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

Assignment 2 - Due date 01/27/21

Yash Doshi

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).

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 Energy~ `Total Renewable Ener~ `Hydroelectric Power~
  <date>                      <dbl>                  <dbl>                  <dbl>
1 1973-01-01                   130.                   404.                   273.
2 1973-02-01                   117.                   361.                   242.
3 1973-03-01                   130.                   400.                   269.
4 1973-04-01                   126.                   380.                   253.
5 1973-05-01                   130.                   392.                   261.
6 1973-06-01                   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().

The starting point for this time series is 1, ending point is 574, and the frequency is 1. The reason why frequency is 1 is because frequency is the number of observations per unit time. In this particular data, there were only one observations per unit time, hence, the frequency is 1.

Below is the output for head() function. It was practically not possible to include all the observations, since it would have become too large. Hence, I only included the first six variables.

```
Time Series:
Start = 1
End = 6
Frequency = 1
  Month Total Biomass Energy Production (Trillion Btu)
1  1096                                          129.787
2  1127                                          117.338
3  1155                                          129.938
4  1186                                          125.636
5  1216                                          129.834
6  1247                                          125.611
  Total Renewable Energy Production (Trillion Btu)
1                                          403.981
2                                          360.900
3                                          400.161
4                                          380.470
5                                          392.141
6                                          377.232
  Hydroelectric Power Consumption (Trillion Btu)
1                                        272.703
2                                        242.199
3                                        268.810
4                                        253.185
5                                        260.770
6                                        249.859
```

Question 3

Compute mean and standard deviation for these three series.

Mean and Standard Deviation of Total Biomass Energy Production

Mean

```
[1] 270.6961
```

Standard Deviation

```
[1] 87.36311
```

Mean and Standard Deviation of Total Renewable Energy Production

Mean

```
[1] 572.7321
```

Standard Deviation

```
[1] 168.4588
```

Mean and Standard Deviation of Hydroelectric Power Consumption

Mean

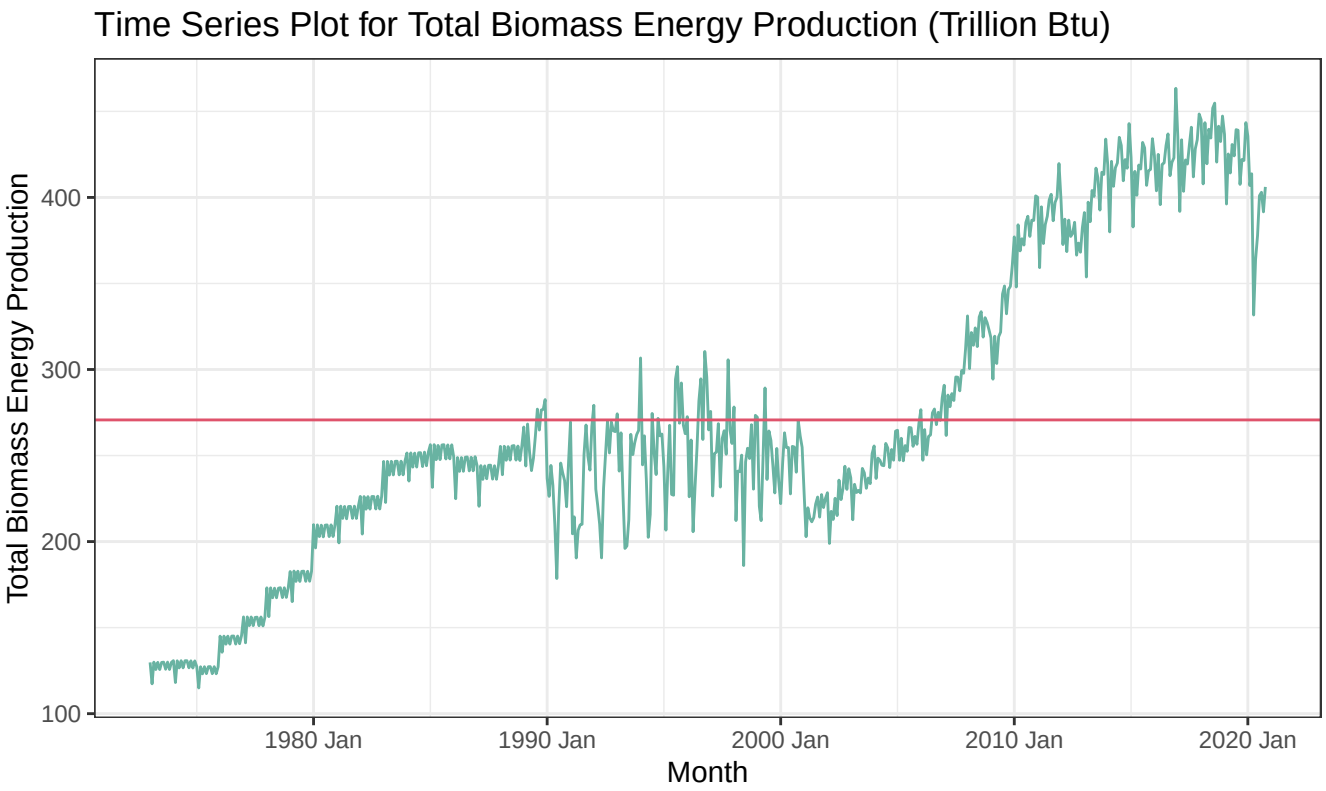
```
[1] 236.9515
```

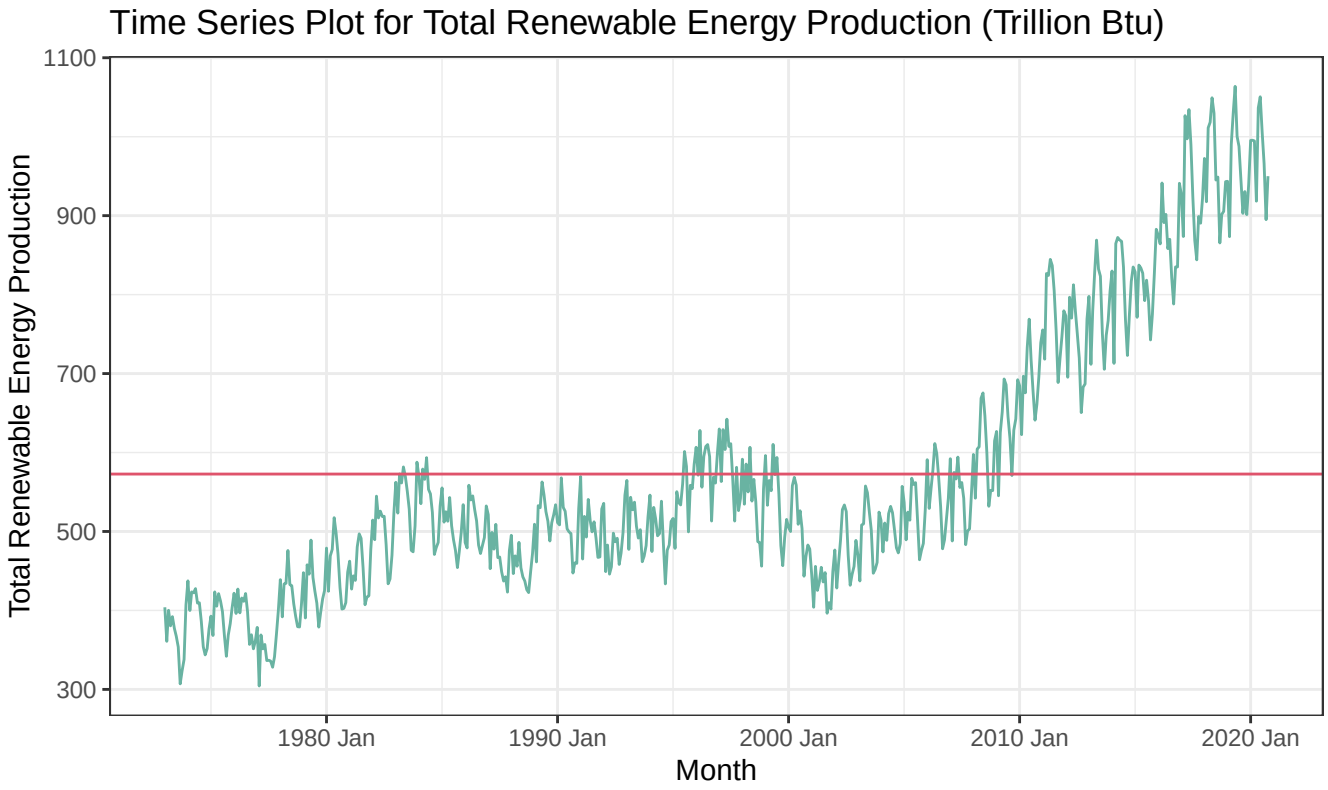
Standard Deviation

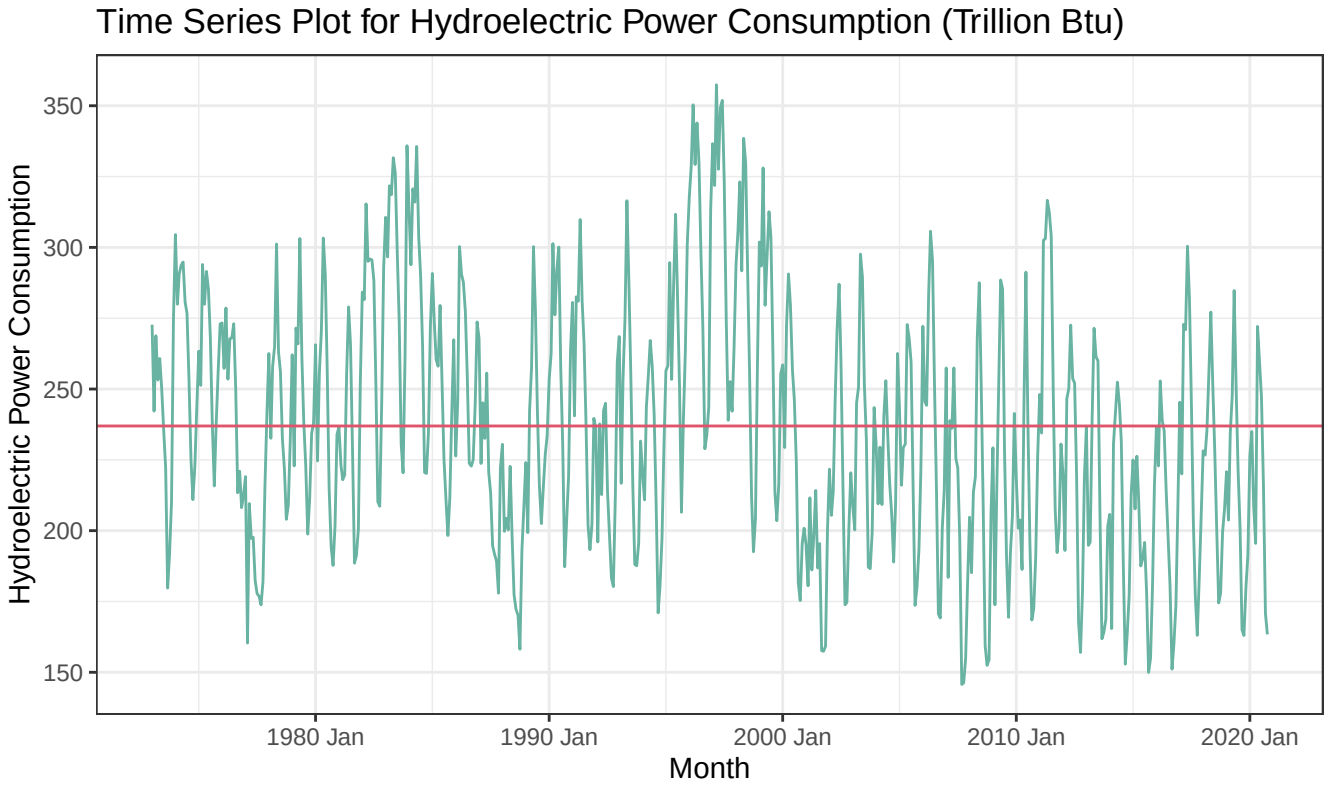
```
[1] 43.90392
```

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.







Question 5

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

Correlation between Total Biomass Energy Production and Total Renewable Energy Production

```
[1] 0.9234609
```

A correlation of 0.92 indicates that there is a strong, positive correlation between Total Biomass Energy Production and Total Renewable Energy Production. It means that both the variables move in the same direction together.

Correlation between Total Biomass Energy Production and Hydroelectric Power Consumption

```
[1] -0.2555675
```

A correlation of -0.255 indicates that there is a weak, negative correlation between Total Biomass Energy Production and Hydroelectric Power Consumption. It means that they do not move in the same direction together. This makes sense, because one variable is a measure of production , whereas, the other variable is a measure of consumption. Both these variables are of two completely different energies (one is of biomass, and the other is of hydroelectric).

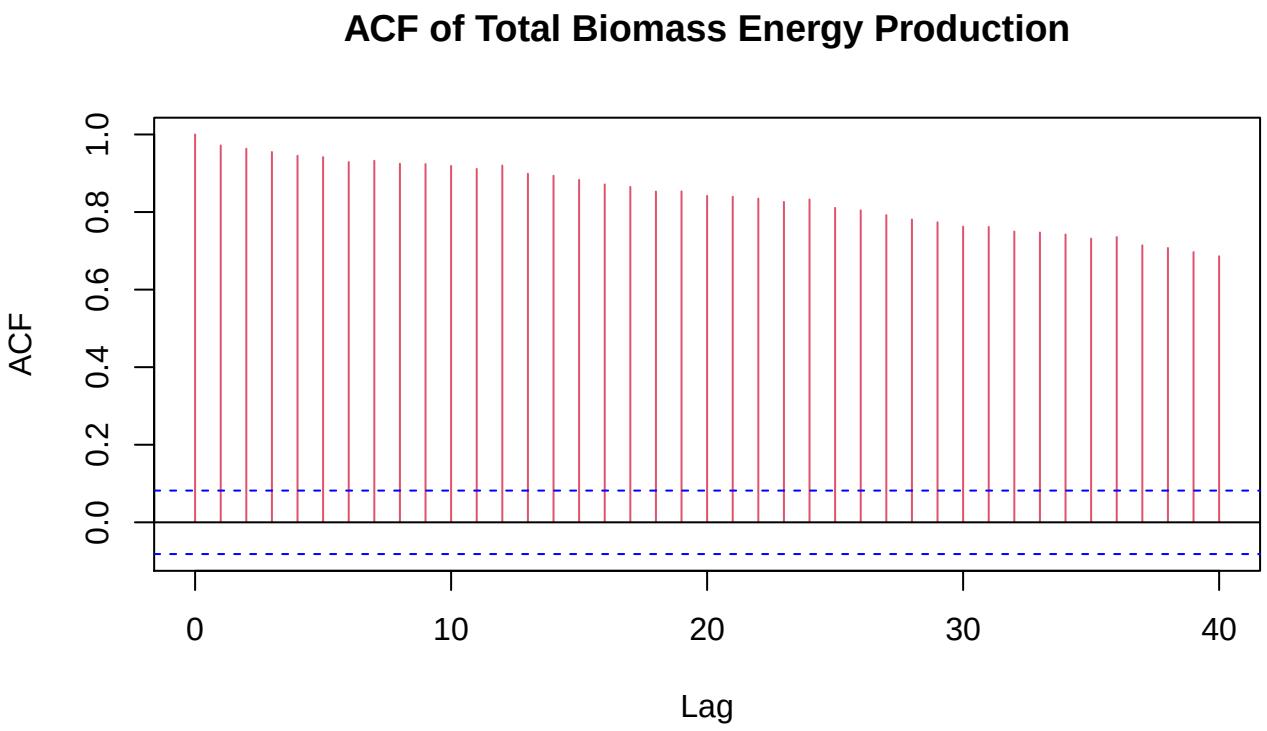
Correlation between Total Renewable Energy Production and Hydroelectric Power Consumption

```
[1] -0.002756852
```

A correlation of -0.0027 indicates a very weak, negative correlation between Total Renewable Energy Production and Hydroelectric Power Consumption. It means that they do not move in the same direction together.

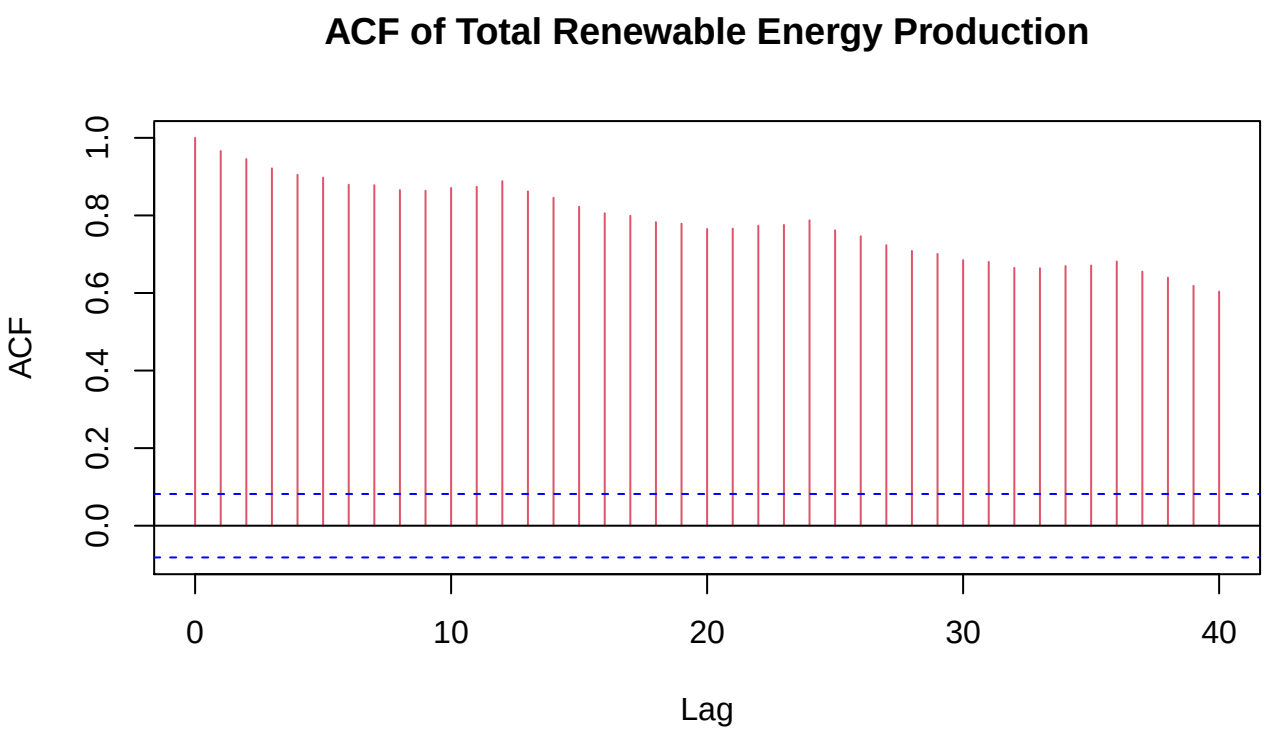
Question 6

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?



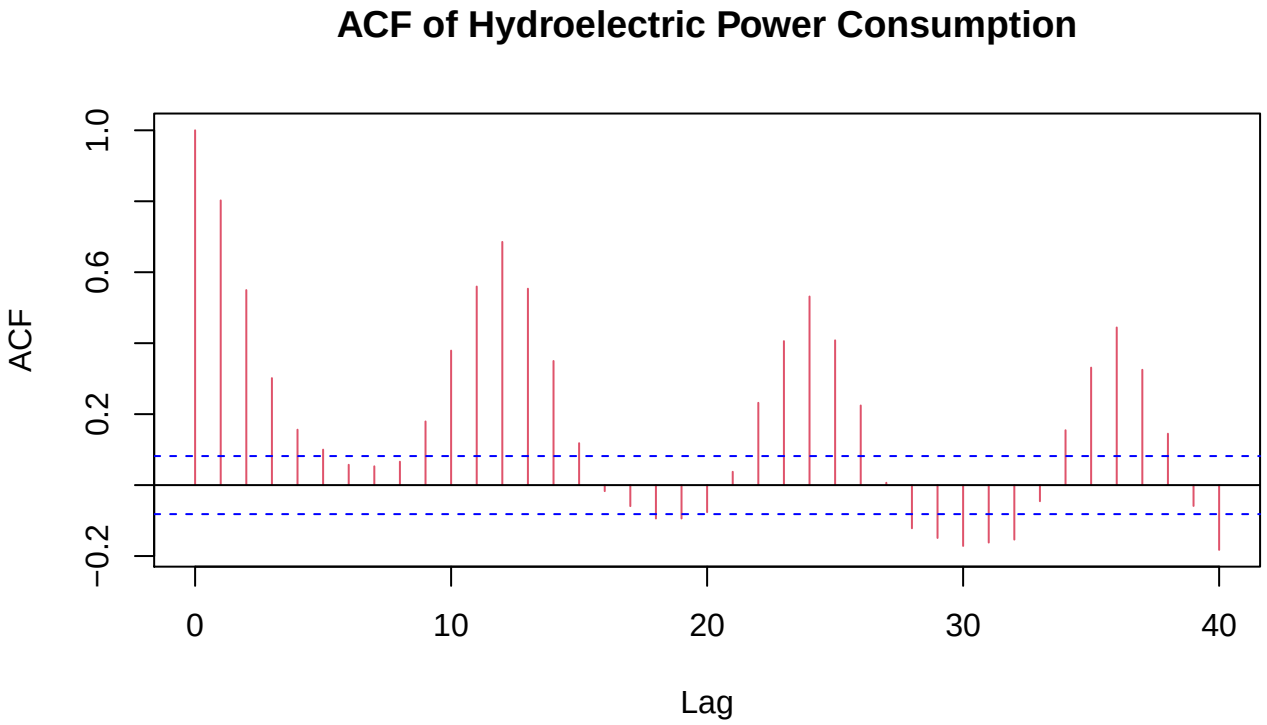
ACF is a measure of dependence between two adjacent values of the same variables. In ACF, we talk about the same variable at different times. This plot tells us how the biomass production at a given time period is related to another time period. For instance, the correlation between first point at lag 1 and second point at lag 2 is 0.972.

Hence, ACF tells us how correlated the points are with each other, based on how many time steps they are separated by. It is how correlated past data points are to the future data points, for different values of time separation.



It can be seen from this plot that there is not much difference in correlation between the total renewable energy production at two separate time periods. In this case, the correlation between first point at lag 1 and second point at lag 2 is 0.966.

If we compare the ACF of total renewable energy production with total biomass energy production, we find that the points at different time intervals are highly autocorrelated in case of biomass energy production than total renewable energy production.



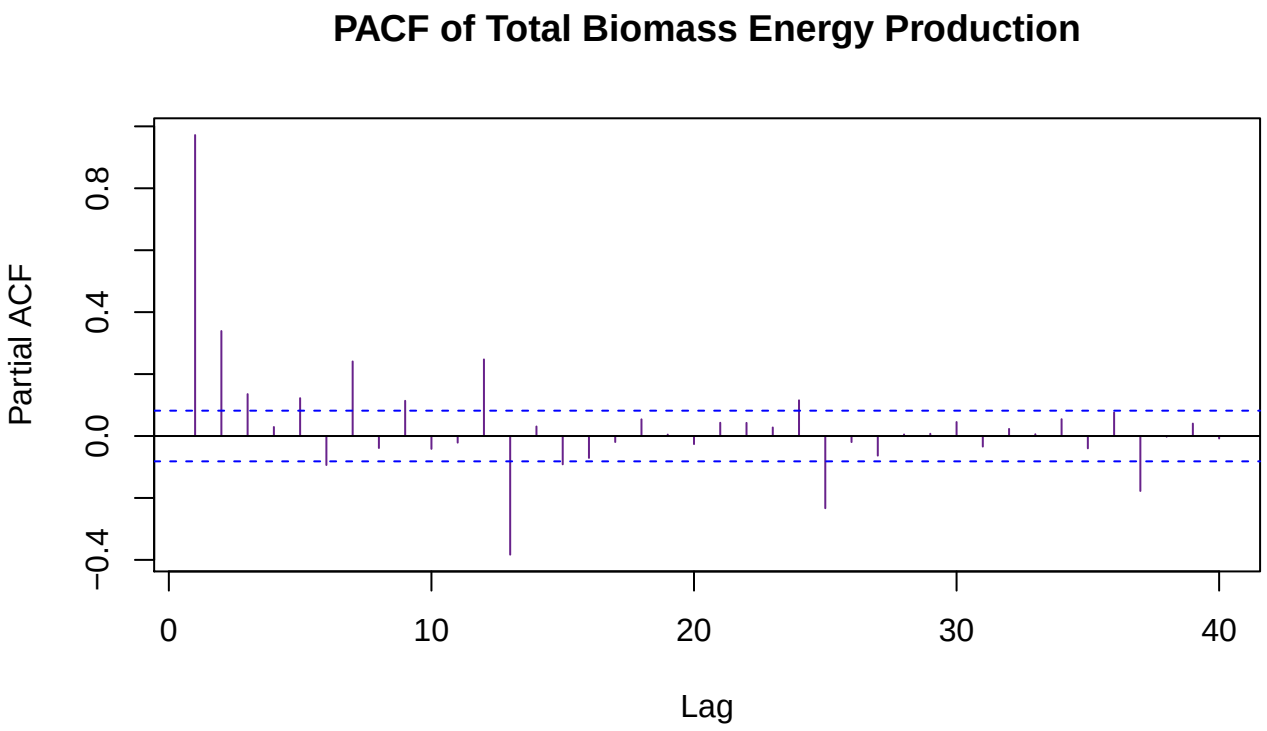
This ACF plot shows that there is very little correlation between the two variables at different time periods. For instance, the correlation between first point at lag 1 and second point at lag 2 is 0.802, and that between the first point and the third point at lag 3 is 0.550.

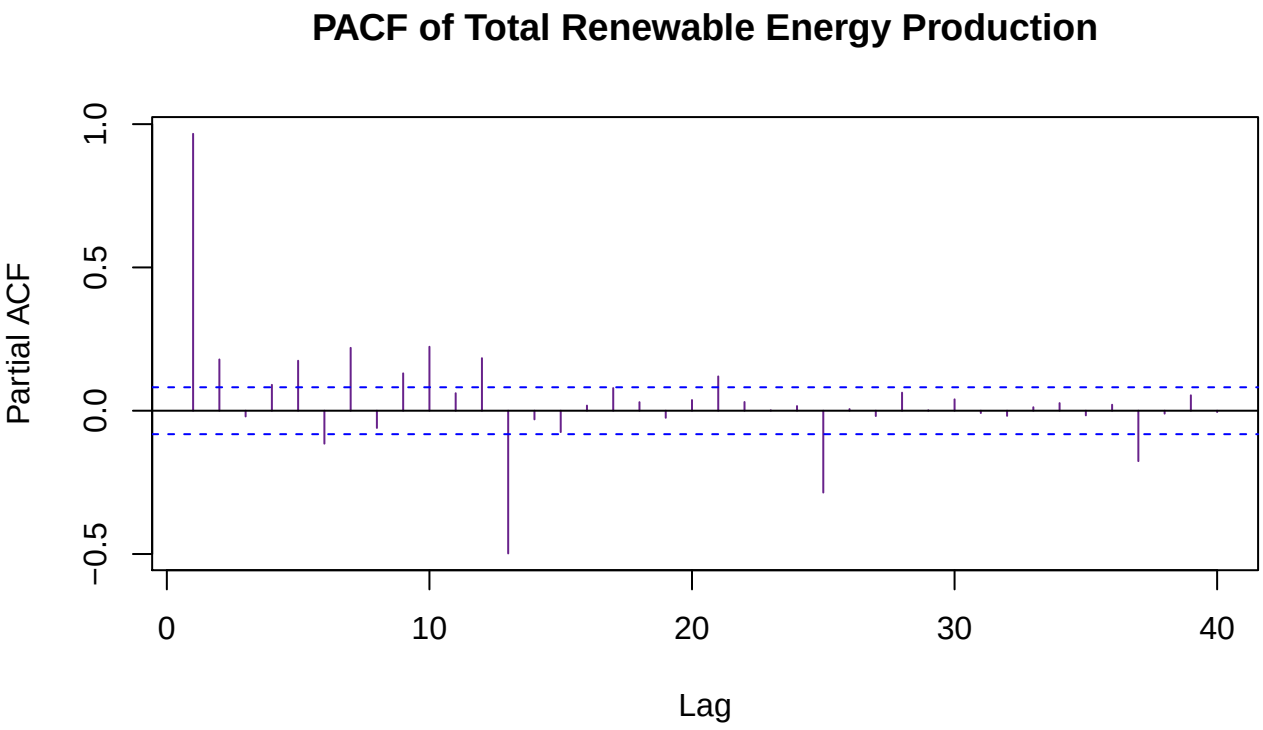
It should also be noted that ACF for Hydroelectric Power Consumption shows a seasonality. That is the reason why the autocorrelation factor drops, increases, and then drops again.

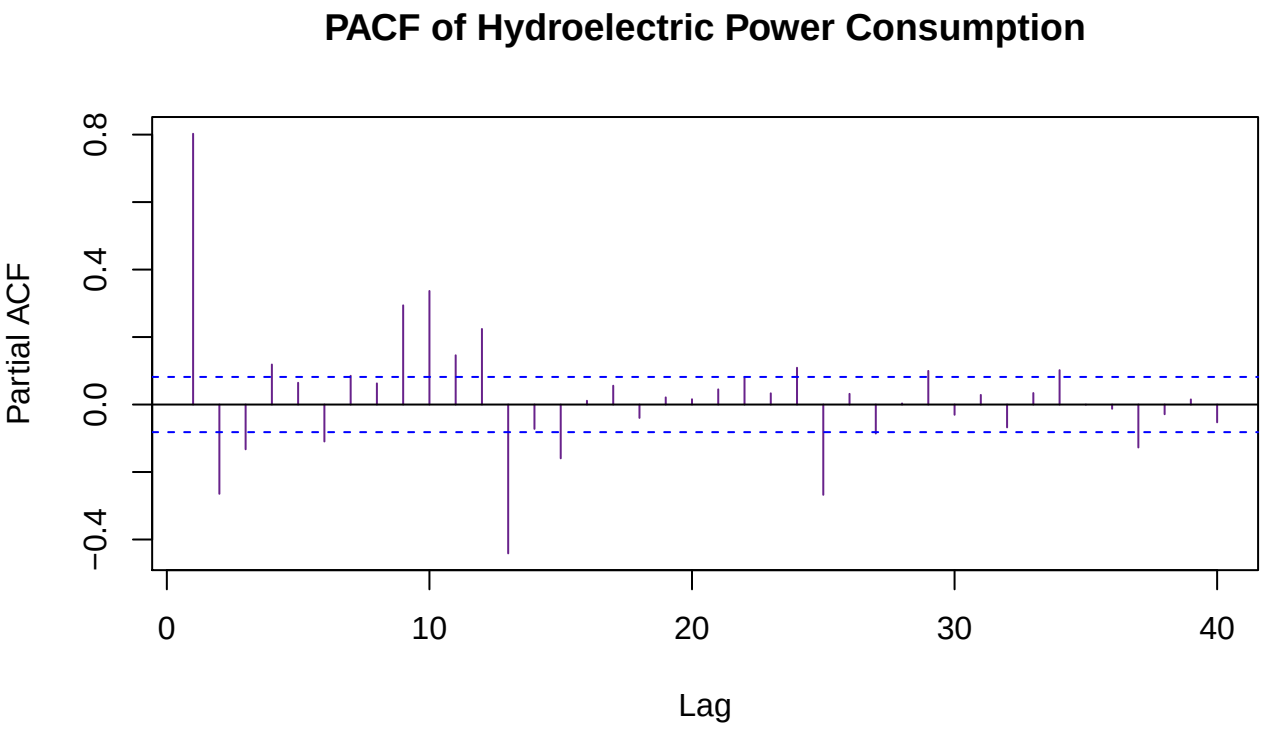
The three graphs do not showcase the same behavior. The ACF for total biomass energy production and total renewable energy production show a similar trend. The only thing is that the correlation of total renewable energy production is slightly lower than the total biomass energy production. However, the graph for hydroelectric power consumption is completely different than the other two. The hydroelectric power consumption shows a seasonal trend. That is the reason why it is rising, dropping, and then rising again.

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 talks about the correlation between two points separated by some time period. PACF does not take into consideration the correlation of the points in-between them. Unlike ACF, the values of PACF try to be as close to zero as possible. PACF is important in order to know how one point is related to some other point in a distant future without the intervening terms.

These plots vary considerably from the plots in Q6. The first plot shows a significant correlation between first and second point, followed by correlations that are not so significant. The pattern of PACF plot for total renewable energy production is similar to the total biomass energy production. The PACF plot for Hydroelectric Power Consumption is the same as its ACF plot.

There is not much we can discover from the PACF plot alone. In order to analyze the data in a better way, we need to use ACF plot too. Together, we can know a lot about the data.

APPENDIX

```
knitr::opts_chunk$set(echo = F, eval = T, comment = NA, message = F,warning = F,
                      fig.width = 7, fig.height = 4.2)
#Load/install required package here
library(forecast)
library(tseries)
library(dplyr)
library(readxl)
library(ggplot2)
library(hrbrthemes)

#Importing data set
setwd("/Users/yashdoshi/Desktop/Duke/Courses/Spring 2021/Time Series Analysis/Labs/Lab Work/ENV790_30_TSA_S2021
eia = read_excel("Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx")
eia$Month = as.Date(eia$Month, format = "%m/%y")
eia[2:3] = NULL
eia[5:12] = NULL

head(eia)
eiats = ts(data = eia, start = 1, end = 574, frequency = 1)
head(eiats)
meantbep = mean(eia$`Total Biomass Energy Production (Trillion Btu)`)
meantbep
sd1 = sd(eia$`Total Biomass Energy Production (Trillion Btu)`)
sd1
meantrep = mean(eia$`Total Renewable Energy Production (Trillion Btu)`)
meantrep
sd2 = sd(eia$`Total Renewable Energy Production (Trillion Btu)`)
sd2
meanhepc = mean(eia$`Hydroelectric Power Consumption (Trillion Btu)`)
meanhepc
sd3 = sd(eia$`Hydroelectric Power Consumption (Trillion Btu)`)
sd3
#Plot for Total Biomass Energy Production
p = ggplot(data = eia, aes(x = Month,
               y = `Total Biomass Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Month") +
  ylab("Total Biomass Energy Production") +
  ggtitle("Time Series Plot for Total Biomass Energy Production (Trillion Btu)") +
  theme_bw()
p + scale_x_date(date_labels = "%Y %b") +
  geom_hline(yintercept = mean(eia$`Total Biomass Energy Production (Trillion Btu)`),
               color = 2)

#Plot for Renewable Energy Production
q = ggplot(eia, aes(x = Month,
                    y = `Total Renewable Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Month") +
  ylab("Total Renewable Energy Production") +
  ggtitle("Time Series Plot for Total Renewable Energy Production (Trillion Btu)") +
```

```
  theme_bw()
q + geom_hline(yintercept = mean(eia$`Total Renewable Energy Production (Trillion Btu)`), color = 2) +
  scale_x_date(date_labels = "%Y %b")

#Plot for Hydroelectric Power Consumption
he = ggplot(eia, aes(x = Month,
                     y = `Hydroelectric Power Consumption (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Month") +
  ylab("Hydroelectric Power Consumption") +
  ggtitle("Time Series Plot for Hydroelectric Power Consumption (Trillion Btu)") +
  theme_bw()
he + geom_hline(yintercept = mean(eia$`Hydroelectric Power Consumption (Trillion Btu)`), color = 2) +
  scale_x_date(date_labels = "%Y %b")

#Correlation between Total Biomass Energy Production and Total Renewable Energy Production
cor1 = cor(eia$`Total Biomass Energy Production (Trillion Btu)`,
    eia$`Total Renewable Energy Production (Trillion Btu)`)
cor1

#Correlation between Total Biomass Energy Production and Hydroelectric Power Consumption
cor2 = cor(eia$`Total Biomass Energy Production (Trillion Btu)`,
           eia$`Hydroelectric Power Consumption (Trillion Btu)`)
cor2

#Correlation between Total Renewable Energy Production and Hydroelectric Power Consumption
cor3 = cor(eia$`Total Renewable Energy Production (Trillion Btu)`,
           eia$`Hydroelectric Power Consumption (Trillion Btu)`)
cor3

#ACF of Total Bimoass Energy Production
acf1 = acf(eia$`Total Biomass Energy Production (Trillion Btu)`, lag.max = 40,
    plot = FALSE)
plot(acf1, main = "ACF of Total Biomass Energy Production", col = 2)

#ACF of Total Renewable Energy Production
acf2 = acf(eia$`Total Renewable Energy Production (Trillion Btu)`, lag.max = 40,
    plot = FALSE)
plot(acf2, main = "ACF of Total Renewable Energy Production", col = 2)

#ACF of Hydroelectric Power Consumption
acf3 = acf(eia$`Hydroelectric Power Consumption (Trillion Btu)`, lag.max = 40,
            plot = FALSE)
plot(acf3, main = "ACF of Hydroelectric Power Consumption", col = 2)

#PACF of Total Biomass Energy Production
pacf1 = pacf(eia$`Total Biomass Energy Production (Trillion Btu)`, lag.max = 40,
     plot = FALSE)
plot(pacf1, main = "PACF of Total Biomass Energy Production", col = "darkorchid4")

#PACF of Total Renewable Energy Production
pacf2 = pacf(eia$`Total Renewable Energy Production (Trillion Btu)`, lag.max = 40,
     plot = FALSE)
plot(pacf2, main = "PACF of Total Renewable Energy Production", col = "darkorchid4")

#PACF of Hydroelectric Power Consumption
pacf3 = pacf(eia$`Hydroelectric Power Consumption (Trillion Btu)`,lag.max = 40,
             plot = FALSE)
plot(pacf3, main = "PACF of Hydroelectric Power Consumption", col = "darkorchid4")
```