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

Assignment 2 - Due date 01/27/26

Yeeun Kim

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.

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(readxl)
library(openxlsx)
library(ggplot2)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sourc 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.

```
#Importing data set
getwd()
```

```
## [1] "/Users/yeeunkim/Library/CloudStorage/OneDrive-DukeUniversity/2026 Spring/ENVIRON 797 Time Serie
```

```
energy_data1 <- read_excel(path="/Users/yeeunkim/Library/CloudStorage/OneDrive-DukeUniversity/2026 Sprin
```

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.

```
# Extract the column names from row 11
read_col_names <- read_excel(path="/Users/yeeunkim/Library/CloudStorage/OneDrive-DukeUniversity/2026 Spr

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

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

```
## # 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>, ...
```

```
#Select the columns
energy_select <- energy_data1 %>%
  select(1,4:6)
head(energy_select)
```

```
## # A tibble: 6 x 4
##   Month               ‘Total Biomass Energy Production‘ Total Renewable Energy~1
##   <dttm>                                          <dbl>                    <dbl>
## 1 1973-01-01 00:00:00                              130.                     220.
## 2 1973-02-01 00:00:00                              117.                     197.
## 3 1973-03-01 00:00:00                              130.                     219.
## 4 1973-04-01 00:00:00                              126.                     209.
## 5 1973-05-01 00:00:00                              130.                     216.
## 6 1973-06-01 00:00:00                              126.                     208.
## # i abbreviated name: 1: ‘Total Renewable Energy Production‘
## # i 1 more variable: ‘Hydroelectric Power Consumption‘ <dbl>
```

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_data <- ts(energy_select[,2:4],start = c(1973,1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
ts_mean <- colMeans(ts_energy_data, na.rm = TRUE)
ts_sd <- apply(ts_energy_data, 2, sd, na.rm = TRUE)

ts_mean
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                         286.04893                         409.19521
##   Hydroelectric Power Consumption
##                          79.35682
```

```
ts_sd
```

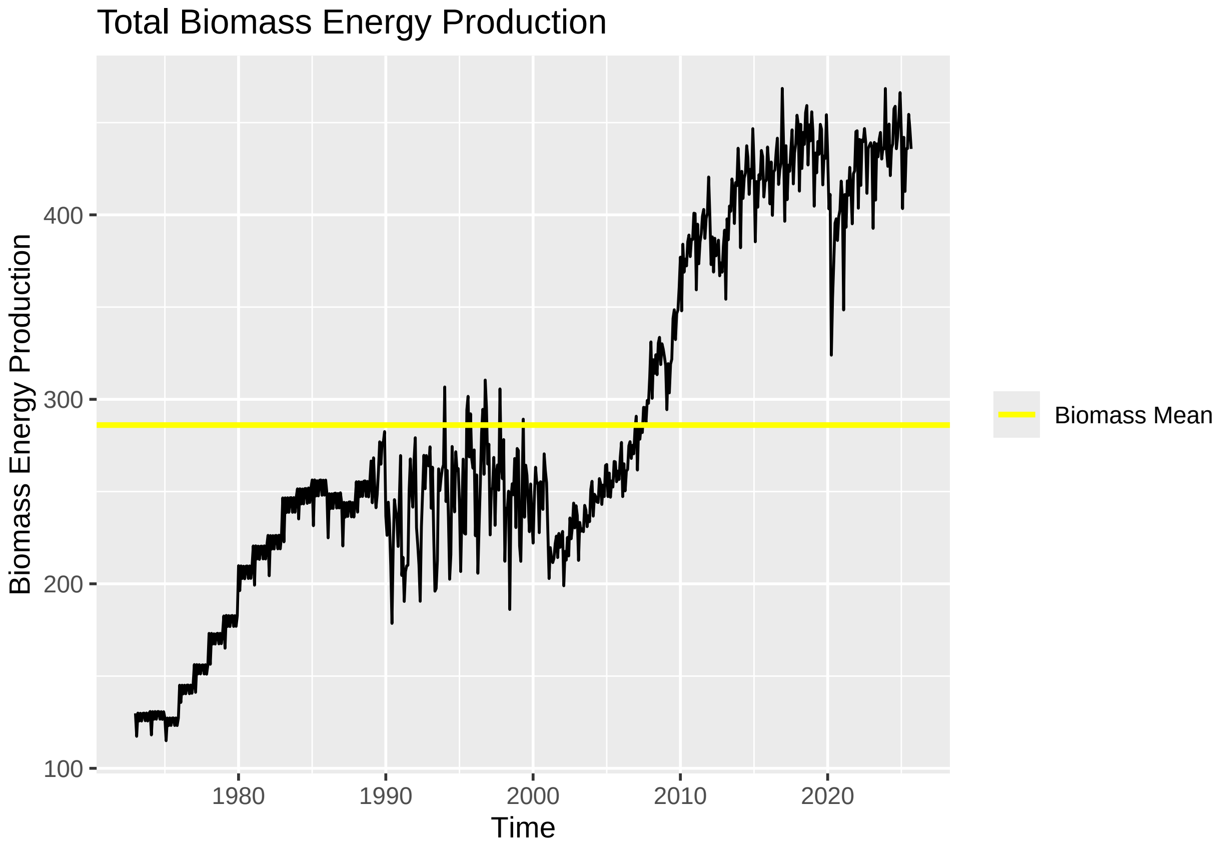
```
##   Total Biomass Energy Production Total Renewable Energy Production
##                          96.21209                         151.42232
##   Hydroelectric Power Consumption
##                          14.12020
```

```
biomass_mean <-ts_mean[1]
renewable_mean <-ts_mean[2]
hydro_mean<-ts_mean[3]
```

Question 4

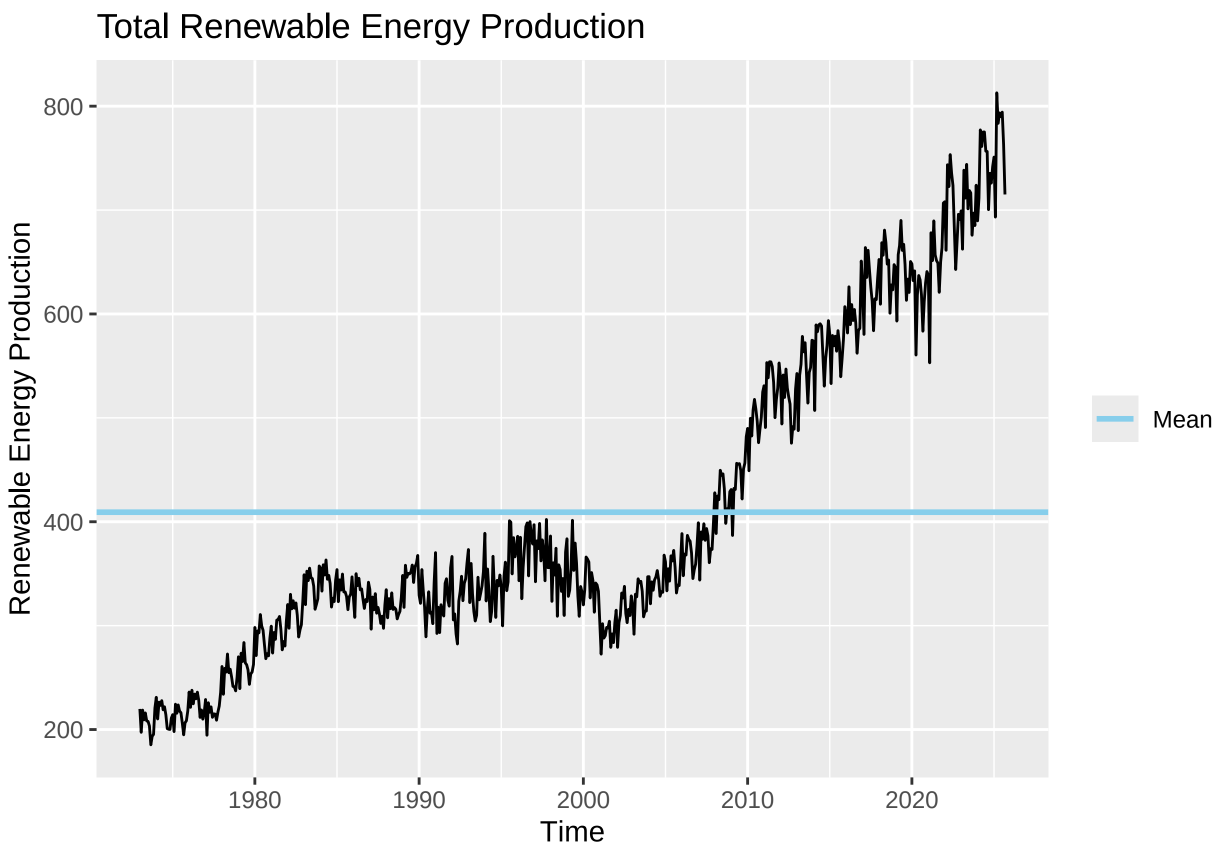
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
#Biomass Energy
autoplot(ts_energy_data[,1]) +
  labs(
    title = "Total Biomass Energy Production",
    x = "Time",
    y = "Biomass Energy Production",
    color = NULL) +
  geom_hline(aes(yintercept = biomass_mean, color="Biomass Mean"), size = 1) +
  scale_color_manual(values = c("Biomass Mean"="yellow"))
```

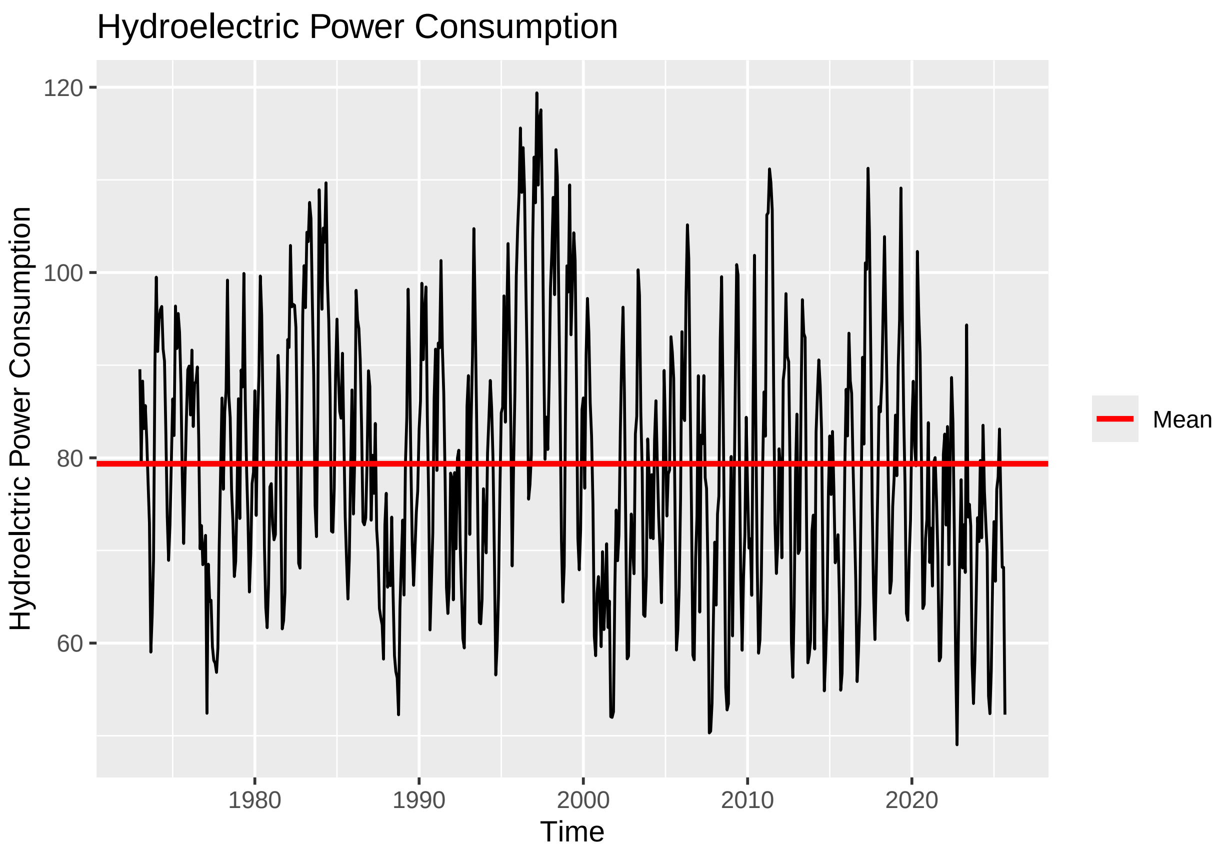


```
#autoplot(ts_energy_data[,1]) +
#  xlab("Time") +
#  ylab("Biomass Energy Production") +
#  geom_hline(aes(yintercept = biomass_mean, color = "Biomass Mean"), size = 1) +
#  scale_color_manual(name = "Legend", values = c("Biomass Mean" = "pink"))

#Renewable Energy
autoplot(ts_energy_data[,2]) +
  labs(
    title = "Total Renewable Energy Production",
    x = "Time",
    y = "Renewable Energy Production",
    color = NULL) +
  geom_hline(aes(yintercept = renewable_mean, color="Mean"), size = 1) +
  scale_color_manual(values = c("Mean"="skyblue"))
```



```r
#Hydroelectric Power Consumption
autoplot(ts_energy_data[,3]) +
  labs(
    title = "Hydroelectric Power Consumption",
    x = "Time",
    y = "Hydroelctric Power Consumption",
    color = NULL) +
  geom_hline(aes(yintercept = hydro_mean, color="Mean"), size = 1) +
  scale_color_manual(values = c("Mean"="red"))
```



Question 5

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

Biomass energy production and renewable energy production have a strong positive correlation, which means they share a trend over time. However, hydroelectric power consumption has weak negative correlations with biomass energy production and renewable energy production, indicating no common trend.

```
#Biomass energy and Renewable energy production
cor(ts_energy_data[,1], ts_energy_data[,2], use="complete.obs")
```

```
## [1] 0.9652985
```

```
#Biomass energy production and hydroelectric power consumption
cor(ts_energy_data[,1], ts_energy_data[,3], use="complete.obs")
```

```
## [1] -0.1347374
```

```
#Renewable energy production and hydroelectric power consumption
cor(ts_energy_data[,2], ts_energy_data[,3], use="complete.obs")
```

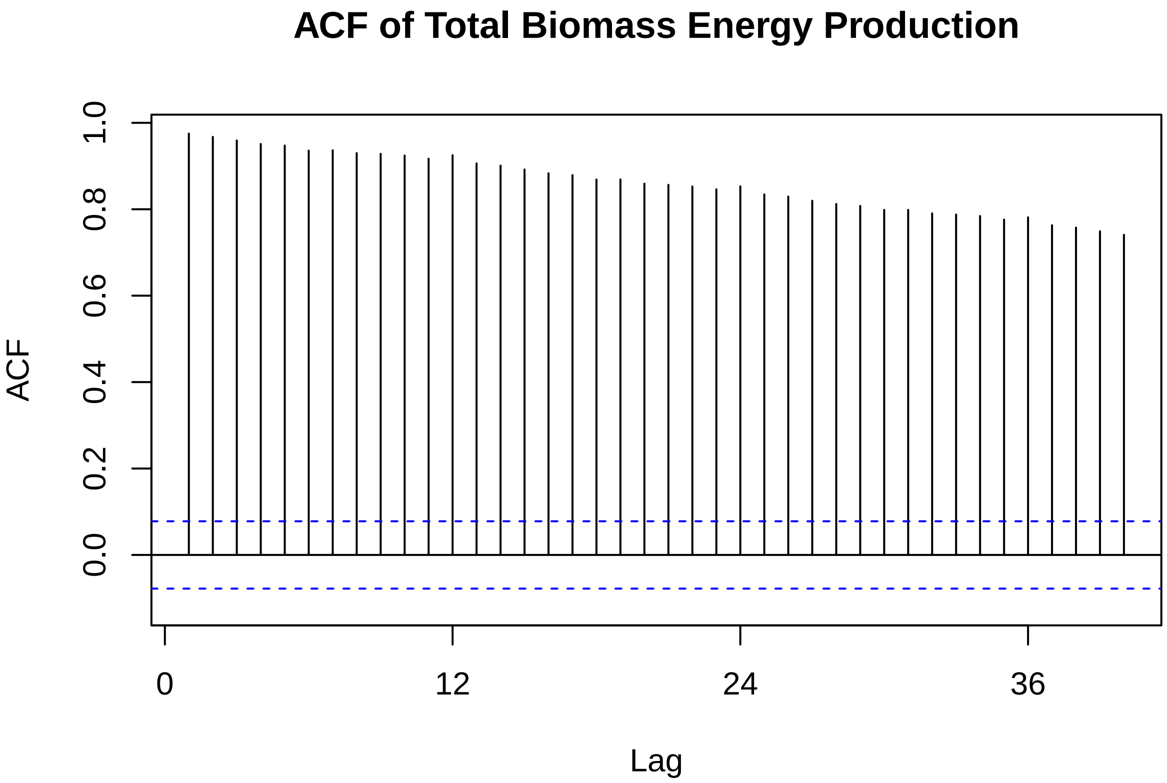
```
## [1] -0.05842436
```

Question 6

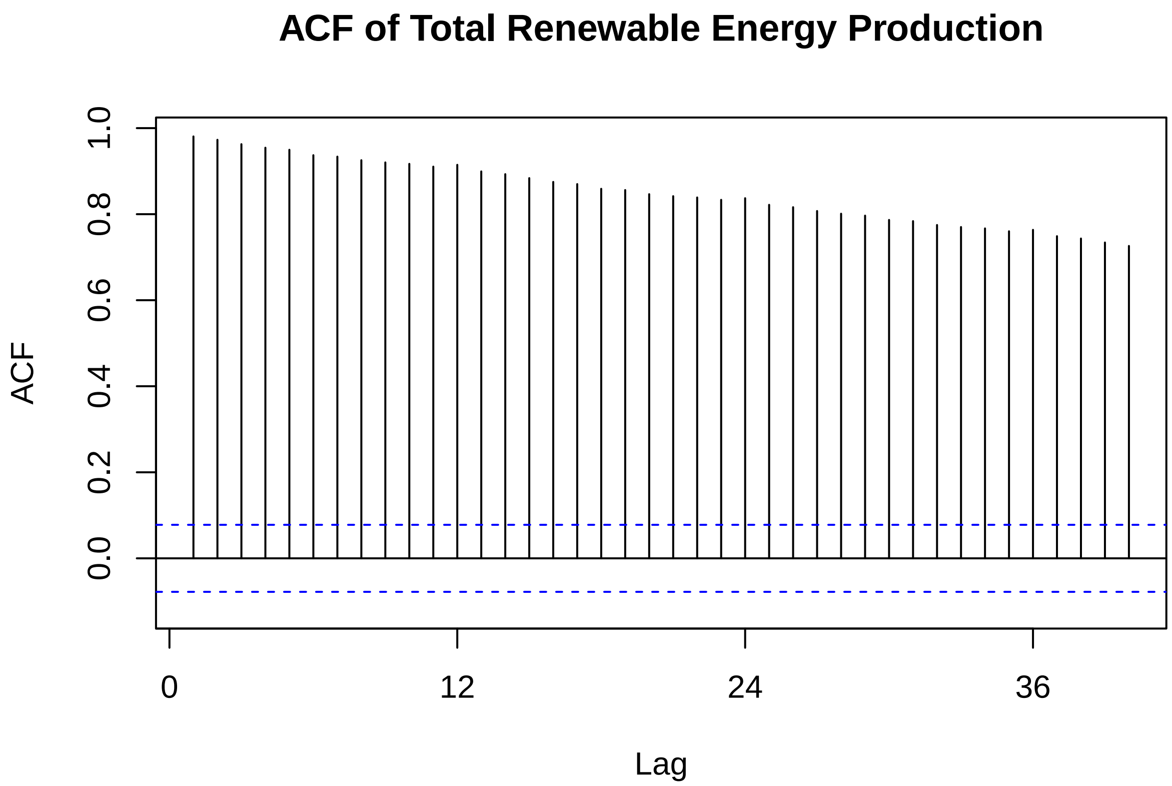
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

The ACF plots of biomass and renewable energy production decay slowly, indicating that these series are non-stationary due to a strong trend. However, hydroelectric power consumption shows seasonality as its autocorrelations have an oscillatory pattern. Biomass and renewable energy production show similar behavior, while hydroelectric power consumption behaves differently.

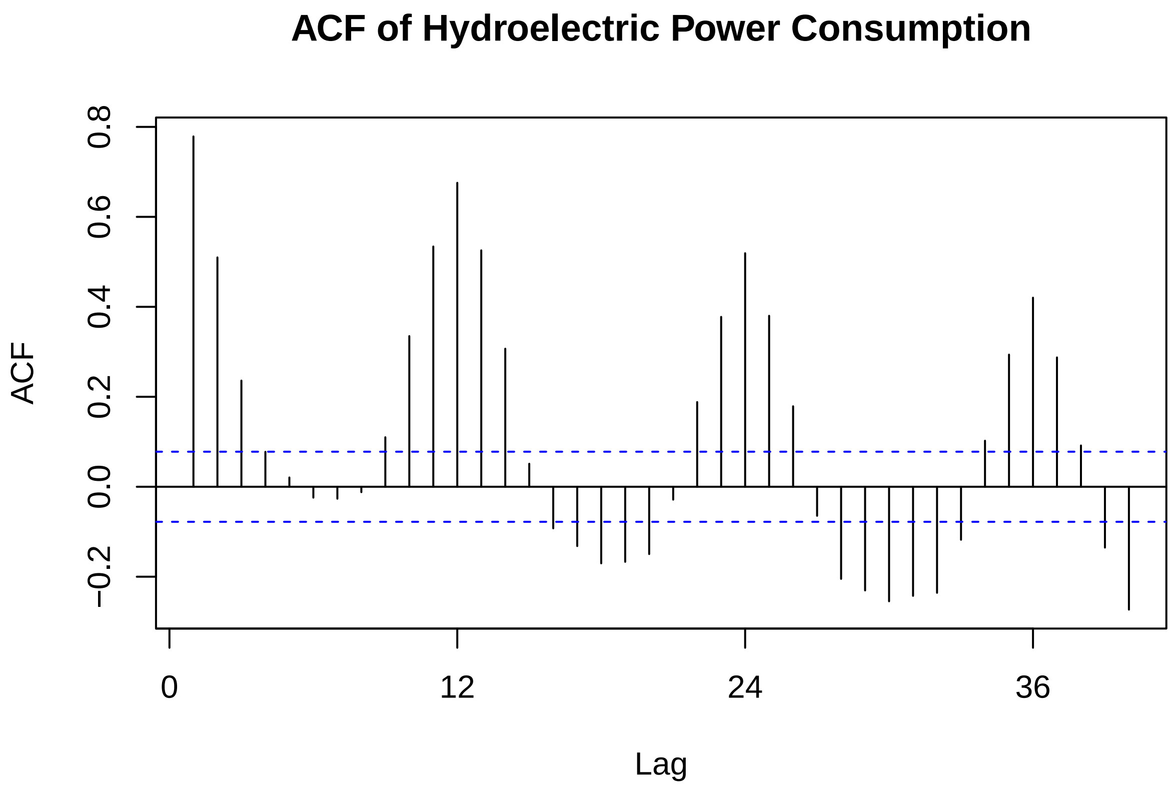
```r
biomass_acf= Acf(ts_energy_data[,1], lag.max = 40, main="ACF of Total Biomass Energy Production")
```



```r
renewable_acf= Acf(ts_energy_data[,2], lag.max = 40, main="ACF of Total Renewable Energy Production")
```



```
hydro_acf= Acf(ts_energy_data[,3], lag.max = 40, main="ACF of Hydroelectric Power Consumption")
```



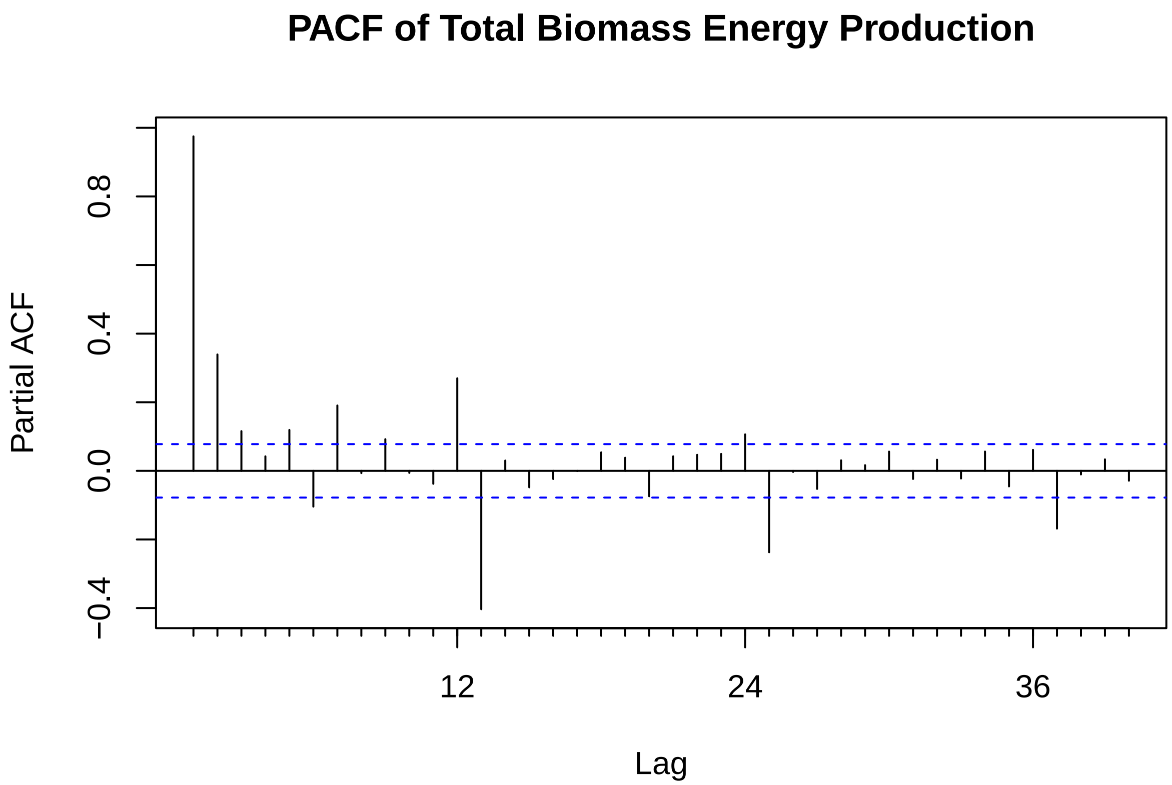
ACF of Hydroelectric Power Consumption

Question 7

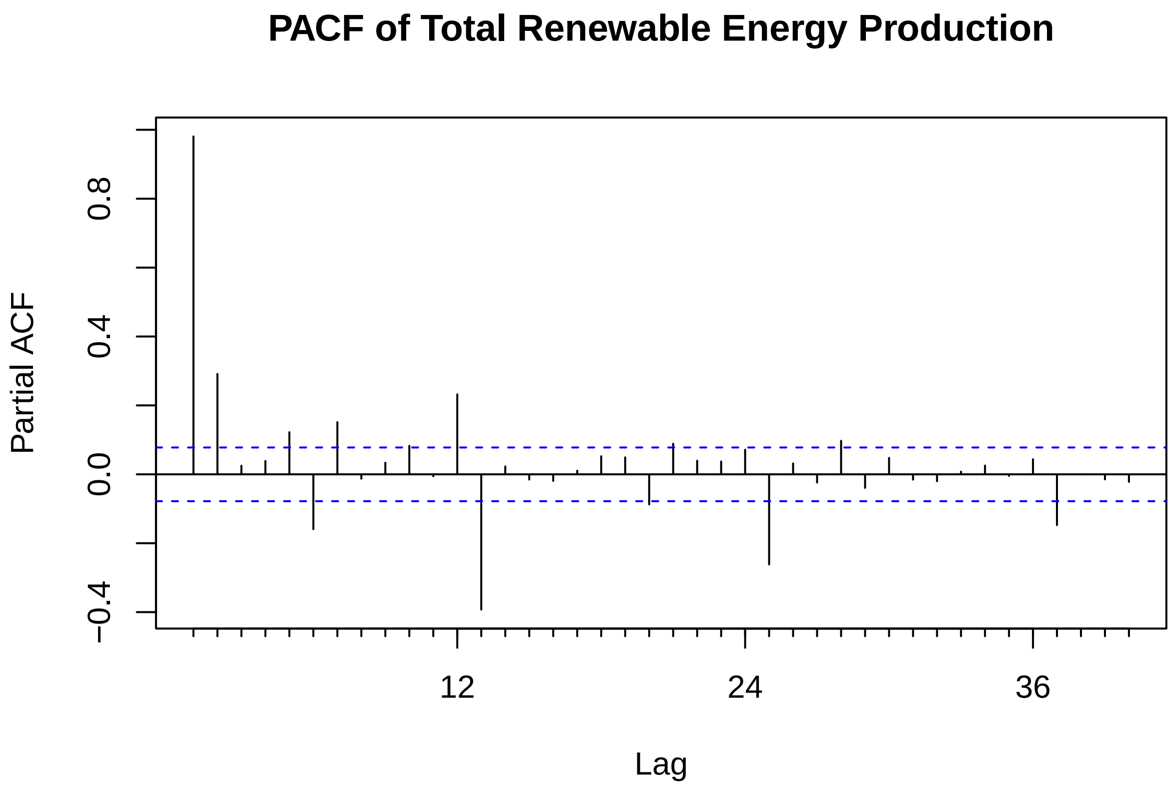
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

Fpr the PACF plots of biomass and renewable energy production, only the first few lags are significant, indicating that the strong persistence in the ACF is caused by indirect effects. However, the PACF plot of hydroelectric power consumption shows significant partial autocorrelations at seasonal lags, proving the seasonality.

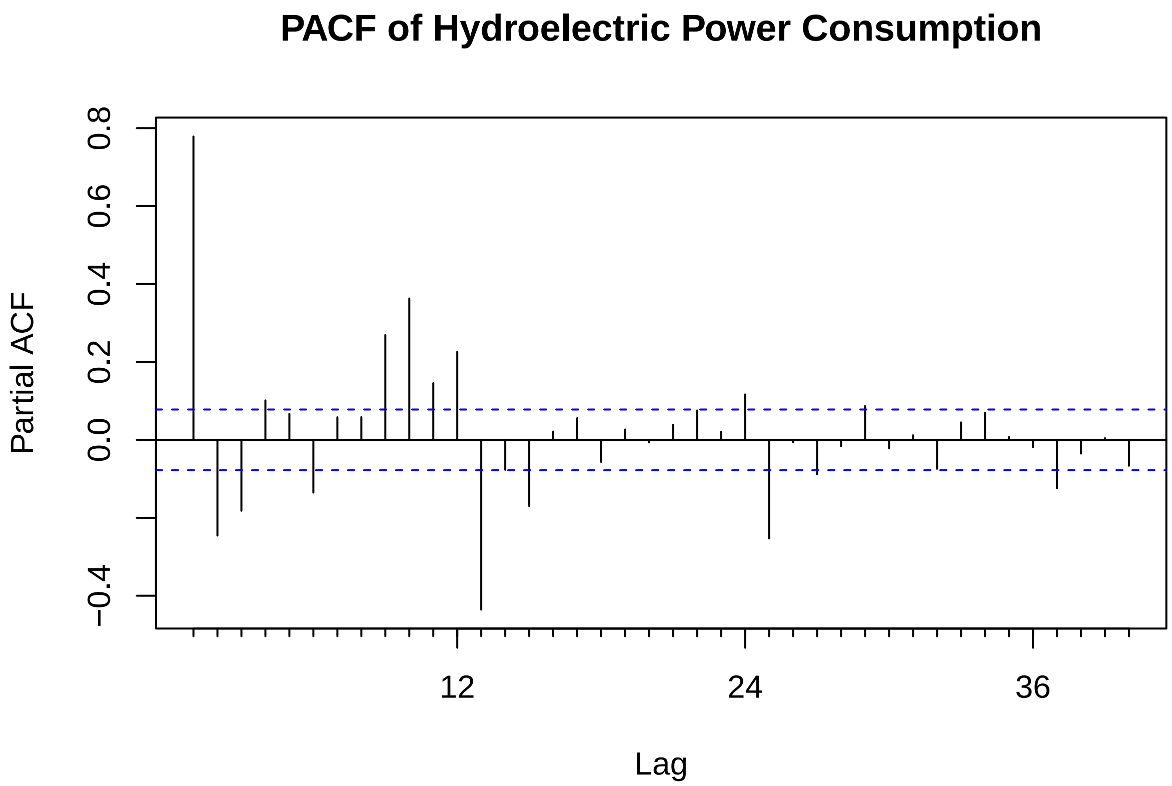
```
biomass_pacf= Pacf(ts_energy_data[,1], lag.max = 40, main="PACF of Total Biomass Energy Production")
```



```
renewable_pacf= Pacf(ts_energy_data[,2], lag.max = 40, main="PACF of Total Renewable Energy Production"
```



```
hydro_pacf= Pacf(ts_energy_data[,3], lag.max = 40, main="PACF of Hydroelectric Power Consumption")
```


PACF of Hydroelectric Power Consumption
Partial ACF
0.8
0.6
0.4
0.2
0.0
−0.4
12
24
36
Lag