ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 3 - Due date 02/12/21

Thomas Hancock

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 project open the first thing you will do is change "Student Name" on line 3 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_A01_Sp21.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_Consumption The data comes from the US Energy Information and Administration and corresponds to the January 2021 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(xlsx)
```

```
raw_data <- read.xlsx("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                      header = TRUE, startRow = 10, sheetIndex = 1)

energy_df <- data.frame(raw_data[-1,1], lapply(raw_data[-1,4:6], as.numeric)) # Convert to numeric valu
```

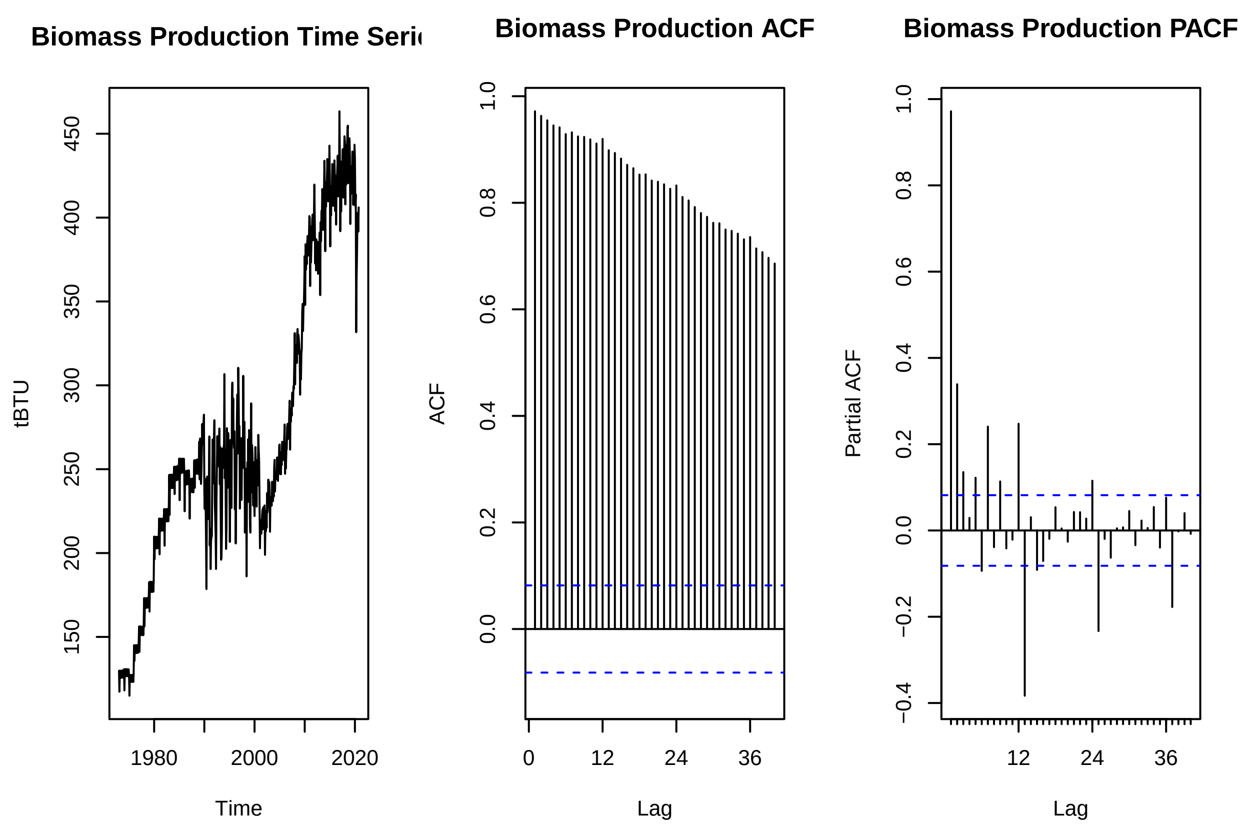
```r
n_series <- ncol(energy_df)-1

n_obs <- nrow(energy_df)

energy_ts <- ts(energy_df[2:(n_series+1)], start=c(1973,1), frequency = 12)

name_list <- c("Biomass Production", "Renewable Production", "Hydroelectric Consumption")
```

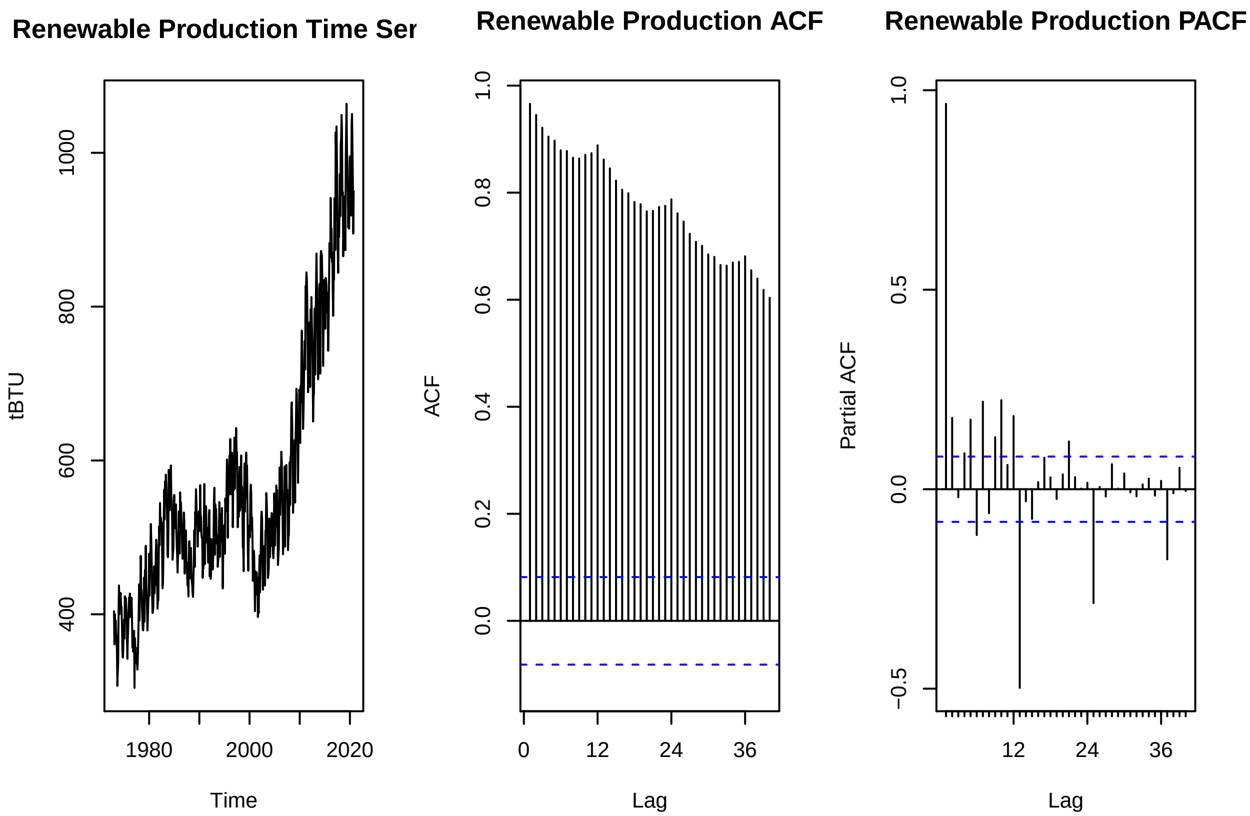
##Trend Component

Q1

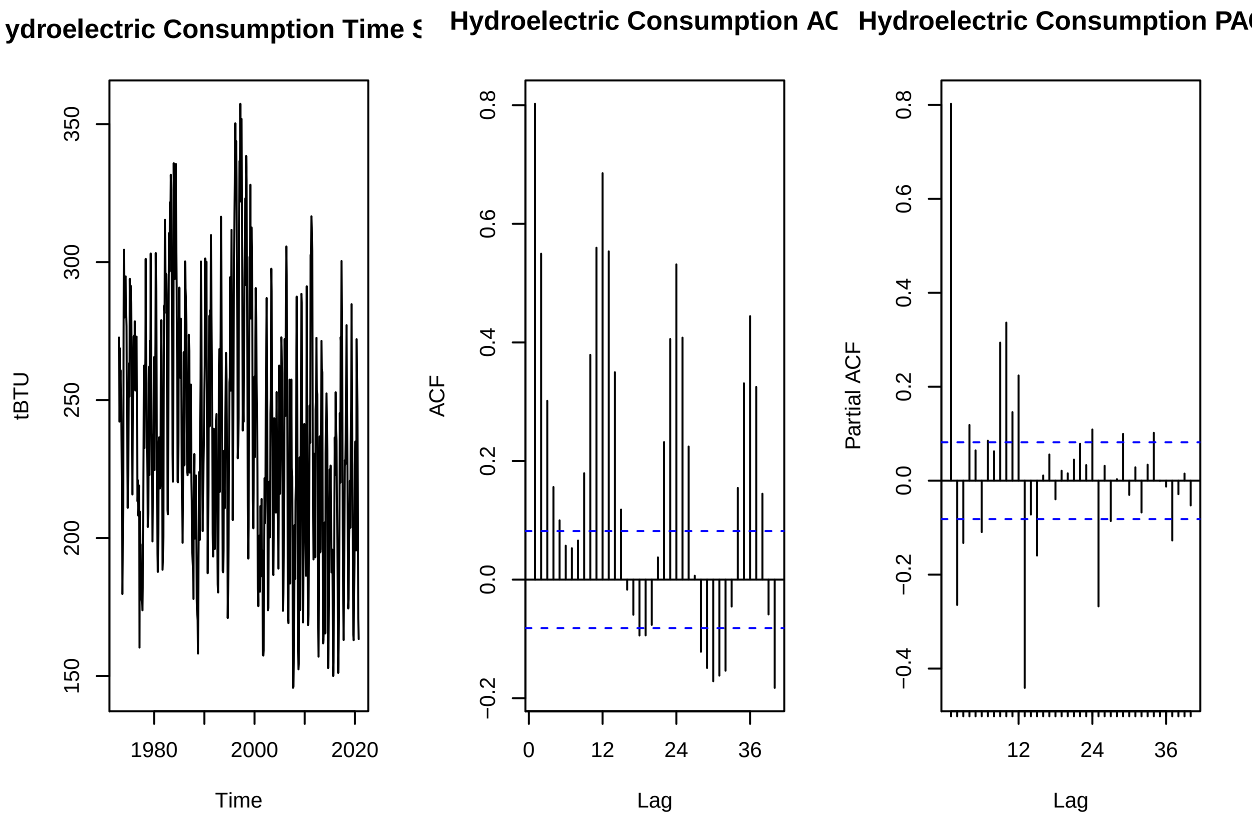
Create a plot window that has one row and three columns. And then for each object on your data frame, fill the plot window with time series plot, ACF and PACF. You may use the some code form A2, but I want all three plots on the same window this time. (Hint: watch videos for M4)

```r
par(mfrow = c(1,3))
for (i in 1:n_series) {
  ts.plot(energy_ts[,i], main = paste0(name_list[i], " Time Series"), ylab = "tBTU")
  Acf(energy_ts[,i], lag.max = 40, main = paste0(name_list[i], " ACF"))
  Pacf(energy_ts[,i], lag.max = 40, main = paste0(name_list[i], " PACF"))
}
```





Renewable Production Time Ser
Renewable Production ACF
Renewable Production PACF
tBTU
400
600
800
1000
1980
2000
2020
Time
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
12
24
36
Lag
Partial ACF
−0.5
0.0
0.5
1.0
12
24
36
Lag



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?

> Answer: The Total Biomas Energy Production appears to have a linear trend, but little to no seasonal trend. The Total Renewable Energy Production series also appears to have a linear trend and, perhaps, a very minor seasonal trend (it looks a little more evident than the previous series, but the seasonal trend may not be significant here). Lastly, the Hydroelectric Power Consumption trend seems to have a fairly evident seasonal trend, but it is not clear if there is a linear trend (there appears to be a slight downward progression, but there is also a lot of noise/variation).

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.

```
t <- c(1:n_obs)

coeffs <- data.frame(row.names = name_list)

for (i in (1:n_series)) {
  model <- lm(energy_ts[,i] ~ t)
  coeffs[i,1] <- model$coefficients[1]
  coeffs[i,2] <- model$coefficients[2]
  print(paste0(name_list[i], ":"))
```

```r
  print(summary(model))
  cat("\n\n")

}
```

```
## [1] "Biomass Production:"
##
## Call:
## lm(formula = energy_ts[, i] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.149  -25.456    4.985   33.353   79.634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.355e+02  3.296e+00   41.11   <2e-16 ***
## t           4.702e-01  9.934e-03   47.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.44 on 572 degrees of freedom
## Multiple R-squared:  0.7966, Adjusted R-squared:  0.7962
## F-statistic:  2240 on 1 and 572 DF,  p-value: < 2.2e-16
##
##
##
## [1] "Renewable Production:"
##
## Call:
## lm(formula = energy_ts[, i] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -224.735  -55.673    5.418   60.453  263.849
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 330.37156    7.86270   42.02   <2e-16 ***
## t             0.84299    0.02369   35.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.07 on 572 degrees of freedom
## Multiple R-squared:  0.6887, Adjusted R-squared:  0.6882
## F-statistic:  1266 on 1 and 572 DF,  p-value: < 2.2e-16
##
##
##
## [1] "Hydroelectric Consumption:"
##
## Call:
## lm(formula = energy_ts[, i] ~ t)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -94.06 -31.57  -1.63  27.73 120.69
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 258.05622    3.52899  73.125  < 2e-16 ***
## t            -0.07341    0.01063  -6.903 1.36e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.22 on 572 degrees of freedom
## Multiple R-squared:  0.07689,    Adjusted R-squared:  0.07528
## F-statistic: 47.64 on 1 and 572 DF,  p-value: 1.361e-11
```

```r
colnames(coeffs) <- c("beta0", "beta1")
coeffs
```

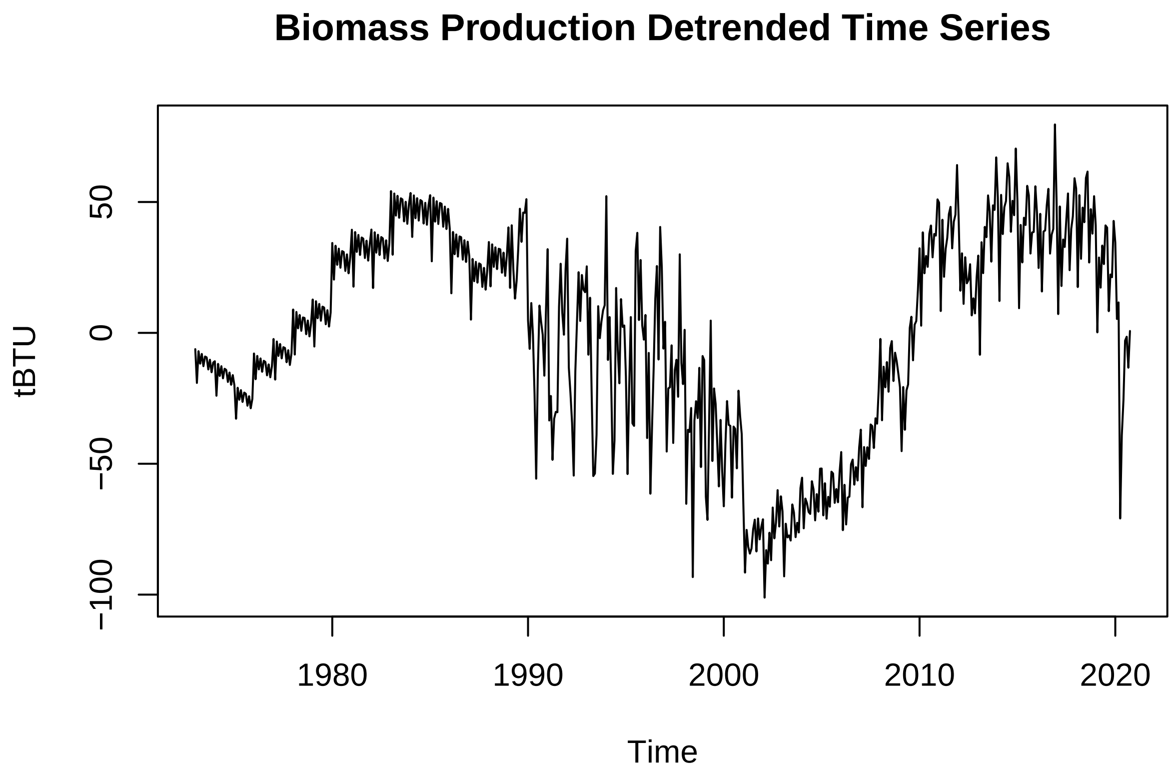
```
##                              beta0       beta1
## Biomass Production        135.5250  0.47016053
## Renewable Production      330.3716  0.84299315
## Hydroelectric Consumption 258.0562 -0.07340757
```

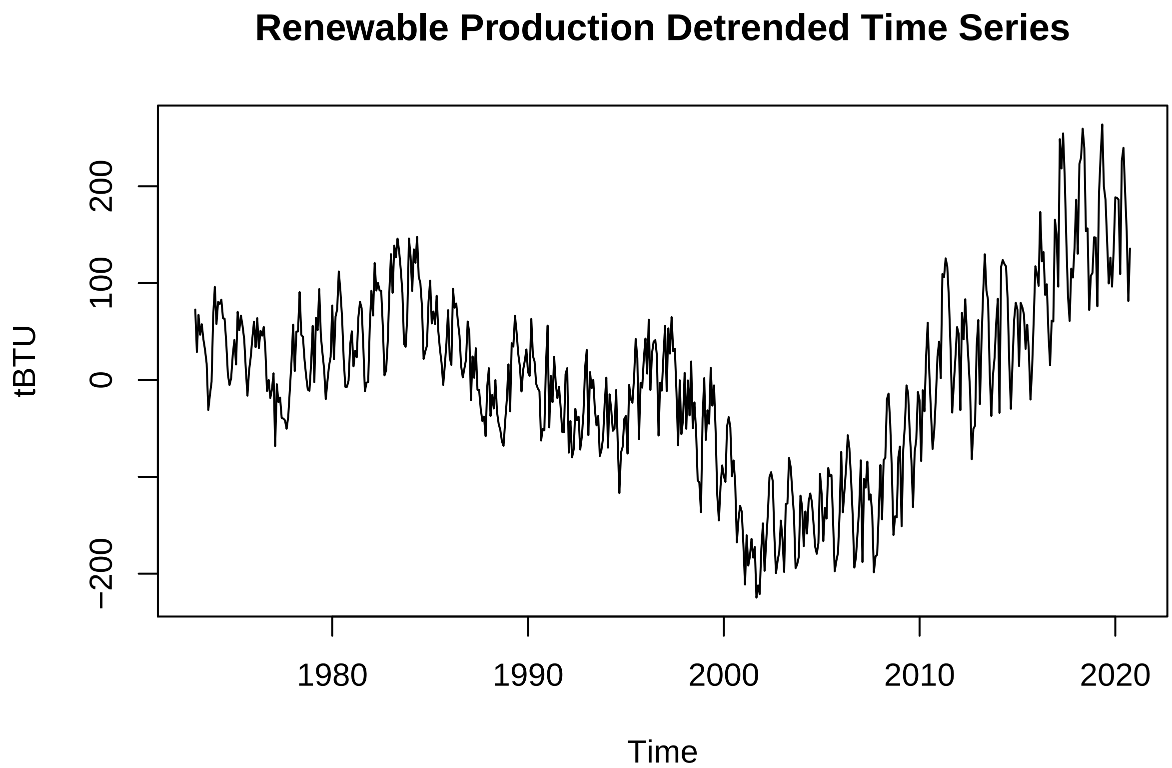
> The first coefficient ("beta0") is the intercept of the linear regression, and the second coefficient ("beta1") is the slope. As such, we can see that the Biomass and Renewable series have positive trends (so they are increasing over time), whereas the Hydroelectric series has a negative (downward-sloping) trend so it is decreasing over time. The intercept units are the same units as the series (tBTU, I believe), and the slope units are tBTU/month. So, for example, the Biomass Energy Production model has an intercept of 135.5 tBTU and increases by 0.47 tBTU every month. These values match the general trends described visually in Question 2. All of them are statistically significant at the 0.001 level.

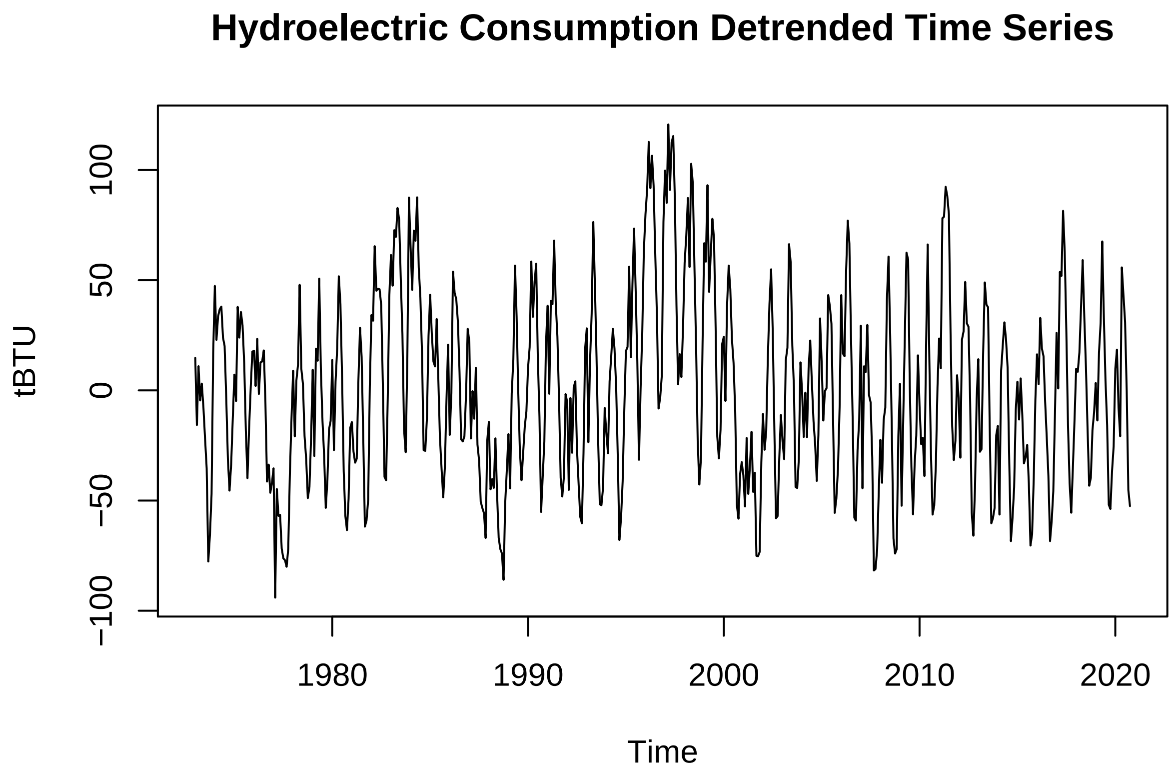
Q4

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
detrend_ts <- energy_ts

for (i in (1:n_series)) {
  detrend_ts[,i] <- energy_ts[,i] - (coeffs[i,"beta0"] + coeffs[i,"beta1"]*t)
  ts.plot(detrend_ts[,i], main = paste0(name_list[i], " Detrended Time Series"), ylab = "tBTU")
}
```


Biomass Production Detrended Time Series
tBTU
50
0
−50
−100
1980
1990
2000
2010
2020
Time


Renewable Production Detrended Time Series
tBTU
200
100
0
−200
1980
1990
2000
2010
2020
Time

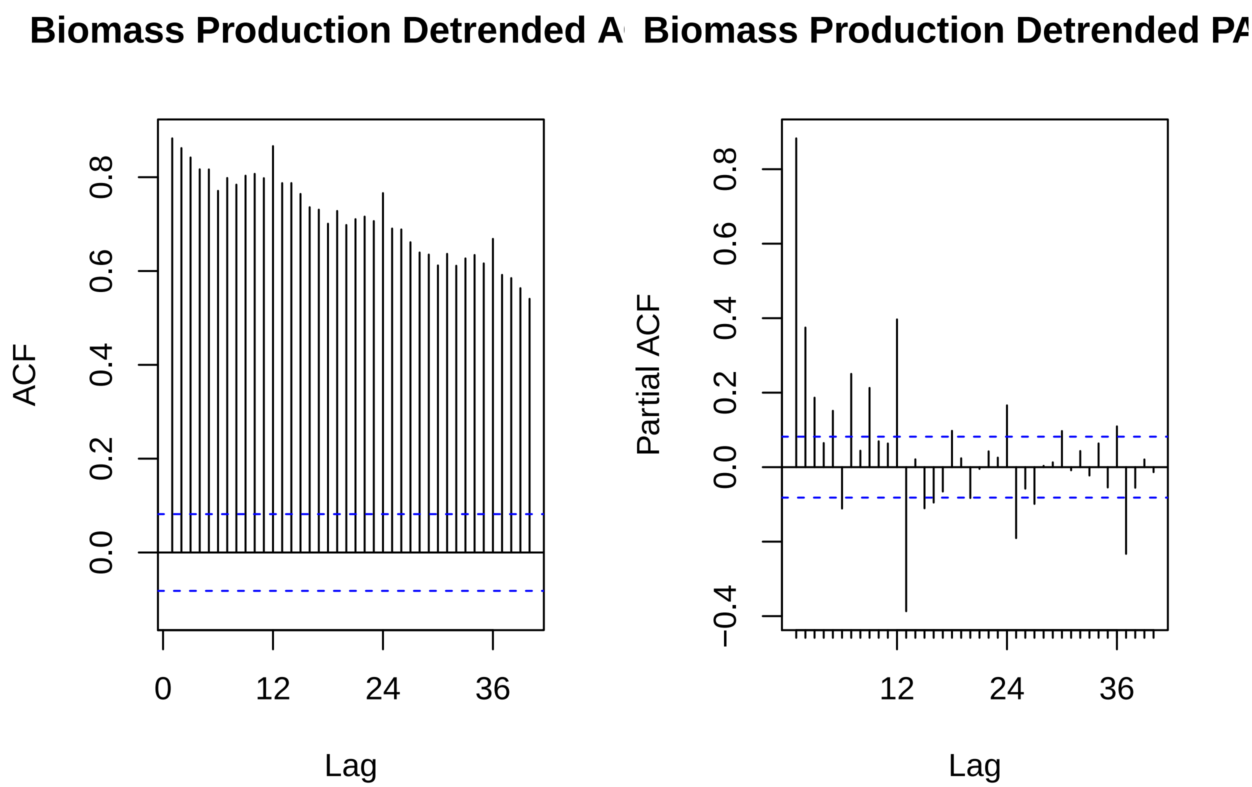


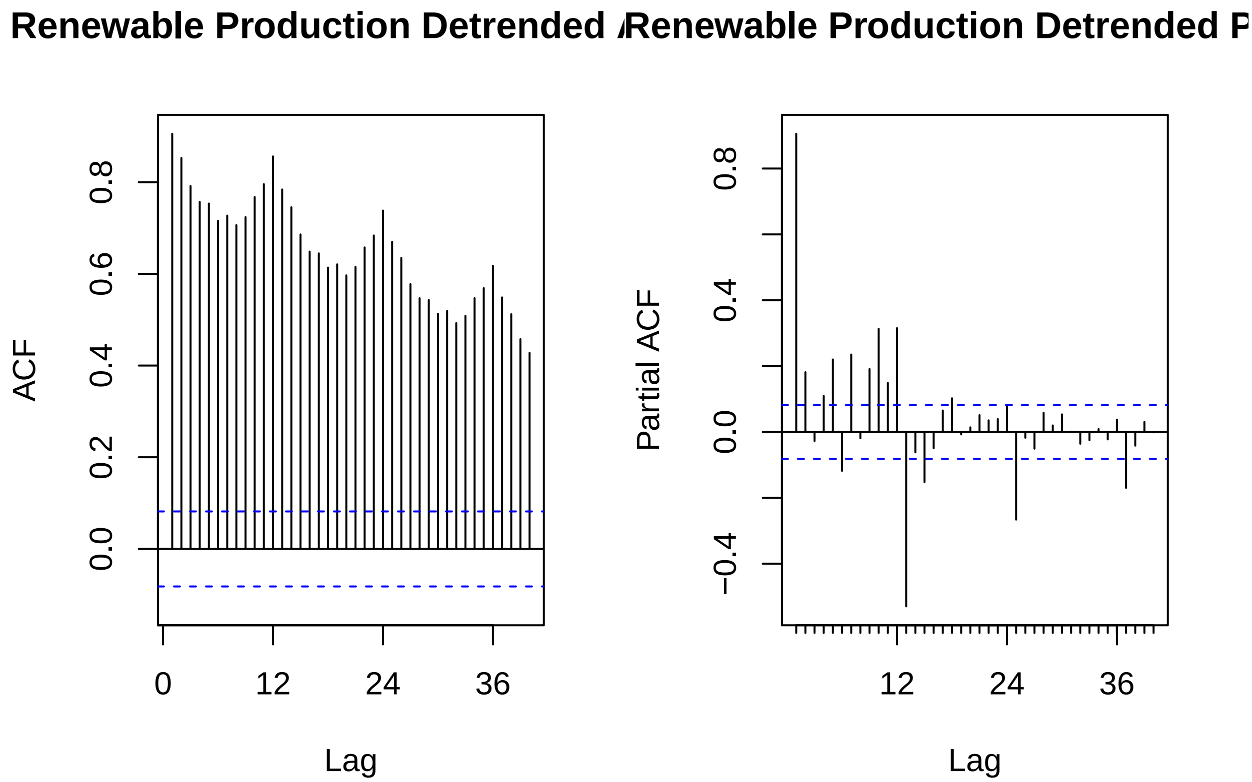
The plots now all seem more centered around 0 on the y-axis (almost as if they were just rotated up or down). Especially the Biomas and Renewable series that had larger slope coefficients in the linear model appear "flatter" now (less rise). There are still seasonal and stochastic variations, but there do not visually appear to be any linear trends in the series now.

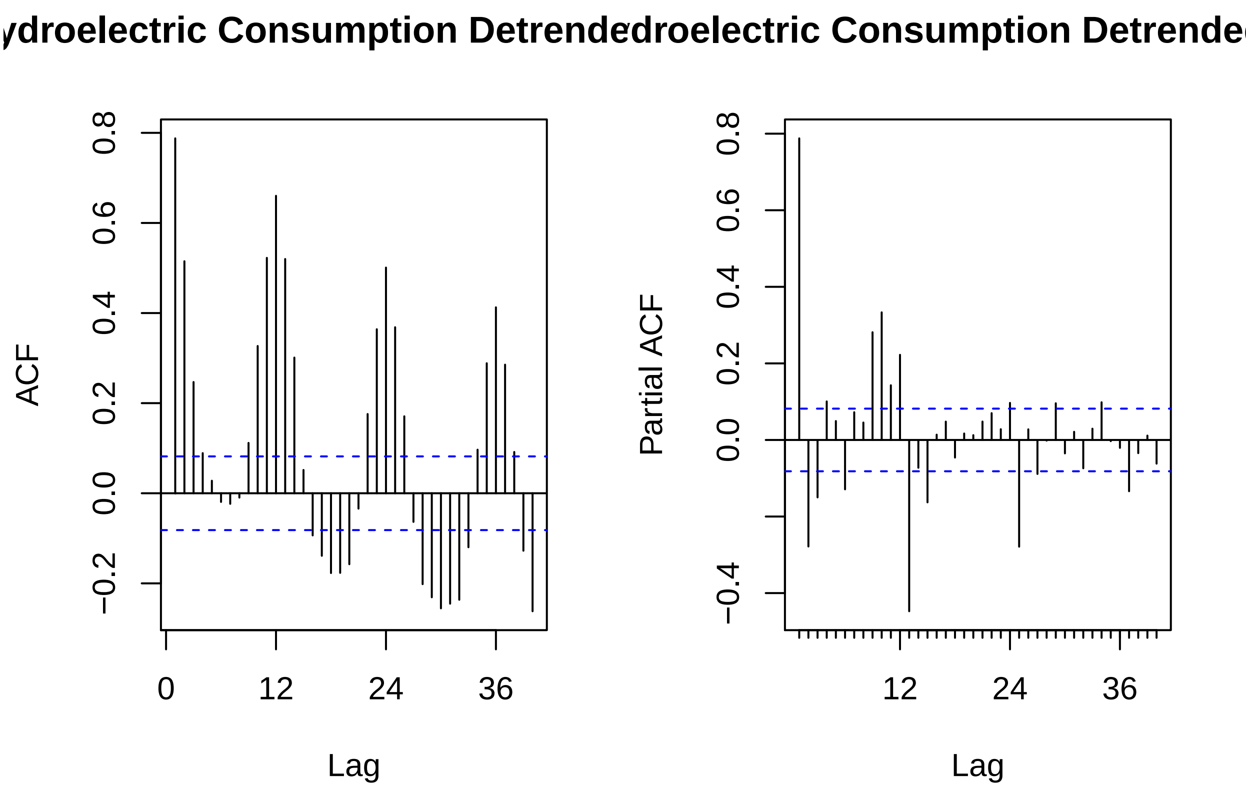
Q5

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

```
par(mfrow = c(1,2))
for (i in 1:n_series) {
  Acf(detrend_ts[,i], lag.max = 40, main = paste0(name_list[i], " Detrended ACF"))
  Pacf(detrend_ts[,i], lag.max = 40, main = paste0(name_list[i], " Detrended PACF"))
}
```


Biomass Production Detrended A
Biomass Production Detrended PA
ACF
Partial ACF
0.0
0.2
0.4
0.6
0.8
−0.4
0
12
24
36
Lag
Lag


Renewable Production Detrended A
Renewable Production Detrended P
ACF
0.0
0.2
0.4
0.6
0.8
0
12
24
36
Lag
Partial ACF
−0.4
0.0
0.4
0.8
12
24
36
Lag



> The ACF and PACF plots changed a little bit, but not much. In particular, there is more evidence of seasonal variability, especially in the ACF plots. This change is particularly notable in the Renewable ACF plot. Presumably this difference is due to the removal of the linear trend, so any seasonal variation has a more prominent effect on the adjusted values (basically, the seasonal trends are less "lost" in the linear trend).

Seasonal Component

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

Q6

Do the series seem to have a seasonal trend? Which serie/series? Use function *lm()* to fit a seasonal means model to this/these time series. Ask R to print the summary of the regression. Interpret the regression output. Save the regression coefficients for further analysis.

```
seas_coeffs <- data.frame(row.names = name_list)

dummies <- seasonaldummy(energy_ts[,1]) # We could use any of the series

for (i in (1:n_series)) {
  model <- lm(energy_ts[,i] ~ dummies)
  seas_coeffs[i,1:12] <- model$coefficients
  print(paste0(name_list[i], ":"))
  print(summary(model))
  cat("\n\n")

}
```

```
## [1] "Biomass Production:"
##
## Call:
## lm(formula = energy_ts[, i] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153.47  -50.56  -20.25   52.13  182.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 280.5693    12.7954  21.927   <2e-16 ***
## dummiesJan   -1.0039    18.0009  -0.056    0.956
## dummiesFeb  -29.3891    18.0009  -1.633    0.103
## dummiesMar   -8.6090    18.0009  -0.478    0.633
## dummiesApr  -20.5046    18.0009  -1.139    0.255
## dummiesMay  -14.0960    18.0009  -0.783    0.434
## dummiesJun  -19.5548    18.0009  -1.086    0.278
## dummiesJul   -3.4306    18.0009  -0.191    0.849
## dummiesAug    0.2220    18.0009   0.012    0.990
## dummiesSep  -11.9821    18.0009  -0.666    0.506
## dummiesOct   -0.5379    18.0009  -0.030    0.976
## dummiesNov   -9.3753    18.0954  -0.518    0.605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.72 on 562 degrees of freedom
## Multiple R-squared:  0.01116,    Adjusted R-squared:  -0.008199
## F-statistic: 0.5764 on 11 and 562 DF,  p-value: 0.8486
##
##
##
## [1] "Renewable Production:"
##
## Call:
## lm(formula = energy_ts[, i] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -263.99 -102.98  -52.33   36.68  453.58
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  580.912     24.406  23.802   <2e-16 ***
## dummiesJan    12.451     34.335   0.363   0.7170
## dummiesFeb   -38.964     34.335  -1.135   0.2569
## dummiesMar    20.515     34.335   0.597   0.5504
## dummiesApr     8.294     34.335   0.242   0.8092
## dummiesMay    36.628     34.335   1.067   0.2865
## dummiesJun    19.560     34.335   0.570   0.5691
## dummiesJul     8.863     34.335   0.258   0.7964
## dummiesAug   -18.480     34.335  -0.538   0.5906
## dummiesSep   -62.410     34.335  -1.818   0.0696 .
## dummiesOct   -42.649     34.335  -1.242   0.2147
```

```
## dummiesNov   -42.516     34.515  -1.232   0.2185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 167.3 on 562 degrees of freedom
## Multiple R-squared:  0.03244,    Adjusted R-squared:  0.01351
## F-statistic: 1.713 on 11 and 562 DF,  p-value: 0.06702
##
##
##
## [1] "Hydroelectric Consumption:"
##
## Call:
## lm(formula = energy_ts[, i] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.064 -22.897  -2.654  20.642  98.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.887      4.863  49.125  < 2e-16 ***
## dummiesJan    13.270      6.841   1.940  0.05291 .
## dummiesFeb    -8.133      6.841  -1.189  0.23499
## dummiesMar    20.442      6.841   2.988  0.00293 **
## dummiesApr    17.199      6.841   2.514  0.01221 *
## dummiesMay    40.726      6.841   5.953 4.64e-09 ***
## dummiesJun    31.764      6.841   4.643 4.28e-06 ***
## dummiesJul    10.858      6.841   1.587  0.11306
## dummiesAug   -17.907      6.841  -2.618  0.00909 **
## dummiesSep   -50.121      6.841  -7.326 8.26e-13 ***
## dummiesOct   -49.165      6.841  -7.187 2.12e-12 ***
## dummiesNov   -32.757      6.877  -4.763 2.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.34 on 562 degrees of freedom
## Multiple R-squared:  0.4345, Adjusted R-squared:  0.4234
## F-statistic: 39.25 on 11 and 562 DF,  p-value: < 2.2e-16
```

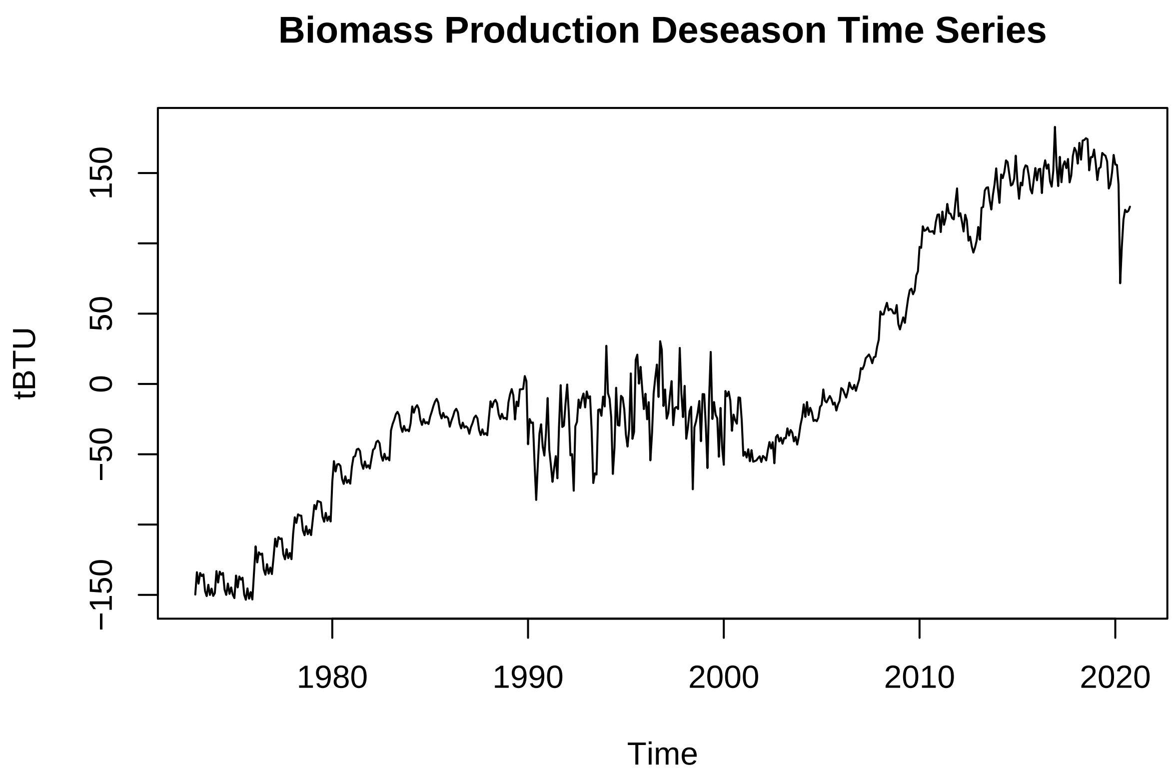
```r
colnames(seas_coeffs) <- names(model$coefficients) # More descriptive names
```

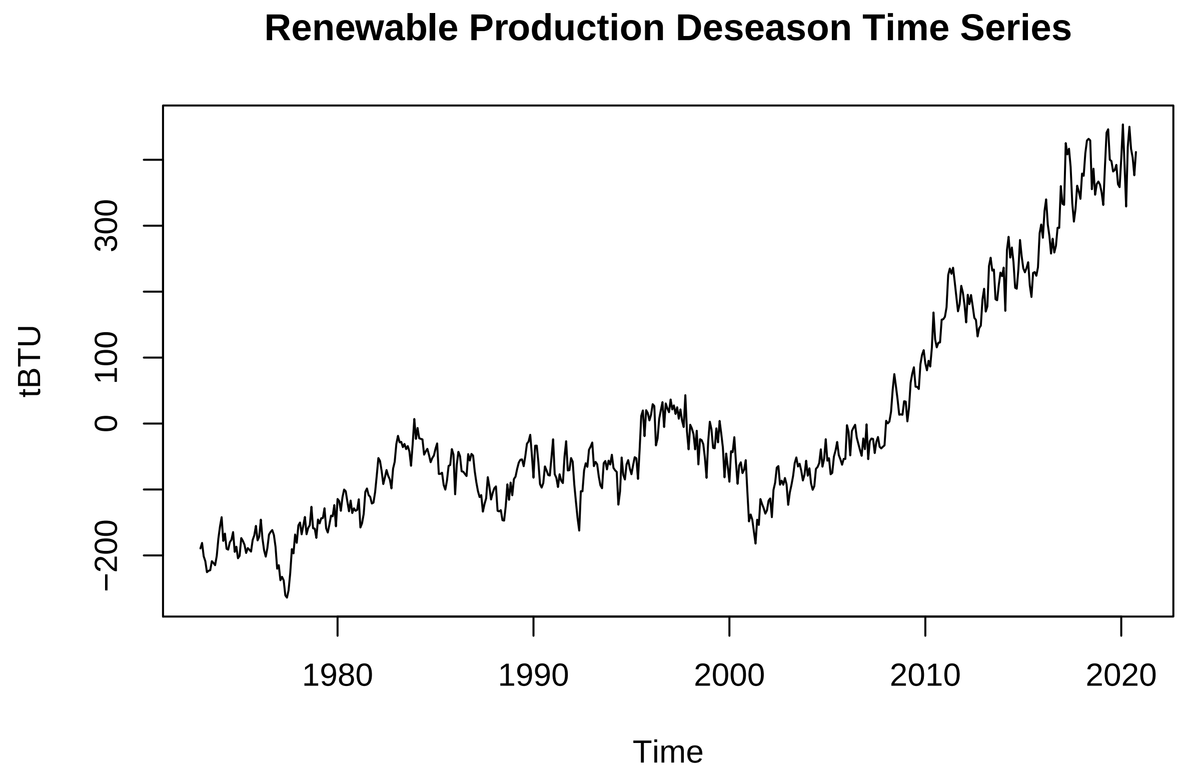
The Hydroelectric series definitely seems to have a seasonal trend. The Renewable series seems to have a slight seasonal trend, at least when examining the ACF plot of the detrended series (Question 5). The Biomass series does not appear to have much seasonality. The regression coefficients in the model summaries show the impact that December (the intercept) has on the data, and then the other dummy variables relative to that. So, for example, January values for the Hydroelectric series model are 13.27 tBTU more than December, but the September values are 50.121 lower. We can also see from the significance indicators that there are not really any significant seasonal variables for the Biomass and Renewable series, as mentioned above. However, most of the seasonal variables for the Hydroelectric series are significant at the 0.05 level.

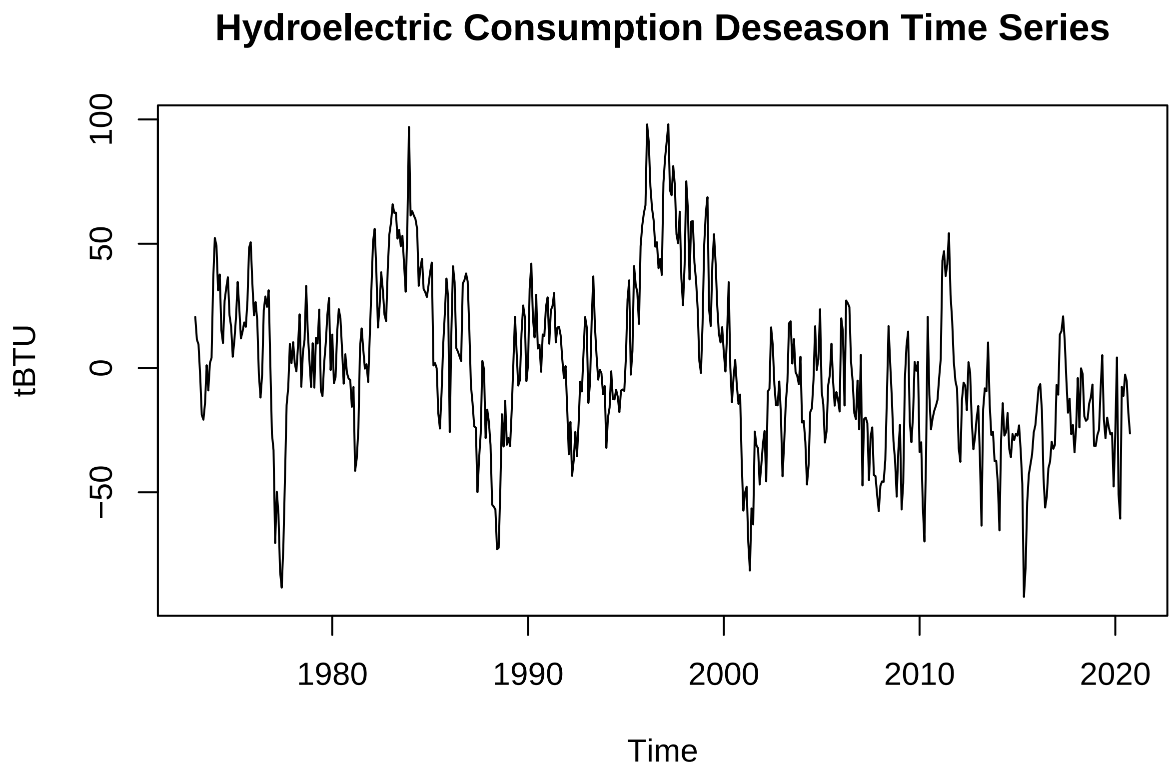
Q7

Use the regression coefficients from Q6 to deseason the series. Plot the deseason series and compare with the plots from part Q1. Did anything change?

```
deseas_ts <- energy_ts


for (i in (1:n_series)) {
  for (k in (1:n_obs)) {
    deseas_ts[k,i] <- energy_ts[k,i] - (seas_coeffs[i,1] + dummies[k,] %*% t(seas_coeffs[i,2:12]))
  }

  ts.plot(deseas_ts[,i], main = paste0(name_list[i], " Deseason Time Series"), ylab = "tBTU")
}
```




Renewable Production Deseason Time Series
tBTU
−200
0
100
300
1980
1990
2000
2010
2020
Time

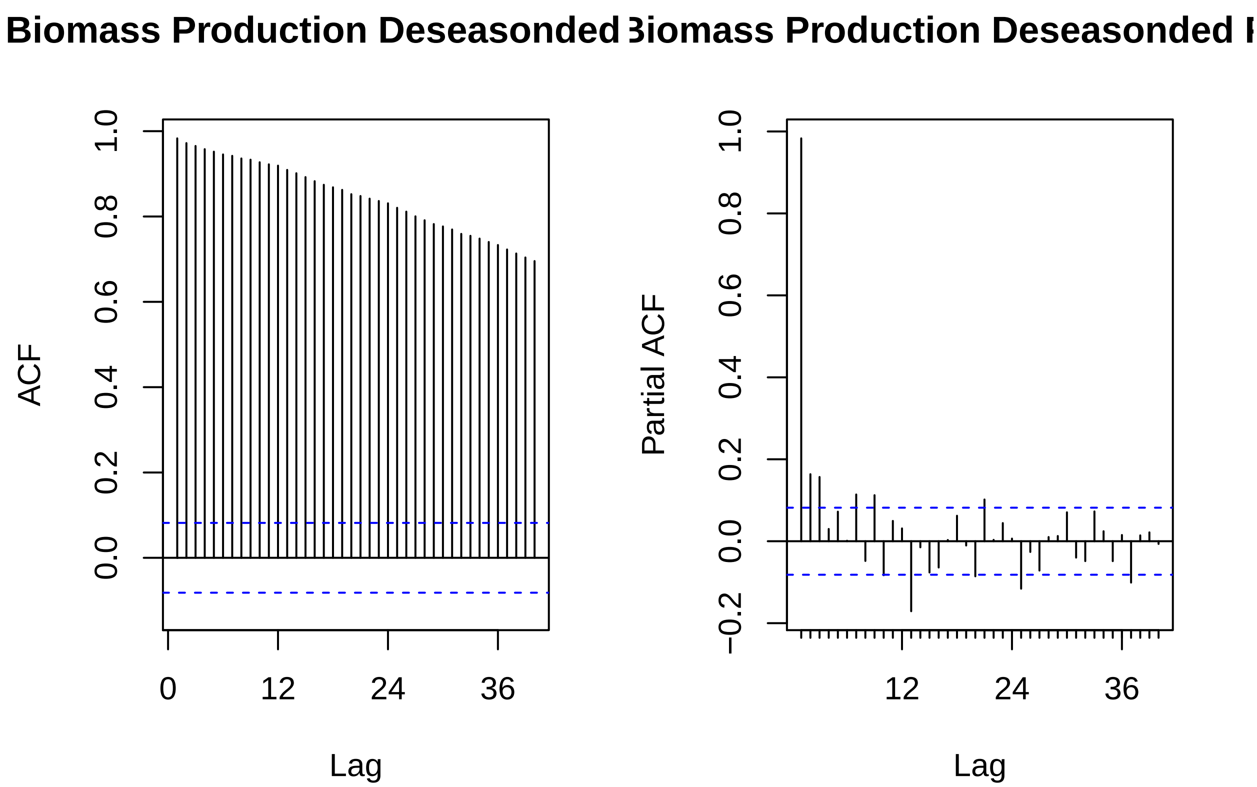


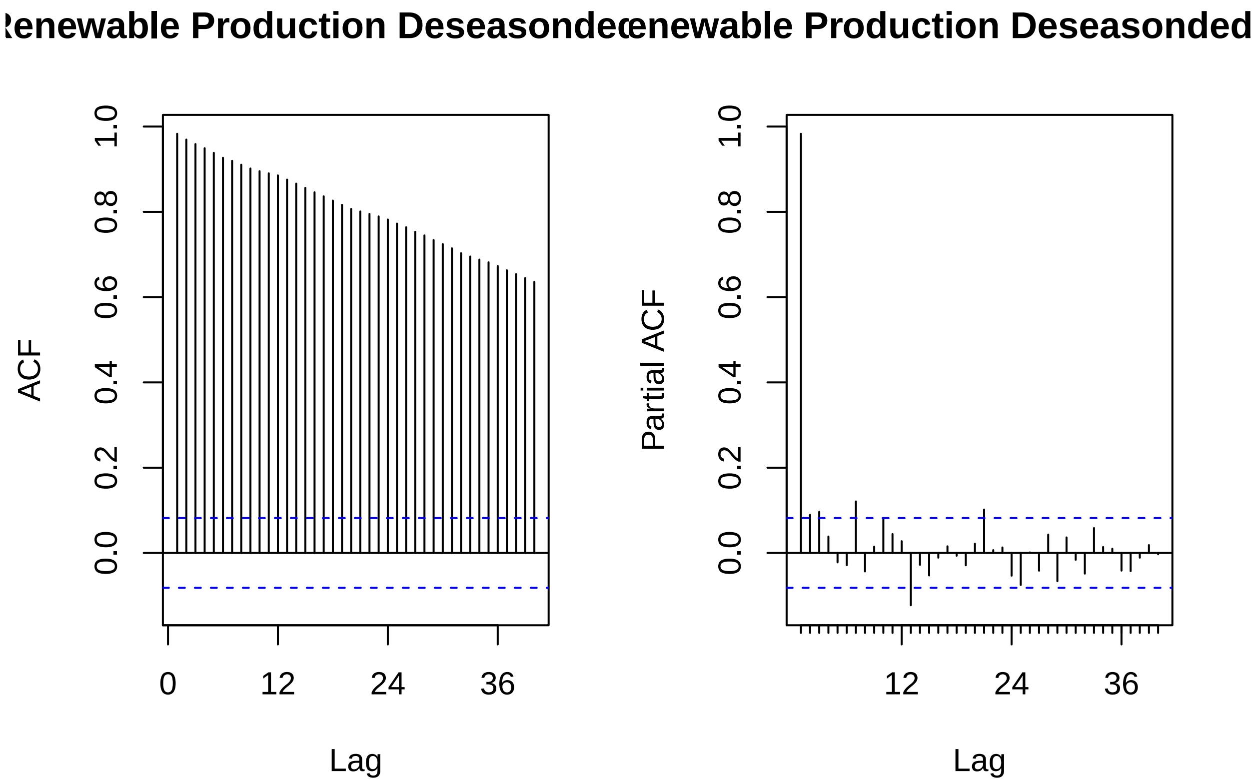
> All of them look a little bit “smoother” with fewer small oscillations (presumably oscillations from month to month). The Biomass and Renewable (first and second) series do not look very different otherwise, but this change is more noticeable for the hydroelectric series, which had much more seasonal variation originally. The large variations are now less frequent - the original data had many oscillations (presumably due to seasonal variations), but with the deseasoning, the variations seem to occur more on a timescale of years.

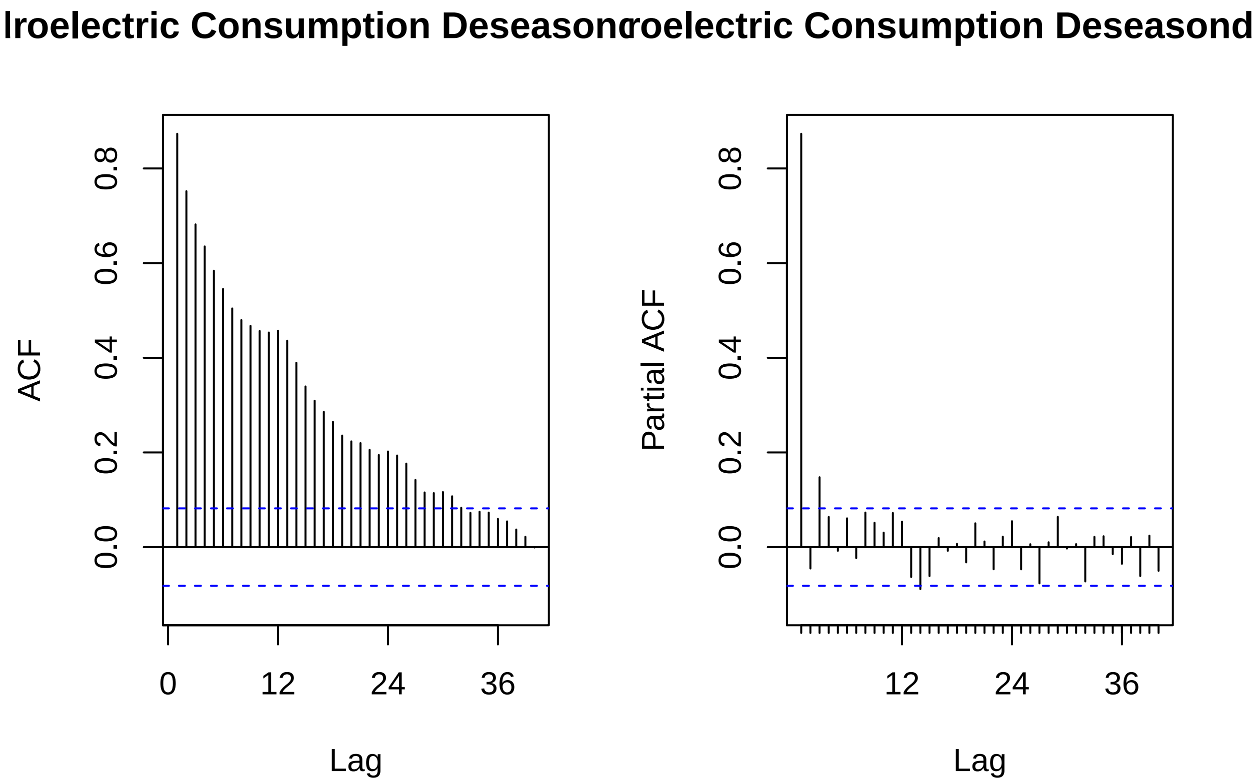
Q8

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

```
par(mfrow = c(1,2))
for (i in 1:n_series) {
  Acf(deseas_ts[,i], lag.max = 40, main = paste0(name_list[i], " Deseasonded ACF"))
  Pacf(deseas_ts[,i], lag.max = 40, main = paste0(name_list[i], " Deseasonded PACF"))
}
```


Biomass Production Deseasonded
Biomass Production Deseasonded F
ACF
Partial ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
12
24
36
Lag
Lag


Renewable Production Deseasonded
Renewable Production Deseasonded
ACF
Partial ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
12
24
36
Lag
Lag



> The Biomass and Renewable (first and second set) are not particularly changed from the original plots. There was a slight seasonal trend in the Renewable (second time series) series, which is not gone. The Hydroelectric series (third series) is notably changed because it had a strong seasonal component that is now removed. For all of them, the ACF now is "smoother" (no periodic "bumps" at regular lag intervals), and the PACF shows little to no significant correlation with lags beyond the first few.