ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 2 - Due date 02/25/24

Jenn McNeill

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_Sp24.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(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(forecast)
```

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

```r
library(tseries)
library(here)
```

```
## here() starts at /Users/jennifermcneill/TSA_Spring2024/TSA_Spring2024
```

```r
library(lubridate)
```

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

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

```r
library(ggplot2)
getwd()
```

```
## [1] "/Users/jennifermcneill/TSA_Spring2024/TSA_Spring2024"
```

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 December 2023 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*.

```r
#Importing data set
Renewable_Energy_Raw <- read.table(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_
```

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.

```r
Renewable_Energy <- select(Renewable_Energy_Raw, Month, Total.Biomass.Energy.Production:Hydroelectric.Po

head(Renewable_Energy)
```

```
##           Month Total.Biomass.Energy.Production
## 1  1973 January                          129.787
## 2 1973 February                          117.338
## 3    1973 March                          129.938
## 4    1973 April                          125.636
## 5      1973 May                          129.834
## 6     1973 June                          125.611
```

```
##   Total.Renewable.Energy.Production Hydroelectric.Power.Consumption
## 1                           219.839                           89.562
## 2                           197.330                           79.544
## 3                           218.686                           88.284
## 4                           209.330                           83.152
## 5                           215.982                           85.643
## 6                           208.249                           82.060
```

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_biomass_production <- ts(Renewable_Energy[,2],start=c(1973,1), frequency=12)
ts_renewable_production <- ts(Renewable_Energy[,3],start=c(1973,1), frequency=12)
ts_hydroelectric_consumption <- ts(Renewable_Energy[,4],start=c(1973,1), frequency=12)

head(ts_biomass_production)
```

```
##          Jan     Feb     Mar     Apr     May     Jun
## 1973 129.787 117.338 129.938 125.636 129.834 125.611
```

```
head(ts_renewable_production)
```

```
##          Jan     Feb     Mar     Apr     May     Jun
## 1973 219.839 197.330 218.686 209.330 215.982 208.249
```

```
head(ts_hydroelectric_consumption)
```

```
##         Jan    Feb    Mar    Apr    May    Jun
## 1973 89.562 79.544 88.284 83.152 85.643 82.060
```

Question 3

Compute mean and standard deviation for these three series.

```
mean_biomass_production <- mean(ts_biomass_production)
sd_biomass_production <- sd(ts_biomass_production)

mean_renewable_production <- mean(ts_renewable_production)
sd_renewable_production <- sd(ts_renewable_production)

mean_hydroelectric_consumption <- mean(ts_hydroelectric_consumption)
sd_hydroelectric_consumption <- sd(ts_hydroelectric_consumption)

mean(ts_biomass_production)
```

```
## [1] 279.8046
```

```r
sd(ts_biomass_production)
```

```
## [1] 92.66504
```

```r
mean(ts_renewable_production)
```

```
## [1] 395.7213
```

```r
sd(ts_renewable_production)
```

```
## [1] 137.7952
```

```r
mean(ts_hydroelectric_consumption)
```

```
## [1] 79.73071
```

```r
sd(ts_hydroelectric_consumption)
```

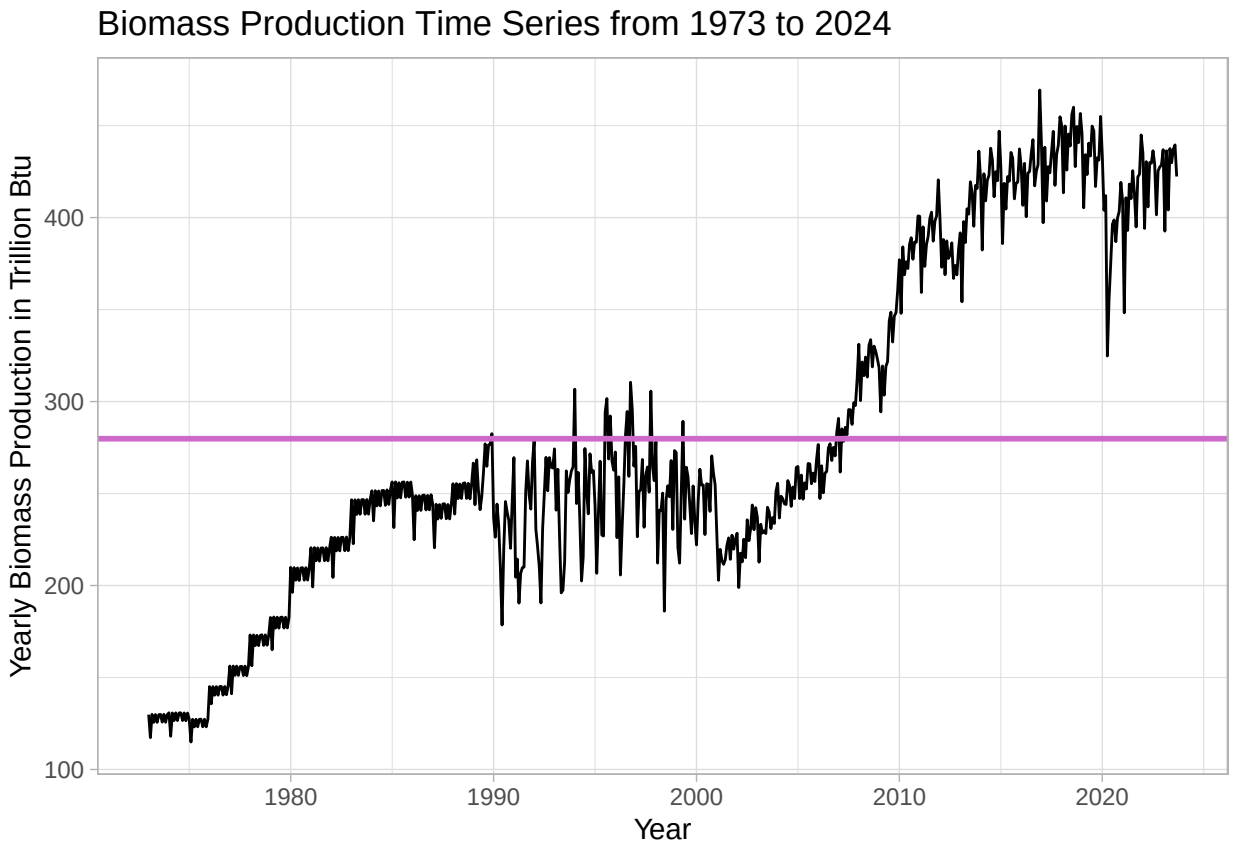
```
## [1] 14.14734
```

Question 4

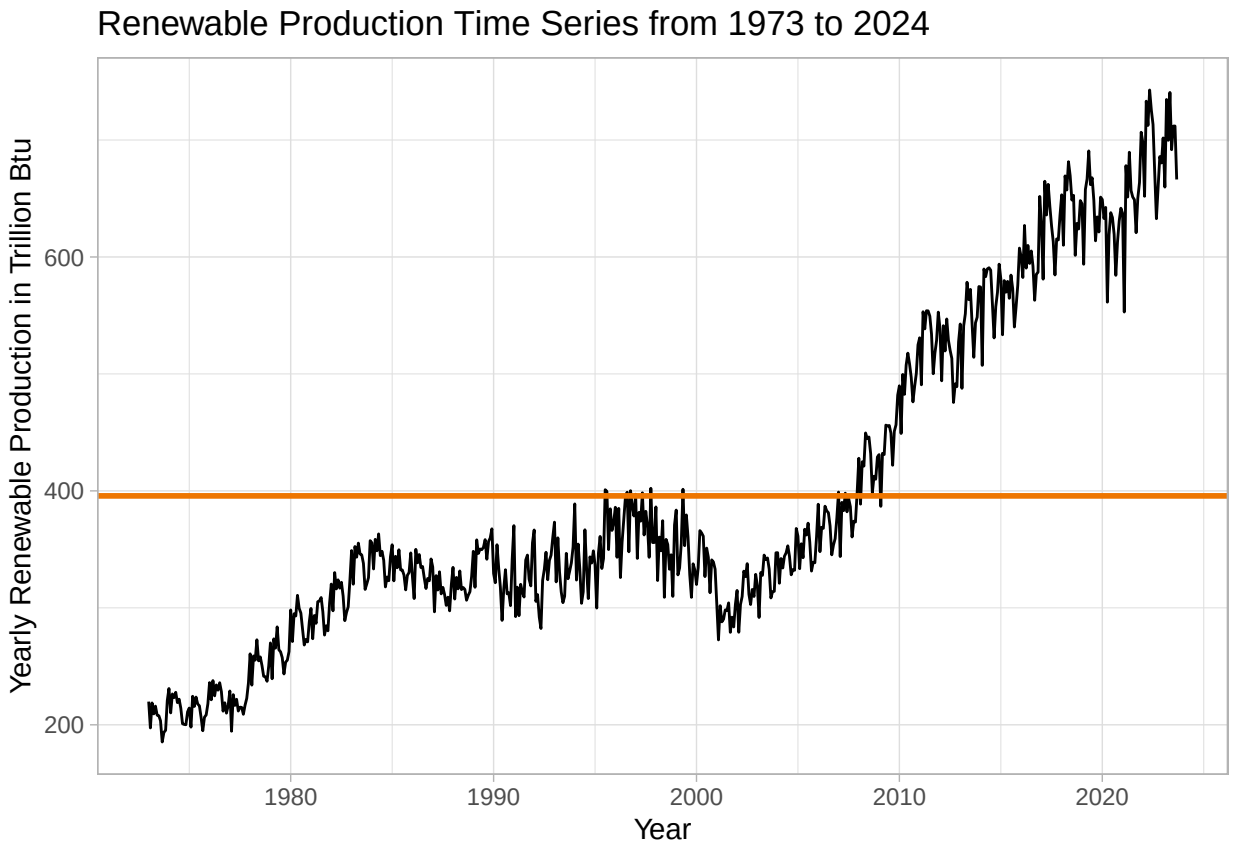
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```r
autoplot(ts_biomass_production)+
  xlab("Year")+
  ylab("Yearly Biomass Production in Trillion Btu")+
  ggtitle("Biomass Production Time Series from 1973 to 2024")+
  geom_hline(yintercept=mean_biomass_production, color = 'orchid3', size=1)+
  theme_light()
```

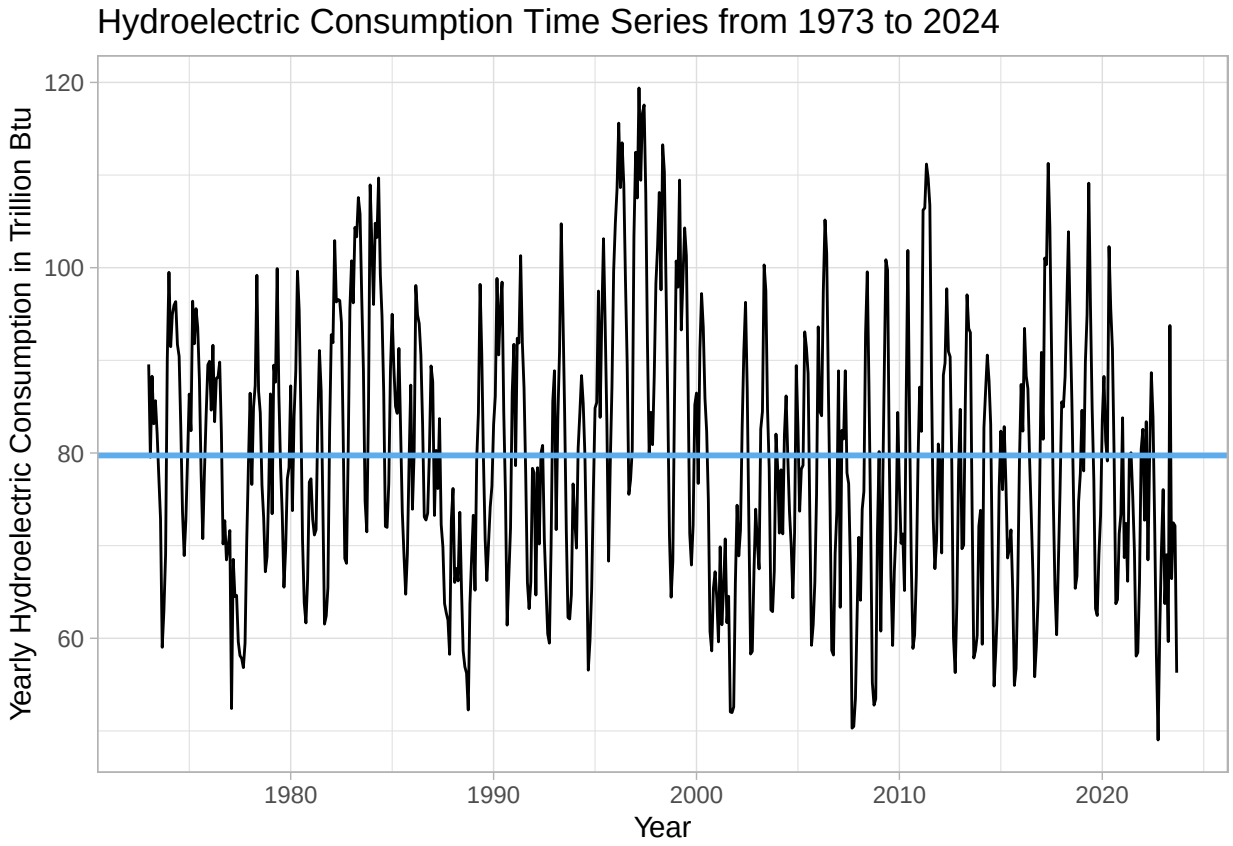
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
autoplot(ts_renewable_production)+
  xlab("Year")+
  ylab("Yearly Renewable Production in Trillion Btu")+
  ggtitle("Renewable Production Time Series from 1973 to 2024")+
  geom_hline(yintercept=mean_renewable_production, color = 'darkorange2', size=1)+
  theme_light()
```




```
autoplot(ts_hydroelectric_consumption)+
  xlab("Year")+
  ylab("Yearly Hydroelectric Consumption in Trillion Btu")+
  ggtitle("Hydroelectric Consumption Time Series from 1973 to 2024")+
  geom_hline(yintercept=mean_hydroelectric_consumption, color = 'steelblue2', size=1)+
  theme_light()
```



Question 5

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

```
Renewable_Energy_correlation <- cor(Renewable_Energy[,2:4])
Renewable_Energy_correlation
```

```
##                                  Total.Biomass.Energy.Production
## Total.Biomass.Energy.Production                       1.00000000
## Total.Renewable.Energy.Production                     0.97074621
## Hydroelectric.Power.Consumption                      -0.09656318
##                                  Total.Renewable.Energy.Production
## Total.Biomass.Energy.Production                        0.970746212
## Total.Renewable.Energy.Production                      1.000000000
## Hydroelectric.Power.Consumption                       -0.001768629
##                                  Hydroelectric.Power.Consumption
## Total.Biomass.Energy.Production                      -0.096563177
## Total.Renewable.Energy.Production                    -0.001768629
## Hydroelectric.Power.Consumption                       1.000000000
```

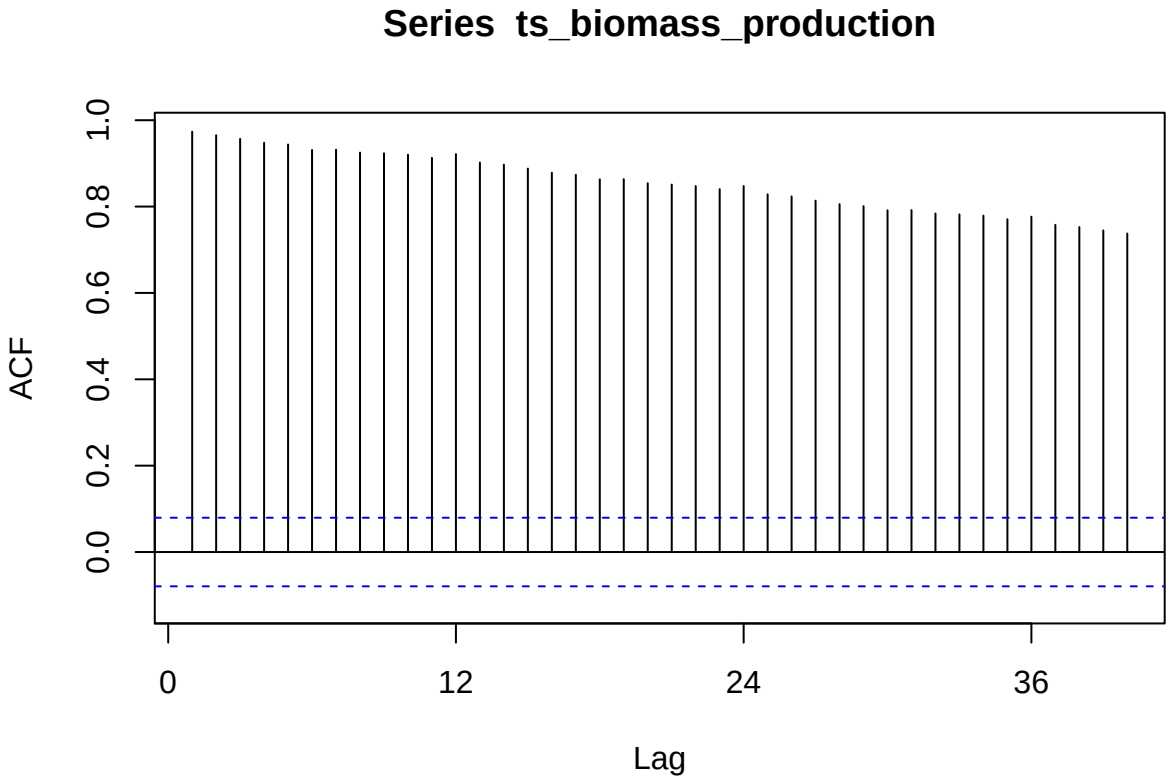
Total biomass energy production and total renewable energy production are significantly correleated with a correlation value of 0.97. It makes sense that these two variables correlate with each other because biomass is a renewable energy, so total biomass energy production is a subset of total renewable energy production. Neither of these series are significantly correlated to hydroelectric power consumption, with correlation values of -0.097 and -0.002, respectively. It makes sense that the power consumption series is

not significantly correlated with energy production series because they are measuring metrics that do not directly depend on one another.

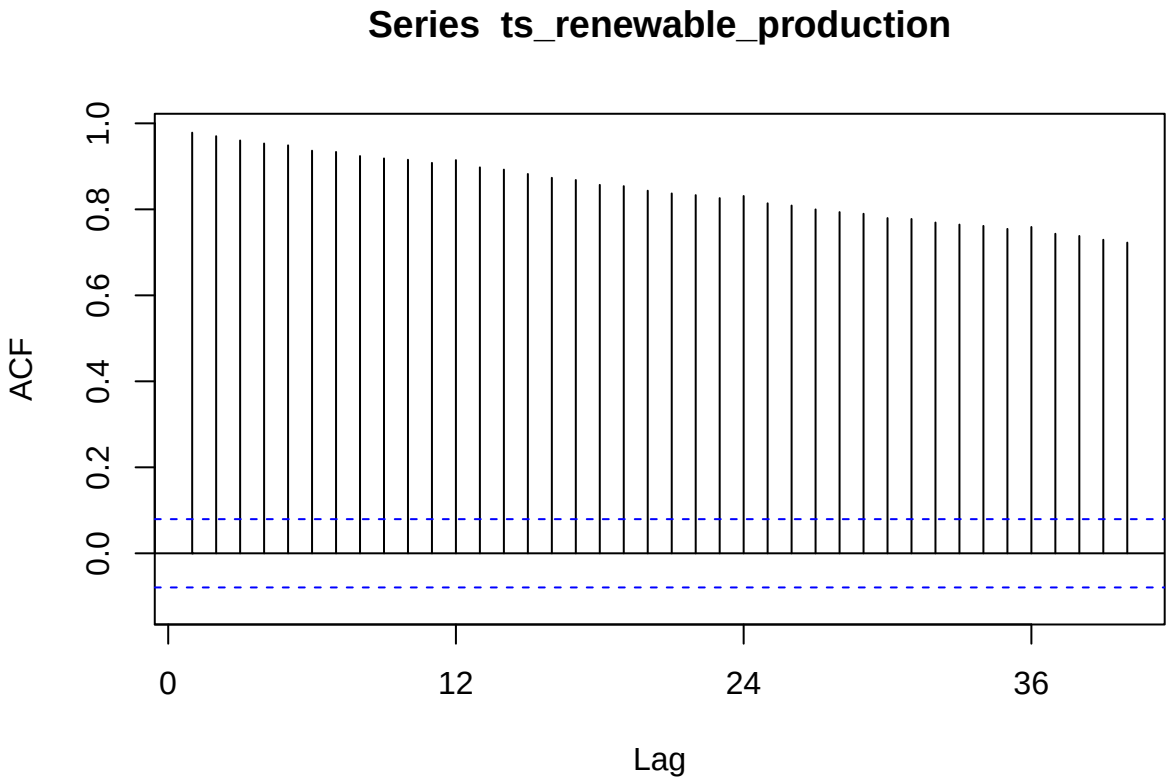
Question 6

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?

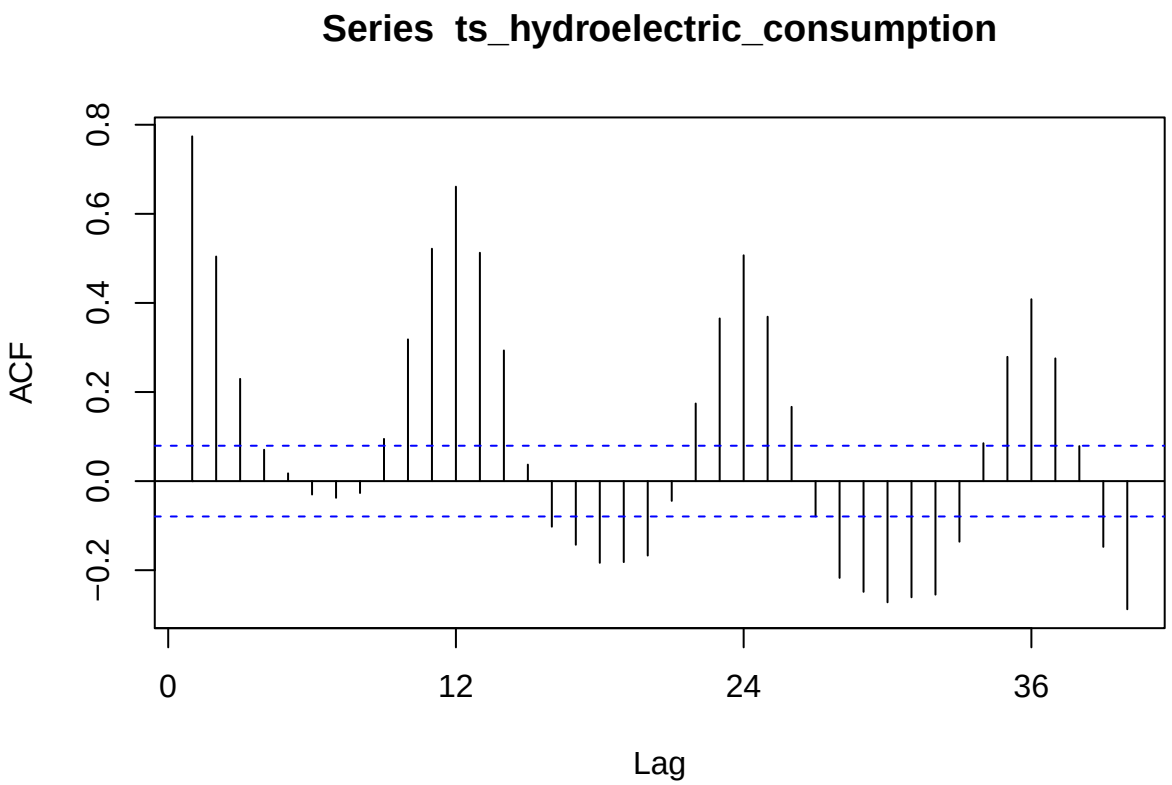
```
Biomass_Production_acf = Acf(ts_biomass_production,lag.max=40,plot=TRUE)
```




```
Renewable_Production_acf = Acf(ts_renewable_production,lag.max=40,plot=TRUE)
```




```
Hydroelectric_Consumption_acf = Acf(ts_hydroelectric_consumption,lag.max=40,plot=TRUE)
```



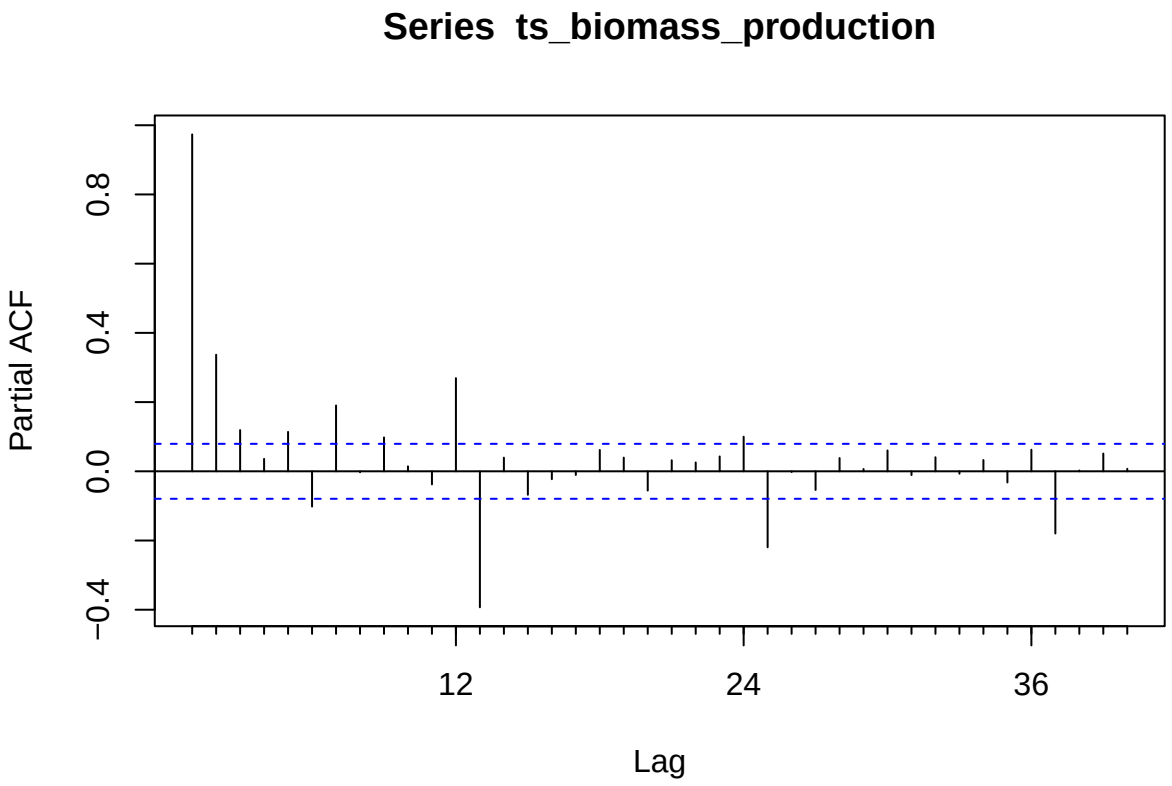
The two types of production, biomass and total renewable, have the same autocorrelation behavior. Both of these autocorrelation plots peak at the first lag, which means that the production data from the previous

month has the largest impact on the production data for the current month. As the lag gets larger, the autocorrelation lessens. The hydroelectric consumption autocorrelation has a seasonality component because we see the autocorrelation function following a pattern that lasts 12 lags. This means that the consumption data from twelve months prior has the largest impact on the consumption data for the current month.

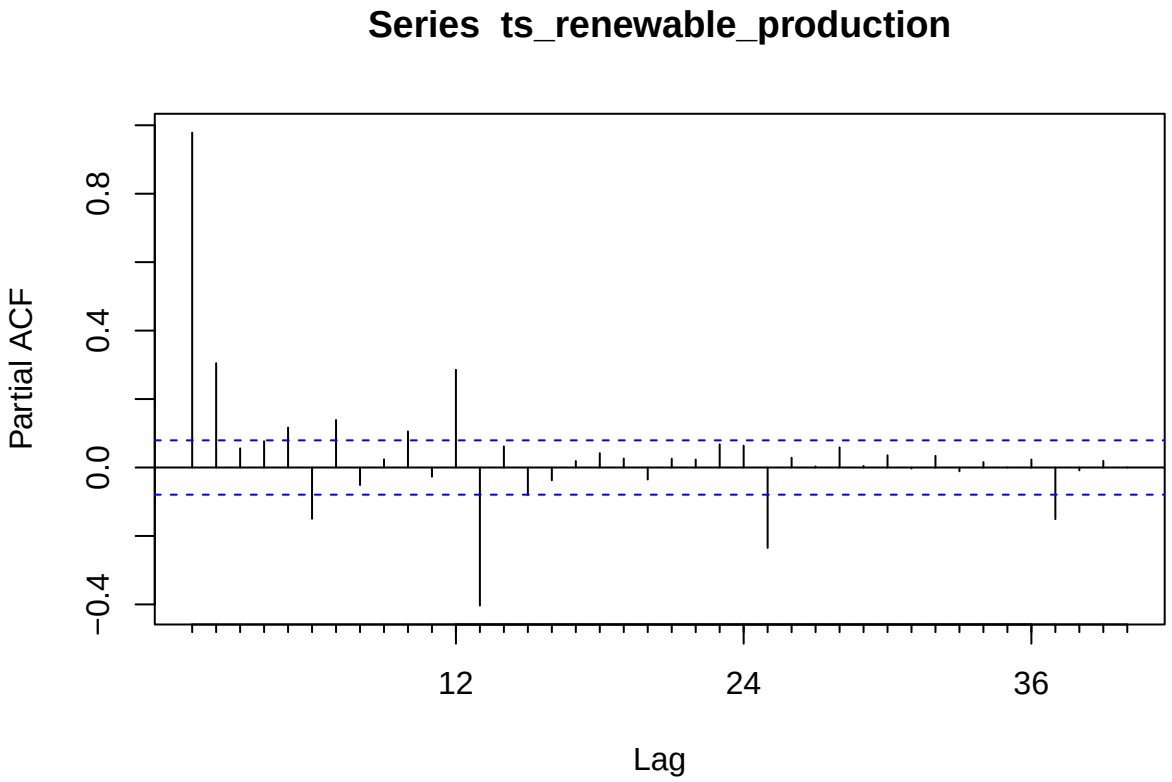
Question 7

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

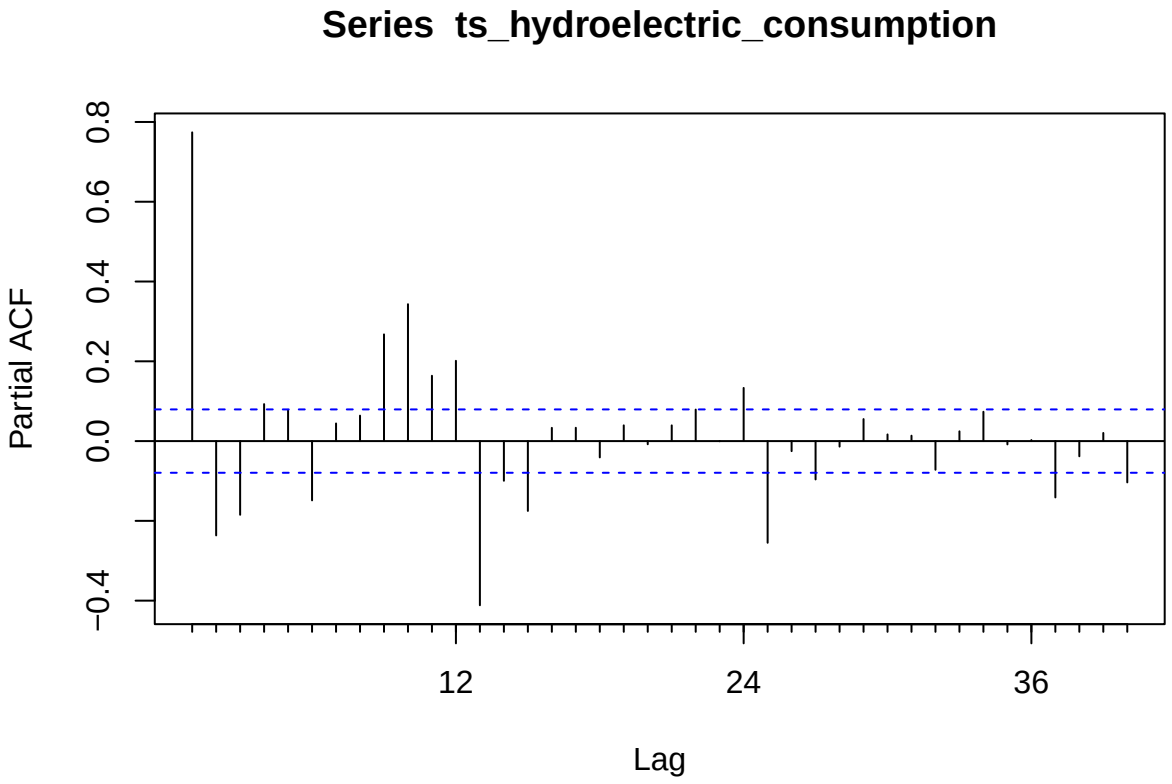
```
Biomass_Production_pacf = Pacf(ts_biomass_production,lag.max=40,plot=TRUE)
```




```
Renewable_Production_pacf = Pacf(ts_renewable_production,lag.max=40,plot=TRUE)
```




```
Hydroelectric_Consumption_pacf = Pacf(ts_hydroelectric_consumption,lag.max=40,plot=TRUE)
```



These plots all have different shapes than the autocorrelation graphs. The biggest difference is that the partial autocorrelation graphs each have a peak at the first lag and then their values drop more significantly

than in the Acf graphs. This difference comes from the fact that the Pacf removes any autocorrelation that is from the lag prior. In interpreting the biomass and total renewable Pacfs, we see that the first lag has the most impact on the correlation and all subsequent lags have far less impact. The Pacf for hydroelectric consumption still has a seasonality component, but we see that lags 12 and 25 impact the correlation far less than the first lag does. These graphs are consistent with what I would assume based on what I know about these variables.