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

Assignment 2 - Due date 01/23/25

Rosie Wu

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
library(lubridate)
library(ggplot2)
library(forecast) #added for Acf and Pacf functions
# install.packages("tseries")
library(tseries)
# install.packages("dplyr")
library(dplyr)
library(here)
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 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*.

```r
#Importing data set
here()
```

```
## [1] "/home/guest/TSA_Sp25"
```

```r
# df <- read_excel(path= "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"
#Importing data set without change the original file using read.xlsx
energy_data1 <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
```

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

```r
#Now let's extract the column names from row 11
read_col_names <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
```

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

```r
#Assign the column names to the data set
colnames(energy_data1) <- read_col_names

#Visualize the first rows of the data set
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.

```r
energy_df_selected <- energy_data1 %>%
  select("Month", "Total Biomass Energy Production",
         "Total Renewable Energy Production",
         "Hydroelectric Power Consumption")
energy_df_selected$Month <- as.Date(energy_df_selected$Month)
head(energy_df_selected)
```

```
## # A tibble: 6 x 4
##   Month      `Total Biomass Energy Production` Total Renewable Energy Producti~1
##   <date>                                 <dbl>                              <dbl>
## 1 1973-01-01                              130.                               220.
## 2 1973-02-01                              117.                               197.
## 3 1973-03-01                              130.                               219.
## 4 1973-04-01                              126.                               209.
## 5 1973-05-01                              130.                               216.
## 6 1973-06-01                              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().

```r
energy_df_ts <- ts(energy_df_selected[,2:4], start=c(2000,1),frequency=12)
biomass_ts <- ts(energy_df_selected[,2],start=c(2000,1),frequency=12)
renewable_ts <- ts(energy_df_selected[,3],start=c(2000,1),frequency=12)
hydro_ts <- ts(energy_df_selected[,4],start=c(2000,1),frequency=12)
```

```r
# Verify the transformation
#print(energy_df_ts)
```

Question 3

Compute mean and standard deviation for these three series.

```r
# Compute mean for each series
mean_biomass <- mean(biomass_ts)
mean_biomass
```

```
## [1] 282.6779
```

```r
mean_renewable <- mean(renewable_ts)
mean_renewable
```

```
## [1] 402.0167
```

```r
mean_hydro <- mean(hydro_ts)
mean_hydro
```

```
## [1] 79.55371
```

```r
# Compute standard deviation for each series
sd_biomass <- sd(biomass_ts)
sd_biomass
```

```
## [1] 94.05815
```

```r
sd_renewable <- sd(biomass_ts)
sd_renewable
```

```
## [1] 94.05815
```

```r
sd_hydro <- sd(biomass_ts)
sd_hydro
```

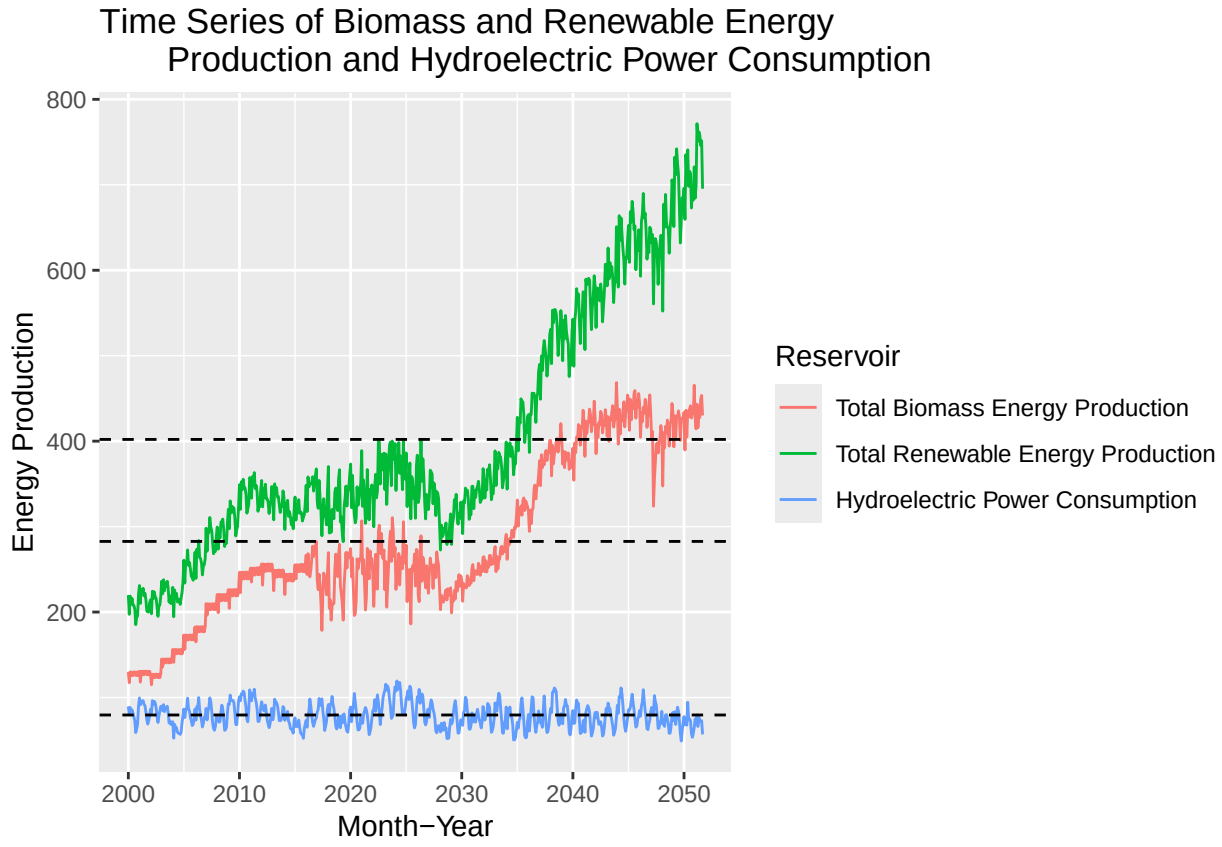
```
## [1] 94.05815
```

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.

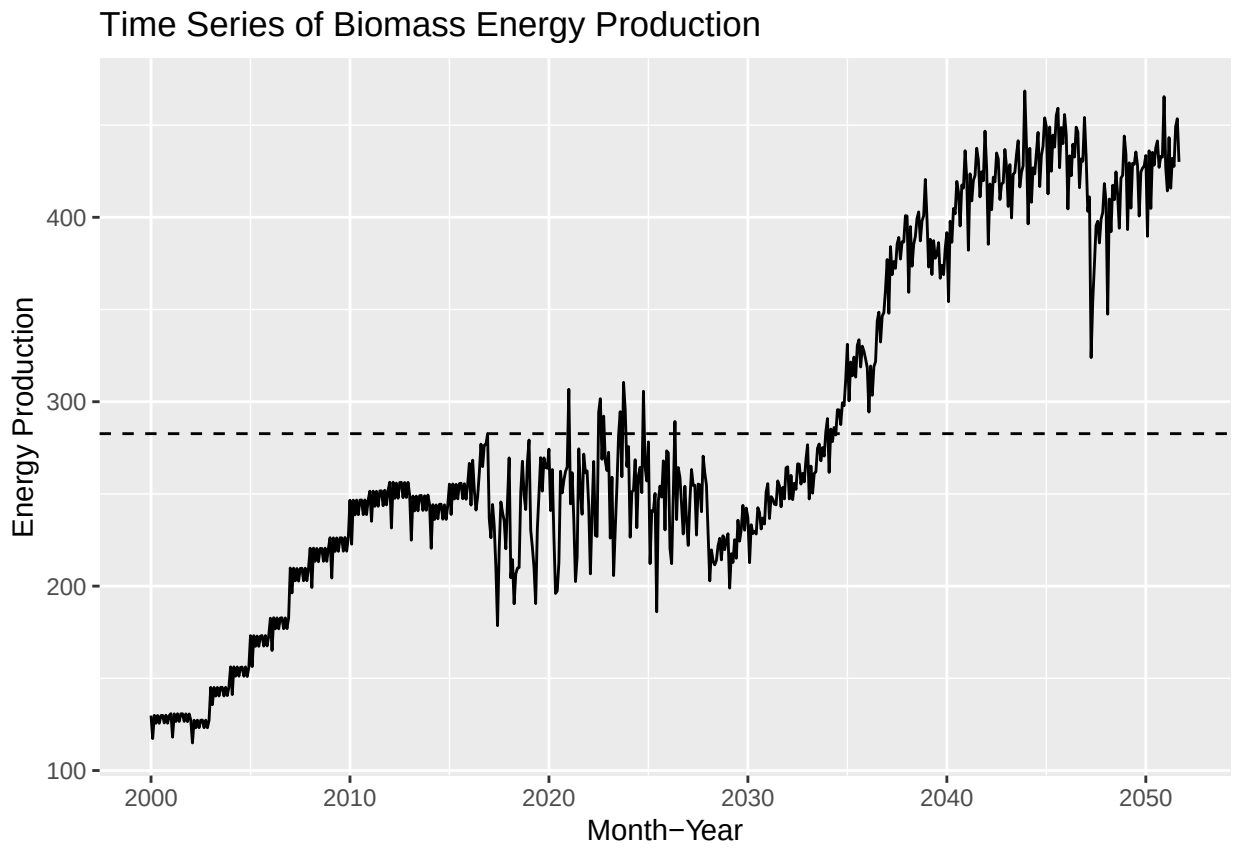
```
# I originally did a all-together graph
autoplot(energy_df_ts) +
  xlab("Month-Year") +
  ylab("Energy Production") +
  labs(color="Reservoir", title = "Time Series of Biomass and Renewable Energy
       Production and Hydroelectric Power Consumption")+
  geom_hline(yintercept = mean(energy_df_ts[, "Total Biomass Energy Production"]),
             color = "black", linetype = "dashed", size = 0.5) +
  geom_hline(yintercept = mean(energy_df_ts[, "Total Renewable Energy Production"]),
             color = "black", linetype = "dashed", size = 0.5) +
  geom_hline(yintercept = mean(energy_df_ts[, "Hydroelectric Power Consumption"]),
             color = "black", linetype = "dashed", size = 0.5)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

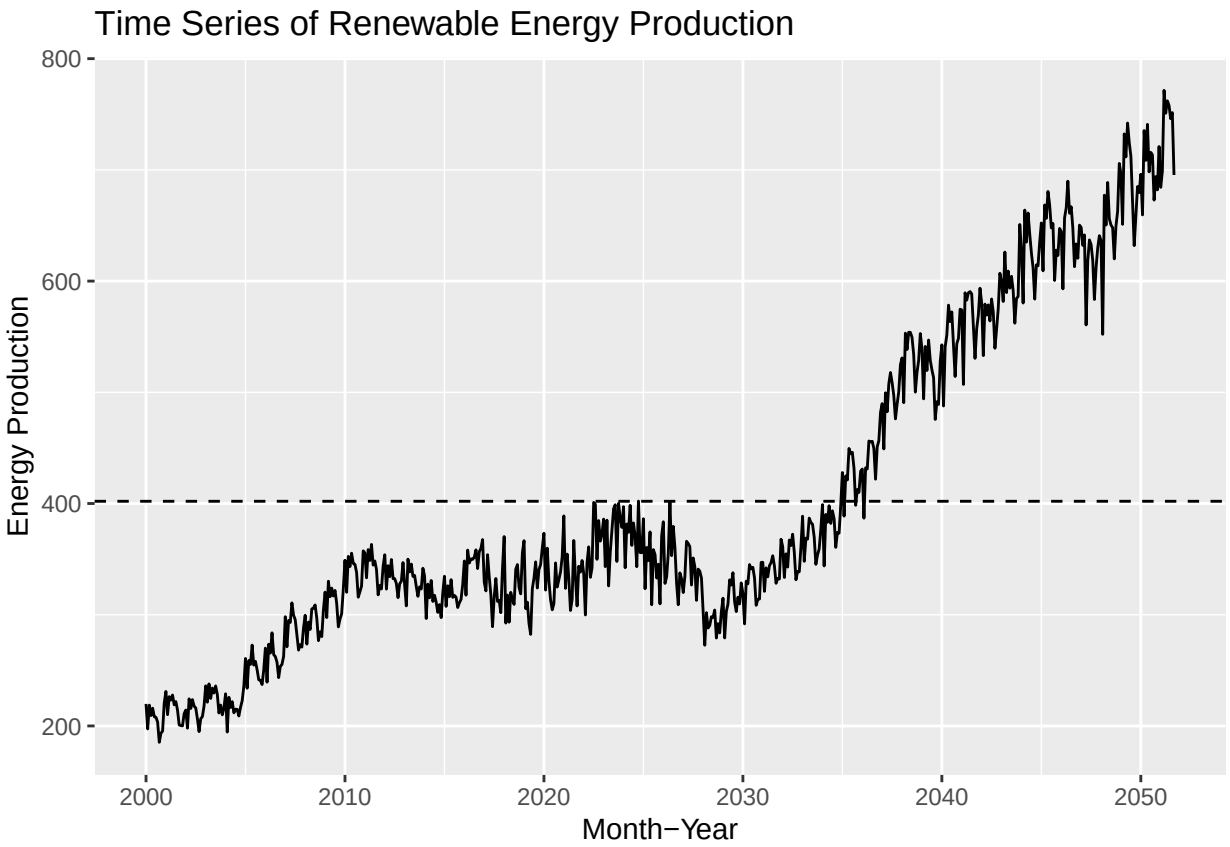



```
# Individual graphs:
autoplot(biomass_ts) +
  xlab("Month-Year") +
  ylab("Energy Production") +
  labs(color="Red", title = "Time Series of Biomass Energy Production")+
```

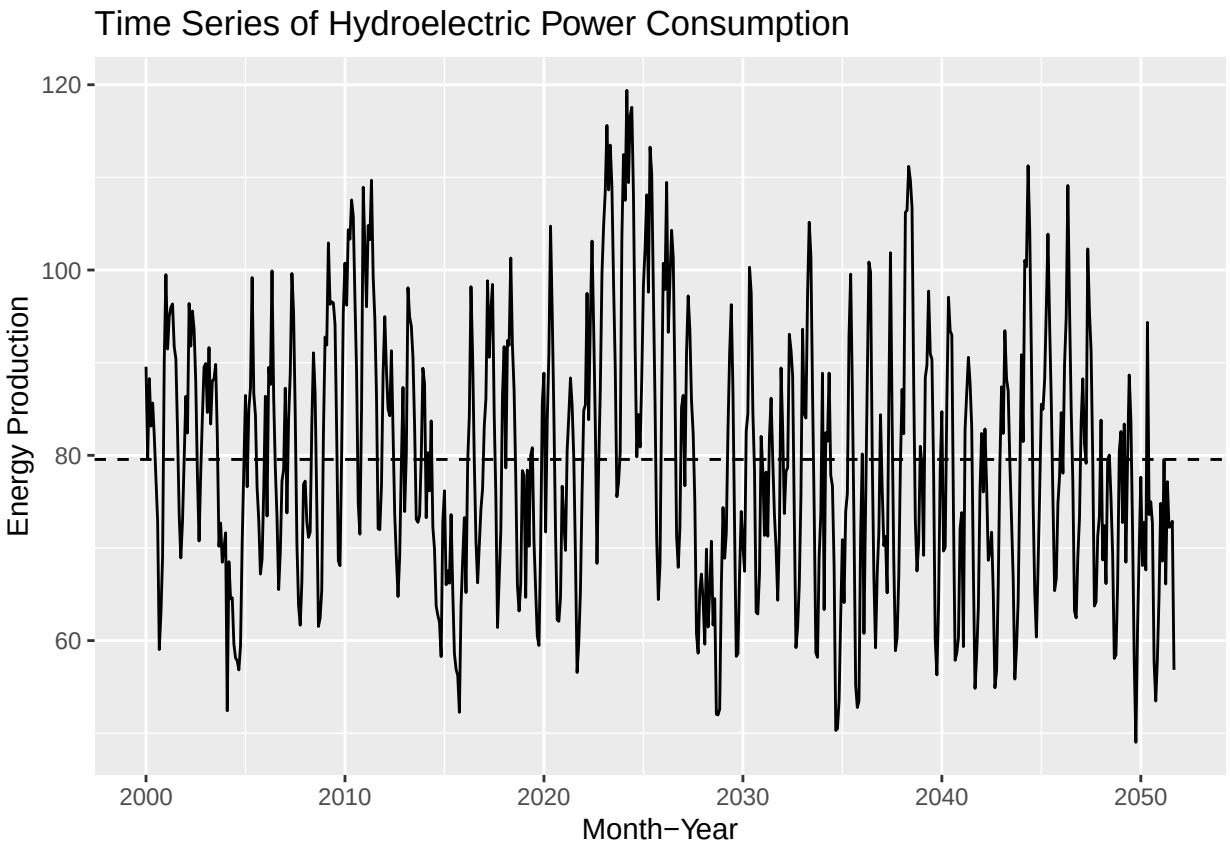
```
  geom_hline(yintercept = mean_biomass,
             color = "black", linetype = "dashed", size = 0.5)
```




```
autoplot(renewable_ts) +
  xlab("Month-Year") +
  ylab("Energy Production") +
  labs(color="Red", title = "Time Series of Renewable Energy Production")+
  geom_hline(yintercept = mean_renewable,
             color = "black", linetype = "dashed", size = 0.5)
```




```
autoplot(hydro_ts) +
  xlab("Month-Year") +
  ylab("Energy Production") +
  labs(color="Red", title = "Time Series of Hydroelectric Power Consumption")+
  geom_hline(yintercept = mean_hydro,
             color = "black", linetype = "dashed", size = 0.5)
```



Question 5

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

```
# can be done in pairs, or create a time series object and have a correlation matrix
cor_matrix <- cor(energy_df_ts)
print(cor_matrix)
```

```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.9678137
## Hydroelectric Power Consumption                       -0.1142927
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                         0.96781371
## Total Renewable Energy Production                       1.00000000
## Hydroelectric Power Consumption                        -0.02916103
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                       -0.11429266
## Total Renewable Energy Production                     -0.02916103
## Hydroelectric Power Consumption                        1.00000000
```

```
# Perform significance test for correlations
cor_test_biomass_renewable <- cor.test(biomass_ts,
                                       renewable_ts)
```

```r
cor_test_biomass_hydro <- cor.test(biomass_ts,
                                   hydro_ts)

cor_test_renewable_hydro <- cor.test(renewable_ts,
                                   hydro_ts)
print(cor_test_biomass_renewable)
```

```
##
##  Pearson's product-moment correlation
##
## data:  biomass_ts and renewable_ts
## t = 95.677, df = 619, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9624198 0.9724443
## sample estimates:
##       cor
## 0.9678137
```

```r
print(cor_test_biomass_hydro)
```

```
##
##  Pearson's product-moment correlation
##
## data:  biomass_ts and hydro_ts
## t = -2.8623, df = 619, p-value = 0.004348
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.19125123 -0.03593747
## sample estimates:
##        cor
## -0.1142927
```

```r
print(cor_test_renewable_hydro)
```

```
##
##  Pearson's product-moment correlation
##
## data:  renewable_ts and hydro_ts
## t = -0.72583, df = 619, p-value = 0.4682
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1075925  0.0496312
## sample estimates:
##         cor
## -0.02916103
```

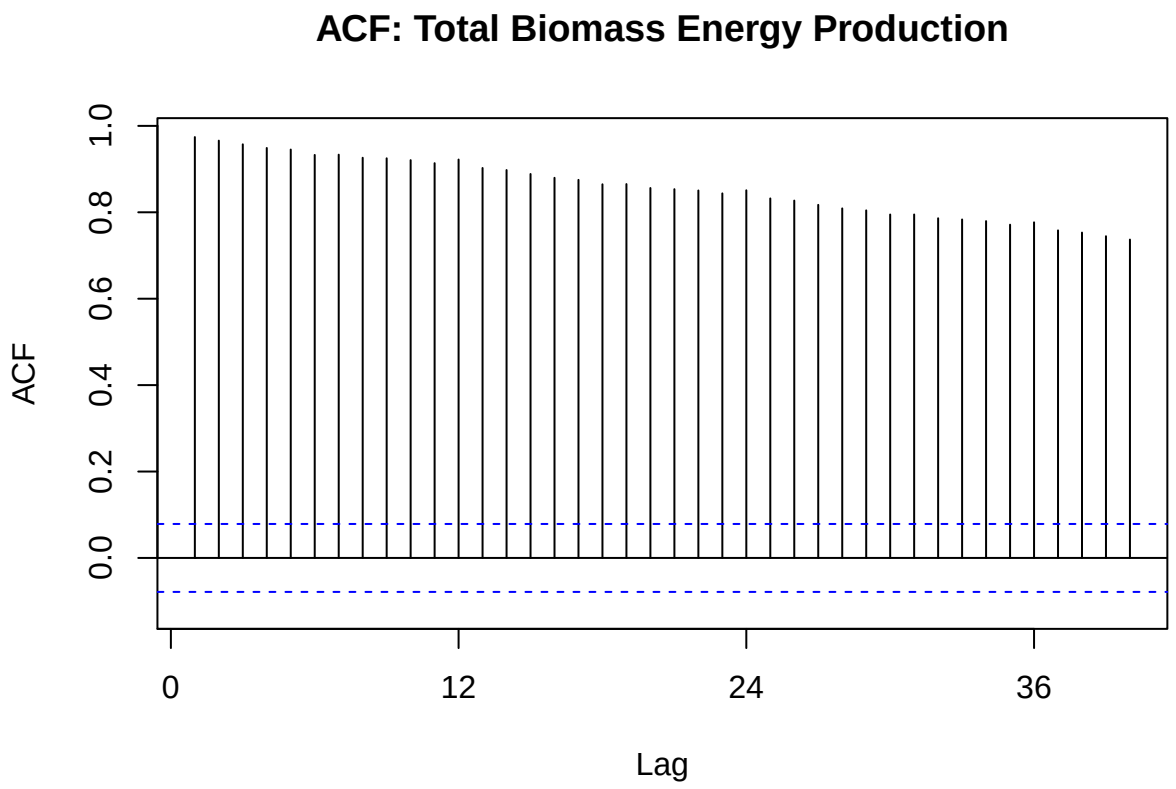
Explain: - The Biomass and Renewable Energy Production are significantly correlated, since t value is extremely high, much higher than 2, and the p-value for the significance test of correlation is extremely close to 0. - The Biomass Energy Production and the Hydro Power Consumption are also significantly correlated, since the t-test stats is $<$ -1.96, and the p value is $<$ 0.05 (basing on

the 95% confidence level). - The renewable energy production and the hydro power consumption are not significantly correlated, since the t-stat is only -0.73, which is not < -1.96 or > 1.96, and the p-value is too high ($0.4682 > 0.05$).

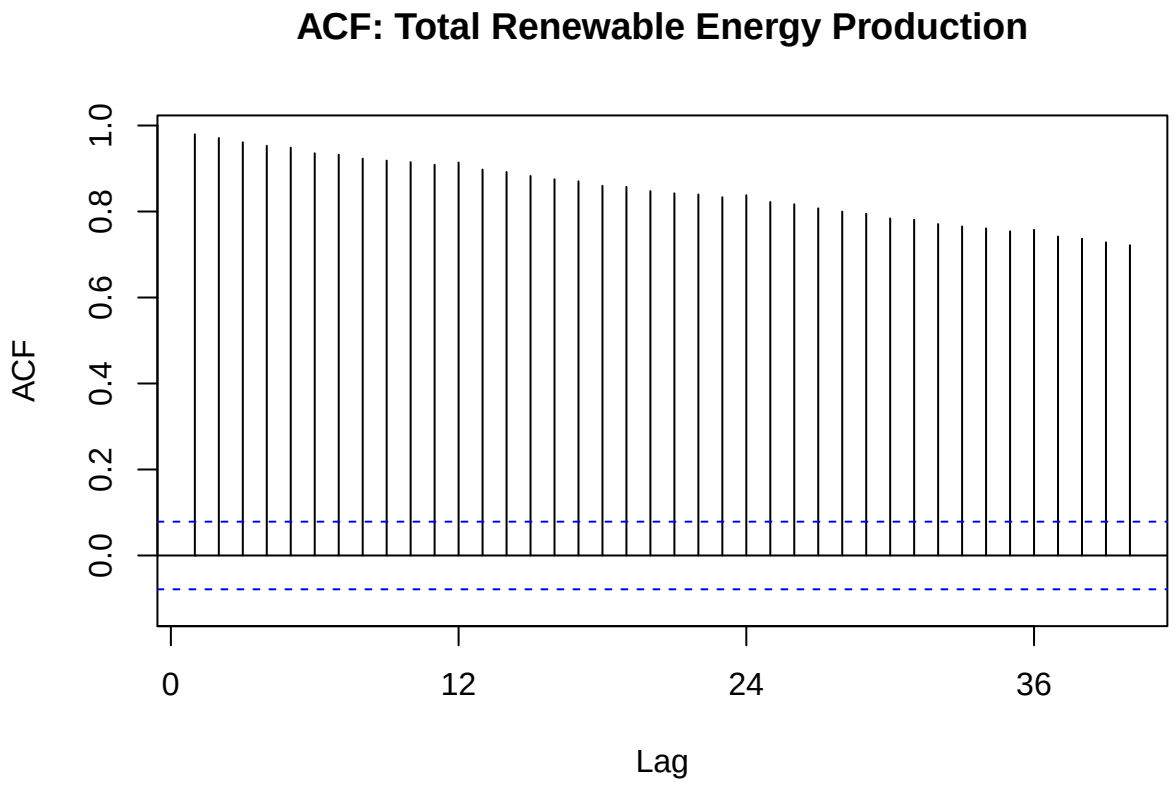
Question 6

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?

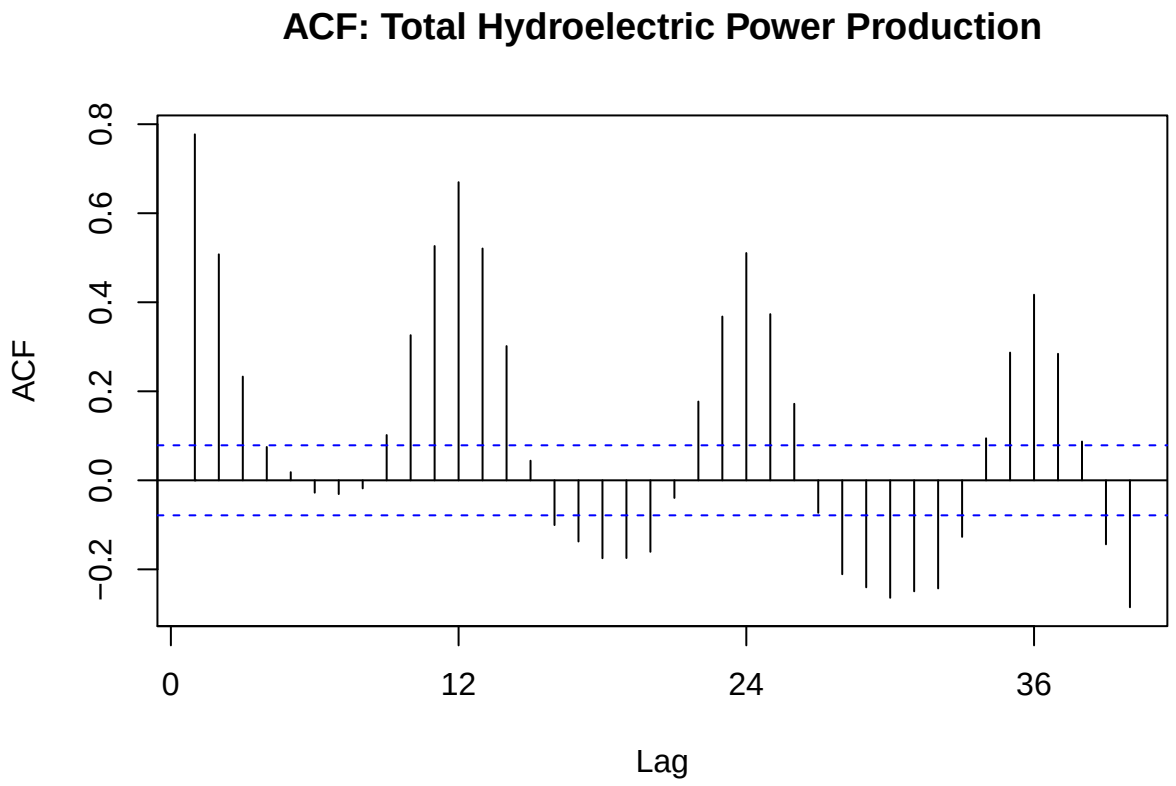
```
biomass_acf = Acf(biomass_ts, lag.max = 40,
                  main = "ACF: Total Biomass Energy Production")
```




```
renewable_acf = Acf(renewable_ts, lag.max = 40,
                  main = "ACF: Total Renewable Energy Production")
```




```
hydro_acf = Acf(hydro_ts, lag.max = 40,
                main = "ACF: Total Hydroelectric Power Production")
```

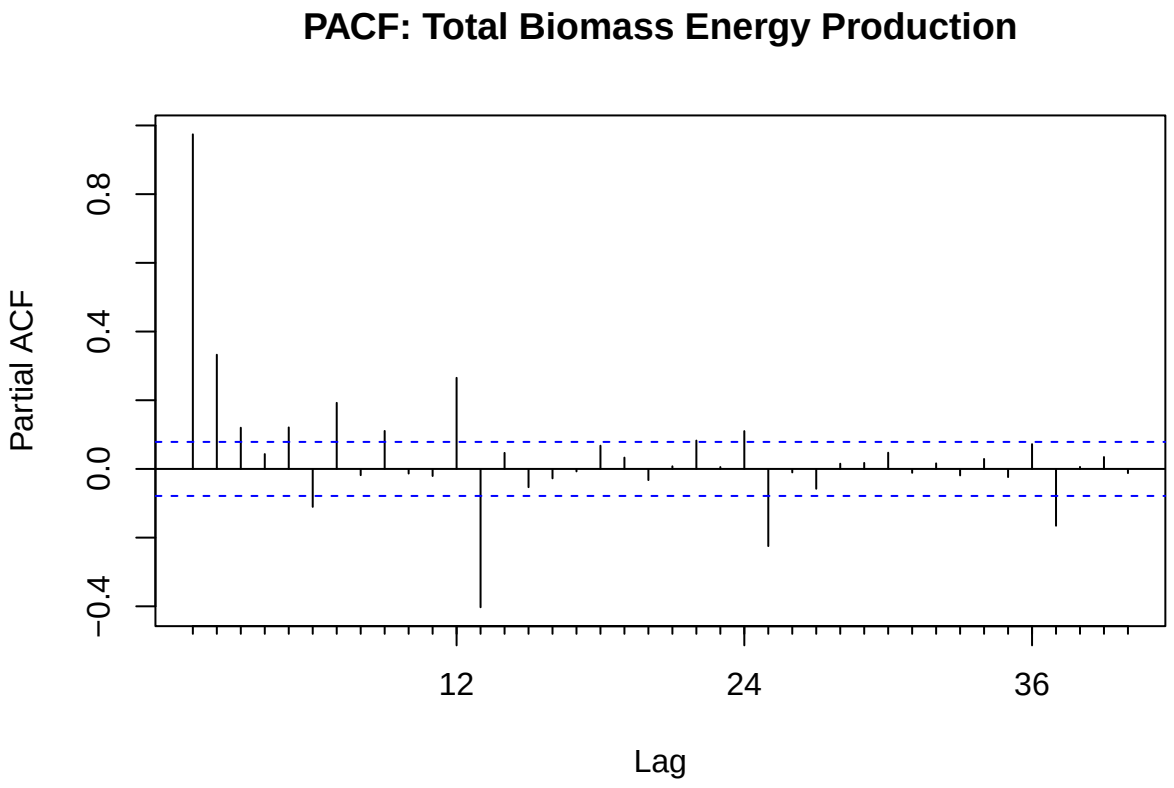


> Answer: The 3 series do show different autocorrelation function trends/ behavoirs, especially Hydropower consumption series is more signficantly different from the other 2. Biomass and Renewable Production Acfs behave relatively similarly: they both slightly decrease as the number of lag increases. As the ACF declines slowly for these two series, it suggests that there is a strong trend in the data, and both of these two series are non-stationary. > However, the acf of Hydropower consumption fluctuates between negative and positive, which potentially reflects seasonal pattern. Also, since the ACF Stays above 0 for many lags, that could imply strong persistence (autoregressive process).

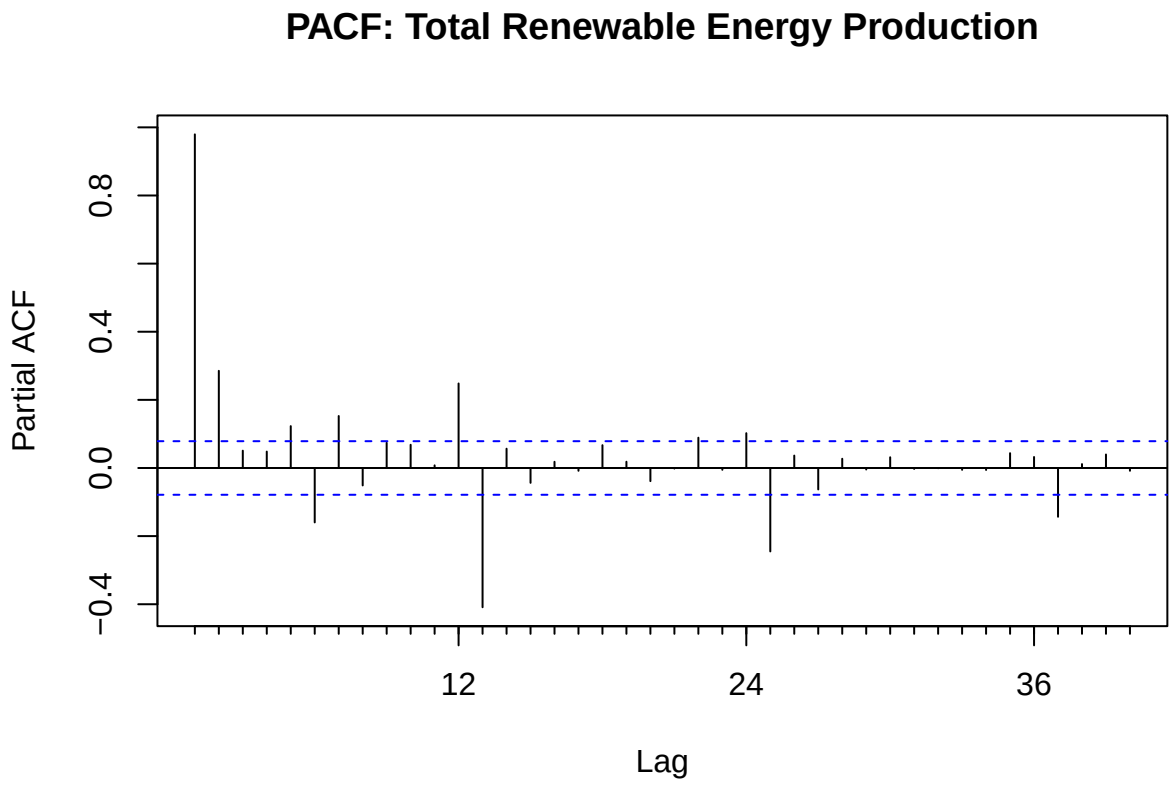
Question 7

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

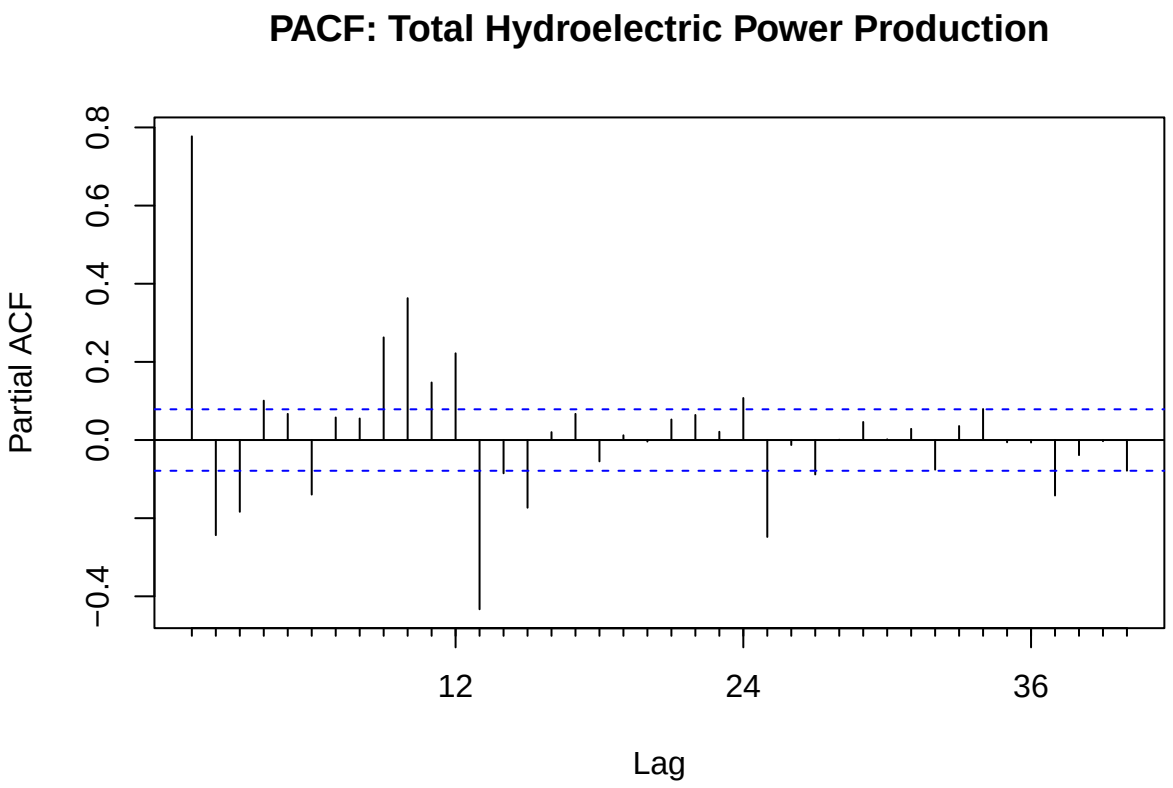
```
biomass_pacf = Pacf(biomass_ts, lag.max = 40,
                main = "PACF: Total Biomass Energy Production")
```




```
renewable_pacf = Pacf(renewable_ts, lag.max = 40,
                      main = "PACF: Total Renewable Energy Production")
```




```
hydro_pacf = Pacf(hydro_ts, lag.max = 40,
                  main = "PACF: Total Hydroelectric Power Production")
```



> Answer: Unlike the previous part, the graph of Pacf of all these three series now show less significant correlation between different time periods. Meanwhile, the partial autocorrelation values tend to be close to zero after lag 1 or 2, it means past values don't strongly impact future values beyond short periods.