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

Assignment 4 - Due date 02/25/21

Stefan Chen

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the project open the first thing you will do is change "Student Name" on line 3 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. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A04_Sp21.Rmd"). Submit this pdf using Sakai.

Questions

Consider the same data you used for A2 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumpti
The data comes from the US Energy Information and Administration and corresponds to the January 2021 Monthly Energy Review.

R packages needed for this assignment:"forecast","tseries", and "Kendall". 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.\

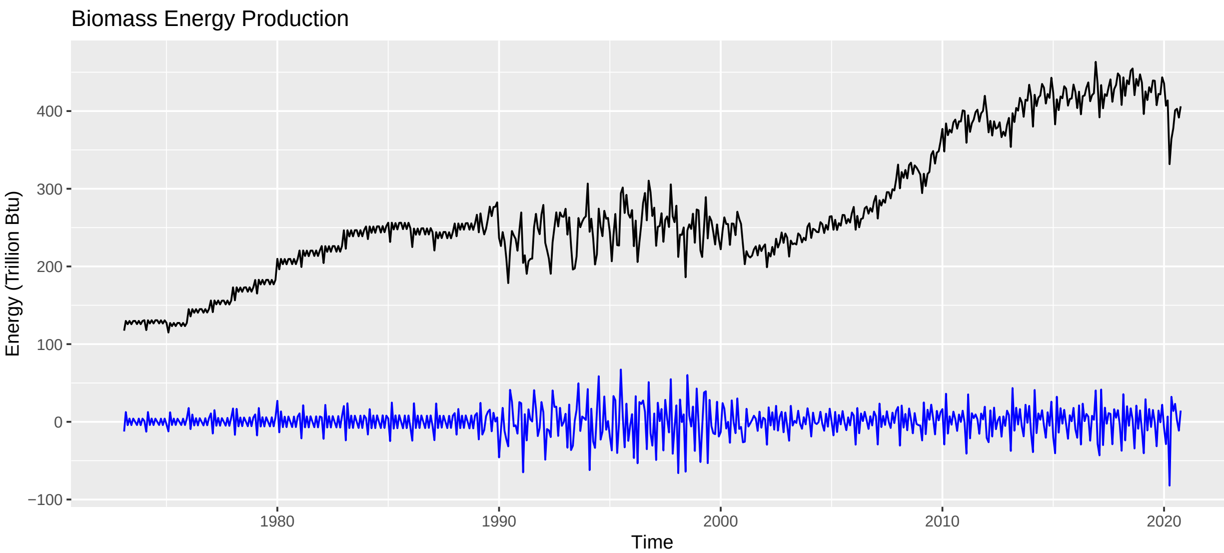
Stochastic Trend and Stationarity Test

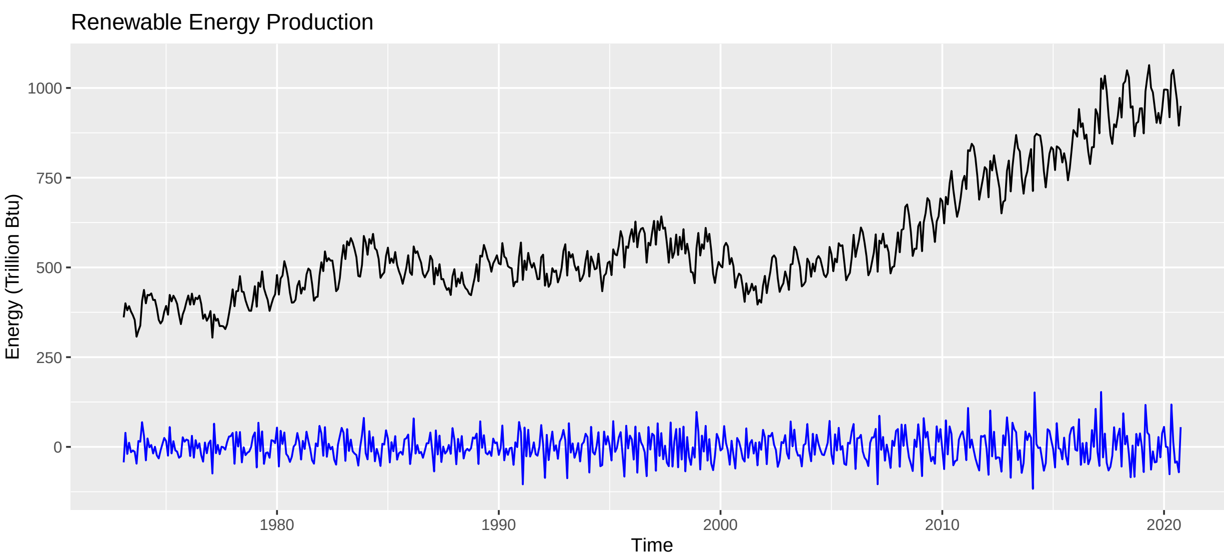
For this part you will once again work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series and the Date column. Don't forget to format the date object.

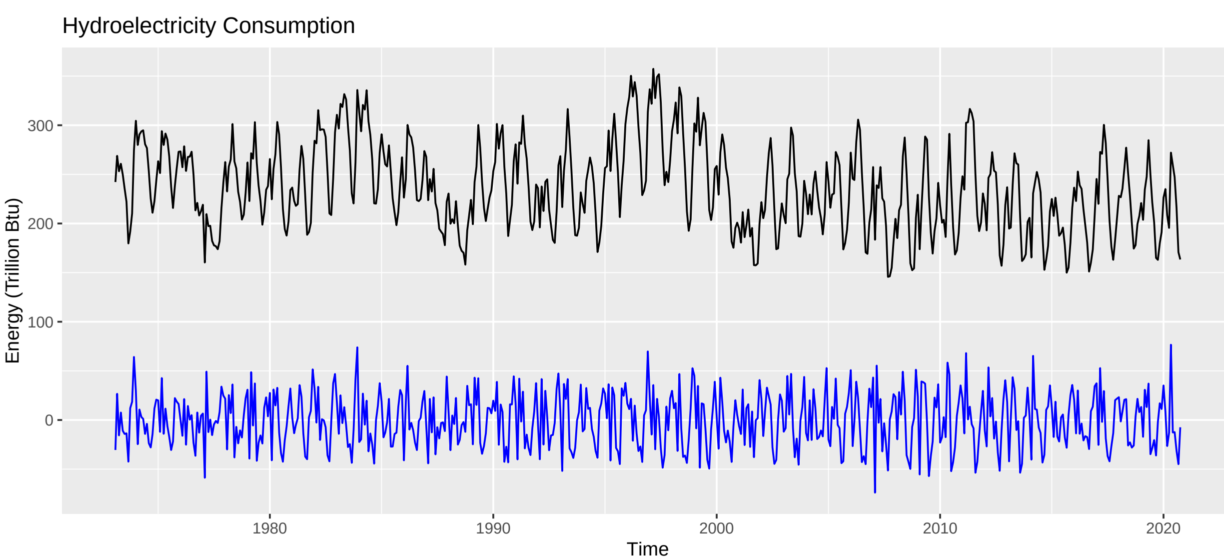
Q1

Now let's try to difference these three series using function diff(). Start with the original data from part (b). Try differencing first at lag 1 and plot the remaining series. Did anything change? Do the series still seem to have trend?

Comapred to the original data, the differenced series seems relatively stable. Based on the graphs, differencing the series appears to remove the trend componenet.


Biomass Energy Production
Energy (Trillion Btu)
400
300
200
100
0
−100
1980
1990
2000
2010
2020
Time


Renewable Energy Production
Energy (Trillion Btu)
1000
750
500
250
0
1980
1990
2000
2010
2020
Time


Hydroelectricity Consumption
Energy (Trillion Btu)
300
200
100
0
1980
1990
2000
2010
2020
Time

Q2

Compute Mann-Kendall and Spearman's Correlation Rank Test for each time series. Ask R to print the results. Interpret the results.

Mann Kendall Test

The Mann Kendall Test resulted in significantly small p-values for biomass energy production, renewable energy production, and hydroelectricity consumption which indicates statistically significant evidence to reject the null hypothesis. Therefore, we can state that the series are not stationary.

```
## [1] "Results for Seasonal Mann Kendall: Biomass Energy Production"

## Score =  9874 , Var(Score) = 150368.7
## denominator =  13442
## tau = 0.735, 2-sided pvalue =< 2.22e-16
## NULL

## [1] "Results for Seasonal Mann Kendall: Renewable Energy Production"

## Score =  9476 , Var(Score) = 150368.7
## denominator =  13442
## tau = 0.705, 2-sided pvalue =< 2.22e-16
## NULL

## [1] "Results for Seasonal Mann Kendall: Hydroelectricity Consumption"

## Score =  -3880 , Var(Score) = 150368.7
## denominator =  13442
## tau = -0.289, 2-sided pvalue =< 2.22e-16
## NULL
```

Spearman's Correlation Rank Test

Spearman's Correlation between each of the three series and the date was conducted. The test resulted in a correlation of 0.865 for biomass energy production, 0.824 for renewable energy production, and -0.285 for hydroelectricity consumption. Results suggest that biomass energy production and renewable energy production have high positive correlation with the date, whereas hydroelectricity consumption has a low negative correlation with the date.

```
## [1] "Results from Spearman Correlation: Biomass, Renewable Energy, and Hydroelectricity"

##
##  Spearman's rank correlation rho
##
## data:  ts_select[, 1] and my_date_num
## S = 4267941, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8645948
```

```
##
##  Spearman's rank correlation rho
##
## data:  ts_select[, 2] and my_date_num
## S = 5552756, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8238326
```

```
##
##  Spearman's rank correlation rho
##
## data:  ts_select[, 3] and my_date_num
## S = 40512802, p-value = 4.078e-12
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2853138
```

Decomposing the series

For this part you will work only with the following columns: Solar Energy Consumption and Wind Energy Consumption.

Q3

Create a data frame structure with these two time series only and the Date column. Drop the rows with *Not Available* and convert the columns to numeric. You can use filtering to eliminate the initial rows or conver to numeric and then use the drop_na() function. If you are familiar with pipes for data wrangling, try using it!

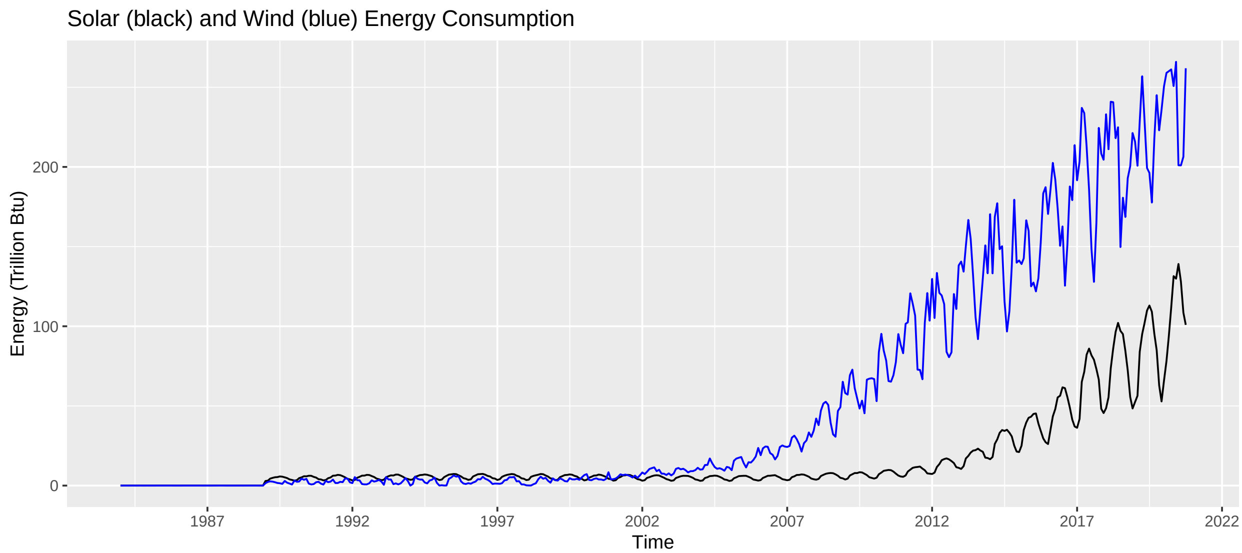
Below are the first five rows of the data frame and the time series.

```
##         Date Solar Energy Consumption Wind Energy Consumption
## 1 1984-01-01                   -0.001                   0.000
## 2 1984-02-01                    0.001                   0.002
## 3 1984-03-01                    0.002                   0.002
## 4 1984-04-01                    0.003                   0.006
## 5 1984-05-01                    0.007                   0.008
## 6 1984-06-01                    0.010                   0.006
```

Q4

Plot the Solar and Wind energy consumption over time using ggplot. Explore the function scale_x_date() on ggplot and see if you can change the x axis to improve your plot.
Hint: use *scale_x_date(date_breaks = "5 years", date_labels = "%Y")")*

Try changing the color of the wind series to blue. Hint: use *color = "blue"*



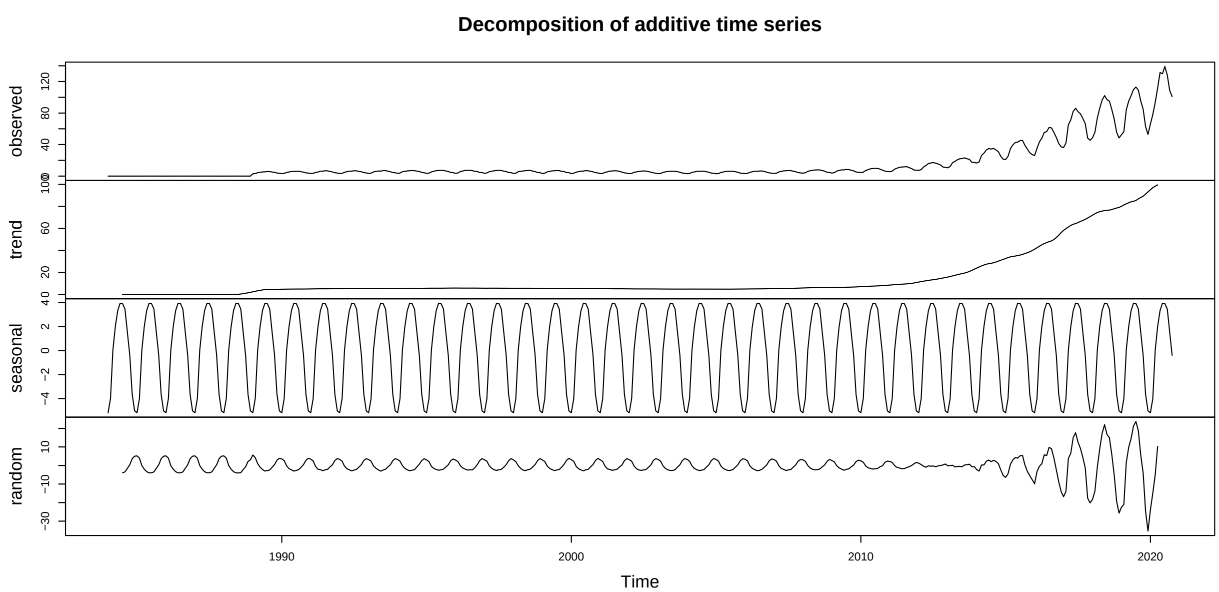
Q5

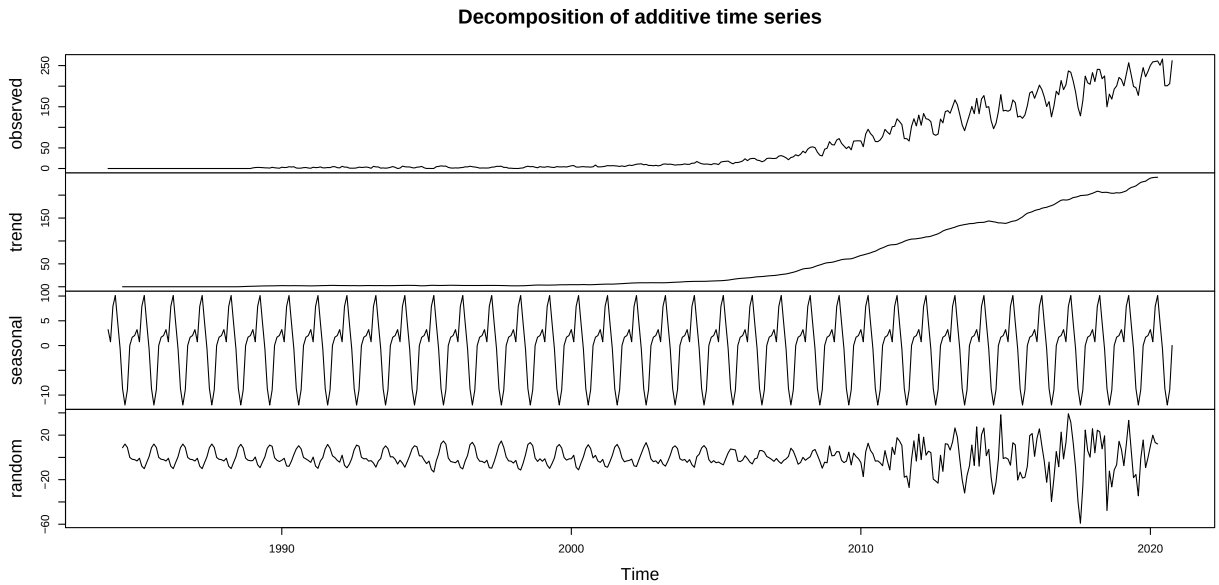
Transform wind and solar series into a time series object and apply the decompose function on them using the additive option. What can you say about the trend component? What about the random component? Does the random component look random? Or does it appear to still have some seasonality on it?

Trend: There appears to be an exponentially growing trend for both wind and solar.

Seasonal: Seasonal trends exist in both the solar and wind time series. However, the seasonal trend in solar series appears to fluctuate less than the seasonal trend in wind series.

Random: The random component seems to still include a seasonal trend that changes in different time periods. Additionally, fluctuation in random trend increased as time approaches present time in both wind and solar time series.

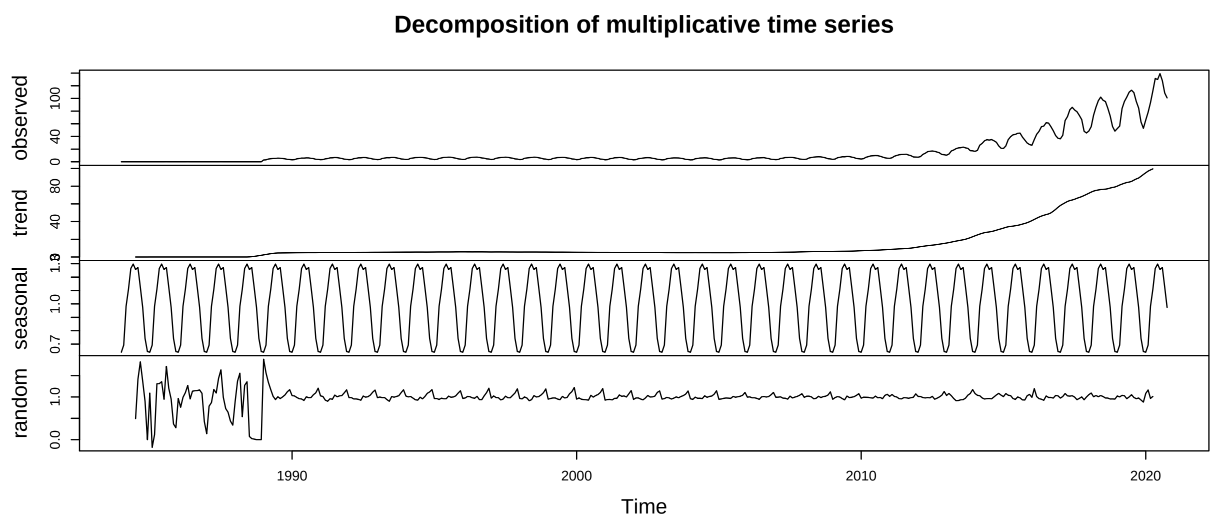


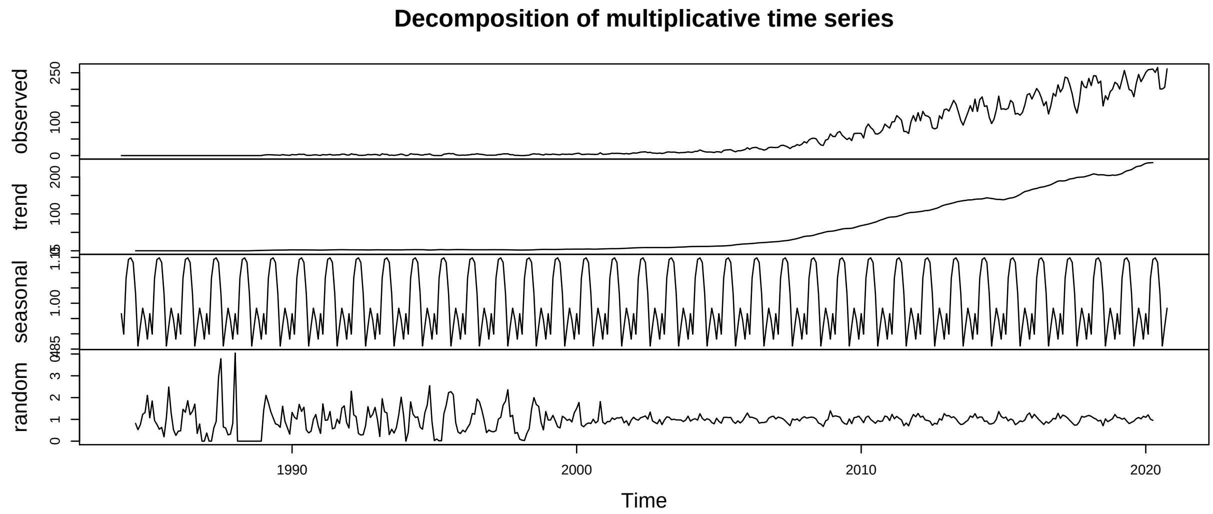


Q6

Use the decompose function again but now change the type of the seasonal component from additive to multiplicative. What happened to the random component this time?

In a multiplicative model, the random component appears to be more random. Additionally, the component for both wind and solar series appear to fluctuate more at the beginning and stabilize as time approaches present time.





Q7

When fitting a model to this data, do you think you need all the historical data? Think about the date from 90s and early 20s. Are there any information from those year we might need to forecast the next six months of Solar and/or Wind consumption. Explain your response.

Since the trend and random components of the two series implies none to extremely low consumption of solar and wind energy before year 2000 compare to now, removing historical data in that time period would not have significant effects on solar and wind energy consumption forecast for the recent months. However, it should be noted that seasonal trend component takes into account monthly consumption which fluctuates based on seasons all years, removing any historical data would have a slight effect on how well the seasonal trend is represented.

Appendix

```
#setup
knitr::opts_chunk$set(echo=FALSE, tidy.opts=list(width.cutoff=80),
                      message=FALSE, warning=FALSE, tidy=FALSE)

#Load/install required package here
library(lubridate)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(outliers)
library(tidyverse)

#Create data frame with the selected column
setwd("/Users/stefanchen/Documents/Duke/Classes/Spring 2021/ENV 790/GitHub/ENV790_30_TSA_S2021/Data")

#Read data
re_select0<-read_xlsx("Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                      skip=12, col_names=FALSE)

#Select columns
re_select<-as.data.frame(re_select0[,c(1,4,5,6)]) #Transform to dataframe
colnames(re_select)=c("Date",
                      "Biomass Energy Production",
                      "Renewable Energy Production",
                      "Hydroelectric Consumption")
re_select$Date<-ymd(re_select$Date)

#Transform dataframe to time series
ts_select<-ts(re_select[,2:4], start=c(1973,1), frequency=12)

my_date<-re_select$Date  #Establish my_date
nobs<-nrow(re_select) #Numbers of rows in ts_select
t<-c(1:nobs) #Set a vector for the number of rows in ts_select

#Data Frame
head(re_select)
#Time Series
head(ts_select)

#Q1
#differencing the series
bio_diff <- diff(ts_select[,"Biomass Energy Production"],lag=1,differences=1)
re_diff <- diff(ts_select[,"Renewable Energy Production"],lag=1,differences=1)
hydro_diff <- diff(ts_select[,"Hydroelectric Consumption"],lag=1,differences=1)

#Check data
head(bio_diff)
head(re_diff)
head(hydro_diff)
```

```r
#Add the new series to our data frame
select_full<-re_select %>%
  cbind(BioDiff=c(NA,as.numeric(bio_diff)),
        ReDiff=c(NA,as.numeric(re_diff)),
        HydroDiff=c(NA,as.numeric(hydro_diff))) %>%
  na.omit(Bio_diff)

head(select_full)

for (i in 2:4){
  print(
  ggplot(select_full, aes(x=Date)) +
      geom_line(aes(x=Date,y=select_full[,i]),color="black") +
      geom_line(aes(x=Date,y=select_full[,i+3]),color="blue") +
      labs(x="Time", y="Energy (Trillion Btu)")
  )}
```

```r
#Q2
#Mann Kendall Test
#Null-it's stationary
#Alternative-it's not stationary (follows a trend)
SMKtest_bio<-SeasonalMannKendall(ts_select[,1])
print("Results for Seasonal Mann Kendall: Biomass")
print(summary(SMKtest_bio))

SMKtest_re<-SeasonalMannKendall(ts_select[,2])
print("Results for Seasonal Mann Kendall: Renewable Energy")
print(summary(SMKtest_re))

SMKtest_hydro<-SeasonalMannKendall(ts_select[,3])
print("Results for Seasonal Mann Kendall: Hydroelectricity")
print(summary(SMKtest_hydro))

#create numeric form my_date
my_date_num<-as.numeric(my_date)

print("Results from Spearman Correlation: Biomass, Renewable Energy, and Hydroelectricity")

#Spearman correlation test
#Deterministic trend with Spearman Correlation Test, with cor.test you can get test statistics
bio_rho<-cor.test(ts_select[,1],my_date_num,method="spearman")
print(bio_rho)

re_rho<-cor.test(ts_select[,2],my_date_num,method="spearman")
print(re_rho)

hydro_rho<-cor.test(ts_select[,3],my_date_num,method="spearman")
print(hydro_rho)

#Q3
#Create a data frame for solar and wind energy consumption
re_select_sw<-as.data.frame(re_select0[,c(1,8,9)]) #Transform to dataframe
colnames(re_select_sw)=c("Date",
```

```r
                      "Solar Energy Consumption",
                      "Wind Energy Consumption")
re_select_sw$Date<-ymd(re_select_sw$Date)

re_select_sw[,2]<-as.numeric(re_select_sw[,2])
re_select_sw[,3]<-as.numeric(re_select_sw[,3])
re_select_sw<-re_select_sw %>% drop_na()

#Data Frame
head(re_select_sw)

#Q4
#plot wind and solar energy consumption
ggplot(re_select_sw, aes(x=Date)) +
    geom_line(aes(x=Date,y=re_select_sw[,2]),color="black") +
    geom_line(aes(x=Date,y=re_select_sw[,3]),color="blue") +
    labs(x="Time",
         y="Energy (Trillion Btu)",
         title="Solar (black) and Wind (blue) Energy Consumption") +
    scale_x_date(date_breaks = "5 years", date_labels = "%Y") #adjusts the time interval
```

```r
#Q5
#Transform dataframe to time series
ts_sw<-ts(re_select_sw[,2:3], start=c(1984,1), frequency=12)

#Decompose-additive
decompose_solar<-decompose(ts_sw[,1],"additive")
decompose_wind<-decompose(ts_sw[,2],"additive")
plot(decompose_solar)
```

```r
plot(decompose_wind)
```

```r
#Q6
#decompose-multiplicative
decompose_solar_mul<-decompose(ts_sw[,1],"multiplicative")
decompose_wind_mul<-decompose(ts_sw[,2],"multiplicative")
plot(decompose_solar_mul)
```

```r
plot(decompose_wind_mul)
```