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

Assignment 2 - Due date 01/26/22

Rob Kravec

## R packages

```
library(tidyverse)
library(forecast)
library(tseries)
library(readxl)
library(lubridate)
library(patchwork)
```

## Data set information

I'll read in the data using the `read_excel` function from the `readxl` package because this function allows me to specify the sheet that I want to read from the `.xlsx` file. Though not specified in the assignment, I'll use the `Monthly Data`. The amount of data is still pretty small, and I can always aggregate to the annual view, if needed.

```
# Read in data
file_path = paste0('../Data/Table_10.1_Renewable_Energy_Production_and',
                   '_Consumption_by_Source.xlsx')
data <- read_excel(path = file_path, sheet = "Monthly Data", skip = 10,
                   na = "Not Available")

# Remove first row, which contains units for each column
# Remove second row, which does not specify a date for the data (though Jan
# 1973 seems likely)
data <- data[-1, ]
```

## 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.

```
# Select columns
data_q1 <- data %>%
  select('Total Biomass Energy Production', 'Total Renewable Energy Production',
         'Hydroelectric Power Consumption')

# Make column names shorter and remove spaces
data_q1 <- data_q1 %>%
  rename(Biomass_prod = 'Total Biomass Energy Production',
         Renewable_prod = 'Total Renewable Energy Production',
         Hydro_consumption = 'Hydroelectric Power Consumption')
```

```r
# Convert data types to numeric
data_q1 <- sapply(data_q1, as.numeric) %>%
  as_tibble()

# Show first six rows
head(data_q1)
```

```
## # A tibble: 6 x 3
##   Biomass_prod Renewable_prod Hydro_consumption
##          <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.
```

## 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
# Frequency of 12 chosen to represent monthly data
data_ts <- ts(data = data_q1, start = c(1973, 1), end = c(2021, 9),
              frequency = 12)

# Display result
head(data_ts)
```

```
##          Biomass_prod Renewable_prod Hydro_consumption
## Jan 1973      129.787        403.981           272.703
## Feb 1973      117.338        360.900           242.199
## Mar 1973      129.938        400.161           268.810
## Apr 1973      125.636        380.470           253.185
## May 1973      129.834        392.141           260.770
## Jun 1973      125.611        377.232           249.859
```

## Question 3

Compute mean and standard deviation for these three series.

```r
# Define function that returns mean and standard deviation
mean_sd <- function(x) {
  c(mean(x), sd(x))
}

# Calculate mean and standard deviation for each column
mean_sd_results <- sapply(data_ts, mean_sd)

# Rename rows
row.names(mean_sd_results) <- c('Mean', 'Standard_deviation')

# Display results
mean_sd_results
```

```
##                       Biomass_prod Renewable_prod Hydro_consumption
## Mean                     273.78392       581.1708         235.96526
## Standard_deviation        89.42852       177.5607          44.01749
```

## Question 4

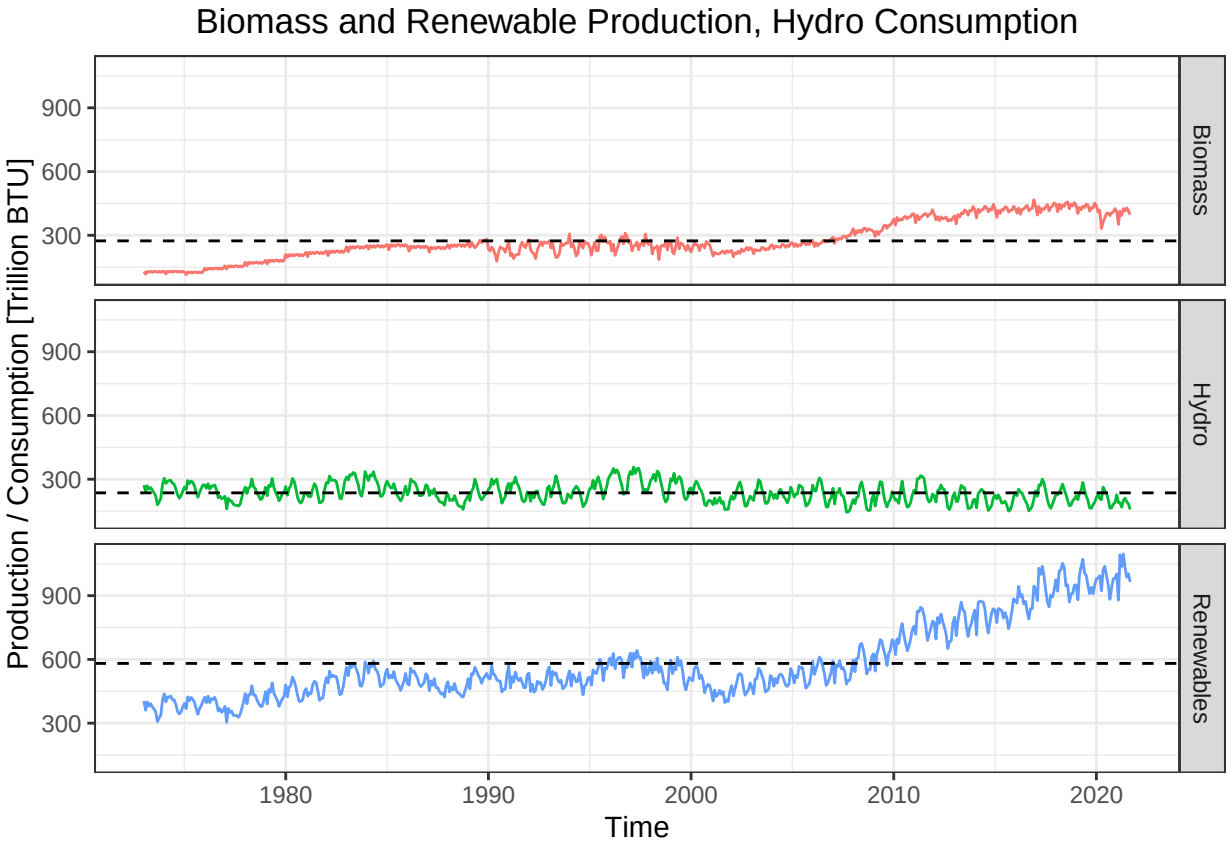
Display and interpret the time series plot for each of these variables.

For this question, I decided to make a faceted plot and denote the mean value of each time series with a dashed black line. A few observations are apparent:

- While `Total Renewable Energy Production` and `Total Biomass Energy Production` rose substantially from 1973-2021, `Hydroelectric Power Consumption` decreased
- Given the rise of `Total Biomass Energy Production` and the stagnation of `Hydroelectric Power Consumption`, the mean biomass production is greater than the mean hydroelectric consumption over the given time period. The mean value of `Total Renewable Energy Production` is obviously the highest of the three categories shown, given that the biomass and hydroelectric variables are two of the inputs of the sum that produces `Total Renewable Energy Production`
- The gap between `Renewable Energy Production` and `Total Biomass Energy Production` + `Hydroelectric Power Consumption` grew over time, denoting the rise of other renewable energy technologies (e.g., wind, solar)

```
# Create dataframe with means
mean_df <- data.frame(name = c('Biomass', 'Renewables', 'Hydro'),
                      means = mean_sd_results[1,])

# Generate requested plot
data %>%
  select(Month, 'Total Biomass Energy Production',
         'Total Renewable Energy Production',
         'Hydroelectric Power Consumption') %>%
  rename(Biomass = 'Total Biomass Energy Production',
         Renewables = 'Total Renewable Energy Production',
         Hydro = 'Hydroelectric Power Consumption') %>%
  pivot_longer(cols = !Month) %>%
  mutate(Month = ymd(Month), value = as.numeric(value)) %>%
  ggplot(mapping = aes(x = Month, y = value, color = name)) +
  geom_line() +
  facet_grid(name~.) +
  geom_hline(data = mean_df, mapping = aes(yintercept = means),
             color = 'black', linetype = 2) +
  labs(x = 'Time', y = 'Production / Consumption [Trillion BTU]',
       title = 'Biomass and Renewable Production, Hydro Consumption') +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "None")
```




## Question 5

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

```
cor(data_ts)
```

```
##                   Biomass_prod Renewable_prod Hydro_consumption
## Biomass_prod         1.0000000     0.92328377       -0.28049970
## Renewable_prod       0.9232838     1.00000000       -0.05680651
## Hydro_consumption   -0.2804997    -0.05680651        1.00000000
```

`Total Renewable Energy Production` and `Total Biomass Energy Production` have a strong positive correlation, which is consistent with our previous observation that `Total Renewable Energy Production` and `Total Biomass Energy Production` both rose substantially from 1973-2021. `Hydroelectric Power Consumption` and `Total Biomass Energy Production` have a weak negative correlation, which also makes sense given moderate rise in biomass production and slow fall in hydroelectric consumption. `Hydroelectric Power Consumption` and `Total Renewable Energy Production` have a very weak negative correlation, bordering on no correlation at all. This result is less intuitive for me, as I would have expected a slightly stronger 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?

For this question, I'll assume that the lags should be in terms of 40 months, not 40 years.

A few observations:

- Lag 1 autocorrelation is quite high for all three variables
- `Total Renewable Energy Production` and `Total Biomass Energy Production` have similar shapes to their acf plots, which show relatively high autocorrelation all the way out to 40 months (with a slow decline in acf as lag increases)
- `Hydroelectric Power Consumption` acf has more of a cyclical pattern to it, suggesting there may be some seasonality at play

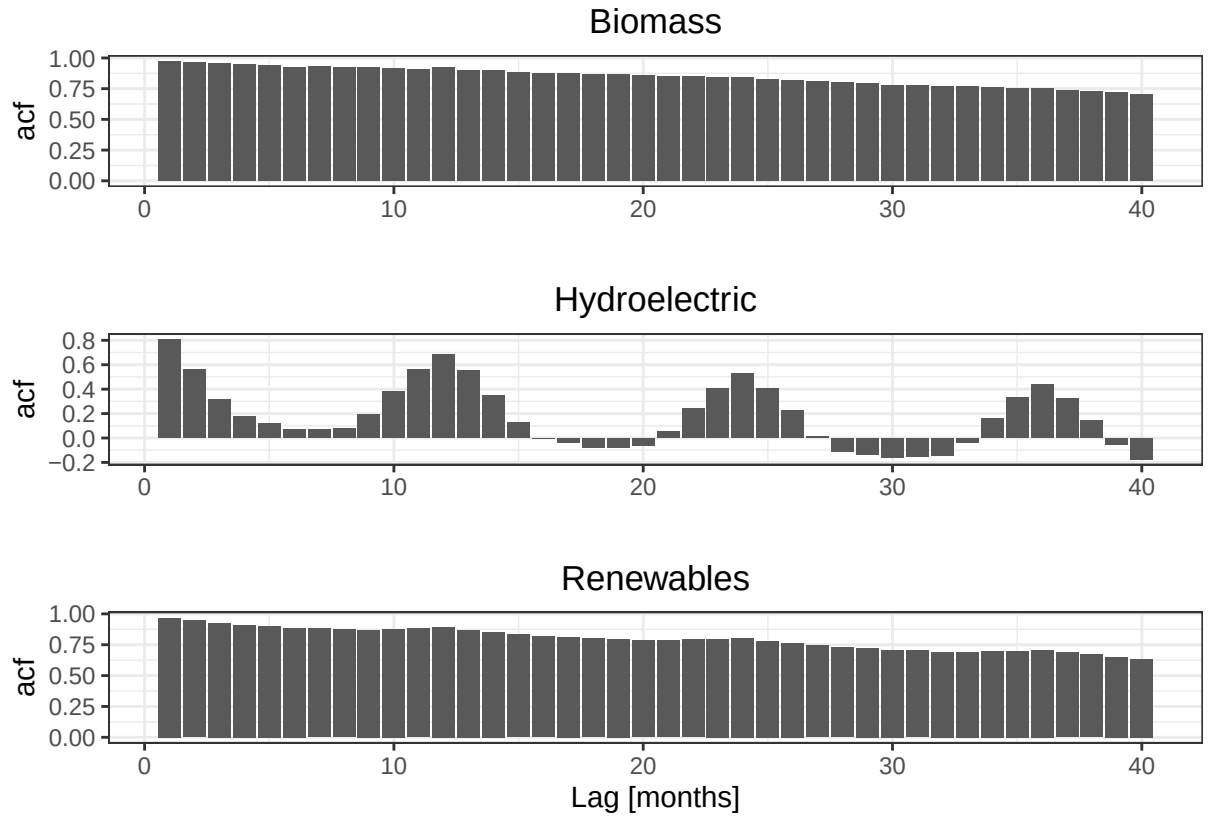
```
# Create function for acf plots
plot_acf <- function(ts, col_num, lag_amt, xlab, title) {
  # Generate acf data
  acf_data <- data.frame(lag = 1:lag_amt,
                         acf = acf(ts[,col_num],
                                   plot = F,
                                   lag.max = lag_amt)$acf[2:(lag_amt + 1)])

  # Generate and return plot
  plt <- ggplot(data = acf_data, mapping = aes(x = lag, y = acf)) +
    geom_bar(stat = 'identity') +
    labs(x = xlab, y = 'acf', title = title) +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5))

  return(plt)
}

# Make acf plots
biomass_plt <- plot_acf(data_ts, 1, 40, '', 'Biomass')
hydro_plt <- plot_acf(data_ts, 3, 40, '', 'Hydroelectric')
renewable_plt <- plot_acf(data_ts, 2, 40, 'Lag [months]', 'Renewables')

# Use patchwork to assemble
biomass_plt / hydro_plt / renewable_plt
```




## 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?

Given that acf and pacf are the same for lag 1, the lag 1 values in the pacf plots are still quite high. However, the magnitude of subsequent lags is much smaller in all three facets due to the nature in which pacf controls for the effect of shorter lags during calculation. We also note that the pacf plots for `Total Renewable Energy Production` and `Total Biomass Energy Production` look most similar, but the pacf plot for `Hydroelectric Power Consumption` does not look so drastically different.

```
# Create function for pacf plots
plot_pacf <- function(ts, col_num, lag_amt, xlab, title) {
  # Generate pacf data
  pacf_data <- data.frame(lag = 1:lag_amt,
                          pacf = pacf(ts[,col_num],
                                      plot = F,
                                      lag.max = lag_amt)$acf)

  # Generate and return plot
  plt <- ggplot(data = pacf_data, mapping = aes(x = lag, y = pacf)) +
    geom_bar(stat = 'identity') +
    labs(x = xlab, y = 'pacf', title = title) +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5))

  return(plt)
```

```
}

# Make acf plots
biomass_p_plt <- plot_pacf(data_ts, 1, 40, '', 'Biomass')
hydro_p_plt <- plot_pacf(data_ts, 3, 40, '', 'Hydroelectric')
renewable_p_plt <- plot_pacf(data_ts, 2, 40, 'Lag [months]', 'Renewables')

# Use patchwork to assemble
biomass_p_plt / hydro_p_plt / renewable_p_plt
```