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

Assignment 2 - Due date 01/27/26

Josh Salzberg

## 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
#install.packages("forecast")
#install.packages("tseries")
#install.packages('readxl')
#install.packages('openxlsx')
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
```

```r
library(readxl)
library(openxlsx)
library(here)
```

```
## here() starts at C:/Users/jhsal/OneDrive - Duke University/797/TSA2026/TSA_Sp26
```

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.1     v stringr   1.6.0
## v lubridate 1.9.4     v tibble    3.3.1
## v purrr     1.2.1     v tidyr     1.3.1
## v readr     2.1.6
```

```
## -- 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
```

## 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.

```r
#Importing data set
energydata <- read_excel(path = here("Data",'Table_10.1_Renewable_Energy_Production_and_Consumption_by_S
                         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
# Extracting column names
lost_col_names <- read_excel(path = here("Data",'Table_10.1_Renewable_Energy_Production_and_Consumption_
                         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 column names
colnames(energydata) <- lost_col_names

head(energydata)
```

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

```
str(energydata)
```

```
## tibble [633 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Month                              : POSIXct[1:633], format: "1973-01-01" "1973-02-01" ...
##  $ Wood Energy Production             : num [1:633] 130 117 130 125 130 ...
##  $ Biofuels Production                : chr [1:633] "Not Available" "Not Available" "Not Available" "
##  $ Total Biomass Energy Production    : num [1:633] 130 117 130 126 130 ...
##  $ Total Renewable Energy Production : num [1:633] 220 197 219 209 216 ...
##  $ Hydroelectric Power Consumption    : num [1:633] 89.6 79.5 88.3 83.2 85.6 ...
##  $ Geothermal Energy Consumption      : num [1:633] 0.49 0.448 0.464 0.542 0.505 0.579 0.614 0.579 0.
##  $ Solar Energy Consumption           : chr [1:633] "Not Available" "Not Available" "Not Available" "
##  $ Wind Energy Consumption            : chr [1:633] "Not Available" "Not Available" "Not Available" "
##  $ Wood Energy Consumption            : num [1:633] 130 117 130 125 130 ...
##  $ Waste Energy Consumption           : num [1:633] 0.157 0.144 0.176 0.174 0.21 0.176 0.17 0.184 0.1
##  $ Biofuels Consumption               : chr [1:633] "Not Available" "Not Available" "Not Available" "
##  $ Total Biomass Energy Consumption   : num [1:633] 130 117 130 126 130 ...
##  $ Total Renewable Energy Consumption: num [1:633] 220 197 219 209 216 ...
```

## 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.

```
tsa_prep <- energydata %>%
  select(TotalBiomassProduction = `Total Biomass Energy Production`,
         TotalRenewablesProduction = `Total Renewable Energy Production`,
         HydroelectricConsumption = `Hydroelectric Power Consumption`)
head(tsa_prep)
```

```
## # A tibble: 6 x 3
##   TotalBiomassProduction TotalRenewablesProduction HydroelectricConsumption
##                    <dbl>                     <dbl>                     <dbl>
## 1                   130.                      220.                      89.6
## 2                   117.                      197.                      79.5
## 3                   130.                      219.                      88.3
## 4                   126.                      209.                      83.2
## 5                   130.                      216.                      85.6
## 6                   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().

```
energy_ts <- ts(data = tsa_prep, start= c(1973, 1), frequency = 12)

head(energy_ts)
```

```
##          TotalBiomassProduction TotalRenewablesProduction
## 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
##          HydroelectricConsumption
## 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.

```
mean(energy_ts[,1])
```

```
## [1] 286.0489
```

```
mean(energy_ts[,2])
```

```
## [1] 409.1952
```

```
mean(energy_ts[,3])
```

```
## [1] 79.35682
```

```r
sd(energy_ts[,1])
```

```
## [1] 96.21209
```

```r
sd(energy_ts[,2])
```

```
## [1] 151.4223
```

```r
sd(energy_ts[,3])
```

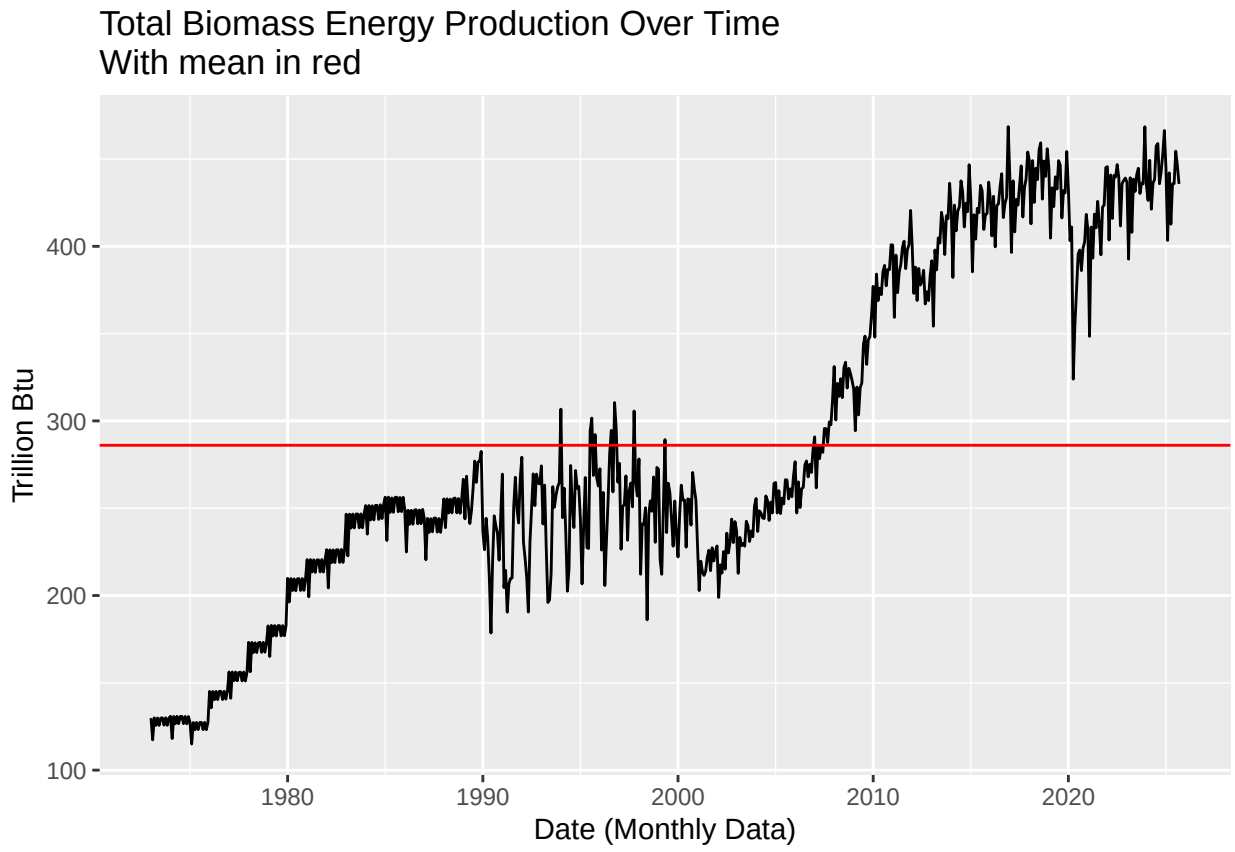
```
## [1] 14.1202
```

Total Biomass Energy Production: Mean = 286.0489; standard deviation = 96.21209 Total Renewables Production: Mean = 409.1952; standard deviation = 151.4223 Total Hydropower Energy Consumption: Mean = 79.35682; standard deviation = 14.1202

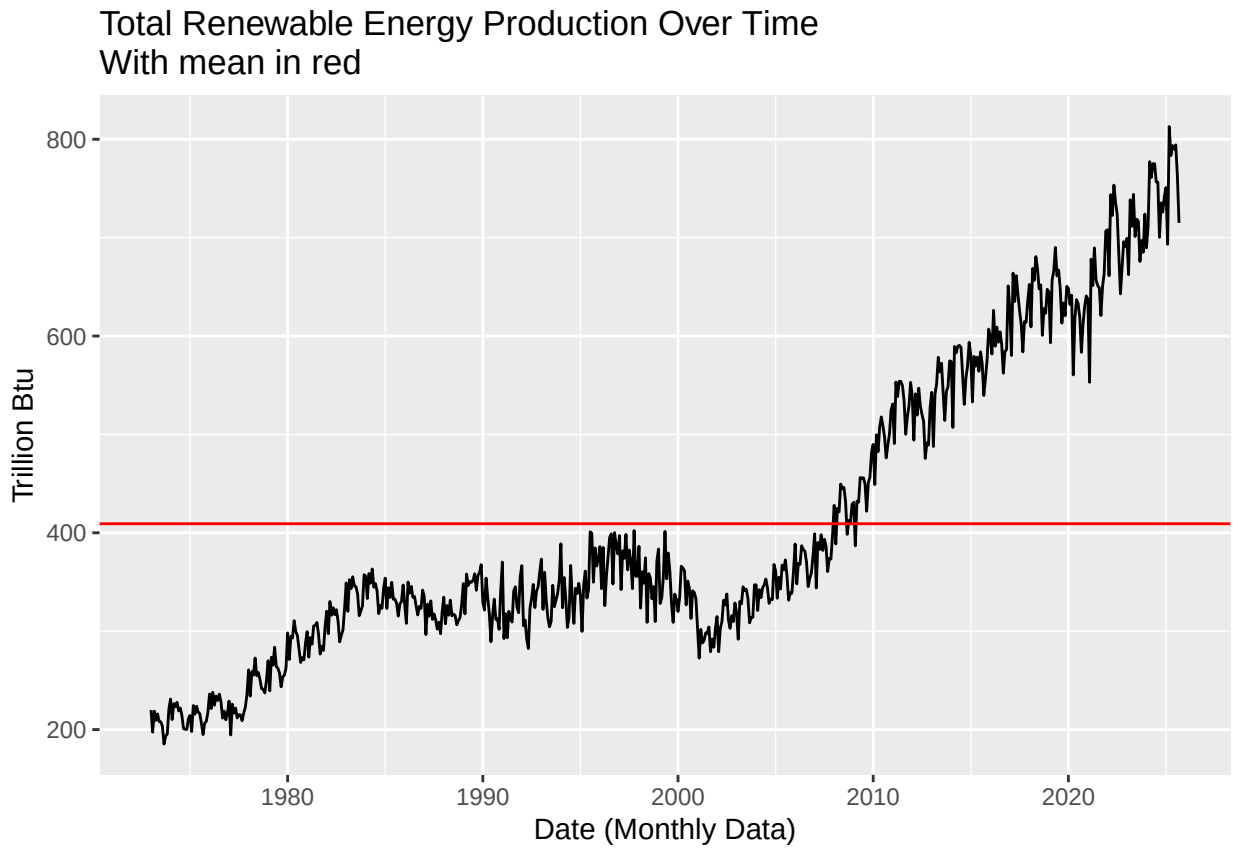
## Question 4

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.

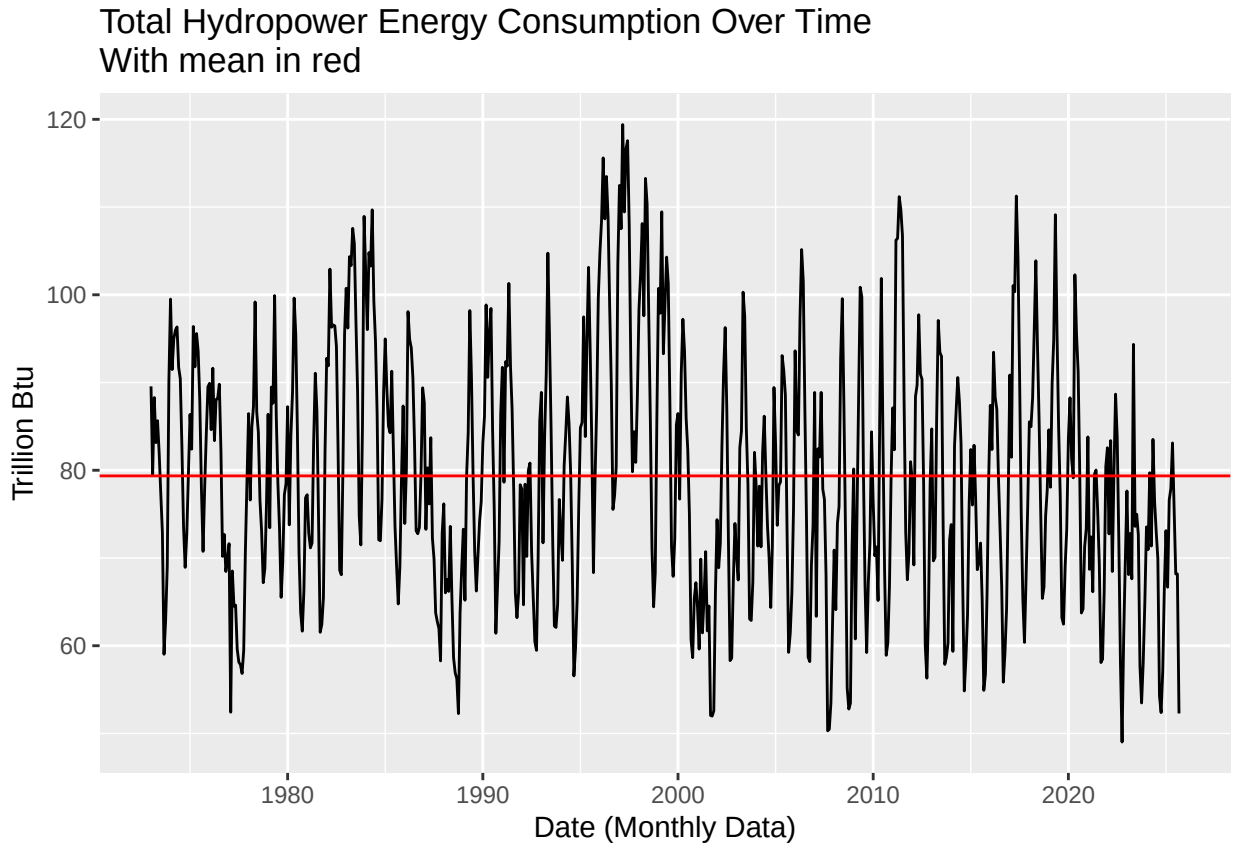
```r
autoplot1 <- autoplot(object = energy_ts[,1],
         main = "Total Biomass Energy Production Over Time\nWith mean in red",
         xlab = "Date (Monthly Data)",
         ylab = "Trillion Btu")+
  geom_hline(yintercept = mean(energy_ts[,1]), color = "red")

autoplot2 <- autoplot(object = energy_ts[,2],
         main = "Total Renewable Energy Production Over Time\nWith mean in red",
         xlab = "Date (Monthly Data)",
         ylab = "Trillion Btu")+
  geom_hline(yintercept = mean(energy_ts[,2]), color = "red")

autoplot3 <- autoplot(object = energy_ts[,3],
         main = "Total Hydropower Energy Consumption Over Time\nWith mean in red",
         xlab = "Date (Monthly Data)",
         ylab = "Trillion Btu")+
  geom_hline(yintercept = mean(energy_ts[,3]), color = "red")

autoplot1
```


Total Biomass Energy Production Over Time
With mean in red
Trillion Btu
400
300
200
100
1980
1990
2000
2010
2020
Date (Monthly Data)

```
autoplot2
```


Total Renewable Energy Production Over Time
With mean in red
800
600
400
200
Trillion Btu
1980
1990
2000
2010
2020
Date (Monthly Data)

```
autoplot3
```




Total Biomass Energy Production shows very regular annual and decadal seasonality from the beginning in 1973, with cycles of growth, until the 1990s. This marks a period of greater variability and irregularity, while still maintaining some seasonal components. From 2000 until the end of the series, there's a marked positive trend, with only a few major disruptions, aligning with the Great Recession and COVID. Total Renewable Energy Production demonstrates less structured seasonality compared to the biomass chart. There is little change in total production until a major trend of growth begins around 2001 or 2002; from there, the seasonality is more dramatic. Total Hydropower Energy Consumption shows a very different story; while there is seasonality, it is of longer duration, lending itself to a longer memory. Overall, there is little discernable trend.

## Question 5

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

```
cor(energy_ts)
```

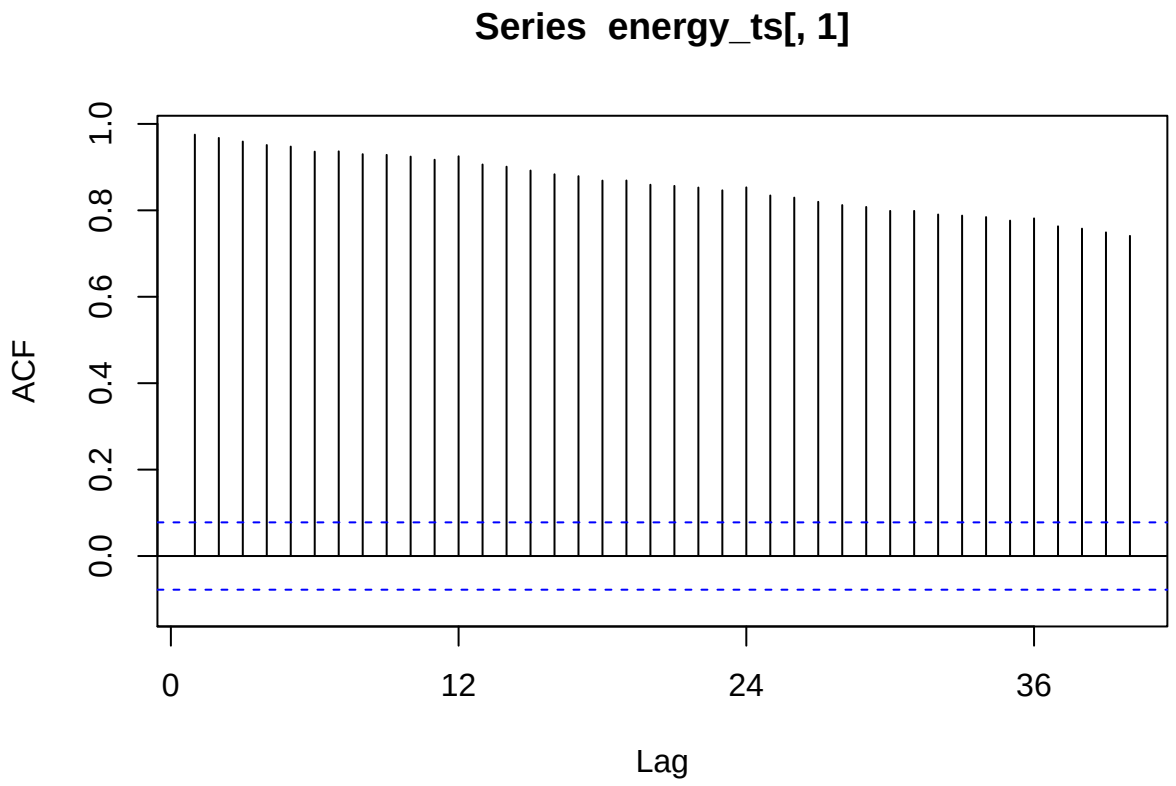
```
##                           TotalBiomassProduction TotalRenewablesProduction
## TotalBiomassProduction                 1.0000000                 0.96529851
## TotalRenewablesProduction              0.9652985                 1.00000000
## HydroelectricConsumption              -0.1347374                -0.05842436
##                           HydroelectricConsumption
## TotalBiomassProduction                -0.13473742
## TotalRenewablesProduction             -0.05842436
## HydroelectricConsumption               1.00000000
```

Biomass and Renewable Energy data show very strong positive correlation to one another, at 0.965. On the other hand, Hydroelectric Energy Consumption shows negligible correlation to Renewable Energy data (-0.058), and extremely weak negative correlation to Biomass Production (-0.135).

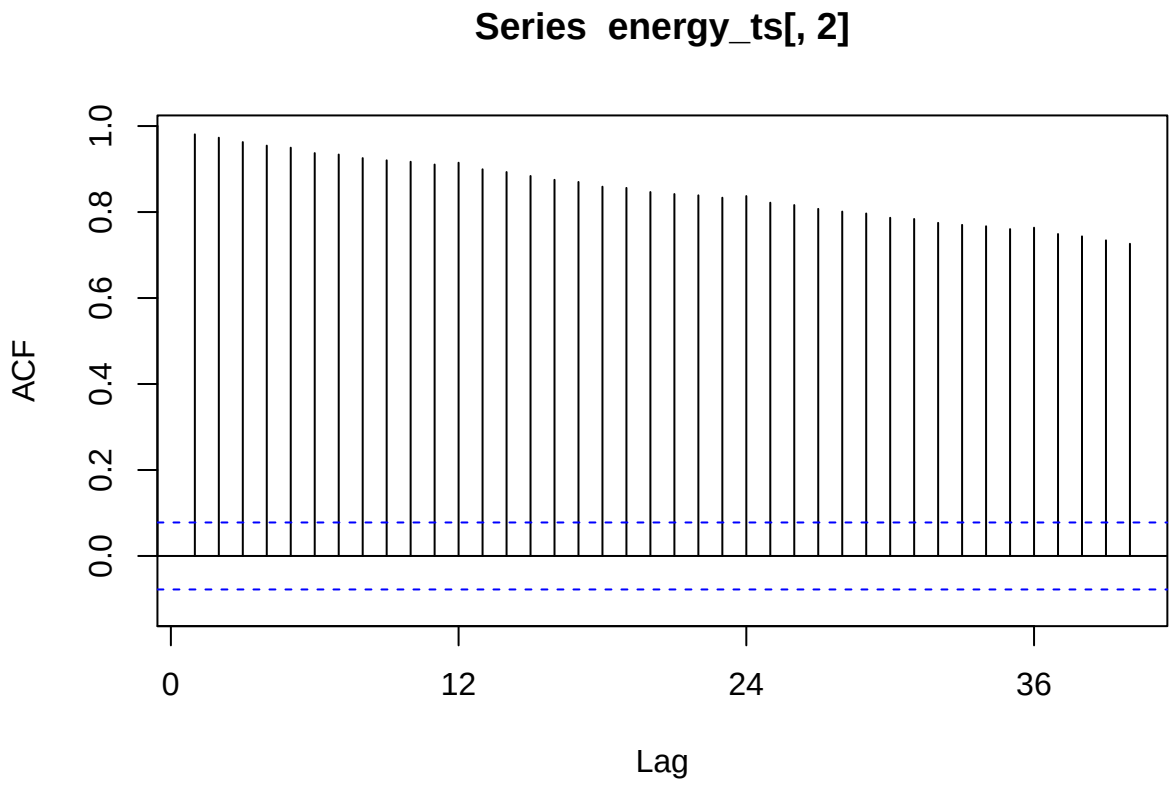
### Question 6

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?

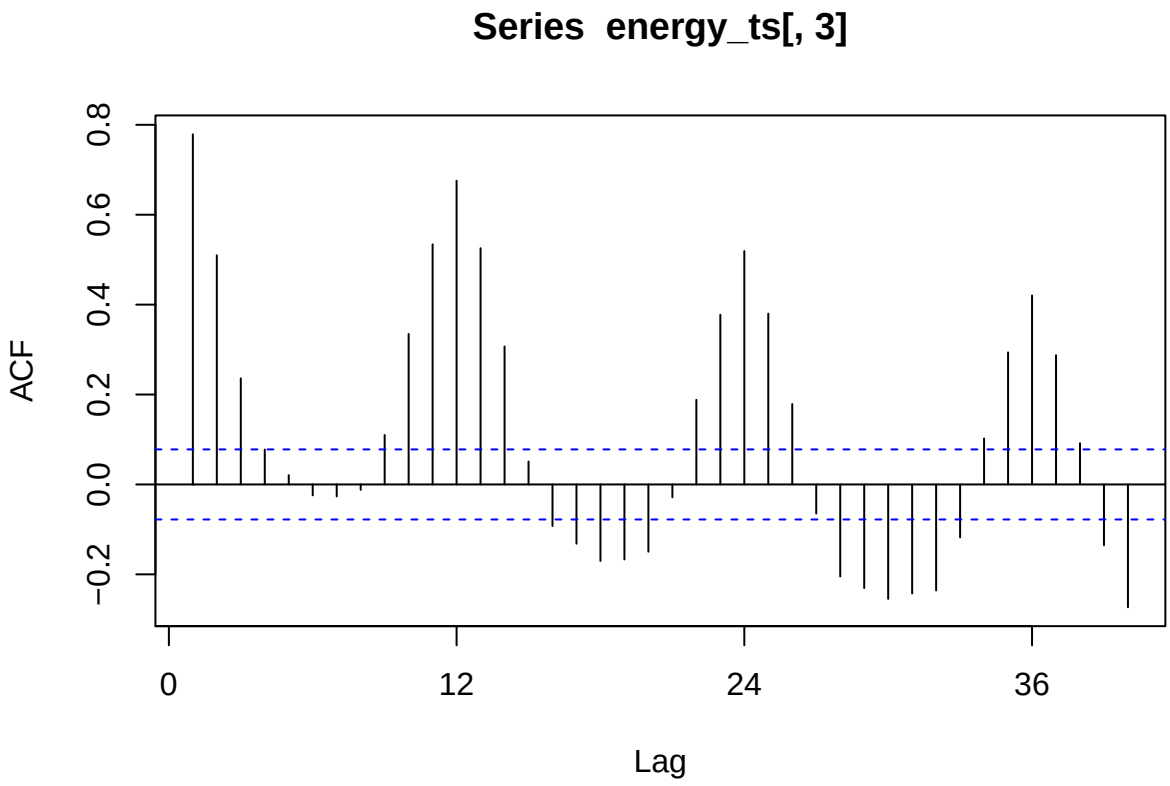
```
Acf1 <- Acf(energy_ts[,1], lag.max = 40)
```




```
Acf2 <- Acf(energy_ts[,2], lag.max = 40)
```




```
Acf3 <- Acf(energy_ts[,3], lag.max = 40)
```

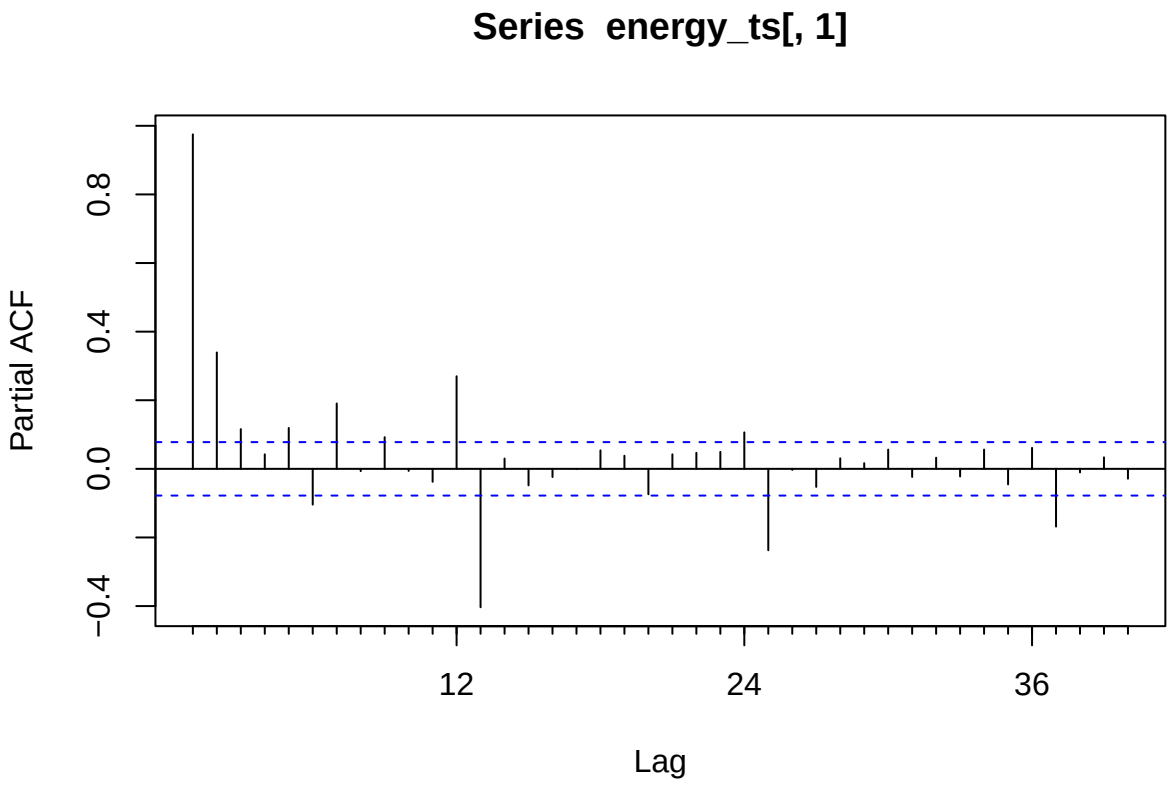



Total Biomass Production and Total Renewable Production both show very strong positive autocorrelation from 1 to 40, with a slow decline. The magnitude is slightly lower for the Total Renewable graph, but still very strong. This implies no major seasonality. Hydropower Consumption shows very intense seasonality and a much more significant decline in autocorrelation as the lags progress. There is also some negative autocorrelation at significant values. This graph tells a very different story.

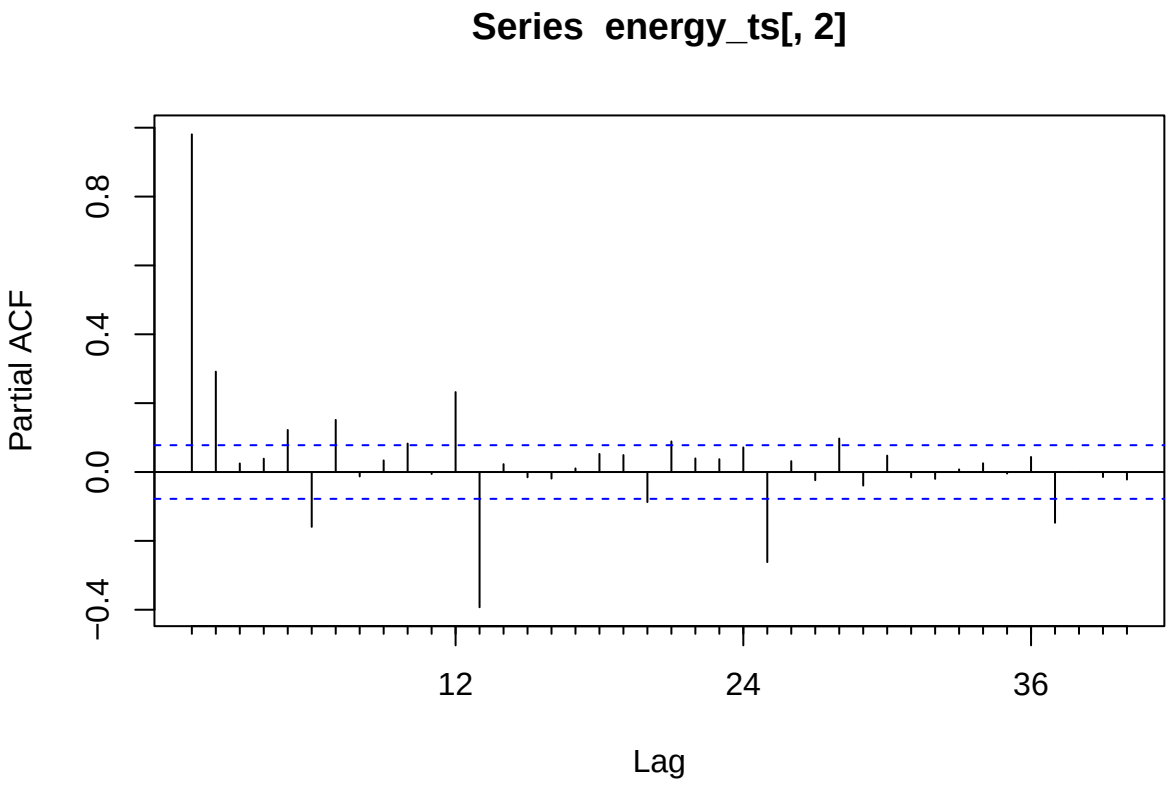
## Question 7

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

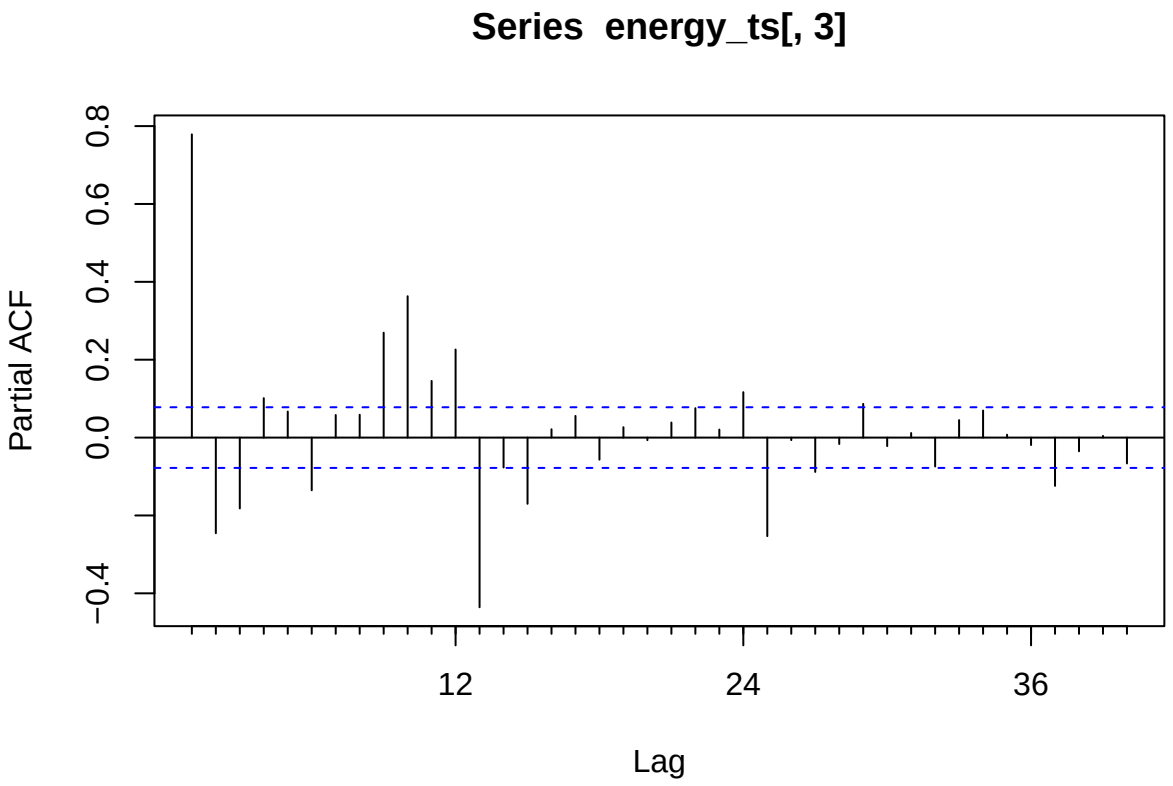
```
Pacf1 <- Pacf(energy_ts[,1], lag.max = 40)
```




```
Pacf2 <- Pacf(energy_ts[,2], lag.max = 40)
```




```
Paf3 <- Pacf(energy_ts[,3], lag.max = 40)
```




While Biomass & Renewable production PACF graphs look extremely similar to one another, they differ greatly from the ACF plots. We see that only a few lags are statistically significant, primarily on 1, 12 and barely 24 indicating some annual seasonality, but overall the autocorrelation dissipates significantly after the 13th lag. The hydropower PACF indicates that there is minimal autocorrelation for lags 5 - 8, while 9 through 12 have some notable correlation values. Lags 13 and 25 show significant negative autocorrelation. There appears to be some remnants of seasonality in the hydropower PACF but overall it shows a short memory.