ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 2 - Due date 02/25/24

David Robinson

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(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.3.2
```

```
## 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(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 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

getwd()
```

```
## [1] "C:/Users/dhr20/OneDrive - Duke University/1 - Academics/1 - First Year/2 - Spring 2024/3 - Time
```

```r
raw_energy_data <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_So
```

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

```r
colnames(raw_energy_data)=c("Month",
                            "Wood Energy Production",
                            "Biofuels Production",
                            "Total Biomass Energy Production",
                            "Total Renewable Energy Production",
                            "Hydroelectric Power Consumption",
                            "Geothermal Energy Consumption",
                            "Solar Energy Consumption",
                            "Wind Energy Consumption",
                            "Wood Energy Consumption",
```

```
                              "Waste Energy Consumption",
                              "Biofuels Consumption",
                              "Total Biomass Energy Consumption",
                              "Total Renewable Energy Consumption")
```

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.

```
raw_energy_data <- raw_energy_data[,1:6]
raw_energy_data_dates <- raw_energy_data[,1]
raw_energy_data_others <- raw_energy_data[,4:6]
raw_energy_data <- cbind(raw_energy_data_dates,raw_energy_data_others)

head(raw_energy_data)
```

```
##         Month Total Biomass Energy Production Total Renewable Energy Production
## 1 1973-01-01                         129.787                           219.839
## 2 1973-02-01                         117.338                           197.330
## 3 1973-03-01                         129.938                           218.686
## 4 1973-04-01                         125.636                           209.330
## 5 1973-05-01                         129.834                           215.982
## 6 1973-06-01                         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().

```
ts_energy_data <- ts(raw_energy_data[,2:4], start=c(1973,1), frequency=12)

head(ts_energy_data)
```

```
##          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
```

```r
tail(ts_energy_data)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Apr 2023                         404.131                           699.747
## May 2023                         437.506                           740.660
## Jun 2023                         429.839                           691.709
## Jul 2023                         437.109                           711.895
## Aug 2023                         439.521                           711.962
## Sep 2023                         422.351                           666.253
##          Hydroelectric Power Consumption
## Apr 2023                          59.646
## May 2023                          93.759
## Jun 2023                          66.434
## Jul 2023                          72.463
## Aug 2023                          72.150
## Sep 2023                          56.284
```

Question 3

Compute mean and standard deviation for these three series.

```r
#For Total Biomass Energy Production
mean_biomass_energy_production <- mean(ts_energy_data[,1])
sd_biomass_energy_production <- sd(ts_energy_data[,1])

mean_biomass_energy_production
```

```
## [1] 279.8046
```

```r
sd_biomass_energy_production
```

```
## [1] 92.66504
```

```r
#For Total Renewable Energy Production
mean_renewable_energy_production <- mean(ts_energy_data[,2])
sd_renewable_energy_production <- sd(ts_energy_data[,2])

mean_renewable_energy_production
```

```
## [1] 395.7213
```

```r
sd_renewable_energy_production
```

```
## [1] 137.7952
```

```r
#For Hydroelectric Power Consumption
mean_hydroelectric_energy_production <- mean(ts_energy_data[,3])
sd_hydroelectric_energy_production <- sd(ts_energy_data[,3])

mean_hydroelectric_energy_production
```

```
## [1] 79.73071
```

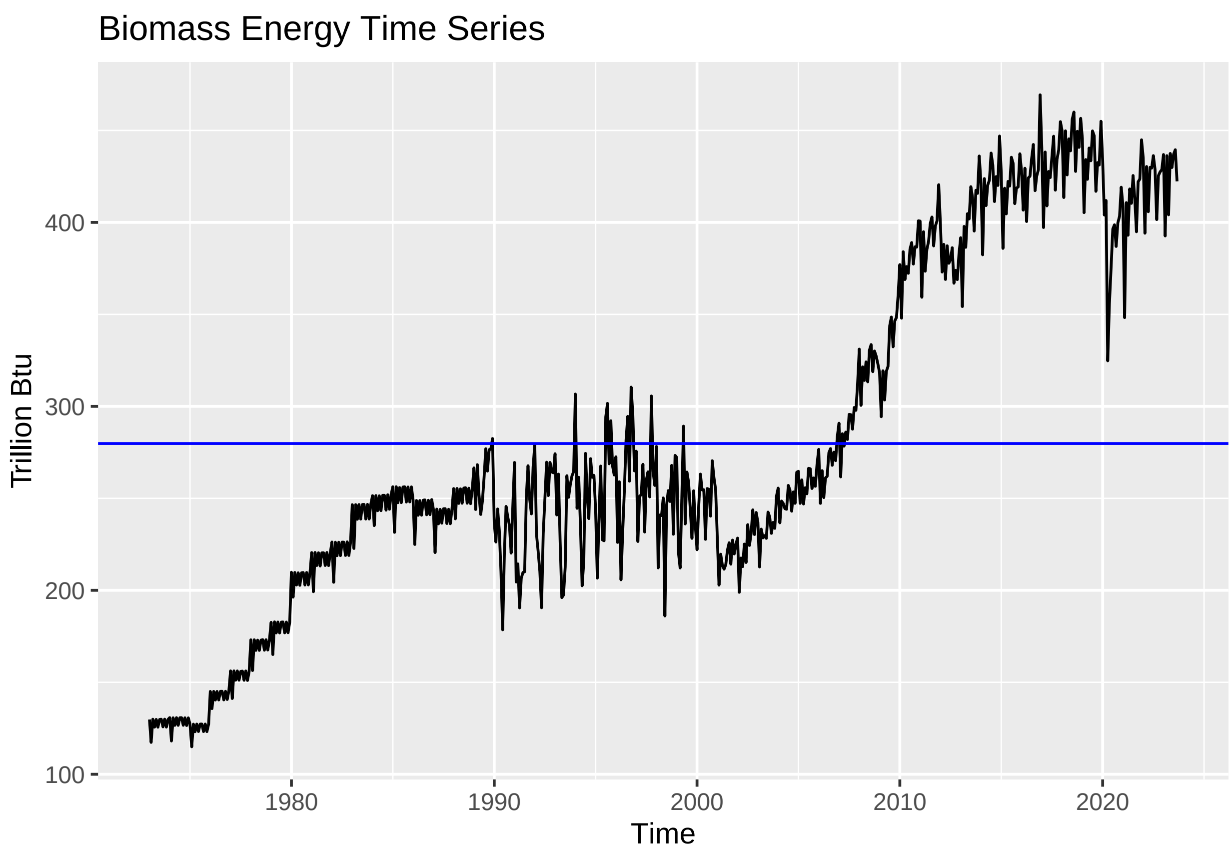
```r
sd_hydroelectric_energy_production
```

```
## [1] 14.14734
```

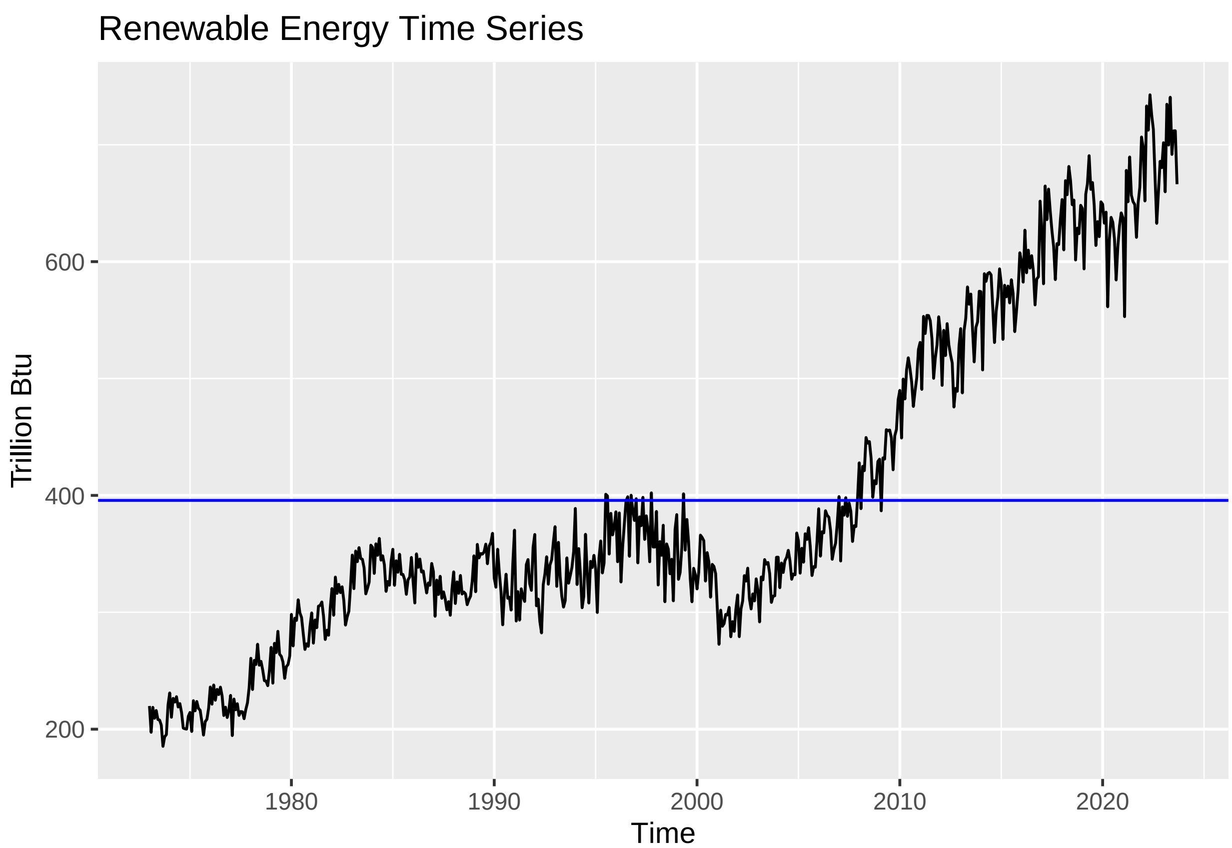
Question 4

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.

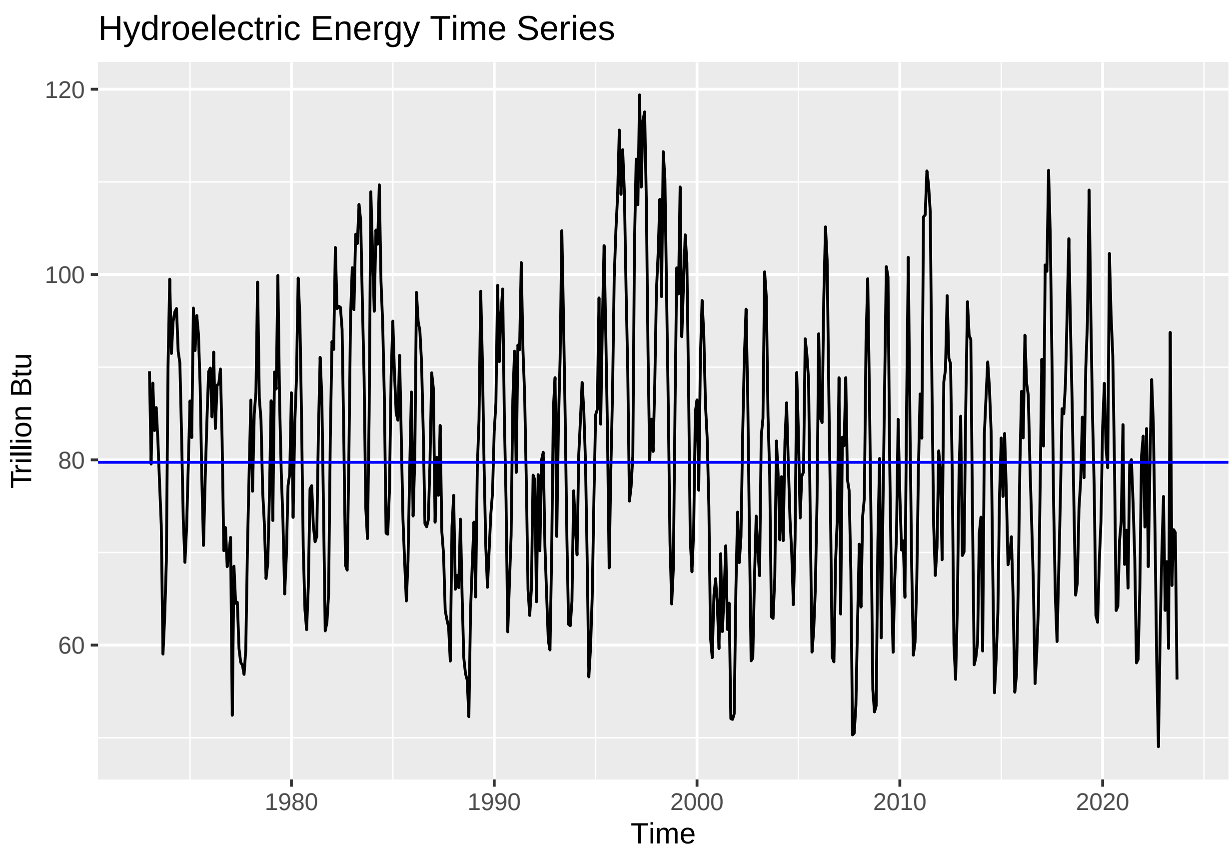
```r
autoplot(ts_energy_data[,1]) +
  ggtitle("Biomass Energy Time Series") +
  xlab("Time") +
  ylab("Trillion Btu") +
  geom_hline(yintercept = mean_biomass_energy_production, color="blue")
```




```r
autoplot(ts_energy_data[,2]) +
  ggtitle("Renewable Energy Time Series") +
  xlab("Time") +
  ylab("Trillion Btu") +
  geom_hline(yintercept = mean_renewable_energy_production, color="blue")
```




```r
autoplot(ts_energy_data[,3]) +
  ggtitle("Hydroelectric Energy Time Series") +
  xlab("Time") +
  ylab("Trillion Btu") +
  geom_hline(yintercept = mean_hydroelectric_energy_production, color="blue")
```



Question 5

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

```
cor.test(ts_energy_data[,1],ts_energy_data[,2])
```

```
##
##  Pearson's product-moment correlation
##
## data:  ts_energy_data[, 1] and ts_energy_data[, 2]
## t = 99.608, df = 607, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9657830 0.9749987
## sample estimates:
##       cor
## 0.9707462
```

```
cor.test(ts_energy_data[,1],ts_energy_data[,3])
```

```
##
##  Pearson's product-moment correlation
##
## data:  ts_energy_data[, 1] and ts_energy_data[, 3]
```

```
## t = -2.3902, df = 607, p-value = 0.01714
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1746734 -0.0172452
## sample estimates:
##         cor
## -0.09656318
```

```
cor.test(ts_energy_data[,2],ts_energy_data[,3])
```

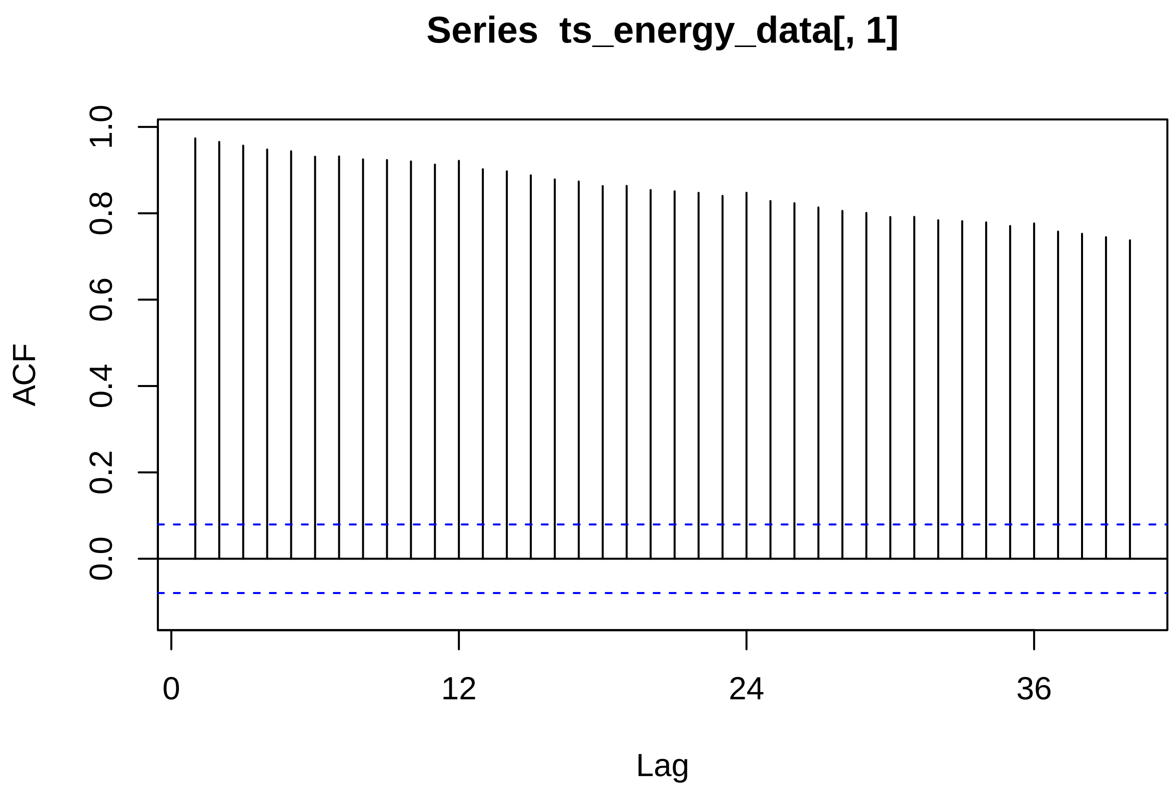
```
##
##  Pearson's product-moment correlation
##
## data:  ts_energy_data[, 2] and ts_energy_data[, 3]
## t = -0.043574, df = 607, p-value = 0.9653
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08120750  0.07769257
## sample estimates:
##          cor
## -0.001768629
```

```
#Biomass Energy is significantly correlated with both Renewable Energy and
#Hydroelectric Energy given the p-values less than 0.05. Renewable Energy and
#Hydroelectric Energy, however, are not significantly correlated given p-value
#greater than 0.05. Note that these are spatial correlations versus time
#correlations in Question 6 and Question 7.
```

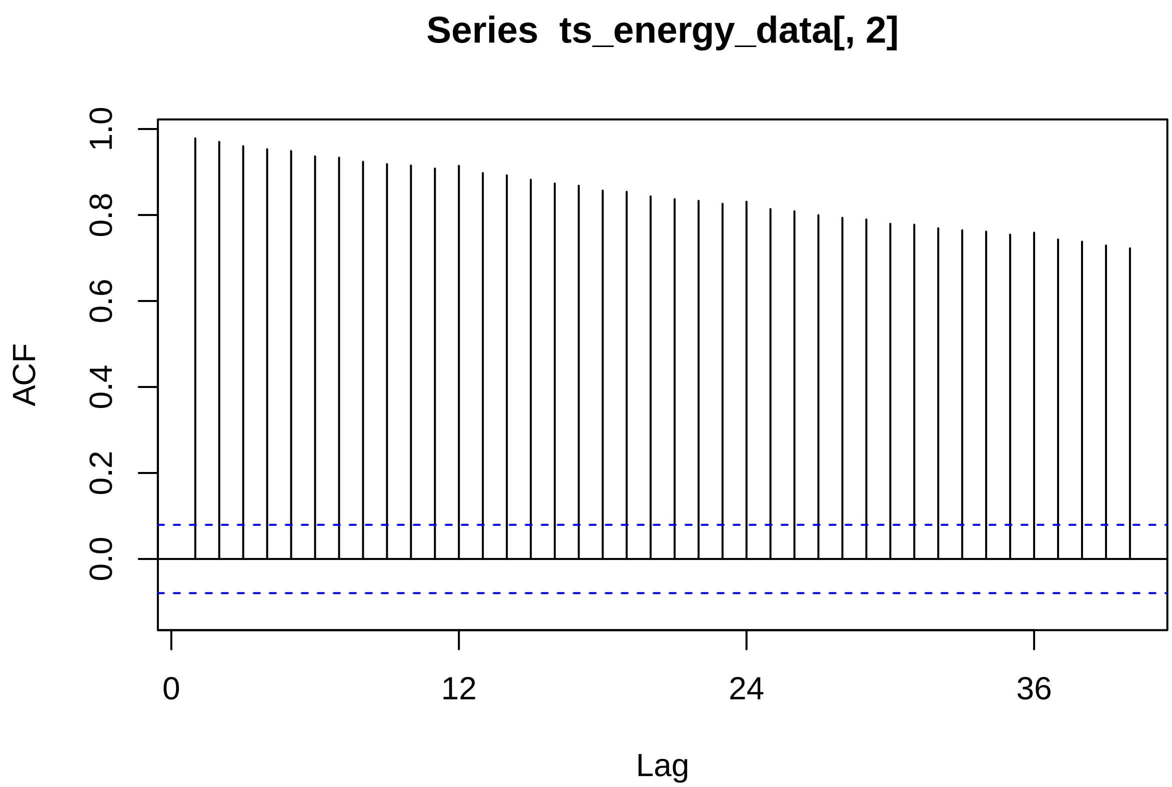
Question 6

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?

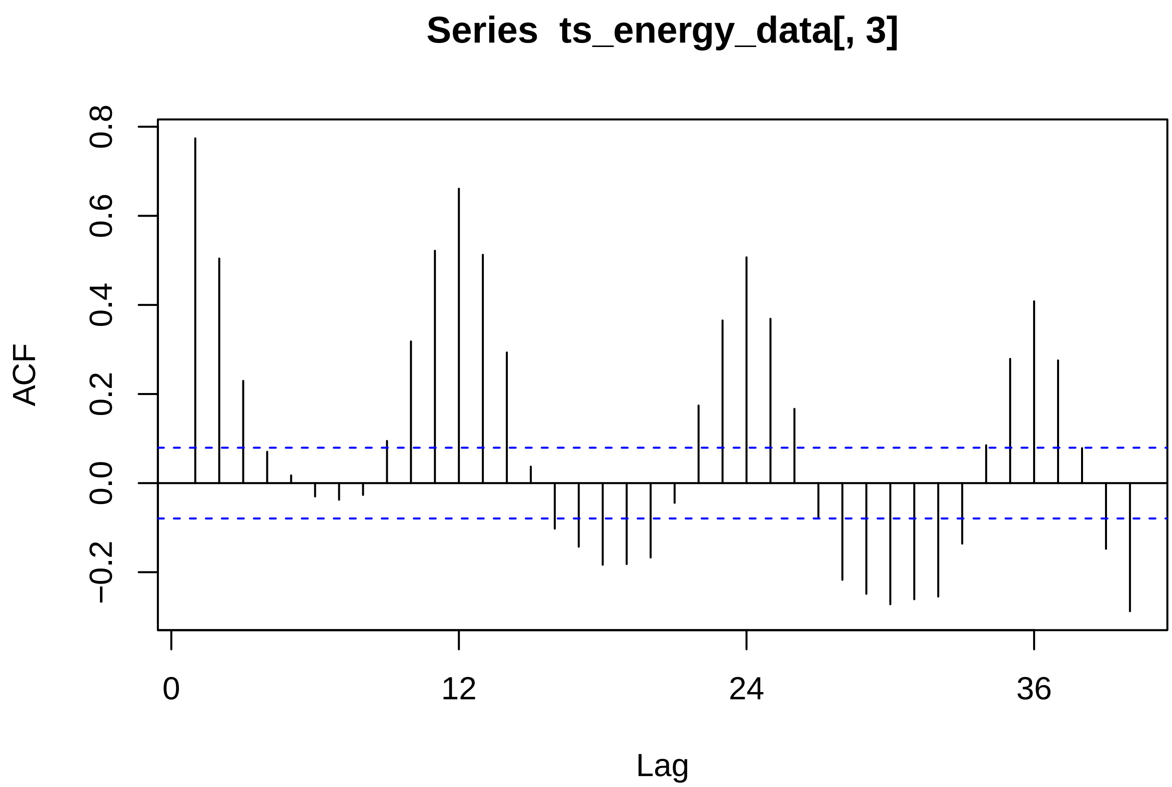
```
biomass_acf = Acf(ts_energy_data[,1],lag.max=40,type="correlation",
                  plot=TRUE)
```




```
renewable_acf = Acf(ts_energy_data[,2],lag.max=40,type="correlation",
                    plot=TRUE)
```




```
hydroelectric_acf = Acf(ts_energy_data[,3],lag.max=40,type="correlation",
                        plot=TRUE)
```

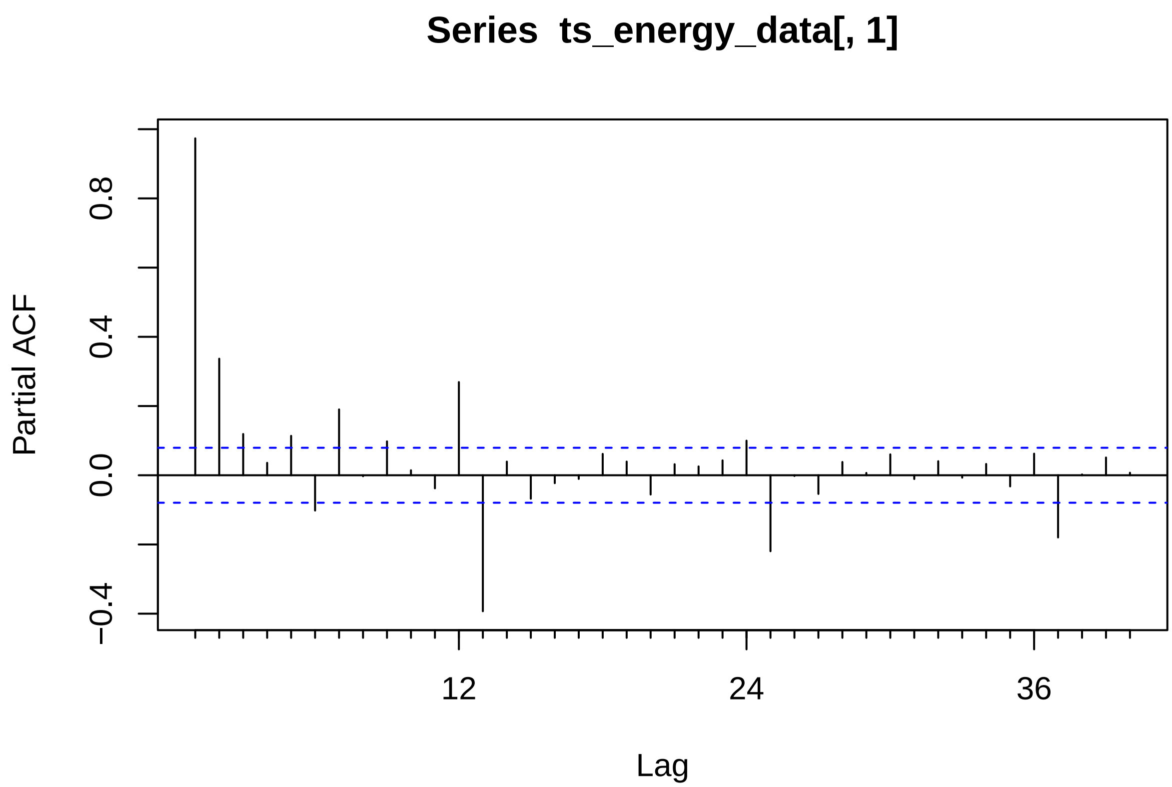



```
#Biomass Energy and Renewable Energy have similar behavior / patterns.
#Hydroelectric, however, has a different behavior and displays as a sinusoidal
#pattern.
```

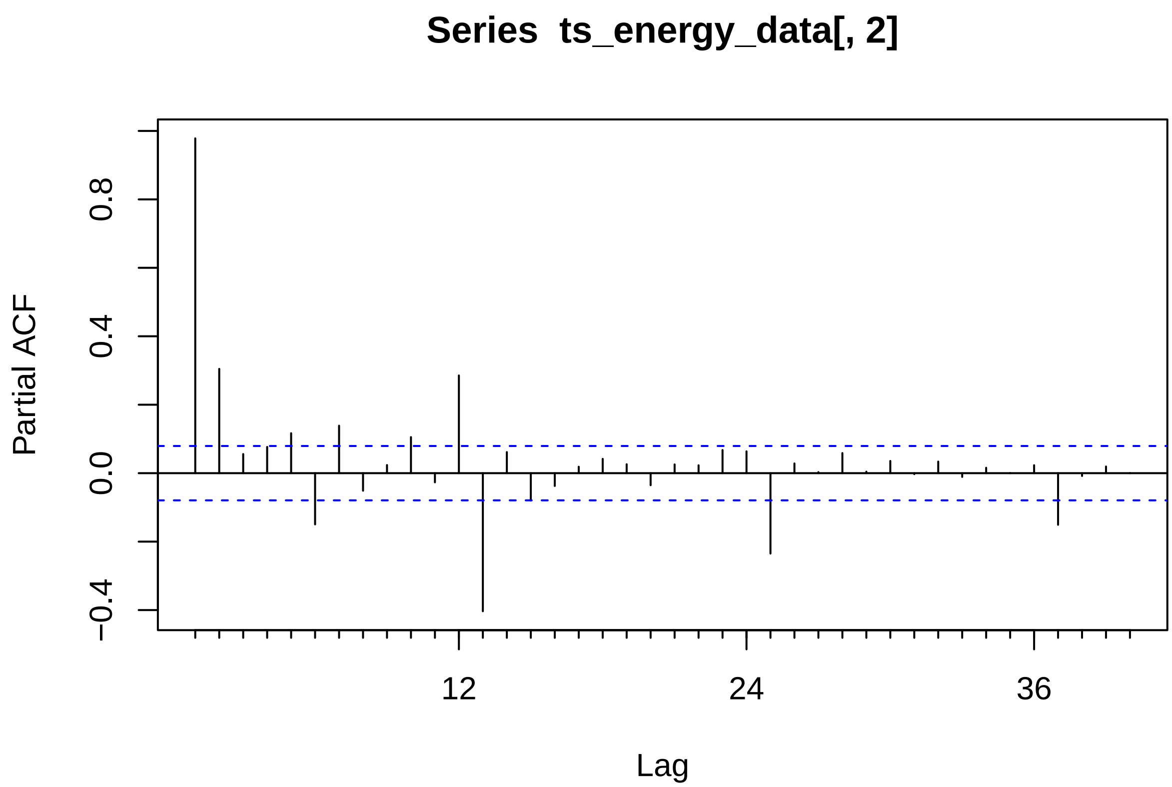
Question 7

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

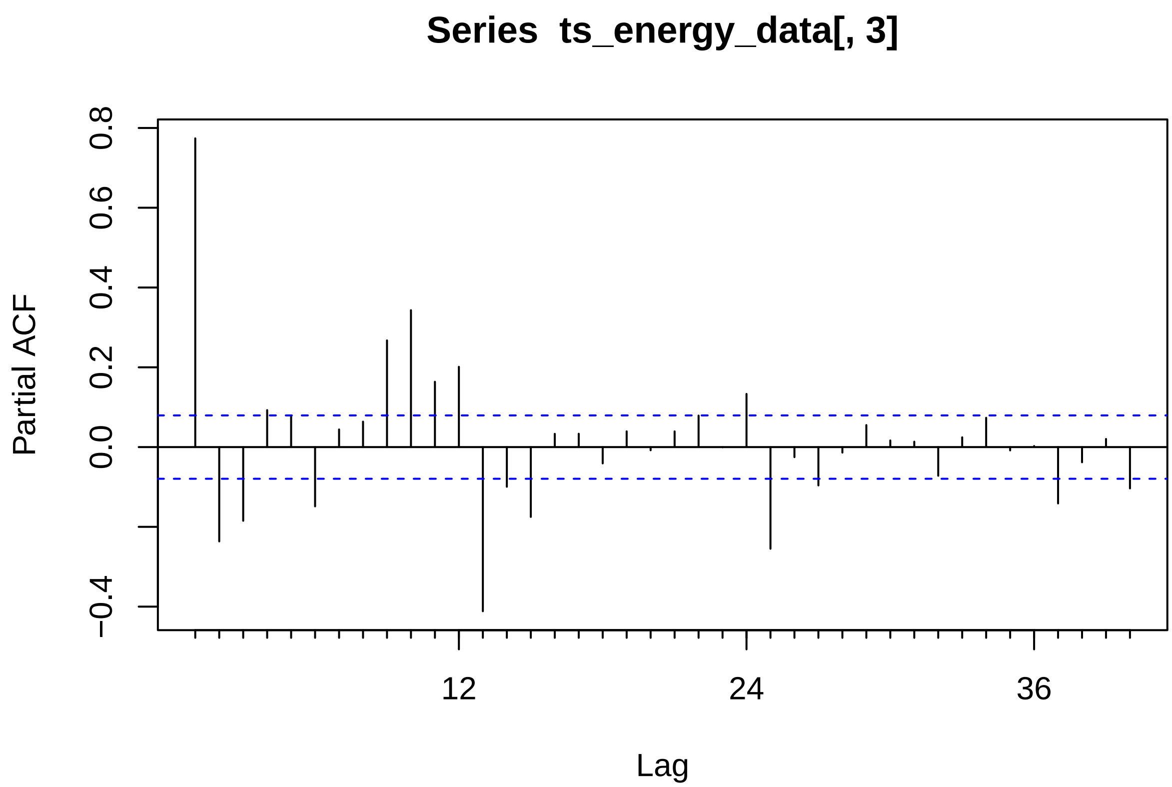
```
biomass_pacf = Pacf(ts_energy_data[,1],lag.max=40,plot=TRUE)
```




```
renewable_pacf = Pacf(ts_energy_data[,2],lag.max=40,plot=TRUE)
```




```
hydroelectric_pacf = Pacf(ts_energy_data[,3],lag.max=40,plot=TRUE)
```




```
#These plots differ from those in Q6 because they remove the effect of the
#influence of intermediate variables between lag 1 and lag 40.
```