ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 2 - Due date 02/03/23

Tony Jiang

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 packages here

library(forecast)
```

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

```
library(tseries)
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 December 2022 Monthly Energy Review. The spreadsheet is ready to be used. You will also find a *.csv* version of the data "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.csv". You may use the function *read.table()* to import the *.csv* data in R. Or refer to the file "M2_ImportingData_CSV_XLSX.Rmd" in our Lessons folder for functions that are better suited for importing the *.xlsx*.

```
#Importing data set and clean the data sets

df1 <- read_xlsx("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx", col_nar

df2 <- read.table(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.csv

df1 = df1[-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.

```
col_names = c("Total Biomass Energy Production",
              "Total Renewable Energy Production",
              "Hydroelectric Power Consumption")

df1_red = subset(df1, select = col_names)

head(df1_red)
```

```
## # A tibble: 6 x 3
##   `Total Biomass Energy Production` `Total Renewable Energy Production` Hydroe~1
##   <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
## # ... with abbreviated variable name 1: `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().

```
ts_df1 = ts(df1_red, start = c(1973, 1), frequency = 12)

head(ts_df1)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Jan 1973                              23                                 84
## Feb 1973                               2                                 49
## Mar 1973                              27                                 79
## Apr 1973                               9                                 63
## May 1973                              25                                 68
## Jun 1973                               8                                 58
##          Hydroelectric Power Consumption
## Jan 1973                             473
## Feb 1973                             346
## Mar 1973                             461
## Apr 1973                             395
## May 1973                             431
## Jun 1973                             374
```

Question 3

Compute mean and standard deviation for these three series.

```
a = c(colnames(ts_df1))

mean_list = data.frame(
  c("Mean of Biomass Energy Production",
```

```
    "Mean of Renewable Energy Production",
    "Mean of Hydroelectric Power Consumption"),
  c(mean(ts_df1[, "Total Biomass Energy Production"]),
    mean(ts_df1[, "Total Renewable Energy Production"]),
    mean(ts_df1[, "Hydroelectric Power Consumption"])
    )
  )

mean_list
```

```
##   c..Mean.of.Biomass.Energy.Production....Mean.of.Renewable.Energy.Production...
## 1                                          Mean of Biomass Energy Production
## 2                                       Mean of Renewable Energy Production
## 3                                    Mean of Hydroelectric Power Consumption
##   c.mean.ts_df1....Total.Biomass.Energy.Production.....mean.ts_df1...
## 1                                                            297.5662
## 2                                                            299.0000
## 3                                                            299.0000
```

```
std_list = data.frame(
  c("Std of Biomass Energy Production",
    "Std of Renewable Energy Production",
    "Std of Hydroelectric Power Consumption"),
  c(sd(ts_df1[, "Total Biomass Energy Production"]),
    sd(ts_df1[, "Total Renewable Energy Production"]),
    sd(ts_df1[, "Hydroelectric Power Consumption"])
    )
  )

std_list
```

```
##   c..Std.of.Biomass.Energy.Production....Std.of.Renewable.Energy.Production...
## 1                                           Std of Biomass Energy Production
## 2                                        Std of Renewable Energy Production
## 3                                     Std of Hydroelectric Power Consumption
##   c.sd.ts_df1....Total.Biomass.Energy.Production.....sd.ts_df1...
## 1                                                         171.9873
## 2                                                         172.4833
## 3                                                         172.4833
```

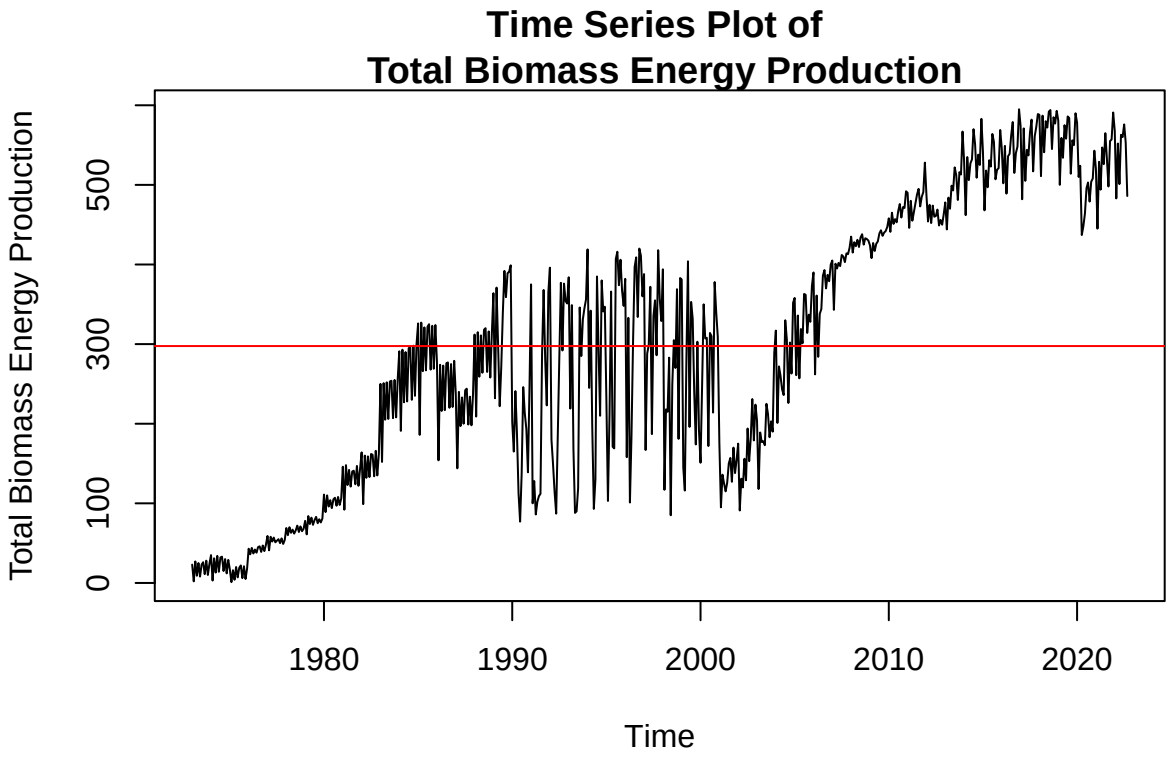
Question 4

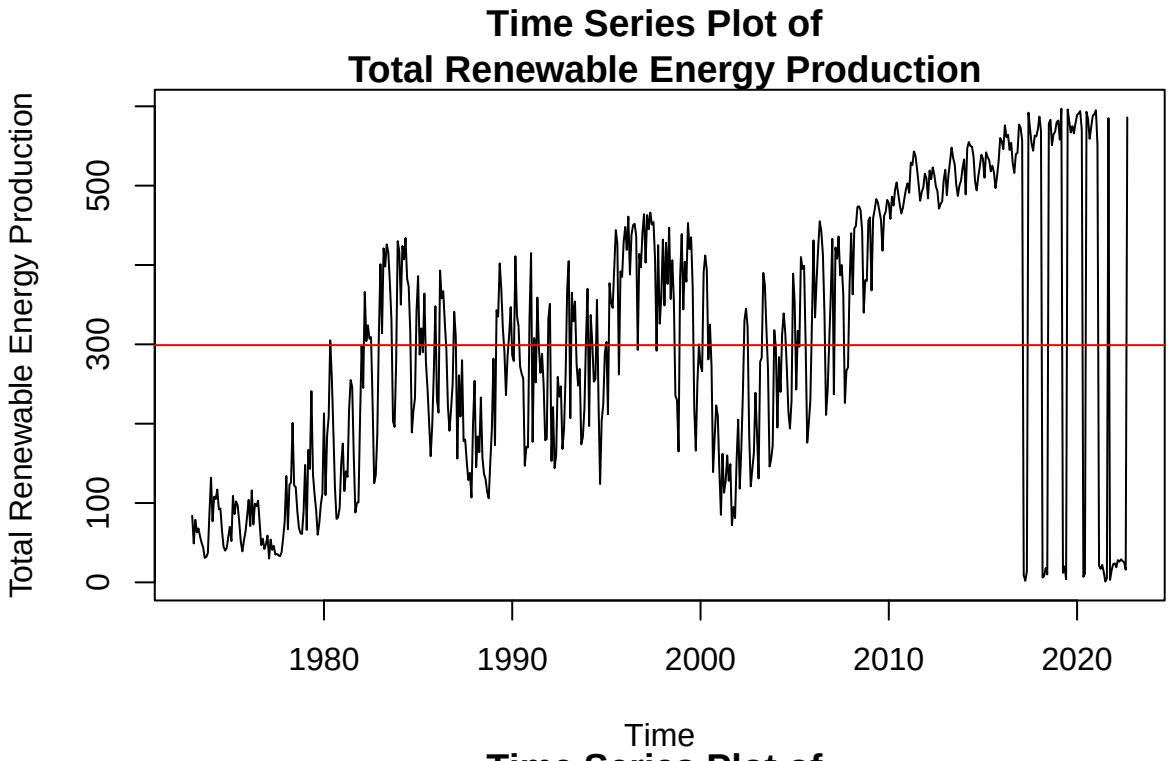
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

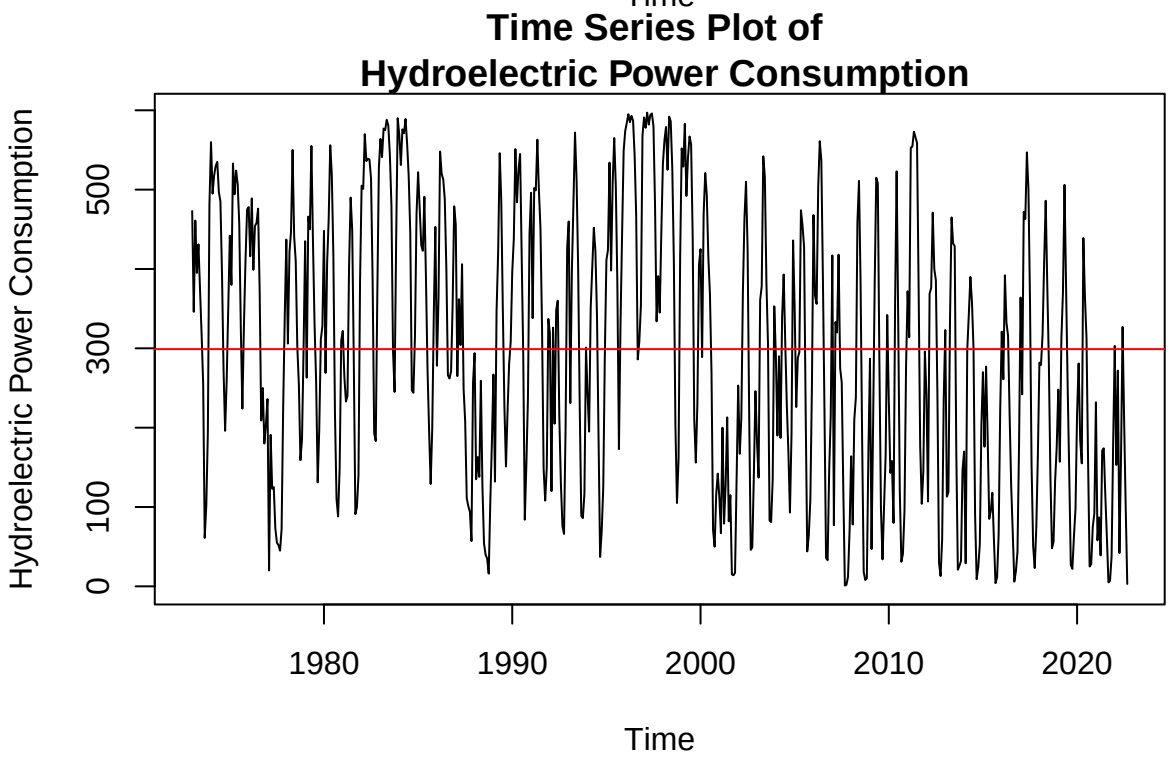
Answers to Q4: 1.The time series plot of biomass energy production generally shows an upward trend over time. But in the middle, the trend is stagnant. This may be caused by the change of regulation environment. 2. The time series plot of renewable energy production also shows an upward trend over time. Meanwhile, it shows seasonality within each year. 3. The time series plot of hydroelectric power consumption shows great seasonality. It is hard to say whether it grows or decreases over time. But I suspect it decreases a little bit because, in recent years, the consumption is generally below the mean value.

```
for (i in c(1:3)) {

plot(ts_df1[ , a[i]], type = "l",
```

```
     xlab = "Time",
     ylab = a[i])

abline(h=mean_list[[i,2]], col = "red")

title(main = paste("Time Series Plot of \n",a[i]), line = 0.2)
}
```







Question 5

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

Answers to Q5: They are not strongly correlated to each other. Among the three, biomass energy production and renewable energy production are the most correlated. This makes sense because biomass energy is a subset of renewable energy. Thus, the production of them may be affected by the same exogenous

factors. Hydroelectric consumption and biomass energy production have a weak negative correlation. This can be explaned as they are competing energy supply in the market. So they crowd our each other in the market. When more biomass energy available, less hydroeletric power will be consumed. Hydroeletric power consumption and renewable energy production are almost not correlated. generally speaking, there are too many factors that can affect these two variables.

```
cor(ts_df1)
```

```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.6471276
## Hydroelectric Power Consumption                       -0.2902430
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                          0.6471276
## Total Renewable Energy Production                        1.0000000
## Hydroelectric Power Consumption                          0.0874112
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                       -0.2902430
## Total Renewable Energy Production                      0.0874112
## Hydroelectric Power Consumption                        1.0000000
```

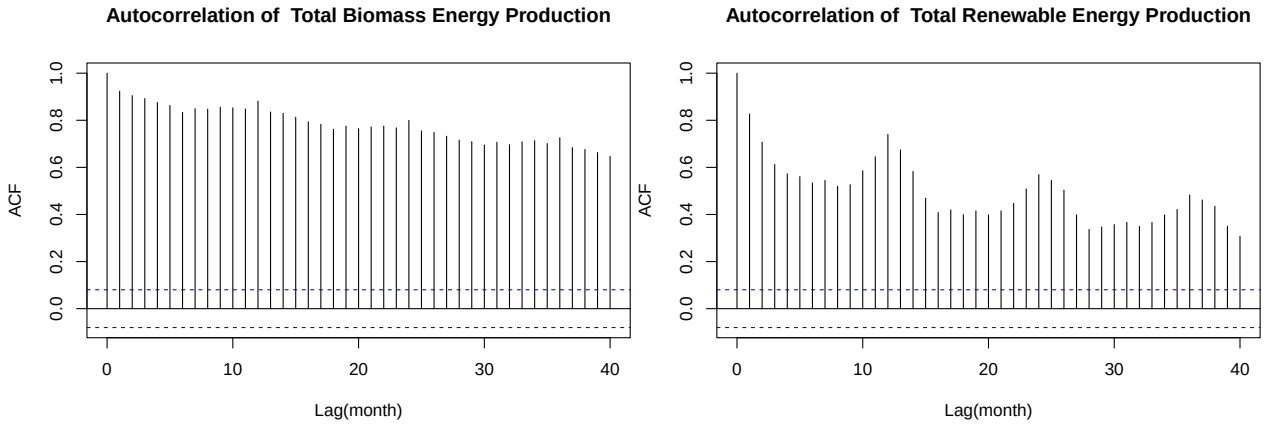
Question 6

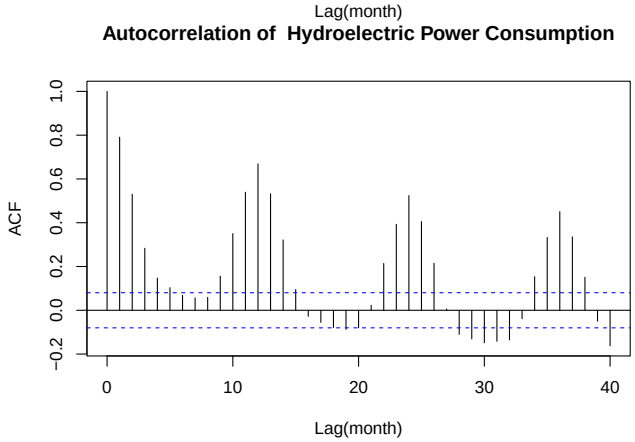
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

Answers to Q6: They all have two similar behaviors. 1. Autocorrelation decreases as lags increase. 2. Their autocorrelation equals 1 at a lag of 0. (Because it's actually self-correlation.)

In addition to similarities, they also have different behaviors. Total biomass energy production has strong autocorrelation for all lags, though its autocorrelation decreases as lag increases. This may indicate its stability over time. Total renewable energy production has a weak but obvious cyclical pattern, with an interval of 12. Hydroelectric power consumption has a strong seasonality, which can be attributed to the seasonality of natural water. E.g. more in summer and less in winter.

```
for (i in c(1:3)) {

fig = acf(ts_df1[, a[i]], lag.max = 40, plot = FALSE)

fig$series <- NULL

fig$lag[1:41] = c(0:40)

fig %>% plot(main = paste("Autocorrelation of ", a[i]),
             xlab = "Lag(month)")
}
```



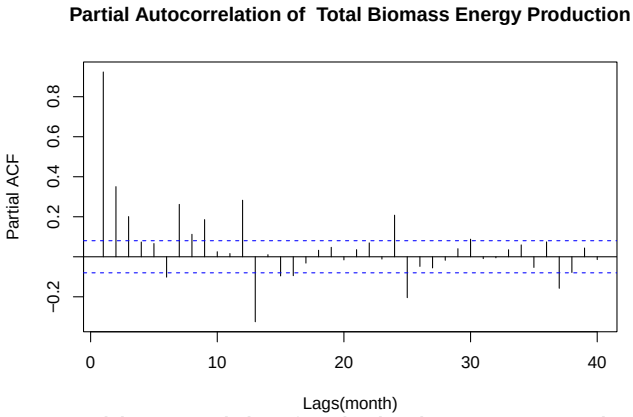


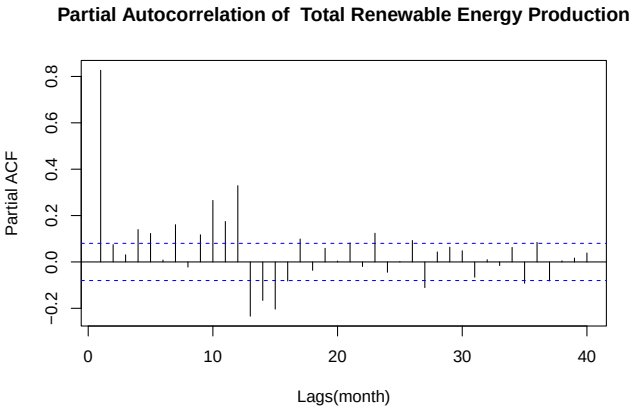
Question 7

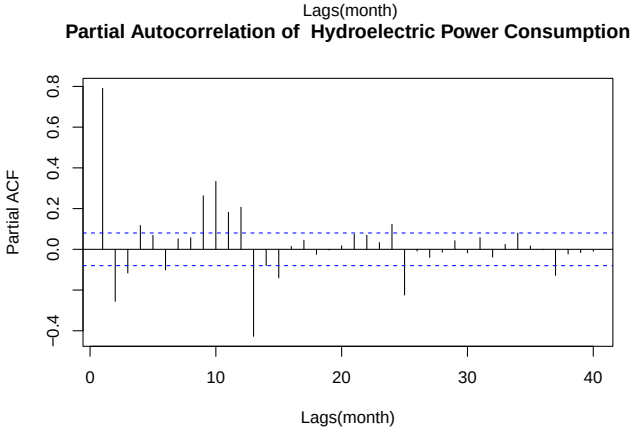
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

Answers to Q7: Partial autocorrelation is consistently less significant than autocorrelation. This characteristic is because the impacts of intermediate series are cleared from the calculation. Also, seasonality in each series becomes less obvious in partial autocorrelation. Moreover, some lags have positive autocorrelation but negative partial correlation, which may indicate that the true relationship between two series are biased by autocorrelation since the effects of intermediate series are included.

```
for (i in c(1:3)) {

fig1 = pacf(ts_df1[, a[i]], lag.max = 40, plot = FALSE)

fig1$series <- NULL

fig1$lag[1:40] <- c(1:40)

fig1 %>% plot(, main = paste("Partial Autocorrelation of ", a[i]),
              xlab = "Lags(month)")
}
```


Partial Autocorrelation of Total Biomass Energy Production
Partial ACF
−0.2
0.2
0.4
0.6
0.8
0
10
20
30
40
Lags(month)


Partial Autocorrelation of Total Renewable Energy Production
Partial ACF
−0.2
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lags(month)


Partial Autocorrelation of Hydroelectric Power Consumption
Partial ACF
−0.4
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lags(month)