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

Assignment 2 - Due date 01/27/21

Rajat Khandelwal

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.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). }

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.

```
## # A tibble: 6 x 4
##   Month               `Total Biomass Ene~ `Total Renewable E~ `Hydroelectric Po~
##   <dttm>              <chr>               <chr>               <chr>
## 1 1973-01-01 00:00:00 129.787             403.981             272.703
## 2 1973-02-01 00:00:00 117.338             360.9               242.199
## 3 1973-03-01 00:00:00 129.938             400.161             268.81
## 4 1973-04-01 00:00:00 125.636             380.47              253.185
## 5 1973-05-01 00:00:00 129.834             392.141             260.77
## 6 1973-06-01 00:00:00 125.611             377.232             249.859
```

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().

```
head(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.
```

```
#Converting dataframe into time-series
data_ts <- as.ts(data[2:4], start = c(1973,1), end = 2021, frequency = 12)
#Convert time column to date format

head(data_ts)
```

```
## Time Series:
## Start = 1
## End = 6
## Frequency = 1
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                           403.981
## 2                         117.338                           360.900
## 3                         129.938                           400.161
## 4                         125.636                           380.470
## 5                         129.834                           392.141
## 6                         125.611                           377.232
##   Hydroelectric Power Consumption
## 1                         272.703
## 2                         242.199
## 3                         268.810
## 4                         253.185
## 5                         260.770
## 6                         249.859
```

Starting point = January, 1973

Frequency = 12

Question 3

Compute mean and standard deviation for these three series.

```
## [1] "Mean"
```

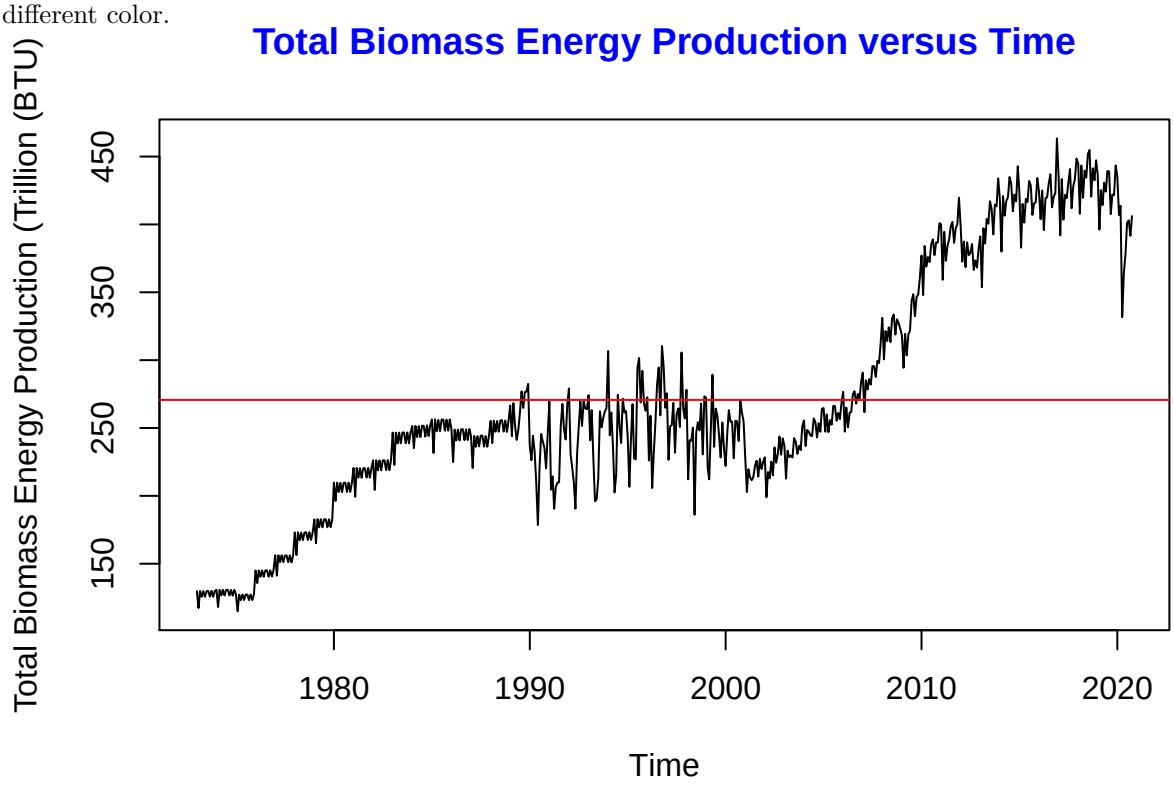
```
##   Total Biomass Energy Production Total Renewable Energy Production
##                           270.6961                           572.7321
##   Hydroelectric Power Consumption
##                           236.9515
```

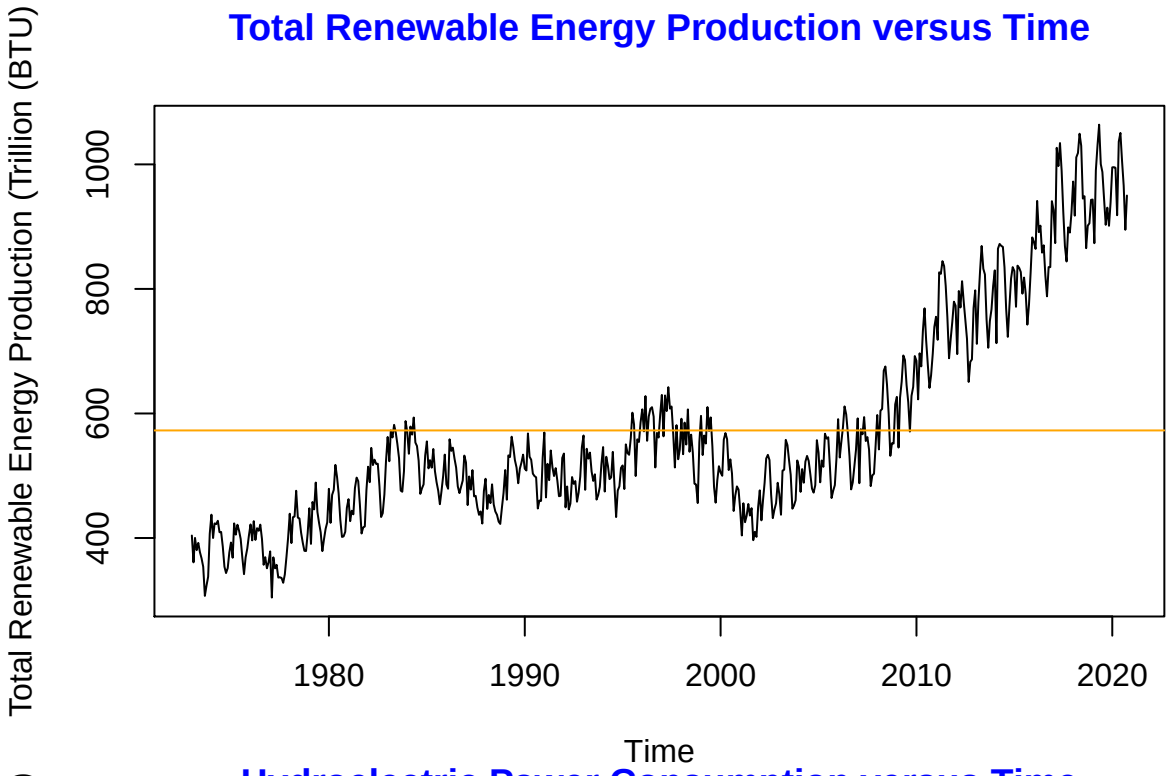
```
## [1] "Standard Deviation"
```

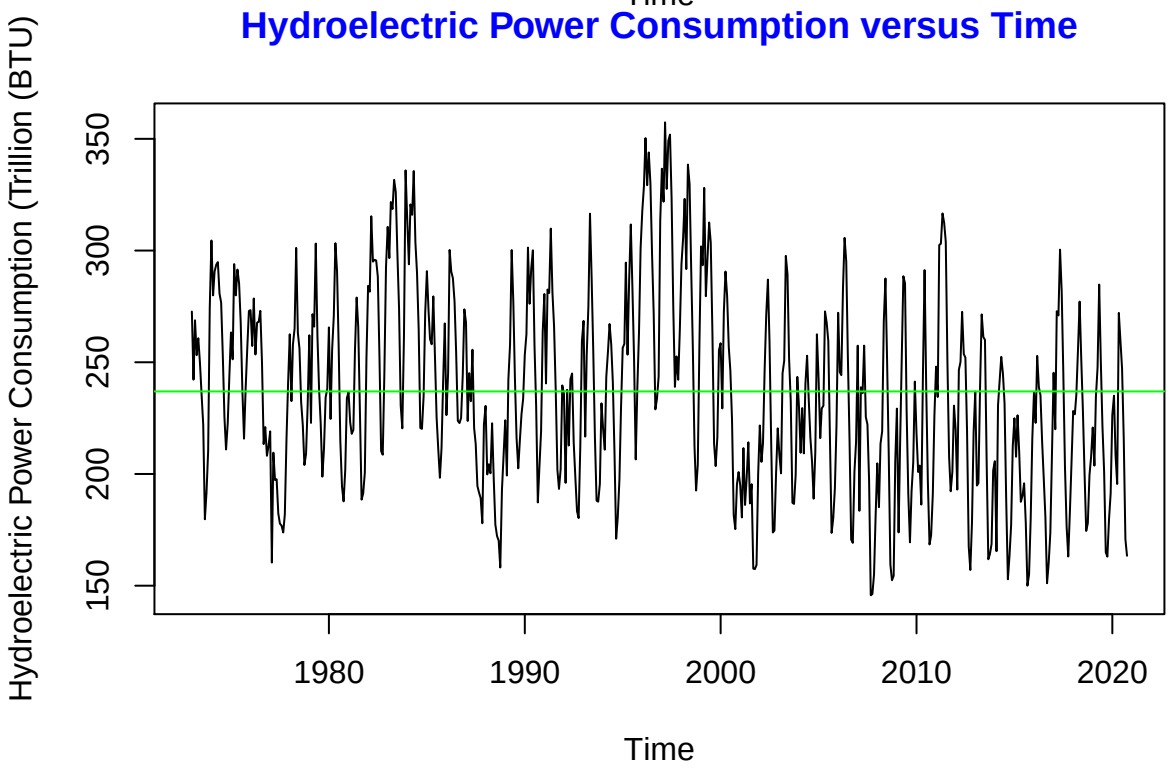
```
##   Total Biomass Energy Production Total Renewable Energy Production
##                          87.36311                          168.45877
##   Hydroelectric Power Consumption
##                          43.90392
```

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.







Question 5

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

```
##                                 Total Biomass Energy Production
## Total Biomass Energy Production                       1.0000000
```

```
## Total Renewable Energy Production                      0.9234609
## Hydroelectric Power Consumption                       -0.2555675
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                       0.923460855
## Total Renewable Energy Production                     1.000000000
## Hydroelectric Power Consumption                      -0.002756852
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                     -0.255567465
## Total Renewable Energy Production                   -0.002756852
## Hydroelectric Power Consumption                      1.000000000
```

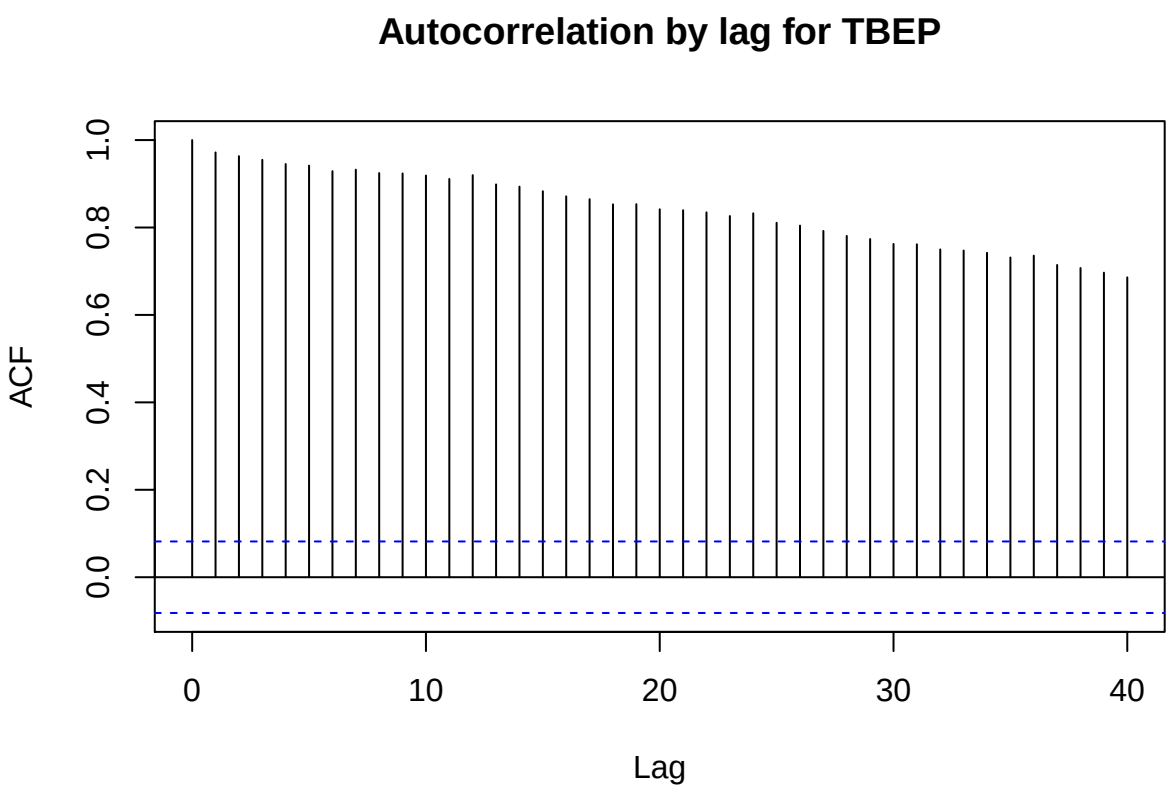
Refer the correlation matrix below.

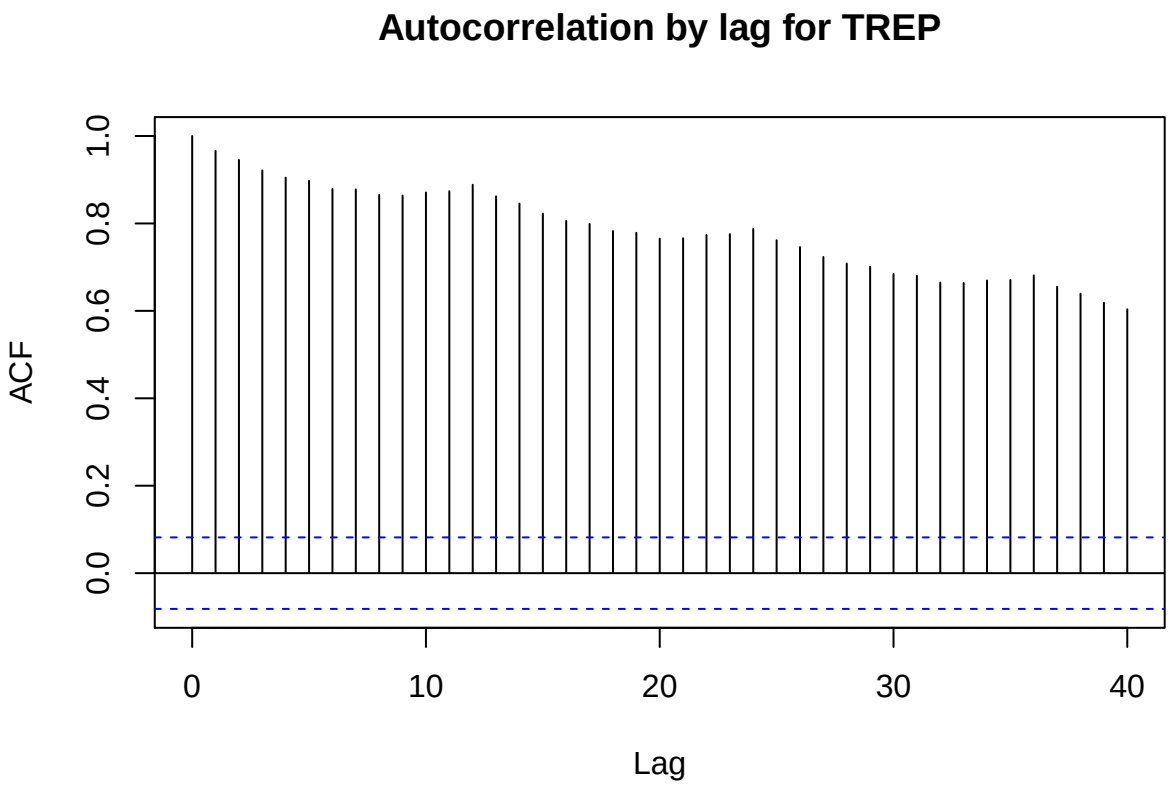
Corr(TBEP, TREP) = +0.923 -> High positive correlation as biomass is a subset of renewables.

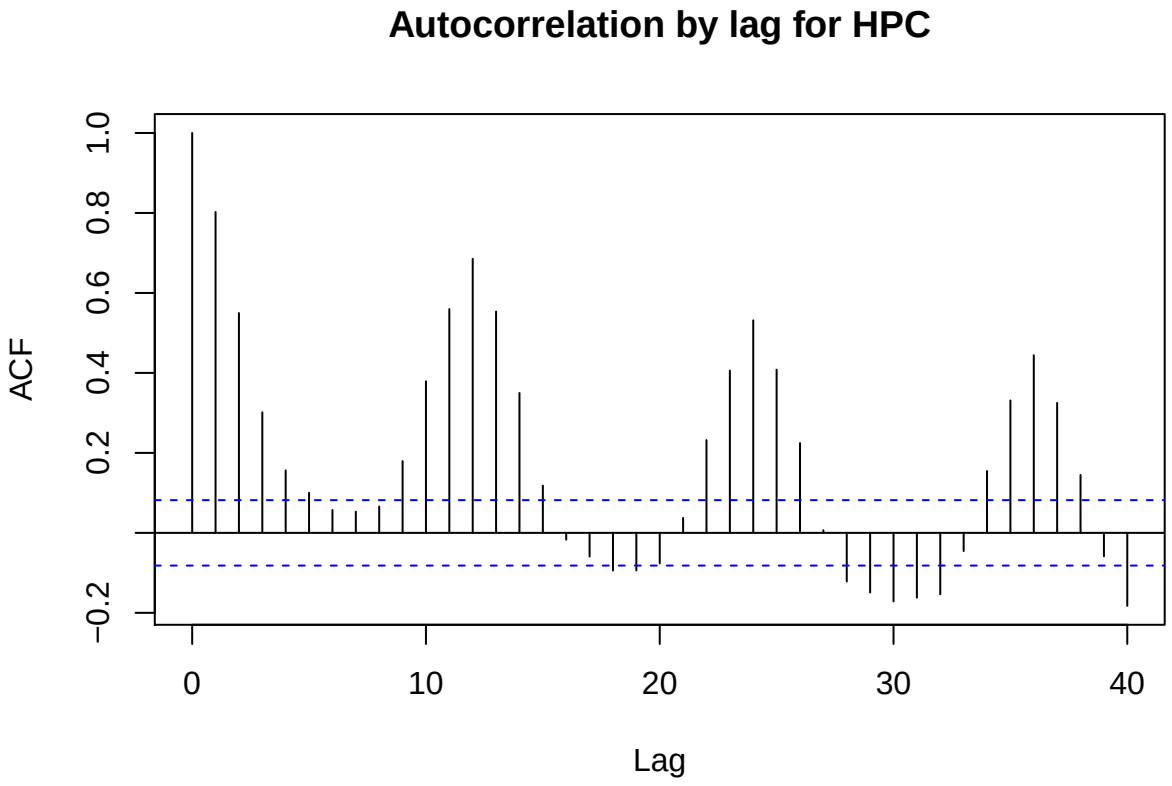
Corr(HPC, TREP) = -0.002 ~ 0 -> Uncorrelated.
Corr(HPC, TBEP) = -0.255 -> Slight negative correlation.

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?







All three plots have different behaviour.

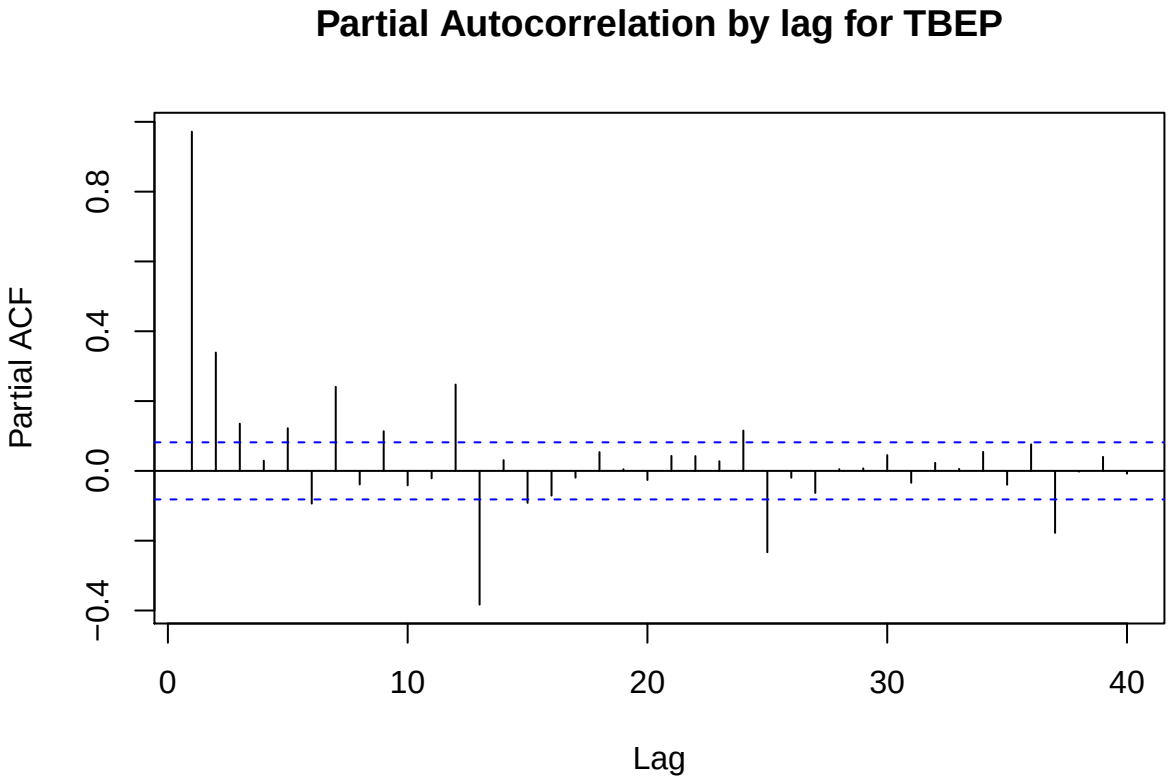
TBEP: ACF is decreasing with lag. This implies that TBEP shows an increasing or decreasing trend.

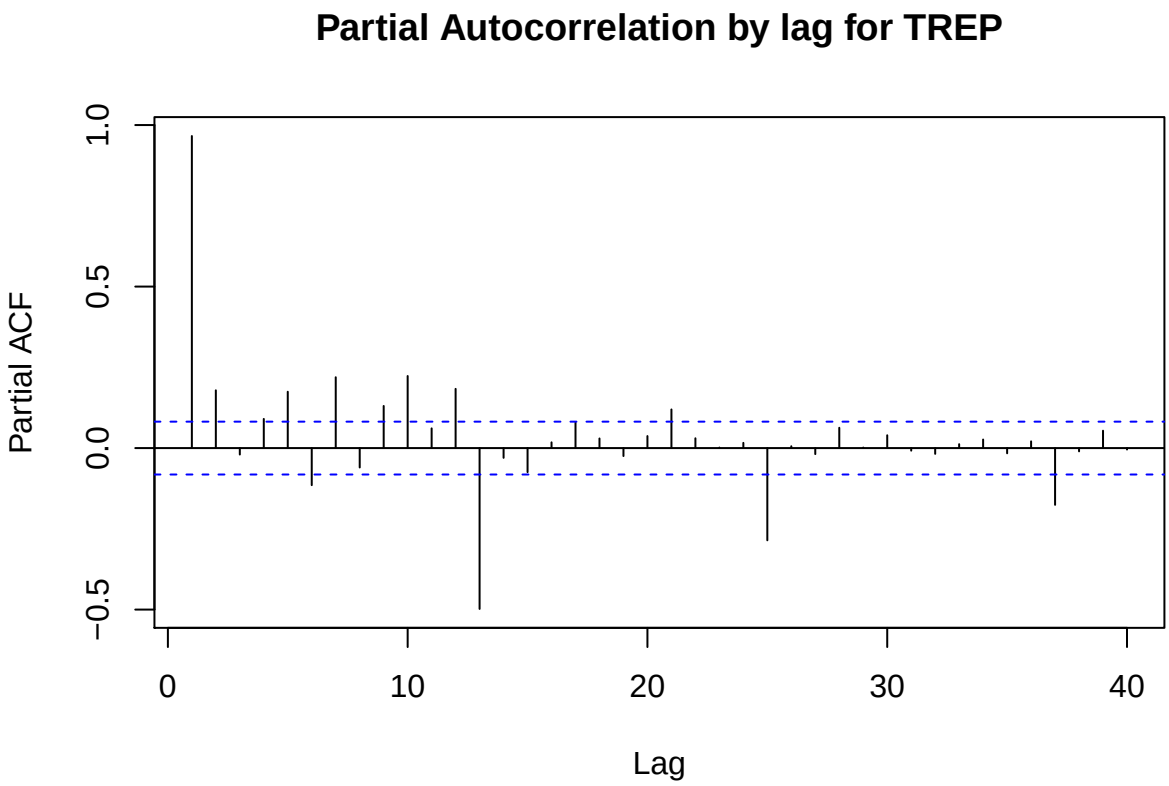
TREP: ACF is decreasing with lag. This implies that TREP shows an increasing or decreasing trend. There is a slight seasonality as well. ACF is higher around the 12 month mark which shows that same months across different years can expect similar TREP. This makes sense as renewable energy production is weather (sun/wind) dependent. Hence, TREP in May '20 will show higher correlation to TREP in May '19 as compared to Jan '19.

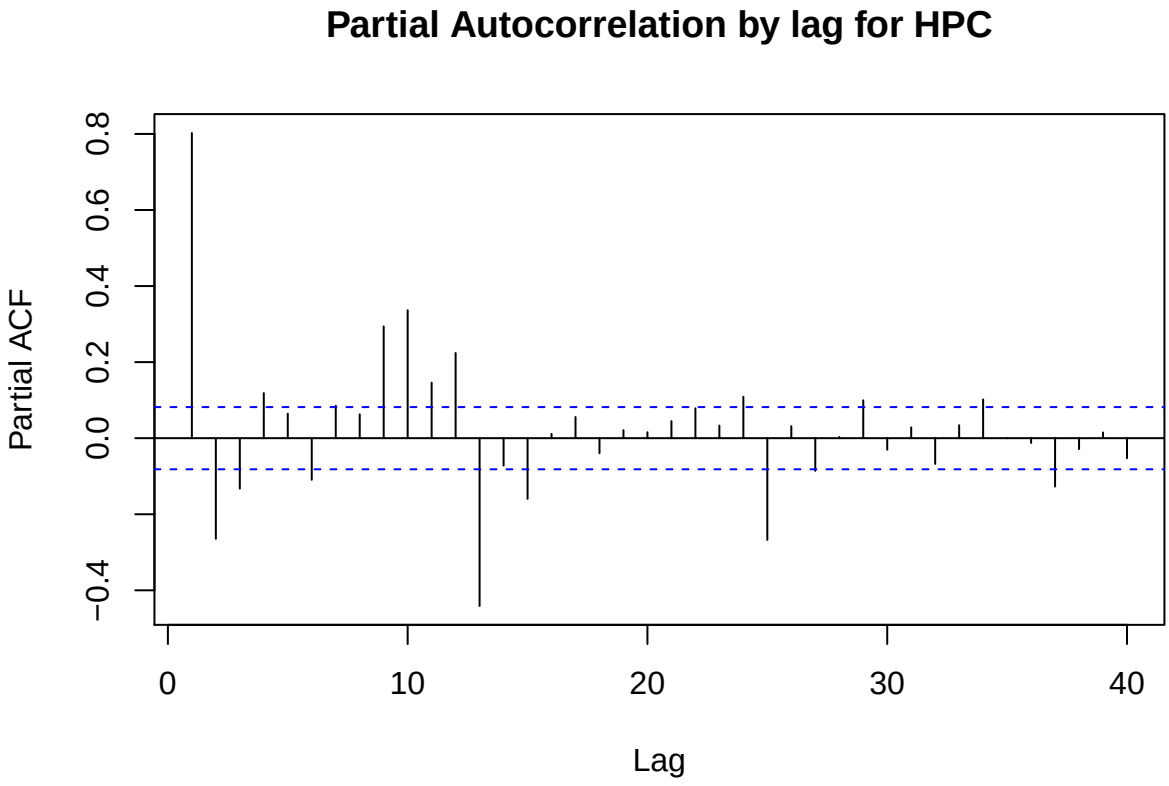
HPC: Shows strong seasonality. HPC is high in the rainy season and low in summer, reflecting seasonal variations in water level for a reservoir.

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?




Partial Autocorrelation by lag for TREP
Partial ACF
1.0
0.5
0.0
−0.5
0
10
20
30
40
Lag


Partial Autocorrelation by lag for HPC
Partial ACF
0.8
0.6
0.4
0.2
0.0
−0.4
0
10
20
30
40
Lag

PACFs for all three series' shows peaks at the lag = 12,24 and 36, illustrating the consistency across the same months in this data. That is, data for January '21 is correlated with data for January '20, January '19 and so on. However, this correlation decreases across years. This shows the seasonality discussed for TREP and HPC in the previous answer. However, using the PACF plot we can see that the TBEP also depicts a slight seasonality.

Appendix

```
#Load/install required package here
library(forecast)
library(tseries)
library(dplyr)
library(utils)

#Importing data set
#NOTE: Locally changed name of XLSX file to "REP_Data.xlsx" for easier reference.
library(readxl)
data <- read_excel("~/Documents/GitHub/ENV790_30_TSA_S2021_Temp/Data/REP_Data.xlsx", skip = 10)

data <- data %>% select(1,4:6) %>% slice(2:n())
head(data)

#Converting dataframe columns into numeric values

data$`Total Biomass Energy Production` <- as.numeric(data$`Total Biomass Energy Production`)
data$`Total Renewable Energy Production` <- as.numeric(data$`Total Renewable Energy Production`)
data$`Hydroelectric Power Consumption` <- as.numeric(data$`Hydroelectric Power Consumption`)

#Converting dataframe into time-series
data_ts <- as.ts(data[2:4], start = c(1973,1), end = 2021, frequency = 12)

#Calculating Mean

print("Mean")
apply(data_ts,2,mean)

#Calculating Standard Deviation

print("Standard Deviation")
apply(data_ts,2,sd)

#Plotting the time-series

plot(x = data$Month, y = data_ts[,1], type = "l", xlab = 'Time',
    ylab = 'Total Biomass Energy Production (Trillion (BTU)')
abline(h = mean(data_ts[,1]), col = "red")
title("Total Biomass Energy Production versus Time",col.main = "blue")

plot(x = data$Month, y = data_ts[,2], type = "l", xlab = 'Time',
     ylab = 'Total Renewable Energy Production (Trillion (BTU)')
abline(h = mean(data_ts[,2]), col = "yellow")
title("Total Renewable Energy Production versus Time", col.main = "blue")

plot(x = data$Month, y = data_ts[,3], type = "l", xlab = 'Time',
     ylab = 'Hydroelectric Power Consumption (Trillion (BTU)')
abline(h = mean(data_ts[,3]), col = "green")
title("Hydroelectric Power Consumption versus Time", col.main = "blue")

#Calculating the correlation matrix
```

```
cor(data_ts, method = "pearson")


#Plotting ACF

acf(data_ts[,1], plot = TRUE, lag = 40, main = "Autocorrelation by lag for TBEP")
acf(data_ts[,2], plot = TRUE, lag = 40, main = "Autocorrelation by lag for TREP")
acf(data_ts[,3], plot = TRUE, lag = 40, main = "Autocorrelation by lag for HyPC")


#Plotting PACF

pacf(data_ts[,1], plot = TRUE, lag = 40, main = "Partial Autocorrelation by lag for TBEP")
pacf(data_ts[,2], plot = TRUE, lag = 40, main = "Partial Autocorrelation by lag for TREP")
pacf(data_ts[,3], plot = TRUE, lag = 40, main = "Partial Autocorrelation by lag for HyPC")
```

‘‘‘