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

Assignment 2 - Due date 01/23/25

Nicole Leines

## 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.\

```
#Load/install required package here

# List of packages to install
#packages <- c("forecast", "tseries", "dplyr")
#install.packages(packages)

library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.3.3
```

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

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.3.3
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tibble' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```r
library(openxlsx)
```

```
## Warning: package 'openxlsx' was built under R version 4.3.3
```

```r
library(ggplot2)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3

##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

## 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
base_dir <- "D:/Geani/Box/Home Folder gnl13/Private/1 Academics/3 Time series/TSA_Sp25"
data_dir <- file.path(base_dir, "Data")
file_name <- "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"
file_path <- file.path(data_dir, file_name)

#Importing data set without change the original file using read.xlsx
energy_data1 <- read_excel(path=file_path,
                           skip = 12,
                           sheet="Monthly Data",
                           col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```r
#Now let's extract the column names from row 11
read_col_names <- read_excel(path=file_path,
                             skip = 10,n_max = 1,
                             sheet="Monthly Data",
                             col_names=FALSE)
```

```
## New names:
```

```
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

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

## 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.

```
column_names <- colnames(energy_data1)
print(column_names)
```

```
##  [1] "Month"                                "Wood Energy Production"
##  [3] "Biofuels Production"                  "Total Biomass Energy Production"
##  [5] "Total Renewable Energy Production"    "Hydroelectric Power Consumption"
##  [7] "Geothermal Energy Consumption"        "Solar Energy Consumption"
##  [9] "Wind Energy Consumption"              "Wood Energy Consumption"
```

```
## [11] "Waste Energy Consumption"           "Biofuels Consumption"
## [13] "Total Biomass Energy Consumption"   "Total Renewable Energy Consumption"
```

```
energy_data2 <- energy_data1[, c("Month",
                                 "Total Biomass Energy Production",
                                  "Total Renewable Energy Production",
                                  "Hydroelectric Power Consumption")]

head(energy_data2)
```

```
## # 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().

```
start_date <- min(energy_data2$Month)
print(start_date)
```

```
## [1] "1973-01-01 UTC"
```

```
ts_energy_data2 <- ts(
  data = energy_data2[,2:4],
  start = c(1973,1),
  frequency = 12
)

head(ts_energy_data2)
```

```
##          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.

```
# Compute mean and standard deviation for each column
means <- apply(ts_energy_data2, 2, mean, na.rm = TRUE)
sds <- apply(ts_energy_data2, 2, sd, na.rm = TRUE)

means <- unname(means)
sds <- unname(sds)

# Combine results into a data frame
results <- data.frame(series=colnames(ts_energy_data2), Mean = means, SD = sds)
print(results)
```

```
##                               series      Mean        SD
## 1   Total Biomass Energy Production 282.67785  94.05815
## 2 Total Renewable Energy Production 402.01667 143.79270
## 3   Hydroelectric Power Consumption  79.55371  14.10737
```

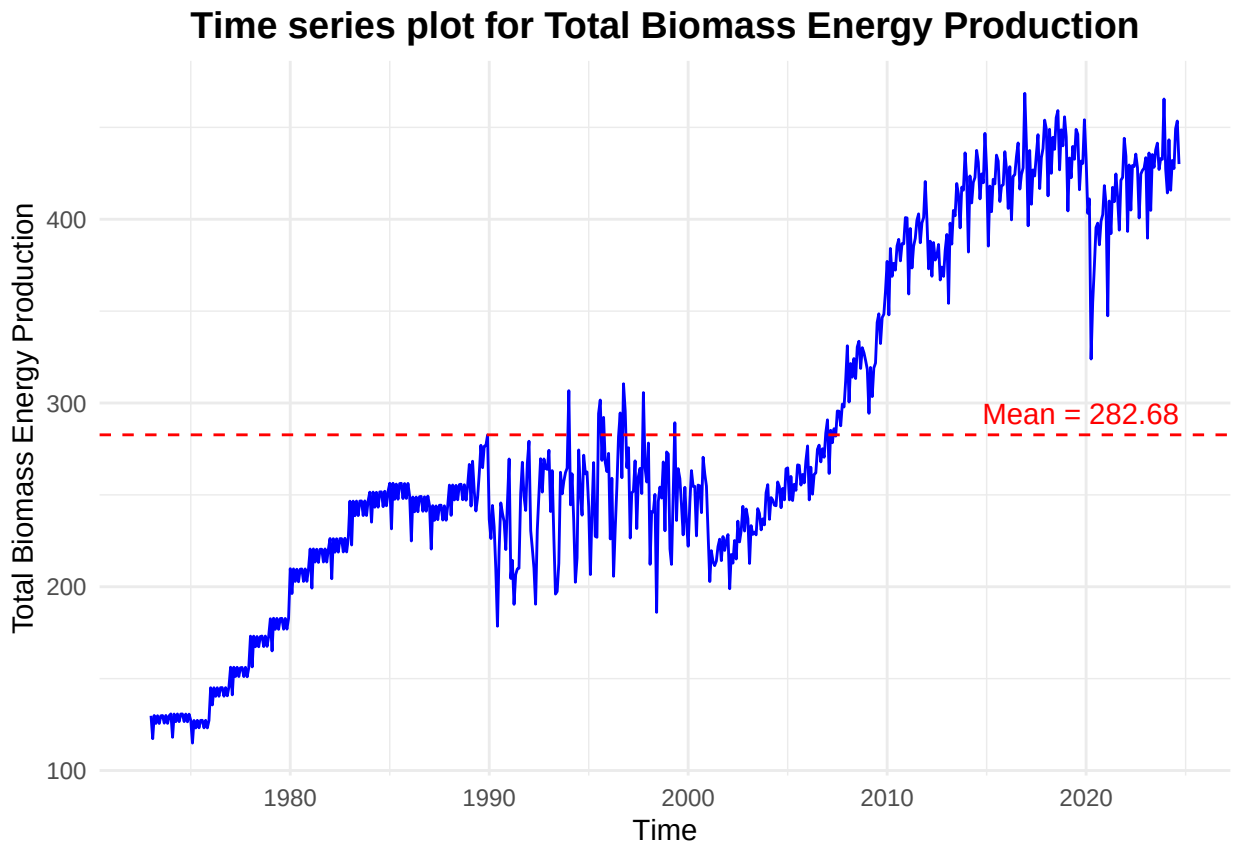
## Question 4

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.

```
p1<-ggplot(energy_data2, aes(x = Month)) +
  geom_line(aes(y = `Total Biomass Energy Production`), color = "blue", size = 0.5) +
  geom_hline(aes(yintercept = mean(`Total Biomass Energy Production`, na.rm = TRUE)),
             color = "red", linetype = "dashed") +
  annotate("text",
           x = max(energy_data2$Month), # Position label at the end of the plot
           y = mean(energy_data2$`Total Biomass Energy Production`, na.rm = TRUE),
           label = paste("Mean =", round(mean(energy_data2$`Total Biomass Energy Production`, na.rm = TR
           color = "red", hjust = 1, vjust = -0.5) +
  xlab("Time") +
  ylab("Total Biomass Energy Production") +
  ggtitle("Time series plot for Total Biomass Energy Production") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
```

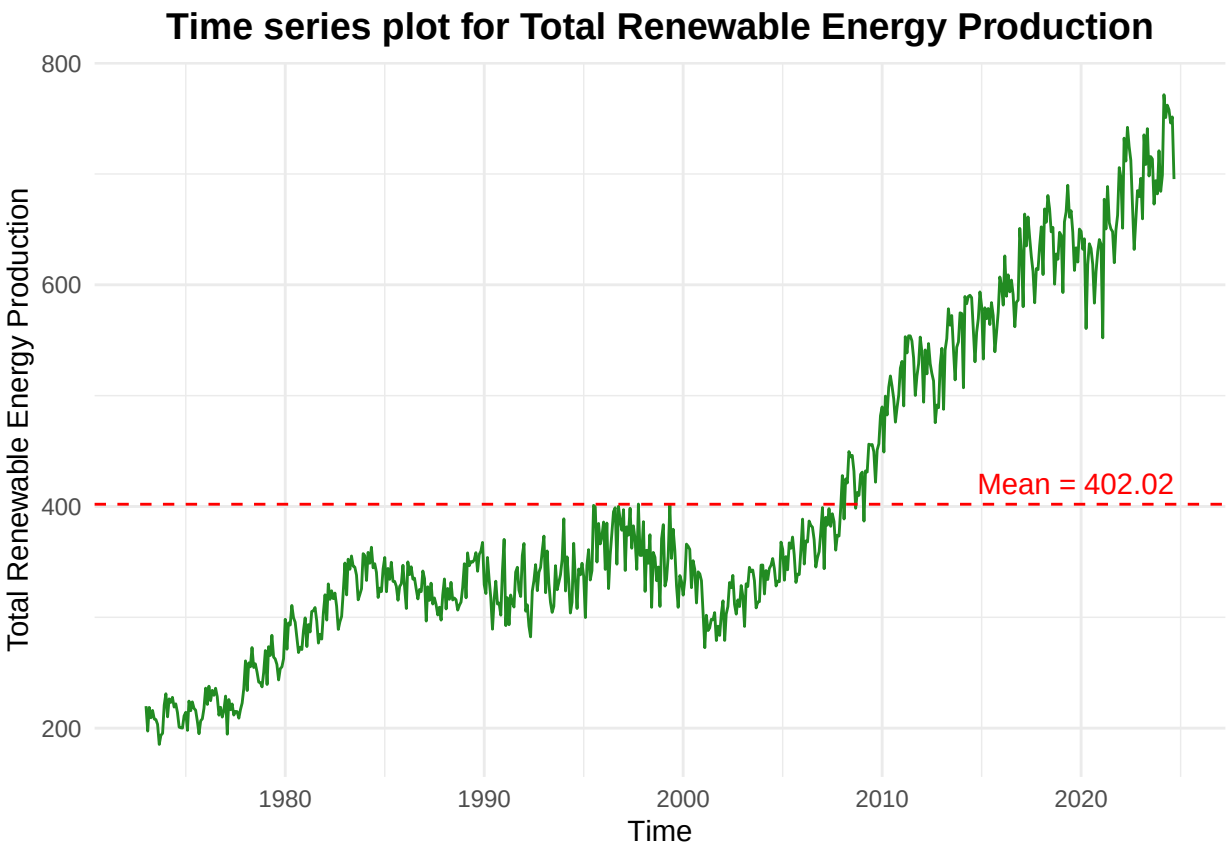
```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
print(p1)
```



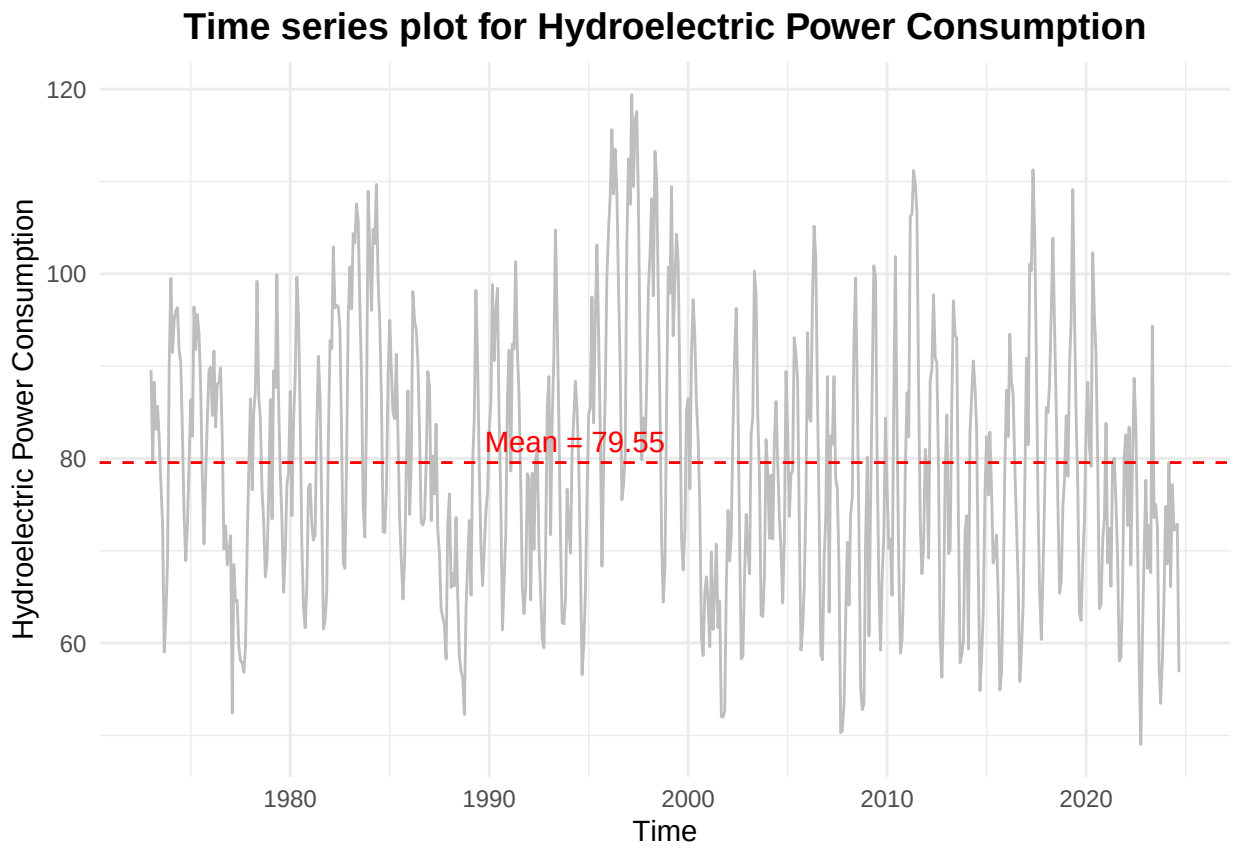

The time series exhibits a positive trend, with biomass energy production increasing significantly from around 1960 to 2024. There are periods of steady growth where production stabilizes before increasing again. The mean of 282.68 represents the average of the entire dataset. Earlier values (pre-1990) are predominantly below the mean, indicating lower production during this period. Post-2000 values are mostly above the mean, reflecting a shift in production to consistently higher levels. This skews the mean downward, making it unrepresentative of the higher values seen after 2010. The dataset is non-stationary, meaning the mean of the series changes over time because of the upward trend. Using the mean as a point of reference for the period after 2010 would misrepresent the central tendency for that period.

```
p2<-ggplot(energy_data2, aes(x = Month)) +
  geom_line(aes(y = `Total Renewable Energy Production`), color = "forestgreen", size = 0.5) +
  geom_hline(aes(yintercept = mean(`Total Renewable Energy Production`, na.rm = TRUE)),
             color = "red", linetype = "dashed") +
  annotate("text",
           x = max(energy_data2$Month), # Position label at the end of the plot
           y = mean(energy_data2$`Total Renewable Energy Production`, na.rm = TRUE),
           label = paste("Mean =", round(mean(energy_data2$`Total Renewable Energy Production`, na.rm =
           color = "red", hjust = 1, vjust = -0.5) +
  xlab("Time") +
  ylab("Total Renewable Energy Production") +
  ggtitle("Time series plot for Total Renewable Energy Production") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
print(p2)
```




Similar to biomass energy production but with fewer stabilizer periods, the renewable energy production time series shows moderate growth with some variability between 1980 and 2000. In general, the time series exhibits a strong upward trend, indicating that renewable energy production has increased substantially from the 1960s to 2024. The mean production value of 402.02 is close to the typical values in the late 1990s and early 2000s, but it is not representative of the higher production levels observed post-2010, where most values exceed the mean. This is also a non-stationary time series due to the evident upward trend. So, the mean production is not a reliable representation of the most recent trends.

```
p3<-ggplot(energy_data2, aes(x = Month)) +
  geom_line(aes(y = `Hydroelectric Power Consumption`), color = "gray", size = 0.5) +
  geom_hline(aes(yintercept = mean(`Hydroelectric Power Consumption`, na.rm = TRUE)),
             color = "red", linetype = "dashed") +
  annotate("text",
           x = median(energy_data2$Month), # Position label at the end of the plot
           y = mean(energy_data2$`Hydroelectric Power Consumption`, na.rm = TRUE),
           label = paste("Mean =", round(mean(energy_data2$`Hydroelectric Power Consumption`, na.rm = TR
           color = "red", hjust = 1, vjust = -0.5) +
  xlab("Time") +
  ylab("Hydroelectric Power Consumption") +
  ggtitle("Time series plot for Hydroelectric Power Consumption") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
print(p3)
```

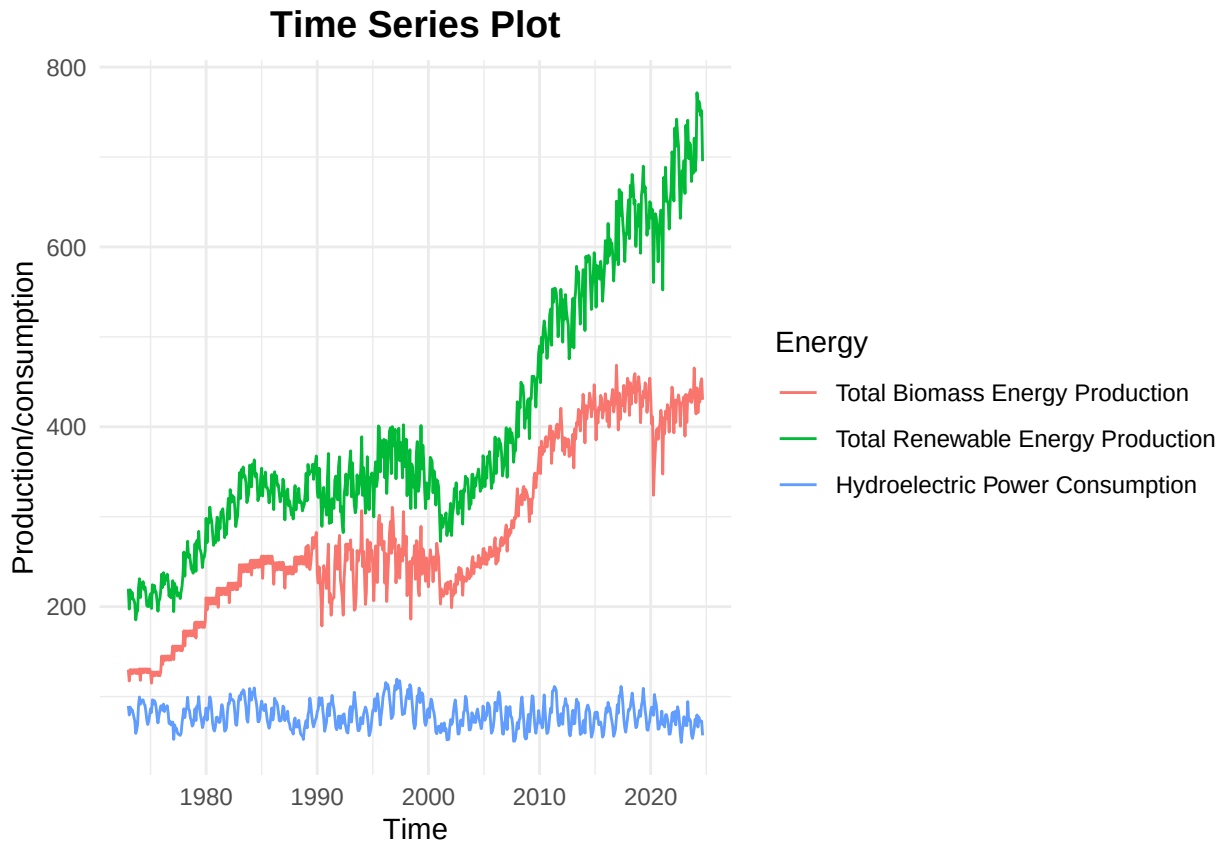



Unlike the other time series with a clear upward trend, this series does not exhibit a strong, consistent trend. Instead, this series shows strong periodic fluctuations over time, which could represent seasonal variability such as rainfall patterns. In this context, the mean consumption value of 79.55 is a reasonable representation of the earlier periods (pre-2000), where values oscillate around this level. However, in more recent years, the series appears to hover below the mean, indicating a potential decline in average consumption. So, the mean becomes less representative of recent years due to the apparent downward shift in consumption levels.

```
p4 <- autoplot(ts_energy_data2, ts.colour = "blue", size = 0.5) +
  labs(
    x = "Time",
    y = "Production/consumption",
    title = "Time Series Plot",
    color="Energy"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
```

```
## Warning in ggplot2::geom_line(na.rm = TRUE, ...): Ignoring unknown parameters:
## 'ts.colour'
```

```
print(p4)
```




## Question 5

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

```
# Compute correlations with significance testing
cor_results <- corr.test(ts_energy_data2)

# Print the correlation matrix with p-values
print(cor_results, short = FALSE)
```

```
## Call:corr.test(x = ts_energy_data2)
## Correlation matrix
##                                    Total Biomass Energy Production
## Total Biomass Energy Production                                1.00
## Total Renewable Energy Production                              0.97
## Hydroelectric Power Consumption                               -0.11
##                                    Total Renewable Energy Production
## Total Biomass Energy Production                                 0.97
## Total Renewable Energy Production                               1.00
## Hydroelectric Power Consumption                                -0.03
##                                    Hydroelectric Power Consumption
## Total Biomass Energy Production                              -0.11
## Total Renewable Energy Production                            -0.03
## Hydroelectric Power Consumption                               1.00
## Sample Size
```

```
## [1] 621
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                                0
## Total Renewable Energy Production                              0
## Hydroelectric Power Consumption                                0
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                               0.00
## Total Renewable Energy Production                             0.00
## Hydroelectric Power Consumption                               0.47
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                             0.01
## Total Renewable Energy Production                           0.47
## Hydroelectric Power Consumption                             0.00
##
##  Confidence intervals based upon normal theory.  To get bootstrapped values, try cor.ci
##             raw.lower raw.r raw.upper raw.p lower.adj upper.adj
## TtBEP-TtREP      0.96  0.97      0.97  0.00      0.96      0.97
## TtBEP-HydPC     -0.19 -0.11     -0.04  0.00     -0.20     -0.02
## TtREP-HydPC     -0.11 -0.03      0.05  0.47     -0.11      0.05
```

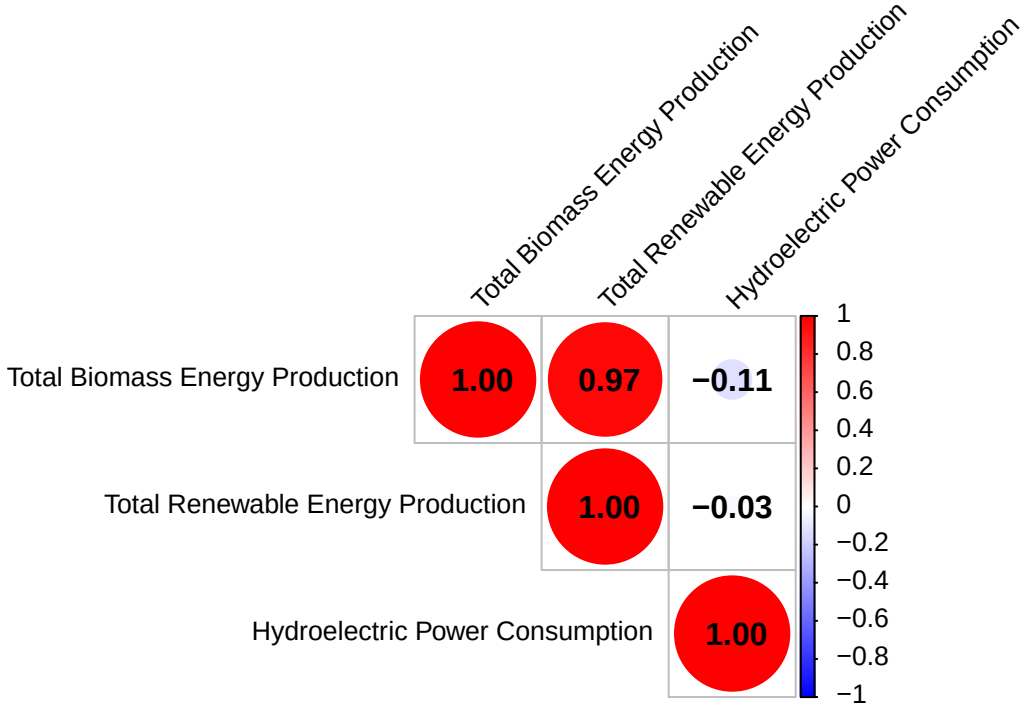
```
# Compute the correlation matrix
cor_matrix <- cor(ts_energy_data2, use = "complete.obs")

# Visualize the correlation matrix
corrplot(cor_matrix, method = "circle", type = "upper", addCoef.col = "black",
        tl.col = "black",
        col = colorRampPalette(c("blue", "white", "red"))(200),# Black text labels
  tl.srt = 45,                          # Rotate text labels for readability
  tl.cex = 0.8,                         # Reduce size of text labels (variable names)
  cl.cex = 0.8,                         # Adjust size of the color legend
  cl.lim = c(-1, 1),                    # Set consistent scale limits (-1 to 1)
  cl.ratio = 0.2,                       # Reduce the height of the color legend
  cl.align.text = "l",                  # Align legend text to the left
  mar = c(1, 1, 2, 1),                  # Add padding for the title
  title = "Correlation Matrix"    # Add a descriptive title
)
```

```
## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt =
## tl.srt, : "cl.lim" is not a graphical parameter
```

```
## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col =
## tl.col, : "cl.lim" is not a graphical parameter
```

```
## Warning in title(title, ...): "cl.lim" is not a graphical parameter
```




Taking into consideration that a p-value $< 0.05$ indicates a statistically significant correlation:

- The correlation between biomass Energy Production and Total Renewable Energy Production is statistically significant, with a 0.97 and a p-value of 0.00. Thus, it shows a strong positive correlation.
- The correlation between Biomass Energy Production and Hydroelectric Power Consumption is statistically significant, with a p-value of 0.01, but its weak magnitude of -0.11 means it is not practically meaningful.
- The correlation between Renewable Energy Production and Hydroelectric Power Consumption is not statistically significant. There is no evidence of a meaningful relationship between these two series, with a -0.03 and a p-value of 0.47.

## Question 6

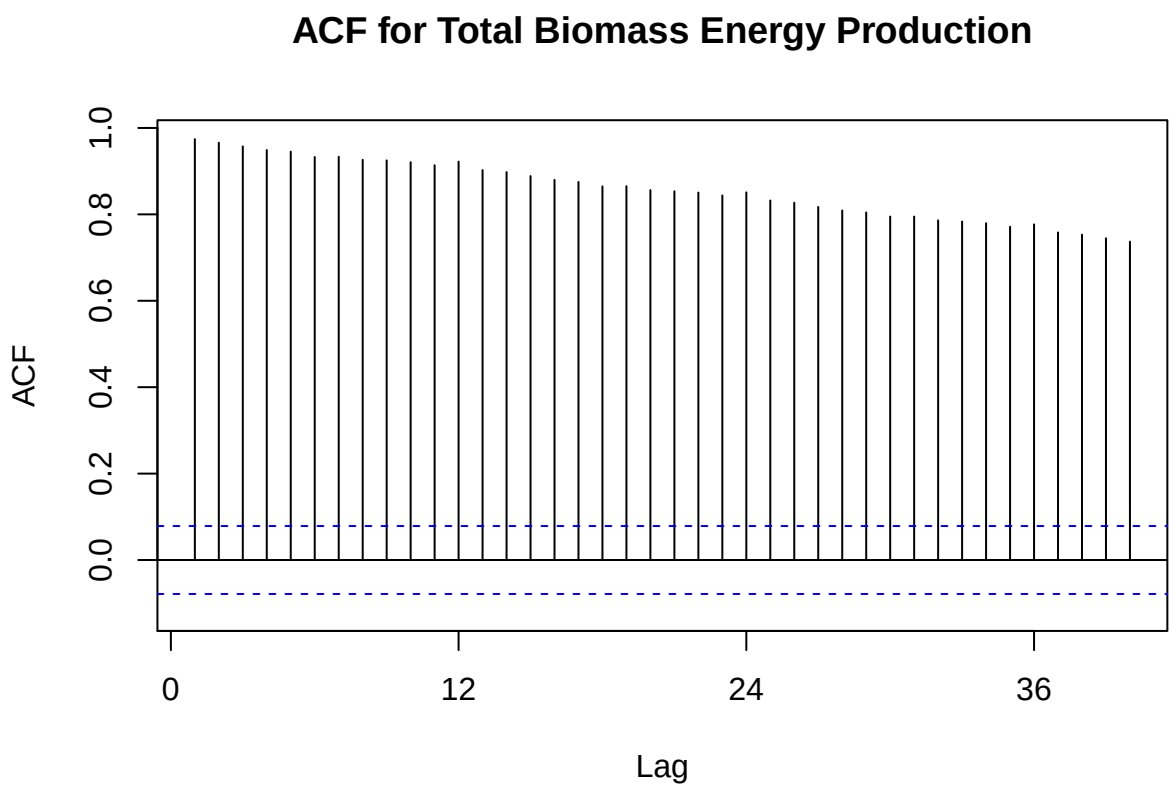
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

Plots 1 and 2 are similar. The ACF starts close to 1 and decays gradually. Considering that the blue dashed lines represent the confidence intervals, significant autocorrelation is present for many lags, with the bars extending beyond them. The only difference is that plot two decays slightly faster than the first series. This behavior suggests a high correlation between current and past values.

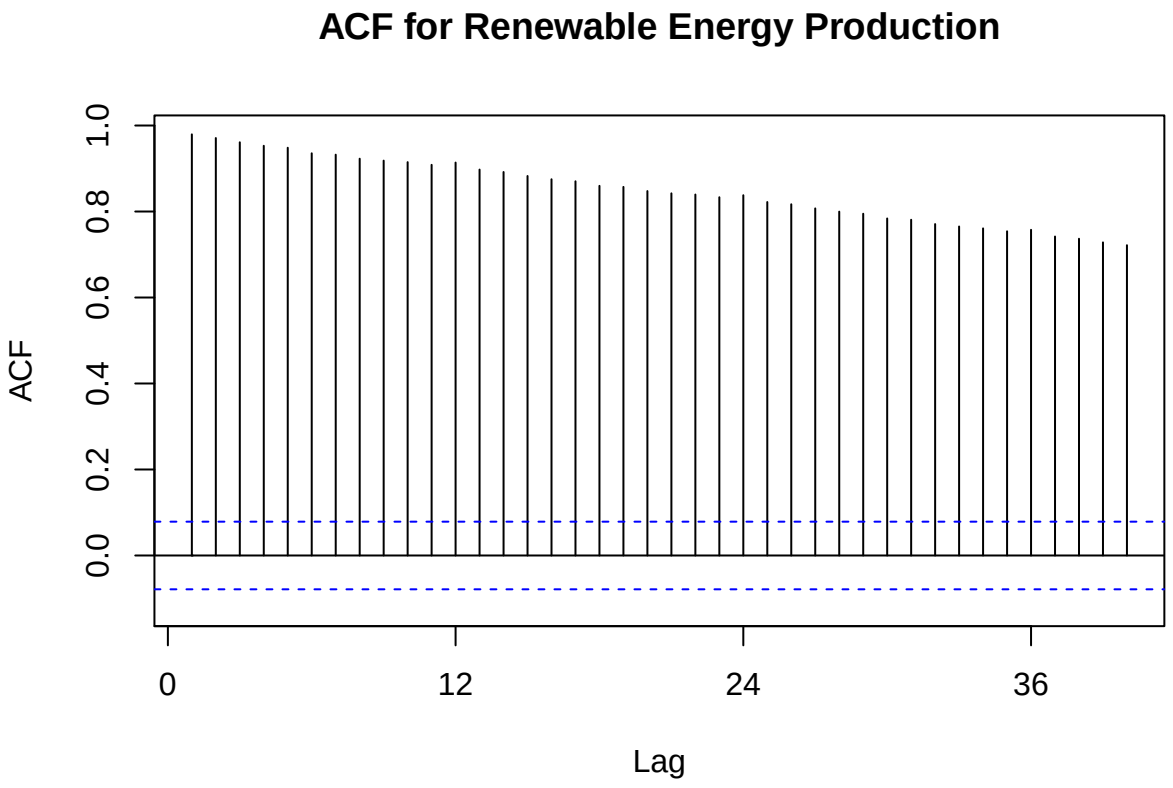
On the other hand, the ACF pattern in plot 3 is very different from the first two plots because there is no gradual decay, and only some of the lags are within the confidence interval, indicating that the fluctuations are not consistently strong. This suggests that the series might not have a strong trend. However, some lags around multiples of 12 exhibit slightly stronger correlations that could indicate a weak seasonal component.

So, in general, the alternating positive and negative values suggest the series might have an irregular or non-periodic component.

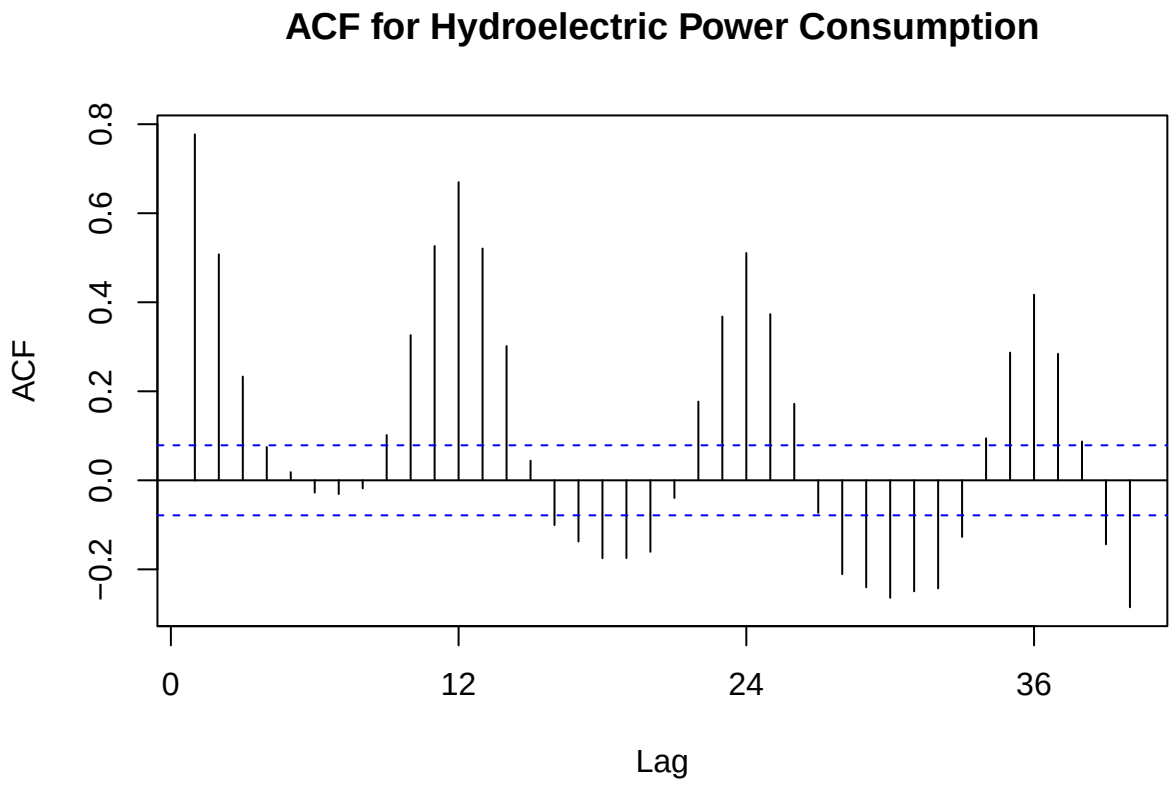
```
BE_acf=Acf(ts_energy_data2[,1],lag.max=40, type="correlation", plot=TRUE, main = "ACF for Total Biomass
```




```
RE_acf=Acf(ts_energy_data2[,2],lag.max=40, type="correlation", plot=TRUE, main = "ACF for Renewable Ene
```




```
HP_acf=Acf(ts_energy_data2[,3],lag.max=40, type="correlation", plot=TRUE, main = "ACF for Hydroelectric
```



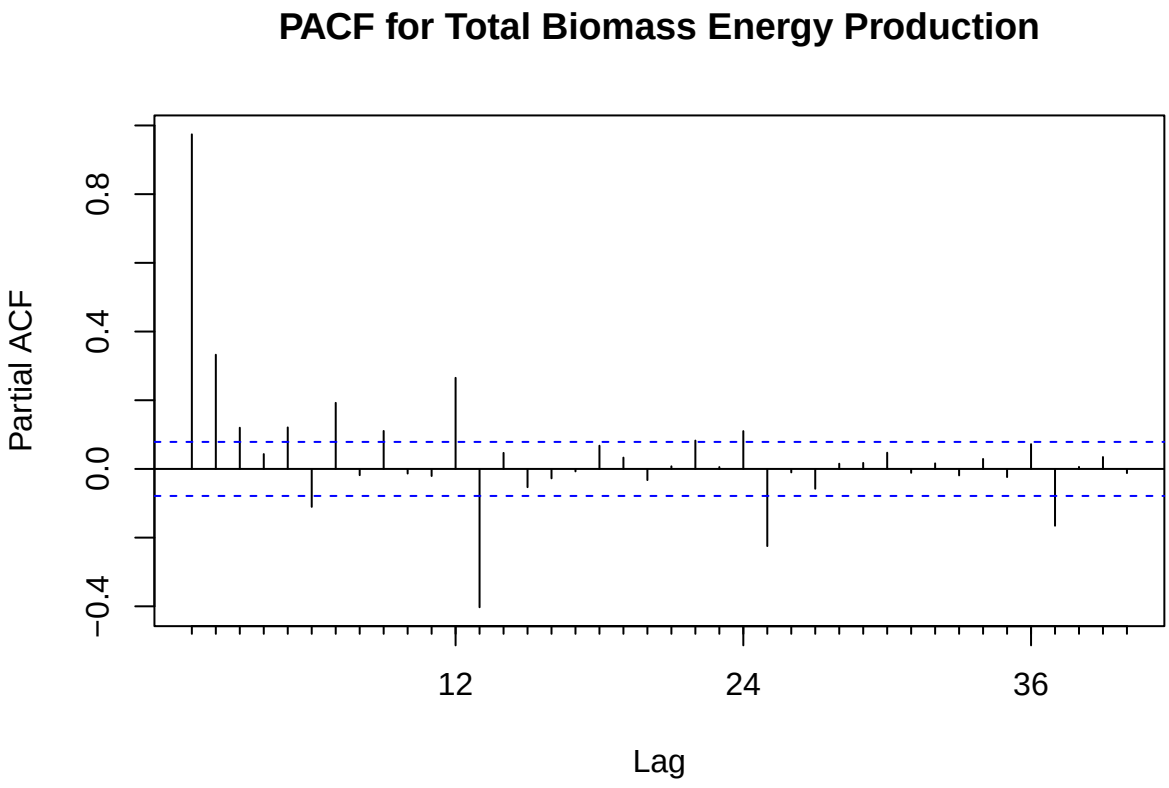

## Question 7 Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

In general, the ACF plots show the combined effects of all lags, and the PACF plots focus on direct relationships.

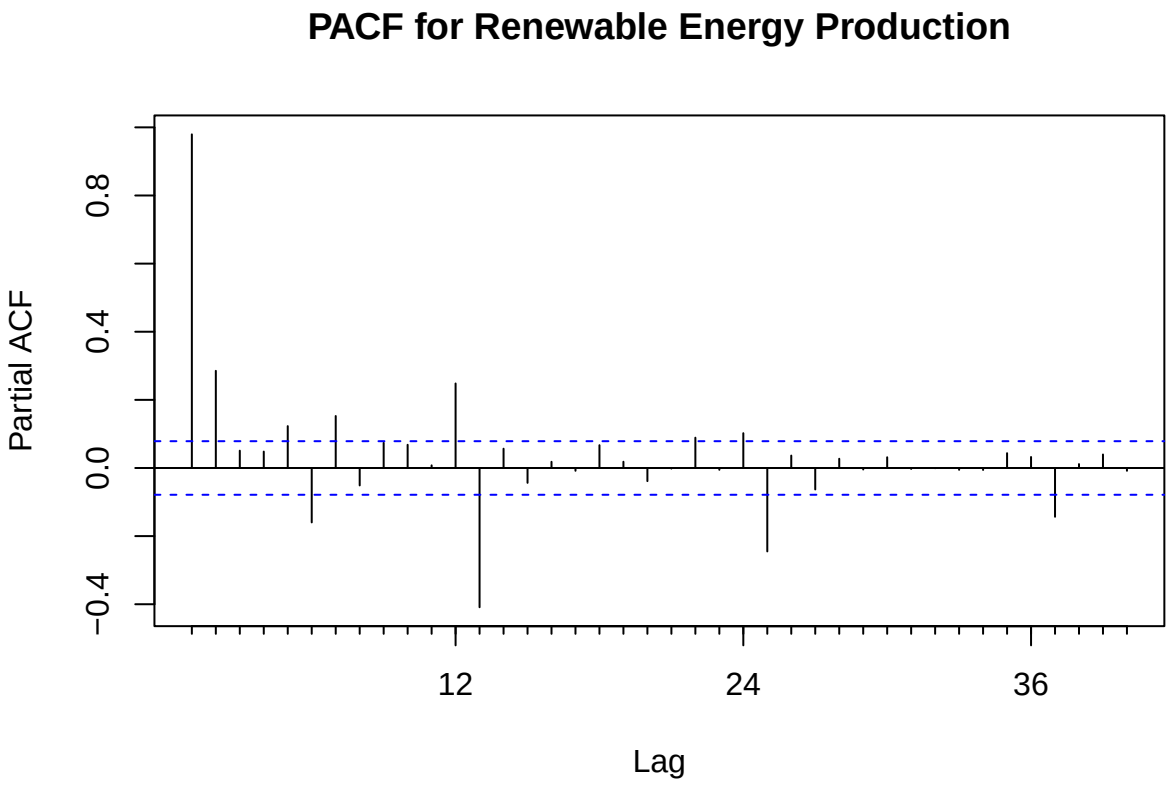
The PACF for plots 1 and 2 shows a significant spike at lag 1, followed by a sharp drop. This indicates that these series have a strong direct relationship with lag one but not with higher lags after accounting for lag 1. In contrast, the ACF plots for these series show a gradual decay, reflecting the combined influence of both direct and indirect correlations across multiple lags, which is typical for trending or persistent series.

The PACF for plot 3, however, shows several significant spikes at higher lags, exceeding the confidence interval. Unlike its ACF plot, which captures weaker overall correlations and appears dominated by noise, the PACF isolates specific lags with meaningful direct correlations.

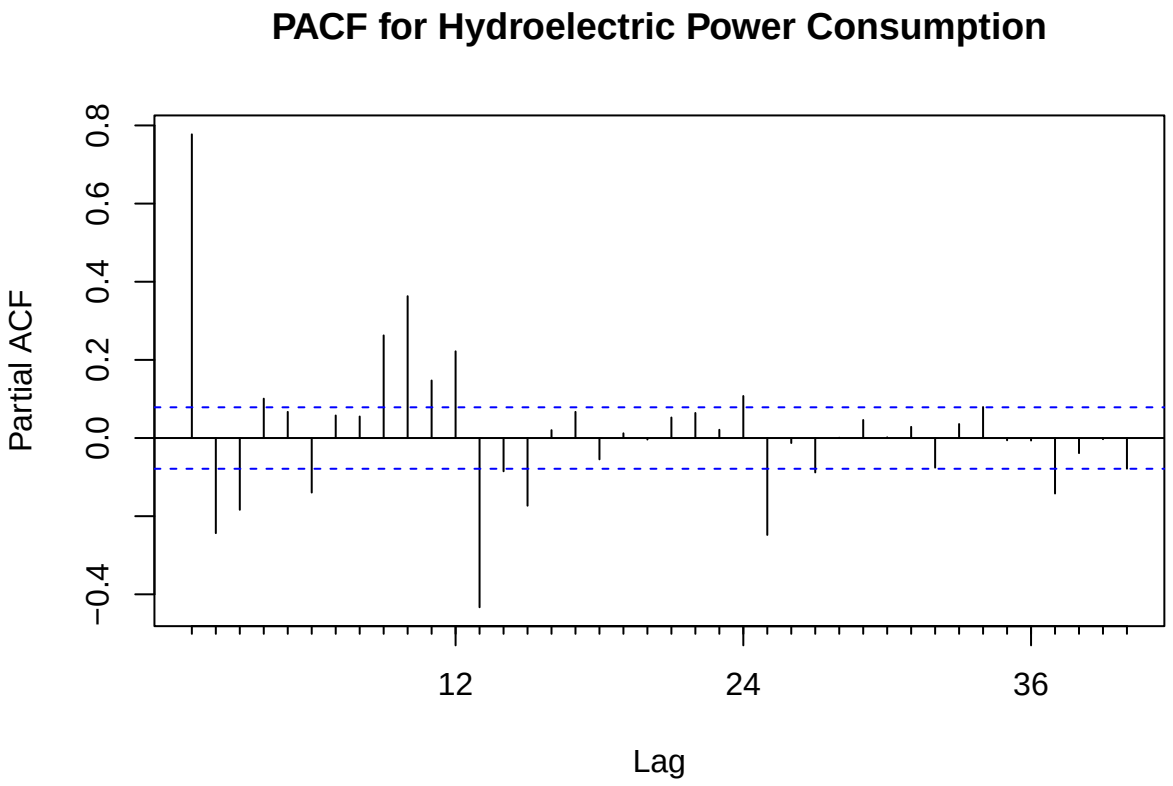
```
BE_pacf=Pacf(ts_energy_data2[,1],lag.max=40, plot=TRUE, main = "PACF for Total Biomass Energy Productio
```




```
RE_pacf=Pacf(ts_energy_data2[,2],lag.max=40, plot=TRUE, main = "PACF for Renewable Energy Production")
```




```
HP_pacf=Pacf(ts_energy_data2[,3],lag.max=40, plot=TRUE, main = "PACF for Hydroelectric Power Consumptio
```




Citation

OpenAI. (2025). ChatGPT (January 2025 Version) [Large language model]. https://chat.openai.com/