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

Assignment 2 - Due date 01/28/25

Aye Nyein Thu

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
library(lubridate)
library(ggplot2)
library(forecast)
library(dplyr)
library(corrplot)
library(knitr)
library(readxl)
library(openxlsx)

#Check working directory
getwd()
```

```
## [1] "/home/guest/TSA_Sp25"
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx" 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*.

```
#Import the data set
Energy <- read_excel(
  path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
  skip = 12, sheet="Monthly Data",col_names=FALSE)

#Extract the column names from row 11
Column_Names <- read_excel(
  path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
  skip = 10,n_max = 1, sheet="Monthly Data", col_names=FALSE)

#Assign the column names to the data set
colnames(Energy) <- Column_Names

#Format the date column
Energy$Month <- as.Date(Energy$Month, format = "%Y-%m-%d")
```

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.

```
#Select columns and change column names
Energy_cleaned <- Energy[,c(1,4:6)] %>%
  rename("Biomass" = "Total Biomass Energy Production",
         "Renewable" = "Total Renewable Energy Production",
         "Hydroelectric" = "Hydroelectric Power Consumption")

#Verify the data
head(Energy_cleaned)
```

```
## # A tibble: 6 x 4
##   Month      Biomass Renewable Hydroelectric
##   <date>       <dbl>     <dbl>         <dbl>
## 1 1973-01-01    130.      220.          89.6
## 2 1973-02-01    117.      197.          79.5
## 3 1973-03-01    130.      219.          88.3
## 4 1973-04-01    126.      209.          83.2
## 5 1973-05-01    130.      216.          85.6
## 6 1973-06-01    126.      208.          82.1
```

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().

```r
#Specify starting points
start_year <- as.numeric(format(min(Energy_cleaned$Month), "%Y"))
start_month <- as.numeric(format(min(Energy_cleaned$Month), "%m"))

#Print starting points
start_year
```

```
## [1] 1973
```

```r
start_month
```

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

```r
#Transform data frame to time series objects
Energy_ts <- ts(Energy_cleaned[,2:4],
                start = c(start_year, start_month), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```r
#Compute mean and standard deviation for three series
Energy_cleaned %>%
  summarise(
    Biomass_Mean = mean(Biomass),
    Biomass_SD = sd(Biomass),
    Renewable_Mean = mean(Renewable),
    Renewable_SD = sd(Renewable),
    Hydroelectric_Mean = mean(Hydroelectric),
    Hydroelectric_SD = sd(Hydroelectric)
  ) %>%
  kable()
```

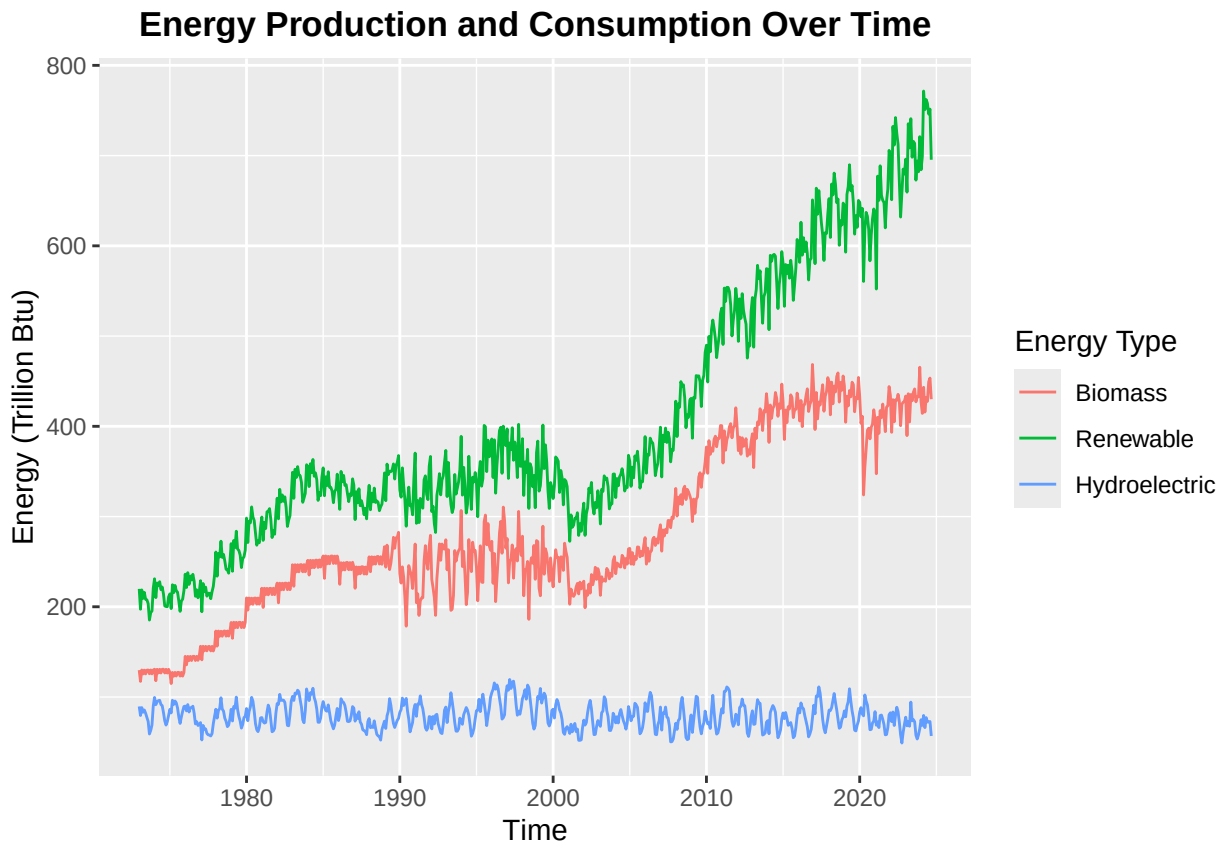
Biomass_Mean	Biomass_SD	Renewable_Mean	Renewable_SD	Hydroelectric_Mean	Hydroelectric_SD
282.6779	94.05815	402.0167	143.7927	79.55371	14.10737

Question 4

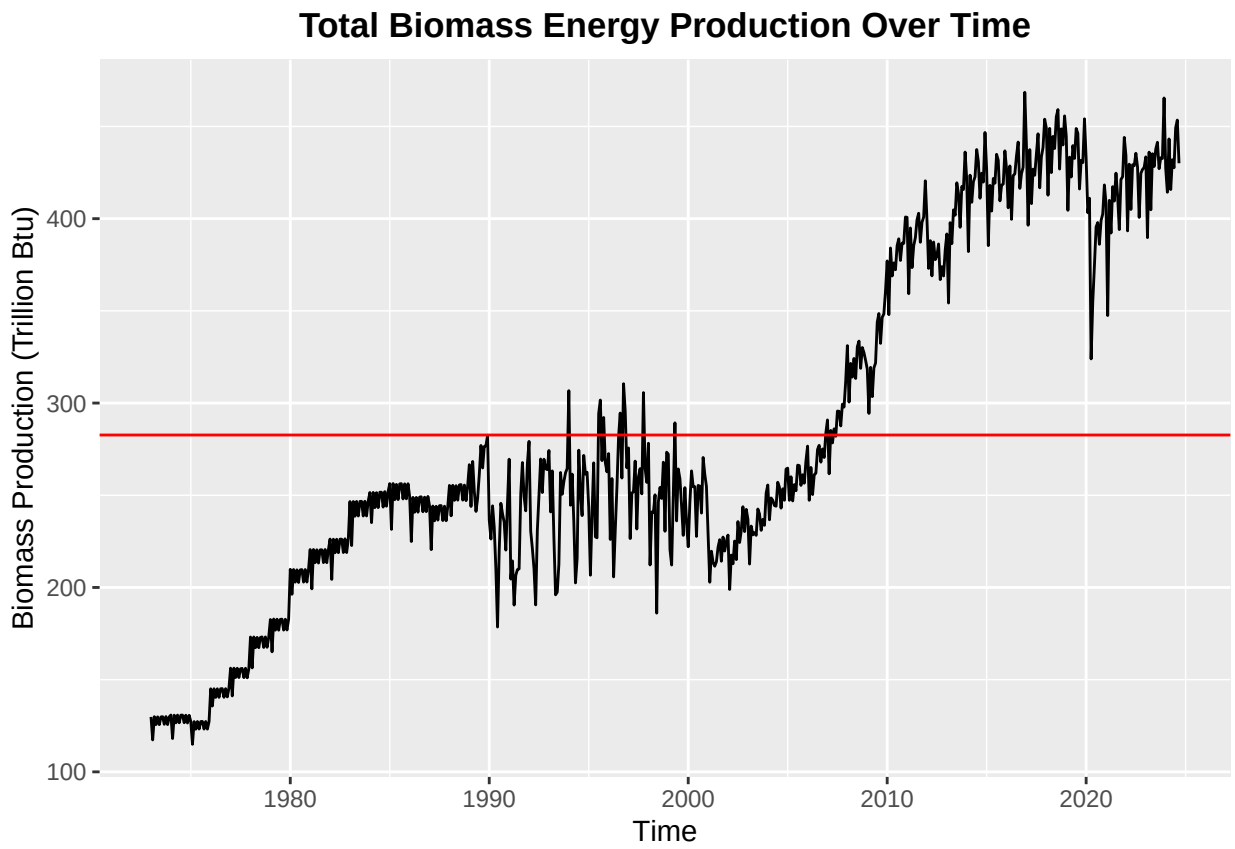
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```r
#Plot comparing all the varies
autoplot(Energy_ts) +
  xlab("Time") +
  ylab("Energy (Trillion Btu)") +
  labs(color="Energy Type") +
  ggtitle("Energy Production and Consumption Over Time") +
```

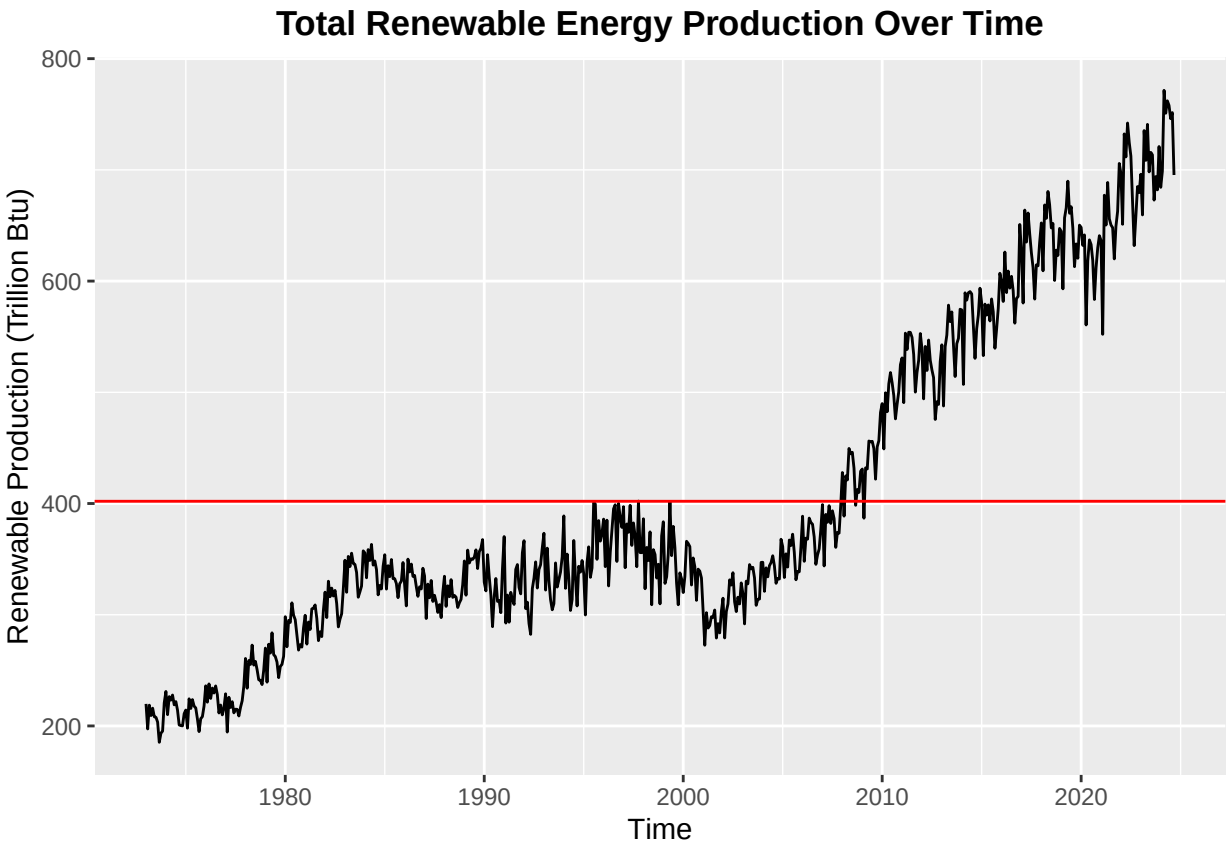
```
  theme(
    plot.title = element_text(hjust = 0.5,
                              face = "bold")
    )
```



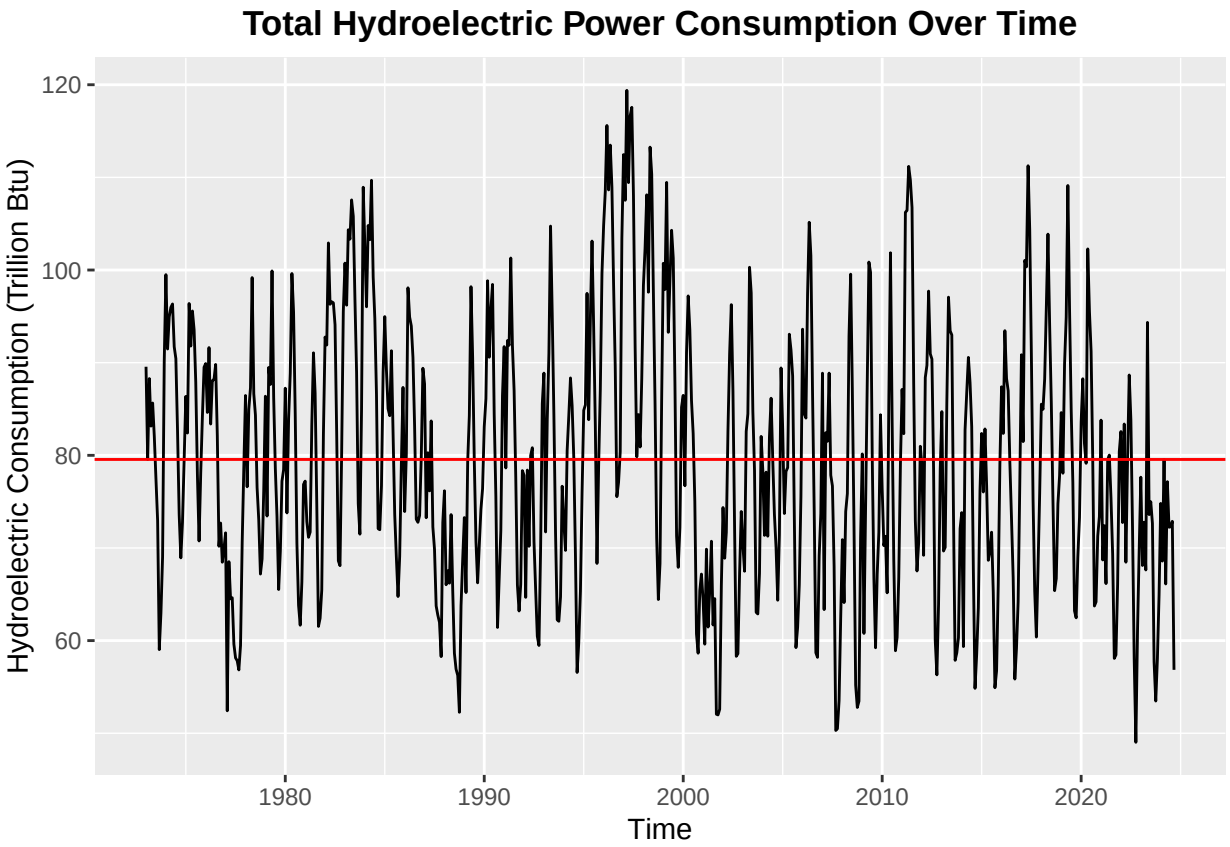

```
#Plot for Total Biomass Energy Production
autoplot(Energy_ts[, "Biomass"]) +
  xlab("Time") +
  ylab("Biomass Production (Trillion Btu)") +
  ggtitle("Total Biomass Energy Production Over Time") +
  theme(
    plot.title = element_text(hjust = 0.5,
                              face = "bold")
    ) +
  geom_hline(
    aes(yintercept = mean(Energy_ts[, "Biomass"])),
    color = "red"
    )
```




```
#Plot for Total Renewable Energy Production
autoplot(Energy_ts[, "Renewable"]) +
  xlab("Time") +
  ylab("Renewable Production (Trillion Btu)") +
  ggtitle("Total Renewable Energy Production Over Time") +
  theme(
    plot.title = element_text(hjust = 0.5,
                              face = "bold")
    ) +
  geom_hline(
    aes(yintercept = mean(Energy_ts[, "Renewable"])),
    color = "red"
    )
```




```
#Plot for Hydroelectric Power Consumption
autoplot(Energy_ts[, "Hydroelectric"]) +
  xlab("Time") +
  ylab("Hydroelectric Consumption (Trillion Btu)") +
  ggtitle("Total Hydroelectric Power Consumption Over Time") +
  theme(
    plot.title = element_text(hjust = 0.5,
                              face = "bold")
    ) +
  geom_hline(
    aes(yintercept = mean(Energy_ts[, "Hydroelectric"])),
    color = "red"
    )
```



When three series of energy production and consumption are plotted together, renewable energy production is large compared to biomass production. Both of them exhibits an increasing trends over time and production seems to raise by the same magnitude of rate until 2010. After 2010, the rate of increase in renewable energy production skyrocketed in magnitude while that in biomass was only moderate. The hydroelectric energy consumption does not exhibit an apparent increasing nor decreasing trends, but rather lots of ups and downs in the data series.

Biomass production was lower than the average until the mid of 1990s. After 2008, it has been enormously increasing. Similarly, renewable energy production was also higher than average starting from 2008. The hydroelectric tells the different story with the apparent episodes of ups and downs throughout the years.

Question 5

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

```
#Conduct correlation test for each series
cor.test (Energy_cleaned$Biomass,Energy_cleaned$Renewable)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Energy_cleaned$Biomass and Energy_cleaned$Renewable
## t = 95.677, df = 619, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9624198 0.9724443
## sample estimates:
##       cor
## 0.9678137
```

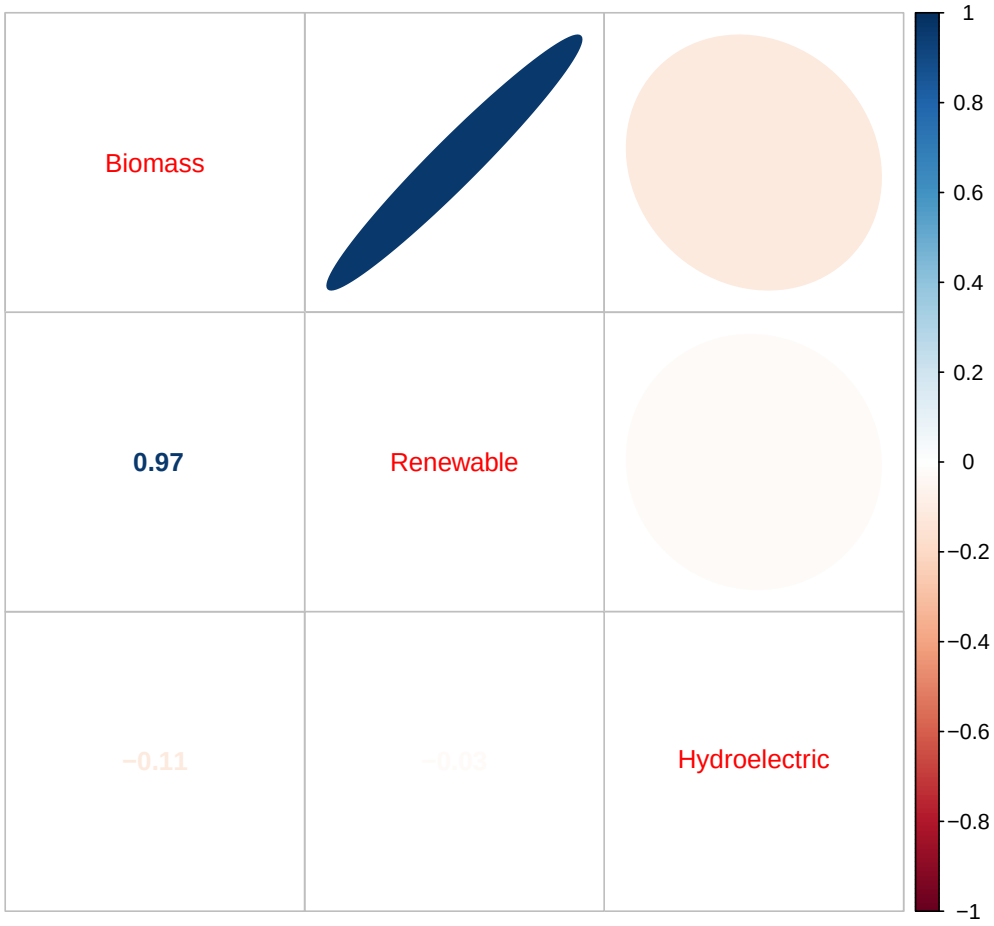
```
cor.test (Energy_cleaned$Biomass,Energy_cleaned$Hydroelectric)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Energy_cleaned$Biomass and Energy_cleaned$Hydroelectric
## t = -2.8623, df = 619, p-value = 0.004348
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.19125123 -0.03593747
## sample estimates:
##        cor
## -0.1142927
```

```
cor.test (Energy_cleaned$Renewable,Energy_cleaned$Hydroelectric)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Energy_cleaned$Renewable and Energy_cleaned$Hydroelectric
## t = -0.72583, df = 619, p-value = 0.4682
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1075925  0.0496312
## sample estimates:
##         cor
## -0.02916103
```

```
#Plot the correlation test
Energy_cor <- Energy_cleaned[,2:4] #subtract data
Cor_matrix <- cor(Energy_cor) #compute the correlation matrix
corrplot.mixed(Cor_matrix, upper = "ellipse") #plot
```

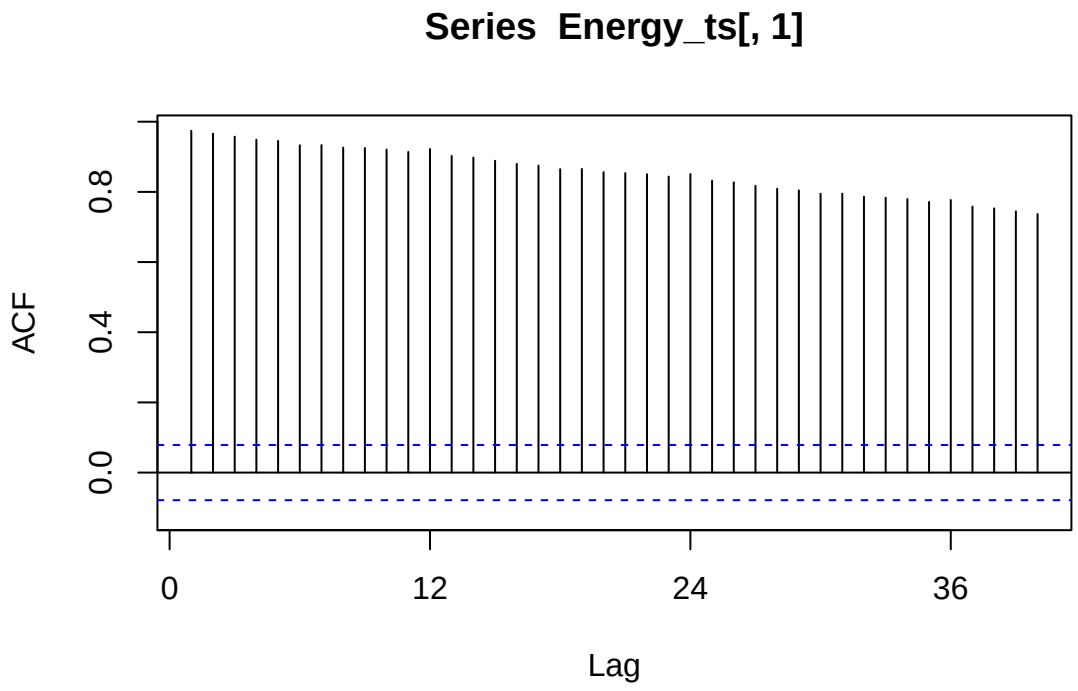


The above correlation tests and plot indicates that there is a strong positive correlation between biomass and renewable energy production and they are statistically significant. The correlation between biomass production and hydroelectric consumption is negative and weak but still significant statistically. Finally, although renewable and hydroelectric are going in a negative direction, it is not statistically significant.

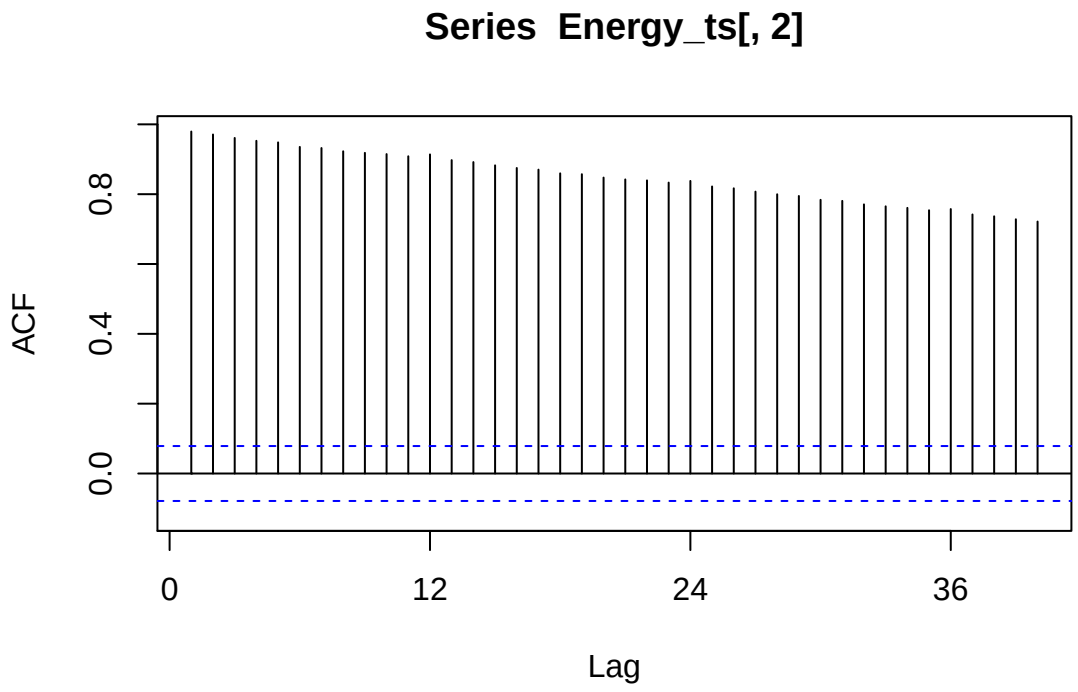
Question 6

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?

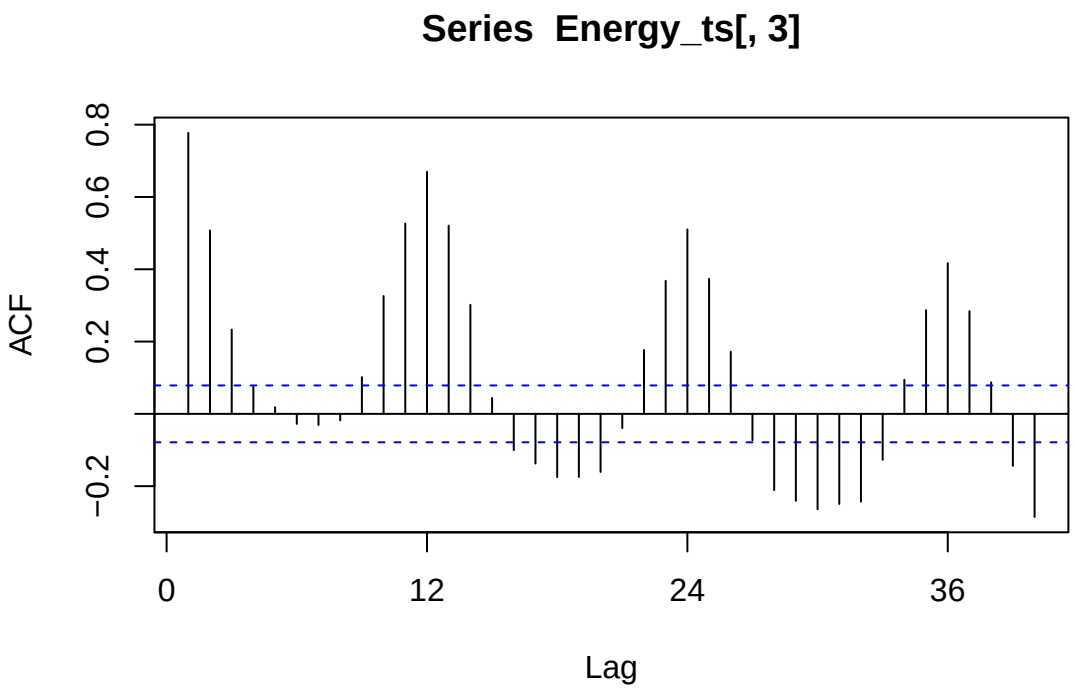
```
#Compute autocorrelation for Total Biomass Energy Production
Biomass_acf=Acf(Energy_ts[,1],lag.max=40, type="correlation", plot=TRUE)
```




```
#Compute autocorrelation for Total Renewable Energy Production
Renewable_acf=Acf(Energy_ts[,2],lag.max=40, type="correlation", plot=TRUE)
```



```r
#Compute autocorrelation for Total Hydroelectric Power Consumption
Hydroelectric_acf=Acf(Energy_ts[,3],lag.max=40, type="correlation", plot=TRUE)
```



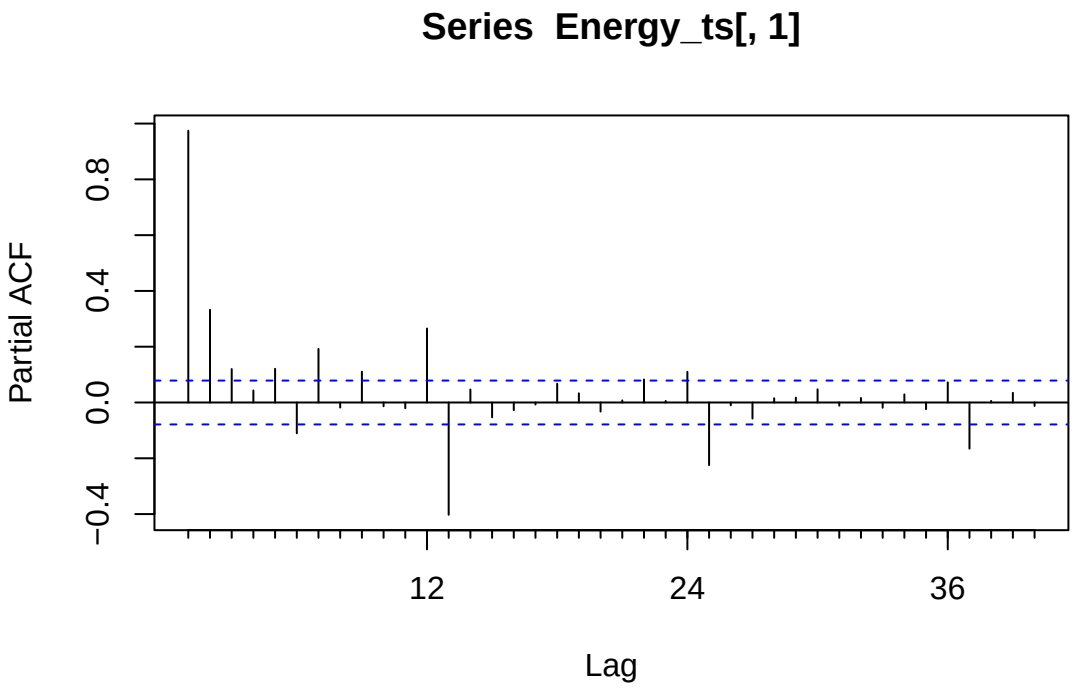
We have seen in the previous plots that biomass and renewable are going in the similar increasing trend and the correlation test indicates that they are strongly and positively correlated. In line with these assumptions, the autocorrelation functions of these two series are similar in nature. Both of them exhibits a high level of autocorrelation and all of them are statistically significant. It means that the present year's energy production relies heavily on the previous year's production for both data series. It is highly likely that both series will have an increasing trend.

On the other hand, the data of hydroelectric consumption is different from the above two. It is more likely to present a seasonal trend. The autocorrelation is not statistically significant in some points.

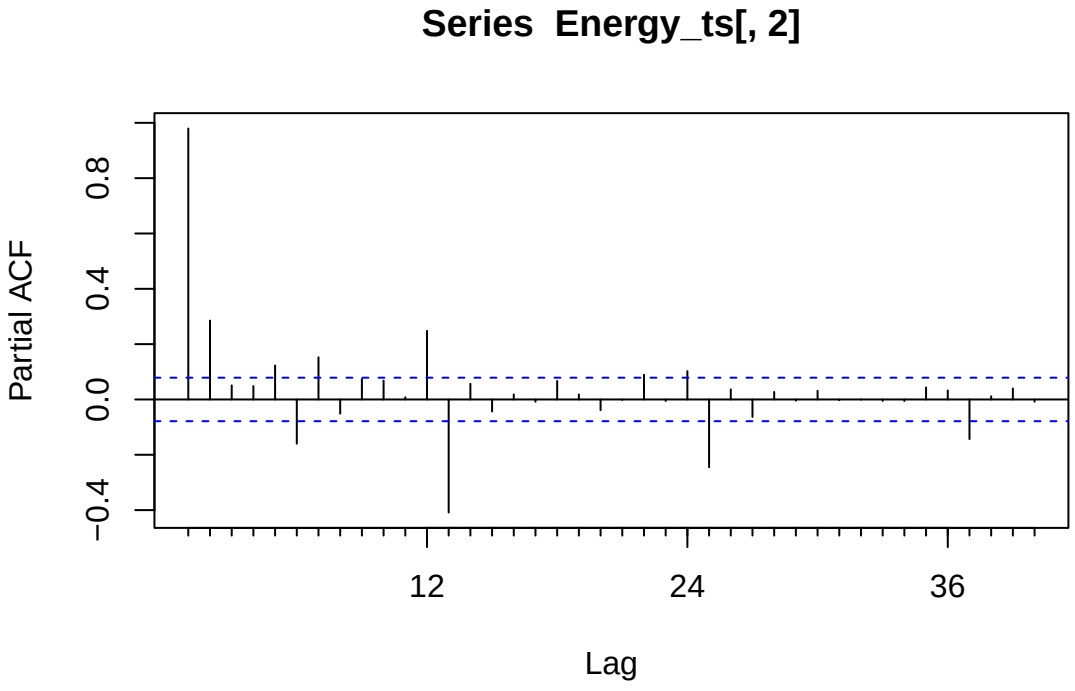
Question 7

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

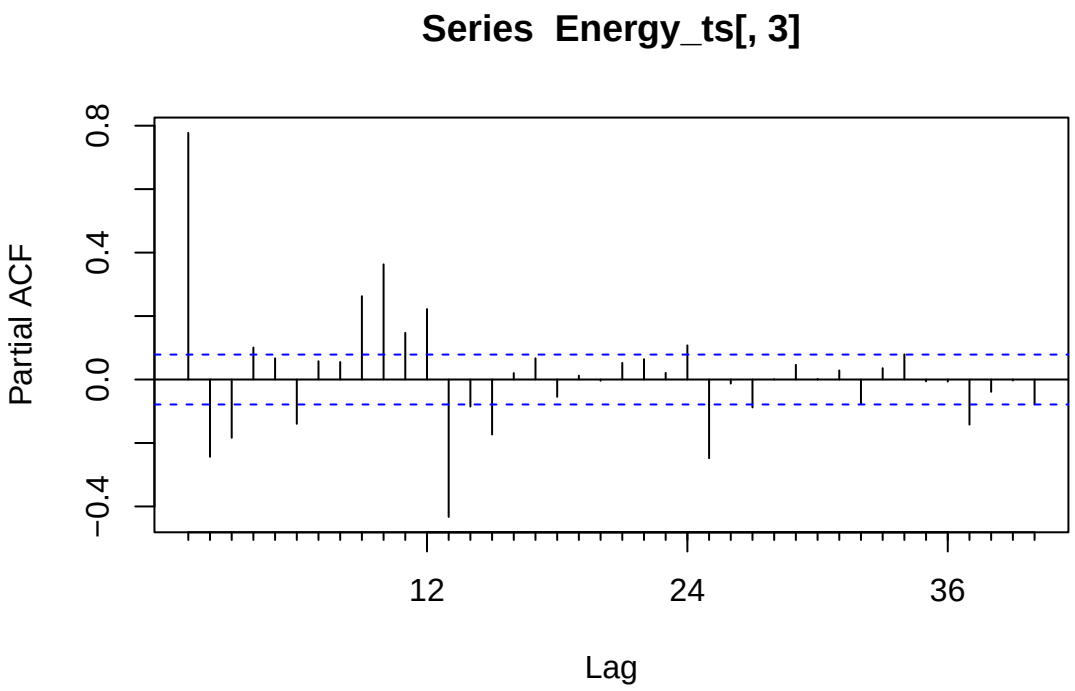
```r
#Compute partial autocorrelation for Total Biomass Energy Production
Biomass_pacf=Pacf(Energy_ts[,1],lag.max=40, plot=TRUE)
```




```
#Compute partial autocorrelation for Total Renewable Energy Production
Renewable_pacf=Pacf(Energy_ts[,2],lag.max=40, plot=TRUE)
```



```r
#Compute partial autocorrelation for Total Hydroelectric Power Consumption
Hydroelectric_pacf=Pacf(Energy_ts[,3],lag.max=40, plot=TRUE)
```



When biomass and renewable energy are plotted on partial autocorrelation function and remove the effect of intermediate variable, it indicated that the partial autocorrelation is not significant starting from year 3 and 2 respectively.

For the hydroelectric consumption, the PACF is not significant starting from year 2.