ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 2 - Due date 01/26/22

Sarah Mansfield

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)
library(tseries)
library(dplyr)
library(readxl)
library(ggfortify)
```

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 January 2021 Monthly Energy Review. The spreadsheet is ready to be used. Use the command *read.table*() to import the data in R or *panda.read_excel*() in Python (note that you will need to import pandas package). }

```
#Importing data set
energy <- read_excel("~/ENVIRON 790/ENV790_TimeSeriesAnalysis_Sp2022/Data/Table_10.1_Renewable_Energy_P
                     skip = 10)
# delete extraneous first row
energy <- energy %>%
  slice(2:nrow(energy))
```

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 <- energy %>%
  select(`Total Biomass Energy Production`,
         `Total Renewable Energy Production`,
         `Hydroelectric Power Consumption`)
head(energy)
```

```
## # A tibble: 6 x 3
##   `Total Biomass Energy Production` `Total Renewable Ener~ `Hydroelectric Power~
##   <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
```

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().

```r
energy_ts <- ts(energy, start = c(1973, 1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

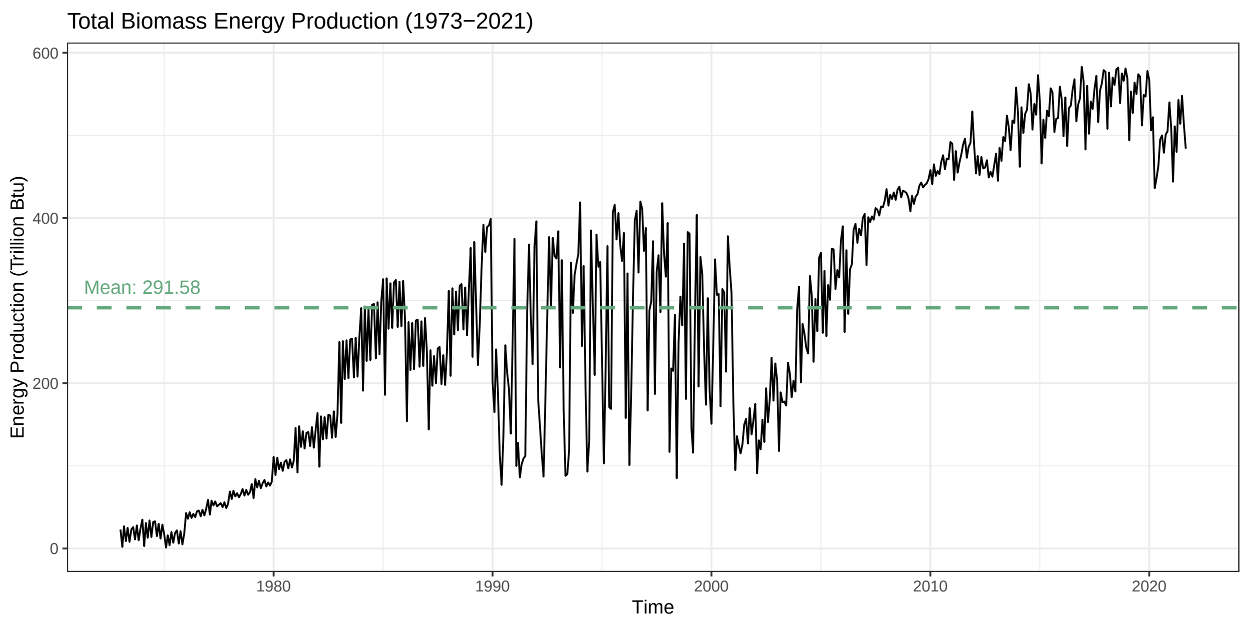
```r
(mean_sd <- sapply(energy_ts, function(x) c(mean = mean(x), sd = sd(x))))
```

```
##      Total Biomass Energy Production Total Renewable Energy Production
## mean                        291.5778                          293.0000
## sd                          168.5232                          169.0192
##      Hydroelectric Power Consumption
## mean                        293.0000
## sd                          169.0192
```

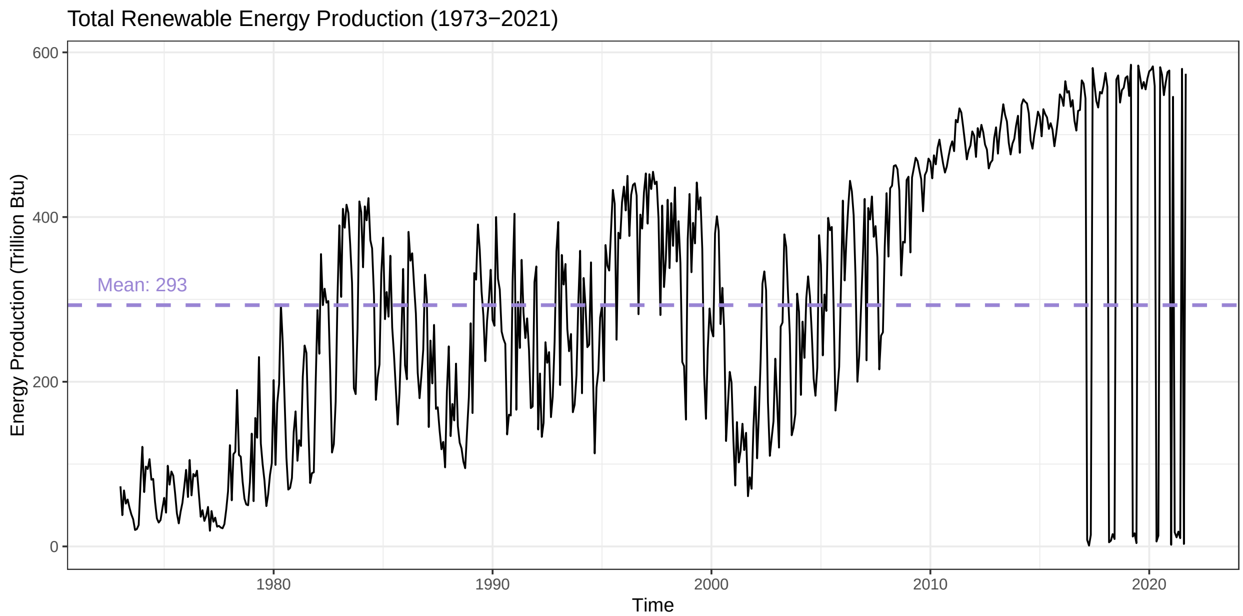
Question 4

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.

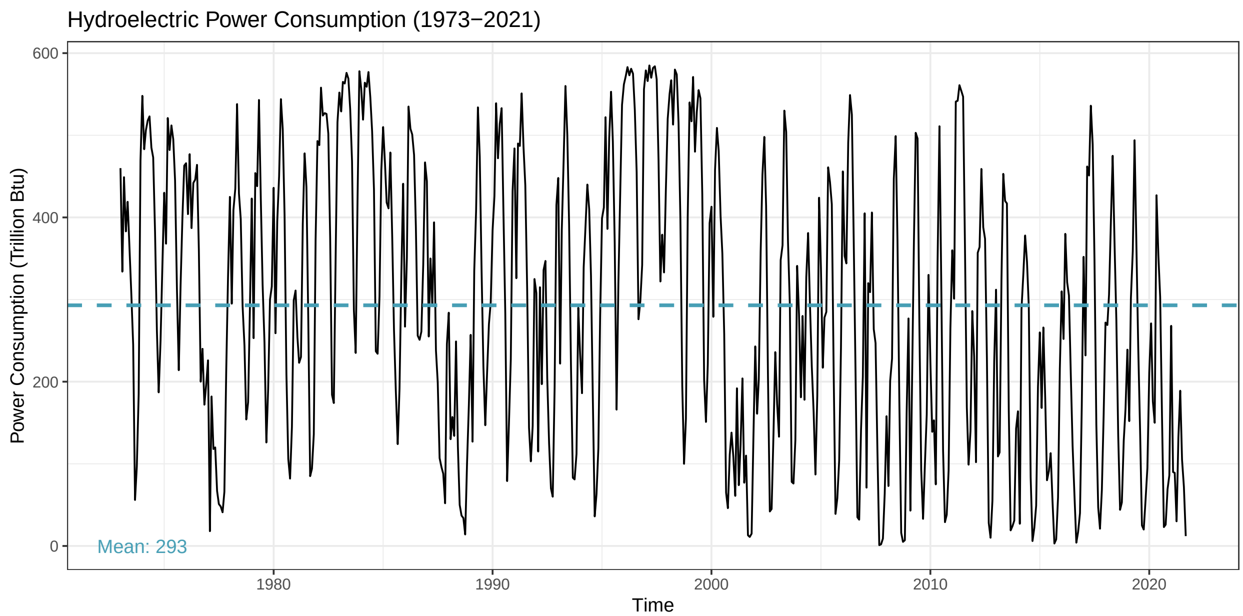
```
autoplot(energy_ts[,1]) +
  geom_hline(yintercept = mean_sd[1,][1],
             size = 1, linetype = "dashed", color = "#62a87c") +
  annotate("text", x = as.Date(0, origin = "1974-01-01"),
           y = mean_sd[1,][1], vjust = -1, color = "#62a87c",
           label = paste0("Mean: ", round(mean_sd[1,][1], 2))) +
  labs(title = "Total Biomass Energy Production (1973-2021)",
       x = "Time",
       y = "Energy Production (Trillion Btu)") +
  theme_bw()
```




```
autoplot(energy_ts[,2]) +
  geom_hline(yintercept = mean_sd[1,][2],
             size = 1, linetype = "dashed", color = "#9984d4") +
  annotate("text", x = as.Date(0, origin = "1974-01-01"),
           y = mean_sd[1,][2], vjust = -1, color = "#9984d4",
           label = paste0("Mean: ", round(mean_sd[1,][2], 2))) +
  labs(title = "Total Renewable Energy Production (1973-2021)",
       x = "Time",
       y = "Energy Production (Trillion Btu)") +
  theme_bw()
```




```r
autoplot(energy_ts[,3]) +
  geom_hline(yintercept = mean_sd[1,][3],
             size = 1, linetype = "dashed", color = "#489fb5") +
  annotate("text", x = as.Date(0, origin = "1974-01-01"),
           y = 0, color = "#489fb5",
           label = paste0("Mean: ", round(mean_sd[1,][3], 2))) +
  labs(title = "Hydroelectric Power Consumption (1973-2021)",
       x = "Time",
       y = "Power Consumption (Trillion Btu)") +
  theme_bw()
```



Question 5

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

```
cor(energy_ts)
```

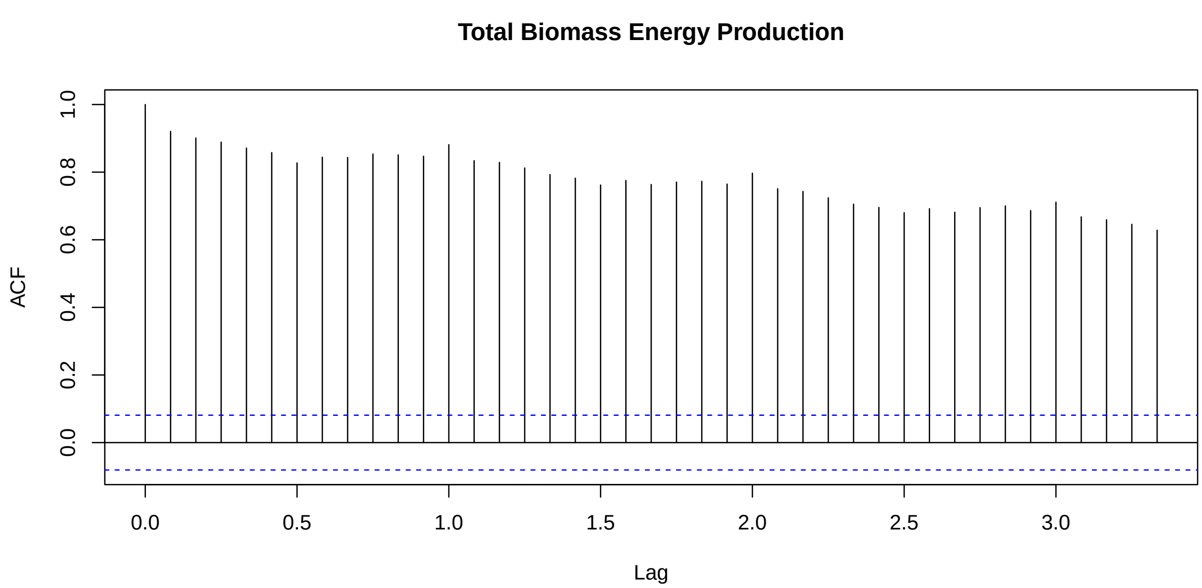
```
##                                    Total Biomass Energy Production
## Total Biomass Energy Production                          1.0000000
## Total Renewable Energy Production                        0.7325836
## Hydroelectric Power Consumption                         -0.2724661
##                                    Total Renewable Energy Production
## Total Biomass Energy Production                           0.73258363
## Total Renewable Energy Production                         1.00000000
## Hydroelectric Power Consumption                           0.06868406
##                                    Hydroelectric Power Consumption
## Total Biomass Energy Production                        -0.27246610
## Total Renewable Energy Production                       0.06868406
## Hydroelectric Power Consumption                         1.00000000
```

Total Biomass Energy Production and Total Renewable Energy Production do seem to be significantly (positively) correlated, as the correlation coefficient between the two is reasonably high at about 0.73. On the other hand, the correlation coefficients calculated between Total Biomass Energy Production and Hydroelectric Power Consumption as well as Total Renewable Energy Production and Hydroelectric Power Consumption are much smaller (about -0.27 and 0.07, respectively), and therefore are not significantly correlated.

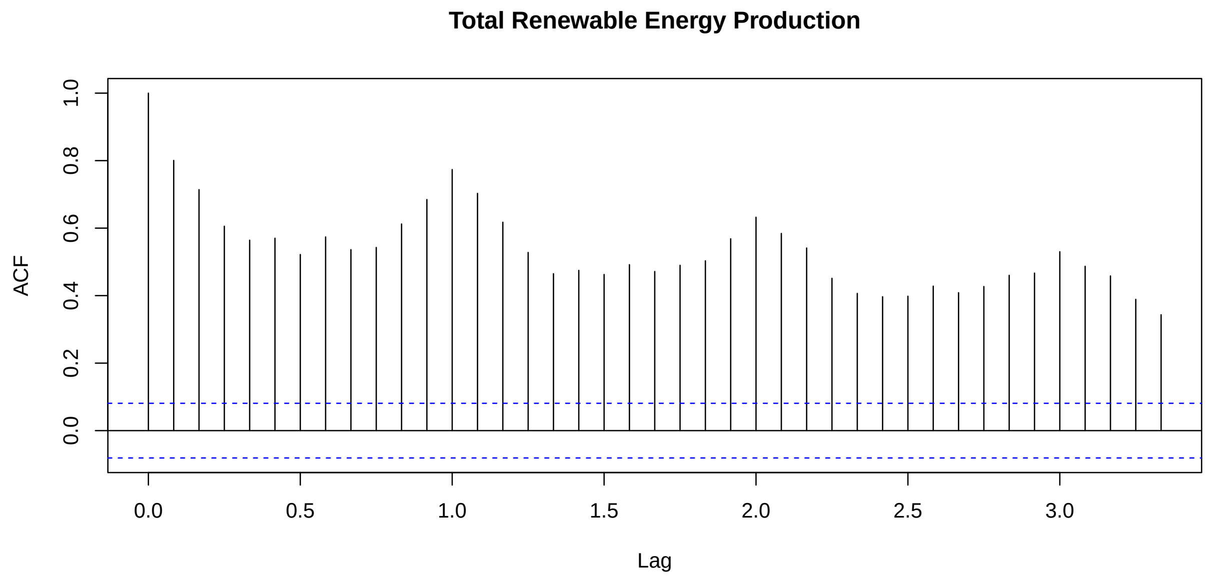
Question 6

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?

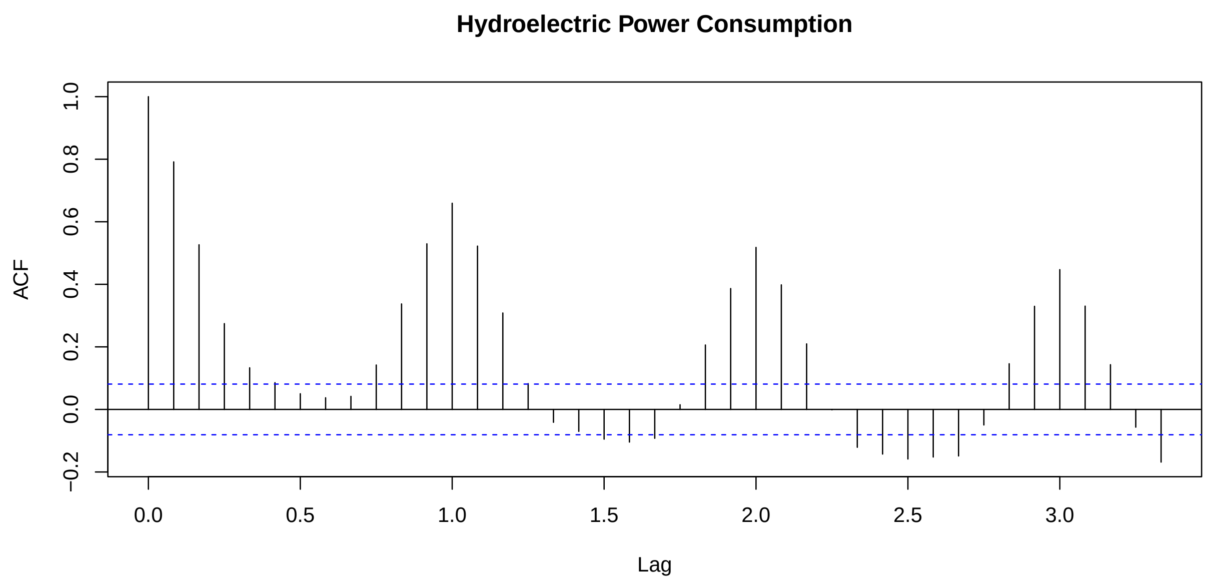
```
acf(energy_ts[,1], lag.max = 40, main = "Total Biomass Energy Production")
```



```r
acf(energy_ts[,2], lag.max = 40, main = "Total Renewable Energy Production")
```




```r
acf(energy_ts[,3], lag.max = 40, main = "Hydroelectric Power Consumption")
```

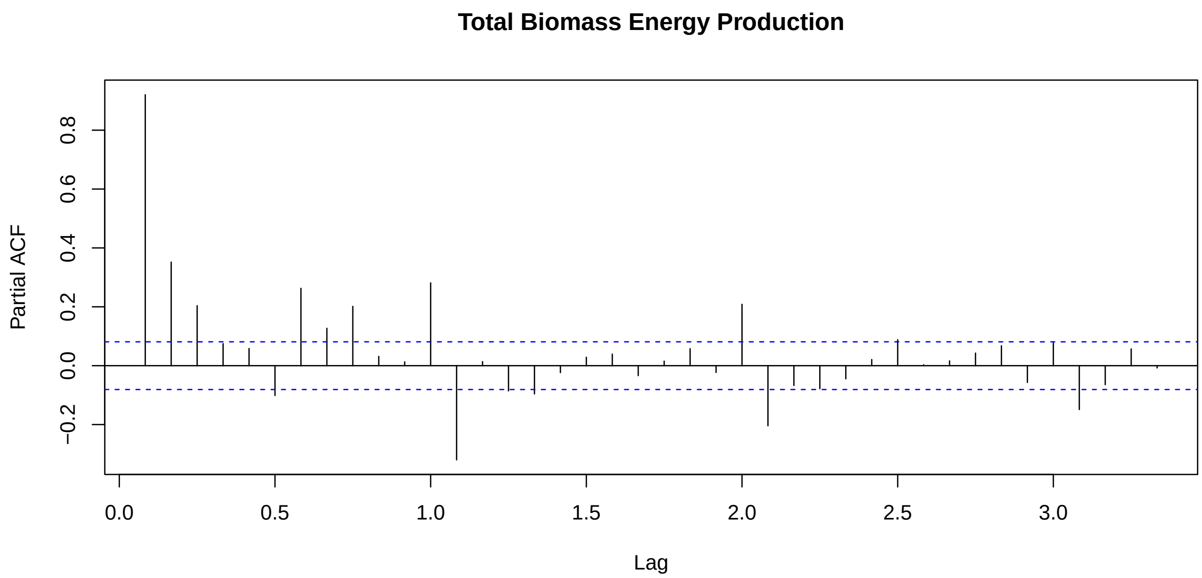


In general, we see an overall decreasing trend in the ACF for each plot as the lag increases, as well as a repeating trend in which the ACF increases, reaches a peak, and then decreases (similar to a bell curve). For Total Biomass Energy Production, this trend is very gradual - the difference in the ACF between each subsequent lag is minimal. For Total Renewable Energy Production, we can see that these trends are more apparent/less gradual. Similarly, for Hydroelectric Power Consumption, the ACF decreases sharply initially before entering that bell curve-like pattern, while also becoming interspersed with periods where the ACF is negative (as the lag increases, we can also see that these values seem to become increasingly negative).

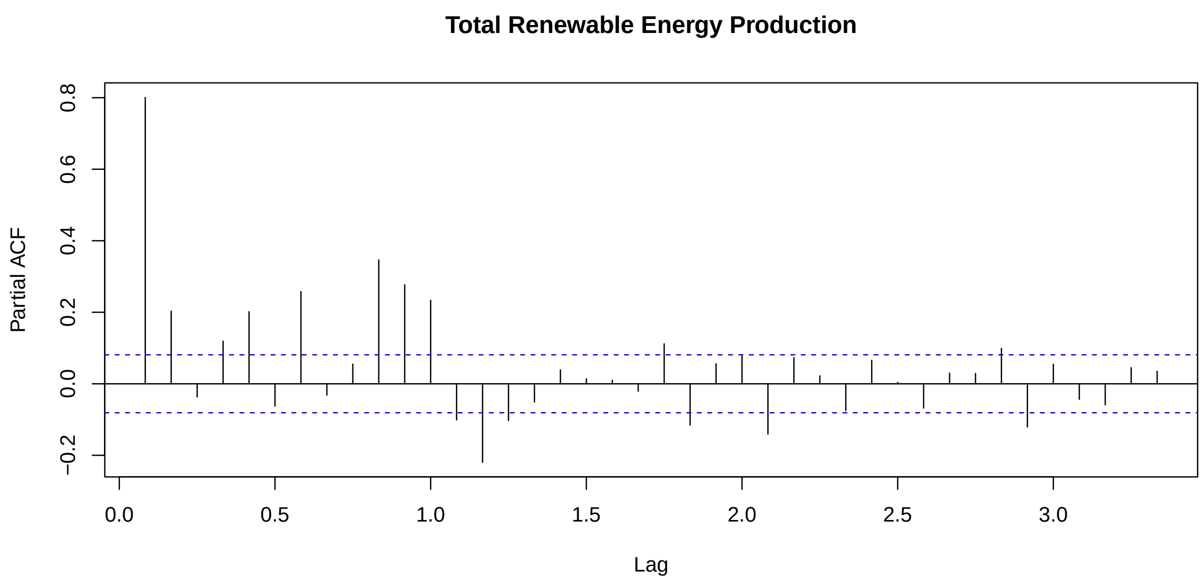
Question 7

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

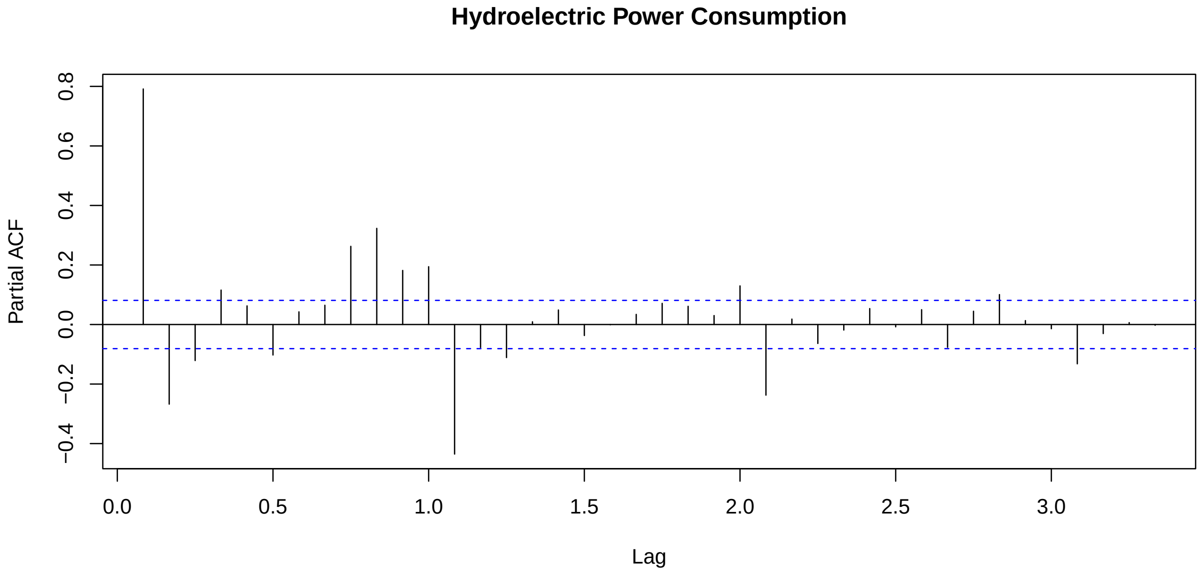
```
pacf(energy_ts[,1], lag.max = 40, main = "Total Biomass Energy Production")
```




```
pacf(energy_ts[,2], lag.max = 40, main = "Total Renewable Energy Production")
```




```
pacf(energy_ts[,3], lag.max = 40, main = "Hydroelectric Power Consumption")
```



Whereas the ACF plots tended to show a more gradual decrease, the PACF plots show an immediate drop in the partial ACF after just one lag, with the partial ACF values quickly approaching zero as the lag increases. We also see more of an alternating pattern between positive and negative values.