ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 2 - Due date 02/25/24

Samantha Pace

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 rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp24.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).

When you have completed the assignment, **Knit** the text and code into a single PDF file. 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

#install.packages("forecast")
#install.packages("tseries")
#install.packages("dplyr")
#install.packages(ggplot2)
#install.packages("tidyverse")


library(forecast)
library(dplyr)
library(tseries)
library(ggplot2)
library(tidyverse)
library(readxl)

# working directory
getwd()
```

```
## [1] "C:/Users/saman/OneDrive/Desktop/Duke Spring 24/GITHUB/TSA_Sp24"
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the December 2023 Monthly Energy Review. The spreadsheet is ready to be used. You will also find a *.csv* version of the data "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.csv". You may use the function *read.table*() to import the *.csv* data in R. Or refer to the file "M2_ImportingData_CSV_XLSX.Rmd" in our Lessons folder for functions that are better suited for importing the *.xlsx*.

```
#Importing data using read.xlsx
energy_data <- read_excel(
  path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                          skip = 12,
                          sheet = "Monthly Data", col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
# Getting column names from row 11
read_col_names <-
  read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
             skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
energy_data <-
  read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
             skip = 12, sheet="Monthly Data",col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
# inputting correct column names
colnames(energy_data) <- read_col_names
head(energy_data)
```

```
## # A tibble: 6 x 14
##   Month               `Wood Energy Production` `Biofuels Production`
##   <dttm>                                 <dbl> <chr>
## 1 1973-01-01 00:00:00                     130. Not Available
## 2 1973-02-01 00:00:00                     117. Not Available
## 3 1973-03-01 00:00:00                     130. Not Available
## 4 1973-04-01 00:00:00                     125. Not Available
## 5 1973-05-01 00:00:00                     130. Not Available
## 6 1973-06-01 00:00:00                     125. Not Available
## # i 11 more variables: `Total Biomass Energy Production` <dbl>,
## #   `Total Renewable Energy Production` <dbl>,
## #   `Hydroelectric Power Consumption` <dbl>,
## #   `Geothermal Energy Consumption` <dbl>, `Solar Energy Consumption` <chr>,
## #   `Wind Energy Consumption` <chr>, `Wood Energy Consumption` <dbl>,
## #   `Waste Energy Consumption` <dbl>, `Biofuels Consumption` <chr>,
## #   `Total Biomass Energy Consumption` <dbl>, ...
```

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.

```
# selecting the 3 columns needed
energy_data_3 <- energy_data %>% select(1,4,5,6)

# checking with head() function
head(energy_data_3)
```

```
## # A tibble: 6 x 4
##   Month               `Total Biomass Energy Production` Total Renewable Energy~1
##   <dttm>                                          <dbl>                    <dbl>
## 1 1973-01-01 00:00:00                              130.                     220.
## 2 1973-02-01 00:00:00                              117.                     197.
## 3 1973-03-01 00:00:00                              130.                     219.
## 4 1973-04-01 00:00:00                              126.                     209.
## 5 1973-05-01 00:00:00                              130.                     216.
## 6 1973-06-01 00:00:00                              126.                     208.
## # i abbreviated name: 1: `Total Renewable Energy Production`
## # i 1 more variable: `Hydroelectric Power Consumption` <dbl>
```

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().

```
# transforming data frame in a time series object
ts_energy_data_3 <- ts(energy_data_3, start = c(1973,1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
# mean for ts of Total Biomass Energy Production column
biomass_mean <- mean(ts_energy_data_3[,2])
biomass_mean
```

```
## [1] 279.8046
```

```
# standard deviation for ts of Total Biomass Energy Production column
sd(ts_energy_data_3[,1])
```

```
## [1] 462705901
```

```
# mean for ts of Total Renewable Energy Production column
renewable_mean <- mean(ts_energy_data_3[,3])
renewable_mean
```

```
## [1] 395.7213
```

```
# standard deviation for ts of Total Renewable Energy Production column
sd(ts_energy_data_3[,2])
```

```
## [1] 92.66504
```

```
# mean for ts of Hydroelectric Power Consumption column
hydroelectric_mean <- mean(ts_energy_data_3[,4])
hydroelectric_mean
```

```
## [1] 79.73071
```

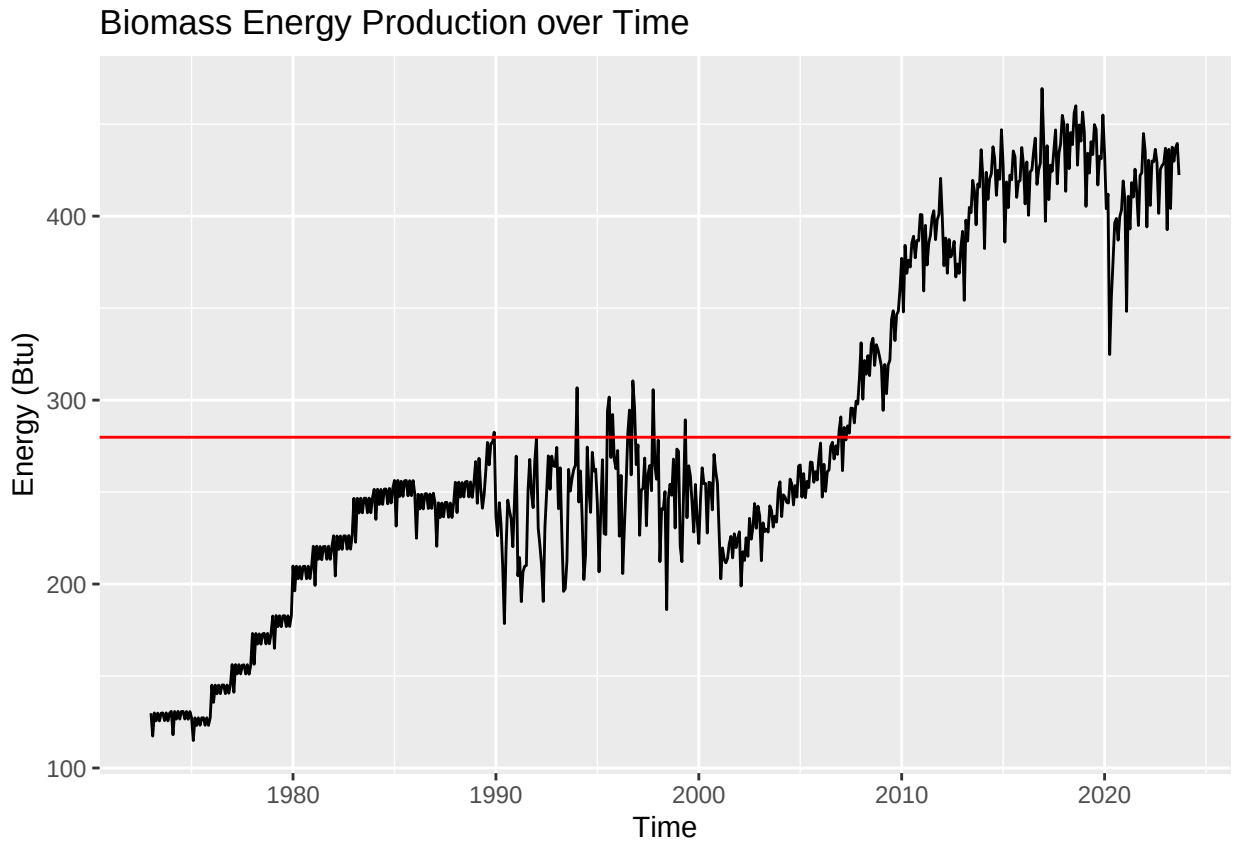
```
# standard deviation for ts of Hydroelectric Power Consumption column
sd(ts_energy_data_3[,3])
```

```
## [1] 137.7952
```

Question 4

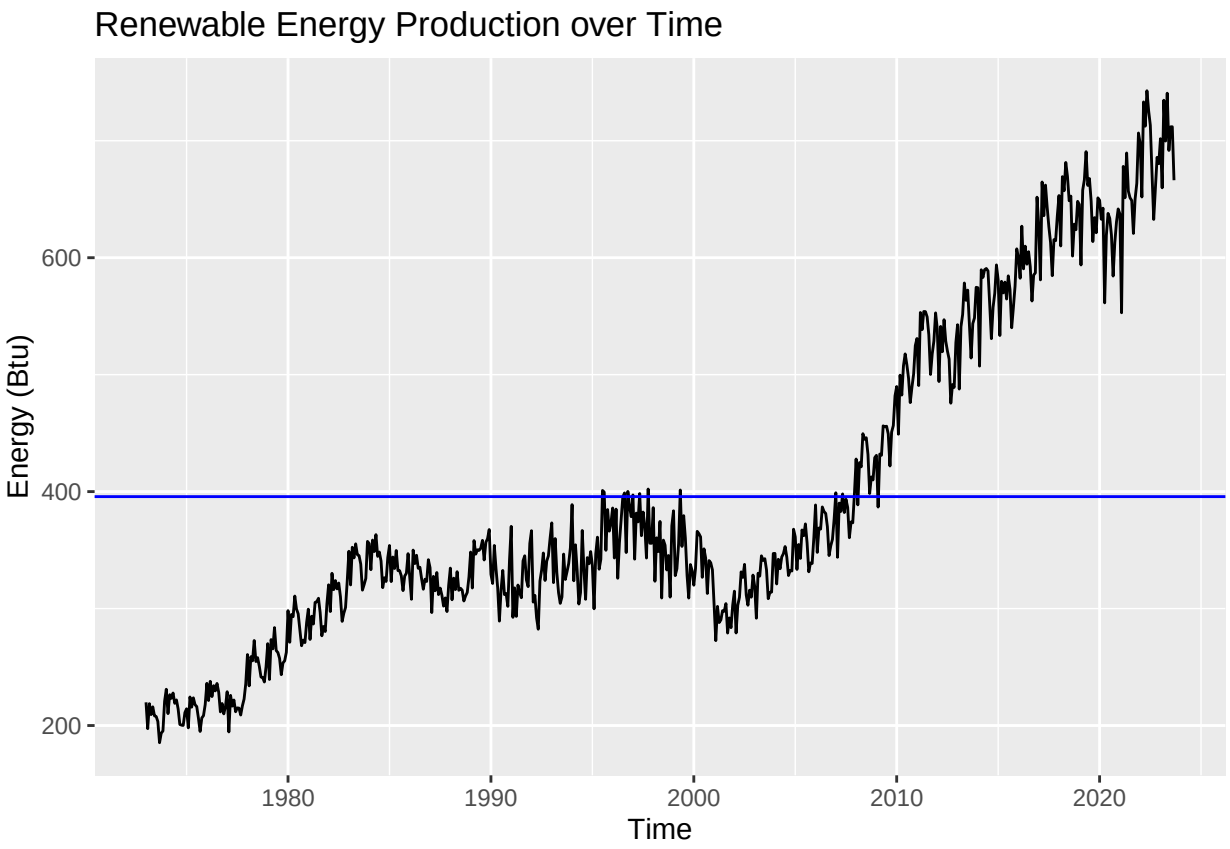
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
# Biomass Energy plot
autoplot(ts_energy_data_3[,2]) +
  ylab("Energy (Btu)") +
  xlab("Time") +
  labs(color="Energy Variable") +
  ggtitle("Biomass Energy Production over Time") +
  geom_hline(yintercept = biomass_mean, col = 'red')
```

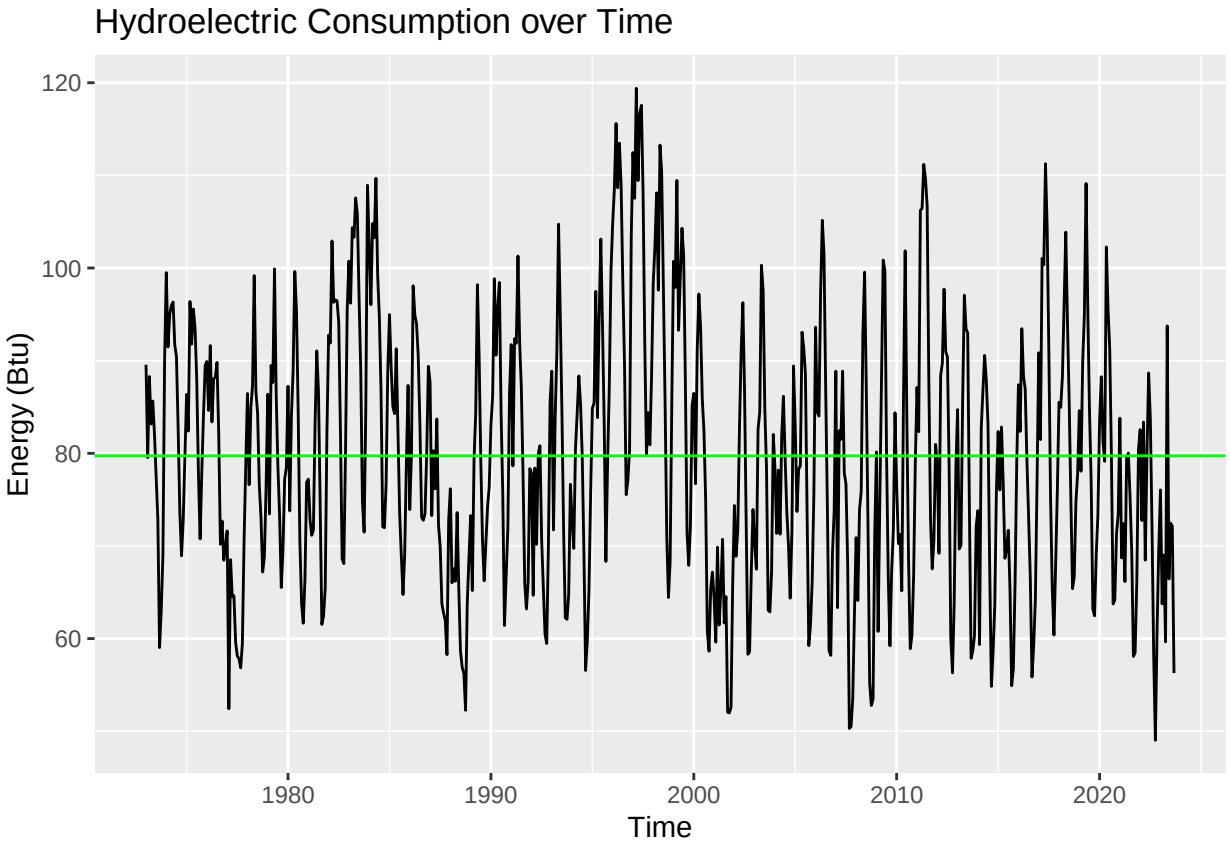



```
#Renewable Energy plot
autoplot(ts_energy_data_3[,3]) +
  ylab("Energy (Btu)") +
```

```
  xlab("Time") +
  labs(color="Energy Variable") +
  ggtitle("Renewable Energy Production over Time") +
  geom_hline(yintercept = renewable_mean, col = 'blue')
```




```
#Hydroelectric Consumption plot
autoplot(ts_energy_data_3[,4]) +
  ylab("Energy (Btu)") +
  xlab("Time") +
  labs(color="Energy Variable") +
  ggtitle("Hydroelectric Consumption over Time") +
  geom_hline(yintercept = hydroelectric_mean, col = 'green')
```



Answer: The Biomass plot shows an upward trend, and the seasonality is hard to decipher. The mean, in the red line, is right at 280. Similarly, the Renewable energy plot shows an upward trend. It shows a little more of a seasonality trend with regular up/down cycles. The mean of the renewable energy production for this time series is 396. Finally, the hydroelectric consumption plot doesn't exhibit much of trend, but may have seasonality as there is somewhat regularly space peaks and valleys in the data. For the duration of the time period, the data is centered on the mean.

Question 5

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

```
# correlation coefficients
cor(energy_data_3[,2:4])
```

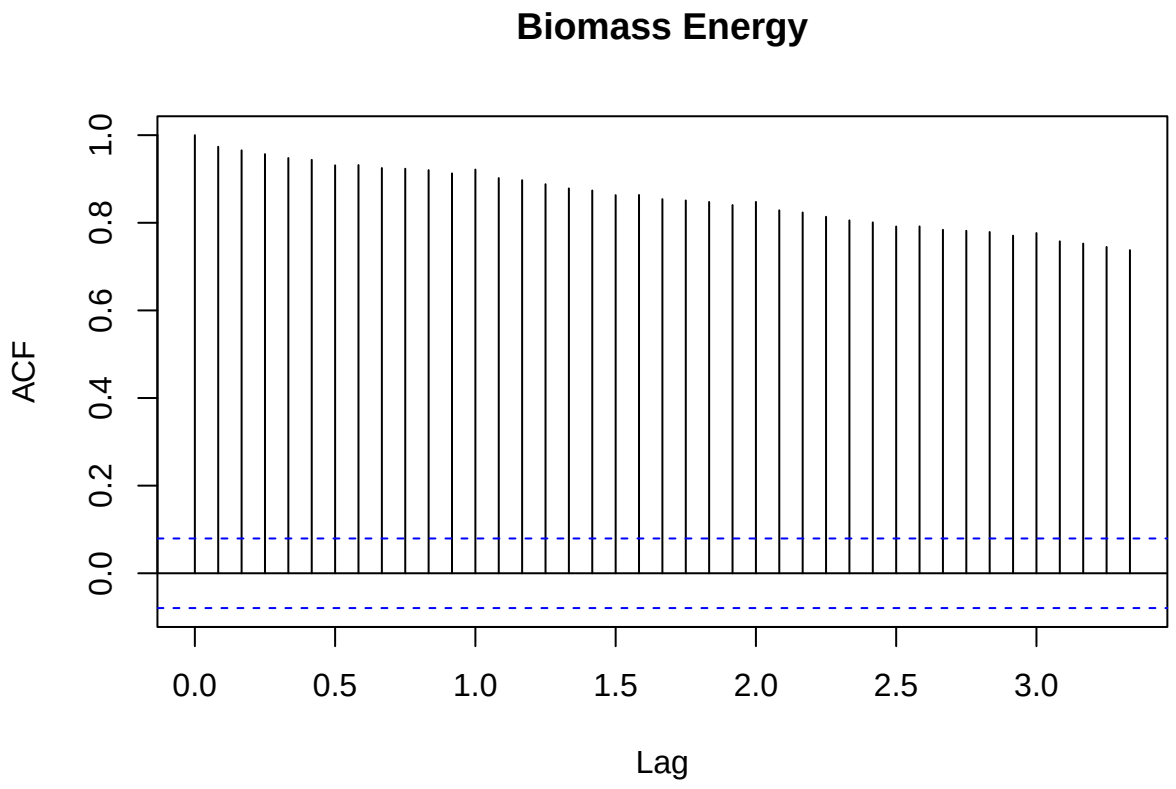
```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                       1.00000000
## Total Renewable Energy Production                     0.97074621
## Hydroelectric Power Consumption                      -0.09656318
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                        0.970746212
## Total Renewable Energy Production                      1.000000000
## Hydroelectric Power Consumption                       -0.001768629
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                      -0.096563177
## Total Renewable Energy Production                    -0.001768629
## Hydroelectric Power Consumption                       1.000000000
```

Answer: Total Biomass Energy Production and Total Renewable Energy Production have a positive correlation coefficient of 0.97. This is very high, and since the scale of co-efficients are between -1 and 1, 0.97 is significantly correlated. Biumass and Hydroelectric have a slight negative correlation at -0.097, but is close to zero, which would indicate an absence of a correlation - they are not significantly correlated. Renewable and Hydroelectric have an even small coefficient of -0.0018, which is very close to zero, meaning they are not significantly correlated.

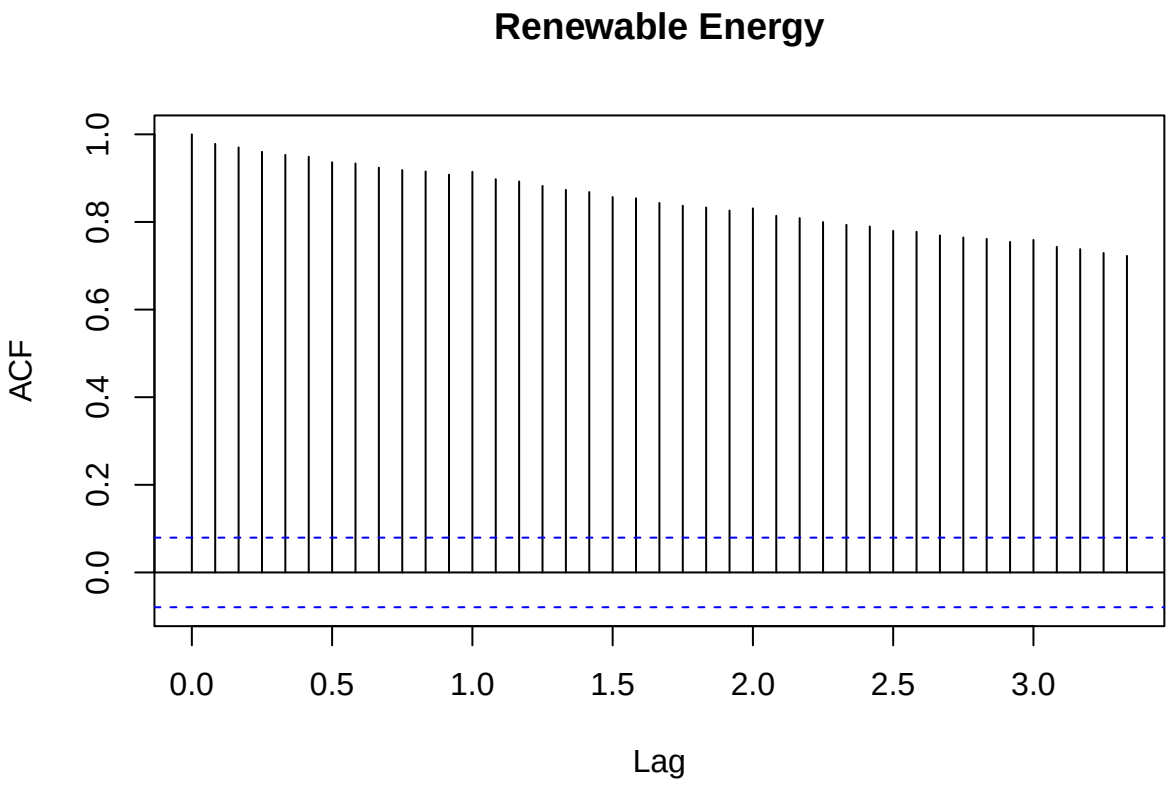
Question 6

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?

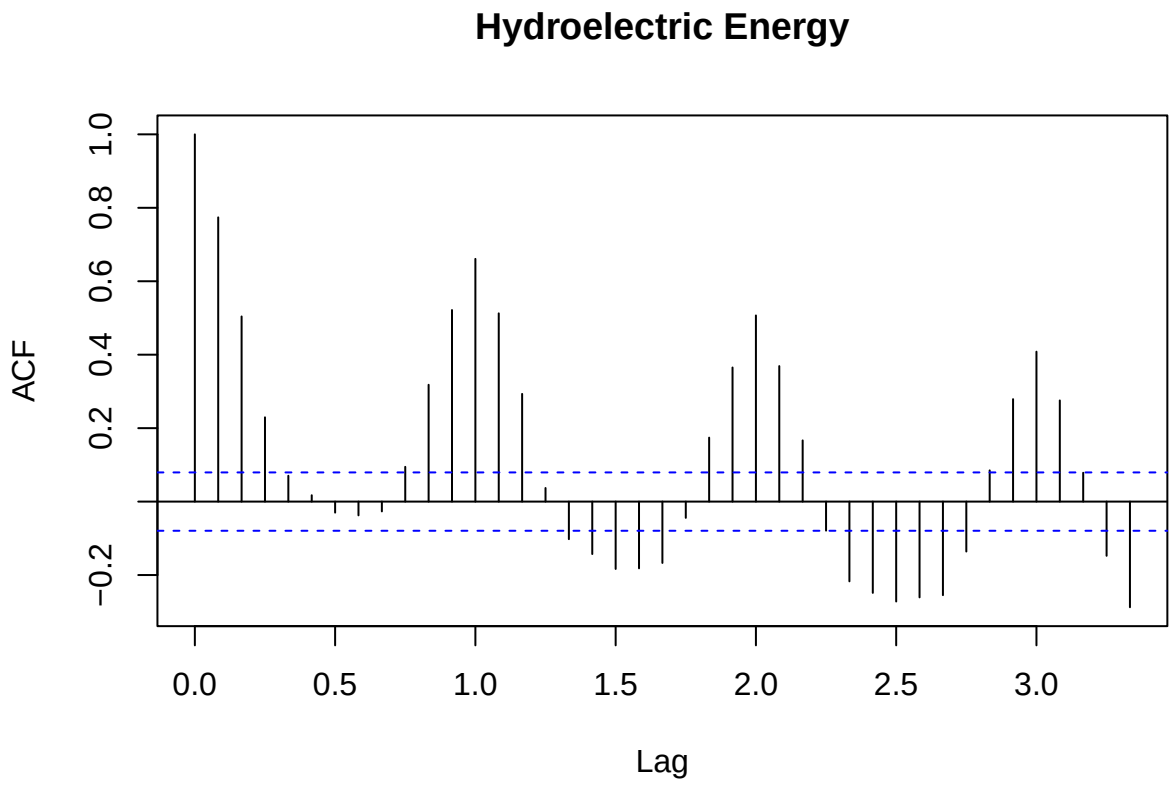
```
# computing the autocorrelation function from lag 1 to 40
acf(ts_energy_data_3[,2], lag.max = 40, type = "correlation", plot = TRUE, main = "Biomass Energy")
```



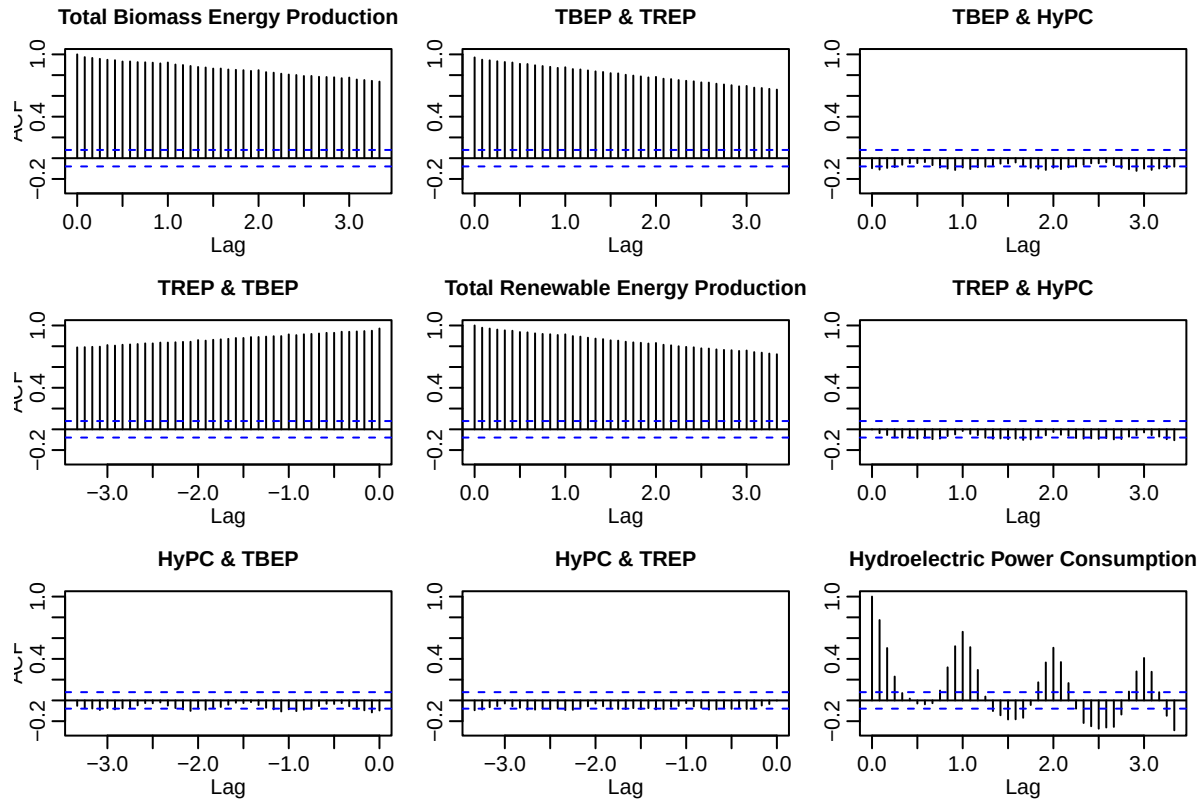

```
acf(ts_energy_data_3[,3], lag.max = 40, type = "correlation", plot = TRUE, main = "Renewable Energy")
```




```
acf(ts_energy_data_3[,4], lag.max = 40, type = "correlation", plot = TRUE, main = "Hydroelectric Energy
```




```
# all three together
acf(ts_energy_data_3[,2:4], lag.max = 40, type = "correlation", plot = TRUE)
```

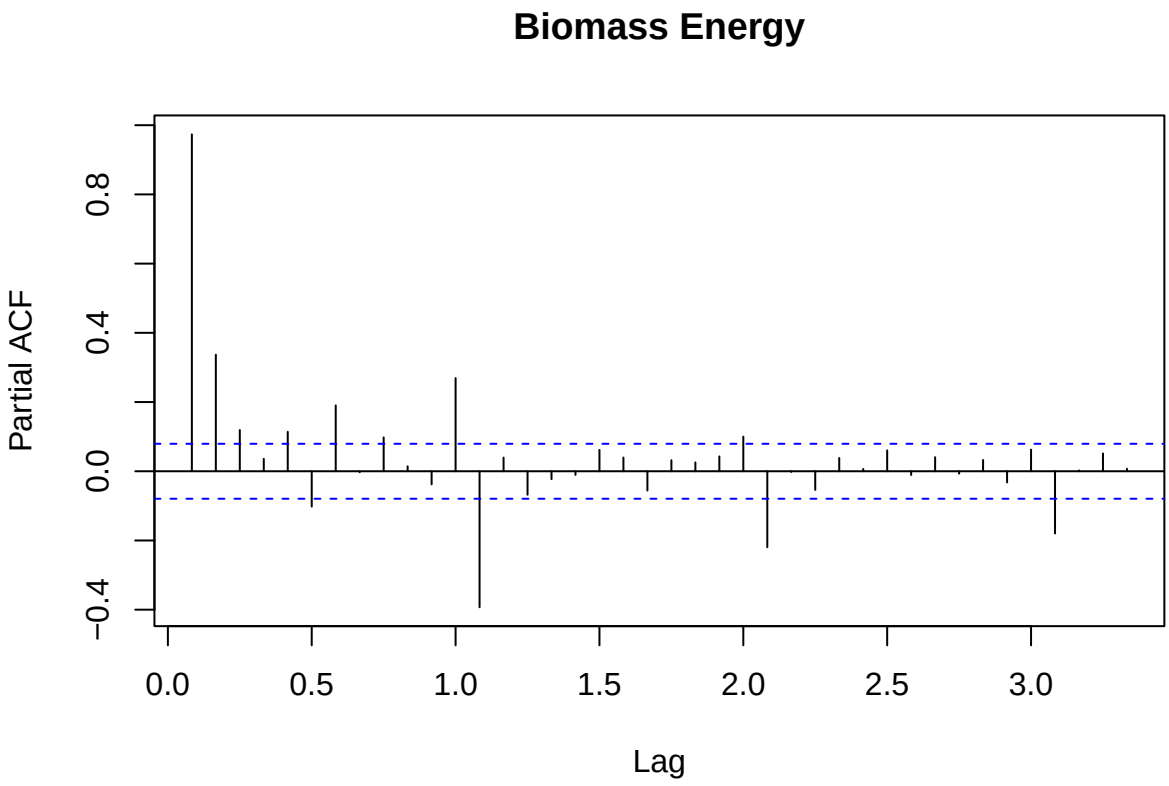


Answer: There is strong correlation in Total Biomass Energy Production and Total Renewable Energy Production; they behave similarly. The Hydroelectric Power Consumption behaves differently from the other two and has a wave-like cyclical pattern behavior.

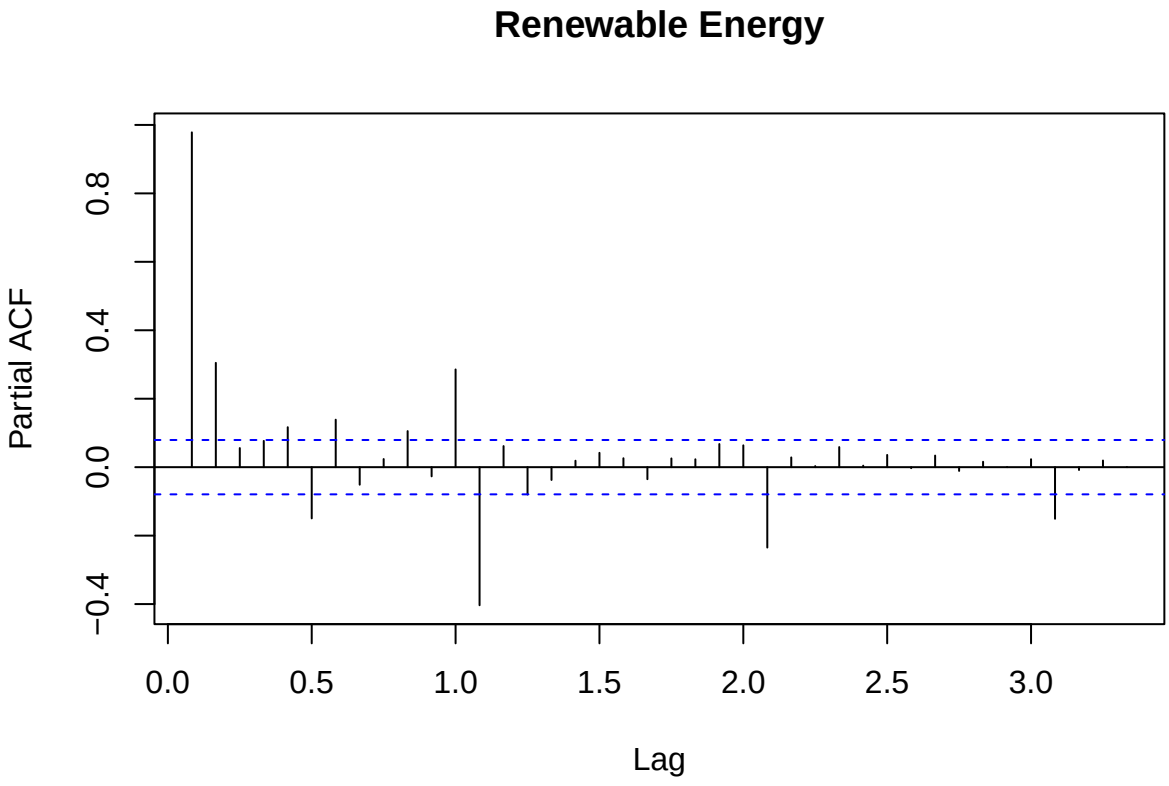
Question 7

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

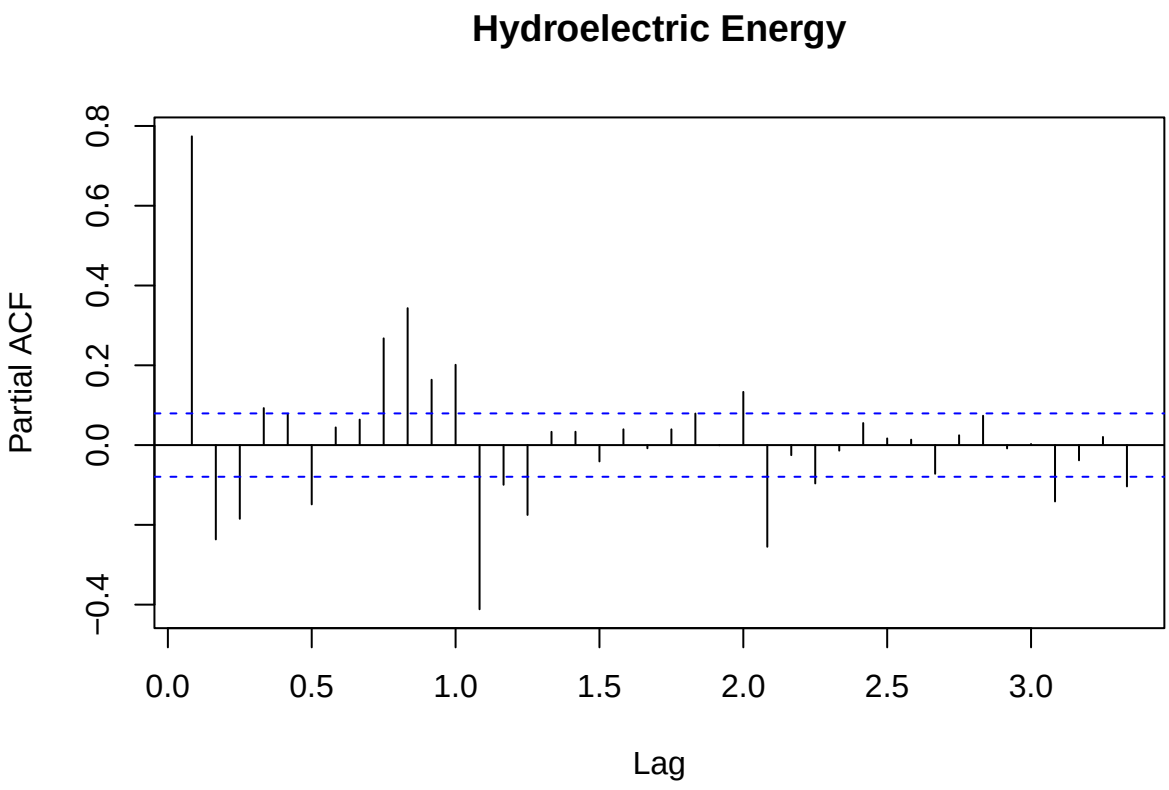
```
# computing the pacf from lag 1 to 40
pacf(ts_energy_data_3[,2], lag.max = 40, plot = TRUE, main = "Biomass Energy")
```



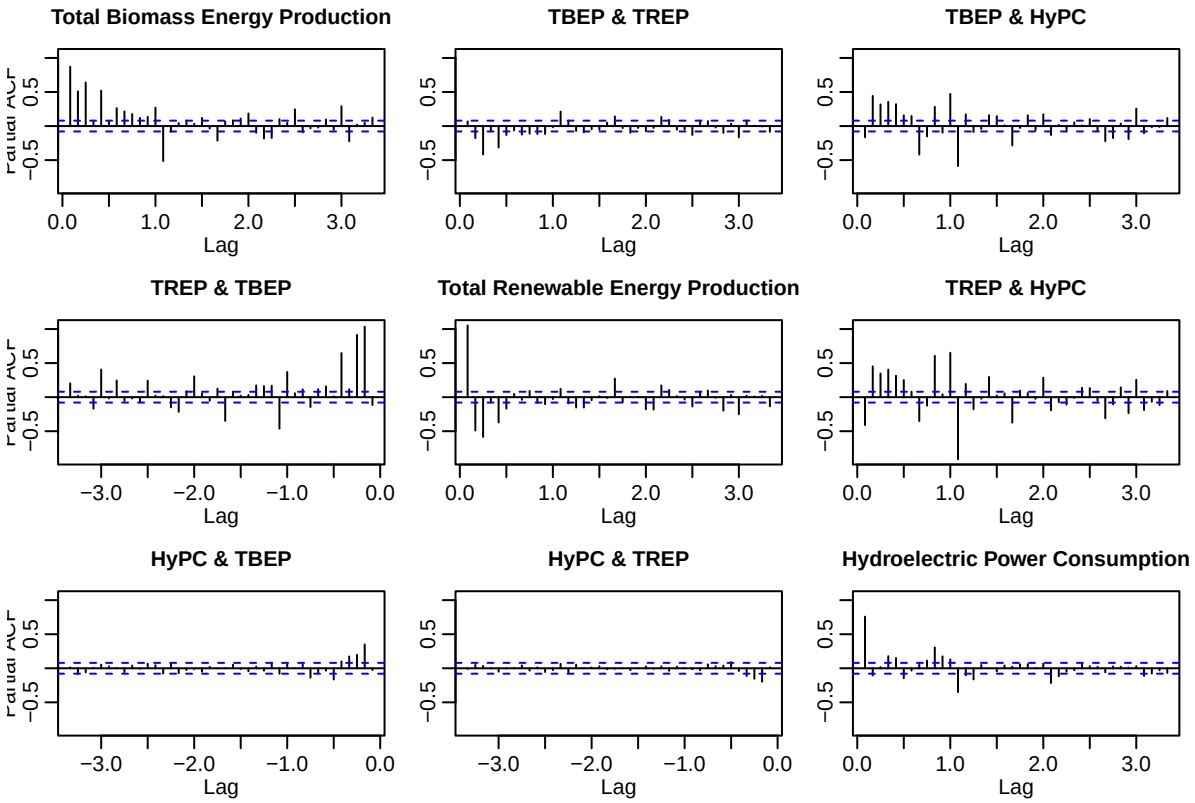

```
pacf(ts_energy_data_3[,3], lag.max = 40, plot = TRUE, main = "Renewable Energy")
```




```
pacf(ts_energy_data_3[,4], lag.max = 40, plot = TRUE, main = "Hydroelectric Energy")
```




```
# all three together
pacf(ts_energy_data_3[,2:4], lag.max = 40, plot = TRUE)
```



Answer: The PACF for Biomass shows that there might be some seasonality, but overall has less linear dependence on the lags. This is a much different story than what is in the plots of Q6 for Biomass - where there is high dependence between the lags, with a decay in the dependence from 1.0 to 0.8 over the lags. The Renewable plot ([,3]) shows similar behavior - the PACF shows there is probably some seasonality in the data, but generally a lower dependence of the lags, unlike the high dependence shown in the plot in Q6.For the hydroelectric ACF in Q6, there is an evenly shaped wave moving through the lags, showing a seasonal trend, but not nearly as high of dependence on the lags as the other two pots. The PACF for hydroelectric also shows some potential seasonality as well.