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

Assignment 2 - Due date 01/23/25

Daniel Whitehead

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.\

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

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
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(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*.

```r
#Importing data set
library(readxl)
Table_10_1_Renewable_Energy_Production_and_Consumption_by_Source <- read_excel(
  "~/ENV797/Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
  skip = 9)
Energy_df <-
  Table_10_1_Renewable_Energy_Production_and_Consumption_by_Source[-1, ]

print(Energy_df)
```

```
## # A tibble: 621 x 14
##    Month               `Wood Energy Production` `Biofuels Production`
##    <dttm>              <chr>                    <chr>
##  1 1973-01-01 00:00:00 129.63                   Not Available
##  2 1973-02-01 00:00:00 117.194                  Not Available
##  3 1973-03-01 00:00:00 129.763                  Not Available
##  4 1973-04-01 00:00:00 125.462                  Not Available
##  5 1973-05-01 00:00:00 129.624                  Not Available
##  6 1973-06-01 00:00:00 125.435                  Not Available
##  7 1973-07-01 00:00:00 129.616                  Not Available
##  8 1973-08-01 00:00:00 129.734                  Not Available
##  9 1973-09-01 00:00:00 125.603                  Not Available
## 10 1973-10-01 00:00:00 129.769                  Not Available
## # i 611 more rows
## # i 11 more variables: `Total Biomass Energy Production` <chr>,
## #   `Total Renewable Energy Production` <chr>,
## #   `Hydroelectric Power Consumption` <chr>,
## #   `Geothermal Energy Consumption` <chr>, `Solar Energy Consumption` <chr>,
## #   `Wind Energy Consumption` <chr>, `Wood Energy Consumption` <chr>,
## #   `Waste Energy Consumption` <chr>, `Biofuels Consumption` <chr>, ...
```

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
selected_energy <- Energy_df %>%
  select(4,5,6)
#realized my data was in character form :(
selected_energy <- as.data.frame(lapply(selected_energy, as.numeric))

head(selected_energy)
```

```
##   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().

```r
energy_ts <- ts(selected_energy, start = c(1973, 1), frequency = 12)

head(energy_ts)
```

```
##          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_biomass <- round(mean(energy_ts[, 1]), 2)
sd_biomass <- round(sd(energy_ts[, 1]), 2)

# Renewable
mean_renewable <- round(mean(energy_ts[, 2]), 2)
sd_renewable <- round(sd(energy_ts[, 2]), 2)

# Hydroelectric
mean_hydroelectric <- round(mean(energy_ts[, 3]), 2)
sd_hydroelectric <- round(sd(energy_ts[, 3]), 2)

print(paste("Biomass Mean and Standard Deviation:"))
```

```
## [1] "Biomass Mean and Standard Deviation:"
```

```r
print(paste("Mean:", mean_biomass, "trillion BTU. Standard Deviation:",
            sd_biomass, "trillion BTU."))
```

```
## [1] "Mean: 282.68 trillion BTU. Standard Deviation: 94.06 trillion BTU."
```

```r
print(paste("Renewable Mean and Standard Deviation:"))
```

```
## [1] "Renewable Mean and Standard Deviation:"
```

```r
print(paste("Mean:", mean_renewable, "trillion BTU. Standard Deviation:",
            sd_renewable, "trillion BTU."))
```

```
## [1] "Mean: 402.02 trillion BTU. Standard Deviation: 143.79 trillion BTU."
```

```r
print(paste("Hydroelectric Mean and Standard Deviation:"))
```

```
## [1] "Hydroelectric Mean and Standard Deviation:"
```

```r
print(paste("Mean:", mean_hydroelectric, "trillion BTU. Standard Deviation:",
            sd_hydroelectric, "trillion BTU."))
```

```
## [1] "Mean: 79.55 trillion BTU. Standard Deviation: 14.11 trillion BTU."
```

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_ts) +
  labs(
    x = "Year",
    y = "Energy Production/Consumption (Trillion BTU)",
    title = "Energy Production/Consumption Over Time",
    caption = "Figure 1: Time series of energy production and consumption,
    with dashed lines indicating the mean values for Total
    Biomass Energy Production (red), Total Renewable Energy
    Production (green), and Hydroelectric Power Consumption (blue)."
  ) +
  geom_hline(yintercept = mean_biomass, color = "red", linetype = "dashed",
             size = 1.2) +
  geom_hline(yintercept = mean_renewable, color = "green", linetype = "dashed",
             size = 1.2) +
  geom_hline(yintercept = mean_hydroelectric, color = "blue",
             linetype = "dashed", size = 1.2) +
  theme(plot.caption = element_text(hjust = 0.5, size = 10,
                                    face = "italic"))
```

```
## 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.
```

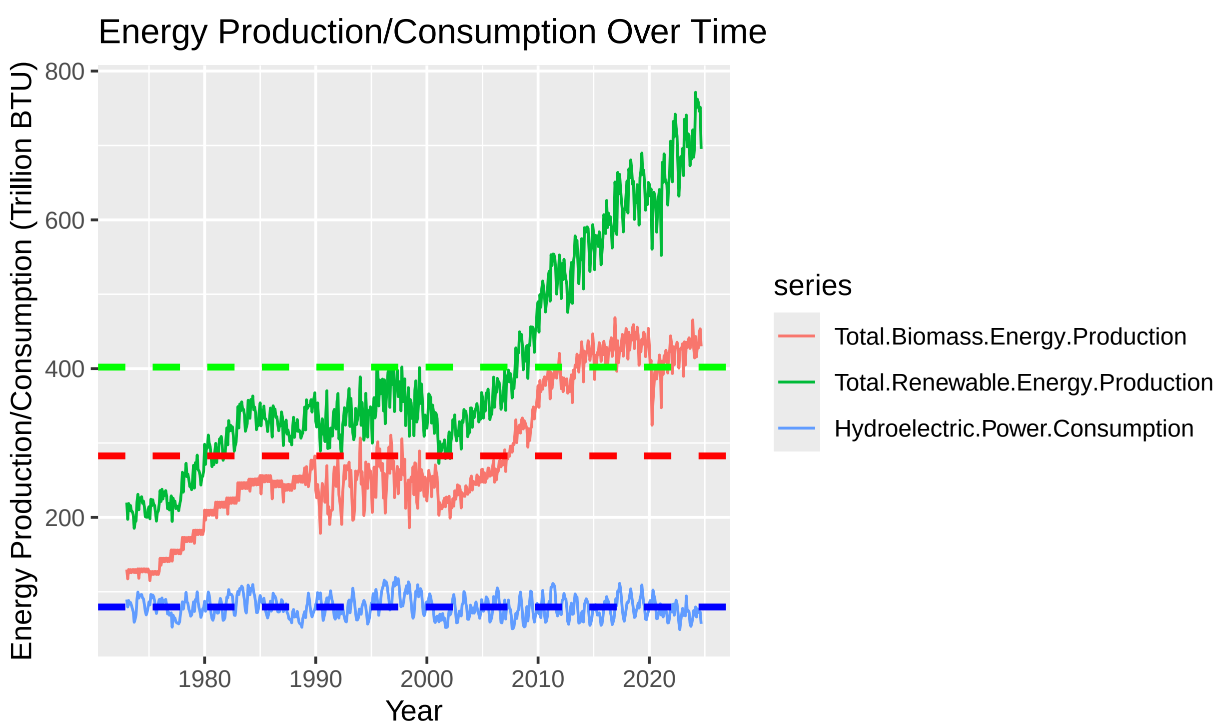


Figure 1: Time series of energy production and consumption, with dashed lines indicating the mean values for Total Biomass Energy Production (red), Total Renewable Energy Production (green), and Hydroelectric Power Consumption (blue).

```r
print("For total renewable energy production, production is at about 200
      trillion BTU in 1973. It increases for about 15 years, stagnates for
      about 20, then hits a (somewhat steep) steady increase that continues to
      grow into the 2020's. Biomass has a very similar trend, though it starts
      closer to 125 trillion BTU in 1973 and stagnates a second time starting
      in about 2015. Hydroelectric consumption remains steady abound 80 trillion
      BTU throughout the entirety of the time series plot.")
```

```
## [1] "For total renewable energy production, production is at about 200 \n      trillion BTU in 1973.
```

Question 5

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

```r
correlation_matrix <- cor(energy_ts, use = "complete.obs")

print("Correlation Matrix:")
```

```
## [1] "Correlation Matrix:"
```

```r
print(correlation_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
```

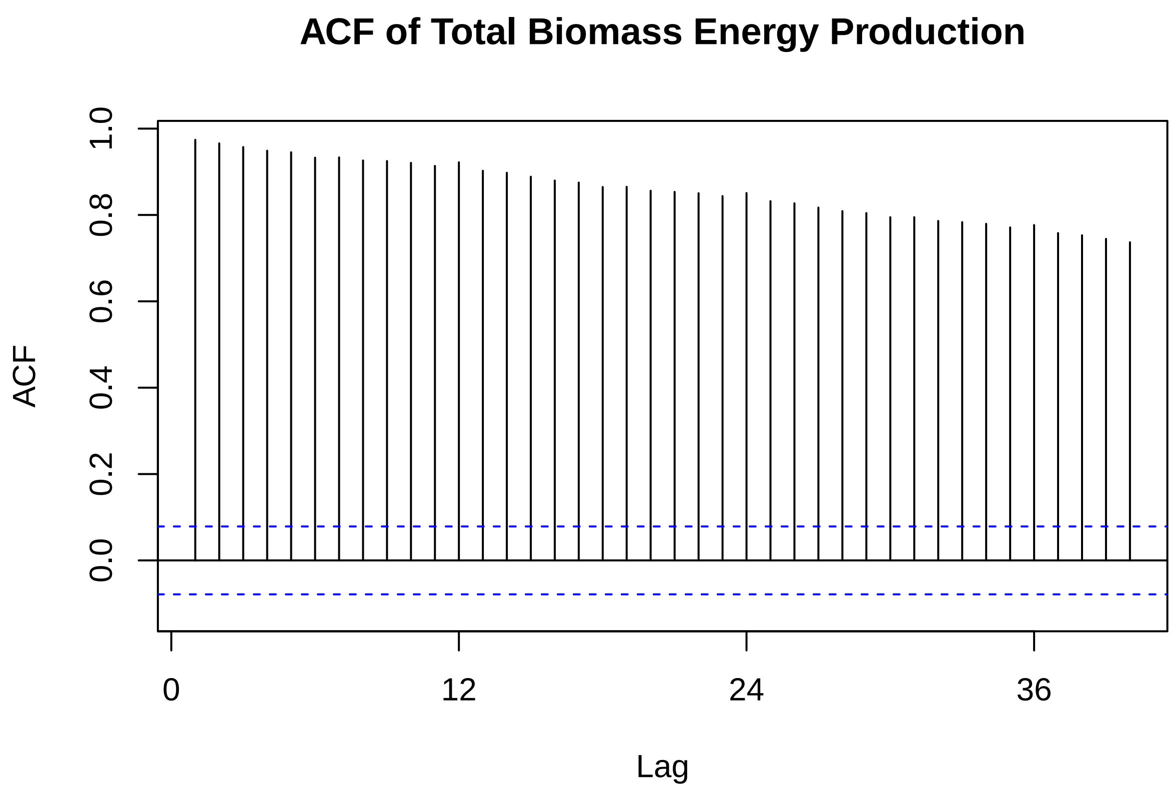
```r
print(paste("Total Biomass Energy Production and Total Renewable Energy Production
      seem to be highly correlated as there is a 0.97 correlation between them.
      However, these series do not seem to have much correlation with
      Hydroelectric Power Consumption, as the correlation values are very low
      and insignificant."))
```

```
## [1] "Total Biomass Energy Production and Total Renewable Energy Production \n      seem to be highly
```

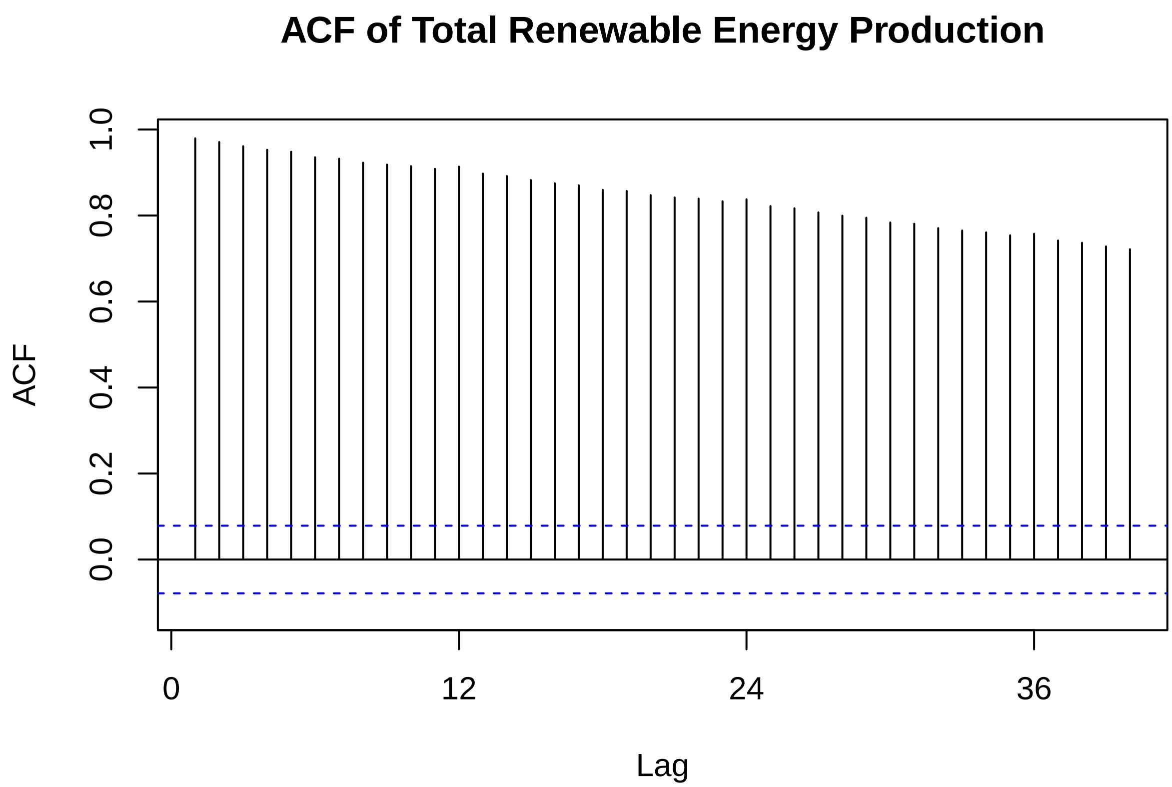
Question 6

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?

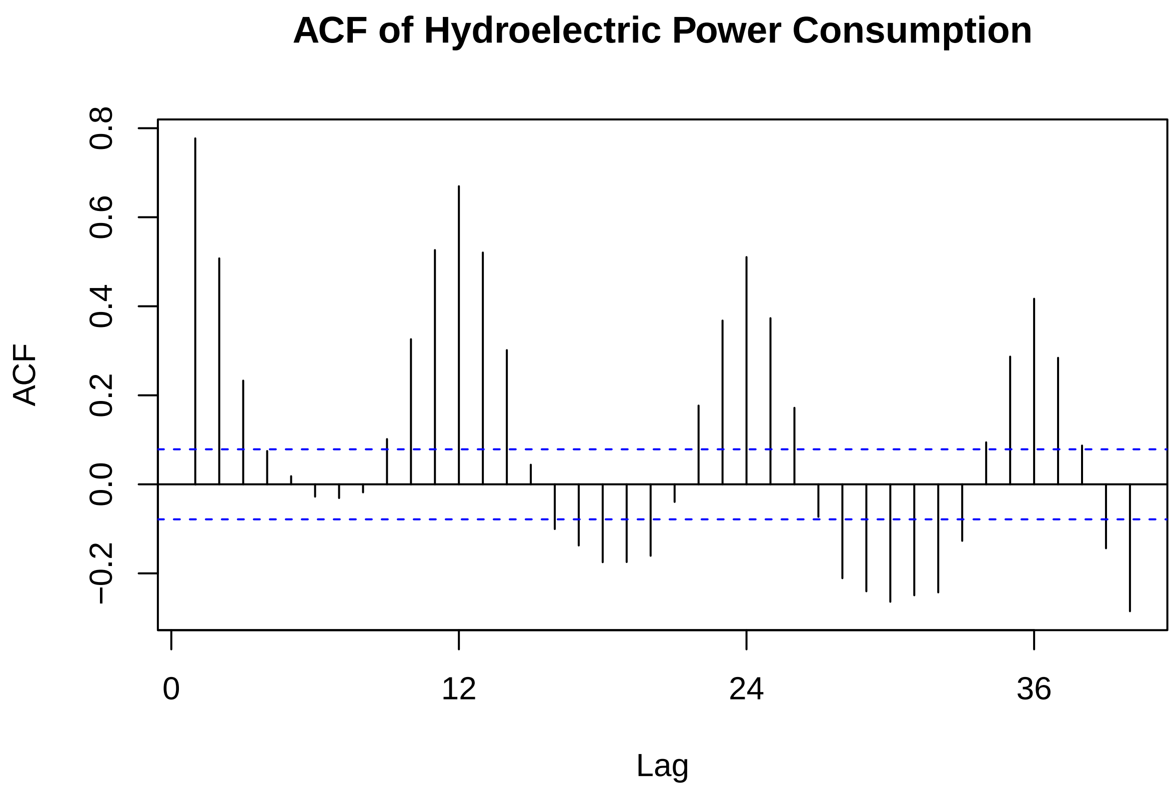
```r
acf_biomass <- Acf(energy_ts[,1], lag.max = 40,
                   main = "ACF of Total Biomass Energy Production",
                   type = "correlation", plot = TRUE)
```




```r
acf_renewable <- Acf(energy_ts[,2], lag.max = 40,
                     main = "ACF of Total Renewable Energy Production",
                     type = "correlation", plot = TRUE)
```




```
acf_hydro <- Acf(energy_ts[,3], lag.max = 40,
                 main = "ACF of Hydroelectric Power Consumption",
                 type = "correlation", plot = TRUE)
```



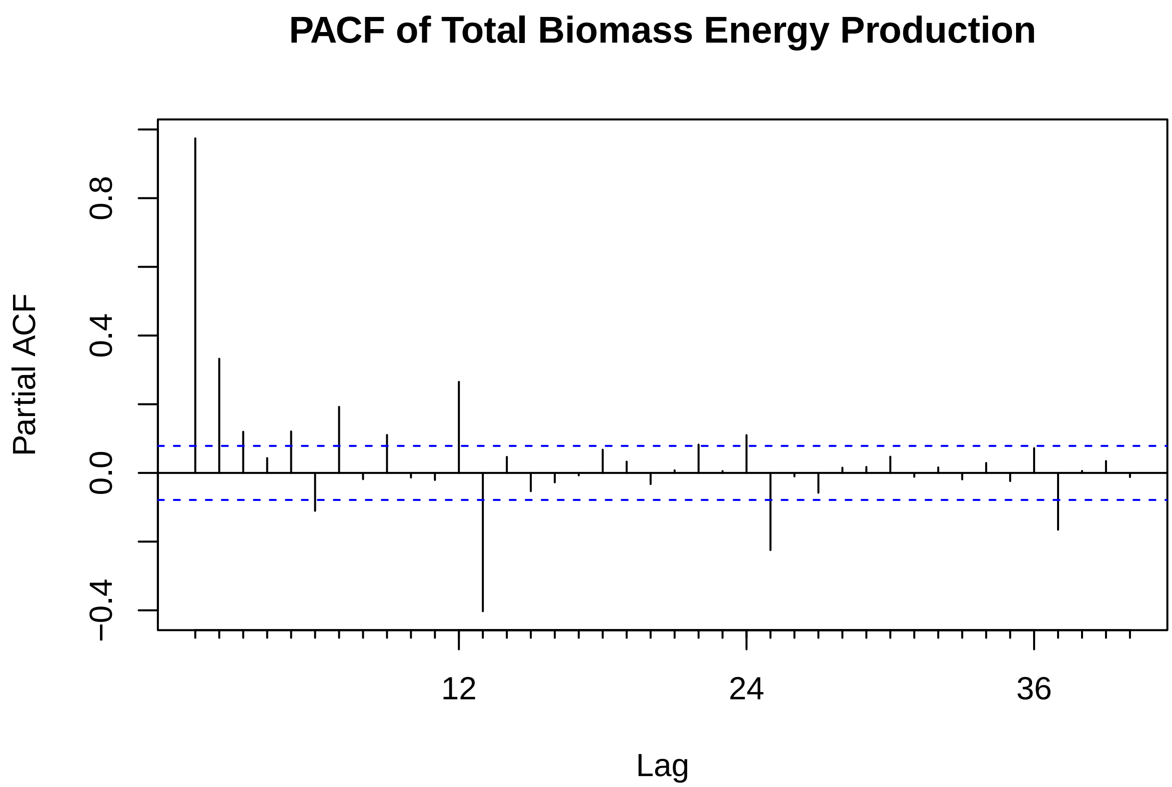

```
print(paste("Total Biomass Energy Production and Total Renewable Energy Production
      have similar ACF plots in terms of behavior. Both start with high values
      and then slowly decline. Hydroelectric Power Consumption is different in
      terms of behavior, as it rises and falls, crossing in and out of negative
      values. This means that for Biomass and Renewables, there is more
      predictability in forecasting the future or longer-term data. Hydro seems
      to be more affected by factors that cause short term fluctuations."))
```

```
## [1] "Total Biomass Energy Production and Total Renewable Energy Production\n      have similar ACF pl
```

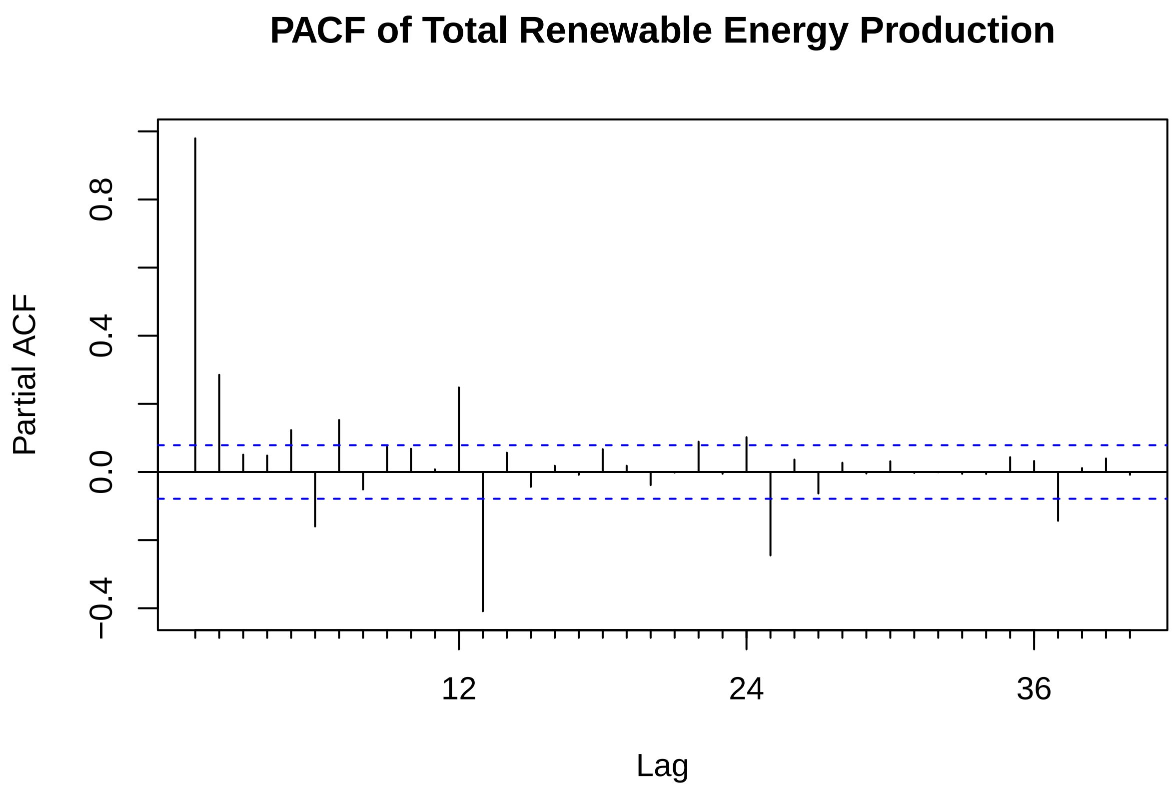
Question 7

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

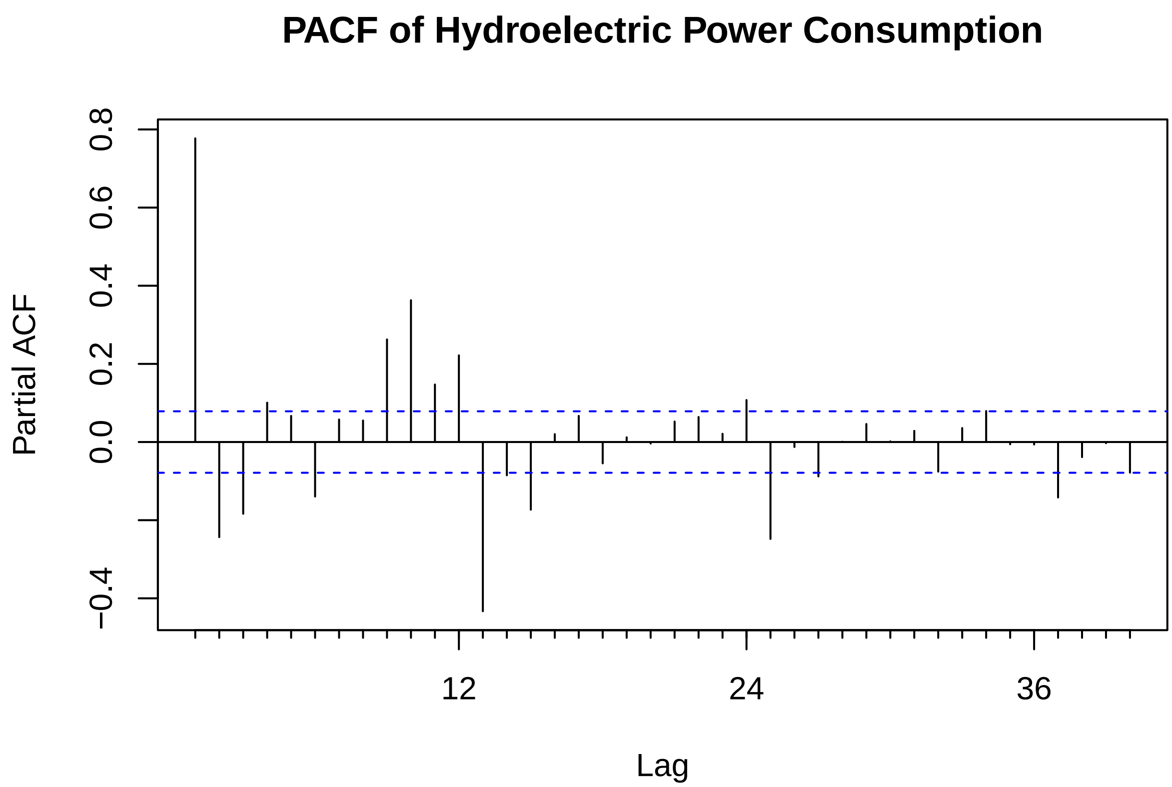
```
pacf_biomass <- Pacf(energy_ts[,1], lag.max = 40,
                     main = "PACF of Total Biomass Energy Production",
                     plot=TRUE)
```




```
pacf_renewable <- Pacf(energy_ts[,2], lag.max = 40,
                       main = "PACF of Total Renewable Energy Production",
                       plot=TRUE)
```




```
pacf_hydro <- Pacf(energy_ts[,3], lag.max = 40,
                   main = "PACF of Hydroelectric Power Consumption", plot=TRUE)
```




```
print(paste("These plots differ because they have much shorter bars. This makes sense
      because in general the PACF plots have shorter bars, as they do not
      account for the complete correlation of the lags and instead do not
      account for shorter lags as the lags get higher in values. Additionally,
      the hydropower plot is much more similar to the biomass and renewable
      plots."))
```

```
## [1] "These plots differ because they have much shorter bars. This makes sense\n      because in gene
```