ENV 790.30 - Time Series Analysis for Energy Data | Spring 2022

Assignment 3 - Due date 02/08/22

Colin Lee

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_A03_Sp22.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. 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)
```

```
## Warning: package 'forecast' was built under R version 4.1.2
```

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

```
library(Kendall)
library(tseries)

library(lubridate)
```

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

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

```
library(ggplot2)
library(xlsx)
```

```
#Importing data set

mydata <- read.xlsx(file = "/Users/colinlee/Documents/Duke/Spring 2022/ENV790/ENV790_TimeSeriesAnalysis_

#cleaning data
mydata <- mydata[,4:6]

colnames(mydata)=c("Total Biomass Energy Production","Total Renewable Energy Production", "Hydroelectri

head(mydata)
```

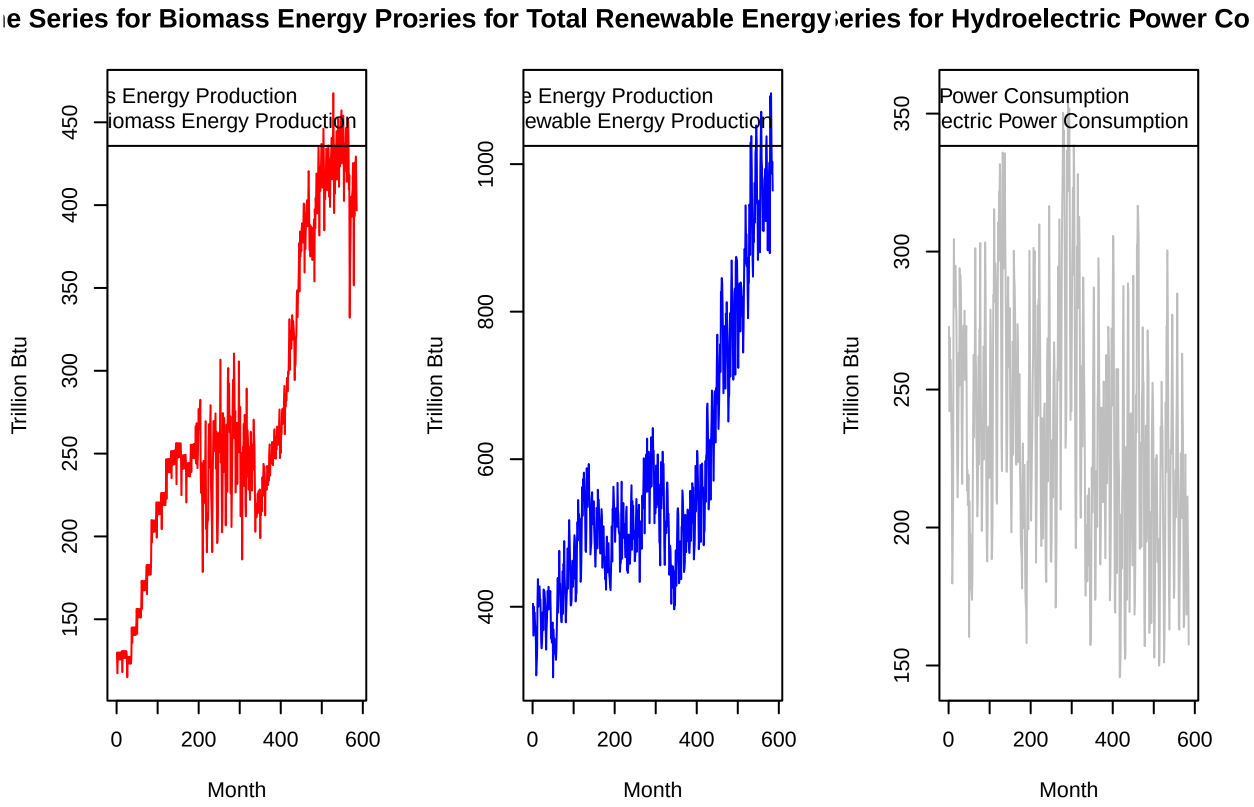
```
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                           403.981
## 2                         117.338                           360.900
## 3                         129.938                           400.161
## 4                         125.636                           380.470
## 5                         129.834                           392.141
## 6                         125.611                           377.232
##   Hydroelectric Power Consumption
## 1                         272.703
## 2                         242.199
## 3                         268.810
## 4                         253.185
## 5                         260.770
## 6                         249.859
```

##Trend Component

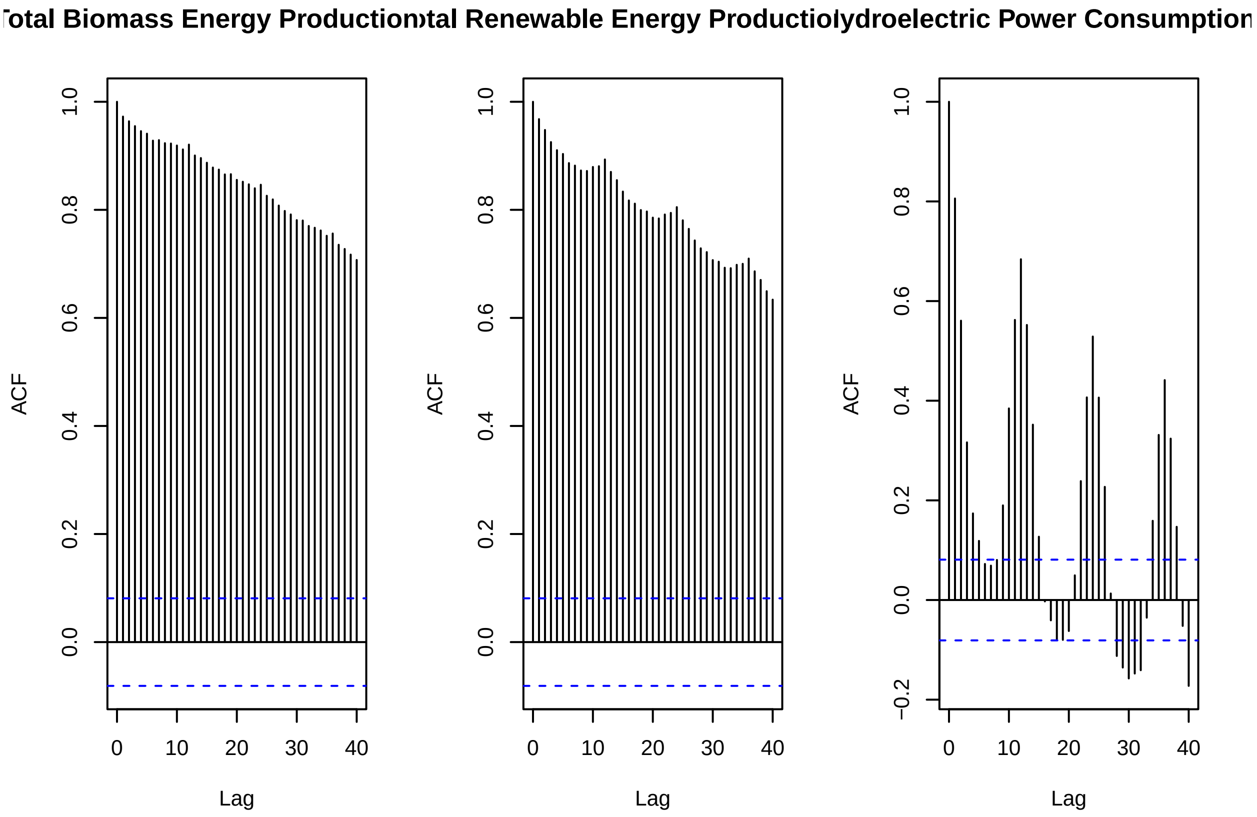
Q1

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: use par() function)

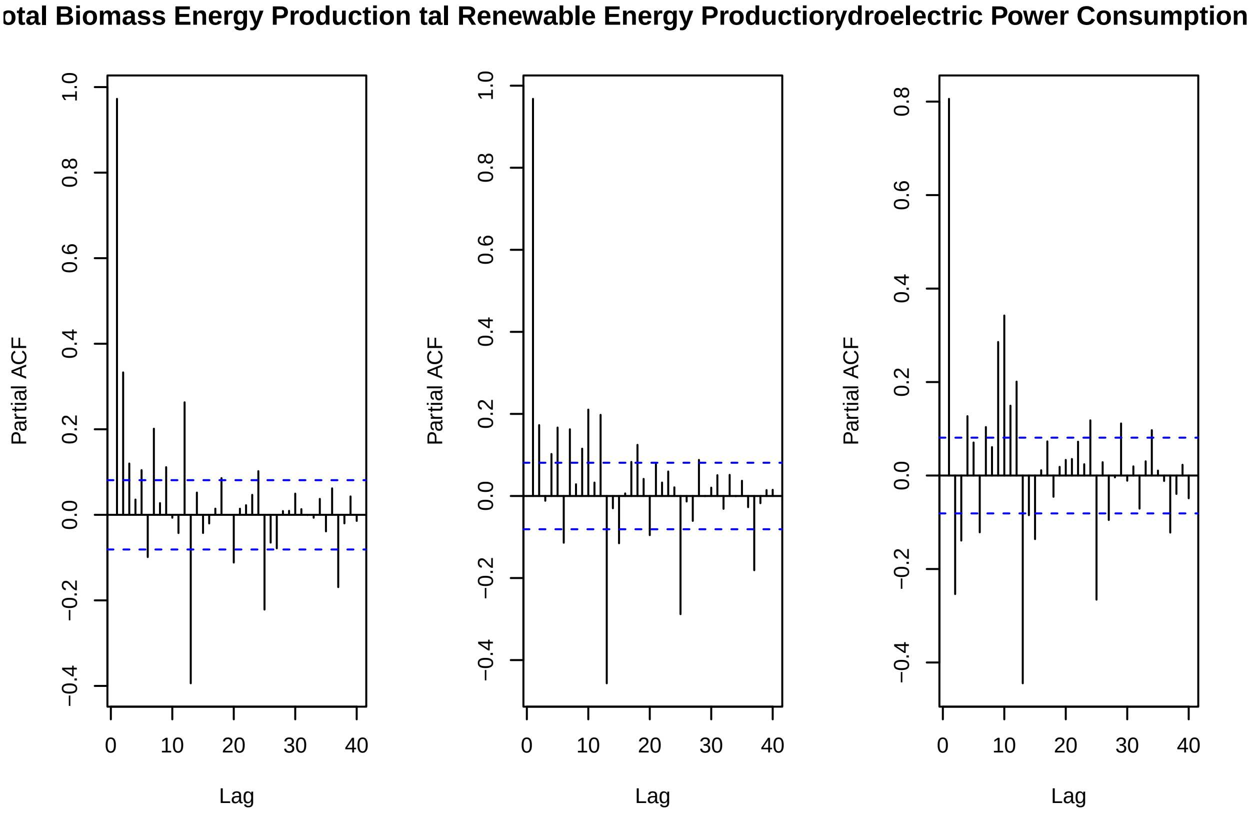
```r
par(mfrow=c(1,3))
plot(mydata[,"Total Biomass Energy Production"],type="l",col="red",ylab="Trillion Btu",xlab = "Month")
title(main="Time Series for Biomass Energy Production")
legend("topright",legend=c("Biomass Energy Production", "Mean Biomass Energy Production"), lty=c("solid

plot(mydata[,"Total Renewable Energy Production"],type="l",col="blue",ylab="Trillion Btu",xlab = "Month
title(main="Time Series for Total Renewable Energy Production")
legend("topright",legend=c("Renewable Energy Production", "Mean Renewable Energy Production"), lty=c("s

plot(mydata[,"Hydroelectric Power Consumption"],type="l",col="grey",ylab="Trillion Btu",xlab = "Month")
title(main="Time Series for Hydroelectric Power Consumption")
legend("topright",legend=c("Hydroelectric Power Consumption", "Mean Hydroelectric Power Consumption"), 
```



```r
par(mfrow=c(1,3))  #place plot side by side
acf(mydata[,1],lag.max=40,main=paste("Total Biomass Energy Production ACF"))
acf(mydata[,2],lag.max=40,main=paste("Total Renewable Energy Production ACF"))
acf(mydata[,3],lag.max=40,main=paste("Hydroelectric Power Consumption ACF"))
```



```
par(mfrow=c(1,3))  #place plot side by side
pacf(mydata[,1],lag.max=40,main=paste("Total Biomass Energy Production PACF"))
pacf(mydata[,2],lag.max=40,main=paste("Total Renewable Energy Production PACF"))
pacf(mydata[,3],lag.max=40,main=paste("Hydroelectric Power Consumption PACF"))
```



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?

From the time series plots, it appears that Total Biomass Energy Production and Total Renewable Energy Production are generally increasing. However, it appears that Hydroelectric Power Consumption is generally decreasing.

Based on the ACF, it appears that Hydroelectric Power Consumption has a seasonal trend. Based on the ACFs for Total Biomass Energy Production and Total Renewable Energy Production, there is no seasonal trend, and the high autocorrelation indicates they are non-stationary.

Additionally, we will need more tests to assess whether the time series are stochastic or deterministic.

We already know there is seasonality in Hydroelectric Power Consumption due to the clearer ACF, but the PACFs may indicate that there may be a little seasonality in Total Biomass Energy Production and Total Renewable Energy Production.

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.

For biomass energy, the slope is positive at 0.47, indicating that biomass energy production follows a trend of increasing by 0.47 trillion Btu each month. Additionally, the intercept is 134.79, indicating that at the start of data collection, the biomass energy production (in accordance with this hypothetical linear model) can be seen as 134.79 trillion Btu. Note this is not the actual starting biomass energy production in JAN 1973, but an estimate given by the linear model.

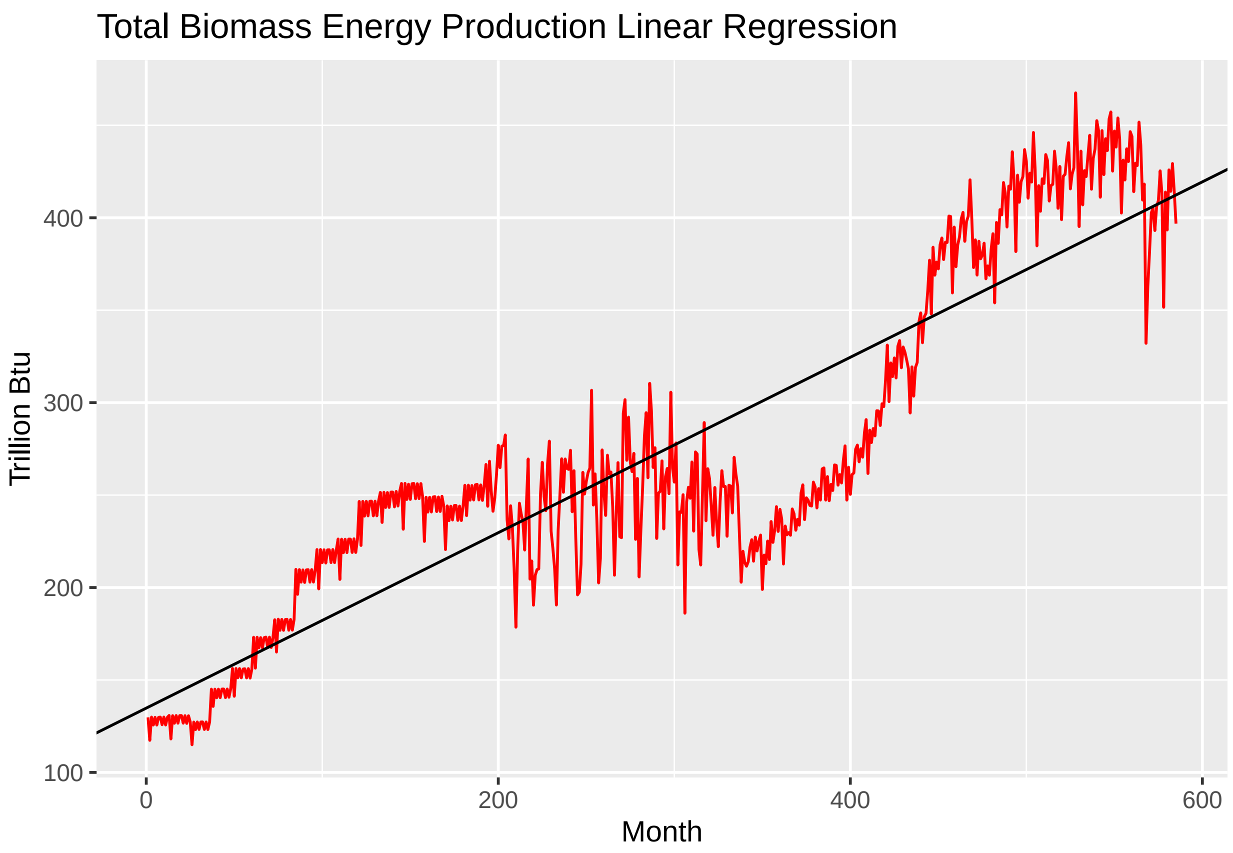
For renewable energy, the slope is positive at 0.88, indicating that renewable energy production follows a trend of increasing by 0.88 trillion Btu each month. Additionally, the intercept is 323.18, indicating that at the start of data collection, the renewable energy production (in accordance with this hypothetical linear model) can be seen as 323.18 trillion Btu. Note this is not the actual starting renewable energy production in JAN 1973, but an estimate given by the linear model.

For hydroelectric power, the slope is negative at -0.079, indicating that hydroelectric power consumption follows a trend of decreasing by 0.079 trillion Btu each month. Additionally, the intercept is 323.18, indicating that at the start of data collection, the hydroelectric power consumption (in accordance with this hypothetical linear model) can be seen as 259.183 trillion Btu. Note this is not the actual starting hydroelectric power consumption in JAN 1973, but an estimate given by the linear model.

```r
t <- c(1:nrow(mydata))
linear_trend_model_1=lm(mydata[,1]~t)
summary(linear_trend_model_1)
```

```
##
## Call:
## lm(formula = mydata[, 1] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.892  -24.306    4.932   33.103   82.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.348e+02  3.282e+00   41.07   <2e-16 ***
## t           4.744e-01  9.705e-03   48.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.64 on 583 degrees of freedom
## Multiple R-squared:  0.8039, Adjusted R-squared:  0.8035
## F-statistic:  2389 on 1 and 583 DF,  p-value: < 2.2e-16
```

```r
beta01=as.numeric(linear_trend_model_1$coefficients[1])  #first coefficient is the intercept term or be
beta11=as.numeric(linear_trend_model_1$coefficients[2])  #second coefficient is the slope or beta1

ggplot(mydata, aes(x = t, y=mydata[,1])) +
            geom_line(color="red") +
            ylab("Trillion Btu") +
            xlab("Month") +
            ggtitle("Total Biomass Energy Production Linear Regression") +
            geom_abline(intercept = beta01, slope = beta11, color="black")
```

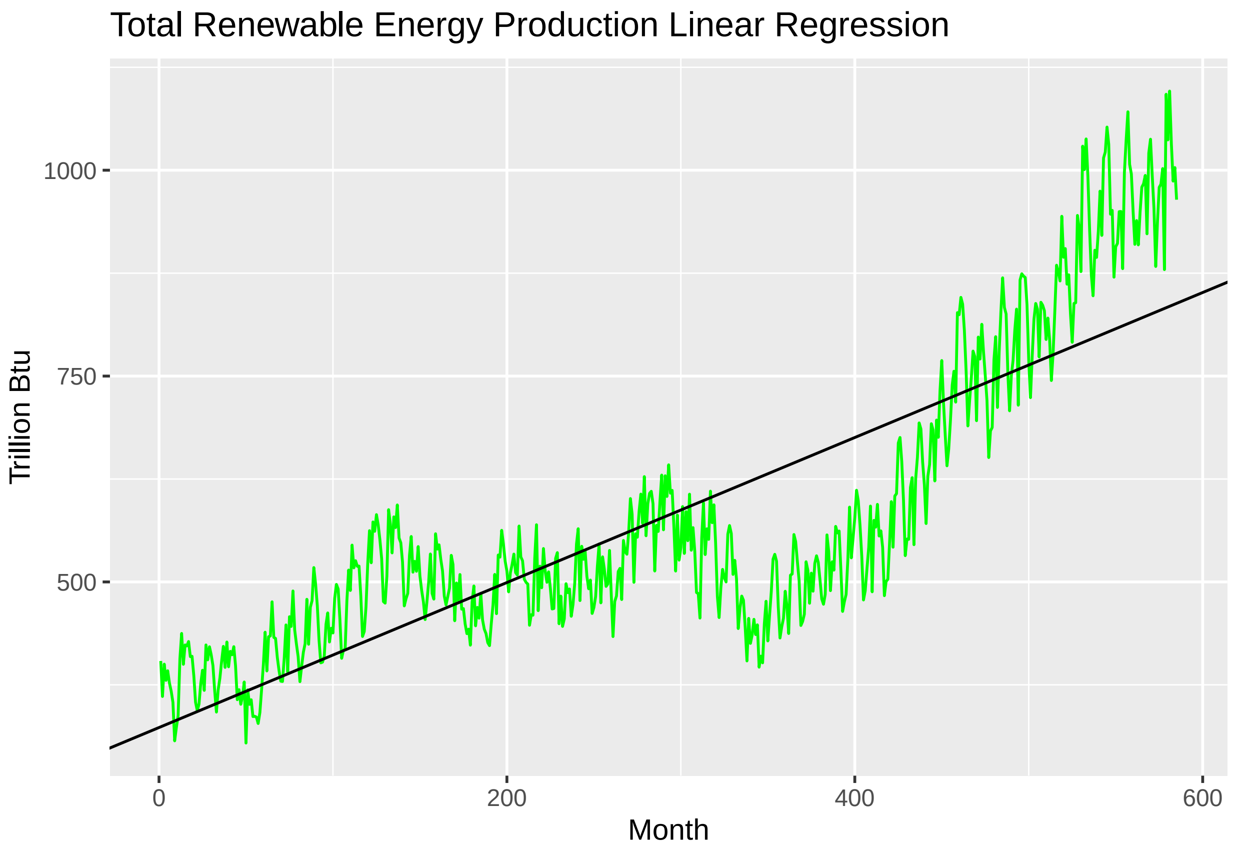


```
linear_trend_model_2=lm(mydata[,2]~t)
summary(linear_trend_model_2)
```

```
##
## Call:
## lm(formula = mydata[, 2] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -230.488  -57.869    5.595   62.090  261.349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 323.18243    8.02555   40.27   <2e-16 ***
## t             0.88051    0.02373   37.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.93 on 583 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.702
## F-statistic:  1377 on 1 and 583 DF,  p-value: < 2.2e-16
```

```
beta02=as.numeric(linear_trend_model_2$coefficients[1])  #first coefficient is the intercept term or be
beta12=as.numeric(linear_trend_model_2$coefficients[2])  #second coefficient is the slope or beta1
```

```
ggplot(mydata, aes(x = t, y=mydata[,2])) +
            geom_line(color="green") +
            ylab("Trillion Btu") +
            xlab("Month") +
            ggtitle("Total Renewable Energy Production Linear Regression") +
            geom_abline(intercept = beta02, slope = beta12, color="black")
```

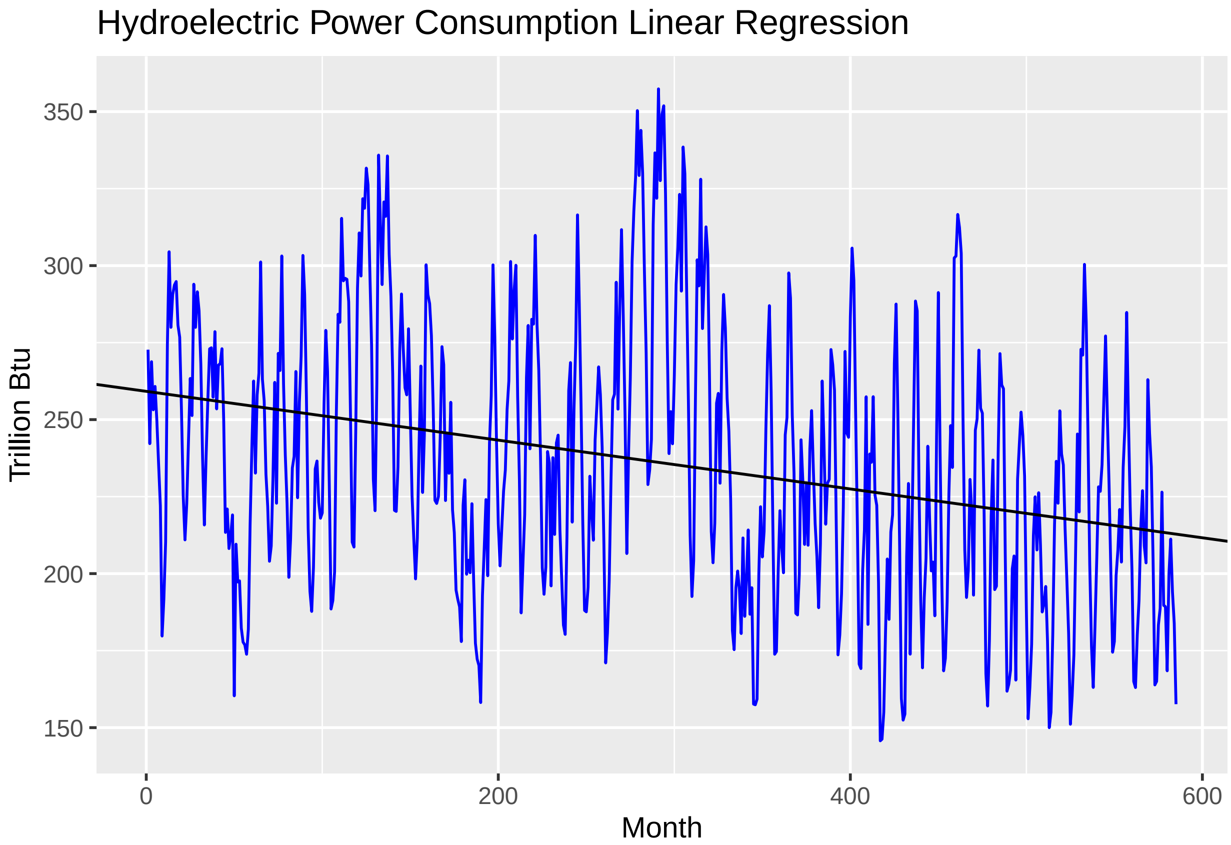


```
linear_trend_model_3=lm(mydata[,3]~t)
summary(linear_trend_model_3)
```

```
##
## Call:
## lm(formula = mydata[, 3] ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.892 -31.300  -2.414  27.876 121.263
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 259.18303    3.47464  74.593  < 2e-16 ***
## t            -0.07924    0.01027  -7.712 5.36e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 41.97 on 583 degrees of freedom
## Multiple R-squared:  0.09258,    Adjusted R-squared:  0.09103
## F-statistic: 59.48 on 1 and 583 DF,  p-value: 5.364e-14
```

```
beta03=as.numeric(linear_trend_model_3$coefficients[1])  #first coefficient is the intercept term or be
beta13=as.numeric(linear_trend_model_3$coefficients[2])  #second coefficient is the slope or beta1

ggplot(mydata, aes(x = t, y=mydata[,3])) +
            geom_line(color="blue") +
            ylab("Trillion Btu") +
            xlab("Month") +
            ggtitle("Hydroelectric Power Consumption Linear Regression") +
            geom_abline(intercept = beta03, slope = beta13, color="black")
```



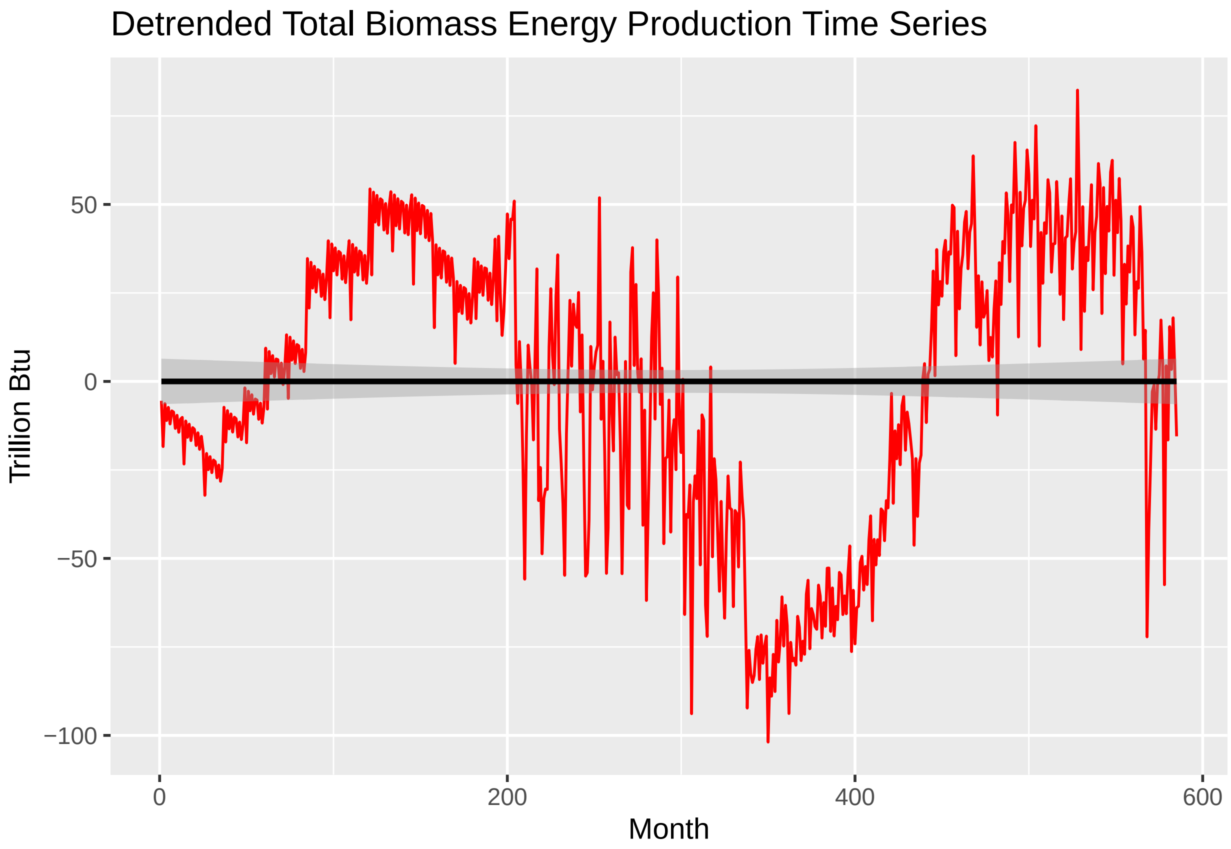
Q4

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?

Yes, we can see that the plots no longer have general trends of increasing or decreasing over time, but rather, remain flatter over long periods of time (save for the random/seasonal fluctuation). In comparison to the plots from Q1, these plots no longer resemble the time series plots as well as these are more flatter in trajectory and no longer have increasing trends (biomass/renewable) or decreasing trend (hydro).

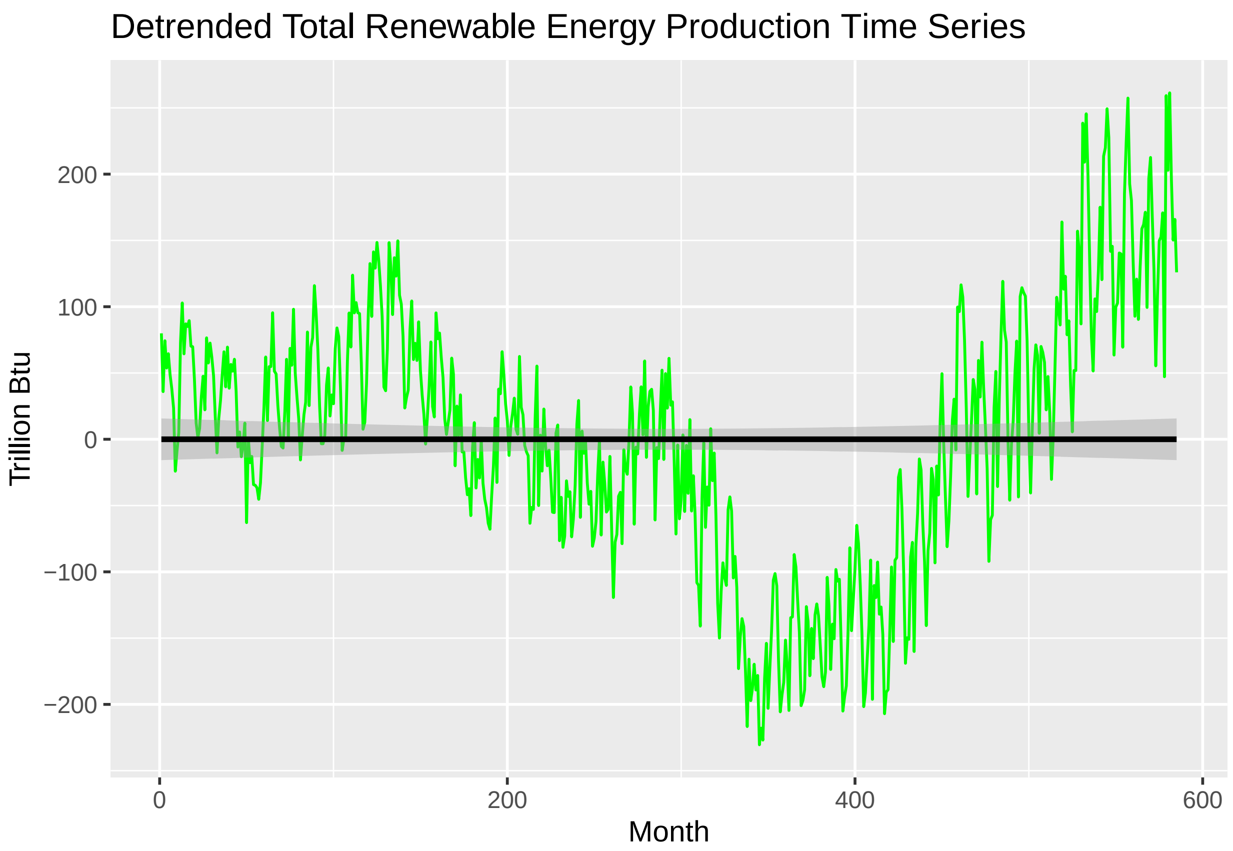
```
detrend_mydata_biomass <- mydata[,(1)]-(beta01+beta11*t)
detrend_mydata_renewable <- mydata[,(2)]-(beta02+beta12*t)
detrend_mydata_hydro <- mydata[,(3)]-(beta03+beta13*t)

ggplot(mydata, aes(x=t, y=detrend_mydata_biomass)) +
            geom_line(color="red") +
            ylab("Trillion Btu") +
            xlab("Month") +
            ggtitle("Detrended Total Biomass Energy Production Time Series") +
            geom_smooth(aes(y=detrend_mydata_biomass),color="black",method="lm")
```

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



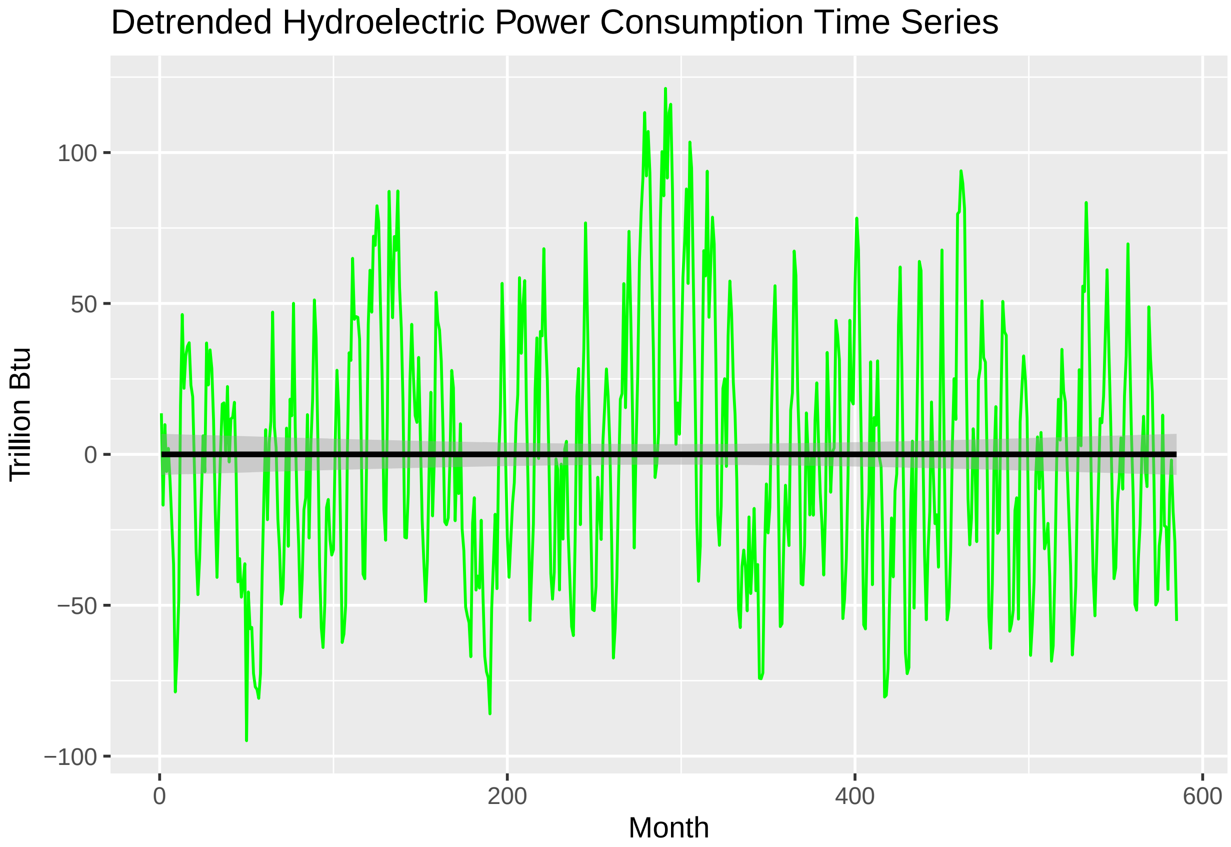
```
ggplot(mydata, aes(x=t, y=detrend_mydata_renewable)) +
            geom_line(color="green") +
            ylab("Trillion Btu") +
            xlab("Month") +
            ggtitle("Detrended Total Renewable Energy Production Time Series") +
            geom_smooth(aes(y=detrend_mydata_renewable),color="black",method="lm")
```

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



```
ggplot(mydata, aes(x=t, y=detrend_mydata_hydro)) +
            geom_line(color="green") +
            ylab("Trillion Btu") +
            xlab("Month") +
            ggtitle("Detrended Hydroelectric Power Consumption Time Series") +
            geom_smooth(aes(y=detrend_mydata_hydro),color="black",method="lm")
```

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



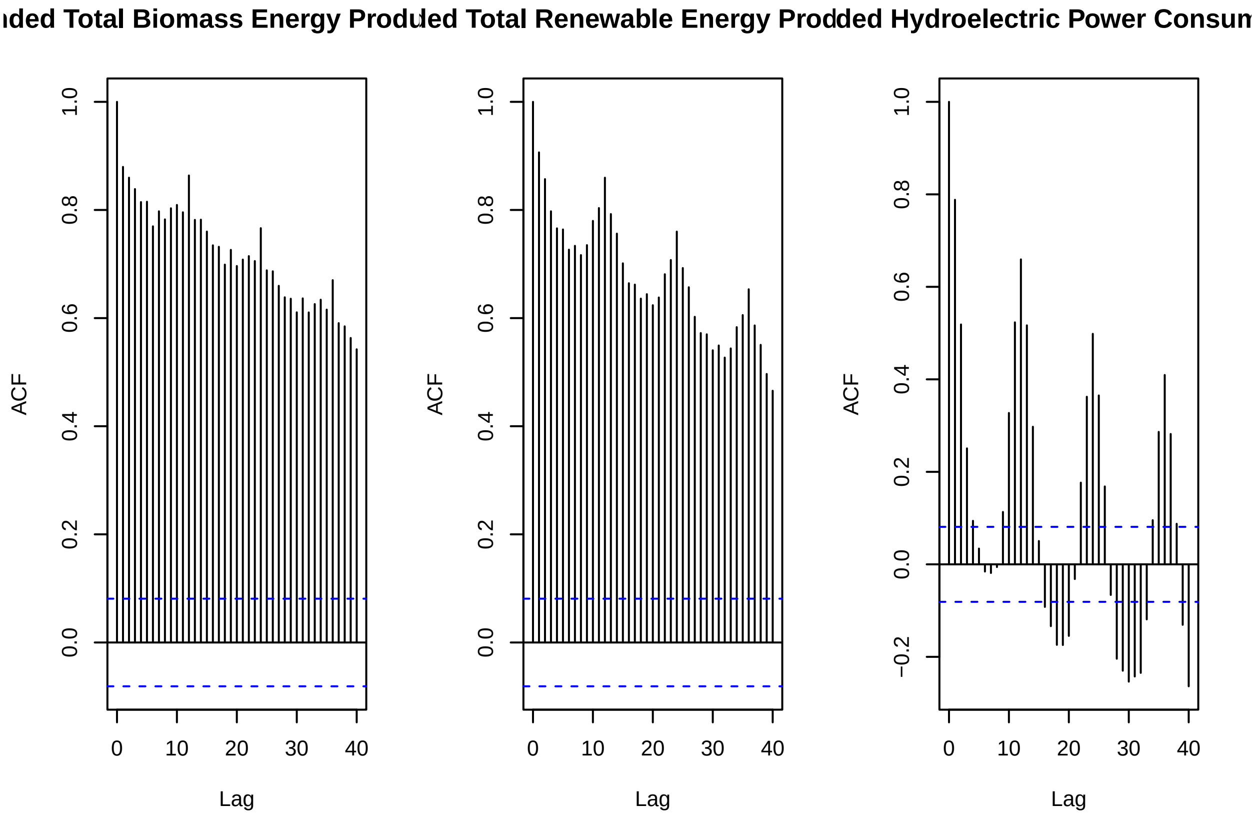
Q5

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

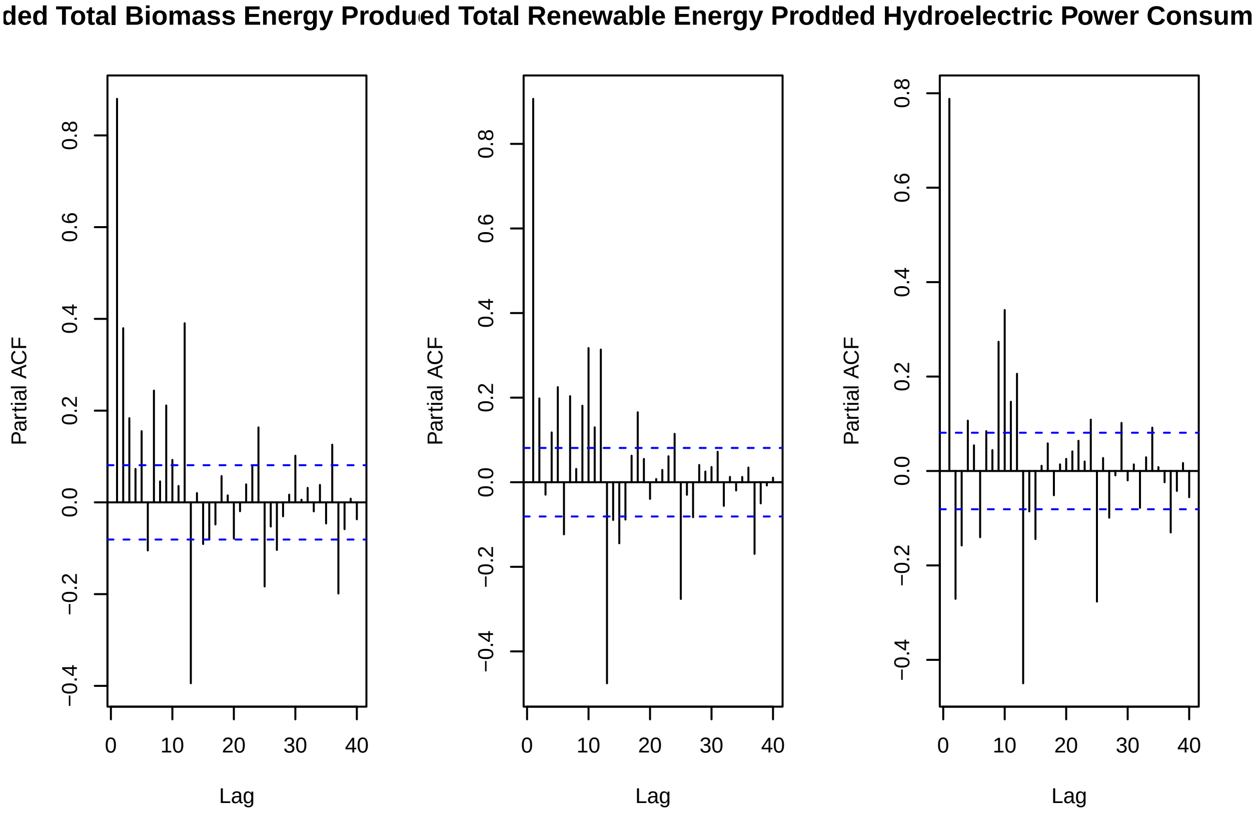
The ACFs look very similar, with high seasonality for Hydroelectric Power Consumption and high autocorrelation among Total Biomass Energy Production and Total Renewable Energy Production. However, in the ACFs for Q5, it is evident that there might be some seasonality as there are some visible peaks and dips, especially in Total Renewable Energy Production.

With regard to the PACFs, it is hard to tell if there is a significant difference and they look almost identical to Q1's.

```
par(mfrow=c(1,3))  #place plot side by side
acf(detrend_mydata_biomass,lag.max=40,main=paste("Detrended Total Biomass Energy Production ACF"))
acf(detrend_mydata_renewable,lag.max=40,main=paste("Detrended Total Renewable Energy Production ACF"))
acf(detrend_mydata_hydro,lag.max=40,main=paste("Detrended Hydroelectric Power Consumption ACF"))
```



```
par(mfrow=c(1,3))  #place plot side by side
pacf(detrend_mydata_biomass,lag.max=40,main=paste("Detrended Total Biomass Energy Production PACF"))
pacf(detrend_mydata_renewable,lag.max=40,main=paste("Detrended Total Renewable Energy Production PACF"))
pacf(detrend_mydata_hydro,lag.max=40,main=paste("Detrended Hydroelectric Power Consumption PACF"))
```



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 (i.e. using the seasonal dummies) 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.

I really only think hydro has seasonality, but I will do this for all the series just to see in case I'm wrong!

Based on the data below: Based on the summary data, it is clear that only the Hydroelectric Power Consumption data is seasonal because of a low p-value of 2.2E-16. The Total Biomass Energy Production is most definitely not seasonal due to a very high p-value of 0.8647, and this was somewhat apparent with the detrended ACF. For the Total Renewable Energy Production, the p-value is relatively low at 0.07, indicating that there may be some seasonality, but this is not statistically significant for conventional p-value of 0.05.

```
ts_mydata <- ts(mydata, start = 1, frequency = 12)

# hydro
dummies_hydro <- seasonaldummy(ts_mydata[,3])
seas_means_model_hydro=lm(ts_mydata[,3]~dummies_hydro)
summary(seas_means_model_hydro)
```

```
##
## Call:
## lm(formula = ts_mydata[, 3] ~ dummies_hydro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.253 -23.017  -3.042  21.487  99.478
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       237.841      4.892  48.616  < 2e-16 ***
## dummies_hydroJan   13.558      6.883   1.970  0.04936 *
## dummies_hydroFeb   -8.090      6.883  -1.175  0.24037
## dummies_hydroMar   20.067      6.883   2.915  0.00369 **
## dummies_hydroApr   16.619      6.883   2.414  0.01607 *
## dummies_hydroMay   39.961      6.883   5.805 1.06e-08 ***
## dummies_hydroJun   31.315      6.883   4.549 6.57e-06 ***
## dummies_hydroJul   10.511      6.883   1.527  0.12732
## dummies_hydroAug  -17.853      6.883  -2.594  0.00974 **
## dummies_hydroSep  -49.852      6.883  -7.242 1.43e-12 ***
## dummies_hydroOct  -48.086      6.919  -6.950 9.96e-12 ***
## dummies_hydroNov  -32.187      6.919  -4.652 4.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.89 on 573 degrees of freedom
## Multiple R-squared:  0.4182, Adjusted R-squared:  0.4071
## F-statistic: 37.45 on 11 and 573 DF,  p-value: < 2.2e-16
```

```
beta_int_hydro=seas_means_model_hydro$coefficients[1]
beta_coeff_hydro=seas_means_model_hydro$coefficients[2:12]

# biomass
dummies_biomass <- seasonaldummy(ts_mydata[,1])
seas_means_model_biomass=lm(ts_mydata[,1]~dummies_biomass)
summary(seas_means_model_biomass)
```

```
##
## Call:
## lm(formula = ts_mydata[, 1] ~ dummies_biomass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.96  -51.40  -22.15   60.65  183.31
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         284.241     12.962  21.928   <2e-16 ***
## dummies_biomassJan   -1.498     18.238  -0.082   0.9346
## dummies_biomassFeb  -30.582     18.238  -1.677   0.0941 .
## dummies_biomassMar   -8.873     18.238  -0.486   0.6268
## dummies_biomassApr  -21.009     18.238  -1.152   0.2498
## dummies_biomassMay  -14.065     18.238  -0.771   0.4409
## dummies_biomassJun  -19.601     18.238  -1.075   0.2829
## dummies_biomassJul   -3.499     18.238  -0.192   0.8479
## dummies_biomassAug   -0.252     18.238  -0.014   0.9890
## dummies_biomassSep  -12.518     18.238  -0.686   0.4928
## dummies_biomassOct   -3.629     18.331  -0.198   0.8432
## dummies_biomassNov   -9.592     18.331  -0.523   0.6010
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 89.81 on 573 degrees of freedom
## Multiple R-squared:  0.01056,    Adjusted R-squared:  -0.008439
## F-statistic: 0.5557 on 11 and 573 DF,  p-value: 0.8647
```

```
beta_int_biomass=seas_means_model_biomass$coefficients[1]
beta_coeff_biomass=seas_means_model_biomass$coefficients[2:12]

# renewable
dummies_renew <- seasonaldummy(ts_mydata[,2])
seas_means_model_renew=lm(ts_mydata[,2]~dummies_renew)
summary(seas_means_model_renew)
```

```
##
## Call:
## lm(formula = ts_mydata[, 2] ~ dummies_renew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -272.95 -111.55  -59.35   65.68  480.41
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       589.971     25.464  23.169   <2e-16 ***
## dummies_renewJan   11.793     35.828   0.329   0.7422
## dummies_renewFeb  -40.992     35.828  -1.144   0.2530
## dummies_renewMar   21.892     35.828   0.611   0.5414
## dummies_renewApr    8.908     35.828   0.249   0.8037
## dummies_renewMay   37.500     35.828   1.047   0.2957
## dummies_renewJun   19.465     35.828   0.543   0.5871
## dummies_renewJul    8.115     35.828   0.227   0.8209
## dummies_renewAug  -18.359     35.828  -0.512   0.6086
## dummies_renewSep  -62.115     35.828  -1.734   0.0835 .
## dummies_renewOct  -51.377     36.012  -1.427   0.1542
## dummies_renewNov  -41.789     36.012  -1.160   0.2464
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 176.4 on 573 degrees of freedom
## Multiple R-squared:  0.03139,    Adjusted R-squared:  0.0128
## F-statistic: 1.688 on 11 and 573 DF,  p-value: 0.07235
```

```
beta_int_renew=seas_means_model_renew$coefficients[1]
beta_coeff_renew=seas_means_model_renew$coefficients[2:12]
```

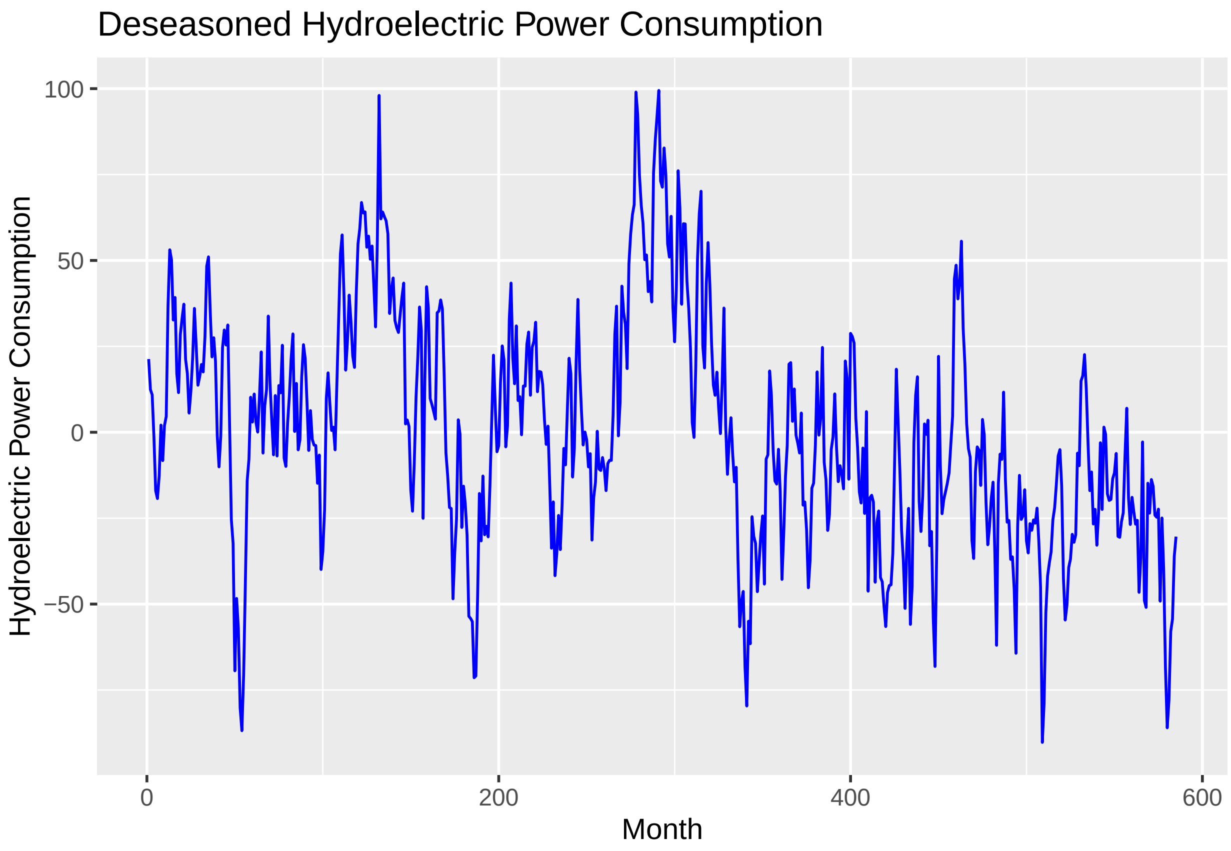
Q7

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?

I will not proceed with Total Biomass Energy Production or Total Renewable Energy Production due to insignificant levels of seasonality.

Compared to the time series plot in Q1 for Hydroelectric Power Consumption, the scale is much lower with ranges oscillating over zero rather than between 200-300. Furthermore, there are less serious oscillations as there are less big peaks and dips that are associated with the annual seasonality, making the this deseasoned time series plot less wavy compared to the Q1 time series plot.

```r
hydro_seas_comp=array(0,nrow(mydata))
for(i in 1:nrow(mydata)){
  hydro_seas_comp[i]=(beta_int_hydro+beta_coeff_hydro%*%dummies_hydro[i,])
}

deseason_hydro <- mydata[,3]-hydro_seas_comp

ggplot(mydata, aes(x=t, y=deseason_hydro)) +
            geom_line(color="blue") +
            ylab("Hydroelectric Power Consumption") +
            xlab("Month") +
            ggtitle("Deseasoned Hydroelectric Power Consumption")
```



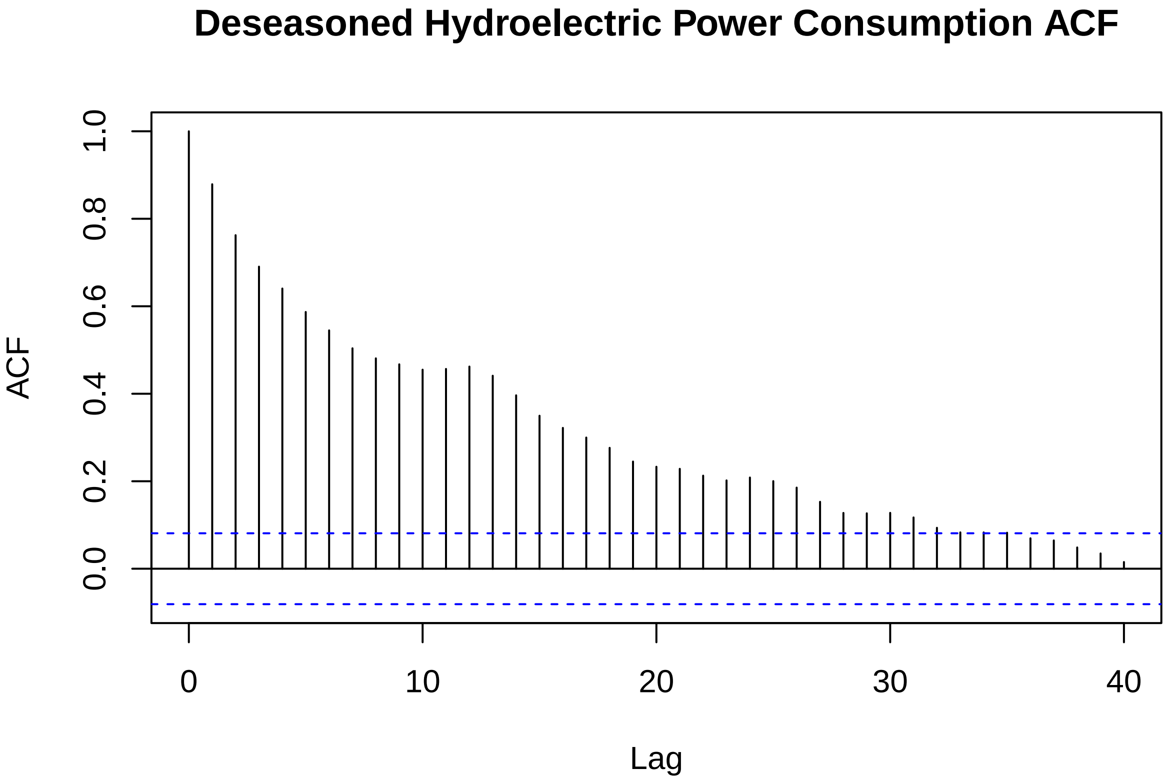
Q8

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

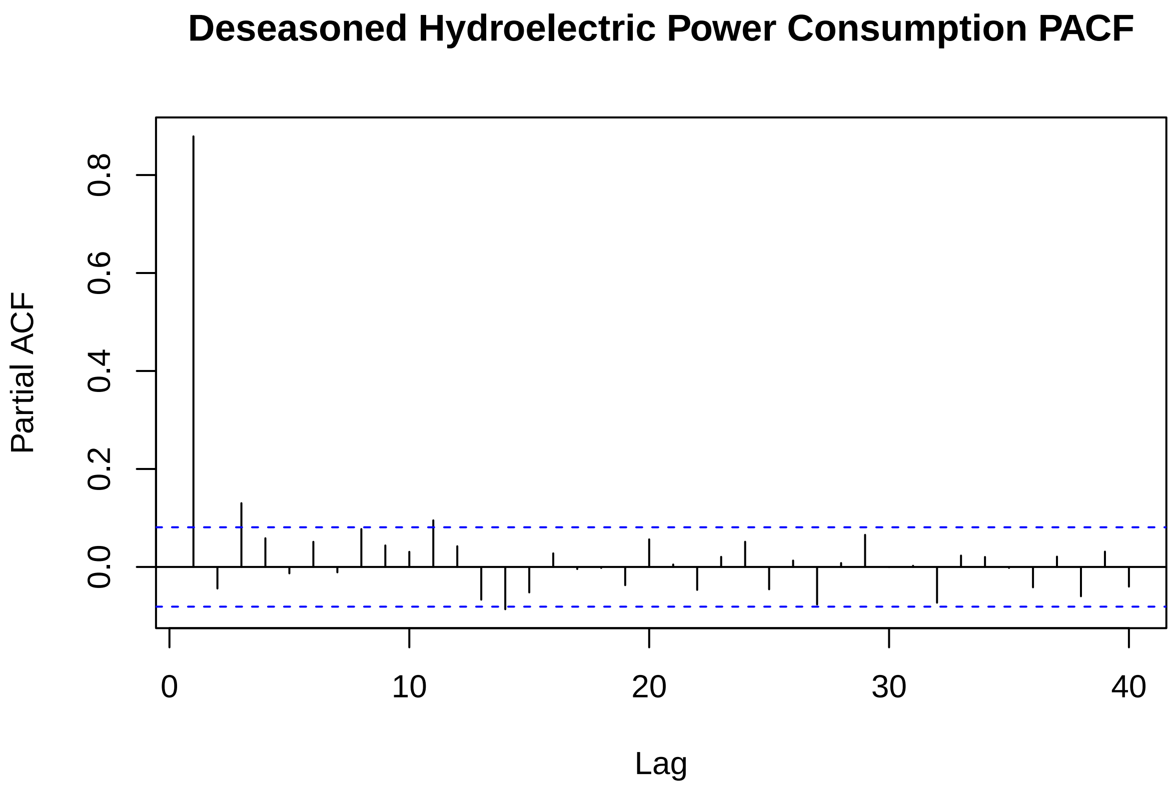
The ACF changed significantly, where the ACF for hydro in Q1 is very seasonal and very hard to assess for autocorrelation. Now, deseasoned, the ACF no longer has this wave pattern, but more resembles a decreasing linear slope.

The PACF is slightly different in that the y-axis range for PACF values is much lower for the deseasoned hydro data compared to the Q1 hydro data (no adjustments made). With regards to changes in how the PACFs look, it is hard to tell any significant differences besides the change in range.

```
acf(deseason_hydro,lag.max=40,main=paste("Deseasoned Hydroelectric Power Consumption ACF"))
```



```
pacf(deseason_hydro,lag.max=40,main=paste("Deseasoned Hydroelectric Power Consumption PACF"))
```


Deseasoned Hydroelectric Power Consumption PACF
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag