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

Assignment 2 - Due date 02/03/23

John Rooney

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(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
```

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 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="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Ed
```

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_data_df<-energy_data %>%
  select(Month, Total.Biomass.Energy.Production, Total.Renewable.Energy.Production, Hydroelectric.Power
head(energy_data_df)
```

```
##           Month Total.Biomass.Energy.Production
## 1  1973 January                         129.787
## 2 1973 February                         117.338
## 3    1973 March                         129.938
## 4    1973 April                         125.636
## 5      1973 May                         129.834
## 6     1973 June                         125.611
##   Total.Renewable.Energy.Production Hydroelectric.Power.Consumption
## 1                           403.981                         272.703
## 2                           360.900                         242.199
## 3                           400.161                         268.810
## 4                           380.470                         253.185
## 5                           392.141                         260.770
## 6                           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().

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

```
##          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
##          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
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(energy_data_df$Total.Biomass.Energy.Production)
```

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

```
sd(energy_data_df$Total.Biomass.Energy.Production)
```

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

```
mean(energy_data_df$Total.Renewable.Energy.Production)
```

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

```
sd(energy_data_df$Total.Renewable.Energy.Production)
```

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

```
mean(energy_data_df$Hydroelectric.Power.Consumption)
```

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

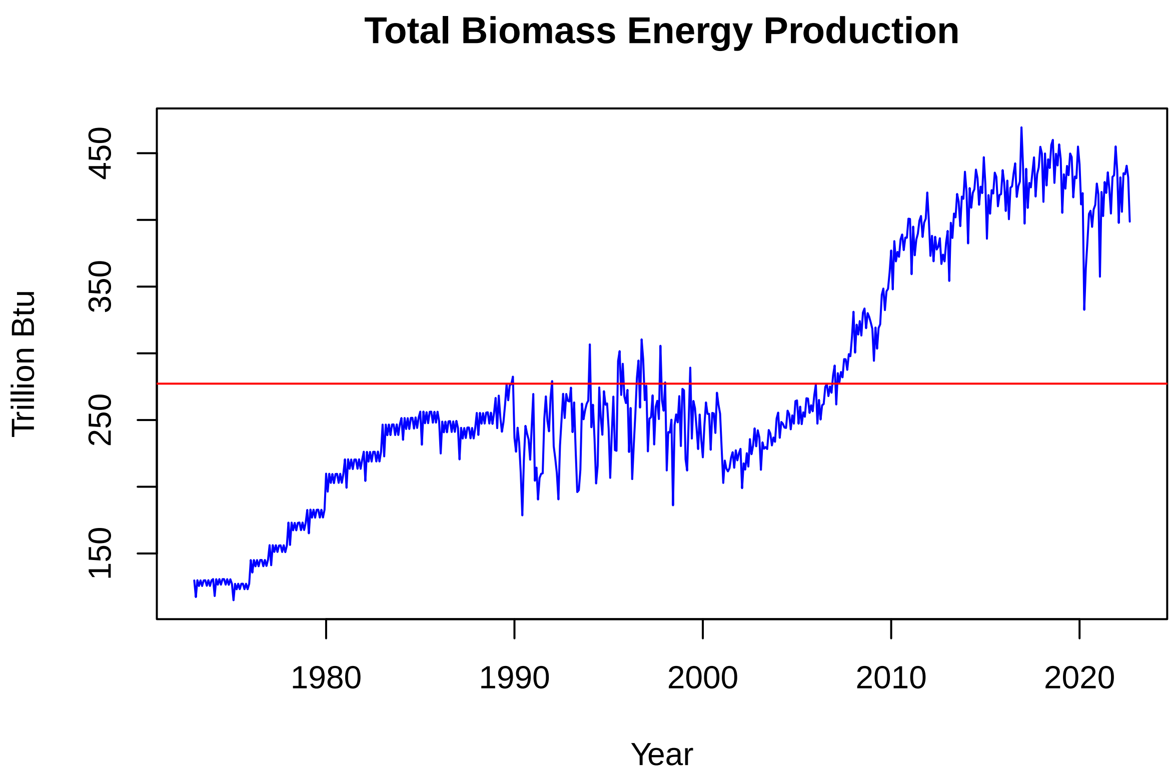
```
sd(energy_data_df$Hydroelectric.Power.Consumption)
```

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

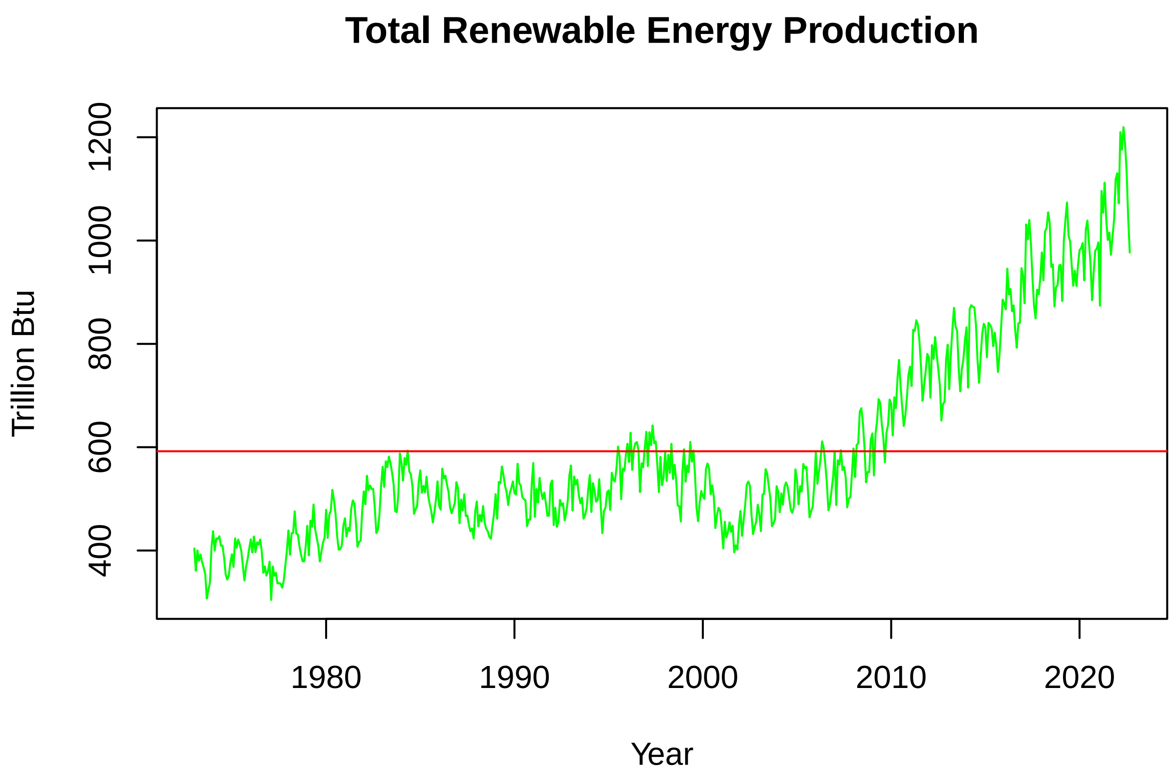
Question 4

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.

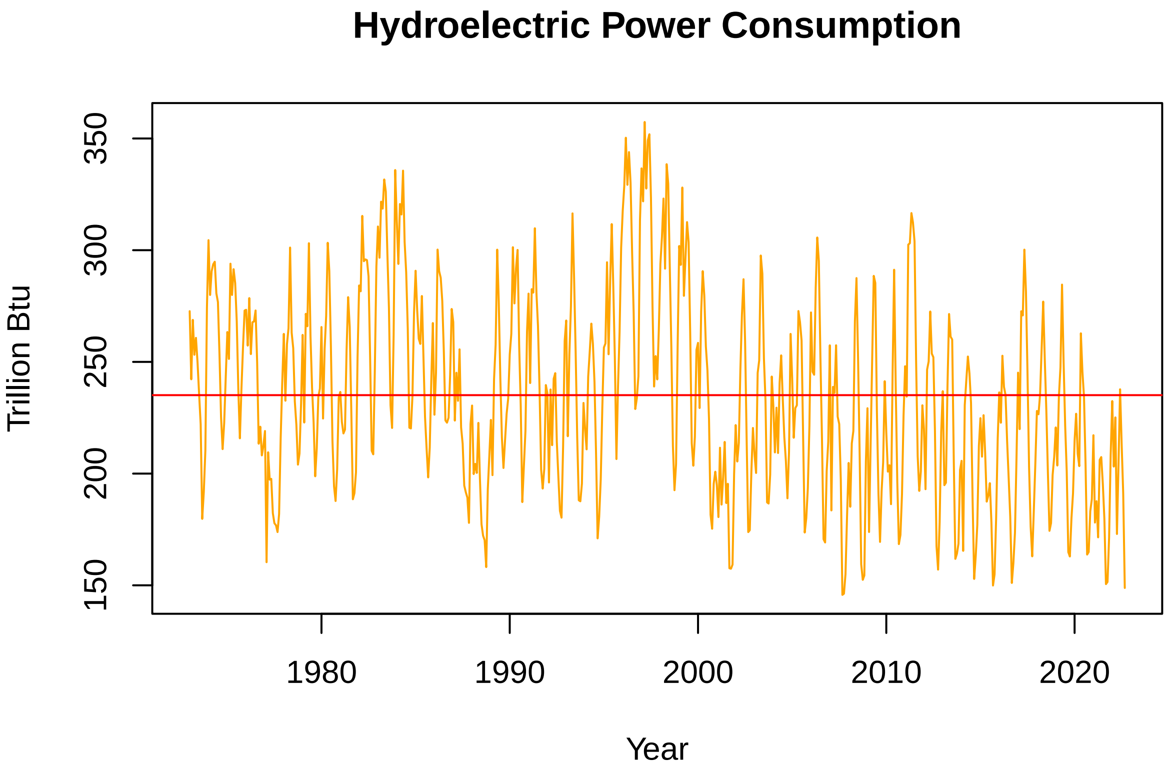
```
plot(ts_energy_data[,"Total.Biomass.Energy.Production"],type="l",col="blue",xlab="Year",ylab="Trillion I
abline(h=mean(ts_energy_data[,"Total.Biomass.Energy.Production"]),col="red")
```




```
plot(ts_energy_data[,"Total.Renewable.Energy.Production"],type="l",col="green",xlab="Year",ylab="Trilli
abline(h=mean(ts_energy_data[,"Total.Renewable.Energy.Production"]),col="red")
```



```
plot(ts_energy_data[,"Hydroelectric.Power.Consumption"],type="l",col="orange",xlab="Year",ylab="Trillion
abline(h=mean(ts_energy_data[,"Hydroelectric.Power.Consumption"]),col="red")
```



Total biomass energy production grew rapidly through the mid-1980s, then stayed relatively stable through 2000 when it again grew rapidly up through 2015. Total renewable energy was relatively stable from 1980 through 2000 when it began to grow rapidly. Total hydroelectric power looks highly seasonal, showing decreases starting around 2000 which would appear to correlate with production of biomass and renewable energy.

Question 5

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

```
cor(ts_energy_data)
```

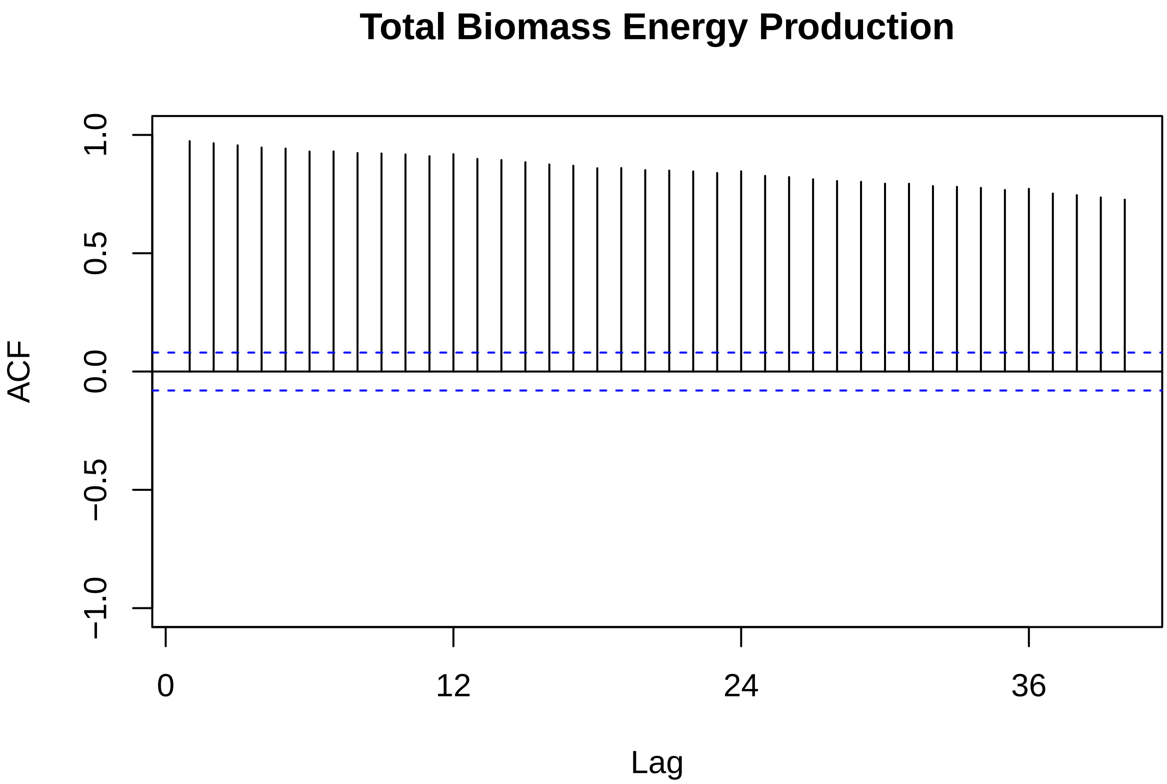
```
##                                     Total.Biomass.Energy.Production
## Total.Biomass.Energy.Production                           1.0000000
## Total.Renewable.Energy.Production                         0.9185941
## Hydroelectric.Power.Consumption                          -0.2998201
##                                     Total.Renewable.Energy.Production
## Total.Biomass.Energy.Production                            0.91859411
## Total.Renewable.Energy.Production                          1.00000000
## Hydroelectric.Power.Consumption                           -0.09958758
##                                     Hydroelectric.Power.Consumption
## Total.Biomass.Energy.Production                         -0.29982013
## Total.Renewable.Energy.Production                       -0.09958758
## Hydroelectric.Power.Consumption                          1.00000000
```

Total biomass energy production and total renewable energy production are strongly positively correlated, with values of 1 to .92 and vice versa. Each is slightly negatively correlated with hydroelectric power consumption, with hydroelectric power consumption to biomass energy production having a correlation value of -0.3 and hydroelectric power consumption to renewable energy production having a correlation value of -0.1.

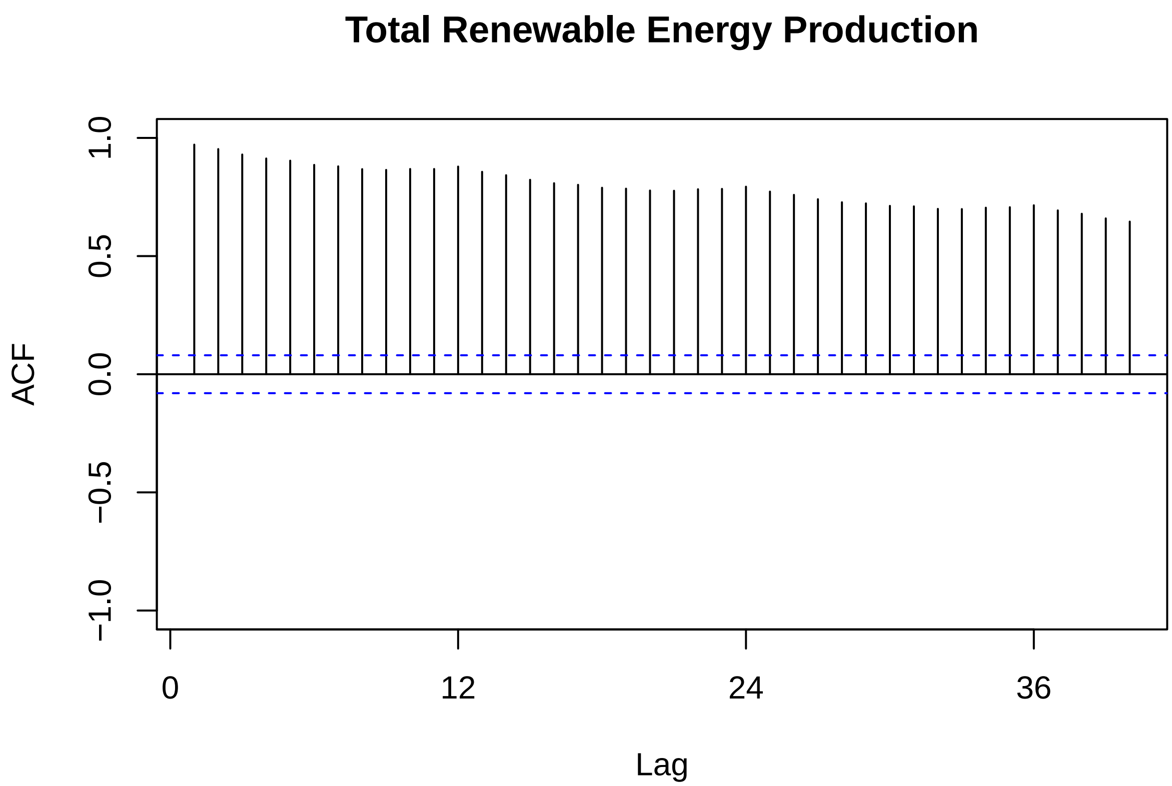
Question 6

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?

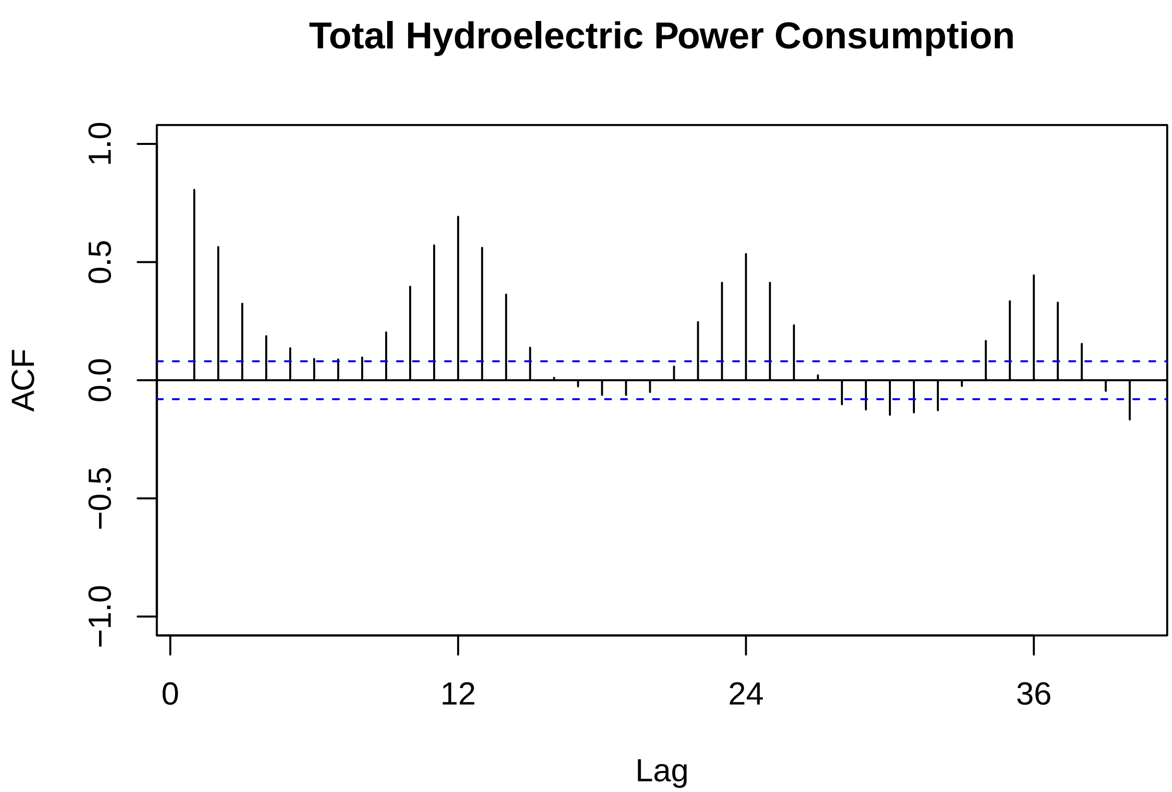
```
Acf(ts_energy_data[,1], lag.max = 40, main="Total Biomass Energy Production", ylim=c(-1,1))
```




```
Acf(ts_energy_data[,2], lag.max = 40, main="Total Renewable Energy Production", ylim=c(-1,1))
```




```
Acf(ts_energy_data[,3], lag.max = 40, main="Total Hydroelectric Power Consumption", ylim=c(-1,1))
```

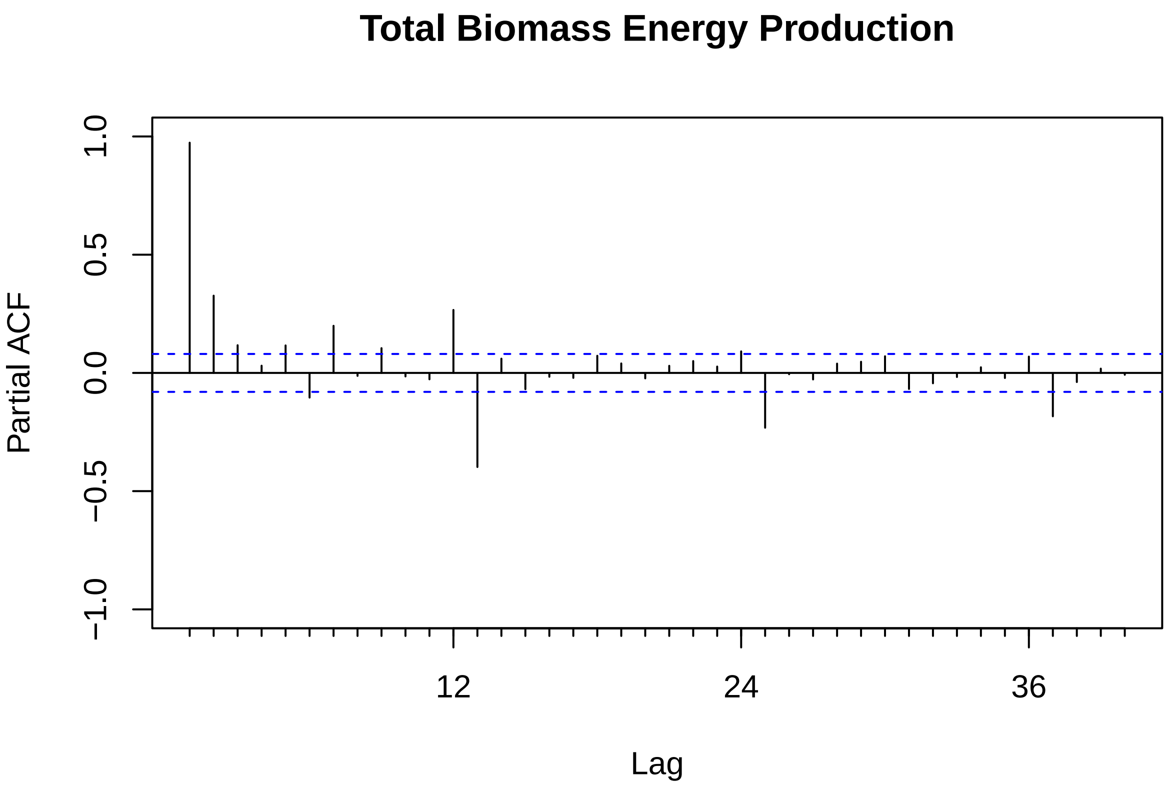


Total biomass energy production and total renewable energy production show similar behavior, both with quite high significance that slowly decreases over time. Total hydroelectric power consumption shows more seasonality, with higher significance at the beginning and end of each year. This significance decreases over time.

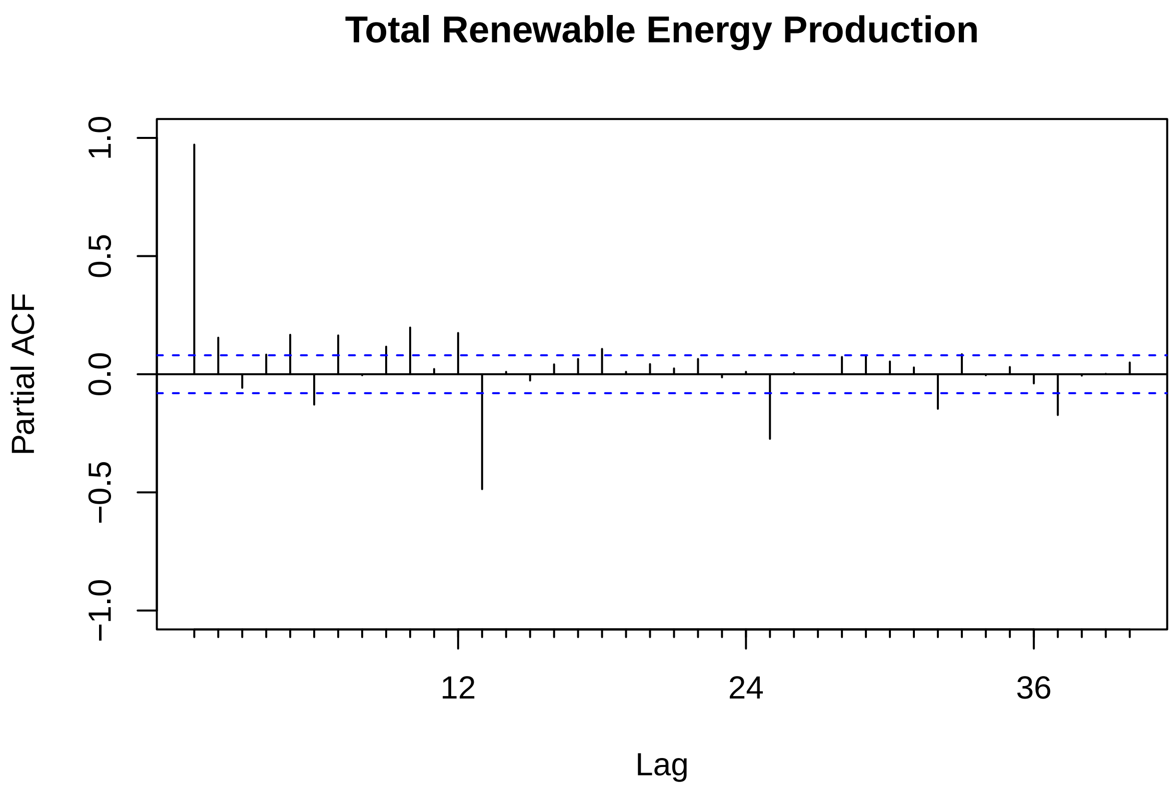
Question 7

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

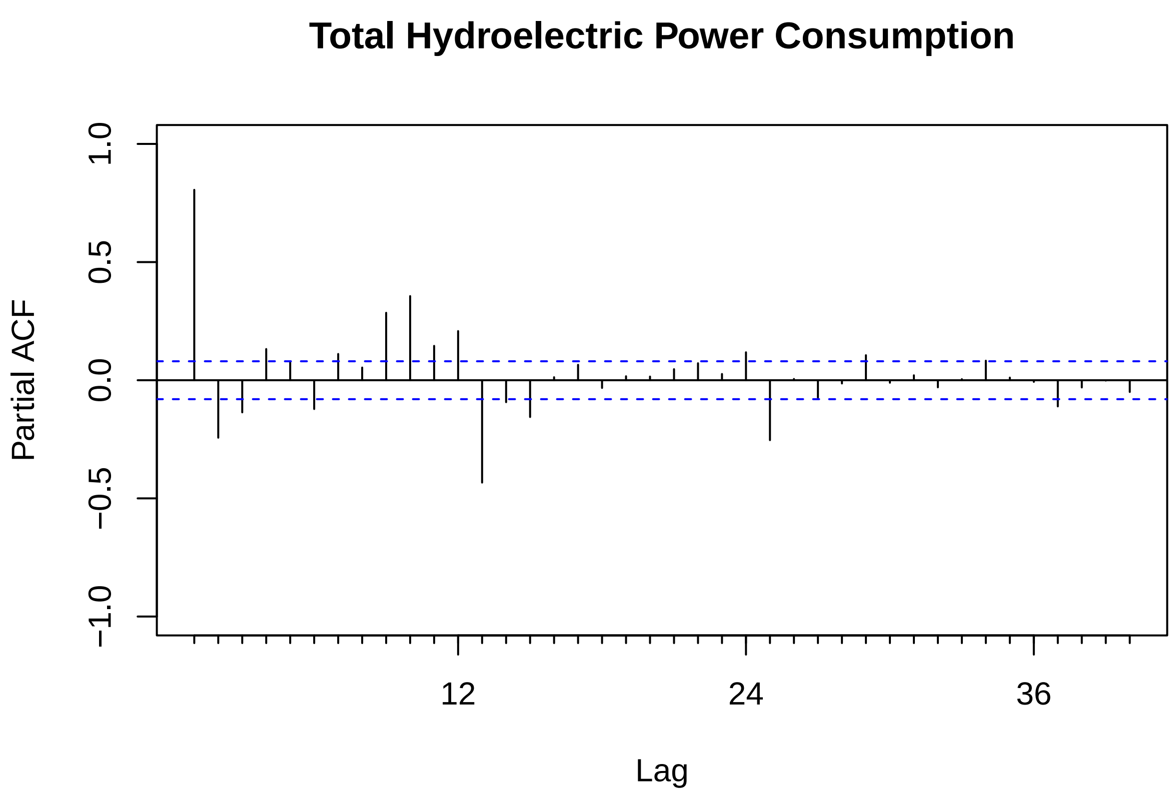
```
Pacf(ts_energy_data[,1], lag.max = 40, main="Total Biomass Energy Production", ylim=c(-1,1))
```




```
Pacf(ts_energy_data[,2], lag.max = 40, main="Total Renewable Energy Production", ylim=c(-1,1))
```




```
Pacf(ts_energy_data[,3], lag.max = 40, main="Total Hydroelectric Power Consumption", ylim=c(-1,1))
```



Total biomass energy production and renewable energy production again show similar behavior, although far fewer data points are significant. In both cases significant moments in time are right at the beginning and end of the year. In both cases significance decreases over time, especially so for total renewable energy production. For total hydroelectric power, data appears significant towards the end of the first first and far less in further years, with significant data showing only at the beginning of each year.