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

Assignment 2 - Due date 02/03/23

Hugh Cipparone

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
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0     v purrr   0.3.5
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.3
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, 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

data<-read.csv("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.csv")
```

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.

```
date<-ym(data$Month)

data.edit<-cbind(data,date)%>%
  select(date,Total.Biomass.Energy.Production, Total.Renewable.Energy.Production, Hydroelectric.Power.C

head(data.edit)
```

```
##         date Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1 1973-01-01                          129.787                             403.981
## 2 1973-02-01                          117.338                             360.900
## 3 1973-03-01                          129.938                             400.161
## 4 1973-04-01                          125.636                             380.470
## 5 1973-05-01                          129.834                             392.141
## 6 1973-06-01                          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 <- ts(data.edit[,(2:4)], frequency=12, start=c(1931,1))
head(ts_energy)
```

```
##          Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## Jan 1931                         129.787                             403.981
## Feb 1931                         117.338                             360.900
## Mar 1931                         129.938                             400.161
## Apr 1931                         125.636                             380.470
## May 1931                         129.834                             392.141
## Jun 1931                         125.611                             377.232
##          Hydroelectric.Power.Consumption
## Jan 1931                         272.703
## Feb 1931                         242.199
## Mar 1931                         268.810
## Apr 1931                         253.185
## May 1931                         260.770
## Jun 1931                         249.859
```

Question 3

Compute mean and standard deviation for these three series.

```
mean.biomass<-mean(data.edit$Total.Biomass.Energy.Production)
sd(data.edit$Total.Biomass.Energy.Production)
```

```
## [1] 91.75367
```

```
mean.renewable<-mean(data.edit$Total.Renewable.Energy.Production)
sd(data.edit$Total.Renewable.Energy.Production)
```

```
## [1] 191.7978
```

```
mean.hydro<-mean(data.edit$Hydroelectric.Power.Consumption)
sd(data.edit$Hydroelectric.Power.Consumption)
```

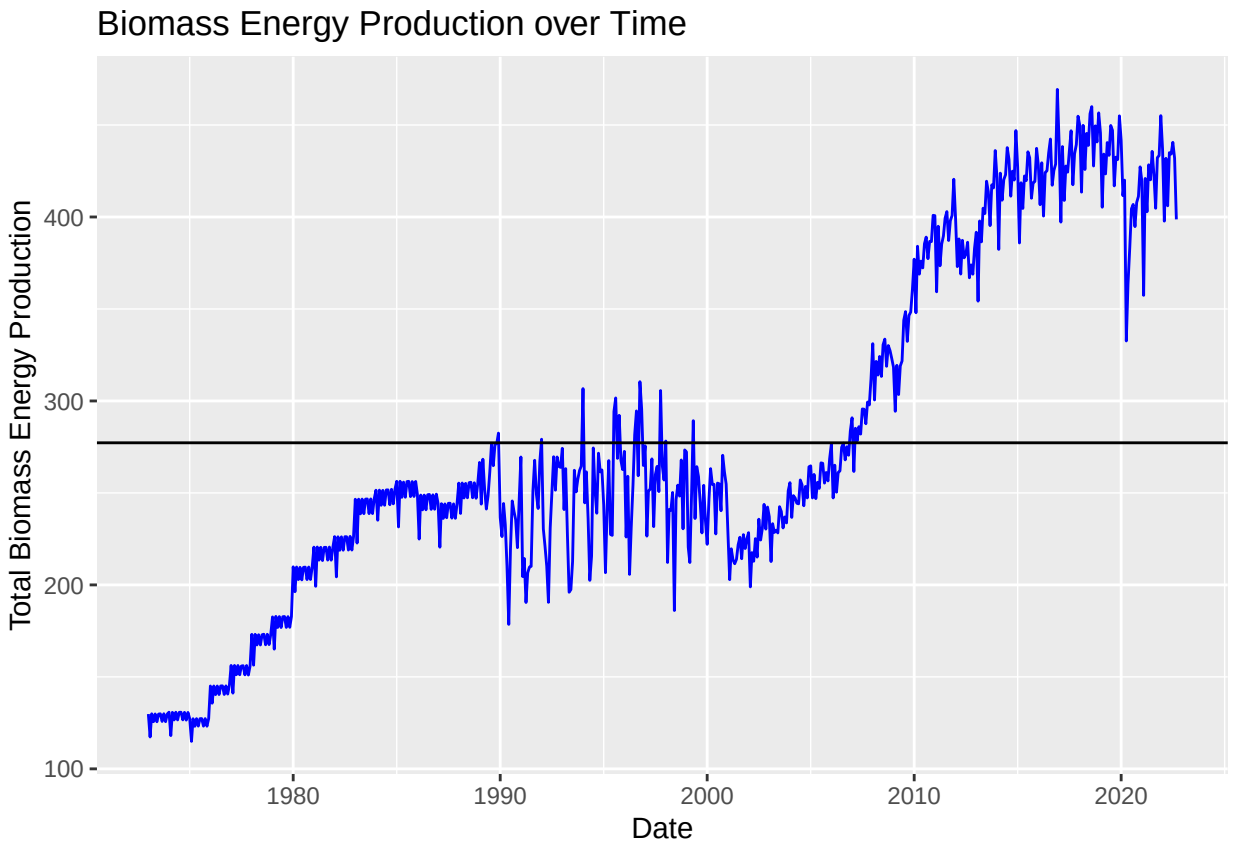
```
## [1] 44.16116
```

Question 4

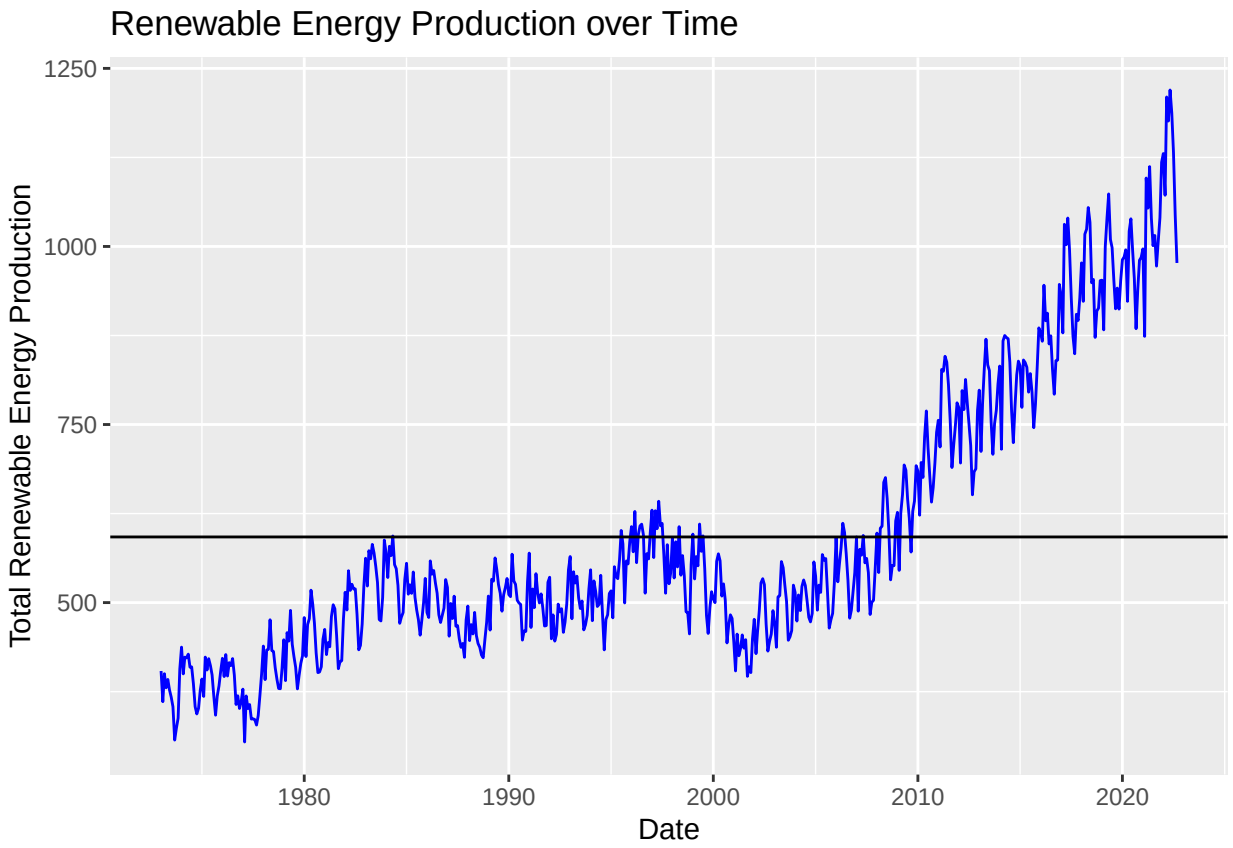
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
ggplot(data.edit, aes(x=date, y=data.edit[,2])) +
            geom_line(color="blue") +
            ylab("Total Biomass Energy Production")+
            xlab("Date")+
```

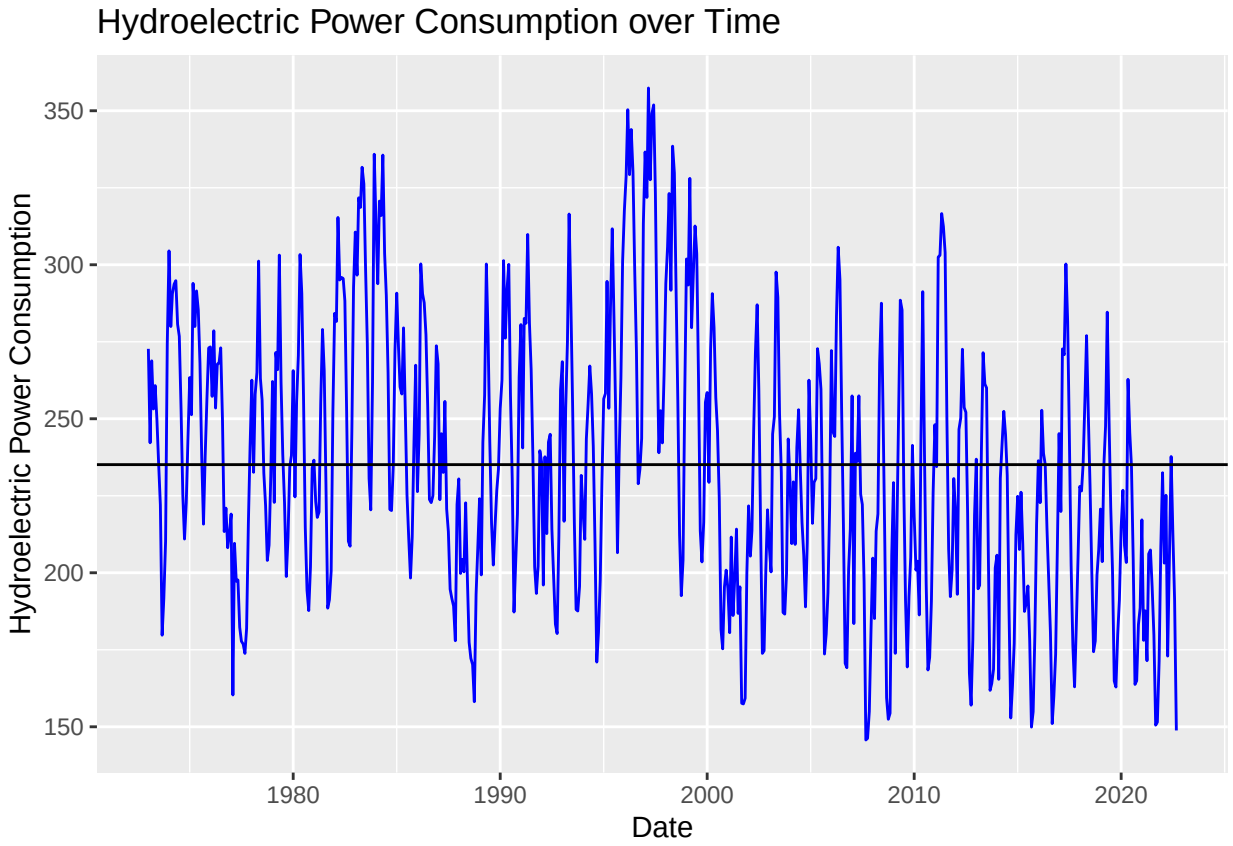
```
            ggtitle("Biomass Energy Production over Time")+
            geom_hline(yintercept=mean.biomass, color="black")
```




```
ggplot(data.edit, aes(x=date, y=data.edit[,3])) +
            geom_line(color="blue") +
            ylab("Total Renewable Energy Production")+
            xlab("Date")+
            ggtitle("Renewable Energy Production over Time")+
            geom_hline(yintercept=mean.renewable, color="black")
```




```
ggplot(data.edit, aes(x=date, y=data.edit[,4])) +
            geom_line(color="blue") +
            ylab("Hydroelectric Power Consumption")+
            xlab("Date")+
            ggtitle("Hydroelectric Power Consumption over Time")+
            geom_hline(yintercept=mean.hydro, color="black")
```



Question 5

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

```
cor.test(data.edit$Total.Biomass.Energy.Production, data.edit$Total.Renewable.Energy.Production)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data.edit$Total.Biomass.Energy.Production and data.edit$Total.Renewable.Energy.Production
## t = 56.697, df = 595, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9050636 0.9302668
## sample estimates:
##       cor
## 0.9185941
```

```
cor.test(data.edit$Total.Biomass.Energy.Production, data.edit$Hydroelectric.Power.Consumption)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data.edit$Total.Biomass.Energy.Production and data.edit$Hydroelectric.Power.Consumption
## t = -7.6661, df = 595, p-value = 7.256e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3711363 -0.2249878
```

```
## sample estimates:
##        cor
## -0.2998201
```

```
cor.test(data.edit$Hydroelectric.Power.Consumption, data.edit$Total.Renewable.Energy.Production)
```

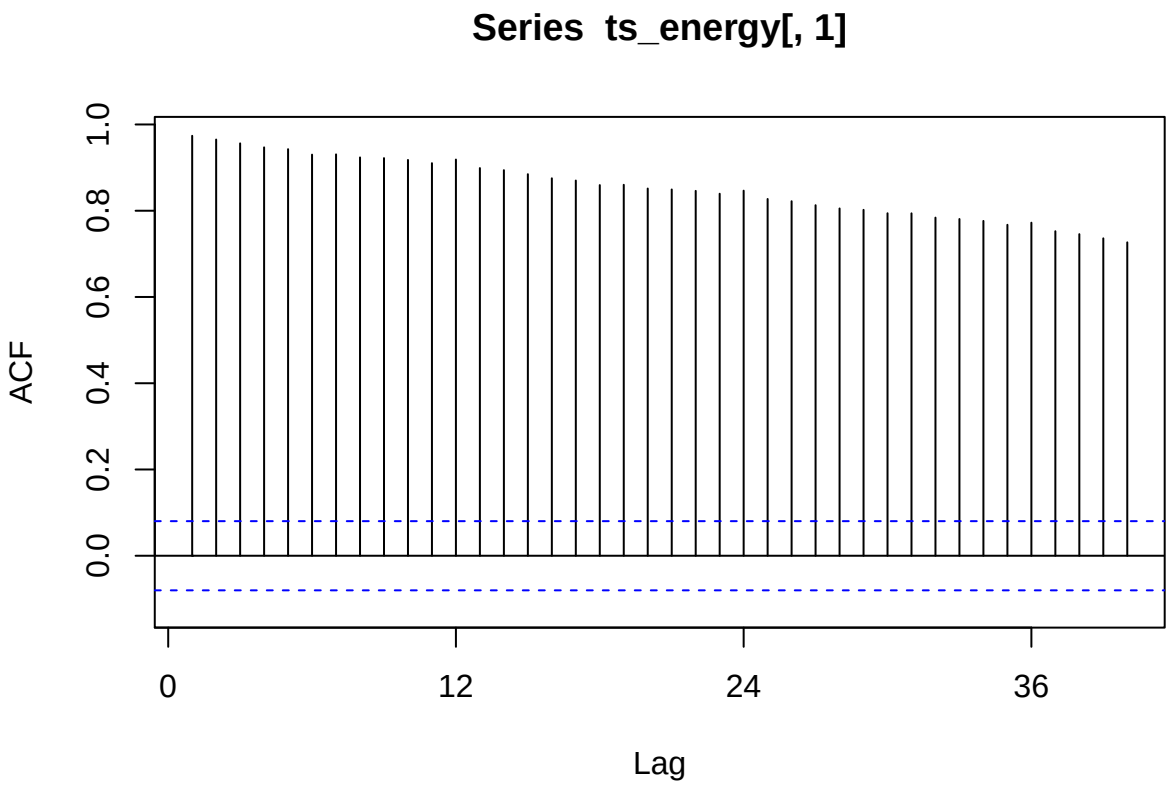
```
##
##  Pearson's product-moment correlation
##
## data:  data.edit$Hydroelectric.Power.Consumption and data.edit$Total.Renewable.Energy.Production
## t = -2.4413, df = 595, p-value = 0.01492
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17840723 -0.01949801
## sample estimates:
##         cor
## -0.09958758
```

Answer: They are all significantly correlated with each other (pearson, p-value < 0.05)

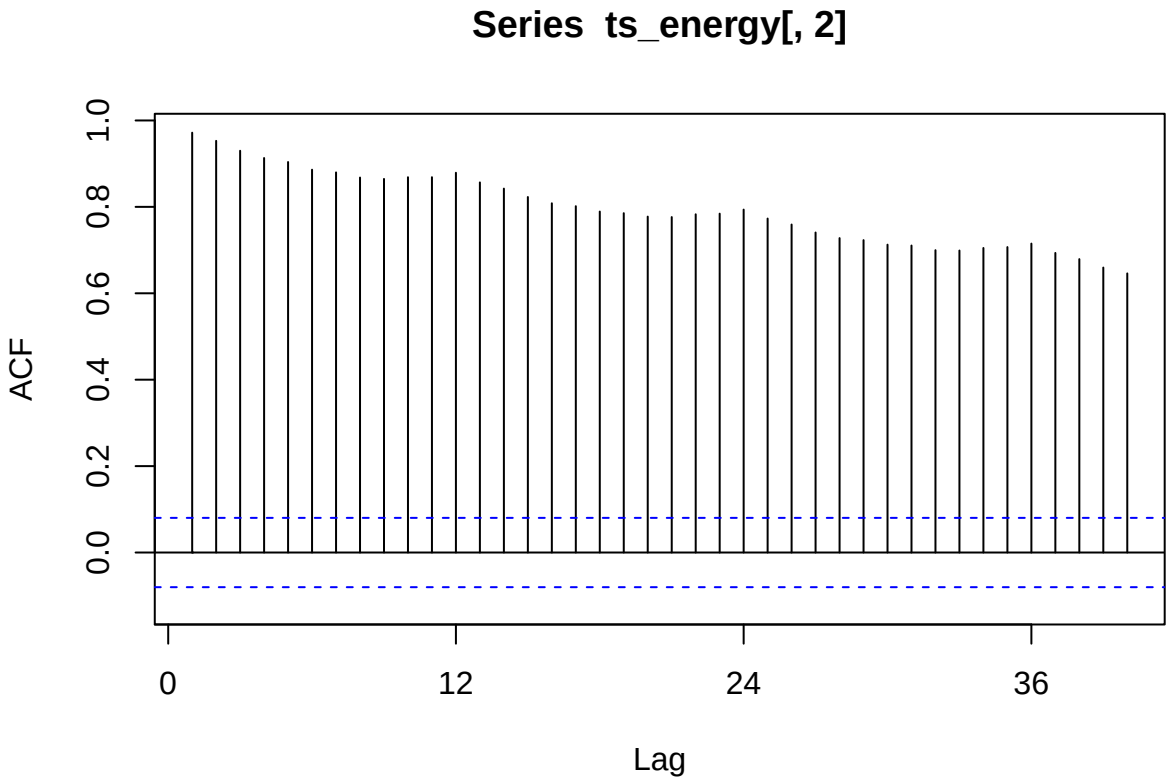
Question 6

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?

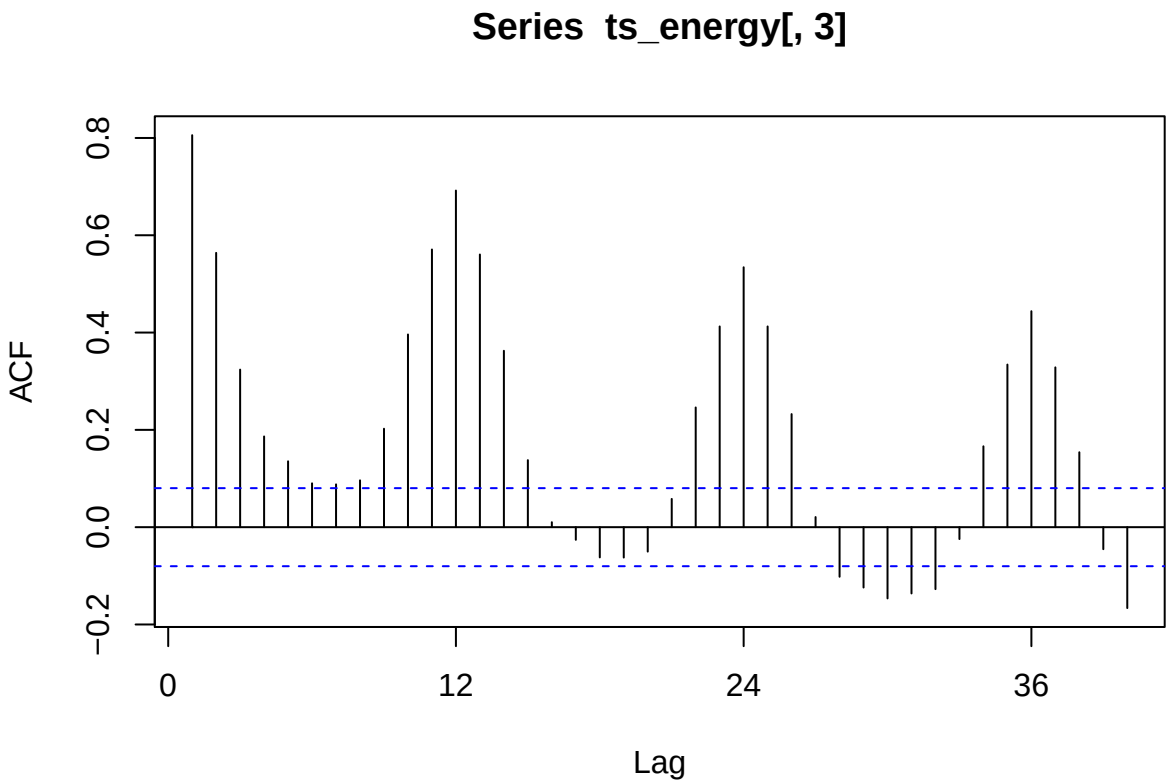
```
Acf(ts_energy[,1],lag.max=40)
```




```
Acf(ts_energy[,2],lag.max=40)
```




```
Acf(ts_energy[,3],lag.max=40)
```



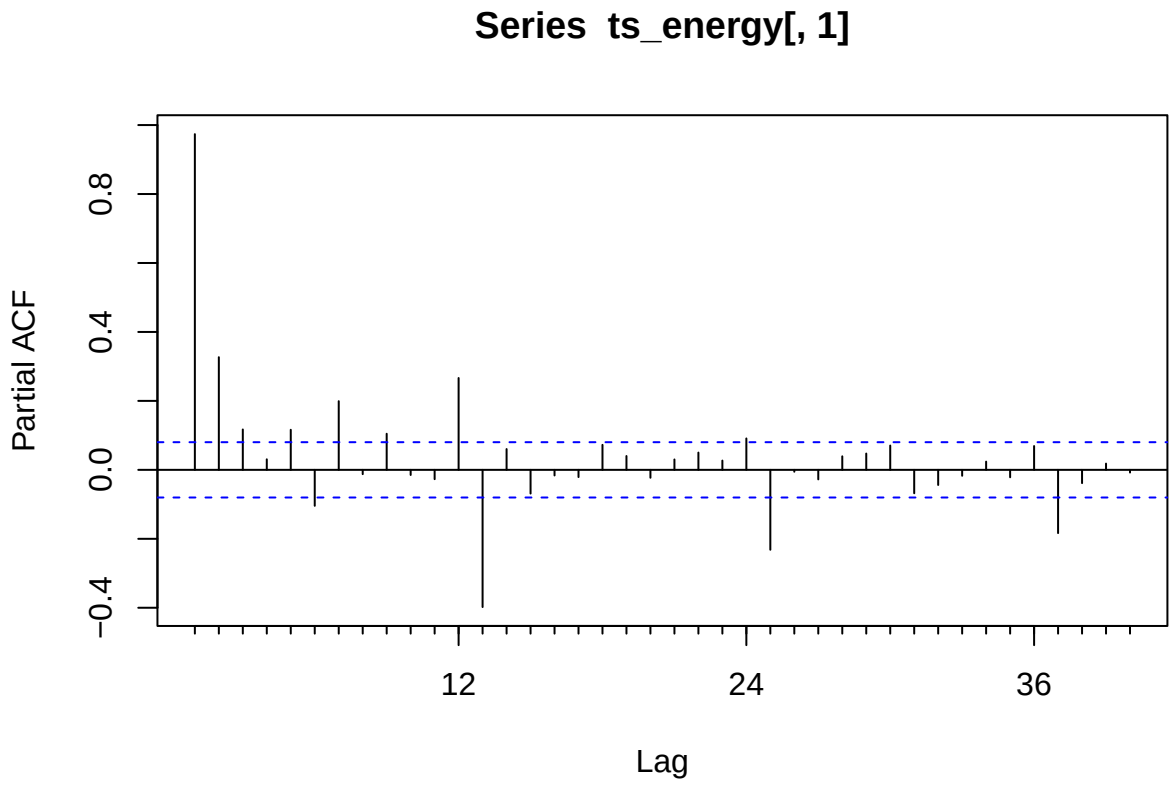
Answer: The first two ACF plots display very little seasonality - instead just declining slowly as the lag increases. However the third plot - the plot representing hydroelectric consumption - displays relatively strong seasonality, with a dip in correlation every six months. Interestingly these dips in correlation - aka the

strength of the relationship between those times and t0 - actually become non-significant. In both of the other plots - and the peak every 12 months in the hydroelectric plot - the correlations over time are always strongly positive.

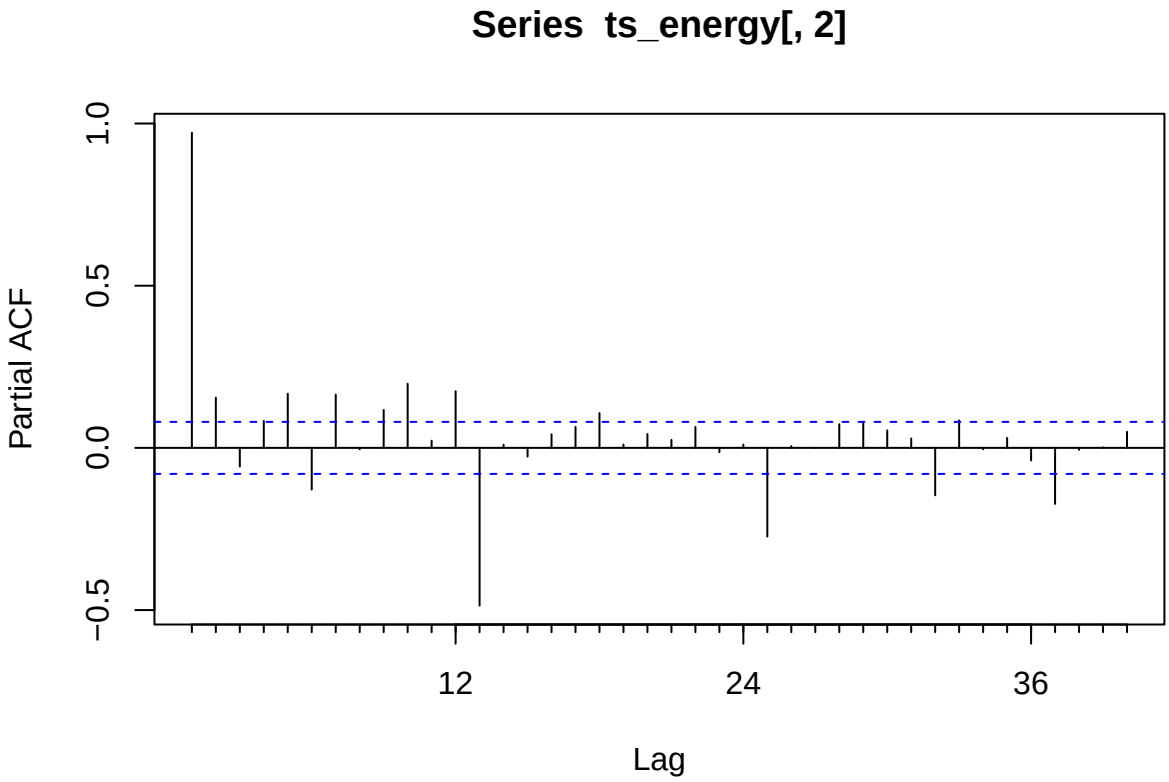
Question 7

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

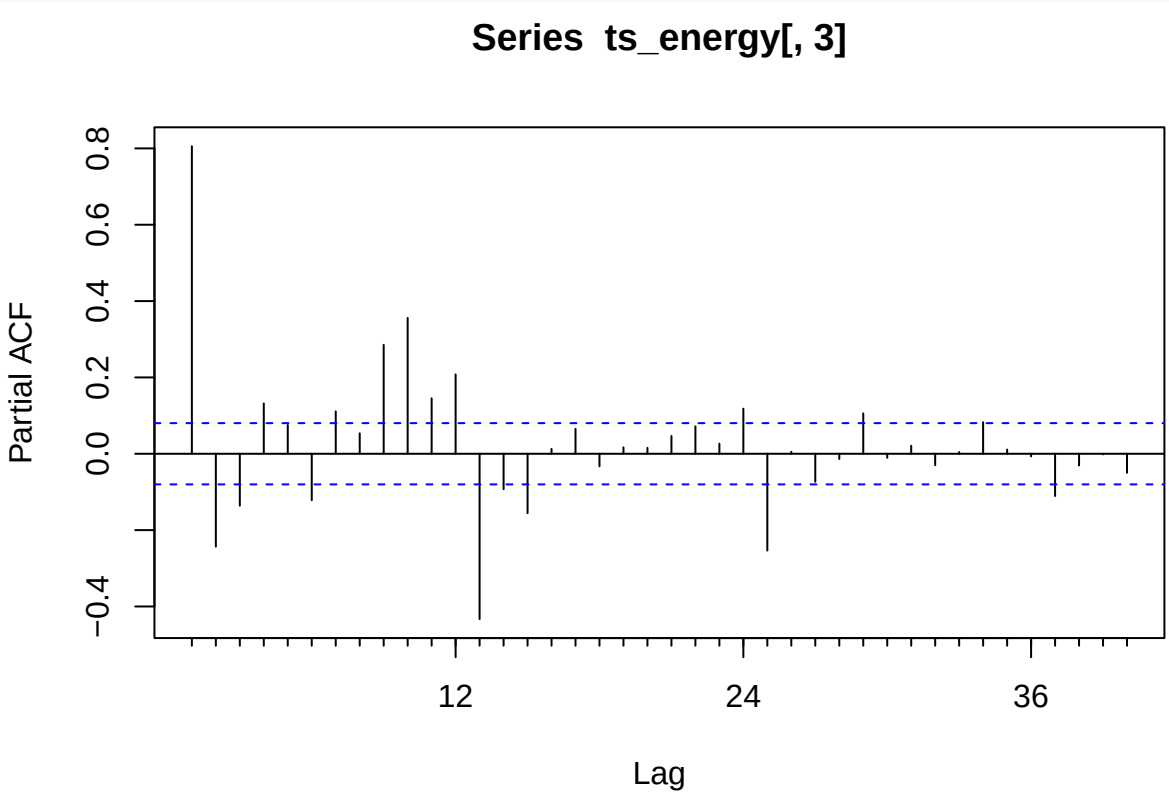
```
Pacf(ts_energy[,1],lag.max=40)
```




```
Pacf(ts_energy[,2],lag.max=40)
```




```
Pacf(ts_energy[,3],lag.max=40)
```



Answer: The interesting thing here is the difference between the ACF and the PACF plots. The PACF - partial autocorrelation plots - for both biomass and renewable energy production show none of the slowly decreasing trend we saw in their ACF plots. Instead, we're seeing that - when you think only of the correlation between

the current time series and the lagged time series - they only rarely have significant correlations, and when we do see significant correlations they match across all plots including the hydroelectric time series - a clear distinction from the ACF plots. This suggests a strong trend over time in both the biomass and renewable plots, because only a strong trend - or linkage in values that is carried over between lags - would show the ACF as being so heavily correlated while the PACF so uncorrelated. Those middle values in the correlation have to be increasing the correlation, because when they're removed in the PACF that strongly positive relationship across all lags disappears.