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

Assignment 3 - Due date 02/01/24

Sai Powar

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. 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 December 2022 **Monthly** Energy Review. Once again you will work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series only.

R packages needed for this assignment:"forecast","tseries", and "Kendall". 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.\

```
#Load/install required package here
library(forecast)
```

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

```
library(tseries)
library(Kendall)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
library(ggplot2)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(ggfortify)
```

```
## Registered S3 methods overwritten by 'ggfortify':
##   method                 from
##   autoplot.Arima         forecast
##   autoplot.acf           forecast
##   autoplot.ar            forecast
##   autoplot.bats          forecast
##   autoplot.decomposed.ts forecast
##   autoplot.ets           forecast
##   autoplot.forecast      forecast
##   autoplot.stl           forecast
##   autoplot.ts            forecast
##   fitted.ar              forecast
##   fortify.ts             forecast
##   residuals.ar           forecast
```

```
#importing dataset
raw_energy_data <- read.table(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_So

#date
energy_date <- ym(raw_energy_data[,1])  #function my from package lubridate
head(energy_date)
```

```
## [1] "1973-01-01" "1973-02-01" "1973-03-01" "1973-04-01" "1973-05-01"
## [6] "1973-06-01"
```

```
energy_data <- cbind(energy_date,raw_energy_data[,(4:6)])
head(energy_data)
```

```
##   energy_date Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1  1973-01-01                         129.787                           219.839
## 2  1973-02-01                         117.338                           197.330
```

```
## 3  1973-03-01                          129.938                              218.686
## 4  1973-04-01                          125.636                              209.330
## 5  1973-05-01                          129.834                              215.982
## 6  1973-06-01                          125.611                              208.249
##   Hydroelectric.Power.Consumption
## 1                          89.562
## 2                          79.544
## 3                          88.284
## 4                          83.152
## 5                          85.643
## 6                          82.060
```

```r
#creating time series object
ts_energy_data <- ts(energy_data[2:4],start = c(1973,1),frequency=12)
head(ts_energy_data)
```

```
##          Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## Jan 1973                         129.787                           219.839
## Feb 1973                         117.338                           197.330
## Mar 1973                         129.938                           218.686
## Apr 1973                         125.636                           209.330
## May 1973                         129.834                           215.982
## Jun 1973                         125.611                           208.249
##          Hydroelectric.Power.Consumption
## Jan 1973                          89.562
## Feb 1973                          79.544
## Mar 1973                          88.284
## Apr 1973                          83.152
## May 1973                          85.643
## Jun 1973                          82.060
```

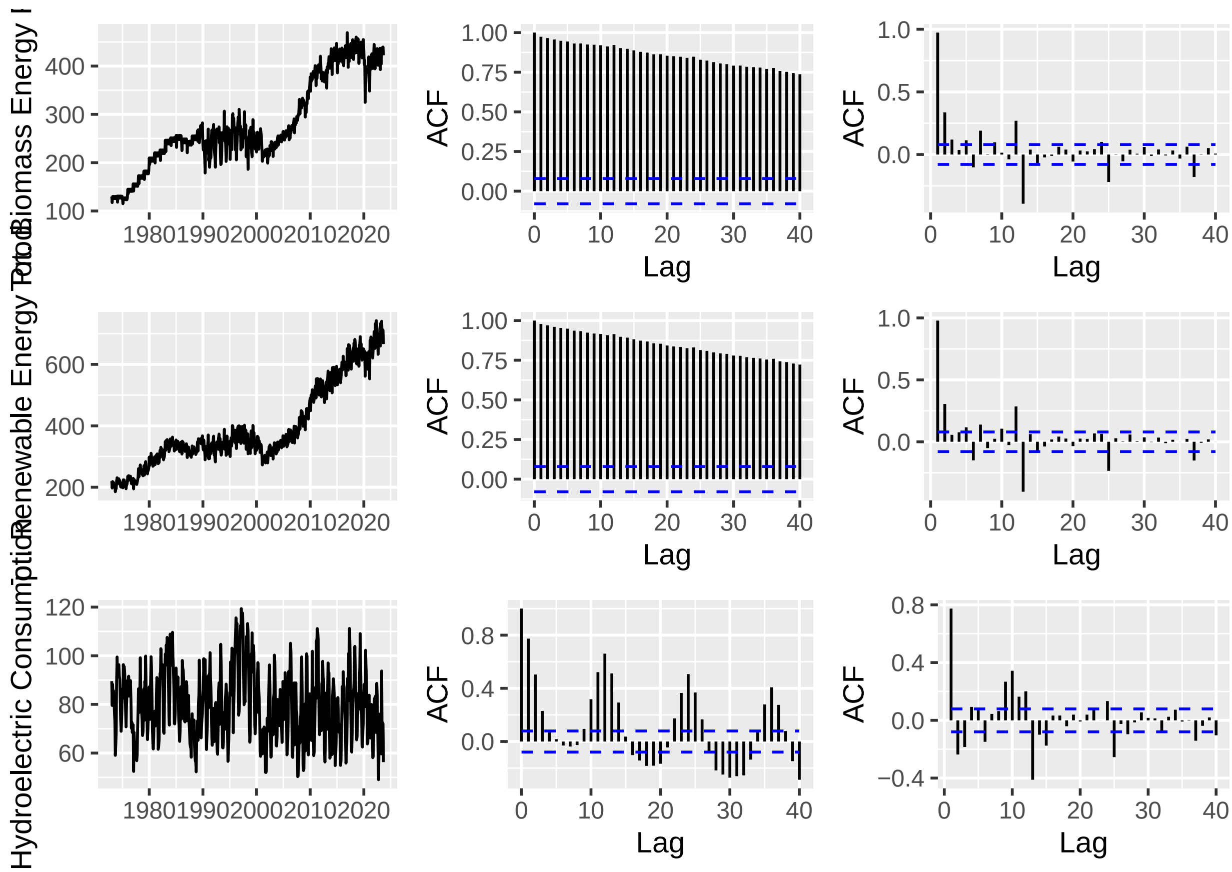
##Trend Component

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)

```r
#creating objects for rows and columns
ncol_energy <- ncol(energy_data)-1
nobs_energy <- nrow(energy_data)

p <- plot_grid(
  autoplot(ts_energy_data[,1],ylab="Tot. Biomass Energy Prod."),
  autoplot(Acf(ts_energy_data[,1],lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_energy_data[,1],lag.max=40,plot=FALSE),main=NULL),
  autoplot(ts_energy_data[,2],ylab="Tot. Renewable Energy Prod."),
  autoplot(Acf(ts_energy_data[,2],lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_energy_data[,2],lag.max=40,plot=FALSE),main=NULL),
  autoplot(ts_energy_data[,3],ylab="Hydroelectric Consumption"),
  autoplot(Acf(ts_energy_data[,3],lag.max=40,plot=FALSE),main=NULL),
```

```
  autoplot(Pacf(ts_energy_data[,3],lag.max=40,plot=FALSE),main=NULL),
  nrow=3,ncol=3
)
p
```



Q2

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

Total Biomass Energy Production and Total Renewable Energy Production appear to have an overall linear trend. However, I do see some possibility of seasonality in the 1990-2000 decade that could be worth exploring. Additionally, there is a more "clear" linear trend from 2000 onwards, compared the years before that.

Hydroelectric Power Consumption series appears to have a seasonal trend. There could be a slight, downward linear trend but we would need to explore the data series more to see if the effect is pronounced.

Q3

Use the *lm()* function to fit a linear trend to the three 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.

```r
#Create vector t
t <- c(1:nobs_energy)

#Fiting a linear trend to TS of the time series
biomass_linear = lm(energy_data[,2]~t)
summary(biomass_linear)
```

```
##
## Call:
## lm(formula = energy_data[, 2] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.344  -23.754    5.491   31.980   83.154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.27841    3.18601   42.15   <2e-16 ***
## t             0.47713    0.00905   52.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.26 on 607 degrees of freedom
## Multiple R-squared:  0.8208, Adjusted R-squared:  0.8205
## F-statistic:  2780 on 1 and 607 DF,  p-value: < 2.2e-16
```

```r
bio_beta0=as.numeric(biomass_linear$coefficients[1])
bio_beta1=as.numeric(biomass_linear$coefficients[2])

re_linear = lm(energy_data[,3]~t)
summary(re_linear)
```

```
##
## Call:
## lm(formula = energy_data[, 3] ~ 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
```

```r
re_beta0=as.numeric(re_linear$coefficients[1])
re_beta1=as.numeric(re_linear$coefficients[2])

hydro_linear = lm(energy_data[,4]~t)
summary(hydro_linear)
```

```
##
## Call:
## lm(formula = energy_data[, 4] ~ 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
```

```r
hydro_beta0=as.numeric(hydro_linear$coefficients[1])
hydro_beta1=as.numeric(hydro_linear$coefficients[2])
```

Biomass - For every unit increase in time (for every month), the biomass energy production increases by 0.477 trillion Btu. When time = 0 (at the beginning of the time series), the biomass energy production is 134.278 trillion Btu.

Renewable - For every unit increase in time (for every month), the renewable energy production increases by 0.704 trillion Btu. When time = 0 (at the beginning of the time series), the renewable energy production is 180.989 trillion Btu.

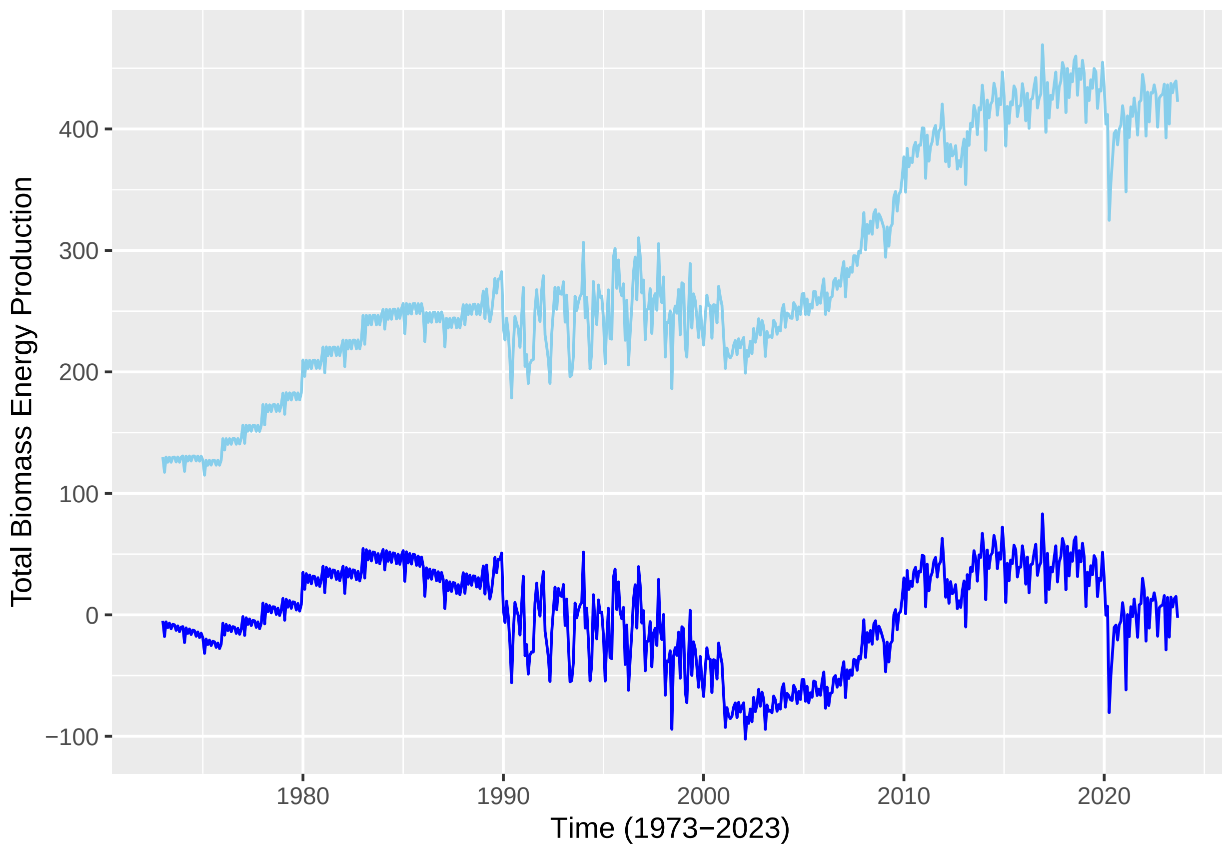
Hydroelectric - For every unit increase in time (for every month), the hydroelectric power consumption decreases by 0.00985 trillion Btu. When time = 0 (at the beginning of the time series), the hydroelectric power consumption is 82.735 trillion Btu.

Q4

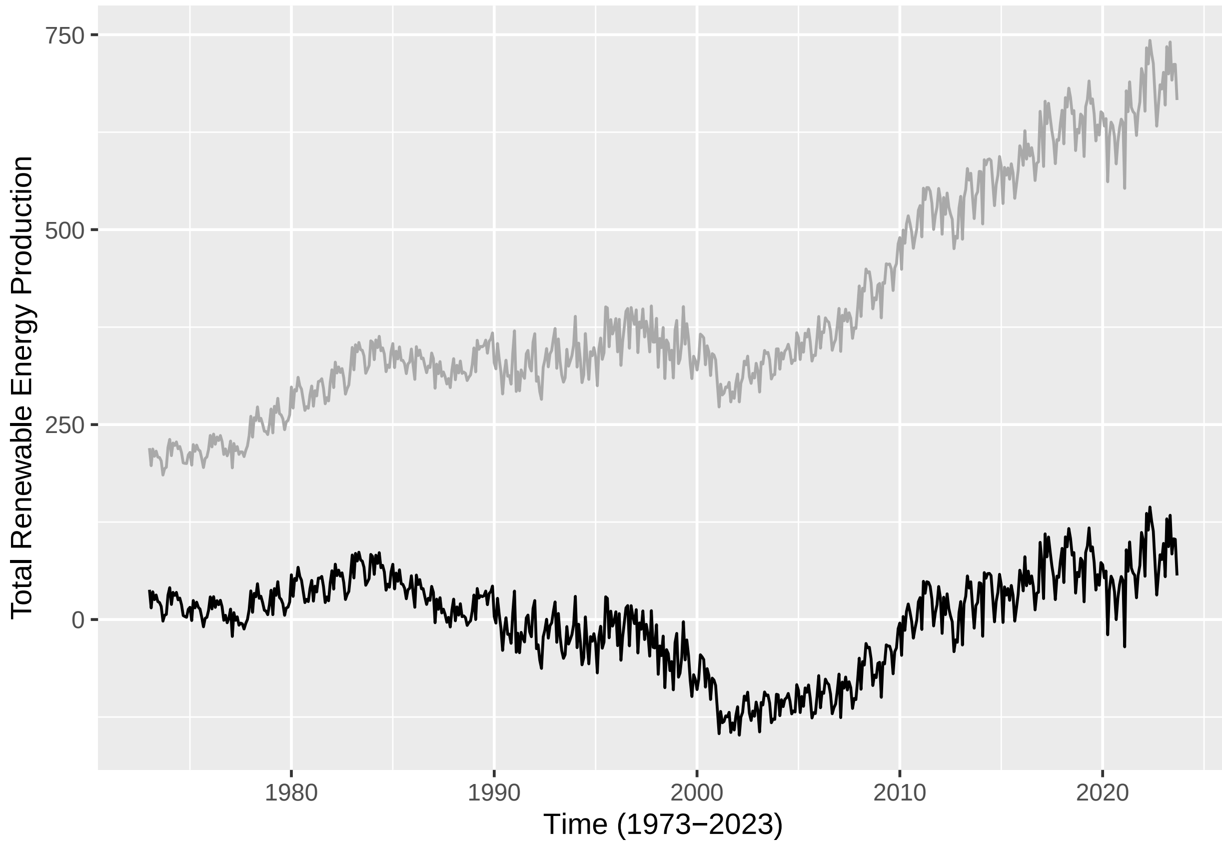
Use the regression coefficients from Q3 to detrend the series. Plot the detrended series and compare with the plots from Q1. What happened? Did anything change?

```r
#removing the trend from series, dt=detrended
dt_biomass_linear <- energy_data[,2]-(bio_beta0+bio_beta1*t)
dt_re_linear <- energy_data[,3]-(re_beta0+re_beta1*t)
dt_hydro_linear <- energy_data[,4]-(hydro_beta0+hydro_beta1*t)

#plotting the detrended series
p1 <- ggplot(energy_data, aes(x=energy_date, y=energy_data[,2])) +
            geom_line(color="skyblue") +
```

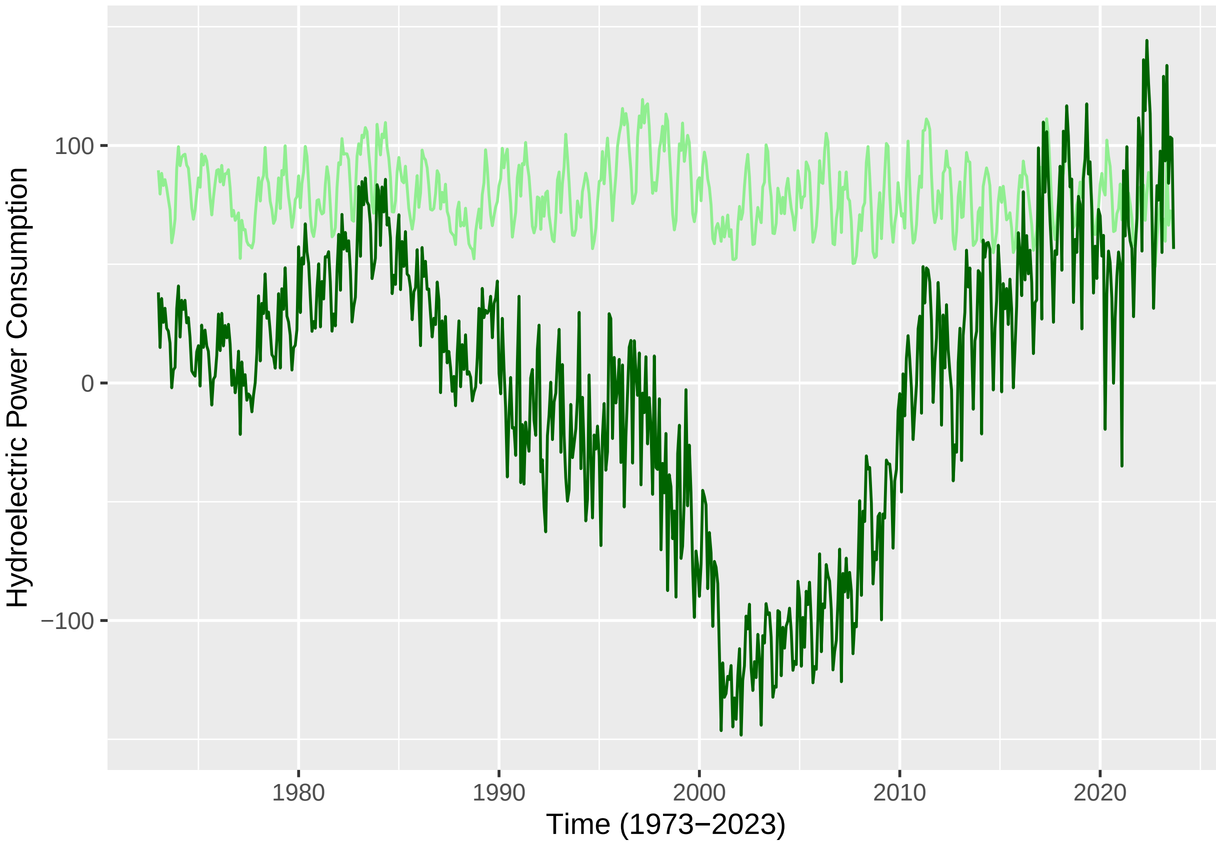
```
            ylab("Total Biomass Energy Production") +
            xlab("Time (1973-2023)")+
            geom_line(aes(y=dt_biomass_linear), col="blue")
p1
```




```
p2 <- ggplot(energy_data, aes(x=energy_date, y=energy_data[,3])) +
            geom_line(color="darkgrey") +
            ylab("Total Renewable Energy Production") +
            xlab("Time (1973-2023)")+
            geom_line(aes(y=dt_re_linear), col="black")
p2
```




```r
p3 <- ggplot(energy_data, aes(x=energy_date, y=energy_data[,4])) +
            geom_line(color="lightgreen") +
            ylab("Hydroelectric Power Consumption") +
            xlab("Time (1973-2023)")+
            geom_line(aes(y=dt_re_linear), col="darkgreen") #this needs to be dt_hydro_linear
p3
```



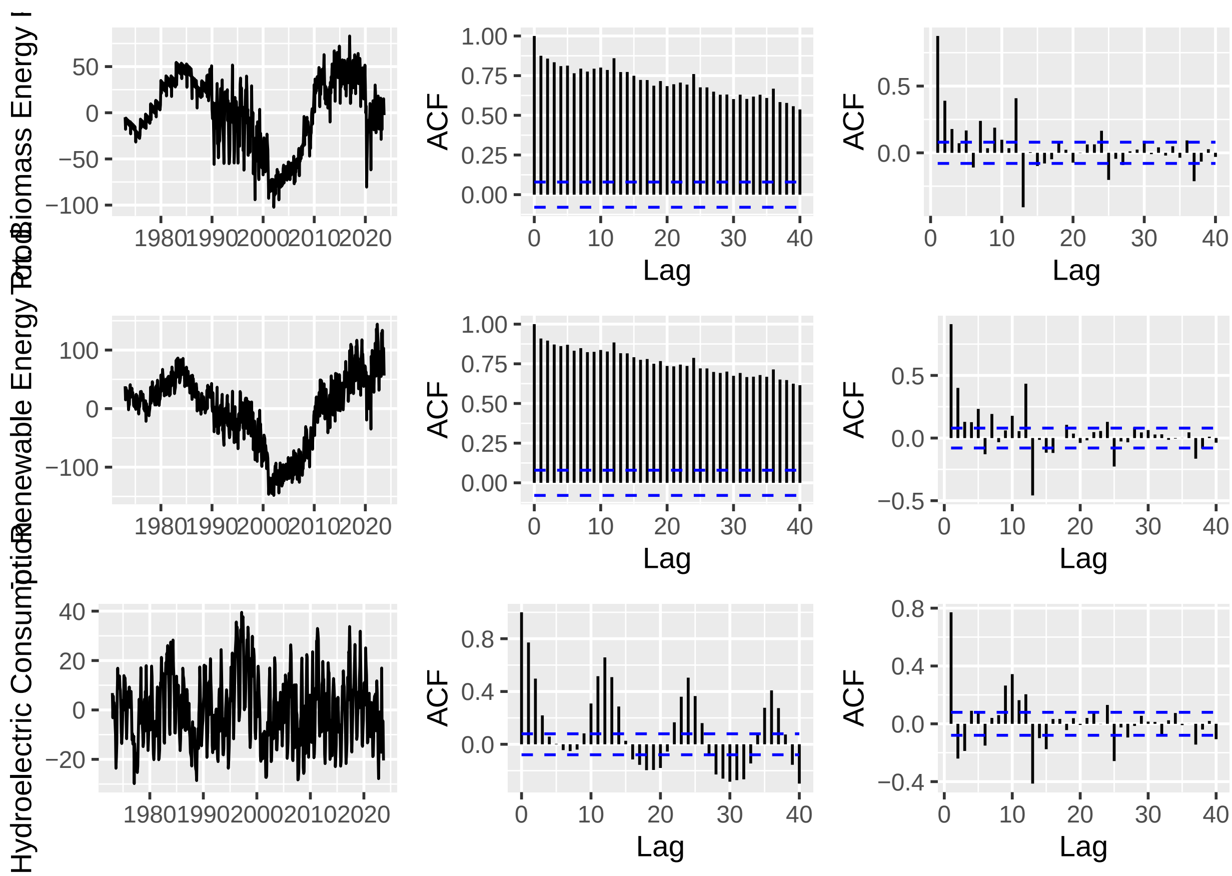
For the biomass and renewable energy production series, the detrended series do not have the upward trend from the original series. However, there is still some variability in the data that is not removed from the detrending. This corresponds to the R-squared value from the linear regression summary. For the hydroelectric power consumption, the detrended series looks odd in a way. The series is negative for some years and it is evident that detrending using the linear component does not really help in understanding the true trend component for this series.

Q5

Plot ACF and PACF for the detrended series and compare with the plots from Q1. You may use plot_grid() again to get them side by side. nut mot mandatory. Did the plots change? How?

```r
#creating time series object for detrended series
ts_dt_bio_lin <- ts(dt_biomass_linear,start=c(1973,1),frequency=12)
ts_dt_re_lin <- ts(dt_re_linear,start=c(1973,1),frequency=12)
ts_dt_hydro_lin <- ts(dt_hydro_linear,start=c(1973,1),frequency=12)


p5 <- plot_grid(
  autoplot(ts_dt_bio_lin,ylab="Tot. Biomass Energy Prod."),
  autoplot(Acf(ts_dt_bio_lin,lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_dt_bio_lin,lag.max=40,plot=FALSE),main=NULL),
  autoplot(ts_dt_re_lin,ylab="Tot. Renewable Energy Prod."),
  autoplot(Acf(ts_dt_re_lin,lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_dt_re_lin,lag.max=40,plot=FALSE),main=NULL),
  autoplot(ts_dt_hydro_lin,ylab="Hydroelectric Consumption"),
```

```
  autoplot(Acf(ts_dt_hydro_lin,lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_dt_hydro_lin,lag.max=40,plot=FALSE),main=NULL),
  nrow=3,ncol=3
)
p5
```



Biomass - For the detrended series the ACF values start falling below 0.75 after lag ~15, whereas this happened at lag 40 in the original series.

Renewable - For the detrended series the ACF values start falling below 0.75 after lag 20, whereas this happened at lag 40 in the original series.

Hydroelectric - There isn't a significant decrease in the ACF and PACF plots between the trended and detrended series.

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.

The hydroelectric power consumption series definitely has a seasonal component. There seems to be a seasonal component in the other two series as well, but in shorter time frames and not the entire series. But, it is not as prominent as it is in the hydroelectric series.

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
#Creating the seasonal dummies
dummies <- seasonaldummy(ts_energy_data[,2])

#Then fit a linear model to the seasonal dummies
bio_seas <- lm(energy_data[,2]~dummies)
summary(bio_seas)
```

```
##
## Call:
## lm(formula = energy_data[, 2] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -163.19  -55.46  -26.30   98.54  178.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 290.4666    13.1583  22.075   <2e-16 ***
## dummiesJan   -1.6748    18.5171  -0.090   0.9280
## dummiesFeb  -31.2863    18.5171  -1.690   0.0916 .
## dummiesMar   -8.8523    18.5171  -0.478   0.6328
## dummiesApr  -21.6024    18.5171  -1.167   0.2438
## dummiesMay  -13.9313    18.5171  -0.752   0.4521
## dummiesJun  -19.3220    18.5171  -1.043   0.2972
## dummiesJul   -3.5675    18.5171  -0.193   0.8473
## dummiesAug   -0.4953    18.5171  -0.027   0.9787
## dummiesSep  -13.1780    18.5171  -0.712   0.4770
## dummiesOct   -4.0129    18.6086  -0.216   0.8293
## dummiesNov   -9.6626    18.6086  -0.519   0.6038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.04 on 597 degrees of freedom
## Multiple R-squared:  0.01007,    Adjusted R-squared:  -0.008173
## F-statistic: 0.5519 on 11 and 597 DF,  p-value: 0.8676
```

```r
re_seas <- lm(energy_data[,3]~dummies)
summary(re_seas)
```

```
##
## Call:
## lm(formula = energy_data[, 3] ~ 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
```

```r
hydro_seas <- lm(energy_data[,4]~dummies)
summary(hydro_seas)
```

```
##
## Call:
## lm(formula = energy_data[, 4] ~ 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
##
```

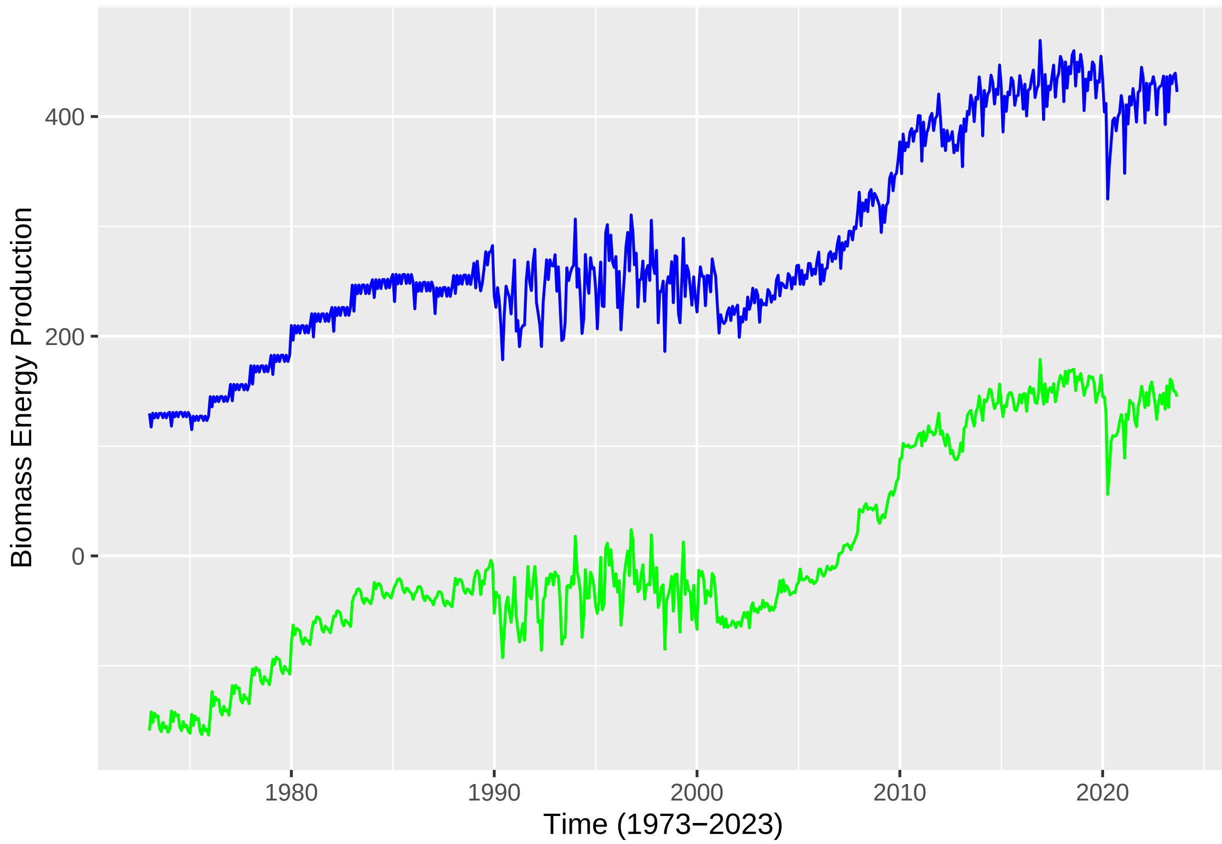
```
## Residual standard error: 10.4 on 597 degrees of freedom
## Multiple R-squared:  0.4697, Adjusted R-squared:  0.4599
## F-statistic: 48.07 on 11 and 597 DF,  p-value: < 2.2e-16
```

For the biomass and renewable energy series, the regression output shows that the series does not have a significant seasonal component. None of the coefficients are significant and the p-values are also high. For the hydroelectric time series, the regression output shows that the series has a seasonal component. All the coefficients are significant and the overall p-value is also low.

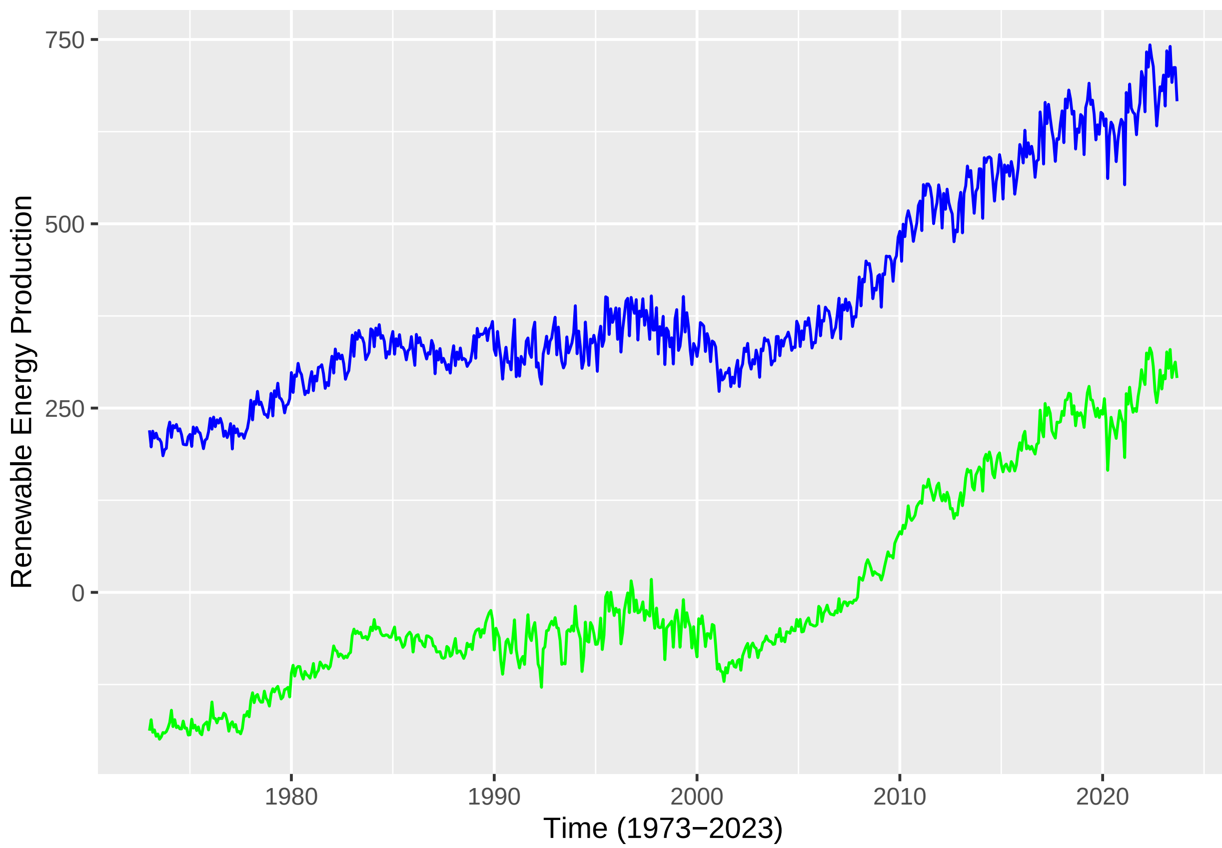
Q8

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?

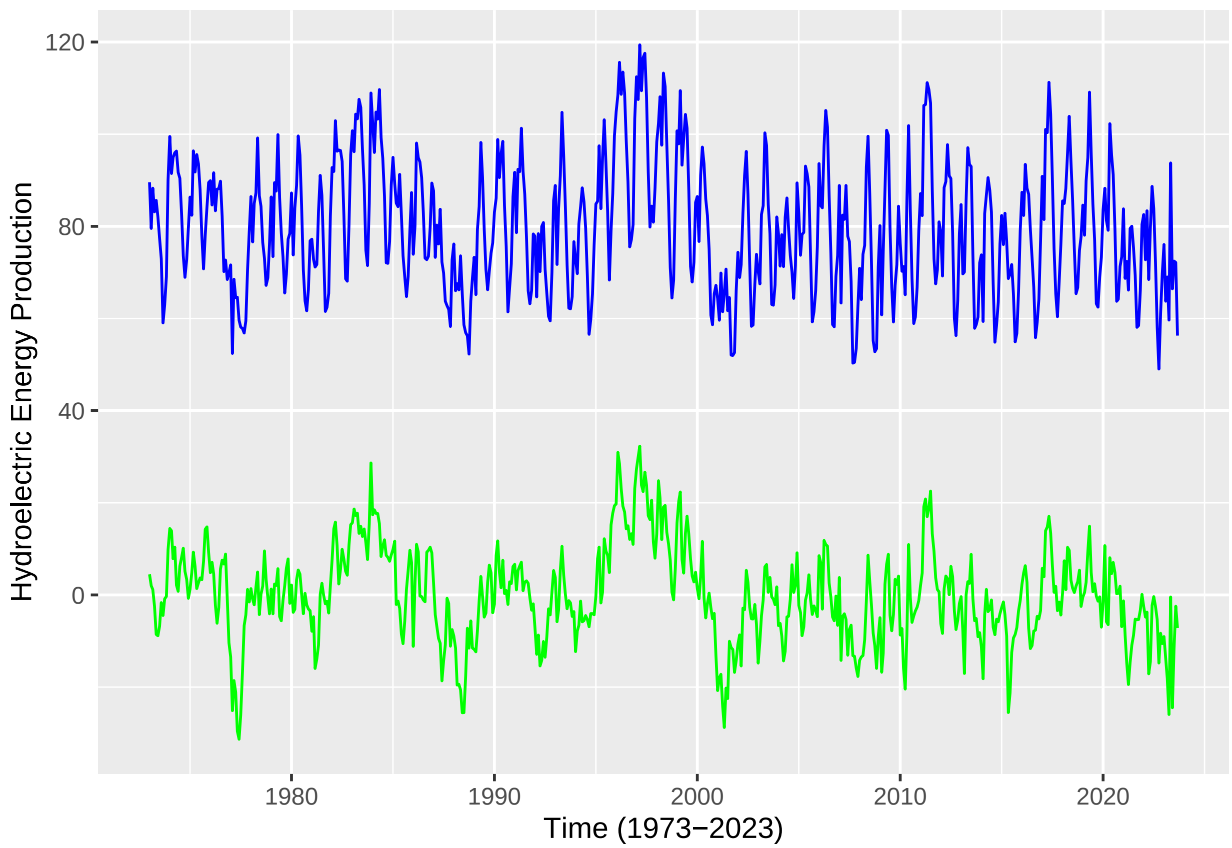
```
#Store regression coefficients
bio_beta_int <- bio_seas$coefficients[1]
bio_beta_coeff <- bio_seas$coefficients[2:12]

re_beta_int <- re_seas$coefficients[1]
re_beta_coeff <- re_seas$coefficients[2:12]

hydro_beta_int <- hydro_seas$coefficients[1]
hydro_beta_coeff <- hydro_seas$coefficients[2:12]

#compute seasonal component
bio_seas_comp <- array(0,nobs_energy)
for(i in 1:nobs_energy){
  bio_seas_comp[i] <- (bio_beta_int+bio_beta_coeff %*% dummies[i,])
}

re_seas_comp <- array(0,nobs_energy)
for(i in 1:nobs_energy){
  re_seas_comp[i] <- (re_beta_int+re_beta_coeff %*% dummies[i,])
}

hydro_seas_comp <- array(0,nobs_energy)
for(i in 1:nobs_energy){
  hydro_seas_comp[i] <- (hydro_beta_int+hydro_beta_coeff %*% dummies[i,])
}

#Removing seasonal component
deseason_bio <- energy_data[,2]-bio_seas_comp
deseason_re <- energy_data[,3]-re_seas_comp
deseason_hydro <- energy_data[,4]-hydro_seas_comp

#Understanding what we did
p6<-ggplot(energy_data, aes(x=energy_date, y=energy_data[,2])) +
            geom_line(color="blue") +
            ylab("Biomass Energy Production") +
            xlab("Time (1973-2023)")+
            geom_line(aes(y=deseason_bio), col="green")
p6
```




```
p7<-ggplot(energy_data, aes(x=energy_date, y=energy_data[,3])) +
            geom_line(color="blue") +
            ylab("Renewable Energy Production") +
            xlab("Time (1973-2023)")+
            geom_line(aes(y=deseason_re), col="green")
p7
```




```r
p8<-ggplot(energy_data, aes(x=energy_date, y=energy_data[,4])) +
            geom_line(color="blue") +
            ylab("Hydroelectric Energy Production") +
            xlab("Time (1973-2023)")+
            geom_line(aes(y=deseason_hydro), col="green")
p8
```



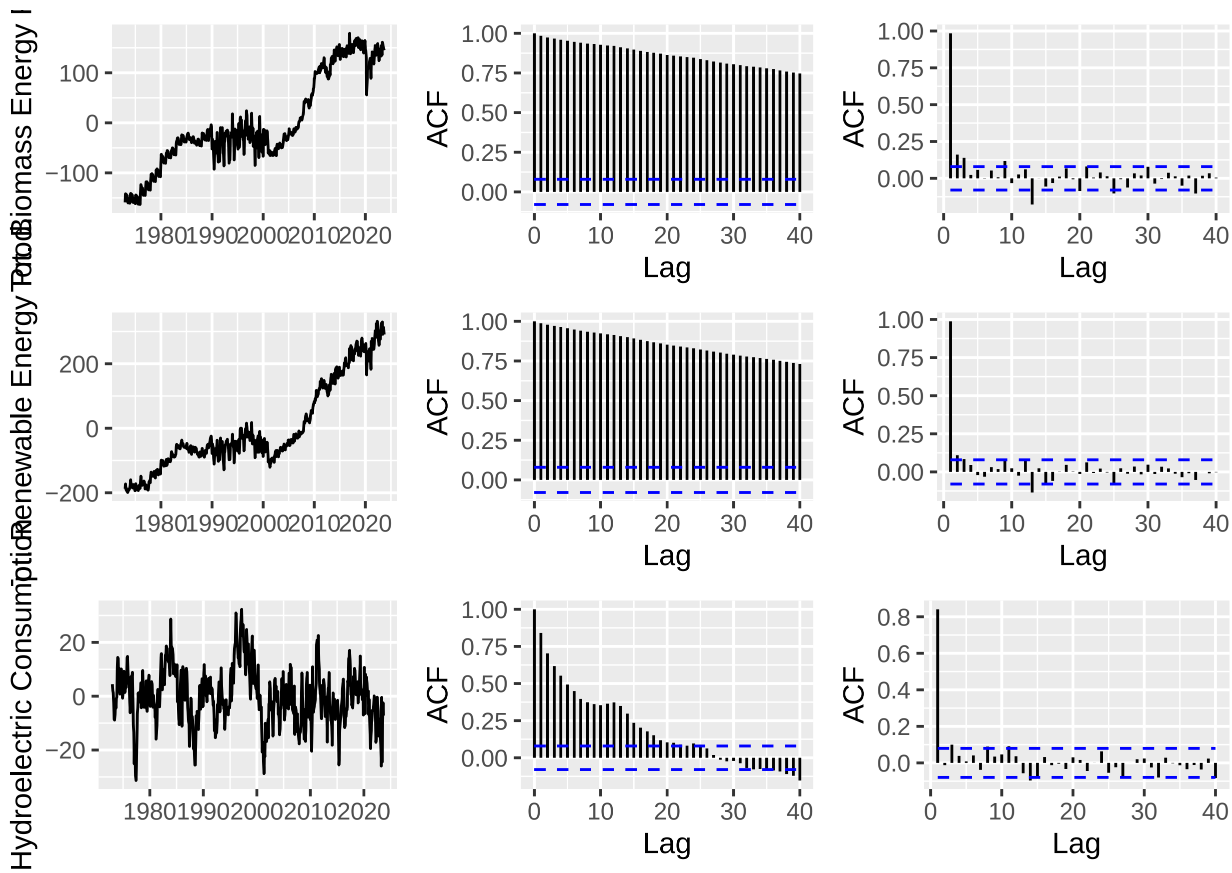
For biomass and hydroelectric, the linear trend still remains. However, for the hydroelectric consumption, removing the seasonal component changed the number of peaks and valleys indicating that the series does not have any (#has less) seasonal variability and the remaining variability from another variable/reason.

Q9

Plot ACF and PACF for the deseason series and compare with the plots from Q1. You may use plot_grid() again to get them side by side. nut mot mandatory. Did the plots change? How?

```
ts_deseason_bio <- ts(deseason_bio,start=c(1973,1),frequency=12)
ts_deseason_re <- ts(deseason_re,start=c(1973,1),frequency=12)
ts_deseason_hydro <- ts(deseason_hydro,start=c(1973,1),frequency=12)


p9 <- plot_grid(
  autoplot(ts_deseason_bio,ylab="Tot. Biomass Energy Prod."),
  autoplot(Acf(ts_deseason_bio,lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_deseason_bio,lag.max=40,plot=FALSE),main=NULL),
  autoplot(ts_deseason_re,ylab="Tot. Renewable Energy Prod."),
  autoplot(Acf(ts_deseason_re,lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_deseason_re,lag.max=40,plot=FALSE),main=NULL),
  autoplot(ts_deseason_hydro,ylab="Hydroelectric Consumption"),
  autoplot(Acf(ts_deseason_hydro,lag.max=40,plot=FALSE),main=NULL),
  autoplot(Pacf(ts_deseason_hydro,lag.max=40,plot=FALSE),main=NULL),
  nrow=3,ncol=3
```

```
)
p9
```



The plot for hydroelectric component has changed a lot. The ACF values fall below the blue line/are not significant after lag 20. The seasonal pattern of the ACF does not exist. Additionally, in the PACF plot, the values are not significant throughout. (Removing the seasonanility helped eliminate some time dependence. But, still some time information is being carried because the ACF is significant at lag 12 as well. More modelling is needed.) The ACF plots for biomass and hydroelectric have not changed.