# ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 3 - Due date 02/10/23

Maggie O'Shea

## 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_Sp23.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)
library(tseries)
library(Kendall)
library(tidyverse)

#Import Data
library(readxl)
energy_data<- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
read_col_names <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
colnames(energy_data) <- read_col_names
```

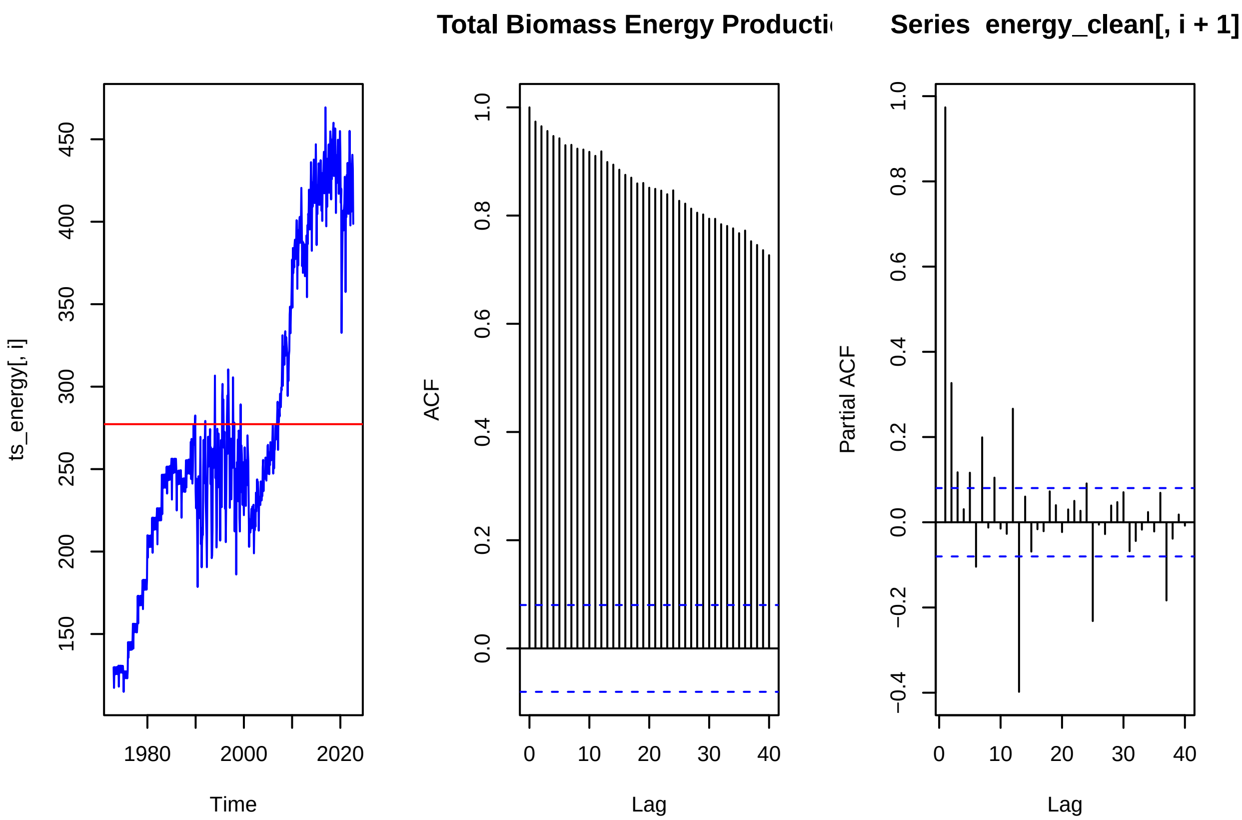
```
energy_clean <- energy_data%>%
  select("Month", "Total Biomass Energy Production", "Total Renewable Energy Production", "Hydroelectric

ts_energy <- ts(energy_clean[,2:4], frequency = 12, start = c(1973, 1))
```

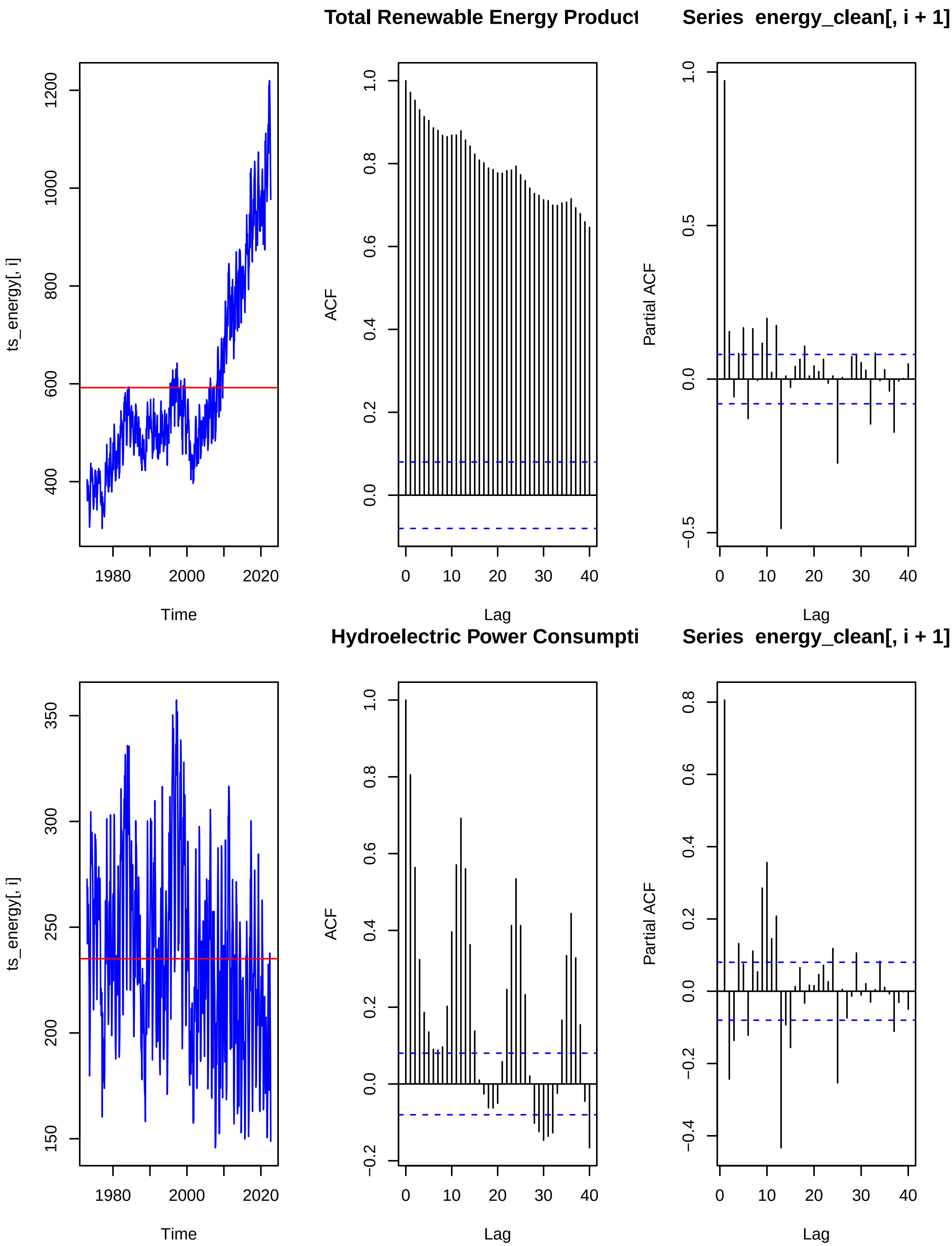
## Trend Component

### 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)

```
for(i in 1:3){
  par(mfrow = c(1,3))
  plot(ts_energy[,i],type="l",col="blue")
  abline(h=mean(ts_energy[,i]),col="red")
  acf(energy_clean[,i+1], lag.max = 40)
  pacf(energy_clean[,i+1], lag.max = 40)
}
```




Total Renewable Energy Product
Series energy_clean[, i + 1]
ts_energy[, i]
Time
ACF
Lag
Partial ACF
Lag
Hydroelectric Power Consumpti
Series energy_clean[, i + 1]
ts_energy[, i]
Time
ACF
Lag
Partial ACF
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?

**Q2 Answer:** The plot for biomass energy production does appear to have a trend. The ACF for both Biomass Energy production and Total Renewable Energy production both show a deterministic trend in the ACF plot which show a high correlation in the early lags that declines consistently. The hydroelectric power plot shows a clear seasonal component that possibly declines as well, though further analysis is needed.

### 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
#Biomass

nobs <- nrow(energy_clean)
t <- 1:nobs

linearTrend_biomass <- lm(energy_clean$`Total Biomass Energy Production` ~ t)
summary(linearTrend_biomass)
```

**Biomass**

```
##
## Call:
## lm(formula = energy_clean$‘Total Biomass Energy Production‘ ~
##     t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.800  -23.994    5.667   32.265   82.192
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.337e+02  3.245e+00   41.22   <2e-16 ***
## t           4.800e-01  9.402e-03   51.05   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 39.59 on 595 degrees of freedom
## Multiple R-squared:  0.8142, Adjusted R-squared:  0.8138
## F-statistic:  2607 on 1 and 595 DF,  p-value: < 2.2e-16
```

```r
beta0_biomass <- linearTrend_biomass$coefficients[1]
beta1_biomass <- linearTrend_biomass$coefficients[2]
```

**Interpret Biomass Linear Trend:** The trend helps to identify the long-term tendency of the data, in this case of Total Biomass Production. The positive coefficients (and statistical significance), both for the intercept and time, indicate that the for ever increase in time (t), there is an associated increase in biomass production. The intercept is the value of Total Biomass production at time 0, and thus the rest of the values of biomass production (at time 1 and on) are in relation to this value of 133.7. Because this and beta1 are positive, the trend is a positive trend with biomass production increasing over time.

```
#Renewable
linearTrend_renew<- lm(energy_clean$`Total Renewable Energy Production` ~ t)
summary(linearTrend_renew)
```

**Renewable**

```
##
## Call:
## lm(formula = energy_clean$`Total Renewable Energy Production` ~
##     t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.75  -61.85    8.59   64.48  352.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 312.2475     8.4902   36.78   <2e-16 ***
## t             0.9362     0.0246   38.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.6 on 595 degrees of freedom
## Multiple R-squared:  0.7088, Adjusted R-squared:  0.7083
## F-statistic:  1448 on 1 and 595 DF,  p-value: < 2.2e-16
```

```
beta0_renew <- linearTrend_renew$coefficients[1]
beta1_renew <- linearTrend_renew$coefficients[2]
```

**Interpret Renewable Energy Linear Trend:** As in the Biomass linear model, time and the intercept are statistically significant and have positive coefficients. The intercept indicates that at time 0 the renewable energy production would be 312.25 and increases from there - an increasing trend over time.

```
linearTrend_hydro <- lm(energy_clean$`Hydroelectric Power Consumption` ~ t)
summary(linearTrend_hydro)
```

**Hydroelectric**

```
##
```

```
## Call:
## lm(formula = energy_clean$‘Hydroelectric Power Consumption‘ ~
##     t)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -95.42 -31.20  -2.56  27.32 121.61
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 259.898013   3.427300  75.832  < 2e-16 ***
## t            -0.082888   0.009931  -8.346 4.94e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 41.82 on 595 degrees of freedom
## Multiple R-squared:  0.1048, Adjusted R-squared:  0.1033
## F-statistic: 69.66 on 1 and 595 DF,  p-value: 4.937e-16
```

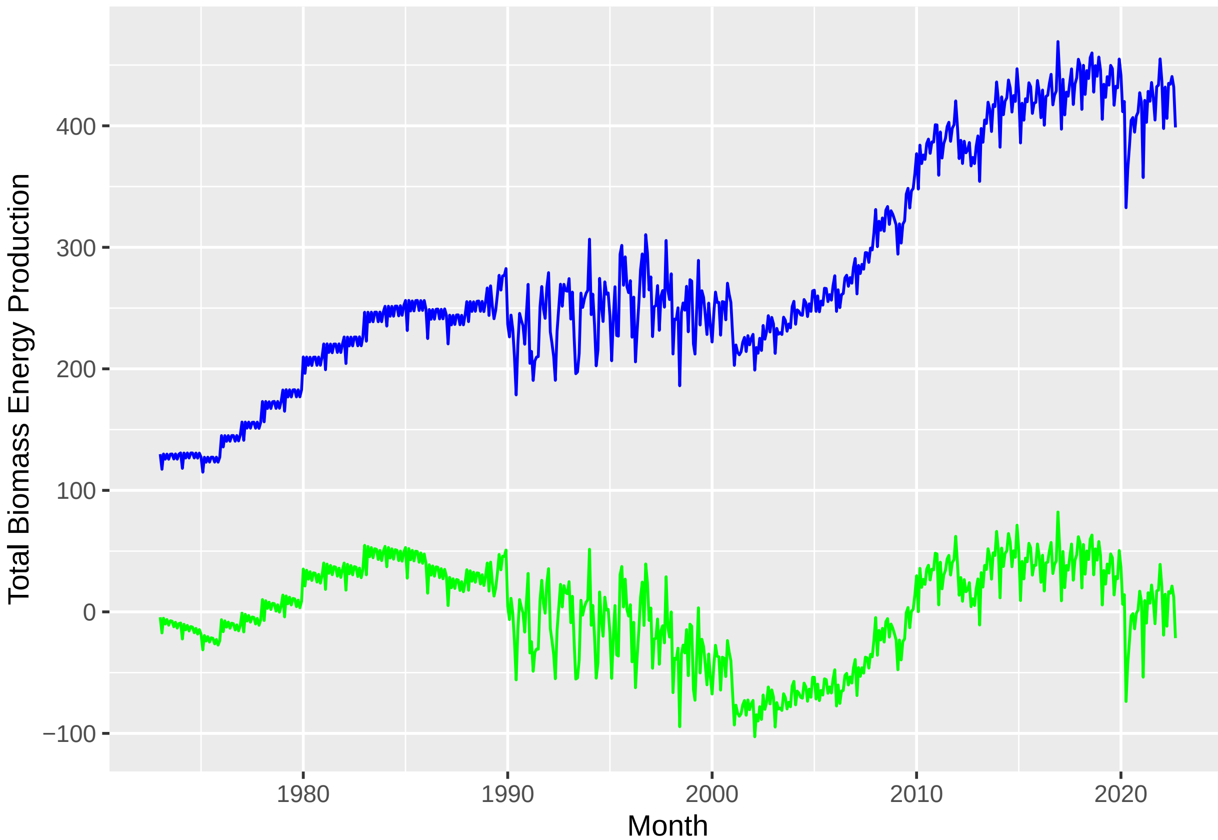
```
beta0_hydro <- linearTrend_hydro$coefficients[1]
beta1_hydro <- linearTrend_hydro$coefficients[2]
```

**Interpret Hydroelectric Power Linear Trend:** Again the intercept and t are statistically significant for hydroelectic power consumption. However, beta-1 for hydroelectric power is negative. The intercept is at 259.898 which is the value of hydroelectric power at time 0, but because the coefficient on time is negative (-0.083), this declines over time.

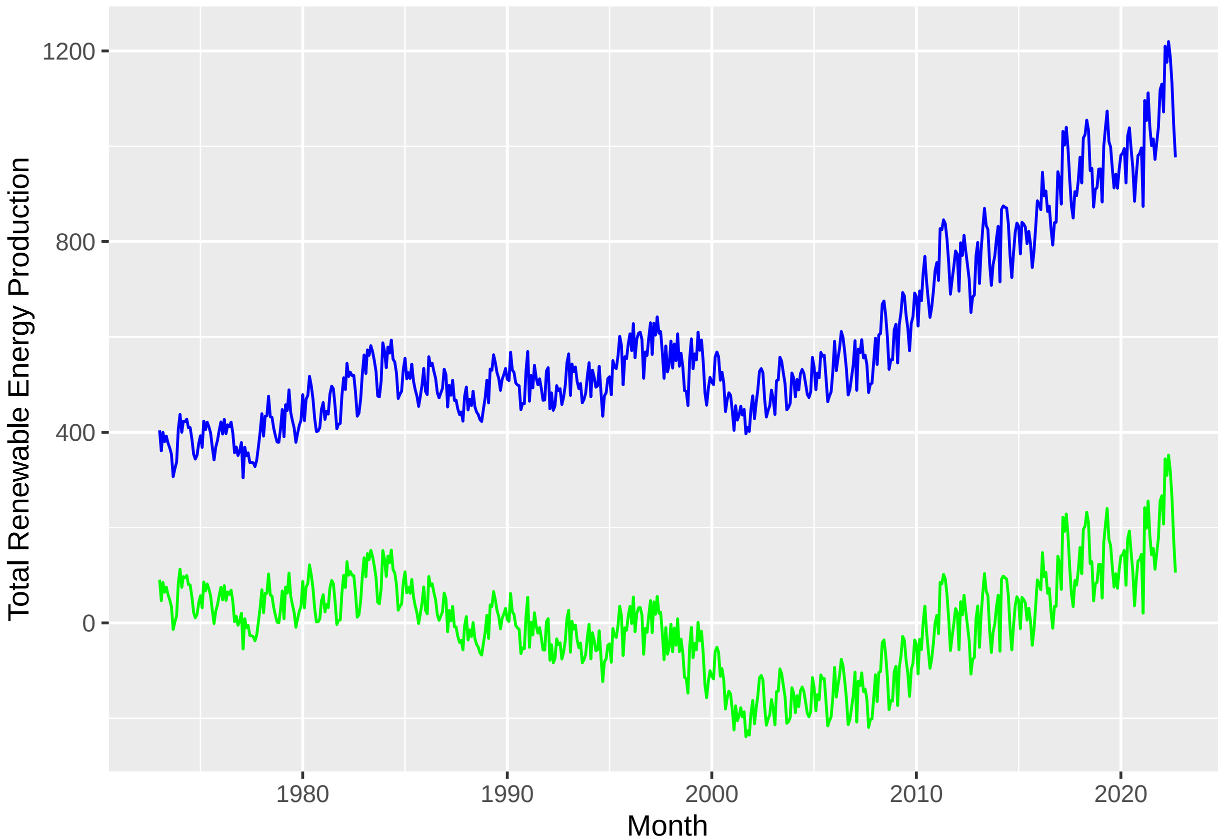
### Q4

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?

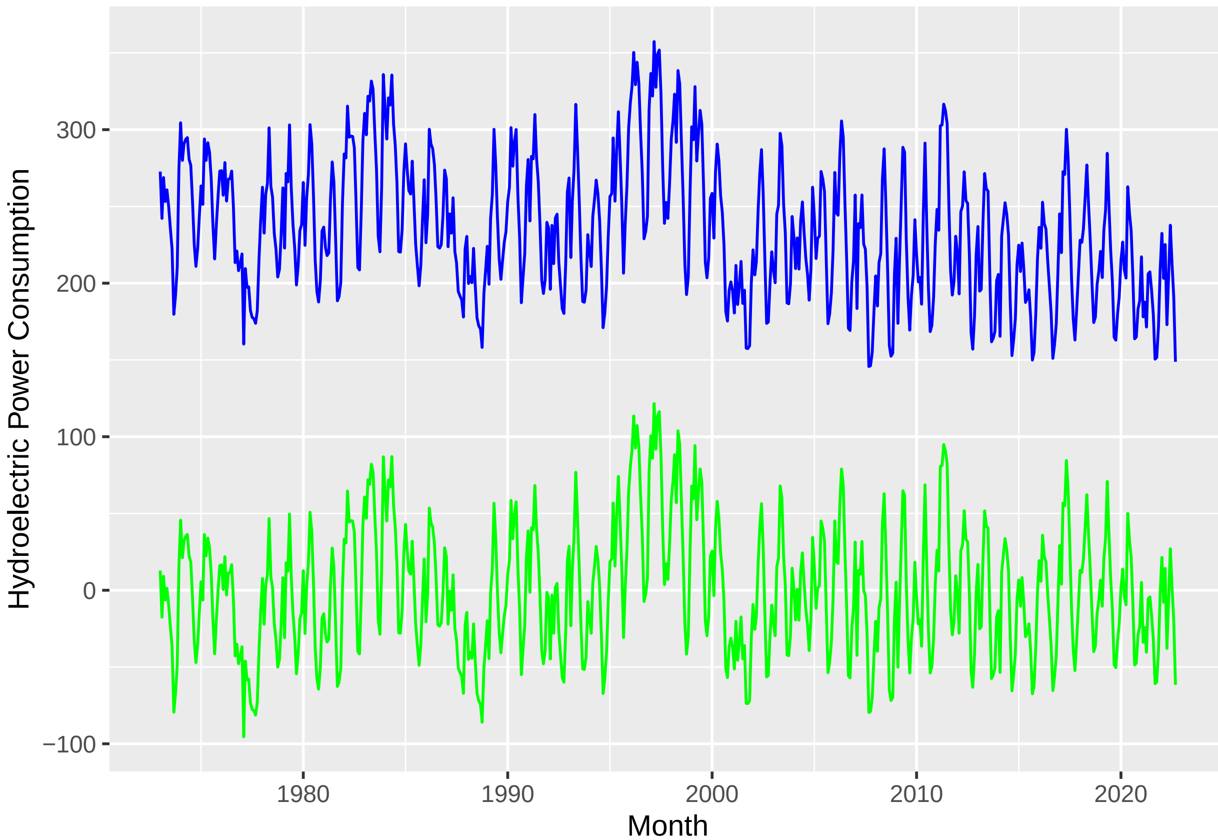
```
#Biomass
detrend_biomass <- energy_clean$`Total Biomass Energy Production` - (beta0_biomass+beta1_biomass*t)

ggplot(energy_clean, aes(x=Month, y=`Total Biomass Energy Production`))+
  geom_line(color = "blue")+
  geom_line(aes(y=detrend_biomass), color = "green")
```




```r
#Renewable Energy
detrend_RE <- energy_clean$`Total Renewable Energy Production` - (beta0_renew+beta1_renew*t)

ggplot(energy_clean, aes(x=Month, y=`Total Renewable Energy Production`))+
  geom_line(color = "blue")+
  geom_line(aes(y=detrend_RE), color = "green")
```




```
#Hydroelectric Power
detrend_hydro <- energy_clean$`Hydroelectric Power Consumption`- (beta0_hydro+beta1_hydro*t)

ggplot(energy_clean, aes(x=Month, y=`Hydroelectric Power Consumption`))+
  geom_line(color = "blue")+
  geom_line(aes(y=detrend_hydro), color = "green")
```

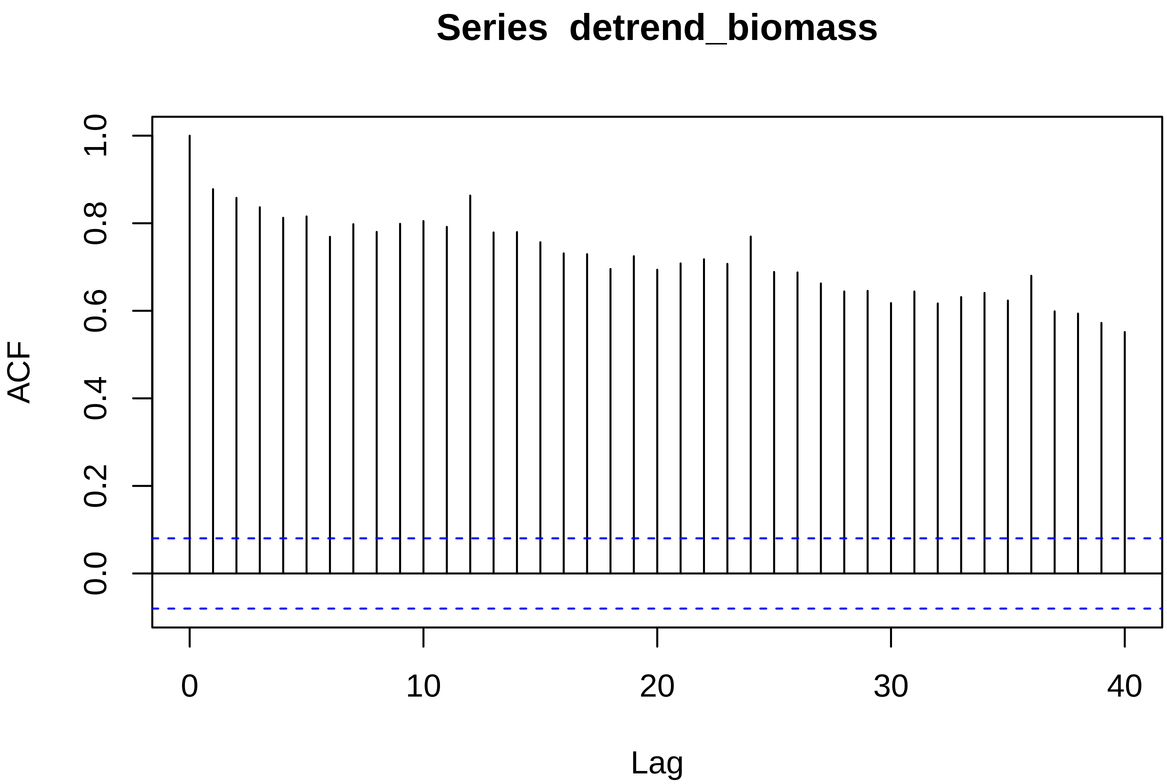



**Q4 Answer:** After removing the trend and plotting the detrended series, all of the plots changed. All of the plots moved down on the graph, with lower values than the non de-trended series. In addition, the first two plots (total renewable production and total biomass production) also were not only lower on the graph but the difference between the start and end values of energy production were closer together in both plots - essentially it is visible that the increasing trend in the data was removed making it closer to a flat trend over time. The decreasing trend in the hydroelectric plot is subtle, and thus this de-trending is more difficult to see in the detrended plot.

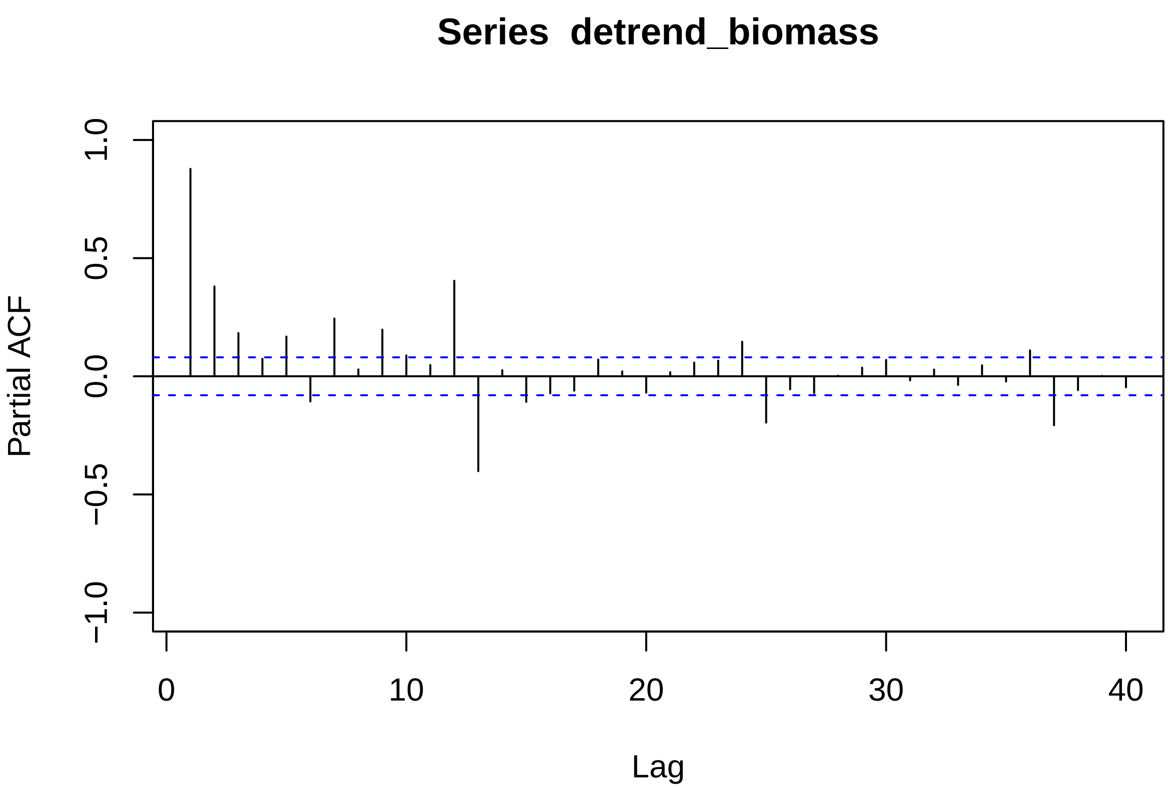
**Q5**

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

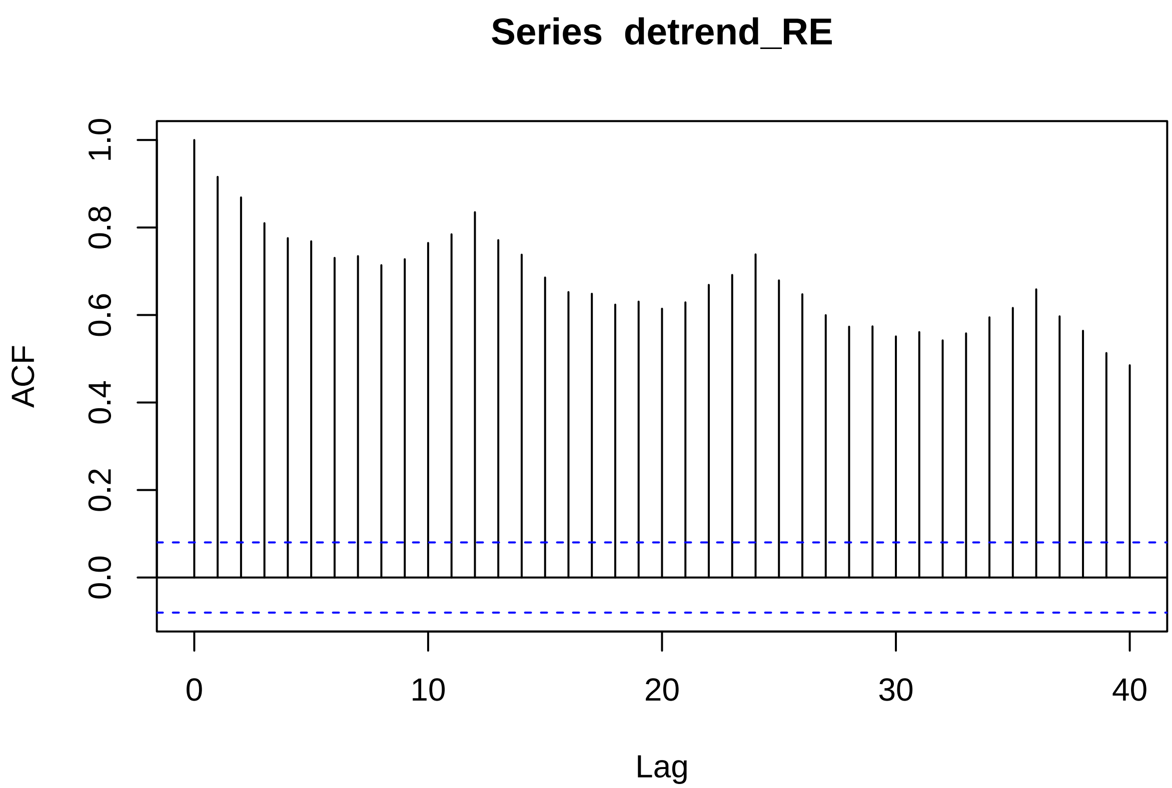
```
#Biomass

acf(detrend_biomass, lag.max=40)
```



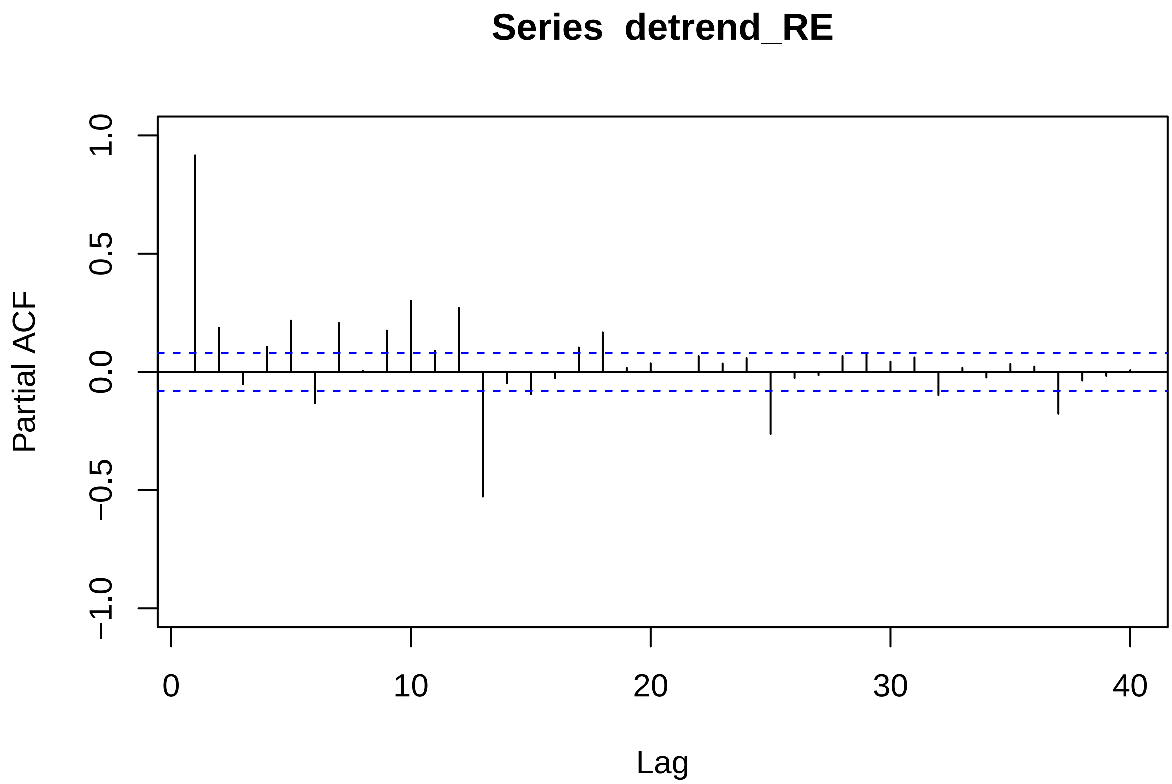

```r
#acf(energy_clean$`Total Biomass Energy Production`, lag.max = 40)
pacf(detrend_biomass, lag.max=40, ylim=c(-1,1))
```



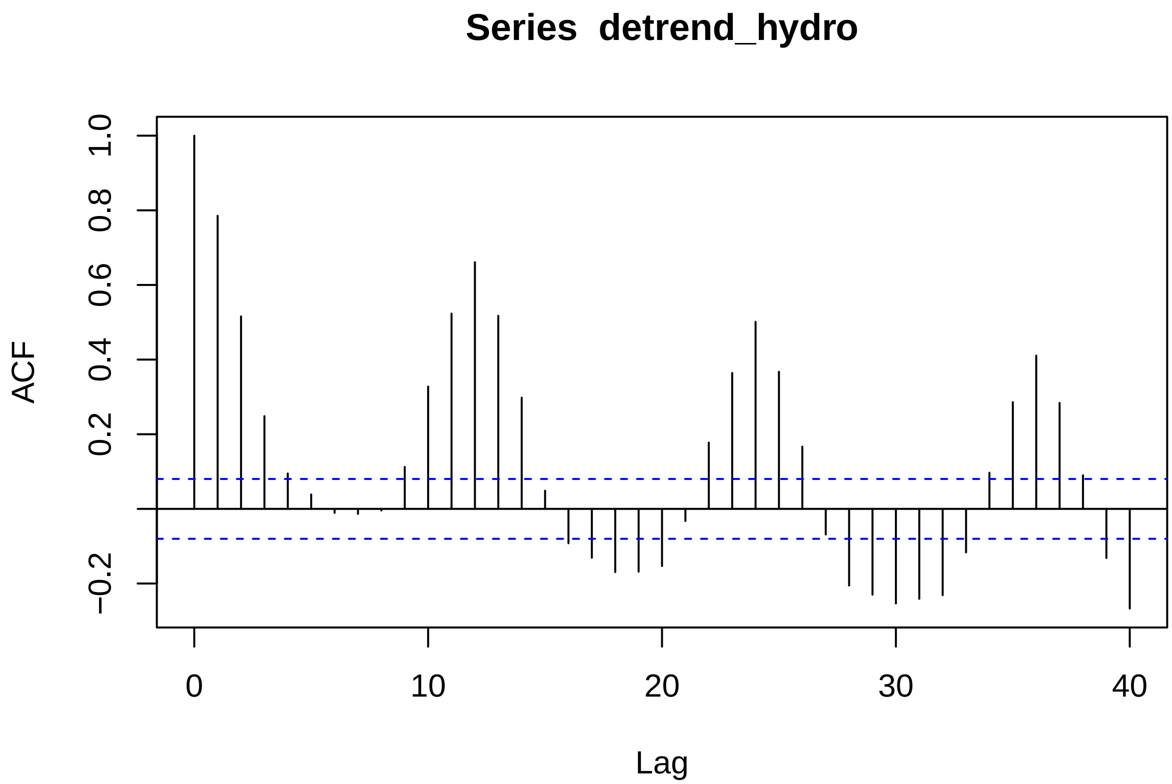
```
#pacf(energy_clean$`Total Biomass Energy Production`, lag.max = 40, ylim=c(-1,1))

#Renewable Energy
acf(detrend_RE, lag.max=40)
```



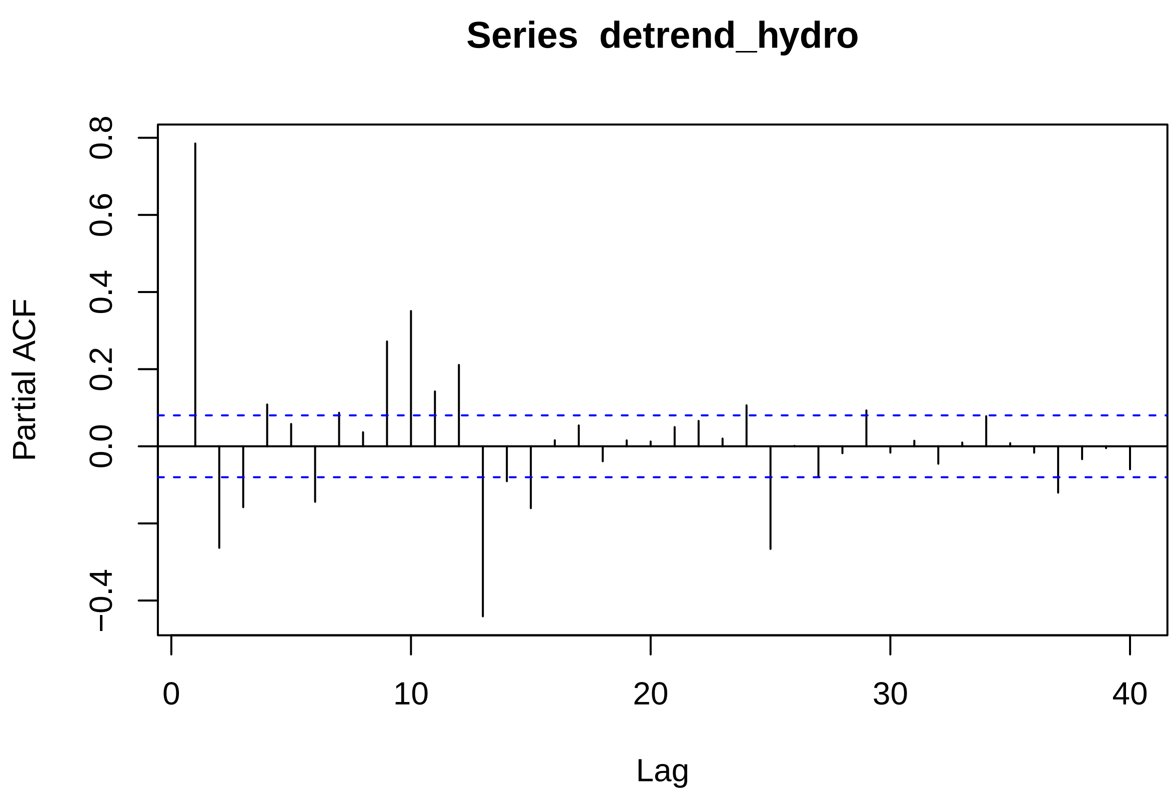

```
#acf(energy_clean$`Total Renewable Energy Production`, lag.max = 40)
pacf(detrend_RE, lag.max=40, ylim=c(-1,1))
```




```
#pacf(energy_clean$`Total Renewable Energy Production`, lag.max = 40, ylim=c(-1,1))


#Hydroelectric
acf(detrend_hydro, lag.max=40)
```




```
#acf(energy_clean$`Hydroelectric Power Consumption`, lag.max = 40)
pacf(detrend_hydro, lag.max=40)
```



```r
#pacf(energy_clean$`Hydroelectric Power Consumption`, lag.max = 40)
```

**Question 5 Answer:** In comparing ACF for biomass and for renewable energy, in both cases the detrended plots did have lower correlations than in the series with the trend. In both cases (biomass and renewable energy), the detrended plots appeared to show a seasonal component that was not as present in the original ACF.The ACFs, in both cases, also maintained their declining trend.

The PACF for renewable energy actually showed greater correlations (in both directions positive/negative) especially in the earlier lags, though looks more similar to the original PACF near lag 25. The same is true for the PACF for biomass production.

Finally, the ACF for hydroelectric without the trend has lower correlations, with more negative correlations as well as just broadly lower correlations even where positive. The PACFs for hydroelectric looked almost exactly the same, with minor differences in few correlations.

## 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.

```r
#Biomass
biomass_dummies <- seasonaldummy(ts_energy[,1])

biomass_seasMeans <- lm(energy_clean$`Total Biomass Energy Production` ~biomass_dummies)
summary(biomass_seasMeans )
```

```
##
## Call:
## lm(formula = energy_clean$‘Total Biomass Energy Production‘ ~
##     biomass_dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -160.74  -53.67  -24.36   90.73  181.34
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          288.020     13.163  21.881   <2e-16 ***
## biomass_dummiesJan    -1.793     18.522  -0.097   0.9229
## biomass_dummiesFeb   -31.102     18.522  -1.679   0.0936 .
## biomass_dummiesMar    -9.104     18.522  -0.492   0.6232
## biomass_dummiesApr   -21.502     18.522  -1.161   0.2462
## biomass_dummiesMay   -14.238     18.522  -0.769   0.4424
## biomass_dummiesJun   -19.602     18.522  -1.058   0.2904
## biomass_dummiesJul    -3.674     18.522  -0.198   0.8428
## biomass_dummiesAug    -0.612     18.522  -0.033   0.9737
## biomass_dummiesSep   -13.335     18.522  -0.720   0.4718
```

```
## biomass_dummiesOct   -4.030     18.615  -0.216   0.8287
## biomass_dummiesNov   -9.849     18.615  -0.529   0.5970
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.14 on 585 degrees of freedom
## Multiple R-squared:  0.01018,	Adjusted R-squared:  -0.008437
## F-statistic: 0.5467 on 11 and 585 DF,  p-value: 0.8714
```

**Biomass Seasonal Means Model: Interpretation** The P-value is 0.8714 on the F-statistic indicating that the model is not statistically significant. This indicates that there is not evidence of seasonality in the dataset. Because of this, I will not de-season the biomass data (and thus did not save the coefficients).

```
#Total Renewable Energy
RE_dummies <- seasonaldummy(ts_energy[,2])

RE_seasMeans <- lm(energy_clean$`Total Renewable Energy Production` ~ RE_dummies)
summary(RE_seasMeans )
```

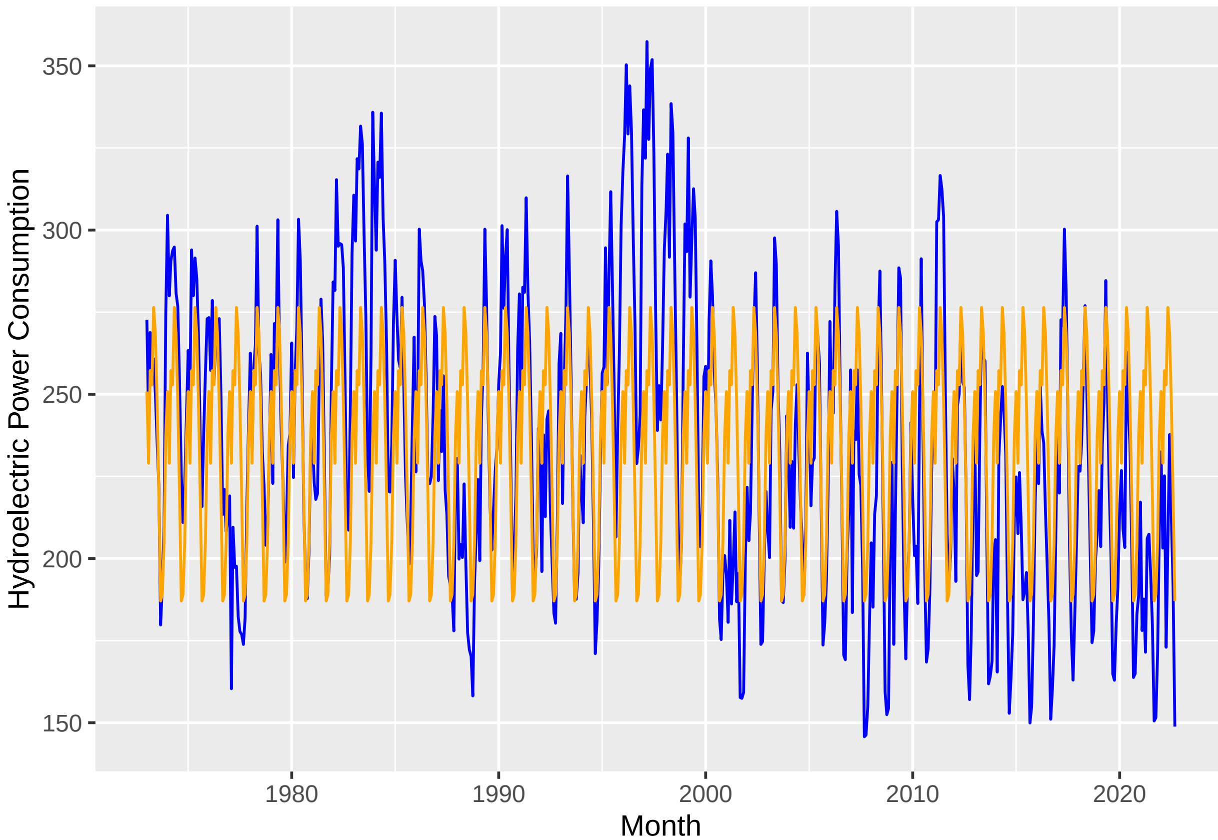
```
##
## Call:
## lm(formula = energy_clean$`Total Renewable Energy Production` ~
##     RE_dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -284.92 -122.23  -68.42   91.22  585.68
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    601.022     27.260  22.048   <2e-16 ***
## RE_dummiesJan   11.468     38.358   0.299    0.765
## RE_dummiesFeb  -41.456     38.358  -1.081    0.280
## RE_dummiesMar   23.130     38.358   0.603    0.547
## RE_dummiesApr    9.959     38.358   0.260    0.795
## RE_dummiesMay   38.853     38.358   1.013    0.312
## RE_dummiesJun   20.378     38.358   0.531    0.595
## RE_dummiesJul    8.298     38.358   0.216    0.829
## RE_dummiesAug  -19.450     38.358  -0.507    0.612
## RE_dummiesSep  -63.770     38.358  -1.662    0.097 .
## RE_dummiesOct  -52.612     38.551  -1.365    0.173
## RE_dummiesNov  -42.537     38.551  -1.103    0.270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 190.8 on 585 degrees of freedom
## Multiple R-squared:  0.02844,	Adjusted R-squared:  0.01017
## F-statistic: 1.557 on 11 and 585 DF,  p-value: 0.1076
```

**Renewable Energy Seasonal Means Model: Interpretation** P-value on F-statistic is 0.1076 which is above the 0.05 threshold for significance (though only slightly), indicating there is not strong statistical evidence of seasonality. For this reason, I will not de-season the renewable energy data.

```r
#Hydroelectric
hydro_dummies <- seasonaldummy(ts_energy[,3])

hydro_seasMeans <- lm(energy_clean$`Hydroelectric Power Consumption` ~hydro_dummies)
summary(hydro_seasMeans)
```

```
##
## Call:
## lm(formula = energy_clean$‘Hydroelectric Power Consumption‘ ~
##     hydro_dummies)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -88.99 -23.47  -2.81  21.99 100.18
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       237.225      4.878  48.634  < 2e-16 ***
## hydro_dummiesJan   13.594      6.864   1.981  0.04811 *
## hydro_dummiesFeb   -8.254      6.864  -1.203  0.22964
## hydro_dummiesMar   19.980      6.864   2.911  0.00374 **
## hydro_dummiesApr   15.649      6.864   2.280  0.02297 *
## hydro_dummiesMay   39.210      6.864   5.713 1.77e-08 ***
## hydro_dummiesJun   31.209      6.864   4.547 6.61e-06 ***
## hydro_dummiesJul   10.436      6.864   1.520  0.12895
## hydro_dummiesAug  -17.909      6.864  -2.609  0.00931 **
## hydro_dummiesSep  -50.173      6.864  -7.310 8.82e-13 ***
## hydro_dummiesOct  -48.262      6.898  -6.996 7.22e-12 ***
## hydro_dummiesNov  -32.285      6.898  -4.680 3.56e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 34.14 on 585 degrees of freedom
## Multiple R-squared:  0.4132, Adjusted R-squared:  0.4022
## F-statistic: 37.45 on 11 and 585 DF,  p-value: < 2.2e-16
```

```r
beta0_hydroS <- hydro_seasMeans$coefficients[1]
beta1_hydroS <- hydro_seasMeans$coefficients[2:12]

hydro_seas_comp <- array(0,nobs)
for (i in 1:nobs){
  hydro_seas_comp[i] = beta0_hydroS + beta1_hydroS %*% hydro_dummies[i,]
}

#Visualize the original data and the seasonal component
ggplot(energy_clean, aes(x=Month, y=`Hydroelectric Power Consumption`))+
  geom_line(color="blue")+
  geom_line(aes(y=hydro_seas_comp), color="orange")
```

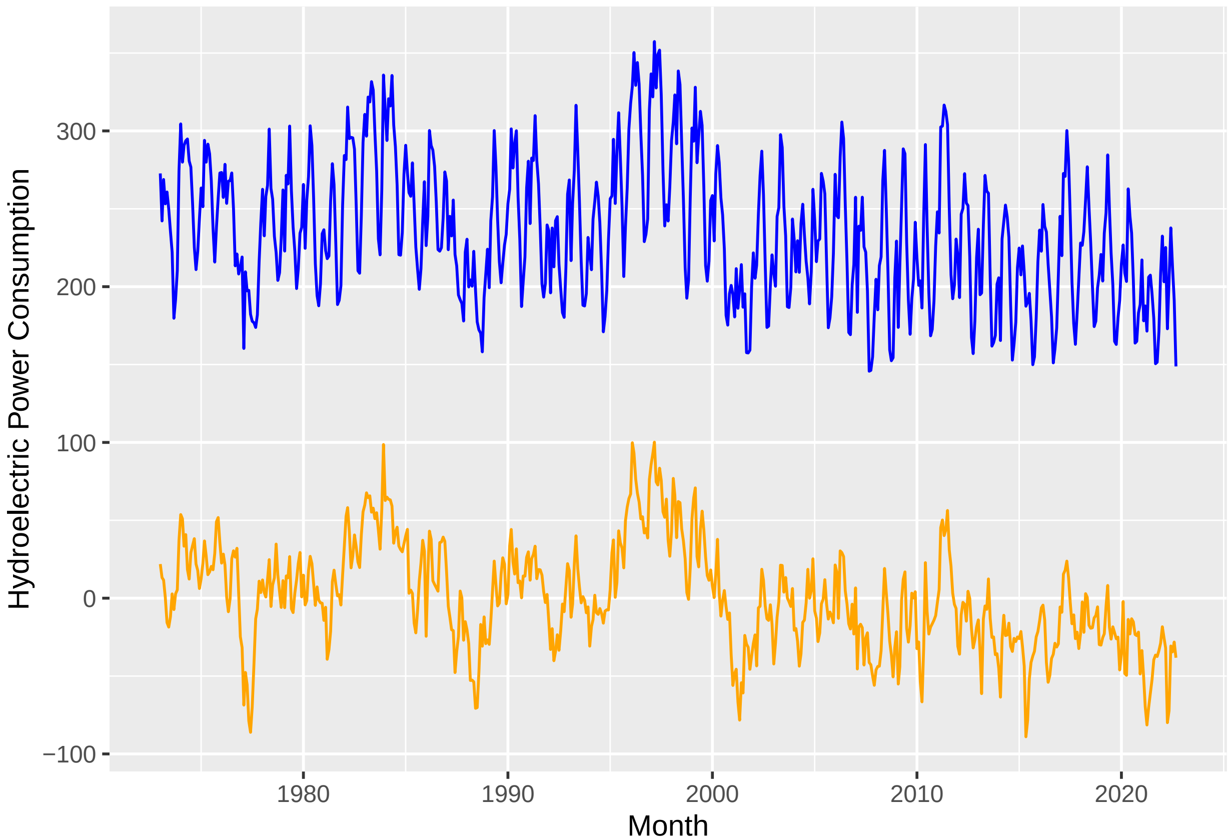



**Hydroelectric Seasonal Means Model: Interpretation** The F-statistic p-value is <0.001 suggesting there is seasonality present. The following analysis on seasonality will only be performed on the hydroelectric consumption data.

### 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?

```
#Removing seasonal component
deSeas_hydro <- energy_clean$`Hydroelectric Power Consumption`- hydro_seas_comp

#Plot
ggplot(energy_clean, aes(x=Month, y=`Hydroelectric Power Consumption`))+
  geom_line(color="blue")+
  geom_line(aes(y=deSeas_hydro), color="orange")
```

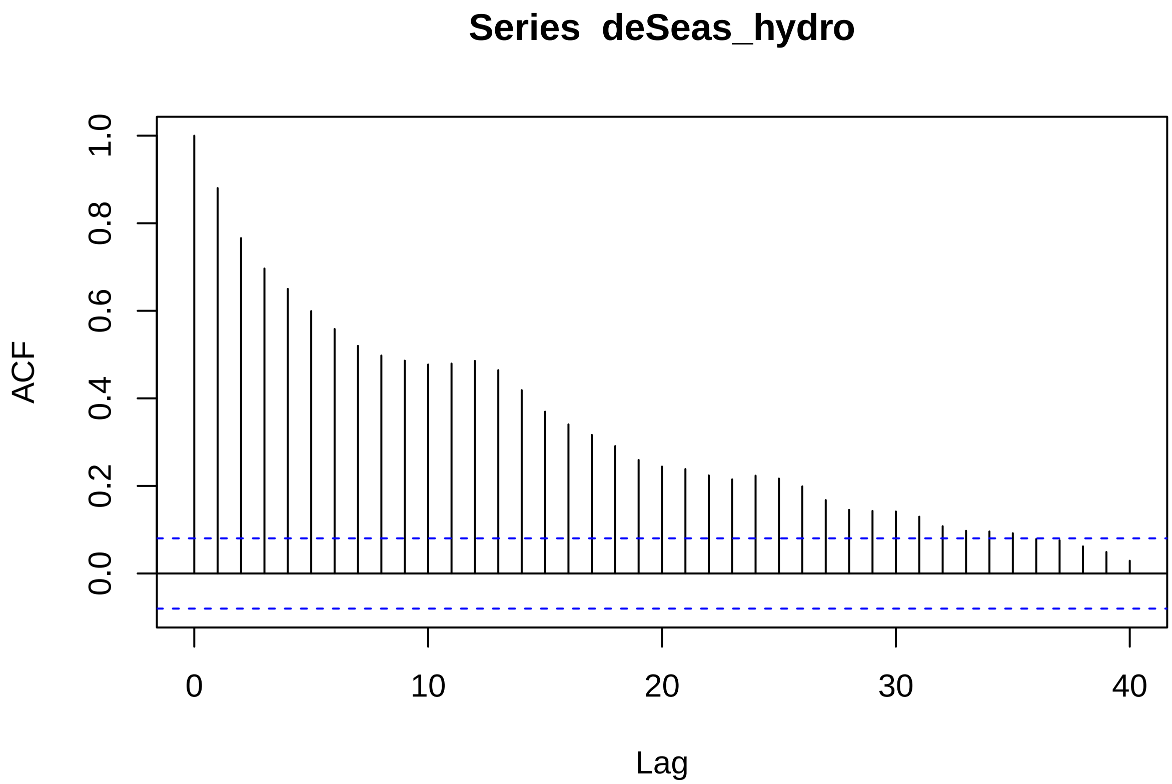



#### Q7 Answer: Yes, the deseason series is different from the original data. In particular there is less of a constant increase/decrease in the line, as well as the data moving lower on the plot ranging from ~-100 to ~100 in the deseason-ed series, whereas the original series ranges from ~150 to ~350.

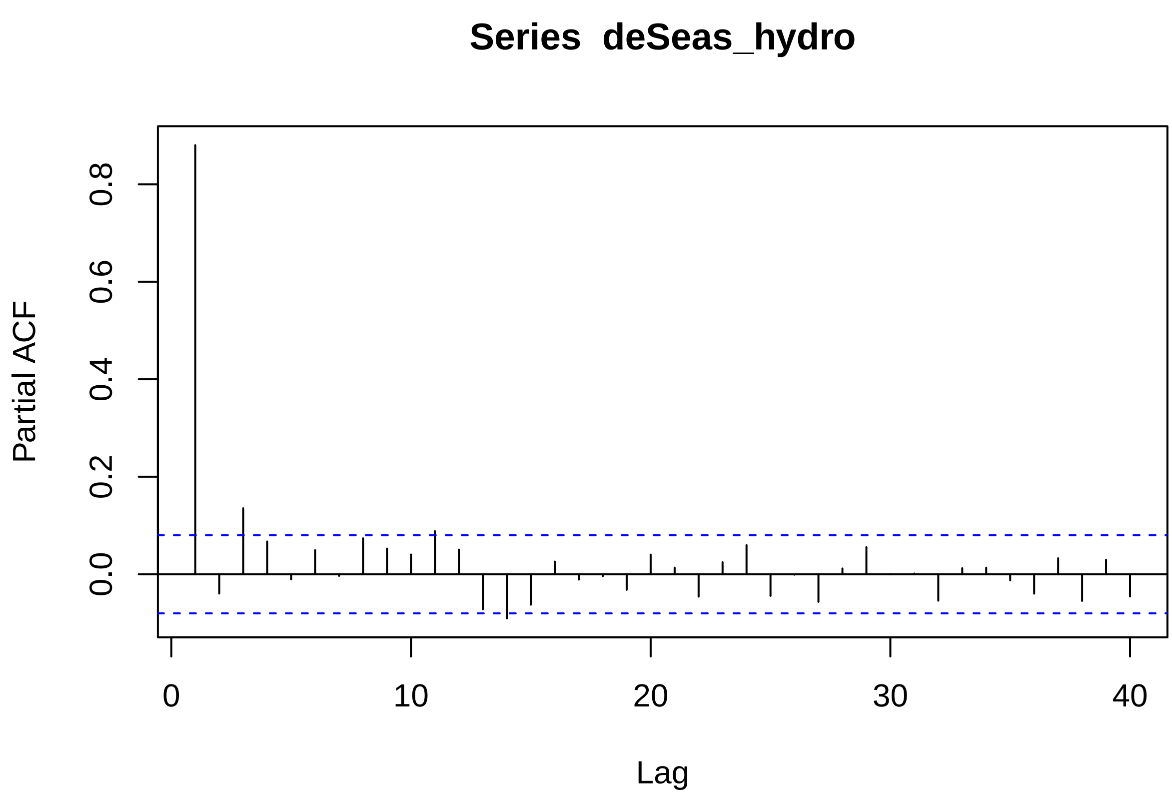
**Q8**

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

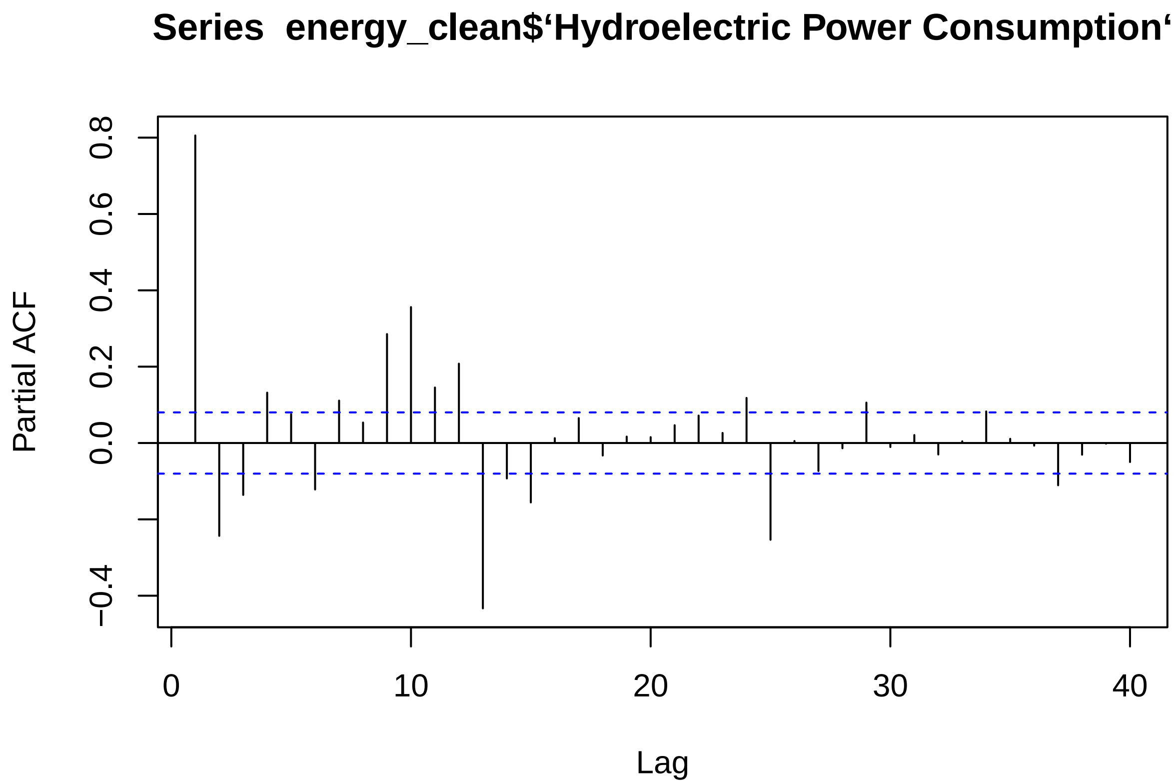
```
#Hydroelectric
acf(deSeas_hydro, lag.max=40)
```




```
#acf(energy_clean$`Hydroelectric Power Consumption`, lag.max = 40)
pacf(deSeas_hydro, lag.max=40)
```



```r
pacf(energy_clean$`Hydroelectric Power Consumption`, lag.max = 40)
```




**Q8 Answer:** Yes, the plots changed significantly – in the original ACF the seasonality was clear, with the correlations changing from positive to negative or simply lower to higher. The deseasoned plot no longer has this effect and instead just shows the declining trend. The PACFs look very different as well, specifically the original PACF had larger correlations (both more positive and more negative) even near lag 40. The deseasoned PACF does not have significant correlations (those that extend beyond the blue line) after lag 10.