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

Assignment 2 - Due date 01/27/21

Peyton Chen (yc451)

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.\

```
#Load/install required package here
library(readxl)
library(forecast)
library(tseries)
library(dplyr)
library(ggplot2)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
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
Data <- read_excel(path = "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                   sheet = "Monthly Data",
                   skip = 9, col_names = TRUE)
Data <- Data[2:nrow(Data),]
```

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.

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

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

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_data <- ts(data = Data, start = c(1973,1), end = c(2020,10),frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
# The mean for these three series
col_mean <- apply(ts_data, 2, mean)
col_mean
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                         286.0889                           287.5000
##   Hydroelectric Power Consumption
##                         287.5000
```

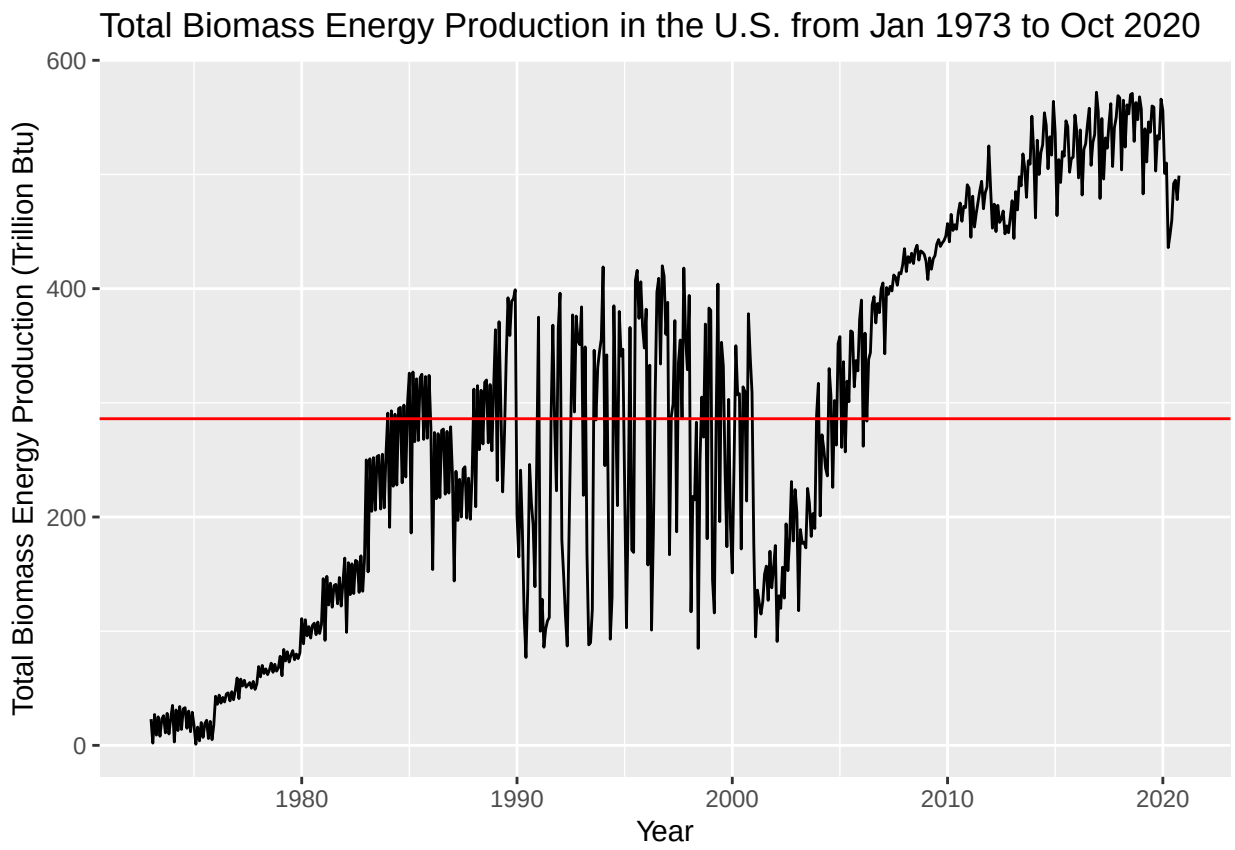
```
# The standard deviation for these three series
col_sd <- apply(ts_data,2, sd)
col_sd
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                         165.3481                           165.8438
##   Hydroelectric Power Consumption
##                         165.8438
```

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.

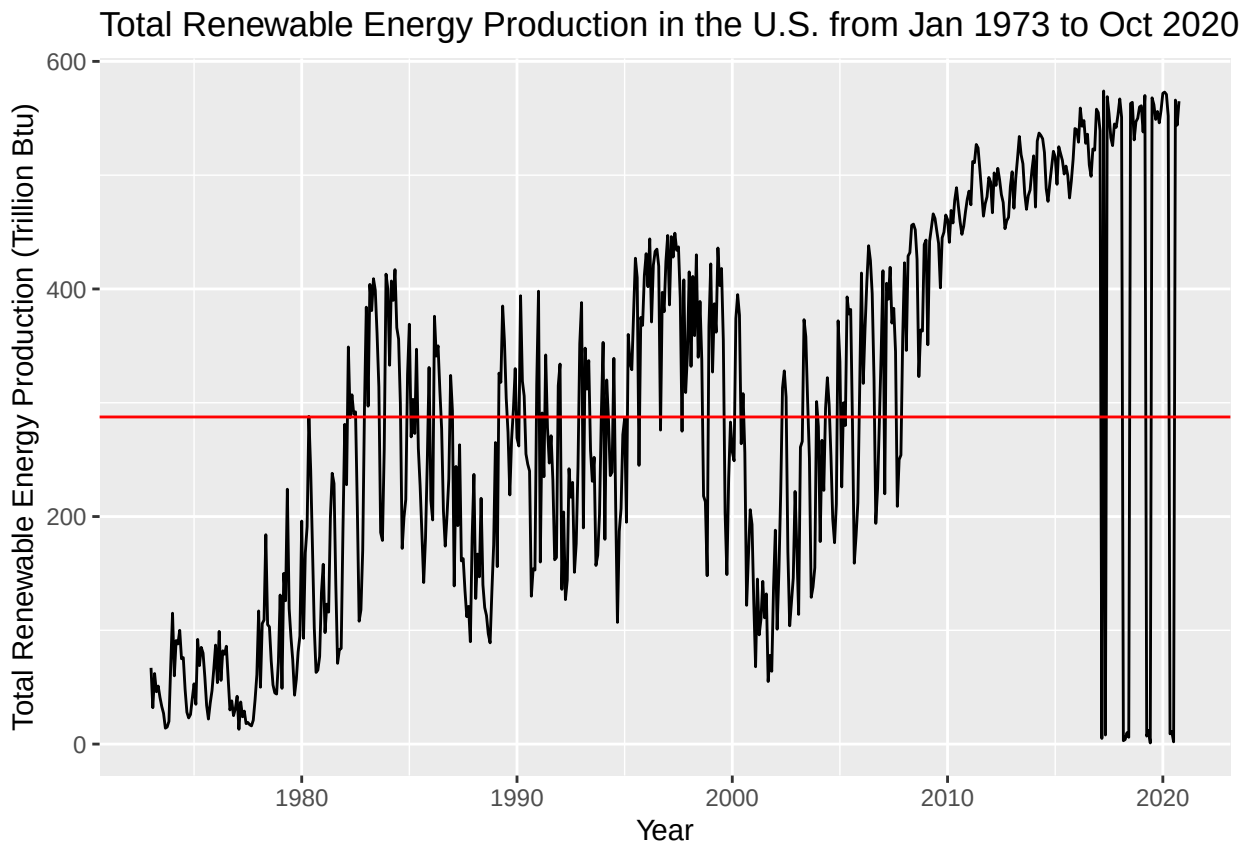
```
col_names <- colnames(ts_data)
p <- 1 # plot 1
autoplot(object = ts_data[,p],
         xlab = "Year",
         ylab = paste0(col_names[p], " (Trillion Btu)"),
         main = paste0(col_names[p]," in the U.S. from Jan 1973 to Oct 2020 ")) +
  geom_abline(intercept = col_mean[p], slope = 0, color = "red")
```



From the time series above, we can see that overall there has been an upward trend for the total biomass energy production (trillion Btu) in the United States based on the data from January 1973 to October 2020. The production level started to increase gradually in 1975 and reached 300 trillion Btu in 1985. The production level fluctuated between the 100 trillion Btu level and the 400 trillion Btu level between the year of 1985 and 2005. But after 2005, we saw the upward trend again. The production level reached a local maximum around 550 trillion Btu level between 2015 and 2020. The overall average for production between 1973 and 2020 is slightly below 300 trillion Btu.

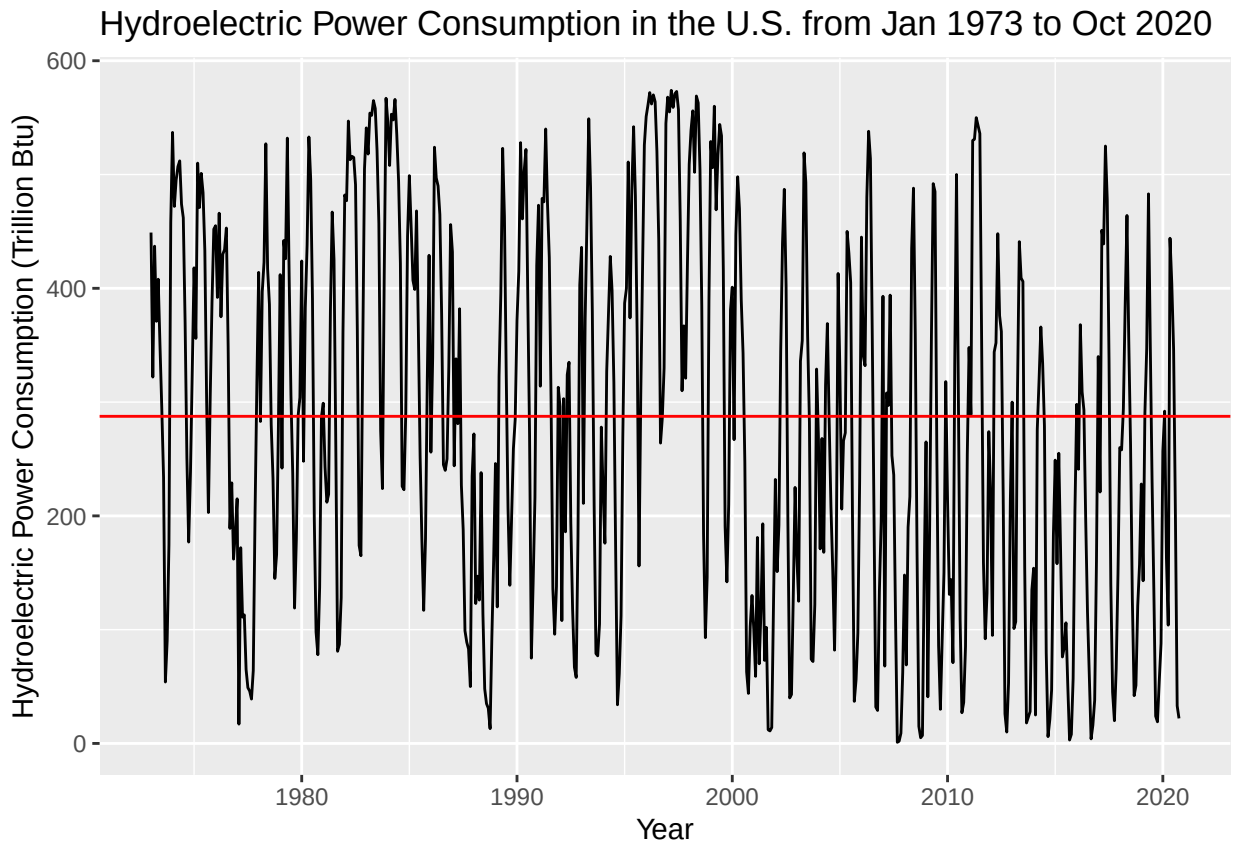
```
p <- 2 # plot 2
autoplot(object = ts_data[,p],
         xlab = "Year",
         ylab = paste0(col_names[p], " (Trillion Btu)"),
```

```
        main = paste0(col_names[p]," in the U.S. from Jan 1973 to Oct 2020 ")) +
  geom_abline(intercept = col_mean[p], slope = 0, color = "red")
```



From the time series above, we can see that overall there has been an upward trend for the total renewable energy production (trillion Btu) in the United States based on the data from January 1973 to October 2020. The overall trend is very similar to the trend of the total renewable energy production. The production level started to increase gradually in 1975 and reached 400 trillion Btu in 1984. The production level fluctuated between the 100 trillion Btu level and the 400 trillion Btu level between the year of 1985 and 2005. After 2005, we saw an upward trend again. The production level reached a local maximum of around 550 trillion Btu level around 2015. However, it seems the renewable energy production paused for a very short period of time for four times between 2016 and 2020. The overall average for production between 1973 and 2020 is slightly below 300 trillion Btu.

```
p <- 3 # plot 3
autoplot(object = ts_data[,p],
         xlab = "Year",
         ylab = paste0(col_names[p], " (Trillion Btu)"),
         main = paste0(col_names[p]," in the U.S. from Jan 1973 to Oct 2020 ")) +
  geom_abline(intercept = col_mean[p], slope = 0, color = "red")
```



From the time series plot above, we can see that the Hydroelectric Power Consumption has been fluctuate significantly throughout the period from 1975 to 2020. The overall average for consumption between 1973 and 2020 is slightly below 300 trillion Btu. No particular trend is observed.

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
round(cor(ts_data[,1], ts_data[,2]),4)
```

```
## [1] 0.7719
```

The correlation between Total Biomass Energy Production and Total Renewable Energy Production is 0.7719. This shows that Total Biomass Energy Production and Total Renewable Energy Production is strongly correlated since there correlation is above 0.75.

```
# Correlation between Total Biomass Energy Production
# and Hydroelectric Power Consumption
round(cor(ts_data[,1], ts_data[,3]),4)
```

```
## [1] -0.2476
```

This shows that there is some negative correlation between Total Biomass Energy Producton and Hydroelectric Power Consumption, but the correlation is weak as the correlation is only -0.2476.

```
# Correlation between Renewable Energy Production
# and Hydroelectric Power Consumption
round(cor(ts_data[,2], ts_data[,3]),4)
```

```
## [1] 0.0807
```

The correlation between Total Renewable Energy Production and Hydroelectric Power Consumption is 0.0807. This shows that Total Renewable Energy Production and Hydroelectric Power Consumption might not be correlated as the correlation is very close to 0.

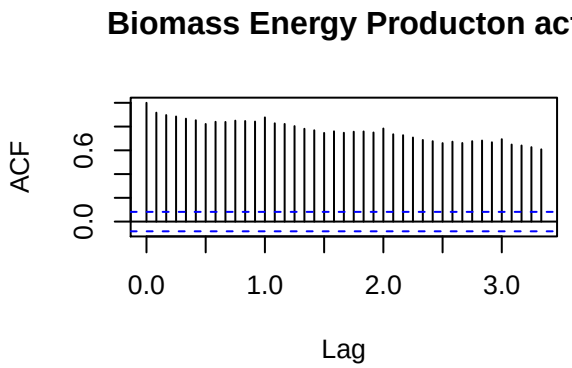
Question 6

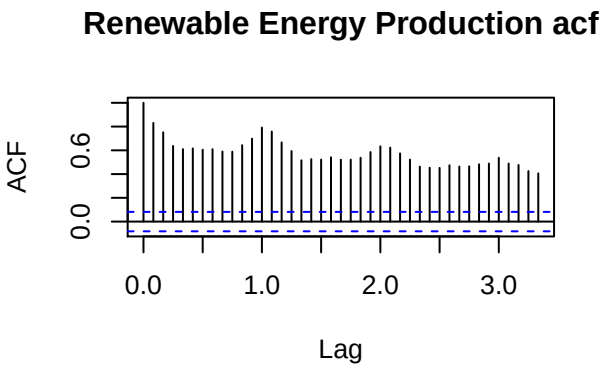
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

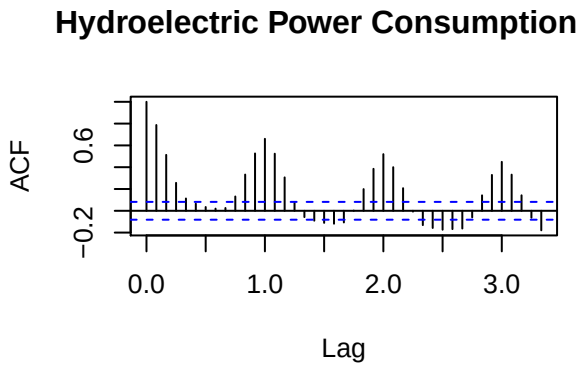
```
par(mfrow = c(2,2))
# Autocorrelation function for Total Biomass Energy Production
autocorr_1 <- acf(x = ts_data[,1], lag.max = 40,
                  type = "correlation", plot = FALSE)
plot(autocorr_1,main = paste0("Biomass Energy Producton acf"))

# Autocorrelation function for Total Renewable Energy Production
autocorr_2 <- acf(x = ts_data[,2], lag.max = 40,
                  type = "correlation", plot = FALSE)
plot(autocorr_2,main = paste0("Renewable Energy Production acf"))

# Autocorrelation function for Hydroelectric Power Consumption
autocorr_3 <- acf(x = ts_data[,3], lag.max = 40,
                  type = "correlation", plot = FALSE)
plot(autocorr_3,main = paste0(col_names[3], " acf"))
```







The autocorrelation functions for Biomass Energy Production and Renewable Energy Production have similar shape. Both functions decreased gradually until lag = 1, 2, or 3 where there is a small increase. Both autocorrelation functions decrease slowly, but the autocorrelation function for Renewable Energy Production decreases faster than that for the Biomass Energy Production. This means that the data points across these two time series are correlated.

The autocorrelation function for Hydroelectric Power Consumption has a different shape compared to the other two time series. It converges to 0 fast. This means that the data points across this time series is not strongly correlated.

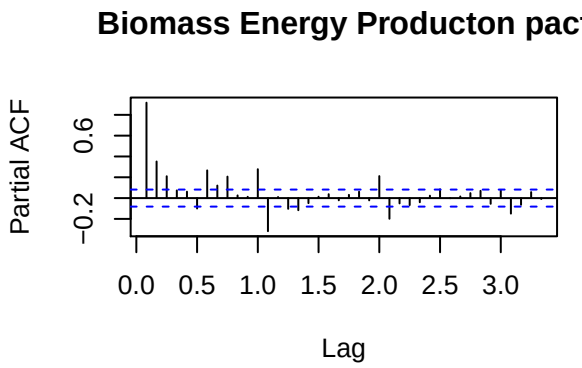
Question 7

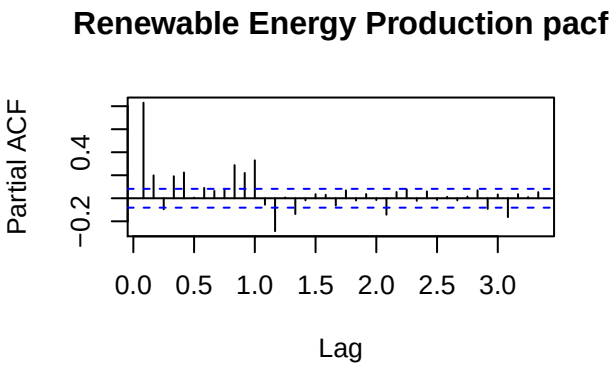
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

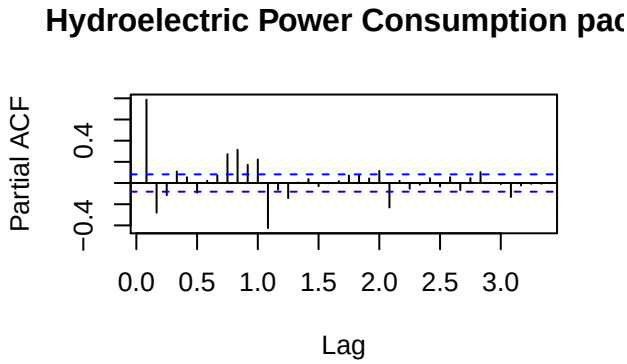
```
par(mfrow = c(2,2))
# Autocorrelation function for Total Biomass Energy Production
autocorr_1 <- acf(x = ts_data[,1], lag.max = 40,
                  type = "partial", plot = FALSE)
plot(autocorr_1,main = paste0("Biomass Energy Producton pacf"))

# Autocorrelation function for Total Renewable Energy Production
autocorr_2 <- acf(x = ts_data[,2], lag.max = 40,
                  type = "partial", plot = FALSE)
plot(autocorr_2,main = paste0("Renewable Energy Production pacf"))
```

```r
# Autocorrelation function for Hydroelectric Power Consumption
autocorr_3 <- acf(x = ts_data[,3], lag.max = 40,
                  type = "partial", plot = FALSE)
plot(autocorr_3,main = paste0(col_names[3], " pacf"))
```







These plots are very different compared to the autocorrelation plots in the Q6. All the partial autocorrelation function converge fast and fluctuate around 0.