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

Assignment 2 - Due date 01/28/25

Yuqi Yang

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(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(openxlsx)
```

Data set information

Consider the data provided in the spreadsheet “Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sourc 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
raw_energy_data <- read.xlsx(xlsxFile = "../Data/Table_10.1_Renewable_Energy_Production_and_Consumption
                             sheet = "Monthly Data",
                             startRow = 13,
                             colNames = FALSE)

read_col_names <- read.xlsx(xlsxFile = "../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_
                            sheet = "Monthly Data",
                            rows = 11,
                            colNames = FALSE)

colnames(raw_energy_data) <- read_col_names
```

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
#Select only three columns of data
energy_data <- raw_energy_data[,4:6]

#Verifying data
head(energy_data)
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                           219.839
## 2                         117.338                           197.330
## 3                         129.938                           218.686
## 4                         125.636                           209.330
## 5                         129.834                           215.982
## 6                         125.611                           208.249
##   Hydroelectric Power Consumption
## 1                          89.562
## 2                          79.544
## 3                          88.284
## 4                          83.152
## 5                          85.643
## 6                          82.060
```

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().

```
#Transform to time series object
ts_energy_data <- ts(energy_data,start = c(1973,1),frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(ts_energy_data[,1],na.rm = TRUE)
```

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

```
sd(ts_energy_data[,1],na.rm = TRUE)
```

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

```
mean(ts_energy_data[,2],na.rm = TRUE)
```

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

```
sd(ts_energy_data[,2],na.rm = TRUE)
```

```
## [1] 143.7927
```

```
mean(ts_energy_data[,3],na.rm = TRUE)
```

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

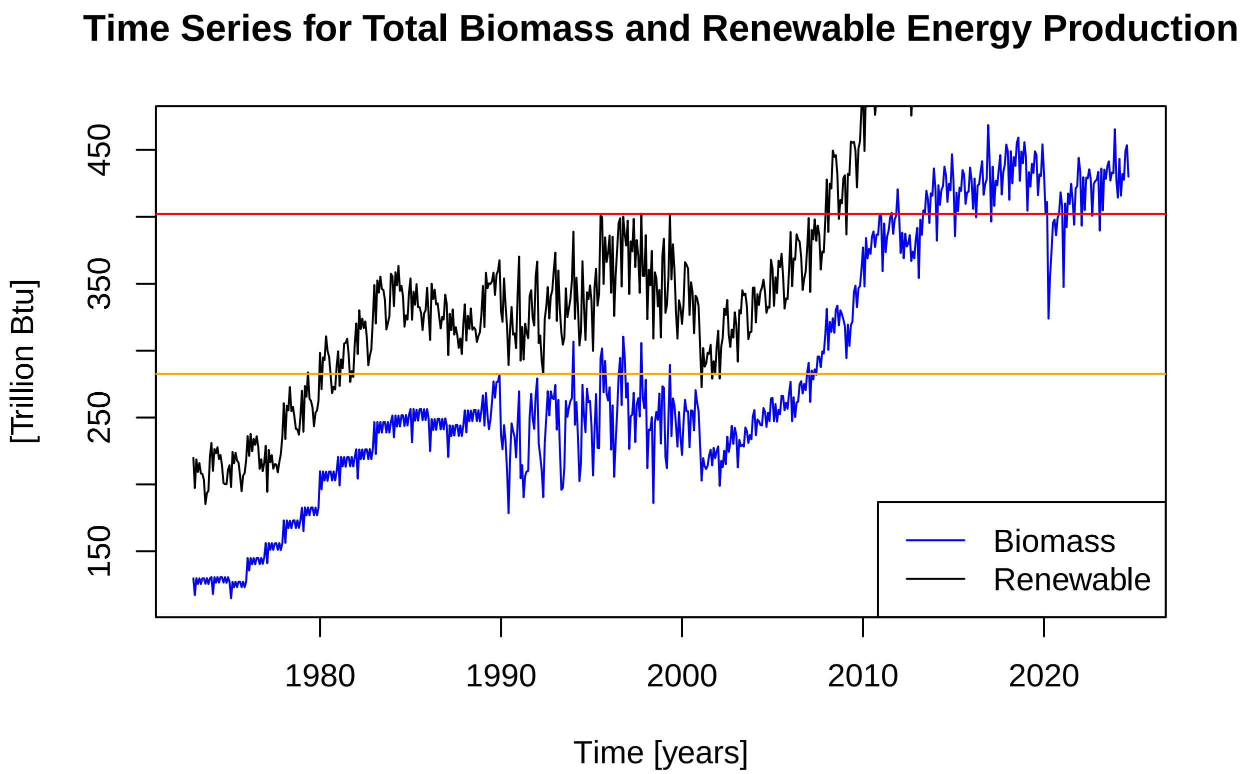
```
sd(ts_energy_data[,3],na.rm = TRUE)
```

```
## [1] 14.10737
```

Question 4

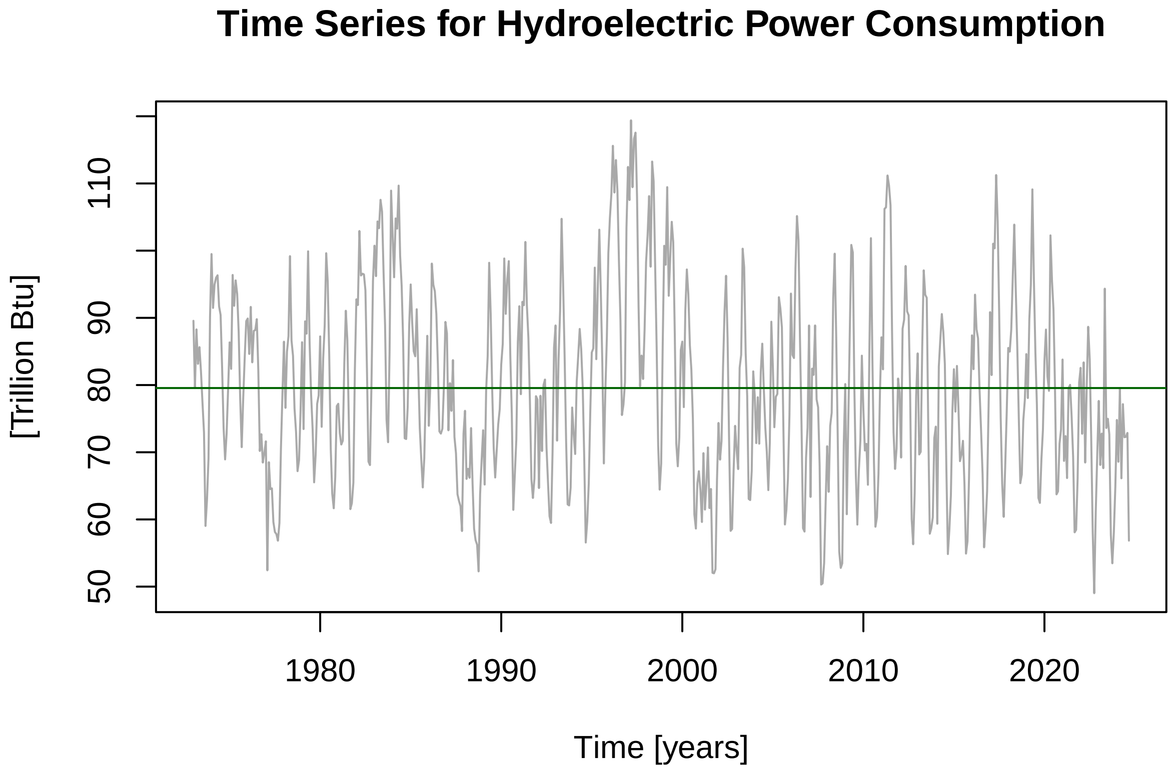
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
#Plot the series for Total Biomass and Renewable Energy Production
plot(ts_energy_data[,"Total Biomass Energy Production"],
     type="l",
     col="blue",
     xlab="Time [years]",
     ylab="[Trillion Btu]")

lines(ts_energy_data[,"Total Renewable Energy Production"],
      col="black")

title(main="Time Series for Total Biomass and Renewable Energy Production")

legend("bottomright",
       legend=c("Biomass","Renewable"),
       lty=c("solid","solid"),
       col=c("blue","black"))

#Adding the mean value lines
abline(h=mean(ts_energy_data[,"Total Biomass Energy Production"],
              na.rm = TRUE),col="orange")
abline(h=mean(ts_energy_data[,"Total Renewable Energy Production"],
              na.rm = TRUE),col="red")
```



```
#Plot the series for Hydroelectric Power Consumption
plot(ts_energy_data[,"Hydroelectric Power Consumption"],
     type="l",
     col="darkgrey",
     xlab="Time [years]",
     ylab="[Trillion Btu]",
     main="Time Series for Hydroelectric Power Consumption")
```

```
#Adding the mean value line
abline(h=mean(ts_energy_data[,"Hydroelectric Power Consumption"],
              na.rm = TRUE), col="darkgreen")
```



Total Biomass Energy Production
Overall, the data shows a steady upward trend, with significant and sustained growth, especially from 2000 to 2020. Between 1990 and 2000, the fluctuations are more significant, and the mean value is close to flat, showing some seasonal or irregular fluctuations, while in 2020 there was a significant decline, but then returned to the original level.

Total Renewable Energy Production
The trend is similar to that of Total Biomass Energy Production, but with consistently higher and more fluctuating values. The upward trend in production was stronger and more pronounced after 2010, possibly driven by policy or technological advances.

Hydroelectric Power Consumption
It shows seasonal fluctuations and remains generally stable with no significant upward or downward trend. Peaks exceed 110 Trillion Btu, troughs are below 50 Trillion Btu, and the average is close to 80 Trillion Btu. Significant troughs occur around 1978, 1989, and 2002, which may be related to weather conditions or policy changes.

Question 5

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

```
#Compute the correlation between these three series
cor_energy_data <- cor(ts_energy_data,use = "complete.obs")

#Test significance of correlation coefficients
cor_test_B_vs_R <- cor.test(ts_energy_data[, 1], ts_energy_data[, 2])
cor_test_B_vs_H <- cor.test(ts_energy_data[, 1], ts_energy_data[, 3])
cor_test_R_vs_H <- cor.test(ts_energy_data[, 2], ts_energy_data[, 3])

#Show results
cor_results <- data.frame(
  Pair = c("Biomass vs Renewable", "Biomass vs Hydroelectric",
           "Renewable vs Hydroelectric"),
  Correlation = c(cor_test_B_vs_R$estimate,
                  cor_test_B_vs_H$estimate,
                  cor_test_R_vs_H$estimate),
  P_Value = c(cor_test_B_vs_R$p.value,
              cor_test_B_vs_H$p.value,
              cor_test_R_vs_H$p.value))

cor_results
```

```
##                         Pair Correlation     P_Value
## 1       Biomass vs Renewable  0.96781371 0.000000000
## 2   Biomass vs Hydroelectric -0.11429266 0.004347652
## 3 Renewable vs Hydroelectric -0.02916103 0.468219410
```

Biomass vs Renewable:A significant strong positive correlation (0.97) as the p value is less than 0.05.

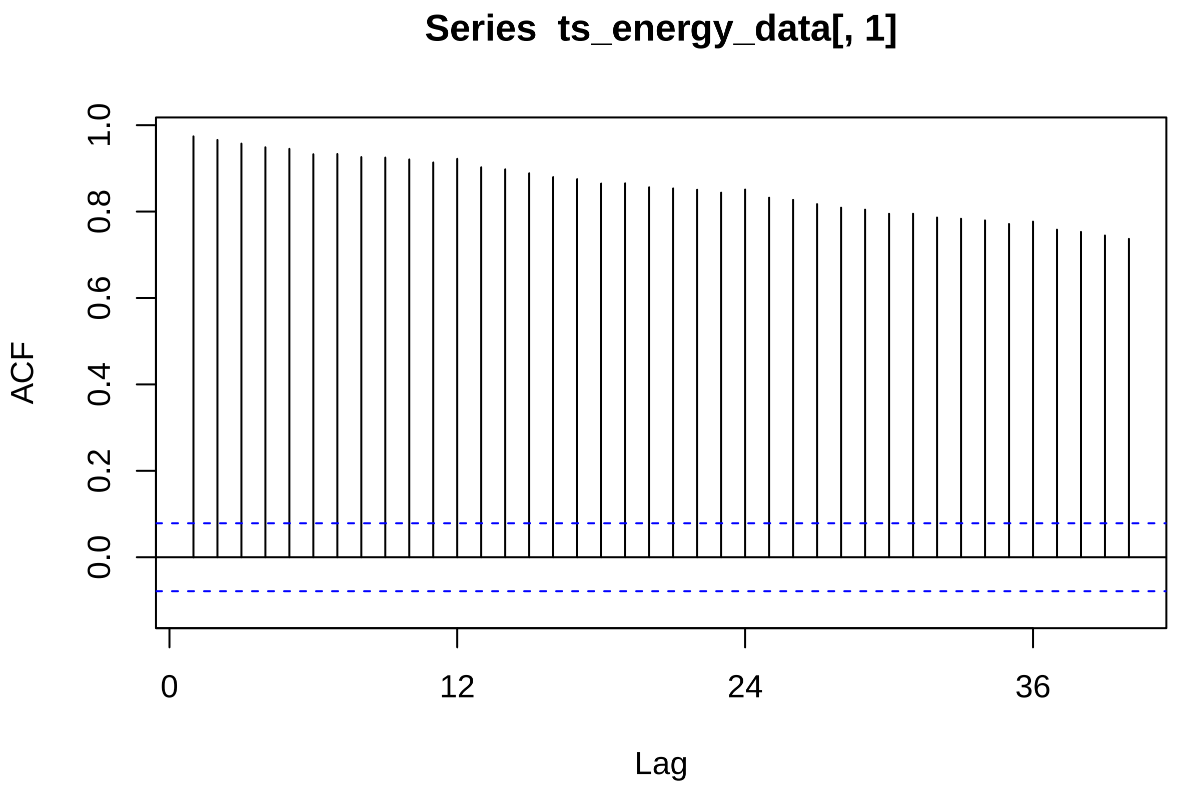
Biomass vs Hydroelectric:There is a weak negative correlation (-0.11) with a p value (0.004) less than 0.05 indicating a significant correlation.

Renewable vs Hydroelectric:The p value (0.47) is greater than 0.05, which is not significant, indicating that there is no meaningful relationship between these two variables.

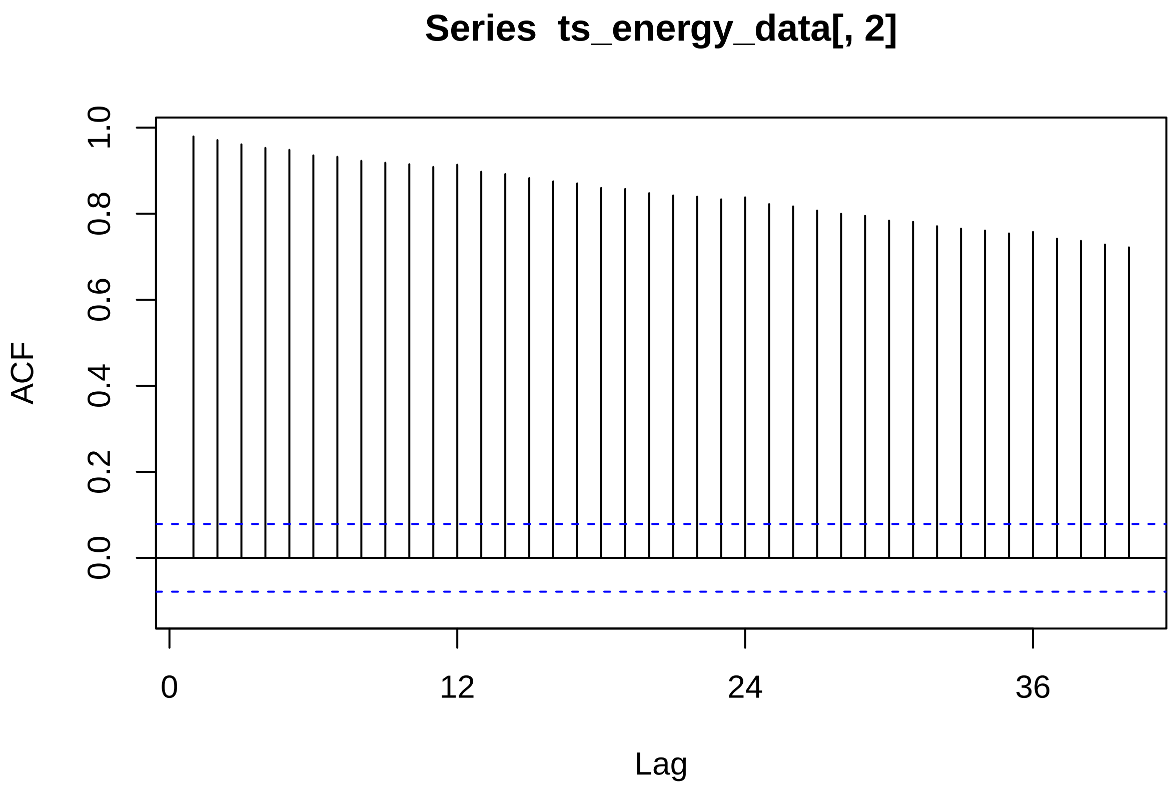
Question 6

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?

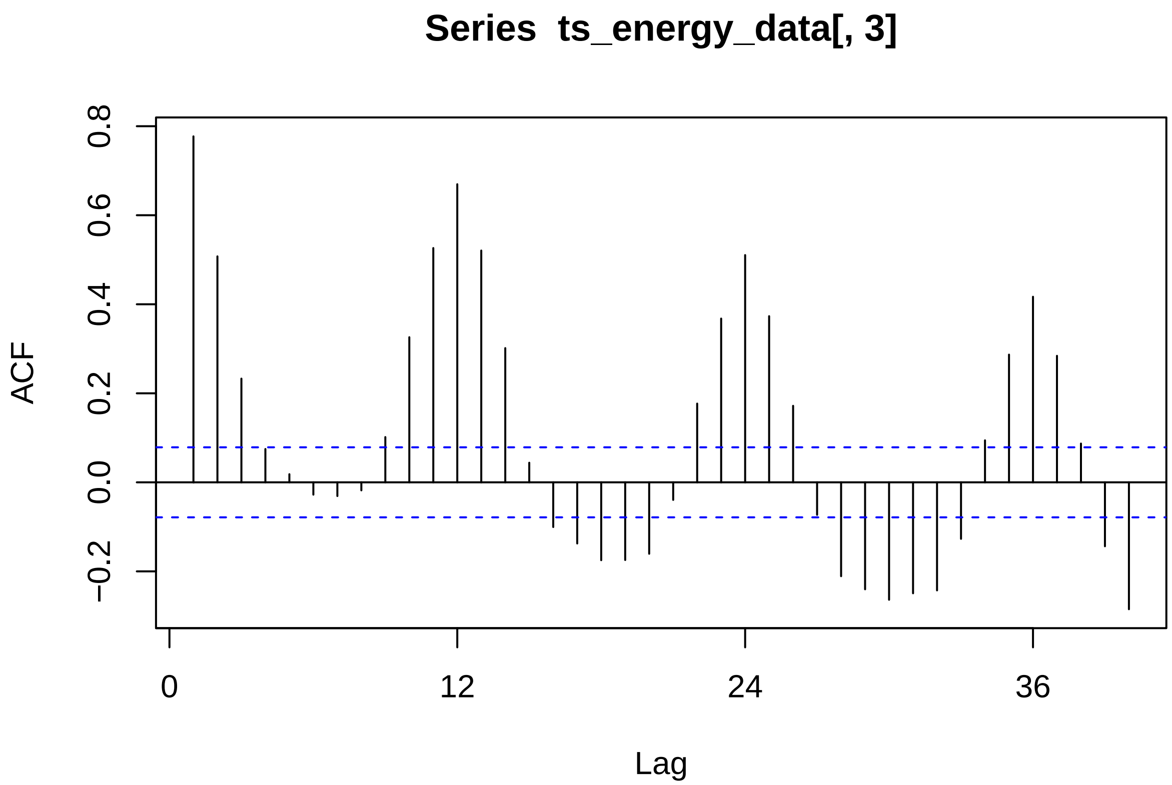
```
Biomass_acf=Acf(ts_energy_data[, 1],lag.max = 40,type = "correlation",plot = TRUE)
```



```
Renewable_acf=Acf(ts_energy_data[, 2],lag.max = 40,type = "correlation",plot = TRUE)
```



```
Hydroelectric_acf=Acf(ts_energy_data[, 3],lag.max = 40,type = "correlation",plot = TRUE)
```

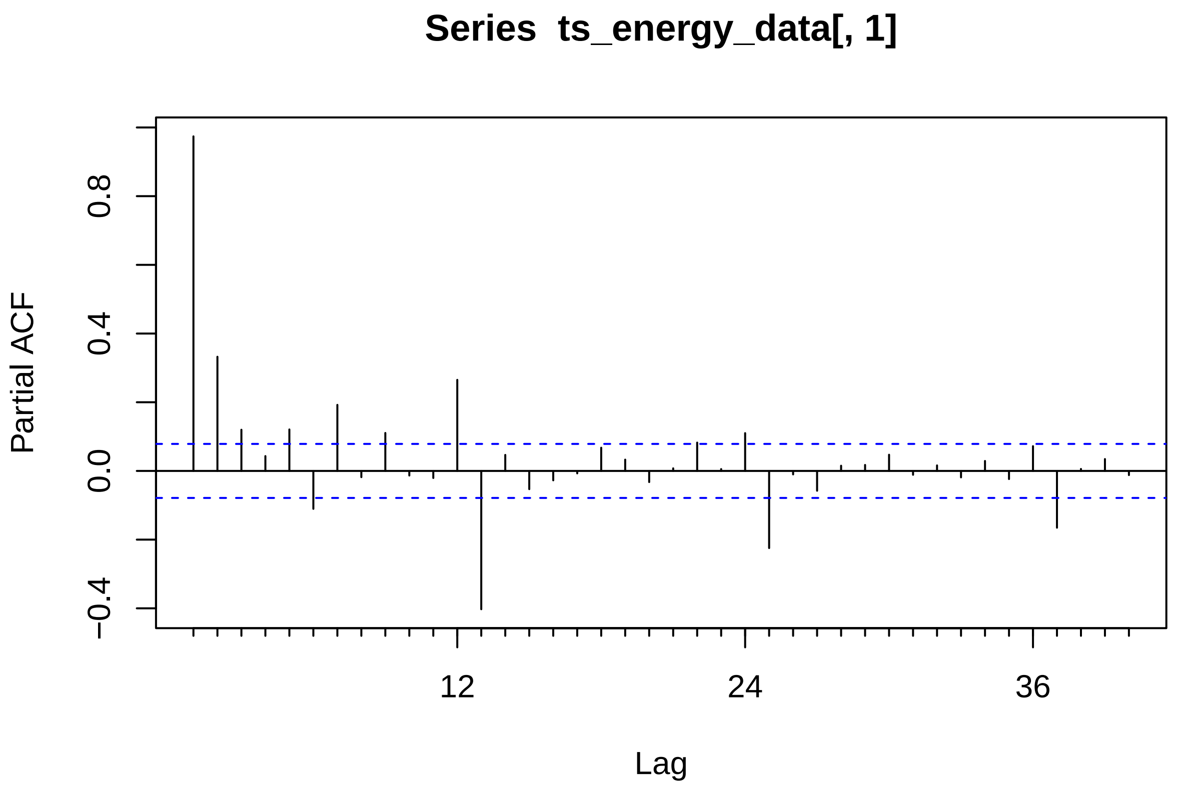


These three variables do not have the same behavior.
The results for **biomass and renewable energy production** are similar in that their ACF decrease very slowly, indicating a strong connection to their past values, with significant trends and strong persistence. While the ACF of **hydroelectric power consumption** has significant periodic fluctuations and shows peaks at lags 12, 24, and 36, indicating that it has a seasonality of a year (12-month) cycle.

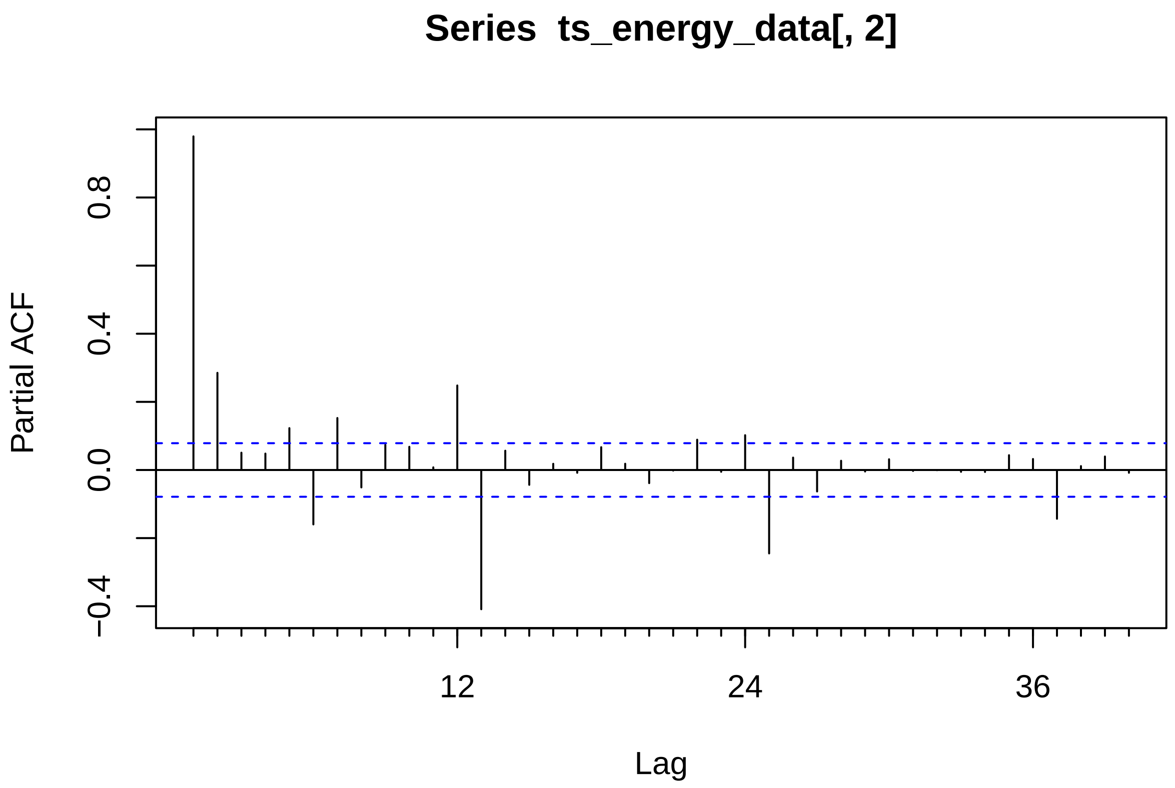
Question 7

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

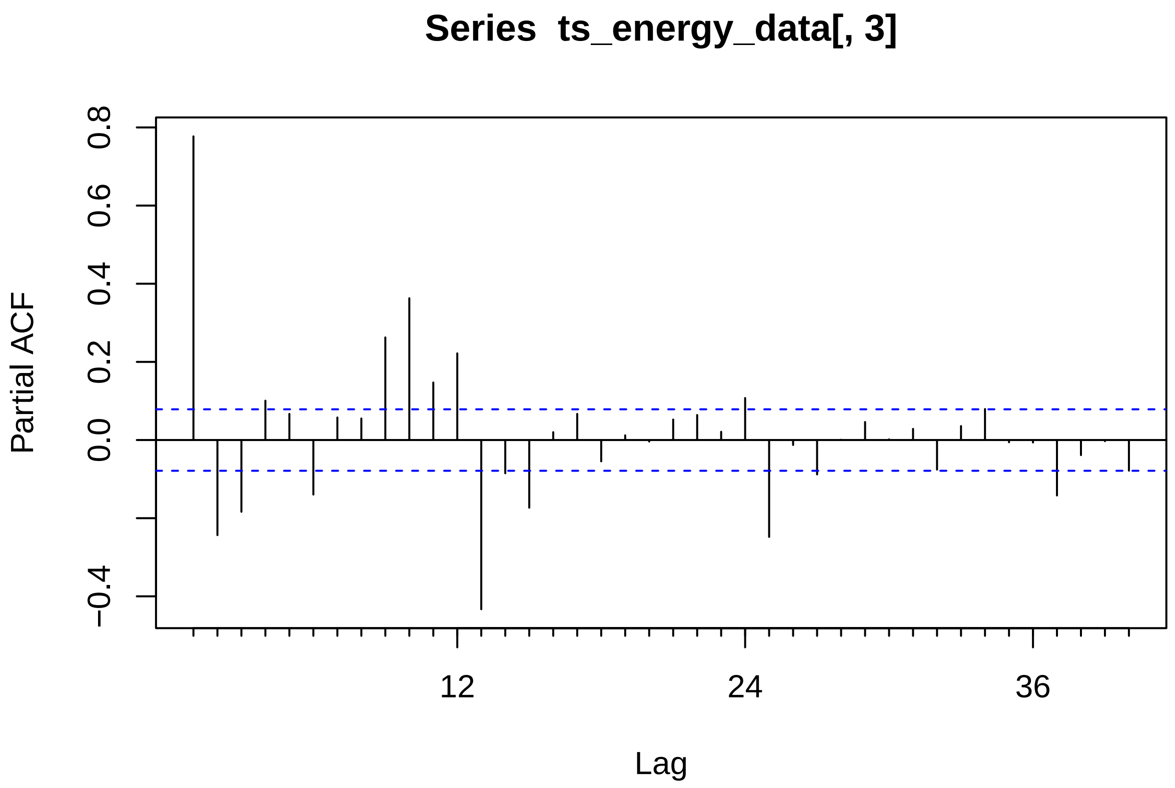
```
Biomass_pacf=Pacf(ts_energy_data[, 1],lag.max = 40,plot = TRUE)
```



```
Renewable_pacf=Pacf(ts_energy_data[, 2],lag.max = 40,plot = TRUE)
```



```
Hydroelectric_pacf=Pacf(ts_energy_data[, 3],lag.max = 40,plot = TRUE)
```



After excluding the effect of intermediate lags, the PACF only indicates direct correlations between current values and specific lags, which is quite different from the Q6 results.

The PACF results for **biomass and renewable energy production** remain similar, with significant correlations only in the first few lags and no obvious seasonal pattern, indicating that the trends in the ACF are due to long-term dependencies rather than direct lags.

The PACF for **hydroelectric power consumption** has fewer significant lags compared to that of ACF, and there is also no clear periodic pattern.