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

Assignment 2 - Due date 02/05/21

Nick Valby

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.\

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx" 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

energy_data <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx", co
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...
```

```
#Organizing the data
energy_data_names <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xls
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...
```

```
#Naming the columns
colnames(energy_data) <- energy_data_names
```

```
head(energy_data)
```

```
## # A tibble: 6 x 14
##   Month               `Wood Energy Pr~ `Biofuels Produ~ `Total Biomass ~
##   <dttm>                         <dbl> <chr>                        <dbl>
## 1 1973-01-01 00:00:00             130. Not Available                 130.
## 2 1973-02-01 00:00:00             117. Not Available                 117.
## 3 1973-03-01 00:00:00             130. Not Available                 130.
## 4 1973-04-01 00:00:00             125. Not Available                 126.
## 5 1973-05-01 00:00:00             130. Not Available                 130.
## 6 1973-06-01 00:00:00             125. Not Available                 126.
## # ... with 10 more variables: `Total Renewable Energy Production` <dbl>,
## #   `Hydroelectric Power Consumption` <dbl>, `Geothermal Energy
## #   Consumption` <dbl>, `Solar Energy Consumption` <chr>, `Wind Energy
## #   Consumption` <chr>, `Wood Energy Consumption` <dbl>, `Waste Energy
## #   Consumption` <dbl>, `Biofuels Consumption` <chr>, `Total Biomass Energy
## #   Consumption` <dbl>, `Total Renewable Energy Consumption` <dbl>
```

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.

Here is the data frame structure with the month/year and three columns:

```
energy_data <- select(energy_data, "Month", "Total Biomass Energy Production", "Total Renewable Energy Produc

head(energy_data)
```

```
## # A tibble: 6 x 4
##   Month               `Total Biomass Ene~ `Total Renewable E~ `Hydroelectric Po~
##   <dttm>                            <dbl>               <dbl>              <dbl>
## 1 1973-01-01 00:00:00                130.                404.               273.
## 2 1973-02-01 00:00:00                117.                361.               242.
## 3 1973-03-01 00:00:00                130.                400.               269.
## 4 1973-04-01 00:00:00                126.                380.               253.
## 5 1973-05-01 00:00:00                130.                392.               261.
## 6 1973-06-01 00:00:00                126.                377.               250.
```

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().

Here is the data frame in a time series format starting in January 1973:

```
#Transforming the data into a time series:
ts_energy_data <- ts(energy_data[,2:4],frequency = 12, start = 1973)

#Here is the data in a time series:
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.

Means of the three time series:

Total Biomass Energy Production: 270.6961324

Total Renewable Energy Production: 572.7320871

Hydroelectric Power Consumption: 236.9515418

Standard deviations of the three time series:

Total Biomass Energy Production: 87.3631136

Total Renewable Energy Production: 168.4587741

Hydroelectric Power Consumption: 43.9039151

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

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

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

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

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

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

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

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

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

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

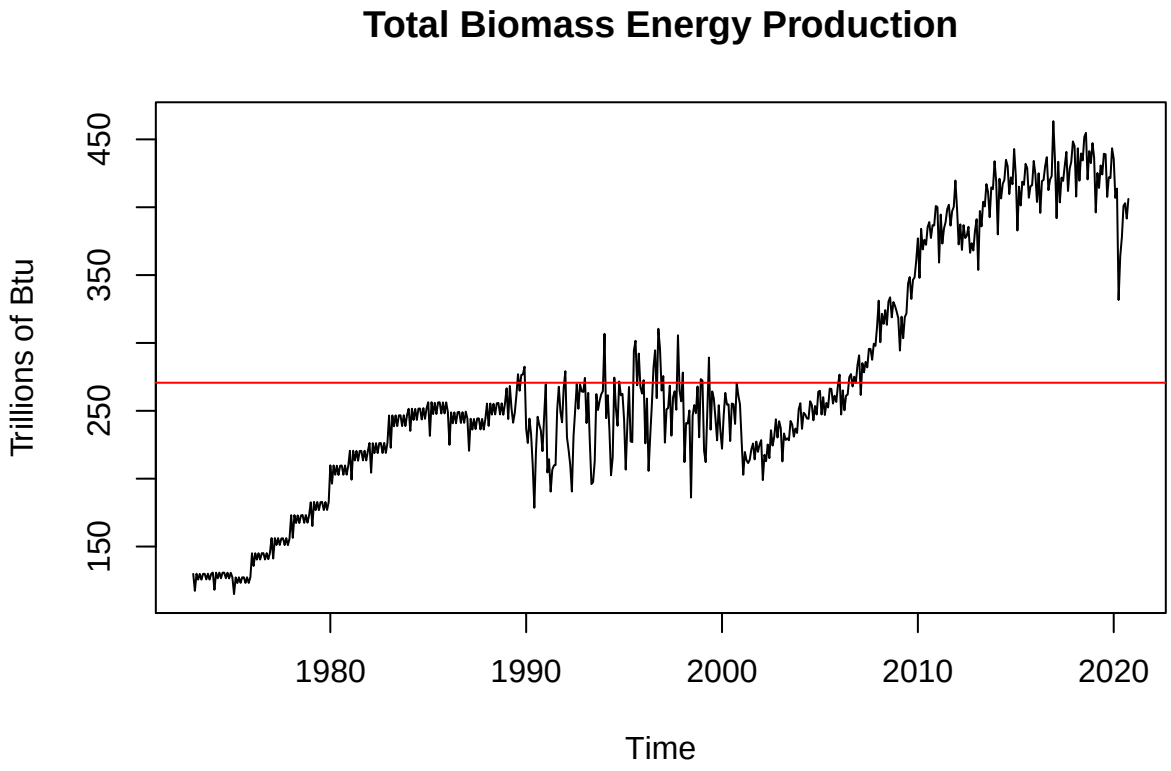
```
sd(ts_energy_data[,3])
```

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

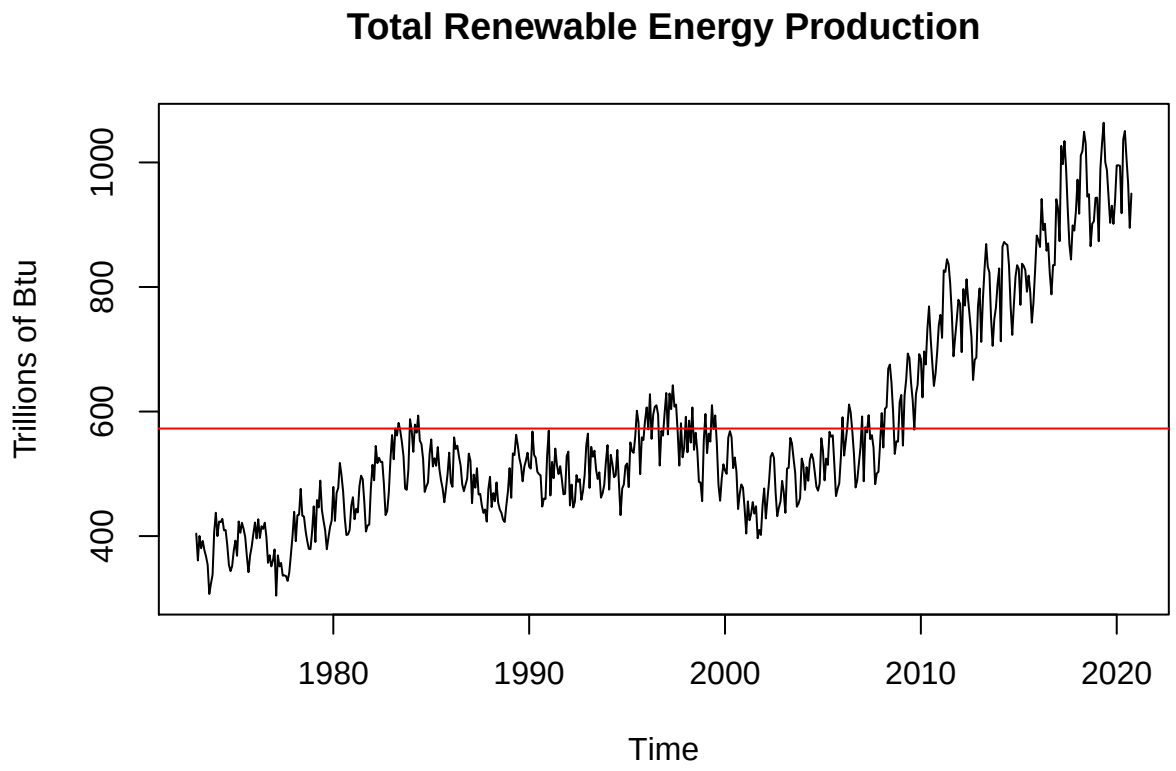
Question 4

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.

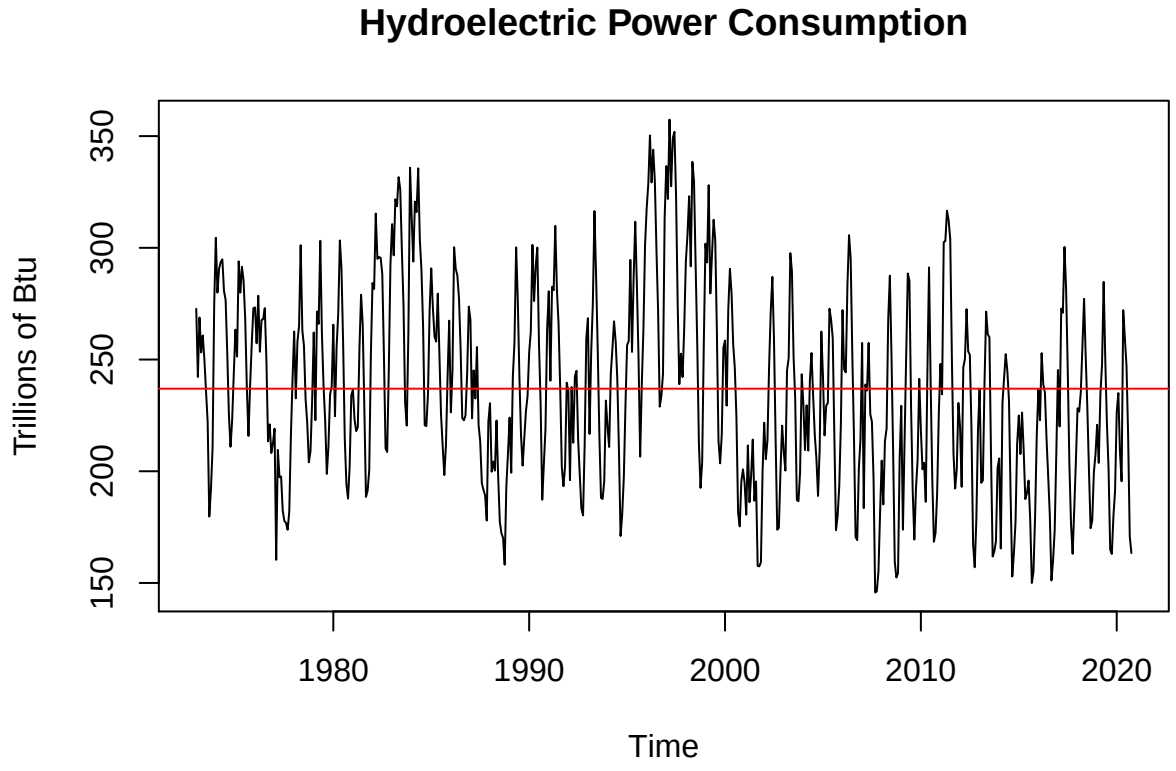
```
plot(ts_energy_data[,1],
     ylab = "Trillions of Btu",
     main = "Total Biomass Energy Production",
     type = "l")
abline(h=270.69, col = "red")
```




```
plot(ts_energy_data[,2],
     ylab = "Trillions of Btu",
     main = "Total Renewable Energy Production",
     type = "l")
abline(h=572.73, col = "red")
```




```
plot(ts_energy_data[,3],
     ylab = "Trillions of Btu",
     main = "Hydroelectric Power Consumption",
     type = "l")
abline(h=236.95, col = "red")
```



Question 5

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

The correlation between Total Biomass Energy Production and Total Renewable Energy Production is 0.9234609 which means the two variables are positively correlated

The correlation between Total Renewable Energy Production and Hydroelectric Power Consumption is -0.0027569 which means the variables are negatively correlated

The correlation between Hydroelectric Power Consumption and Total Biomass Energy Production is -0.2555675 which means the variables are weakly correlated in a negative direction

```
cor(ts_energy_data[,1],ts_energy_data[,2])
```

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

```
cor(ts_energy_data[,2],ts_energy_data[,3])
```

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

```
cor(ts_energy_data[,3],ts_energy_data[,1])
```

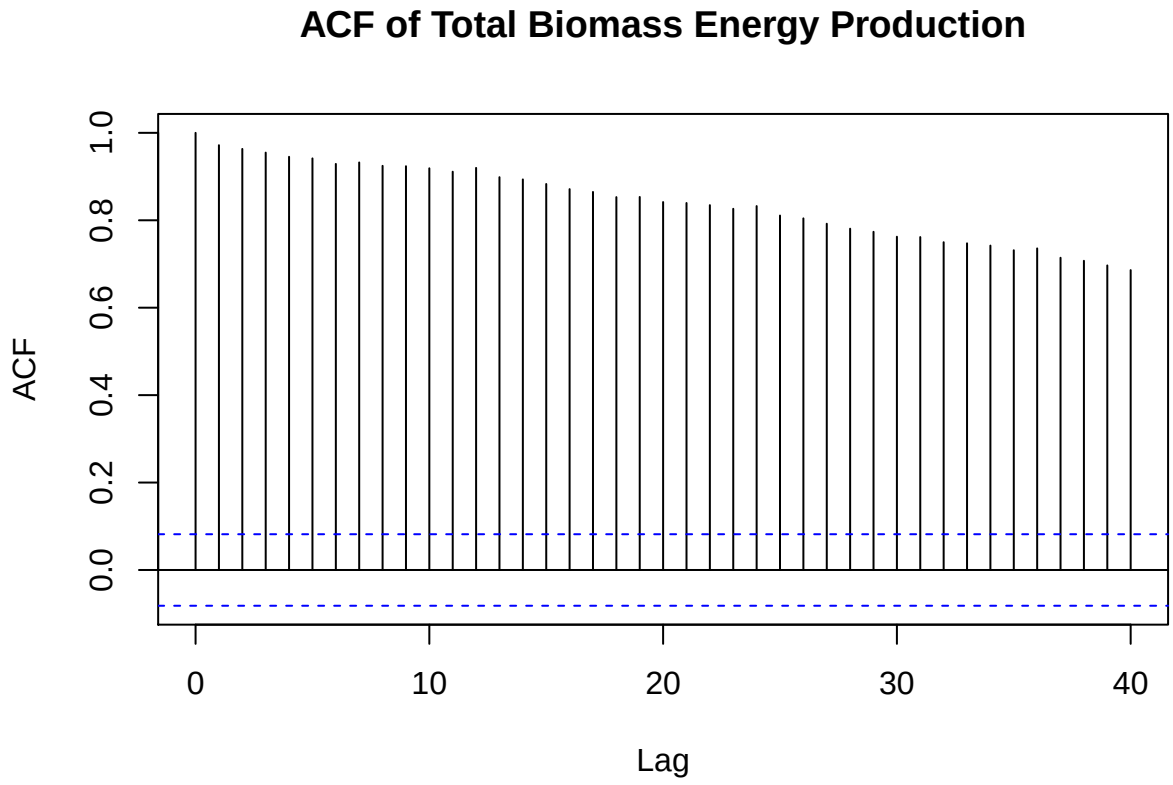
```
## [1] -0.2555675
```

Question 6

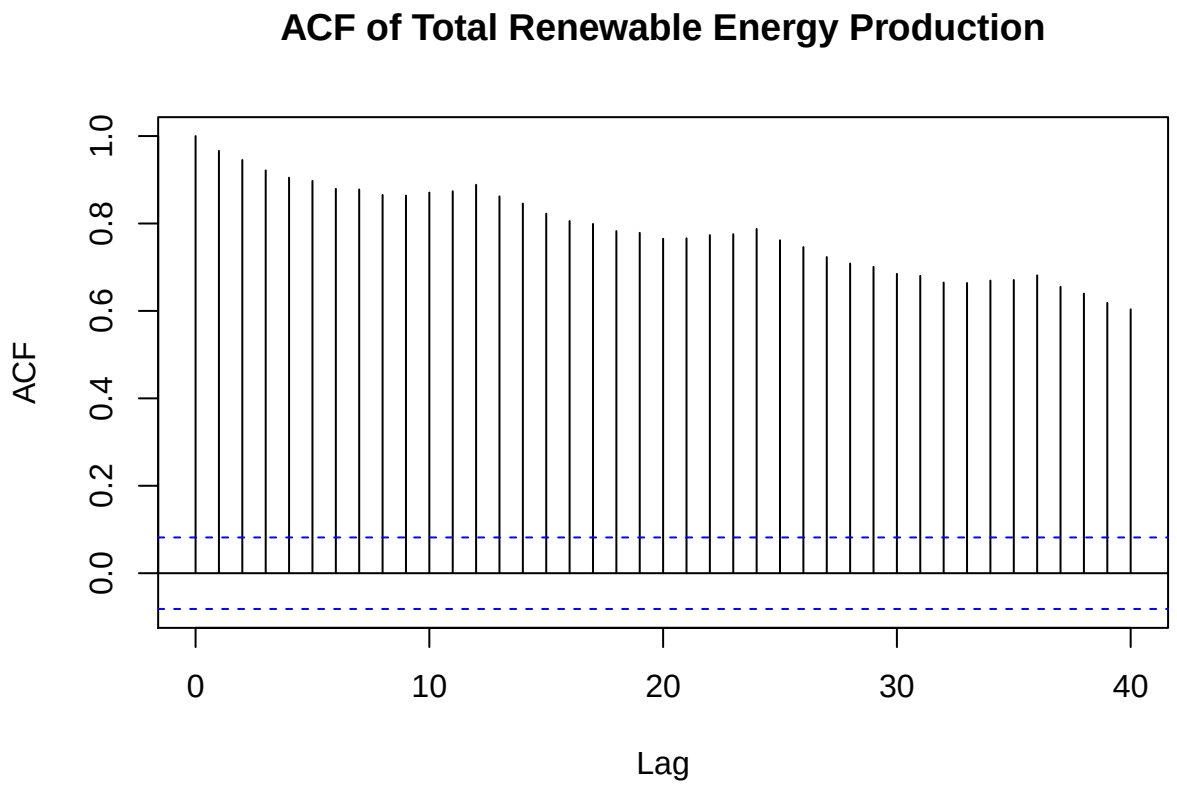
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

The plots below do not have the same behavior. The first two have similar downward trends over time, while the third is obviously seasonal/cyclical.

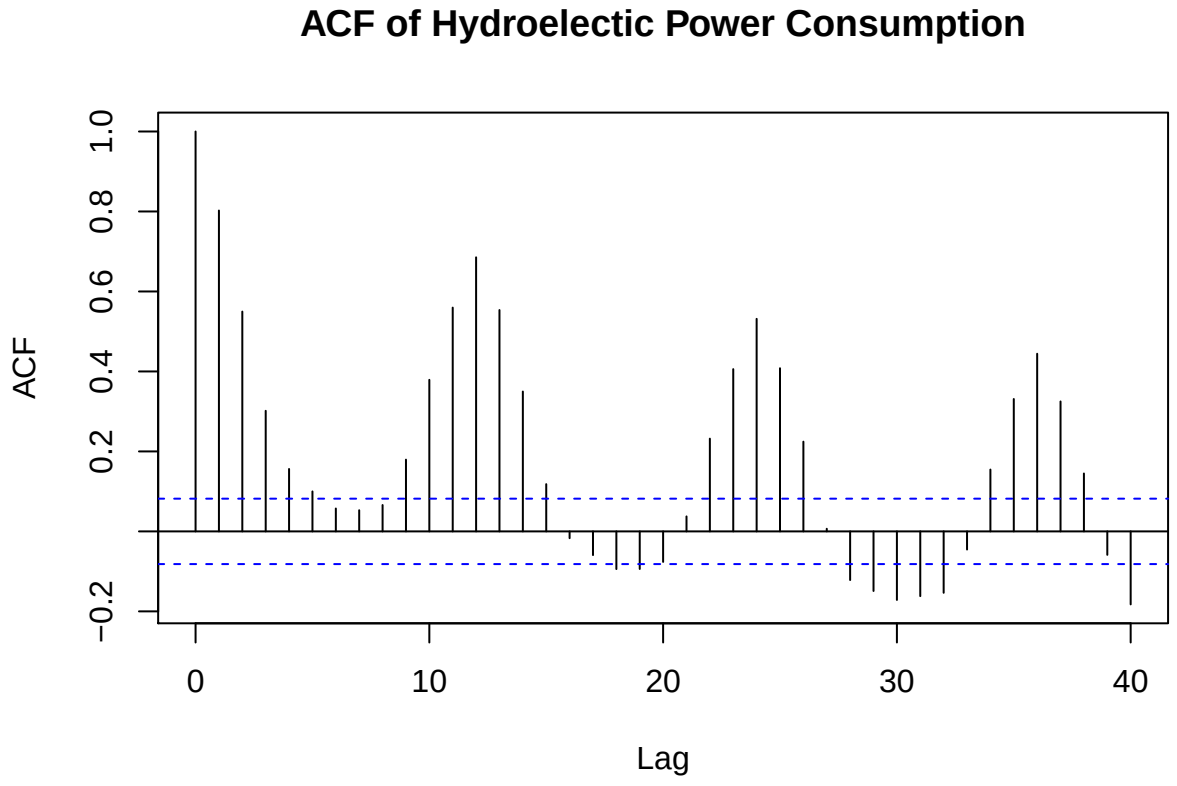
```
acf(energy_data[,2], lag.max = 40, main = "ACF of Total Biomass Energy Production")
```




```
acf(energy_data[,3], lag.max = 40, main = "ACF of Total Renewable Energy Production")
```




```
acf(energy_data[,4], lag.max = 40, main = "ACF of Hydroelectic Power Consumption")
```

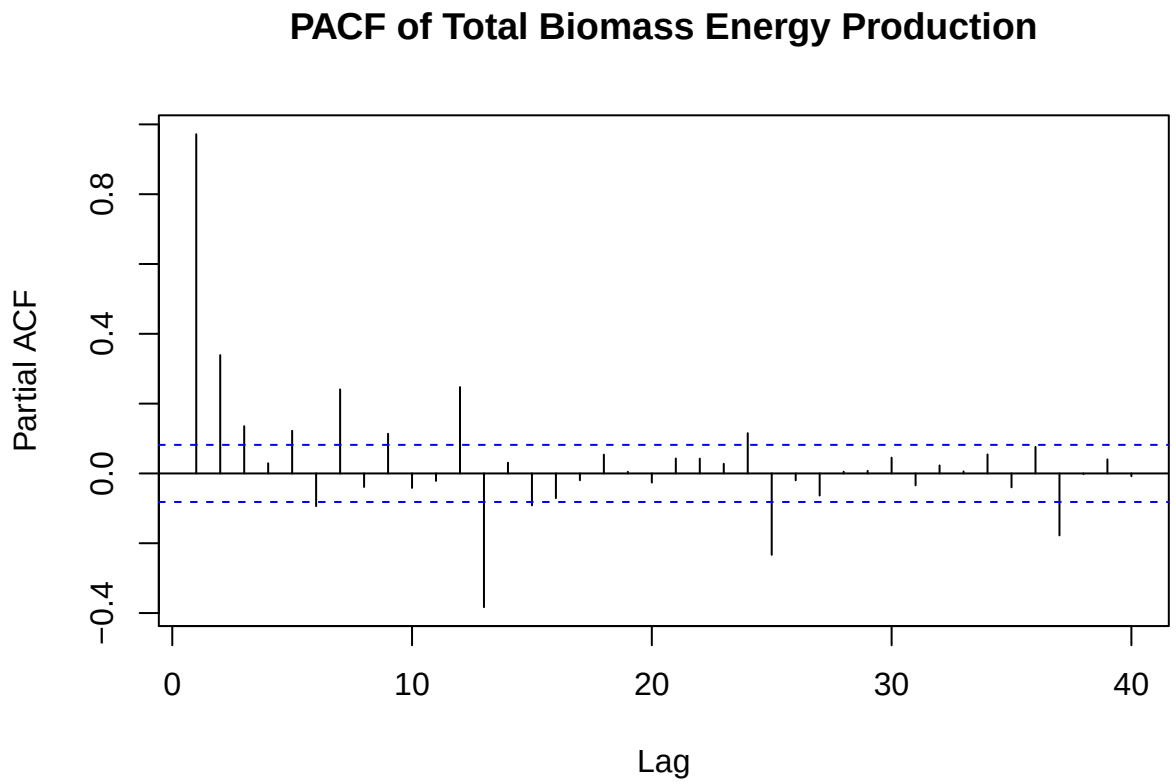


Question 7

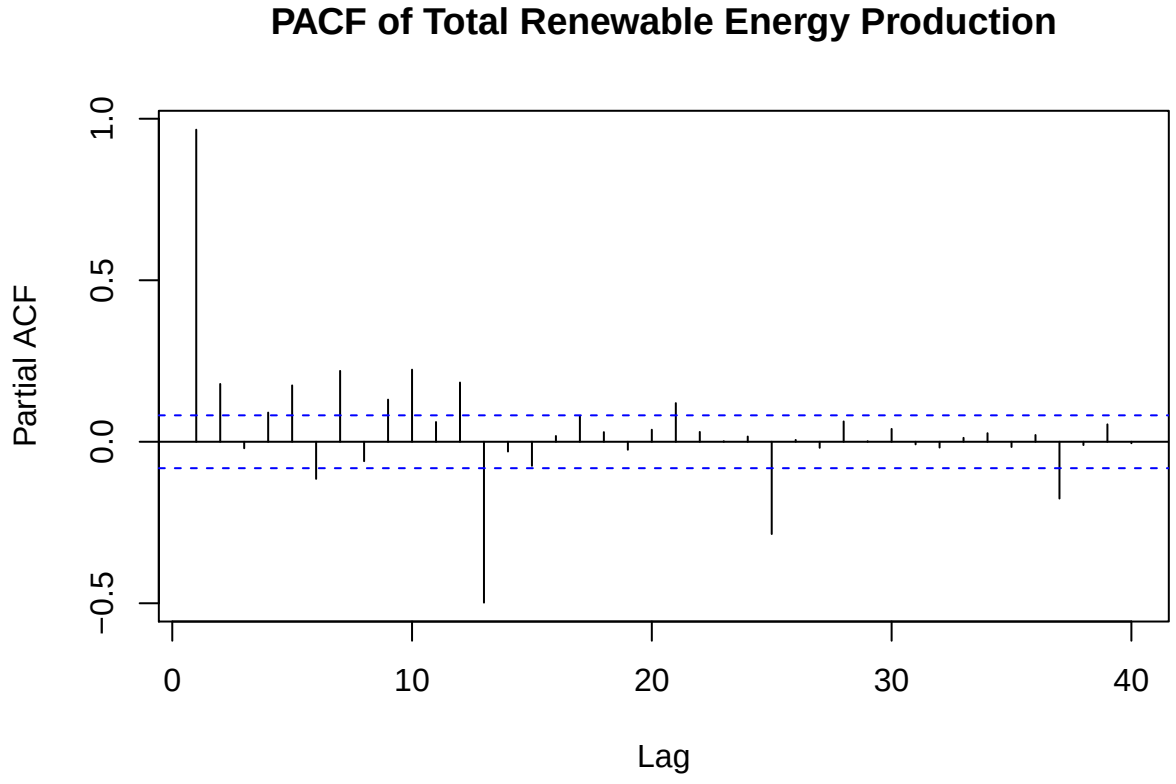
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

The plots below are different from the ones in Q6 because these plots represent the difference between between a variable at two different times WITH the "white noise" removed.

```
pacf(energy_data[,2], lag.max = 40, main = "PACF of Total Biomass Energy Production")
```




```
pacf(energy_data[,3], lag.max = 40, main = "PACF of Total Renewable Energy Production")
```



```r
pacf(energy_data[,4], lag.max = 40, main = "PACF of Hydroelectic Power Consumption")
```