...8‘
## * ‘‘ -> ‘...9‘
## * ‘‘ -> ‘...10‘
## * ‘‘ -> ‘...11‘
## * ‘‘ -> ‘...12‘
## * ‘‘ -> ‘...13‘
## * ‘‘ -> ‘...14‘
```

```
#Now let's extract the column names from row 11 only
read_col_names <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
                             skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)
```

```
## New names:
## * ‘‘ -> ‘...1‘
## * ‘‘ -> ‘...2‘
## * ‘‘ -> ‘...3‘
## * ‘‘ -> ‘...4‘
## * ‘‘ -> ‘...5‘
## * ‘‘ -> ‘...6‘
## * ‘‘ -> ‘...7‘
## * ‘‘ -> ‘...8‘
## * ‘‘ -> ‘...9‘
## * ‘‘ -> ‘...10‘
## * ‘‘ -> ‘...11‘
## * ‘‘ -> ‘...12‘
## * ‘‘ -> ‘...13‘
## * ‘‘ -> ‘...14‘
```

```
colnames(energy_data) <- read_col_names
#head(energy_data)

data <- energy_data[,c(1,5:6)]
nobs <- nrow(data)

#Create vector t - time index
t <- 1:nobs

#transforming just the two columns of interest into ts object
tsdata <- ts(data[t,2:3], frequency=12,start=c(1973,1))
#frequency=12 because of monthly data. Data starts from Jan 1973.

head(tsdata)
```

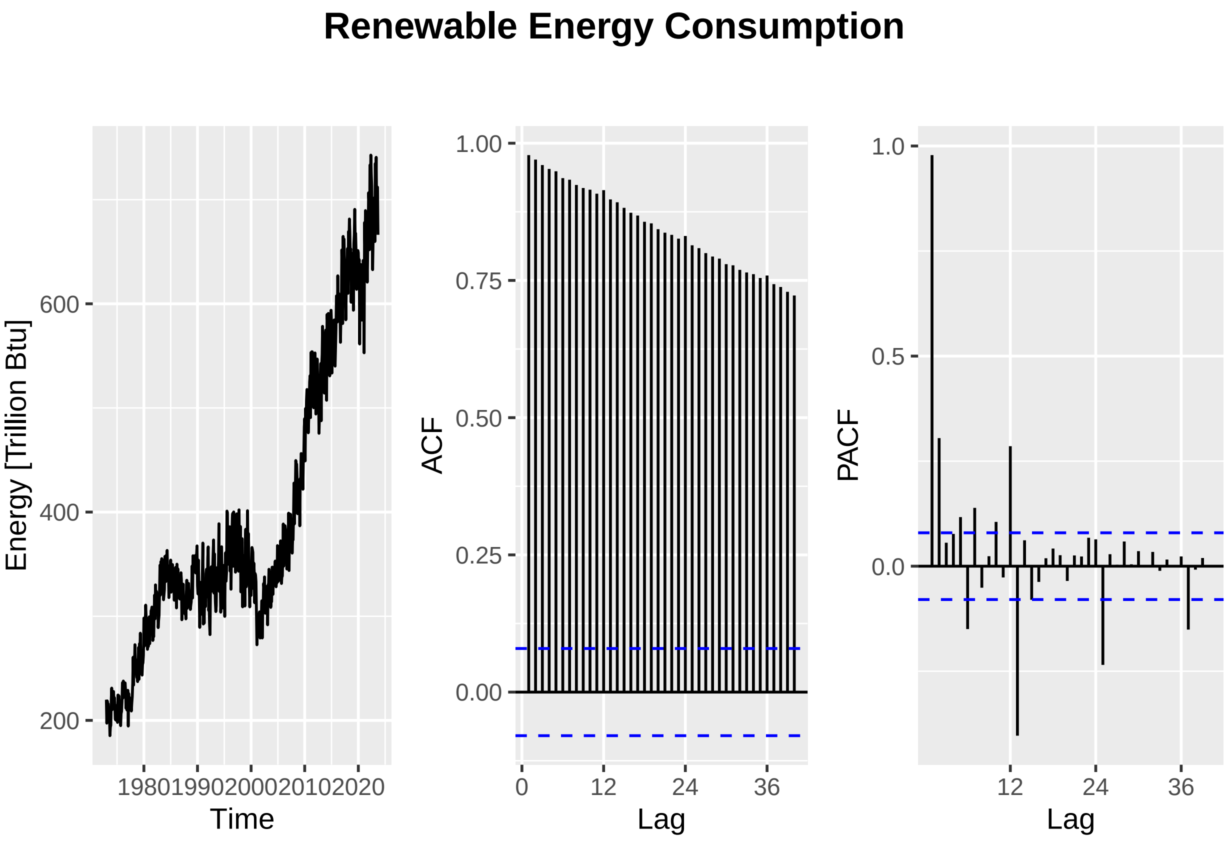
```
##          Total Renewable Energy Production Hydroelectric Power Consumption
## Jan 1973                           219.839                          89.562
## Feb 1973                           197.330                          79.544
## Mar 1973                           218.686                          88.284
## Apr 1973                           209.330                          83.152
## May 1973                           215.982                          85.643
## Jun 1973                           208.249                          82.060
```

##Trend Component

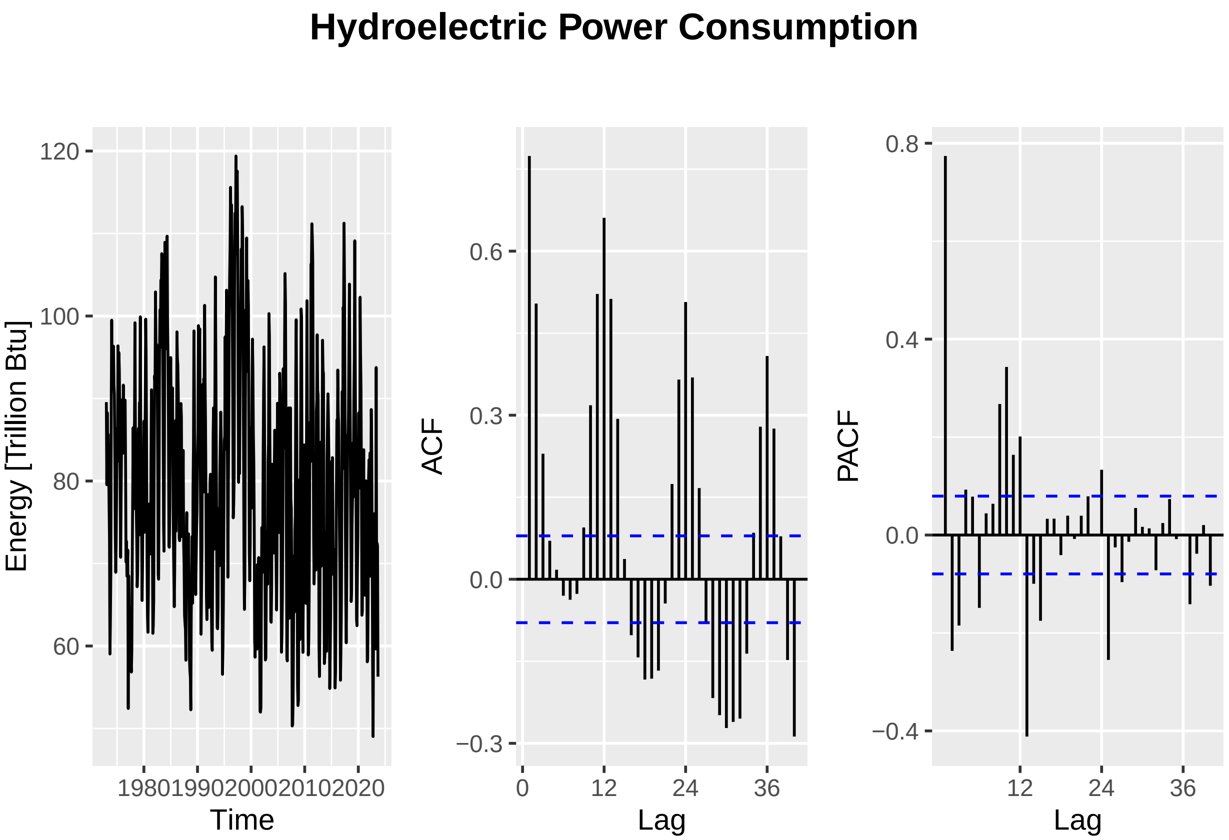
Q1

For each time series, i.e., Renewable Energy Production and Hydroelectric Consumption create three plots: one with time series, one with the ACF and with the PACF. You may use the some code form A2, but I want all the three plots side by side as in a grid. (Hint: use function `plot_grid()` from the `cowplot` package)

```
#Renewable Energy Production
ts_plot <- autoplot(tsdata[,1])+
            ylab("Energy [Trillion Btu]")+
            ggtitle("")

acf_plot <- autoplot(Acf(tsdata[,1],lag.max=40,plot=FALSE))+
  ggtitle("")

pacf_plot <- autoplot(Pacf(tsdata[,1],lag.max=40,plot=FALSE))+
  ggtitle("")

#Adding title
plot_row <- plot_grid(ts_plot,acf_plot,pacf_plot,nrow=1,ncol=3)
title <- ggdraw() + draw_label("Renewable Energy Consumption", fontface='bold')

plot_grid(title,plot_row,nrow=2,ncol=1,rel_heights = c(0.1,1))
```



```r
#Hydroelectric Power Consumption

ts_plot <- autoplot(tsdata[,2])+
            ylab("Energy [Trillion Btu]")+
            ggtitle("")

acf_plot <- autoplot(Acf(tsdata[,2],lag.max=40,plot=FALSE))+
  ggtitle("")

pacf_plot <- autoplot(Pacf(tsdata[,2],lag.max=40,plot=FALSE))+
  ggtitle("")

#Adding title
plot_row <- plot_grid(ts_plot,acf_plot,pacf_plot,nrow=1,ncol=3)
title <- ggdraw() + draw_label("Hydroelectric Power Consumption", fontface='bold')

plot_grid(title,plot_row,nrow=2,ncol=1,rel_heights = c(0.1,1))
```



Q2

From the plots in Q1, do the series Total Renewable Energy Production and Hydroelectric Power Consumption appear to have a trend? If yes, what kind of trend?

> Answer: For Renewable Energy Production, from the time series plot we can see an increasing trend. The ACF shows a slow decay on correlation coefficients meaning strong time dependence, i.e., current observation highly dpeendent on previous observations of the series.

> Answer: For the Hydroelectric Consumption it is hard to identify any trends from the time series plot. The ACF shows a slow decay, i.e., even higher lags correlation coefficients still significant. And it has a wave like pattern indicating we may have a seasonal component.

Q3

Use the *lm()* function to fit a linear trend to the two time series. Ask R to print the summary of the regression. Interpret the regression output, i.e., slope and intercept. Save the regression coefficients for further analysis.

Regression results for total renewable energy production:

```
regmodel_renewable=lm(tsdata[,1]~t,cbind(tsdata[,1],t))
beta0_renewable=regmodel_renewable$coefficients[1]
beta1_renewable=regmodel_renewable$coefficients[2]
print(summary(regmodel_renewable))
```

```
##
## Call:
## lm(formula = tsdata[, 1] ~ t, data = cbind(tsdata[, 1], t))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.27  -35.63   11.58   41.51  144.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.98940    4.90151   36.92   <2e-16 ***
## t             0.70404    0.01392   50.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.41 on 607 degrees of freedom
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.8078
## F-statistic:  2557 on 1 and 607 DF,  p-value: < 2.2e-16
```

> Answer: Regression results show a strong increasing linear trend with a slope of 0.704 as we identified in Q2.

Regression results for hydroelectric power consumption:

```
regmodel_hydroelec=lm(tsdata[,2]~t,cbind(tsdata[,2],t))
beta0_hydroelec=regmodel_hydroelec$coefficients[1]
beta1_hydroelec=regmodel_hydroelec$coefficients[2]
print(summary(regmodel_hydroelec))
```

```
##
## Call:
## lm(formula = tsdata[, 2] ~ t, data = cbind(tsdata[, 2], t))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.818 -10.620  -0.669   9.357  39.528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 82.734747   1.140265  72.557  < 2e-16 ***
## t           -0.009849   0.003239  -3.041  0.00246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 607 degrees of freedom
## Multiple R-squared:  0.015,  Adjusted R-squared:  0.01338
## F-statistic: 9.247 on 1 and 607 DF,  p-value: 0.002461
```

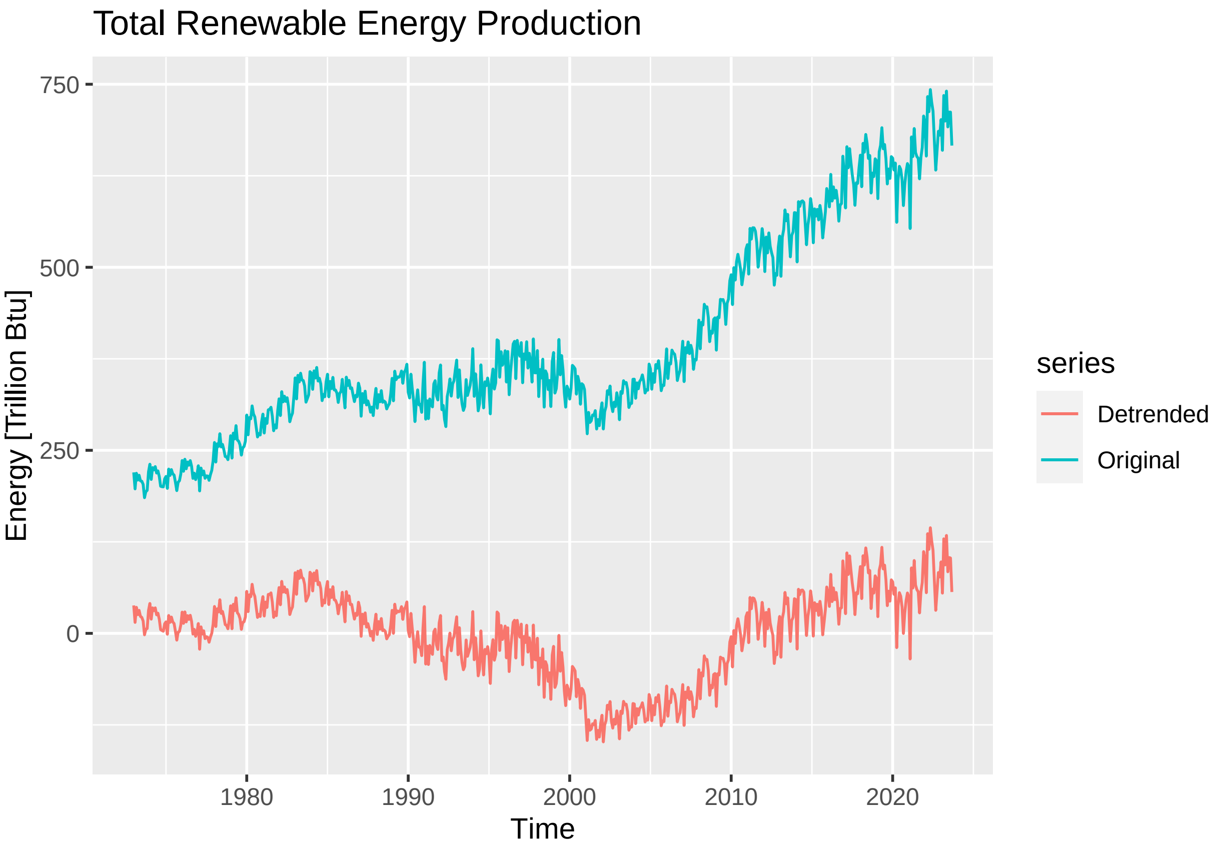
Answer: Regression results show a slight decreasing linear trend with a slope of -0.00985.

Q4

Use the regression coefficients from Q3 to detrend the series. Plot the detrended series and compare with the plots from Q1. What happened? Did anything change?

Total Renewable Energy Production

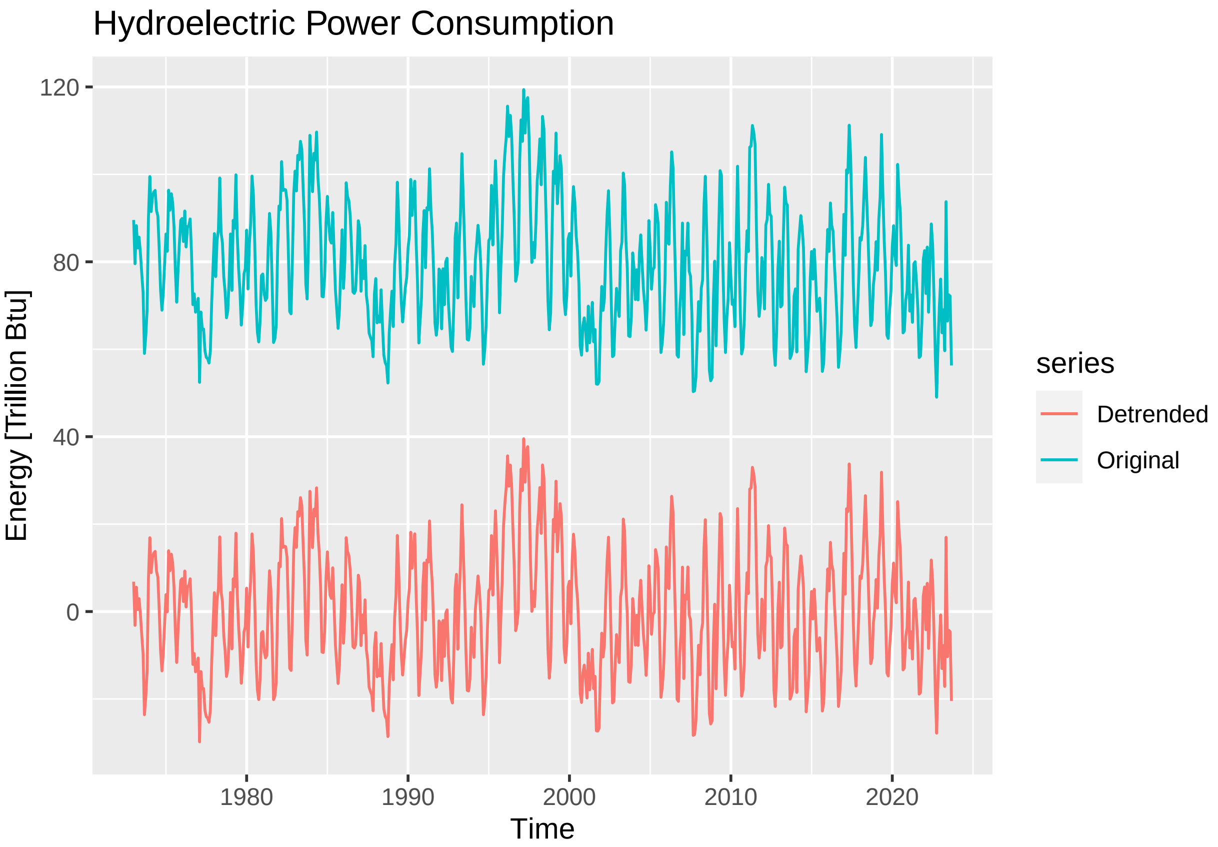
```r
renewable_detrend <- tsdata[,1] - (beta0_renewable+beta1_renewable*t)

renewable_detrend=ts(renewable_detrend, frequency=12,start=c(1973,1))

#the plot from part (d)
autoplot(tsdata[,1],series="Original") +
      autolayer(renewable_detrend,series="Detrended") +
      ylab("Energy [Trillion Btu]") +
      ggtitle("Total Renewable Energy Production")
```



Answer: The detrended series still seems to be trending over time, i.e., the linear model was not able to completely remove the trend.

Hydroelectric Power Consumption

```
hydroelec_detrend <- array(0,nobs)
for(i in t){
  hydroelec_detrend[i]=tsdata[i,2]-(beta0_hydroelec+beta1_hydroelec*i)
}
hydroelec_detrend=ts(hydroelec_detrend,frequency=12,start=c(1973,1))

#the plot from part (d)
autoplot(tsdata[,2],series="Original") +
      autolayer(hydroelec_detrend,series="Detrended") +
      ylab("Energy [Trillion Btu]") +
      ggtitle("Hydroelectric Power Consumption")
```

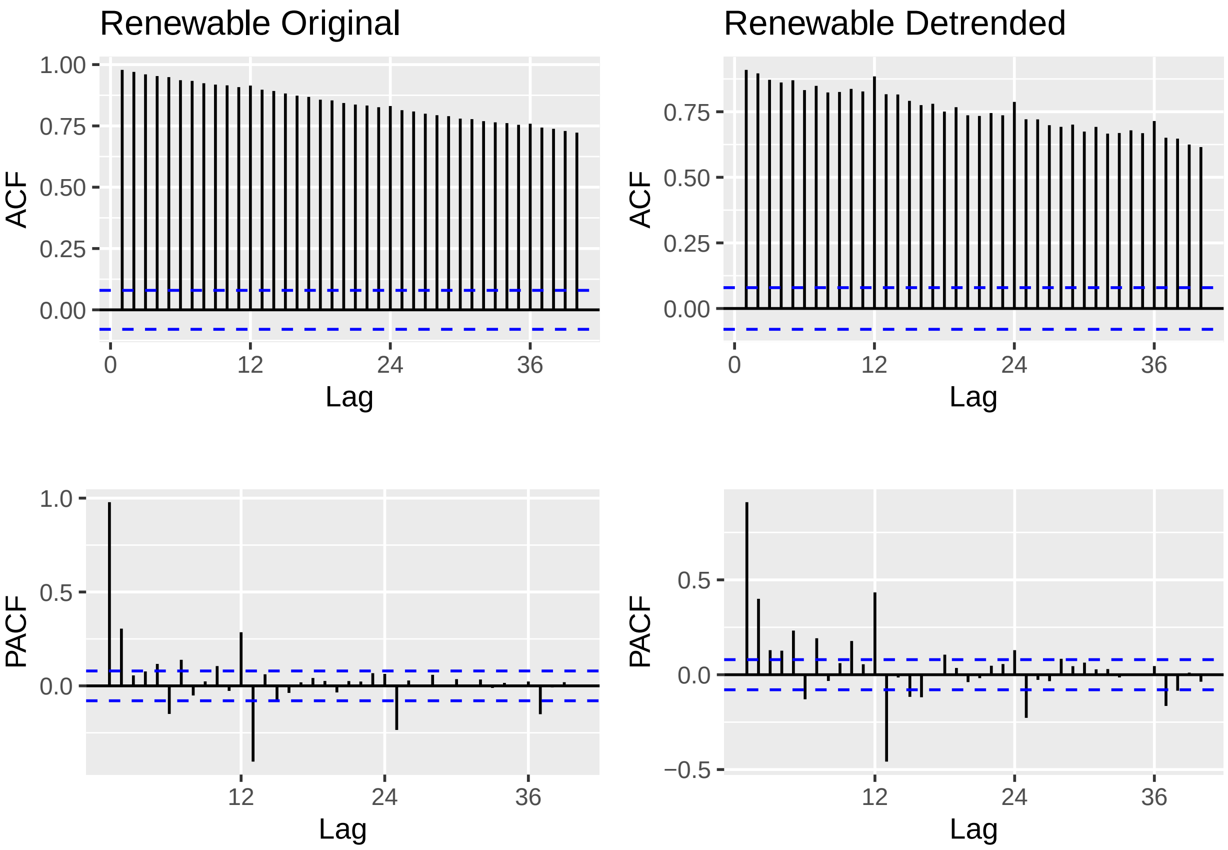


Answer: The detrended series does not seem to have a trend.

Q5

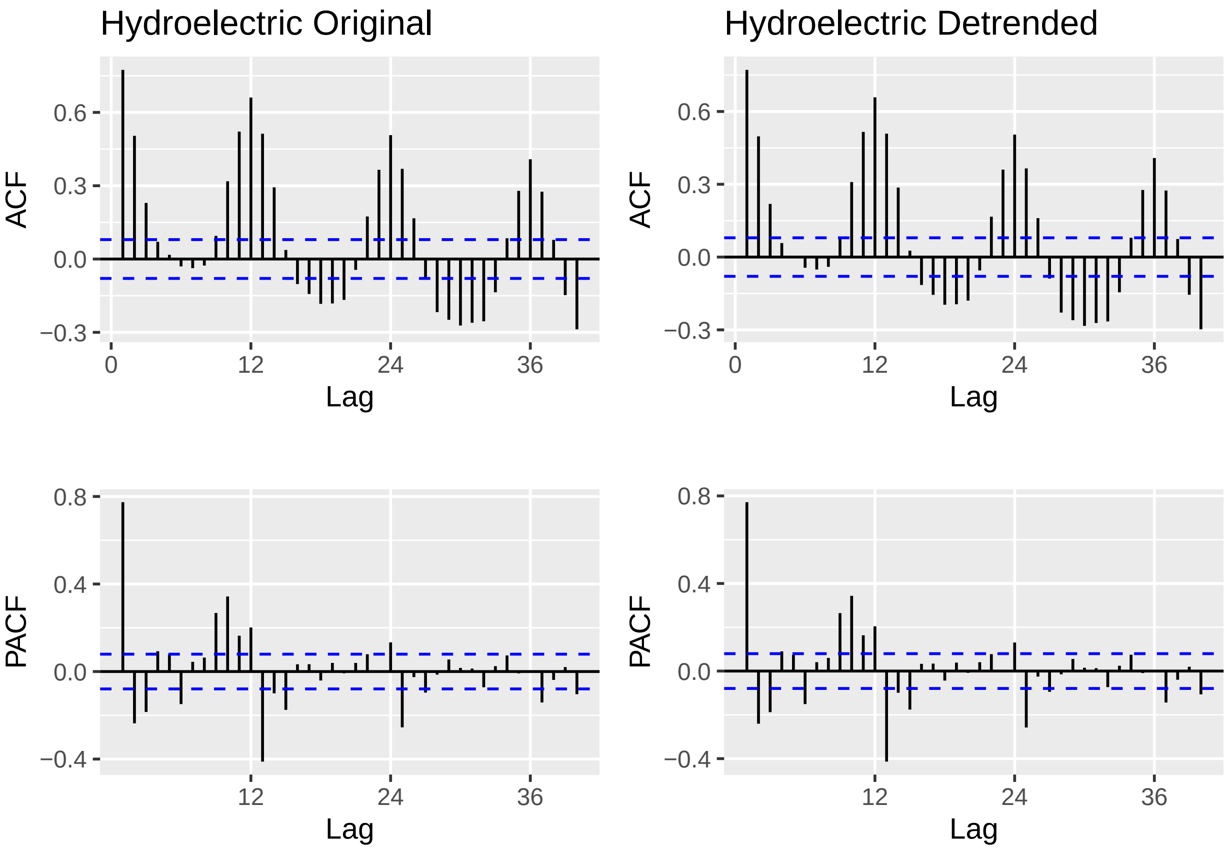
Plot ACF and PACF for the detrended series and compare with the plots from Q1. Did the plots change? How?

```
#lag.max: the maximum lag at which to calculate

#Comparing ACF and PACF for renewable production
plot_grid(
  autoplot(Acf(tsdata[,1],lag.max=40,plot=FALSE),main="Renewable Original"), #ylim=c(-0.5,1)
  autoplot(Acf(renewable_detrend,lag.max=40,plot=FALSE),main="Renewable Detrended"),
  autoplot(Pacf(tsdata[,1],lag.max=40,plot=FALSE),main=" ",),
  autoplot(Pacf(renewable_detrend,lag.max=40,plot=FALSE),main=" "),
  nrow=2,ncol=2
)
```



Answer: Looking at the ACF plots, the correlation coefficients seems to have decreased - look at the y axis limits. Also the detrended series ACF seems to be peaking at the seasonal lags, so after detrending the seasonal component is more well defined.

```
#Comparing ACF and PACF for hydroelectric consumption

plot_grid(
  autoplot(Acf(tsdata[,2],lag.max=40,plot=FALSE),main="Hydroelectric Original"),
  autoplot(Acf(hydroelec_detrend,lag.max=40,plot=FALSE),main="Hydroelectric Detrended"),
  autoplot(Pacf(tsdata[,2],lag.max=40,plot=FALSE),main=" "),
  autoplot(Pacf(hydroelec_detrend,lag.max=40,plot=FALSE),main=" "),
  nrow=2,ncol=2
)
```



Answer: Not much changed since there is only a slight decreasing trend.

Seasonal Component

Set aside the detrended series and consider the original series again from Q1 to answer Q6 to Q8.

Q6

Just by looking at the time series and the acf plots, do the series seem to have a seasonal trend? No need to run any code to answer your question. Just type in you answer below.

> Answer: The hydroelectric series seems to have a seasonal component from the ACF plot we can see wave-like patterns. But for renewable production it's not clear from the plots if there is seasonality.

Q7

Use function *lm()* to fit a seasonal means model (i.e. using the seasonal dummies) the two time series. Ask R to print the summary of the regression. Interpret the regression output. From the results which series have a seasonal trend? Do the results match you answer to Q6?

```r
# Renewable Production
dummies=seasonaldummy(tsdata[,1])   #This function only
#works if Y is a ts object and if you specify the frequency | precondition
```

```
reg_dummies_1=lm(tsdata[,1]~dummies)
print(summary(reg_dummies_1))
```

```
##
## Call:
## lm(formula = tsdata[, 1] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -199.19  -86.35  -48.84  113.18  331.58
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  404.526     19.574  20.666   <2e-16 ***
## dummiesJan     2.962     27.546   0.108    0.914
## dummiesFeb   -34.476     27.546  -1.252    0.211
## dummiesMar     3.929     27.546   0.143    0.887
## dummiesApr    -8.695     27.546  -0.316    0.752
## dummiesMay     6.645     27.546   0.241    0.809
## dummiesJun    -4.198     27.546  -0.152    0.879
## dummiesJul     2.460     27.546   0.089    0.929
## dummiesAug    -5.026     27.546  -0.182    0.855
## dummiesSep   -29.119     27.546  -1.057    0.291
## dummiesOct   -20.068     27.682  -0.725    0.469
## dummiesNov   -20.346     27.682  -0.735    0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.4 on 597 degrees of freedom
## Multiple R-squared:  0.009296,   Adjusted R-squared:  -0.008958
## F-statistic: 0.5093 on 11 and 597 DF,  p-value: 0.8976
```

```
#Store the regression coefficients
beta_int_1=reg_dummies_1$coefficients[1]
beta_coeff_1=reg_dummies_1$coefficients[2:12]
```

> Answer: The renewable production regression results show that there is not enough evidence to state there is a seasonal component in the original series. But keep in mind that this results might just be saying a seasonal means model is not a goog representtaion of the seasonal component.

What if we run the regression on the detrended series?

```
# Renewable Production
dummies=seasonaldummy(renewable_detrend)   #This function only
#works if Y is a ts object and if you specify the frequency | precondition
reg_dummies_test=lm(renewable_detrend~dummies)
print(summary(reg_dummies_test))
```

```
##
## Call:
## lm(formula = renewable_detrend ~ dummies)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -146.18  -37.88   14.16   42.18  128.82 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)    8.101      8.397   0.965  0.33506   
## dummiesJan     6.482     11.816   0.549  0.58350   
## dummiesFeb   -31.660     11.816  -2.679  0.00758 **
## dummiesMar     6.041     11.816   0.511  0.60937   
## dummiesApr    -7.287     11.816  -0.617  0.53766   
## dummiesMay     7.349     11.816   0.622  0.53422   
## dummiesJun    -4.198     11.816  -0.355  0.72253   
## dummiesJul     1.755     11.816   0.149  0.88195   
## dummiesAug    -6.434     11.816  -0.544  0.58630   
## dummiesSep   -31.231     11.816  -2.643  0.00843 **
## dummiesOct   -18.660     11.875  -1.571  0.11662   
## dummiesNov   -19.642     11.875  -1.654  0.09864 . 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 59.37 on 597 degrees of freedom
## Multiple R-squared:  0.04977,    Adjusted R-squared:  0.03226 
## F-statistic: 2.843 on 11 and 597 DF,  p-value: 0.001226
```

Answer: Still very similar results.

```r
# Hydroelectric Consumption
dummies=seasonaldummy(tsdata[,2])
reg_dummies_2=lm(tsdata[,2]~dummies)
print(summary(reg_dummies_2))
```

```
## 
## Call:
## lm(formula = tsdata[, 2] ~ dummies)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -31.323  -5.849  -0.468   6.243  32.290 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   80.282      1.470  54.601  < 2e-16 ***
## dummiesJan     4.807      2.069   2.323  0.02050 *  
## dummiesFeb    -2.725      2.069  -1.317  0.18831    
## dummiesMar     6.825      2.069   3.298  0.00103 ** 
## dummiesApr     5.319      2.069   2.571  0.01039 *  
## dummiesMay    13.922      2.069   6.729 4.02e-11 ***
## dummiesJun    10.650      2.069   5.147 3.60e-07 ***
## dummiesJul     3.912      2.069   1.891  0.05914 .  
## dummiesAug    -5.677      2.069  -2.744  0.00626 ** 
## dummiesSep   -16.797      2.069  -8.118 2.72e-15 ***
## dummiesOct   -16.468      2.079  -7.920 1.17e-14 ***
## dummiesNov   -10.885      2.079  -5.235 2.29e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.4 on 597 degrees of freedom
## Multiple R-squared:  0.4697, Adjusted R-squared:  0.4599
## F-statistic: 48.07 on 11 and 597 DF,  p-value: < 2.2e-16
```

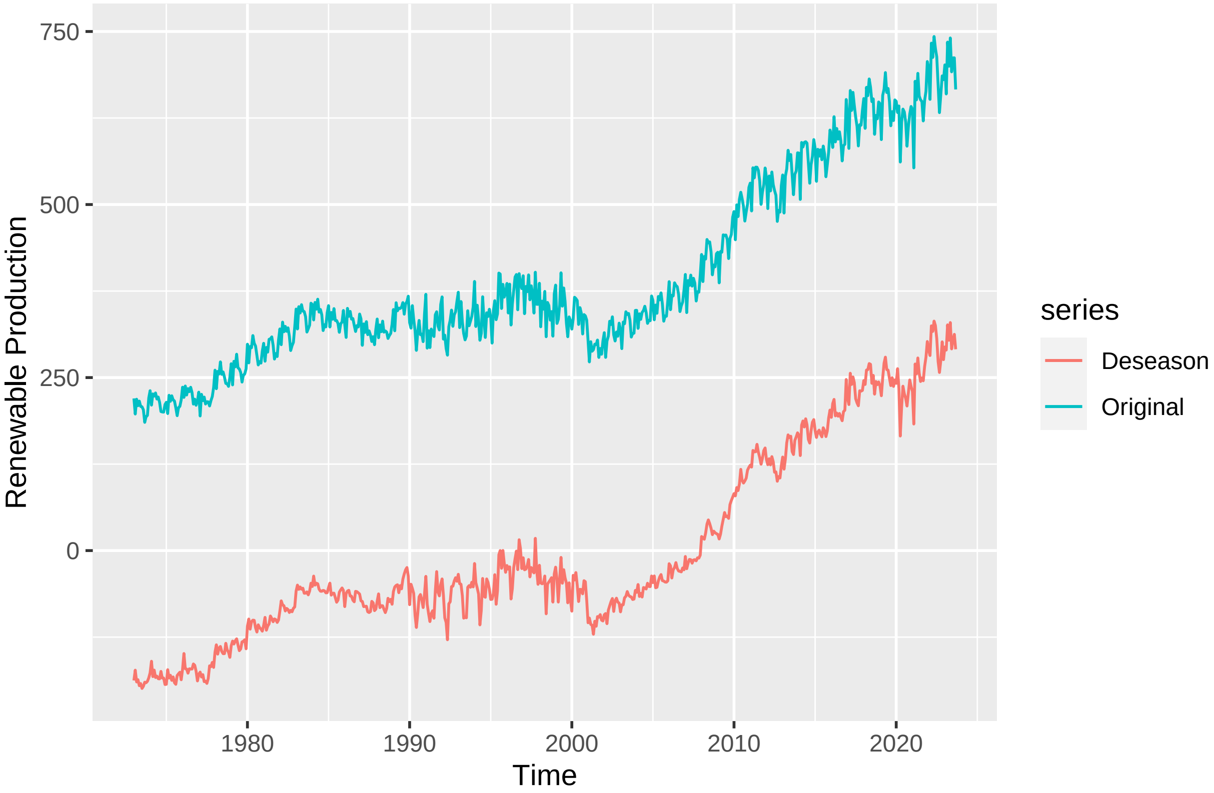
```r
#Store the regression coefficients
beta_int_2=reg_dummies_2$coefficients[1]
beta_coeff_2=reg_dummies_2$coefficients[2:12]
```

> Answer: For hydroelectric consumption we observe the regression is showing significant results and seasonality explains 45.99% of the series variability.

Q8

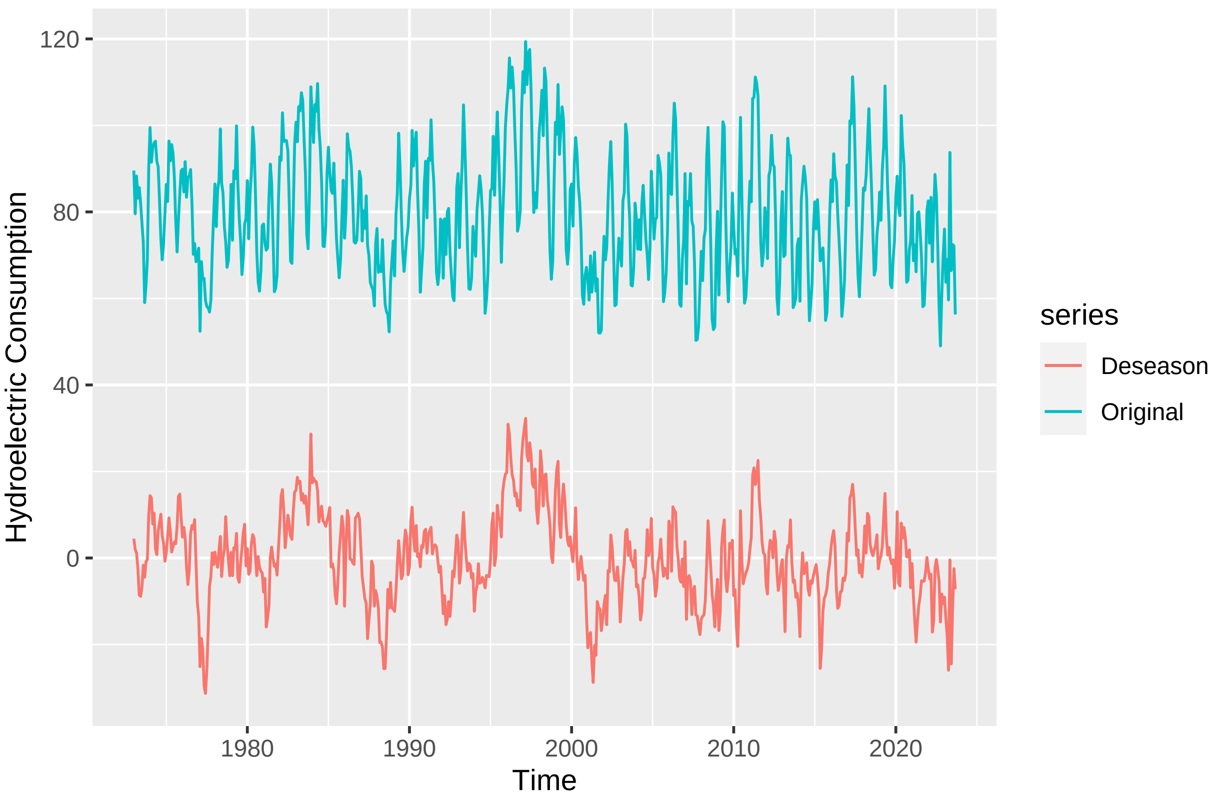
Use the regression coefficients from Q7 to deseason the series. Plot the deseason series and compare with the plots from part Q1. Did anything change?

```r
# Renewable Production
renew_deseason=array(0,nobs)
for(i in 1:nobs){
  renew_deseason[i]=tsdata[i,1]-(beta_int_1+beta_coeff_1%*%dummies[i,])
  #The symbol %*% means inner product
}
renew_deseason=ts(renew_deseason,frequency=12,start=c(1973,1))

autoplot(tsdata[,1],series="Original",ylab="Renewable Production")+
  autolayer(renew_deseason,series="Deseason") #+
```



Answer: You do not need to deseason renewable production, but if you did you should get this results. Not much has changed from the original series.

```
# Hydroelectric consumption
hydro_deseason=array(0,nobs)
for(i in 1:nobs){
  hydro_deseason[i]=tsdata[i,2]-(beta_int_2+beta_coeff_2%*%dummies[i,])
  #The symbol %*% means inner product
}
hydro_deseason=ts(hydro_deseason,frequency=12,start=c(1973,1))

autoplot(tsdata[,2],series="Original",ylab="Hydroelectric Consumption")+
  autolayer(hydro_deseason,series="Deseason")
```

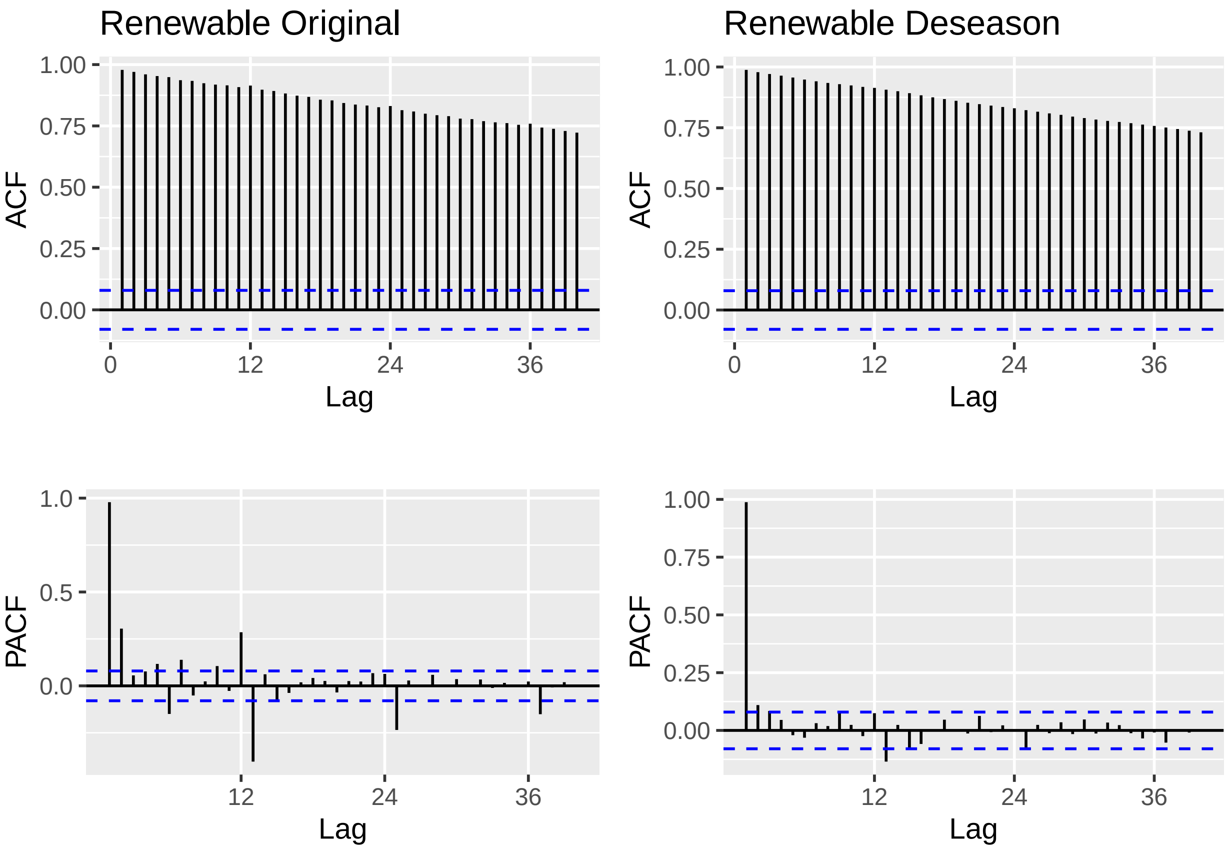


Answer: For teh hydroelectric consumption series we seems to have eliminated the equally spaced wave pattern and have more random movements.

Q9

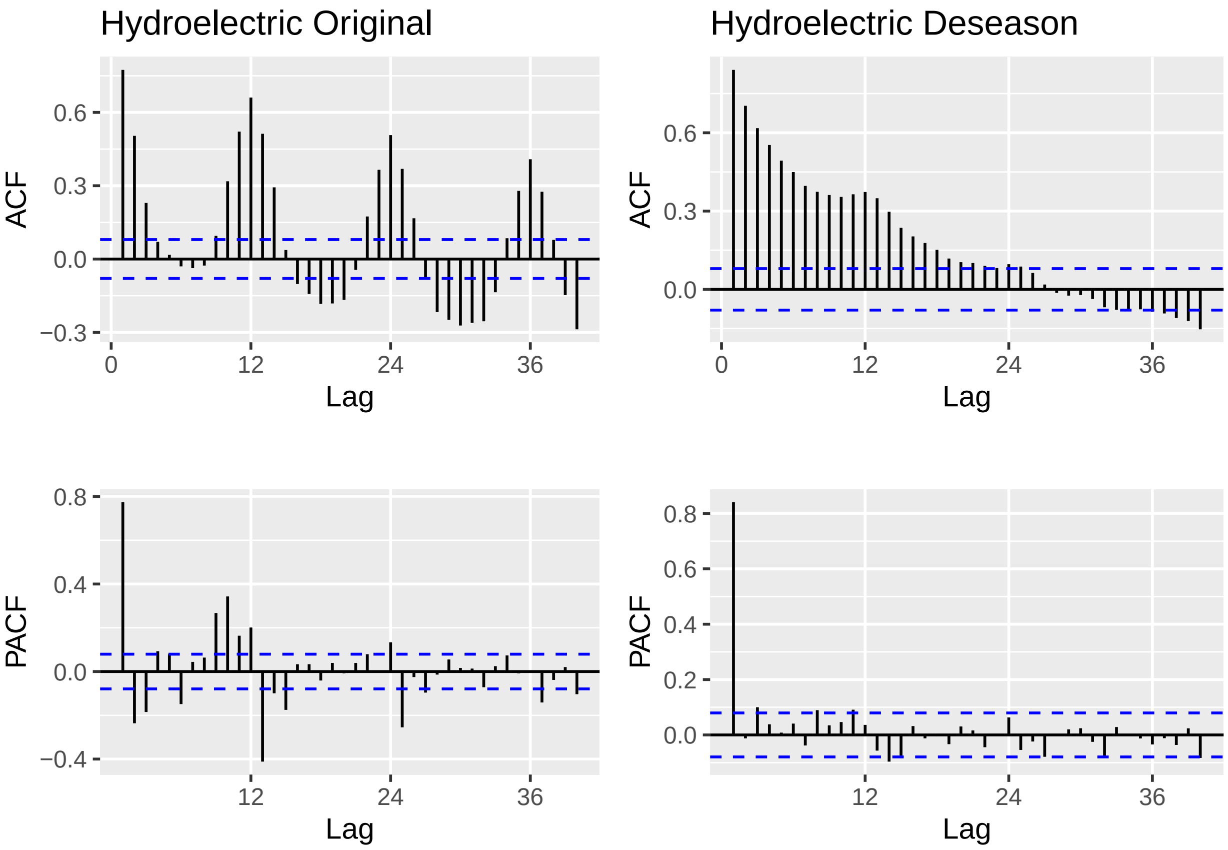
Plot ACF and PACF for the deseason series and compare with the plots from Q1. Did the plots change? How?

```
plot_grid(
  autoplot(Acf(tsdata[,1],lag.max=40,plot=FALSE),main="Renewable Original"),
  autoplot(Acf(renew_deseason,lag.max=40,plot=FALSE),main="Renewable Deseason"),
  autoplot(Pacf(tsdata[,1],lag.max=40,plot=FALSE),main=" "),
  autoplot(Pacf(renew_deseason,lag.max=40,plot=FALSE),main=" "),
  nrow=2,ncol=2
)
```



Answer: For renewable production not much has changed in the ACF since we did not have a significant seasonal component. But the PACF does seem to have decreased the correlation at seasonal lags 12,24 and 36.

```
plot_grid(
  autoplot(Acf(tsdata[,2],lag.max=40,plot=FALSE),main="Hydroelectric Original"),
  autoplot(Acf(hydro_deseason,lag.max=40,plot=FALSE),main="Hydroelectric Deseason"),
  autoplot(Pacf(tsdata[,2],lag.max=40,plot=FALSE),main=" "),
  autoplot(Pacf(hydro_deseason,lag.max=40,plot=FALSE),main=" "),
  nrow=2,ncol=2
)
```



Answer: For the hydroelectric consumption we see significant changes. Not the wave-like pattern on teh ACF is gone after deseasoning the series and also the correlation coefficients now stop bein significant after a few lags. The PACf also shows significant cganfes, no more significant lags at 12, 24 and 36.