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

Assignment 2 - Due date 01/28/25

Jessalyn Chuang

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
```

```
library(lubridate)
```

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

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

```
library(ggplot2)
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*.

```
#Importing data set
energy_data <- 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`
```

```
#Extract column names from 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_data) <- read_col_names

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

```
## # 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_subset <- energy_data[, 4:6]
head(energy_subset)
```

```
## # 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_subset_ts <- ts(energy_subset[,1:3],start=c(1973,1),frequency=12)
```

Question 3

Compute mean and standard deviation for these three series.

```
# Find mean for each column
column_means <- apply(energy_subset_ts, 2, mean, na.rm = TRUE)

# Find the standard deviation for each column
column_sd <- apply(energy_subset_ts, 2, sd, na.rm = TRUE)

# Print results
print("Means of each series:")
```

```
## [1] "Means of each series:"
```

```
print(column_means)
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                         282.67785                          402.01667
##   Hydroelectric Power Consumption
##                          79.55371
```

```
print("Standard deviations of each series:")
```

```
## [1] "Standard deviations of each series:"
```

```
print(column_sd)
```

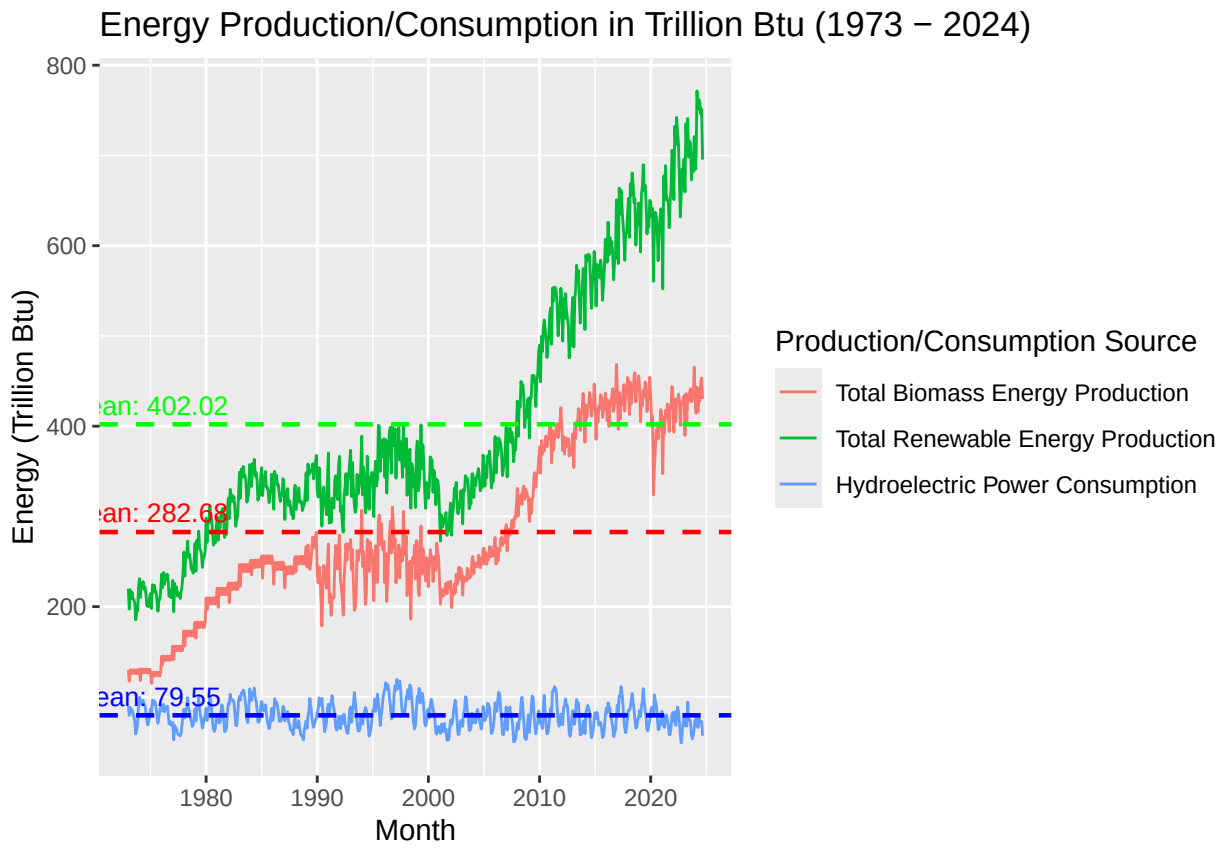
```
##   Total Biomass Energy Production Total Renewable Energy Production
##                         94.05815                          143.79270
##   Hydroelectric Power Consumption
##                         14.10737
```

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.

```r
autoplot(energy_subset_ts) +
  labs(
    x = "Month",
    y = "Energy (Trillion Btu)",
    title = "Energy Production/Consumption in Trillion Btu (1973 - 2024)",
    color = "Production/Consumption Source"
  ) +
  #Adding horizontal lines at the means with labels
  geom_hline(yintercept = column_means[1], color = "red", linetype = "dashed", size = 0.8) +
  geom_hline(yintercept = column_means[2], color = "green", linetype = "dashed", size = 0.8) +
  geom_hline(yintercept = column_means[3], color = "blue", linetype = "dashed", size = 0.8) +
  annotate("text", x = max(time(energy_subset_ts)) - 50, y = column_means[1],
           label = paste("mean:", round(column_means[1], 2)), color = "red", vjust = -0.5, size = 3.5) +
  annotate("text", x = max(time(energy_subset_ts)) - 50, y = column_means[2],
           label = paste("mean:", round(column_means[2], 2)), color = "green", vjust = -0.5, size = 3.5)
  annotate("text", x = max(time(energy_subset_ts)) - 50, y = column_means[3],
           label = paste("mean:", round(column_means[3], 2)), color = "blue", vjust = -0.5, size = 3.5)
```

```
## 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.
```



The plot shows the trends in Total Biomass Energy Production, Total Renewable Energy Production, and Hydroelectric Power Consumption from 1973 to 2024. Biomass and total renewable energy show significant

upward trends, with production accelerating quickly after the 2000s. Hydroelectric power consumption remains relatively stable over time, fluctuating seasonally around its mean of 79.55 Trillion BTU. This divergence alludes to the shift in energy systems toward diversified renewable sources, with biomass and other renewables driving the overall growth while hydroelectric power keeps a steady role.

Question 5

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

```
str(energy_subset)
```

```
## tibble [621 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Total Biomass Energy Production  : num [1:621] 130 117 130 126 130 ...
##  $ Total Renewable Energy Production: num [1:621] 220 197 219 209 216 ...
##  $ Hydroelectric Power Consumption  : num [1:621] 89.6 79.5 88.3 83.2 85.6 ...
```

```
# Perform correlation tests for each pair of series
cor_test_biomass_renewable <- cor.test(
                            energy_subset$`Total Biomass Energy Production`,
                             energy_subset$`Total Renewable Energy Production`)
cor_test_biomass_hydro <- cor.test( energy_subset$`Total Biomass Energy Production`,
                                    energy_subset$`Hydroelectric Power Consumption`)
cor_test_renewable_hydro <- cor.test(energy_subset$`Total Renewable Energy Production`,
                                     energy_subset$`Hydroelectric Power Consumption`)

# Print the results
print("Correlation Test between Total Biomass Energy Production and Total Renewable Energy Production:")
```

```
## [1] "Correlation Test between Total Biomass Energy Production and Total Renewable Energy Production:"
```

```
print(cor_test_biomass_renewable )
```

```
##
##  Pearson's product-moment correlation
##
## data:  energy_subset$`Total Biomass Energy Production` and energy_subset$`Total Renewable Energy Pro
## 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
```

```
print("Correlation Test between Total Biomass Energy Production and Hydroelectric Power Consumption:")
```

```
## [1] "Correlation Test between Total Biomass Energy Production and Hydroelectric Power Consumption:"
```

```r
print(cor_test_biomass_hydro)
```

```
##
##  Pearson's product-moment correlation
##
## data:  energy_subset$`Total Biomass Energy Production` and energy_subset$`Hydroelectric Power Consump
## 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("Correlation Test between Total Renewable Energy Production and Hydroelectric Power Consumption:")
```

```
## [1] "Correlation Test between Total Renewable Energy Production and Hydroelectric Power Consumption:"
```

```r
print(cor_test_renewable_hydro)
```

```
##
##  Pearson's product-moment correlation
##
## data:  energy_subset$`Total Renewable Energy Production` and energy_subset$`Hydroelectric Power Consu
## 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
```

The correlation tests reveal varying relationships between the energy production and consumption variables. The strongest and most significant correlation was found between Total Biomass Energy Production and Total Renewable Energy Production (r = 0.9678, p < 2.2e-16). This indicates a very strong positive relationship, suggesting that increases in biomass energy production are closely tied to increases in overall renewable energy production.

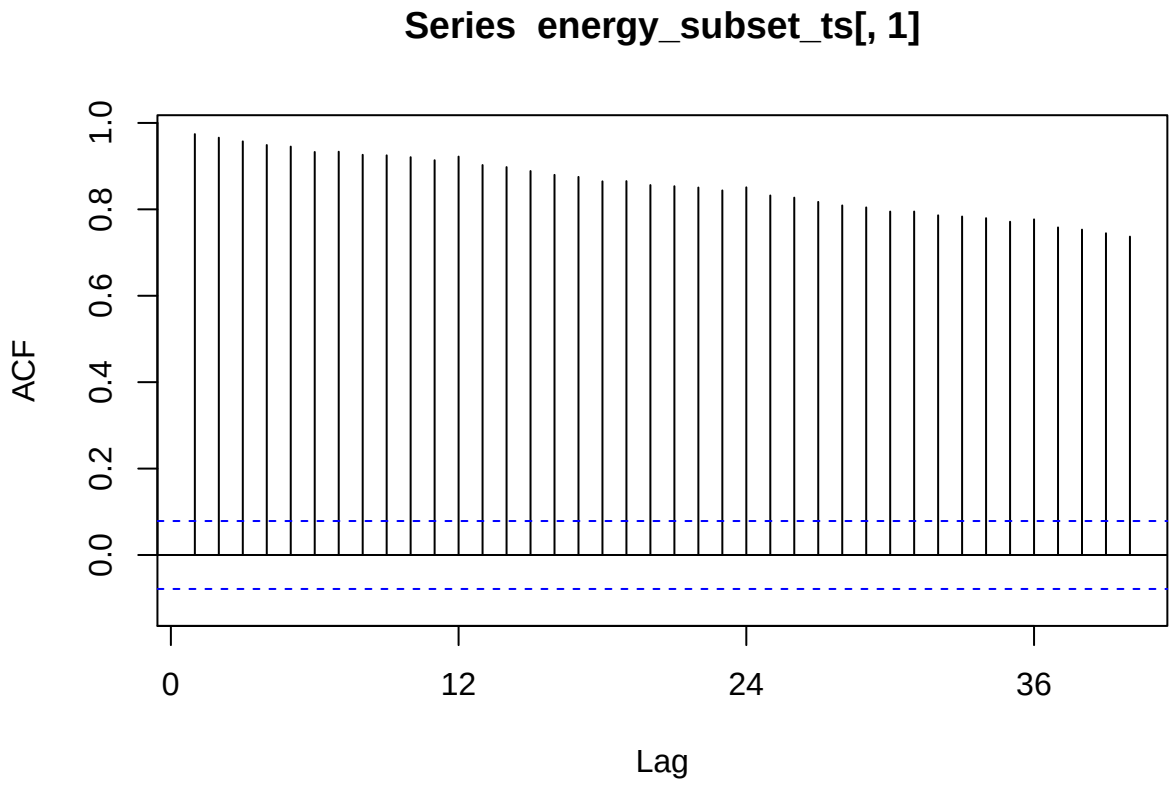
In contrast, the correlation between Total Biomass Energy Production and Hydroelectric Power Consumption is weakly negative (r = -0.1143) but statistically significant (p = 0.0043). Although the correlation is weak, the negative value suggests a slight inverse relationship, potentially indicating that variations in biomass energy production and hydroelectric consumption are not strongly aligned and may operate independently or inversely influenced by other factors.

The correlation between Total Renewable Energy Production and Hydroelectric Power Consumption is very weak (r = -0.0292) and not statistically significant (p = 0.4682). This suggests that there is no meaningful relationship between these two variables.

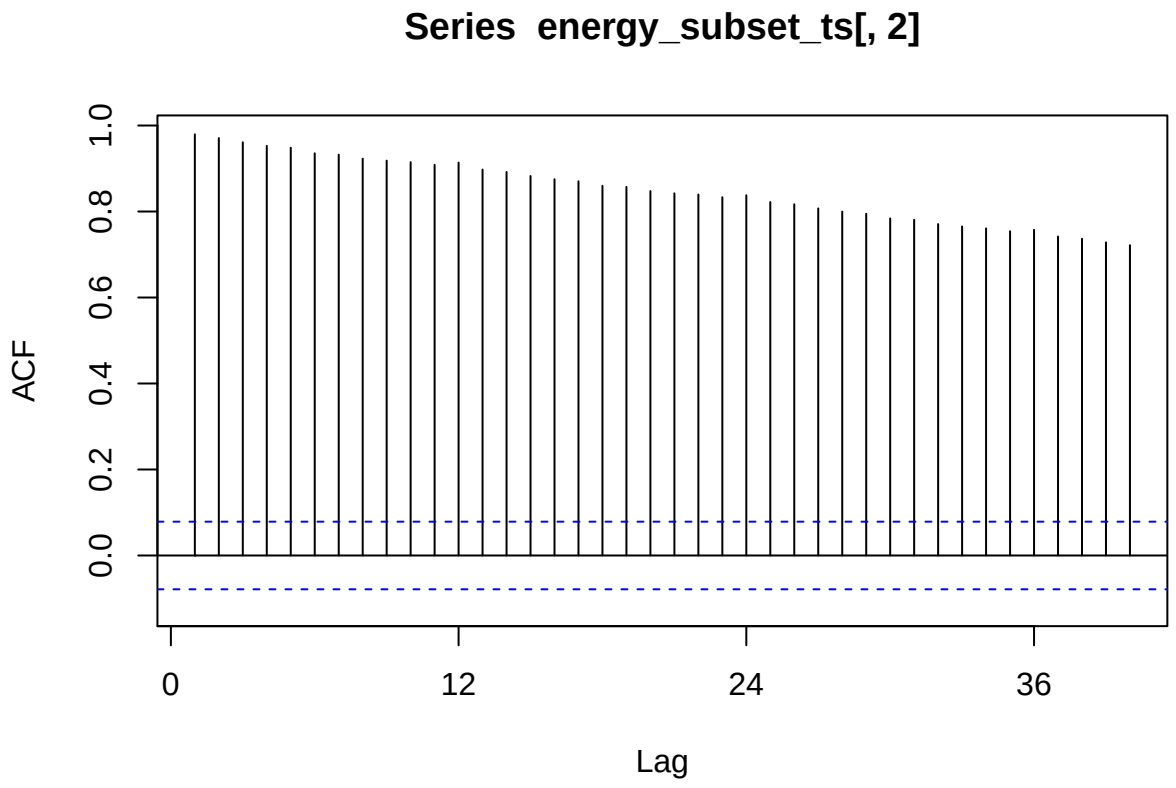
Question 6

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?

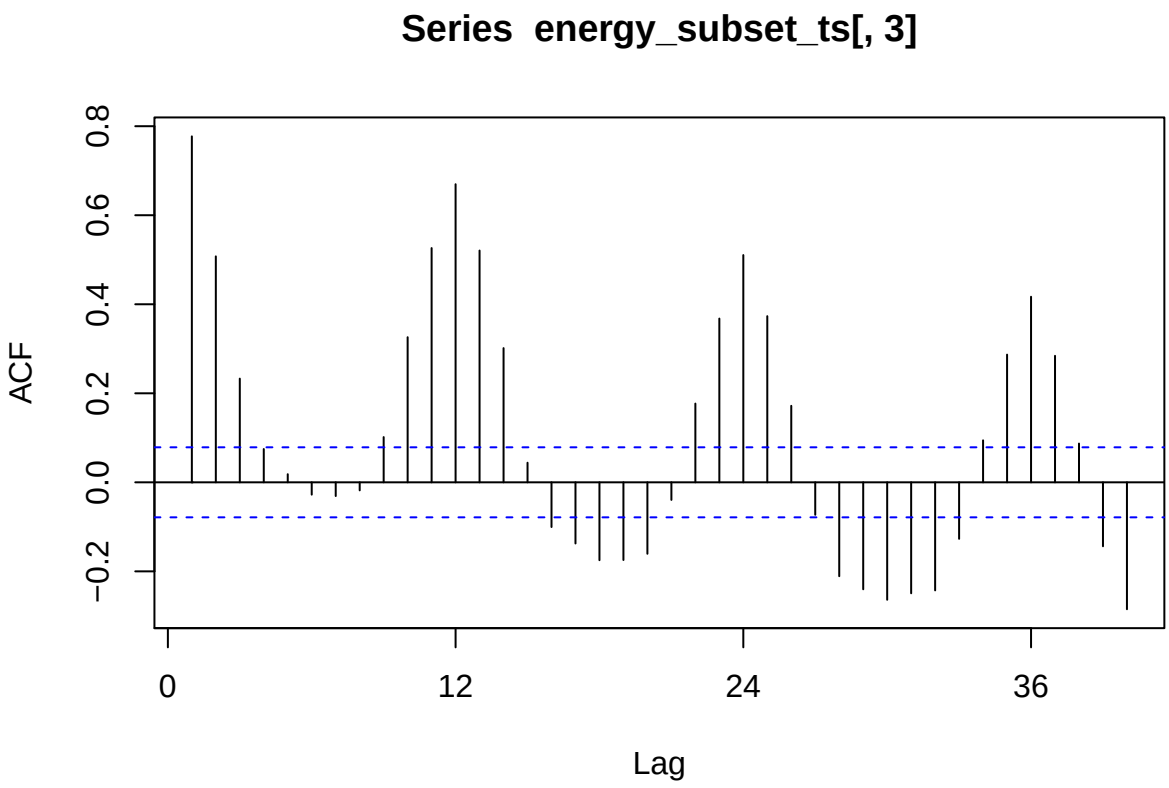
```r
biomass_acf=Acf(energy_subset_ts[,1],lag=40)
```




```r
renewable_acf=Acf(energy_subset_ts[,2],lag=40)
```




```
hydro_acf=Acf(energy_subset_ts[,3],lag=40)
```



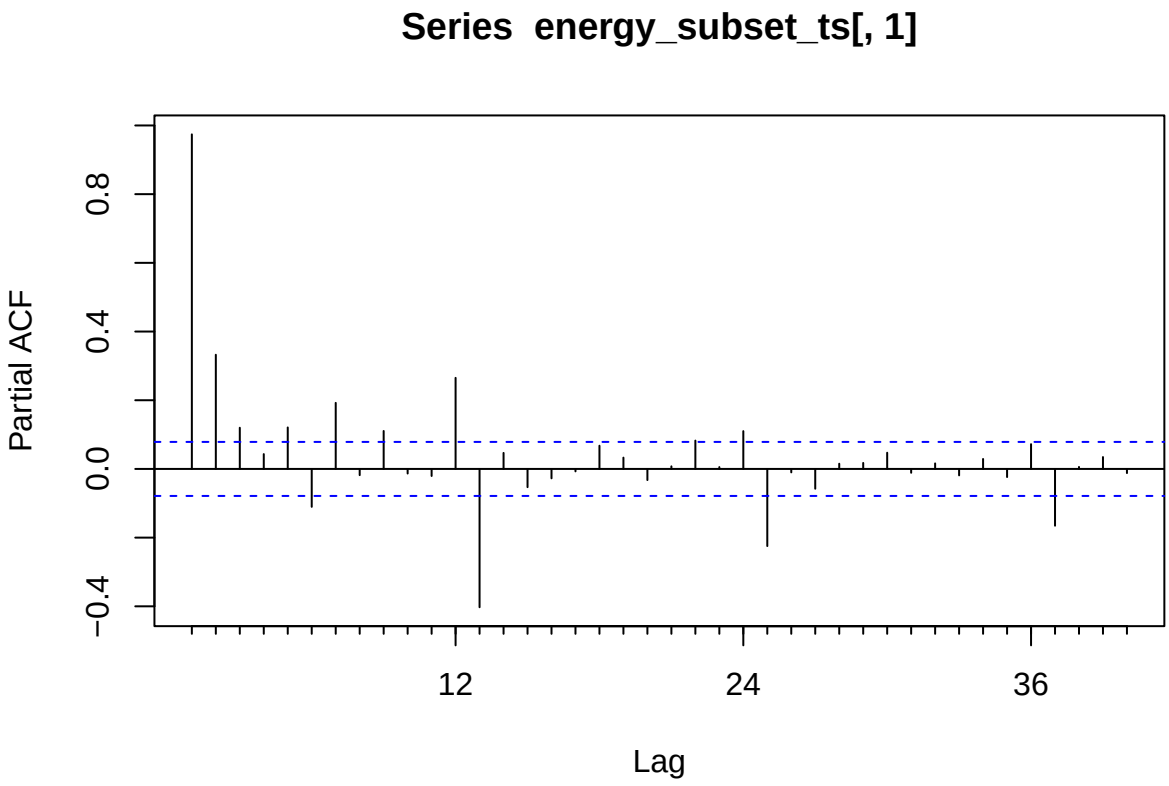
The ACF for both Biomass Energy Production and Renewable Energy Production shows very high autocorrelation at lag 1 that gradually declines but remains significant across many lags. This pattern suggests strong temporal dependence and potential non-stationarity in the series. It seems like past values heavily influence future values. This is reasonable as renewable energy production includes biomass as a component, and the two are likely influenced by similar long-term growth trends.

The ACF for Hydroelectric Power Consumption behaves differently. It has oscillations, and the correlations are not consistently high across lags. This behavior suggests different underlying dynamics, possibly driven by seasonal or cyclical factors. The fluctuating correlations imply that hydroelectric power consumption may be closer to stationarity or less dominated by long-term growth trends compared to biomass and renewable energy production.

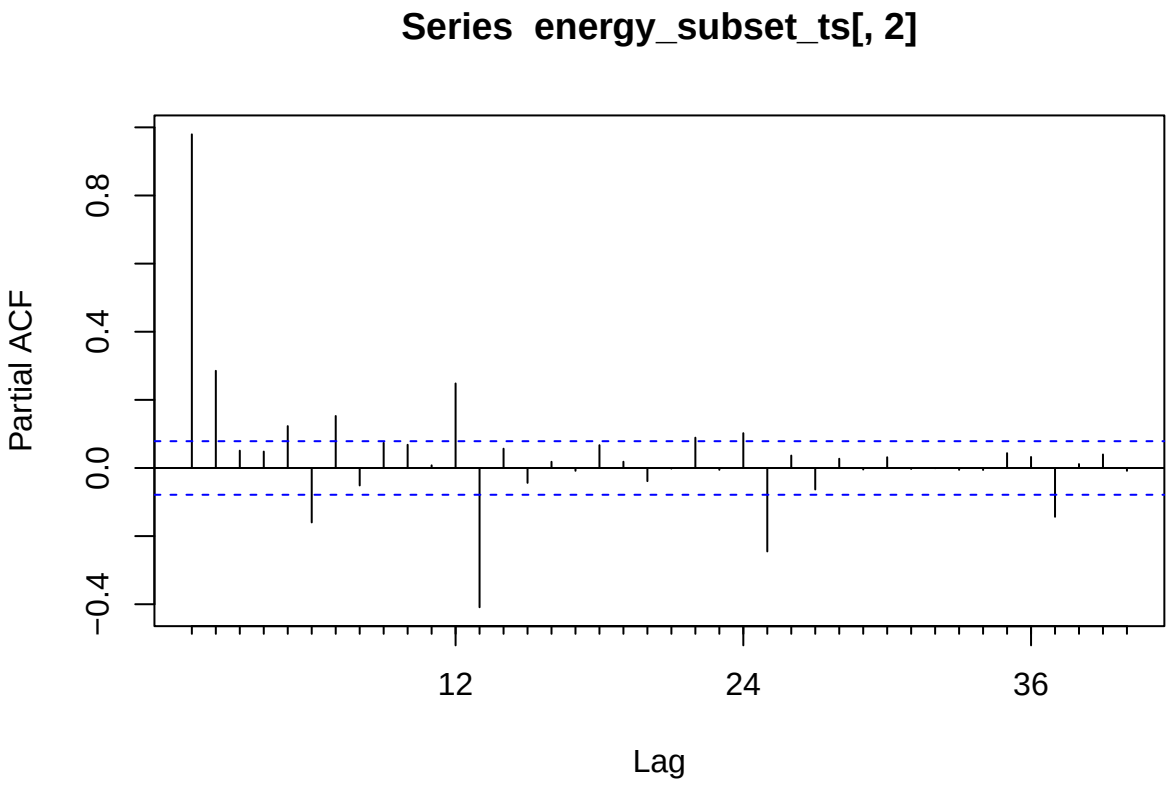
Question 7

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

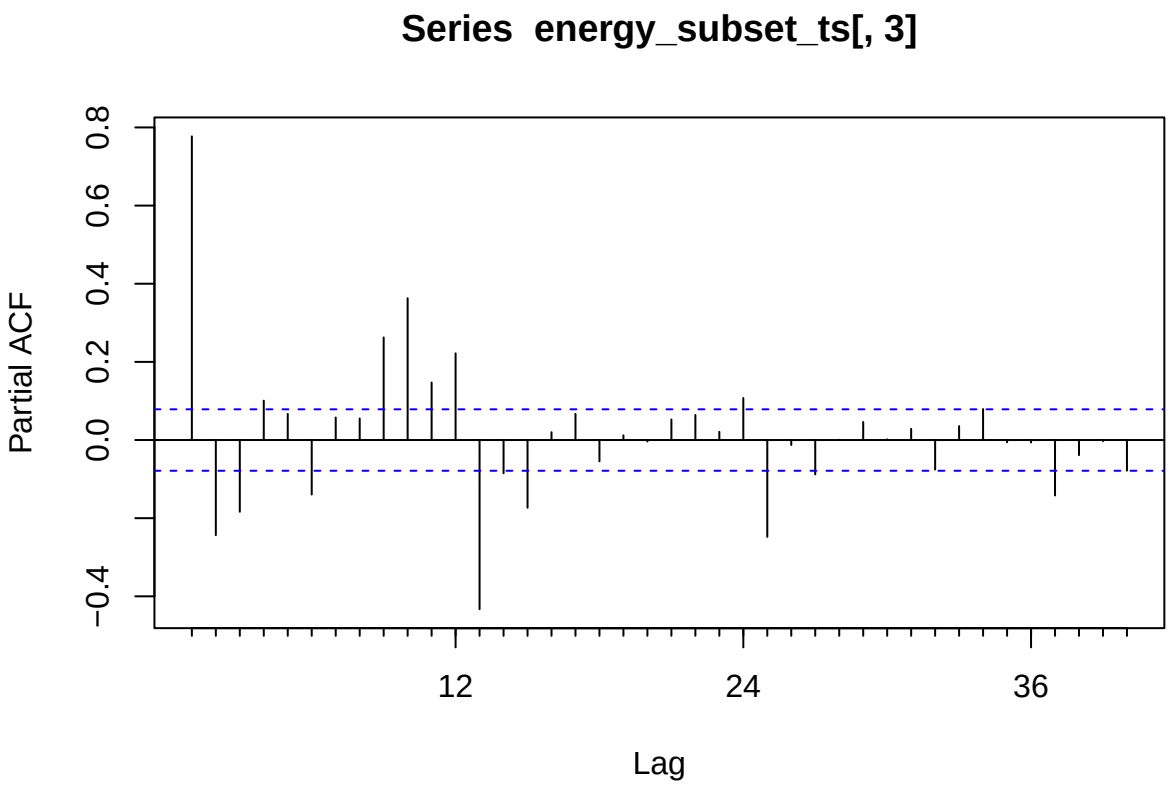
```
biomass_pacf=Pacf(energy_subset_ts[,1],lag=40)
```




```
renewable_pacf=Pacf(energy_subset_ts[,2],lag=40)
```




```
hydro_pacf=Pacf(energy_subset_ts[,3],lag=40)
```



For Biomass and Renewable Energy Production, the PACF shows a strong spike at lag 1, with much smaller values in following lags. This indicates that both series are mostly influenced by their immediate past values. Unlike the ACF, which showed high autocorrelation over many lags, the PACF reveals that these longer-term correlations are largely indirect effects of lag 1. Meanwhile, the PACF for Hydroelectric Power Consumption shows a strong spike at lag 1, along with smaller but significant spikes at higher lags. These oscillations suggest the presence of seasonal behavior, which aligns with the patterns observed in its ACF plot.