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

Assignment 2 - Due date 02/03/23

Katherine Burley

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(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) # Causing error: package or namespace failed to load... DLL 'lmtest' not found
library(tseries)
```

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

```
library(readxl)
library(ggplot2)
```

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
energy_data <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
                          skip = 12, sheet="Monthly Data",col_names=FALSE)
```

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

```
#Extract the column names from row 11
read_col_names <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
                          skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)
```

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

```
colnames(energy_data) <- read_col_names
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>   <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.

```
# Preserve Original Dataframe
energy_data_orig <- energy_data

# Subset to Columns We Need
energy_data <- energy_data %>%
  select(c("Total Biomass Energy Production", "Total Renewable Energy Production", "Hydroelectric Power
head(energy_data)
```

```
## # A tibble: 6 x 3
##   `Total Biomass Energy Production` `Total Renewable Energy Production` Hydroe~1
##                               <dbl>                                <dbl>    <dbl>
## 1                              130.                                 404.     273.
## 2                              117.                                 361.     242.
## 3                              130.                                 400.     269.
## 4                              126.                                 380.     253.
## 5                              130.                                 392.     261.
## 6                              126.                                 377.     250.
## # ... 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().

```
# Convert to time series object
ts_energy <- ts(energy_data, start=1973, frequency=12) # Monthly data, freq=12
```

Question 3

Compute mean and standard deviation for these three series.

```
# Biomass Energy Production
mean(ts_energy[,"Total Biomass Energy Production"])
```

```
## [1] 277.2525
```

```
sd(ts_energy[,"Total Biomass Energy Production"])
```

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

```
# Renewable Energy Production
mean(ts_energy[,"Total Renewable Energy Production"])
```

```
## [1] 592.1583
```

```
sd(ts_energy[,"Total Renewable Energy Production"])
```

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

```
# Hydroelectric Power Consumption
mean(ts_energy[,"Hydroelectric Power Consumption"])
```

```
## [1] 235.1146
```

```
sd(ts_energy[,"Hydroelectric Power Consumption"])
```

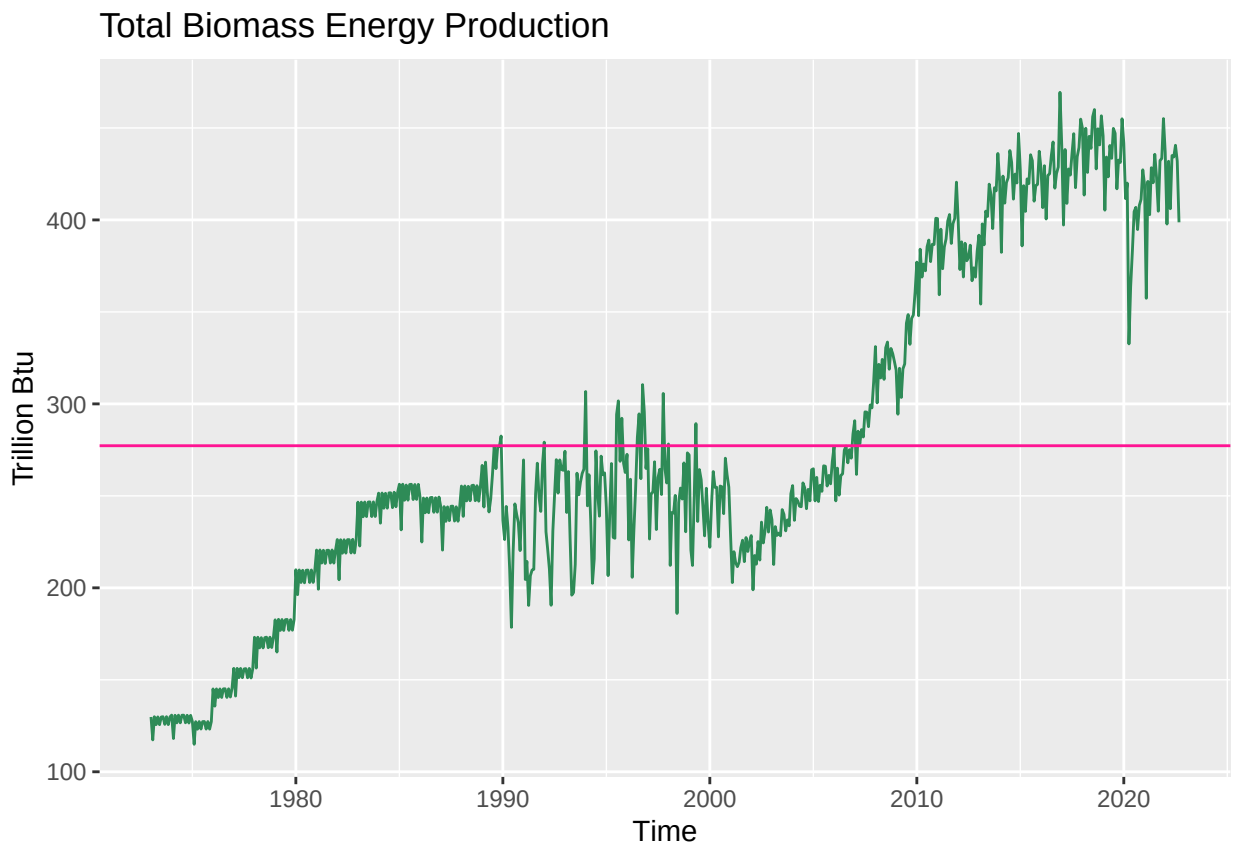
```
## [1] 44.16116
```

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.

Total Biomass Energy Production

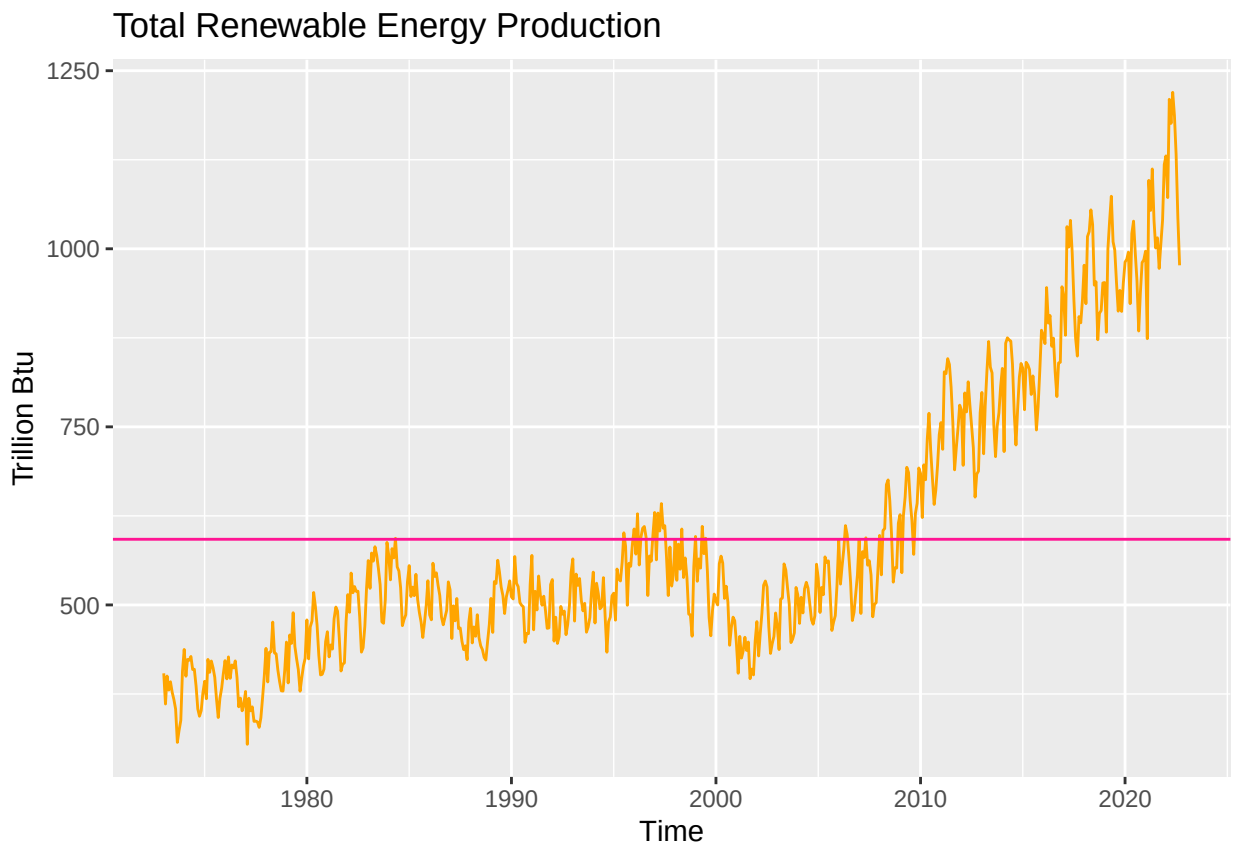
```
# Biomass Energy
# plot(ts_energy[,1],
#      main = "Total Biomass Energy Production")

ggplot(energy_data_orig, aes(x=Month, y=`Total Biomass Energy Production`)) +
  geom_line(color="seagreen4") +
  xlab("Time") +
  ylab("Trillion Btu") +
  ggtitle("Total Biomass Energy Production") +
  geom_hline(yintercept = mean(ts_energy[,"Total Biomass Energy Production"]), color="deeppink1")
```



Total biomass energy production increases steadily from the beginning of the time series in 1973 to around 1990, with what appears to be steady seasonal variation within each year. From 1990 to around 2000, total biomass energy production remains fairly steady, with increased and more erratic variation. From a little after 2000, total biomass energy production again starts increasing with smaller variation, with the increase slowing down and beginning to flatten out around 2015. There appears to be a significant dip in production in 2020, possibly due to COVID-19.

Total Renewable Energy Production

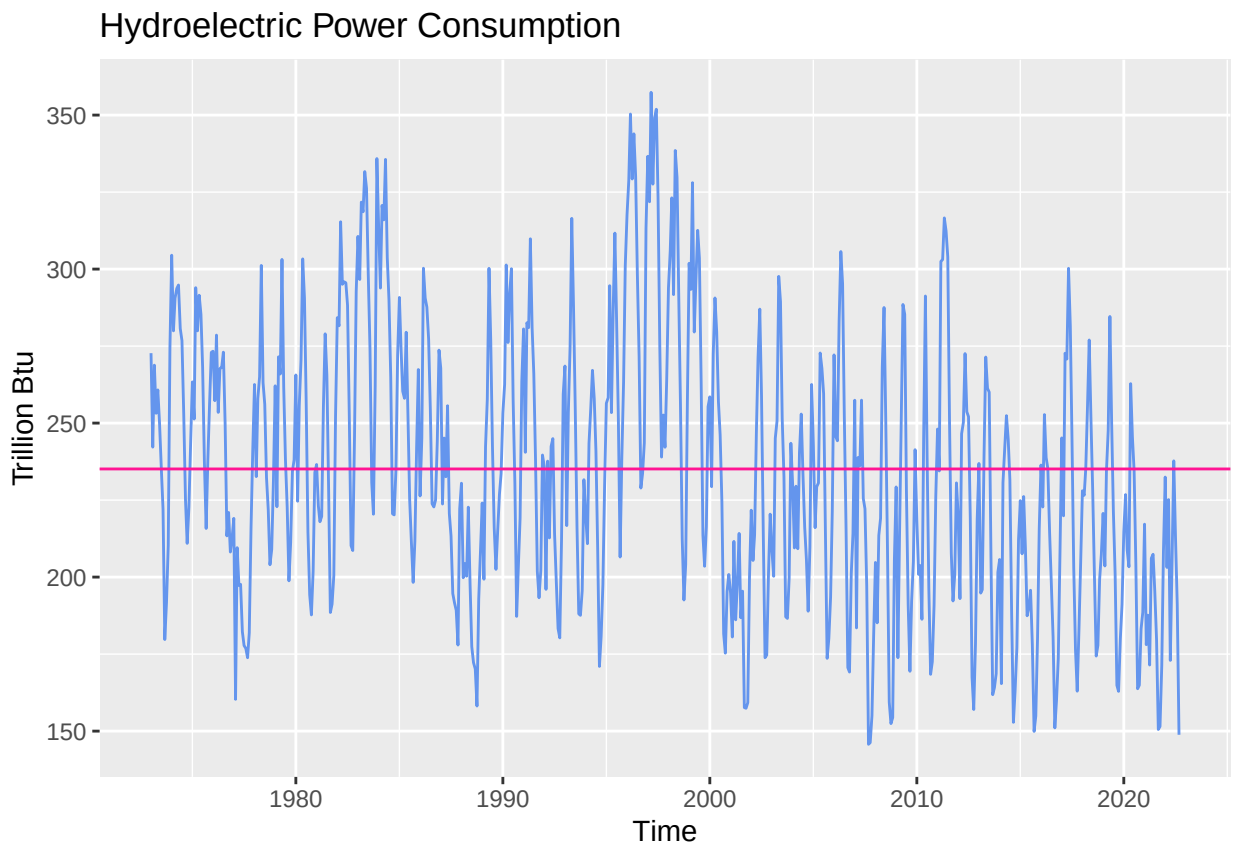
```
# Renewable Energy
# plot(ts_energy[,2],
#      main = "Total Renewable Energy Production")

ggplot(energy_data_orig, aes(x=Month, y=`Total Renewable Energy Production`)) +
  geom_line(color="orange") +
  xlab("Time") +
  ylab("Trillion Btu") +
  ggtitle("Total Renewable Energy Production") +
  geom_hline(yintercept = mean(ts_energy[,"Total Renewable Energy Production"]), color="deeppink1")
```



From the early 1970s to the early 2000s, total renewable energy production increases overall with some periods of decline. From the early 2000s, total RE production begins to increase steadily until 2023. Throughout the series, variation around the trend is fairly steady, with a slight increase in variation towards the end of the series.

Hydroelectric Power Consumption

```
# Hydroelectric Consumption
ggplot(energy_data_orig, aes(x=Month, y=`Hydroelectric Power Consumption`)) +
  geom_line(color="cornflowerblue") +
  xlab("Time") +
  ylab("Trillion Btu") +
  ggtitle("Hydroelectric Power Consumption") +
  geom_hline(yintercept = mean(ts_energy[,"Hydroelectric Power Consumption"]), color="deeppink1")
```



Throughout the time series, Hydroelectric Power Consumption declines slightly from 1973 to 2023, with significant variation around the mean from year to year. There appear to be some periods of slightly higher hydroelectric power consumption in the mid 1980s and late 1990s.

Question 5

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

```
# Correlation Between Series
cor(ts_energy[, c('Total Biomass Energy Production', 'Total Renewable Energy Production', 'Hydroelectri
```

```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.9185941
## Hydroelectric Power Consumption                       -0.2998201
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                         0.91859411
## Total Renewable Energy Production                       1.00000000
## Hydroelectric Power Consumption                        -0.09958758
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                      -0.29982013
## Total Renewable Energy Production                    -0.09958758
## Hydroelectric Power Consumption                       1.00000000
```

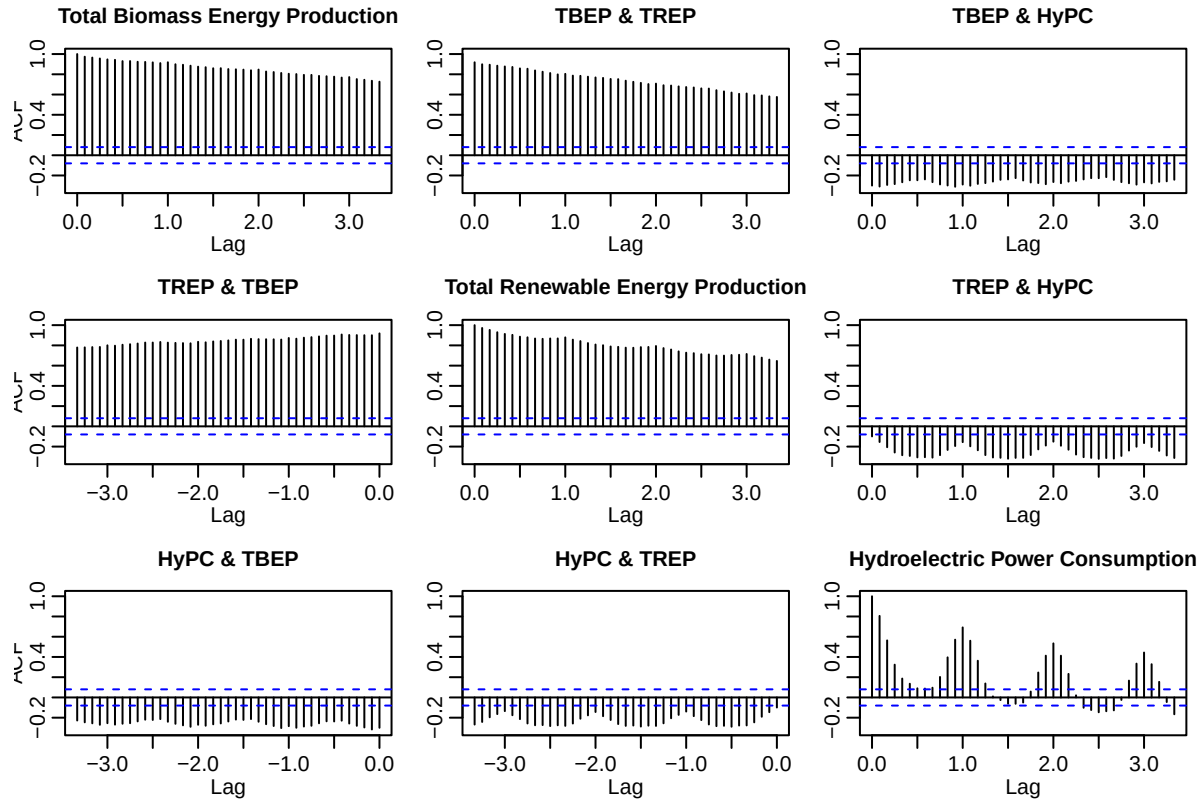
Correlation describes the strength of the linear dependence between two series, ranging from -1 to 1. Values closer to the extremes (-1 and 1) indicate a strong linear dependence, while values close to 0 indicate weaker or no linear dependence. Total Biomass Energy Production and Total Renewable Energy Production appear to be strongly, positively correlated, with a correlation of 0.9186. The plots affirm this, as these two series

follow fairly similar trajectories over the time frame. Hydroelectric Power Consumption does not appear to have a significant correlation with either of the other two series, with correlation values closer to zero (-0.3 and -0.1 for Total Biomass Energy Production and Total Renewable Energy Production, respectively). These numbers indicate a weak negative correlation.

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?

```
# Plot ACF with Lags 1-40
acf(ts_energy, lag.max=40)
```

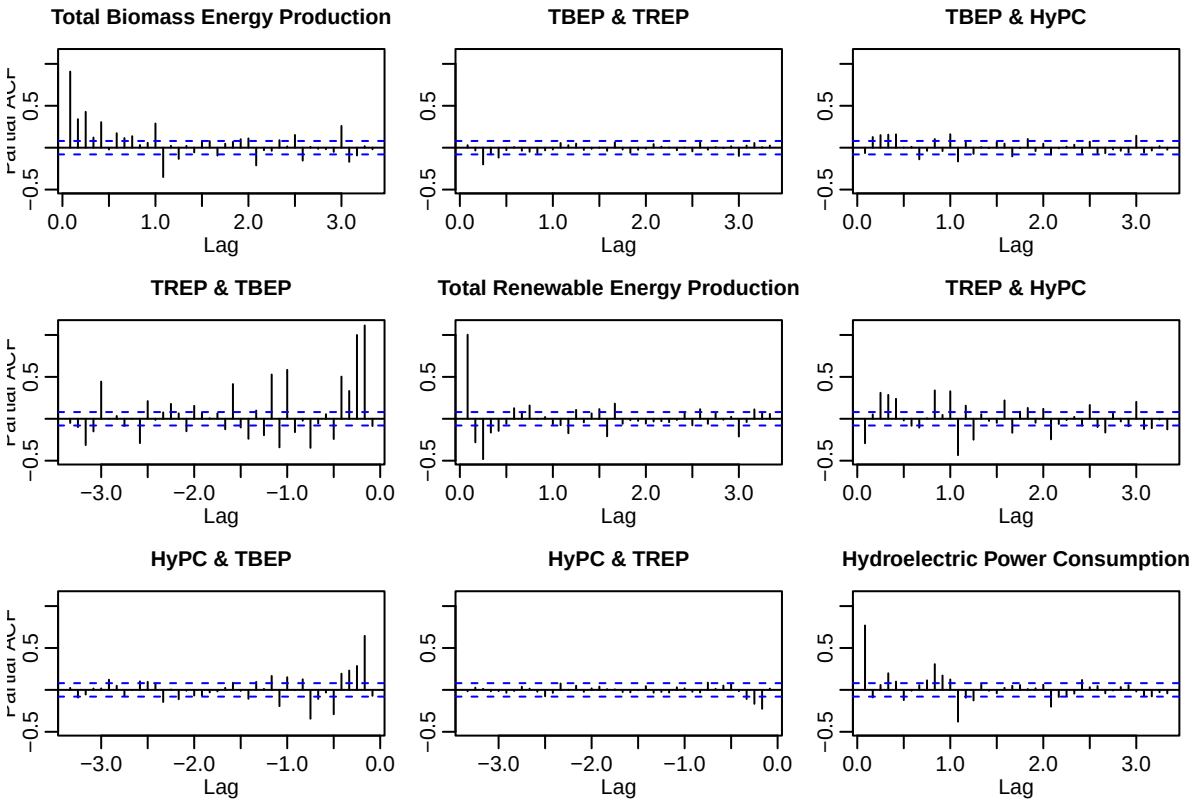


Total Biomass Energy Production and Total Renewable Energy Production seem similarly behaved with their ACF values very highest with a lag of 1 and then decreasing slightly as the number of lags increases. On the other hand, hydroelectric power consumption behaves quite differently from the other series, with the ACF demonstrating some seasonality with the highest ACF values with one lag and then higher ACF values around 10, 20, and 30 lags.

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?

```
# Plot PACF with Lags 1-40
pacf(ts_energy, lag.max=40)
```



The Partial Autocorrelation Function (PACF) tells us the linear dependency between a series and lagged values within the same series, while removing the influence of intermediate values between the value of interest and the lagged value. In the ACF plots, the ACF values for TBEP and TREP remained fairly high with slight decline as the number of lags increased, while for HyPC the ACF value spiked periodically with the number of lags. In each series, the PACF plot looks very different from the ACF plots. The PACF plots for each series show a high PACF value for one lag, that drops and fluctuates around zero for higher lags. I don't detect any distinguishable pattern in the PACF value for higher lags on any of the series.