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

Assignment 2 - Due date 01/28/25

Ananya Aggarwal

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.

Setting R code chunk options

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(openxlsx)
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
#any kind of cleaning up of downloaded data needs to be done in R itself to make the code reproducible
#avoid manual adjustments in the excel/csv
getwd()
```

```
## [1] "C:/Users/nancy/Desktop/Time Series Analysis/TSA_Sp25/Assignments"
```

```r
energy_data <-
  read.xlsx(xlsxFile="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                          sheet="Monthly Data",
                          startRow=13,
                          colNames=FALSE)

#this dataset has sort of two headers (name of the variable and then the units)
read_colnames <- read.xlsx(xlsxFile="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_S
                          sheet="Monthly Data",
                          rows=11,
                          colNames=FALSE)

#converting the date-time format from the excel format to month-year
energy_data[,1] <- as.Date(energy_data[,1], origin = "1899-12-30")

colnames(energy_data) <- read_colnames

head(energy_data)
```

```
##        Month Wood Energy Production Biofuels Production
## 1 1973-01-01                129.630       Not Available
## 2 1973-02-01                117.194       Not Available
## 3 1973-03-01                129.763       Not Available
## 4 1973-04-01                125.462       Not Available
## 5 1973-05-01                129.624       Not Available
## 6 1973-06-01                125.435       Not Available
##   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 Geothermal Energy Consumption
## 1                          89.562                          0.490
## 2                          79.544                          0.448
## 3                          88.284                          0.464
## 4                          83.152                          0.542
## 5                          85.643                          0.505
## 6                          82.060                          0.579
##   Solar Energy Consumption Wind Energy Consumption Wood Energy Consumption
## 1          Not Available          Not Available                129.630
## 2          Not Available          Not Available                117.194
## 3          Not Available          Not Available                129.763
## 4          Not Available          Not Available                125.462
## 5          Not Available          Not Available                129.624
```

```
## 6            Not Available           Not Available                 125.435
##   Waste Energy Consumption Biofuels Consumption
## 1                    0.157        Not Available
## 2                    0.144        Not Available
## 3                    0.176        Not Available
## 4                    0.174        Not Available
## 5                    0.210        Not Available
## 6                    0.176        Not Available
##   Total Biomass Energy Consumption Total Renewable Energy Consumption
## 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
```

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.

```
#filtering the dataset
energy_data_trim <- energy_data[,c("Month","Total Biomass Energy Production",
                                   "Total Renewable Energy Production",
                                   "Hydroelectric Power Consumption")]
head(energy_data_trim)
```

```
##        Month Total Biomass Energy Production Total Renewable Energy Production
## 1 1973-01-01                          129.787                             219.839
## 2 1973-02-01                          117.338                             197.330
## 3 1973-03-01                          129.938                             218.686
## 4 1973-04-01                          125.636                             209.330
## 5 1973-05-01                          129.834                             215.982
## 6 1973-06-01                          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().

```
#creating one object with all three time series
ts_energy_data <- ts(energy_data_trim[,2:4], start = c(1973, 1), frequency = 12)
head(ts_energy_data)
```

```
##          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
series <- c("Total Biomass Energy Production",
            "Total Renewable Energy Production",
            "Hydroelectric Power Consumption")

means <- c()
stdevs <- c()

for (i in series) {
  means[i] <- mean(ts_energy_data[, i])
  stdevs[i] <- sd(ts_energy_data[, i])
}

for (i in series) {
  cat(i, "\n")
  cat("Mean: ", means[i], "\n")
  cat("Std.Deviation: ", stdevs[i], "\n\n")
}
```

```
## Total Biomass Energy Production
## Mean:  282.6779
## Std.Deviation:  94.05815
##
## Total Renewable Energy Production
## Mean:  402.0167
## Std.Deviation:  143.7927
##
## Hydroelectric Power Consumption
## Mean:  79.55371
## Std.Deviation:  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

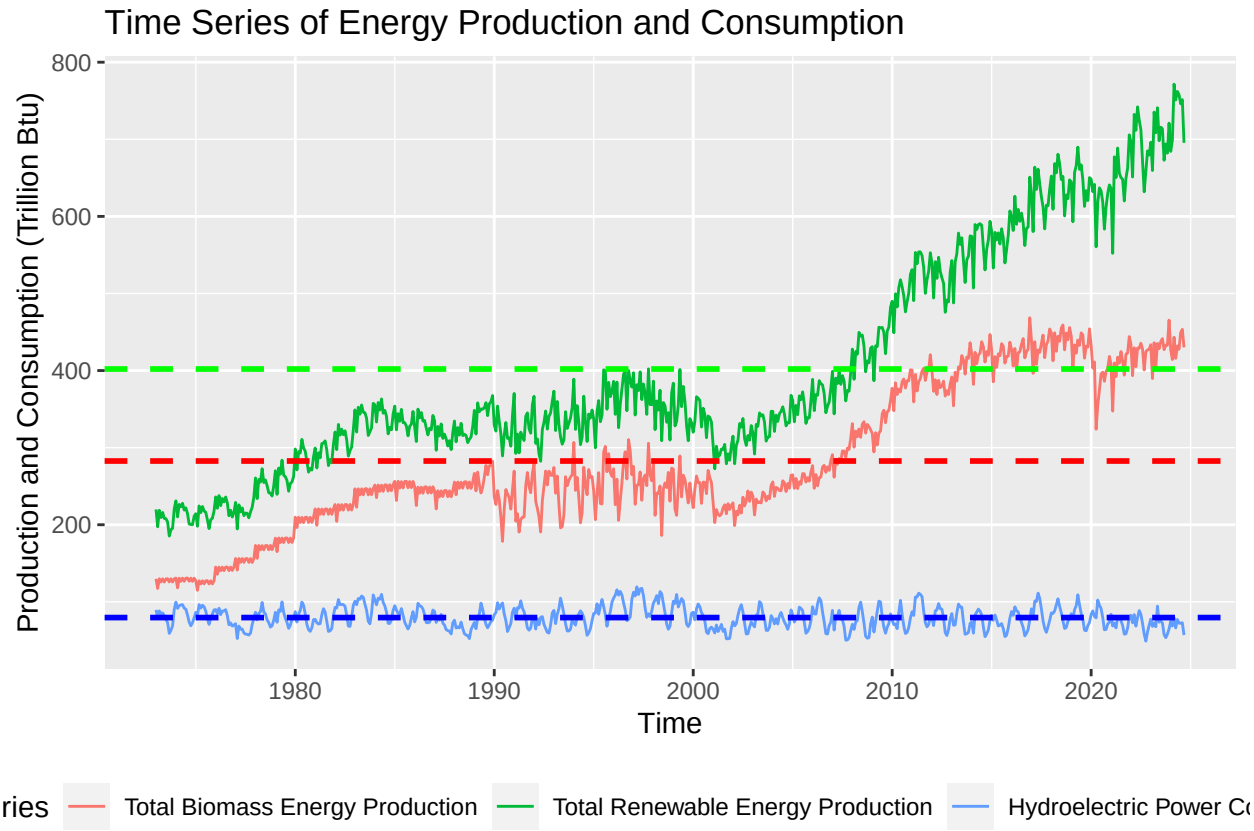
different color.

```
autoplot(ts_energy_data) +
  xlab("Time") +
  ylab("Production and Consumption (Trillion Btu)") +

  ggtitle("Time Series of Energy Production and Consumption") +

  geom_hline(aes(yintercept = means["Total Biomass Energy Production"]),
             color = "red", linetype = "dashed", size = 1) +
  geom_hline(aes(yintercept = means["Total Renewable Energy Production"]),
             color = "green", linetype = "dashed", size = 1) +
  geom_hline(aes(yintercept = means["Hydroelectric Power Consumption"]),
             color = "blue", linetype = "dashed", size = 1) +

  theme(legend.position = "bottom")
```



Answer 4 Trend: - Total Energy Production for both Biomass and Renewable Energy shows a non-linear increasing trend, with the period between 1985-2000 showing a stable trend on average. - Total Renewable Energy Production, in specific, has grown rapidly post 2002 (from around 300 Trillion Btu to more than 700 Trillion Btu). - Hydroelectric Power Consumption shows a stable trend throughout the study period, which can also be seen from a relatively low standard deviation (std.dev. of 14.11 around a mean of 79.55).

Seasonality: All three time series seem to have a strong seasonality component to them

Variations: - We see the effect of Covid-19 on energy production, where both biomass and renewable energy production saw a significant dip in the beginning of 2020. - Hydroelectric power consumption has been lower than average in the past few years (2021-23), probably owing to changes in the precipitation patterns.

Question 5

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

```
cor(ts_energy_data)
```

```
##                                    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
```

Answer 5

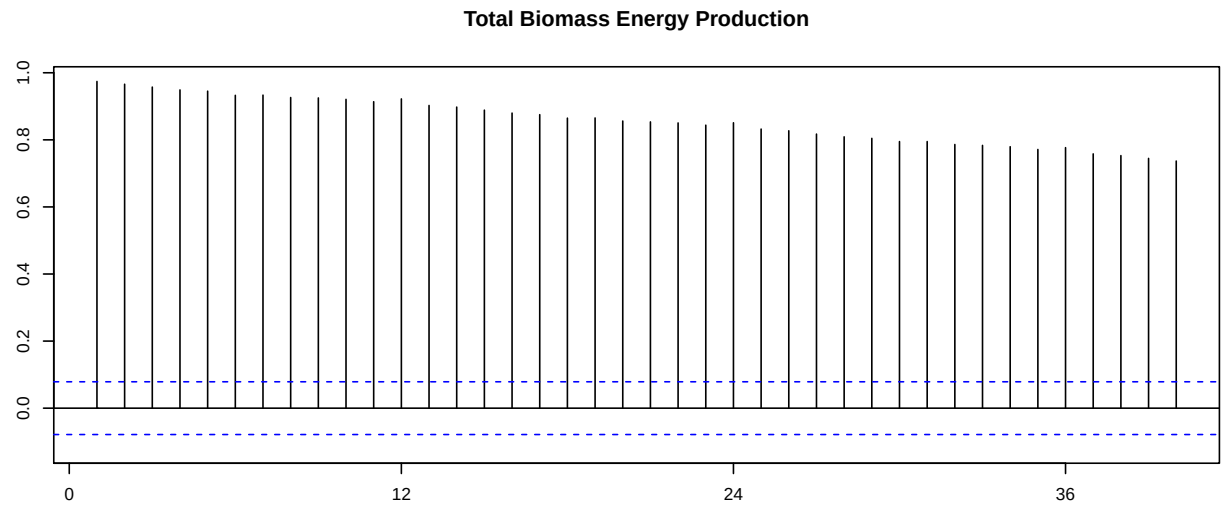
- Total Biomass and Renewable Energy Production show a very strong positive correlation, which suggests that as biomass energy production increase, renewable energy production tends to increase as well, and vice versa.
- On the other hand, there is a very weak negative correlation between these energy production datasets and hydroelectric power consumption, indicating changes in hydroelectric power consumption are not influenced by biomass and renewable energy production.

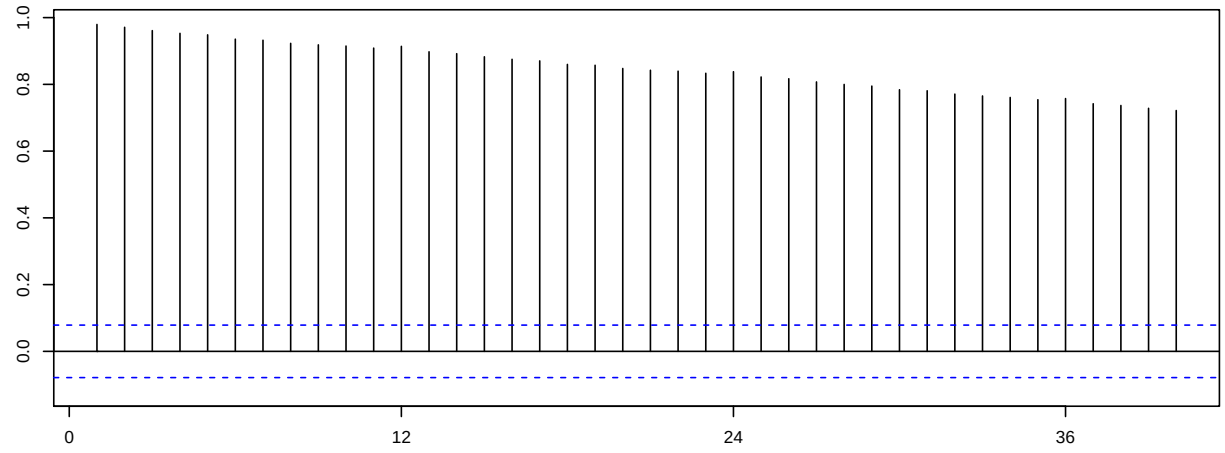
Question 6

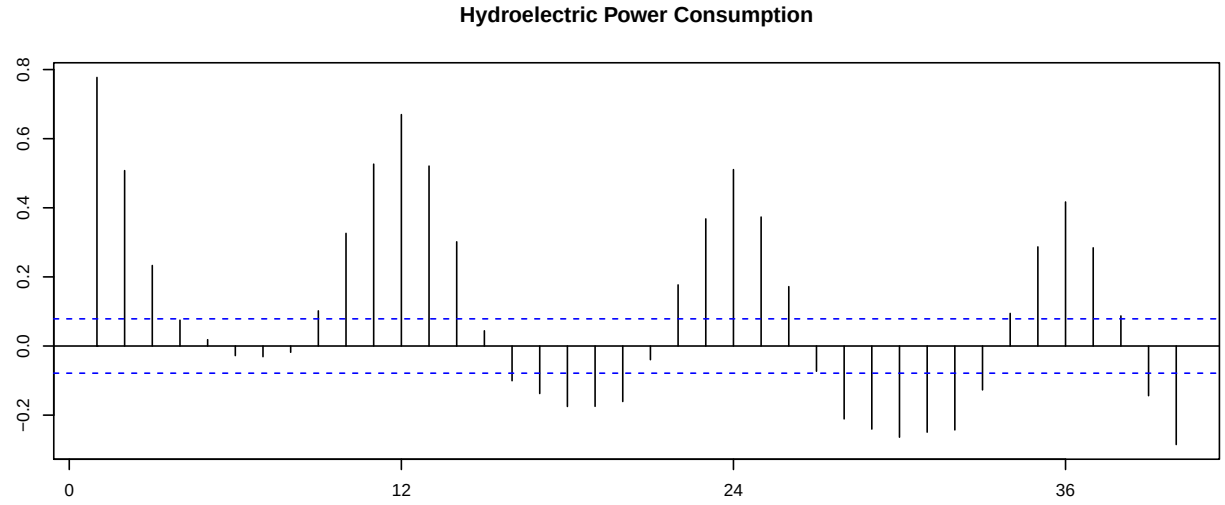
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

```
par(mar = c(3,3,3,1)) #adjusting the bottom, left, top, right in-line margins
par(mfrow=c(3,1))

Acf(ts_energy_data[, "Total Biomass Energy Production"], lag.max = 40, main = "Total Biomass Energy Pro
Acf(ts_energy_data[, "Total Renewable Energy Production"], lag.max = 40, main = "Total Renewable Energy
Acf(ts_energy_data[, "Hydroelectric Power Consumption"], lag.max = 40, main = "Hydroelectric Power Cons
```







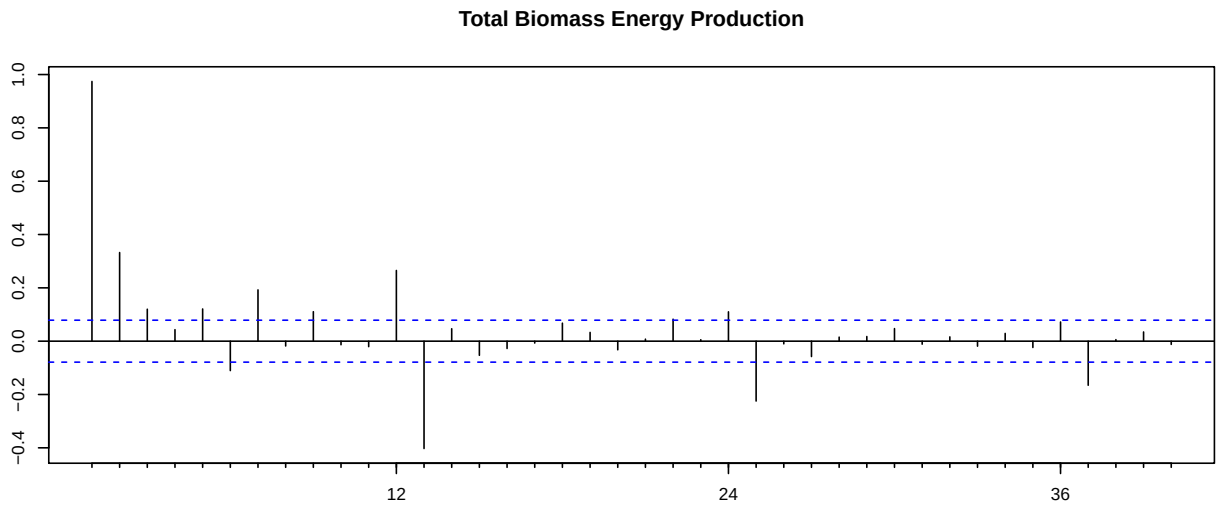
Answer 6 - For the Total Biomass and Renewable Energy Production timeseries, we see a high positive autocorrelation. However, there is a gradual decrease as the lag increases, suggesting that the influence of past values diminishes over time. - The autocorrelation plot for hydroelectric power consumption shows a wave-like pattern, with periodic spikes at a lag of 12. This indicates that there is a seasonal component to the dataset.

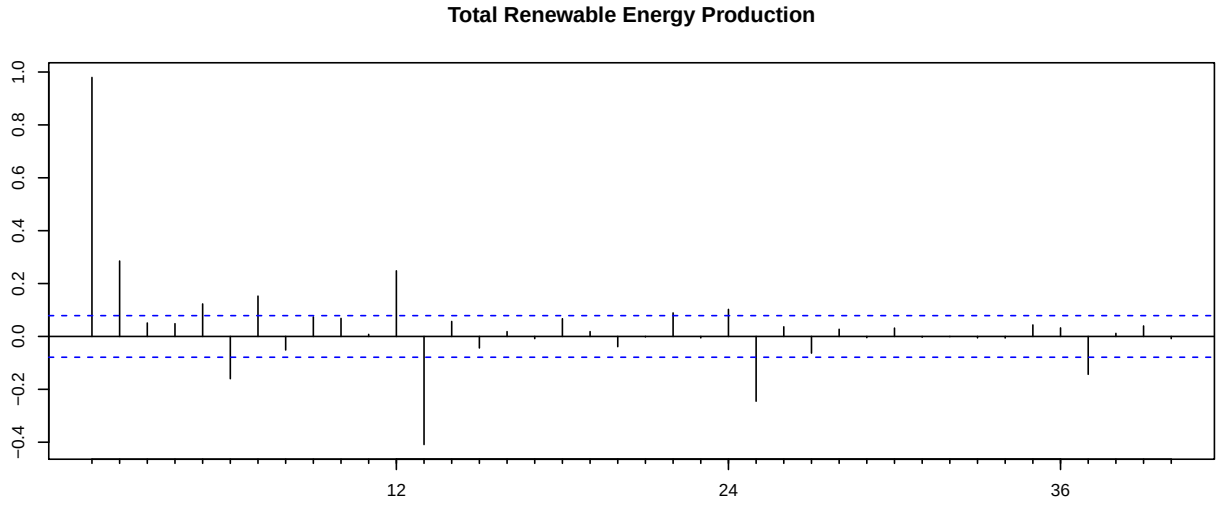
Question 7

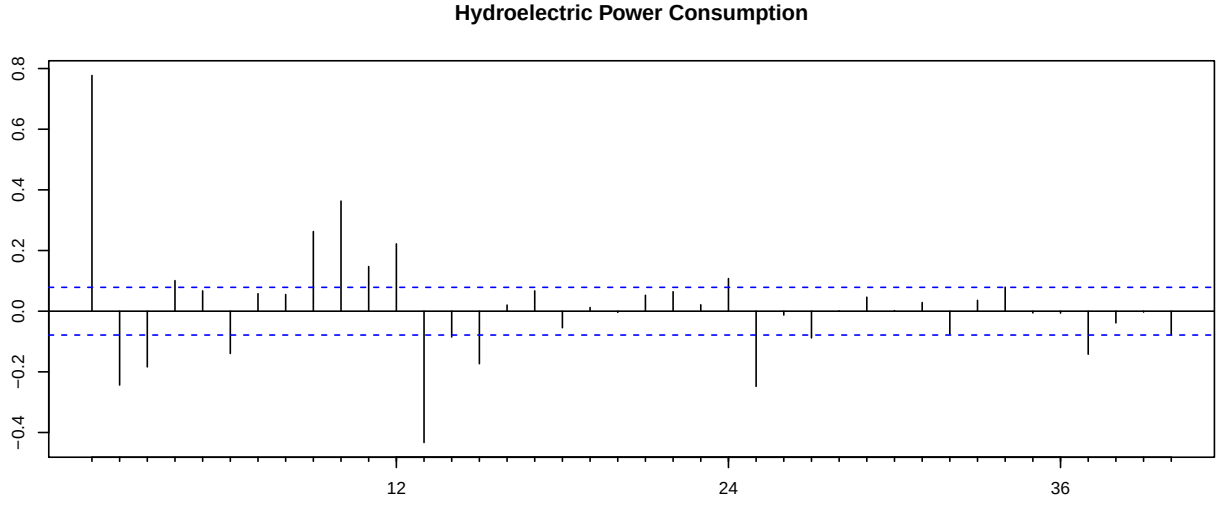
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

```
par(mar = c(3,3,3,1)) #adjusting the bottom, left, top, right in-line margins
par(mfrow=c(3,1))

Pacf(ts_energy_data[, "Total Biomass Energy Production"], lag.max = 40, main = "Total Biomass Energy Pr
Pacf(ts_energy_data[, "Total Renewable Energy Production"], lag.max = 40, main = "Total Renewable Energ
Pacf(ts_energy_data[, "Hydroelectric Power Consumption"], lag.max = 40, main = "Hydroelectric Power Cons
```







Answer 7 To get a clearer picture of the relationship between a time series and its past values, we remove the effect of intermediate lags using the Pacf function - The high positive value at lag 1 for all three indicates that the time series are highly dependent on their immediate past values, while the subsequent lags have a weaker correlation. - Most lags beyond the first one fall within the confidence bands (the blue-dashed lines), which suggests that the correlations are not statistically significant. - Some negative values at later

lag values might suggest some cyclical pattern in the data. Especially for hydroelectric power consumption, most of the first 12 values are relevant, and their wave-like pattern suggest that the data is seasonal.