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

Assignment 2 - Due date 01/27/26

Mingjie Wei

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)
```

```
## 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(readxl)
library(openxlsx)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.x 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.

```
#Import the dataset
energy_data <- read.xlsx(xlsxFile="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou

#Extract the column names
read_col_names <- read.xlsx(xlsxFile="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_

#Assign the column names to the dataset
colnames(energy_data) <- read_col_names

#Visualize the first rows of the dataset
energy_data$Month <- convertToDate(energy_data$Month)

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.

```
energy_3df <- energy_data |>
  select('Total Biomass Energy Production',
         'Total Renewable Energy Production',
         'Hydroelectric Power Consumption')

head(energy_3df)
```

```
##   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().

```
# Extract start year and month
start_date <- min(energy_data$Month)
start_year <- as.numeric(format(start_date, "%Y"))
start_month <- as.numeric(format(start_date, "%m"))

# Create time series
energy_ts <- ts(
  data = energy_3df,
  start = c(start_year, start_month),
  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.

```
means <- colMeans(energy_ts)
sds <- apply(energy_ts, 2, sd)

summary_stats <- data.frame(
  Variable = colnames(energy_ts),
  Mean = means,
  SD = sds
)

print(summary_stats)
```

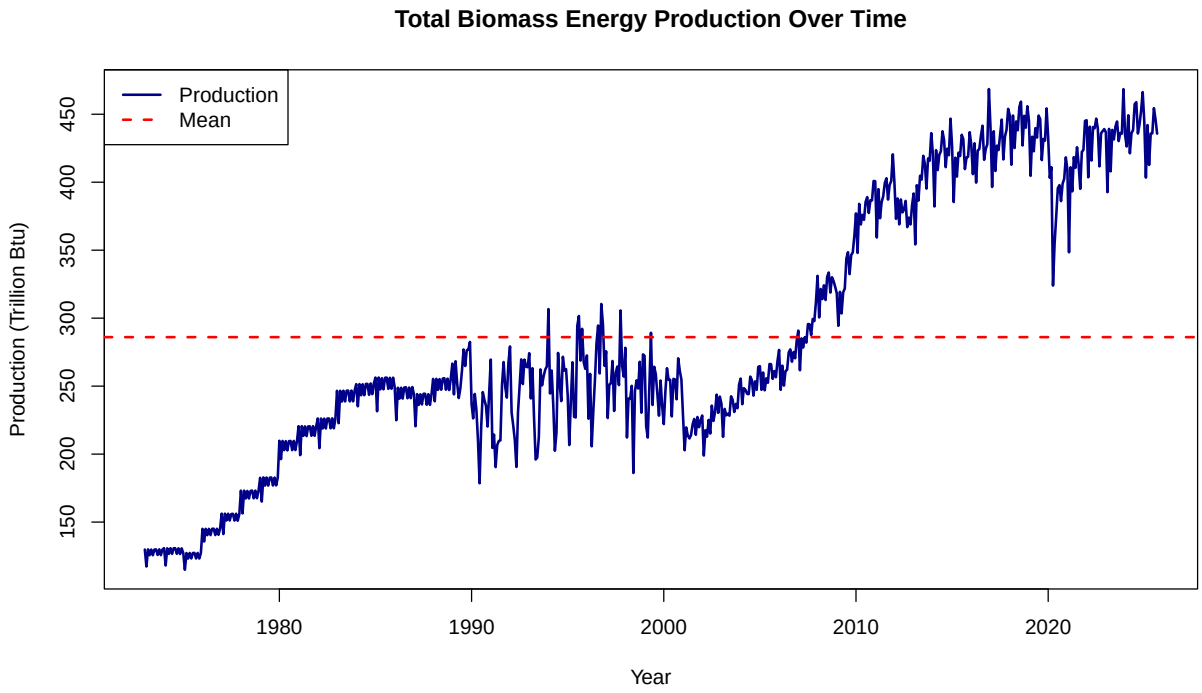
```
##                                                                 Variable      Mean
## Total Biomass Energy Production     Total Biomass Energy Production 286.04893
## Total Renewable Energy Production Total Renewable Energy Production 409.19521
## Hydroelectric Power Consumption     Hydroelectric Power Consumption  79.35682
##                                        SD
## Total Biomass Energy Production    96.21209
## Total Renewable Energy Production 151.42232
## Hydroelectric Power Consumption    14.12020
```

Question 4

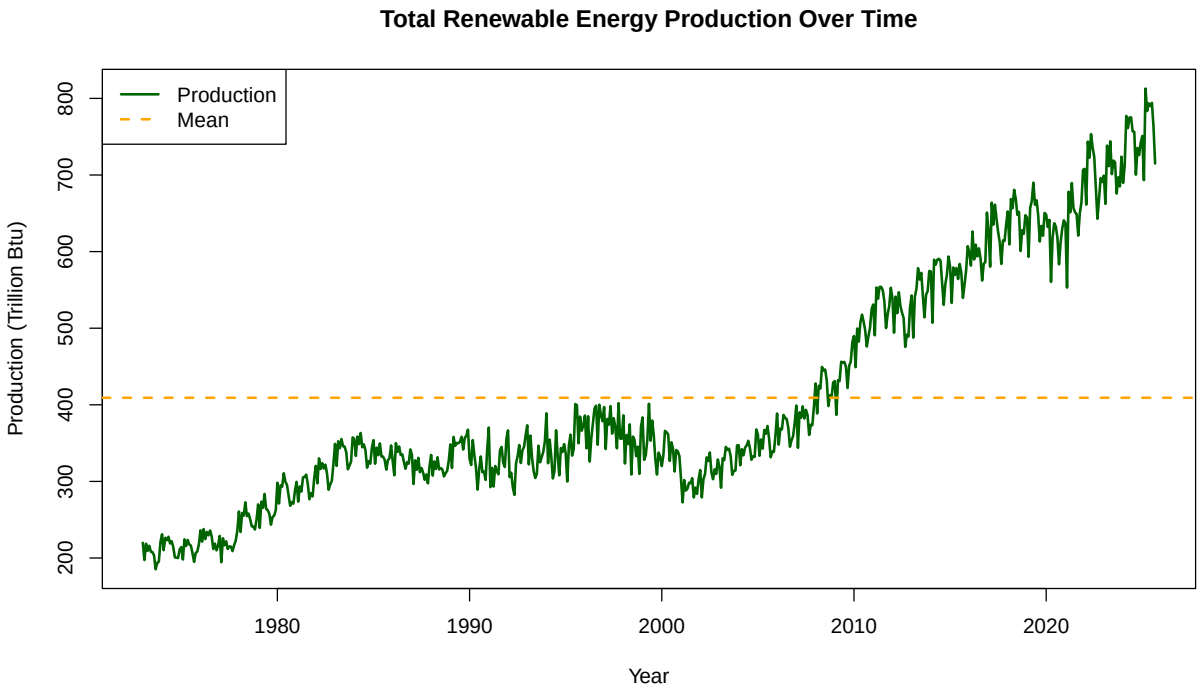
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
# Plot 1: Total Biomass Energy Production
plot(energy_ts[, "Total Biomass Energy Production"],
     main = "Total Biomass Energy Production Over Time",
     ylab = "Production (Trillion Btu)",
     xlab = "Year",
     col = "darkblue",
     lwd = 2)
abline(h = mean(energy_ts[, "Total Biomass Energy Production"]),
       col = "red", lwd = 2, lty = 2)
legend("topleft",
```

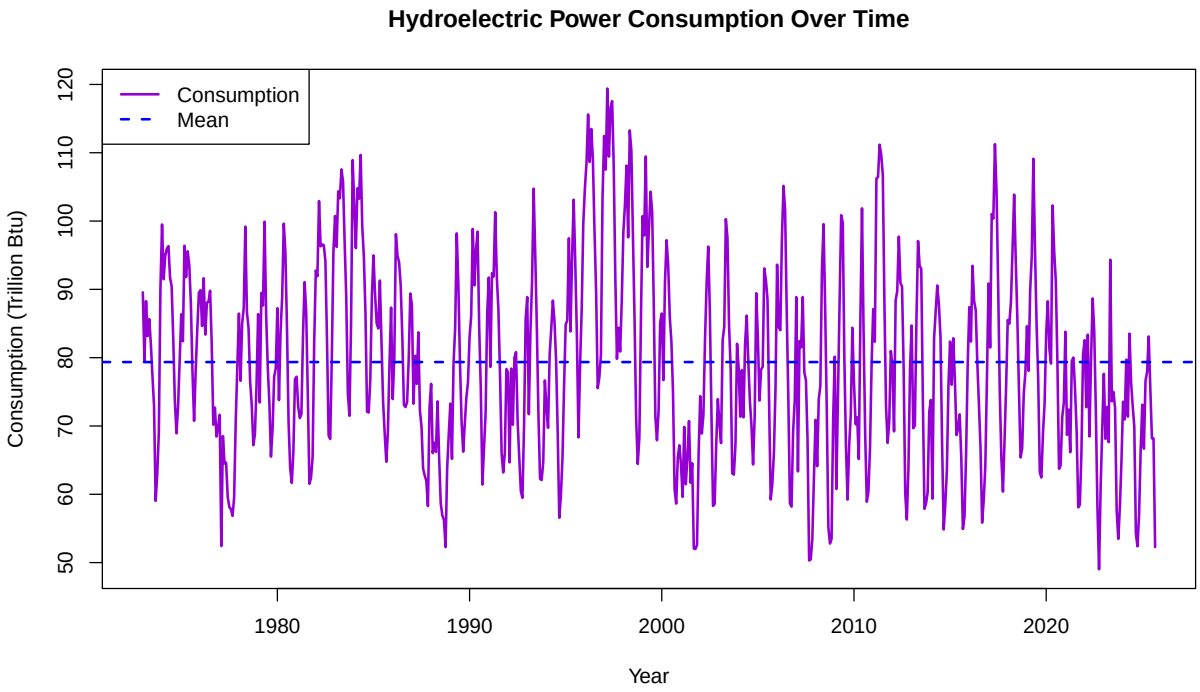
```
       legend = c("Production", "Mean"),
       col = c("darkblue", "red"),
       lty = c(1, 2),
       lwd = 2)
```




```
# Plot 2: Total Renewable Energy Production
plot(energy_ts[, "Total Renewable Energy Production"],
     main = "Total Renewable Energy Production Over Time",
     ylab = "Production (Trillion Btu)",
     xlab = "Year",
     col = "darkgreen",
     lwd = 2)
abline(h = mean(energy_ts[, "Total Renewable Energy Production"]),
       col = "orange", lwd = 2, lty = 2)
legend("topleft",
       legend = c("Production", "Mean"),
       col = c("darkgreen", "orange"),
       lty = c(1, 2),
       lwd = 2)
```




```r
# Plot 3: Hydroelectric Power Consumption
plot(energy_ts[, "Hydroelectric Power Consumption"],
     main = "Hydroelectric Power Consumption Over Time",
     ylab = "Consumption (Trillion Btu)",
     xlab = "Year",
     col = "darkviolet",
     lwd = 2)
abline(h = mean(energy_ts[, "Hydroelectric Power Consumption"]),
       col = "blue", lwd = 2, lty = 2)
legend("topleft",
       legend = c("Consumption", "Mean"),
       col = c("darkviolet", "blue"),
       lty = c(1, 2),
       lwd = 2)
```



Total Biomass Energy Production

- Trend: Clear upward trend over time, with faster growth after around 2008.
- Seasonality: Some short-term fluctuations, but the long-run increase is the main pattern.
- Mean line: Mostly below the mean in earlier years; mostly above the mean in recent years.
- Notable changes: A sharp drop around 2020, followed by a quick rebound.

Total Renewable Energy Production

- Trend: Strong upward trend, especially after about 2008–2010 when growth speeds up.
- Seasonality: Fluctuations are visible, and the size of swings becomes larger as the series rises.
- Mean line: Mostly below the mean in early years; mostly above the mean after 2010.
- Notable changes: A clear "jump" to a higher level around 2008–2010, then continued increase.

Hydroelectric Power Consumption

- Trend: No strong long-term upward trend; it stays within a broad range over time.
- Seasonality: Very high variability with frequent peaks and dips; changes are not smooth.
- Mean line: Crosses the mean many times, suggesting it often moves around its average level.
- Notable changes: Some periods are noticeably lower than average, but there is no sustained growth like the other two series.

Question 5

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

```r
# Correlation matrix
cor_mat <- cor(energy_ts, use = "pairwise.complete.obs")
cor_mat
```

```
##                                   Total Biomass Energy Production
## Total Biomass Energy Production                         1.0000000
## Total Renewable Energy Production                       0.9652985
## Hydroelectric Power Consumption                        -0.1347374
##                                   Total Renewable Energy Production
## Total Biomass Energy Production                          0.96529851
## Total Renewable Energy Production                        1.00000000
## Hydroelectric Power Consumption                         -0.05842436
##                                   Hydroelectric Power Consumption
## Total Biomass Energy Production                       -0.13473742
## Total Renewable Energy Production                     -0.05842436
## Hydroelectric Power Consumption                        1.00000000
```

```r
# Pairwise correlation tests (with p-values)
x1 <- energy_ts[, "Total Biomass Energy Production"]
x2 <- energy_ts[, "Total Renewable Energy Production"]
x3 <- energy_ts[, "Hydroelectric Power Consumption"]

test_12 <- cor.test(x1, x2)
test_13 <- cor.test(x1, x3)
test_23 <- cor.test(x2, x3)

list(
  biomass_vs_totalrenew = c(r = unname(test_12$estimate), p = test_12$p.value),
  biomass_vs_hydro      = c(r = unname(test_13$estimate), p = test_13$p.value),
  totalrenew_vs_hydro   = c(r = unname(test_23$estimate), p = test_23$p.value)
)
```

```
## $biomass_vs_totalrenew
##         r         p
## 0.9652985 0.0000000
##
## $biomass_vs_hydro
##             r             p
## -0.1347374219  0.0006769838
##
## $totalrenew_vs_hydro
##           r           p
## -0.05842436  0.14202918
```

- **Total Biomass Energy Production vs. Total Renewable Energy Production:**

  The correlation is very strong and positive (r = 0.9653) with p is approximately equal to 0. This means they are significantly positively correlated. This is expected because both series increase over time.

- **Total Biomass Energy Production vs. Hydroelectric Power Consumption:**

  The correlation is weak and negative (r = -0.1347) but the p-value is 0.0006769, which is below 0.05. So it is statistically significant, but the relationship is very small in magnitude (practically weak).

- **Total Renewable Energy Production vs. Hydroelectric Power Consumption:**

The correlation is very weak (r = -0.0584) and the p-value is 0.1420, which is greater than 0.05. Therefore, this pair is not significantly correlated.

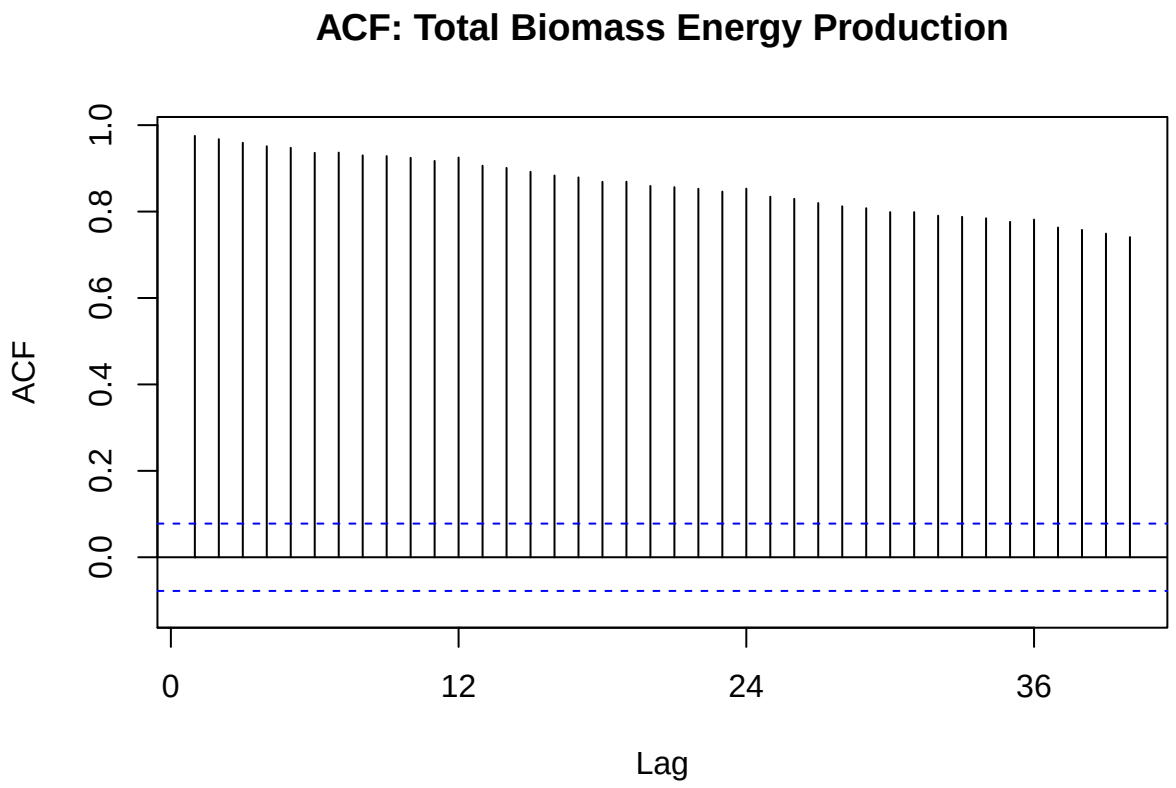
- **Conclusion:**

  Only biomass and total renewable production are strongly and significantly correlated. Hydroelectric consumption does not move closely with the other two series. Also, since biomass and total renewables both have strong upward trends, their high correlation may be partly driven by the shared trend rather than a direct short-term relationship.

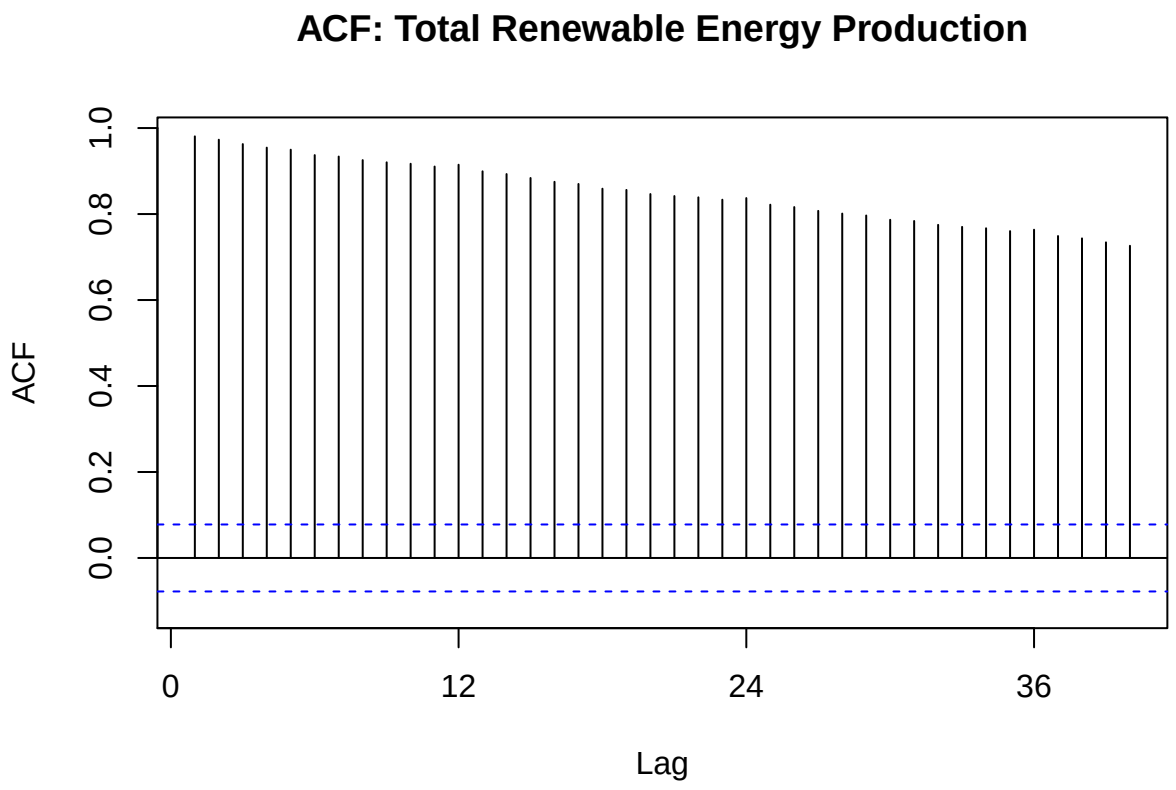
Question 6

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?

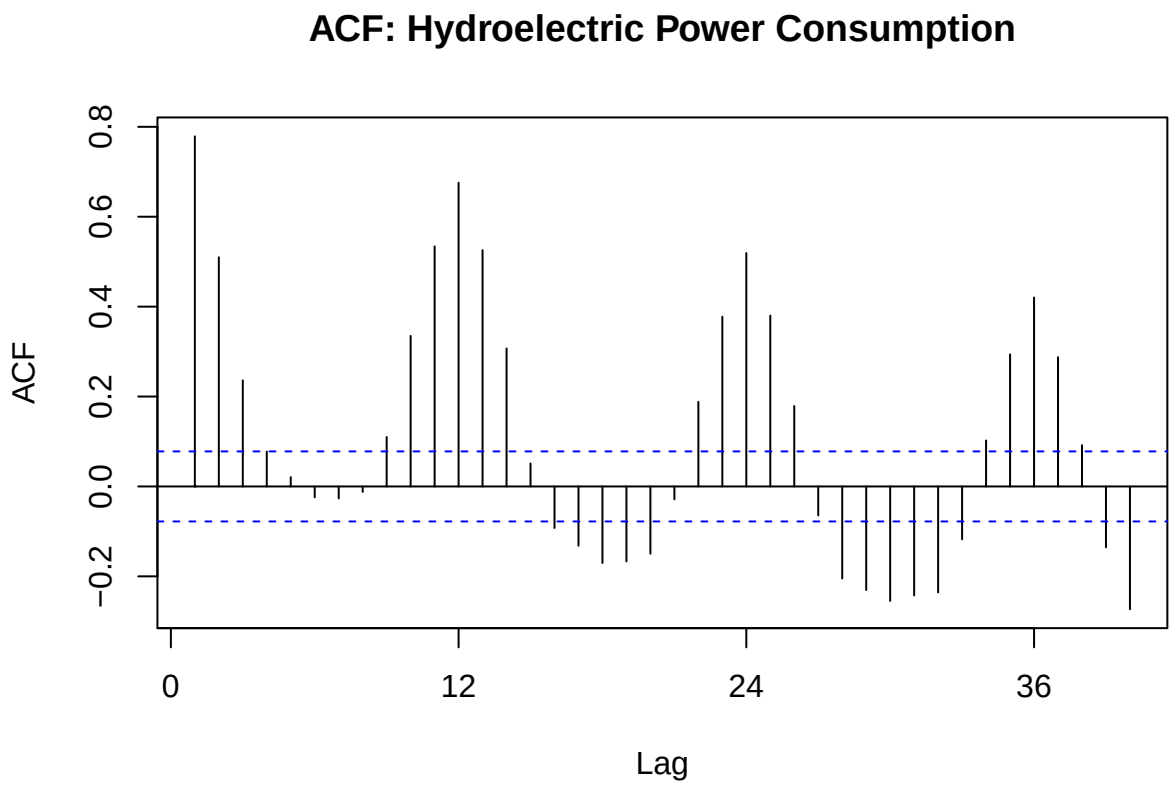
```
Acf(energy_ts[, "Total Biomass Energy Production"], lag.max = 40,
    main = "ACF: Total Biomass Energy Production")
```




```
Acf(energy_ts[, "Total Renewable Energy Production"], lag.max = 40,
    main = "ACF: Total Renewable Energy Production")
```




```r
Acf(energy_ts[, "Hydroelectric Power Consumption"], lag.max = 40,
    main = "ACF: Hydroelectric Power Consumption")
```



Total Biomass Energy Production

- The ACF is very high at small lags (close to 1) and decreases very slowly as lag increases.
- Many autocorrelations stay well above the blue significance bands, even up to lag 40.
- This pattern suggests strong persistence over time and likely non-stationarity (often caused by a strong trend).

Total Renewable Energy Production

- The ACF looks very similar to biomass: it starts near 1 and decays slowly.
- Most lags are significant (above the confidence bands).
- This also indicates strong persistence and likely non-stationarity due to trend.

Hydroelectric Power Consumption

- The ACF behavior is different. It is positive at the first few lags, but then it oscillates (some positive, some negative).
- There are clear seasonal peaks around lag 12, 24, and 36 (about 1, 2, 3 years for monthly data), which suggests annual seasonality.
- Compared to the other two series, hydro has weaker long-run persistence and a stronger seasonal pattern.

The three variables don't have the same behavior

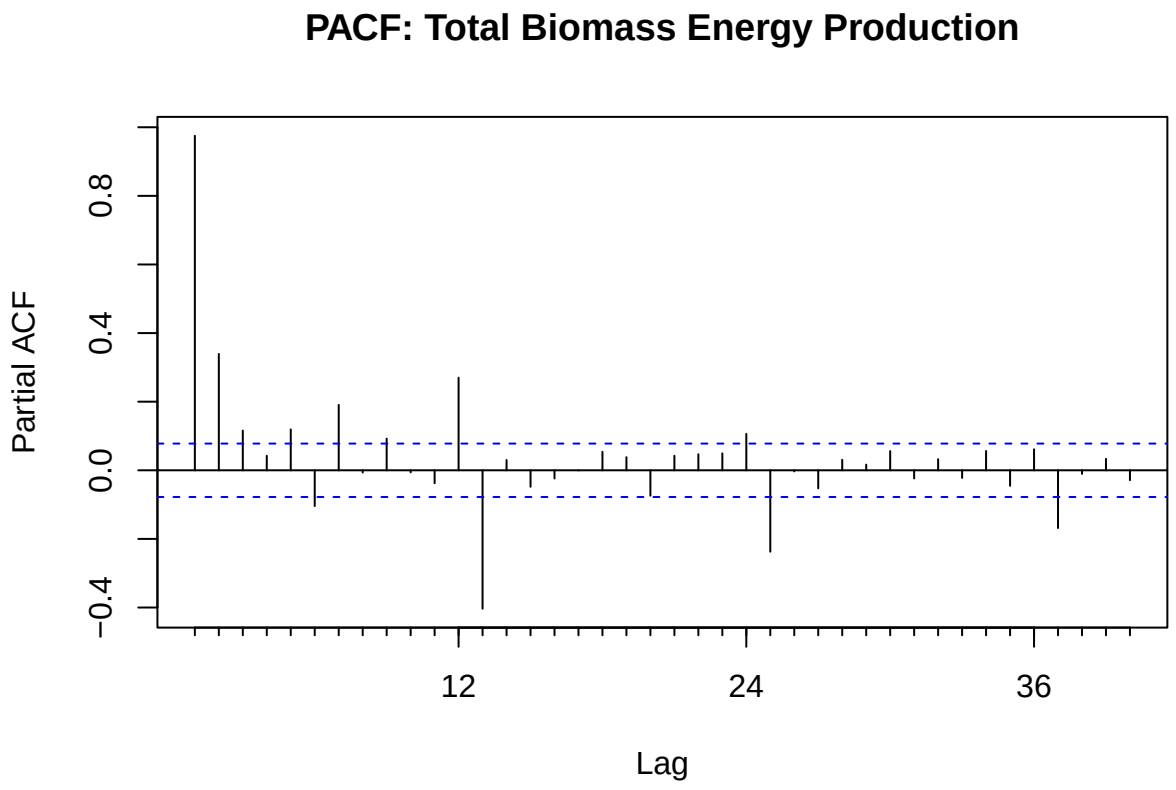
- Biomass and Total Renewables show very similar ACF patterns (slow decay, strong persistence, trend-like / non-stationary behavior).

- Hydroelectric consumption shows seasonal and oscillating behavior, which is different from the trend-driven patterns in the other two.

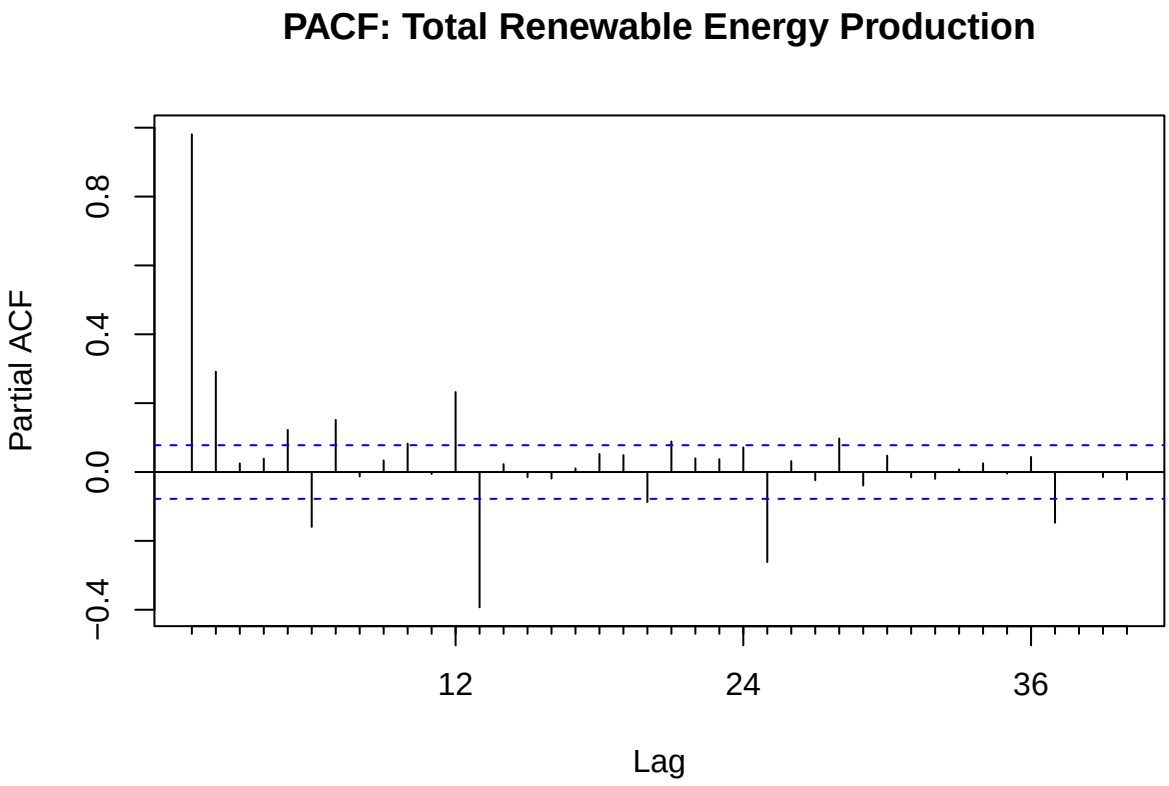
Question 7

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

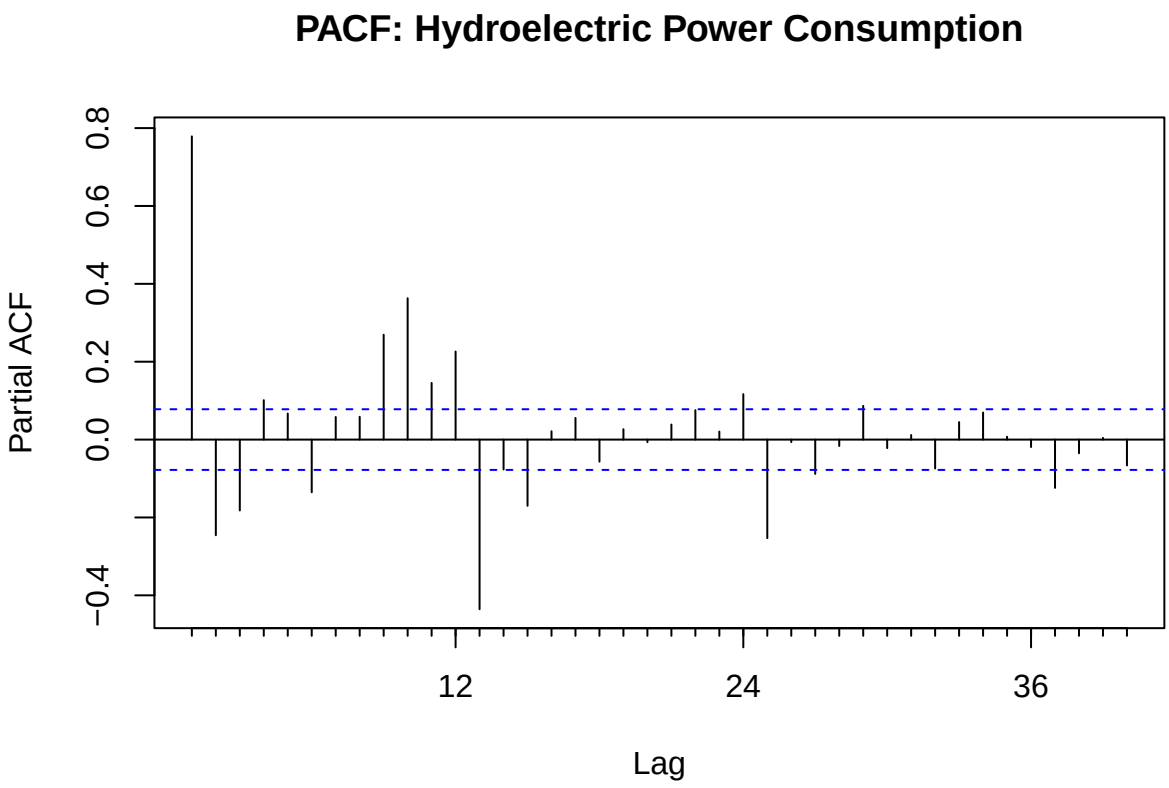
```
Pacf(energy_ts[, "Total Biomass Energy Production"], lag.max = 40,
     main = "PACF: Total Biomass Energy Production")
```




```
Pacf(energy_ts[, "Total Renewable Energy Production"], lag.max = 40,
     main = "PACF: Total Renewable Energy Production")
```




```
Pacf(energy_ts[, "Hydroelectric Power Consumption"], lag.max = 40,
     main = "PACF: Hydroelectric Power Consumption")
```



Total Biomass Energy Production

- In Q6 (ACF), the autocorrelations stayed very high for many lags and decayed slowly, which shows strong persistence (trend-like behavior).
- In Q7 (PACF), we see a very large spike at lag 1, but after controlling for lag 1, most other lags become small (many are near zero or within the bands).
- This means the high ACF at many lags in Q6 is largely due to indirect effects through lag 1 (and the trend), while the PACF shows that the main direct dependence is on the previous month.

Total Renewable Energy Production

- In Q6 (ACF), the series also showed very high correlations across many lags with slow decay, again suggesting strong persistence and a trend.
- In Q7 (PACF), the pattern is much simpler: there is a strong spike at lag 1, and most other lags are much smaller after we control for the shorter lags.
- So, compared with Q6, the PACF suggests that the long "tail" in the ACF is mainly from the strong lag-1 effect plus trend, rather than many separate direct lag effects.

Hydroelectric Power Consumption

- In Q6 (ACF), the autocorrelation oscillated and showed clear seasonal peaks around lag 12, 24, 36, which suggests annual seasonality.
- In Q7 (PACF), the plot highlights which lags have a direct effect: we still see meaningful spikes (especially around lag 12), but fewer lags are important once shorter lags are controlled for.
- This means Q6 shows overall seasonal dependence (direct + indirect), while Q7 helps identify specific seasonal lags that matter directly.