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

Assignment 2 - Due date 02/05/21

Xueying Feng

Submission Instructions

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 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).

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_A02_Sp21.Rmd"). Submit this pdf using Sakai.

R packages

R packages needed for this assignment:"forecast","tseries", and "dplyr". 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
getwd()
```

```
## [1] "/Users/ethel/Desktop/ENV 790 Time Series/ENV790_TSA_S2021/Assignments"
```

```
library(lubridate)
```

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

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

```
library(ggplot2)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("forecast")
library(forecast)
```

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

```
#install.packages("tseries")
library(tseries)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.x on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the January 2021 Monthly Energy Review. The spreadsheet is ready to be used. Use the command *read.table*() to import the data in R or *panda.read_excel*() in Python (note that you will need to import pandas package). }

```
#Importing data set

#install.packages("readxl")
library("readxl")
MonthlyData <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx

#AnnualData <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xls
#AnnualData
```

Question 1

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. Use the command head() to verify your data.

```
#specific rows and columns just trim the table
MonthlyData_subset<- MonthlyData[2:575, c(1, 4, 5, 6)]

#Checking data
head(MonthlyData_subset)
```

```
## # A tibble: 6 x 4
##   Month               `Total Biomass Ene~ `Total Renewable E~ `Hydroelectric Po~
##   <dttm>              <chr>               <chr>               <chr>
## 1 1973-01-01 00:00:00 129.787             403.981             272.703
## 2 1973-02-01 00:00:00 117.338             360.9               242.199
## 3 1973-03-01 00:00:00 129.938             400.161             268.81
## 4 1973-04-01 00:00:00 125.636             380.47              253.185
## 5 1973-05-01 00:00:00 129.834             392.141             260.77
## 6 1973-06-01 00:00:00 125.611             377.232             249.859
```

```
str(MonthlyData_subset)
```

```
## tibble [574 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Month                            : POSIXct[1:574], format: "1973-01-01" "1973-02-01" ...
##  $ Total Biomass Energy Production  : chr [1:574] "129.787" "117.338" "129.938" "125.636" ...
##  $ Total Renewable Energy Production: chr [1:574] "403.981" "360.9" "400.161" "380.47" ...
##  $ Hydroelectric Power Consumption  : chr [1:574] "272.703" "242.199" "268.81" "253.185" ...
```

Question 2

Transform your data frame in a time series object and specify the starting point and frequency of the time series using the function ts().

```
# change character format to numeric format
MonthlyData_subset[,2:4] <- sapply(MonthlyData_subset[,2:4],as.numeric)
str(MonthlyData_subset)
```

```
## tibble [574 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Month                            : POSIXct[1:574], format: "1973-01-01" "1973-02-01" ...
##  $ Total Biomass Energy Production  : num [1:574] 130 117 130 126 130 ...
##  $ Total Renewable Energy Production: num [1:574] 404 361 400 380 392 ...
##  $ Hydroelectric Power Consumption  : num [1:574] 273 242 269 253 261 ...
```

```
#add column names
colnames(MonthlyData_subset)=c("Month","Biomass","Renewable","Hydroelectric")


# from Jan 1973 to Oct 2020 as a time series object
MonthlyData_subset_ts <- ts(MonthlyData_subset[2:4], frequency = 12, start = c(1973, 1, 1), end = c(2020
#MonthlyData_subset_ts
```

Question 3

Compute mean and standard deviation for these three series.

```
Biomass_mean <- mean(MonthlyData_subset_ts[,"Biomass"])
Biomass_sd <- sd(MonthlyData_subset_ts[,"Biomass"])
Biomass_mean
```

```
## [1] 270.6961
```

```
str(Biomass_mean)
```

```
##  num 271
```

```
Biomass_sd
```

```
## [1] 87.36311
```

```
Renewable_mean <- mean(MonthlyData_subset_ts[,"Renewable"])
Renewable_sd <- sd(MonthlyData_subset_ts[,"Renewable"])
Renewable_mean
```

```
## [1] 572.7321
```

```
Renewable_sd
```

```
## [1] 168.4588
```

```
Hydroelectric_mean <- mean(MonthlyData_subset_ts[,"Hydroelectric"])
Hydroelectric_sd <- sd(MonthlyData_subset_ts[,"Hydroelectric"])
Hydroelectric_mean
```

```
## [1] 236.9515
```

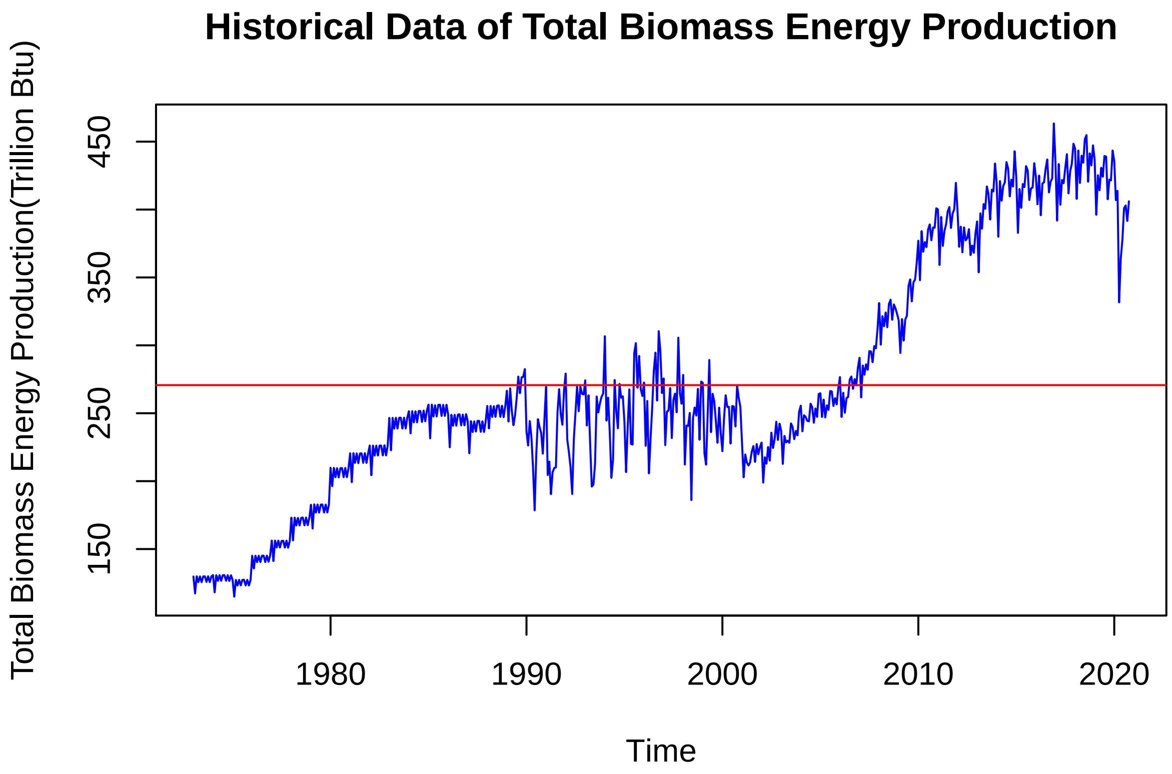
```
Hydroelectric_sd
```

```
## [1] 43.90392
```

Question 4

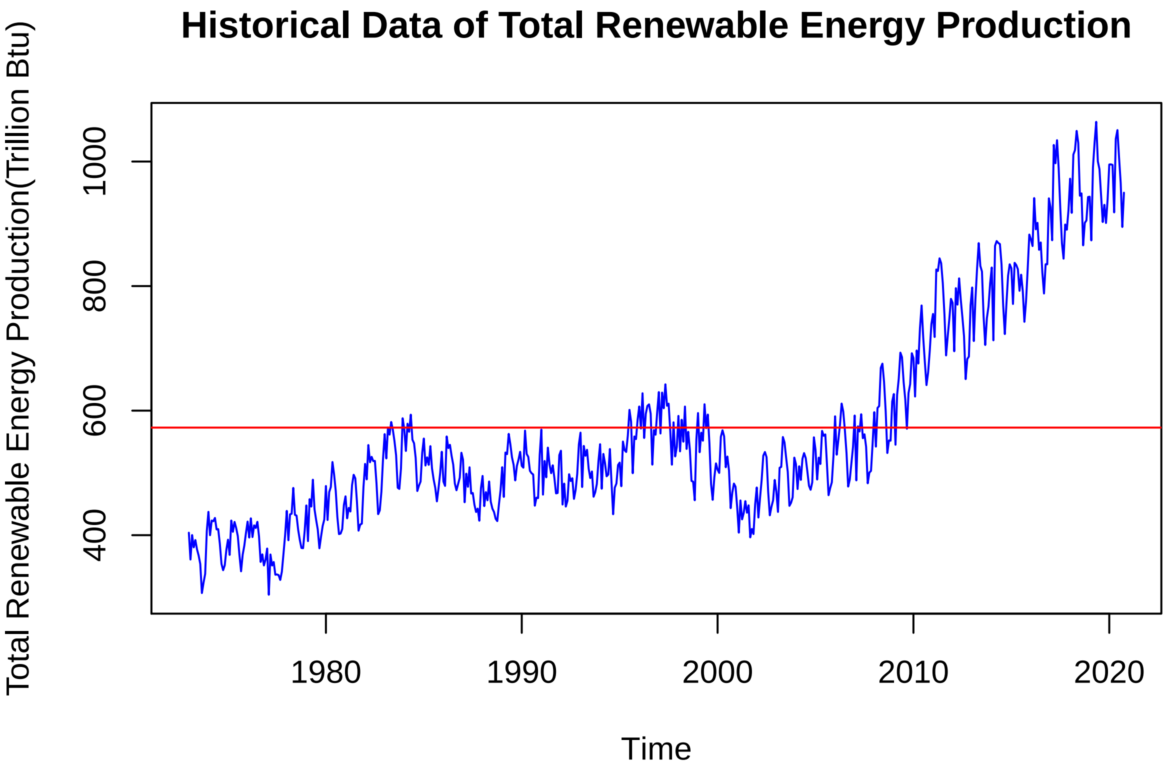
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
#library(ggplot2)
plot(MonthlyData_subset_ts[,"Biomass"],col="blue",
     ylab = "Total Biomass Energy Production(Trillion Btu)",
     main = "Historical Data of Total Biomass Energy Production"
     )
#add a line with the mean
abline(h=Biomass_mean,col="red")
```



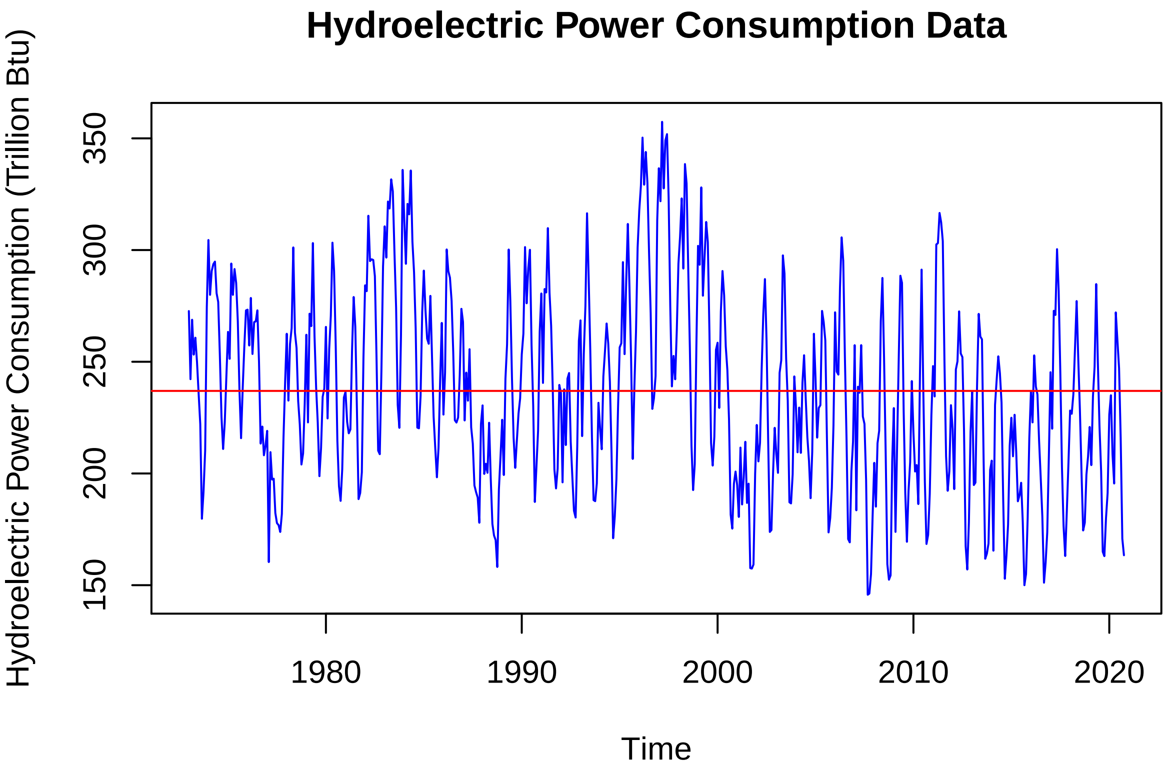
>This time series plot shows a drastic shift in 1975 and 2000 and shows a clear upward trend. Between 1990 and 2000, it shows random variation.

```
plot(MonthlyData_subset_ts[,"Renewable"],col="blue",
     ylab = "Total Renewable Energy Production(Trillion Btu)",
     main = "Historical Data of Total Renewable Energy Production"
     )
#add a line with the mean
abline(h=Renewable_mean,col="red")
```



>This plot shows increase in the data values seems to accelerate after 2000. Before the 2000, the cycles do not repeat at regular intervals and do not have the same shape.

```
plot(MonthlyData_subset_ts[,"Hydroelectric"],col="blue",
     ylab = "Hydroelectric Power Consumption (Trillion Btu)",
     main = "Hydroelectric Power Consumption Data"
     )
#add a line with the mean
abline(h=Hydroelectric_mean,col="red")
```

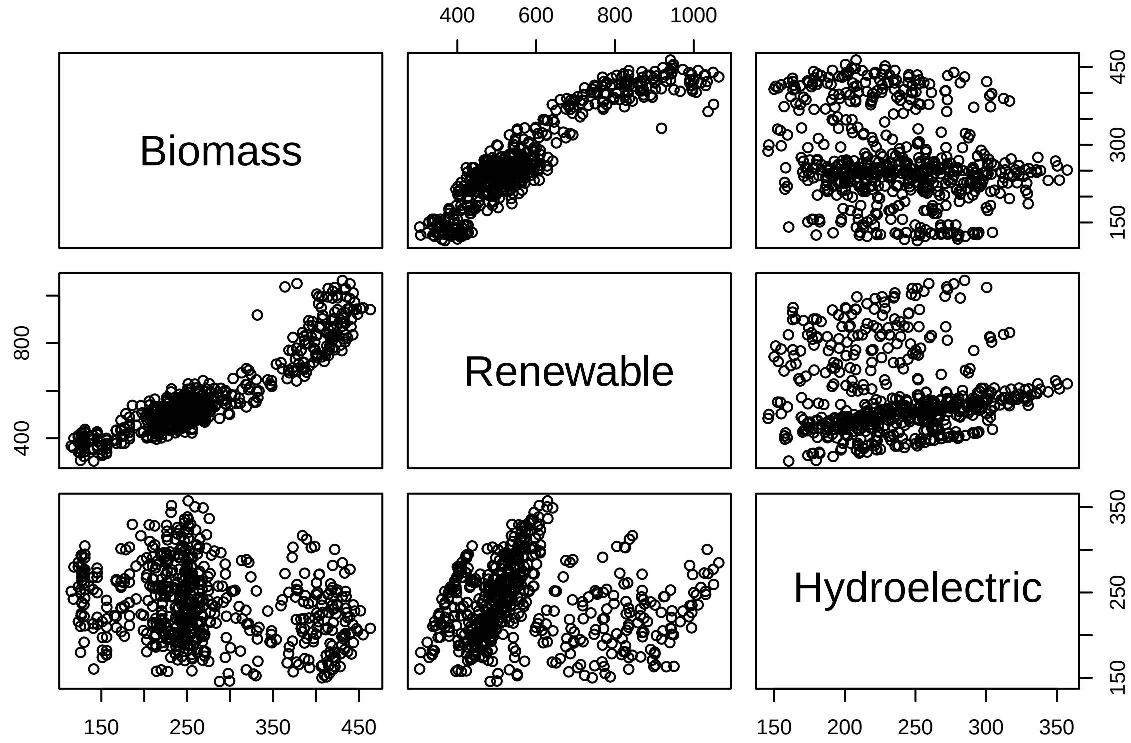


>These plot displays the pattern of Hydroelectric Power Consumption increasing and decreasing from 1980 to 2020, which seems changing seasonal.

Question 5

Compute the correlation between these three series. Are they significantly correlated? Explain your answer.

```
pairs(MonthlyData_subset[, c(2,3,4)])
```



```
cor(MonthlyData_subset_ts)
```

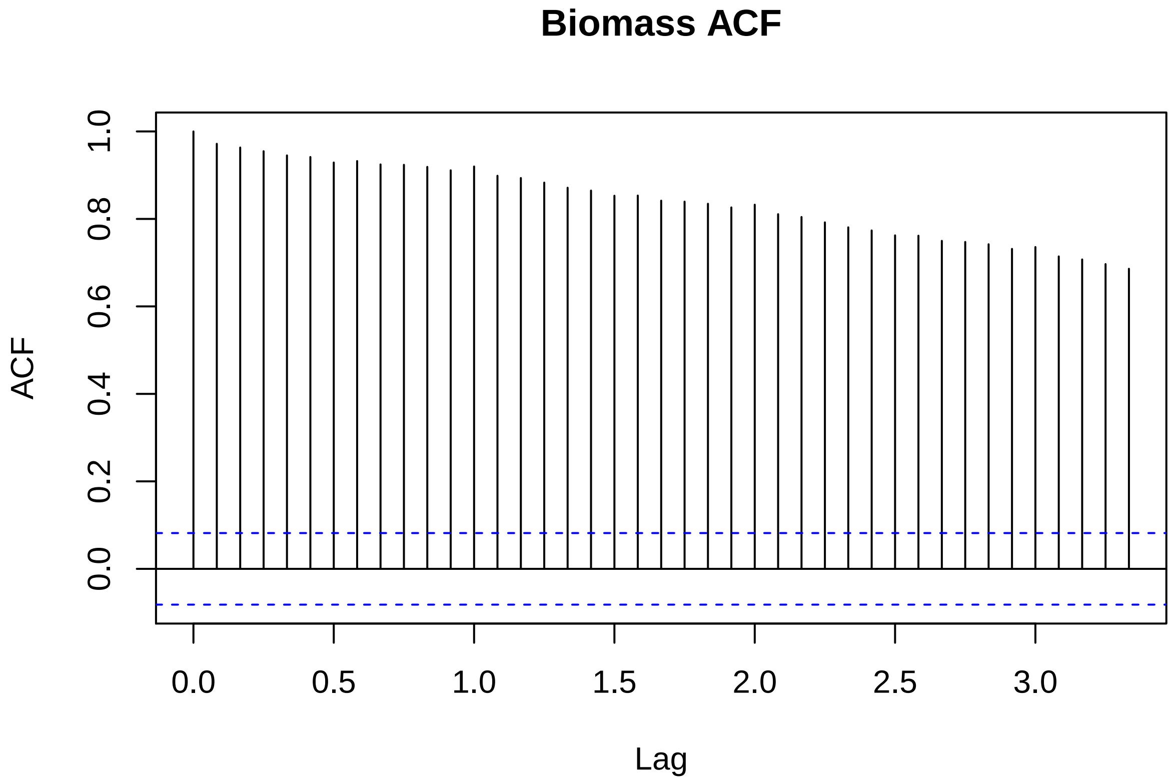
```
##                  Biomass    Renewable Hydroelectric
## Biomass        1.0000000  0.923460855  -0.255567465
## Renewable      0.9234609  1.000000000  -0.002756852
## Hydroelectric -0.2555675 -0.002756852   1.000000000
```

Due to the results of correlation and plots, Total Biomass Energy Production and Total Renewable Energy Production are significatly correlated, whereas the other two are not.

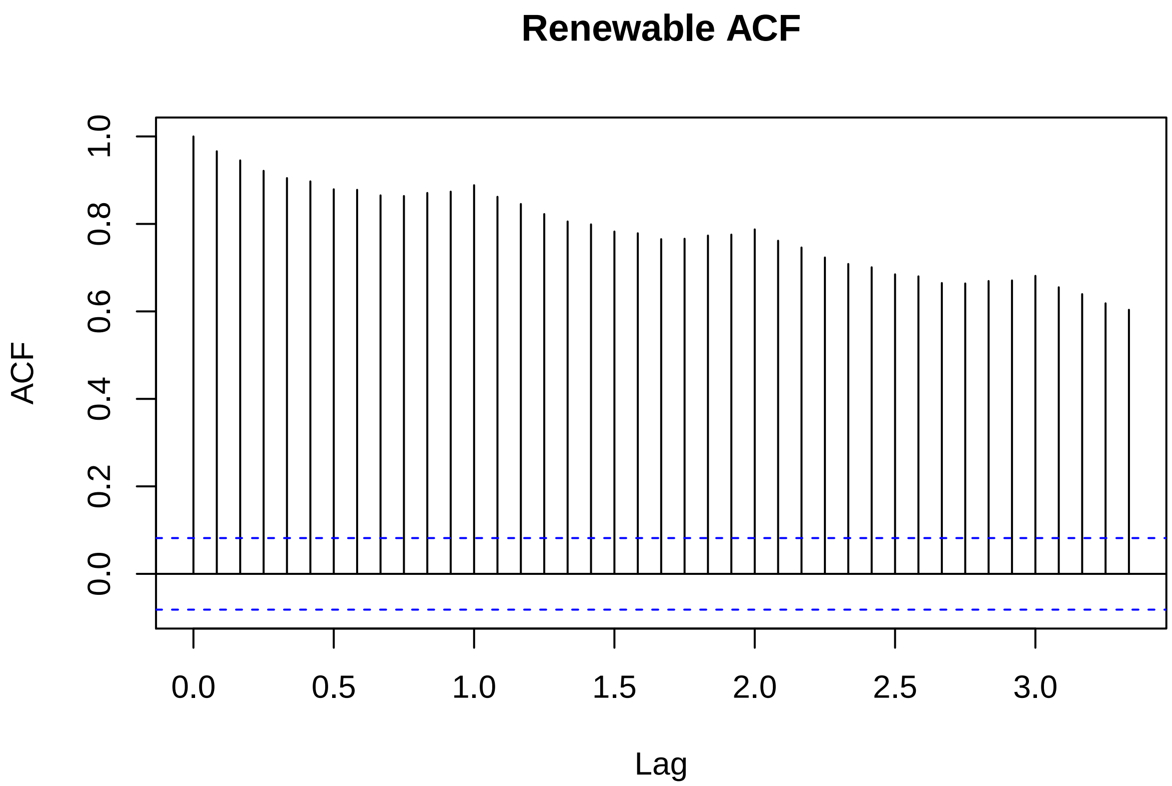
Question 6

Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

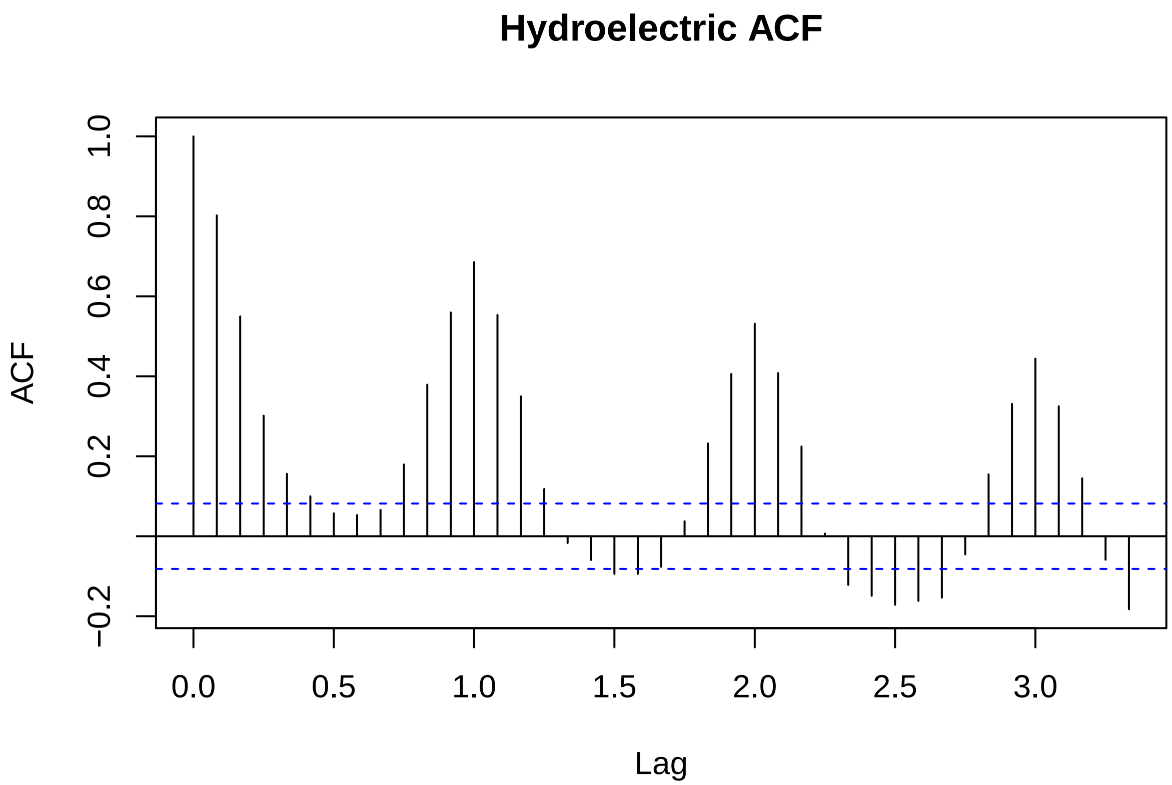
```
Biomass_acf = acf(MonthlyData_subset_ts[,1], lag.max=40, type = "correlation", plot=TRUE, main="Biomass
```



```
Renewable_acf = acf(MonthlyData_subset_ts[,2], lag.max=40, type = "correlation", plot=TRUE, main="Renewa
```



```
Hydroelectric_acf = acf(MonthlyData_subset_ts[,3], lag.max=40, type = "correlation", plot=TRUE, main="Hy
```

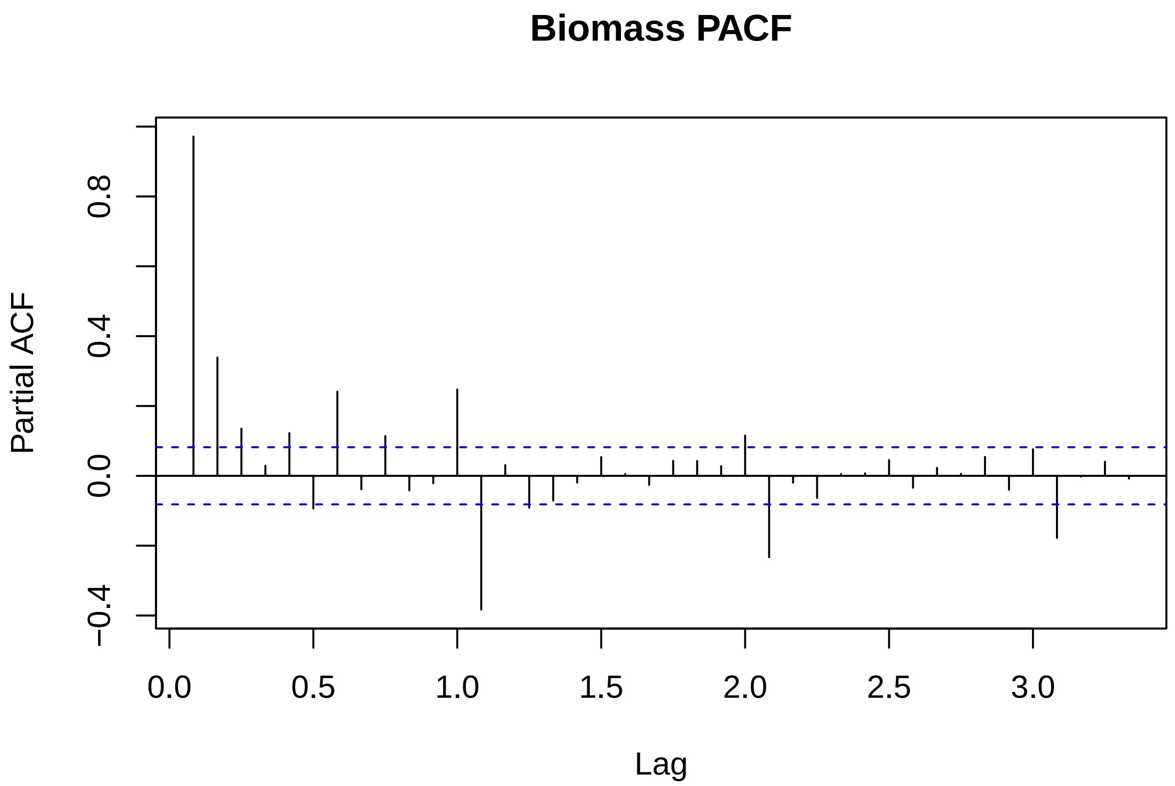


> The Biomass and Renewable plots have very similar trend, but Hydroelectric one is kind of a periodic trend. They do not have same behavior.

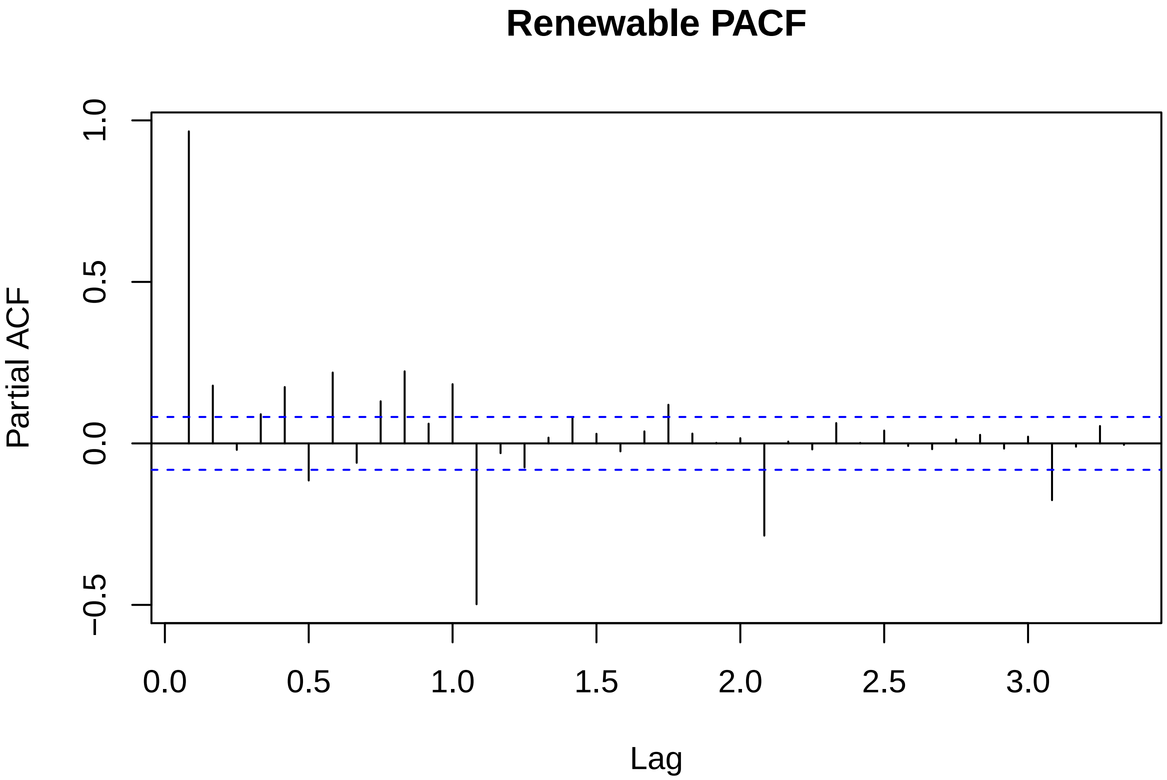
Question 7

Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

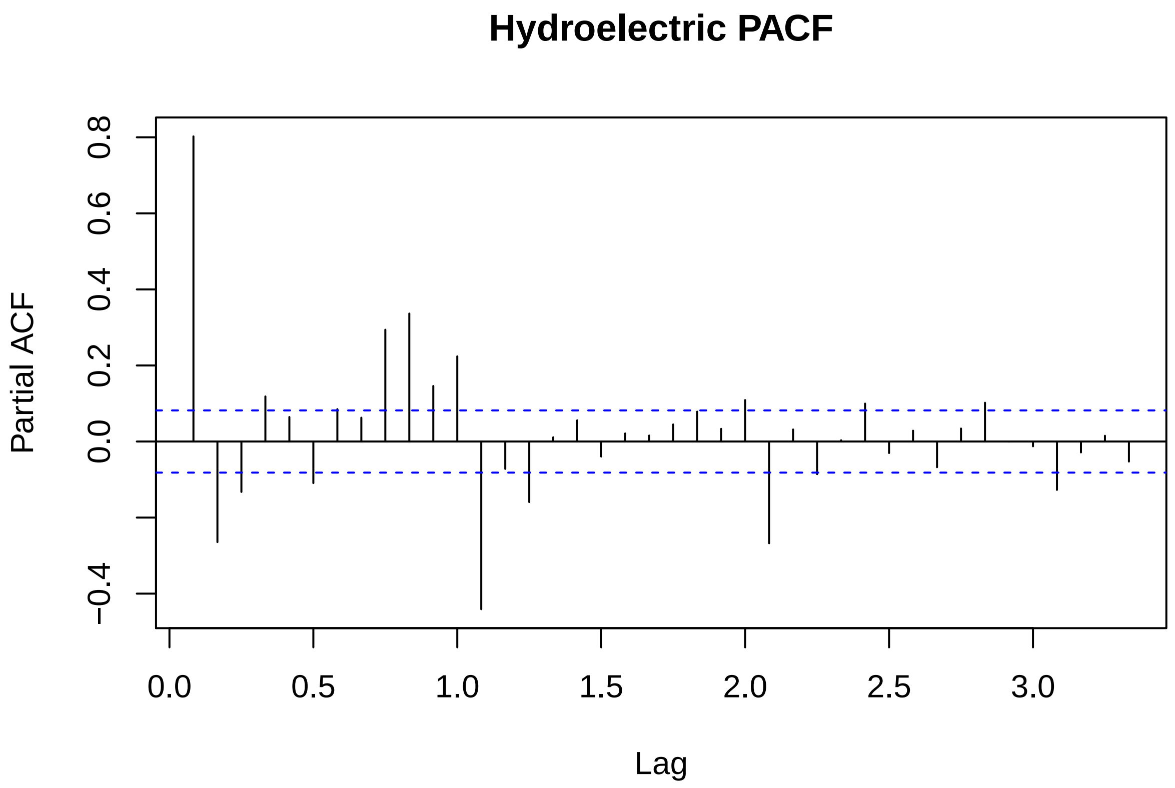
```
Biomass_pacf = pacf(MonthlyData_subset_ts[,1], lag.max=40,plot=TRUE, main="Biomass PACF")
```



```
Renewable_pacf = pacf(MonthlyData_subset_ts[,2], lag.max=40, plot=TRUE, main="Renewable PACF")
```



```
Hydroelectric_pacf = pacf(MonthlyData_subset_ts[,3], lag.max=40, plot=TRUE, main="Hydroelectric PACF")
```



> The three plots in Q7 is much more similar, but Hydroelectric plot is total different than Biomass and Renewable plots in Q6.