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

Assignment 3 - Due date 02/12/21

Keyang Xue

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_Consumpti 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(tseries)
library(Kendall)
library(forecast)
library(readxl)
library(ggplot2)
```

```
#import data
raw_energy <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
```

```
## * ‘‘ -> ...3
## * ‘‘ -> ...4
## * ‘‘ -> ...5
## * ...
```

```r
#import column names
col_names <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"
```

```
## New names:
## * ‘‘ -> ...1
## * ‘‘ -> ...2
## * ‘‘ -> ...3
## * ‘‘ -> ...4
## * ‘‘ -> ...5
## * ...
```

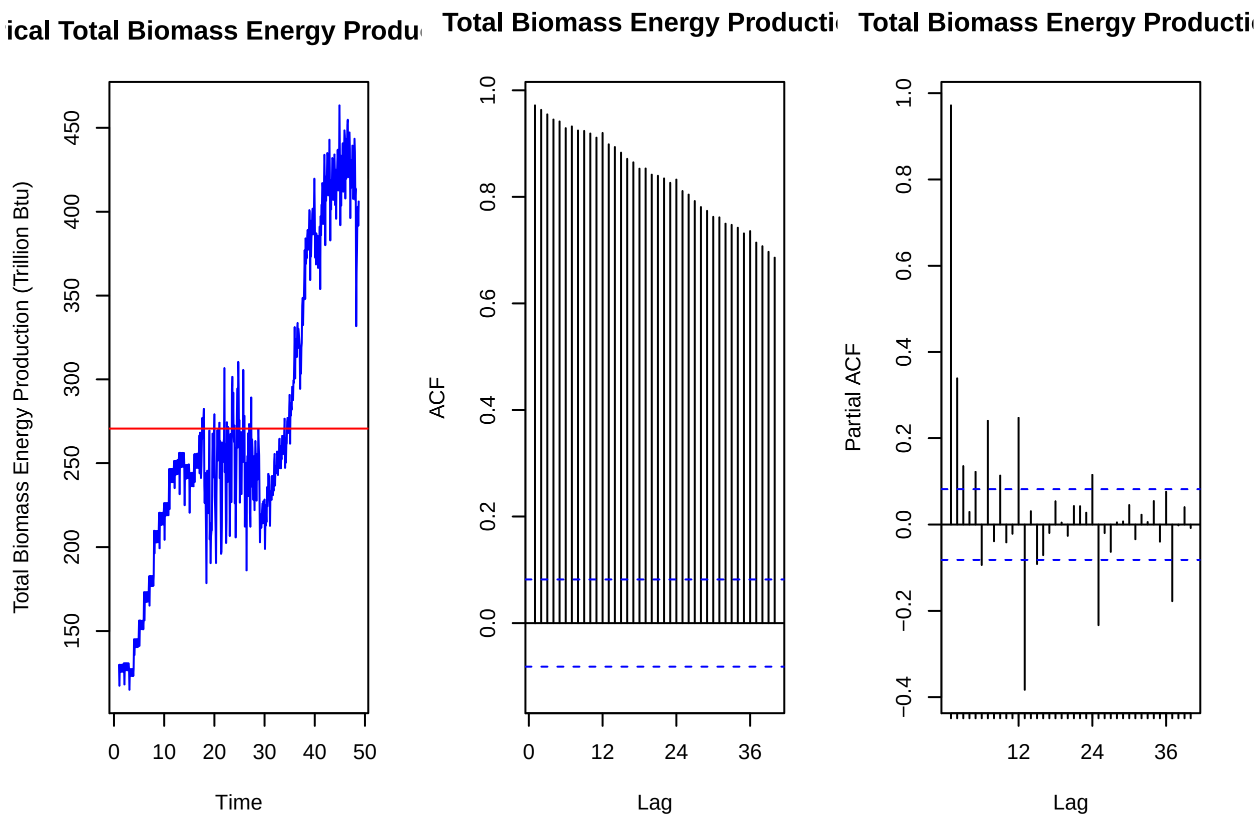
```r
#extract required data
energy <- cbind.data.frame(raw_energy[,1],raw_energy[,4:6])

#transform data to ts
ts_energy <- ts(energy[2:4], frequency = 12)
```

##Trend Component

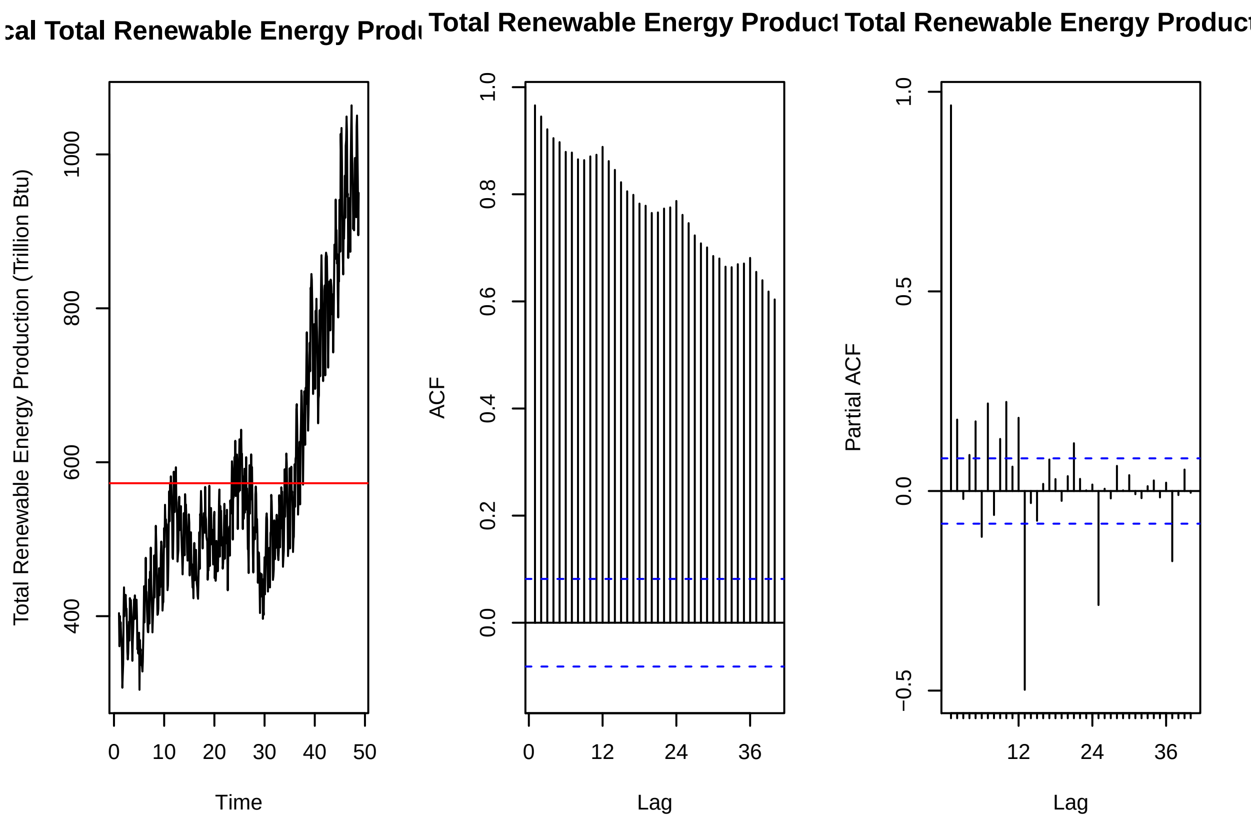
Q1

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)

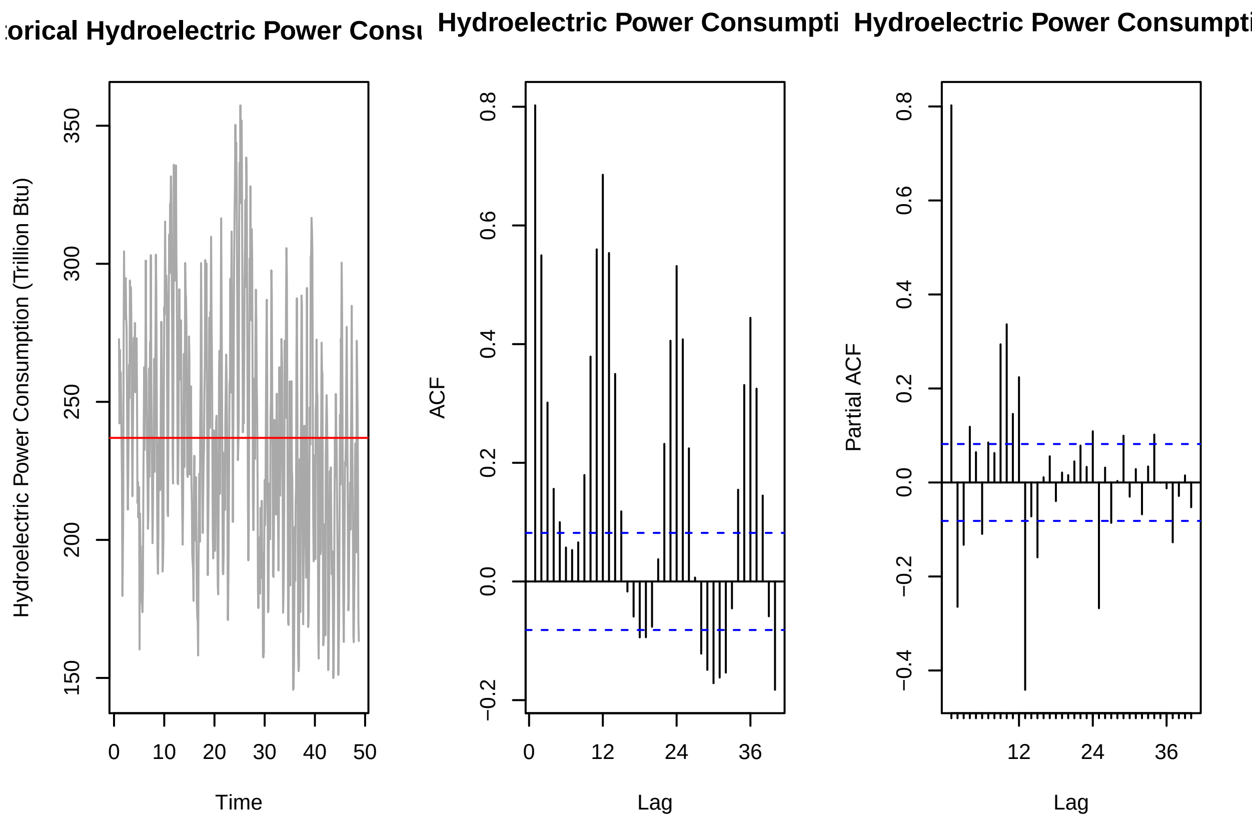
```r
par(mfrow=c(1,3))
#plot biomass
# # ggplot(energy, aes(x= energy[,1], y=energy[,2])) +
# #   geom_line(color="blue")+ xlab("Time") +
#     ylab("Total Biomass Energy Production (Trillion Btu)") +
#     labs(title = 'Historical Total Biomass Energy Production Data') +   geom_hline(yintercept=mean(ts_

plot(ts_energy[,1], type="l",col="blue",
     ylab="Total Biomass Energy Production (Trillion Btu)",
     main='Historical Total Biomass Energy Production Data')
abline(h=mean(ts_energy[,1]),col="red")

biomass_acf <- Acf(ts_energy[,1],lag.max=40, type="correlation", plot=TRUE,
                   main = paste(col_names[,4]))

biomass_pacf <- Pacf(ts_energy[,1],lag.max=40, plot=TRUE,
                     main = paste(col_names[,4]))
```




```r
#plot renewable
# ggplot(energy, aes(x= energy[,1], y=energy[,3])) +
#              geom_line(color="black")+ xlab("Time") + ylab("Total Renewable Energy Production (Trillio

plot(ts_energy[,2], type="l",col="black",
     ylab="Total Renewable Energy Production (Trillion Btu)",
     main='Historical Total Renewable Energy Production Data')
abline(h=mean(ts_energy[,2]),col="red")

renewable_acf <- Acf(ts_energy[,2],lag.max=40, type="correlation", plot=TRUE,
                     main = paste(col_names[,5]))

renewable_pacf <- Pacf(ts_energy[,2],lag.max=40, plot=TRUE,
                       main = paste(col_names[,5]))
```




```
#plot hydroelectric
# ggplot(energy, aes(x= energy[,1], y=energy[,4])) +
#              geom_line(color="dark gray")+ xlab("Time") + ylab("Hydroelectric Power Consumption (Trill

plot(ts_energy[,3], type="l",col="dark gray",
     ylab="Hydroelectric Power Consumption (Trillion Btu)",
     main='Historical Hydroelectric Power Consumption')
abline(h=mean(ts_energy[,3]),col="red")

hydro_acf <- Acf(ts_energy[,3],lag.max=40, type="correlation", plot=TRUE, main = paste(col_names[,6]))

hydro_pacf <- Pacf(ts_energy[,3],lag.max=40, plot=TRUE,
                   main = paste(col_names[,6]))
```



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?

The total biomass production series and the total renewable energy production series show an increasing trend while the hydroelectric power consumption time series plot shows a decreasing trend. All three series graphs show a linear trend.

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:nrow(ts_energy))

#biomass lm
biomass_model <- lm(energy[,2]~t)
summary(biomass_model)
```

```
##
## Call:
## lm(formula = energy[, 2] ~ 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
```

```
#renewable lm
renewable_model <- lm(energy[,3]~t)
summary(renewable_model)
```

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

```
#hydroelectric lm
hydro_model <- lm(energy[,4]~t)
summary(hydro_model)
```

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

The results of the biomass linear model show that the intercept is 135.5 and the slope is 0.47, which means that the predicted total biomass energy production at the starting time (1973-01-01) is 135.5, and the total biomass energy production tends to increase 0.47 for every one unit change in month. This model accounts for 79.62% of variations in this dataset.

The results of the renewable linear model show that the intercept is 330.37 and the slope is 0.84. The intercept means that the predicted total renewable energy production at the starting time is 330.37, and the total renewable energy production tends to increase 0.84 for every one unit change in month. This model accounts for 68.82% of variations in the dataset.

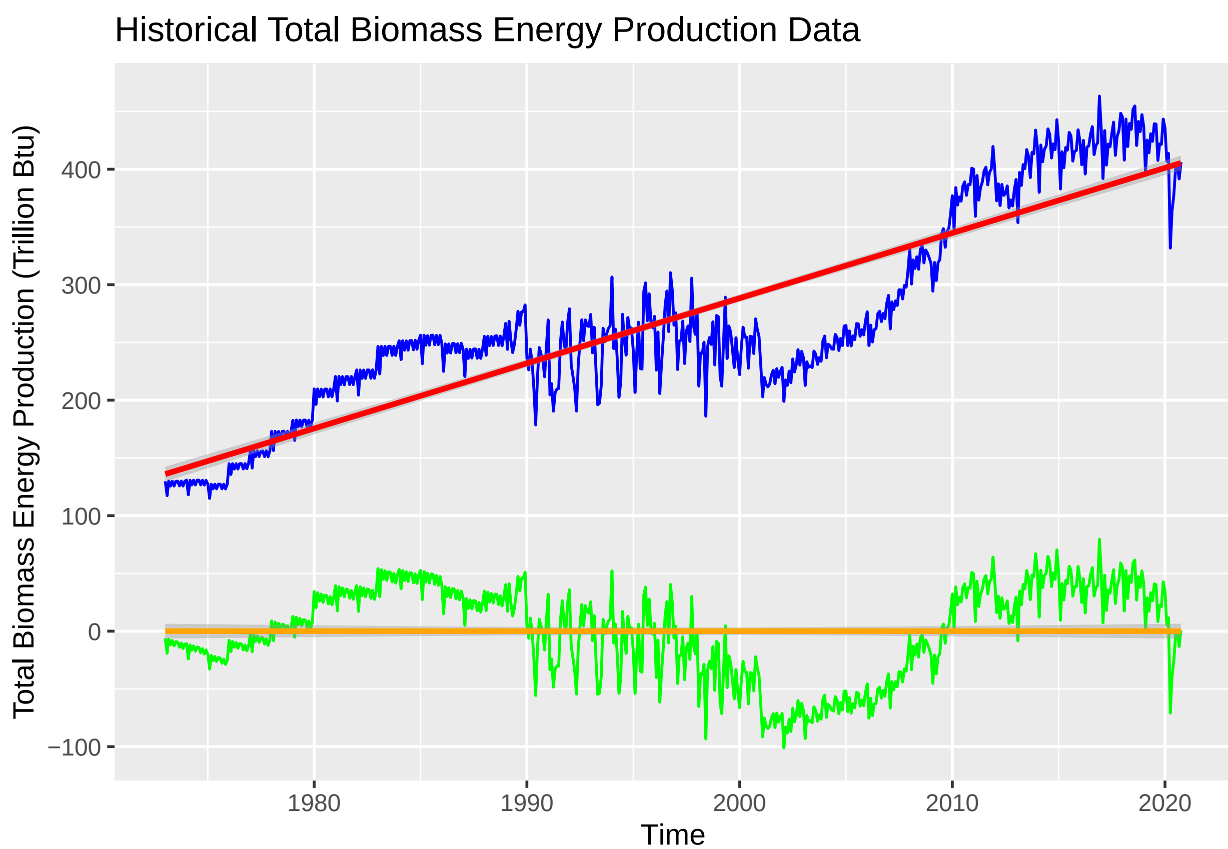
The results of the hydroelectric linear model show that the intercept is 258.06 and the slope is -0.07, which means that the predicted hydroelectric energy consumption at the starting time is 258.06 while the consumption tends to decrease 0.07 for every one nuit change in month. This model accounts for 7.53% of variations in the dataset.

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?

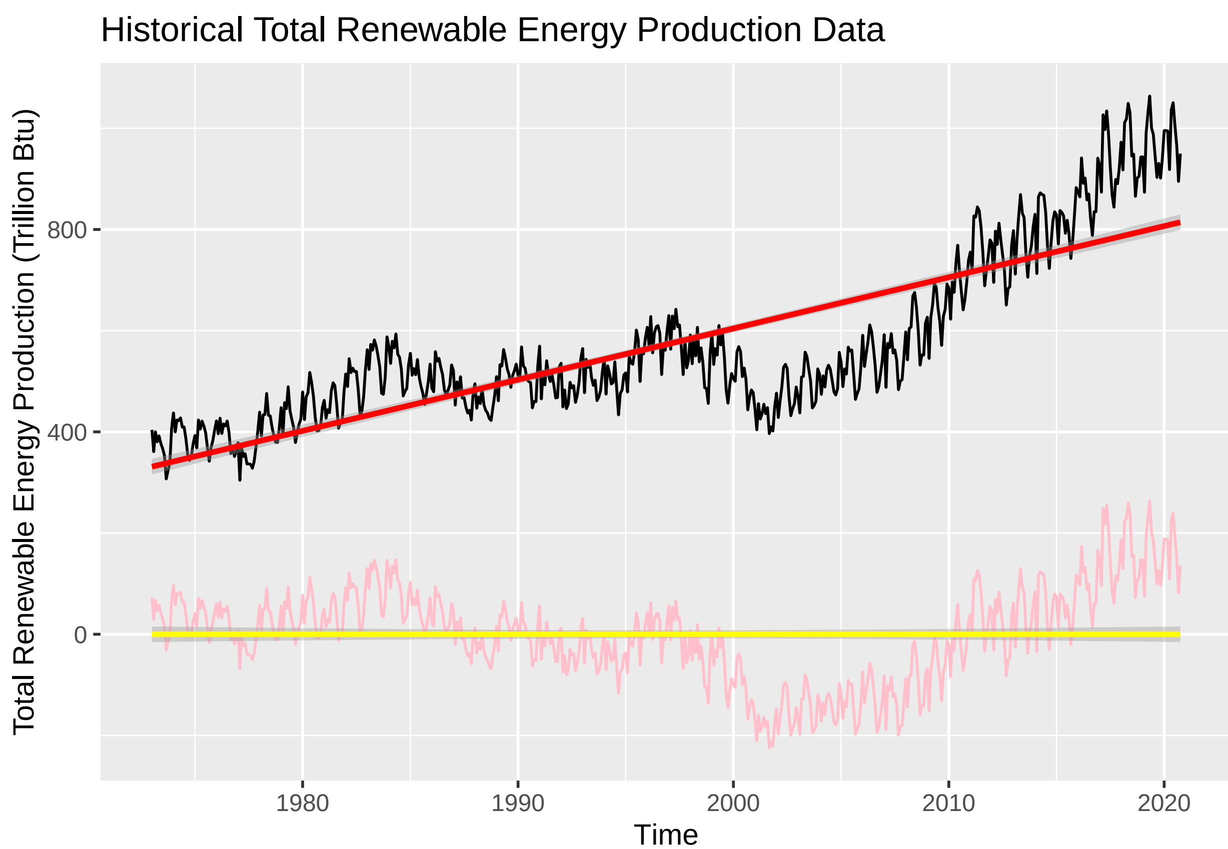
```r
#biomass coef
biomass_beta0 <- as.numeric(biomass_model$coefficients[1])  #intercept term
biomass_beta1 <- as.numeric(biomass_model$coefficients[2])  #slope

#renewable coef
renewable_beta0 <- as.numeric(renewable_model$coefficients[1])
renewable_beta1 <- as.numeric(renewable_model$coefficients[2])

#hydro coef
hydro_beta0 <- as.numeric(hydro_model$coefficients[1])
hydro_beta1 <- as.numeric(hydro_model$coefficients[2])

#detrend
detrend_biomass_data <- energy[,2]-(biomass_beta0+biomass_beta1*t)
detrend_renewable_data <- energy[,3]-(renewable_beta0+renewable_beta1*t)
detrend_hydro_data <- energy[,4]-(hydro_beta0+hydro_beta1*t)

#compare original and detrended plots
ggplot(energy, aes(x= energy[,1], y=energy[,2])) +
      geom_line(color="blue")+ xlab("Time") +
      ylab("Total Biomass Energy Production (Trillion Btu)") +
      labs(title = 'Historical Total Biomass Energy Production Data') +     geom_smooth(col="red",metho
      geom_line(aes(y=detrend_biomass_data), color="green") +      geom_smooth(aes(y=detrend_biomass_da
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



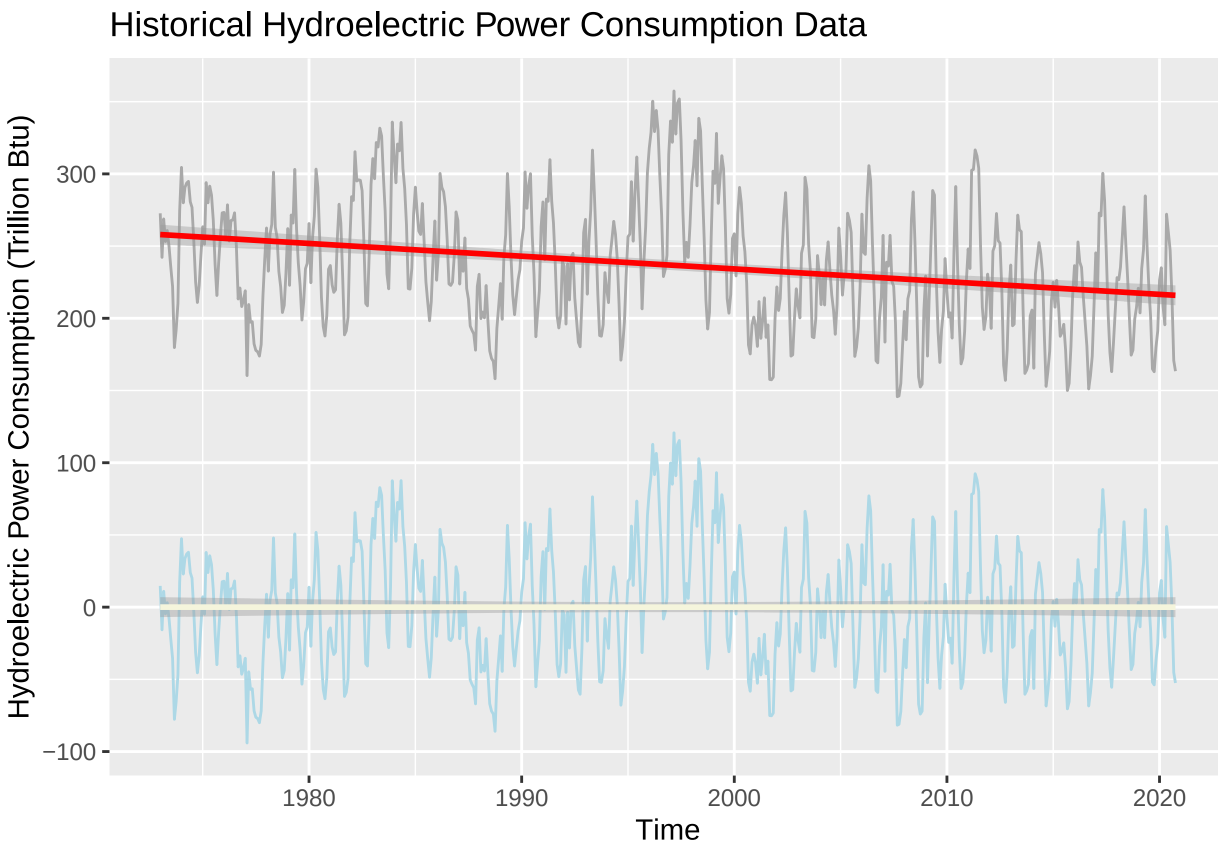

```
ggplot(energy, aes(x= energy[,1], y=energy[,3])) +
  geom_line(color="black")+ xlab("Time") +
  ylab("Total Renewable Energy Production (Trillion Btu)") +
  labs(title = 'Historical Total Renewable Energy Production Data') +    geom_smooth(col="red",method =
  geom_line(aes(y = detrend_renewable_data), color="pink") +
  geom_smooth(aes(y = detrend_renewable_data),col="yellow",method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(energy, aes(x= energy[,1], y=energy[,4])) +
      geom_line(color="dark gray")+ xlab("Time") +
      ylab("Hydroelectric Power Consumption (Trillion Btu)") +
      labs(title = 'Historical Hydroelectric Power Consumption Data') + geom_smooth(col="red",method =
      geom_line(aes(y=detrend_hydro_data), color="light blue") +
      geom_smooth(aes(y=detrend_hydro_data), col="beige",method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

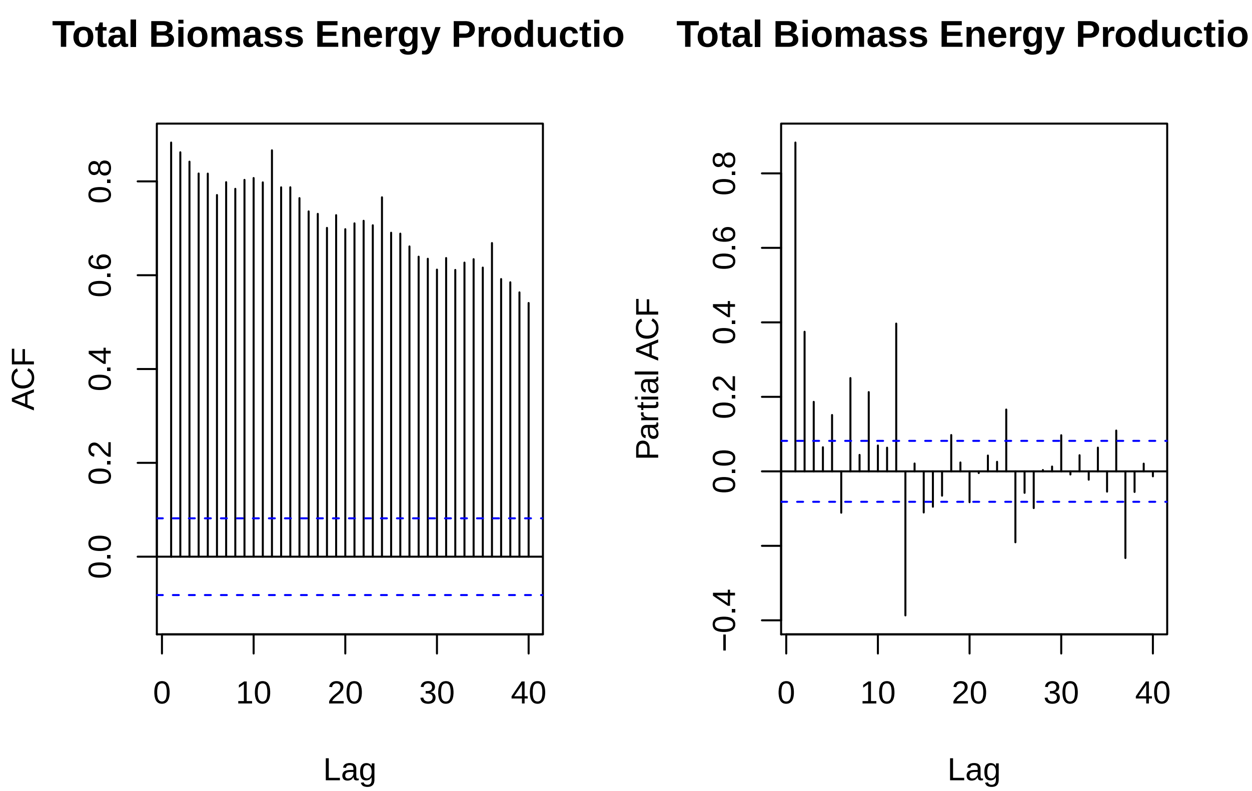


The trendlines are constant at 0 after detrending. While the oscillating patterns in the original time series plots are the same, the values were brought down to oscillate around 0.

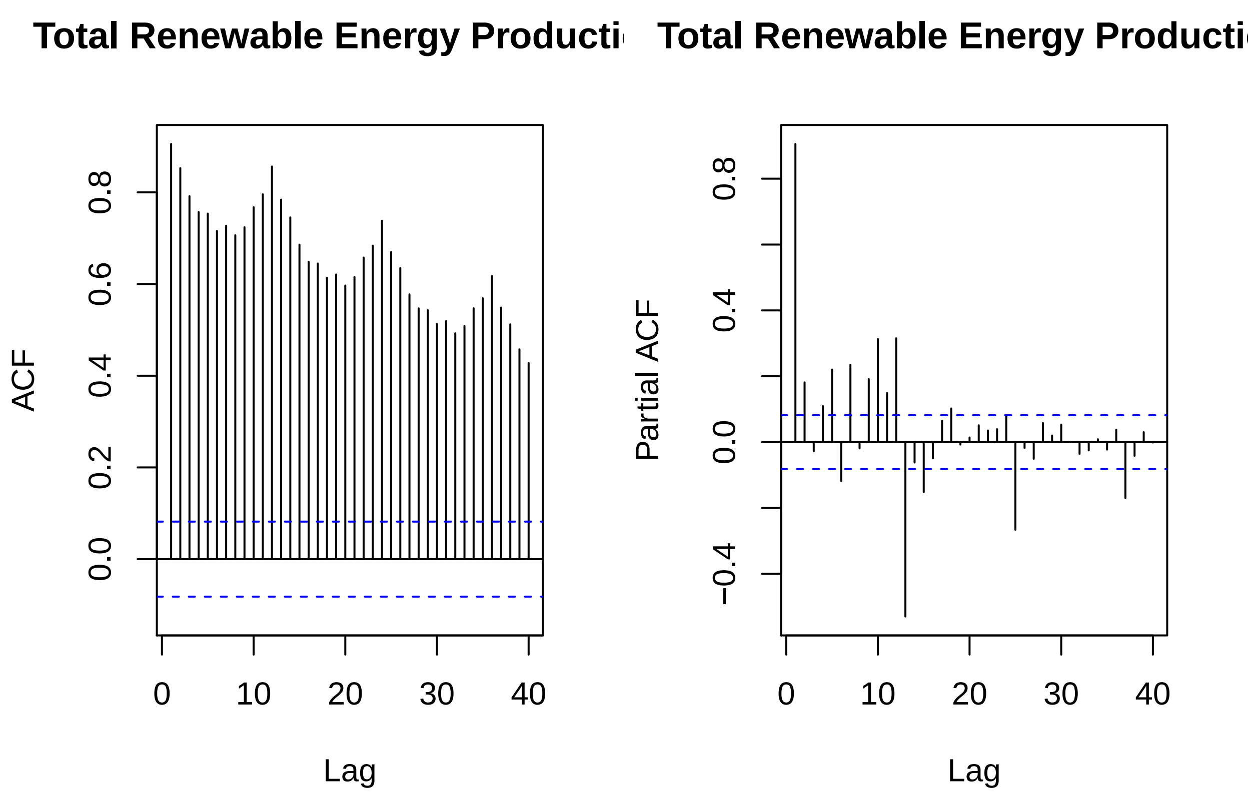
Q5

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

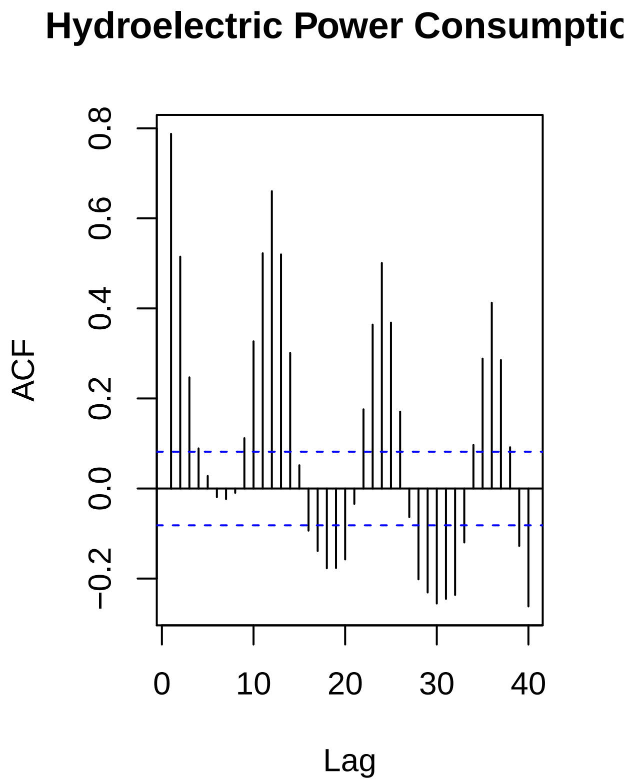
```
par(mfrow = c(1,2))
#biomass
detrend_biomass_acf <- Acf(detrend_biomass_data,lag.max=40, type="correlation",
                           plot=TRUE, main = paste(col_names[,4]))
detrend_biomass_pacf <- Pacf(detrend_biomass_data,lag.max=40, plot=TRUE, main = paste(col_names[,4]))
```

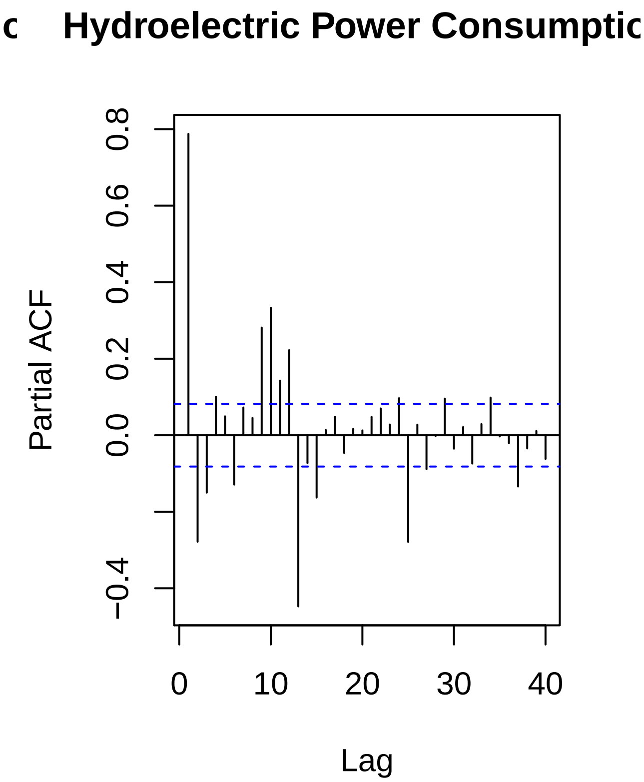



```
#renewable
detrend_renewable_acf <- Acf(detrend_renewable_data,lag.max=40, type="correlation",
                             plot=TRUE, main = paste(col_names[,5]))
detrend_renewable_pacf <- Pacf(detrend_renewable_data,lag.max=40, plot=TRUE, main = paste(col_names[,5]
```



```r
#hydro
detrend_hydro_acf <- Acf(detrend_hydro_data,lag.max=40, type="correlation",
                         plot=TRUE, main = paste(col_names[,6]))
detrend_hydro_pacf <- Pacf(detrend_hydro_data,lag.max=40, plot=TRUE, main = paste(col_names[,6]))
```





The PACF plots barely change, and the ACF plot for hydroelectric energy consumption did not change much. The ACF plots for renewable energy production and biomass energy production show a more pronounced cyclical/seasonal change (the curvy pattern) with less ACF values.

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.

```r
#set dummies
dummies <- seasonaldummy(ts_energy[,1])

#fit seasonal means model for hydroelectric data
seas_hydro_model <- lm(energy[,4]~dummies)
summary(seas_hydro_model)
```

```
##
```

```
## Call:
## lm(formula = energy[, 4] ~ 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
#hydro season coef
hydro_beta_int=seas_hydro_model$coefficients[1]
hydro_beta_coeff=seas_hydro_model$coefficients[2:12]

#fit seasonal means model for renewable data
seas_renew_model <- lm(energy[,3]~dummies)
summary(seas_renew_model)
```

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

```
#renewable season coef
renew_beta_int=seas_renew_model$coefficients[1]
renew_beta_coeff=seas_renew_model$coefficients[2:12]

#fit seasonal means model for biomass data
seas_biomass_model <- lm(energy[,2]~dummies)
summary(seas_biomass_model)
```

```
##
## Call:
## lm(formula = energy[, 2] ~ 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
```

```
#biomass season coef
biomass_beta_int=seas_biomass_model$coefficients[1]
biomass_beta_coeff=seas_biomass_model$coefficients[2:12]
```

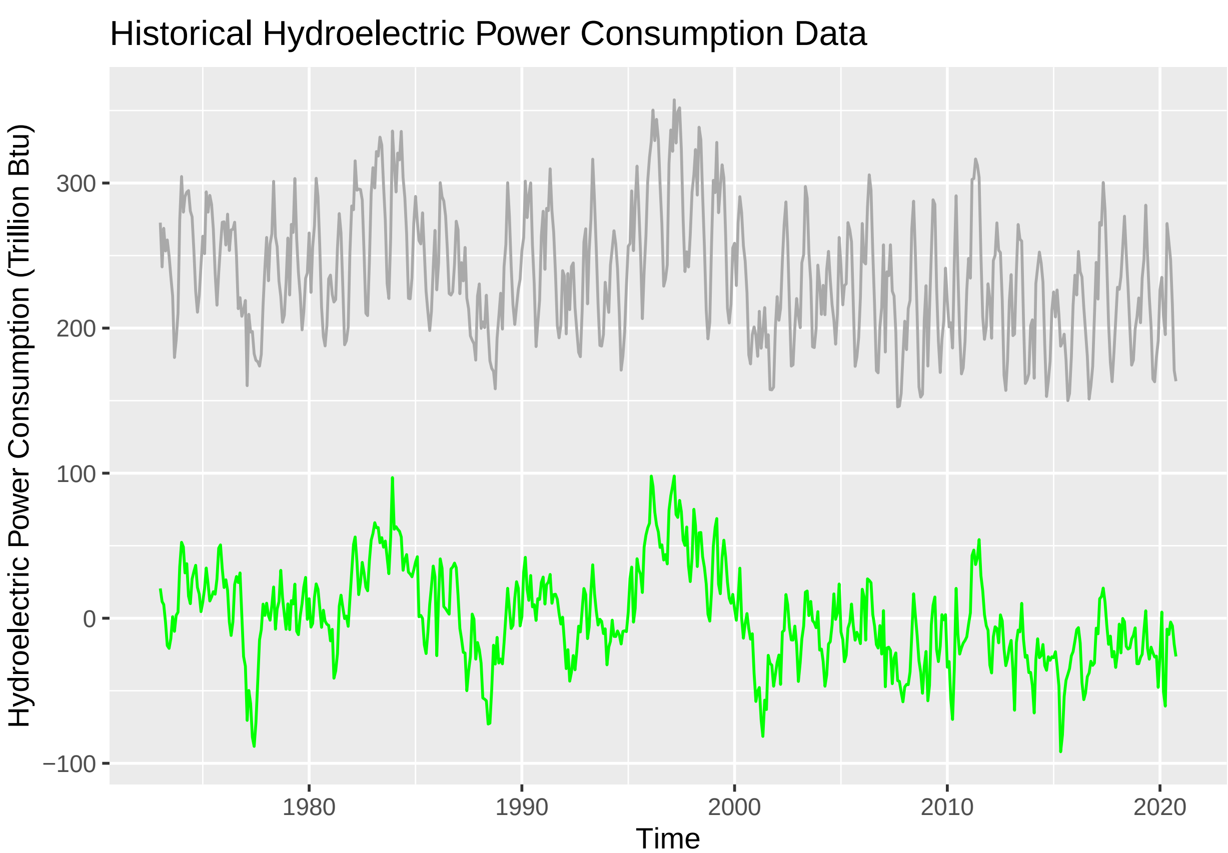
The hydroelectric energy consumption series plot shows the most noticeable seasonal trend while the seasonality in renewable energy production and biomass energy production series plots are not that obvious.

The regression results show that differences between each month and December are not significant for total biomass energy production and total renewable energy production.
The intercept in the hydroelectric energy consumption models represents the mean consumption for December, which is 238.89 trillion Btu. The dummy coefficients for other 11 months mean that the mean consumption is 13.27 more, 8.13 less, 20.44 more, 17.2 more, 40.73 more, 31.76 more, 10.86 more, 17.91 less, 50.12 less, 49.17 less, 32.76 less than that of December for Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct, and Nov respectively.

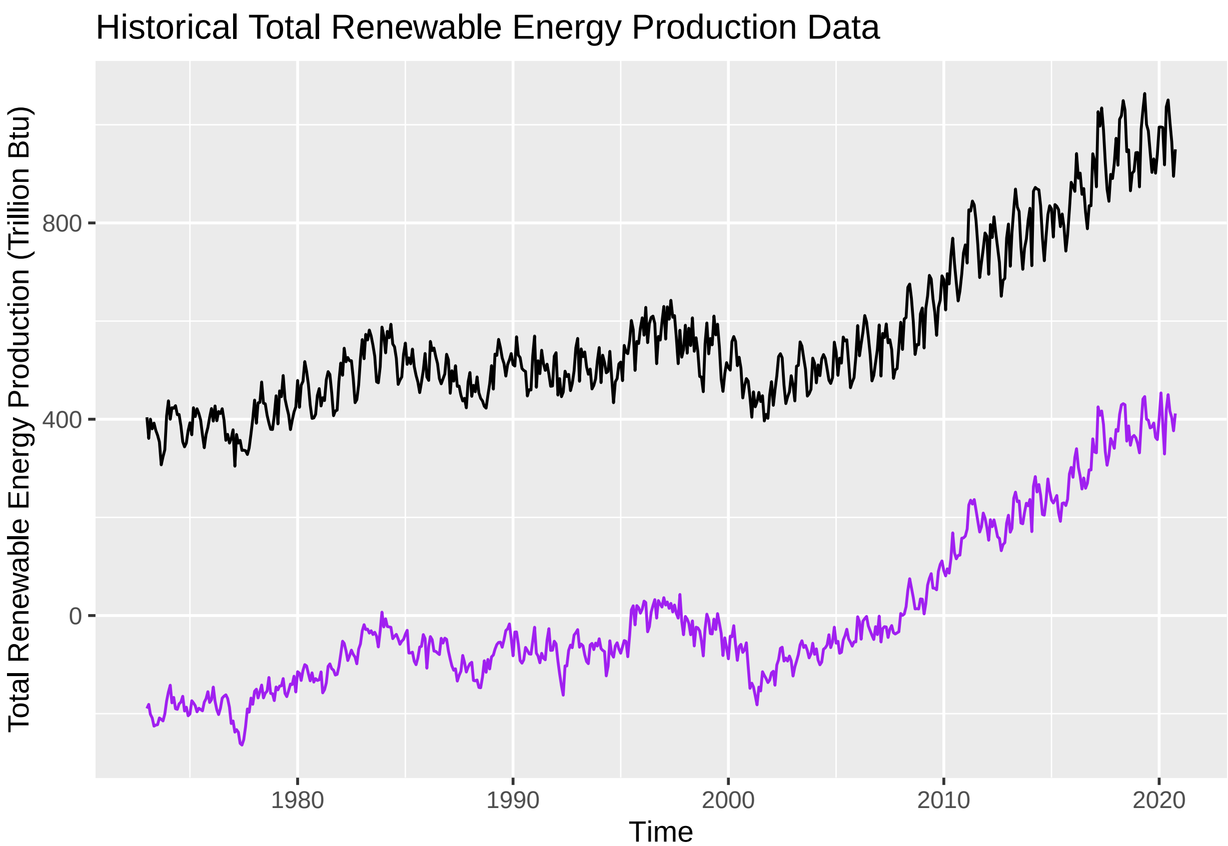
Q7

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?

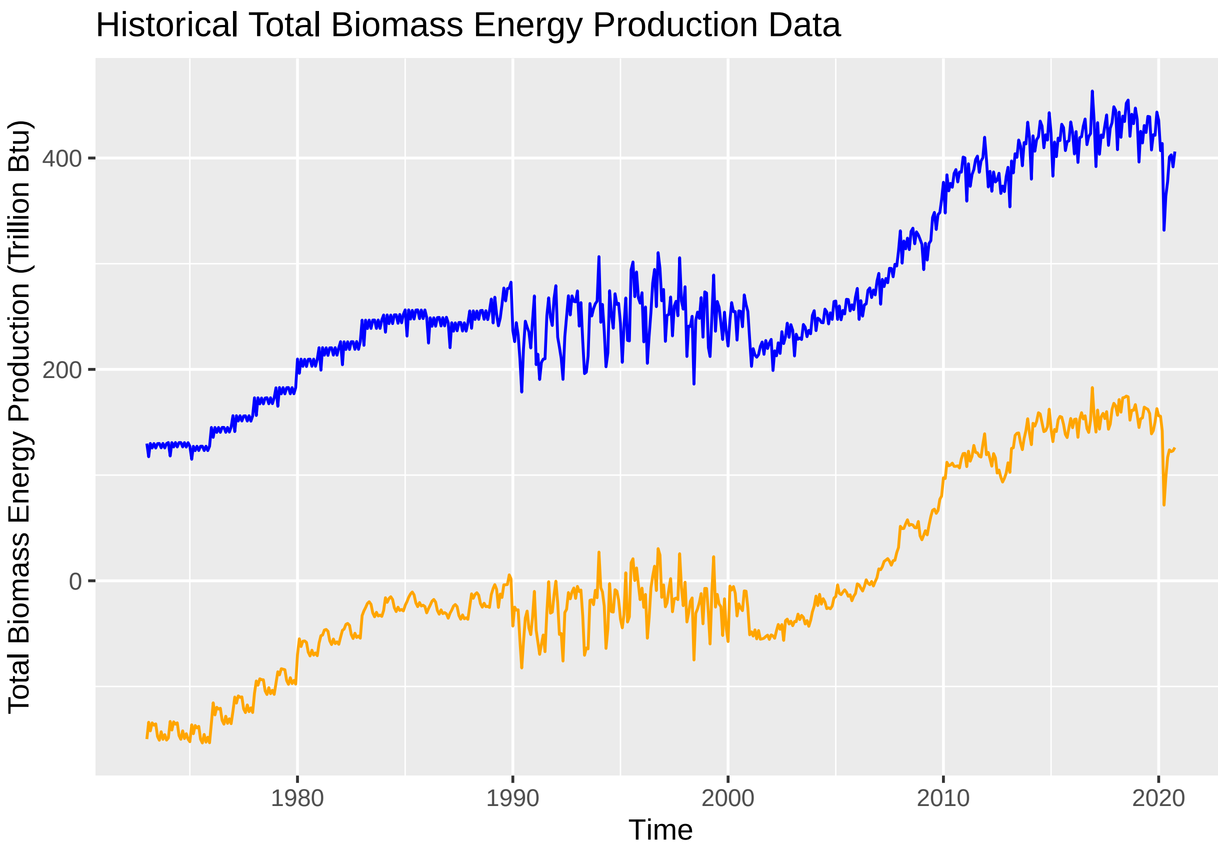
```
#deseason hydro
hydro_seas_comp=array(0,nrow(ts_energy))
for(i in 1:nrow(ts_energy)){
  hydro_seas_comp[i]=(hydro_beta_int+hydro_beta_coeff%*%dummies[i,])
}
deseason_hydro_data <- ts_energy[,3]-hydro_seas_comp
ggplot(energy, aes(x= energy[,1], y=energy[,4])) +
  geom_line(color="dark gray")+ xlab("Time") +
  ylab("Hydroelectric Power Consumption (Trillion Btu)") +
  labs(title = 'Historical Hydroelectric Power Consumption Data') +
  geom_line(aes(y = deseason_hydro_data),col="green")
```



```
#deseason renewable
renew_seas_comp=array(0,nrow(ts_energy))
for(i in 1:nrow(ts_energy)){
  renew_seas_comp[i]=(renew_beta_int+renew_beta_coeff%*%dummies[i,])
}
deseason_renew_data <- ts_energy[,2]-renew_seas_comp
ggplot(energy, aes(x= energy[,1], y=energy[,3])) +
  geom_line(color="black")+ xlab("Time") +
  ylab("Total Renewable Energy Production (Trillion Btu)") +
  labs(title = 'Historical Total Renewable Energy Production Data') +
  geom_line(aes(y = deseason_renew_data),col="purple")
```




```
#deseason biomass
biomass_seas_comp=array(0,nrow(ts_energy))
for(i in 1:nrow(ts_energy)){
  biomass_seas_comp[i]=(biomass_beta_int+biomass_beta_coeff%*%dummies[i,])
}
deseason_biomass_data <- ts_energy[,1]-biomass_seas_comp
ggplot(energy, aes(x= energy[,1], y=energy[,2])) +
  geom_line(color="blue")+ xlab("Time") +
  ylab("Total Biomass Energy Production (Trillion Btu)") +
  labs(title = 'Historical Total Biomass Energy Production Data') +
  geom_line(aes(y = deseason_biomass_data),col="orange")
```

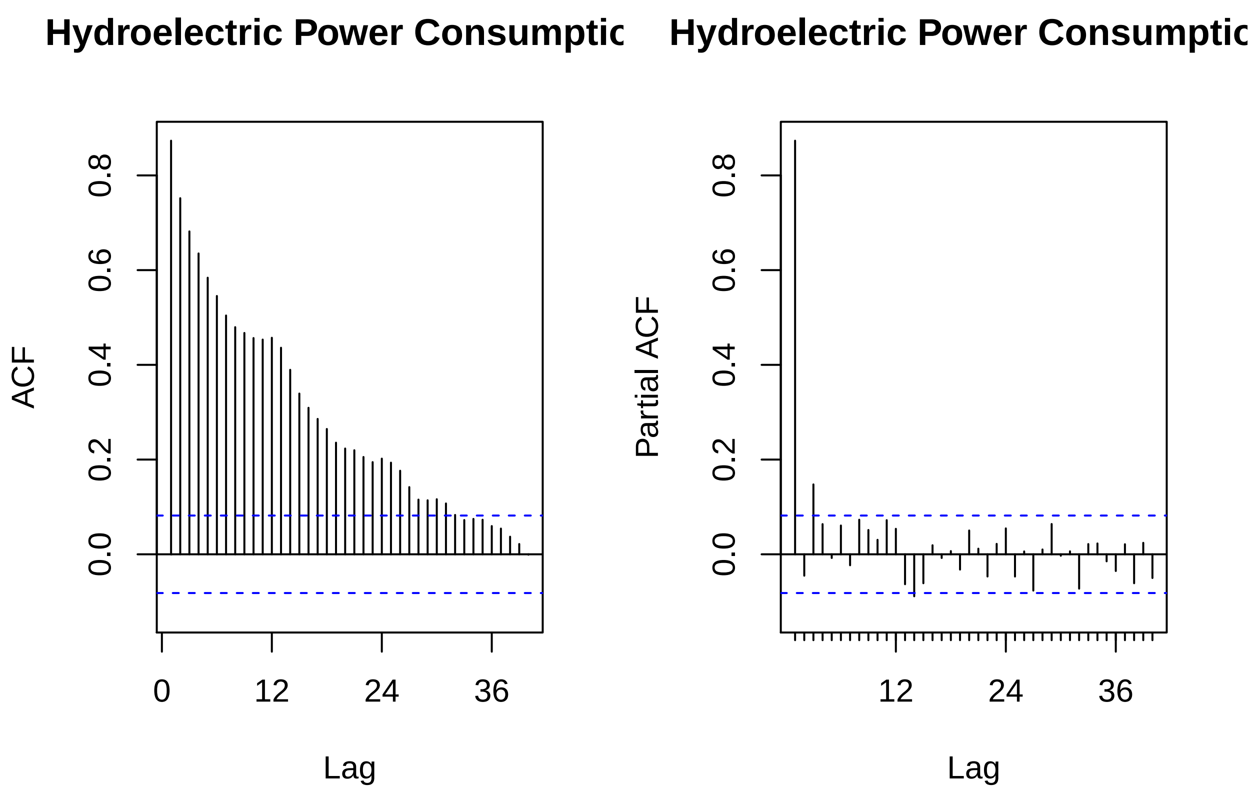


The values after deseasoning are less than that of original one. The seasonal trend is weakened as shown by more irregular oscillations in the deseasoned time series plot.

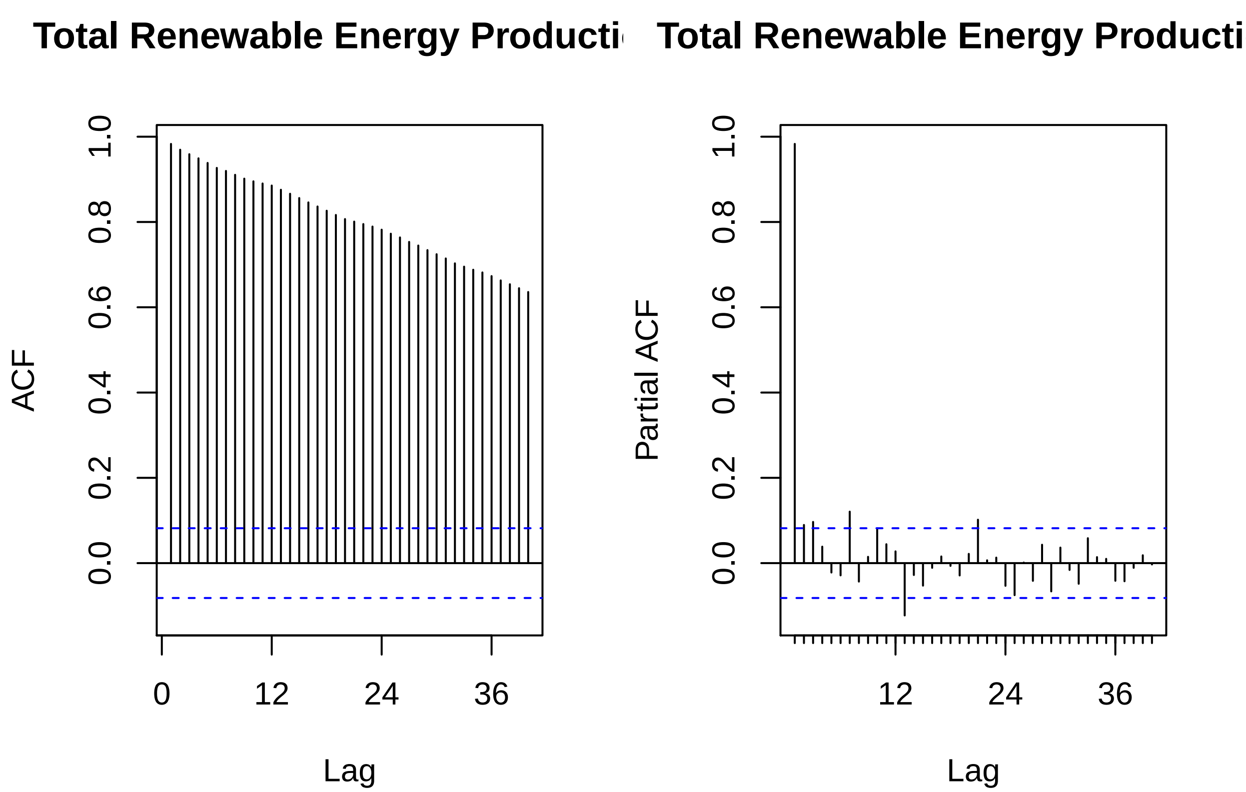
Q8

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

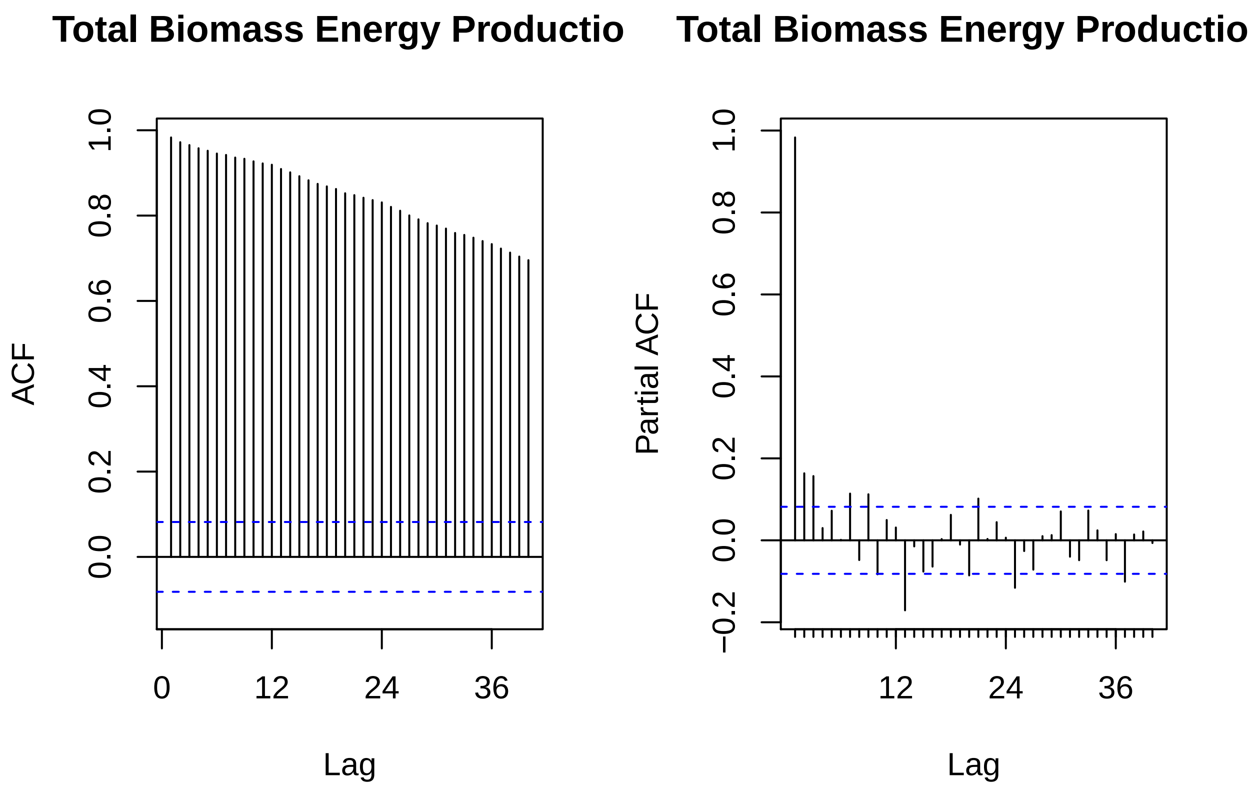
```
par(mfrow = c(1,2))
deseason_hydro_acf <- Acf(deseason_hydro_data,lag.max=40, type="correlation",
                          plot=TRUE, main = paste(col_names[,6]))

deseason_hydro_pacf <- Pacf(deseason_hydro_data,lag.max=40, plot=TRUE, main = paste(col_names[,6]))
```




```
deseason_renew_acf <- Acf(deseason_renew_data,lag.max=40, type="correlation",
                          plot=TRUE, main = paste(col_names[,5]))

deseason_renew_pacf <- Pacf(deseason_renew_data,lag.max=40, plot=TRUE, main = paste(col_names[,5]))
```



```
deseason_biomass_acf <- Acf(deseason_biomass_data,lag.max=40, type="correlation",
                          plot=TRUE, main = paste(col_names[,4]))

deseason_biomass_pacf <- Pacf(deseason_biomass_data,lag.max=40, plot=TRUE, main = paste(col_names[,4]))
```



The curvy pattern in ACF is weakened, and negative ACF no longer exist. The absolute values of PACF were less than the original ones.The range of both ACF and PACF appear to be less than before.