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

Assignment 3 - Due date 02/12/21

Yash Doshi

---

## 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.

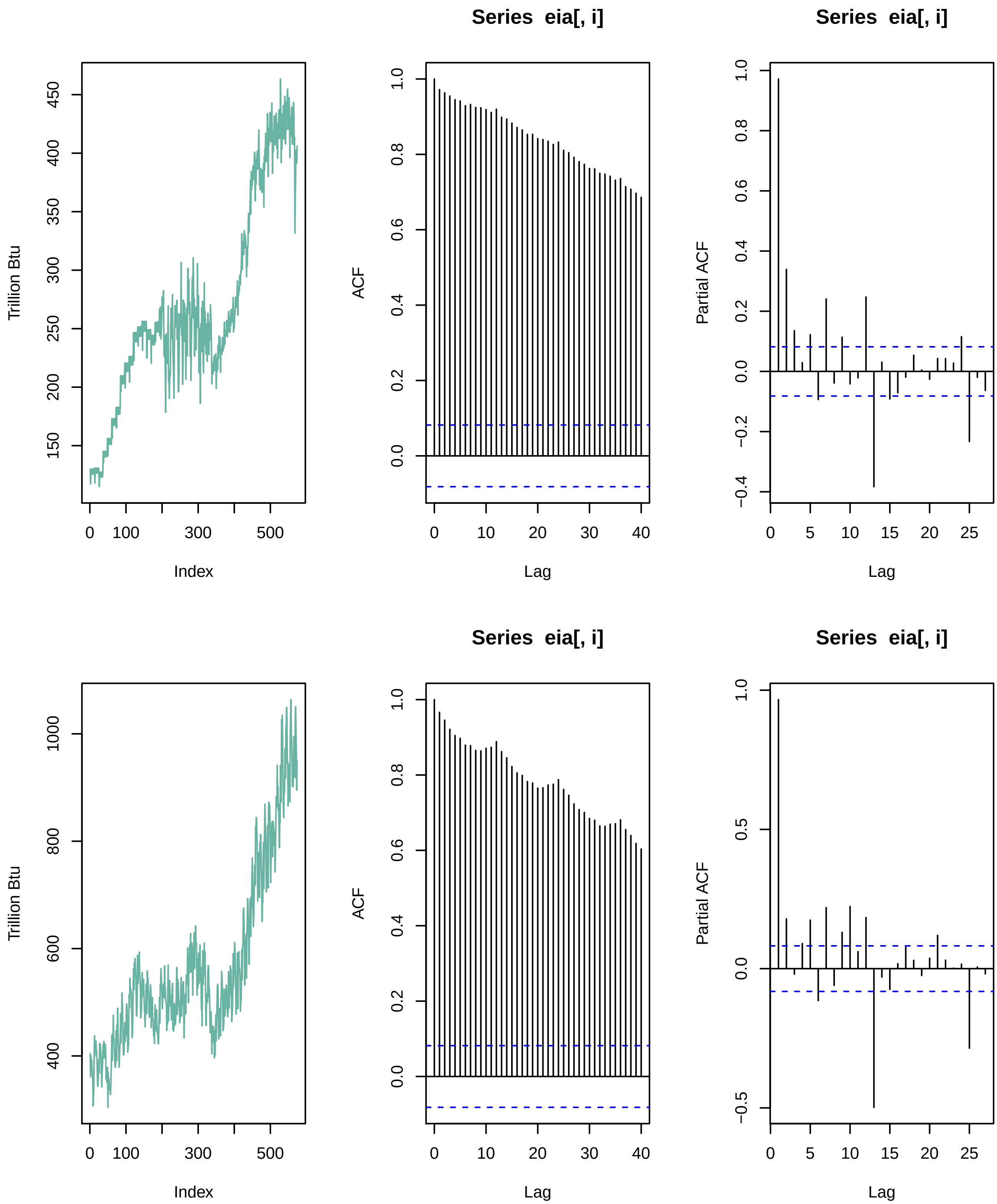
R packages needed for this assignment:"forecast","tseries", and "Kendall". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

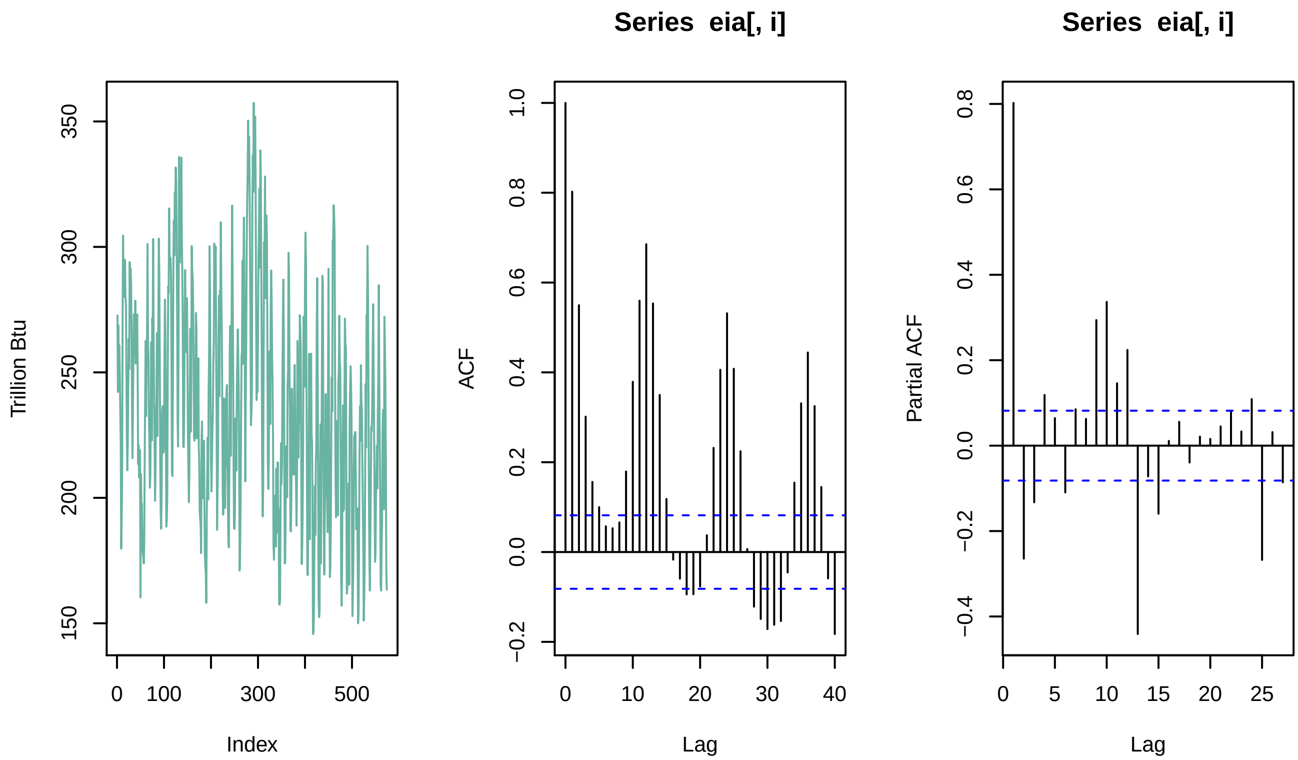
---

# 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 same code form A2, but I want all three plots on the same window this time. (Hint: watch videos for M4)**


Series eia[, i]
Series eia[, i]
Trillion Btu
Index
ACF
Lag
Partial ACF
Lag
Series eia[, i]
Series eia[, i]
Trillion Btu
Index
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?**

*Total Biomas Energy Production*

This plot shows that there is some sort of trend. It can be seen that the graph is initially rising, then it becomes constant for osme time, then it rises steeply again. During the end (2020), it can be seen that the trend dropped. This might be due to COVID.

*Total Renewable Enegry Production*

This plot shows that there is a trend. The trend observed in this plot is very similar to that observed in the, "Total Biomass Energy Production." It can be seen that the graph is initially rising, then it becomes constant for osme time, then it rises steeply again.

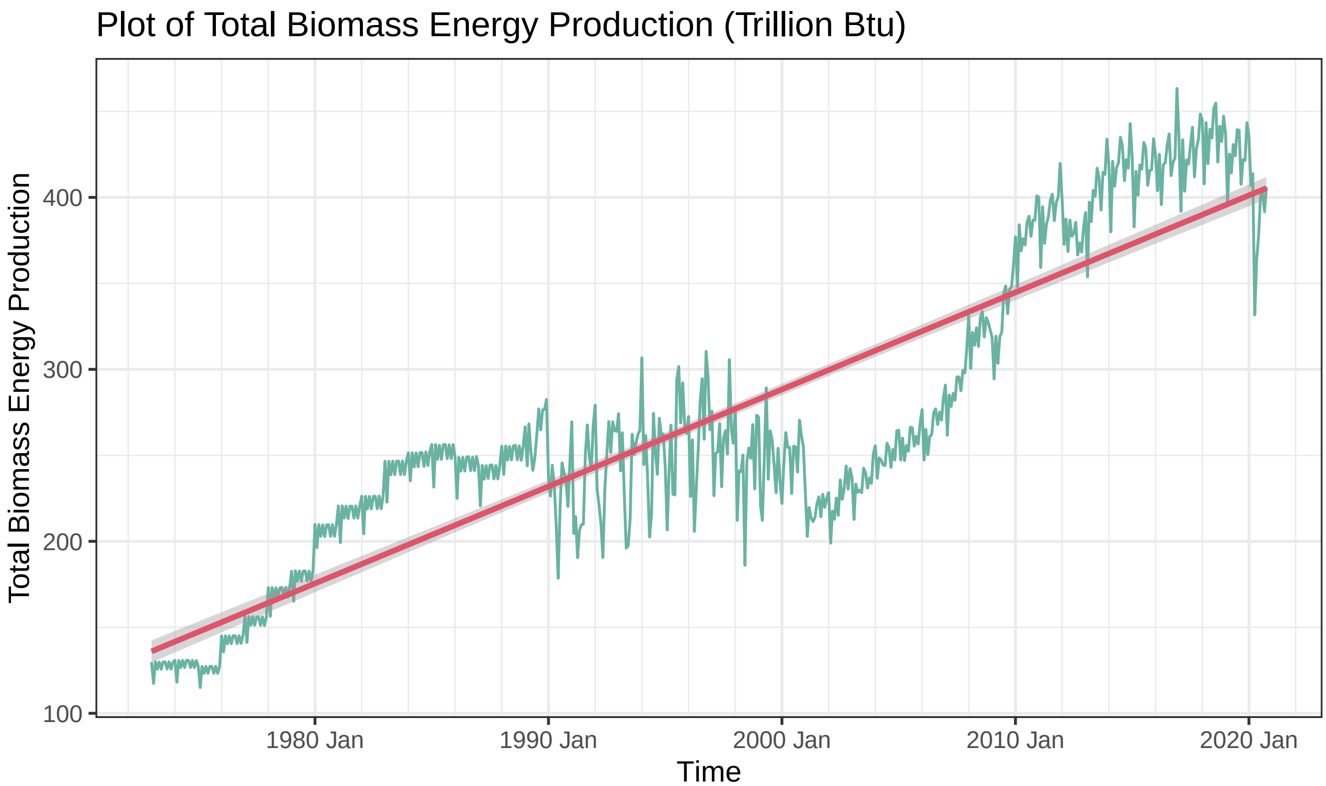
*Hydroelectric Power Consumption*

This plot shows very minimal trend. If observed carefully, it can be seen that the trend is decreasing in the later years.

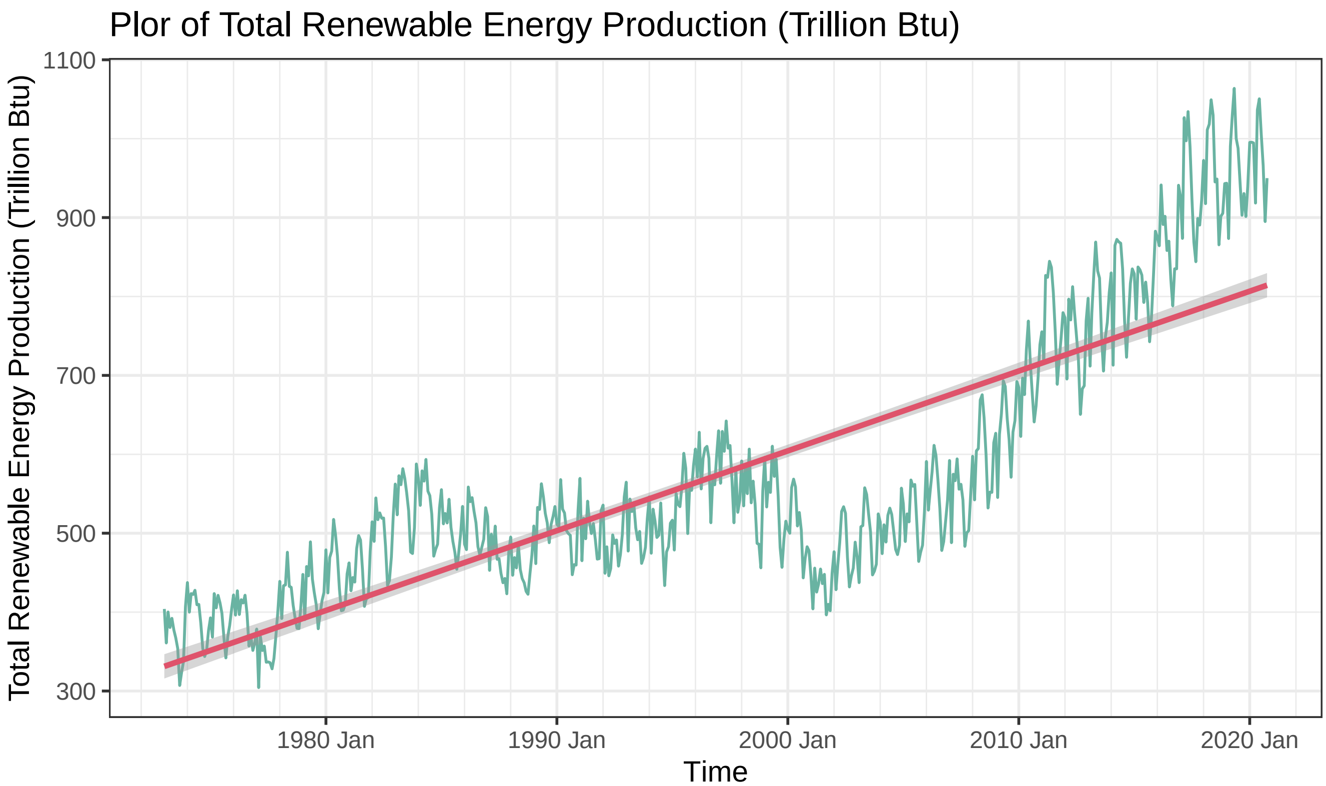
# Q3

**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.**

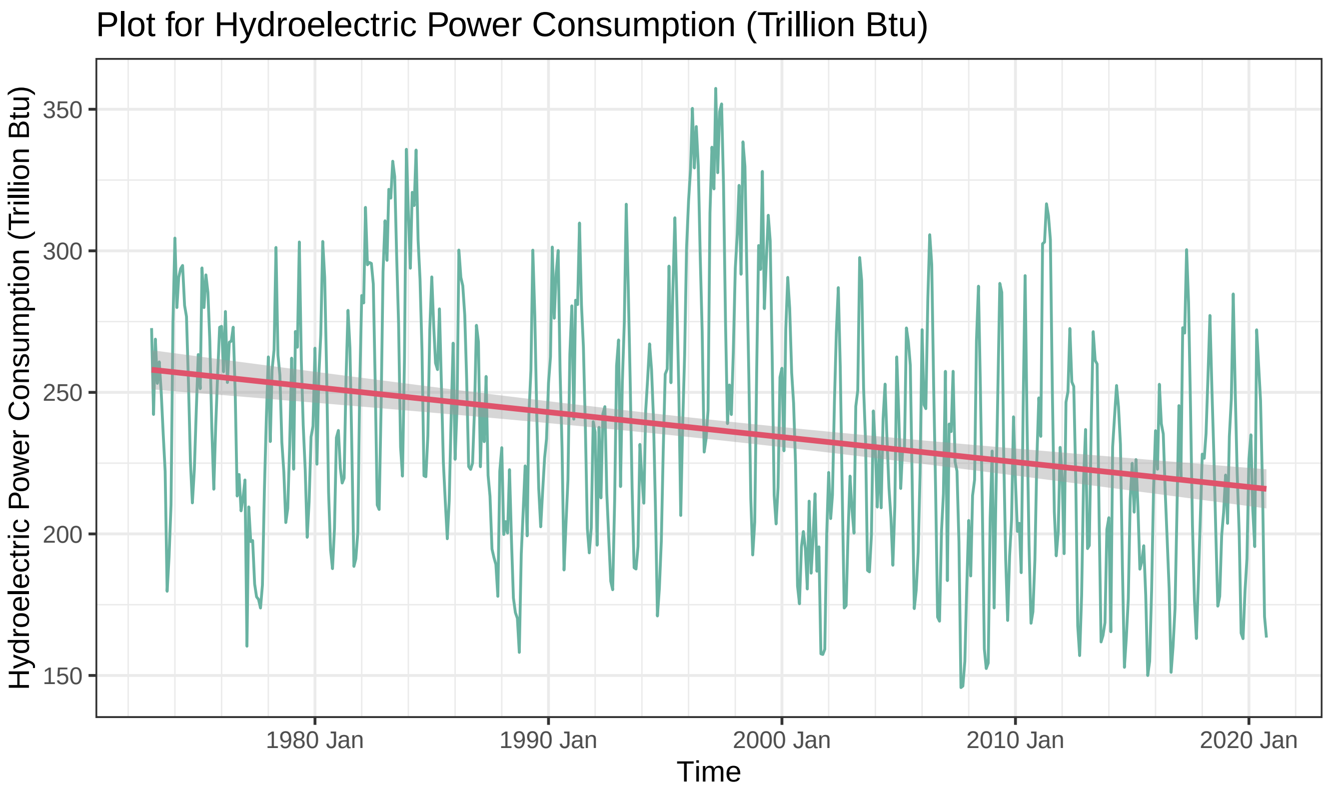
Below are the plots and linear regression summaries of all three columns.




```
Call:
lm(formula = eia$`Total Biomass Energy Production (Trillion Btu)` ~
    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
```



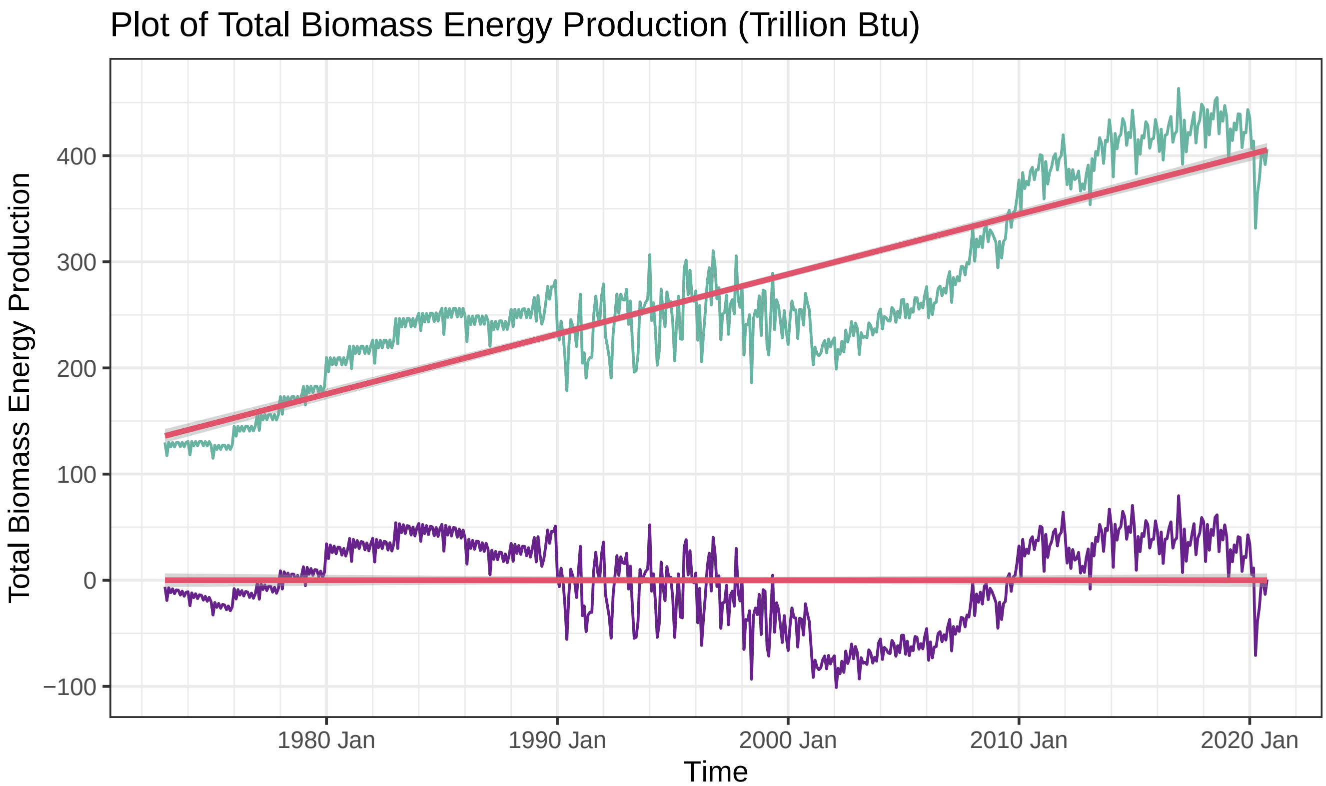

```
Call:
lm(formula = eia$‘Total Renewable Energy Production (Trillion Btu)‘ ~
    t)

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
```



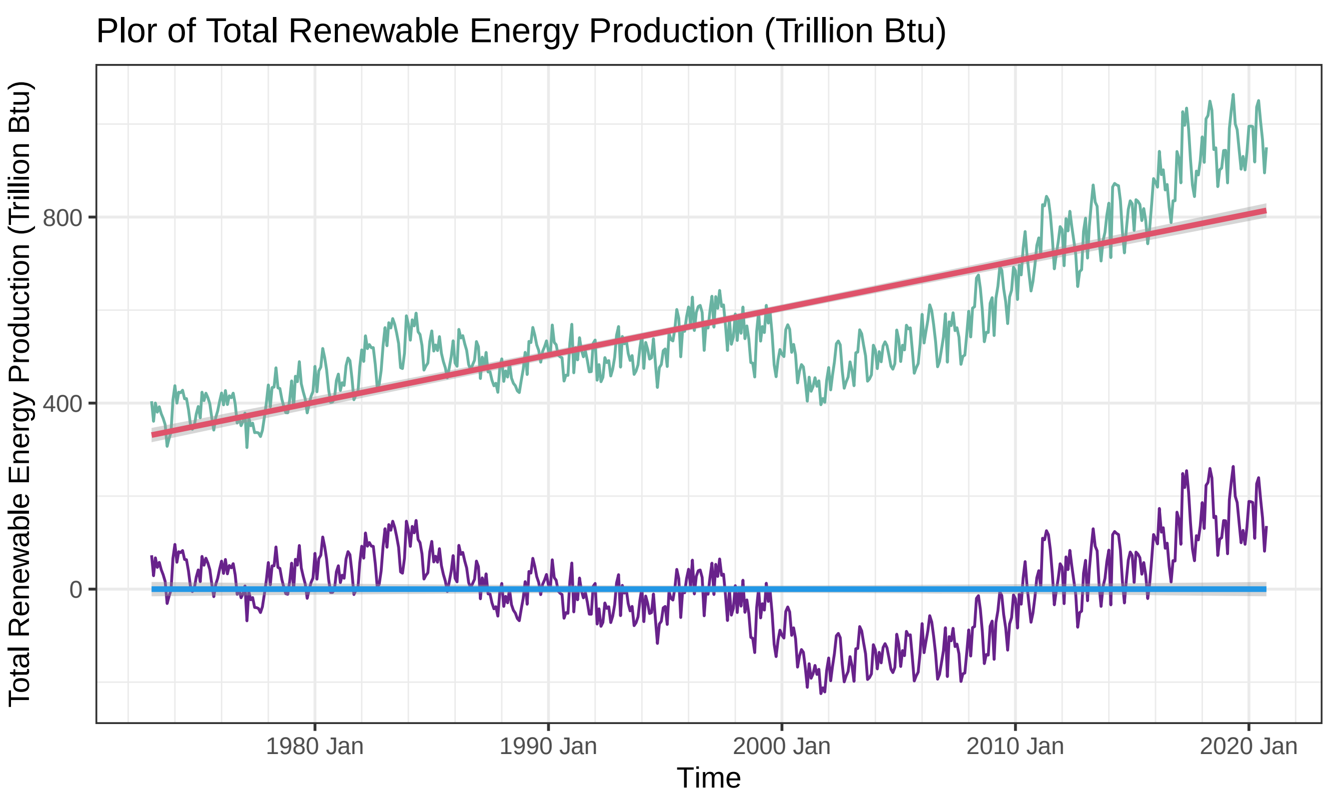

```
Call:
lm(formula = eia$`Hydroelectric Power Consumption (Trillion Btu)` ~
    t)

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

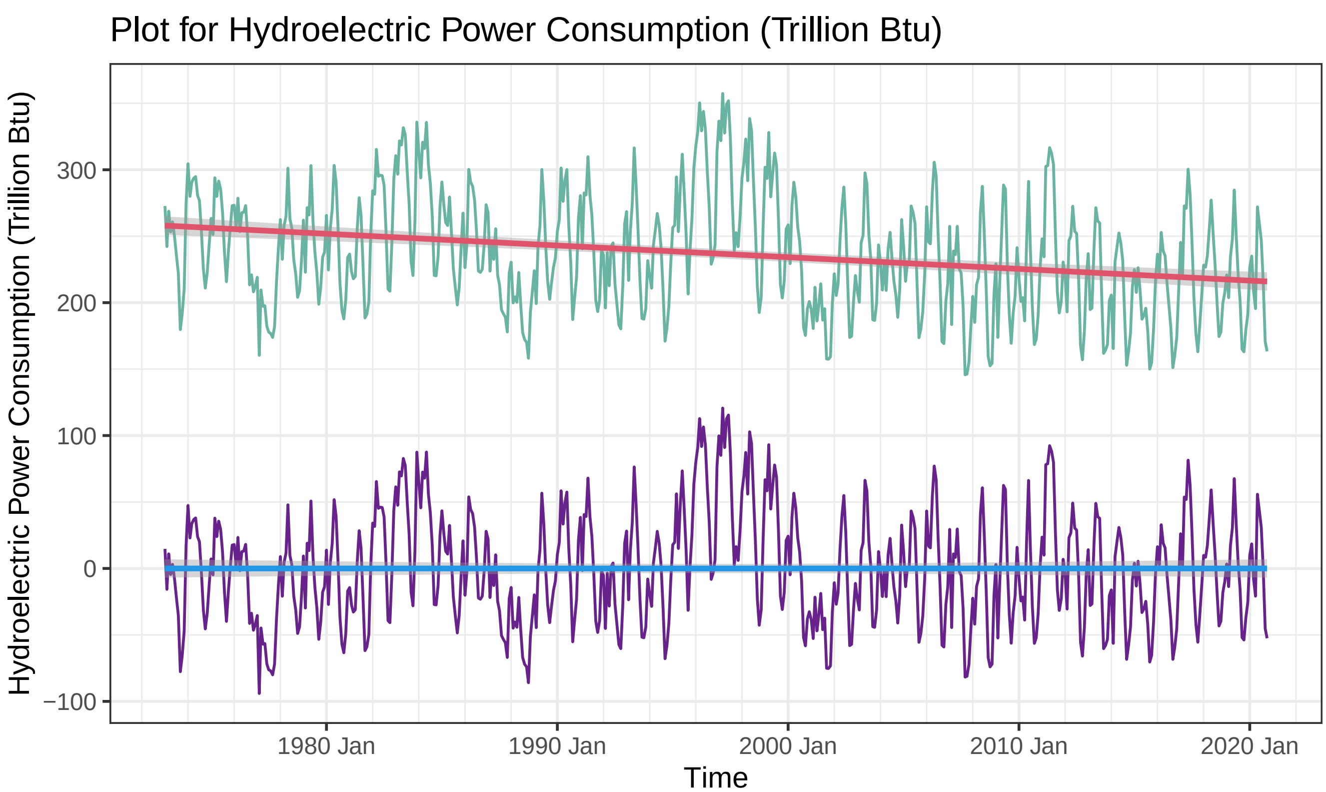
**Use the regression coefficients from Q3 to detrend the series. Plot the detrended series and compare with the plots from Q1. What happened? Did anything change?**




The blue plot shows the original plot with the trend, whereas, the purple one shows no trend. Total Biomass Energy Production showed an increasing trend. It means that as time passed, the amount of biomass produced was increasing. After detrending, we can see that the trend is at 0 and is remaining constant. Detrending also helps us in identifying any seasonality (if there is any). In thi case there is no visible trend.



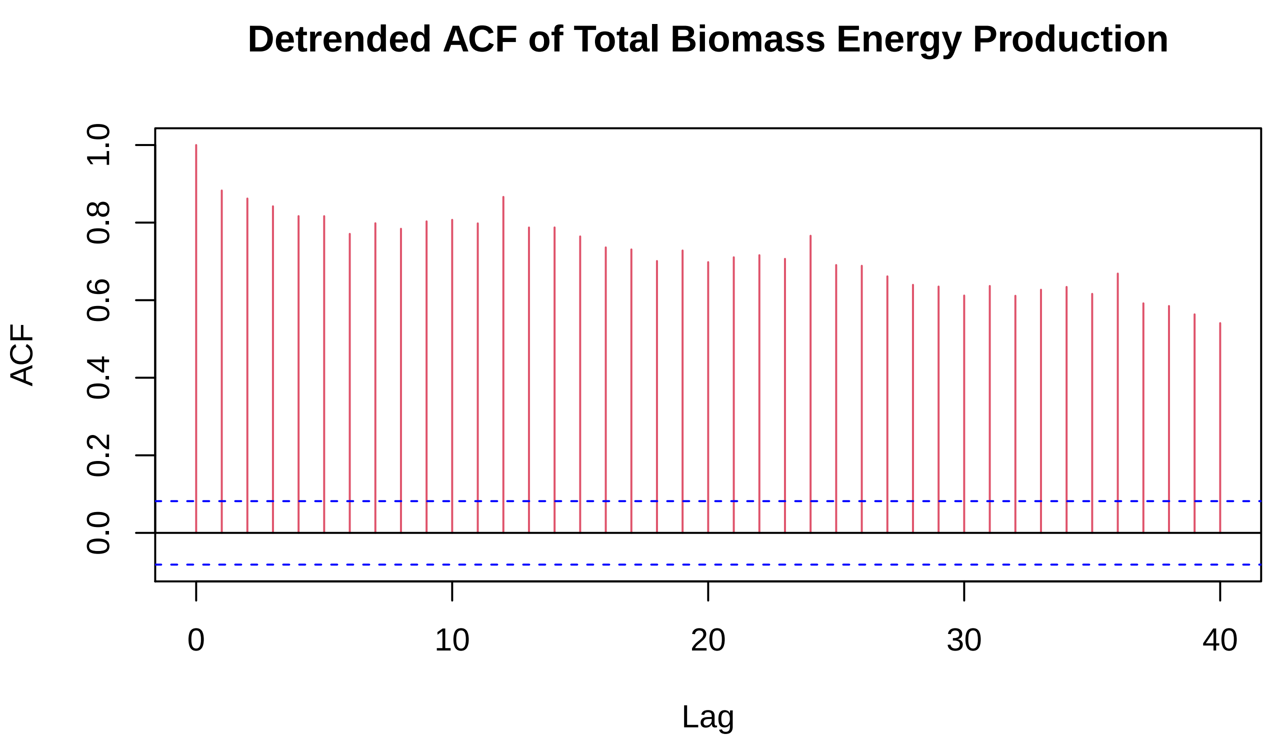

The above plot on Total Renewable Energy Production shows us the trend (blue plot) and detrend (purple plot). After detrending, we can see that the trend is at 0 and is remaining constant. Detrending also helps us in identifying any seasonality (if there is any). In thi case there is no visible trend.

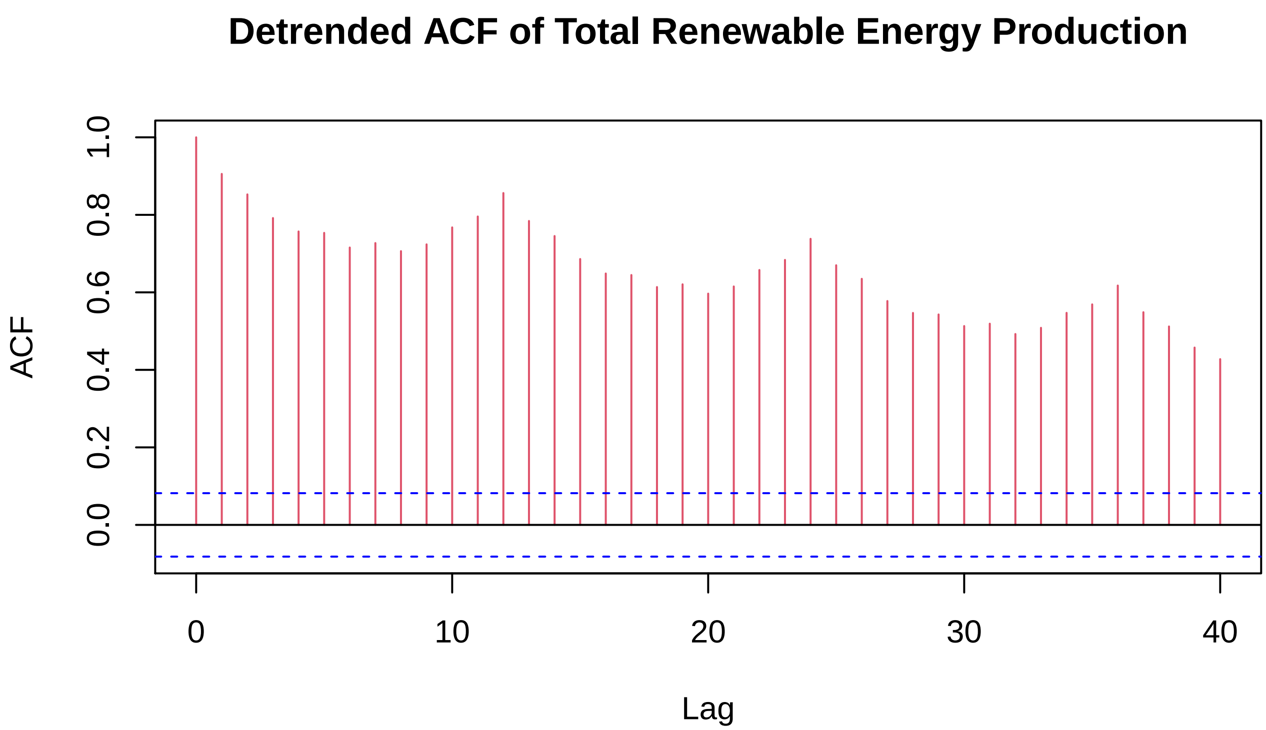


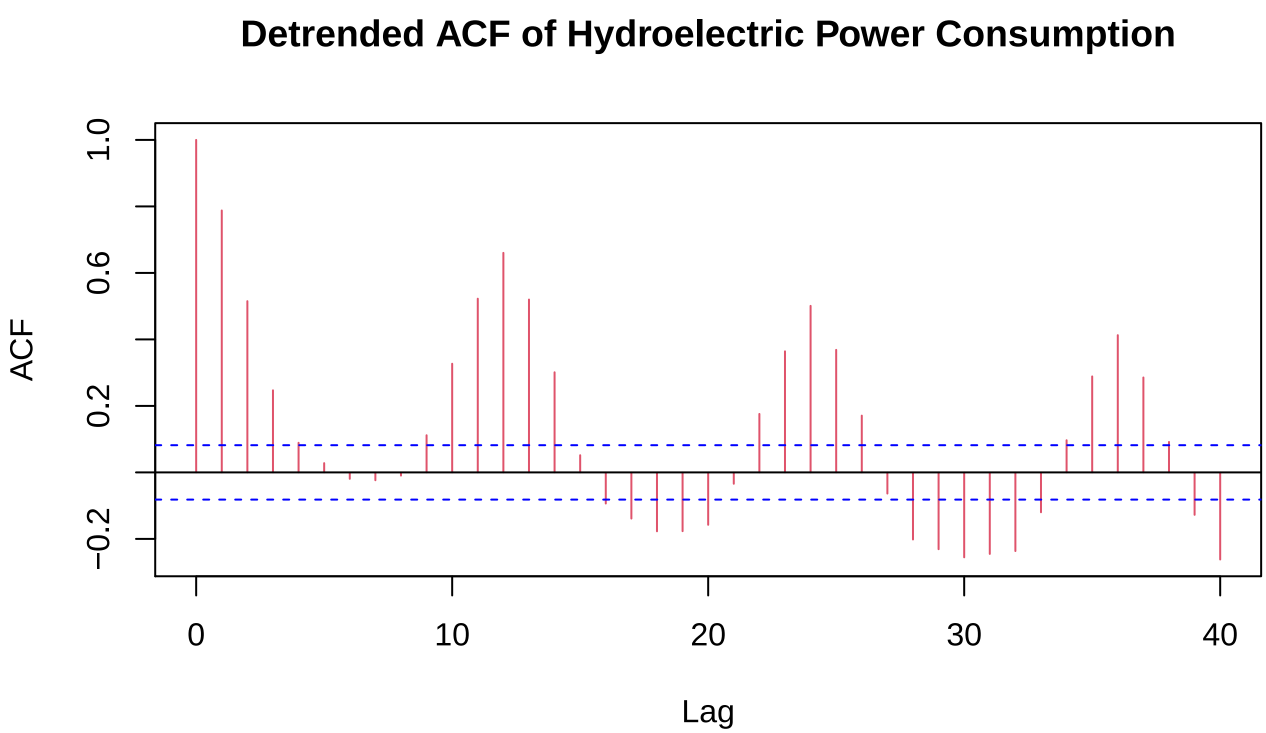

In the above Hydrolic Power Consumption plot, we can clearly see that the detrended line (purple color) has seasonality. Therefore, detrending helped us in analyzing whether a partiuclar data is seasonal or not.

## Q5

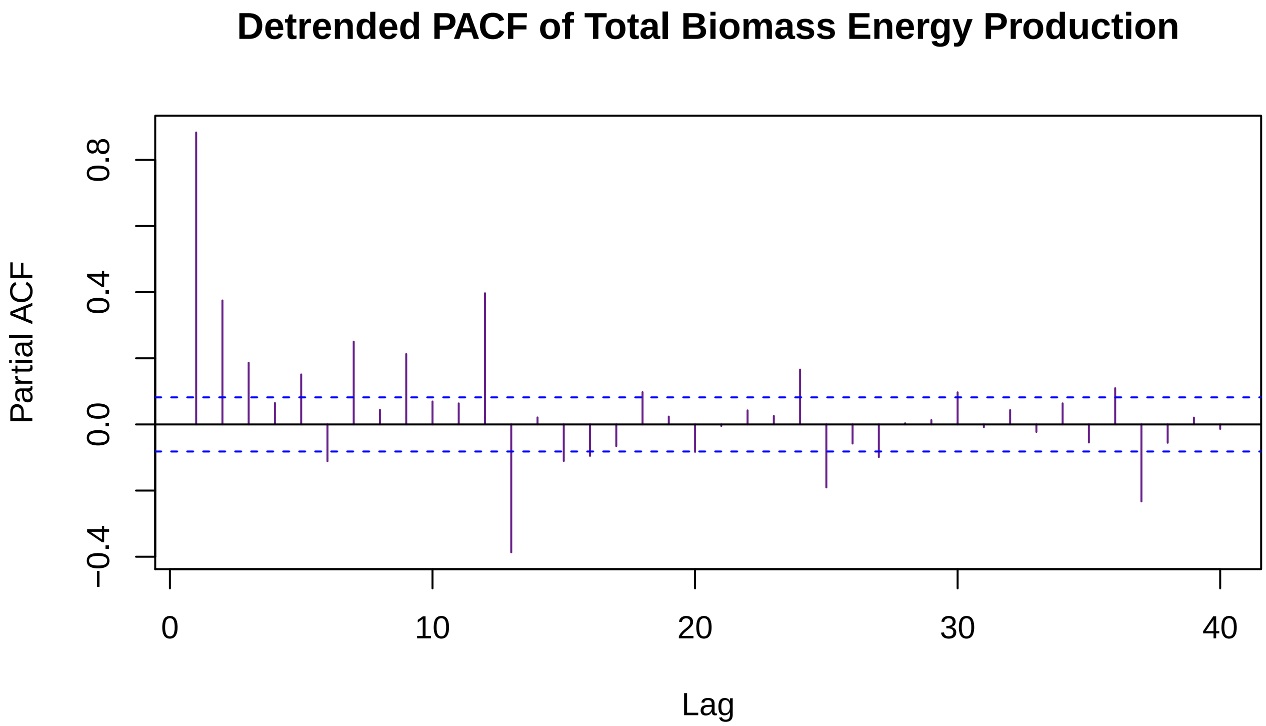
**Plot ACF and PACF for the detrended series and compare with the plots from Q1. Did the plots change? How?**

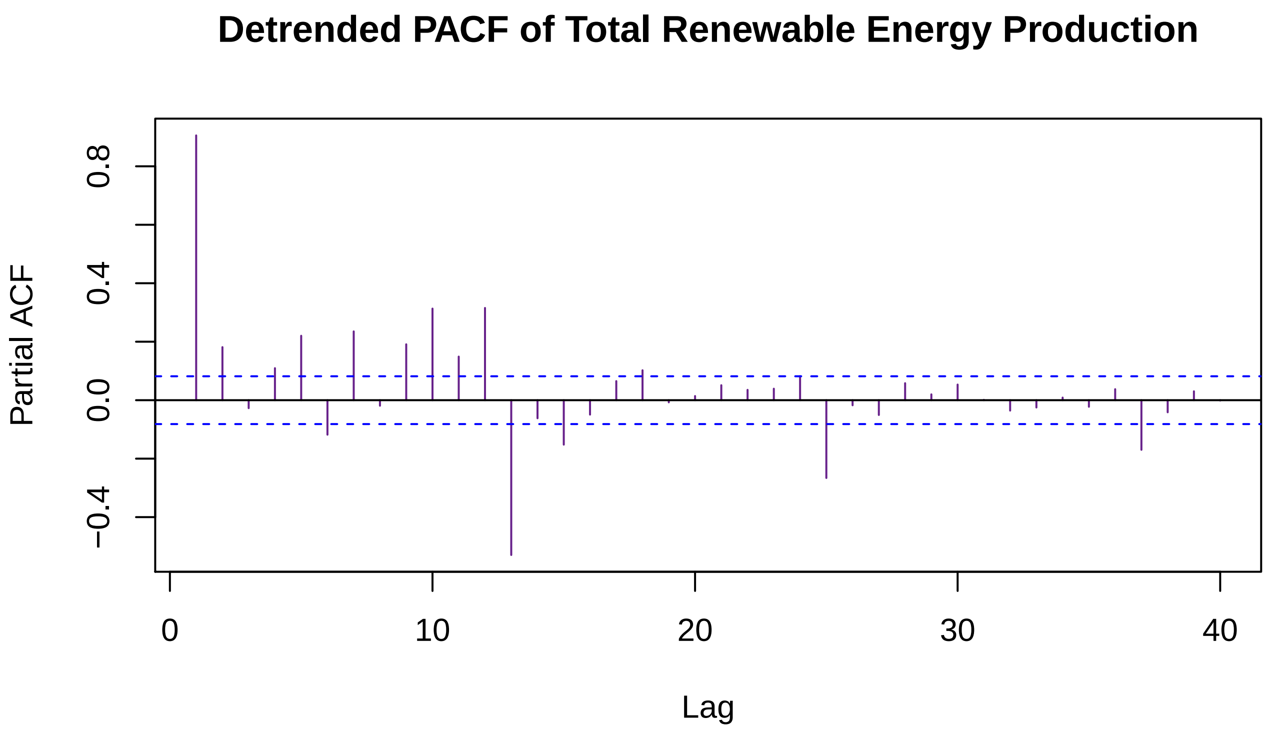


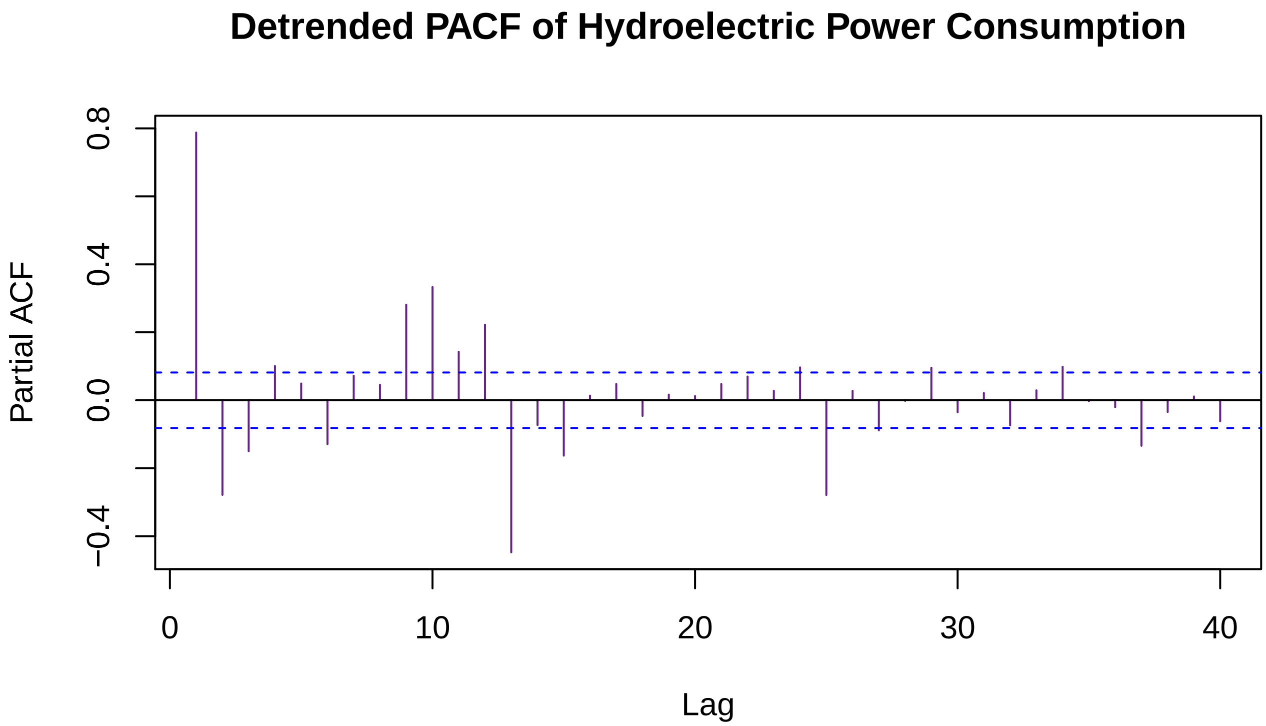



IN ACF, there was a minor change in the correlation of the data points with each other. The difference in correlation between the first point and the next increased.


Detrended PACF of Total Biomass Energy Production
Partial ACF
0.8
0.4
0.0
−0.4
0
10
20
30
40
Lag


Detrended PACF of Total Renewable Energy Production
Partial ACF
0.8
0.4
0.0
−0.4
0
10
20
30
40
Lag




The PACF plot, however, remained the same. There was no difference in the PACF of this detrended data and the data in Question 1.

---

## Seasonal Component

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

## Q6

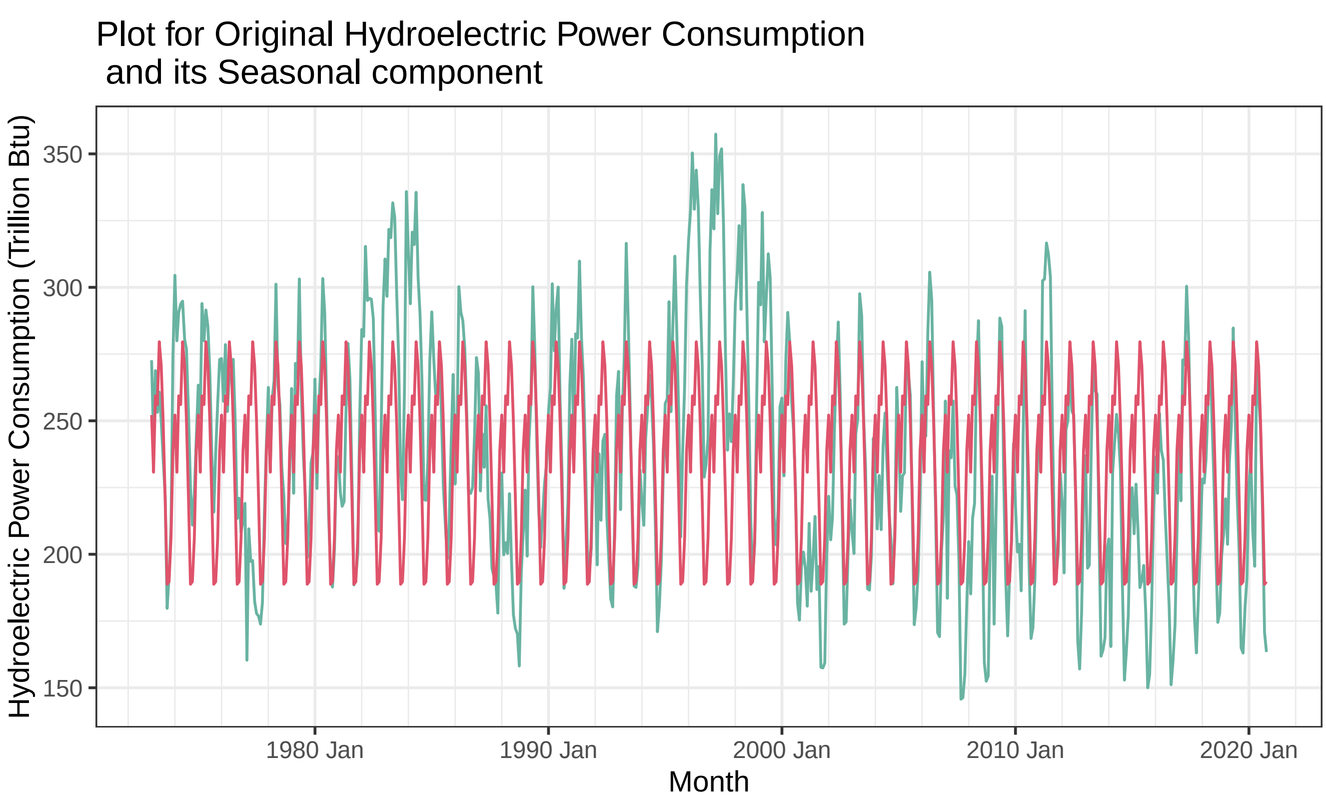
**Do the series seem to have a seasonal trend? Which serie/series? Use function *lm()* to fit a seasonal means model to this/these time series. Ask R to print the summary of the regression. Interpret the regression output. Save the regression coefficients for further analysis.**

Yes, Hydroelectric power consumption is a serie with seasonal trend.

Below is the summary of linear regression of this time serie. It will be seen from the below summary table that there are some coefficients which have a negative value, whereas, the others have a positive value. It means that the hydroelectric power consumption on certain months decreases, whereas, in certain months, it increases. From August to November, the coefficient is negative. It means that in these months, the consumption decreases. Another thing to note is the P-value of the coefficients. The p-values need to be less than 0.05. At the bottom, we can also see that the adjusted R-squared value is 0.4234. It means that 42.34% of our data is being explained by our regression model. I am unsure of whether this can be considered good or bad in our case, but we have to make do with the kind of data we have.

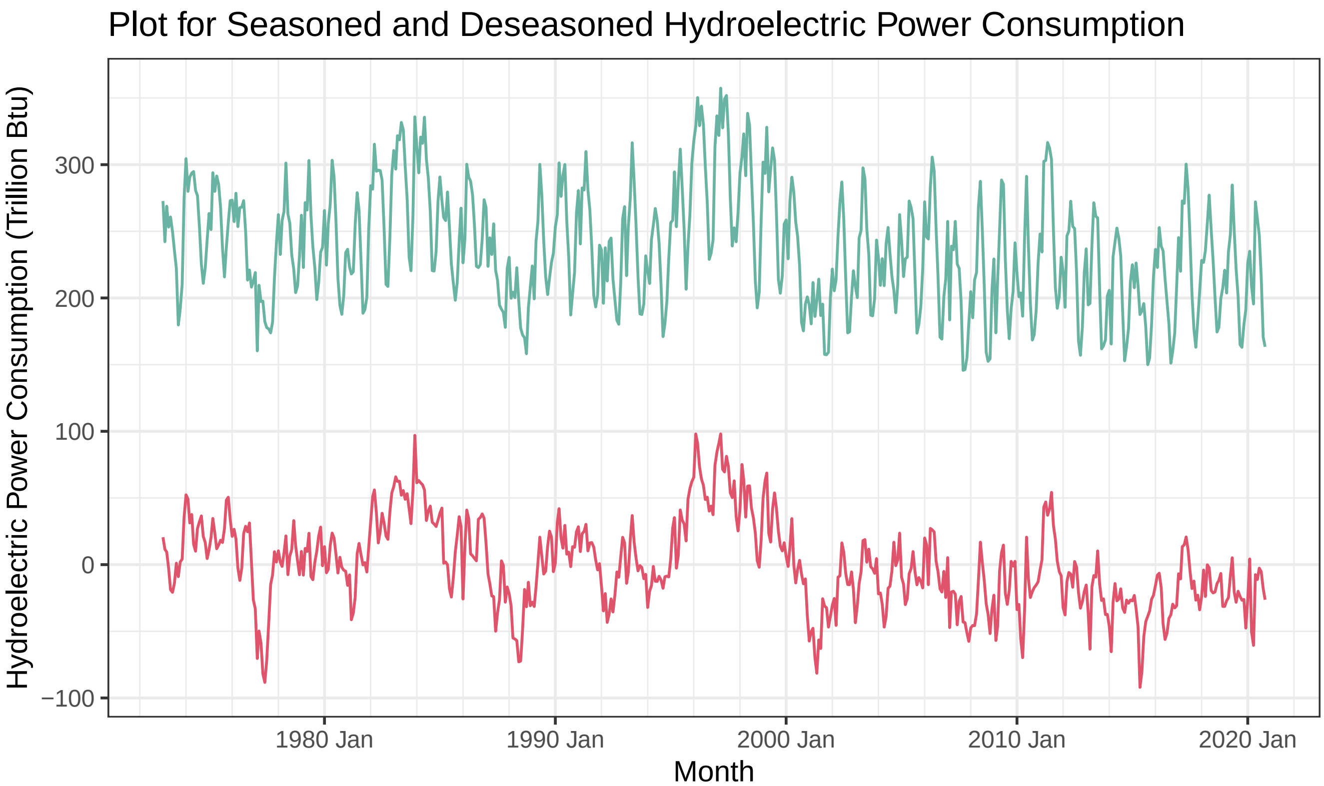
```
Call:
lm(formula = eia$`Hydroelectric Power Consumption (Trillion Btu)` ~
    dummies)

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
```

Below is the plot for original Hydroelectric Power Consumption and its seasonal component.



## 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?**



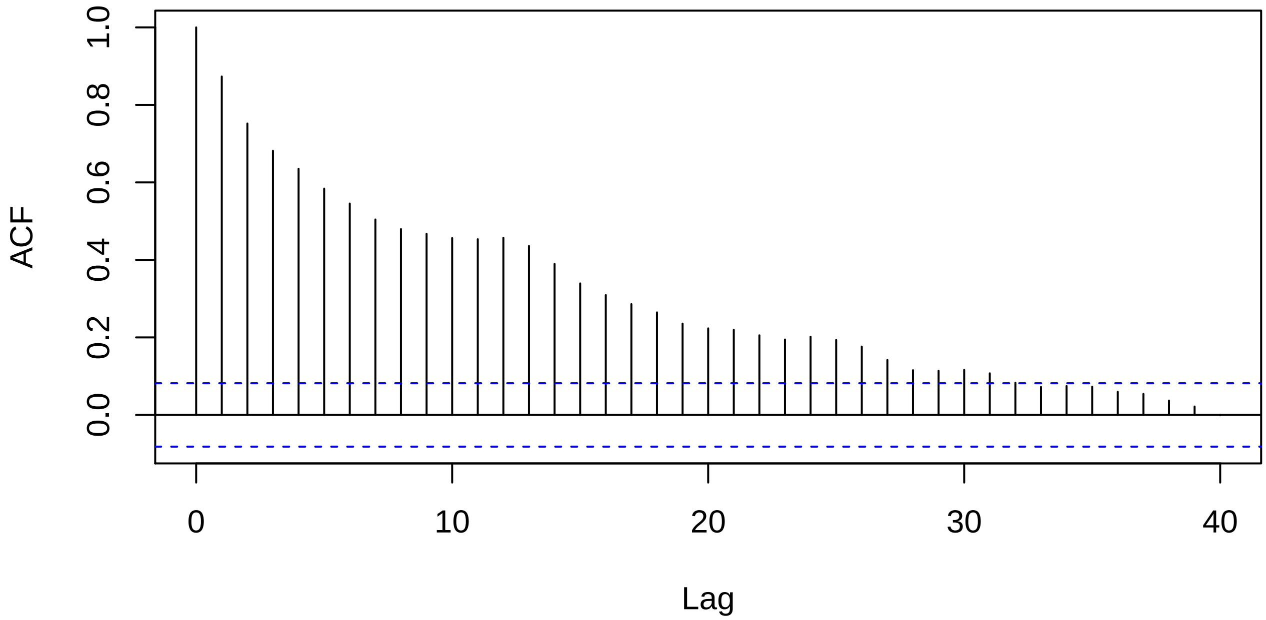

It can be seen from the above plot that the deseasoned graph (purple graph) has shifted down. The points are fluctuaitng around zero. The original plot (blue plot) is plotted at its normal points. It is, therefore, clear that the data is deseasoned. However, if we look at the way points are plotted, we will see no difference. When a point in the original plot goes up, the point in the deseasoned graph also goes up. The only way we can really tell that the purple plot is the deseasoned one is by looking at the y-axis. As mentioned above, the deseasoned graph is fluctuating around zero.

## Q8

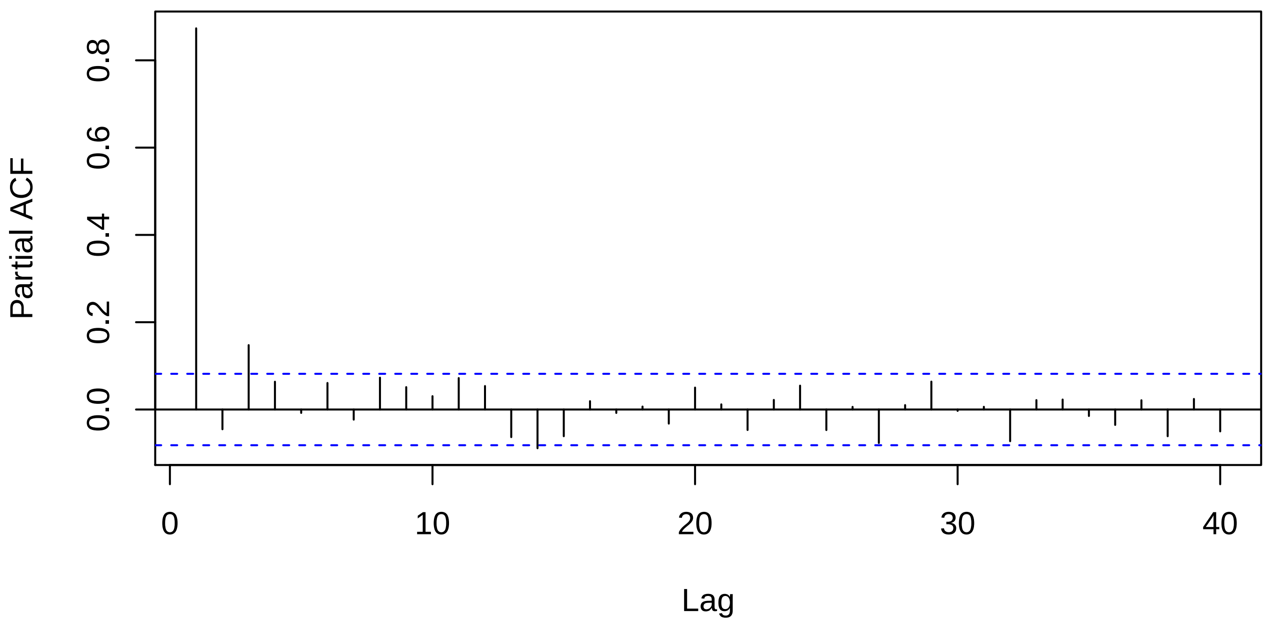
**Plot ACF and PACF for the deseason series and compare with the plots from Q1. Did the plots change? How?**

Yes, the plots did change. If we look at the original ACF plot, we can clearly identify the seasonality. This is because ACF plots help us in identifying whether there is any seasonality or not. After deseasonality, that seasonal fluctuation is completely gone from the ACF plot. Moreover, in the deseasoned PACF plot, the lines are all between the dotted blue threshold lines. In the original plot, however, this was not the case.

**Deseasoned ACF of Hydroelectric Power Consumption**




**Deseasoned PACF of Hydroelectric Power Consumption**



## APPENDIX

```r
knitr::opts_chunk$set(echo = F, eval = T, comment = NA, message = F,warning = F,
                      fig.width = 7, fig.height = 4.2)
#Load/install required package here
library(forecast)
library(tseries)
library(Kendall)
library(dplyr)
library(readxl)
library(ggplot2)

setwd("/Users/yashdoshi/Desktop/Duke/Courses/Spring 2021/Time Series Analysis/Labs/Lab Work/ENV790_30_TSA_S202
library(readxl)
eia = read_excel("Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx")
eia$Month = as.Date(eia$Month, format = "%m/%y")
eia[2:3] = NULL
eia[5:12] = NULL
eia = as.data.frame(eia)
#Converting to TS object
eiats = ts(data = eia[,2:4], start = 1973, frequency = 12)

btu = ncol(eia)

#Plotting Graph, ACF, and PACF
for(i in 2:btu){
  par(mfrow = c(1,3))
  plot(eia[,i], col = "#69b3a2", ylab = "Trillion Btu", type = "l")
  acf(eia[,i], lag.max = 40)
  pacf(eia[,i])
}

#Plot
p = ggplot(data = eia, aes(x = Month,
                           y = `Total Biomass Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Total Biomass Energy Production") +
  theme_bw() +
  ggtitle("Plot of Total Biomass Energy Production (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_smooth(method = "lm", color = 2)
p

t = c(1:nrow(eia))

#Linear Model
lm1 = lm(eia$`Total Biomass Energy Production (Trillion Btu)` ~ t)
summary(lm1)

#Plot
q = ggplot(data = eia, aes(x = Month,
                           y = `Total Renewable Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Total Renewable Energy Production (Trillion Btu)") +
  theme_bw() +
  ggtitle("Plor of Total Renewable Energy Production (Trillion Btu)") +
  geom_smooth(method = "lm", col = 2) +
```

```r
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b")
q

#Linear Model
lm2 = lm(eia$`Total Renewable Energy Production (Trillion Btu)` ~ t)
summary(lm2)
#Plot
he = ggplot(data = eia, aes(x = Month, y = `Hydroelectric Power Consumption (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Hydroelectric Power Consumption (Trillion Btu)") +
  theme_bw() +
  geom_smooth(method = "lm", col = "2") +
  ggtitle("Plot for Hydroelectric Power Consumption (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b")
he

#Linear Model
lm3 = lm(eia$`Hydroelectric Power Consumption (Trillion Btu)` ~ t)
summary(lm3)

#Plot for Total Biomass Energy Production
b0 = as.numeric(lm1$coefficients[1])
b1 = as.numeric(lm1$coefficients[2])


detrend_data = eia$`Total Biomass Energy Production (Trillion Btu)` - (b0 + b1*t)

pe = ggplot(data = eia, aes(x = Month,
                            y = `Total Biomass Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Total Biomass Energy Production") +
  theme_bw() +
  ggtitle("Plot of Total Biomass Energy Production (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_smooth(method = "lm", color = 2) +
  geom_line(mapping = aes(y = detrend_data), col = "darkorchid4") +
  geom_smooth(aes(y = detrend_data), col = "2", method = "lm")
pe

#Plot for total renewable energy consumption
b00 = as.numeric(lm2$coefficients[1])
b11 = as.numeric(lm2$coefficients[2])

detrend_data2 = eia$`Total Renewable Energy Production (Trillion Btu)` - (b00 + b11*t)

qe = ggplot(data = eia, aes(x = Month,
                            y = `Total Renewable Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Total Renewable Energy Production (Trillion Btu)") +
  theme_bw() +
  ggtitle("Plor of Total Renewable Energy Production (Trillion Btu)") +
  geom_smooth(method = "lm", col = 2) +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_line(mapping = aes(y = detrend_data2), col = "darkorchid4") +
  geom_smooth(aes(y = detrend_data2), col = "4", method = "lm")
qe
```

```r
#Plot for hydroelectric power consumption
b000 = as.numeric(lm3$coefficients[1])
b111 = as.numeric(lm3$coefficients[2])

detrend_data3 = eia$`Hydroelectric Power Consumption (Trillion Btu)` - (b000 + b111*t)

hee = ggplot(data = eia, aes(x = Month, y = `Hydroelectric Power Consumption (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Hydroelectric Power Consumption (Trillion Btu)") +
  theme_bw() +
  geom_smooth(method = "lm", col = "2") +
  ggtitle("Plot for Hydroelectric Power Consumption (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_line(mapping = aes(y = detrend_data3), col = "darkorchid4") +
  geom_smooth(aes(y = detrend_data3), col = "4", method = "lm")
hee

#Detrended ACF of Total Biomass Energy Production
acf1 = acf(detrend_data, lag.max = 40, plot = FALSE)
plot(acf1, main = "Detrended ACF of Total Biomass Energy Production", col = "2")

#Detrended ACF of Total Renewable Energy Production
acf2= acf(detrend_data2, lag.max = 40, plot = FALSE)
plot(acf2, main = "Detrended ACF of Total Renewable Energy Production", col = "2")

#Detrended ACF of Hydroelectric Power Consumption
acf3 = acf(detrend_data3, lag.max = 40, plot = FALSE)
plot(acf3, main = "Detrended ACF of Hydroelectric Power Consumption", col = "2")

#Detrended PACF of Total Biomass Energy Production
pacf1 = pacf(detrend_data, lag.max = 40, plot = FALSE)
plot(pacf1, main = "Detrended PACF of Total Biomass Energy Production",
     col = "darkorchid4")

#Detrended PACF of Total Renewable Energy Production
pacf2 = pacf(detrend_data2, lag.max = 40, plot = FALSE)
plot(pacf2, main = "Detrended PACF of Total Renewable Energy Production",
     col = "darkorchid4")

#Detrended PACF of Hydroelectric Power Consumption
pacf3 = pacf(detrend_data3, lag.max = 40, plot = FALSE)
plot(pacf3, main = "Detrended PACF of Hydroelectric Power Consumption",
     col = "darkorchid4")

dummies = seasonaldummy(eiats[,3])
seaslm = lm(eia$`Hydroelectric Power Consumption (Trillion Btu)` ~ dummies)
summary(seaslm)

beta0 = seaslm$coefficients[1]
beta1 = seaslm$coefficients[2:12]

nobs = nrow(eia)

seas_comp = array(0,nobs)

for(i in 1:nobs){
  seas_comp[i] = (beta0 + beta1 %*% dummies[i,])
}
```

```r
#Plot
ggplot(eia, aes(x = Month, y = `Hydroelectric Power Consumption (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  theme_bw() +
  ggtitle("Plot for Original Hydroelectric Power Consumption  \n and its Seasonal component") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_line(aes(y = seas_comp), col = "2")

deseason = eia[,4] - seas_comp

ggplot(eia, aes(x = Month, y = `Hydroelectric Power Consumption (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  theme_bw() +
  ggtitle("Plot for Seasoned and Deseasoned Hydroelectric Power Consumption") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_line(aes(y = deseason), col = "2")

#Deseasoned ACF of Hydroelectric Power Consumption
acf4 = acf(deseason,lag.max = 40, plot = FALSE)
plot(acf4, main = "Deseasoned ACF of Hydroelectric Power Consumption")
#Deseasoned PACF of Hydroelectric Power Consumption
pacf4 = pacf(deseason, lag.max = 40, plot = FALSE)
plot(pacf4, main = "Deseasoned PACF of Hydroelectric Power Consumption")
```