ENV 790.30 - Time Series Analysis for Energy Data | Spring 2025

Assignment 2 - Due date 01/27/26

Lingyue Hao

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_Sp26.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.\

```
library(forecast)
```

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

```
library(tseries)
library(dplyr)
```

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

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```r
library(readxl)
library(openxlsx)
```

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 2025 Monthly Energy Review. The spreadsheet is ready to be used. Refer to the file "M2_ImportingData_XLSX.Rmd" in our Lessons folder for instructions on how to read *.xlsx* files.

```r
energy_data1 <- read_excel(
  path = "../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
  sheet = "Monthly Data",
  skip = 12,
  col_names = FALSE
)
```

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

```r
read_col_names <- read_excel(
  path = "../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
  sheet = "Monthly Data",
  skip = 10,
  n_max = 1,
  col_names = FALSE
)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
```

```
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
colnames(energy_data1) <- as.character(read_col_names[1, ])
head(energy_data1)
```

```
## # 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.

```
energy_q1 <- energy_data1[
  , c(
    "Total Biomass Energy Production",
    "Total Renewable Energy Production",
    "Hydroelectric Power Consumption"
  )
]

head(energy_q1)
```

```
## # A tibble: 6 x 3
##   Total Biomass Energy Productio~1 Total Renewable Ener~2 Hydroelectric Power ~3
##                              <dbl>                  <dbl>                   <dbl>
## 1                             130.                   220.                    89.6
## 2                             117.                   197.                    79.5
## 3                             130.                   219.                    88.3
## 4                             126.                   209.                    83.2
## 5                             130.                   216.                    85.6
## 6                             126.                   208.                    82.1
## # i abbreviated names: 1: `Total Biomass Energy Production`,
```

```
## #   2: 'Total Renewable Energy Production',
## #   3: 'Hydroelectric Power Consumption'
```

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().

```
energy_ts <- ts(
  energy_q1,
  start = c(1973, 1),
  frequency = 12
)
head(energy_ts)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Jan 1973                         129.787                             219.839
## Feb 1973                         117.338                             197.330
## Mar 1973                         129.938                             218.686
## Apr 1973                         125.636                             209.330
## May 1973                         129.834                             215.982
## Jun 1973                         125.611                             208.249
##          Hydroelectric Power Consumption
## Jan 1973                          89.562
## Feb 1973                          79.544
## Mar 1973                          88.284
## Apr 1973                          83.152
## May 1973                          85.643
## Jun 1973                          82.060
```

Question 3

Compute mean and standard deviation for these three series.

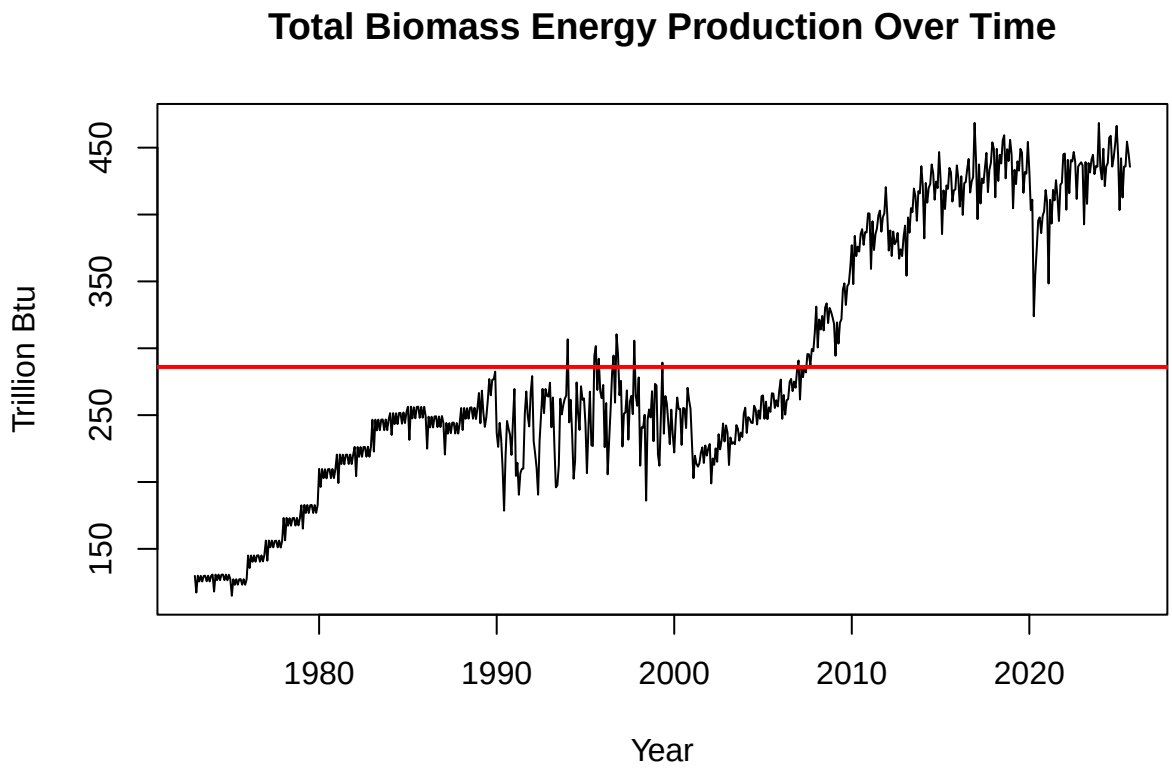
```
mean_energy <- apply(energy_ts, 2, mean, na.rm = TRUE)
sd_energy <- apply(energy_ts, 2, sd,na.rm = TRUE)

energy_stats <- data.frame(
  Mean = mean_energy,
  SD   = sd_energy
)

print(energy_stats)
```

```
##                                       Mean        SD
## Total Biomass Energy Production   286.04893  96.21209
## Total Renewable Energy Production 409.19521 151.42232
## Hydroelectric Power Consumption    79.35682  14.12020
```

Question 4

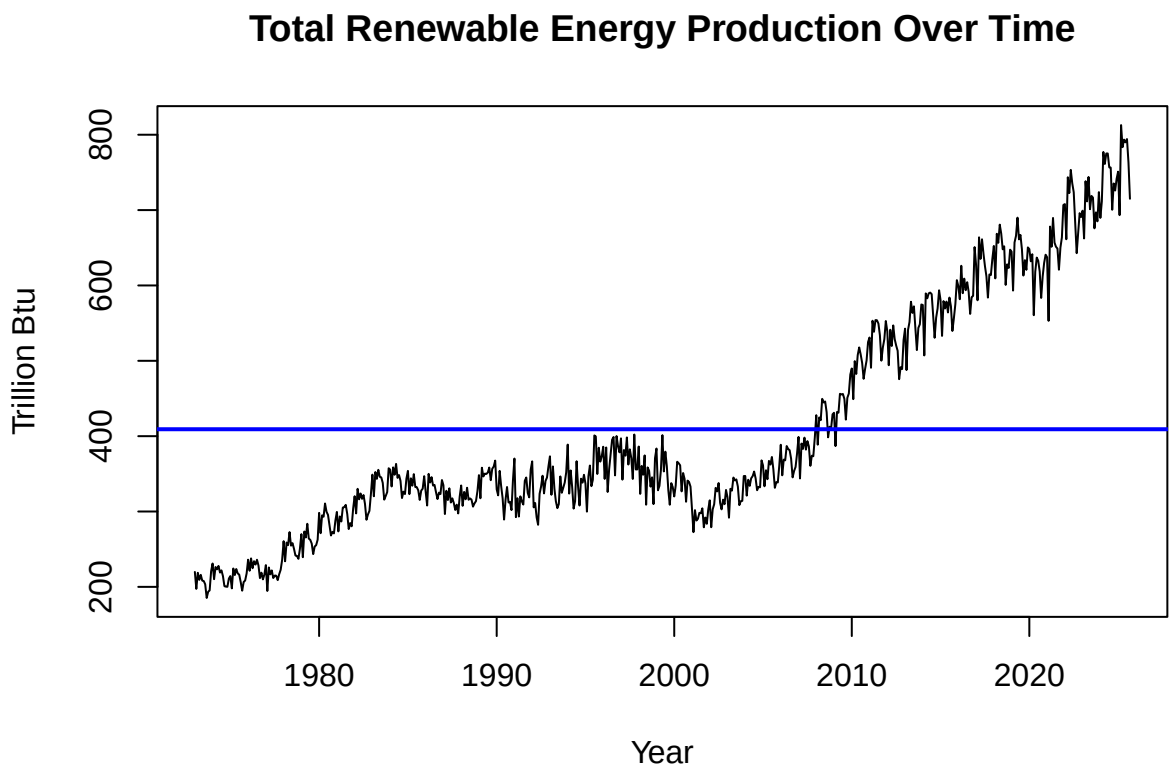
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
plot(
  energy_ts[, "Total Biomass Energy Production"],
  main = "Total Biomass Energy Production Over Time",
  xlab = "Year",
  ylab = "Trillion Btu",
  col  = "black"
)

abline(
  h = mean(energy_ts[, "Total Biomass Energy Production"], na.rm = TRUE),
  col = "red",
  lwd = 2
)
```

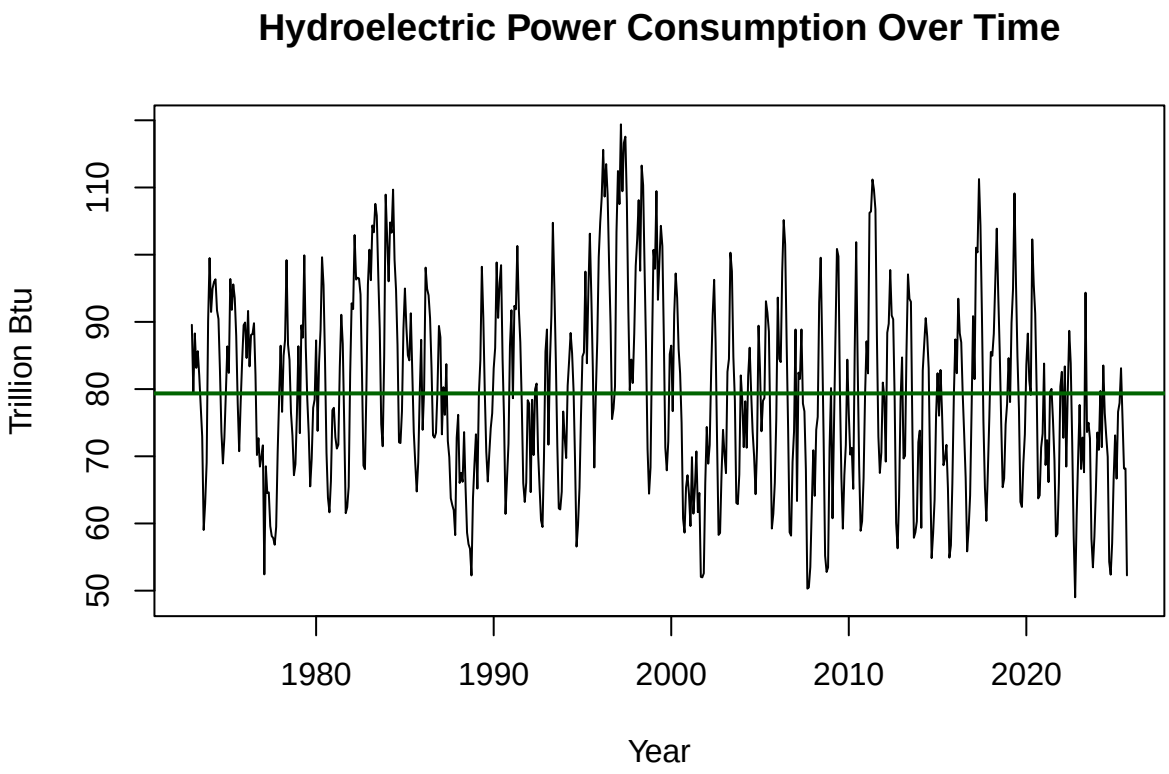



```
plot(
  energy_ts[, "Total Renewable Energy Production"],
  main = "Total Renewable Energy Production Over Time",
  xlab = "Year",
  ylab = "Trillion Btu",
```

```
  col  = "black"
)

abline(
  h = mean(energy_ts[, "Total Renewable Energy Production"], na.rm = TRUE),
  col = "blue",
  lwd = 2
)
```




```
plot(
  energy_ts[, "Hydroelectric Power Consumption"],
  main = "Hydroelectric Power Consumption Over Time",
  xlab = "Year",
  ylab = "Trillion Btu",
  col  = "black"
)

abline(
  h = mean(energy_ts[, "Hydroelectric Power Consumption"], na.rm = TRUE),
  col = "darkgreen",
  lwd = 2
)
```



Total biomass energy production and total renewable energy production both display clear upward trends, especially after the mid-2000s, with most recent values lying well above their historical means. In contrast, hydroelectric power consumption does not show a strong long-term trend and instead fluctuates around its mean with noticeable short-term variability, likely due to seasonal and environmental factors. So biomass and overall renewable energy production have expanded significantly over time, hydroelectric power consumption has remained relatively stable.

Question 5

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

```
cor_energy <- cor(energy_ts, use = "complete.obs")
cor_energy
```

```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.9652985
## Hydroelectric Power Consumption                       -0.1347374
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                          0.96529851
## Total Renewable Energy Production                        1.00000000
## Hydroelectric Power Consumption                         -0.05842436
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                       -0.13473742
## Total Renewable Energy Production                     -0.05842436
## Hydroelectric Power Consumption                        1.00000000
```

```r
cor.test(
  energy_ts[, "Total Biomass Energy Production"],
  energy_ts[, "Total Renewable Energy Production"]
)
```

```
##
##  Pearson's product-moment correlation
##
## data:  energy_ts[, "Total Biomass Energy Production"] and energy_ts[, "Total Renewable Energy Produc
## t = 92.851, df = 631, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9595516 0.9702413
## sample estimates:
##       cor
## 0.9652985
```

```r
cor.test(
  energy_ts[, "Total Biomass Energy Production"],
  energy_ts[, "Hydroelectric Power Consumption"]
)
```

```
##
##  Pearson's product-moment correlation
##
## data:  energy_ts[, "Total Biomass Energy Production"] and energy_ts[, "Hydroelectric Power Consumptio
## t = -3.4157, df = 631, p-value = 0.000677
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.21045616 -0.05741173
## sample estimates:
##        cor
## -0.1347374
```

```r
cor.test(
  energy_ts[, "Total Renewable Energy Production"],
  energy_ts[, "Hydroelectric Power Consumption"]
)
```

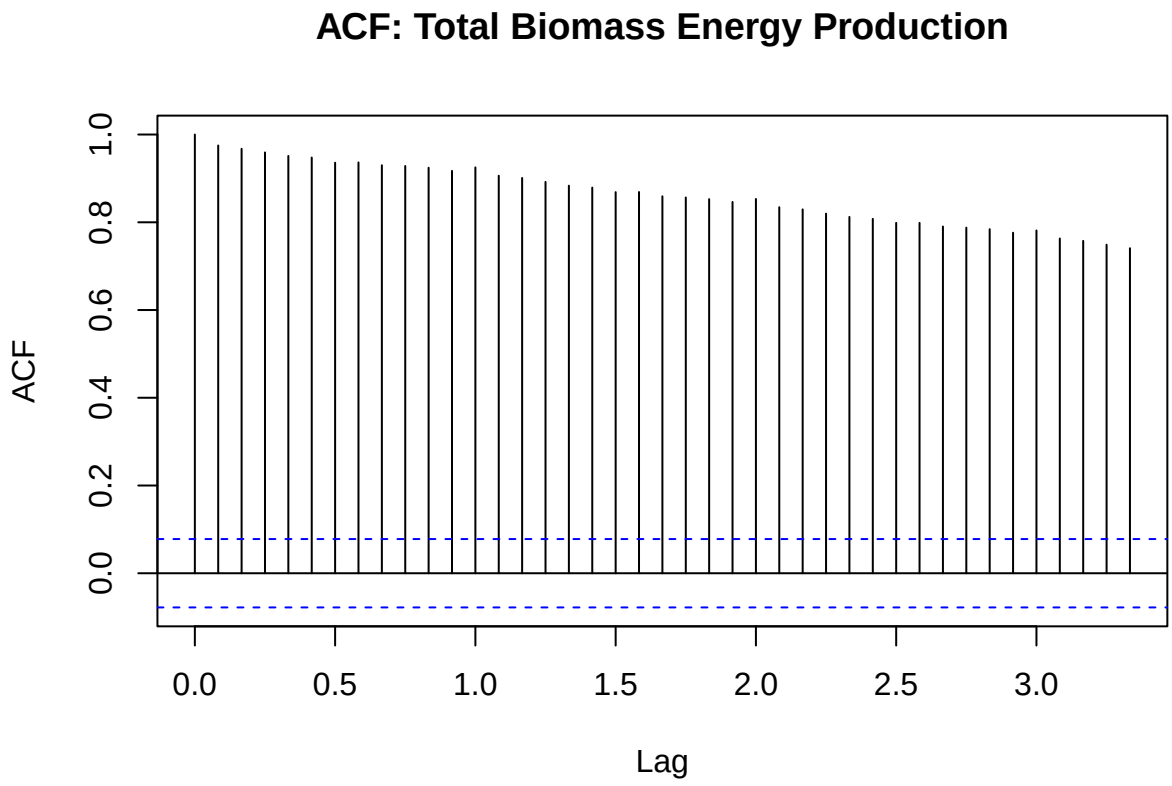
```
##
##  Pearson's product-moment correlation
##
## data:  energy_ts[, "Total Renewable Energy Production"] and energy_ts[, "Hydroelectric Power Consump
## t = -1.4701, df = 631, p-value = 0.142
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13573488  0.01959335
## sample estimates:
##         cor
## -0.05842436
```

The results show that Total Biomass Energy Production and Total Renewable Energy Production are strongly and significantly correlated, with a very high correlation coefficient (about 0.97) and an extremely small p-value, which indicates a clear positive relationship. This makes sense because biomass energy is an important part of total renewable energy. In contrast, Hydroelectric Power Consumption does not show a strong relationship with the other two series. Its correlation with biomass energy is weakly negative, and although it is statistically significant, the magnitude is small. Its correlation with total renewable energy is also very weak and not statistically significant, meaning there is no strong evidence of a linear relationship.

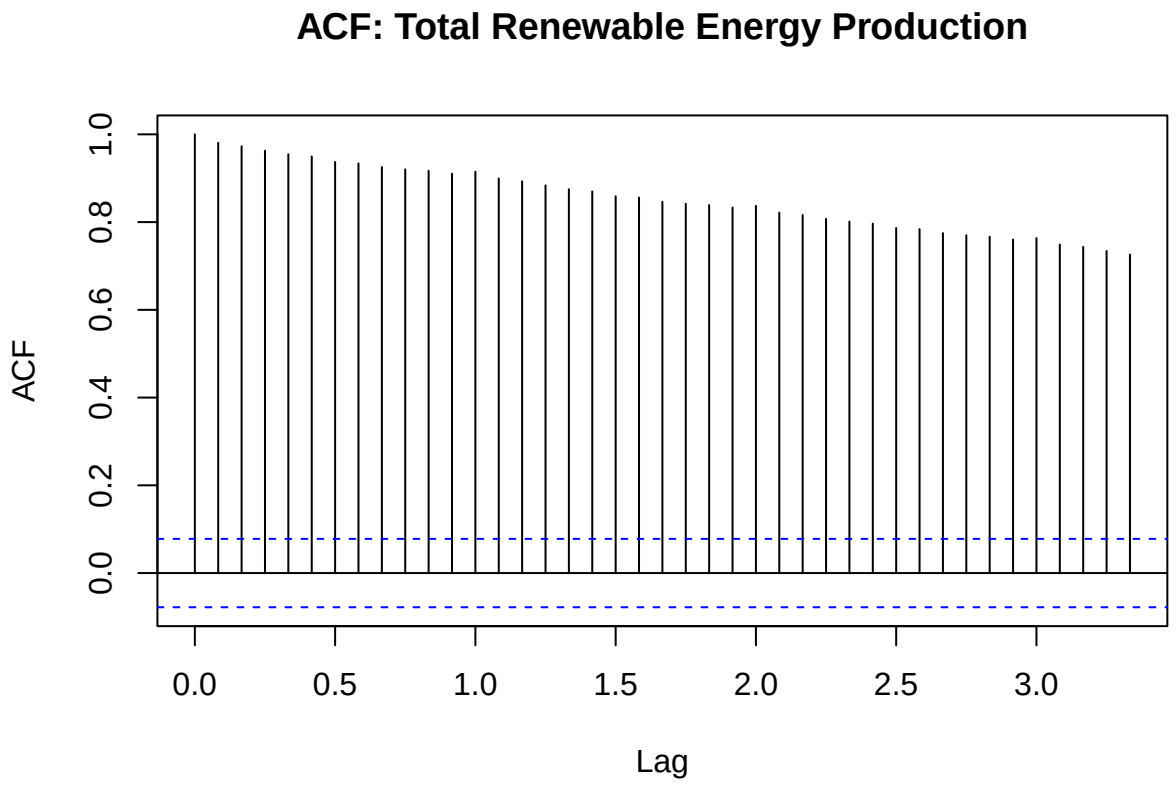
Question 6

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?

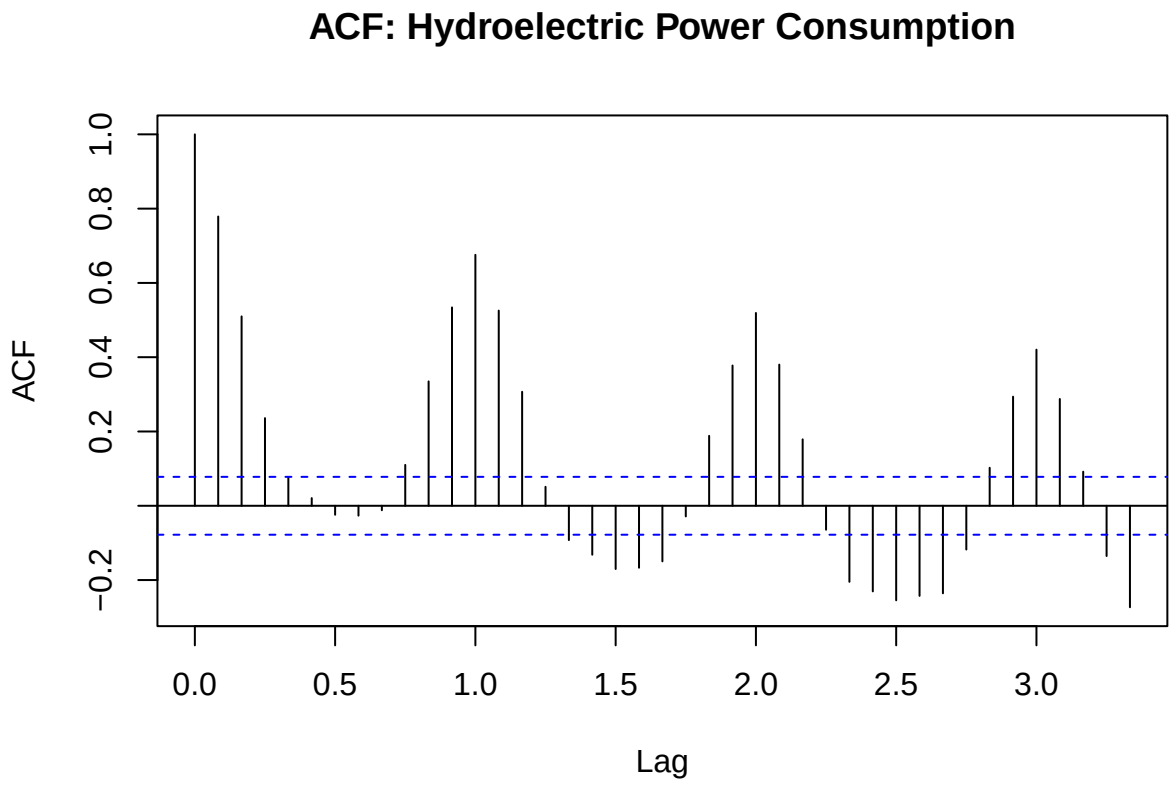
```
acf(
  energy_ts[, "Total Biomass Energy Production"],
  lag.max = 40,
  main = "ACF: Total Biomass Energy Production"
)
```




```
acf(
  energy_ts[, "Total Renewable Energy Production"],
  lag.max = 40,
  main = "ACF: Total Renewable Energy Production"
)
```




```
acf(
  energy_ts[, "Hydroelectric Power Consumption"],
  lag.max = 40,
  main = "ACF: Hydroelectric Power Consumption"
)
```

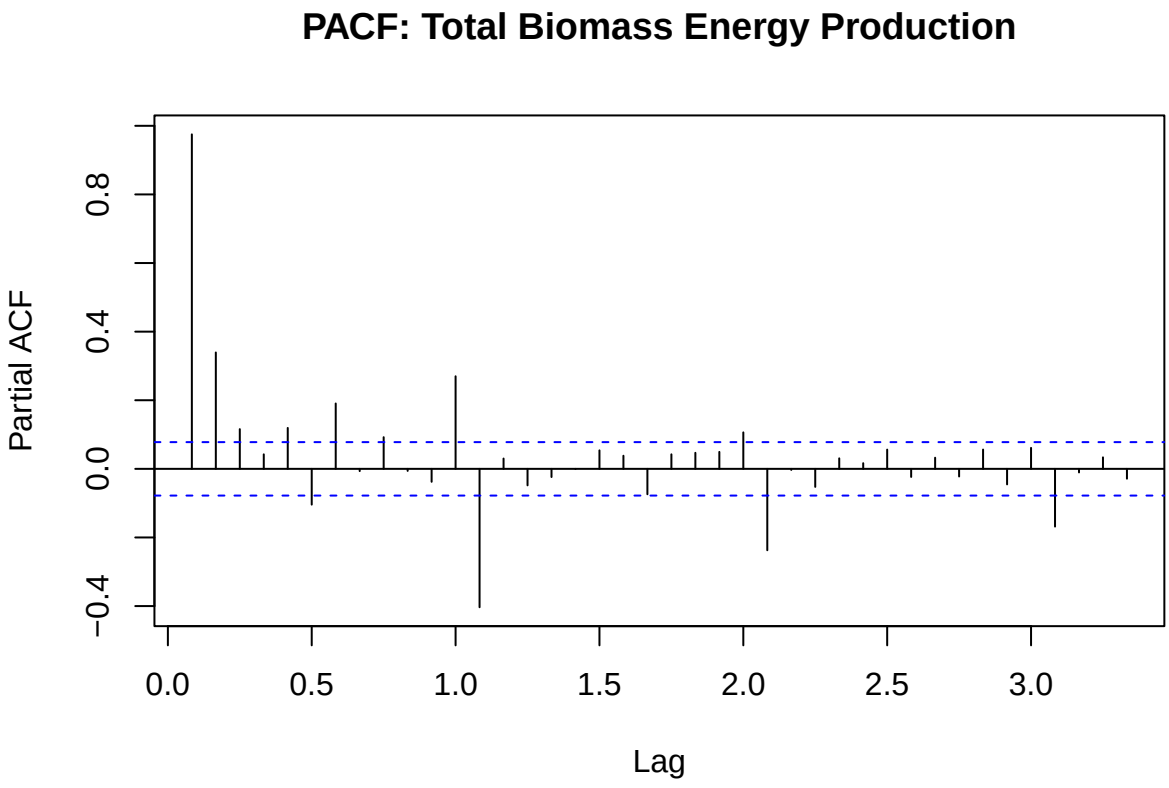


Three series do not show the same behavior. Total Biomass Energy Production and Total Renewable Energy Production both have very high autocorrelations that decay slowly over many lags, which suggests a clear long-term trend. In contrast, Hydroelectric Power Consumption shows autocorrelations that drop off much faster and alternate between positive and negative values, which suggests a more short-term dependence and cyclical behavior.

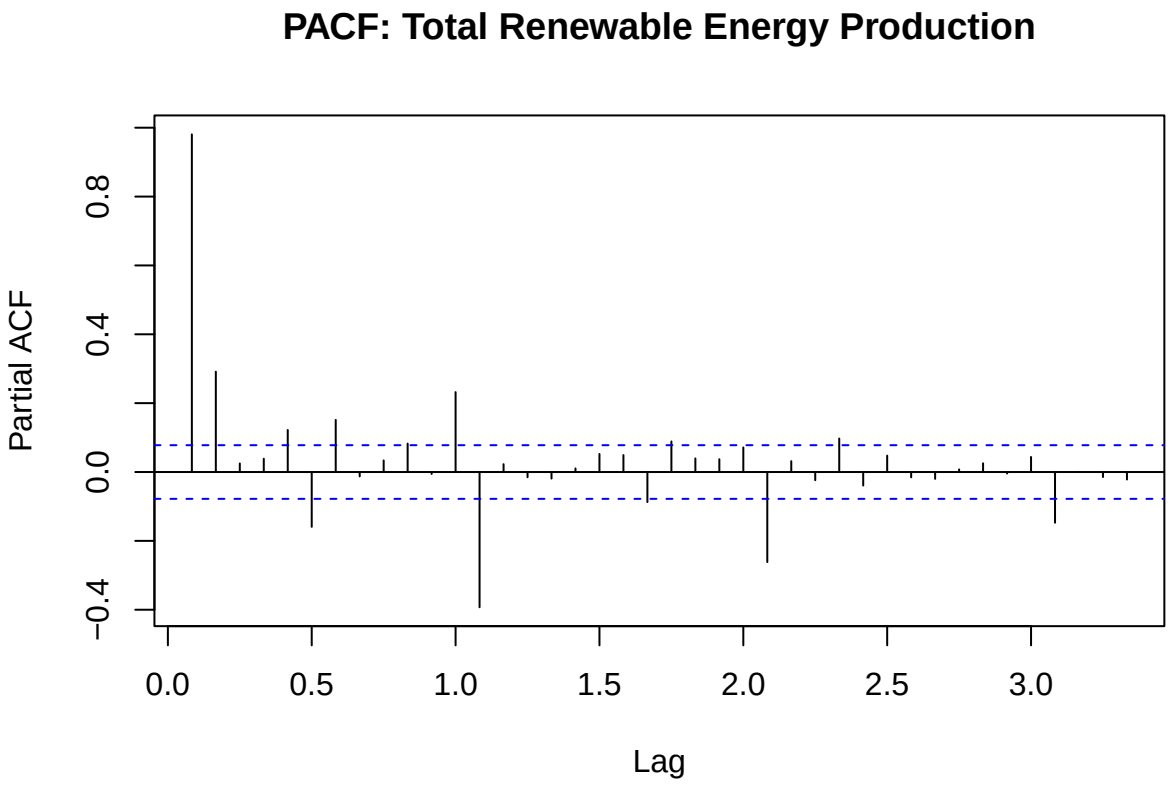
Question 7

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

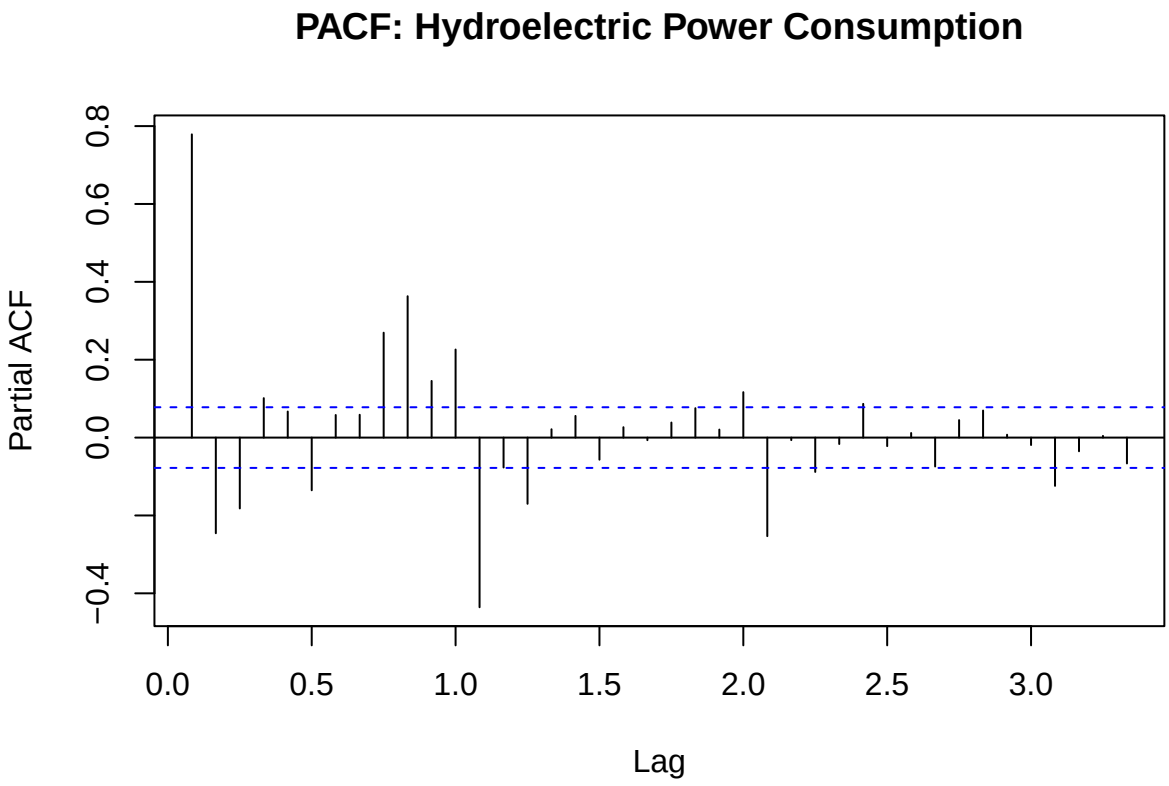
```
pacf(energy_ts[, "Total Biomass Energy Production"], lag.max = 40,
     main = "PACF: Total Biomass Energy Production")
```




```
pacf(energy_ts[, "Total Renewable Energy Production"], lag.max = 40,
     main = "PACF: Total Renewable Energy Production")
```




```
pacf(energy_ts[, "Hydroelectric Power Consumption"], lag.max = 40,
     main = "PACF: Hydroelectric Power Consumption")
```



Compared to Q6, the PACF plots in Q7 are much sparser. In Q6, the ACF shows strong and persistent correlations across many lags, especially for total biomass and total renewable energy. In contrast, the PACF shows that once earlier lags are controlled for, many of these correlations shrink back into the confidence bounds. For hydroelectric power consumption, although the ACF shows several significant correlations across different lags, the PACF suggests that only a limited number of lags have independent effects, with most others being explained by earlier lags.