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

Assignment 2 - Due date 01/22/26

Sylvia Hipp

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.

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)

# additional packages
library(readxl)
library(lubridate)
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 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
filepath <- "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"

# read data
```

```
energy_data_raw <- read_excel(path = filepath,
                              sheet = "Monthly Data",
                              skip = 12,
                              col_names = FALSE) %>%
  tibble()

# read row with column names
energy_data_colnames <- read_excel(path = filepath,
                                   sheet = "Monthly Data",
                                   skip = 10,
                                   col_names = FALSE,
                                   n_max = 1)

# assign column names
colnames(energy_data_raw) <- energy_data_colnames
```

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_data_clean <- energy_data_raw %>%
  janitor::clean_names() %>%     # for easier column names to work with
  select(month, c(total_biomass_energy_production:hydroelectric_power_consumption))

energy_data_clean %>% head()
```

```
## # A tibble: 6 x 4
##   month               total_biomass_energy_production total_renewable_energy_p~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().

```
# get starting point
start_year <- year(first(energy_data_clean$month))
start_month <- month(first(energy_data_clean$month))

# create time series object
energy_data_ts <- ts(energy_data_clean,
                     start = c(start_year, start_month),
                     frequency = 12)  # monthly time series
```

Question 3

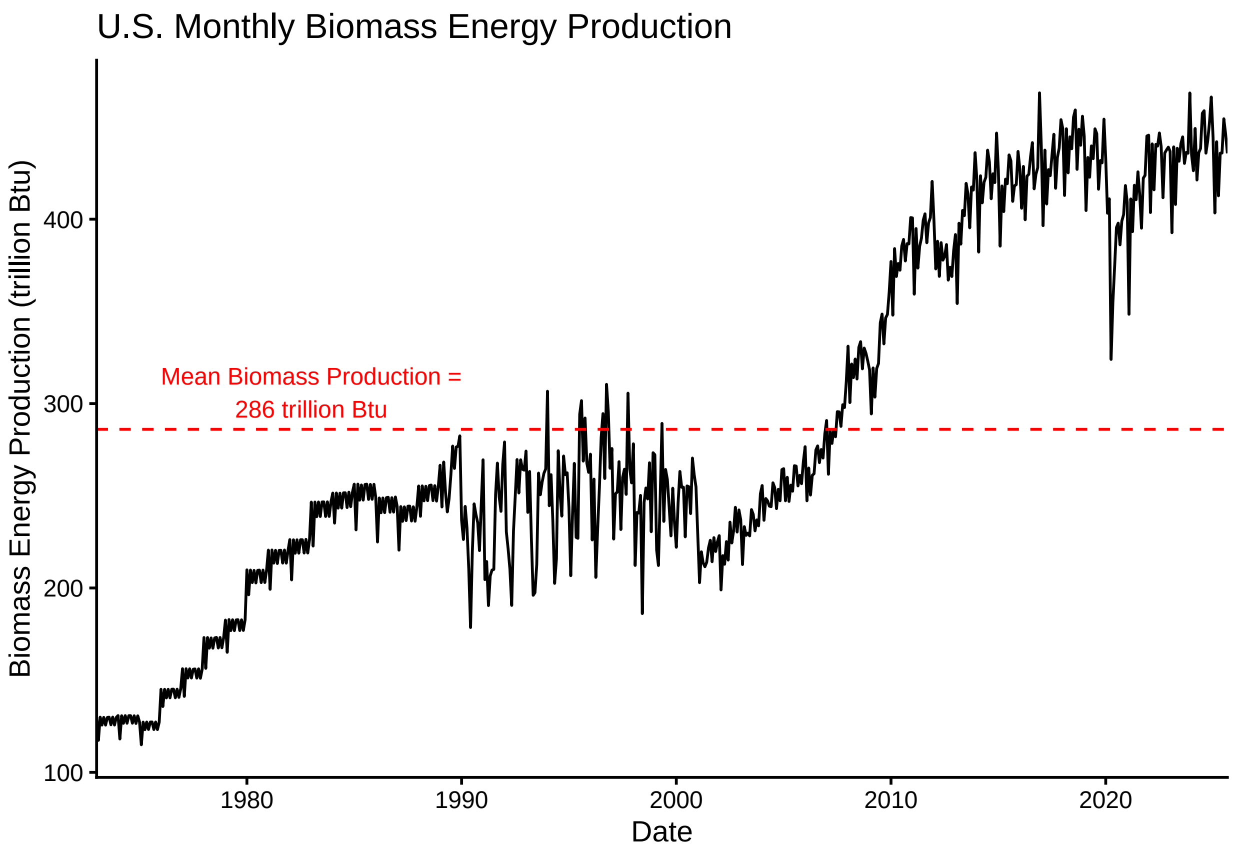
Compute mean and standard deviation for these three series.

```
# biomass energy production
biomass_mean <- mean(energy_data_ts[,"total_biomass_energy_production"])  # 286.05 trillion Btu
biomass_sd <- sd(energy_data_ts[,"total_biomass_energy_production"])    # 96.21 trillion Btu

# total renewable energy production
re_mean <- mean(energy_data_ts[,"total_renewable_energy_production"])  # 409.20 trillion Btu
re_sd <- sd(energy_data_ts[,"total_renewable_energy_production"])    # 151.42 trillion Btu

# hydroelectric power consumption
hydro_mean <- mean(energy_data_ts[,"hydroelectric_power_consumption"])  # 79.36 trillion Btu
hydro_sd <- sd(energy_data_ts[,"hydroelectric_power_consumption"])    # 14.12 trillion Btu
```

Question 4

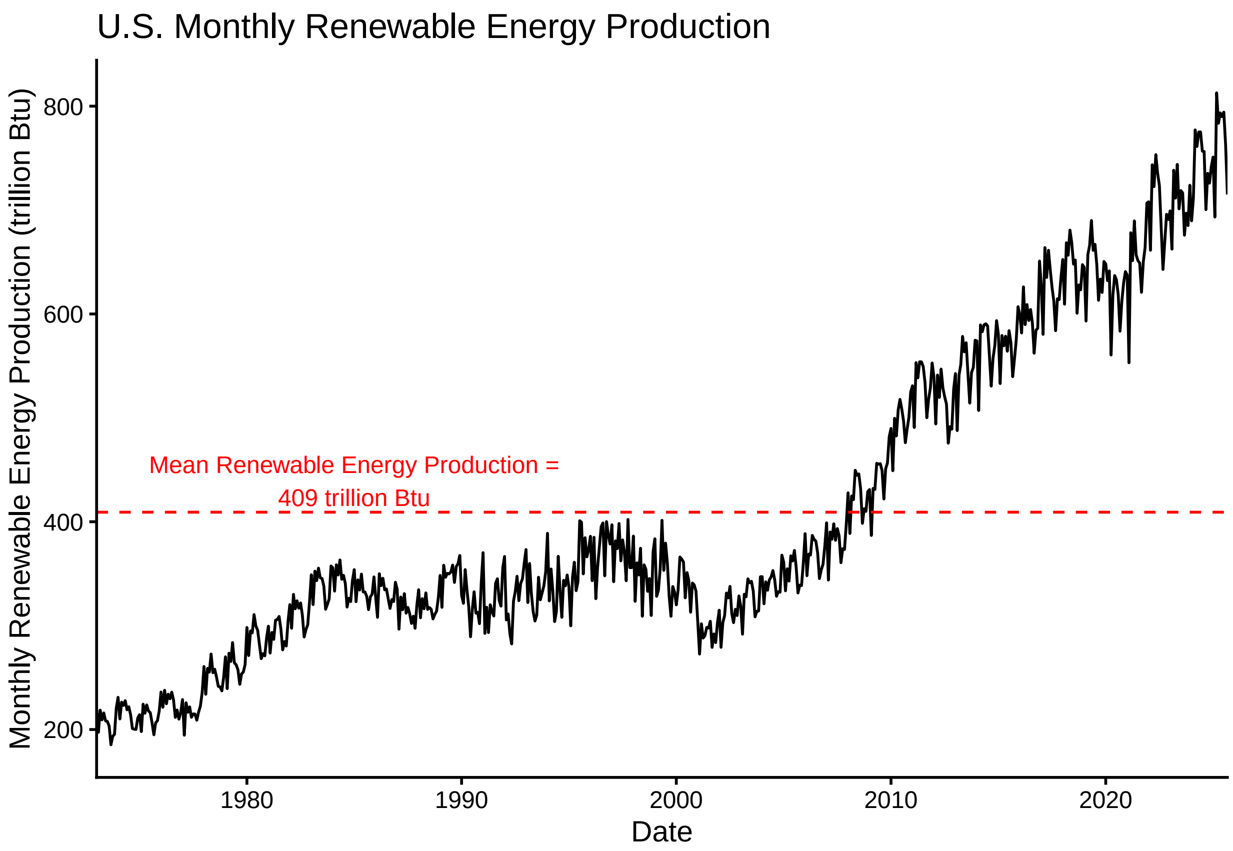
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
# Biomass
autoplot(energy_data_ts[,"total_biomass_energy_production"]) +
  geom_hline(yintercept = biomass_mean, color = "red", linetype = 2) +
  annotate("text", x = 1983, y = biomass_mean+20,
           size = 3, color = "red",
           label = "Mean Biomass Production =\n286 trillion Btu") +
  theme_classic() +
  scale_x_continuous(expand = c(0,0)) +
  labs(x = "Date",
       y = "Biomass Energy Production (trillion Btu)",
       title = "U.S. Monthly Biomass Energy Production")
```



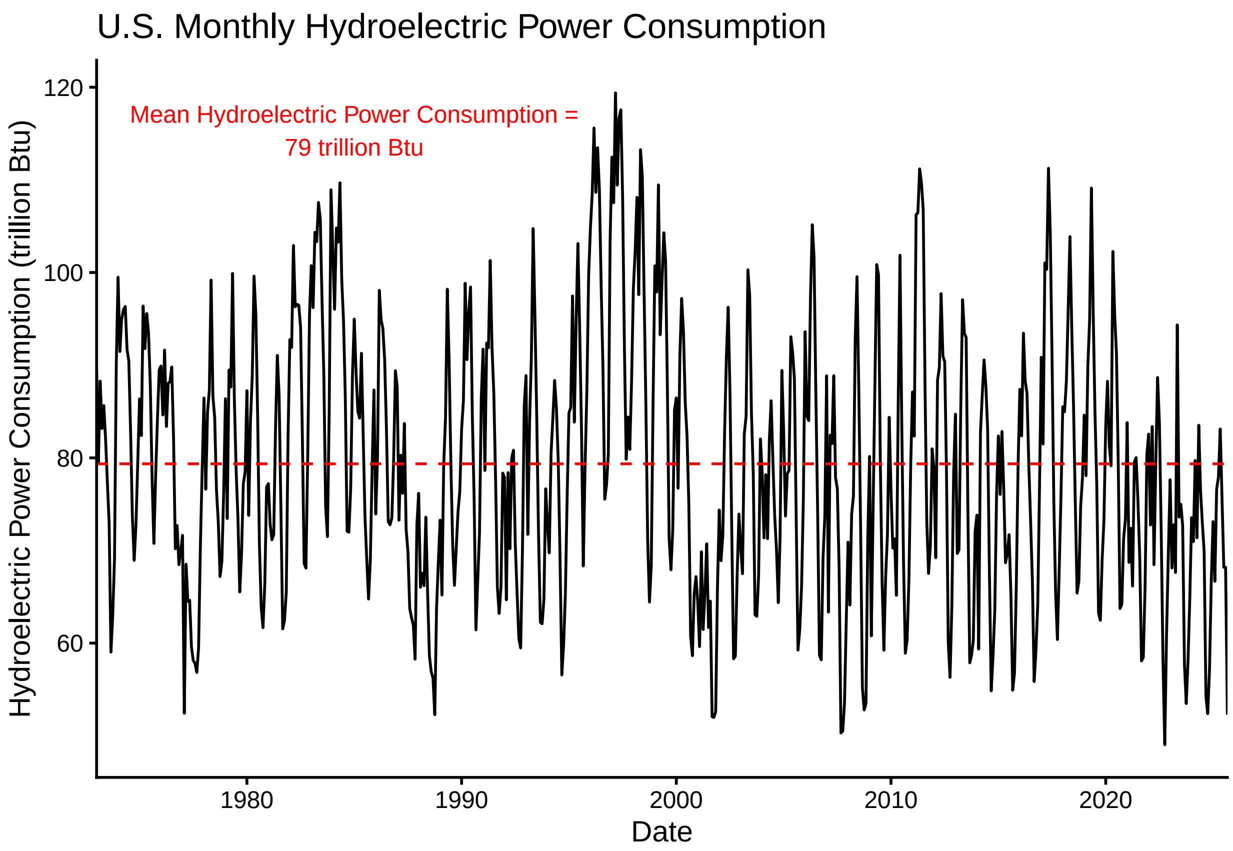
Biomass energy production has been increasing steadily over time, with some month-to-month volatility between 1990-2000 and some drop in production around 2020, likely due to demand shifts around the pandemic.

```r
# Renewable Energy
autoplot(energy_data_ts[,"total_renewable_energy_production"]) +
  geom_hline(yintercept = re_mean, color = "red", linetype = 2) +
  annotate("text", x = 1985, y = re_mean+30,
           size = 3, color = "red",
           label = "Mean Renewable Energy Production =\n409 trillion Btu") +
  theme_classic() +
  scale_x_continuous(expand = c(0,0)) +
  labs(x = "Date",
       y = "Monthly Renewable Energy Production (trillion Btu)",
       title = "U.S. Monthly Renewable Energy Production")
```



> Renewable energy production has also been increasing steadily over time, with an increase in month production accelerating greatly beginning around 2000. There appears to be less volatility month-to-month in this series than we saw in the biomass production series. There is also overall a much greater range in monthly renewable production observed over the period compared to biomass production.

```
# Hydroelectric Power Consumption
autoplot(energy_data_ts[,"hydroelectric_power_consumption"]) +
  geom_hline(yintercept = hydro_mean, color = "red", linetype = 2) +
  annotate("text", x = 1985, y = hydro_mean+36,
           size = 3, color = "red",
           label = "Mean Hydroelectric Power Consumption =\n79 trillion Btu") +
  theme_classic() +
  scale_x_continuous(expand = c(0,0)) +
  labs(x = "Date",
       y = "Hydroelectric Power Consumption (trillion Btu)",
       title = "U.S. Monthly Hydroelectric Power Consumption")
```

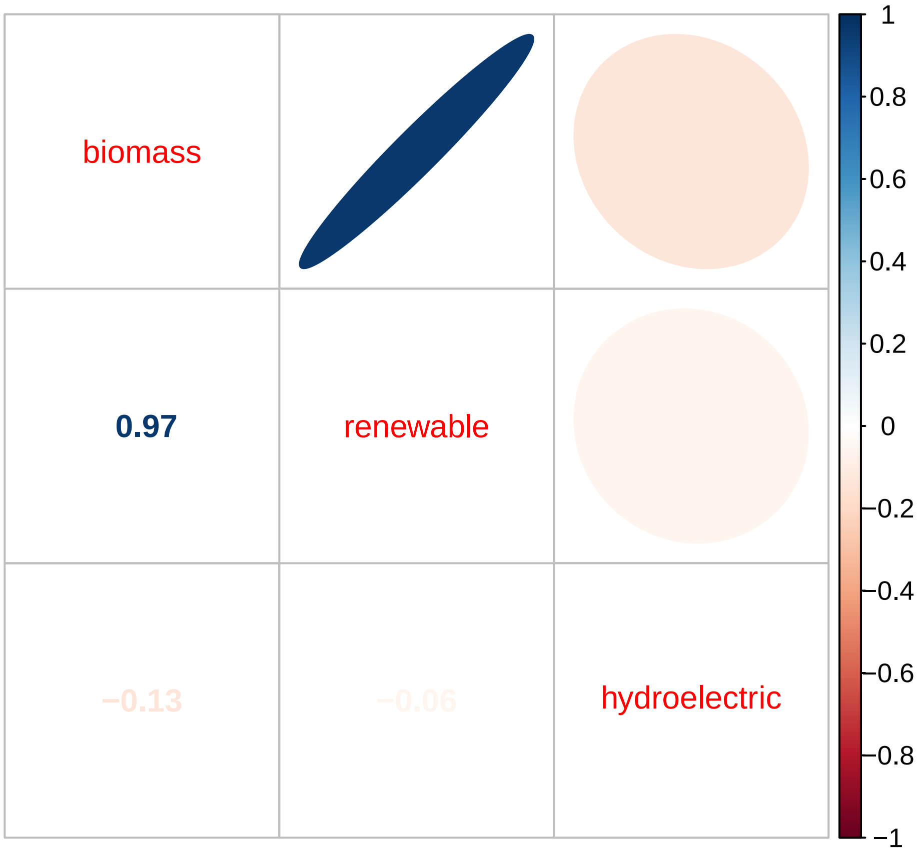


Hydroelectric power consumption has been relatively constant over the whole period of the data (1973-2025) with some fluctuation month-to-month. There is a much smaller range and standard deviation in this series compared to biomass and renewable energy production. Monthly fluctuations could potentially be caused by seasonal changes affecting reservoir levels and thus hydroelectric power supply.

Question 5

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

```
# calc correlation
energy_corr <- cor(
  energy_data_clean %>%
    select(c(total_biomass_energy_production:hydroelectric_power_consumption)) %>%
    rename("biomass" = 1, "renewable" = 2, "hydroelectric" = 3)
  )

# plot correlation
corrplot::corrplot.mixed(energy_corr, upper = "ellipse")
```

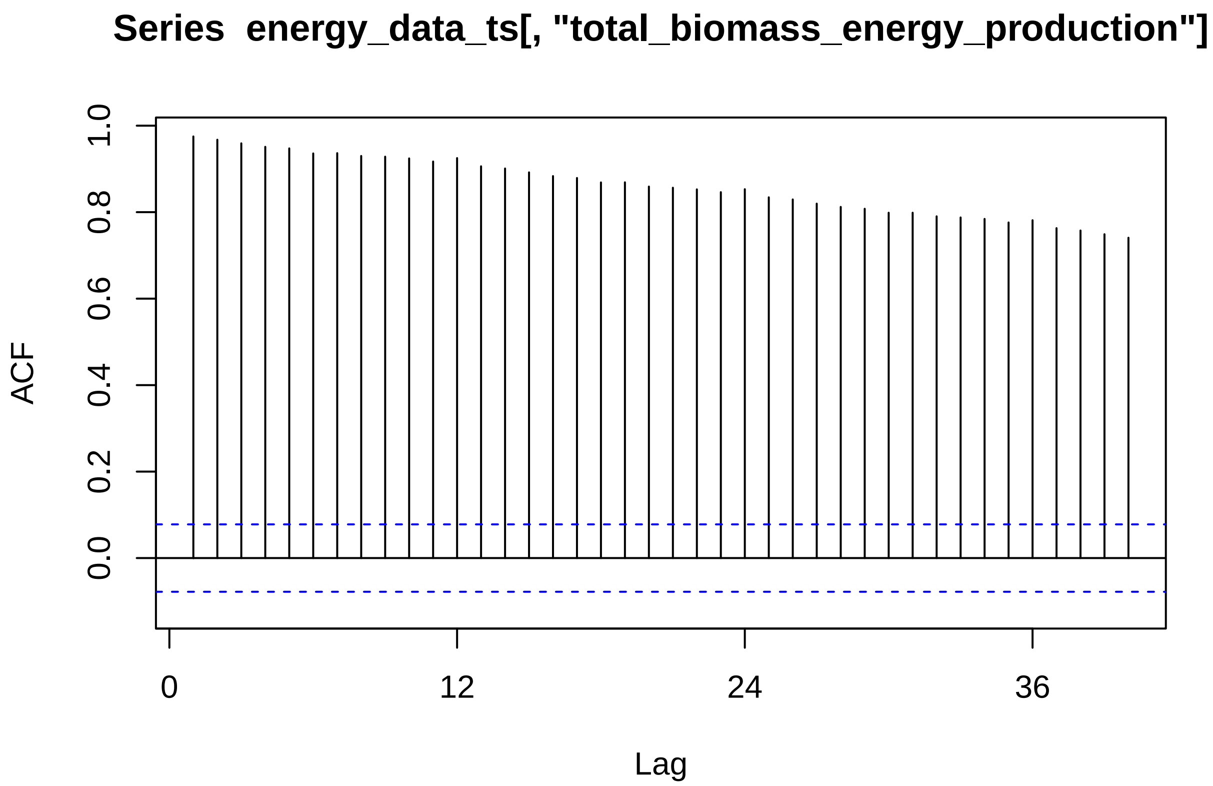


Monthly biomass energy production has a very strong positive correlation with renewable energy production (coefficient = 0.97). This makes sense as both energy sources may be increasing production to meet growing energy demand over time. Neither biomass nor renewable production is significantly correlated with hydropower consumption (coefficient = -0.13 and -0.06, respectively). This could potentially be because as production of some resources grows, consumption of a specific energy resource will not necessarily change and may even shift away to other more readily available resources. Additionally, as mentioned above, the hydroelectric consumption series shows a seasonal component that isn't as clearly present in the other series, which may impact their correlation with one another.

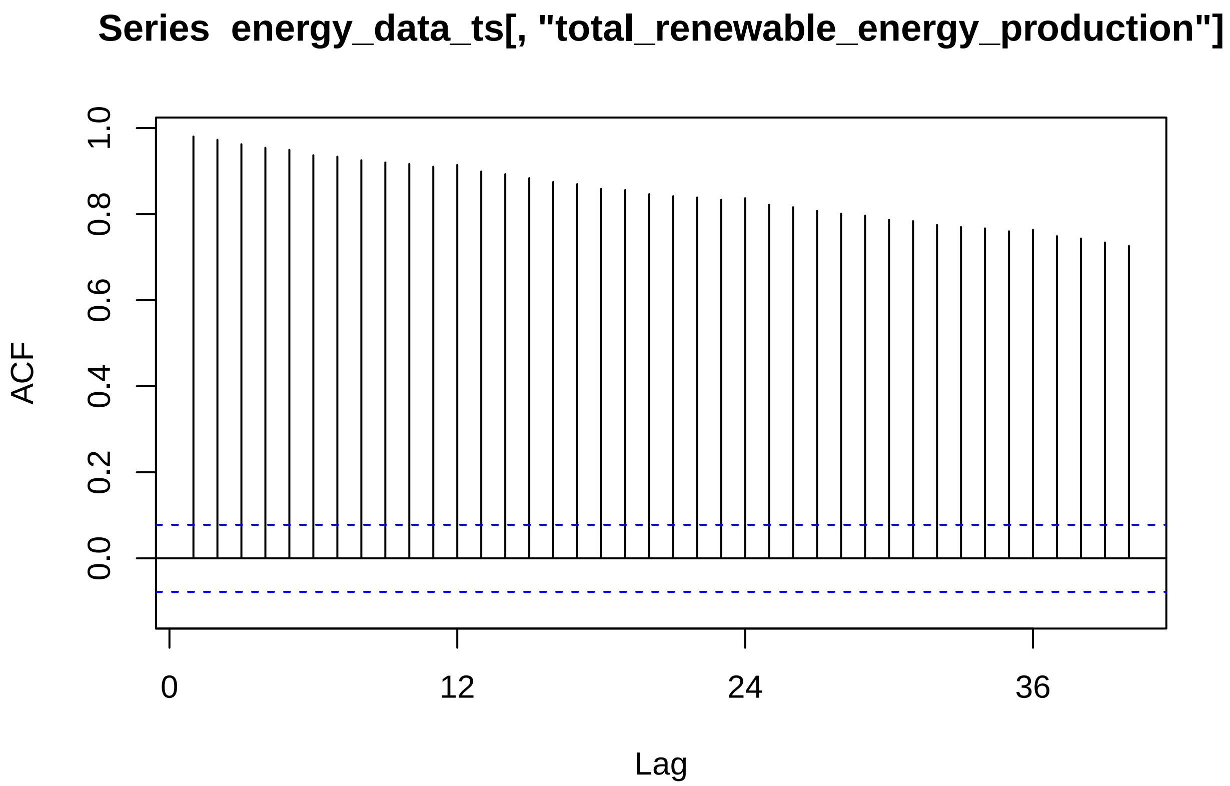
Question 6

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?

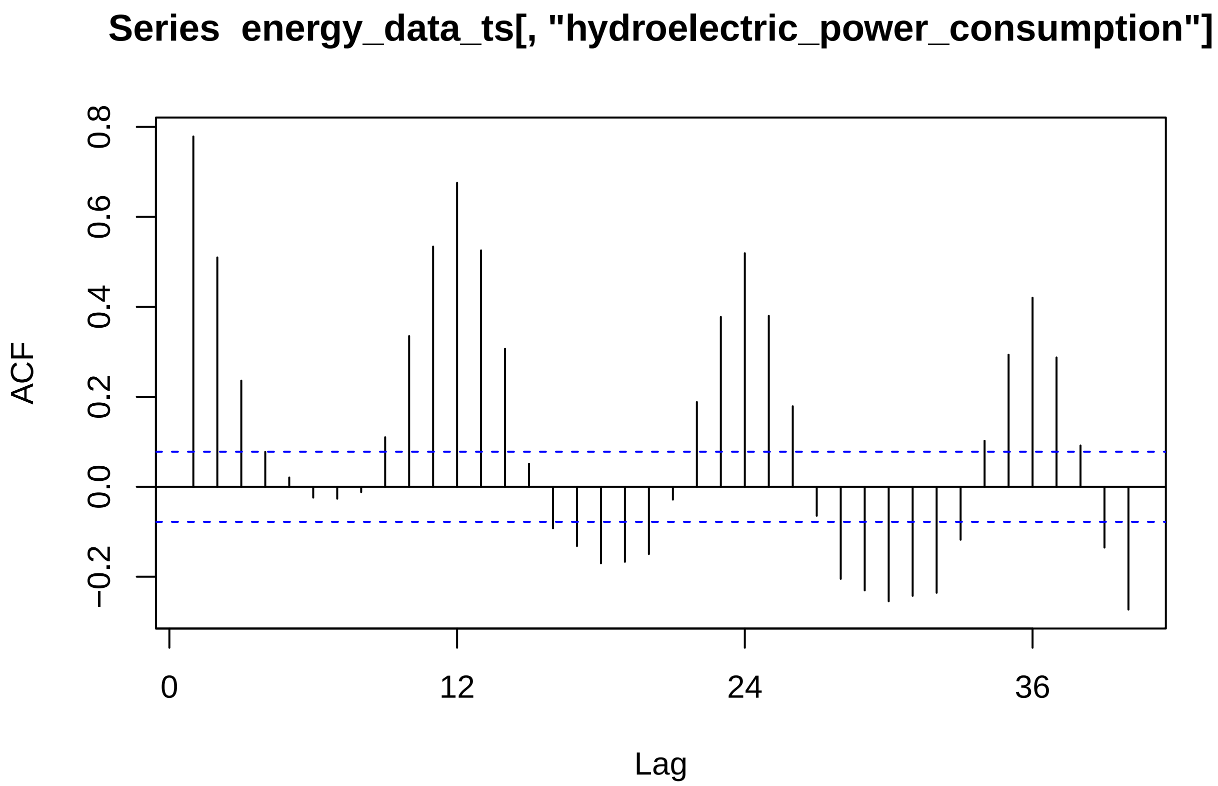
```r
biomass_acf <- Acf(energy_data_ts[,"total_biomass_energy_production"], lag.max = 40)
```



```r
renewable_acf <- Acf(energy_data_ts[,"total_renewable_energy_production"], lag.max = 40)
```



```
hydro_acf <- Acf(energy_data_ts[,"hydroelectric_power_consumption"], lag.max = 40)
```

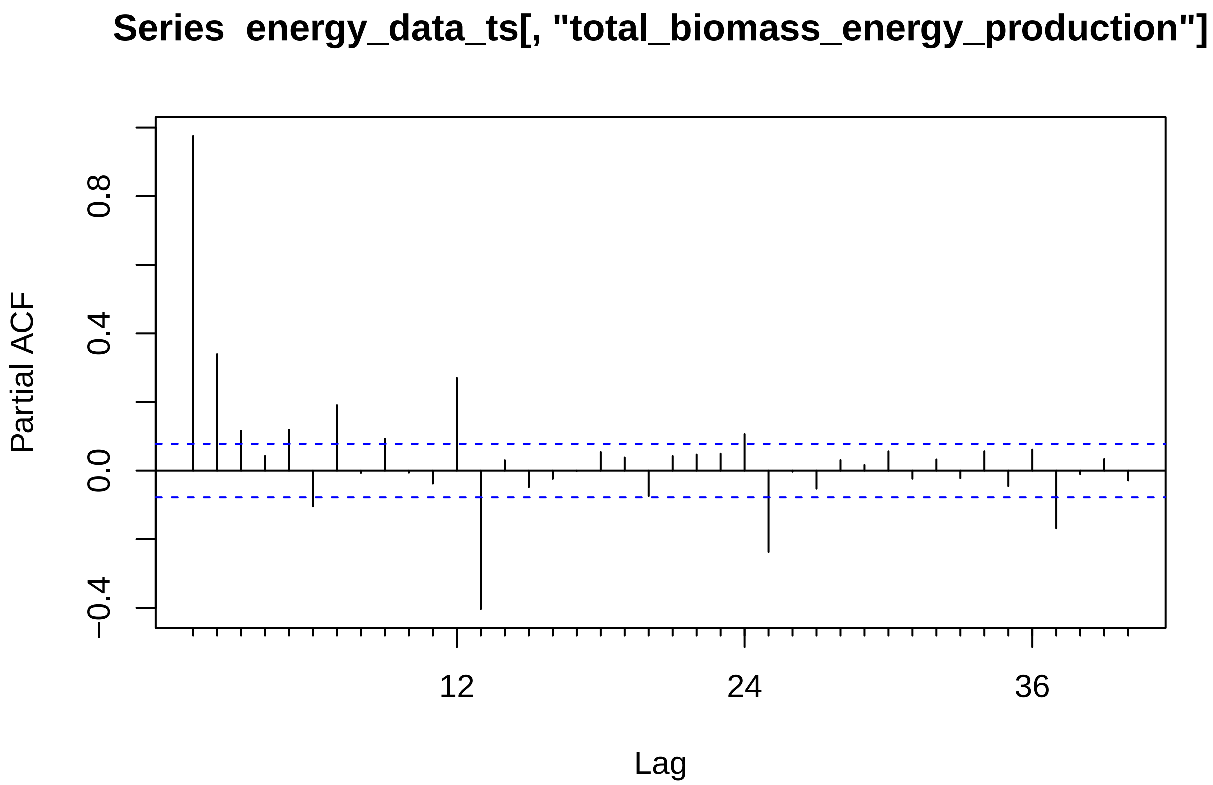


Biomass ACF shows very high autocorrelation up to 40 lags with steady but small decaying between each lag. This might suggest that biomass energy production has a strong memory month-to-month. Renewable energy production shows the same pattern as biomass, with very high autocorrelation and steady decaying. Strong memory may suggest that as capacity is added over time, production will be similar month-to-month without strong variation. Hydroelectric power consumption shows a very different trend: the ACF suggests seasonality in the data as autocorrelation oscillates between positive and negative values every ~12 lags. This could be explained by seasonal variation in reservoir levels used for hydroelectric production.

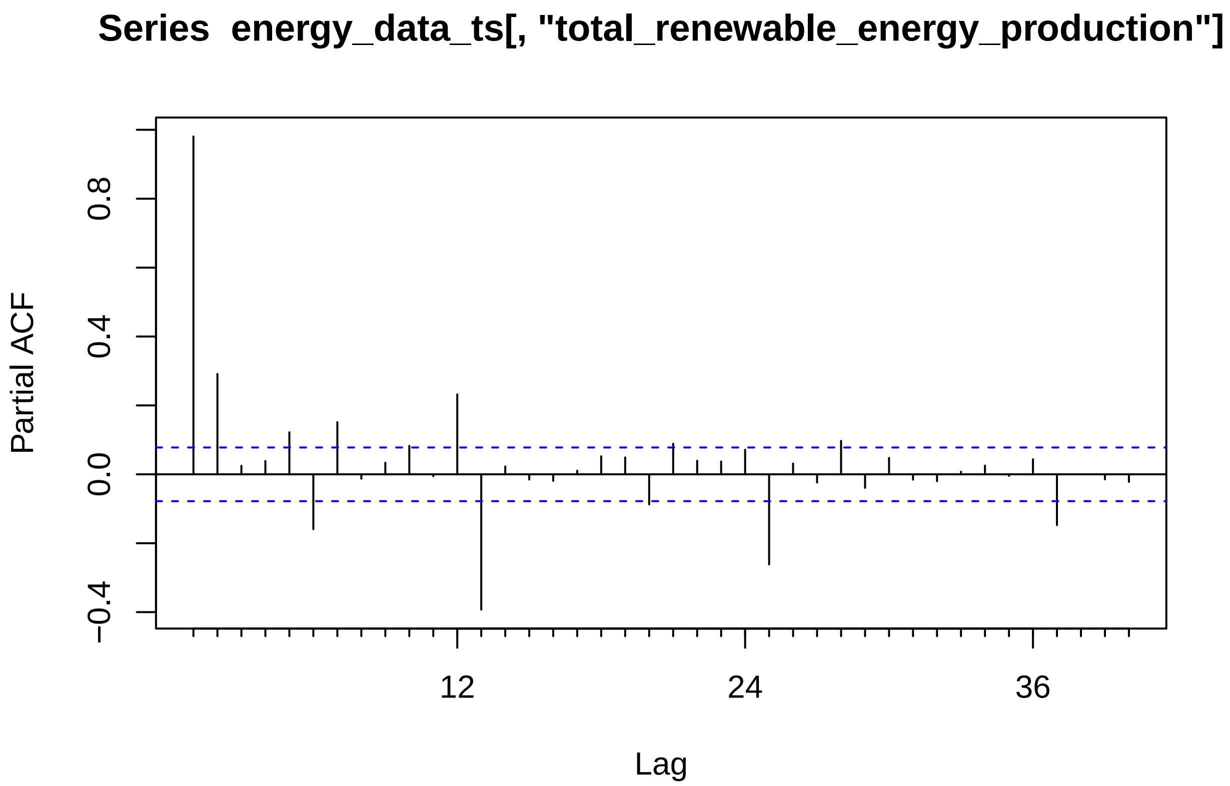
Question 7

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

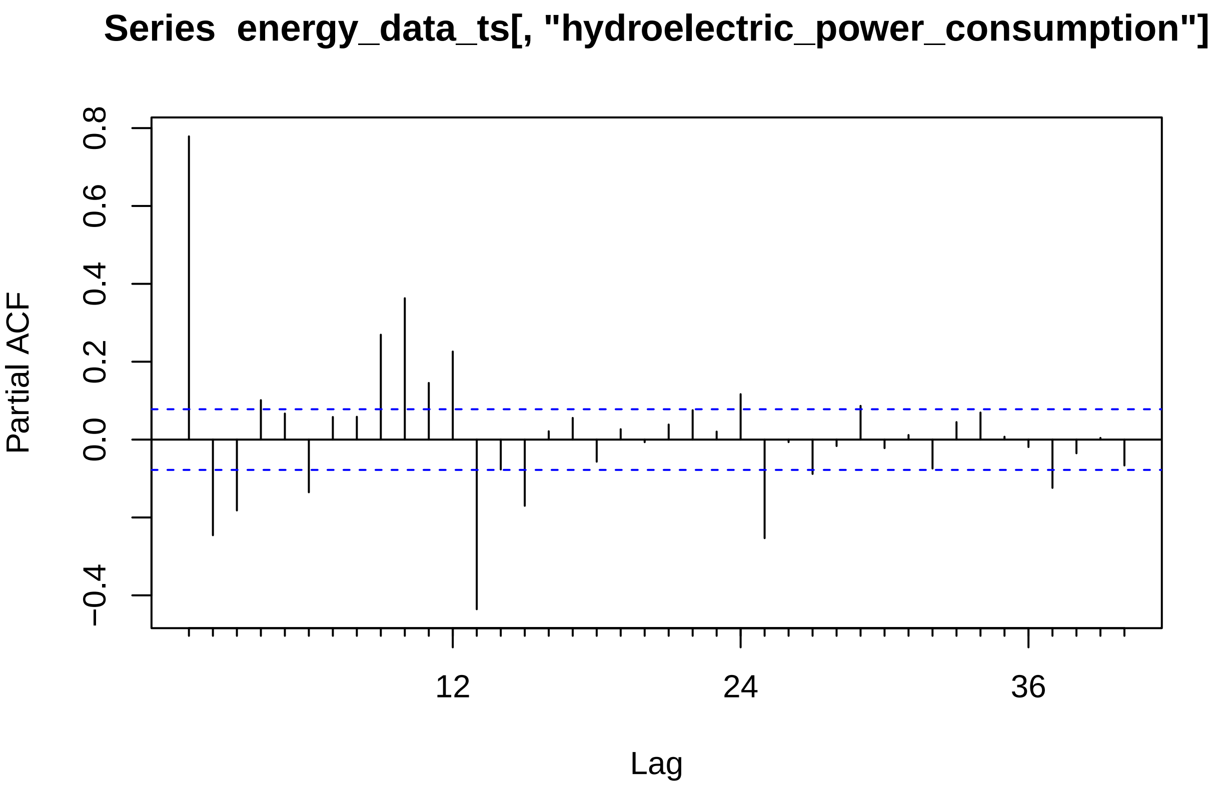
```
biomass_pacf <- Pacf(energy_data_ts[,"total_biomass_energy_production"], lag.max = 40)
```



```r
renewable_pacf <- Pacf(energy_data_ts[,"total_renewable_energy_production"], lag.max = 40)
```



```
hydro_pacf <- Pacf(energy_data_ts[,"hydroelectric_power_consumption"], lag.max = 40)
```



These PACF plots for biomass and renewable energy production no longer show high values at lags beyond t-1. This suggests that the autocorrelation of the first lag is very high, which may have influenced the ACF of later lags in these series. The PACF for hydroelectric power consumption shows a weaker seasonality pattern than the ACF of this series, though the partial autocorrelation of this series is stronger at later lags than for the other two series, which may suggest that this series has a long memory (and also confirms our class discussion that stream flow / dam reservoir levels have a long memory).