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

Assignment 2 - Due date 02/05/21

Ben Culberson

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.\

```r
#Load/install required package here
library(forecast)
```

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

```r
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
```

```r
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
Renewable.df <- read_excel("/Users/benculberson/Documents/Duke /Spring 2021/Time Series Analysis/ENV790
```

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ...
```

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.

```
Renewable_altered.df<-Renewable.df[12:585,4:6]
colnames(Renewable_altered.df)=c("Total Biomass Energy Production","Total Renewable Energy Production",
```

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().

```
Renewable_ts.df<-ts(Renewable_altered.df, start=1, frequency=1)
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(Renewable_ts.df[,1])
```

```
## [1] 286.0889
```

```
sd(Renewable_ts.df[,1])
```

```
## [1] 165.3481
```

```
mean(Renewable_ts.df[,2])
```

```
## [1] 287.5
```

```
sd(Renewable_ts.df[,2])
```

```
## [1] 165.8438
```

```
mean(Renewable_ts.df[,3])
```

```
## [1] 287.5
```

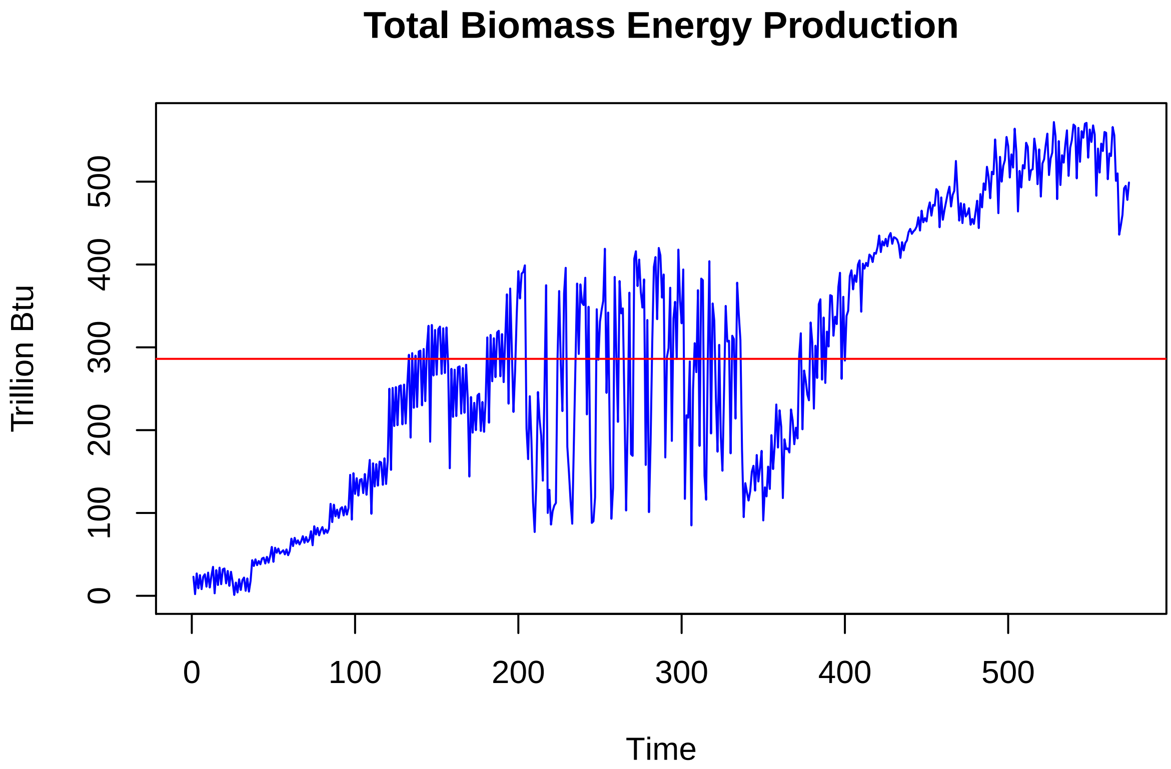
```
sd(Renewable_ts.df[,3])
```

```
## [1] 165.8438
```

Question 4

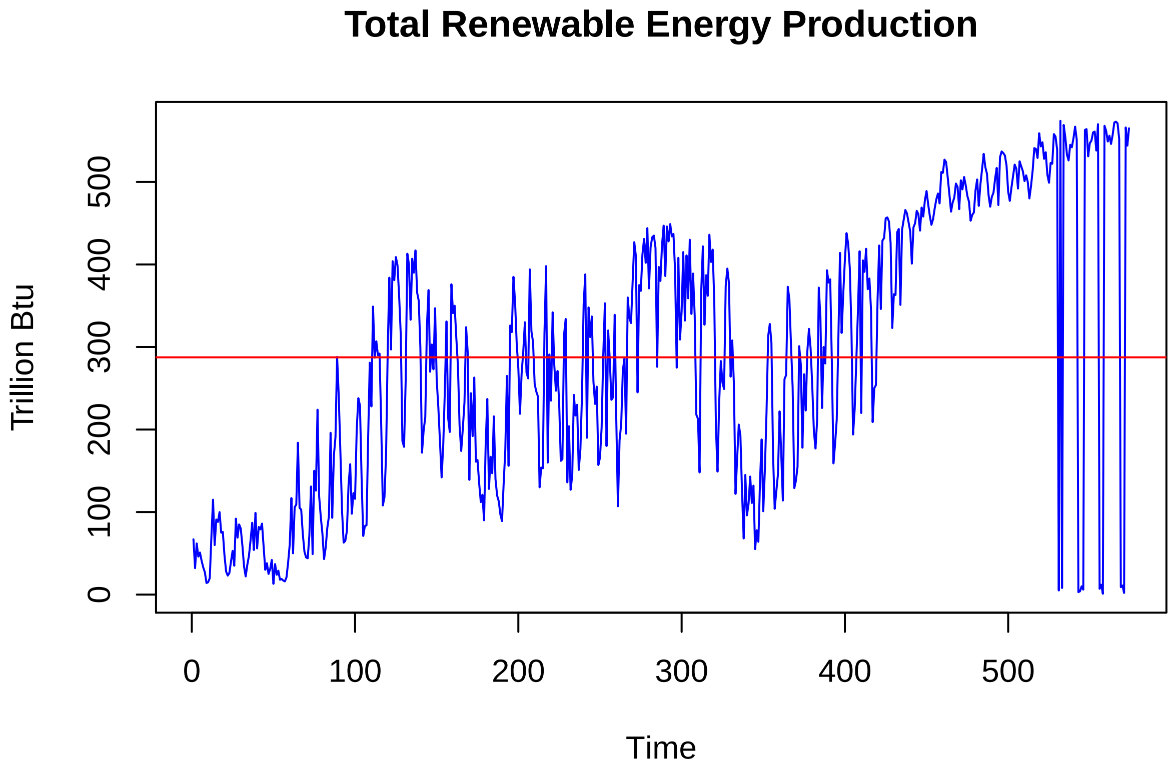
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
plot(Renewable_ts.df[,1], ylab="Trillion Btu", col=c("blue"))+abline(h=mean(Renewable_ts.df[,1]), col=c
```



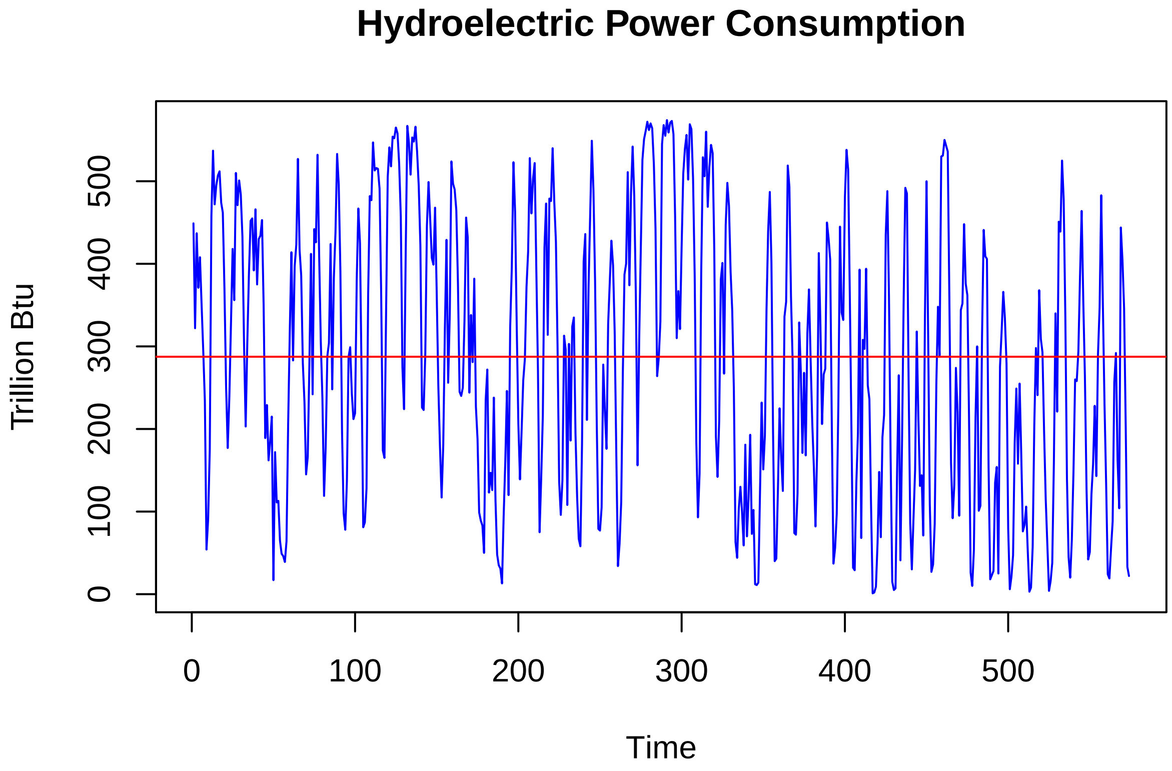
```
## integer(0)
```

```
plot(Renewable_ts.df[,2], ylab="Trillion Btu", col=c("blue"))+abline(h=mean(Renewable_ts.df[,2]), col=c
```



```
## integer(0)
```

```
plot(Renewable_ts.df[,3], ylab="Trillion Btu", col=c("blue"))+abline(h=mean(Renewable_ts.df[,3]), col=c
```



```
## integer(0)
```

Question 5

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

```
cor(Renewable_ts.df)
```

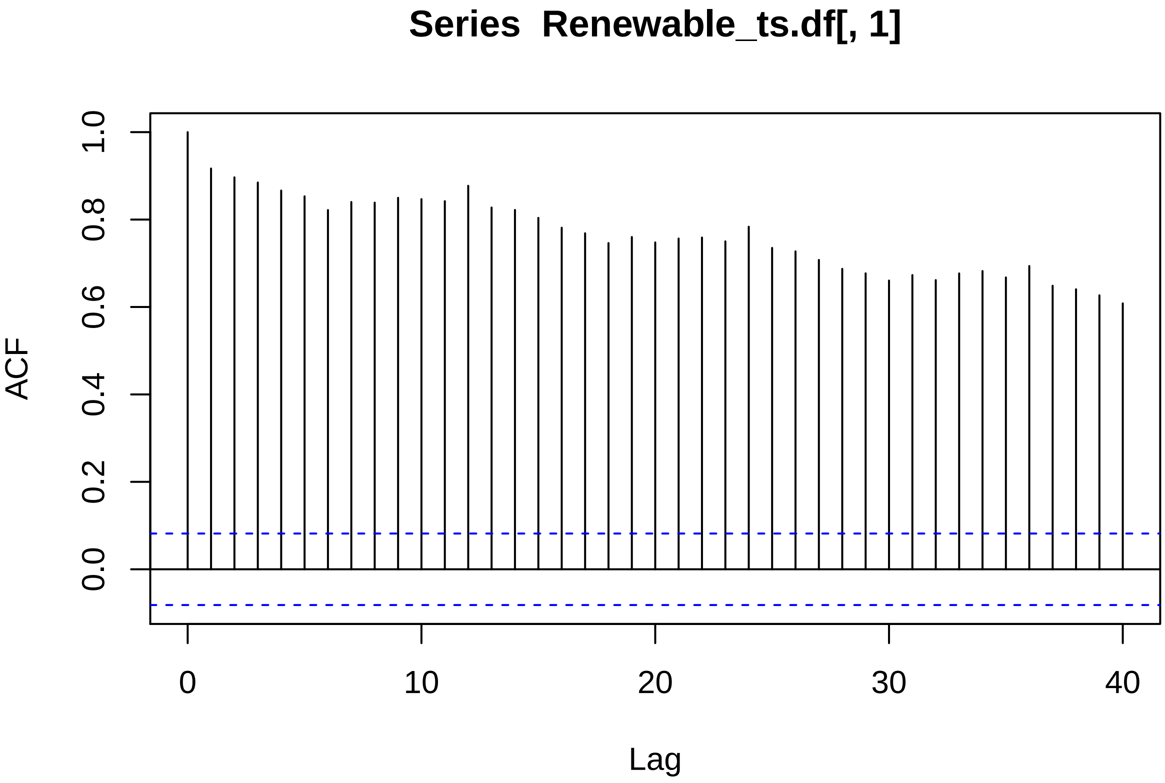
```
##                                    Total Biomass Energy Production
## Total Biomass Energy Production                          1.0000000
## Total Renewable Energy Production                        0.7719172
## Hydroelectric Power Consumption                         -0.2476318
##                                    Total Renewable Energy Production
## Total Biomass Energy Production                           0.77191719
## Total Renewable Energy Production                         1.00000000
## Hydroelectric Power Consumption                           0.08066507
##                                    Hydroelectric Power Consumption
## Total Biomass Energy Production                        -0.24763182
## Total Renewable Energy Production                       0.08066507
## Hydroelectric Power Consumption                         1.00000000
```

Total Biomass Energy Production and Total Renewable Energy Production are highly positively correlated to one another with correlation coefficients of roughly 0.8. Hydroelectric power consumption was much less correlated with either of the other two series, with correlation coefficients varying between roughly -0.25 and 0.1.

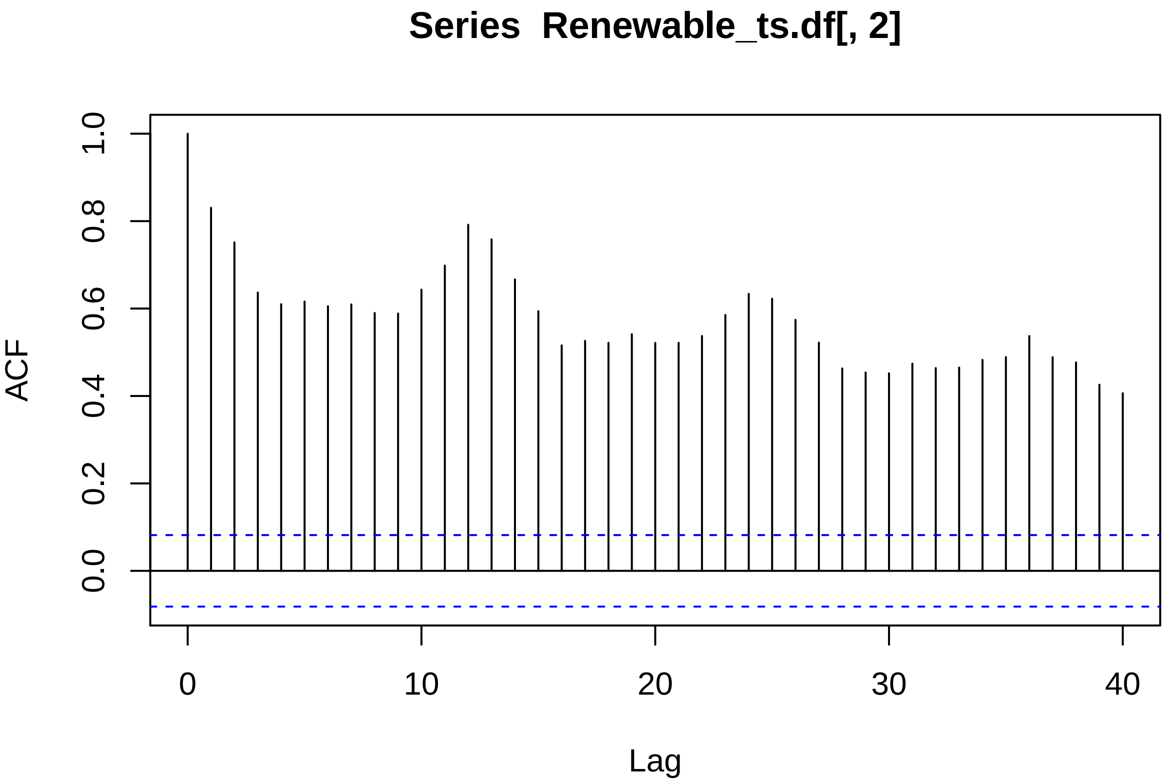
Question 6

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?

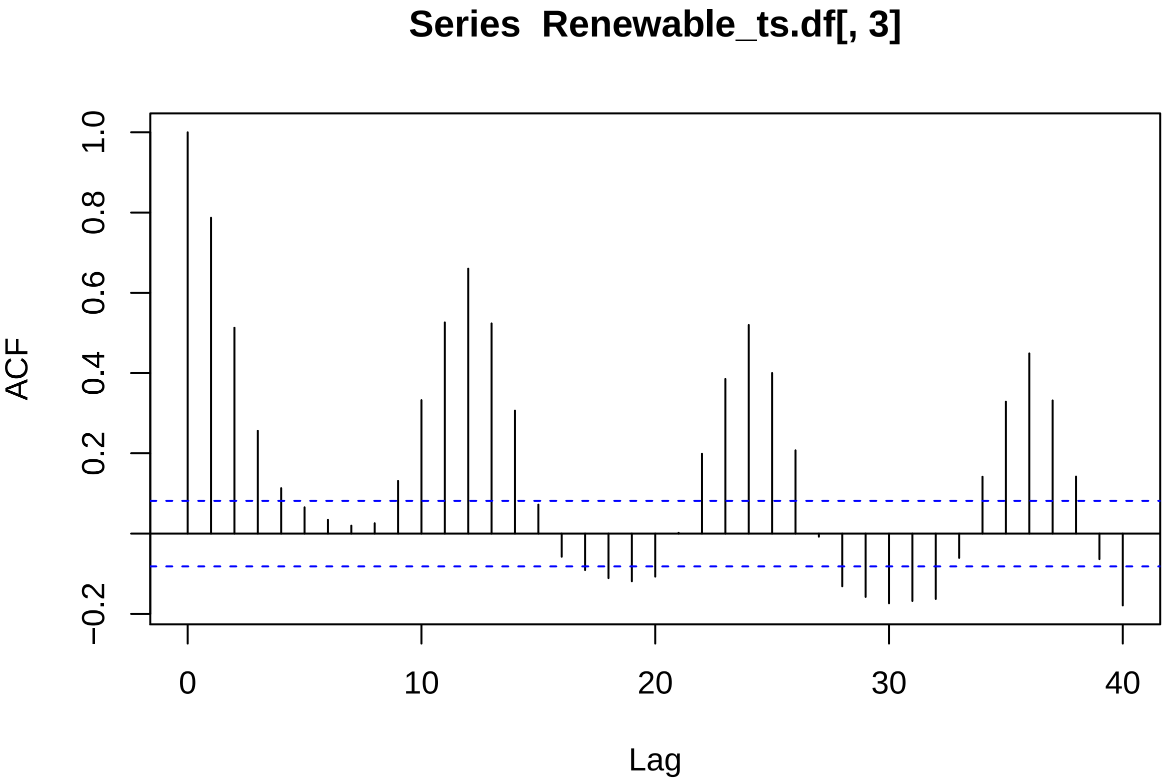
```
acf(Renewable_ts.df[,1], lag.max=40)
```



```r
acf(Renewable_ts.df[,2], lag.max=40)
```



```r
acf(Renewable_ts.df[,3], lag.max=40)
```

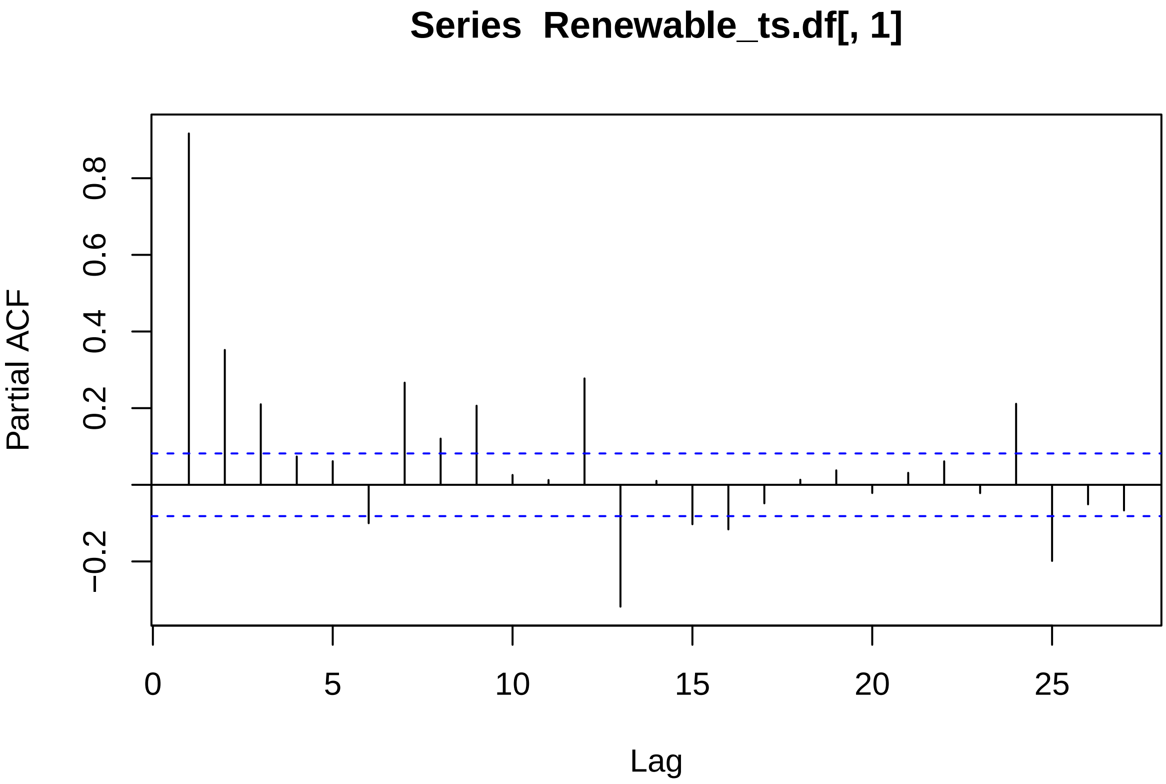


These three plots do not exhibit the same behavior. Both Total Biomass and Total Renewable production plots show a roughly linear decrease in ACF as the lag increases from 1 to 40. There is some seasonality variations but for the most part, this relationship is linear. The Total Hydroelectric Power Consumption on the other hand oscillates significantly as lag increases, with the correlation rising and falling several times as the lag moves from 1 to 40.

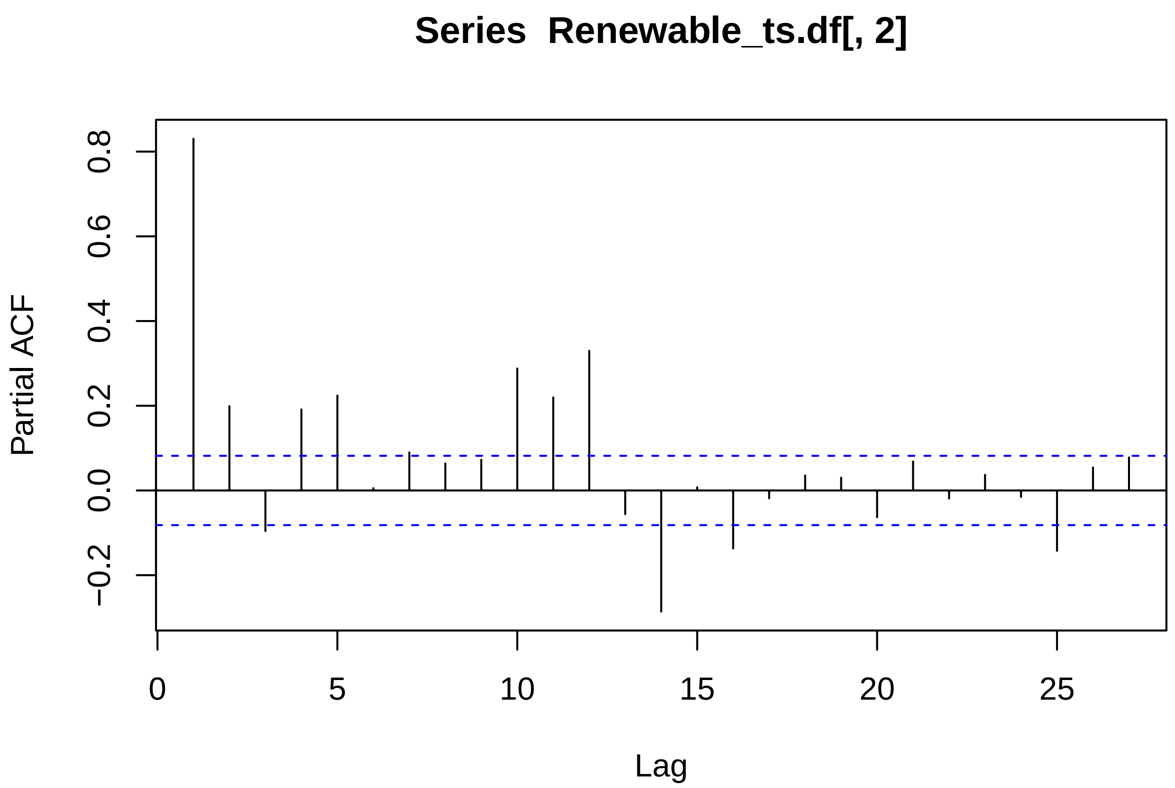
Question 7

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

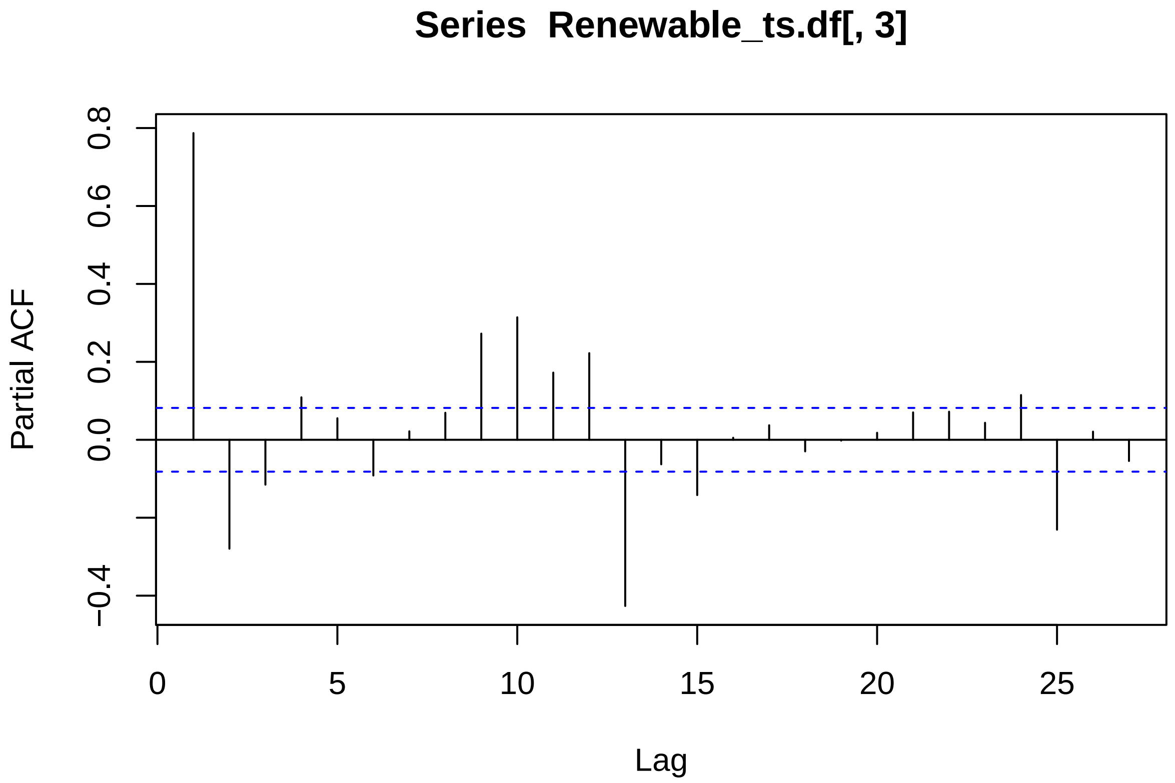
```
pacf(Renewable_ts.df[,1])
```



```
pacf(Renewable_ts.df[,2])
```



```
pacf(Renewable_ts.df[,3])
```



These three plots differ from the ones in question 6 in that they remove the influence of intermediate values in the correlation calculations. Obviously, this will not alter the lag 1 values, but as lag increases to 40, we see much more seasonality in the total Biomass and Total Renewable production correlation plots. The ACF obscured this seasonality with a slight linear decrease in correlation as lag increased, but these PACF plots clearly show oscillations in the correlations over all three of our series.