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

Assignment 2 - Due date 02/25/24

Cara Kuuskvere

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(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
```

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*.

```
#Importing data set
raw_energy_data <- read.csv(
  "Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.csv",
  header=FALSE,skip=12)

#as.data.frame(raw_energy_data)
```

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.

```
#trim the table
energy_data <- raw_energy_data[,4:6] #want all rows
                                 #all columns from 4 to 6
n_energy_sources <- ncol(energy_data) #number of variables
n_obs <- nrow(energy_data) #number observations

#Adding column names
colnames(energy_data)=c("Total Biomass Energy Production (Trillion Btu)",
                        "Total Renewable Energy Production (Trillion Btu)",
                        "Hydroelectric Power Consumption (Trillion Btu)")

head(energy_data)
```

```
##   Total Biomass Energy Production (Trillion Btu)
## 1                                        129.824
## 2                                        130.807
## 3                                        118.091
## 4                                        130.727
## 5                                        126.583
## 6                                        130.789
##   Total Renewable Energy Production (Trillion Btu)
## 1                                          220.755
## 2                                          231.010
## 3                                          210.188
## 4                                          226.384
## 5                                          223.218
## 6                                          227.793
##   Hydroelectric Power Consumption (Trillion Btu)
## 1                                         90.131
## 2                                            99.500
## 3                                            91.476
## 4                                            94.950
## 5                                            95.969
## 6                                            96.337
```

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_data <- ts(energy_data,start=c(1973,1),frequency = 12)
# frequency is 12 because it's monthly data, and months repeat every 12 entries!
# starts in year 1973
head(ts_energy_data)
```

```
##          Total Biomass Energy Production (Trillion Btu)
## Jan 1973                                        129.824
## Feb 1973                                        130.807
## Mar 1973                                        118.091
## Apr 1973                                        130.727
## May 1973                                        126.583
## Jun 1973                                        130.789
##          Total Renewable Energy Production (Trillion Btu)
## Jan 1973                                          220.755
## Feb 1973                                          231.010
## Mar 1973                                          210.188
## Apr 1973                                          226.384
## May 1973                                          223.218
## Jun 1973                                          227.793
##          Hydroelectric Power Consumption (Trillion Btu)
## Jan 1973                                         90.131
## Feb 1973                                         99.500
## Mar 1973                                         91.476
## Apr 1973                                         94.950
## May 1973                                         95.969
## Jun 1973                                         96.337
```

Question 3

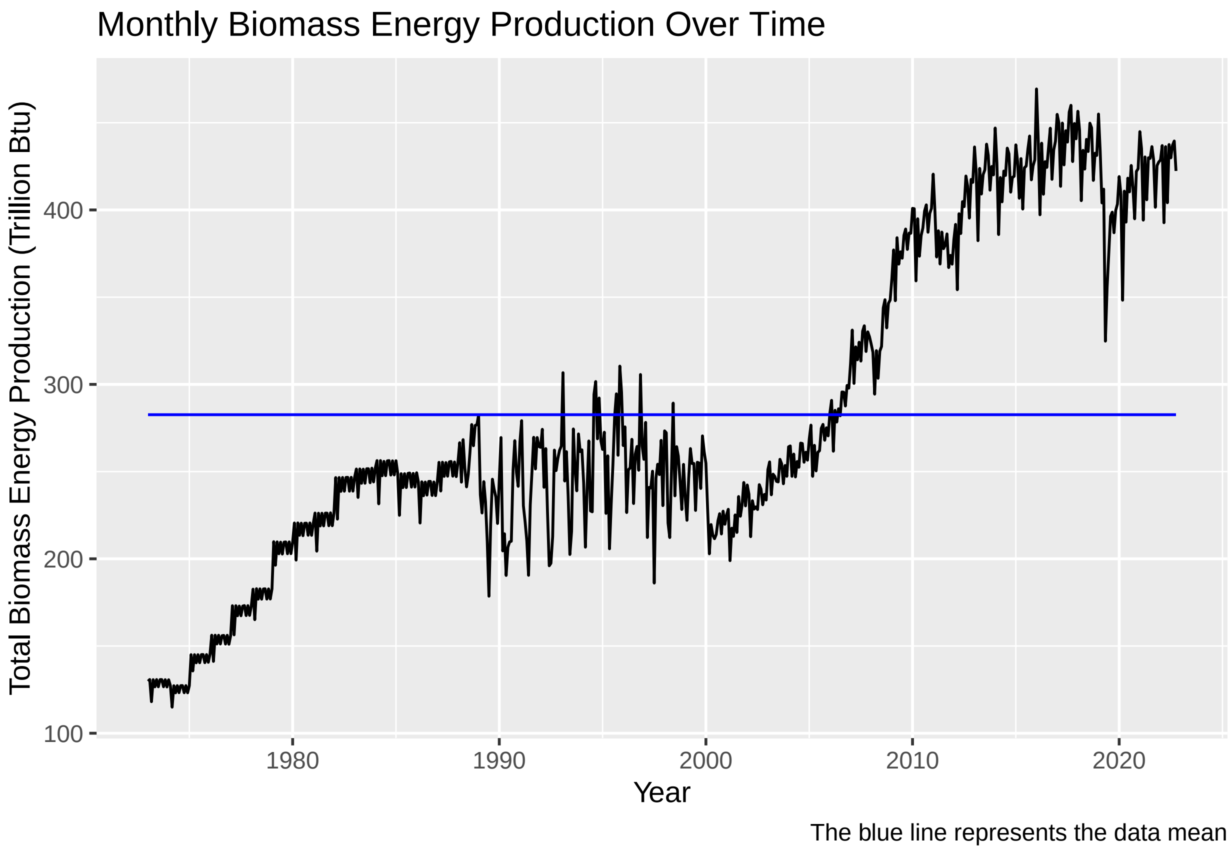
Compute mean and standard deviation for these three series.

```
#Means of each series
mean_ts_biomass <- mean(ts_energy_data[,1])
mean_ts_RE <- mean(ts_energy_data[,2])
mean_ts_hydro <- mean(ts_energy_data[,3])

#Standard Dev of each series
#...data is clean so not worried about missing values
sd_ts_biomass <- sd(ts_energy_data[,1])
sd_ts_RE <- sd(ts_energy_data[,2])
sd_ts_hydro <- sd(ts_energy_data[,3])
```

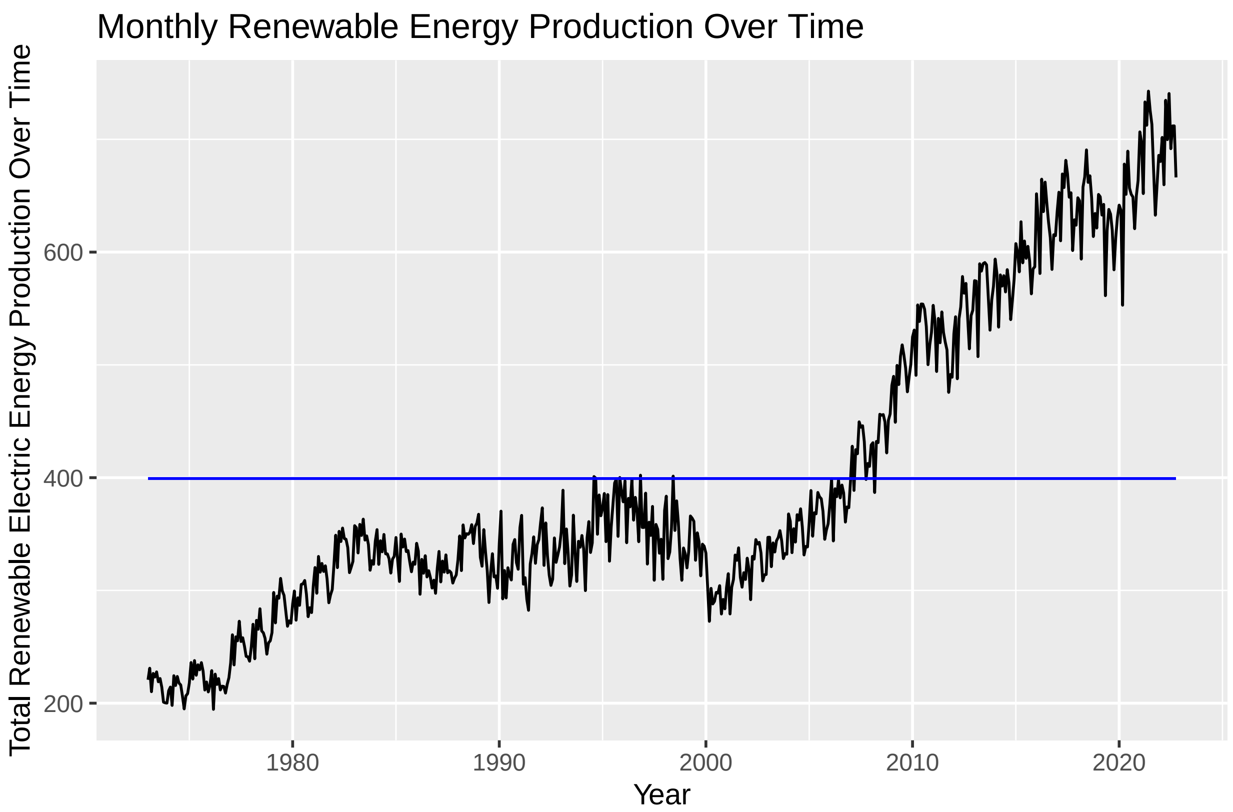
Question 4

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.

```
library(ggplot2)
biomass_ts_plt <- autoplot(ts_energy_data[,1]) +
  labs(x="Year",y= "Total Biomass Energy Production (Trillion Btu)",
    title = "Monthly Biomass Energy Production Over Time") +
  geom_line(aes(y=mean_ts_biomass),color="blue")+
  labs(caption="The blue line represents the data mean")

biomass_ts_plt
```

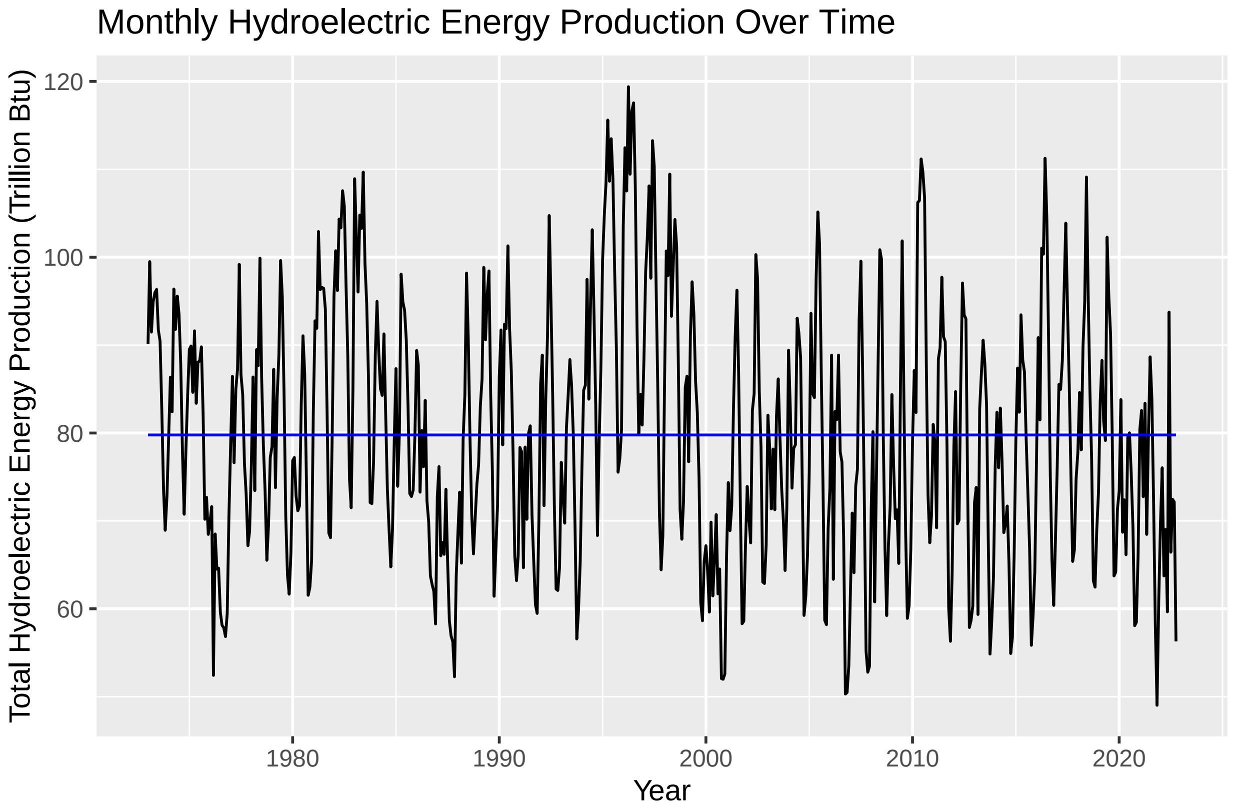


```
RE_ts_plt <- autoplot(ts_energy_data[,2]) +
  labs(x="Year",y= "Total Renewable Electric Energy Production Over Time",
  title = "Monthly Renewable Energy Production Over Time")+
  geom_line(aes(y=mean_ts_RE),color="blue")+
  labs(caption="The blue line represents the data mean")

RE_ts_plt
```



The blue line represents the data mean

```
hydro_ts_plt <- autoplot(ts_energy_data[,3]) +
  labs(x="Year",y= "Total Hydroelectric Energy Production (Trillion Btu)",
    title = "Monthly Hydroelectric Energy Production Over Time") +
  geom_line(aes(y=mean_ts_hydro),color="blue")+
  labs(caption="The blue line represents the data mean")

hydro_ts_plt
```



The blue line represents the data mean

Question 5

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

```
correlation <- cor(ts_energy_data)
correlation
```

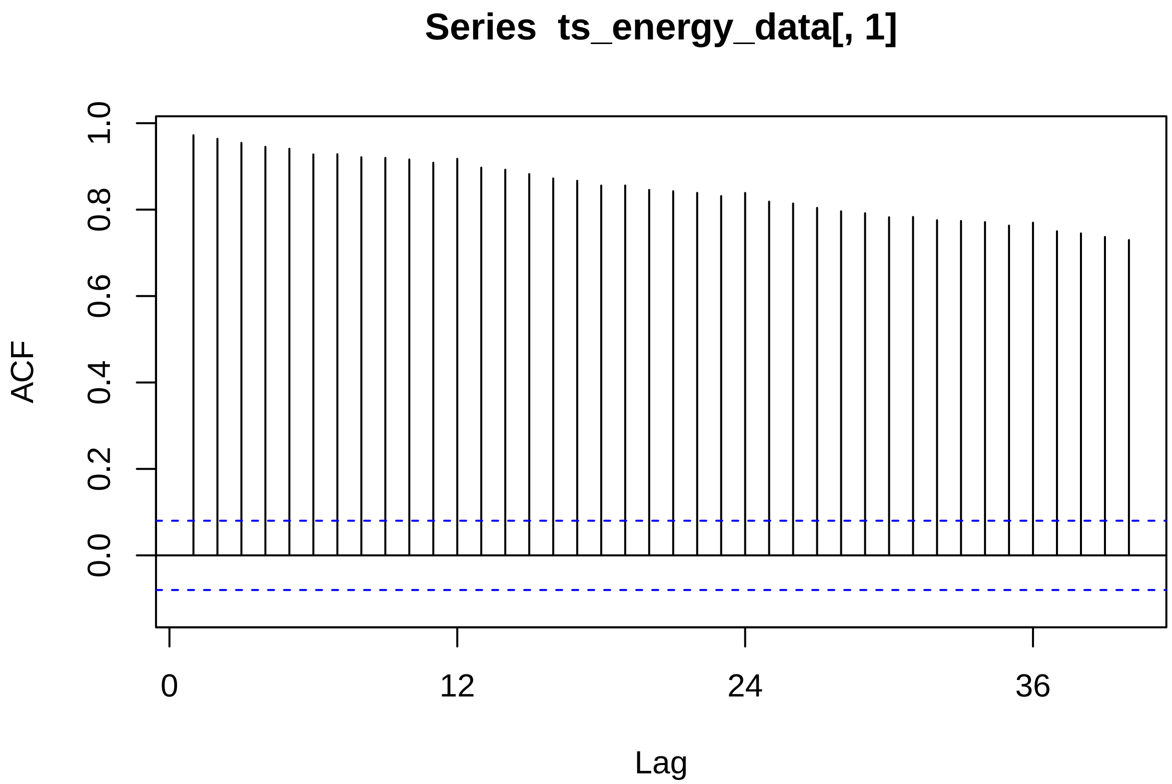
```
##                                                   Total Biomass Energy Production (Trillion Btu)
## Total Biomass Energy Production (Trillion Btu)                                          1.0000000
## Total Renewable Energy Production (Trillion Btu)                                        0.9701793
## Hydroelectric Power Consumption (Trillion Btu)                                         -0.1050690
##                                                   Total Renewable Energy Production (Trillion Btu)
## Total Biomass Energy Production (Trillion Btu)                                          0.970179323
## Total Renewable Energy Production (Trillion Btu)                                        1.000000000
## Hydroelectric Power Consumption (Trillion Btu)                                         -0.007315341
##                                                   Hydroelectric Power Consumption (Trillion Btu)
## Total Biomass Energy Production (Trillion Btu)                                     -0.105068970
## Total Renewable Energy Production (Trillion Btu)                                   -0.007315341
## Hydroelectric Power Consumption (Trillion Btu)                                      1.000000000
```

Biomass and renewable energy are strongly correlated with one another (their values are close to 1), whereas hydro is not strongly correlated with renewable energy and biomass due to its smaller values. Hydro is more strongly correlated to biomass than it is to renewable energy consumption as the absolute value of the correlation of biomass is greater than .05, and the absolute value fo the correlation coefficient of hydro and RE is very small.

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?

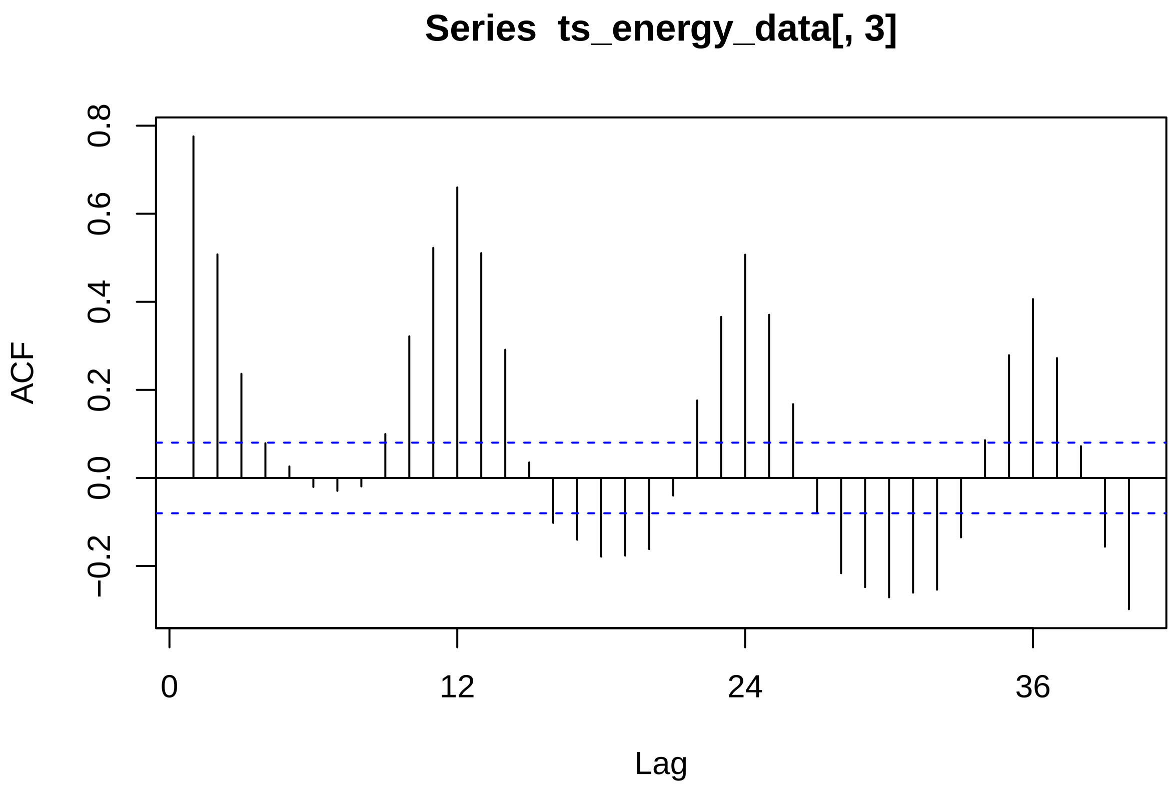
```
biomass_acf <- Acf(ts_energy_data[,1],lag.max=40)
```



```
biomass_acf
```

```
##
## Autocorrelations of series 'ts_energy_data[, 1]', by lag
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12
## 1.000 0.972 0.964 0.954 0.945 0.941 0.928 0.928 0.921 0.920 0.916 0.909 0.918
##    13    14    15    16    17    18    19    20    21    22    23    24    25
## 0.897 0.892 0.882 0.872 0.867 0.856 0.856 0.846 0.843 0.839 0.831 0.839 0.819
##    26    27    28    29    30    31    32    33    34    35    36    37    38
## 0.814 0.804 0.796 0.792 0.782 0.783 0.775 0.774 0.771 0.763 0.770 0.750 0.745
##    39    40
## 0.737 0.730
```

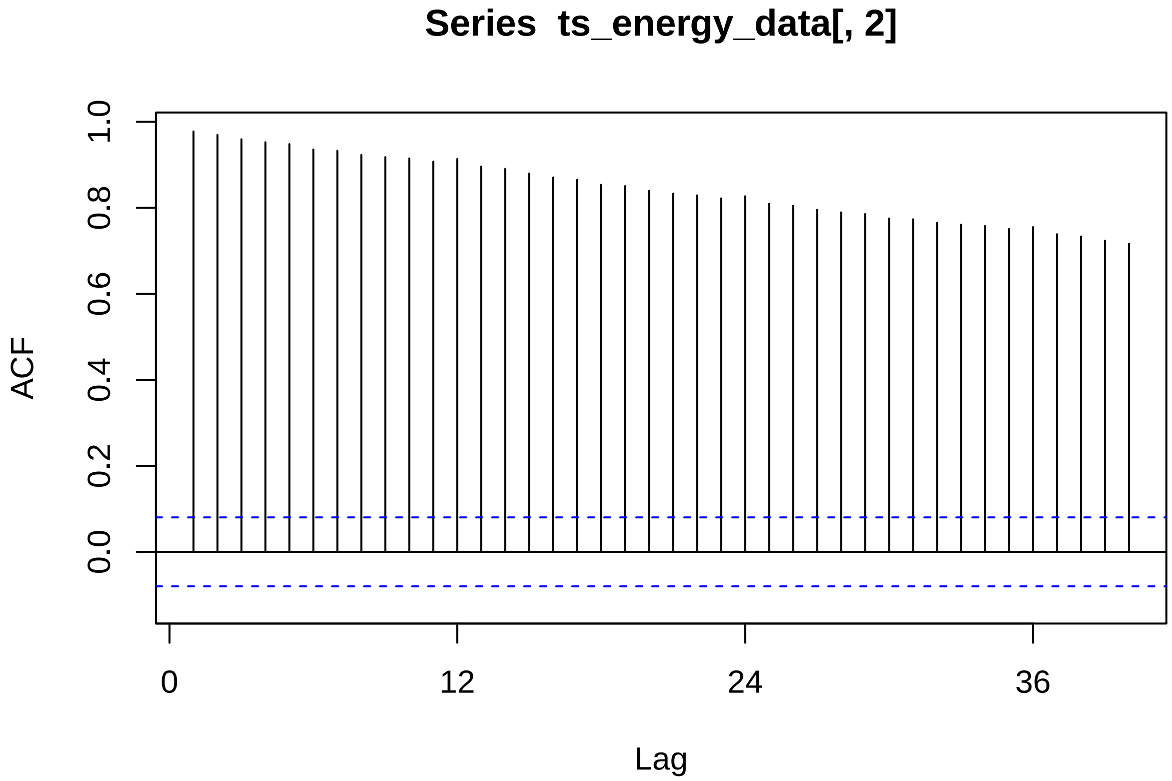
```
hydro_acf <- Acf(ts_energy_data[,3],lag.max=40)
```



```
hydro_acf
```

```
##
## Autocorrelations of series 'ts_energy_data[, 3]', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.776  0.508  0.237  0.079  0.026 -0.021 -0.029 -0.019  0.100  0.322
##     11     12     13     14     15     16     17     18     19     20     21
##  0.523  0.660  0.511  0.291  0.035 -0.102 -0.140 -0.179 -0.176 -0.162 -0.040
##     22     23     24     25     26     27     28     29     30     31     32
##  0.176  0.366  0.507  0.371  0.168 -0.078 -0.216 -0.248 -0.271 -0.260 -0.254
##     33     34     35     36     37     38     39     40
## -0.135  0.086  0.279  0.406  0.272  0.072 -0.156 -0.298
```

```
RE_acf <- Acf(ts_energy_data[,2],lag.max=40)
```



```
RE_acf
```

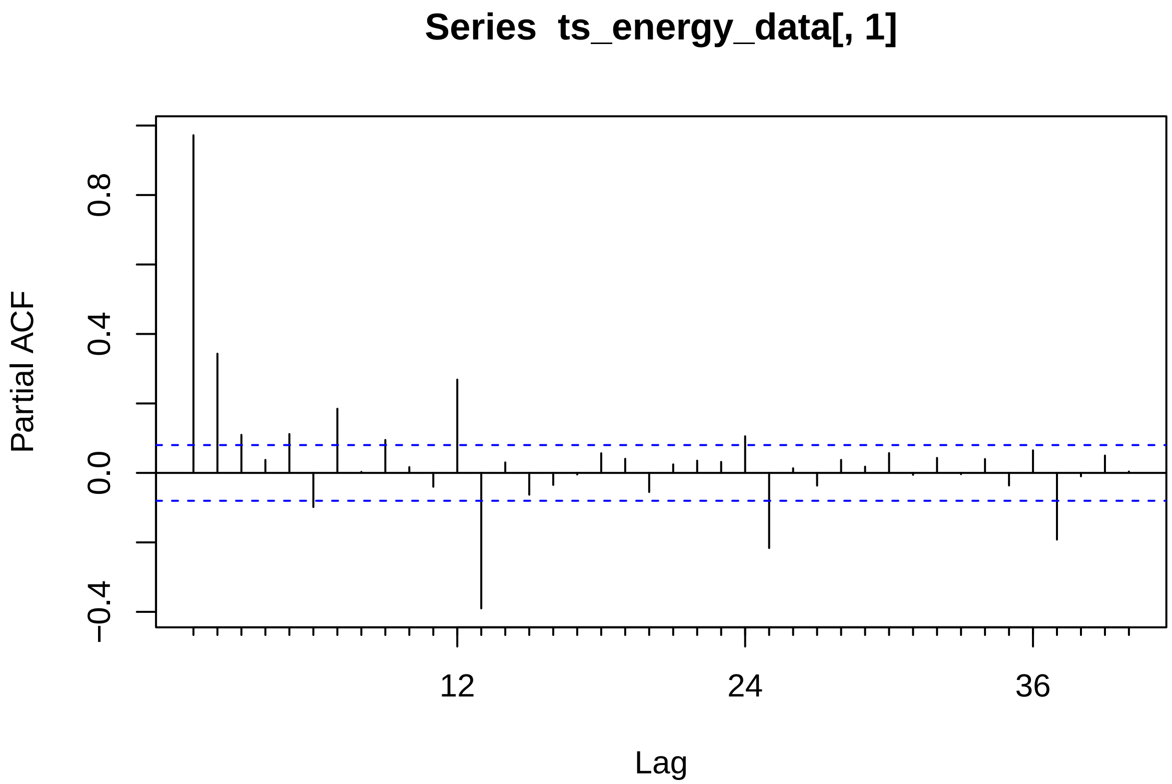
```
##
## Autocorrelations of series ’ts_energy_data[, 2]’, by lag
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12
## 1.000 0.978 0.970 0.959 0.953 0.948 0.936 0.933 0.923 0.918 0.915 0.908 0.914
##    13    14    15    16    17    18    19    20    21    22    23    24    25
## 0.896 0.891 0.880 0.871 0.865 0.854 0.851 0.840 0.833 0.829 0.822 0.827 0.809
##    26    27    28    29    30    31    32    33    34    35    36    37    38
## 0.805 0.795 0.789 0.785 0.775 0.773 0.765 0.761 0.758 0.751 0.755 0.739 0.733
##    39    40
## 0.724 0.717
```

Biomass and renewable energy follow a similar acf, whereas hydro varies much more over time. Biomass and renewable energy show a strong (close to 1) acf which falls over time, whereas it looks like hydro has some seasonality to its acf. All the biomass and renewable values, along with many of the biomass values, are outside the blue indication of statistical significance. The ACF shows you that you have an autoregressive component to your model.

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?

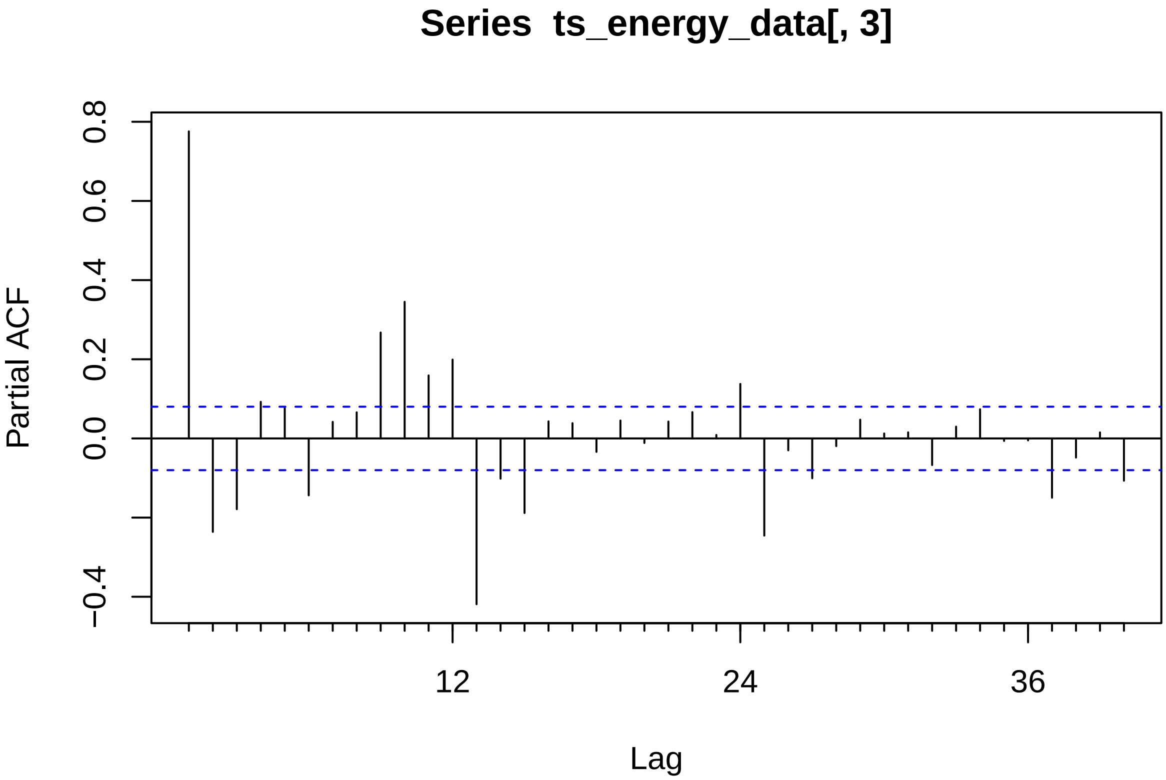
```
biomass_pacf <- Pacf(ts_energy_data[,1],lag.max=40)
```



```
biomass_pacf
```

```
##
## Partial autocorrelations of series 'ts_energy_data[, 1]', by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
##  0.972  0.343  0.110  0.037  0.112 -0.098  0.185  0.003  0.095  0.017 -0.040
##     12     13     14     15     16     17     18     19     20     21     22
##  0.269 -0.390  0.030 -0.063 -0.035 -0.005  0.057  0.041 -0.055  0.025  0.035
##     23     24     25     26     27     28     29     30     31     32     33
##  0.032  0.106 -0.216  0.013 -0.037  0.038  0.018  0.057 -0.006  0.043 -0.004
##     34     35     36     37     38     39     40
##  0.040 -0.036  0.065 -0.192 -0.010  0.050  0.004
```

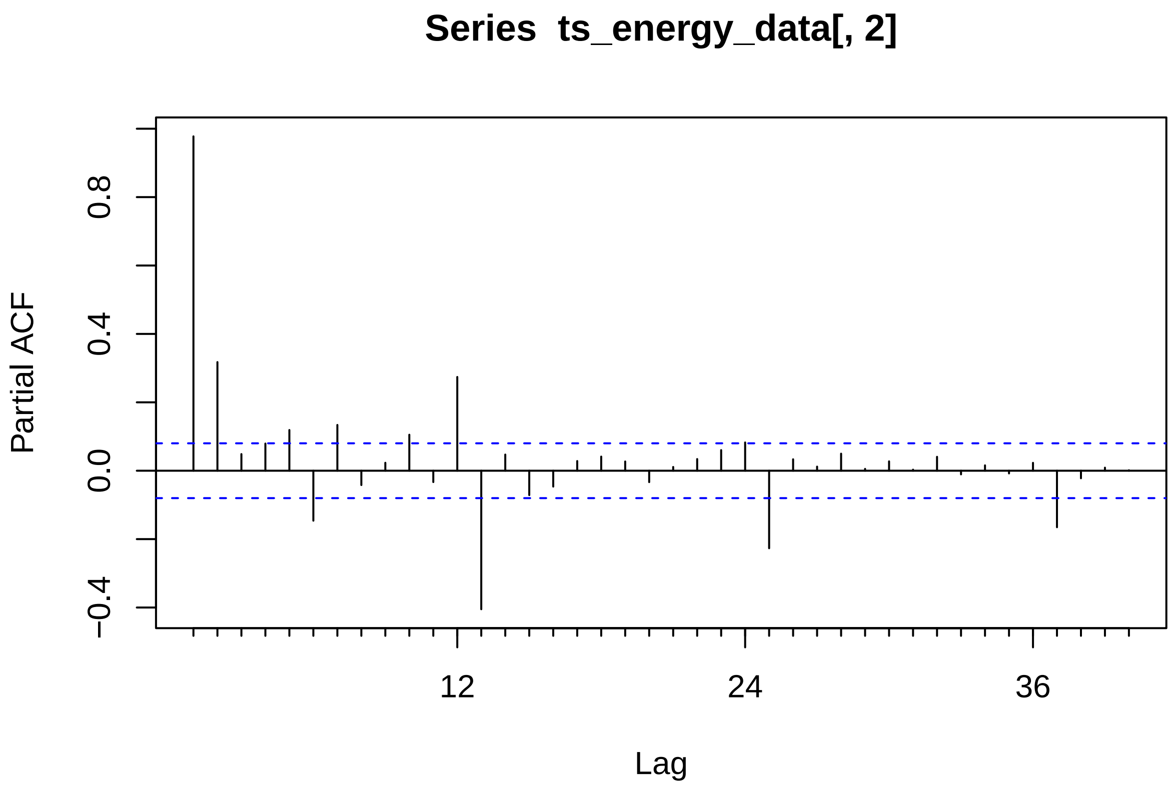
```
hydro_pacf <- Pacf(ts_energy_data[,3],lag.max=40)
```



```
hydro_pacf
```

```
##
## Partial autocorrelations of series 'ts_energy_data[, 3]', by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
##  0.776 -0.236 -0.179  0.092  0.079 -0.144  0.042  0.066  0.268  0.345  0.159
##     12     13     14     15     16     17     18     19     20     21     22
##  0.199 -0.419 -0.102 -0.188  0.043  0.039 -0.034  0.045 -0.012  0.043  0.067
##     23     24     25     26     27     28     29     30     31     32     33
##  0.009  0.138 -0.245 -0.030 -0.101 -0.019  0.047  0.013  0.015 -0.067  0.030
##     34     35     36     37     38     39     40
##  0.074 -0.006 -0.005 -0.150 -0.048  0.015 -0.107
```

```
RE_pacf <- Pacf(ts_energy_data[,2],lag.max=40)
```



```
RE_pacf
```

```
##
## Partial autocorrelations of series ’ts_energy_data[, 2]’, by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
##  0.978  0.318  0.049  0.079  0.119 -0.146  0.134 -0.042  0.023  0.105 -0.033
##     12     13     14     15     16     17     18     19     20     21     22
##  0.274 -0.405  0.047 -0.071 -0.046  0.028  0.041  0.027 -0.033  0.011  0.034
##     23     24     25     26     27     28     29     30     31     32     33
##  0.060  0.083 -0.227  0.033  0.012  0.050  0.005  0.027  0.003  0.041 -0.011
##     34     35     36     37     38     39     40
##  0.016 -0.008  0.023 -0.165 -0.022  0.009  0.001
```

These plots look far more similar to one another than the acf plots. They show far fewer correlations over time, with the strength of correlation falling as the lag increases. They also show far fewer plots beyond the blue indication of statistical significance. This is due to the partial autocorrelation removing the influence of intermediate variables to give us which lags are most important in the model.