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

Assignment 2 - Due date 01/26/22

Emre Yurtbay

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(tseries)
```

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

```
library(forecast)
```

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

```
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
tab <- read_excel("/Users/emreyurtbay/Documents/Duke/env790/ENV790_TimeSeriesAnalysis_Sp2022/Data/Table_
```

```
# remove the first row
tab = tab[-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.

```
df = tab[, 4:6]
```

```
df$`Total Biomass Energy Production` = as.numeric(df$`Total Biomass Energy Production`)
df$`Total Renewable Energy Production` = as.numeric(df$`Total Renewable Energy Production`)
df$`Hydroelectric Power Consumption` = as.numeric(df$`Hydroelectric Power Consumption`)
```

```
head(df)
```

```
## # A tibble: 6 x 3
##   `Total Biomass Energy Production` `Total Renewable Ener~ `Hydroelectric Power~
##                               <dbl>                  <dbl>                  <dbl>
## 1                              130.                   404.                   273.
## 2                              117.                   361.                   242.
## 3                              130.                   400.                   269.
## 4                              126.                   380.                   253.
## 5                              130.                   392.                   261.
## 6                              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().

```
ts_df <- ts(df, frequency = 12, start = c(1973, 1))
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(ts_df[,c("Total Biomass Energy Production")])
```

```
## [1] 273.7839
```

```
mean(ts_df[,c("Total Renewable Energy Production")])
```

```
## [1] 581.1708
```

```
mean(ts_df[,c("Hydroelectric Power Consumption")])
```

```
## [1] 235.9653
```

```
sd(ts_df[,c("Total Biomass Energy Production")])
```

```
## [1] 89.42852
```

```
sd(ts_df[,c("Total Renewable Energy Production")])
```

```
## [1] 177.5607
```

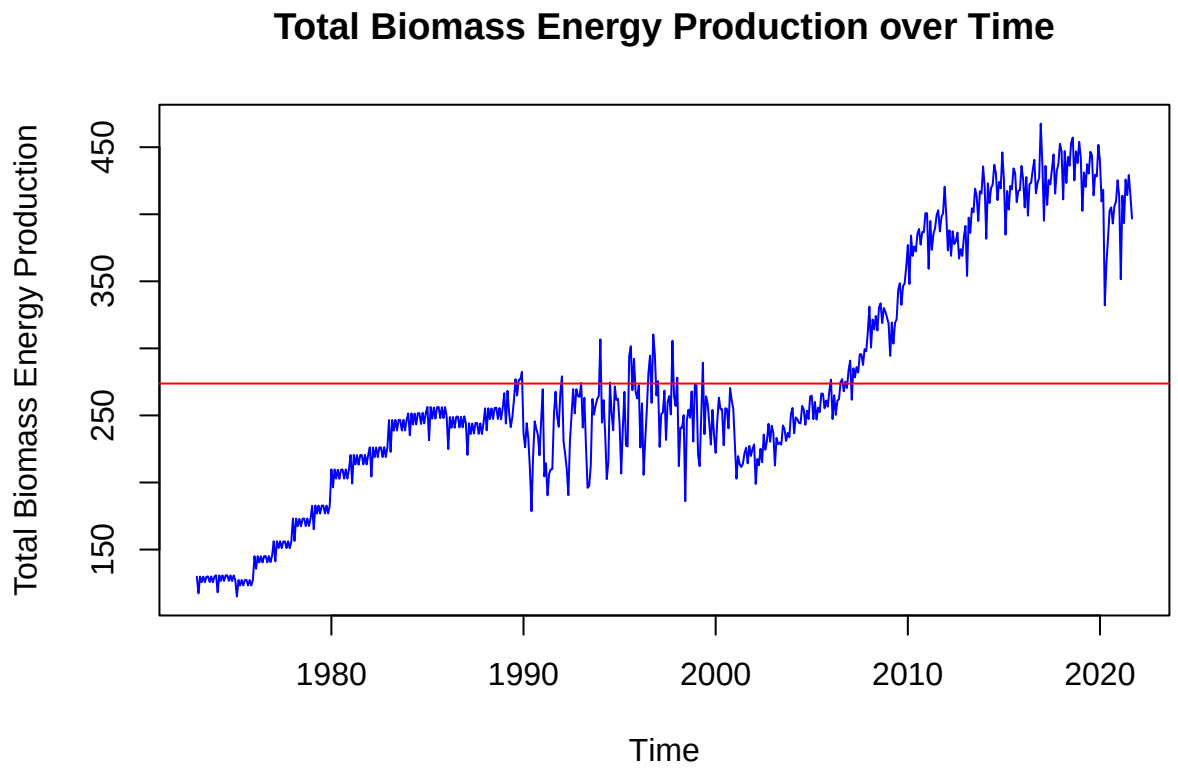
```
sd(ts_df[,c("Hydroelectric Power Consumption")])
```

```
## [1] 44.01749
```

Question 4

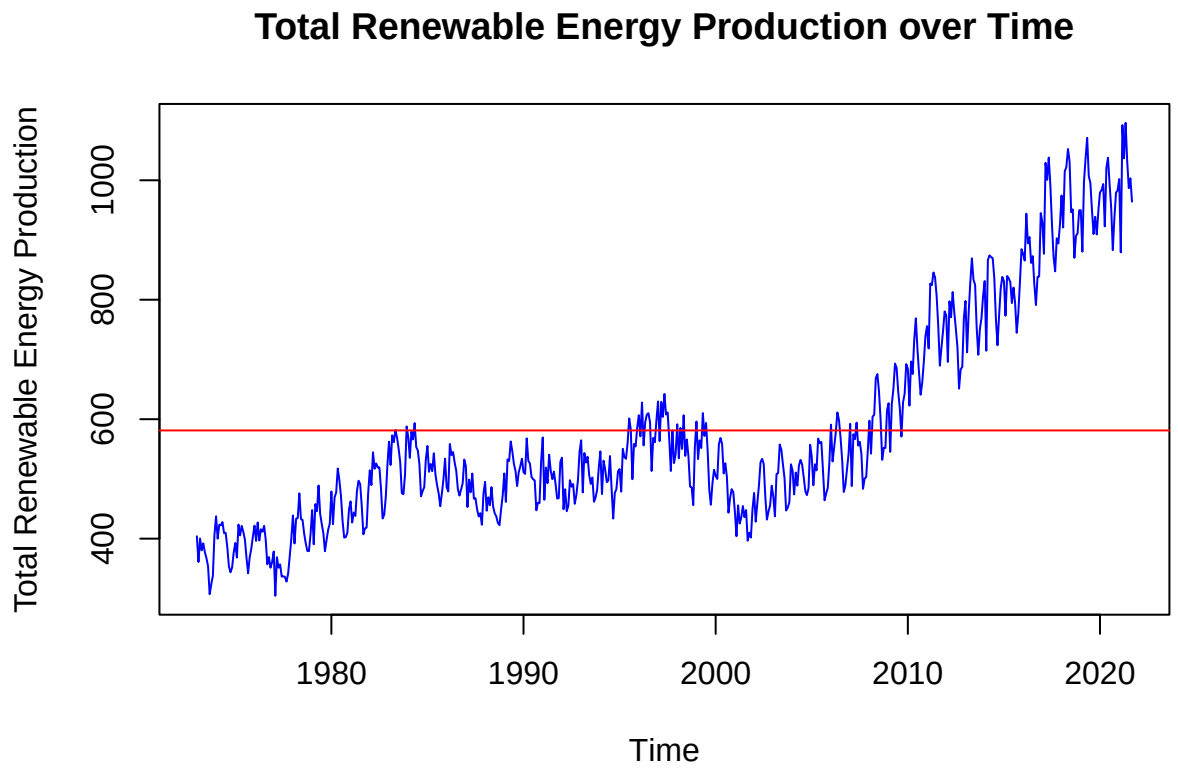
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
plot(ts_df[,c("Total Biomass Energy Production")], type="l", col = "blue",
     main = "Total Biomass Energy Production over Time", xlab = "Time",
     ylab ="Total Biomass Energy Production")
abline(h = mean(ts_df[,c("Total Biomass Energy Production")]), col = "red")
```

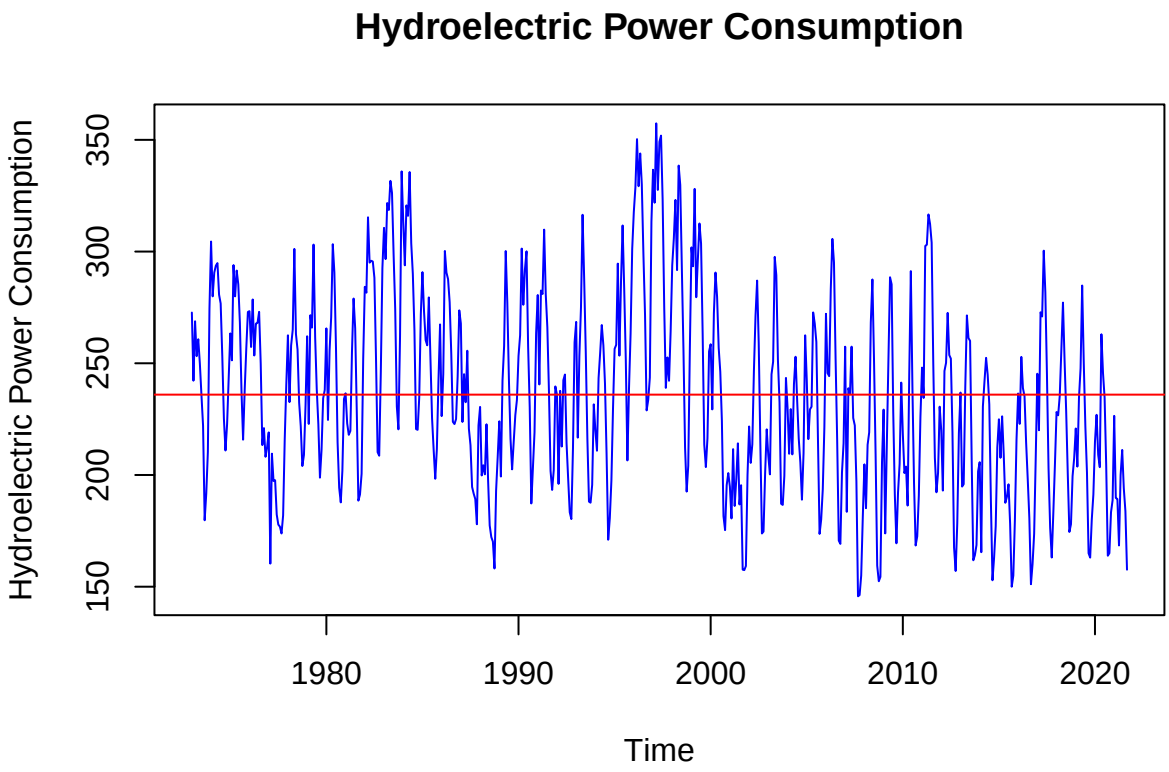



```
plot(ts_df[,c("Total Renewable Energy Production")], type="l", col = "blue",
     main = "Total Renewable Energy Production over Time", xlab = "Time",
     ylab ="Total Renewable Energy Production")
```

```
abline(h = mean(ts_df[,c("Total Renewable Energy Production")]), col = "red")
```




```
plot(ts_df[,c("Hydroelectric Power Consumption")], type="l", col = "blue",
     main = "Hydroelectric Power Consumption", xlab = "Time",
     ylab ="Hydroelectric Power Consumption")
abline(h = mean(ts_df[,c("Hydroelectric Power Consumption")]), col = "red")
```



Question 5

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

```
cor(x = ts_df[,c("Total Biomass Energy Production")],
    y = ts_df[,c("Total Renewable Energy Production")])
```

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

```
cor(x = ts_df[,c("Total Biomass Energy Production")],
    y = ts_df[,c("Hydroelectric Power Consumption")])
```

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

```
cor(x = ts_df[,c("Total Renewable Energy Production")],
    y = ts_df[,c("Hydroelectric Power Consumption")])
```

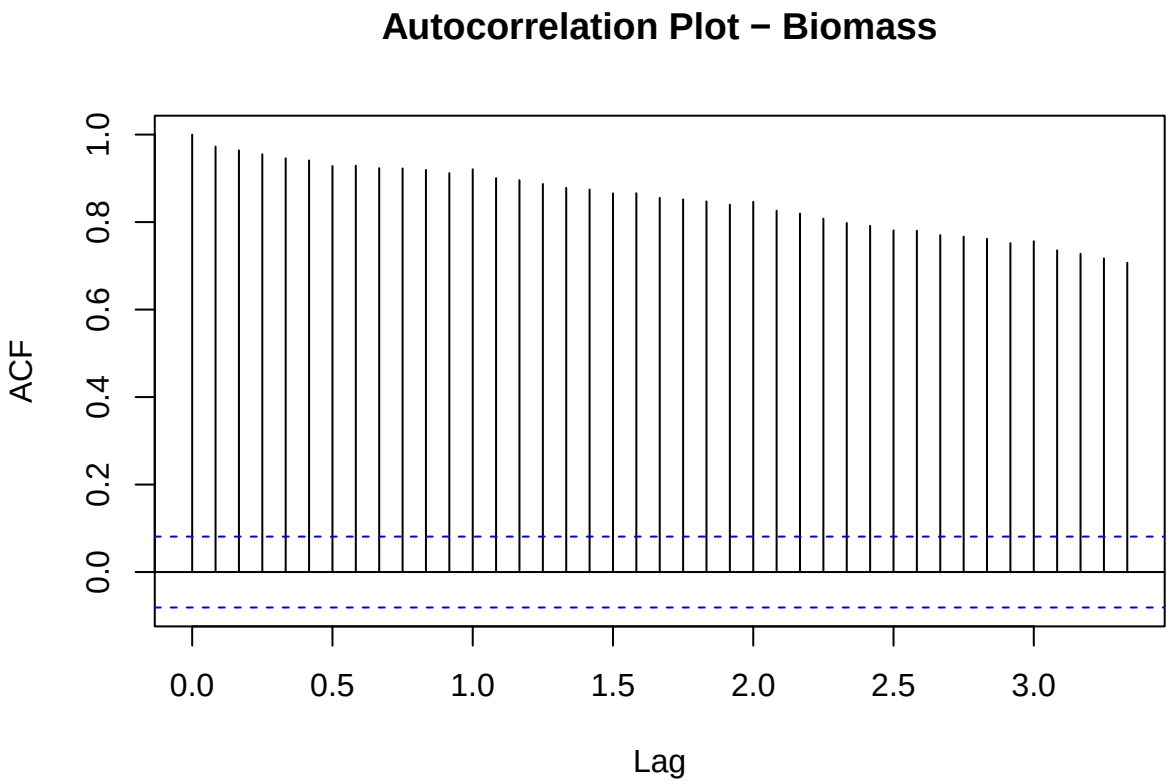
```
## [1] -0.05680651
```

The series "Total Biomass Energy Production" is highly positively correlated with "Total Renewable Energy Production", with a correlation of 0.92. This means as "Total Biomass Energy Production" increases, "Total Renewable Energy Production" increases with it. Neither of the previous 2 series is particularly strongly correlated with "Hydroelectric Power Consumption," as the correlations are fairly close to 0.

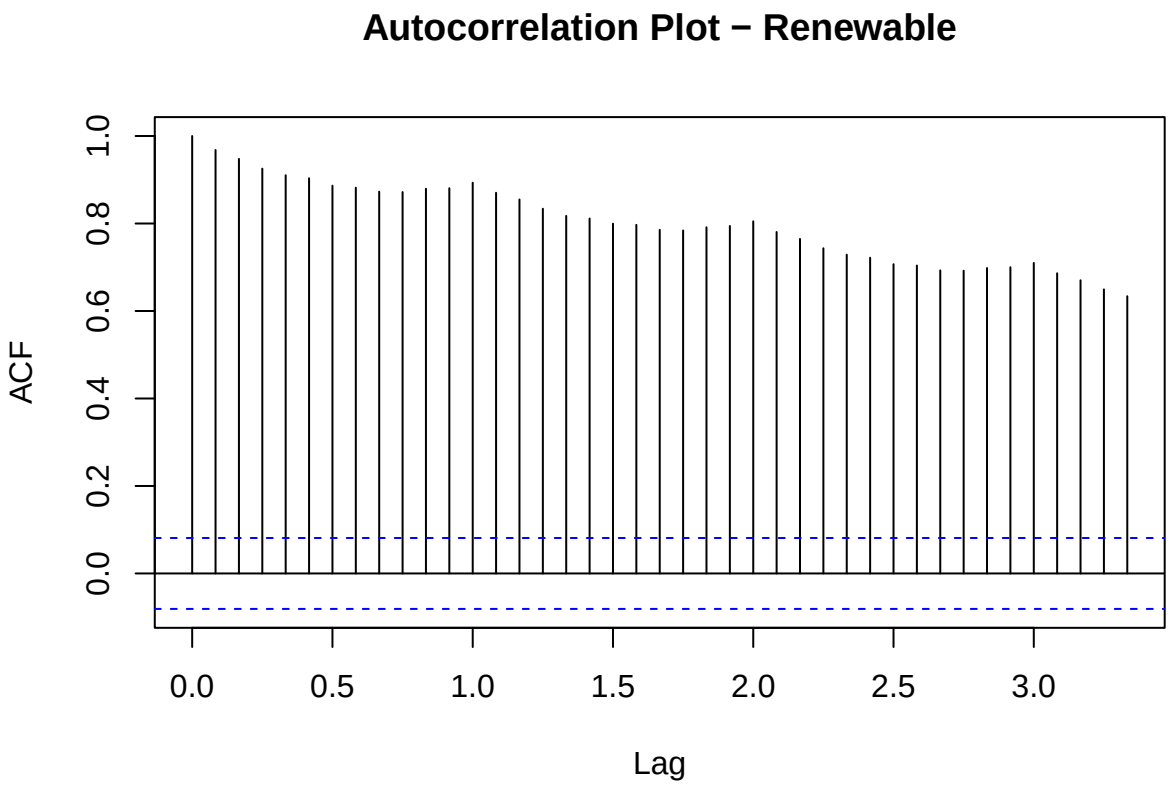
Question 6

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?

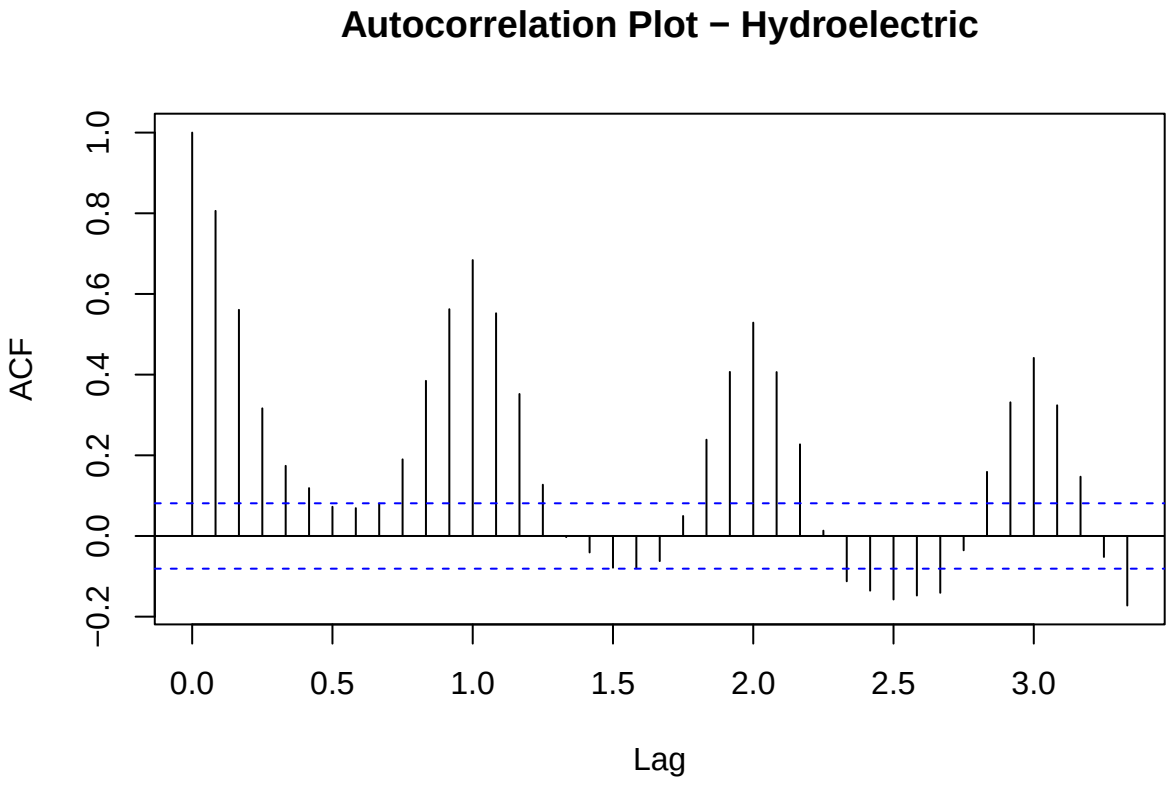
```r
ts_df[,c("Total Biomass Energy Production")] %>%
  acf(lag.max = 40,
      main = "Autocorrelation Plot - Biomass")
```




```r
ts_df[,c("Total Renewable Energy Production")] %>%
  acf(lag.max = 40,
      main = "Autocorrelation Plot - Renewable")
```




```
ts_df[,c("Hydroelectric Power Consumption")] %>%
  acf(lag.max = 40,
      main = "Autocorrelation Plot - Hydroelectric")
```



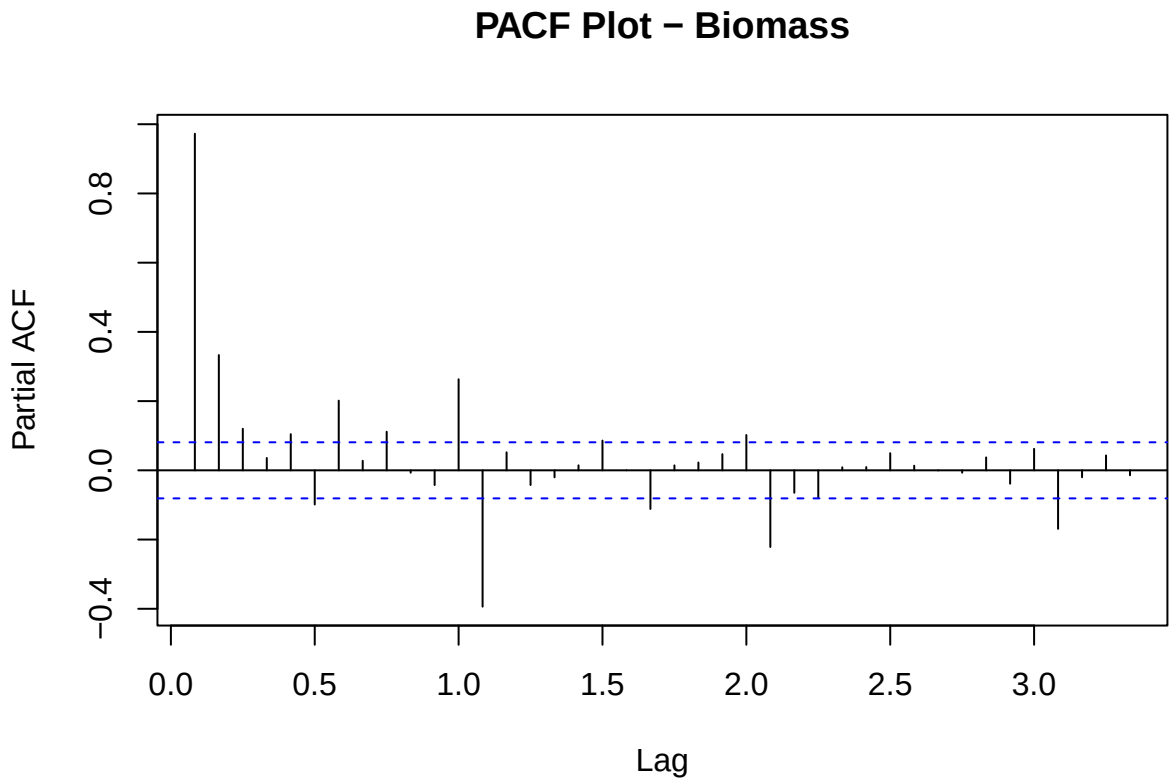
The acf plots for Biomass and Renewable look very similar. Both show very strong autocorrelations for lags

up to 40; we see a slow decrease in autocorrelation as the lag increases. This means autocorrelations for small lags are in general higher than autocorrelations for large lags. The Renewable Energy acf shows a scalloped shape in the plot that the Biomass acf does not, which indicates perhaps some weak seasonality in the Renewable Energy series. The acf for Hydroelectric looks very different from the other 2; we see both positive and negative autocorrelations in the acf for Hydroelectric. The shape of the acf plot for Hydroelectric is oscillating and dampening, indicating a reasonable amount of seasonality.

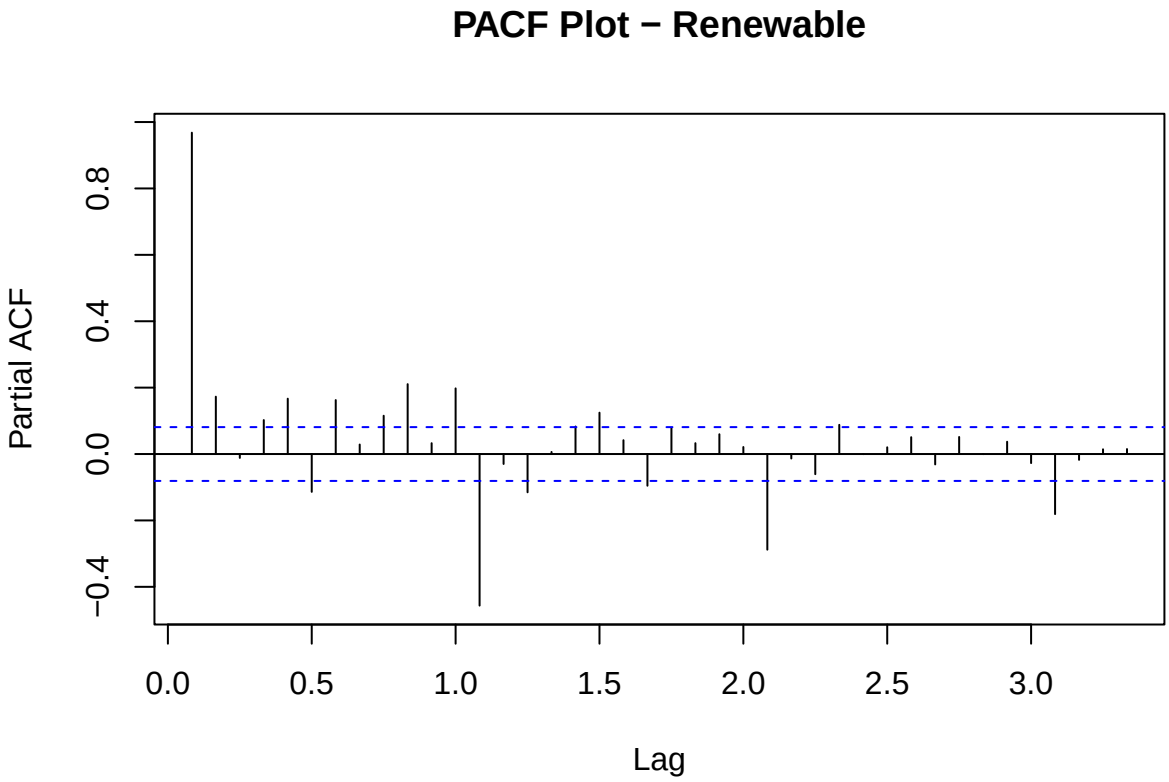
Question 7

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

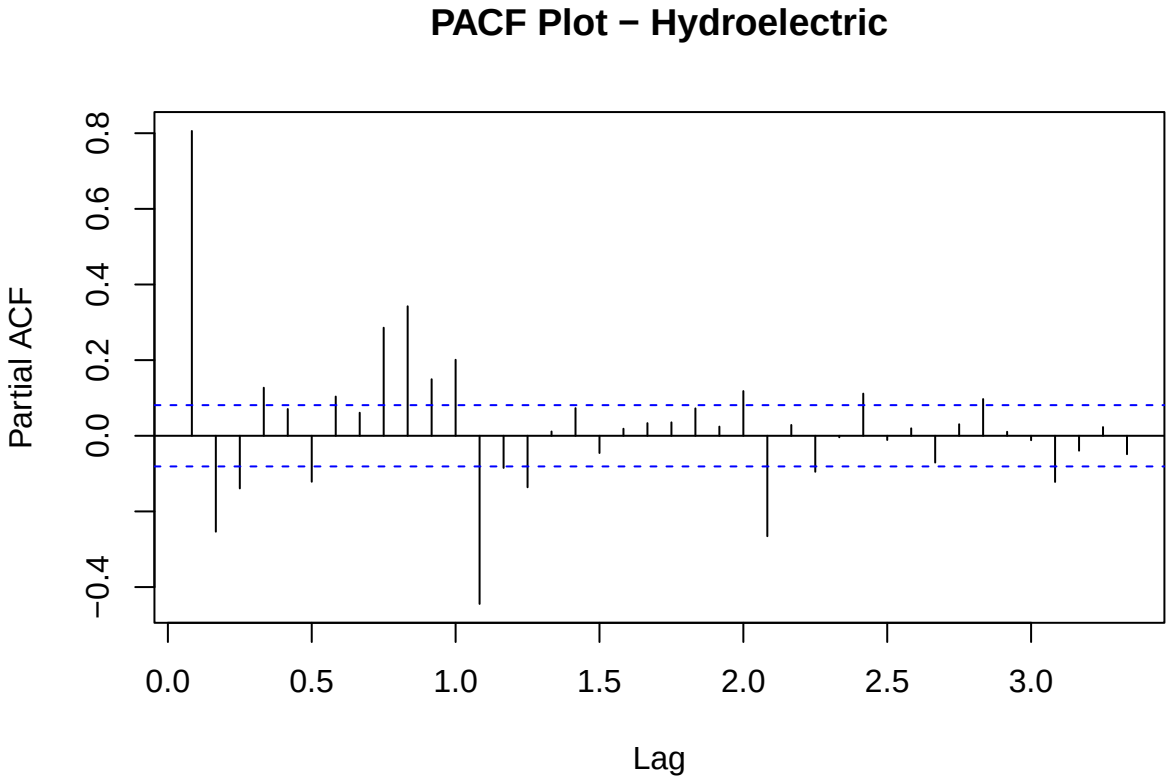
```
ts_df[,c("Total Biomass Energy Production")] %>%
  pacf(lag.max = 40,
       main = "PACF Plot - Biomass")
```




```
ts_df[,c("Total Renewable Energy Production")] %>%
  pacf(lag.max = 40,
       main = "PACF Plot - Renewable")
```




```
ts_df[,c("Hydroelectric Power Consumption")] %>%
  pacf(lag.max = 40,
       main = "PACF Plot - Hydroelectric")
```



The PACF plots do look different from the ACF plots because when we calculate PACF, we remove the effect

of intermediate lags. We actually see some negative values for PACF in the Biomass and Renewable series, while the ACF for the first 40 lags were all positive.