ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 2 - Due date 01/26/22

Yu Hai

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 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. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp22.Rmd"). 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.\

```
library(forecast)
```

```
## Warning:   'forecast' R 4.1.2
```

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

```
library(tseries)
```

```
## Warning:   'tseries' R 4.1.2
```

```
library(dplyr)
```

```
##
##    '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)
```

```
## Warning:   'readxl' R 4.1.2
```

```
library(knitr)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sourc
on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the January 2022 Monthly Energy Review. The spreadsheet is ready to be used. Use the command *read.table*() to import the data in R or *panda.read_excel*() in Python (note that you will need to import pandas package). }

```
getwd()
```

```
## [1] "C:/Users/lenovo/Desktop/Spring_2022/ENV790/ENV790_TimeSeriesAnalysis_Sp2022"
```

```
#Importing data set
raw_RE_data<-read_excel(path="C:/Users/lenovo/Desktop/Spring_2022/ENV790/ENV790_TimeSeriesAnalysis_Sp20
raw_RE_data
```

```
## # A tibble: 586 x 14
##    Month               `Wood Energy Prod~ `Biofuels Product~ `Total Biomass Ene~
##    <dttm>              <chr>              <chr>              <chr>
##  1 NA                  (Trillion Btu)     (Trillion Btu)     (Trillion Btu)
##  2 1973-01-01 00:00:00 129.63             Not Available      129.787
##  3 1973-02-01 00:00:00 117.194            Not Available      117.338
##  4 1973-03-01 00:00:00 129.763            Not Available      129.938
##  5 1973-04-01 00:00:00 125.462            Not Available      125.636
##  6 1973-05-01 00:00:00 129.624            Not Available      129.834
##  7 1973-06-01 00:00:00 125.435            Not Available      125.611
##  8 1973-07-01 00:00:00 129.616            Not Available      129.787
##  9 1973-08-01 00:00:00 129.734            Not Available      129.918
## 10 1973-09-01 00:00:00 125.603            Not Available      125.782
## # ... with 576 more rows, and 10 more variables:
## #   Total Renewable Energy Production <chr>,
## #   Hydroelectric Power Consumption <chr>, Geothermal Energy Consumption <chr>,
## #   Solar Energy Consumption <chr>, Wind Energy Consumption <chr>,
## #   Wood Energy Consumption <chr>, Waste Energy Consumption <chr>,
## #   Biofuels Consumption <chr>, Total Biomass Energy Consumption <chr>,
## #   Total Renewable Energy Consumption <chr>
```

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.

```
sub_RE_data <- raw_RE_data[-c(1),4:6]
head(sub_RE_data)
```

```
## # A tibble: 6 x 3
##   `Total Biomass Energy Production` `Total Renewable Ener~ `Hydroelectric Power~
##   <chr>                             <chr>                  <chr>
## 1 129.787                           403.981                272.703
## 2 117.338                           360.9                  242.199
## 3 129.938                           400.161                268.81
## 4 125.636                           380.47                 253.185
## 5 129.834                           392.141                260.77
## 6 125.611                           377.232                249.859
```

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().

```
RE_data <- cbind(raw_RE_data[-c(1),1],sub_RE_data[,])
ts_RE_data <- ts(RE_data[,2:4], start=c(1973, 1), end=c(2021, 09), frequency=12)
head(ts_RE_data)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Jan 1973                              23                                73
## Feb 1973                               2                                38
## Mar 1973                              27                                68
## Apr 1973                               9                                52
## May 1973                              25                                57
## Jun 1973                               8                                47
##          Hydroelectric Power Consumption
## Jan 1973                             460
## Feb 1973                             334
## Mar 1973                             449
## Apr 1973                             383
## May 1973                             419
## Jun 1973                             362
```

Question 3

Compute mean and standard deviation for these three series.

```
RE_data$`Total Biomass Energy Production`<-as.numeric(RE_data$`Total Biomass Energy
↪ Production`)
RE_data$`Total Renewable Energy Production`<-as.numeric(RE_data$`Total Renewable Energy
↪ Production`)
```

```r
RE_data$`Hydroelectric Power Consumption`<-as.numeric(RE_data$`Hydroelectric Power
  Consumption`)
mean(RE_data$`Total Biomass Energy Production`)
```

```
## [1] 273.7839
```

```r
mean(RE_data$`Total Renewable Energy Production`)
```

```
## [1] 581.1708
```

```r
mean(RE_data$`Hydroelectric Power Consumption`)
```

```
## [1] 235.9653
```

```r
sd(RE_data$`Total Biomass Energy Production`)
```

```
## [1] 89.42852
```

```r
sd(RE_data$`Total Renewable Energy Production`)
```

```
## [1] 177.5607
```

```r
sd(RE_data$`Hydroelectric Power Consumption`)
```

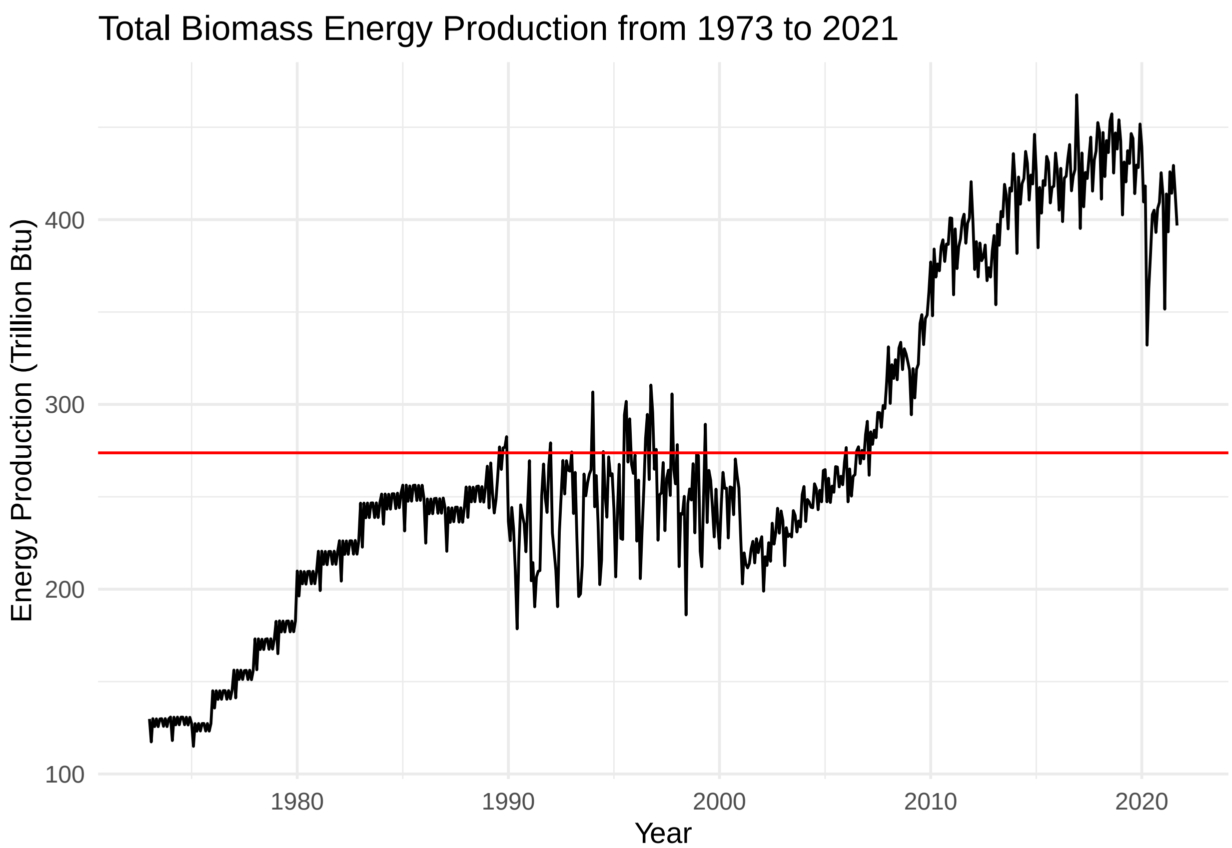
```
## [1] 44.01749
```

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.

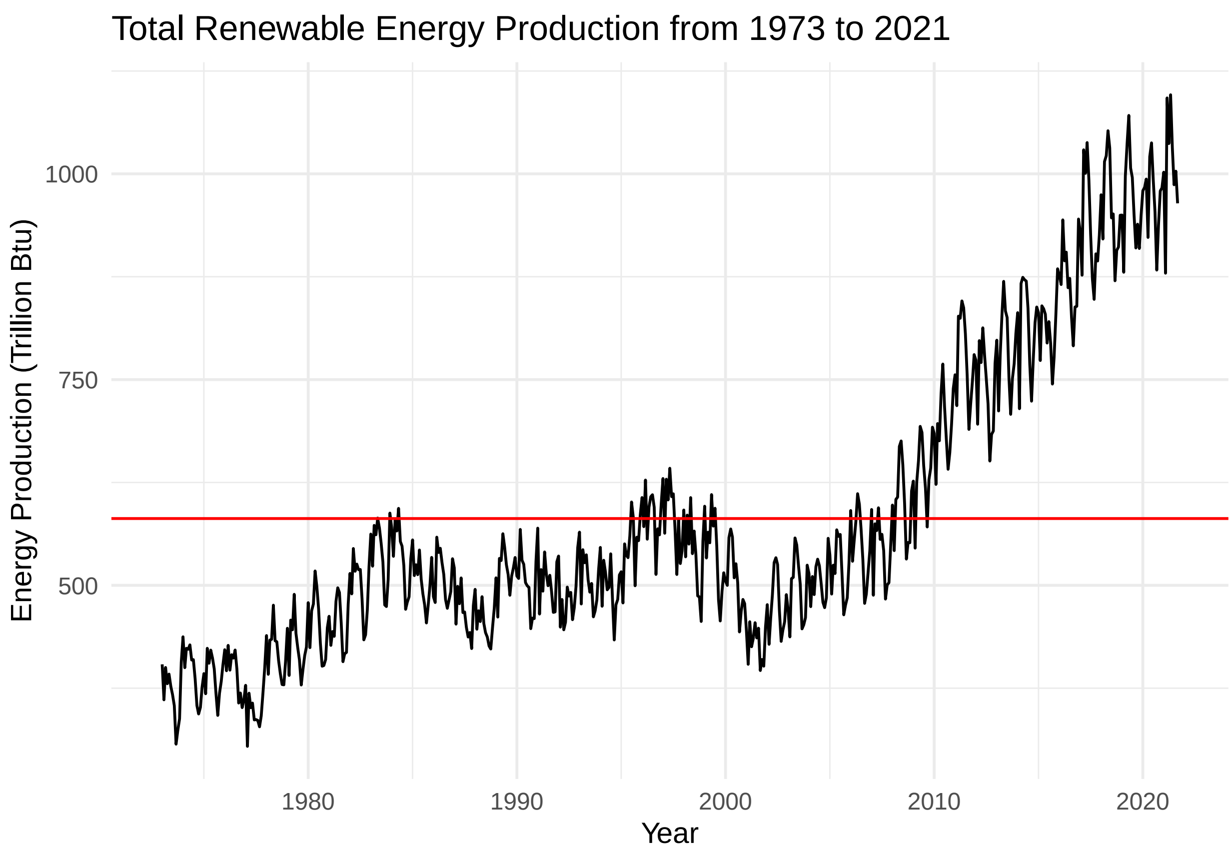
```r
library(ggplot2)
ggplot(RE_data, aes(x=Month)) +
            geom_line(aes(y=RE_data$`Total Biomass Energy Production`)) +xlab("Year") +
                ylab("Energy Production (Trillion Btu)") + labs(title="Total Biomass
                Energy Production from 1973 to 2021")+geom_hline(yintercept =
                mean(RE_data$`Total Biomass Energy Production`),
                color="red")+theme_minimal()
```

```
## Warning: Use of `RE_data$`Total Biomass Energy Production`` is discouraged. Use
## `Total Biomass Energy Production` instead.
```



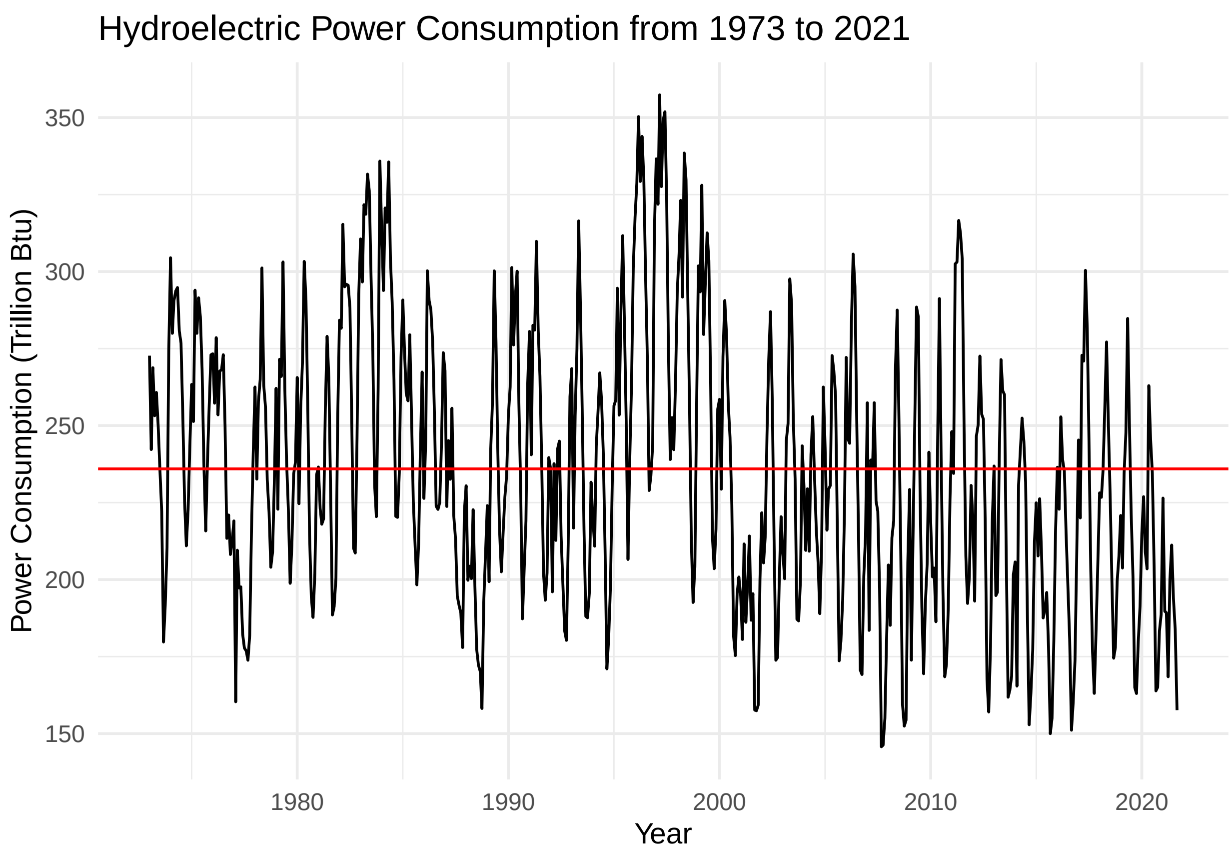

```
ggplot(RE_data, aes(x=Month)) +
            geom_line(aes(y=RE_data$`Total Renewable Energy Production`)) +xlab("Year") +
            ↪   ylab("Energy Production (Trillion Btu)")+labs(title="Total Renewable
            ↪   Energy Production from 1973 to 2021")+geom_hline(yintercept =
            ↪   mean(RE_data$`Total Renewable Energy Production`),
            ↪   color="red")+theme_minimal()
```

```
## Warning: Use of `RE_data$`Total Renewable Energy Production`` is discouraged.
## Use `Total Renewable Energy Production` instead.
```




```
ggplot(RE_data, aes(x=Month)) +
            geom_line(aes(y=RE_data$`Hydroelectric Power Consumption`)) +xlab("Year") +
             ↪   ylab("Power Consumption (Trillion Btu)")+labs(title="Hydroelectric Power
             ↪   Consumption from 1973 to 2021")+geom_hline(yintercept =
             ↪   mean(RE_data$`Hydroelectric Power Consumption`),
             ↪   color="red")+theme_minimal()
```

```
## Warning: Use of `RE_data$`Hydroelectric Power Consumption`` is discouraged. Use
## `Hydroelectric Power Consumption` instead.
```



Total biomass energy production shows a overall increasing trend. The most significant increase is from 2000 to 2010, while between 1990 and 2000 the increasing trend is weak but the variation within each year is large.

Total renewable energy production shows an overall increasing trend. This trend is not so obvious before 2000 and becomes significant since 2000. The seasonality is significant and relatively constant from 1973 to 2021.

The hydroelectric power consumption doesn't show a clear increasing or decreasing trend over the years of observation, though there is always a great fluctuation between observations within each year.

Question 5

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

```
cor(RE_data[,2:4])
```

```
##                                   Total Biomass Energy Production
## Total Biomass Energy Production                         1.0000000
## Total Renewable Energy Production                       0.9232838
## Hydroelectric Power Consumption                        -0.2804997
##                                   Total Renewable Energy Production
## Total Biomass Energy Production                          0.92328377
## Total Renewable Energy Production                        1.00000000
## Hydroelectric Power Consumption                         -0.05680651
##                                   Hydroelectric Power Consumption
```

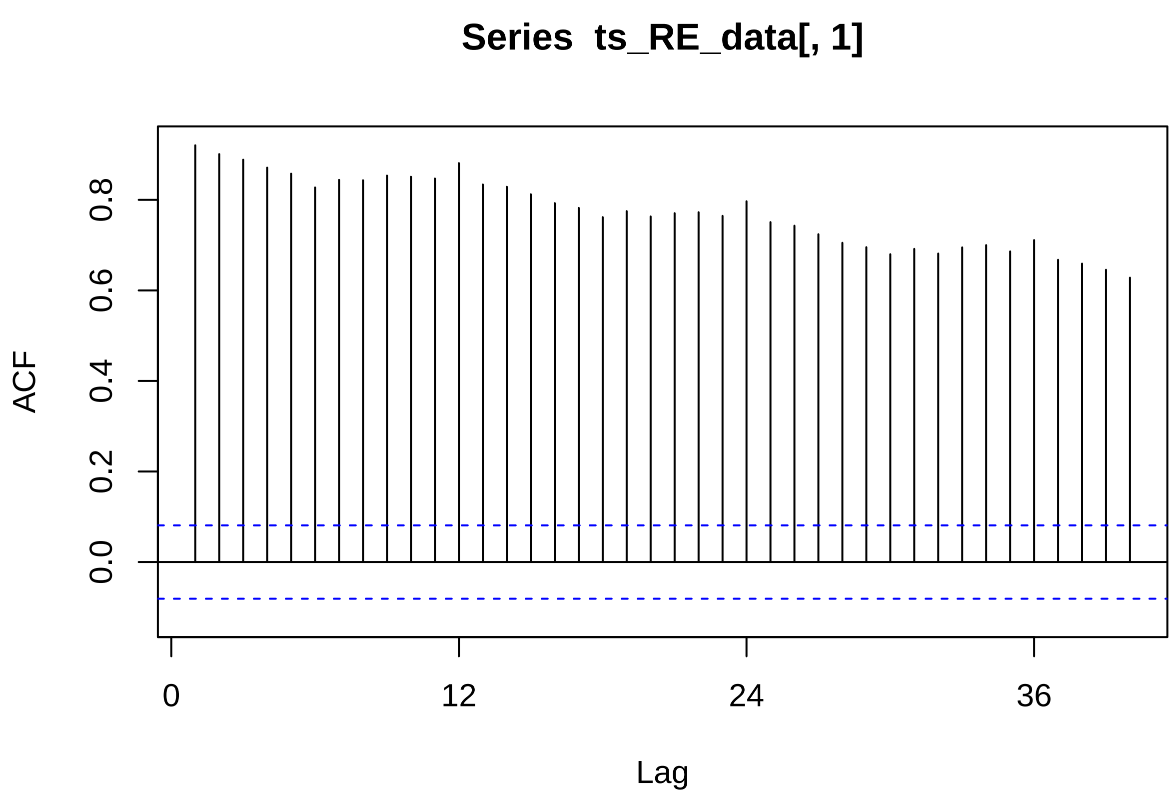
```
## Total Biomass Energy Production                          -0.28049970
## Total Renewable Energy Production                        -0.05680651
## Hydroelectric Power Consumption                           1.00000000
```

According to the correlation coefficients, total biomass energy production shows a strong positive correlation with renewable energy production (0.92), while it shows a weak negative correlation with hydroelectric power consumption(-0.28). The renewable energy production also shows a weak negative correlation with hydroelectric power consumption(-0.0057).

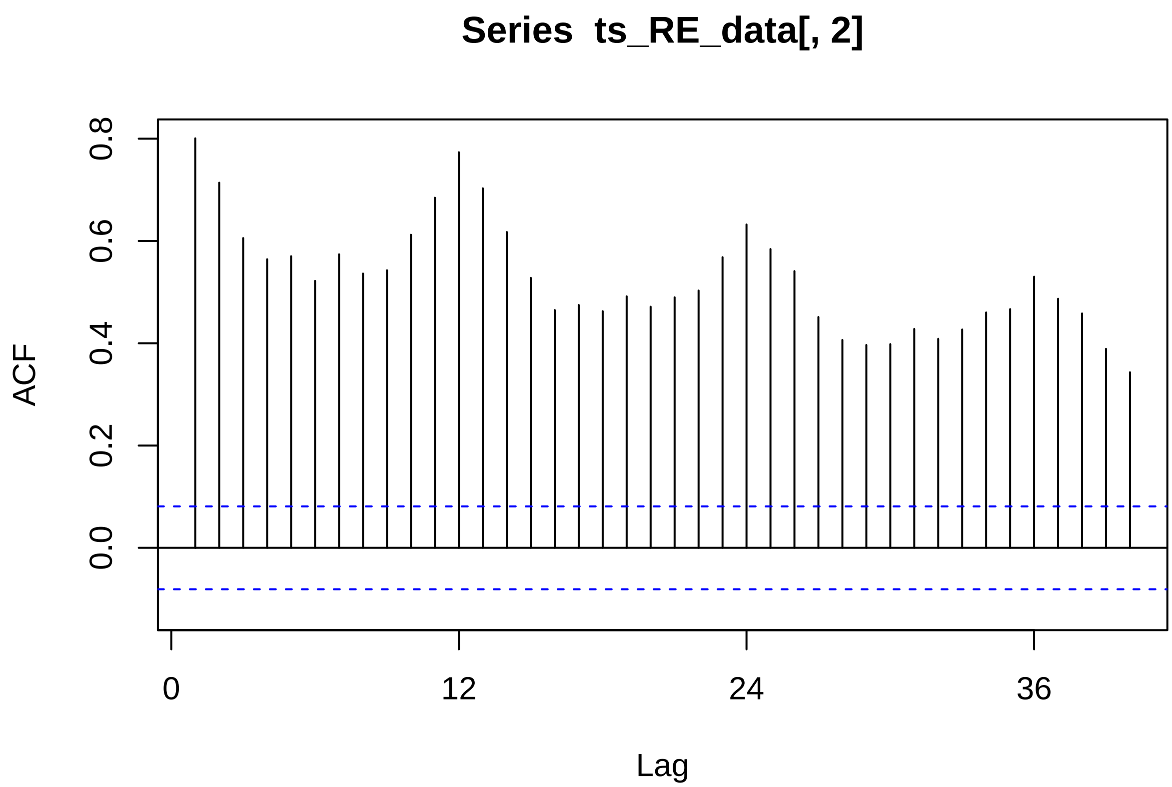
Question 6

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?

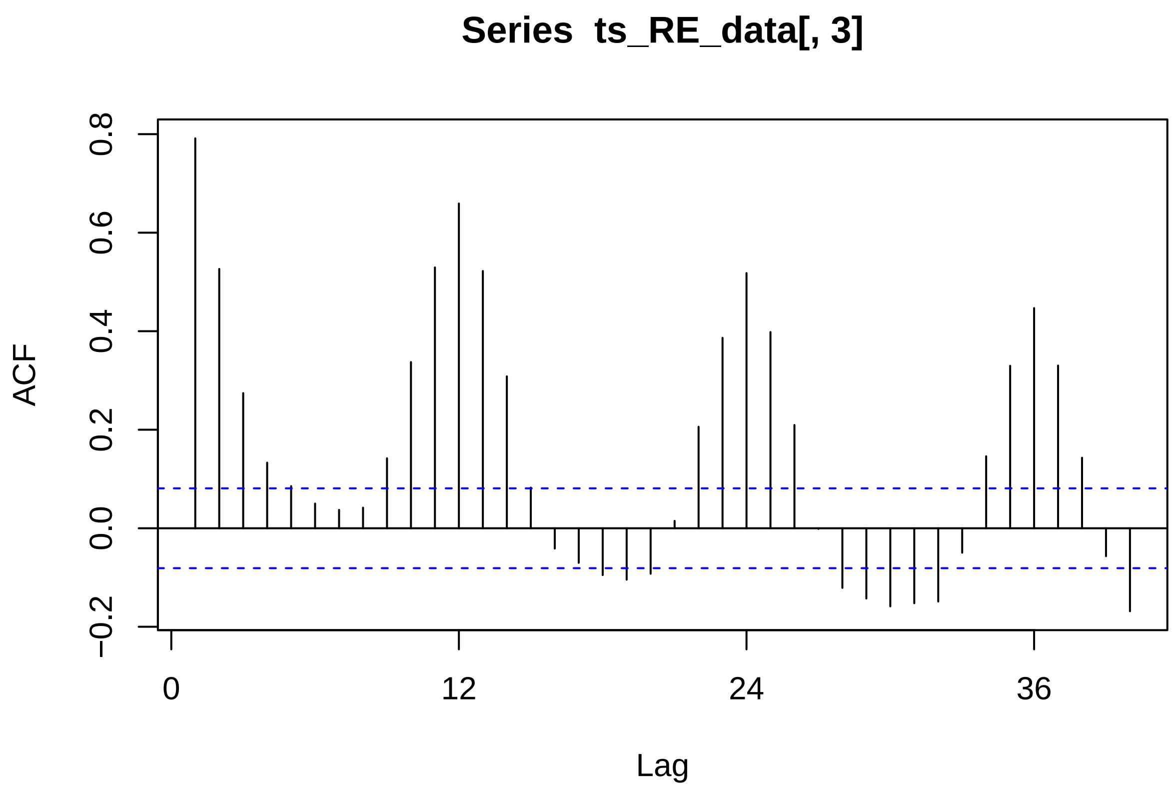
```
Biomass_acf=Acf(ts_RE_data[,1],lag.max=40, type="correlation", plot=TRUE)
```




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




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



Biomass; ACF at all lags from 1 to 40 are positive (i.e. the correlation between Y1 and Y2,Y2…,Y40 are all positive), and there is a weak seasonality observed.

Renewable: Similar as the graph for Biomass energy production, ACFs at all lags are positive, and there is a stronger seasonality observed.

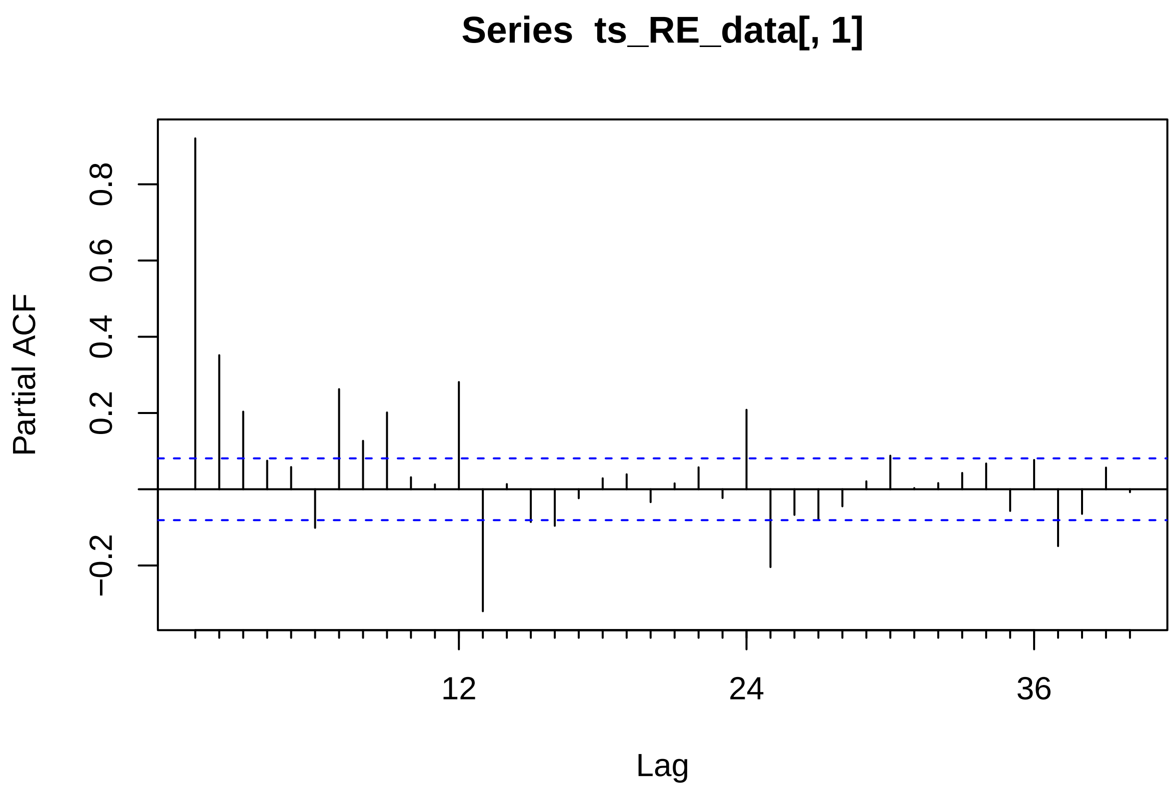
Hydroelectric: There are both positive and negative ACFs and there is a strong seasonality observed.

The three ACF graphs show different behaviors, but the common thing is that the absolute values of ACFs become smaller as the lag time goes up, and this is because the autocorrelation is weaker between the variables that are further away in time.

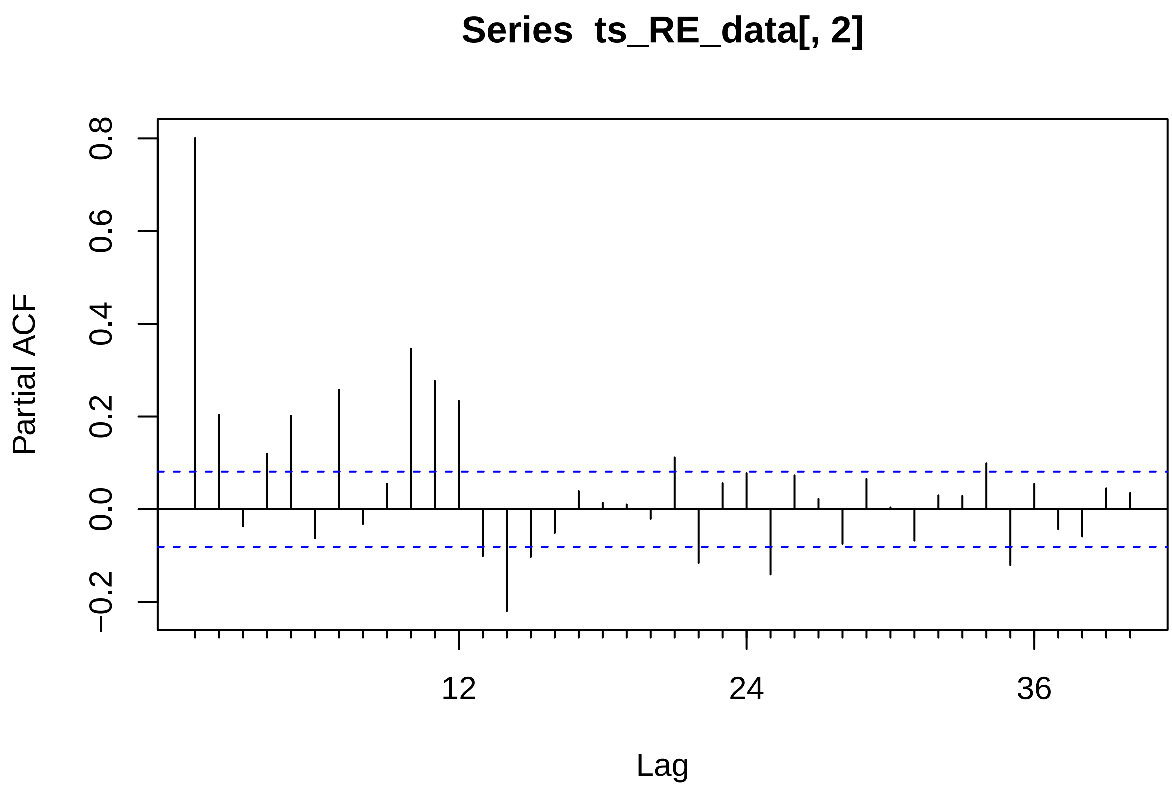
Question 7

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

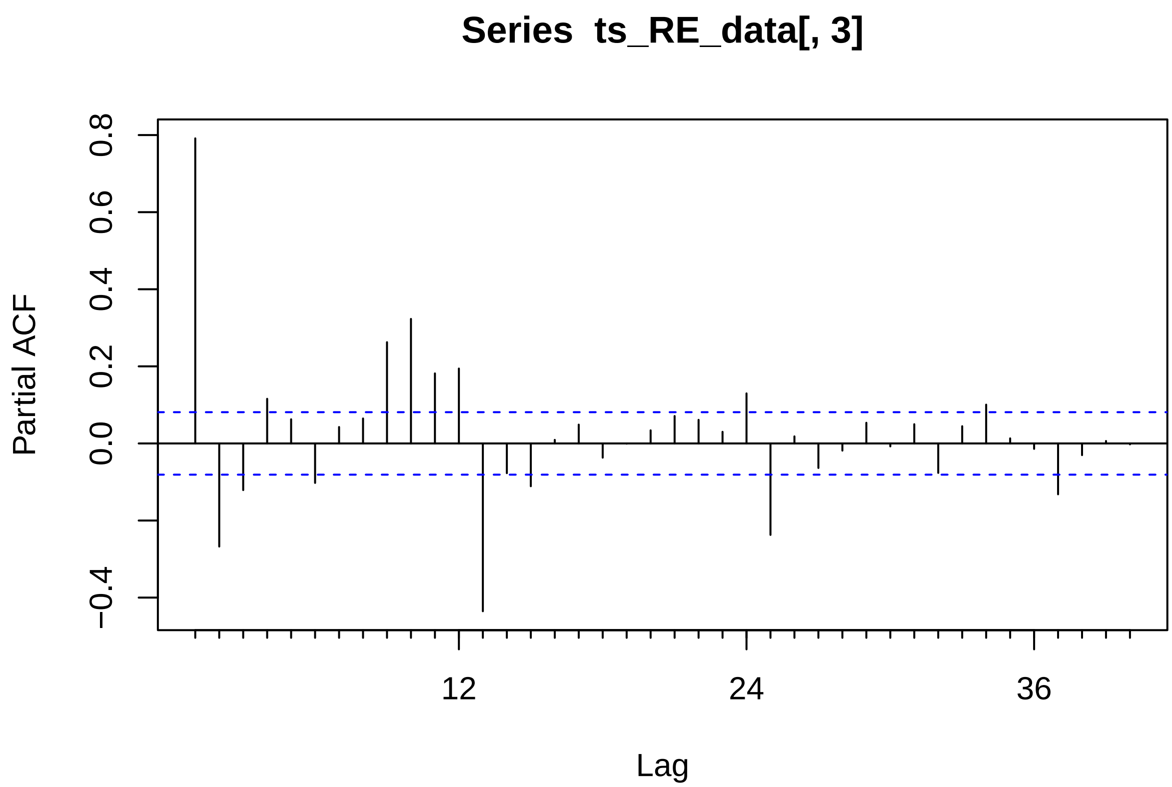
```
Biomass_pacf=Pacf(ts_RE_data[,1],lag.max=40, plot=TRUE)
```




```
Renewable_pacf=Pacf(ts_RE_data[,2],lag.max=40, plot=TRUE)
```




```
Hydro_pacf=Pacf(ts_RE_data[,3],lag.max=40, plot=TRUE)
```



The values of PACF are smaller than ACFs (except the lag of 1) because the calculation of PACF removes the influence of all these intermediate variables and only leaves the directly correlation between Yt and Yt-h. In all three graphs, there are positive ACFs at some lags with corresponding negative PACFs.