ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 2 - Due date 02/03/23

Zhengqi Jiao

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_Sp23.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)
```

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

```
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
```

```
library(tseries)
```

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 2022 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*.

```
#Importing data set
energy_data <- read.csv(file="/Users/christine/Documents/TimeSeriesAnalysis/TimeSeriesAnalysis_Jiao/Data
```

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.

```
energy_data1 <- energy_data[,4:6]
head(energy_data1)
```

```
##   Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1                         129.787                           403.981
## 2                         117.338                           360.900
## 3                         129.938                           400.161
## 4                         125.636                           380.470
## 5                         129.834                           392.141
## 6                         125.611                           377.232
##   Hydroelectric.Power.Consumption
## 1                         272.703
## 2                         242.199
## 3                         268.810
## 4                         253.185
## 5                         260.770
## 6                         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().

```
ts_energy_data1 <- ts(energy_data1,frequency=12,start=c(1973,1))
head(ts_energy_data1,20)
```

```
##          Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## Jan 1973                         129.787                           403.981
## Feb 1973                         117.338                           360.900
## Mar 1973                         129.938                           400.161
## Apr 1973                         125.636                           380.470
## May 1973                         129.834                           392.141
## Jun 1973                         125.611                           377.232
## Jul 1973                         129.787                           367.325
## Aug 1973                         129.918                           353.757
## Sep 1973                         125.782                           307.006
## Oct 1973                         129.970                           323.453
## Nov 1973                         125.643                           337.817
## Dec 1973                         129.824                           406.694
## Jan 1974                         130.807                           437.467
## Feb 1974                         118.091                           399.942
```

```
## Mar 1974                        130.727                            423.474
## Apr 1974                        126.583                            422.323
## May 1974                        130.789                            427.657
## Jun 1974                        126.611                            409.281
## Jul 1974                        130.756                            409.719
## Aug 1974                        130.763                            386.101
##          Hydroelectric.Power.Consumption
## Jan 1973                        272.703
## Feb 1973                        242.199
## Mar 1973                        268.810
## Apr 1973                        253.185
## May 1973                        260.770
## Jun 1973                        249.859
## Jul 1973                        235.670
## Aug 1973                        222.077
## Sep 1973                        179.733
## Oct 1973                        191.723
## Nov 1973                        210.285
## Dec 1973                        274.435
## Jan 1974                        304.506
## Feb 1974                        279.950
## Mar 1974                        290.582
## Apr 1974                        293.702
## May 1974                        294.828
## Jun 1974                        280.695
## Jul 1974                        276.772
## Aug 1974                        253.175
```

Question 3

Compute mean and standard deviation for these three series.

```
Biomass_mean <- mean(ts_energy_data1[,"Total.Biomass.Energy.Production"])
Biomass_mean
```

```
## [1] 277.2525
```

```
Biomass_sd  <- sd(ts_energy_data1[,"Total.Biomass.Energy.Production"])
Biomass_sd
```

```
## [1] 91.75367
```

```
Reweable_mean <- mean(ts_energy_data1[,"Total.Renewable.Energy.Production"])
Reweable_mean
```

```
## [1] 592.1583
```

```
Reweable_sd <- sd(ts_energy_data1[,"Total.Renewable.Energy.Production"])
Reweable_sd
```

```
## [1] 191.7978
```

```
Hydroelectric_mean <- mean(ts_energy_data1[,"Hydroelectric.Power.Consumption"])
Hydroelectric_mean
```

```
## [1] 235.1146
```

```
Hydroelectric_sd <- sd(ts_energy_data1[,"Hydroelectric.Power.Consumption"])
Hydroelectric_sd
```

```
## [1] 44.16116
```

The mean and the standard deviation of the Total.Biomass.Energy.Production is 277.2525226 and 91.7536727, respectively. The mean and the standard deviation of the Total.Renewable.Energy.Production is 592.1582948 and 191.7978345, respectively. The mean and the standard deviation of the Hydroelectric.Power.Consumption is 235.1146499 and 44.161163, respectively.

Question 4

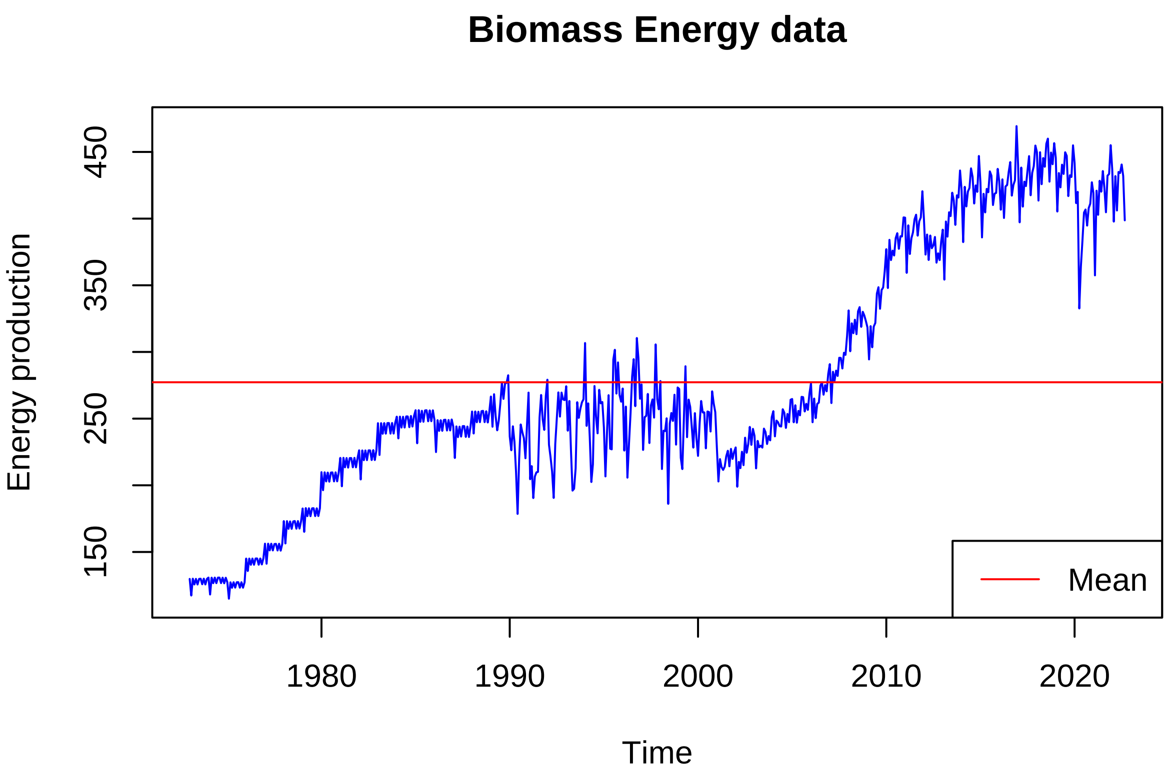
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
library(lubridate)
```

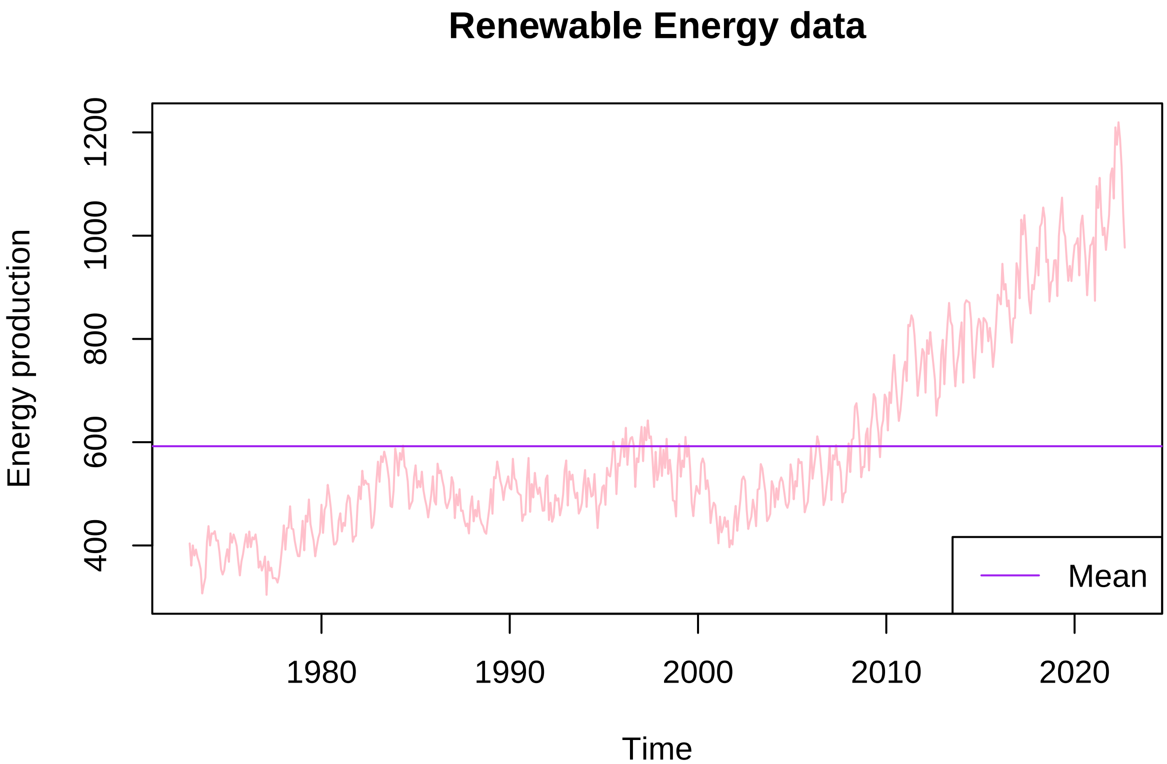
```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

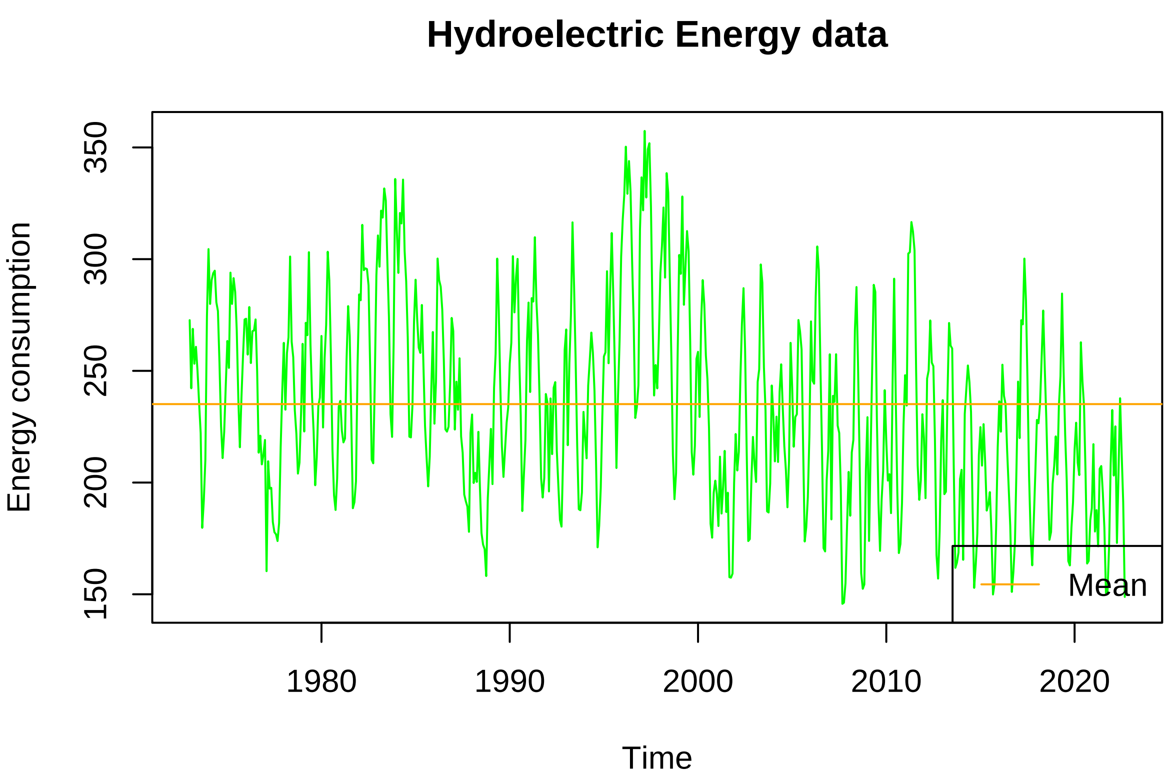
```
plot(ts_energy_data1[,"Total.Biomass.Energy.Production"],type="l",col="blue",ylab="Energy production",ma
abline(h=mean(ts_energy_data1[,"Total.Biomass.Energy.Production"]),col="red")
legend("bottomright", legend="Mean",col=c("red"), lty=1)
```




```
plot(ts_energy_data1[,"Total.Renewable.Energy.Production"],type="l",col="pink",ylab="Energy production"
abline(h=mean(ts_energy_data1[,"Total.Renewable.Energy.Production"]),col="purple")
legend("bottomright", legend="Mean",col=c("purple"), lty=1)
```




```
plot(ts_energy_data1[,"Hydroelectric.Power.Consumption"],type="l",col="green",ylab="Energy consumption"
abline(h=mean(ts_energy_data1[,"Hydroelectric.Power.Consumption"]),col="orange")
legend("bottomright", legend="Mean",col=c("orange"), lty=1)
```



Question 5

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

```
corr <- cor(ts_energy_data1, use = "everything", method = c("spearman"))
corr
```

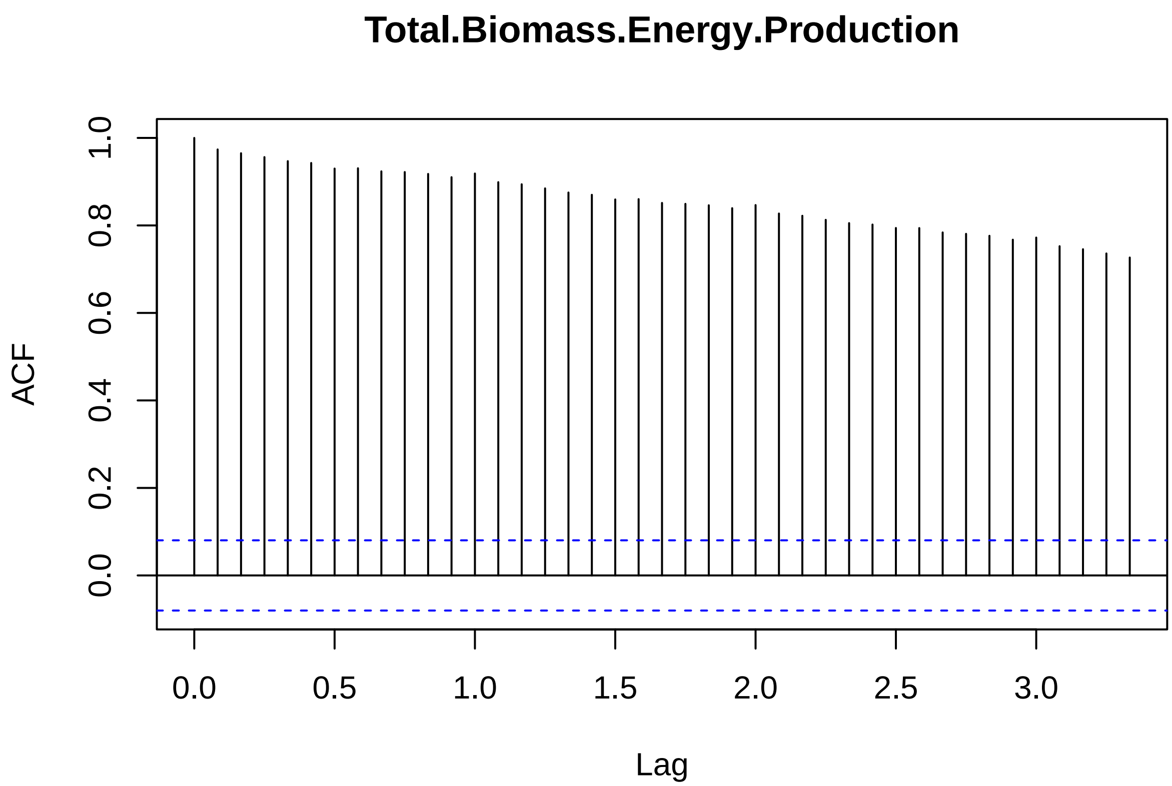
```
##                                    Total.Biomass.Energy.Production
## Total.Biomass.Energy.Production                          1.0000000
## Total.Renewable.Energy.Production                        0.8868431
## Hydroelectric.Power.Consumption                         -0.2902982
##                                    Total.Renewable.Energy.Production
## Total.Biomass.Energy.Production                           0.88684308
## Total.Renewable.Energy.Production                         1.00000000
## Hydroelectric.Power.Consumption                           0.05020665
##                                    Hydroelectric.Power.Consumption
## Total.Biomass.Energy.Production                        -0.29029824
## Total.Renewable.Energy.Production                       0.05020665
## Hydroelectric.Power.Consumption                         1.00000000
```

Since the data relationship is not linear, so use the Spearman's correlation instead of Pearson.The range of correlation is between +1 to -1. Closer to +1 means a stronger positive correlation. Closer to -1 means a stronger negative correlation. When the correlation is closer to 0, it means there is no trend. In this question, Total.Biomass.Energy.Production has a strong positive correlation with Total.Renewable.Energy.Production because the value is 0.8868431. Total.Renewable.Energy.Production has a strong negative correlation with Hydroelectric.Power.Consumption because the value is 0.0502066. Hydroelectric.Power.Consumption has a weak negative correlation with Total.Biomass.Energy.Production because the value is -0.2902982.

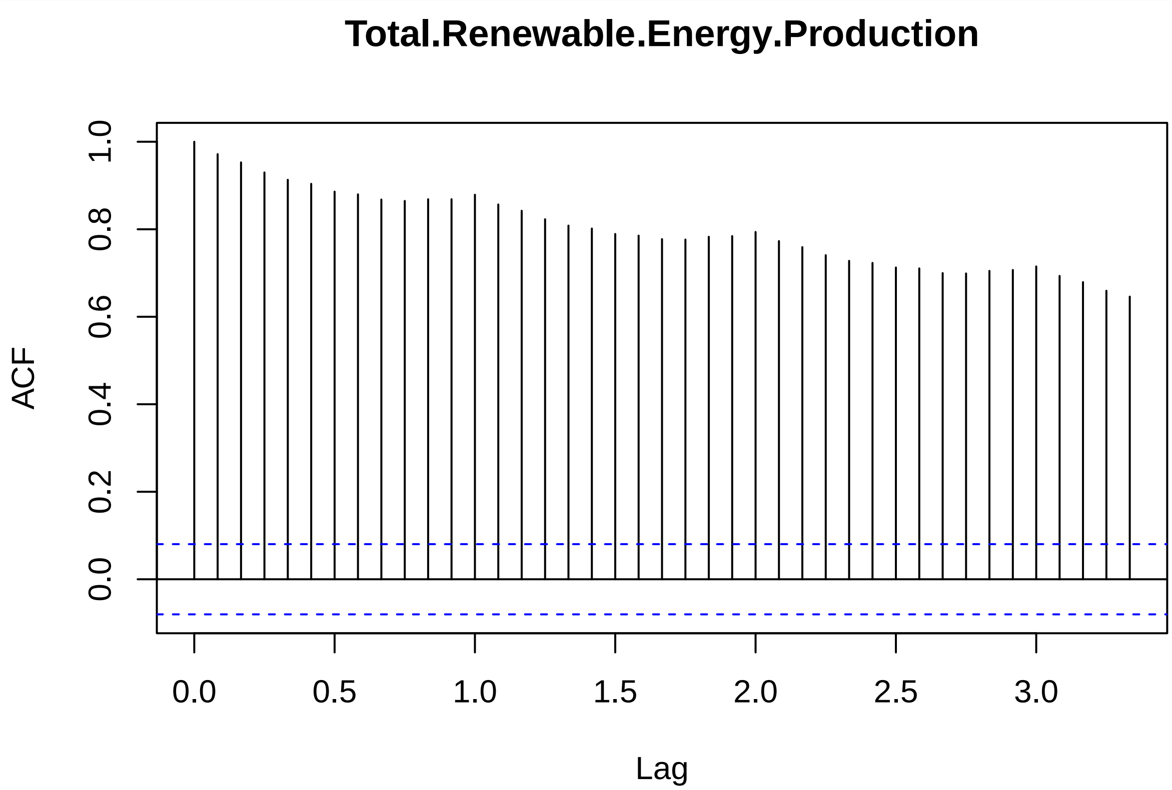
Question 6

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?

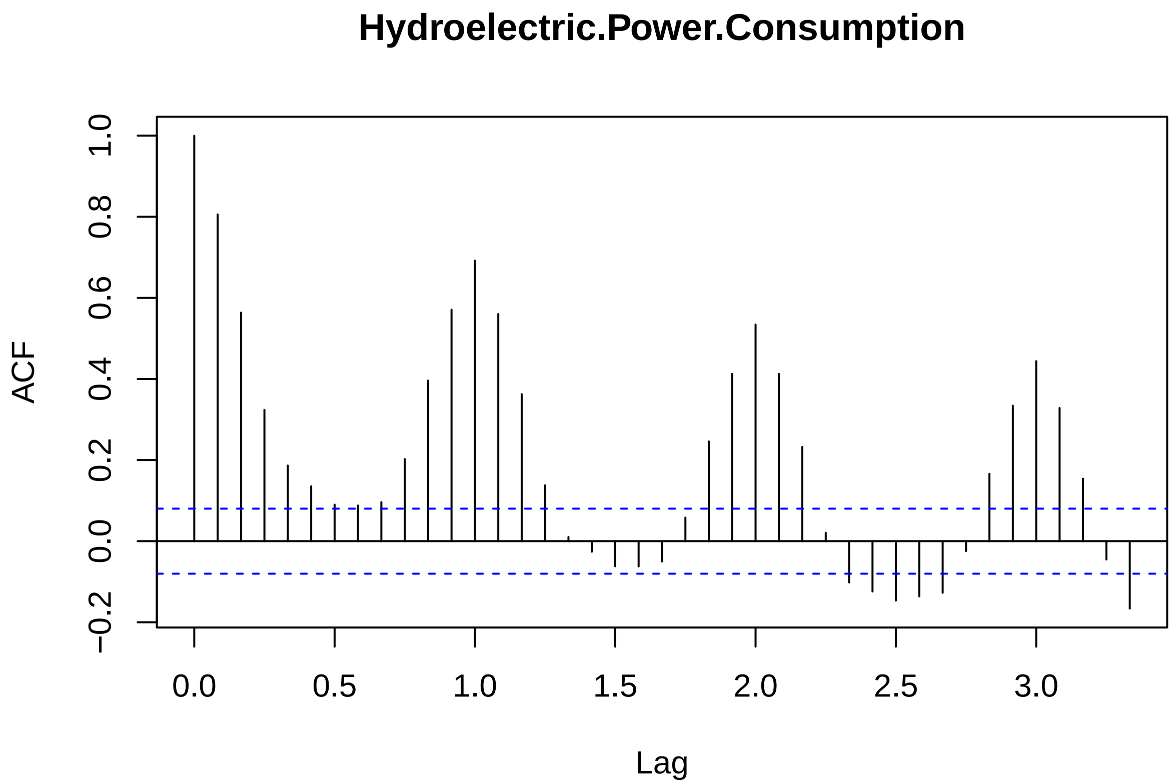
```
acf(ts_energy_data1[,1],lag.max = 40, main = "Total.Biomass.Energy.Production")
```




```
acf(ts_energy_data1[,2],lag.max = 40, main = "Total.Renewable.Energy.Production")
```




```
acf(ts_energy_data1[,3],lag.max = 40, main = "Hydroelectric.Power.Consumption")
```

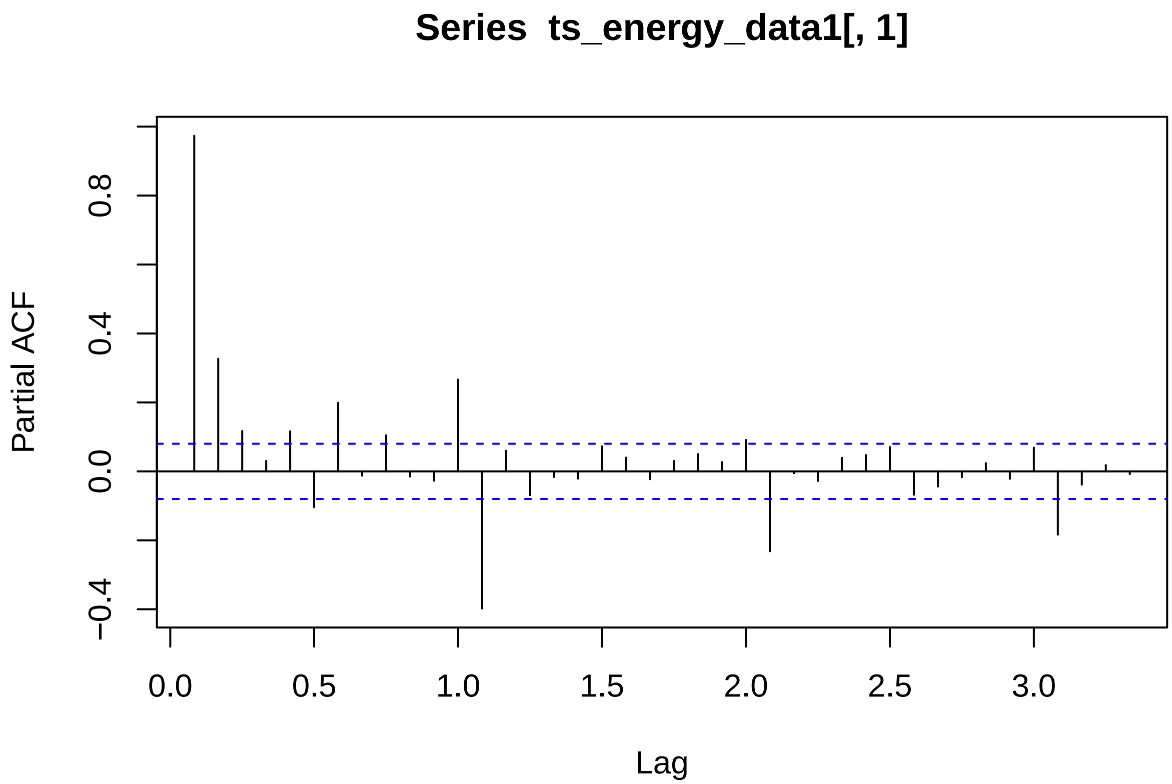


In the first pot,for lags up to 3.4,values are statistically significant. This represents that two adjacent values of the Total.Biomass.Energy.Production are highly correlated. In the second plot, for lags up to 3.4,values are statistically significant This represents that two adjacent values of the Total.Renewable.Energy.Production are highly correlated. In the third plot, the autocorrelation plot for Hydroelectric.Power.Consumption shows that the most spikes are outside the dotted line area, which means they are statistically significant. However, there are some spikes inside the dotted line area, which means they are not statistically significant. This represents most of the two adjacent values of that the Hydroelectric.Power.Consumption are highly correlated, but some are not.

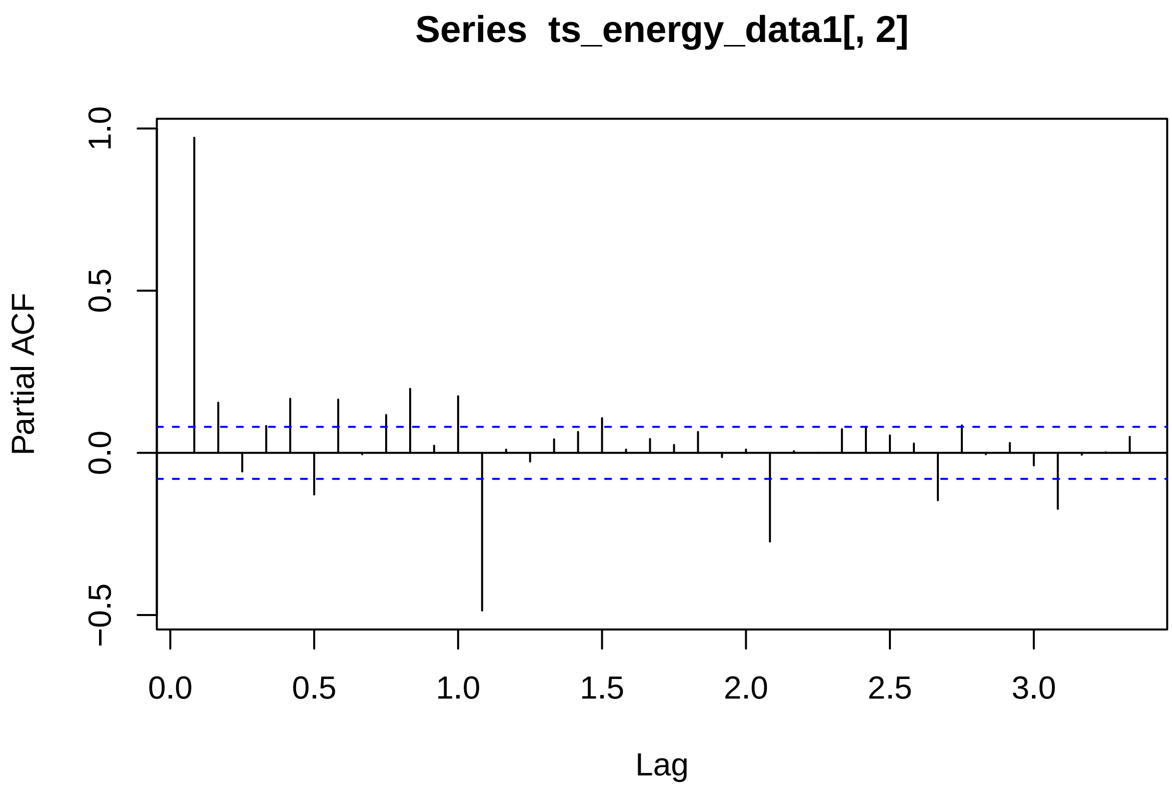
Question 7

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

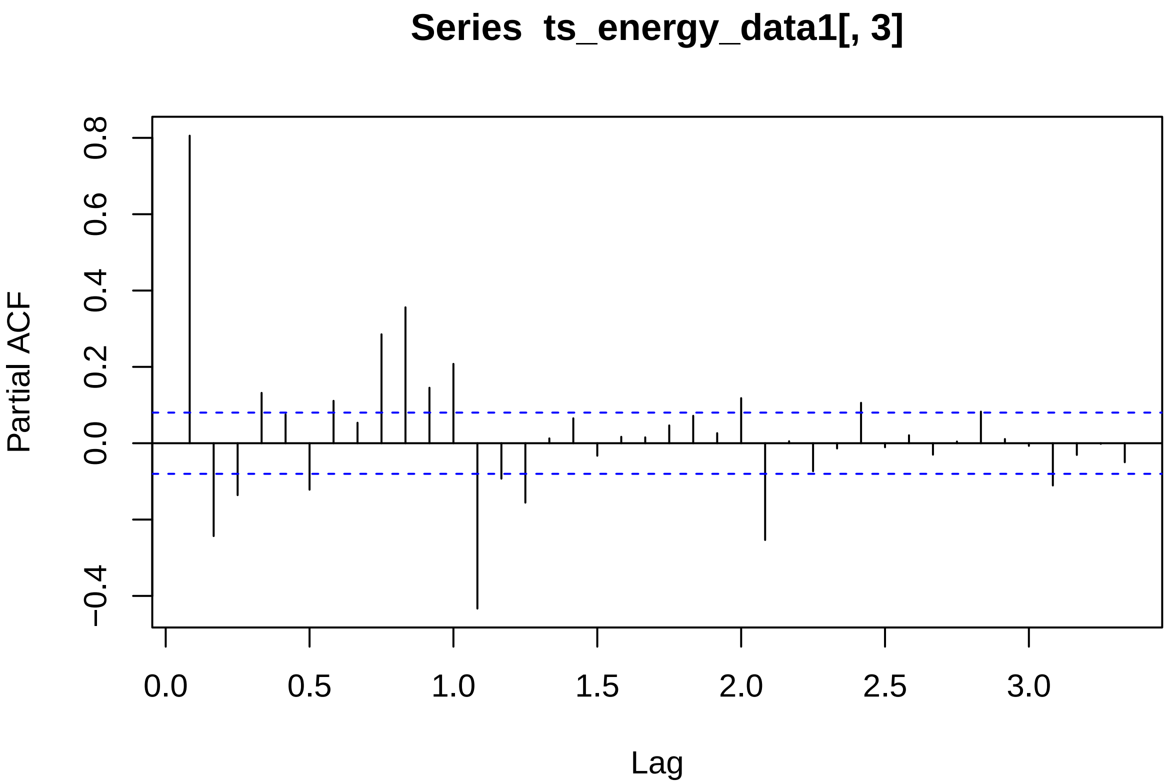
```
pacf(ts_energy_data1[,1],lag.max = 40)
```




```
pacf(ts_energy_data1[,2],lag.max = 40)
```




```
pacf(ts_energy_data1[,3],lag.max = 40)
```



PACF correlation is always smaller than ACF. PACF is about the directly correlation by removing all intermediate variables. Same with the ACF in Question6, spikes outside the dotted line area are statistically significant. Spikes inside the dotted line area are not statistically significant. Otherwise, ACF and PACF have similar ways for deciding correlations.