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

Assignment 3 - Due date 02/15/21

Stefan Chen

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

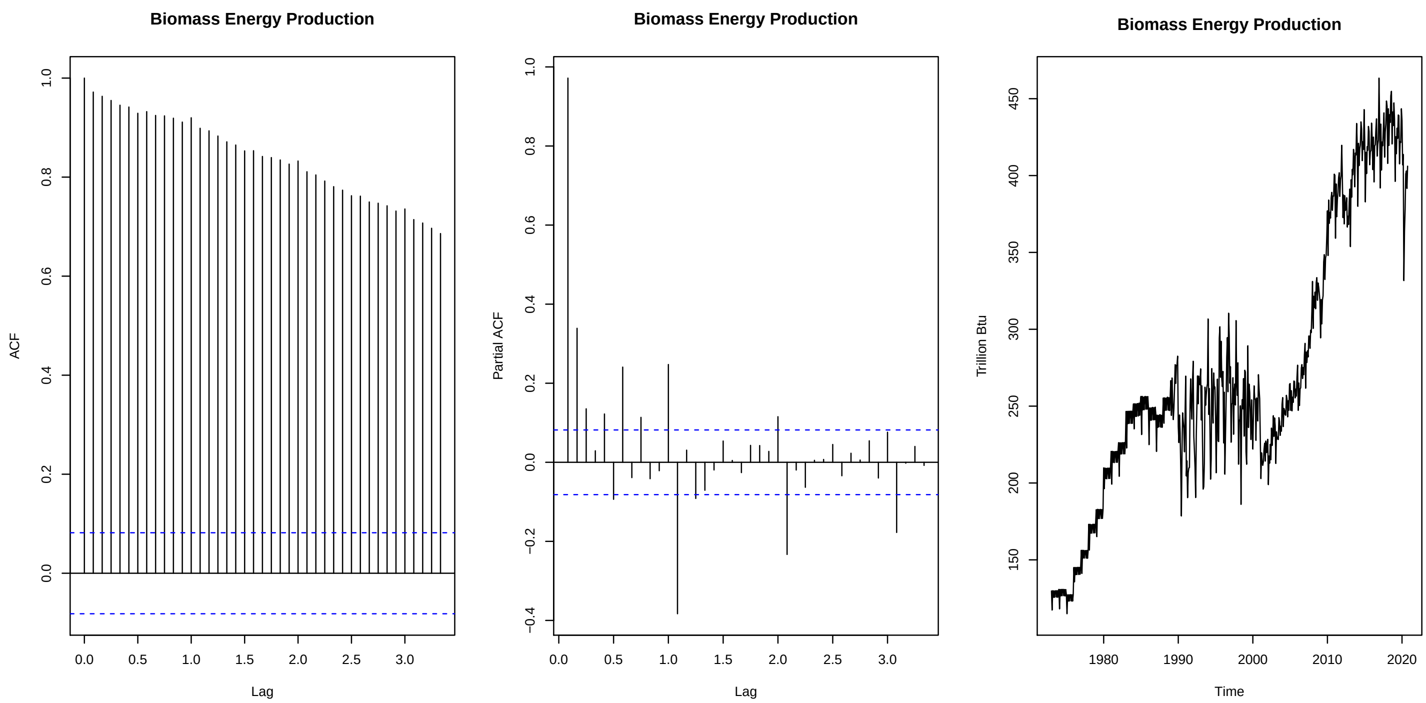
Consider the same data you used for A2 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_ The data comes from the US Energy Information and Administration and corresponds to the January 2021 Monthly Energy Review. Once again you will work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series only.

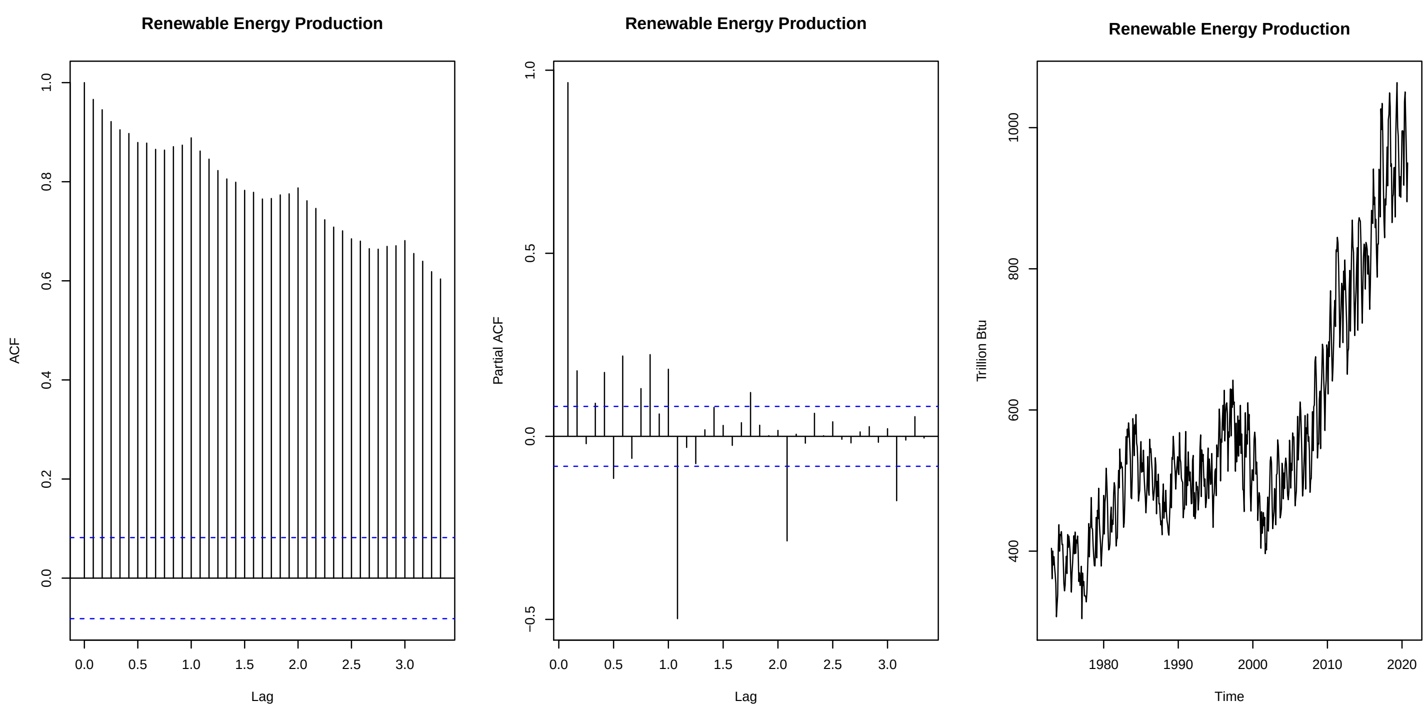
R packages needed for this assignment:"forecast","tseries", and "Kendall". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

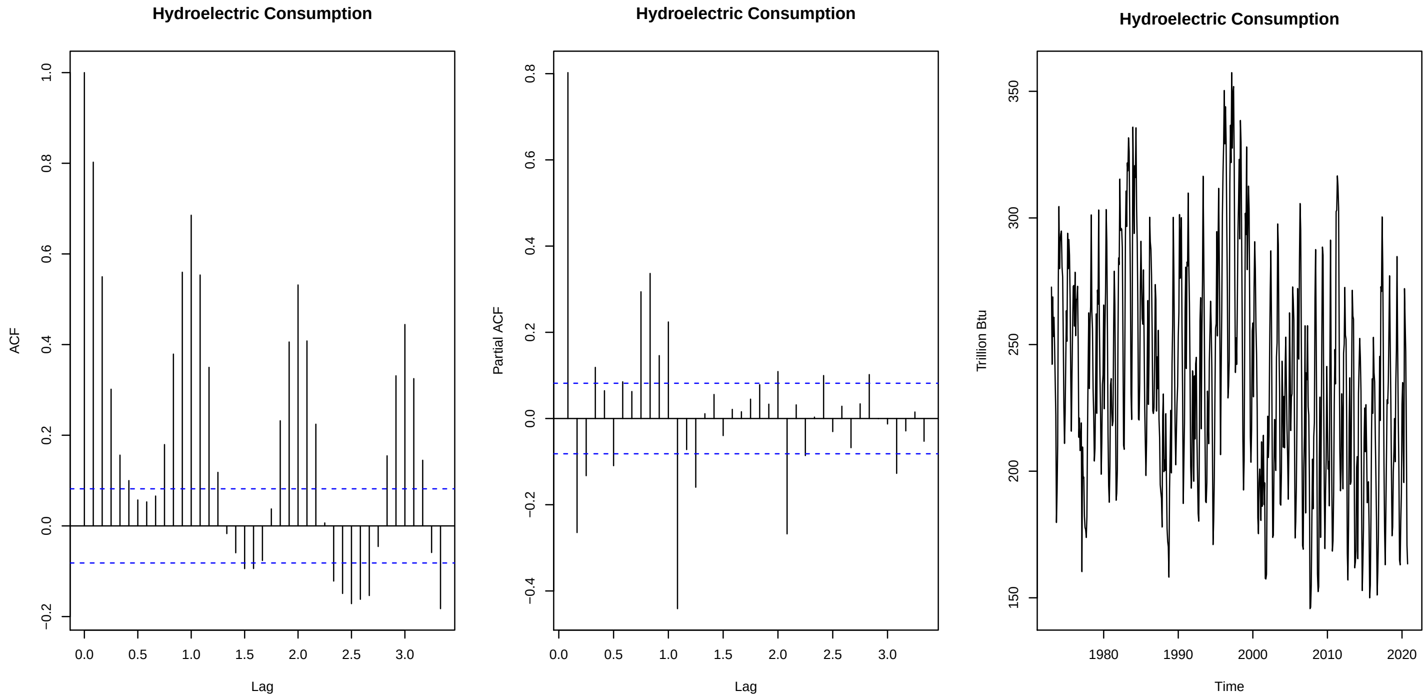
Trend Component

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)







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 show a growing trend between 1973 and 2020. However, both energy production appear to be stagnant between 1985 and 2000. On the other hand, hydroelectric power consumption shows a slight decreasing 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.

Summary of linear model: biomass energy production

```
##
## Call:
## lm(formula = ts_select[, 1] ~ 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
```

Summary of linear model: renewable energy production

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

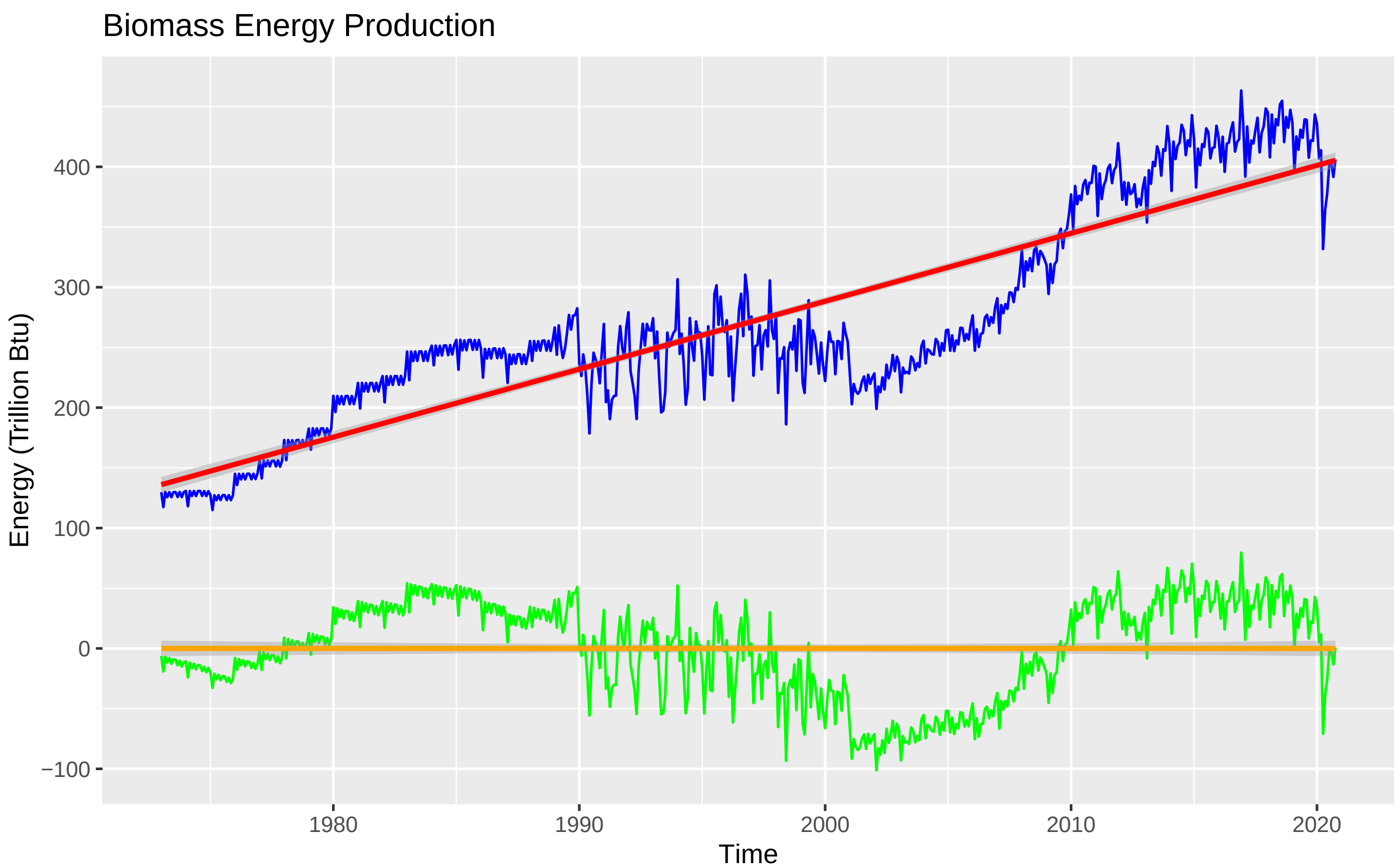
Summary of linear model: renewable energy production

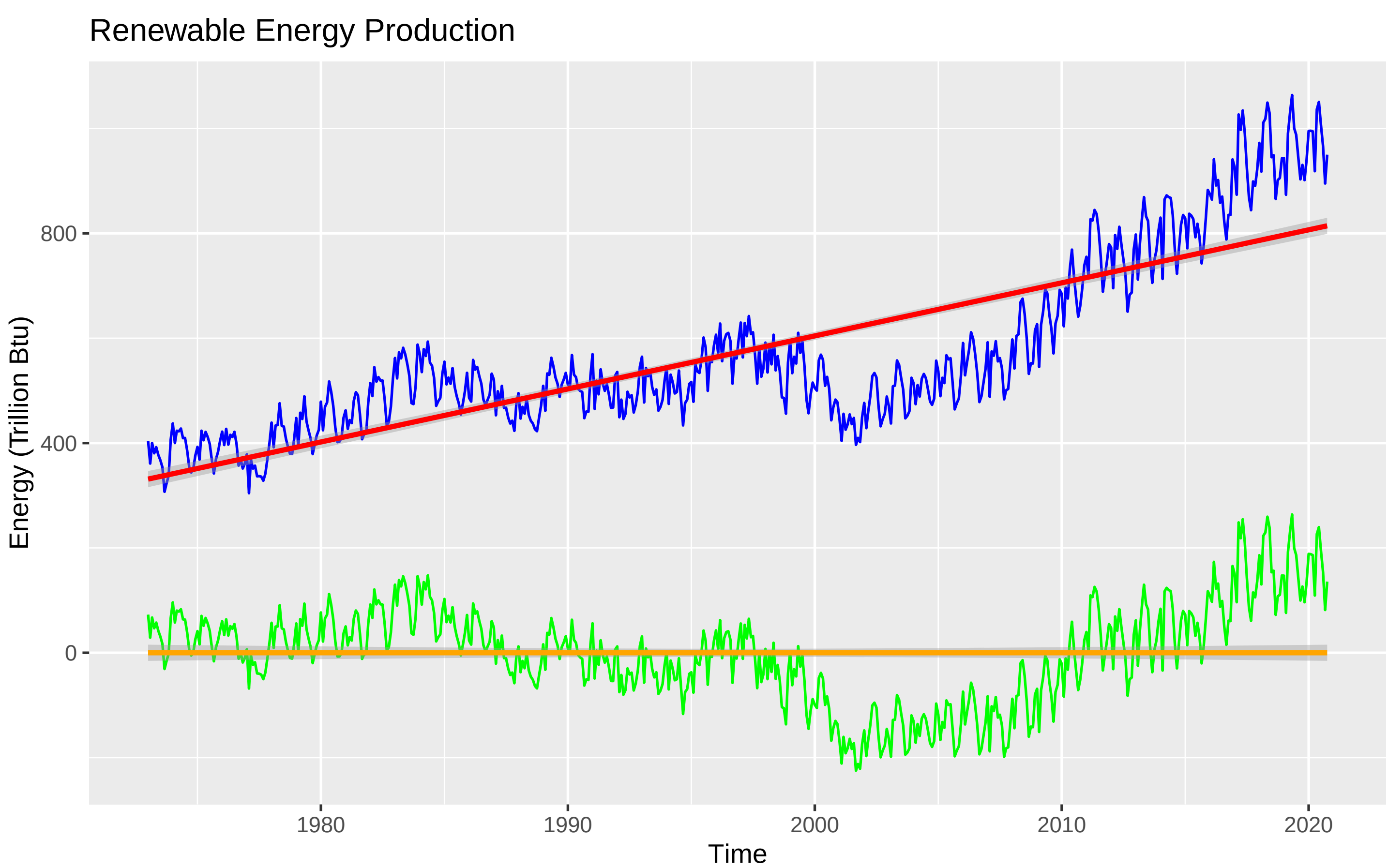
```
##
## Call:
## lm(formula = ts_select[, 3] ~ t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -94.06 -31.57  -1.63  27.73 120.69
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 258.05622    3.52899  73.125  < 2e-16 ***
## t            -0.07341    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
```

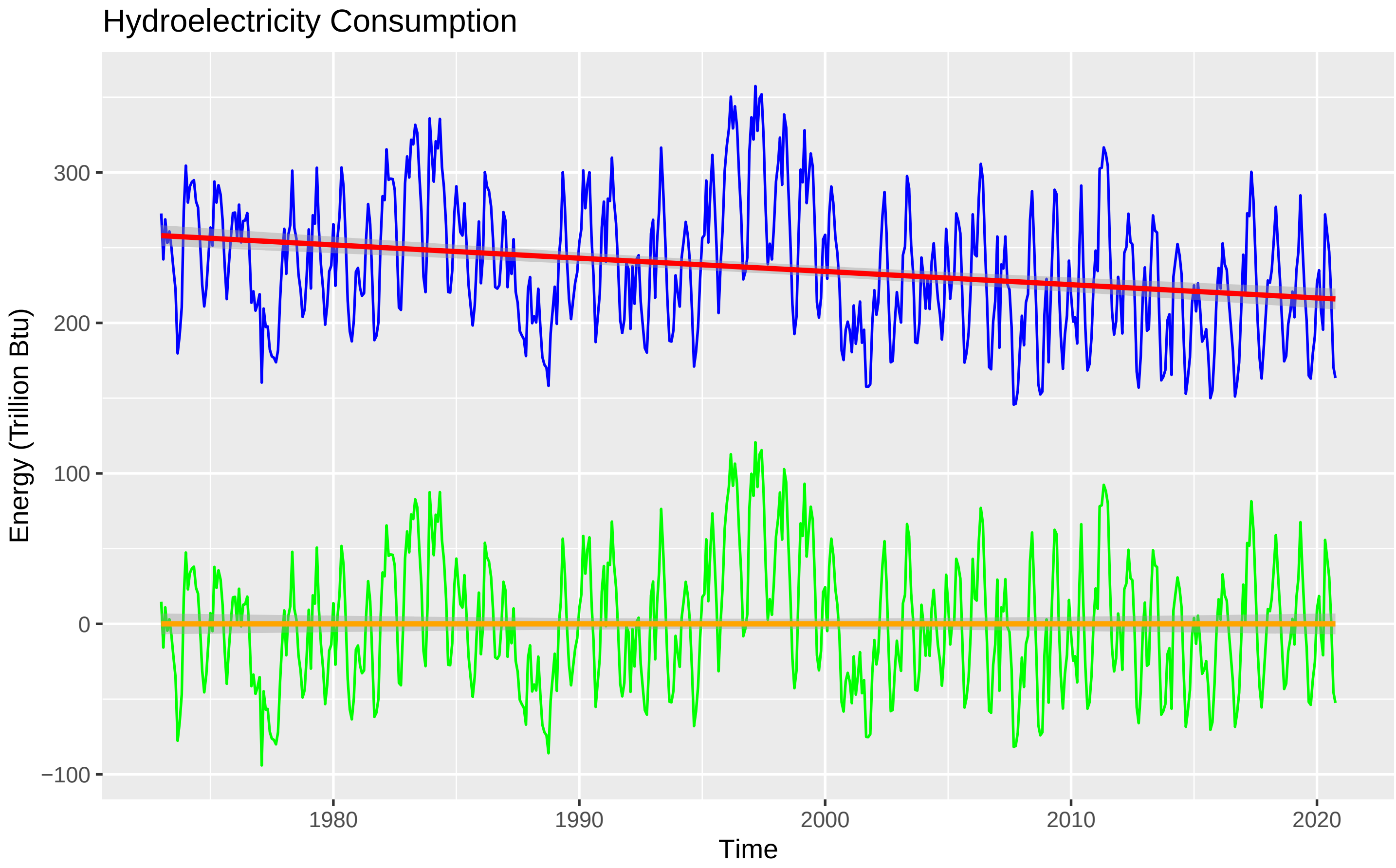
Q4

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?

New plots eliminated the upward trends in biomass and renewable energy production, and the downward trend in hydroelectricity consumption.


Biomass Energy Production
Energy (Trillion Btu)
400
300
200
100
0
−100
1980
1990
2000
2010
2020
Time

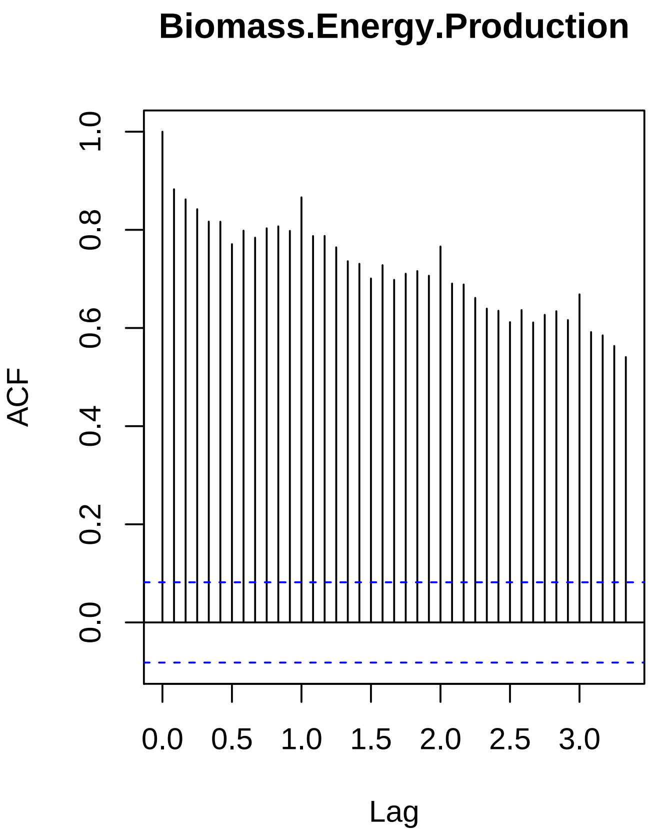

Renewable Energy Production
Energy (Trillion Btu)
800
400
0
1980
1990
2000
2010
2020
Time

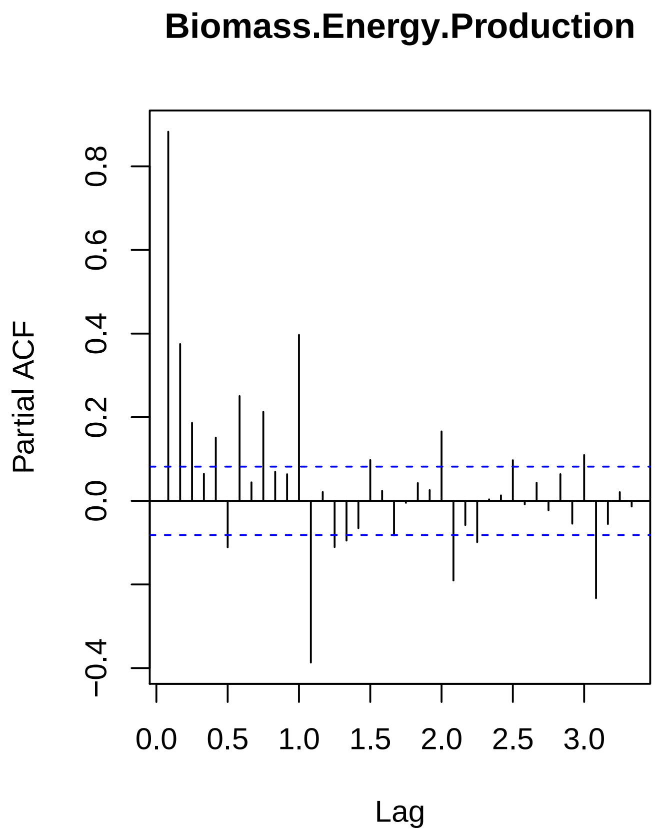


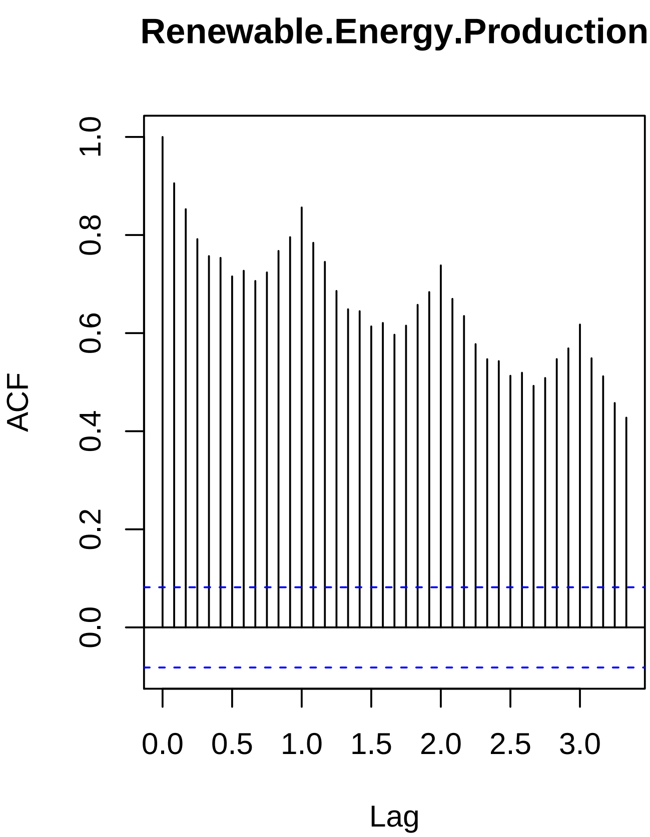
Q5

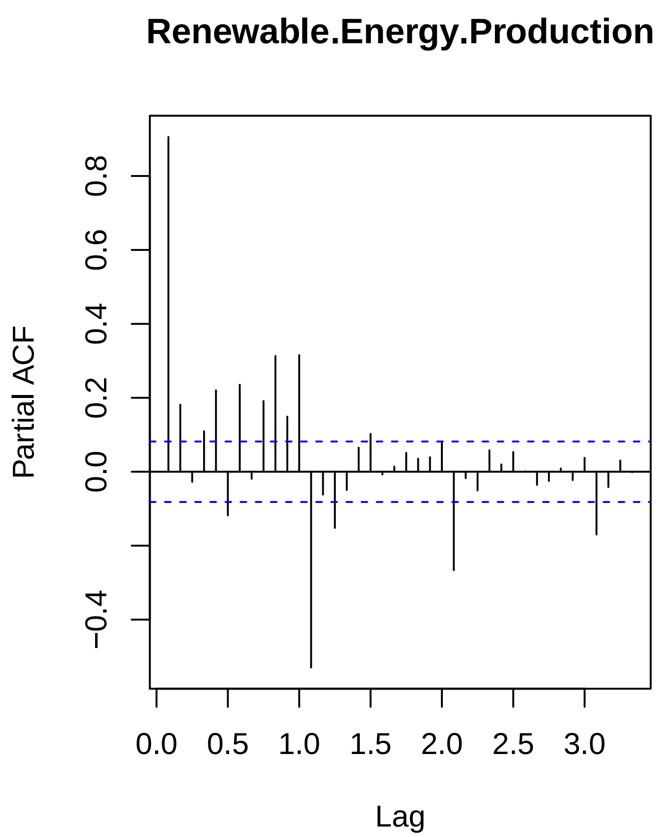
Plot ACF and PACF for the detrended series and compare with the plots from Q1. Did the plots change? How?

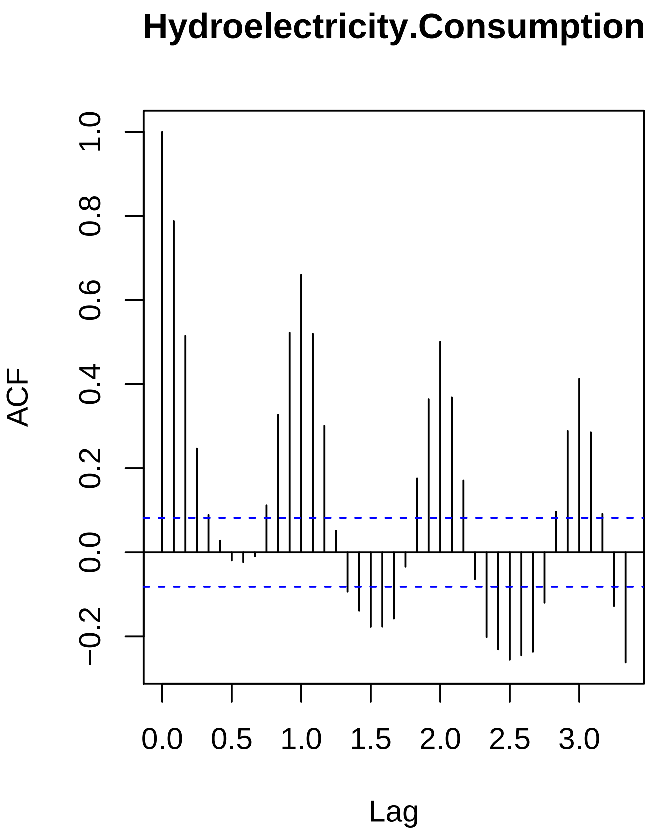
The autoregression of biomass and renewable energy production appears to fluctuate slightly more than the original autoregression plot although still not passing autoregression of zero.

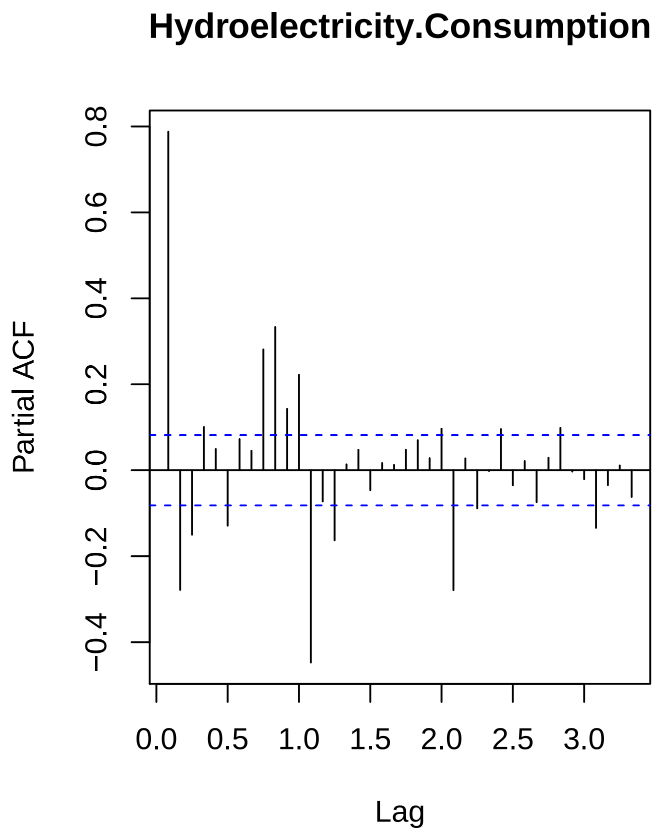

Biomass.Energy.Production
ACF
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0


Biomass.Energy.Production
Partial ACF
Lag
−0.4
0.0
0.2
0.4
0.6
0.8
0.0 0.5 1.0 1.5 2.0 2.5 3.0


Renewable.Energy.Production
ACF
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0


Renewable.Energy.Production
Partial ACF
Lag
−0.4
0.0
0.2
0.4
0.6
0.8
0.0 0.5 1.0 1.5 2.0 2.5 3.0





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.

Biomass Energy Production

Based on the regression summary of biomass energy production, only coefficient for December appears to show statistically significant effect on the model. Therefore, it can be concluded that there's no significant seasonal trend.

```
##
## Call:
## lm(formula = re_select[, 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
```

Renewable Energy Production

Based on the regression summary of renewable energy production, only coefficient for December appears to show statistically significant effect on the model. Therefore, it can be concluded that there's no significant seasonal trend.

```
##
## Call:
## lm(formula = re_select[, 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
```

Hydroelectricity Consumption

Based on the regression summary of hdroelectricity consumption, nine out of twelve months appears to show statistically significant effect on the model. Therefore, it can be concluded that there is a significant seasonal trend.

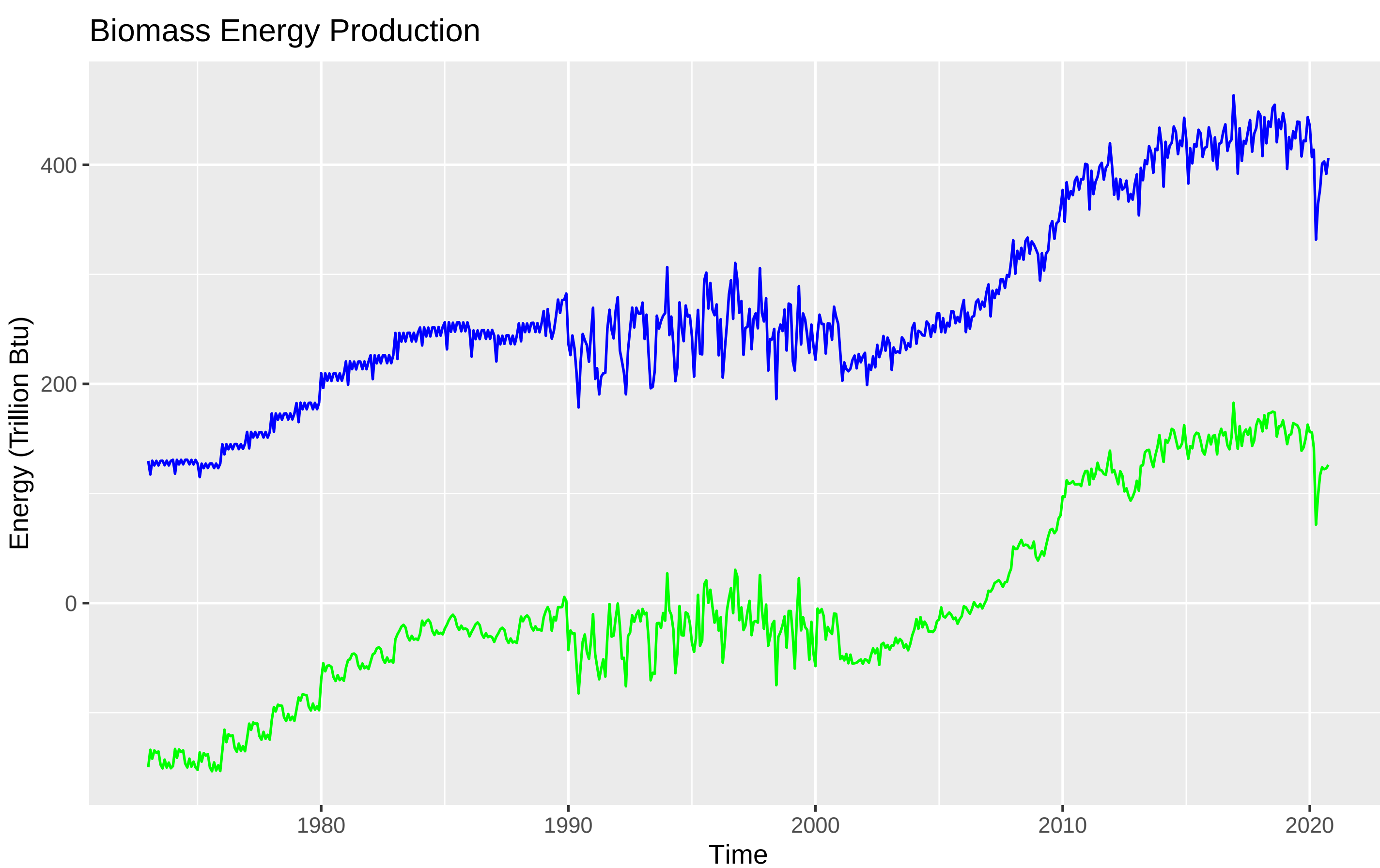
```
##
## Call:
## lm(formula = re_select[, 4] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.064 -22.897  -2.654  20.642  98.058
```

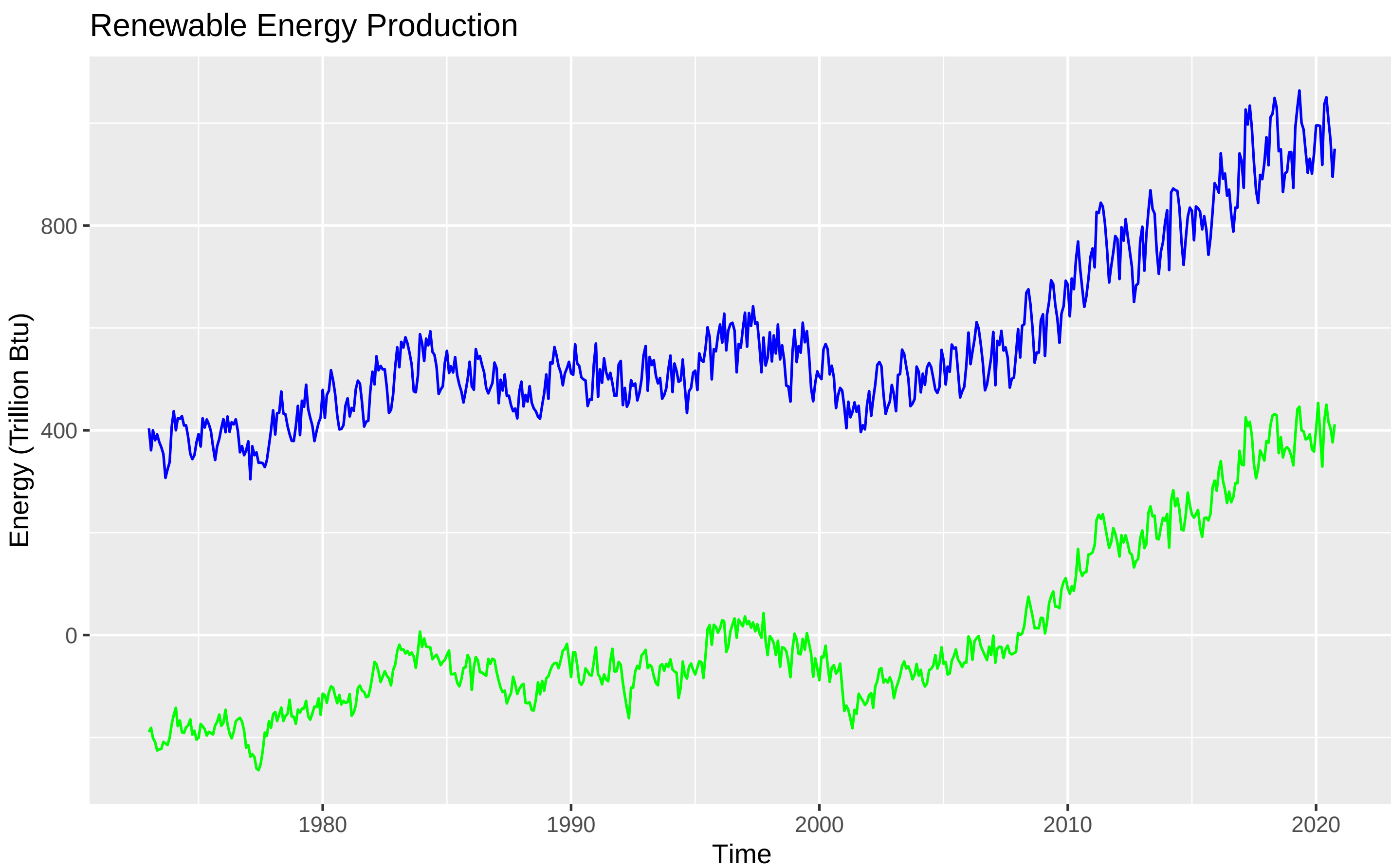
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  238.887      4.863  49.125  < 2e-16 ***
## dummiesJan    13.270      6.841   1.940  0.05291 .
## dummiesFeb    -8.133      6.841  -1.189  0.23499
## dummiesMar    20.442      6.841   2.988  0.00293 **
## dummiesApr    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
```

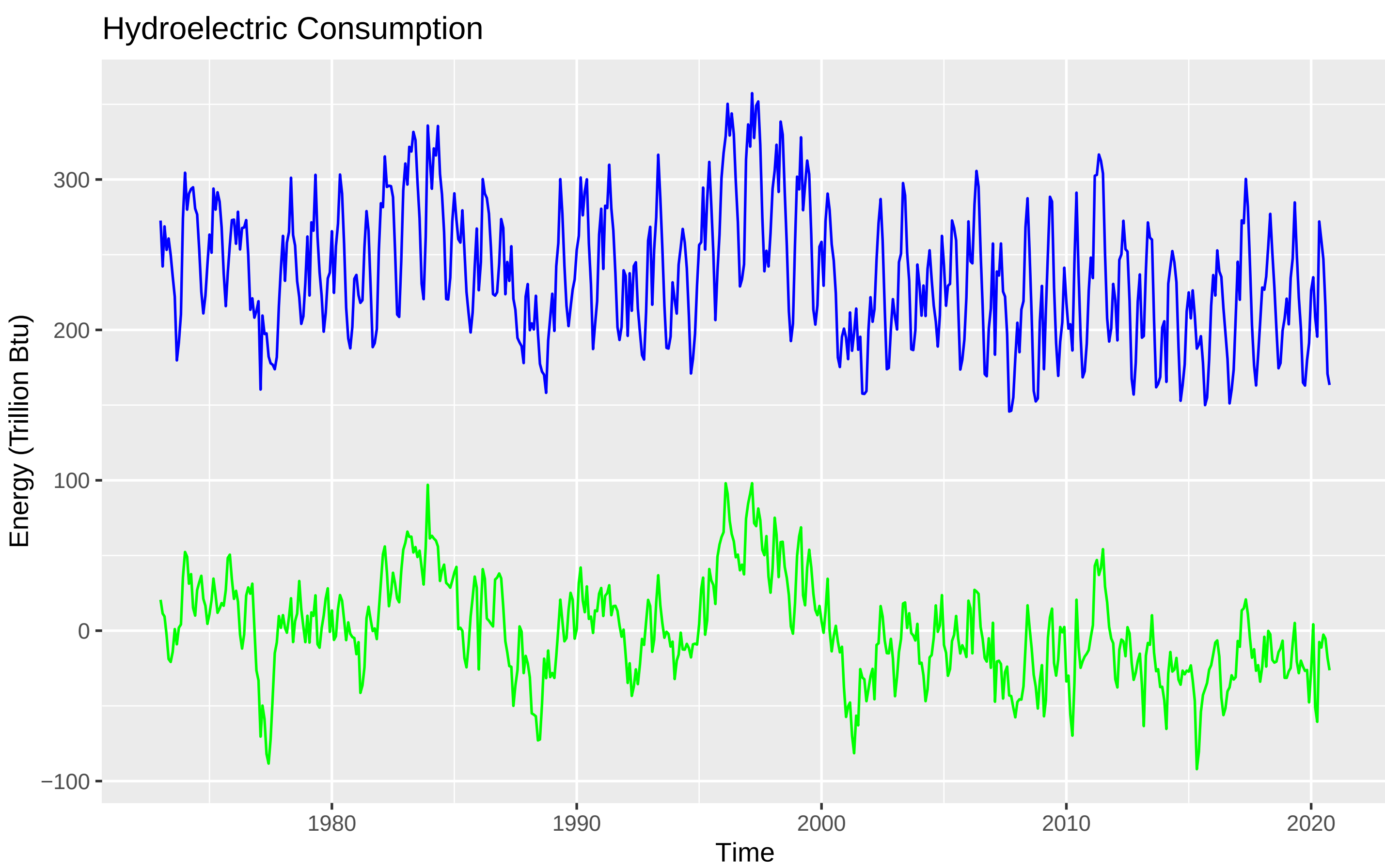
Q7

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?

Overall flunctuation appears to be smoothened for all plots. However, the general upward and downward trend observed from the original plots remain observale.



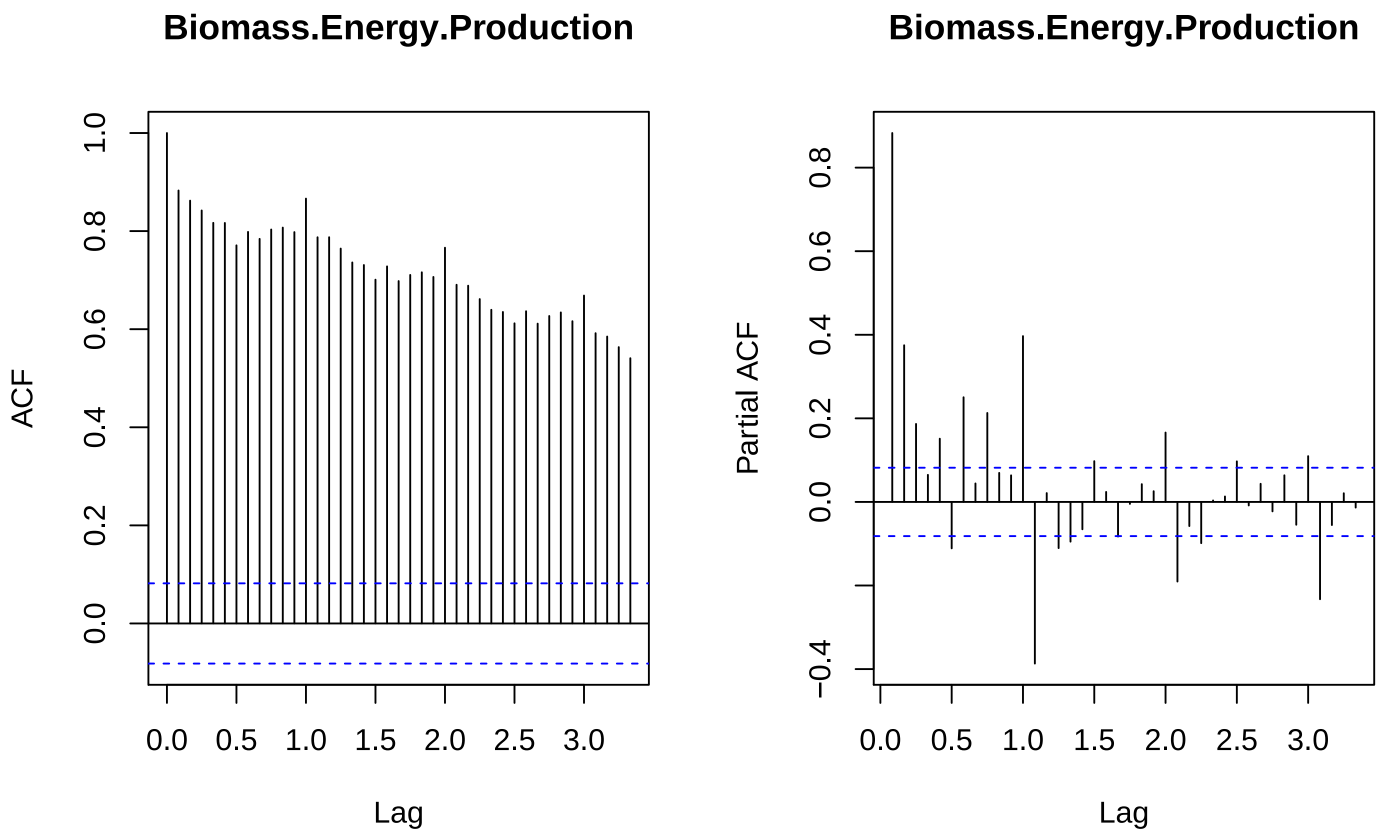

Renewable Energy Production
Energy (Trillion Btu)
800
400
0
1980
1990
2000
2010
2020
Time

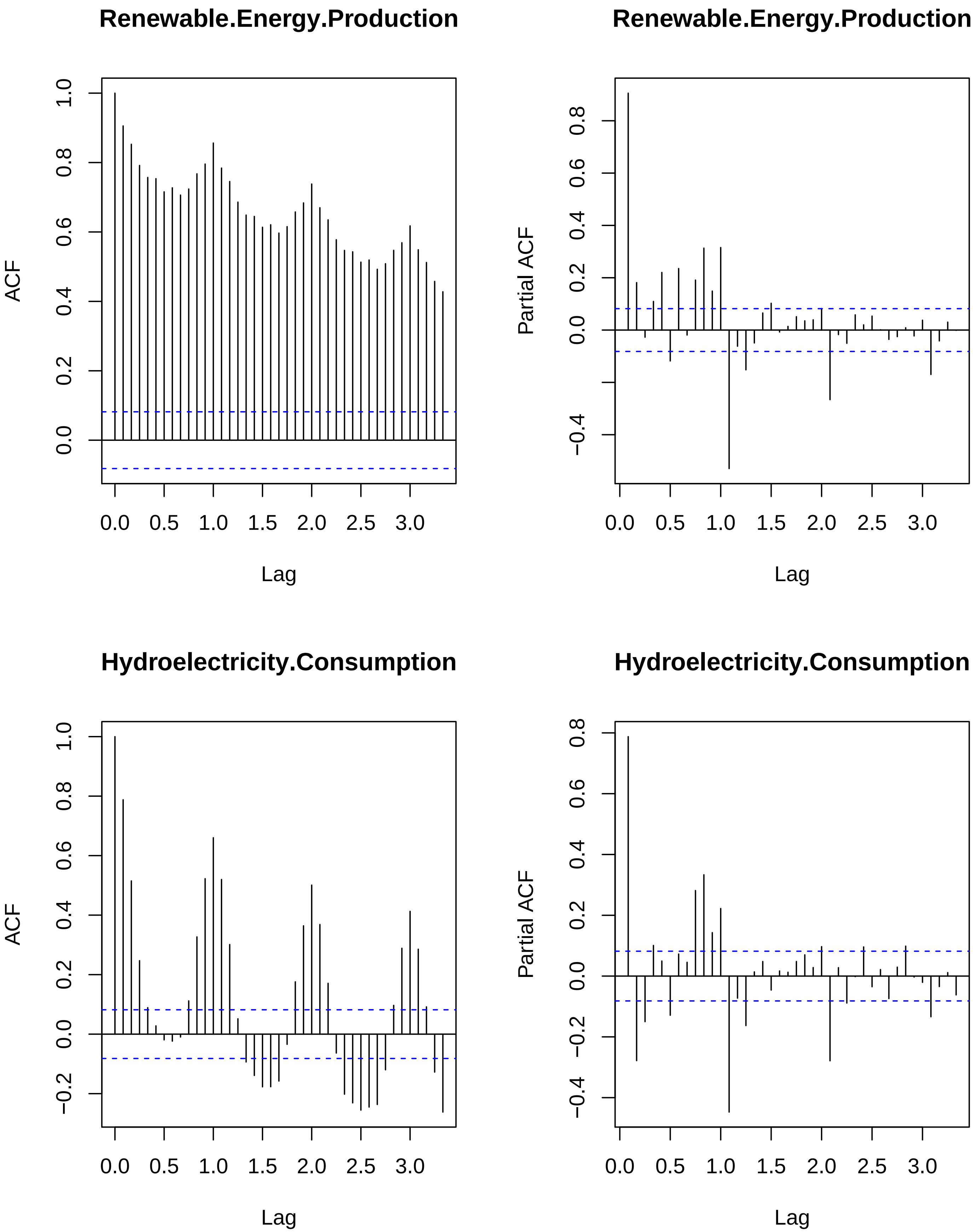

Hydroelectric Consumption
Energy (Trillion Btu)
300
200
100
0
−100
1980
1990
2000
2010
2020
Time

Q8

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

The autoregression of biomass and renewable energy production appears to fluctuate slightly more than the original autoregression plot although still not passing autoregression of zero.




Renewable.Energy.Production
ACF
Lag
Renewable.Energy.Production
Partial ACF
Lag
Hydroelectricity.Consumption
ACF
Lag
Hydroelectricity.Consumption
Partial ACF
Lag

Appendix

```
knitr::opts_chunk$set(echo=FALSE, fig.width=8, fig.height=6, tidy.opts=list(width.cutoff=80), message=FALSE, wa

#Load/install required package here
library(tseries)
library(forecast)
library(dplyr)
library(readxl)
library(lubridate)
library(ggplot2)

#Create data frame with the selected column
setwd("/Users/stefanchen/Documents/Duke/Classes/Spring 2021/ENV 790/GitHub/ENV790_30_TSA_S2021/Data")

#Read data
re_select0<-read_xlsx("Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx", skip=12, col_nam

#Select columns
re_select<-as.data.frame(re_select0[,c(1,4,5,6)]) #Transform to dataframe
colnames(re_select)=c("Date",
                      "Biomass Energy Production",
                      "Renewable Energy Production",
                      "Hydroelectric Consumption")

#Checking data
head(re_select)
str(re_select)

#Transform dataframe to time series
ts_select<-ts(re_select[,2:4], start=c(1973,1), frequency=12)
ts_select

my_date<-re_select$Date  #Establish my_date
nobs<-nrow(re_select) #Numbers of rows in ts_select
t<-c(1:nobs) #Set a vector for number of rows in ts_select

#Acf, Pacf, and time series
for(i in 1:3){
  par(mfrow=c(1,3))  #Place plot side by side
  acf(ts_select[,i], lag.max = 40, main=colnames(ts_select)[i])
  pacf(ts_select[,i],lag.max=40, main=colnames(ts_select)[i])
  plot(ts_select[,i],
     main=colnames(ts_select)[i],
     ylab="Trillion Btu",
     type="l")
}
```

```
#Biomass linear model
lmodel_bio<-lm(ts_select[,1]~t)
summary(lmodel_bio)
beta0_bio=as.numeric(lmodel_bio$coefficients[1])  #First coefficient is the intercept term or beta0
beta1_bio=as.numeric(lmodel_bio$coefficients[2])  #Second coefficient is the slope or beta1

#Renewable energy linear model
lmodel_re<-lm(ts_select[,2]~t)
summary(lmodel_re)
beta0_re=as.numeric(lmodel_re$coefficients[1])
beta1_re=as.numeric(lmodel_re$coefficients[2])
```

```r
#Hydropower linear model
lmodel_hydro<-lm(ts_select[,3]~t)
summary(lmodel_hydro)
beta0_hydro=as.numeric(lmodel_hydro$coefficients[1])
beta1_hydro=as.numeric(lmodel_hydro$coefficients[2])

#Remove the trend from series
detrend_bio<-re_select[,2]-(beta0_bio+beta1_bio*t)

#Comparison plot for biomass
ggplot(re_select, aes(x=my_date, y=re_select[,2])) +
            geom_line(color="blue") +
            labs(x="Time", y="Energy (Trillion Btu)", title="Biomass Energy Production") +
            geom_smooth(color="red",method="lm") +
            geom_line(aes(y=detrend_bio), col="green")+
            geom_smooth(aes(y=detrend_bio),color="orange",method="lm")
```

```r
#Remove the trend from series
detrend_re<-re_select[,3]-(beta0_re+beta1_re*t)

#Comparison plot for biomass
ggplot(re_select, aes(x=my_date, y=re_select[,3])) +
            geom_line(color="blue") +
            labs(x="Time", y="Energy (Trillion Btu)", title="Renewable Energy Production") +
            geom_smooth(color="red",method="lm") +
            geom_line(aes(y=detrend_re), col="green")+
            geom_smooth(aes(y=detrend_re),color="orange",method="lm")
```

```r
#Remove the trend from series
detrend_hydro<-re_select[,4]-(beta0_hydro+beta1_hydro*t)

#Comparison plot for biomass
ggplot(re_select, aes(x=my_date, y=re_select[,4])) +
            geom_line(color="blue") +
            labs(x="Time", y="Energy (Trillion Btu)", title="Hydroelectricity Consumption") +
            geom_smooth(color="red",method="lm") +
            geom_line(aes(y=detrend_hydro), col="green")+
            geom_smooth(aes(y=detrend_hydro),color="orange",method="lm")
```

```r
detrend<-data.frame("Biomass Energy Production"=detrend_bio,
                    "Renewable Energy Production"=detrend_re,
                    "Hydroelectricity Consumption"=detrend_hydro) #Combine all the detrend model

ts_detrend<-ts(detrend, start=c(1973,1), frequency=12) #time series of the detrend models

#Acf, Pacf, and time series
for(i in 1:3){
  par(mfrow=c(1,2))  #Place plot side by side
  acf(ts_detrend[,i], lag.max=40, main=colnames(ts_detrend)[i])
  pacf(ts_detrend[,i], lag.max=40, main=colnames(ts_detrend)[i])
}
```

```r
#Use seasonal means model
dummies<-seasonaldummy(ts_select[,1])  #Create the seasonal dummies

#Fit a linear model to the seasonal dummies - bio
sm_model_bio<-lm(re_select[,2]~dummies)
summary(sm_model_bio)
```

```r
#Fit a linear model to the seasonal dummies - re
sm_model_re<-lm(re_select[,3]~dummies)
summary(sm_model_re)

#Fit a linear model to the seasonal dummies - hydro
sm_model_hydro<-lm(re_select[,4]~dummies)
summary(sm_model_hydro)

#Store regression coefficients
sm_int_bio<-sm_model_bio$coefficients[1]
sm_int_re<-sm_model_re$coefficients[1]
sm_int_hydro<-sm_model_hydro$coefficients[1]
sm_coeff_bio<-sm_model_bio$coefficients[2:12]
sm_coeff_re<-sm_model_re$coefficients[2:12]
sm_coeff_hydro<-sm_model_hydro$coefficients[2:12]

#Create regression coefficient dataframes
sm_int<-data.frame("bio"=sm_int_bio, "re"=sm_int_re, "hydro"=sm_int_hydro)
sm_coeff<-data.frame("bio"=sm_int_bio, "re"=sm_int_re, "hydro"=sm_int_hydro)

#Compute seasonal component
seas_comp_bio<-array(0,nobs) #bio
for(i in 1:nobs){
  seas_comp_bio[i]=(sm_int_bio+sm_coeff_bio%*%dummies[i,])}

seas_comp_re<-array(0,nobs) #re
for(i in 1:nobs){
  seas_comp_re[i]=(sm_int_re+sm_coeff_re%*%dummies[i,])}

seas_comp_hydro<-array(0,nobs) #hydro
for(i in 1:nobs){
  seas_comp_hydro[i]=(sm_int_hydro+sm_coeff_hydro%*%dummies[i,])}

#Dataframe of seas_comp
seas_comp<-data.frame("bio"=seas_comp_bio, "re"=seas_comp_re, "hydro"=seas_comp_hydro)

#Removing seasonal component
deseason_bio<-re_select[,2]-seas_comp[1]
deseason_re<-re_select[,3]-seas_comp[2]
deseason_hydro<-re_select[,4]-seas_comp[3]

#Dataframe of deseasoned
deseason<-data.frame("bio"=deseason_bio, "re"=deseason_re, "hydro"=deseason_hydro)

#Plotting comparison graph
 for(n in 1:3){
   print(
     ggplot(re_select, aes(x=my_date, y=re_select[,n+1])) +
            geom_line(color="blue") +
            labs(x="Time", y="Energy (Trillion Btu)", title=colnames(re_select)[n+1]) +
            geom_line(aes(y=deseason[,n]), col="green")
   )
 }
```

```r
#Create time series of deseasoned model
ts_deseason<-ts(deseason, start=c(1973,1), frequency=12)

#Acf and Pacf
for(i in 1:3){
```

```r
  par(mfrow=c(1,2))  #Place plot side by side
  acf(ts_detrend[,i], lag.max=40, main=colnames(ts_detrend)[i])
  pacf(ts_detrend[,i],lag.max=40, main=colnames(ts_detrend)[i])
}
```