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

Assignment 4 - Due date 02/25/21

Yash Doshi

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_Consumption_by_ 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.

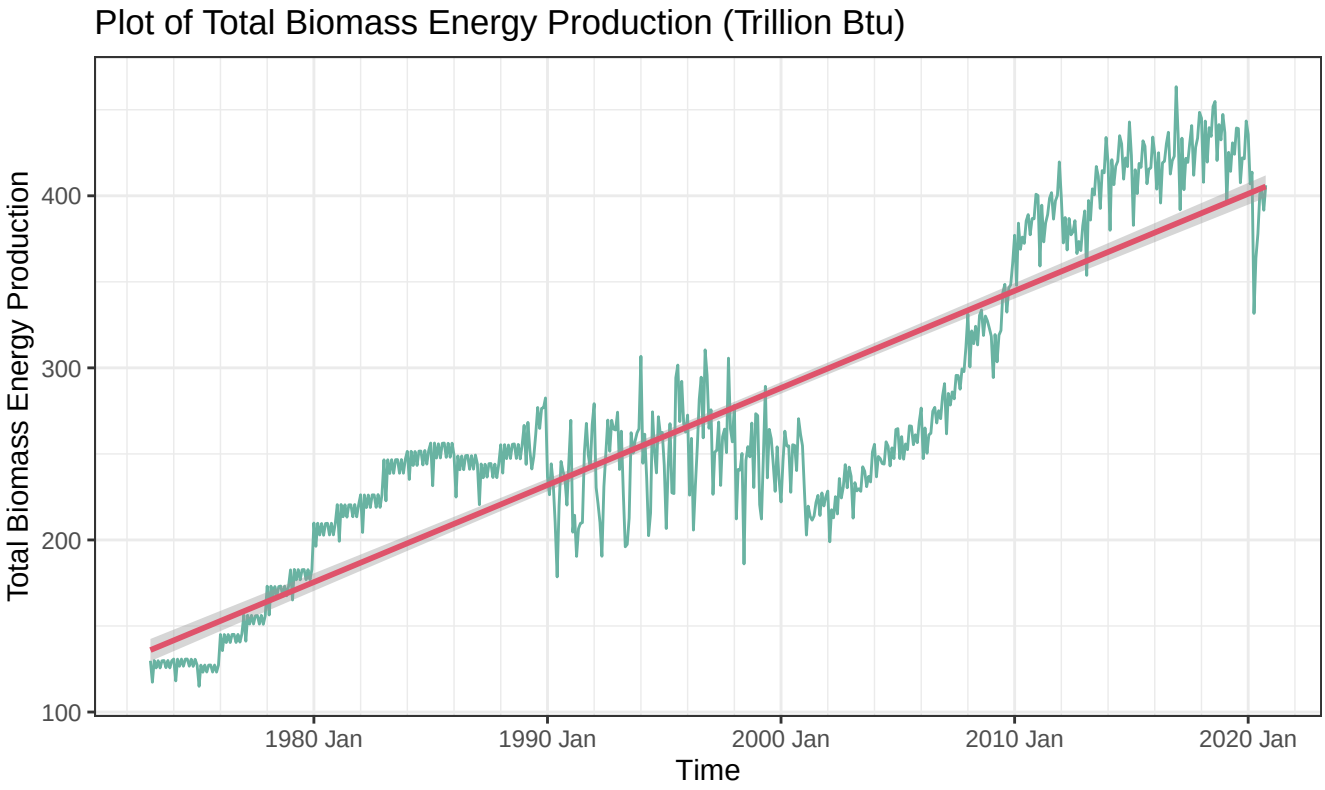
Stochastic Trend and Stationarity Test

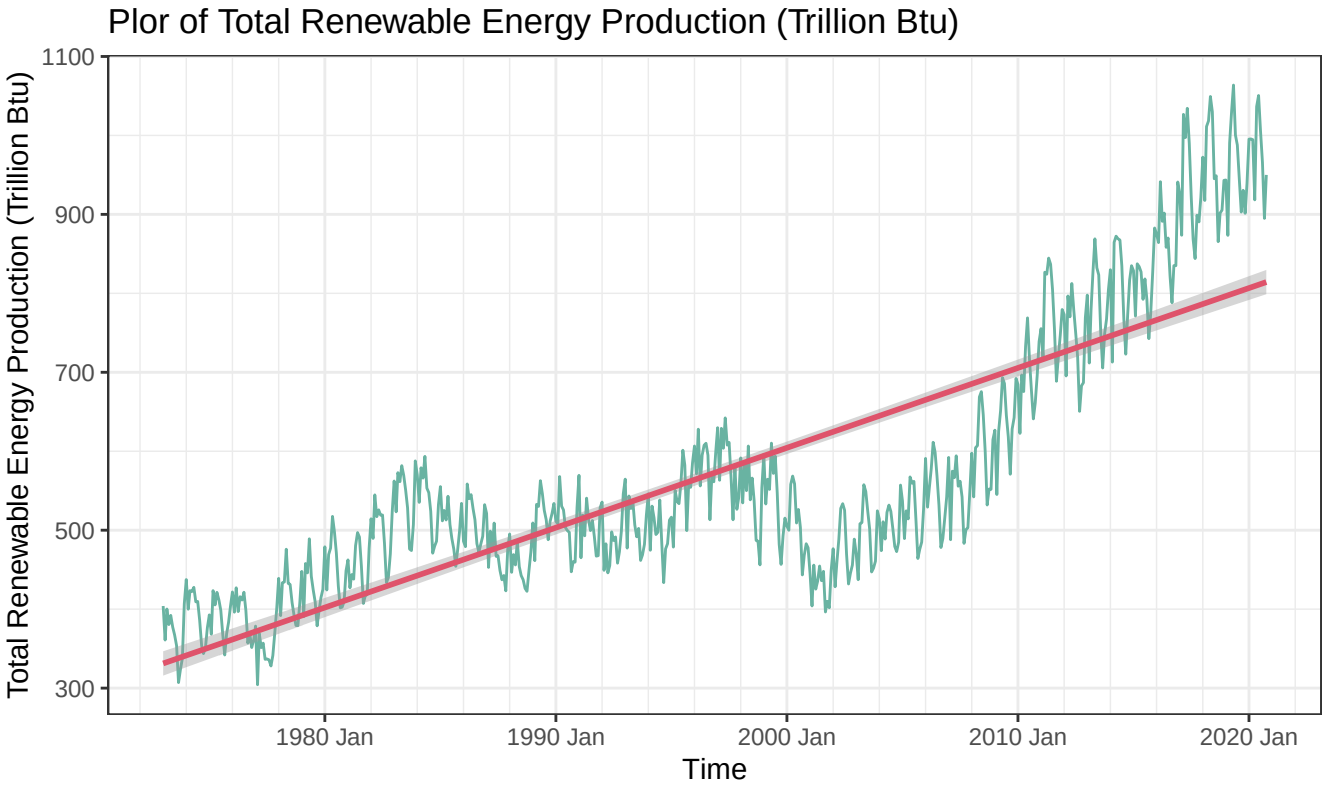
For this part you will once again work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series and the Date column. Don't forget to format the date object.

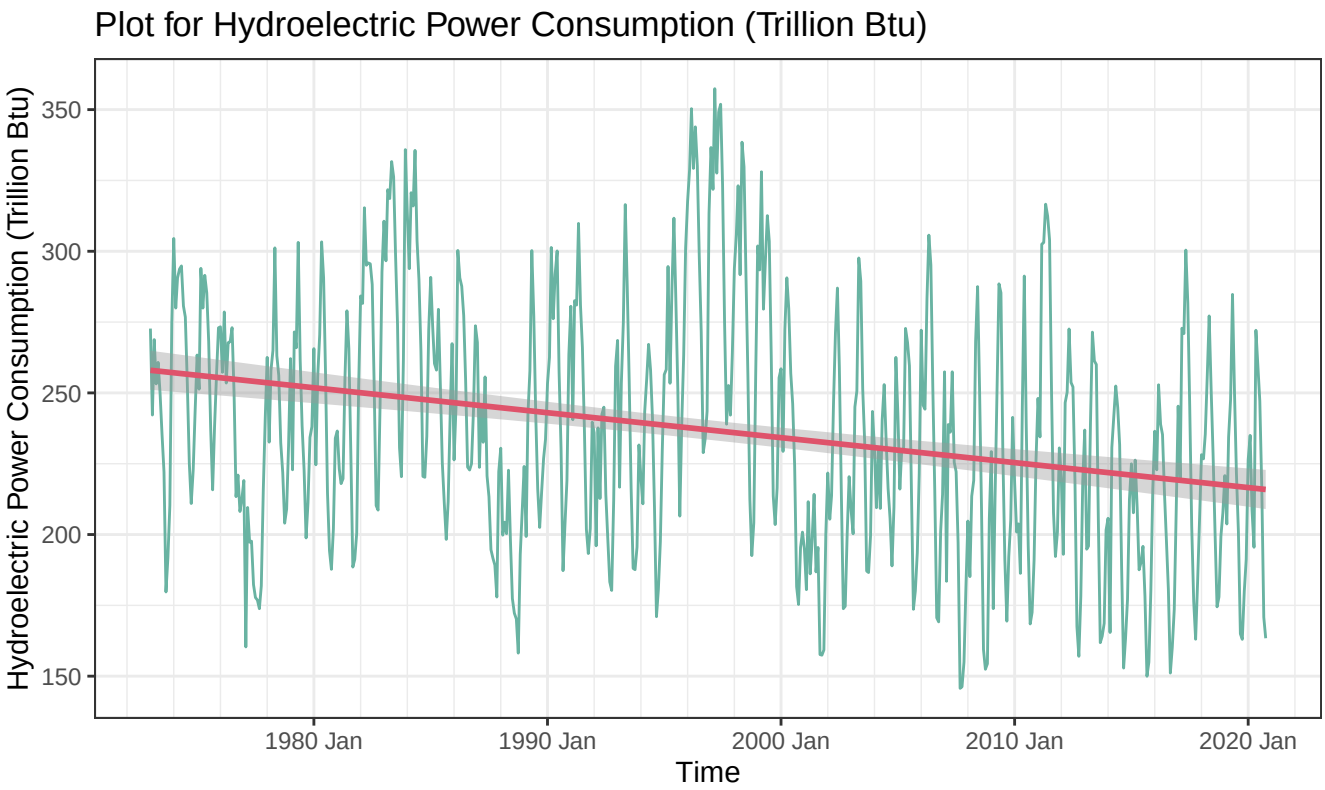
Q1

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?

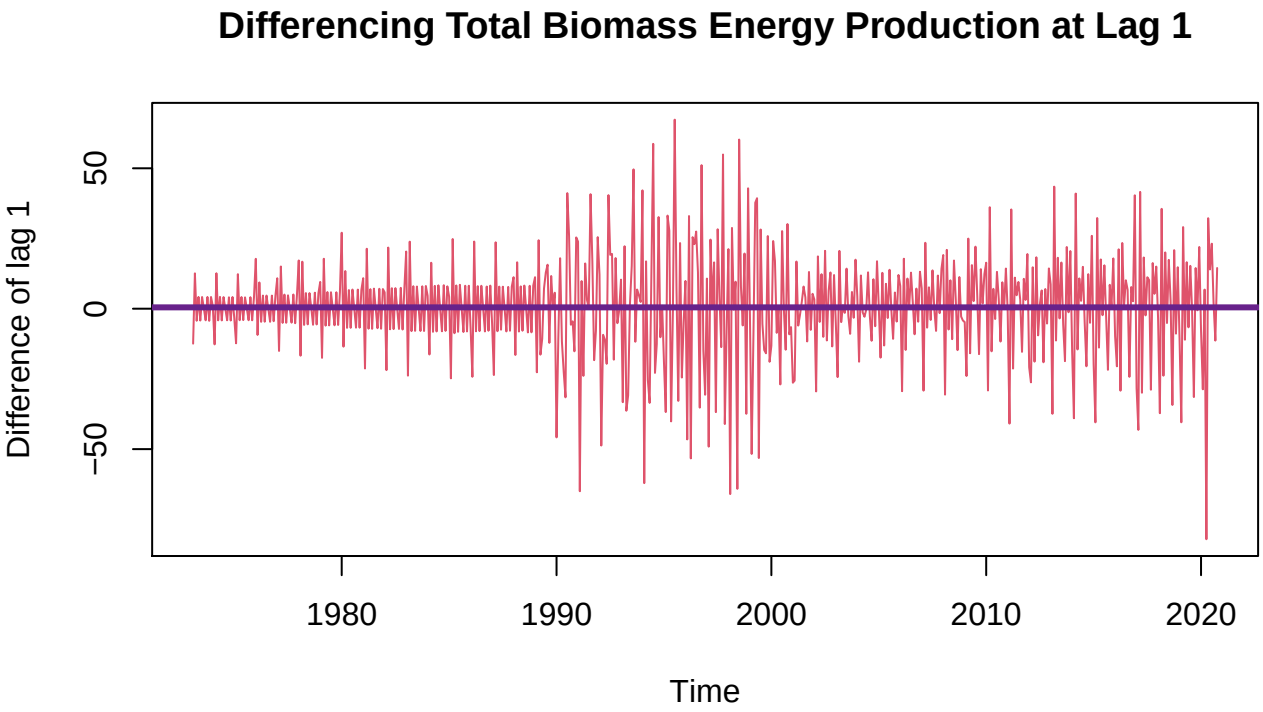
First, I will plot the initial plots along with a trend line to see if there is a trend or not. Below are the plots for initial time series with the trend line.

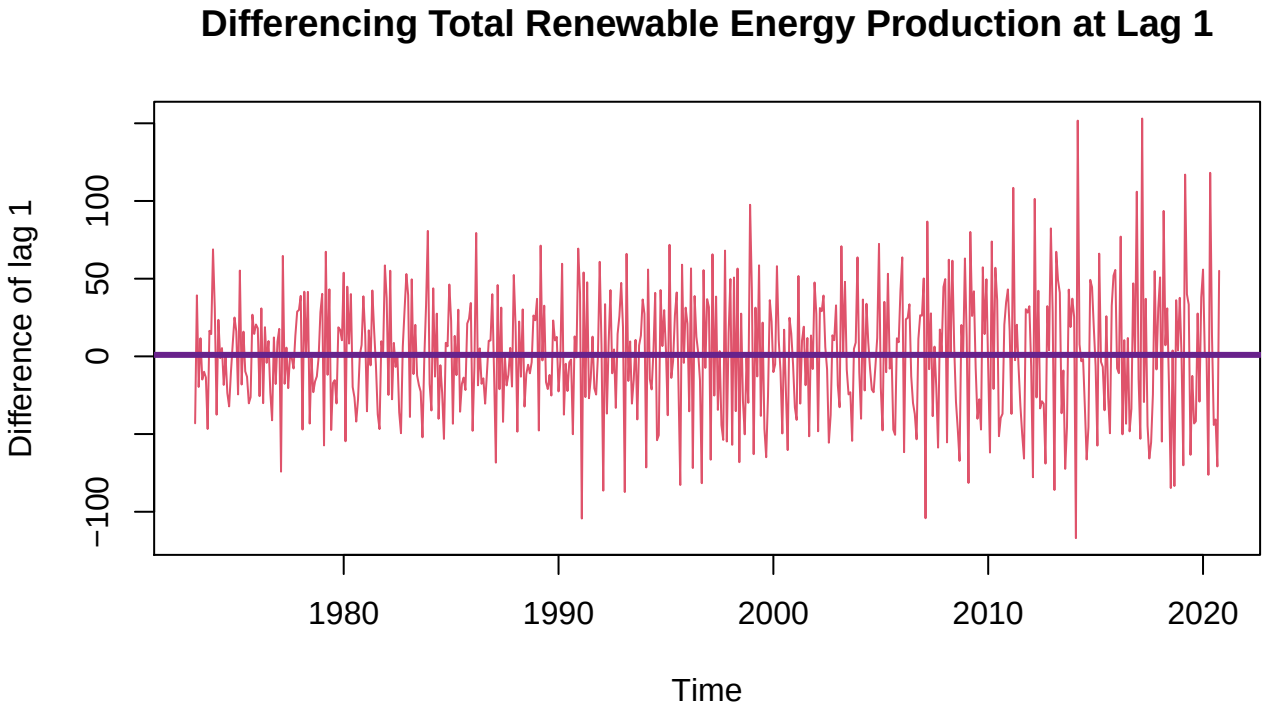

Plot of Total Biomass Energy Production (Trillion Btu)
Total Biomass Energy Production
400
300
200
100
1980 Jan
1990 Jan
2000 Jan
2010 Jan
2020 Jan
Time

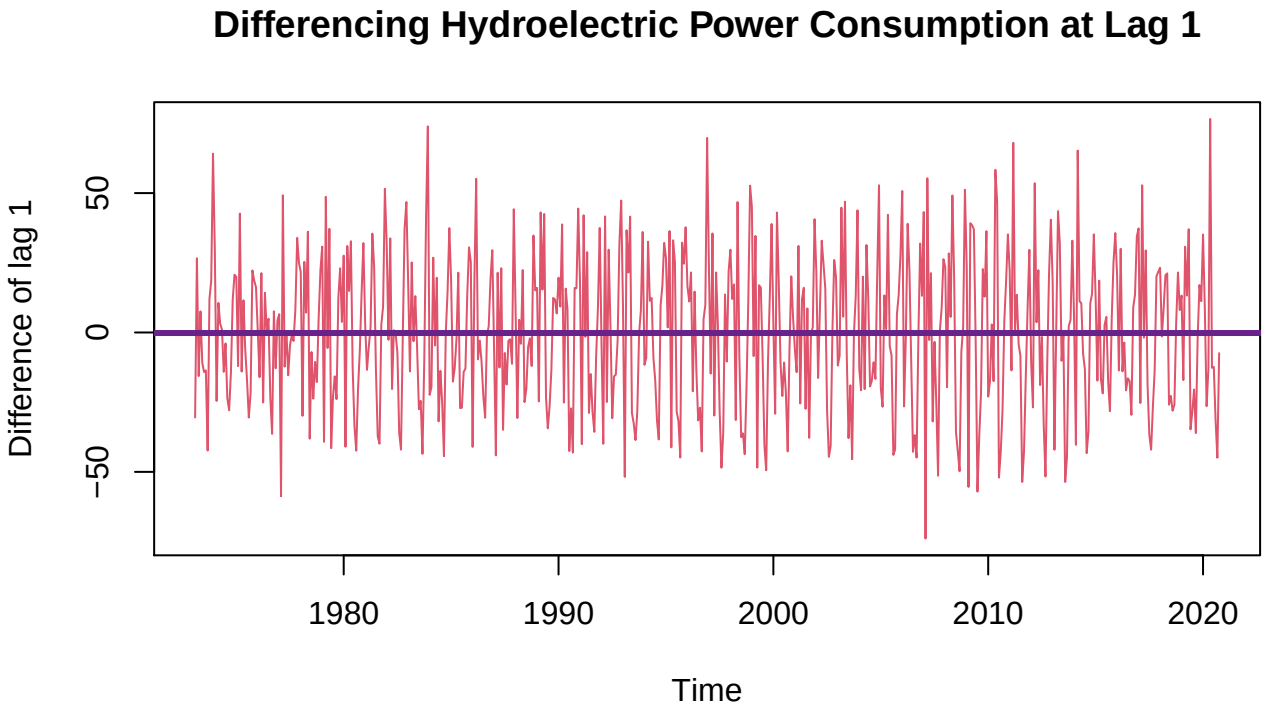

Plor of Total Renewable Energy Production (Trillion Btu)
Total Renewable Energy Production (Trillion Btu)
1100
900
700
500
300
1980 Jan
1990 Jan
2000 Jan
2010 Jan
2020 Jan
Time



It can be seen from the above initial plots that there is clealy a trend. Now, I will use the difference function to the time series objects and plot new graphs. Below are the plots for differenced time series.







It can be seen from the above plots that there is no more trend. All the plots have a mean straight line. Therefore, the trend that we observed in the initial plots is now gone. Hence, we can move forward with the tests.

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

Below are the results for Mann-Kendall test.

For the Mann-Kendall test, I used SeasonalMannKendall() test for all, because all the three variables showed seasonality.

1. Mann-Kendall for Biomass

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

It can be seen from the above result that the score is 9,874. It tells us how many times the observation increased from one time step to another. Hence, higher the score, the better. Furthermore, the null hypothesis of Mann-Kendall states that the data is stationary, where as the alternate hypothesis states that there is a trend. The p-value in the above result is 2.22e-16, which is way less than our threshold of 0.05. Therefore, we reject the null hypothesis in favor of alternate hypothesis. It means that the time series data is not stationary, and hence, there is a trend.

2. Mann-Kendall test for Renewable

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

It can be seen from the above result that the score is 9,476. It tells us how many times the observation increased from one time step to another. Hence, higher the score, the better. Furthermore, the null hypothesis of Mann-Kendall states that the data is stationary, where as the alternate hypothesis states that there is a trend. The p-value in the above result is 2.22e-16, which is way less than our threshold of 0.05. Therefore, we reject the null hypothesis in favor of alternate hypothesis. It means that the time series data is not stationary, and hence, there is a trend.

3. Mann-Kendall test for Hydroelectric

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

The above results show that we have a very low negative score of -3,880. It means that the trend is decreasing over time. This score is still high for the 574 number of observations that we have. Furthermore, the null hypothesis of Mann-Kendall states that the data is stationary, where as the alternate hypothesis states that there is a trend. The p-value in the above result is 2.22e-16, which is way less than our threshold of 0.05. Therefore, we reject the null hypothesis in favor of alternate hypothesis. It means that the time series data is not stationary, and hence, there is a trend.

Spearman's Rank Correlation Coefficient

Below are the test results for Spearman's Rank Correlation Coefficient.

1. Spearman's Rank Correlation Coefficient for Biomass

```
    Spearman's rank correlation rho

data:  eia$`Total Biomass Energy Production (Trillion Btu)` and my_date
S = 4267941, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8645948
```

From the above result it can be seen that the rho is 0.86459, which means that there is a strong positive correlation between Biomass Energy Production and Time. Furthermore, because biomass production and time have a positive relationship, we can say that there is also trend between biomass production and time. As me move forward in time, the biomass energy production will also increase.

2. Spearman's Rank Correlation Coefficient for Renewables

```
    Spearman's rank correlation rho

data:  eia$`Total Renewable Energy Production (Trillion Btu)` and my_date
S = 5552756, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8238326
```

From the above result it can be seen that the rho is 0.82383, which means that there is a strong positive correlation between Renewable Energy Production and Time. Furthermore, because renewable production and time have a positive relationship, we can say that there is also trend between renewable production and time. As me move forward in time, the renewable energy production will also increase.

3. Spearman's Rank Correlation Coefficient for Hydroelectric

```
    Spearman's rank correlation rho

data:  eia$`Hydroelectric Power Consumption (Trillion Btu)` and my_date
S = 40512802, p-value = 3.26e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2853138
```

From the above result it can be seen that the rho is 0.86459, which means that there is a weak negative correlation between Hydroelectric Energy Consumption and Time. Furthermore, because hydroelectric consumption and time have a weak negative relationship, there is also negative trend between hydroelectric consumption and time. As me move forward in time, the hydroelectric energy consumption will decrease.

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 convert to numeric and then use the drop_na() function. If you are familiar with pipes for data wrangling, try using it!

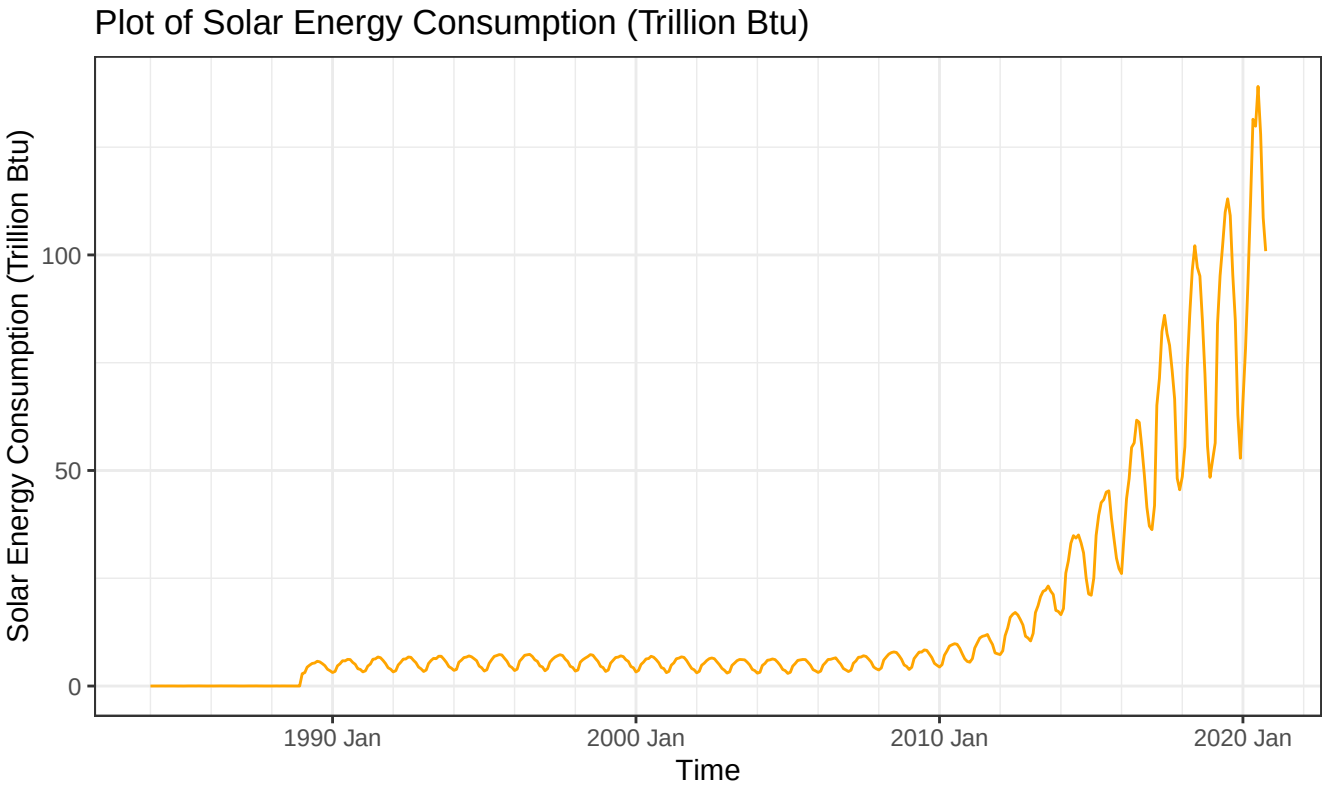
```
# A tibble: 6 x 3
  Month      `Solar Energy Consumption (Trilli~ `Wind Energy Consumption (Trill~
  <date>                                  <dbl>                            <dbl>
1 1984-01-01                             -0.001                            0
2 1984-02-01                              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                              0.01                             0.006
```

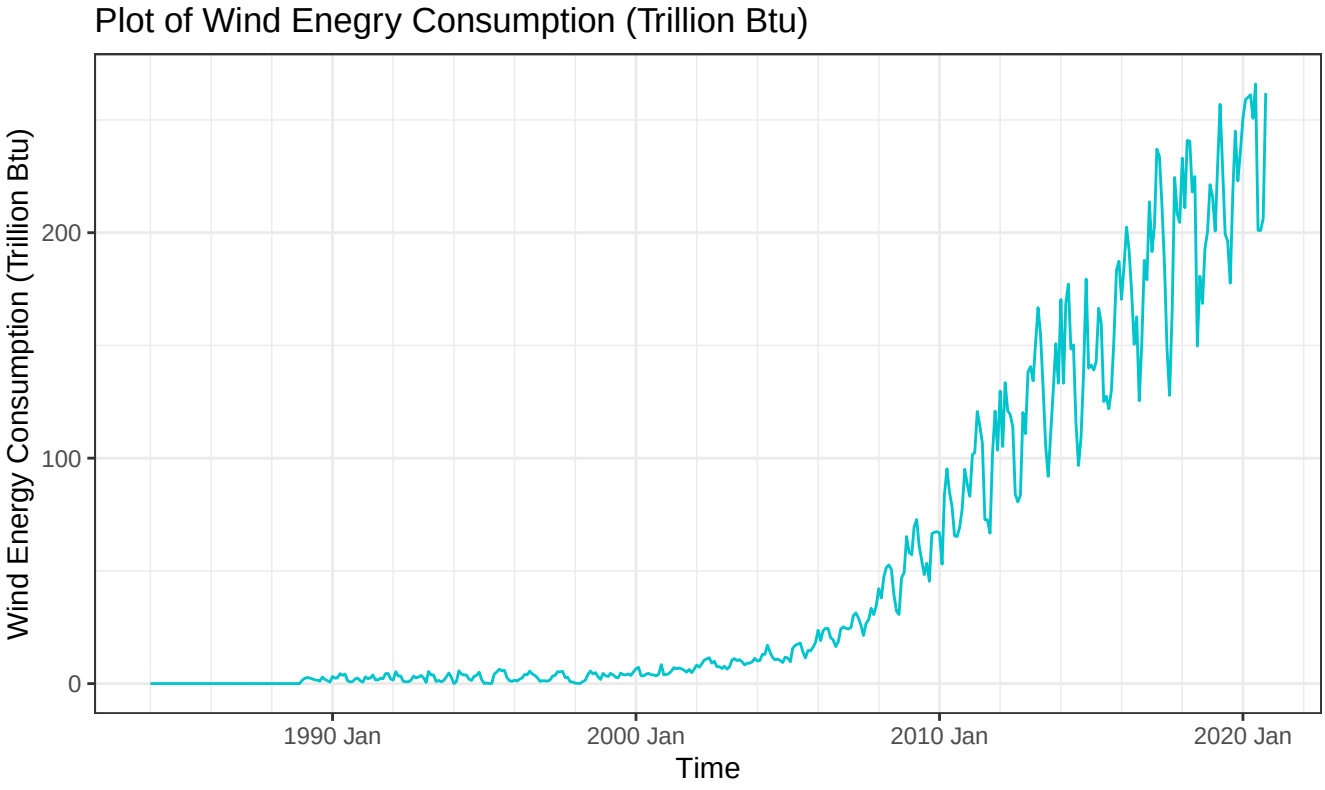
Q4

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"*

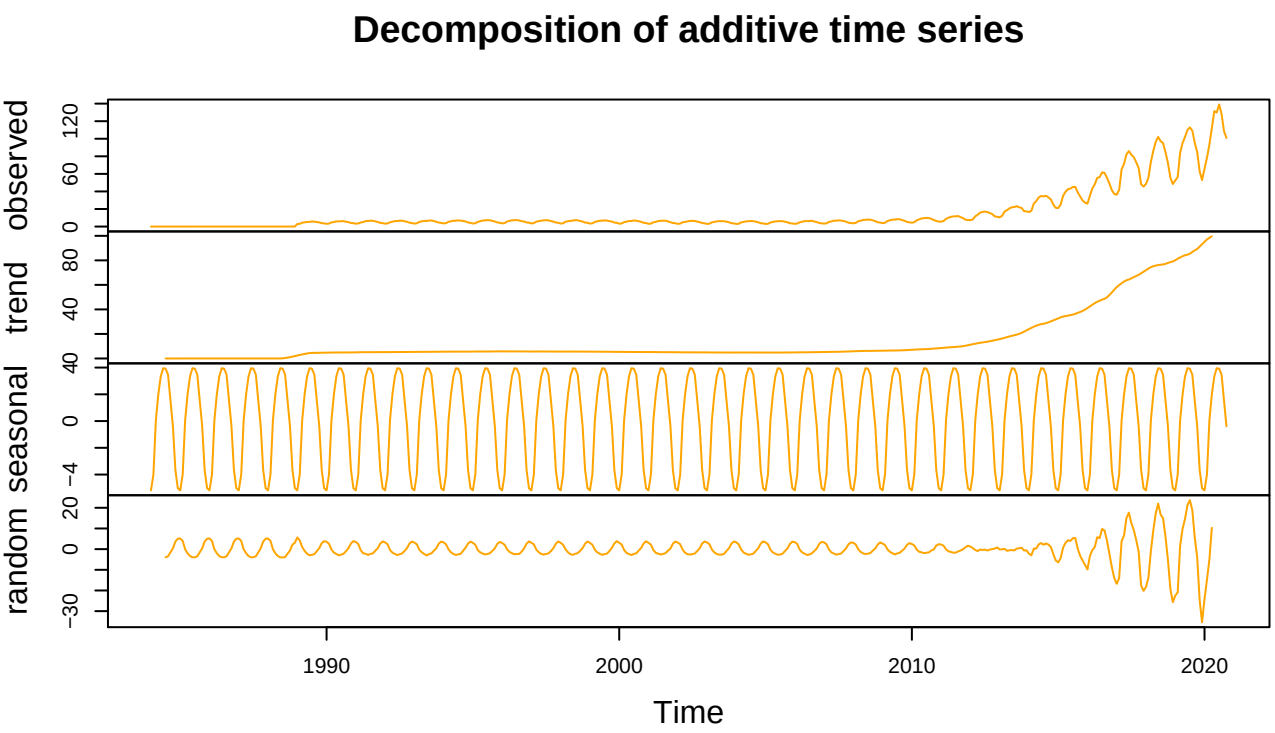




Q5

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?

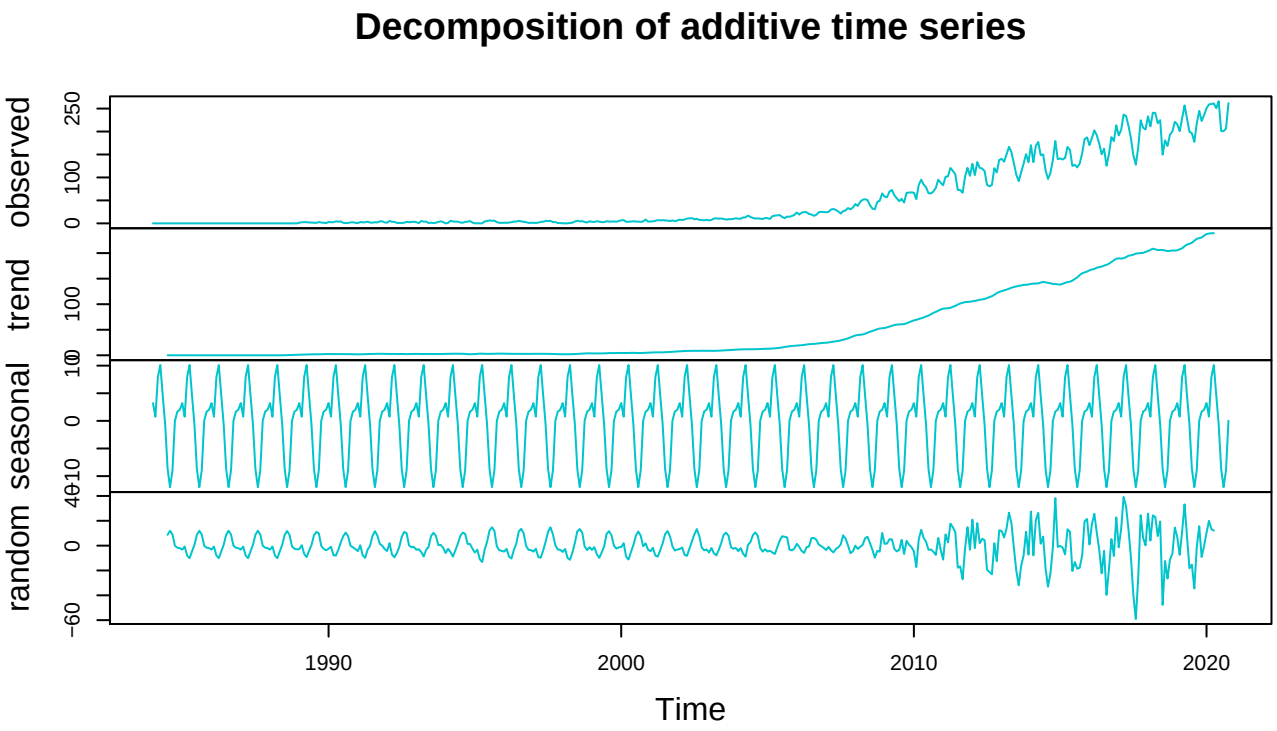
	Solar Energy Consumption (Trillion Btu)
Jan 1984	-0.001
Feb 1984	0.001
Mar 1984	0.002
Apr 1984	0.003
May 1984	0.007
Jun 1984	0.010
	Wind Energy Consumption (Trillion Btu)
Jan 1984	0.000
Feb 1984	0.002
Mar 1984	0.002
Apr 1984	0.006
May 1984	0.008
Jun 1984	0.006



Trend - It can be seen from the above trend plot that there is indeed some trend in the later years. At first, there is no trend. Hence, there is a straight line. But after the 2010, the trend line is increasing, meaning that there has been an increasing trend in the Solar Energy Consumption (which is a good thing). The main reason for the increase in solar consumption is because it was when the countries understood that the renewable energy was the need of the hour (due to increasing GHG and rapid climate change), and that this change was necessary for a sustainable future.

Random - There is no randomness till the year 2010. This is because the solar energy consumption itself was close to 0 (almost negligible). After 2010, however, there is some randomness observed. The amplitude of this randomness is also increasing as the time is moving forward.

There is some sort of seasonality attached to the random component. It is not completely random.



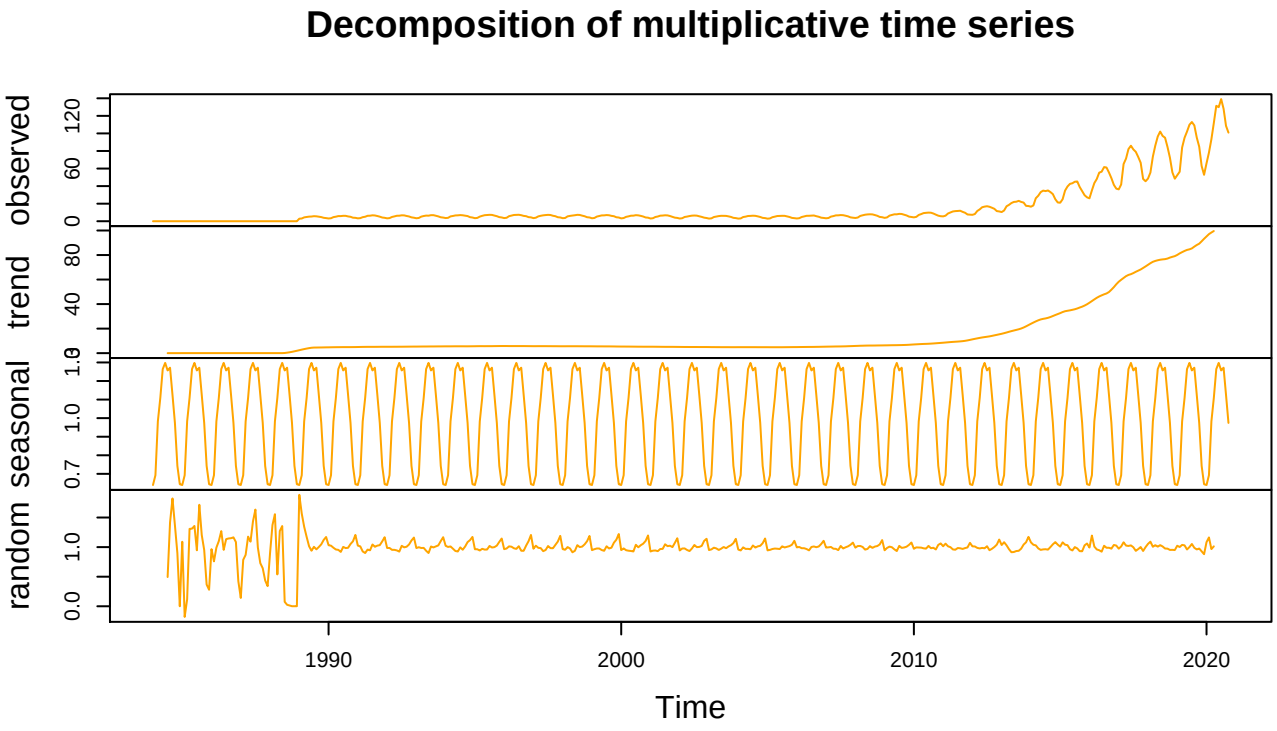
Trend - It can be seen from the above trend plot that there is indeed some trend in the later years. At first, there is no trend. Hence, there is a straight line. But after the 2010, the trend line is increasing, meaning that there has been an increasing trend in the Wind Energy Consumption (which is a good thing).

Random - There is almost no randomness till the year 2010. This is because the wind energy consumption itself was close to 0 (almost negligible). After 2010, however, very high randomness is observed.

In this case, there is no seasonality attached to the random component.

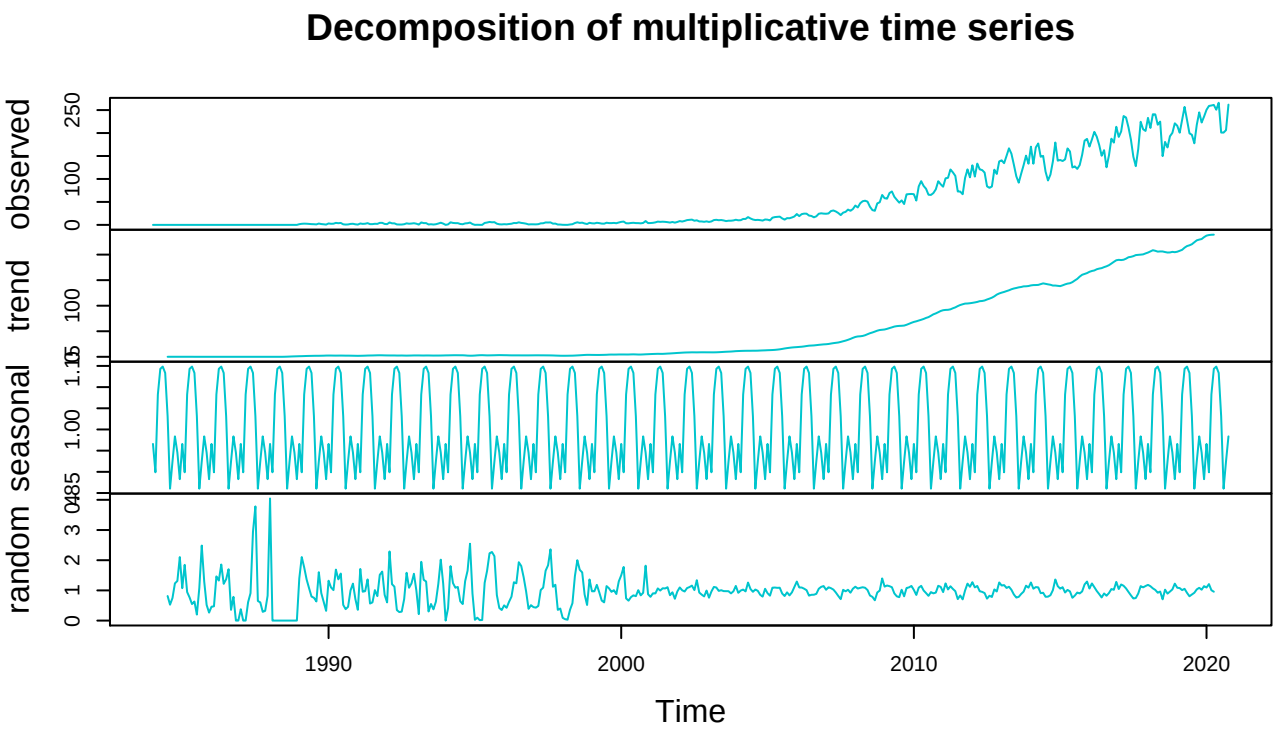
Q6

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?



Multiplicative model is normally not used for energy data. This is because the energy data does not change over time.

When we use multiplicative model on the energy data, it will yield results such as that shown in the above plot. It can be seen from the above model that randomness is very high when the solar energy consumption is 0 and constant. After 1990, when the consumption increases a little, the randomness decreases. Furthermore, after 2010, when the solar consumption jumps, the randomness decreases even further to a bare minimum.



When we use multiplicative model on the energy data, it will yield results such as that shown in the above plot. It can be seen from the above model that randomness is very high when the wind energy consumption is almost 0 and constant. This goes on till early 20s. In the early 20s when the wind consumption jumps, the randomness decreases further to a bare minimum. Therefore, multiplicative model on energy data does not yield accurate results.

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.

No, I do not think that the historical data is needed to forecast the next six months of solar or wind energy consumption. One of the main reason for this is from 1980s to 2010, the solar and wind energy consumption was almost negligible. Wind energy consumption especially was almost close to 0. Furthermore, there is no significant event that took place in terms of solar and wind consumption. Had there been a sudden increase or drop in the energy consumption, it would have been vital to consider that too. However, since that is not the case, we do not need to take into account the 90s and early 20s for forecasting. There is a possibility that taking them into consideration might not yield accurate forecast. The past years will try to diminish the effect (because of almost negligible solar and wind consumption) of sudden jump in consumption from 2010 onward. The increase in renewable energy consumption started rapidly in 2010 and has continued since. The trend is that the dependence on fossil fuels will slowly decrease and the use of renewable energy sources will increase.

APPENDIX

```
knitr::opts_chunk$set(echo = F, eval = T, comment = NA, message = F,warning = F,
                      fig.width = 7, fig.height = 4.2)
#Load/install required package here
library(readxl)
library(forecast)
library(tseries)
library(Kendall)
library(dplyr)
library(readxl)
library(ggplot2)
library(stats)

#Importing the data set
setwd("/Users/yashdoshi/Desktop/Duke/Courses/Spring 2021/Time Series Analysis/Labs/Lab Work/ENV790_30_TSA_S2021

eia = read_excel("Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx")
eia$Month = as.Date(eia$Month, format = "%m/%y")
eia[2:3] = NULL
eia[5:12] = NULL

#Converting to time series
eia_ts = ts(eia[,2:4], start = 1973, frequency = 12)

#Initial Plot with trendline

#Total Biomass Energy Production
p = ggplot(data = eia, aes(x = Month,
                           y = `Total Biomass Energy Production (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Total Biomass Energy Production") +
  theme_bw() +
  ggtitle("Plot of Total Biomass Energy Production (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b") +
  geom_smooth(method = "lm", color = 2)
p

#Total Renewable Energy Production
q = ggplot(data = eia, aes(x = Month,
                           y = `Total Renewable Energy Production (Trillion Btu)`)) + geom_line(color = "#69b3a
  xlab("Time") +
  ylab("Total Renewable Energy Production (Trillion Btu)") +
  theme_bw() +
  ggtitle("Plor of Total Renewable Energy Production (Trillion Btu)") +
  geom_smooth(method = "lm", col = 2) +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b")
q
#Hydroelectric Power Consumption
he = ggplot(data = eia, aes(x = Month, y = `Hydroelectric Power Consumption (Trillion Btu)`)) +
  geom_line(color = "#69b3a2") +
  xlab("Time") +
  ylab("Hydroelectric Power Consumption (Trillion Btu)") +
  theme_bw() +
  geom_smooth(method = "lm", col = "2") +
  ggtitle("Plot for Hydroelectric Power Consumption (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b")
he
```

```r
#Differencing the time series
#Biomass
diff1 = diff(eia_ts[,1], lag = 1)
plot(diff1, col = "2", ylab = "Difference of lag 1",
     main = "Differencing Total Biomass Energy Production at Lag 1")
abline(h = mean(diff1), col = "darkorchid4", lwd = 3)

#Renewable
diff2 = diff(eia_ts[,2], lag = 1)
plot(diff2, col = "2", ylab = "Difference of lag 1",
     main = "Differencing Total Renewable Energy Production at Lag 1")
abline(h = mean(diff2), col = "darkorchid4", lwd = 3)

#Hydroelectric
diff3 = diff(eia_ts[,3], lag = 1)
plot(diff3, col = "2", ylab = "Difference of lag 1",
     main = "Differencing Hydroelectric Power Consumption at Lag 1")
abline(h = mean(diff3), col = "darkorchid4", lwd = 3)

#Mann-Kendall Test
#Biomass
mk1 = SeasonalMannKendall(eia_ts[,1])
summary(mk1)
#Renewables
mk2 = SeasonalMannKendall(eia_ts[,2])
summary(mk2)

#Hydroelectric
mk3 = SeasonalMannKendall(eia_ts[,3])
summary(mk3)

#Spearman's Rank Correlation Coefficient
my_date = as.numeric(eia$Month)

#Biomass
sp1 = cor.test(eia$`Total Biomass Energy Production (Trillion Btu)`, my_date,
               method = "spearman", exact = FALSE)
sp1

#Renewables
sp2 = cor.test(eia$`Total Renewable Energy Production (Trillion Btu)`, my_date,
               method = "spearman", exact = FALSE)
sp2

#Hydroelectric
sp3 = cor.test(eia$`Hydroelectric Power Consumption (Trillion Btu)`, my_date,
               method = "spearman", exact = FALSE)
sp3

#New data with Wind and Solar

setwd("/Users/yashdoshi/Desktop/Duke/Courses/Spring 2021/Time Series Analysis/Labs/Lab Work/ENV790_30_TSA_S2021

eia2 = read_excel("Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx")
eia2$Month = as.Date(eia2$Month, format = "%m/%y")
eia2[2:7] = NULL
eia2[4:8] = NULL
eia2[,2] = as.numeric(eia2$`Solar Energy Consumption (Trillion Btu)`)
eia2[,3] = as.numeric(eia2$`Wind Energy Consumption (Trillion Btu)`)
```

```r
eia2 = na.omit(eia2)

head(eia2)

#Solar Energy Consumption
a = ggplot(data = eia2, aes(x = Month,
                            y = `Solar Energy Consumption (Trillion Btu)`)) +
  geom_line(color = "orange1") +
  xlab("Time") +
  ylab("Solar Energy Consumption (Trillion Btu)") +
  theme_bw() +
  ggtitle("Plot of Solar Energy Consumption (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b")
a

#Wind Energy Consumption
b = ggplot(data = eia2, aes(x = Month,
                            y = `Wind Energy Consumption (Trillion Btu)`)) +
  geom_line(color = "turquoise3") +
  xlab("Time") + ylab("Wind Energy Consumption (Trillion Btu)") +
  theme_bw() +
  ggtitle("Plot of Wind Enegry Consumption (Trillion Btu)") +
  scale_x_date(date_minor_breaks = "2 years", date_labels = "%Y %b")
b

#Converting to time series
eia2_ts = ts(eia2[,2:3], start = 1984, frequency = 12)
head(eia2_ts)

#Decompose on solar (Additive)
decompose_solar = decompose(eia2_ts[,1], type = "additive")
plot(decompose_solar, col = "orange1")

#Decompose on wind (Additive)
decompose_wind = decompose(eia2_ts[,2], type = "additive")
plot(decompose_wind, col = "turquoise3")

#Decompose on Solar (Multiplicative)
decomp_solar = decompose(eia2_ts[,1], type = "multiplicative")
plot(decomp_solar, col = "orange1")

#Decompose on Wind (Multiplicative)
decomp_wind = decompose(eia2_ts[,2], type = "multiplicative")
plot(decomp_wind, col = "turquoise3")
```