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

Assignment 2 - Due date 01/23/25

Chloe Young

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(c('readxl', 'openxlsx', 'forecast', 'tseries', 'dplyr'))

library(readxl)
library(openxlsx)
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
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
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 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 set

#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`
```

```
#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`
```

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

```
energy_data1 <- energy_data1[ ,4:6] #the space before the comma means you want all rows
                                    #and 4:6 means all columns from 4 to 6

#Visualize the first rows of the data set
head(energy_data1)
```

```
## # 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().

```
ts_energy_data1 <- ts(energy_data1, start=c(1973,1),frequency=12)
```

Question 3

Compute mean and standard deviation for these three series.

```
#Calculating the means for each column
mean_biomass <- mean(ts_energy_data1[, 1], na.rm = TRUE)
mean_renewable <- mean(ts_energy_data1[, 2], na.rm = TRUE)
mean_hydroelectric <- mean(ts_energy_data1[, 3], na.rm = TRUE)

print(mean_biomass)
```

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

```
print(mean_renewable)
```

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

```
print(mean_hydroelectric)
```

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

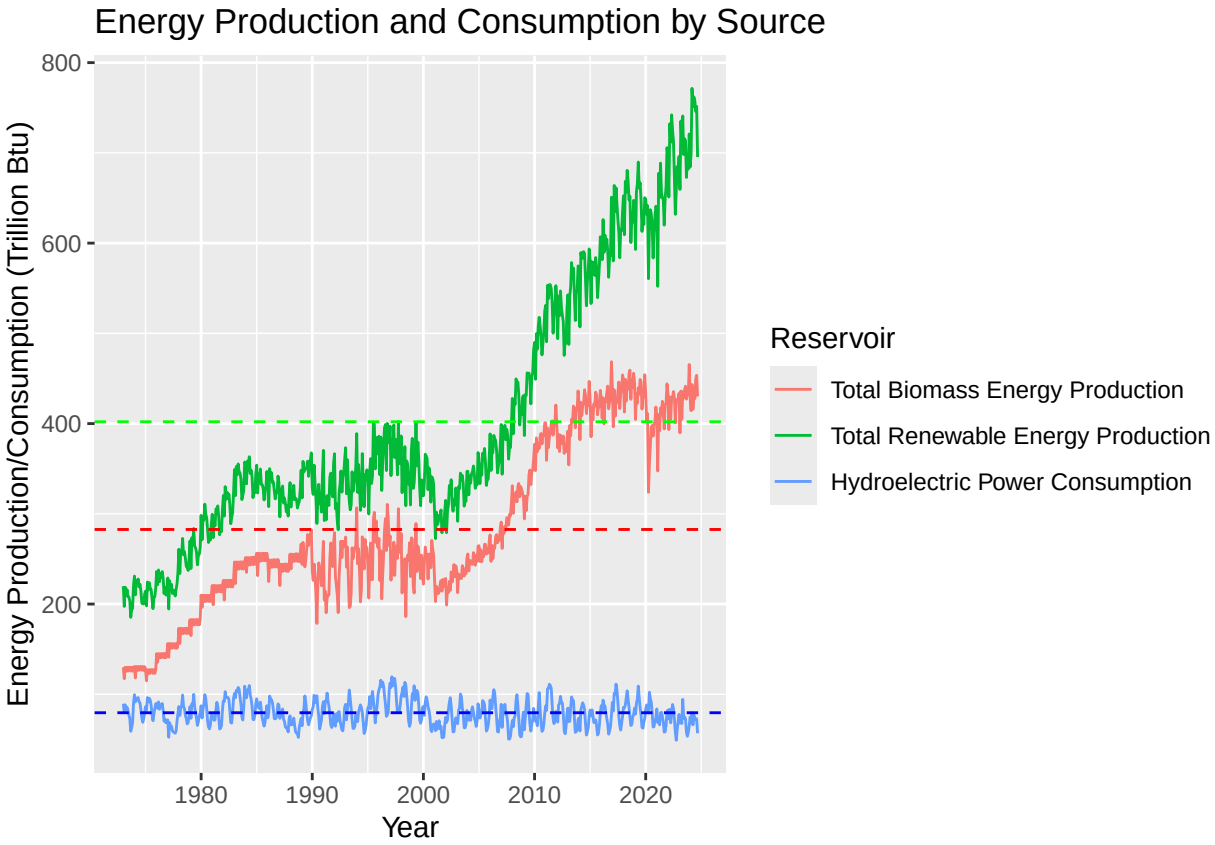
```
#Calculating the standard deviations for each column
apply(ts_energy_data1, 2, sd, na.rm = TRUE)
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                           94.05815                           143.79270
##   Hydroelectric Power Consumption
##                           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.

```
autoplot(ts_energy_data1) +
  ggtitle("Energy Production and Consumption by Source") +
  xlab("Year") +
  ylab("Energy Production/Consumption (Trillion Btu)") +
  labs(color="Reservoir") +
  geom_hline(yintercept = mean_biomass, linetype = "dashed", color = "red") +
  geom_hline(yintercept = mean_renewable, linetype = "dashed", color = "green") +
  geom_hline(yintercept = mean_hydroelectric, linetype = "dashed", color = "blue")
```



This plot shows that Total Biomass Production and Total Renewable Production have increased overtime, with renewable energy having increased the most given that it has the highest end point and highest mean. Renewable energy production and biomass production follow a relatively similar pattern, indicating that the two may be correlated. Hydroelectric power consumption follows a seemingly different pattern to the other two as it stays relatively constant over the years. It also has the lowest mean indicating that hydroelectric consumption is a less utilized source of energy compared to biomass and renewables.

Question 5

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

```
cor(ts_energy_data1)
```

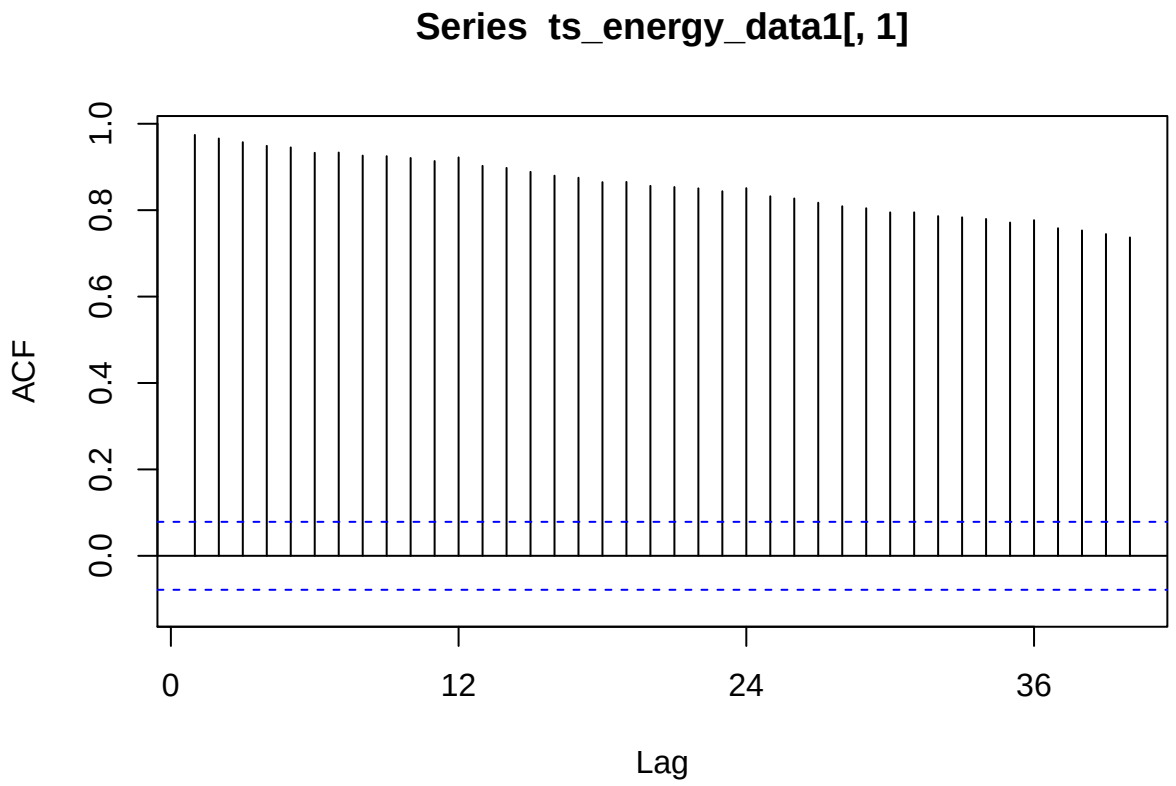
```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.9678137
## Hydroelectric Power Consumption                       -0.1142927
##                                  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
```

Biomass Production and Renewable Energy Production are significantly correlated as the correlation value is close to 1 at 0.97, meaning there is a positive correlation. Hydroelectric Power and Biomass Production are very slightly negatively correlated with a value of -0.11, while hydroelectric power and renewable energy production are nearly not correlated at all with a very slight negative correlation of -0.029. However, given how close this value is to 0, it is reasonable to state that there is no significant correlation between them.

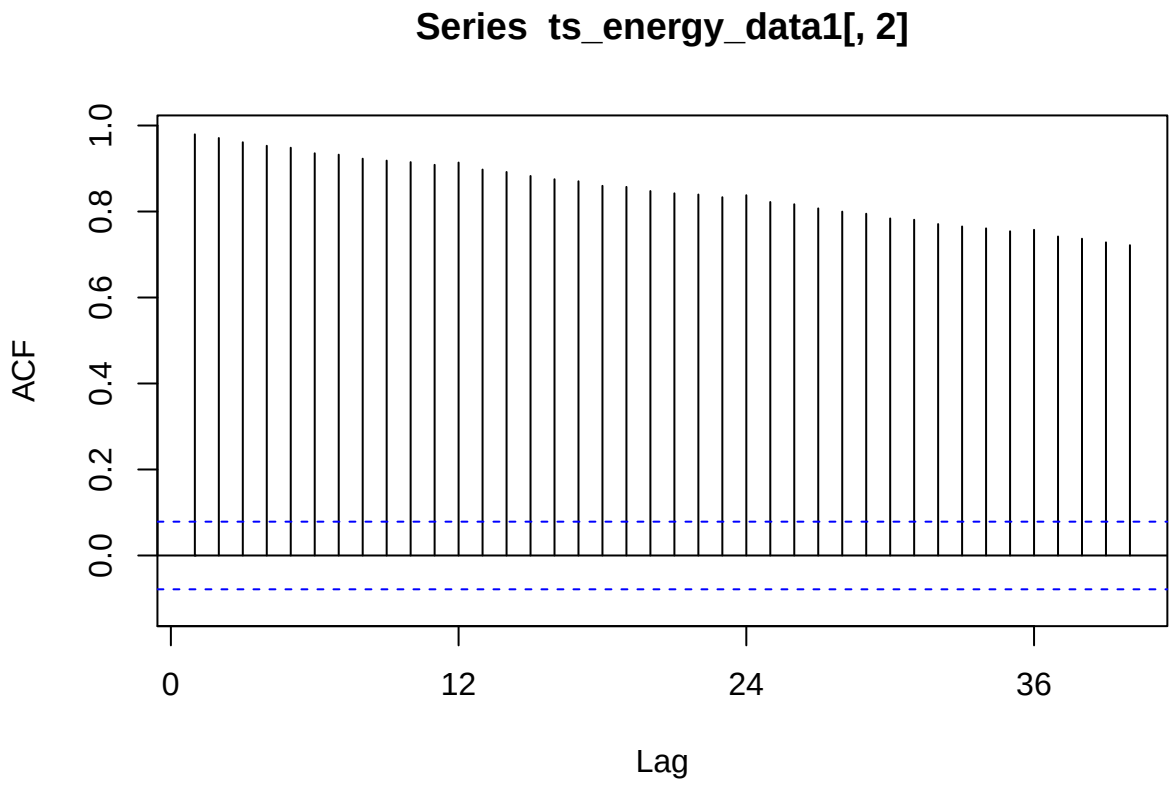
Question 6

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?

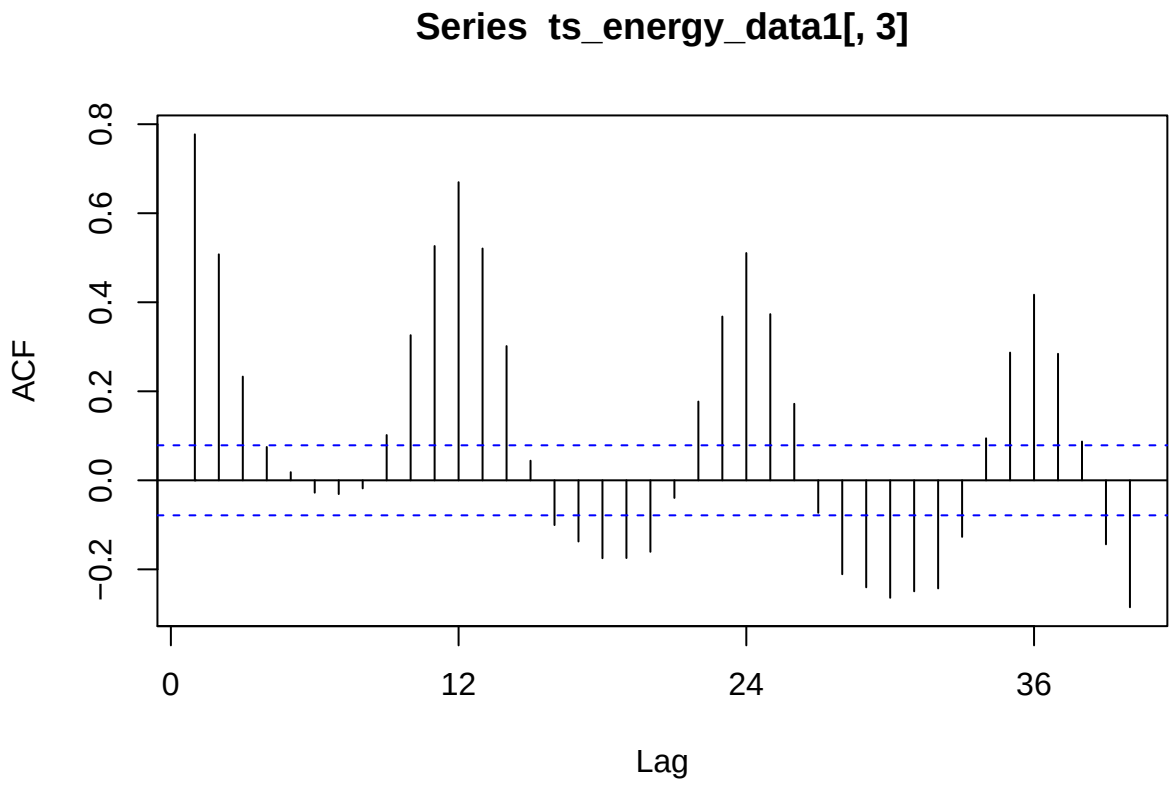
```
Biomass_acf=Acf(ts_energy_data1[,1],lag=40)
```




```
Renewable_acf=Acf(ts_energy_data1[,2],lag=40)
```




```
Hydroelectric_acf=Acf(ts_energy_data1[,3],lag=40)
```



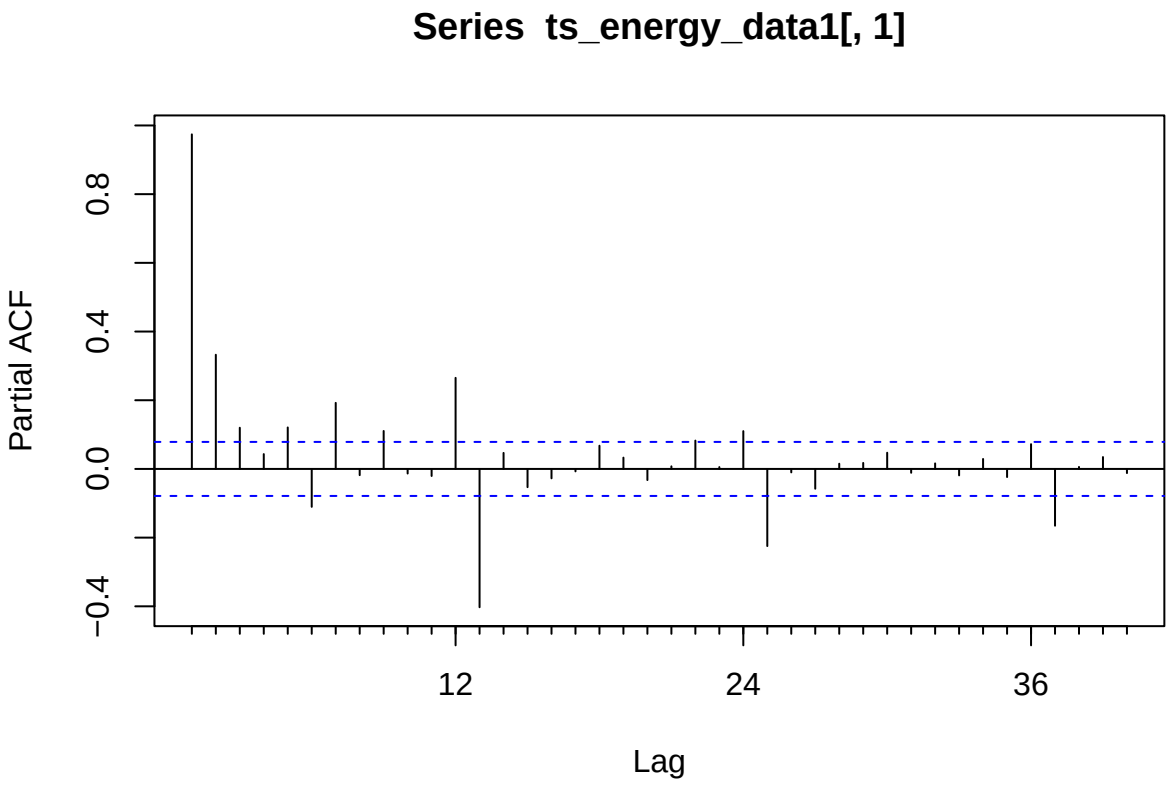
The three plots don't have the same behavior. For biomass and renewable production, each value is statistically significant meaning that there is autocorrelation. There is no seasonal variation in the data, rather there is a gradual decline which indicates long term dependency in the data. For hydroelectric consumption, however, there appears to be seasonal variation in the data as the ACF spikes every 12 months. This means that hydroelectric consumption is seasonal, which makes sense given that water availability from precipitation varies in different seasons. Biomass and renewable production, one the other hand, don't vary with seasons.

Source: https://medium.com/@kis.andras.nandor/understanding-autocorrelation-and-partial-autocorrelation-functions-acf-and-pacf-2998e7e1bcb5

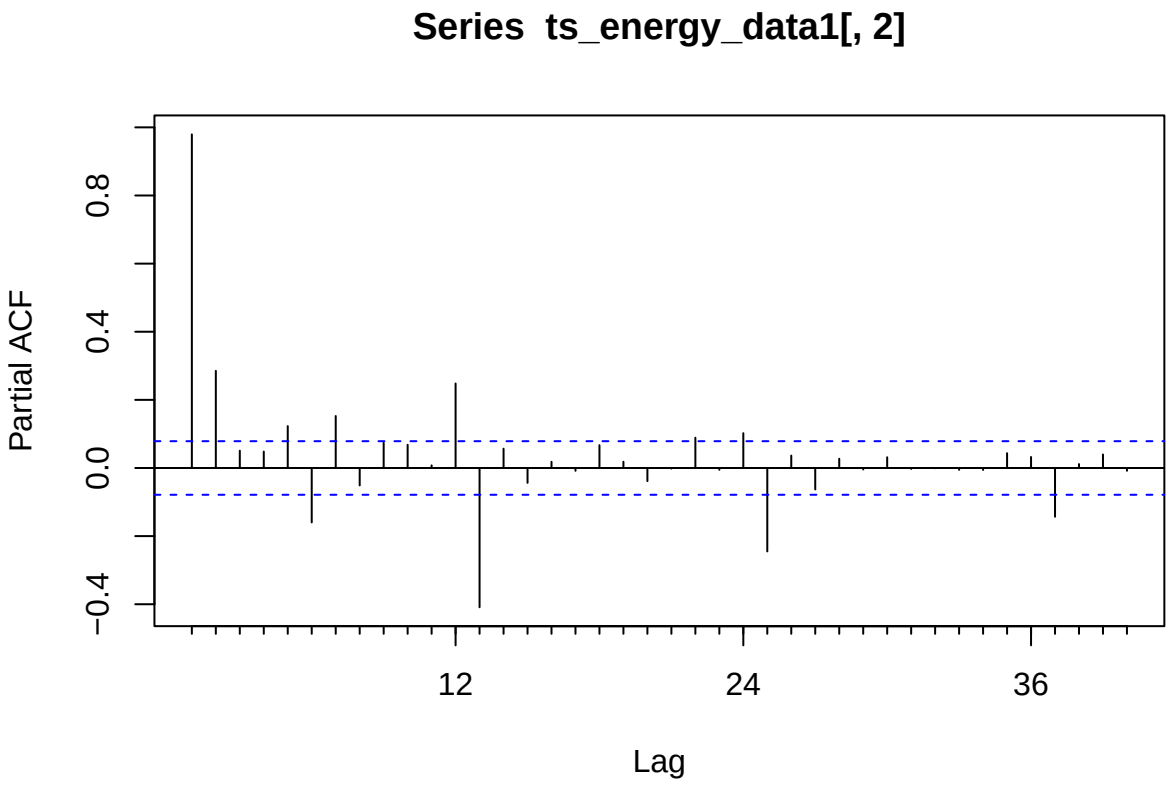
Question 7

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

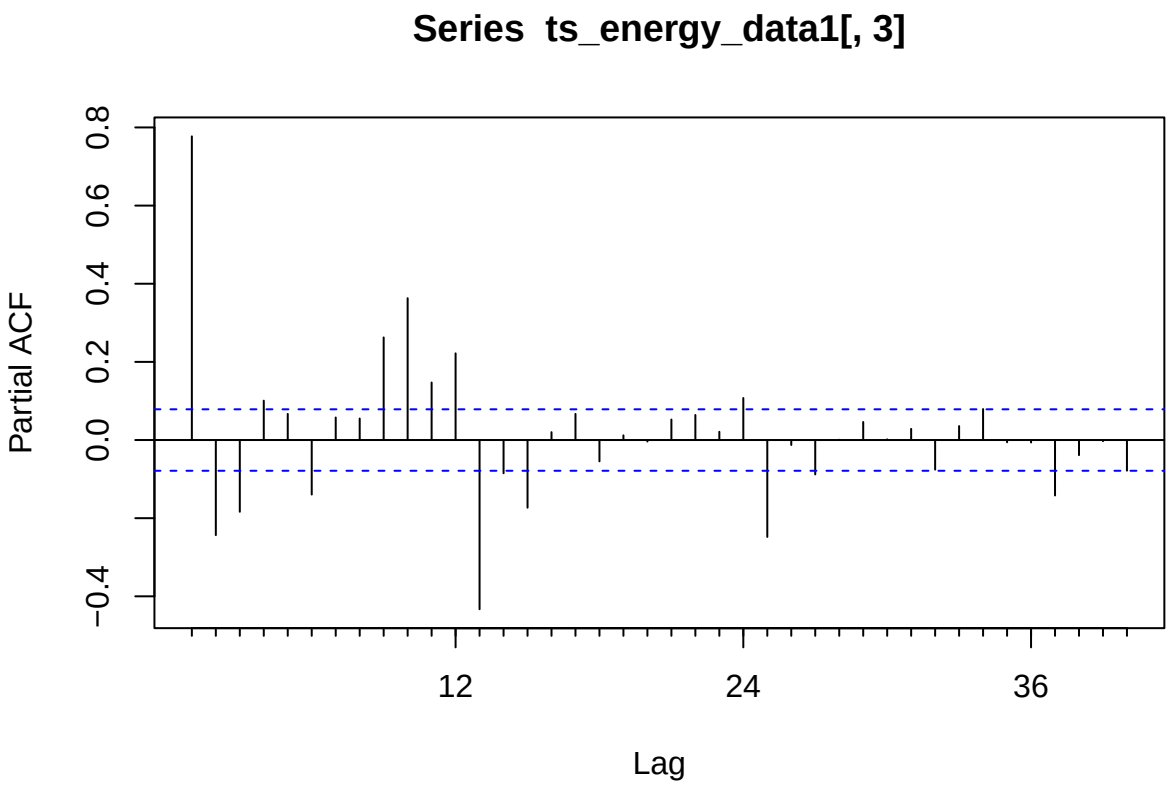
```
Biomass_pacf=Pacf(ts_energy_data1[,1],lag=40)
```




```
Renewable_pacf=Pacf(ts_energy_data1[,2],lag=40)
```




```
Hydroelectric_pacf=Pacf(ts_energy_data1[,3],lag=40)
```



The PACF plots differ vastly from the ACF plots since it only measures the direct correlation between the time series and lagged value. Plots 1 and 2 for Biomass and Renewable production show significant spikes towards the beginning, up until lags 2 or 3, and then most of the spikes become insignificant except for around the 13, 25, and 27 month marks, so every 12 months. For hydroelectric power, the seasonal variation isn't shown in the PACF plot, rather more significant correlations are in the first 12 lags and then it decreases in significant from then onward.