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

Assignment 2 - Due date 01/27/26

Lauren Shohan

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.\

```
#Load/install required package here
library(forecast)
library(tseries)
library(dplyr)

library(lubridate)
library(ggplot2)
```

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.

```
getwd()
```

```
## [1] "/home/guest/TSA_Sp26/Assignments"
```

```r
#Importing data set
library(readxl)
library(openxlsx)

#extract data
energydata <- read_excel(path=
"/home/guest/TSA_Sp26/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`
```

```r
#extract columns
energydata_cols <- read_excel(path=
"/home/guest/TSA_Sp26/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`
```

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
#assign column names to data
colnames(energydata) <- energydata_cols

#select columns i want - columns 4,5,6
energydata_3cols <- energydata[,4:6]

head(energydata_3cols)
```

```
## # 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().

```r
#starting in 1973 and going monthly, thus freq is 12
ts_energydata <- ts(energydata_3cols,start=c(1973,1), frequency=12)
head(ts_energydata)
```

```
##          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.

```r
biomass_mean <- mean(ts_energydata[,1])
biomass_sd <- sd(ts_energydata[,1])

renewable_mean <- mean(ts_energydata[,2])
renewable_sd <- sd(ts_energydata[,2])

hydro_mean <- mean(ts_energydata[,3])
hydro_sd <- sd(ts_energydata[,3])

cat('Biomass Mean and SD =',biomass_mean, 'and',biomass_sd, '\n')
```

```
## Biomass Mean and SD = 286.0489 and 96.21209
```

```r
cat('Renewable Energy Mean and SD =',renewable_mean, 'and',renewable_sd, '\n')
```

```
## Renewable Energy Mean and SD = 409.1952 and 151.4223
```

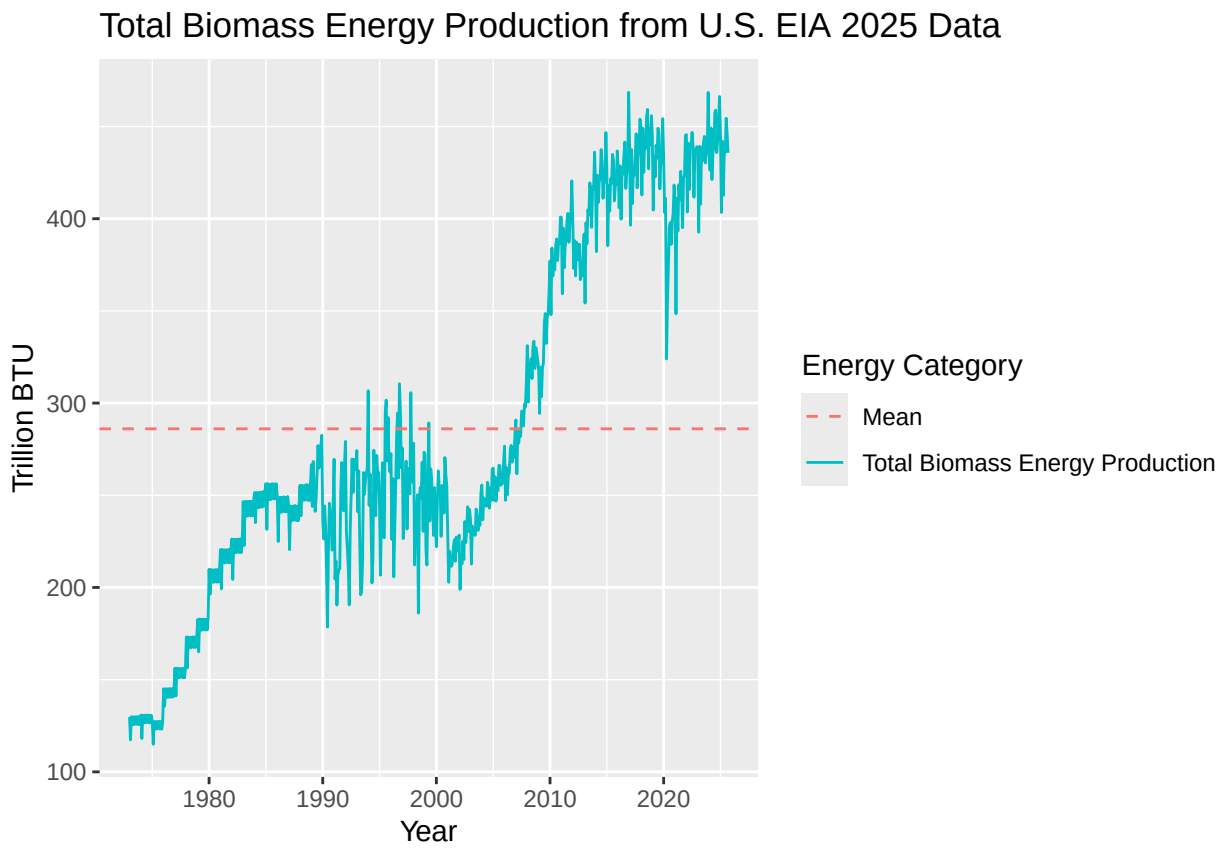
```r
cat('Hydroelectric Mean and SD =',hydro_mean, 'and',hydro_sd)
```

```
## Hydroelectric Mean and SD = 79.35682 and 14.1202
```

Question 4

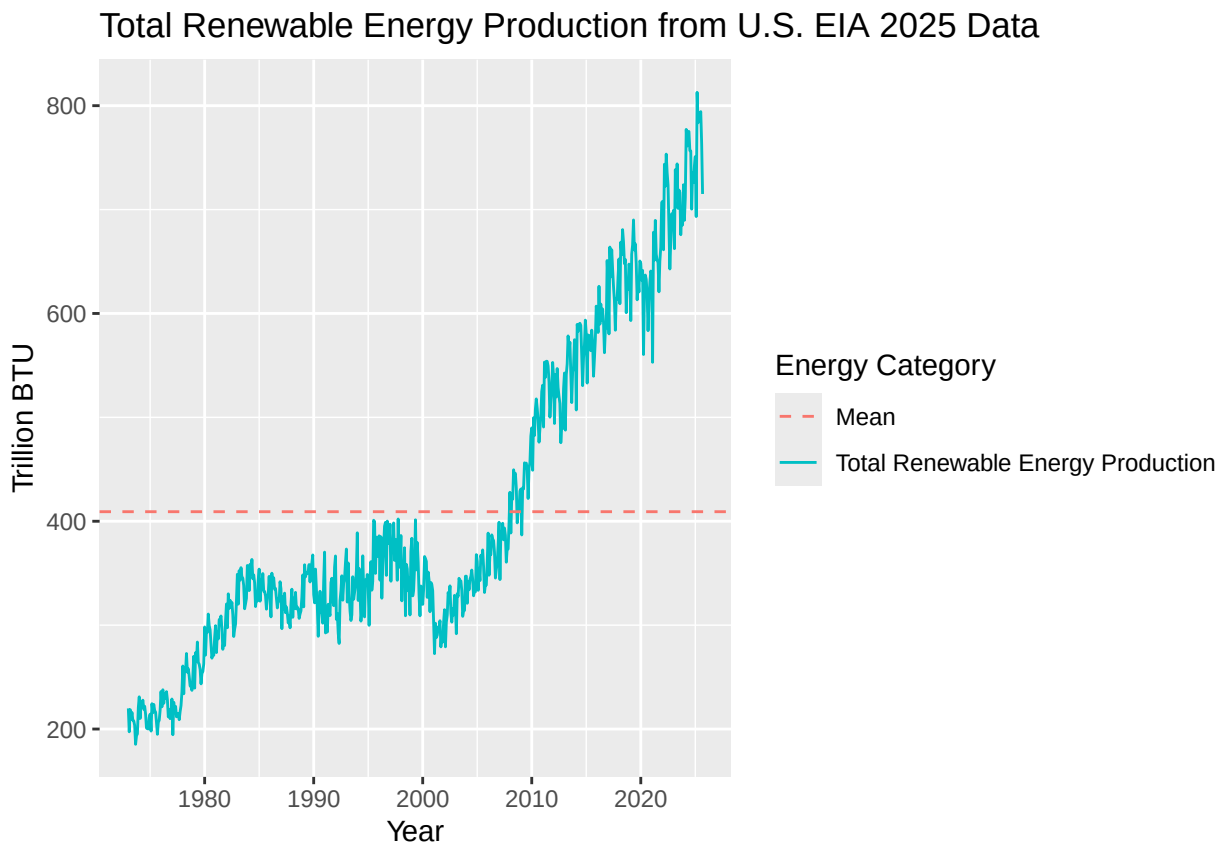
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```r
#biomass
autoplot(ts_energydata[,1], series = 'Total Biomass Energy Production') +
  geom_hline(aes(yintercept = biomass_mean, color = 'Mean'), linetype = 'dashed') +
  xlab("Year") +
  ylab("Trillion BTU") +
  labs(color="Energy Category") +
  ggtitle('Total Biomass Energy Production from U.S. EIA 2025 Data')
```



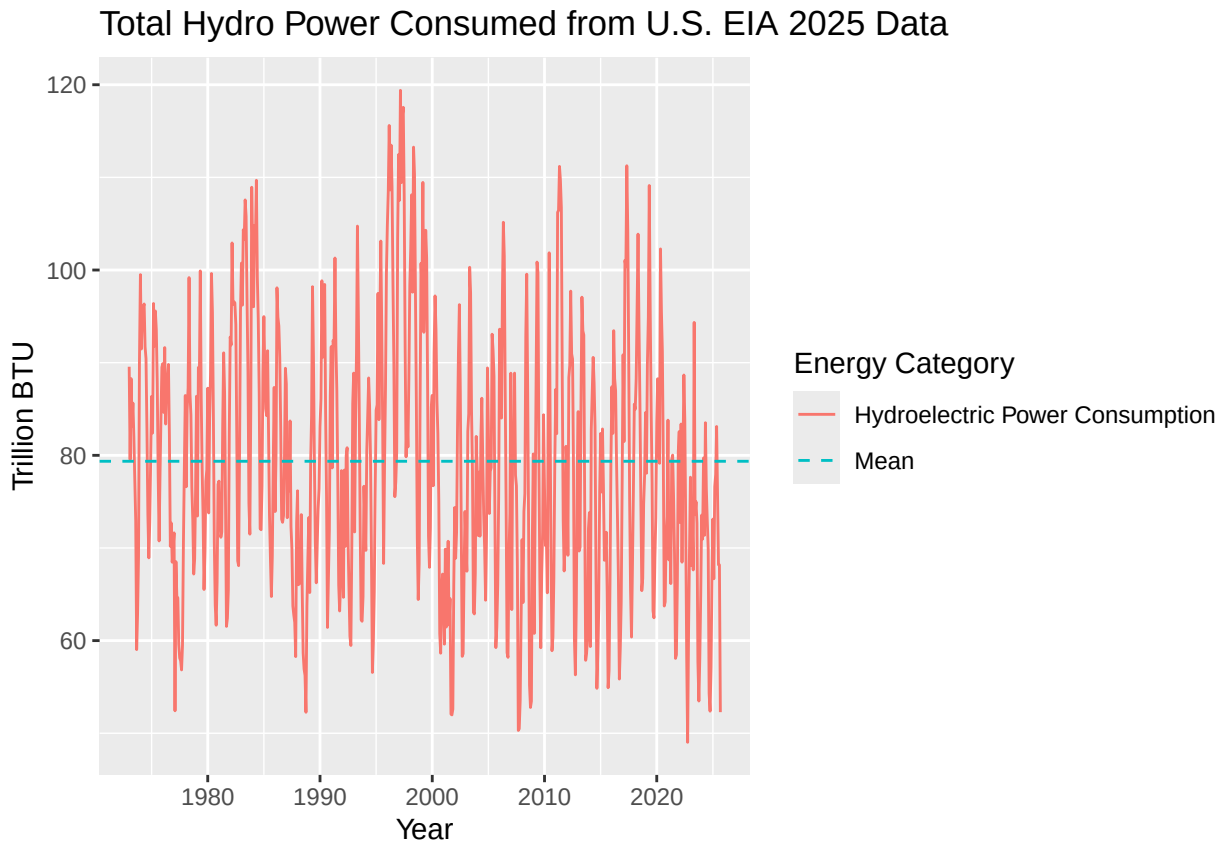
This ts plot of total biomas energy production illustrates an increasing trend over the decades, with a stagnant period between 1990-2000 that varied quite alot between 200-300 trillion bto, but then slowy climbed from 2000 onward all the way up to near 500 trillion.

```
#renewable energy
autoplot(ts_energydata[,2], series = 'Total Renewable Energy Production') +
  geom_hline(aes(yintercept = renewable_mean, color = 'Mean'), linetype = 'dashed') +
  xlab("Year") +
  ylab("Trillion BTU") +
  labs(color="Energy Category") +
  ggtitle('Total Renewable Energy Production from U.S. EIA 2025 Data')
```



This graph shows a similar story as the biomass ts graph, a slow steady increase with an odd period between 1990-2000 that caused a slight dip post-2000 but then rapidly increased from around 300 trillion all the way to 800 trillion. This one saw a much sharper increase.

```
#hydro
autoplot(ts_energydata[,3], series = 'Hydroelectric Power Consumption') +
  geom_hline(aes(yintercept = hydro_mean, color = 'Mean'), linetype = 'dashed') +
  xlab("Year") +
  ylab("Trillion BTU") +
  labs(color="Energy Category") +
  ggtitle('Total Hydro Power Consumed from U.S. EIA 2025 Data')
```



This ts graph is much different than the other two, the hydroelectric ts graph showcases a steady up and down pattern that looks seasonal. During the period 1990-2000 there is the largest spike up to 120 trillion, but besides that it seems to stay steady beween 50/60 trillion for low points and around 100/110 trillion for high points.

Question 5

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

```
cor(ts_energydata)
```

```
##                                    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
```

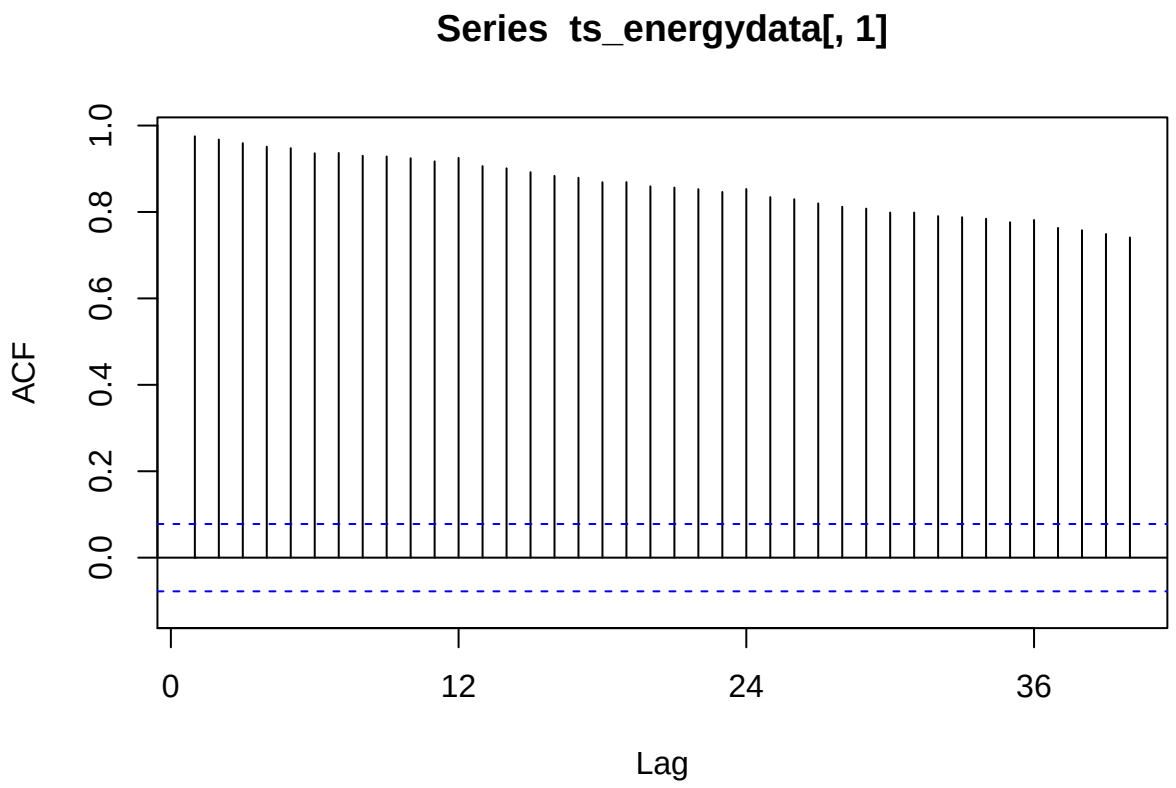
This data shows a weak, negative correlation between biomass and hydroelectric with a correlation = -0.134 and a weak, negative correlation between total renewable energy and hydroelectric with a -0.058 correlation.

However, it shows a very strong and positive correlation between biomass and total renewable power at 0.965, indicating a strong linear dependence.

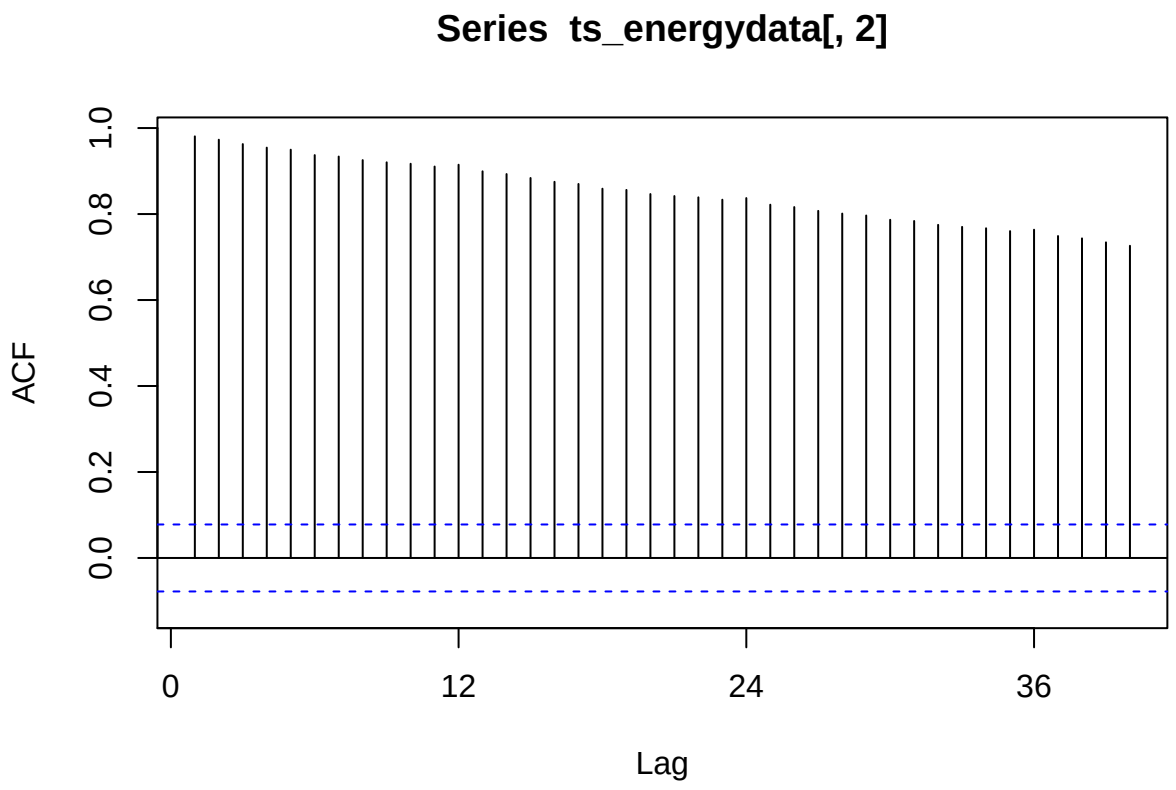
Question 6

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?

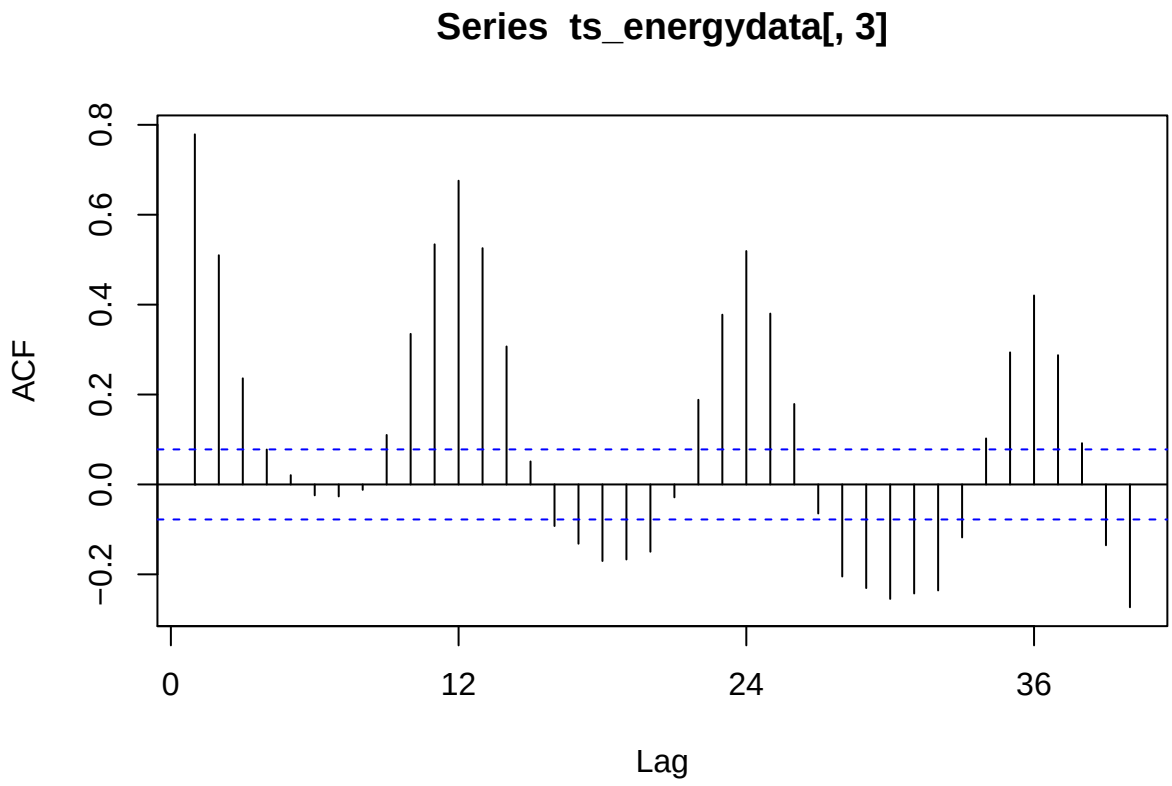
```
biomass_acf = Acf(ts_energydata[,1], lag.max = 40, plot = TRUE)
```




```
totalrenew_acf = Acf(ts_energydata[,2], lag.max = 40, plot = TRUE)
```




```
hydro_acf = Acf(ts_energydata[,3], lag.max = 40, plot = TRUE)
```

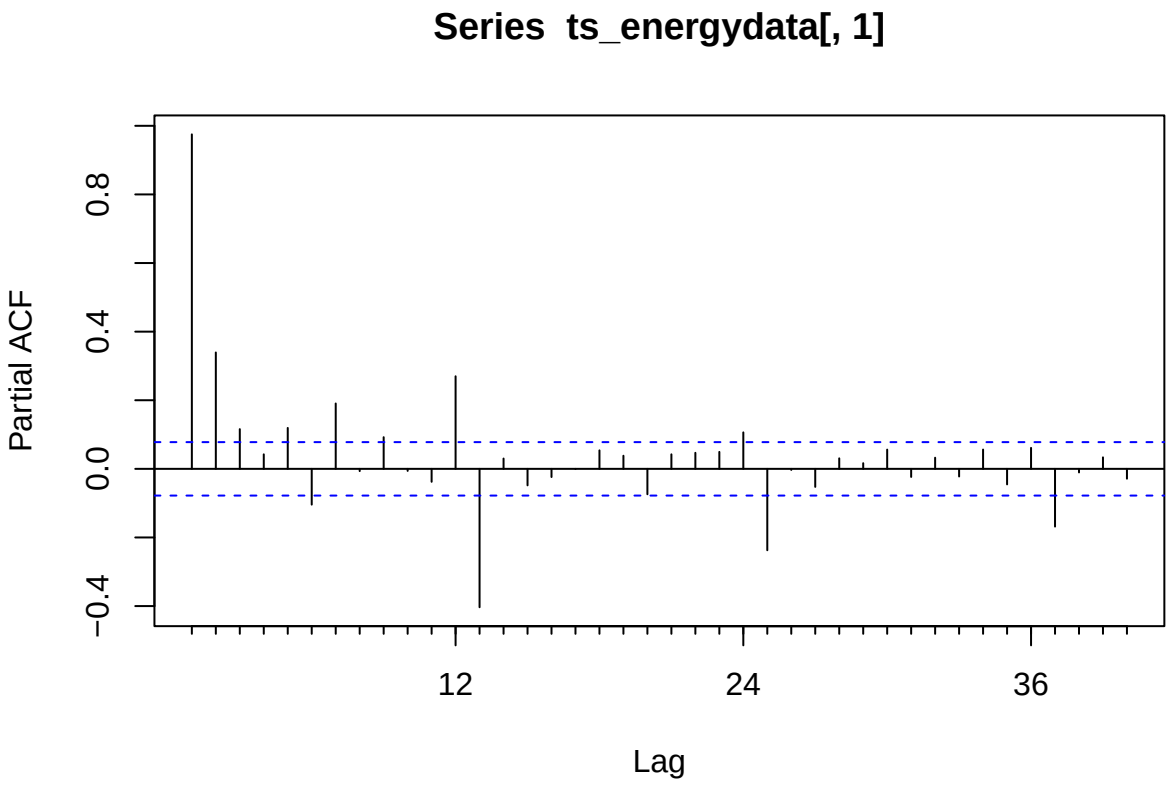


Biomass and total renewable have almost identical ACF graphs, with the biomass one having slightly taller bars, but both following the same general downward trend over a certain amount of lags. This slow decay likely indicates strong persistence or memory of the series. However, the ACF graph for hydroelectric is what is to be expected for hydro data and goes up and down, following a seasonal trend of negative lags in what I can assume to be warmer months (less water) and large positive lags during the colder months. However, the negative ACFs are getting more negative with time, while the positive ACFs are decreasing as well.

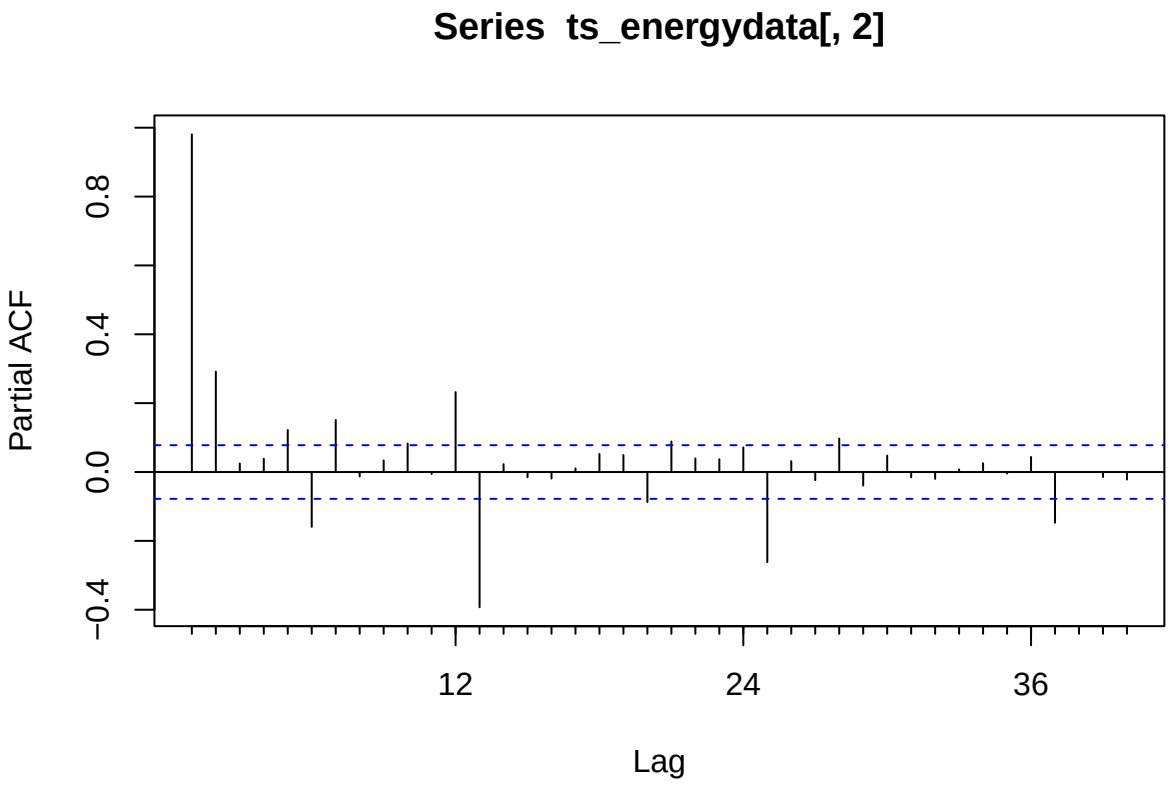
Question 7

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

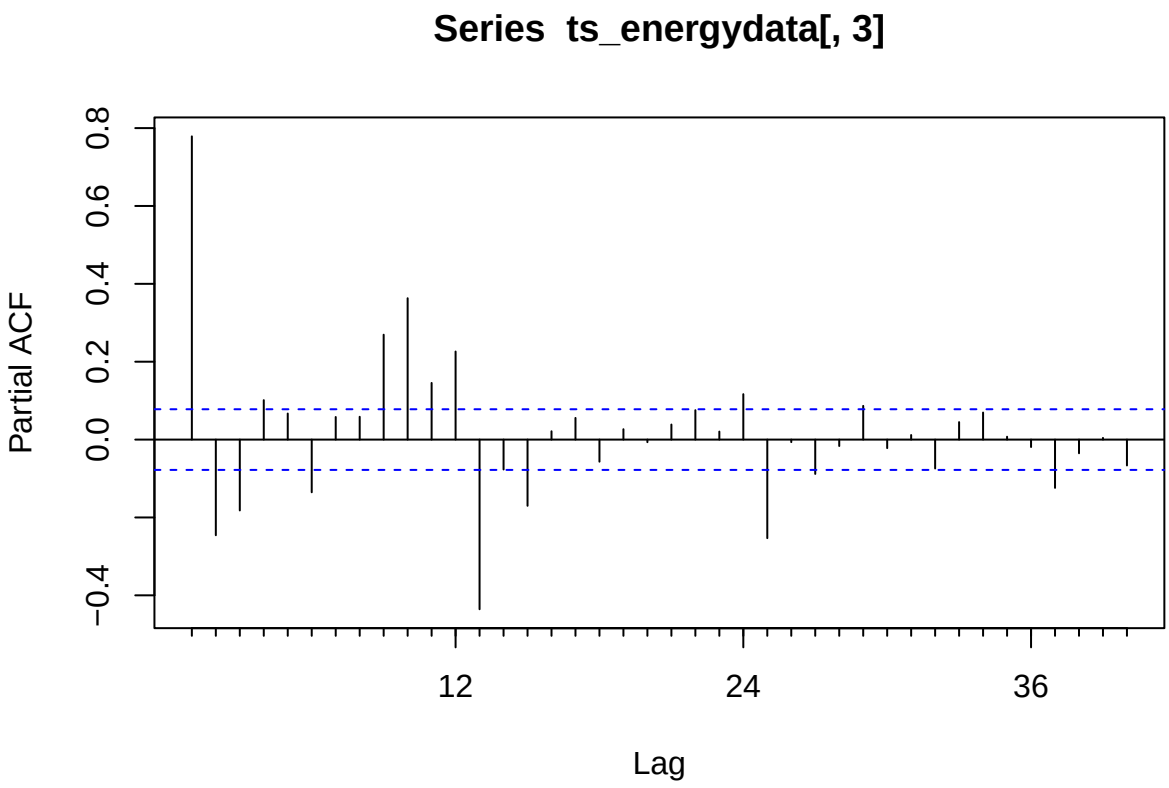
```
biomass_pacf = Pacf(ts_energydata[,1], lag.max = 40, plot = TRUE)
```




```
totalrenew_pacf = Pacf(ts_energydata[,2], lag.max = 40, plot = TRUE)
```




```
hydro_pacf = Pacf(ts_energydata[,3], lag.max = 40, plot = TRUE)
```



The PACF graphs are very different than the ACF graphs from the previous question. Once the influence of the intermediate variables were removed, the direct correlation was not as visibly strong. The first two look like completely different graphs that now contain some negative PACFs and doesn't show a visible downward trend any longer. The hydro graph has a similar up and down trend for at least the first 14 or so lags, but is lost after. All three have lost the majority of their bars and length.