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

Assignment 2 - Due date 02/03/21

Thomas Hancock

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
#install.packages("forecast", "tseries")
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)
```

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 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 (I used read_excel because read.table wasn't working for me)
Energy <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx", c

Energy <- Energy[-1,] #Remove row that has units (they are all the same - Trillion BTU)
```

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_small <- Energy[,c(4:6)] #Create data frame with subset of data
head(Energy_small) #Check subset
```

```
## # A tibble: 6 x 3
##   `Total Biomass Energy Pr~ `Total Renewable Energy Pr~ `Hydroelectric Power Co~
##   <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
```

```
Energy_small <- data.frame(lapply(Energy_small, as.numeric)) # Set data to numeric so we can do numeric
```

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().

```
Energy_ts <- ts(Energy_small, start=c(1973,1), frequency = 12)

# Bio_ts <- ts(Energy_small$`Total Biomass Energy Production`, start=c(1973,1), frequency = 12) # Bioma
# RE_ts <- ts(Energy_small$`Total Renewable Energy Production`, start=c(1973,1), frequency = 12) # Rene
# Hydro_ts <- ts(Energy_small$`Hydroelectric Power Consumption`, start=c(1973,1), frequency = 12) # Hyd
```

Question 3

Compute mean and standard deviation for these three series.

```
# Calculate means (requires conversion to numeric class)
Bio_avg <- mean(Energy_ts[,1])
RE_avg <- mean(Energy_ts[,2])
Hydro_avg <- mean(Energy_ts[,3])

# Calculate standard deviations
Bio_sd <- sd(Energy_ts[,1])
RE_sd <- sd(Energy_ts[,2])
Hydro_sd <- sd(Energy_ts[,3])
```

```
# Display results
Bio_avg
```

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

```
RE_avg
```

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

```
Hydro_avg
```

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

```
Bio_sd
```

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

```
RE_sd
```

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

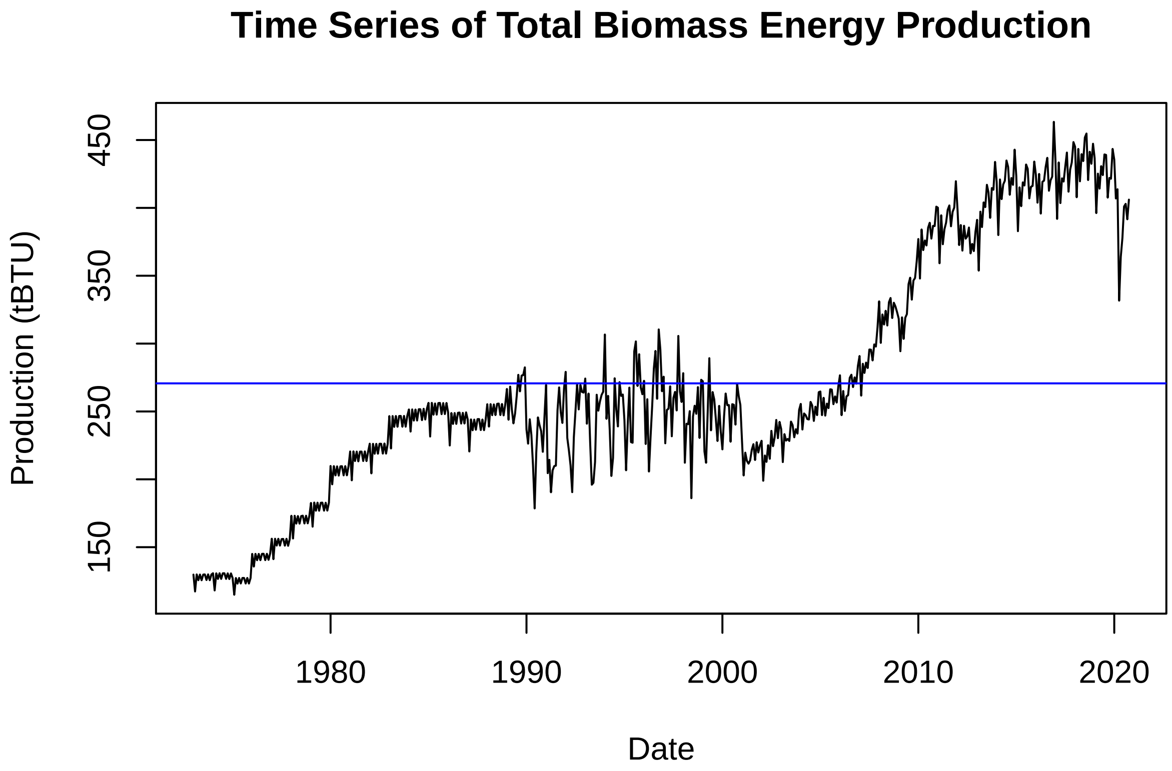
```
Hydro_sd
```

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

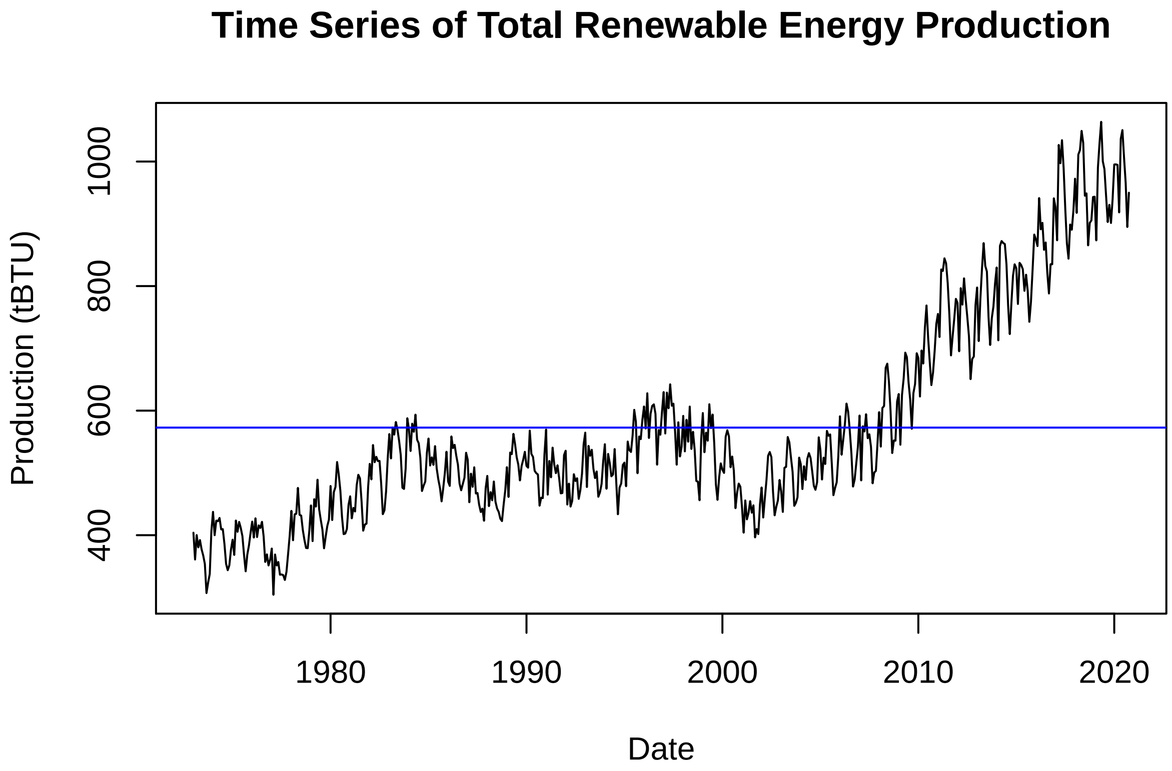
Question 4

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.

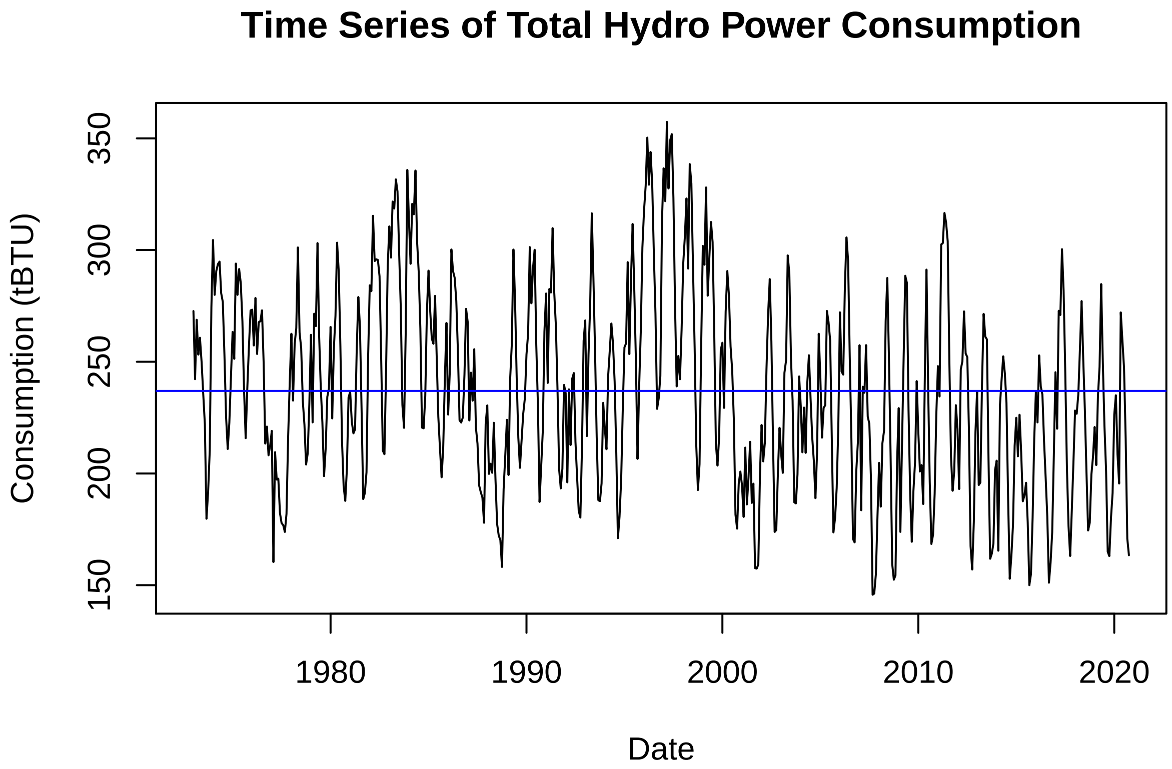
```
#Biomass Plot
plot.ts(Energy_ts[,1], main = "Time Series of Total Biomass Energy Production", axes = TRUE, xlab = "Da
abline(h = Bio_avg, col = "Blue")
```



```
# Renewable Energy Plot
plot.ts(Energy_ts[,2], main = "Time Series of Total Renewable Energy Production", axes = TRUE, xlab = "
abline(h = RE_avg, col = "Blue")
```



```
# Hydro Plot
plot.ts(Energy_ts[,3], main = "Time Series of Total Hydro Power Consumption", axes = TRUE, xlab = "Date
abline(h = Hydro_avg, col = "Blue")
```



> Answer: In the first plot, we see that Biomass Production has generally increased from 1970 to present, with a period of growth in the late '70s until the late '80s or so, followed by a stagnation which included wide variations in production until the '00s. Biomass Production then grew for another decade or so, before starting to have large variations and potential stagnation again. There is potentially a large drop in Biomass Production around 2020, but it is unclear if that is an anomaly since it occurs at the end of the dataset.

In the second plot, we see that Total Renewable Production has grown significantly since the early 2000s. While there was a slight growth trend, largely obscured by regular variation, between the '70s and '00s, the recent growth has been quite pronounced.

In the third plot, we see a lot of variation between neighboring years as well as perhaps a cyclical pattern every decade or so. There is a slight downward trend after 2000 (these decades seem to usually be below the average), but it is much less pronounced than the later trends of the other two datasets.

Question 5

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

```
cor(Energy_ts[,1], Energy_ts[,2]) # Correlation between biomass and total renewables
```

```
## [1] 0.9234609
```

```
cor(Energy_ts[,1], Energy_ts[,3]) # Correlation between biomass and hydro
```

```
## [1] -0.2555675
```

```
cor(Energy_ts[,2], Energy_ts[,3]) # Correlation between total renewables and hydro
```

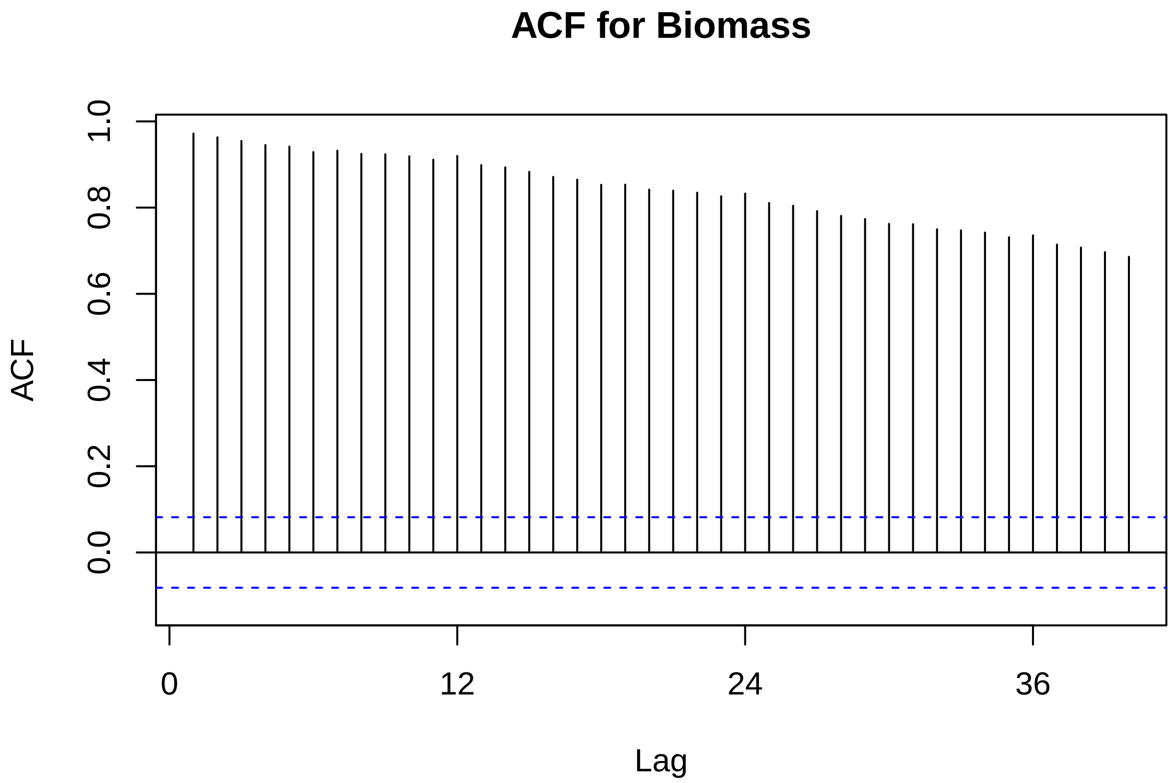
```
## [1] -0.002756852
```

Answer: We can see that there is high correlation between biomass and total renewable generation (0.923), showing that these two follow similar trends. There is a slight negative correlation between biomass and hydro power (-0.256), so when one increases, the other one tends to decrease. There is an almost negligible negative correlation between total renewables and hydro (-0.003), so one is not a useful predictor of the behavior of the other.

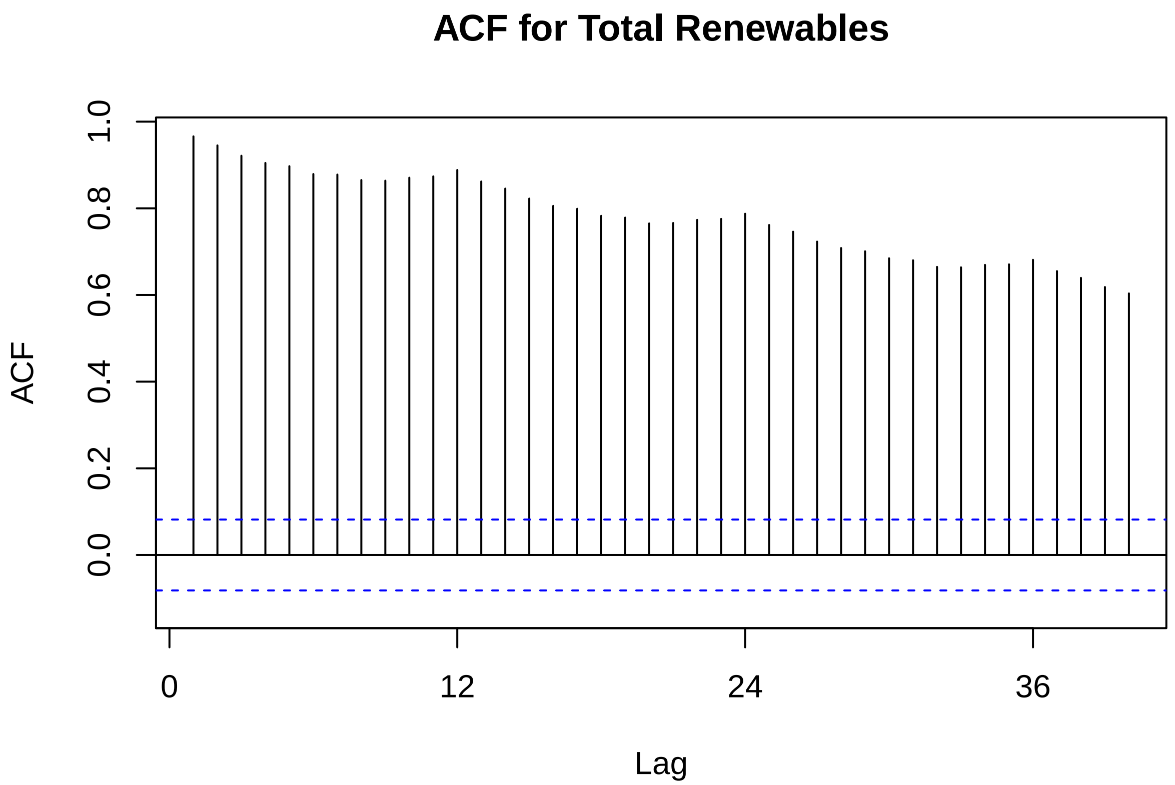
Question 6

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?

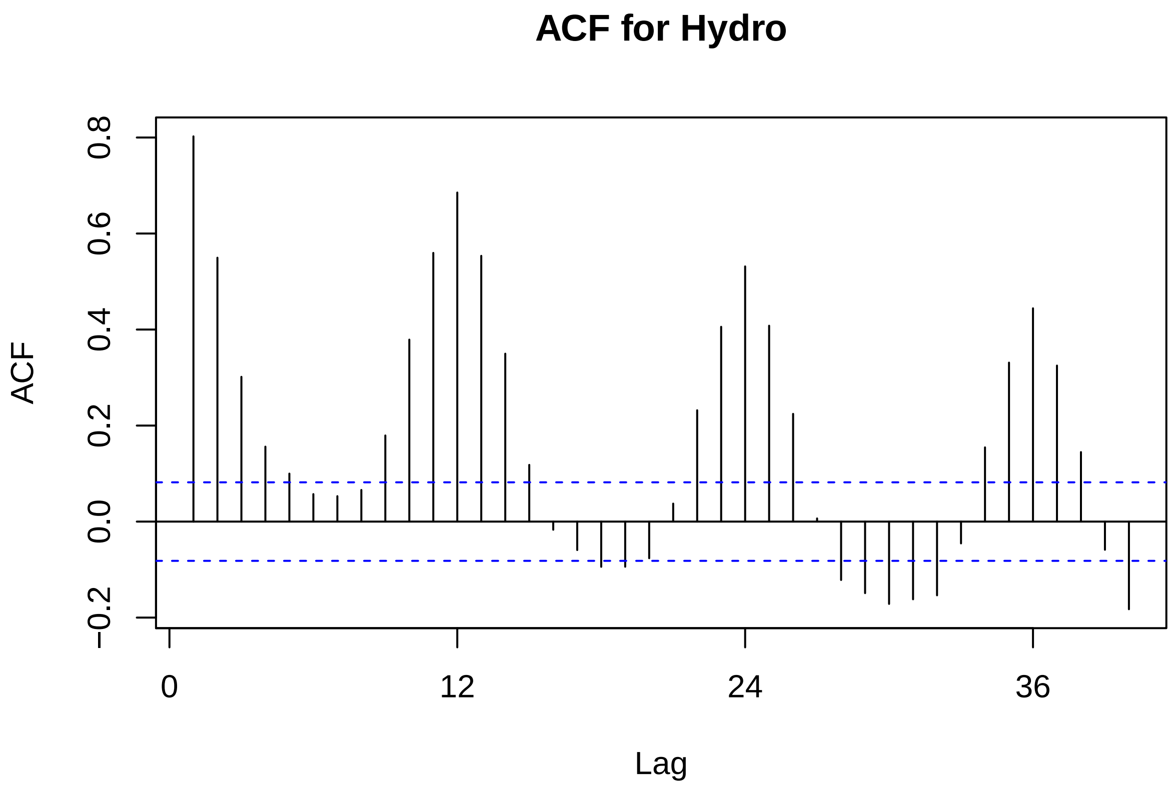
```
Acf(Energy_ts[,1], lag.max = 40, main = "ACF for Biomass")
```



```
Acf(Energy_ts[,2], lag.max = 40, main = "ACF for Total Renewables")
```



```
Acf(Energy_ts[,3], lag.max = 40, main = "ACF for Hydro")
```

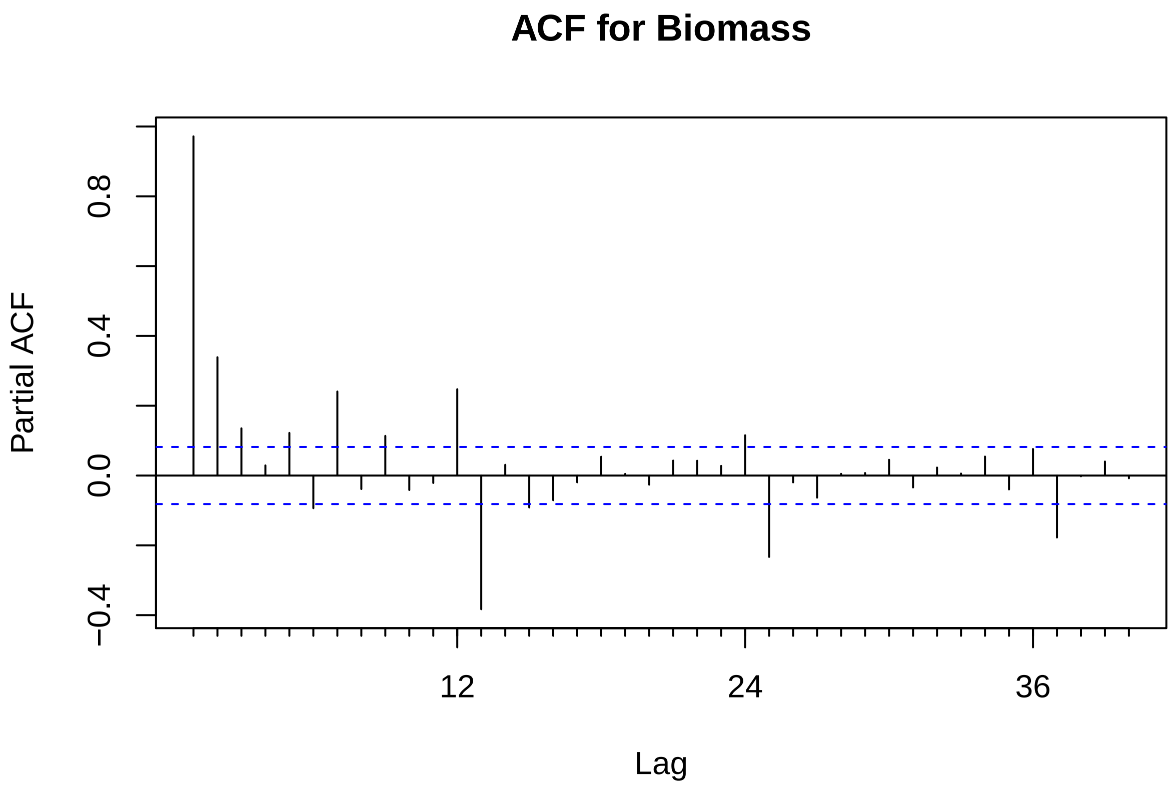


> Answer: The first two plots (Biomass and Total Renewables) are similar, with autocorrelation decreasing fairly regularly as lag increases (i.e., the impact of closer numbers is more significant than numbers that are farther away). However, the ACF plot for hydropower is different, showing a periodic (seasonal) trend in the ACF significance. Notably, there is a cycle that seems to be 12 months of high correlation.

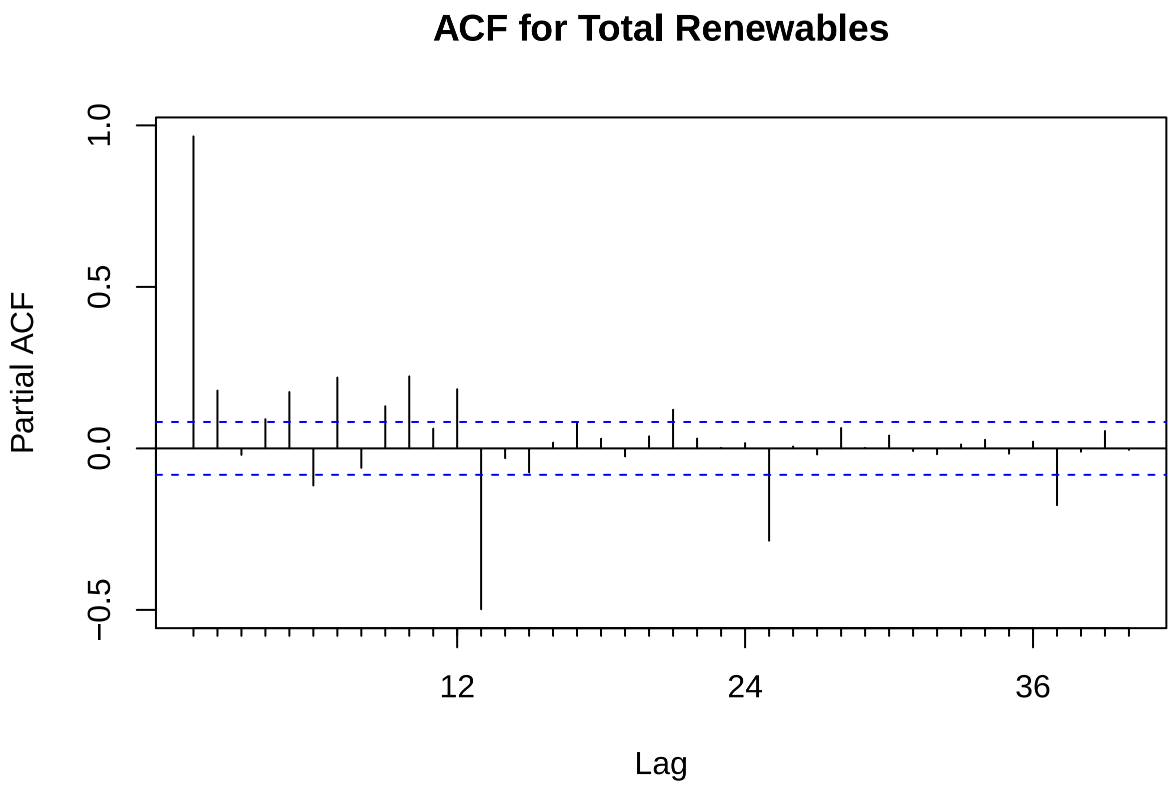
Question 7

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

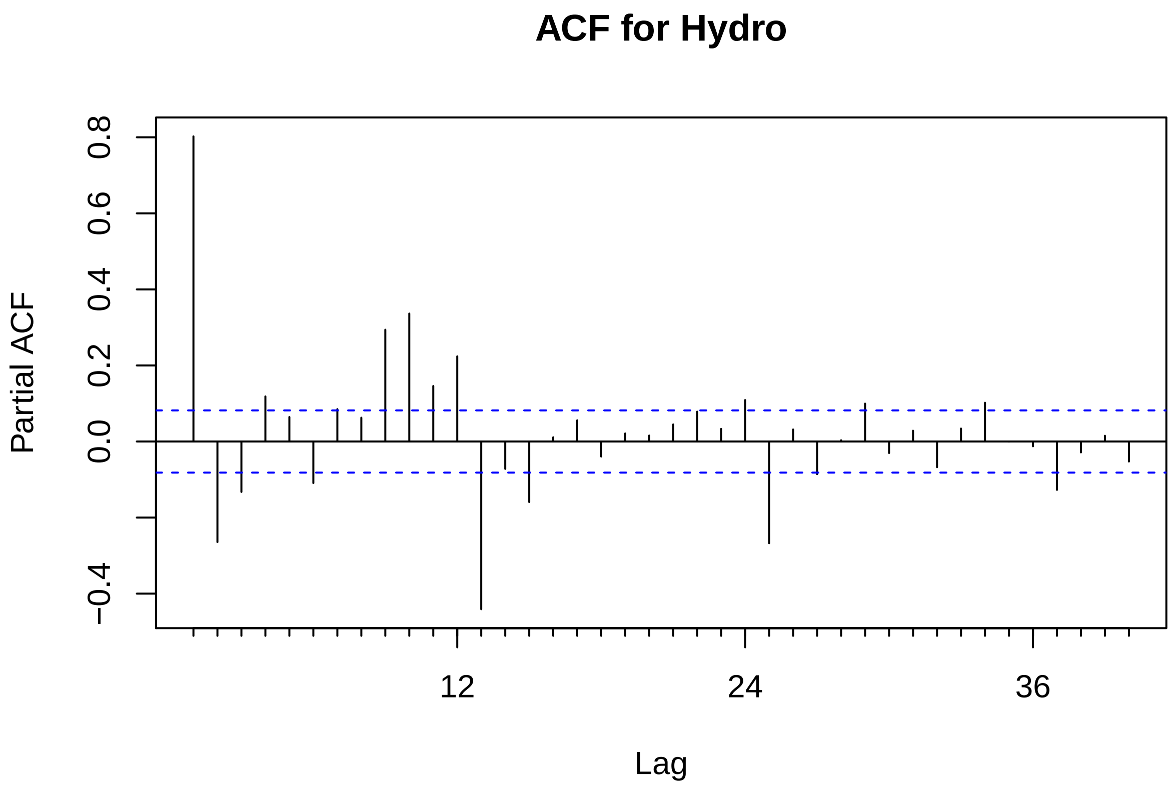
```
Pacf(Energy_ts[,1], lag.max = 40, main = "ACF for Biomass")
```



```
Pacf(Energy_ts[,2], lag.max = 40, main = "ACF for Total Renewables")
```



```
Pacf(Energy_ts[,3], lag.max = 40, main = "ACF for Hydro")
```



> Answer: The PACF values for each variable are much lower than the ACF values (as we would expect). Most of the values are much closer to the 95% significance level (blue dashed lines), with many values being statisically insignificant (at the 95% confidence level). There is also more of a cyclical pattern amongst the first two plots now (although it is a little hard to tell if there is a pattern or just a random distribution around 0), that was not present in the ACF plots. The last plot still looks a bit cyclical, but it is not as clear as the ACF.