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

Assignment 2 - Due date 02/25/24

Emma Kaufman

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.\

Data set information

Consider the data provided in the spreadsheet

"Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"

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*.

```
#Importing data set without change the original file using read.xlsx
energy_data1 <- read_excel(path=
                        "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                          skip = 12,
                          sheet="Monthly Data",
                          col_names=FALSE)

#Now let's extract the column names from row 11
read_col_names <- read_excel(path=
```

```
                          "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
                             skip = 10,
                             n_max = 1,
                             sheet="Monthly Data",
                             col_names=FALSE)

colnames(energy_data1) <- read_col_names
head(energy_data1)
```

```
## # A tibble: 6 x 14
##   Month               `Wood Energy Production` `Biofuels Production`
##   <dttm>                                 <dbl> <chr>
## 1 1973-01-01 00:00:00                     130. Not Available
## 2 1973-02-01 00:00:00                     117. Not Available
## 3 1973-03-01 00:00:00                     130. Not Available
## 4 1973-04-01 00:00:00                     125. Not Available
## 5 1973-05-01 00:00:00                     130. Not Available
## 6 1973-06-01 00:00:00                     125. Not Available
## # i 11 more variables: `Total Biomass Energy Production` <dbl>,
## #   `Total Renewable Energy Production` <dbl>,
## #   `Hydroelectric Power Consumption` <dbl>,
## #   `Geothermal Energy Consumption` <dbl>, `Solar Energy Consumption` <chr>,
## #   `Wind Energy Consumption` <chr>, `Wood Energy Consumption` <dbl>,
## #   `Waste Energy Consumption` <dbl>, `Biofuels Consumption` <chr>,
## #   `Total Biomass Energy Consumption` <dbl>, ...
```

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.

```
#selecting desired columns
energy_interest <- select(energy_data1,
                          'Total Biomass Energy Production',
                          'Total Renewable Energy Production',
                          'Hydroelectric Power Consumption')

#previewing new dataframe
head(energy_interest)
```

```
## # A tibble: 6 x 3
##   Total Biomass Energy Productio~1 Total Renewable Ener~2 Hydroelectric Power ~3
##                              <dbl>                  <dbl>                   <dbl>
## 1                             130.                   220.                    89.6
## 2                             117.                   197.                    79.5
## 3                             130.                   219.                    88.3
## 4                             126.                   209.                    83.2
## 5                             130.                   216.                    85.6
## 6                             126.                   208.                    82.1
## # i abbreviated names: 1: `Total Biomass Energy Production`,
## #   2: `Total Renewable Energy Production`,
## #   3: `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().

```
#creating a timeseries object
energy_ts <- ts(energy_interest, start = c(1973,1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
#extracting each time series
biomass <- energy_ts[,1]
renewable <- energy_ts[,2]
hydro <- energy_ts[,3]

#computing mean
mean_bio <- mean(biomass)
mean_bio
```

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

```
mean_renewable <- mean(renewable)
mean_renewable
```

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

```
mean_hydro <- mean(hydro)
mean_hydro
```

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

```
#computing std dev
sd_bio <- sd(biomass)
sd_bio
```

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

```
sd_renewable <- sd(renewable)
sd_renewable
```

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

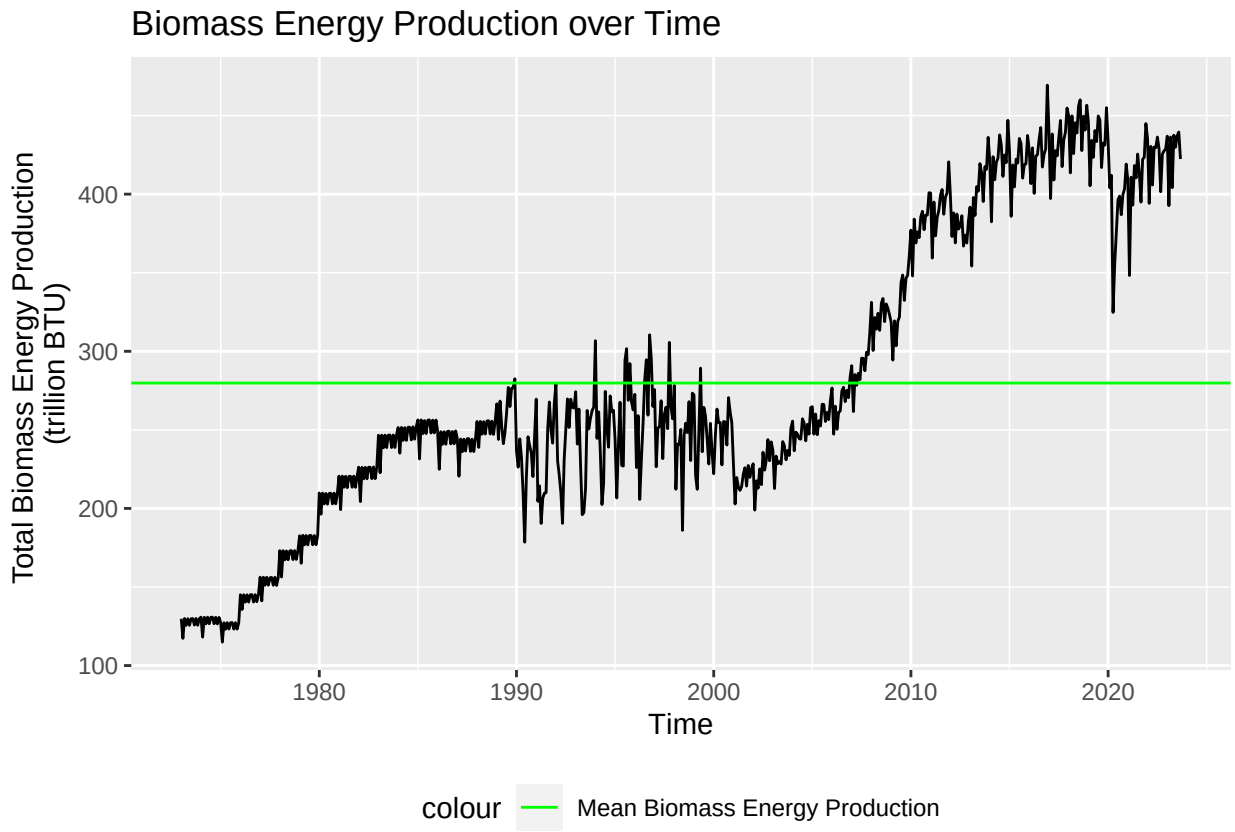
```
sd_hydro <- sd(hydro)
sd_hydro
```

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

Question 4

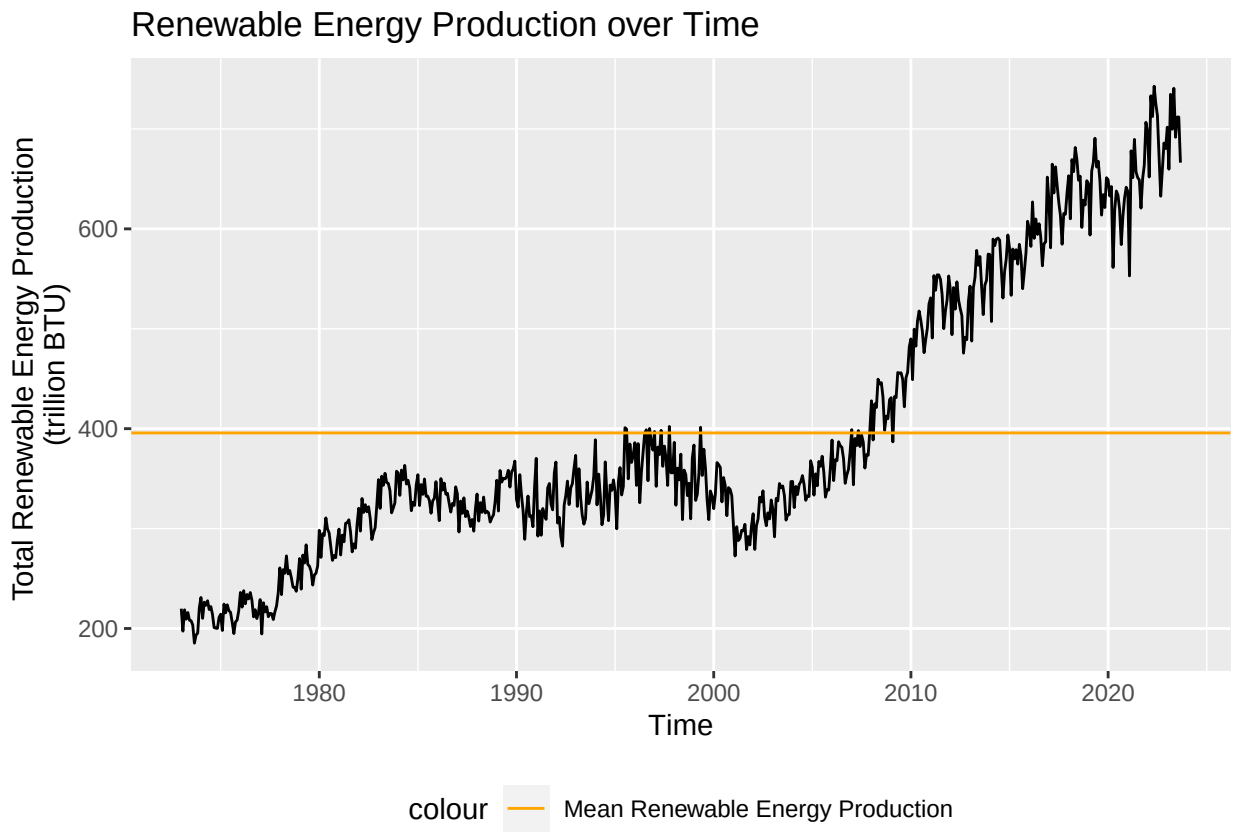
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
autoplot(biomass) +
  labs(x = "Time",
       y = "Total Biomass Energy Production \n(trillion BTU)",
       title = "Biomass Energy Production over Time") +
  #xlab("Time") +
  #ylab("Total Biomass Energy Production \n(trillion BTU)") +
  #ggtitle("Biomass Energy Production over Time")+
  geom_hline(aes(yintercept = mean_bio, color = "Mean Biomass Energy Production")) +
  scale_color_manual(values = c("Mean Biomass Energy Production" = "green")) +
  theme(legend.position = "bottom")
```

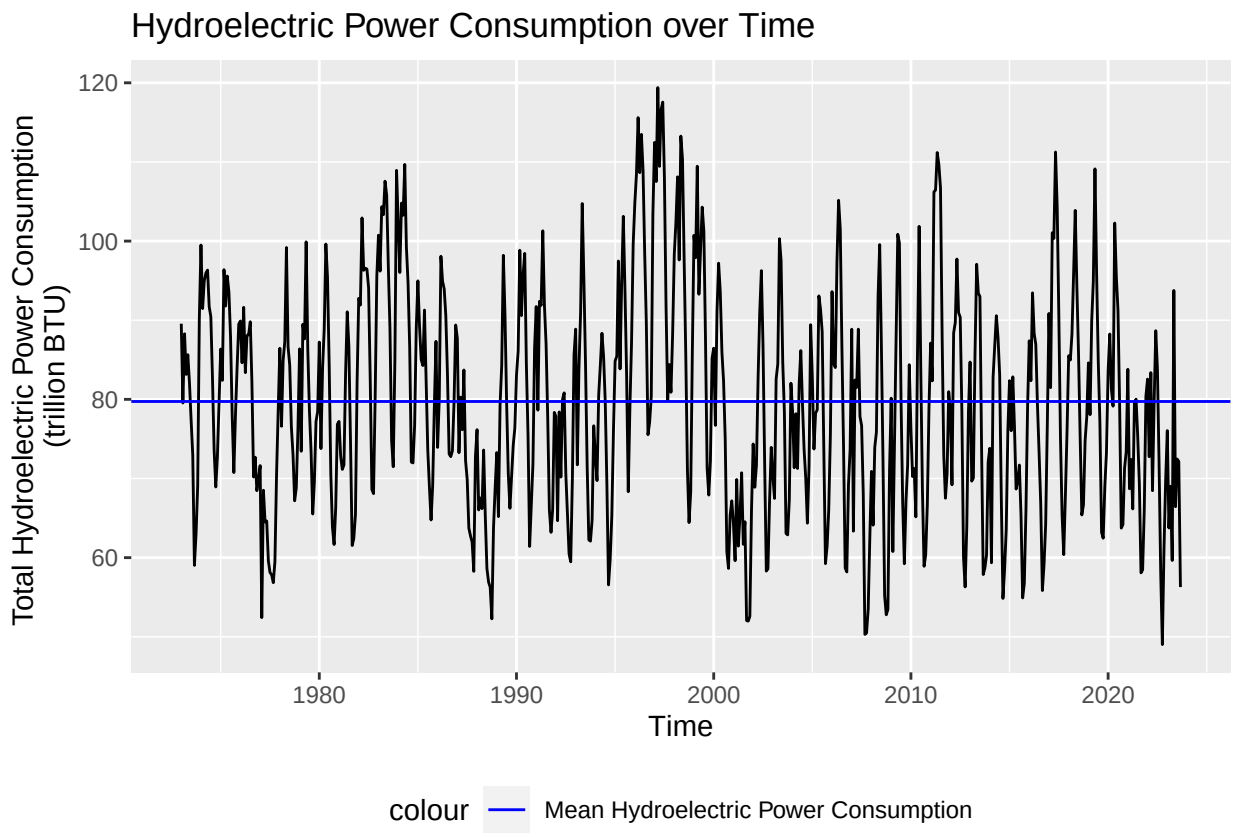



```
autoplot(renewable) +
  labs(x = "Time",
       y = "Total Renewable Energy Production \n(trillion BTU)",
       title = "Renewable Energy Production over Time") +
  #xlab("Time") +
  #ylab("Total Renewable Energy Production \n(trillion BTU)") +
  #ggtitle("Renewable Energy Production over Time")+
  geom_hline(aes(yintercept = mean_renewable, color = "Mean Renewable Energy Production")) +
```

```
  scale_color_manual(values = c("Mean Renewable Energy Production" = "orange")) +
  theme(legend.position = "bottom")
```




```
autoplot(hydro) +
  labs(x = "Time",
       y = "Total Hydroelectric Power Consumption \n(trillion BTU)",
       title = "Hydroelectric Power Consumption over Time") +
  #xlab("Time") +
  #ylab("Total Hydroelectric Power Consumption \n(trillion BTU)") +
  #ggtitle("Hydroelectric Power Consumption over Time")+
  geom_hline(aes(yintercept = mean_hydro, color = "Mean Hydroelectric Power Consumption")) +
  scale_color_manual(values = c("Mean Hydroelectric Power Consumption" = "blue")) +
  theme(legend.position = "bottom")
```



Interpretation: Hydroelectric power consumption doesn't follow a clear trend. The consumption is very noisy and overall I predict if we removed the noise the trend would be relatively flat over time. On the other hand, renewable energy and biomass energy production both show clear increases over time. They both show the rate of production increasing around the year 2000 as well, which is interesting. They both show relative dips in the year 2020.

Question 5

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

```
correlation <- cor(energy_ts)
correlation
```

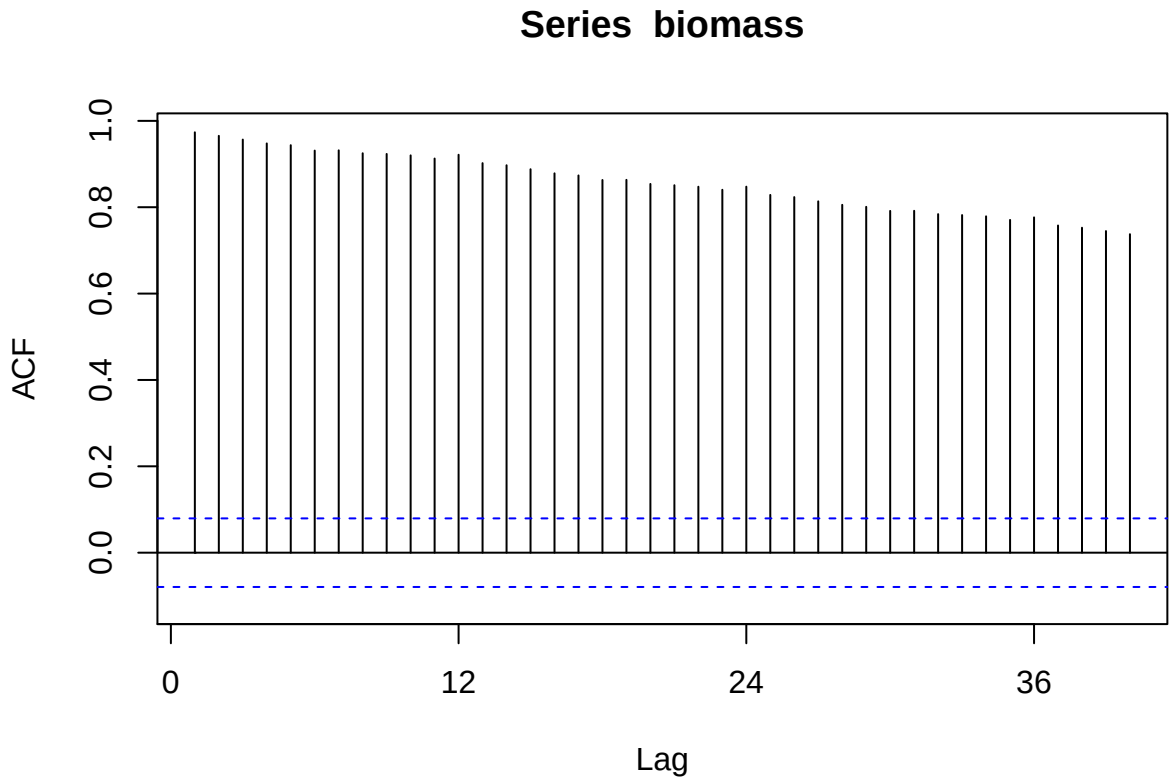
```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                       1.00000000
## Total Renewable Energy Production                     0.97074621
## Hydroelectric Power Consumption                      -0.09656318
##                                  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
```

Biomass energy production and total renewable energy production are significantly correlated (0.97). This makes sense if biomass is considered a renewable and is encompassed within the total renewable energy production. Hydroelectric power consumption is not significantly correlated with total renewable or total biomass energy production (-0.0017 and -0.0965). Consumption is not necessarily related to production if this hydroelectric power is not being used.

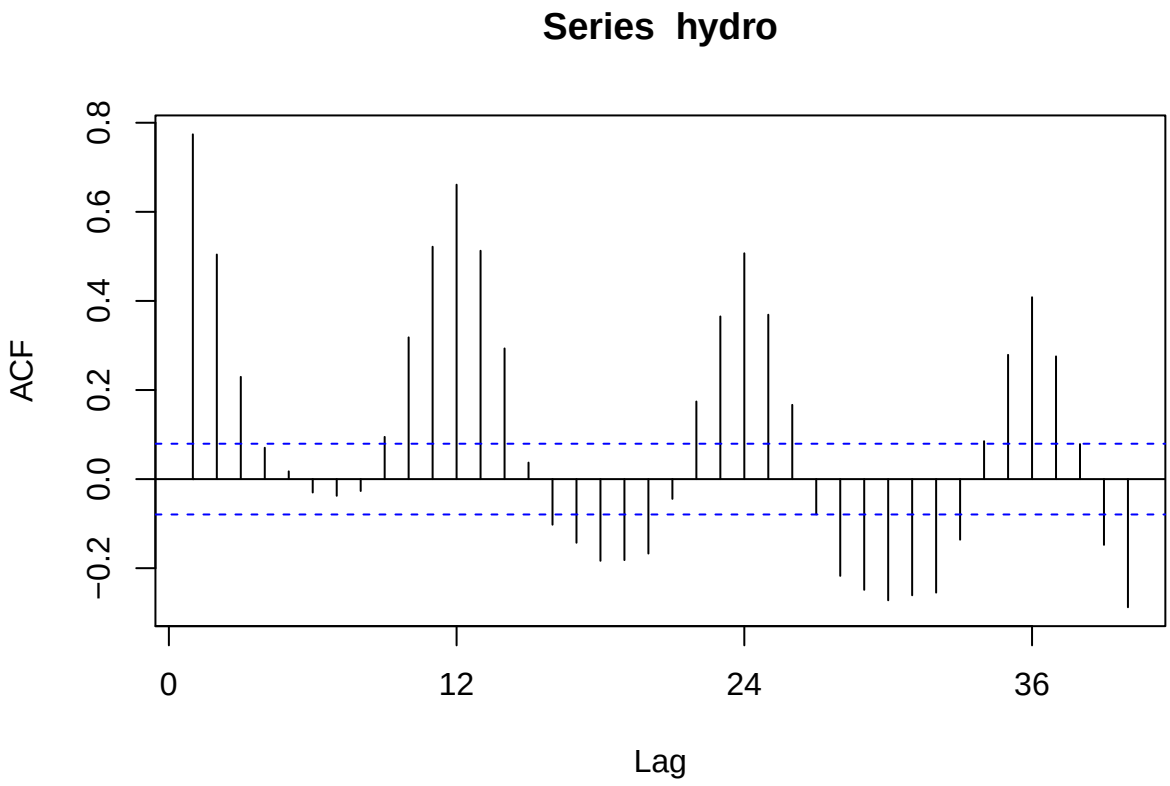
Question 6

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?

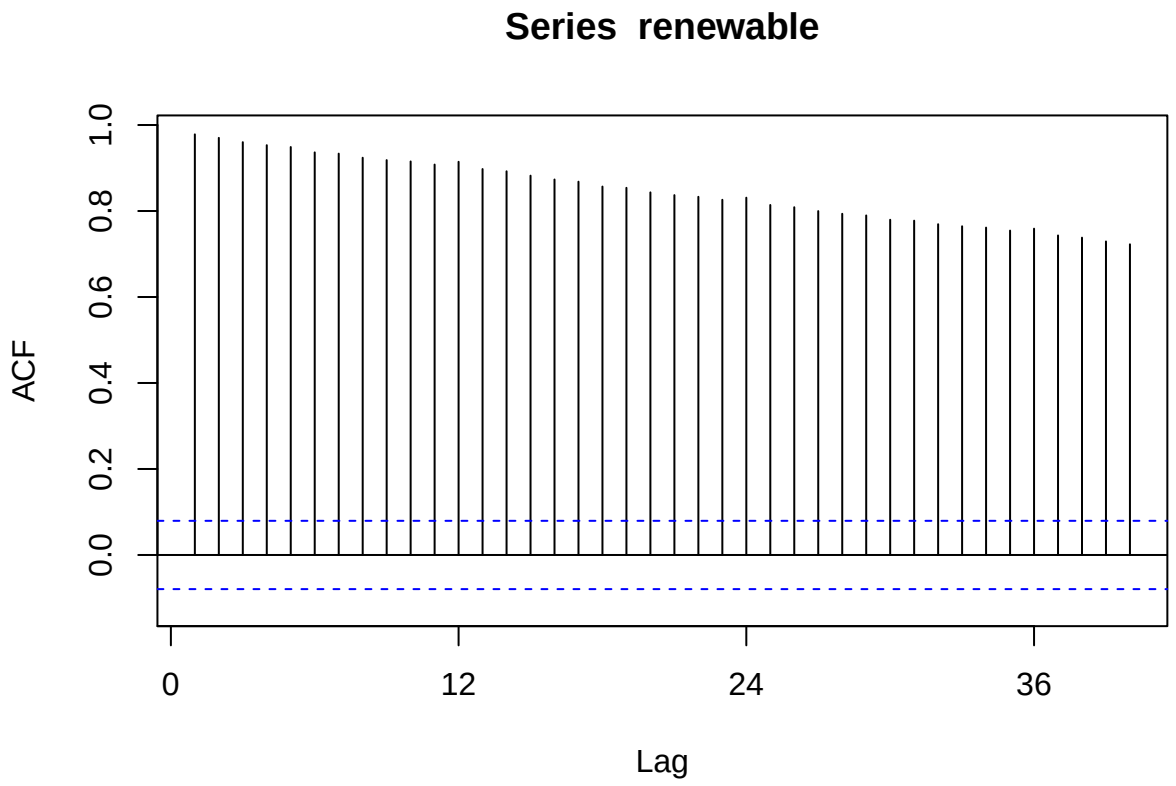
```
#acf plots with lags 1-40
Bio_acf= Acf(biomass,lag.max=40)
```




```
Hydro_acf= Acf(hydro,lag.max=40)
```




```
Renewable_acf= Acf(renewable,lag.max=40)
```

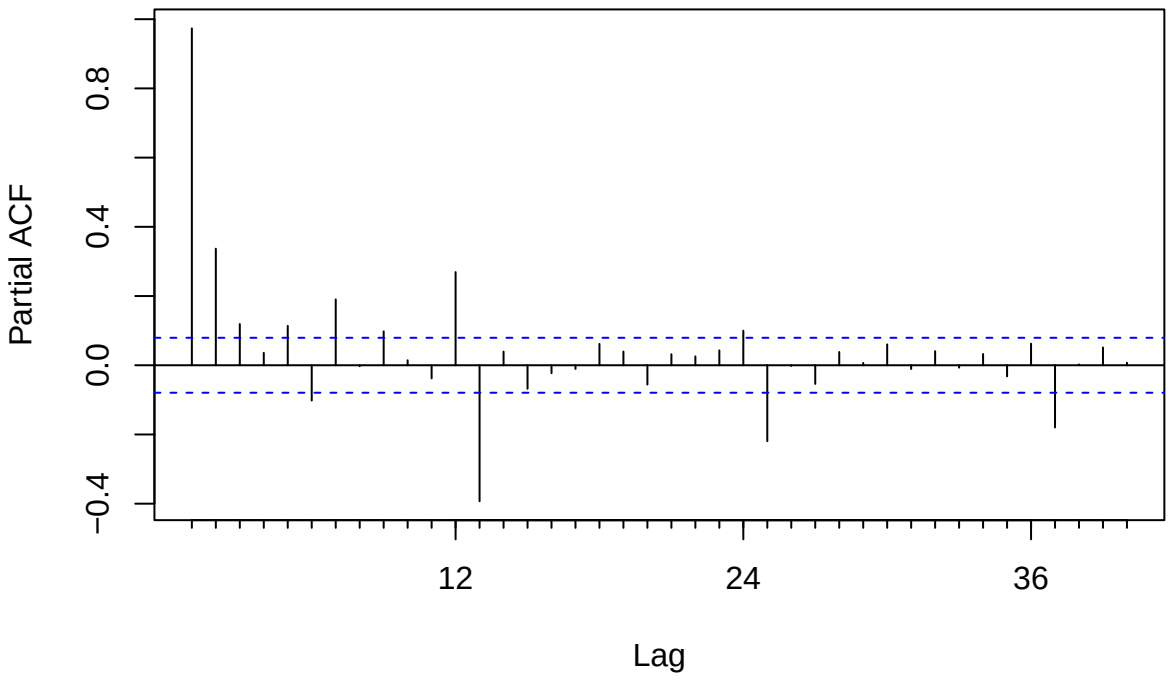


Biomass and renewable have very similar behavior (and were highly correlated above so this makes sense). The ACF plots show that biomass and renewable energy production are auto correlated. The highest auto correlation for both is at a 1 month lag; this auto correlation decreases with each additional lag. The hydroelectric power consumption shows a periodic autocorrelation, with the highest autocorrelation displaying at a one month lag, then the next highest at a year lag.

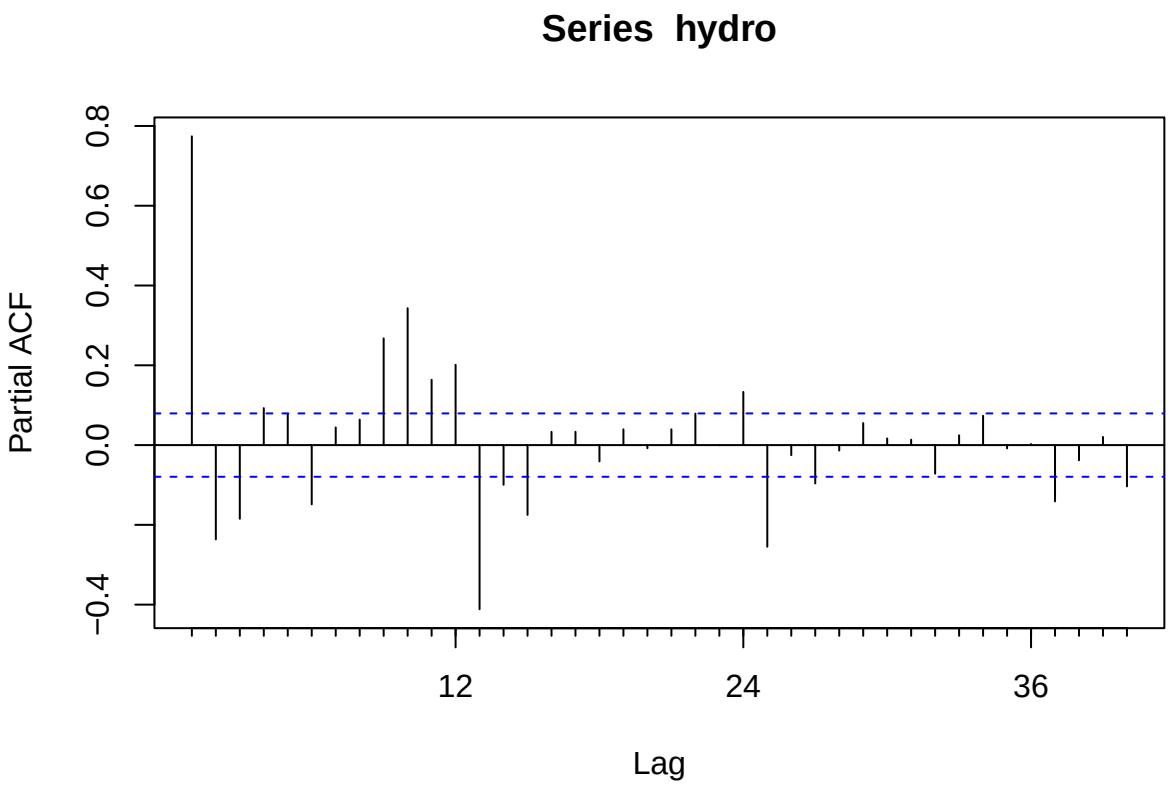
Question 7

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

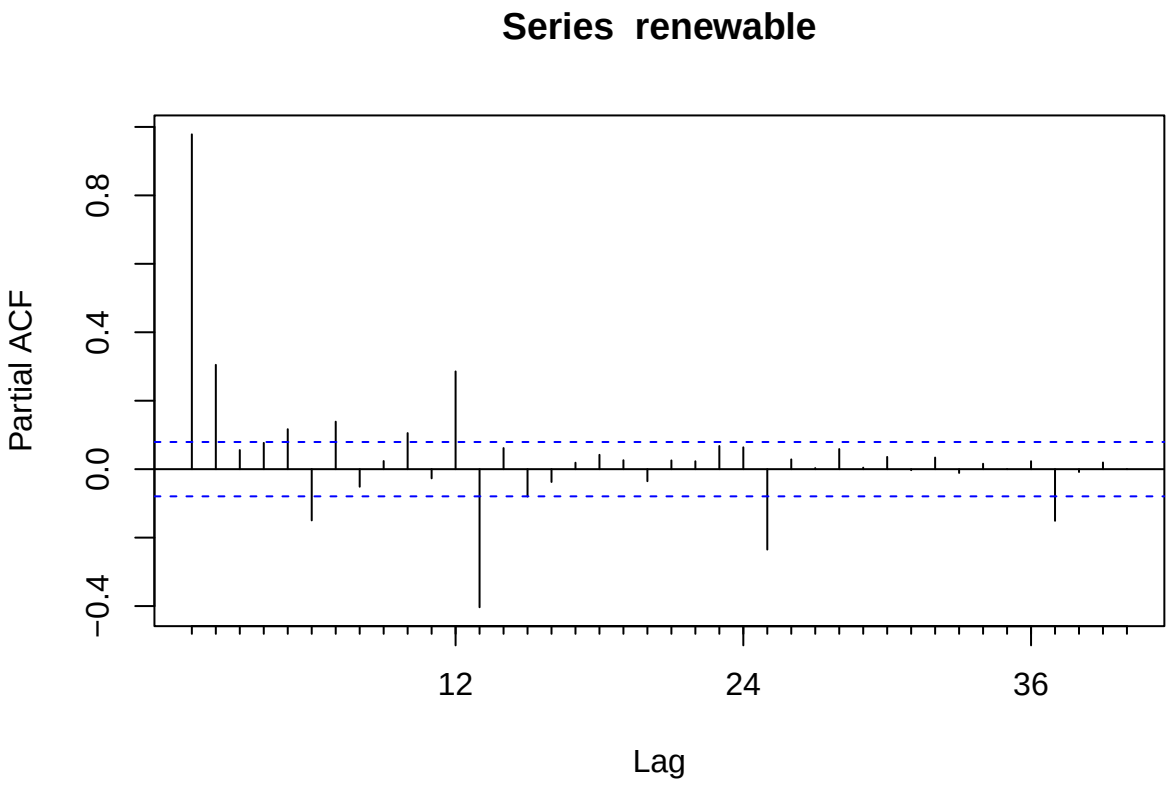
```
#pacf plots with lags 1-40
Bio_pacf= Pacf(biomass,lag.max=40)
```




```
Hydro_pacf= Pacf(hydro,lag.max=40)
```




```
Renewable_pacf= Pacf(renewable,lag.max=40)
```



The PACF plots are very different from the ACF plots. The three pacf plots look somewhat similar. They all show the highest pacf at the first lag. For biomass energy the pac decreases by 50% or more after the first lag. Compared to the acf biomass plot which almost has the same ac at the second lag. This means that most of this autocorrelatio is coming from the intermediate relationship. The same pattern is seen for the total renewable energy production. For the hydro power consumption there is still somewhat of a periodic relationship between pac and the lag, though not nearly as pronounced as the ACF plot. The second lag on the PACF hydro plot is negative and lower in magnitude than the second lag on the ACF plot, which also suggests that most of the autocorrelation is coming from the intermediate relationship. This is why the partial autocorrelation is important!