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

Assignment 2 - Due date 01/26/22

Aasha Reddy

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_Sp22.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(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.1.2
```

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

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.1.2
```

```
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
repc <- read_excel("/Users/Aasha Reddy/Documents/Statistics - Duke University/2022 Spring/Time Series An
repc <- repc[-1,]
```

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.

```
repc <- repc %>%
  select(`Total Biomass Energy Production`,
         `Total Renewable Energy Production`,
         `Hydroelectric Power Consumption`)

head(repc)
```

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

```
# check data types
apply(repc, 2, typeof)
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                         "character"                         "character"
##   Hydroelectric Power Consumption
##                         "character"
```

```
# change variables to numeric

repc <- repc %>%
  mutate(`Total Biomass Energy Production` = as.numeric(`Total Biomass Energy Production`),
         `Total Renewable Energy Production` = as.numeric(`Total Renewable Energy Production`),
         `Hydroelectric Power Consumption` = as.numeric(`Hydroelectric Power Consumption`))
```

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().

```
repc <- ts(data = repc, start = c(1973, 1), frequency = 12)
```

```
is.ts(repc)
```

```
## [1] TRUE
```

```
head(repc)
```

```
##          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.

Total Biomass Energy Production:

```
# Total Biomass Energy Production
paste0("mean: ", mean(repc[,"Total Biomass Energy Production"]))
```

```
## [1] "mean: 273.783924786325"
```

```
paste0("sd: ", sd(repc[,"Total Biomass Energy Production"]))
```

```
## [1] "sd: 89.4285220898559"
```

Total Renewable Energy Production:

```
# Total Renewable Energy Production
paste0("mean: ", mean(repc[,"Total Renewable Energy Production"]))
```

```
## [1] "mean: 581.170830769231"
```

```
paste0("sd: ", sd(repc[,"Total Renewable Energy Production"]))
```

```
## [1] "sd: 177.560723786192"
```

Hydroelectric Power Consumption:

```
# Hydroelectric Power Consumption
paste0("mean: ", mean(repc[,"Hydroelectric Power Consumption"]))
```

```
## [1] "mean: 235.96525982906"
```

```
paste0("sd: ", sd(repc[,"Hydroelectric Power Consumption"]))
```

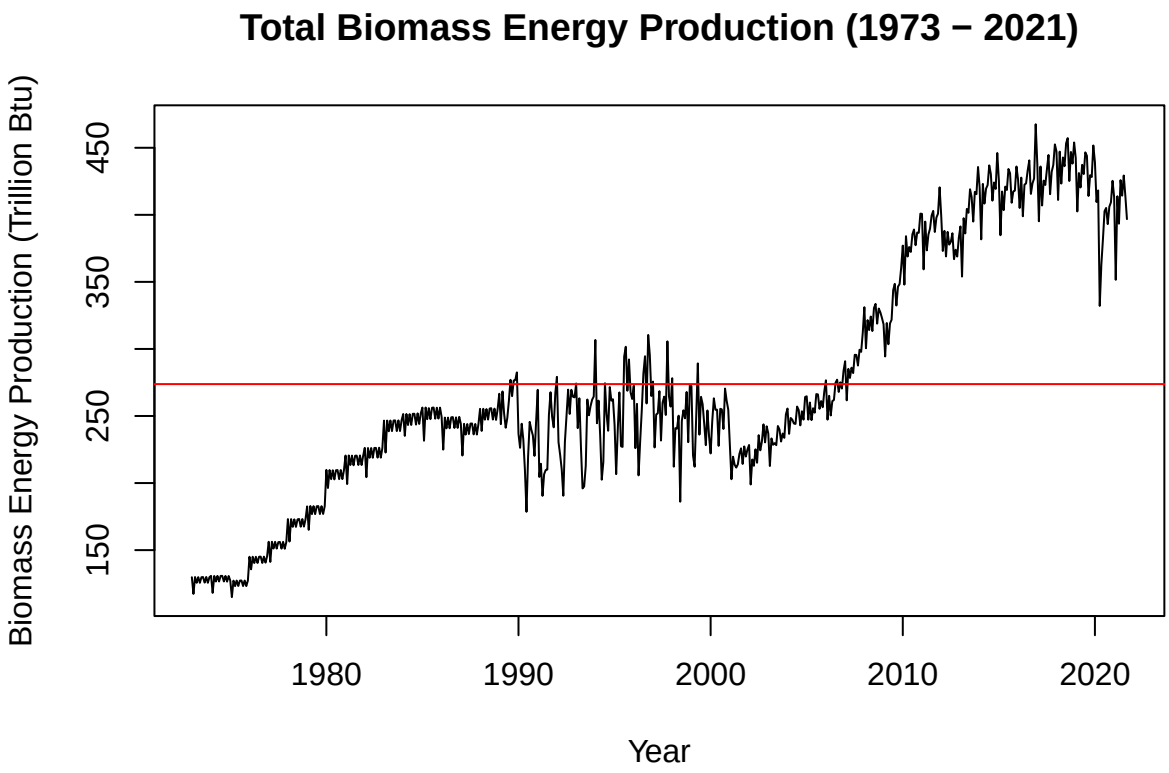
```
## [1] "sd: 44.0174921003646"
```

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.

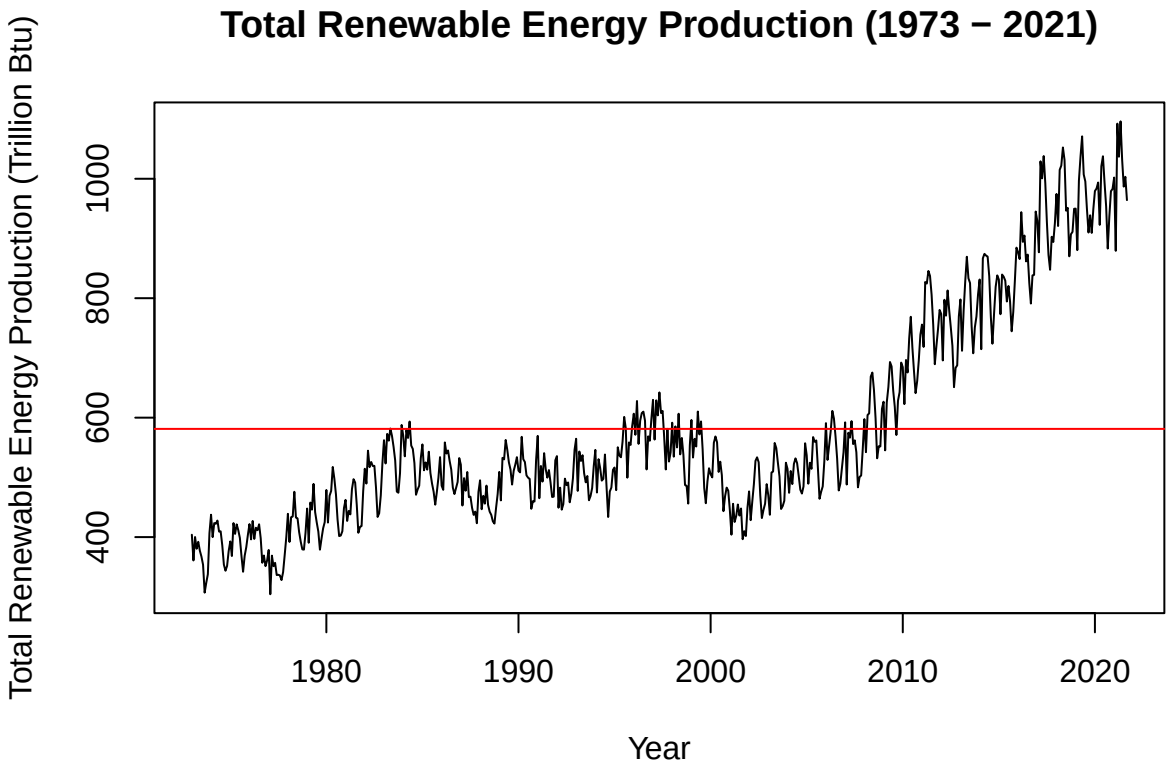
Total Biomass Energy Production:

```
# Total Biomass Energy Production
plot(repc[, "Total Biomass Energy Production"],
     main = "Total Biomass Energy Production (1973 - 2021)",
     ylab = "Biomass Energy Production (Trillion Btu)",
     xlab = "Year")
abline(h = mean(repc[,"Total Biomass Energy Production"]), col = "red")
```



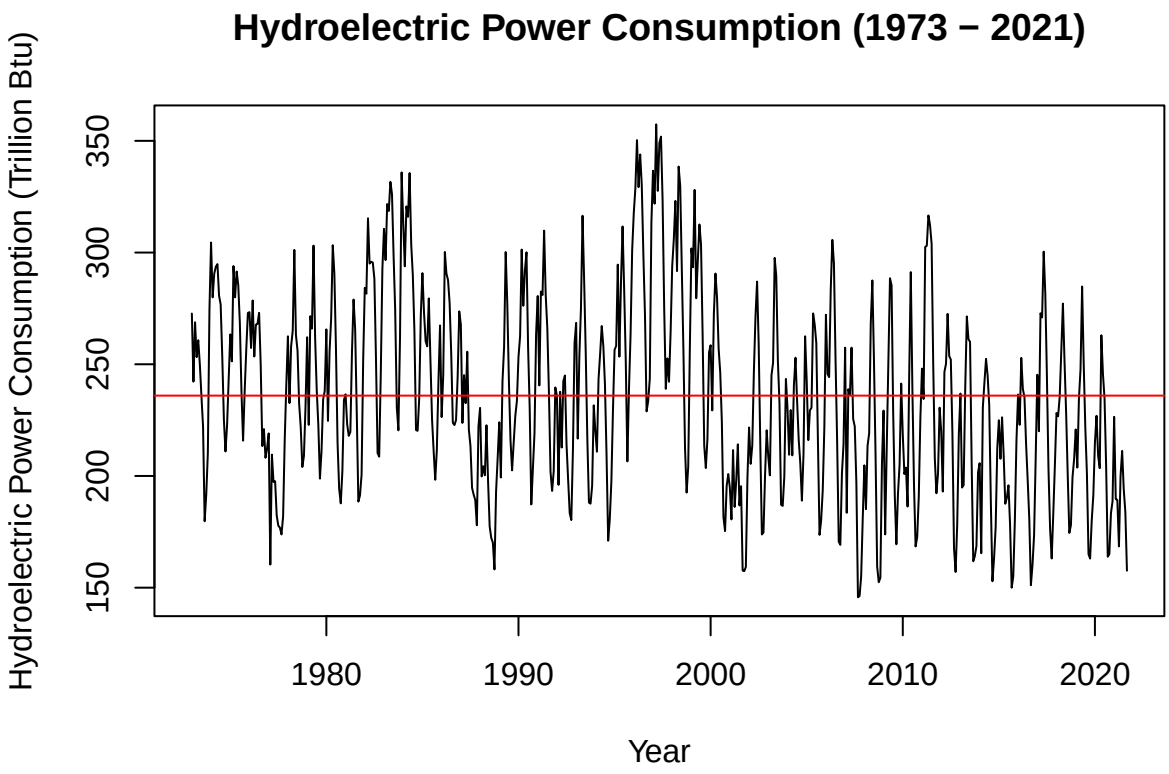
Total Renewable Energy Production:

```
# Total Renewable Energy Production
plot(repc[, "Total Renewable Energy Production"],
     main = "Total Renewable Energy Production (1973 - 2021)",
     ylab = "Total Renewable Energy Production (Trillion Btu)",
     xlab = "Year")
abline(h = mean(repc[,"Total Renewable Energy Production"]), col = "red")
```



Hydroelectric Power Consumption:

```
# Hydroelectric Power Consumption
plot(repc[, "Hydroelectric Power Consumption"],
     main = "Hydroelectric Power Consumption (1973 - 2021)",
     ylab = "Hydroelectric Power Consumption (Trillion Btu)",
     xlab = "Year")
abline(h = mean(repc[,"Hydroelectric Power Consumption"]), col = "red")
```



Question 5

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

From the below, we can see that only Total Renewable Energy Production and Total Biomass Energy are significantly positively correlated. This means that as the value of one variable increases, the other tends to increase as well. Both of those time series correlation with Hydroelectric Power Consumption were negatively, but insignificantly correlated. The negative correlation indicates that as the value of one variable increases the value of the other tends to decrease.

Hydroelectric Power Consumption and Total Renewable Energy Production

```
cor(repc[,"Hydroelectric Power Consumption"], repc[,"Total Renewable Energy Production"])
```

```
## [1] -0.05680651
```

Hydroelectric Power Consumption and Total Biomass Energy Production

```
cor(repc[,"Hydroelectric Power Consumption"], repc[,"Total Biomass Energy Production"])
```

```
## [1] -0.2804997
```

Total Renewable Energy Production and Total Biomass Energy Production

```
cor(repc[,"Total Renewable Energy Production"], repc[,"Total Biomass Energy Production"])
```

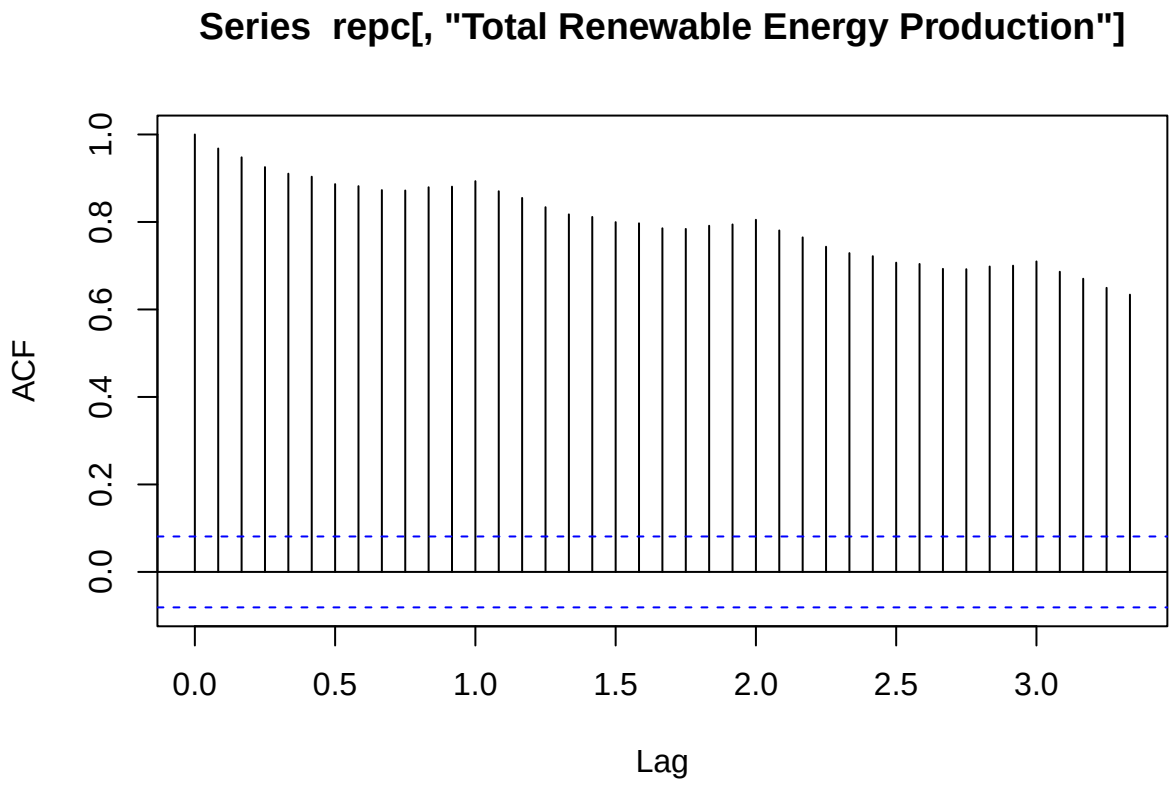
```
## [1] 0.9232838
```

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?

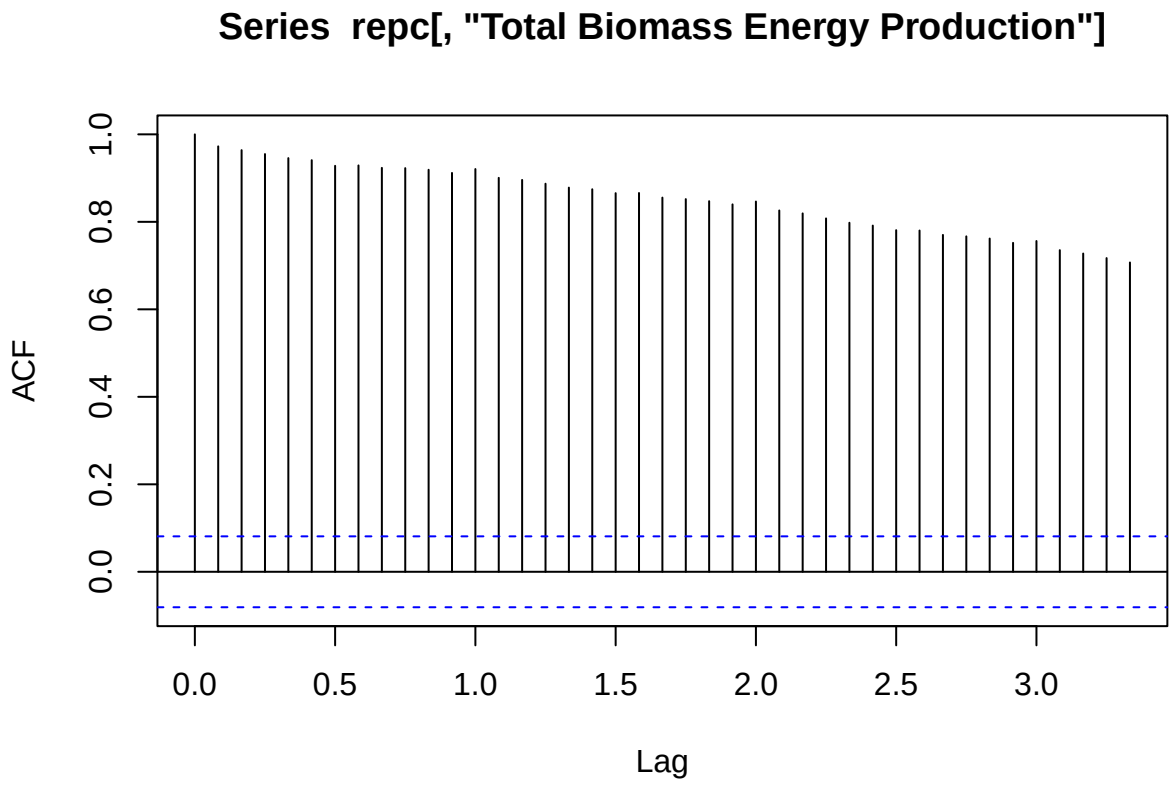
Total Renewable Energy Production:

```
acf(repc[,"Total Renewable Energy Production"], lag = 40)
```



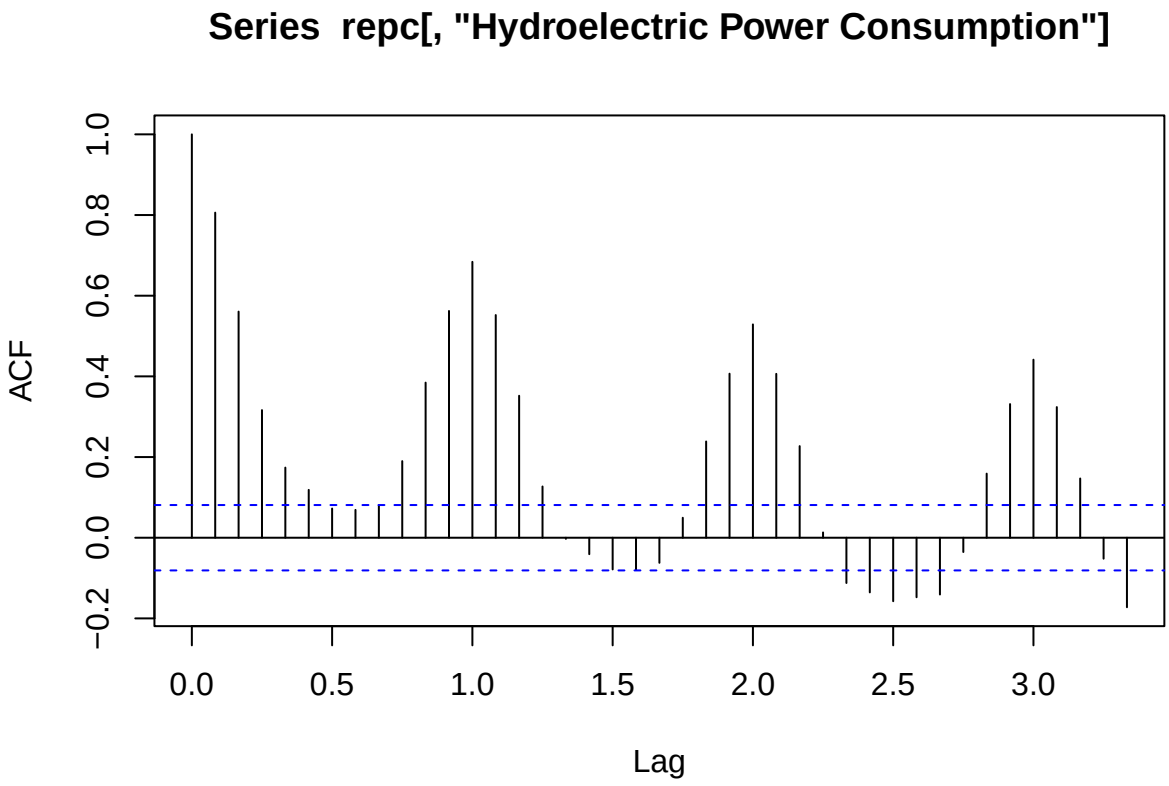
Total Biomass Energy Production:

```
acf(repc[,"Total Biomass Energy Production"], lag = 40)
```



Hydroelectric Power Consumption:

```
acf(repc[,"Hydroelectric Power Consumption"], lag = 40)
```



For both Total Renewable Energy Production and Total Biomass Energy Production have high correlation between time periods. As expected, for these plots the ACF is decaying as we increase the lag.

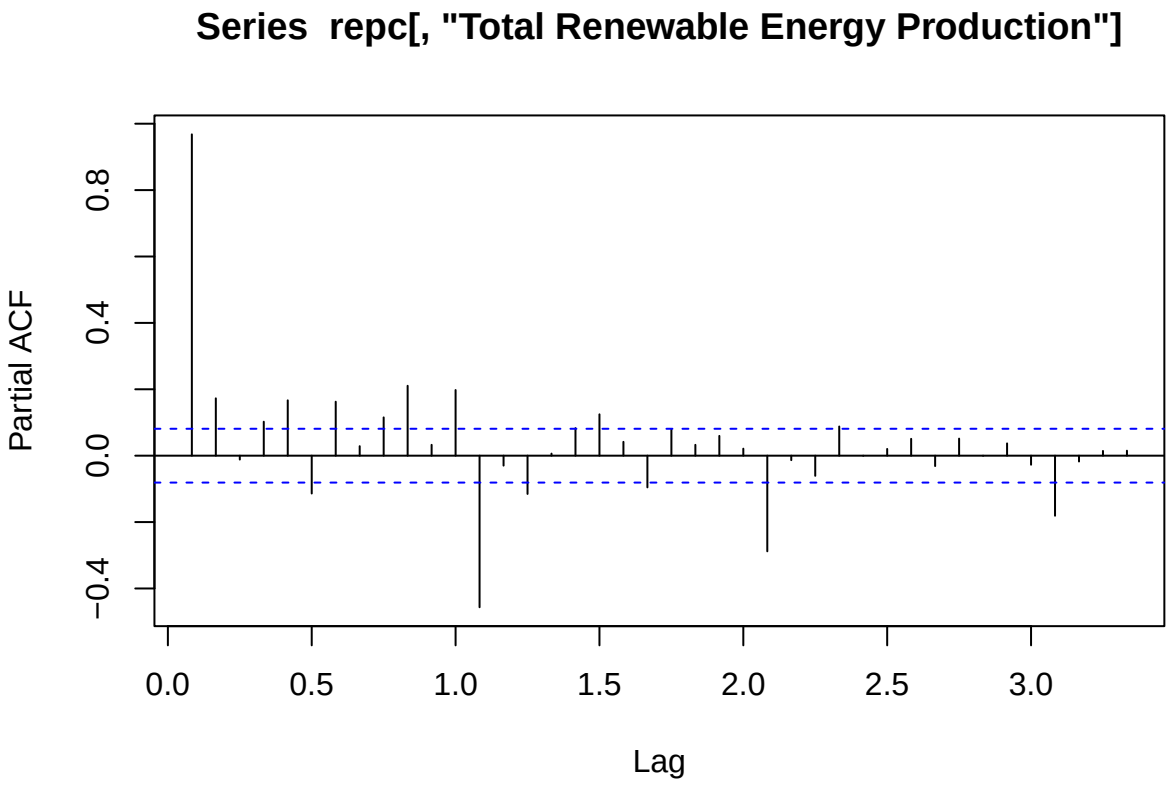
However, the plot for Hydroelection Power Consumption exhibits a different behavior - the ACF is increasing and decreasing, and this is due to the seasonality of the variable.

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?

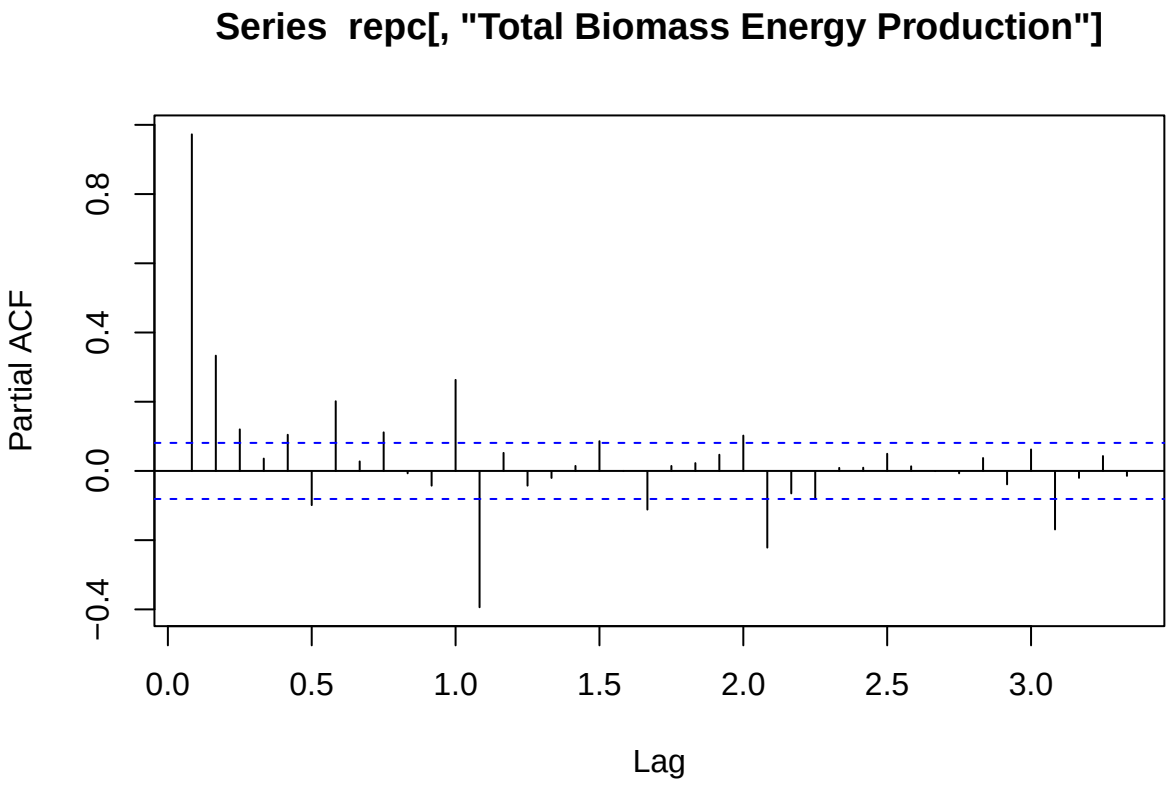
Total Renewable Energy Production:

```
pacf(repc[,"Total Renewable Energy Production"], lag = 40)
```



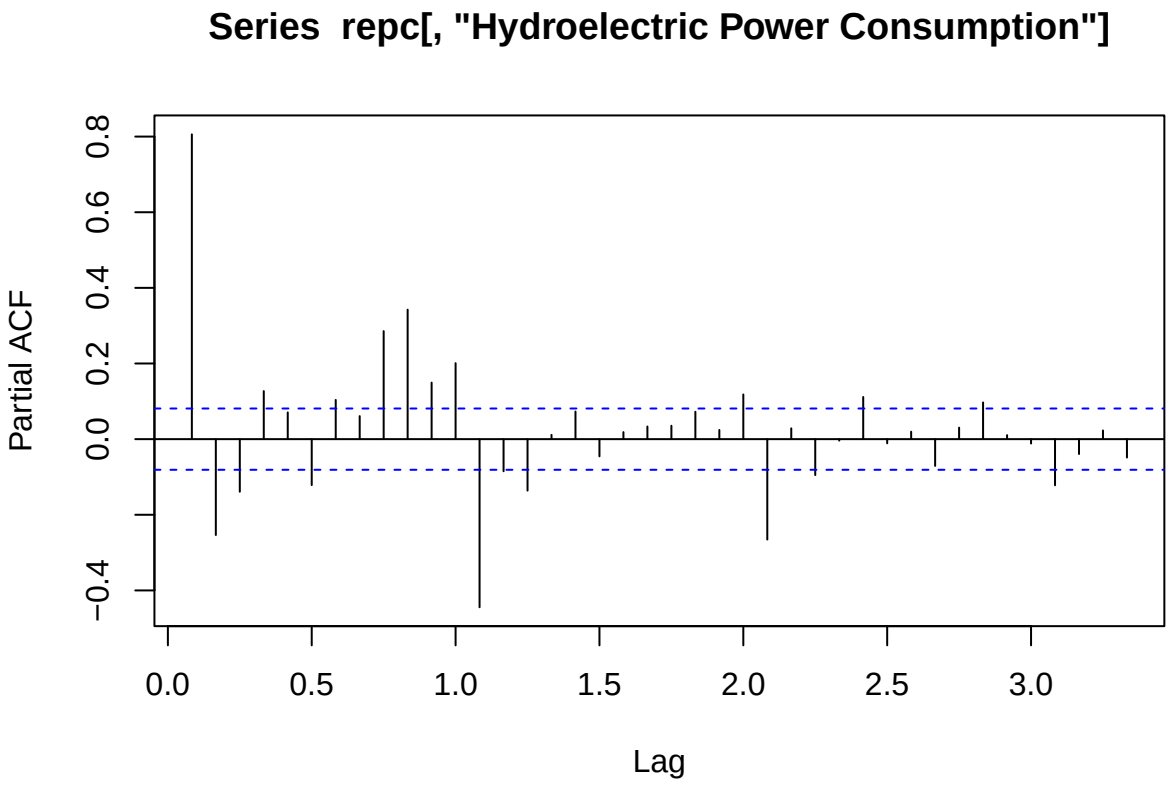
Total Biomass Energy Production:

```
pacf(repc[,"Total Biomass Energy Production"], lag = 40)
```



Hydroelectric Power Consumption:

```
pacf(repc[,"Hydroelectric Power Consumption"], lag = 40)
```



We can see that the PACF plots look different from the ACF plots, they also show smaller correlations as we would expect. The PACF plots look different because they are removing the intermediate dependence.