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

Assignment 2 - Due date 02/05/21

Keyang Xue

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_Sp21.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.\

```
#Load/install required package here
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
```

```
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 January 2021 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). }

```
#Importing data set
raw_energy <- read_excel(path='/Users/apple/Desktop/790 Time Series Analysis/ENV790_30_TSA_S2021/Data/Ta
```

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 <- cbind.data.frame(raw_energy[,1],raw_energy[,4:6])
head(energy)
```

```
##        Month Total Biomass Energy Production Total Renewable Energy Production
## 1       <NA>                   (Trillion Btu)                    (Trillion Btu)
## 2 1973-01-01                          129.787                           403.981
## 3 1973-02-01                          117.338                             360.9
## 4 1973-03-01                          129.938                           400.161
## 5 1973-04-01                          125.636                            380.47
## 6 1973-05-01                          129.834                           392.141
##   Hydroelectric Power Consumption
## 1                  (Trillion Btu)
## 2                         272.703
## 3                         242.199
## 4                          268.81
## 5                         253.185
## 6                          260.77
```

```
num_energy <- sapply(energy[2:575,2:4],as.numeric) %>% as.data.frame()
head(num_energy)
```

```
##   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 <- ts(num_energy)
head(ts_energy)
```

```
## Time Series:
## Start = 1
## End = 6
## Frequency = 1
##   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 3

Compute mean and standard deviation for these three series.

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

```
## [1] 270.6961
```

```
sd(ts_energy[,1])
```

```
## [1] 87.36311
```

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

```
## [1] 572.7321
```

```
sd(ts_energy[,2])
```

```
## [1] 168.4588
```

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

```
## [1] 236.9515
```

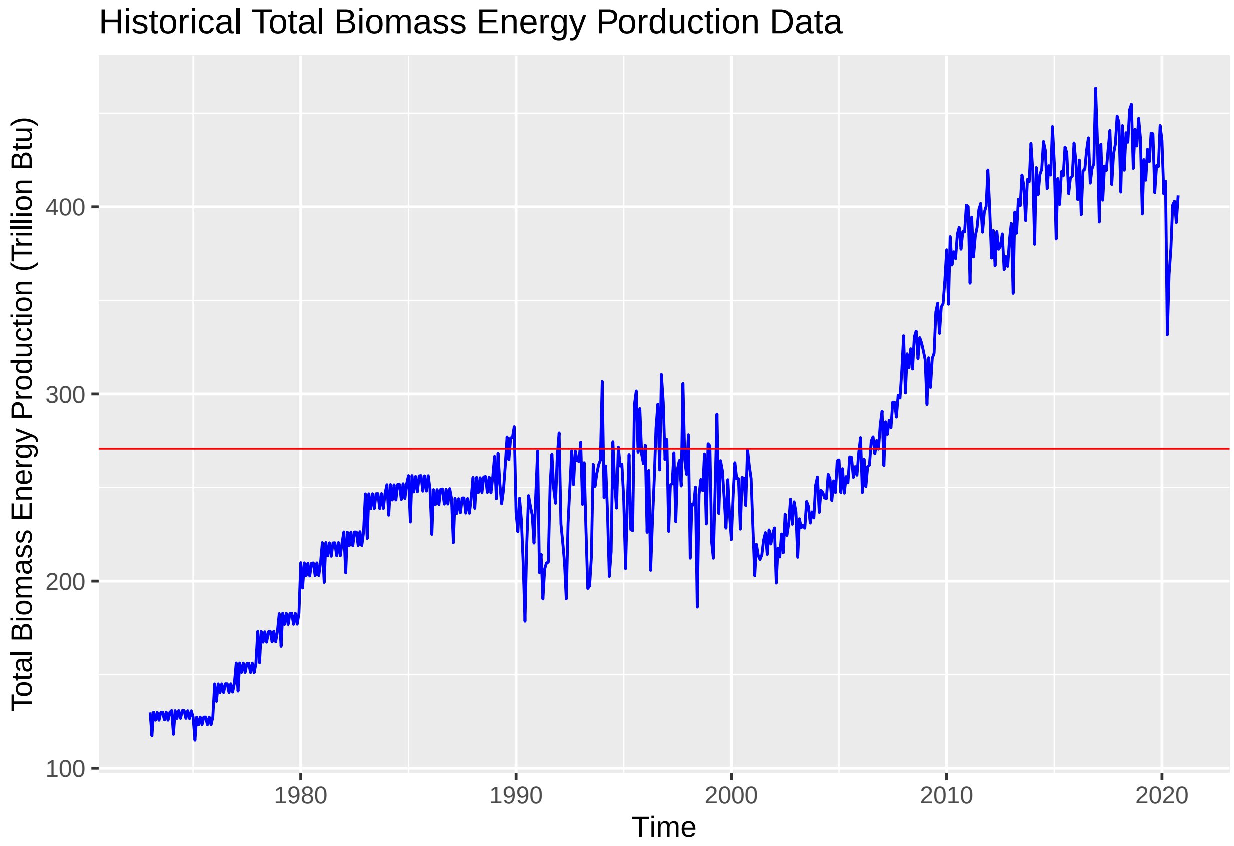
```
sd(ts_energy[,3])
```

```
## [1] 43.90392
```

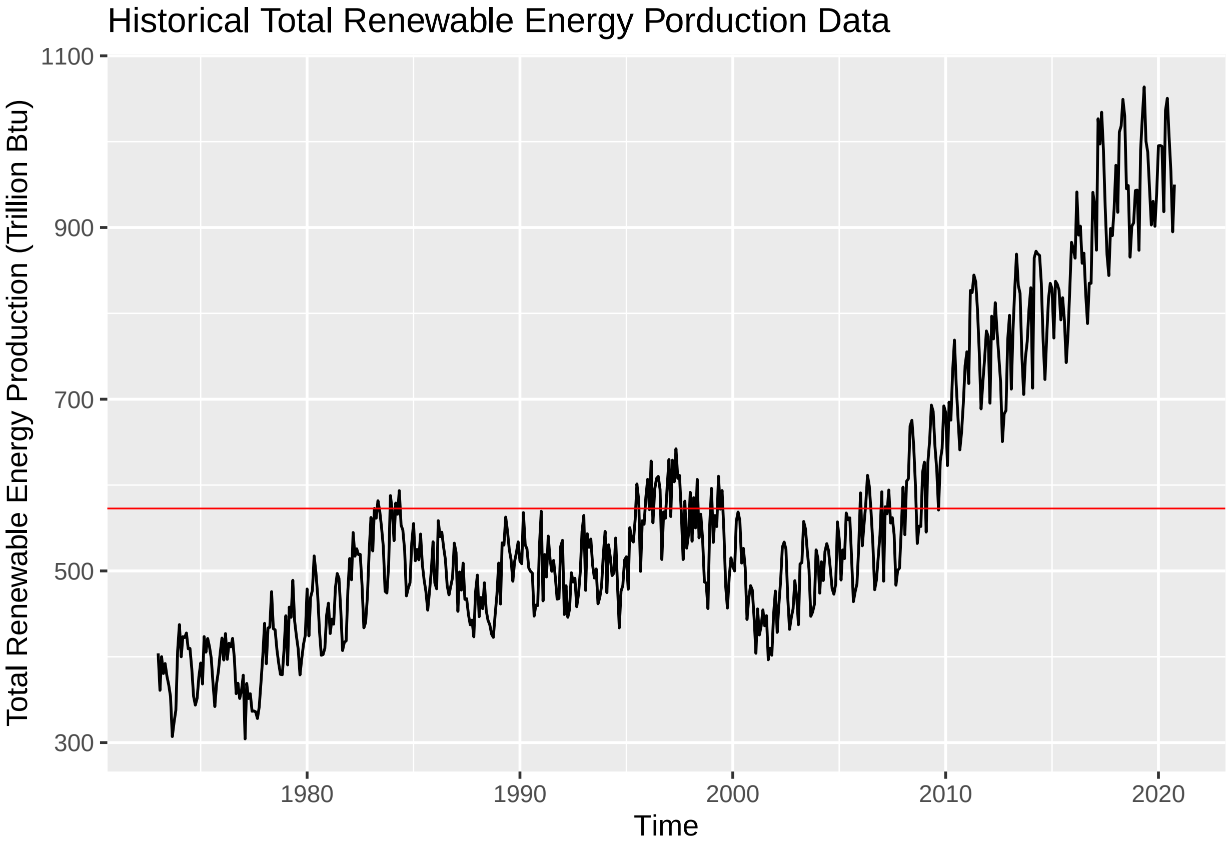
Question 4

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.

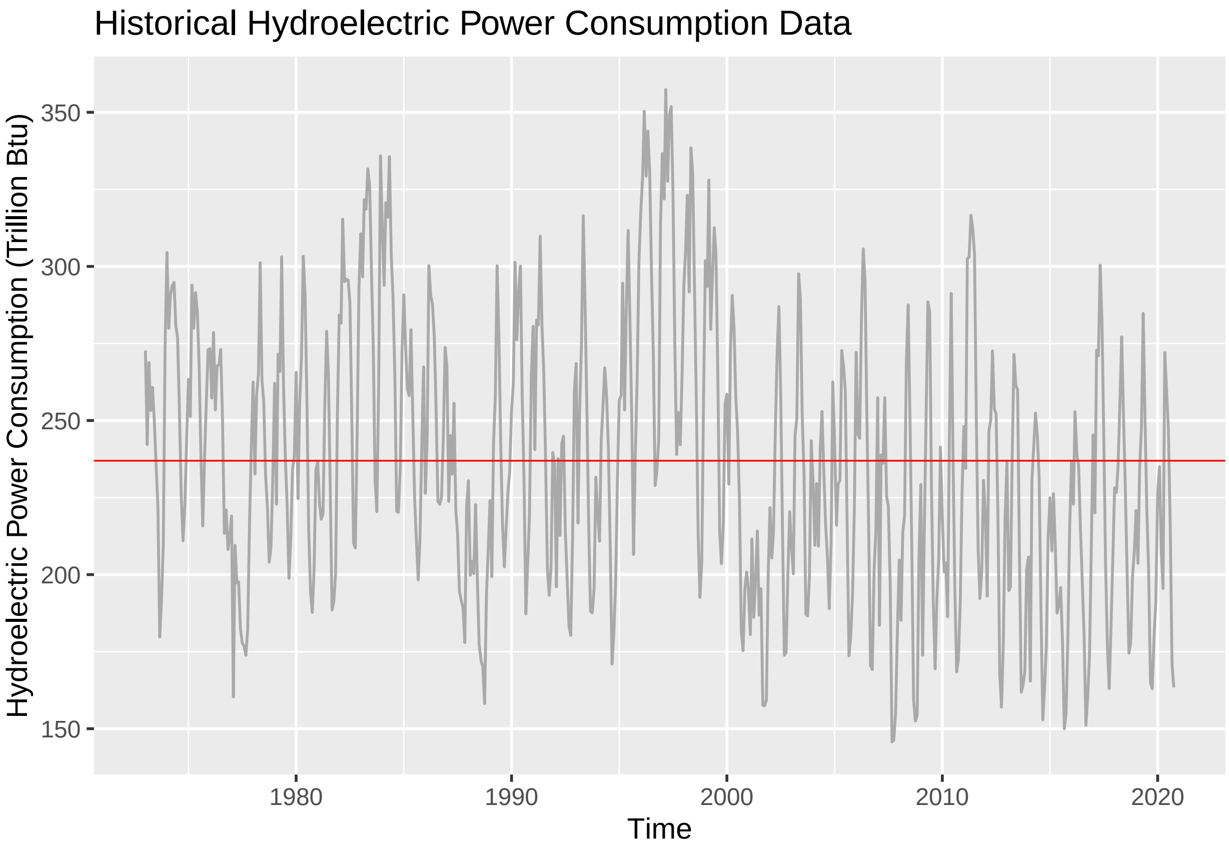
```
ggplot(num_energy, aes(x= energy$Month[2:575], y=`Total Biomass Energy Production`)) +
            geom_line(color="blue")+ xlab("Time") + ylab("Total Biomass Energy Production (Trillion Btu
  labs(title = 'Historical Total Biomass Energy Porduction Data') +
  geom_hline(yintercept=mean(ts_energy[,1]),col="red",lwd=0.3)
```



```
ggplot(num_energy, aes(x= energy$Month[2:575], y=`Total Renewable Energy Production`)) +
            geom_line(color="black")+ xlab("Time") + ylab("Total Renewable Energy Production (Trillion B
  labs(title = 'Historical Total Renewable Energy Porduction Data') +
  geom_hline(yintercept=mean(ts_energy[,2]),col="red",lwd=0.3)
```



```
ggplot(num_energy, aes(x= energy$Month[2:575], y=`Hydroelectric Power Consumption`)) +
            geom_line(color="dark gray")+ xlab("Time") + ylab("Hydroelectric Power Consumption (Trillion
  labs(title = 'Historical Hydroelectric Power Consumption Data') +
  geom_hline(yintercept=mean(ts_energy[,3]),col="red",lwd=0.3)
```



Total biomass energy production experienced a sharp increase before 1990. It stayed around 200 to 300 trillion Btu from 1990 to 2001 with fluctuations bigger than before. Since around 2001,it has shown an increasing trend again. This trend then stopped at 2020 followed by a sharp drop, but it then seemed to increase again. Total renewable energy production shows a similar trend, in which it shows an increasing trend before 1985, and it remained stagnant until about 2002. It has started to increase ever since then. Hydroelectric power consumption shows more fluctuations from time to time (probably due to seasonality), but its general trend is relatively more stable than both biomass and renewable energy production. Hydroelectric power consumption seemed to be larger from 1973 to 2000 compared to now, showing by smaller minimum and maximum values from 2000 to 2020.

Question 5

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

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

```
##
##  Pearson's product-moment correlation
##
## data:  ts_energy[, 1] and ts_energy[, 2]
## t = 57.562, df = 572, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9104276 0.9346626
## sample estimates:
##       cor
## 0.9234609
```

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

```
##
##  Pearson’s product-moment correlation
##
## data:  ts_energy[, 2] and ts_energy[, 3]
## t = -0.065935, df = 572, p-value = 0.9475
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08457627  0.07909949
## sample estimates:
##          cor
## -0.002756852
```

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

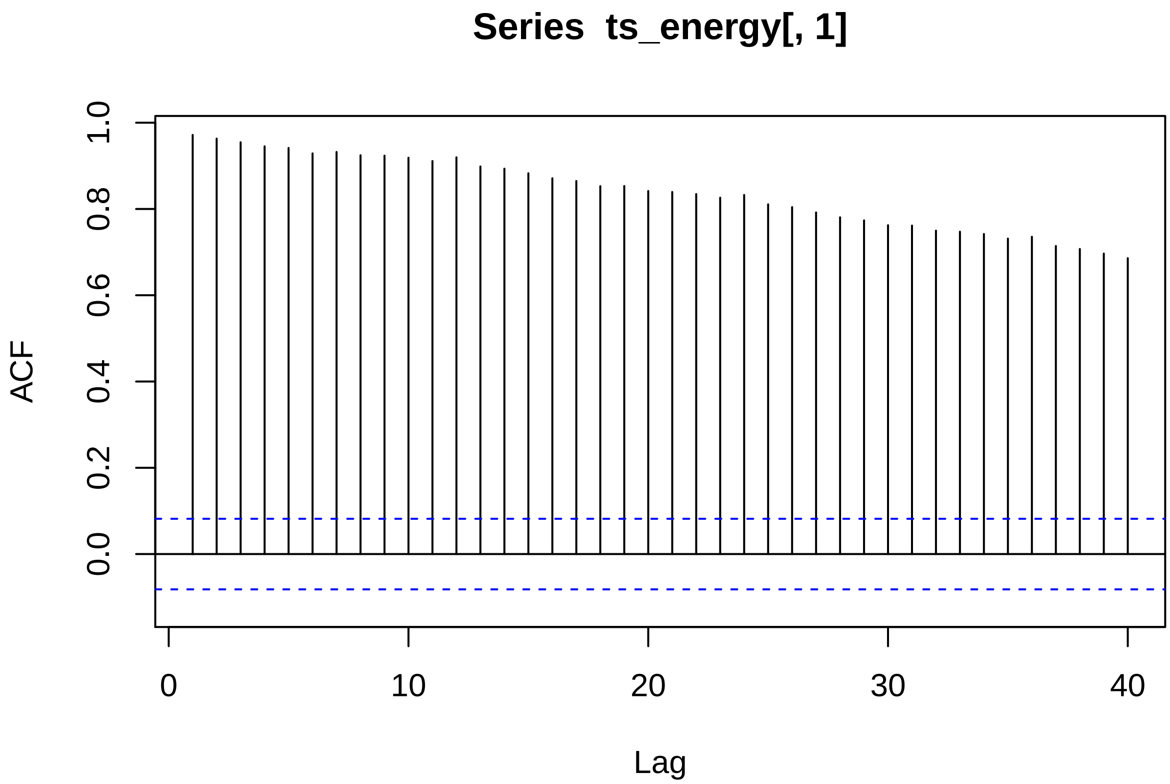
```
##
##  Pearson’s product-moment correlation
##
## data:  ts_energy[, 1] and ts_energy[, 3]
## t = -6.3222, df = 572, p-value = 5.195e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3304936 -0.1774402
## sample estimates:
##        cor
## -0.2555675
```

The correlation significance test results show that the correlation between total biomass energy production and total renewable energy production, and total biomass energy production and hydroelectric power consumption are significant because they have p-values < 0.05. However, the correlation between total renewable energy production and hydroelectric power consumption is not significant since the p-value for its correlation significance test is 0.9475, which is way bigger than 0.05(alpha).

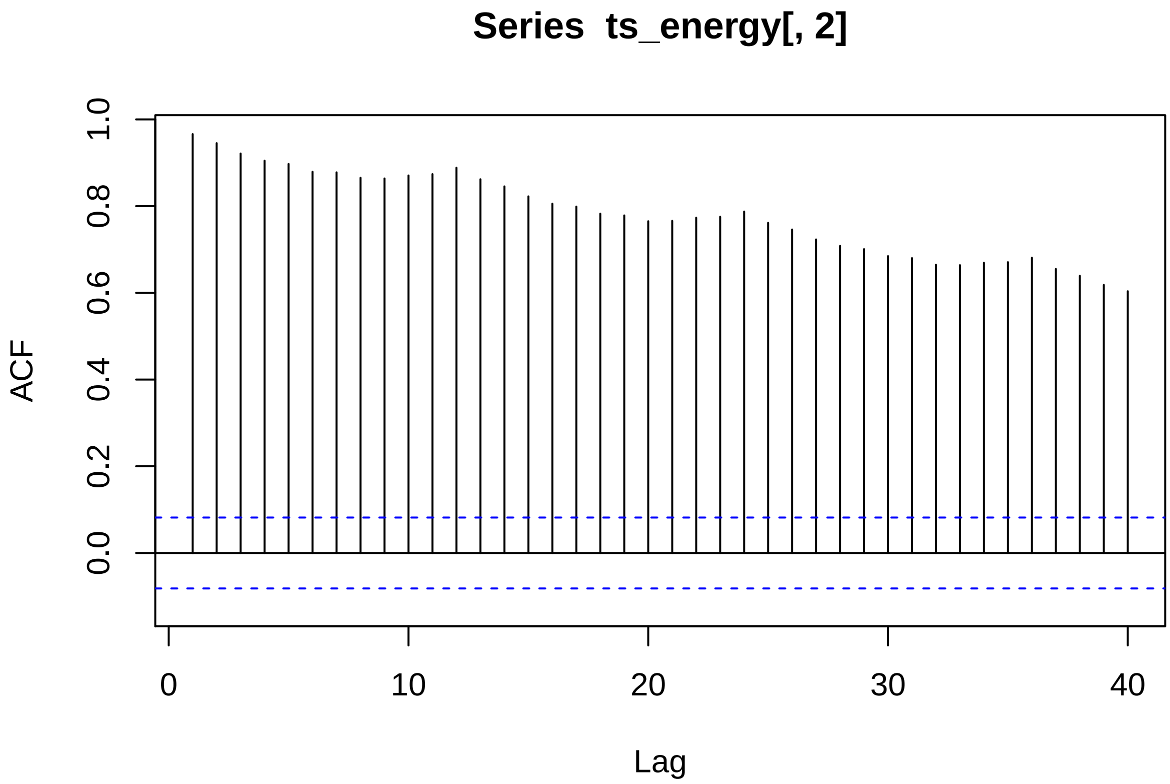
Question 6

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?

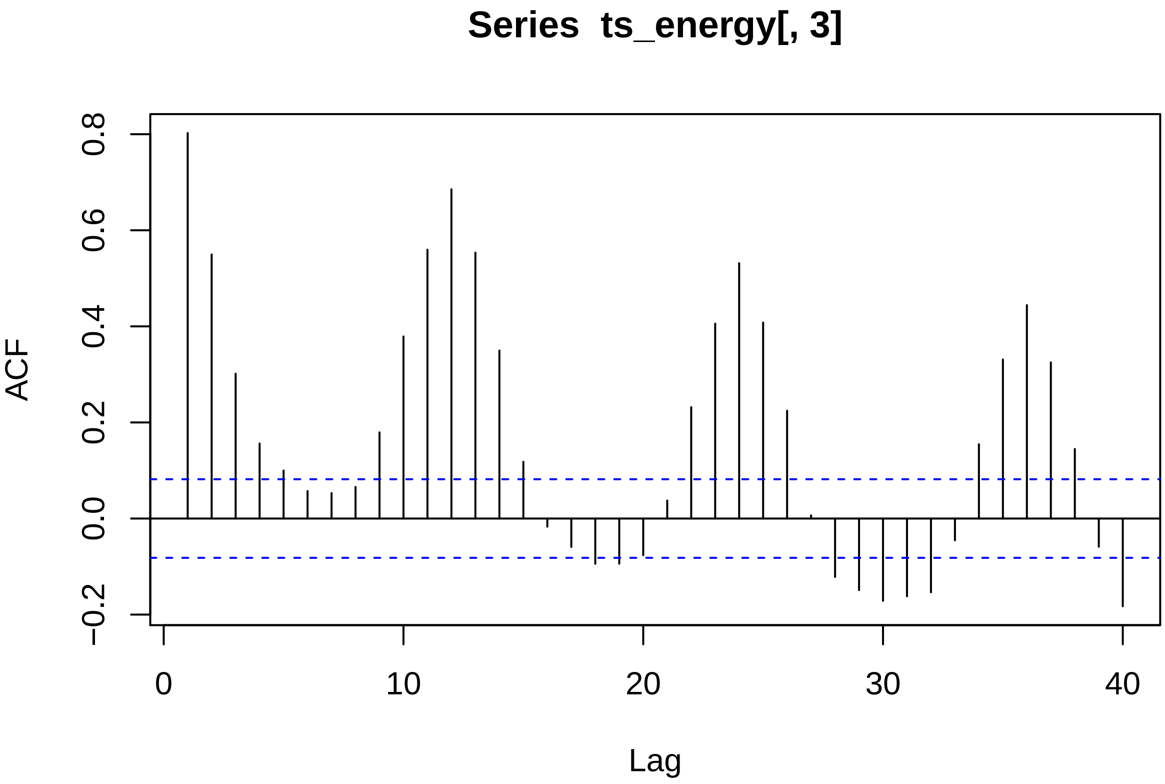
```
biomass_acf=Acf(ts_energy[,1],lag.max=40, type="correlation", plot=TRUE)
```



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



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

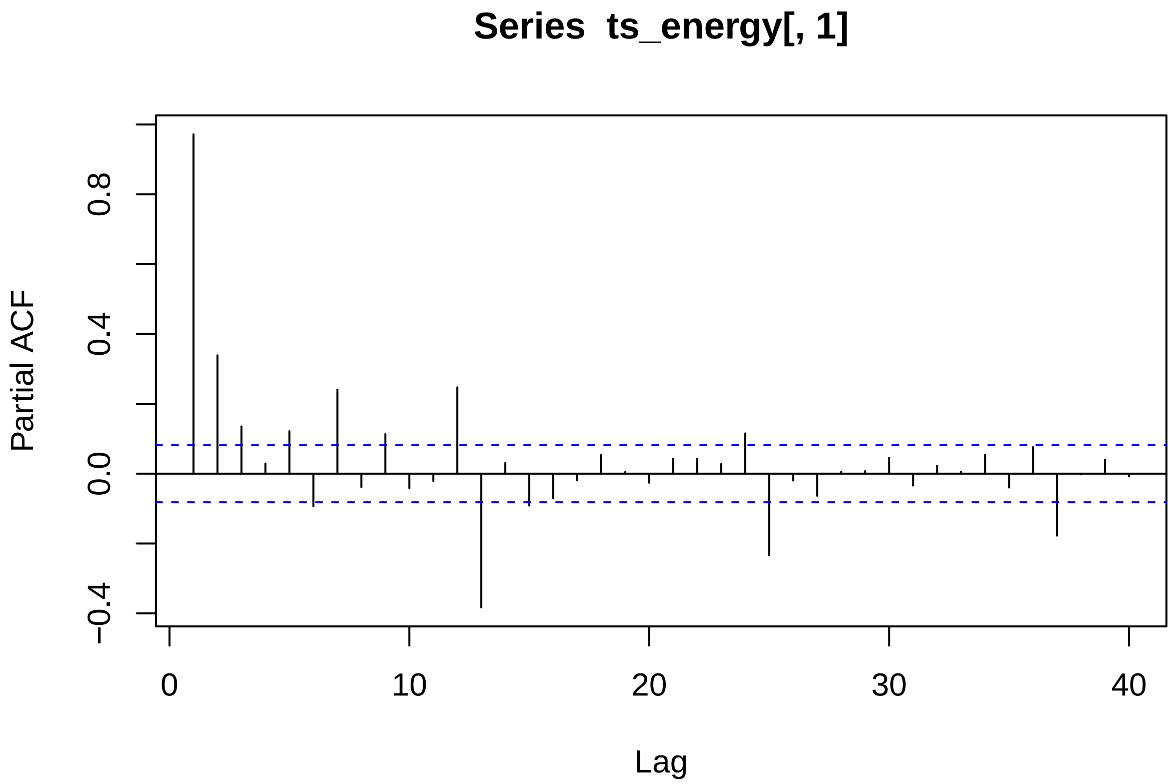


There is a same cyclical pattern in all three plots (12 lags/cycle), with the hydro one has the most obvious cyclical pattern, followed by renewable energy production and biomass energy production. The biomass ACF and renewable energy ACF are all positive from lag 1 to lag 40 while some of the low ACF for hydroelectric power consumption hit negative values. Moreover,all three plots show a decreasing trend in ACF.

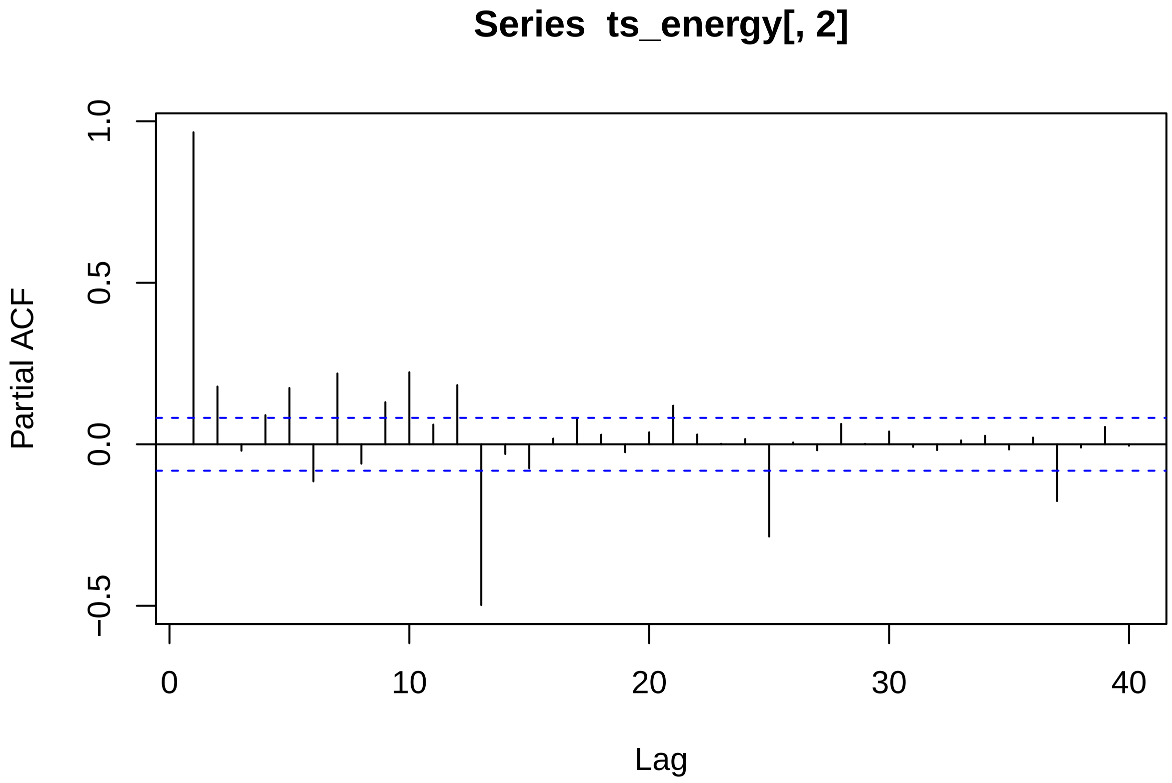
Question 7

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

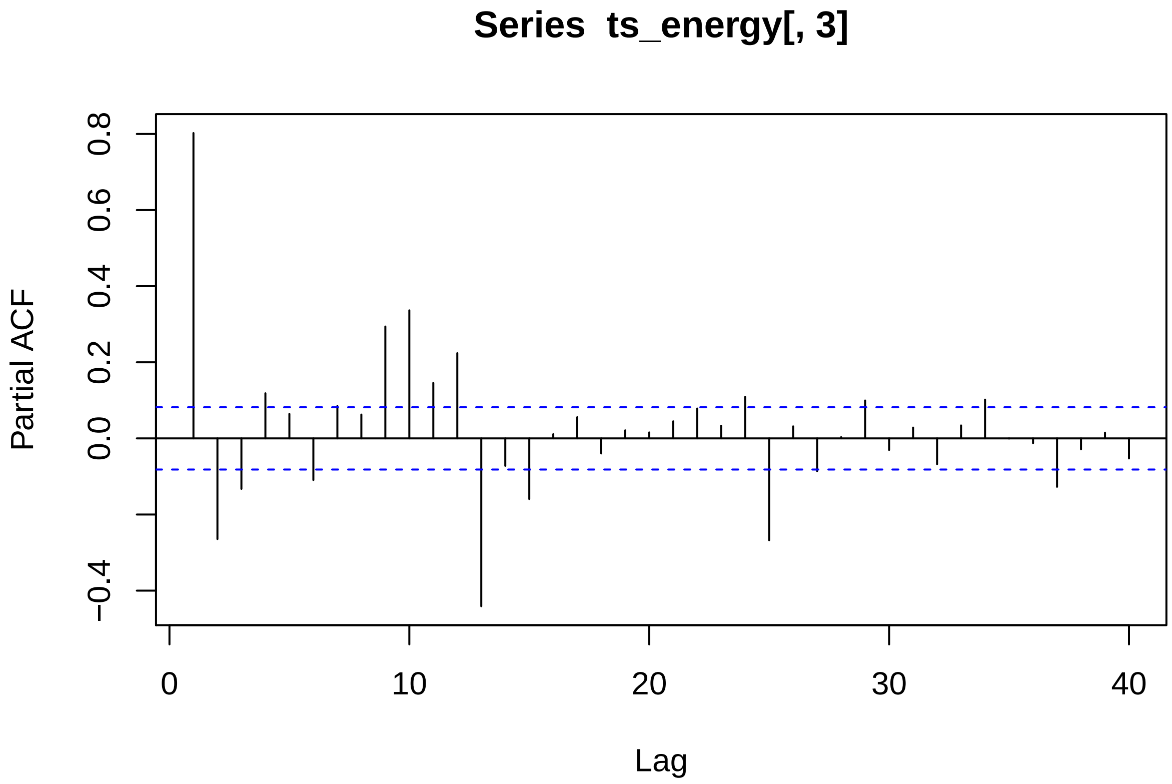
```
biomass_pacf=Pacf(ts_energy[,1],lag.max=40, plot=TRUE)
```



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



```
hydro_pacf=Pacf(ts_energy[,3],lag.max=40, plot=TRUE)
```



Plots for biomass energy production and renewable energy production now show both positive and negative PACF instead of just positive values like shown in the ACF plots. PACF plots also have more values fall into the confidence interval (blue dotted line range) than ACF plots do. PACF plots also show a decreasing trend (positive and negative values tend to gradually getting closer to 0) like the ACF plots do.