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

Assignment 2 - Due date 02/03/23

Kassie Huang

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 rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp23.Rmd"). Then 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. 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(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
```

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
library(readxl)
energy_data <- read_excel(path='./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
colnames(energy_data) <- read_excel(path='./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
head(energy_data)
```

```
## # A tibble: 6 x 14
##   Month               Wood Ene~1 Biofu~2 Total~3 Total~4 Hydro~5 Geoth~6 Solar~7
##   <dttm>                   <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl> <chr>
## 1 1973-01-01 00:00:00       130. Not Av~    130.    404.    273.    1.49 Not Av~
## 2 1973-02-01 00:00:00       117. Not Av~    117.    361.    242.    1.36 Not Av~
## 3 1973-03-01 00:00:00       130. Not Av~    130.    400.    269.    1.41 Not Av~
## 4 1973-04-01 00:00:00       125. Not Av~    126.    380.    253.    1.65 Not Av~
## 5 1973-05-01 00:00:00       130. Not Av~    130.    392.    261.    1.54 Not Av~
## 6 1973-06-01 00:00:00       125. Not Av~    126.    377.    250.    1.76 Not Av~
## # ... with 6 more variables: `Wind Energy Consumption` <chr>,
## #   `Wood Energy Consumption` <dbl>, `Waste Energy Consumption` <dbl>,
## #   `Biofuels Consumption` <chr>, `Total Biomass Energy Consumption` <dbl>,
## #   `Total Renewable Energy Consumption` <dbl>, and abbreviated variable names
## #   1: `Wood Energy Production`, 2: `Biofuels Production`,
## #   3: `Total Biomass Energy Production`,
## #   4: `Total Renewable Energy Production`, ...
```

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.

```
energy_data2 <- as.data.frame(energy_data[,4:6])
datanames <- colnames(energy_data2)
head(energy_data2)
```

```
##   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
```

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_energy_data2 <- ts(energy_data2,start = c(1973,1), frequency = 12)

#check the incomplete 2022 data are correctly imported
tail(ts_energy_data2,10)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Dec 2021                         455.075                          1118.251
## Jan 2022                         436.735                          1130.434
## Feb 2022                         397.763                          1071.921
## Mar 2022                         431.900                          1209.837
## Apr 2022                         406.080                          1176.084
## May 2022                         434.982                          1219.656
## Jun 2022                         434.199                          1187.064
## Jul 2022                         440.585                          1132.137
## Aug 2022                         431.969                          1043.955
## Sep 2022                         398.696                           976.813
##          Hydroelectric Power Consumption
## Dec 2021                         208.358
## Jan 2022                         232.468
## Feb 2022                         203.135
## Mar 2022                          225.165
## Apr 2022                          172.956
## May 2022                          204.198
## Jun 2022                          237.735
## Jul 2022                          213.172
## Aug 2022                          191.155
## Sep 2022                          148.808
```

Question 3

Compute mean and standard deviation for these three series.

```
#create empty vectors to store means and SDs
means <- c()
SDs <- c()
#indexs
ndata <- ncol(energy_data2)

for(i in 1:ndata){
  means <- c(means,mean(energy_data2[,i]))
  SDs <- c(SDs,sd(energy_data2[,i]))
}

mean_SD <- rbind(means,SDs)
colnames(mean_SD) <- datanames
mean_SD
```

```
##       Total Biomass Energy Production Total Renewable Energy Production
## means                       277.25252                          592.1583
## SDs                          91.75367                          191.7978
##       Hydroelectric Power Consumption
## means                       235.11465
## SDs                          44.16116
```

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.

```
#installed ggfortify previously in console following
#http://www.sthda.com/english/wiki/ggfortify-extension-to-ggplot2-to-handle-some-popular-packages-r-sof

library(ggfortify)
```

```
## Loading required package: ggplot2
```

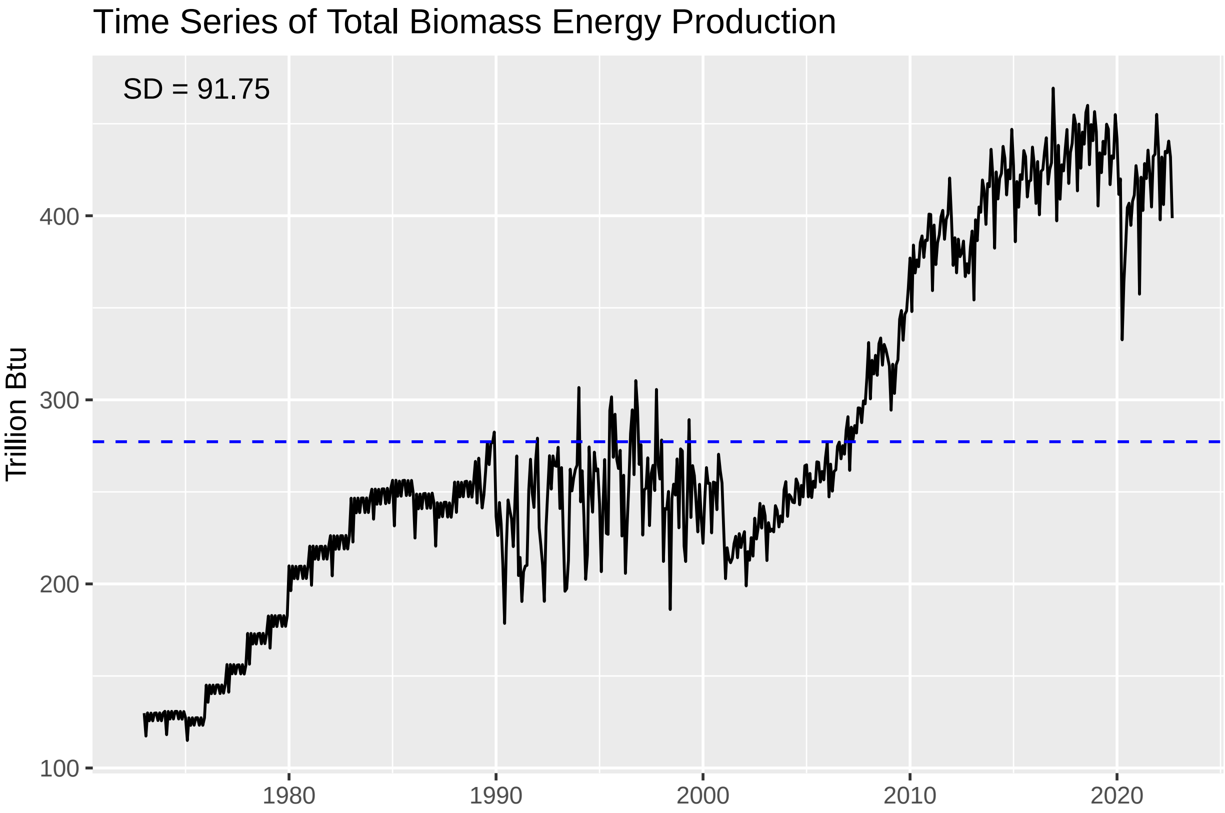
```
## Registered S3 methods overwritten by 'ggfortify':
##   method                 from
##   autoplot.Arima         forecast
##   autoplot.acf           forecast
##   autoplot.ar            forecast
##   autoplot.bats          forecast
##   autoplot.decomposed.ts forecast
##   autoplot.ets           forecast
##   autoplot.forecast      forecast
##   autoplot.stl               forecast
##   autoplot.ts                forecast
##   fitted.ar                  forecast
##   fortify.ts                 forecast
##   residuals.ar               forecast
```

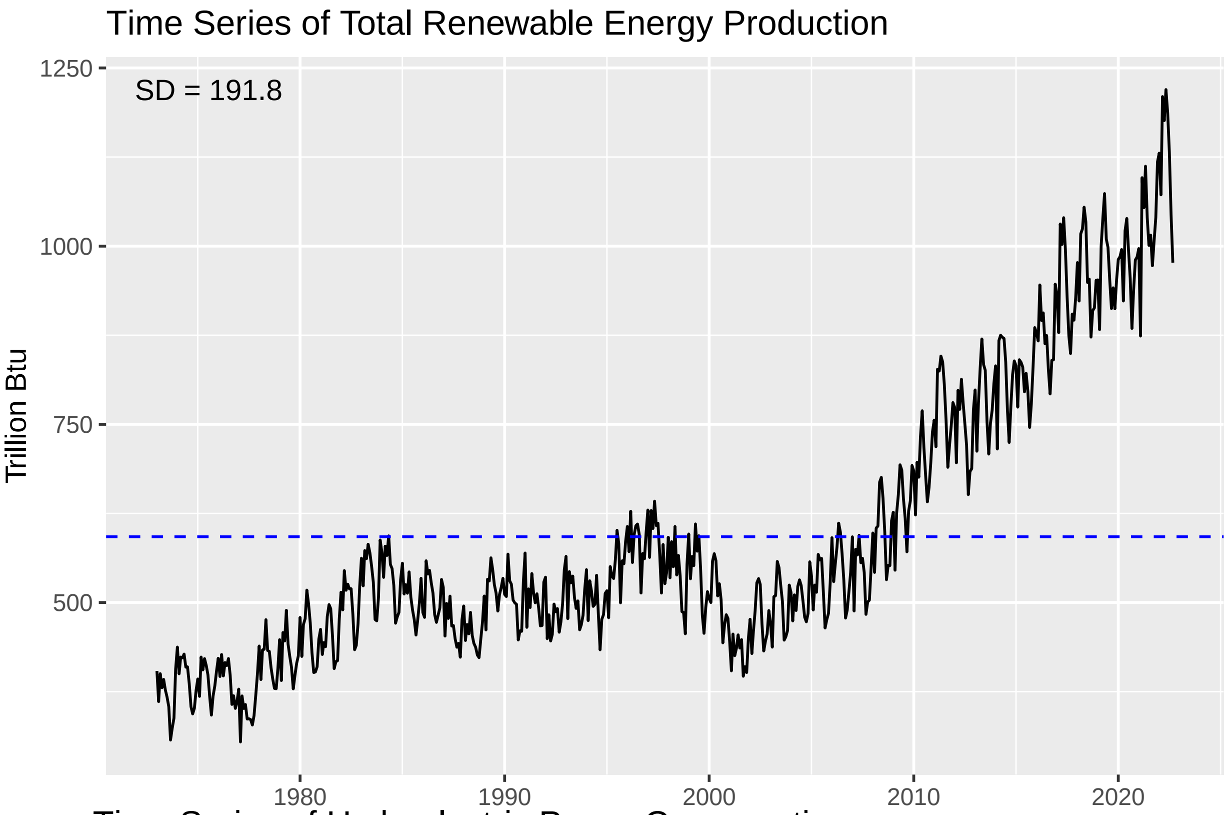
```
library(lubridate)
```

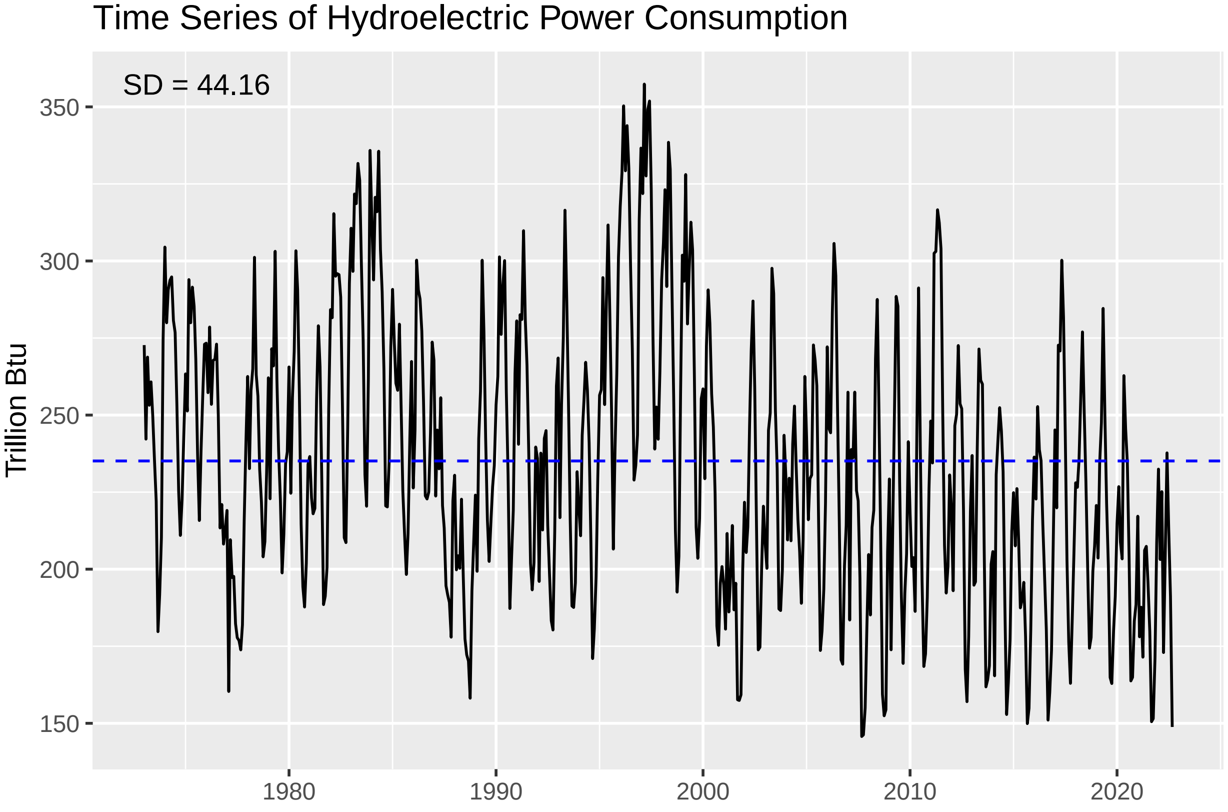
```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
for(i in 1:ndata){
  print(autoplot(ts_energy_data2[,i])+
          geom_hline(yintercept = means[i],linetype="dashed",color='blue')+
          labs(title = paste('Time Series of',datanames[i],sep=' '),
               y='Trillion Btu')+
        annotate("text", x=as_date(2020) ,y = max(ts_energy_data2[,i]), label = paste("SD =",round(SDs[
}
```







The time series plot of the total biomass energy production showed a non-linear increase trend. The pre-1990 data show some subtle seasonality with a more pronounced valley and several smaller waveform, while the post-1990 data become more volatile and it is difficult to observe any seasonality. The time series plot of the total renewable energy production showed a non-linear increase trend. There may also be some seasonal components in this data series, which may be related to the seasonality of sunlight and wind energy, but it's not clear in looking the plot. The time series plot of the hydroelectric power consumption didn't show a clear trend and it is dominated by strong seasonal patterns.

Question 5

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

```
#Total Biomass Energy Production ~ Renewable Energy Production
cor_Biomass_Renew <- lm(`Total Biomass Energy Production`~`Total Renewable Energy Production`,data=energ
summary(cor_Biomass_Renew)
```

```
##
## Call:
## lm(formula = `Total Biomass Energy Production` ~ `Total Renewable Energy Production`,
##     data = energy_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.775  -23.753    3.474   26.182   79.099
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         17.032186   4.823999   3.531 0.000446 ***
## `Total Renewable Energy Production`  0.439444   0.007751  56.697  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.29 on 595 degrees of freedom
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8436
## F-statistic:  3215 on 1 and 595 DF,  p-value: < 2.2e-16
```

The P-value is 2.2e-16, which is less than 0.001, indicating that the total renewable energy production is significantly correlated with the total biomass energy production. Meanwhile, the coefficient is equal to 0.44, which is greater than 0, indicating that the total renewable energy production is positively correlated with the biomass energy production.

```
#Total Renewable energy production ~ Hydroelectric Power Consumption
cor_Renew_Hydro <- lm(`Total Renewable Energy Production`~`Hydroelectric Power Consumption`,data=energy_
summary(cor_Renew_Hydro)
```

```
##
## Call:
## lm(formula = `Total Renewable Energy Production` ~ `Hydroelectric Power Consumption`,
##     data = energy_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -320.18 -138.68  -56.52   85.59  614.13
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       693.8506    42.3815  16.372   <2e-16 ***
## `Hydroelectric Power Consumption`  -0.4325     0.1772  -2.441   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 191 on 595 degrees of freedom
## Multiple R-squared:  0.009918,   Adjusted R-squared:  0.008254
## F-statistic:  5.96 on 1 and 595 DF,  p-value: 0.01492
```

The P-value is 0.01492, which is less than 0.05, indicating that the total renewable energy production is significantly correlated with the hydroelectric power consumption. Meanwhile, the coefficient is equal to -0.43, which is less than 0, indicating that the total renewable energy production is positively correlated with the hydroelectric power consumption.

```
#Total Biomass Energy Production ~ Hydroelectric Power Consumption
cor_Biomass_Hydro <- lm(`Total Biomass Energy Production`~`Hydroelectric Power Consumption`,data=energy_
summary(cor_Biomass_Hydro)
```

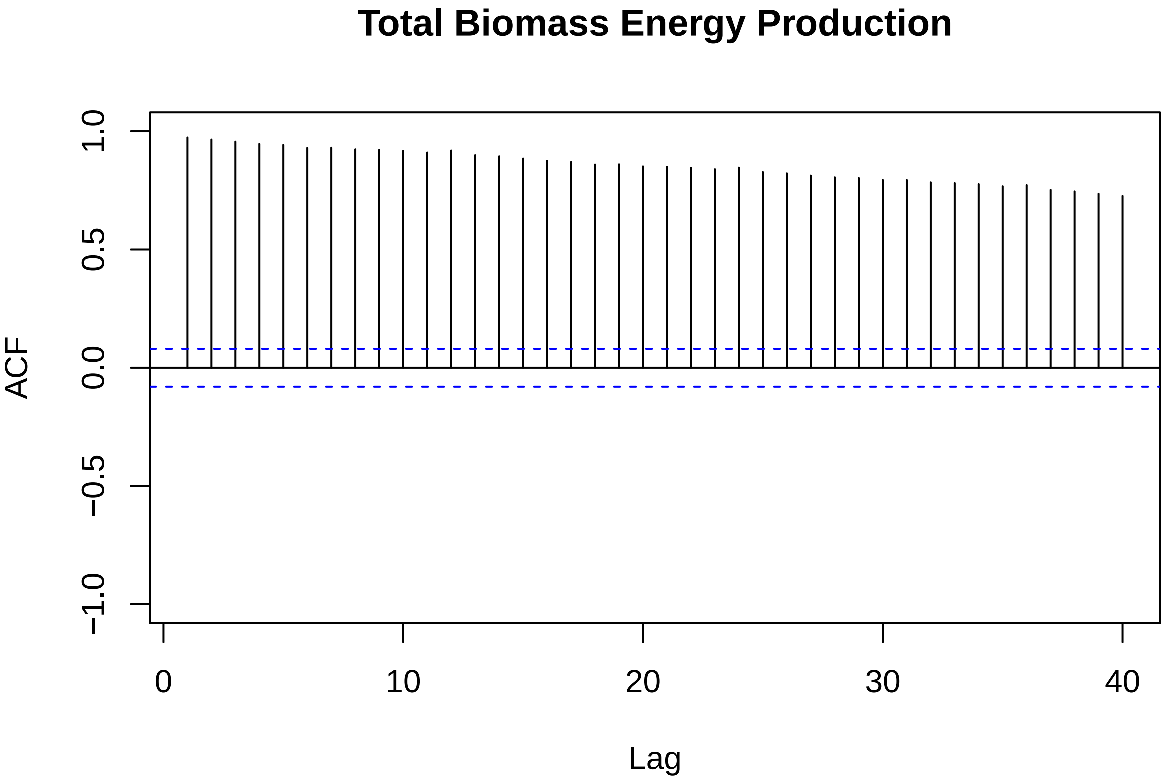
```
##
## Call:
## lm(formula = `Total Biomass Energy Production` ~ `Hydroelectric Power Consumption`,
##     data = energy_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185.97  -56.43  -15.37   62.30  194.26
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       423.71401   19.43869  21.797  < 2e-16 ***
## `Hydroelectric Power Consumption`  -0.62294    0.08126  -7.666 7.26e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.61 on 595 degrees of freedom
## Multiple R-squared:  0.08989,    Adjusted R-squared:  0.08836
## F-statistic: 58.77 on 1 and 595 DF,  p-value: 7.256e-14
```

The P-value is 7.26e-14, which is less than 0.001, indicating that the total biomass energy production is significantly correlated with the hydroelectric power consumption. Meanwhile, the coefficient is equal to -0.62, which is less than 0, indicating that the total biomass energy production is positively correlated with the hydroelectric power consumption.

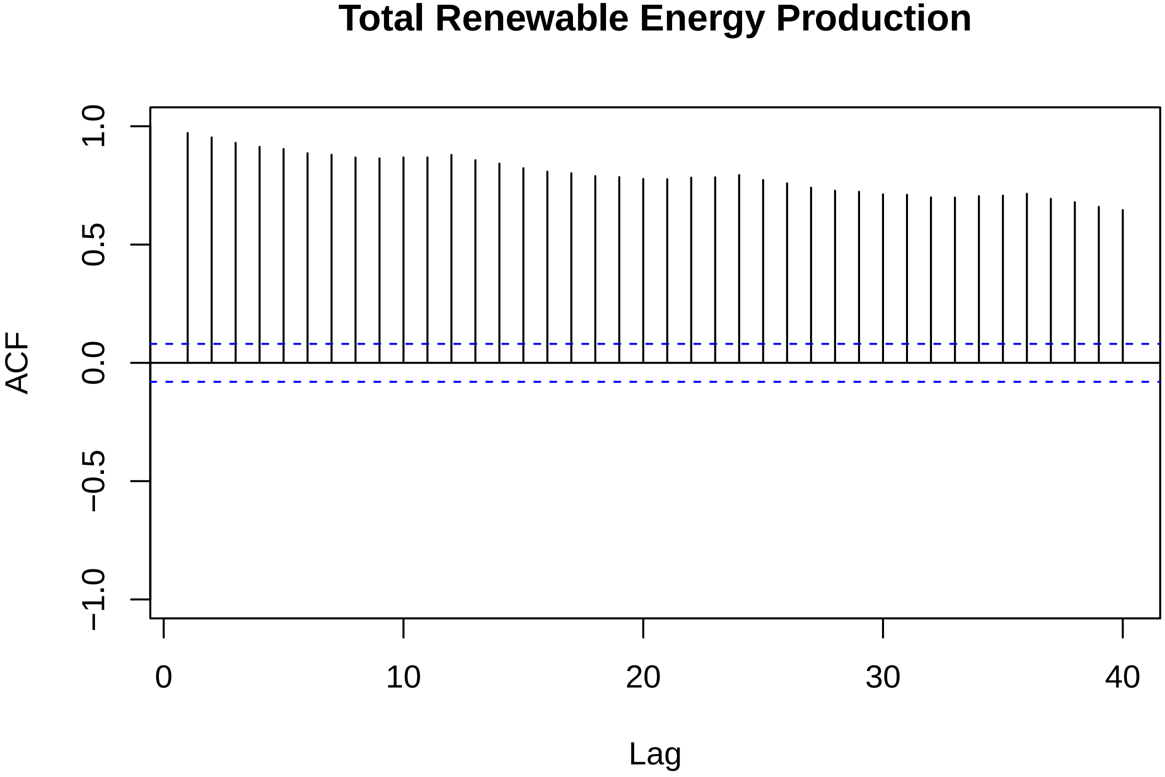
Question 6

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?

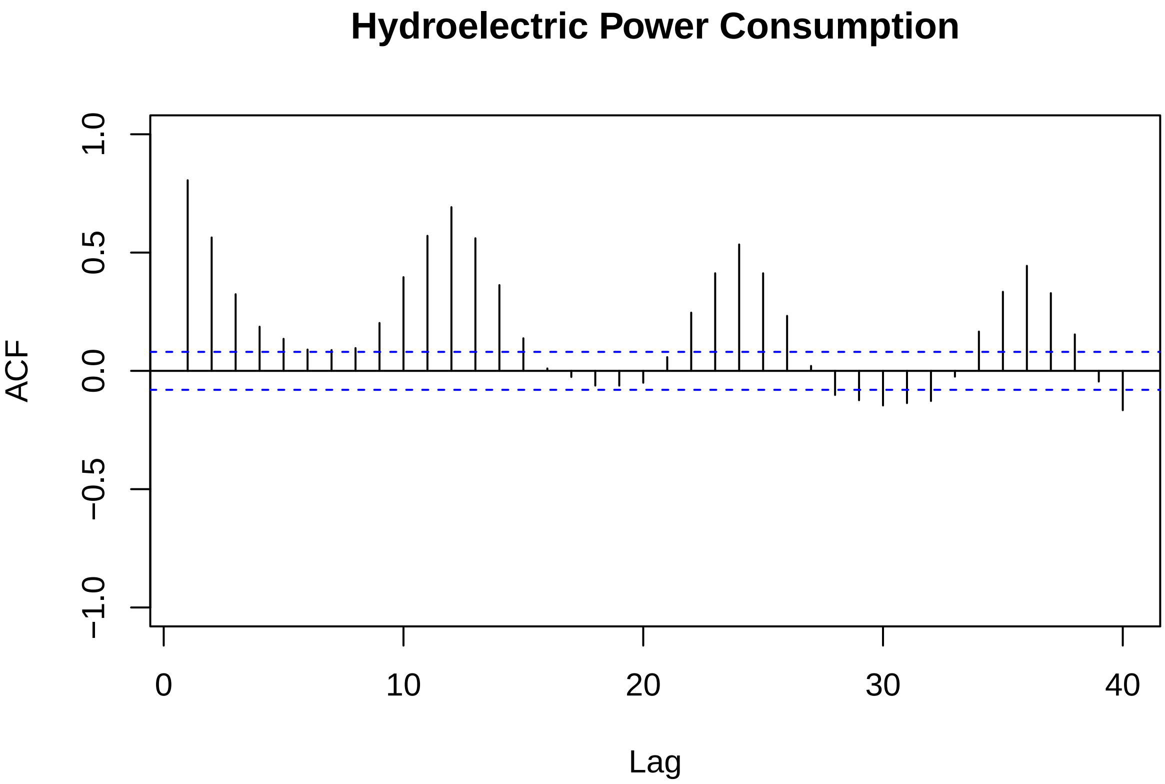
```
for(i in 1:ndata){
  Acf(energy_data2[,i],lag.max = 40,main=datanames[i],ylim=c(-1,1))
}
```



Total Biomass Energy Production



Total Renewable Energy Production

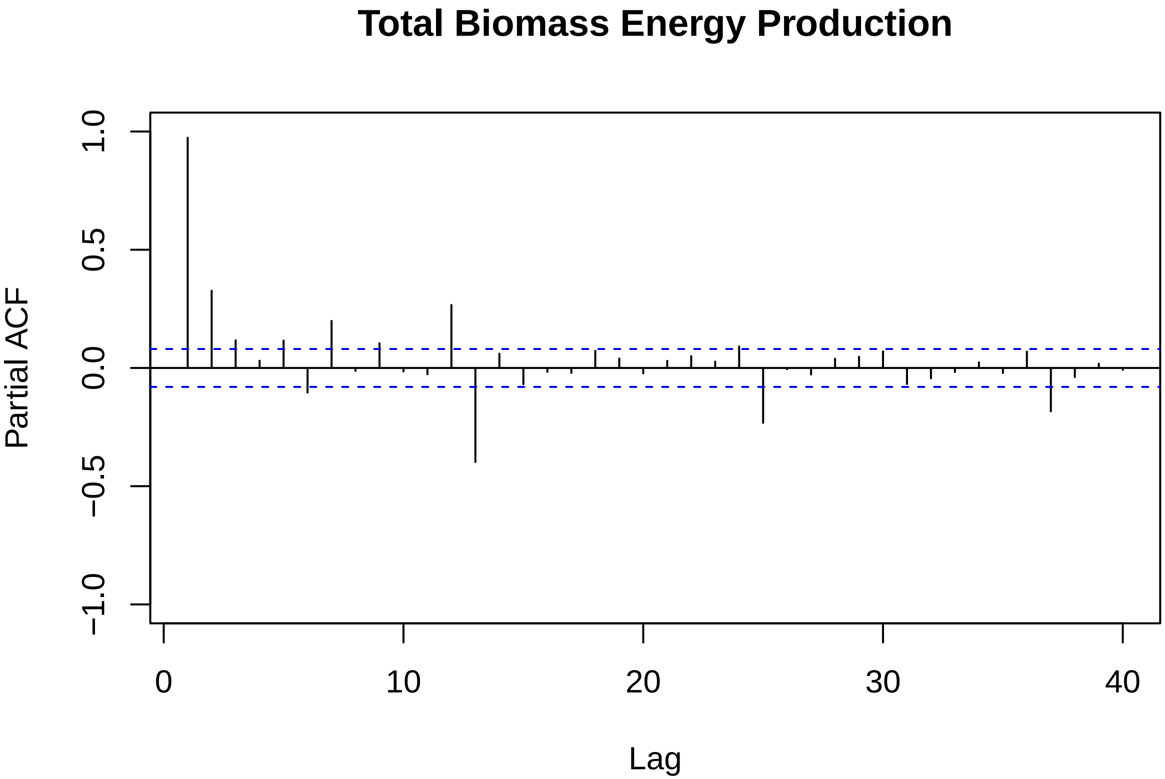


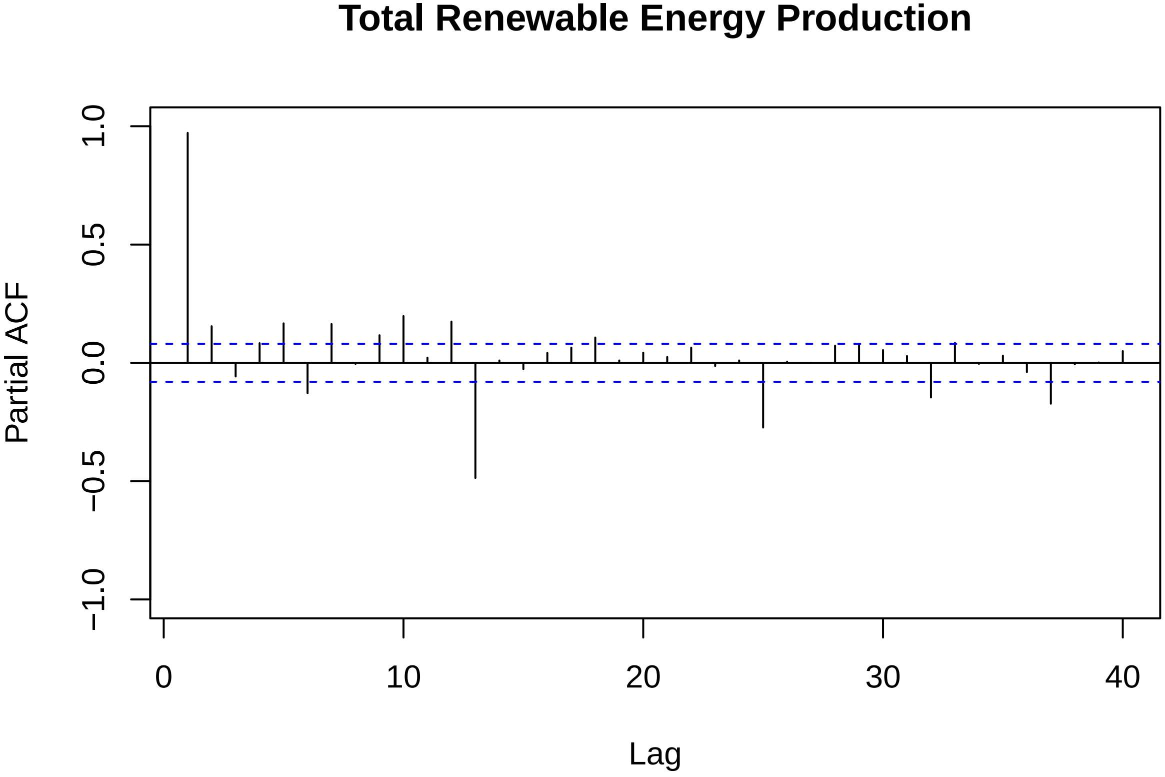
From the ACF plot of the total biomass energy production, this time series exhibits a strong auto-correlation. The ACF decays as lag increases, but even at lag=40, the value of ACF is still as high as ~0.7. Similarly, the second plot of the total renewable energy production also showed a stronger auto-correlation, although smaller than the first one. The hydroelectric power consumption, observed from its ACF plot, shows a different patter. It shows a clear waveform, but the value of the crest gradually decreases, which means that there is a seasonality in the data.

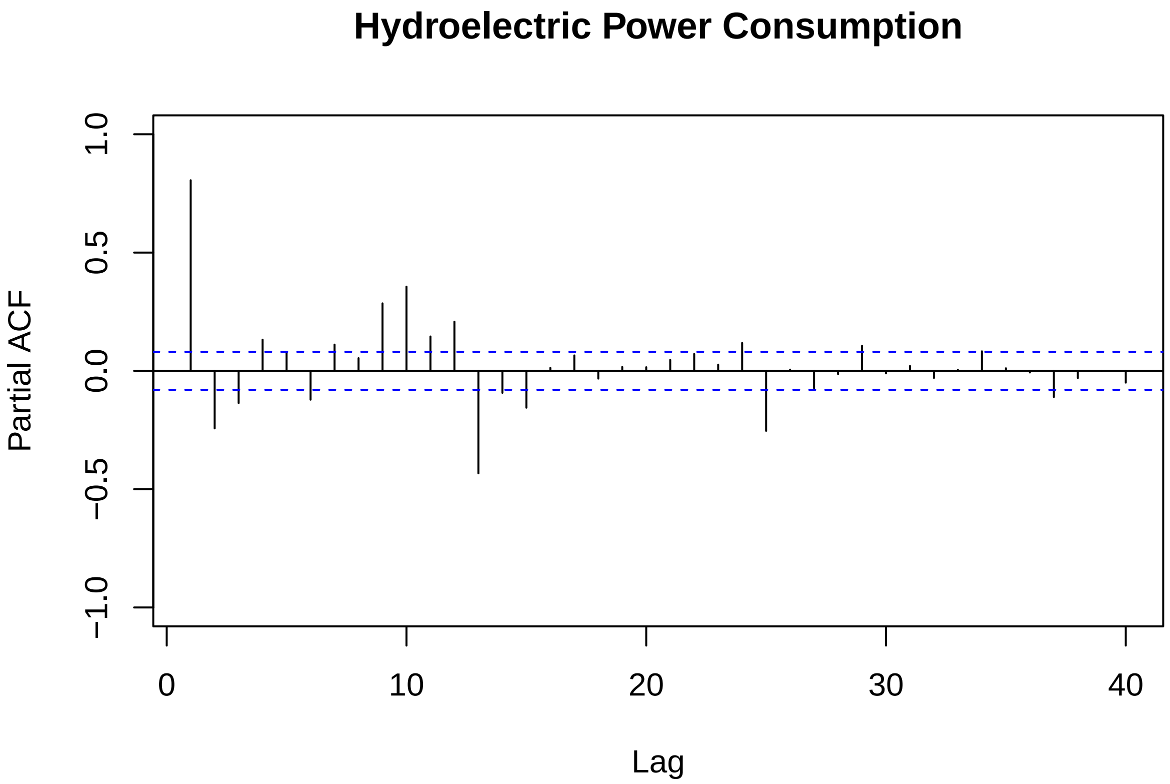
Question 7

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

```
for(i in 1:ndata){
  Pacf(energy_data2[,i],lag.max = 40,main=datanames[i],ylim=c(-1,1))
}
```


Total Biomass Energy Production
Partial ACF
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
Lag


Total Renewable Energy Production
Partial ACF
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
Lag



Compared to Q6, all three plots show the same change: the values of Partial ACF are significantly reduced to within the significant interval, except for lag=1 which remains unchanged. For biomass energy and renewable energy production, which previously showed only positive values in the ACF plot, their PACFs both showed several significant negative values. These negative auto-correlations may have been previously masked by the strong positive correlation for lag=1. In addition, the waveform presented in the ACF plot of hydroelectric power consumption almost disappear.