ENV 790.30 - Time Series Analysis for Energy Data | Spring 2022

Assignment 4 - Due date 02/17/22

Grace Choi

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.

R packages needed for this assignment: "xlsx" or "readxl", "ggplot2", "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.\

```
#Load/install required package here
library(xlsx)
```

```
## Warning: package 'xlsx' was built under R version 4.0.5
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.5
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(Kendall)
library(tseries)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.5

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.0.5
```

```r
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.0.5

##
## Attaching package: 'cowplot'

## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(dplyr)
```

Questions

Consider the same data you used for A3 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. For this assignment you will work only with the column "Total Renewable Energy Production".

```
#Importing data set - using xlsx package
Ener <- read.csv(file="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.cs
raw_data <- read.csv(file="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edi
raw_data <-raw_data[c(1,4:6)]
colnames(raw_data)=c("Date","Biomass", "Original", "Hydroelectric")
raw_data$Date <- ym(raw_data$Date)

Ener <- Ener[c(1:5)]
colnames(Ener)=c("Date","a","b","c","Renewable")

Ener_processed <-
  Ener %>%
  mutate(Date = ym(Date)) %>%
  arrange(Date)

head(Ener,15)
```

```
##              Date       a             b       c Renewable
## 1    1973 January 129.630 Not Available 129.787   403.981
## 2   1973 February 117.194 Not Available 117.338   360.900
## 3      1973 March 129.763 Not Available 129.938   400.161
## 4      1973 April 125.462 Not Available 125.636   380.470
## 5        1973 May 129.624 Not Available 129.834   392.141
## 6       1973 June 125.435 Not Available 125.611   377.232
## 7       1973 July 129.616 Not Available 129.787   367.325
## 8     1973 August 129.734 Not Available 129.918   353.757
## 9  1973 September 125.603 Not Available 125.782   307.006
## 10   1973 October 129.769 Not Available 129.970   323.453
## 11  1973 November 125.492 Not Available 125.643   337.817
## 12  1973 December 129.690 Not Available 129.824   406.694
## 13   1974 January 130.655 Not Available 130.807   437.467
## 14  1974 February 117.949 Not Available 118.091   399.942
## 15     1974 March 130.579 Not Available 130.727   423.474
```

```
tail(Ener,15)
```

```
##                Date       a       b       c Renewable
## 571       2020 July 179.419 187.423 402.748   993.568
## 572     2020 August 183.247 185.444 405.123   953.474
## 573 2020 September 176.758 182.066 393.075   883.110
## 574    2020 October 181.003 188.717 406.043   937.063
## 575   2020 November 180.663 192.958 409.352   979.210
```

```
## 576  2020 December 191.478 195.958 425.380   982.997
## 577   2021 January 189.420 185.553 413.246  1002.052
## 578  2021 February 169.804 147.769 351.551   879.302
## 579     2021 March 186.832 189.077 413.877  1092.268
## 580     2021 April 176.045 181.624 393.384  1036.825
## 581       2021 May 187.412  201.68 425.888  1096.106
## 582      2021 June 184.377 195.807 414.266  1031.691
## 583      2021 July 190.908  202.84 429.335   986.802
## 584    2021 August 189.282 189.777 414.181  1003.261
## 585 2021 September 182.256 179.835 396.794   964.228
```

Stochastic Trend and Stationarity Tests

Q1

Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should be differenced.

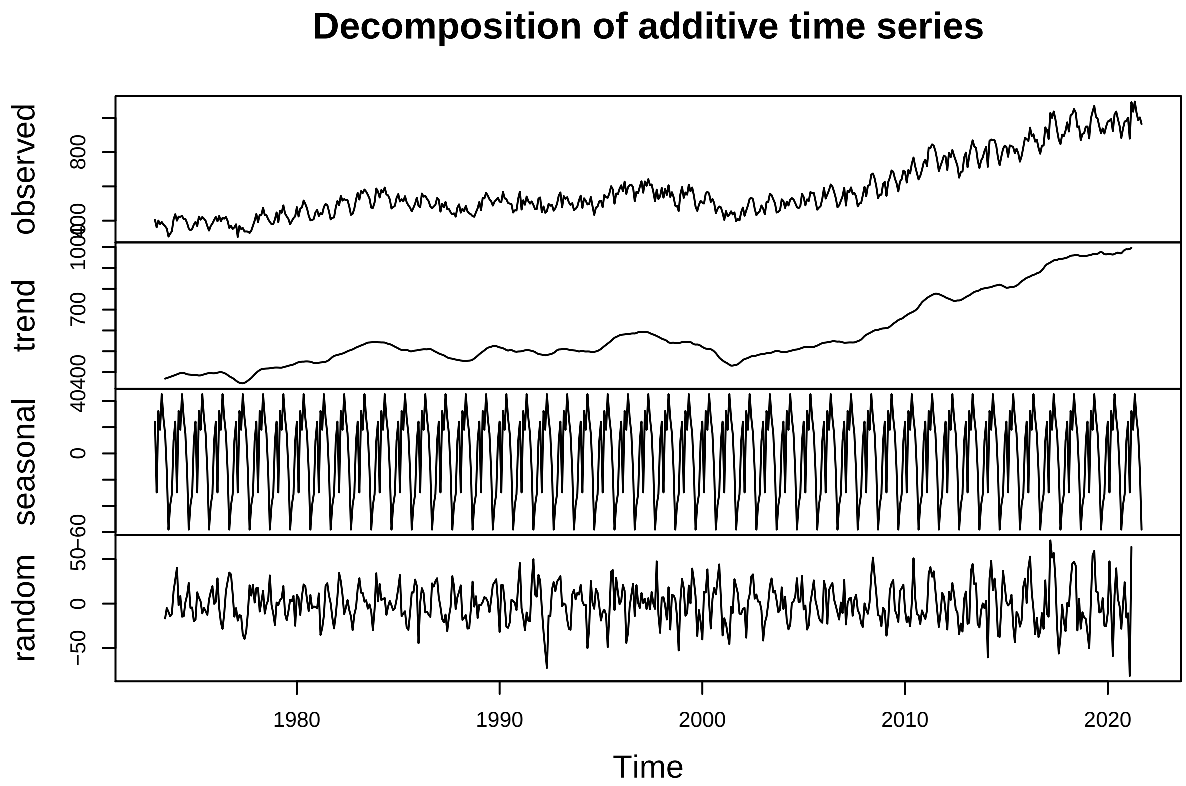
Try differencing at lag 1 only once, i.e., make `lag=1` and `differences=1`. Plot the differenced series Do the series still seem to have trend?

- The trend alters from an increasing to a decreasing trend. Therefore, a changing direction trend seems to exist.

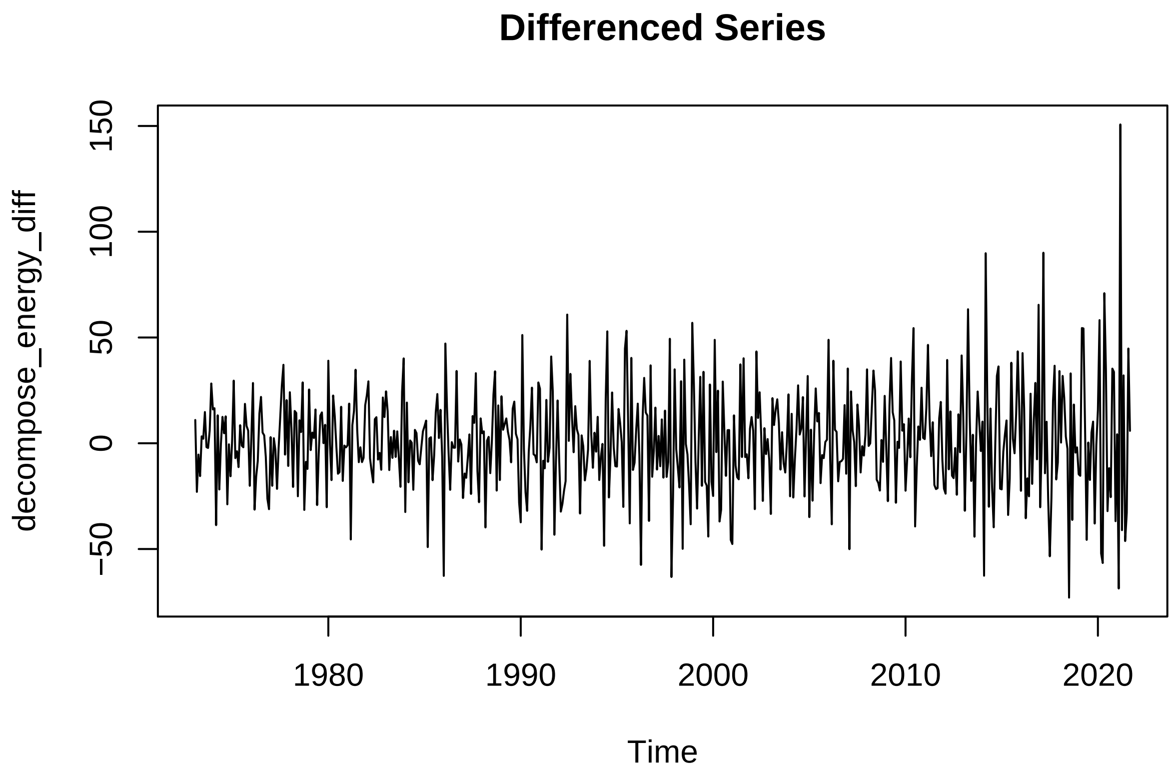
```
nvar <- ncol(Ener_processed) - 1
nobs <- nrow(Ener_processed)

ts_energy <- ts(Ener_processed[,2:(nvar+1)],
               start=c(year(Ener_processed$Date[1]),month(Ener_processed$Date[1])),
               frequency=12)

head(ts_energy,15)
```

```
##                a   b       c Renewable
## Jan 1973 129.630 382 129.787   403.981
## Feb 1973 117.194 382 117.338   360.900
## Mar 1973 129.763 382 129.938   400.161
## Apr 1973 125.462 382 125.636   380.470
## May 1973 129.624 382 129.834   392.141
## Jun 1973 125.435 382 125.611   377.232
## Jul 1973 129.616 382 129.787   367.325
## Aug 1973 129.734 382 129.918   353.757
## Sep 1973 125.603 382 125.782   307.006
## Oct 1973 129.769 382 129.970   323.453
## Nov 1973 125.492 382 125.643   337.817
## Dec 1973 129.690 382 129.824   406.694
## Jan 1974 130.655 382 130.807   437.467
## Feb 1974 117.949 382 118.091   399.942
## Mar 1974 130.579 382 130.727   423.474
```

```
decompose_energy <- decompose(ts_energy[,"Renewable"],"additive")
plot(decompose_energy)
```




```
deseasonal_energy <- seasadj(decompose_energy)
decompose_energy_diff <- diff(deseasonal_energy,differences=1)
plot(decompose_energy_diff, main="Differenced Series")
```



Q2

Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in A3 using linear regression. (Hint: Just copy and paste part of your code for A3)

Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

```
nobs_raw <- nrow(raw_data)
nenergy_raw<- ncol(raw_data)-1
ienergy = 1

t = c(1:nobs_raw)
beta = matrix(ncol=2, nrow=3)

colnames(beta)=c("beta0", "beta1")
for(i in 1:nenergy_raw){
linear_trend_model=lm(raw_data[,3]~t)
print(summary(linear_trend_model))
beta[i,1] = as.numeric(linear_trend_model$coefficients[1])
beta[i,2] = as.numeric(linear_trend_model$coefficients[2])
}
```

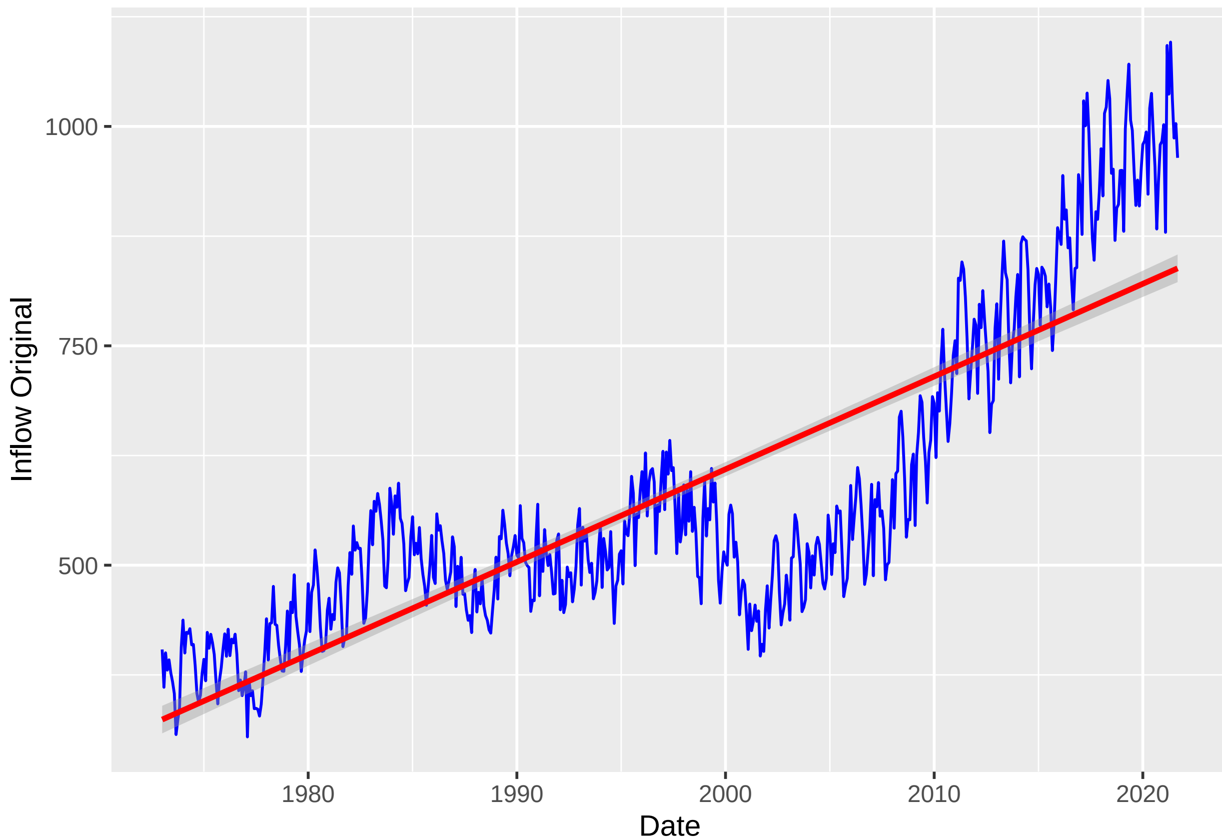
```
##
```

```
## Call:
## lm(formula = raw_data[, 3] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -230.488  -57.869    5.595   62.090  261.349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 323.18243    8.02555   40.27   <2e-16 ***
## t             0.88051    0.02373   37.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.93 on 583 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.702
## F-statistic:  1377 on 1 and 583 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = raw_data[, 3] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -230.488  -57.869    5.595   62.090  261.349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 323.18243    8.02555   40.27   <2e-16 ***
## t             0.88051    0.02373   37.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.93 on 583 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.702
## F-statistic:  1377 on 1 and 583 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = raw_data[, 3] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -230.488  -57.869    5.595   62.090  261.349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 323.18243    8.02555   40.27   <2e-16 ***
## t             0.88051    0.02373   37.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.93 on 583 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.702
```

```
## F-statistic:  1377 on 1 and 583 DF,  p-value: < 2.2e-16
```

```
ggplot(raw_data, aes(x=Date, y=raw_data[,(3)])) +
            geom_line(color="blue") +
            ylab(paste0("Inflow ",colnames(raw_data)[(3)],sep="")) +
            #geom_abline(intercept = beta0, slope = beta1, color="red")
            geom_smooth(color="red",method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

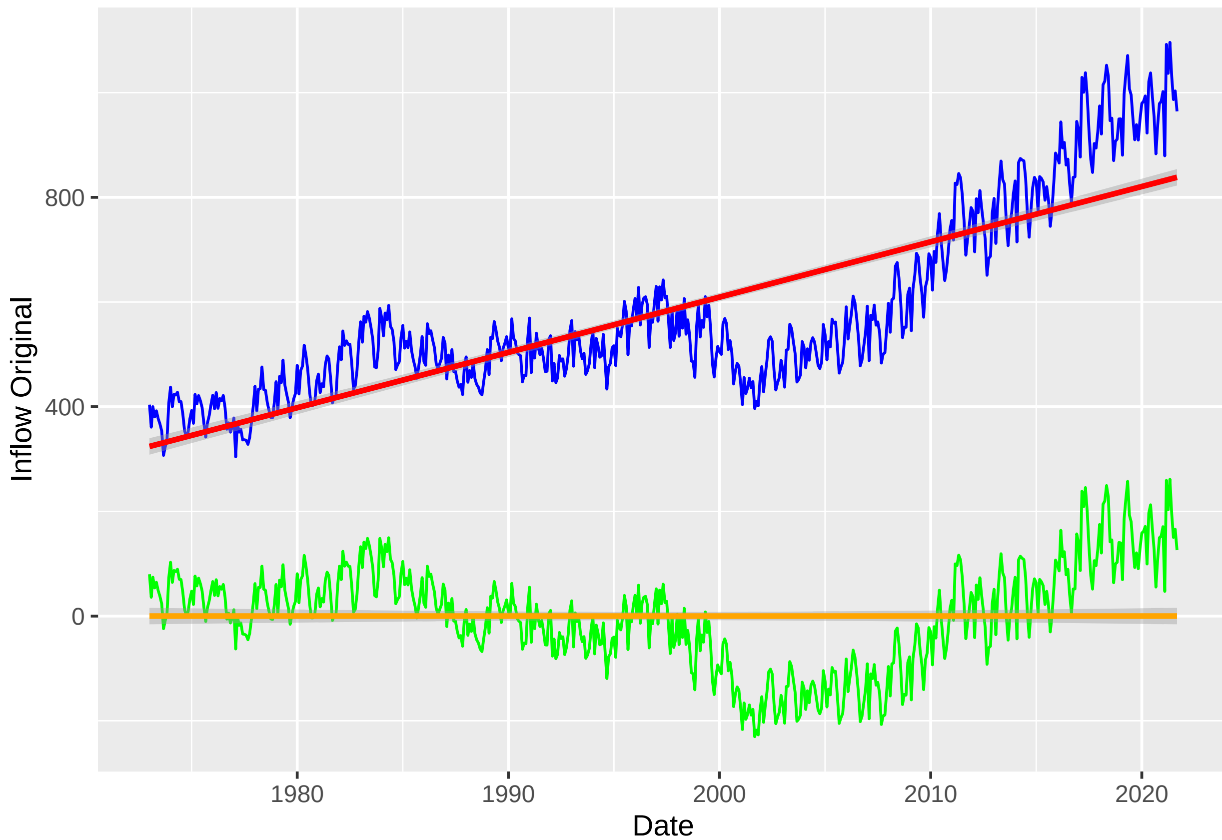



```
print("Regression detrended Total Energy Production")
```

```
## [1] "Regression detrended Total Energy Production"
```

```
detrend_raw_data2 <- raw_data[,ienergy+2]-(beta[2,1]+beta[2,2]*t)
prep_data <- data.frame("Detrend"=detrend_raw_data2)


ggplot(raw_data, aes(x=Date, y=raw_data[,ienergy+2])) +
geom_line(color="blue") +
ylab(paste0("Inflow ",colnames(raw_data)[ienergy+2],sep="")) +
geom_smooth(color="red",method="lm") +
geom_line(aes(y=detrend_raw_data2), col="green")+
geom_smooth(aes(y=detrend_raw_data2),color="orange",method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Q3

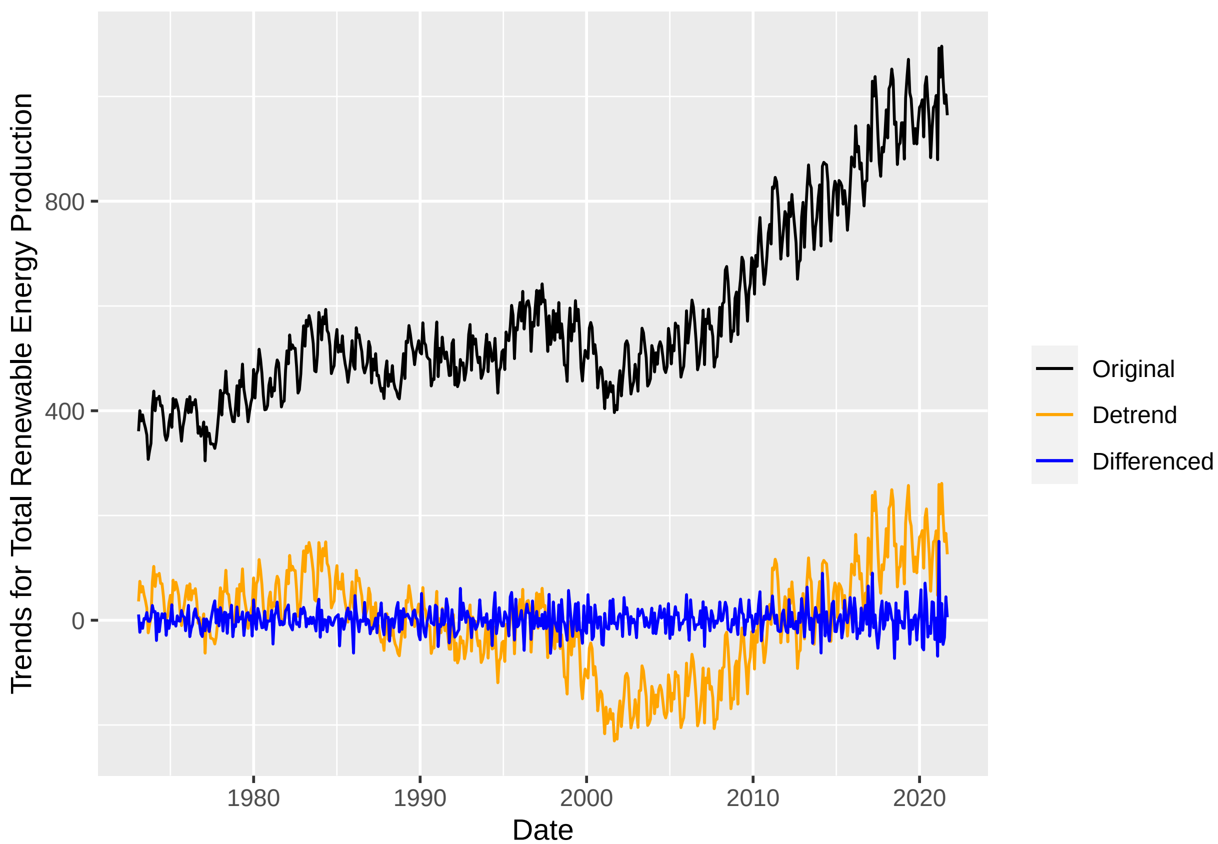
Create a data frame with 4 columns: month, original series, detrended by Regression Series and differenced series. Make sure you properly name all columns. Also note that the differenced series will have only 584 rows because you loose the first observation when differencing. Therefore, you need to remove the first observations for the original series and the detrended by regression series to build the new data frame.

```
#Data frame - remember to note include January 1973

df <- raw_data[c(1,3)]

df4<-
  df %>%
  cbind(Detrended=prep_data) %>%
  cbind(Differenced = c(NA,as.numeric(decompose_energy_diff))) %>%
  na.omit(residentialDiff)
```

Q4

Using ggplot() create a line plot that shows the three series together. Make sure you add a legend to the plot.

```r
#Use ggplot

ggplot(df4, aes(x=Date))+
  geom_line(aes(y=Original,color="Original"))+
  geom_line(aes(y=Detrend,color="Detrend"))+
  geom_line(aes(y=Differenced,color="Differenced"))+
  labs(color="")+
  scale_color_manual(values=c("Original" = "black",
                              "Detrend" =" orange",
                              "Differenced"="blue"),
                     labels=c("Original", "Detrend", "Differenced")) +
  theme(legend.position = "right")+
  ylab(label="Trends for Total Renewable Energy Production")
```

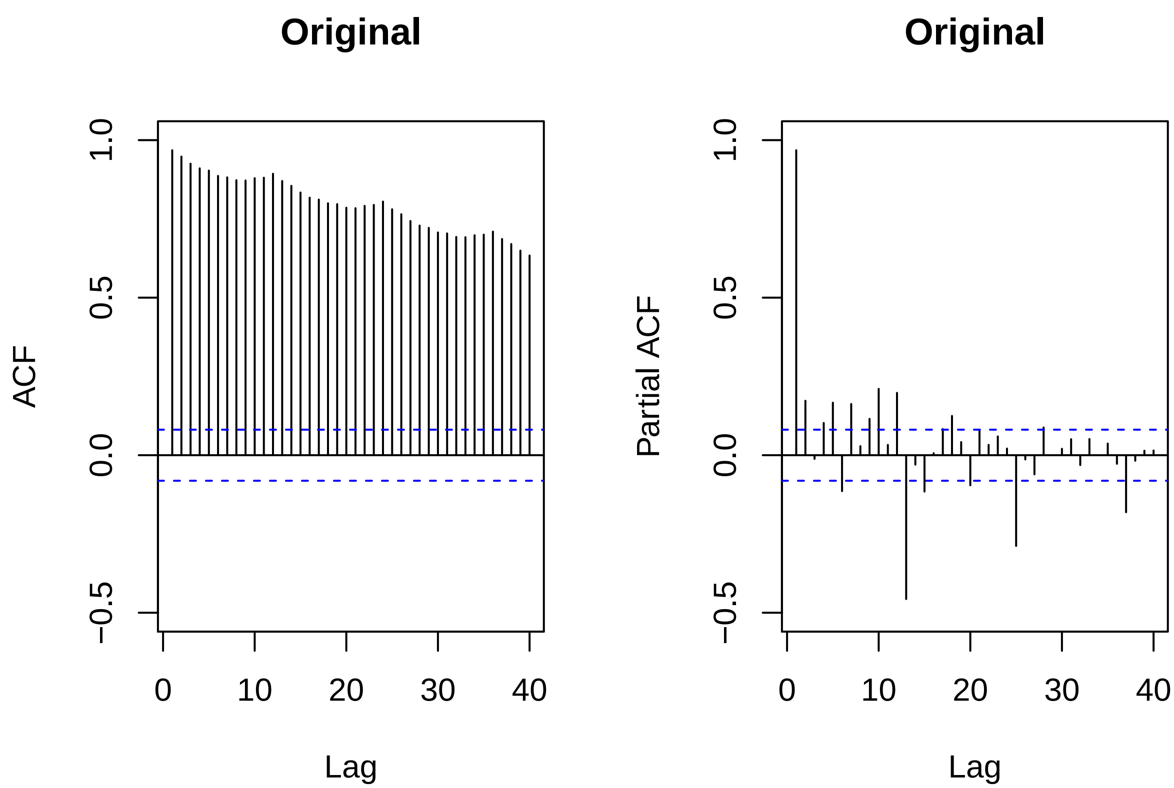


Q5

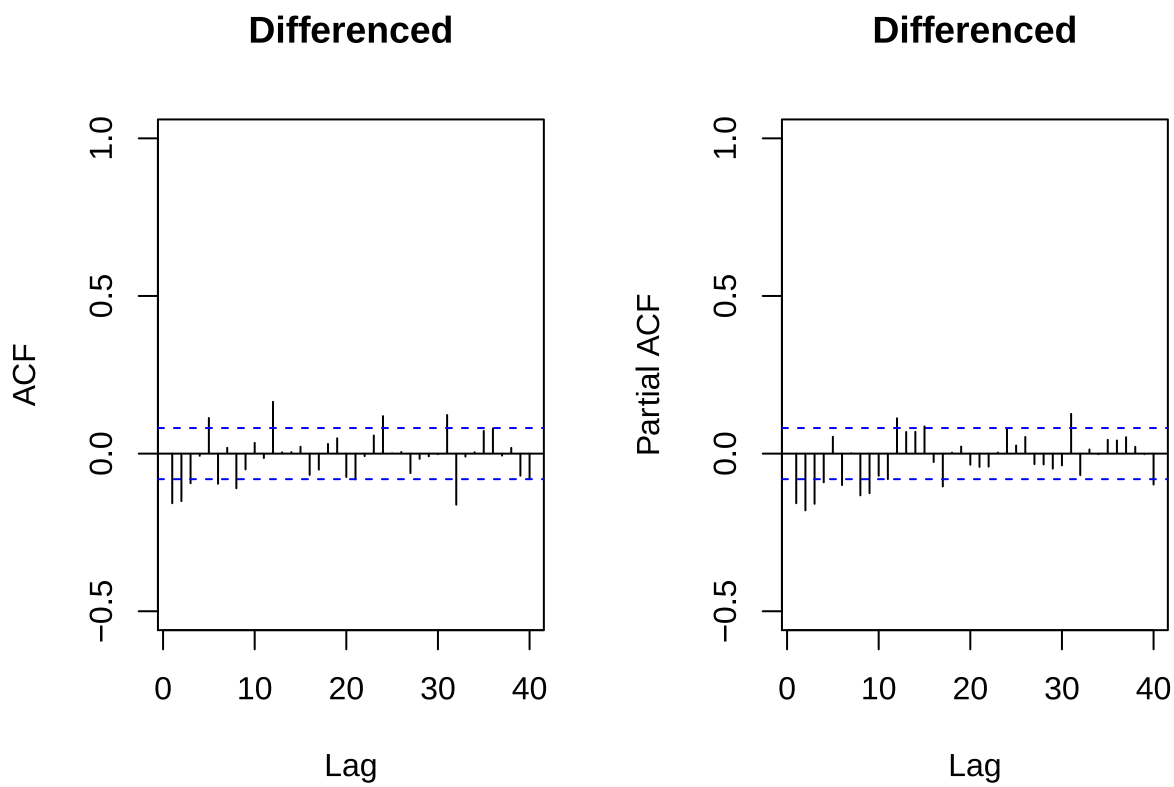
Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the Acf() function to make sure all three y axis have the same limits. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

- Differencing was more efficient in eliminating the trend because there are not many significant spikes for the Differencing Series and everything fall within the blue boundaries of the ACF and PACF plots for the Differencing Series.

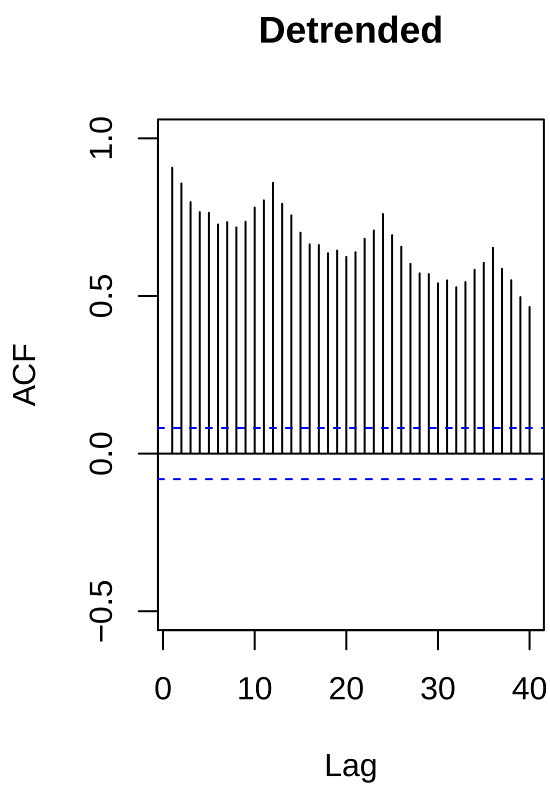
```
#Compare ACFs
par(mfrow=c(1,2))
Acf(df$Original,lag.max=40,main="Original",ylim=c(-0.5,1))
Pacf(df$Original,lag.max=40,main="Original",ylim=c(-0.5,1))
```

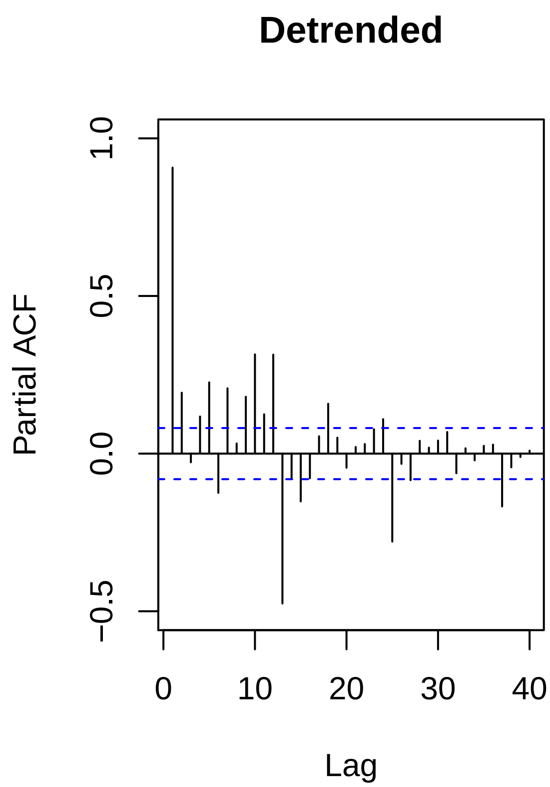



```
par(mfrow=c(1,2))
Acf(df4$Differenced,lag.max=40,main="Differenced",ylim=c(-0.5,1))
Pacf(df4$Differenced,lag.max=40,main="Differenced",ylim=c(-0.5,1))
```




```
par(mfrow=c(1,2))
Acf(df4$Detrend,lag.max=40,main="Detrended", ylim=c(-0.5,1))
Pacf(df4$Detrend,lag.max=40,main="Detrended", ylim=c(-0.5,1))
```





Q6

Compute the Seasonal Mann-Kendall and ADF Test for the original "Total Renewable Energy Production" series. Ask R to print the results. Interpret the results for both test. Whats the conclusion from the Seasonal Mann Kendall test? What's the conclusion for the ADF test? Do they match what you observed in Q2? Recall that having a unit root means the series has a stochastic trend. And when a series has stochastic trend we need to use a different procedure to remove the trend.

- Results from the Seasonal Mann-Kendall : A high Score of 9984 and small p value, indicating a trend. The positive Score value indicates a positive trend./
- Results from the ADF Test : The p-value =0.9554 which is greater than 0.05. Therefore, we accept the null hypothesis. The null hypothesis from the ADF test states that our time series has a unit root. Therefore, it has a stochastic trend./
- The results from the tests are in agreement with Q2. Q2's lm geom_smooth line increases over time. The slope of the trend line is positive. The graphs from Q2 also indicate that there is a trend.

```
ts_ener_data <- ts(Ener[,2:5],frequency=12)

SMKtest <- SeasonalMannKendall(ts_ener_data[,4])
print(summary(SMKtest))
```

```
## Score =  9984 , Var(Score) = 159104
## denominator =  13968
```

```
## tau = 0.715, 2-sided pvalue =< 2.22e-16
## NULL
```

```
print((adf.test(deseasonal_energy,alternative="stationary")))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  deseasonal_energy
## Dickey-Fuller = -0.86903, Lag order = 8, p-value = 0.9554
## alternative hypothesis: stationary
```

Q7

Aggregate the original "Total Renewable Energy Production" series by year. You can use the same procedure we used in class. Store series in a matrix where rows represent months and columns represent years. And then take the columns mean using function colMeans(). Recall the goal is the remove the seasonal variation from the series to check for trend.

```
ener_data_matrix<- matrix(Ener_processed[,5],byrow=FALSE,nrow=12)
```

```
## Warning in matrix(Ener_processed[, 5], byrow = FALSE, nrow = 12): data length
## [585] is not a sub-multiple or multiple of the number of rows [12]
```

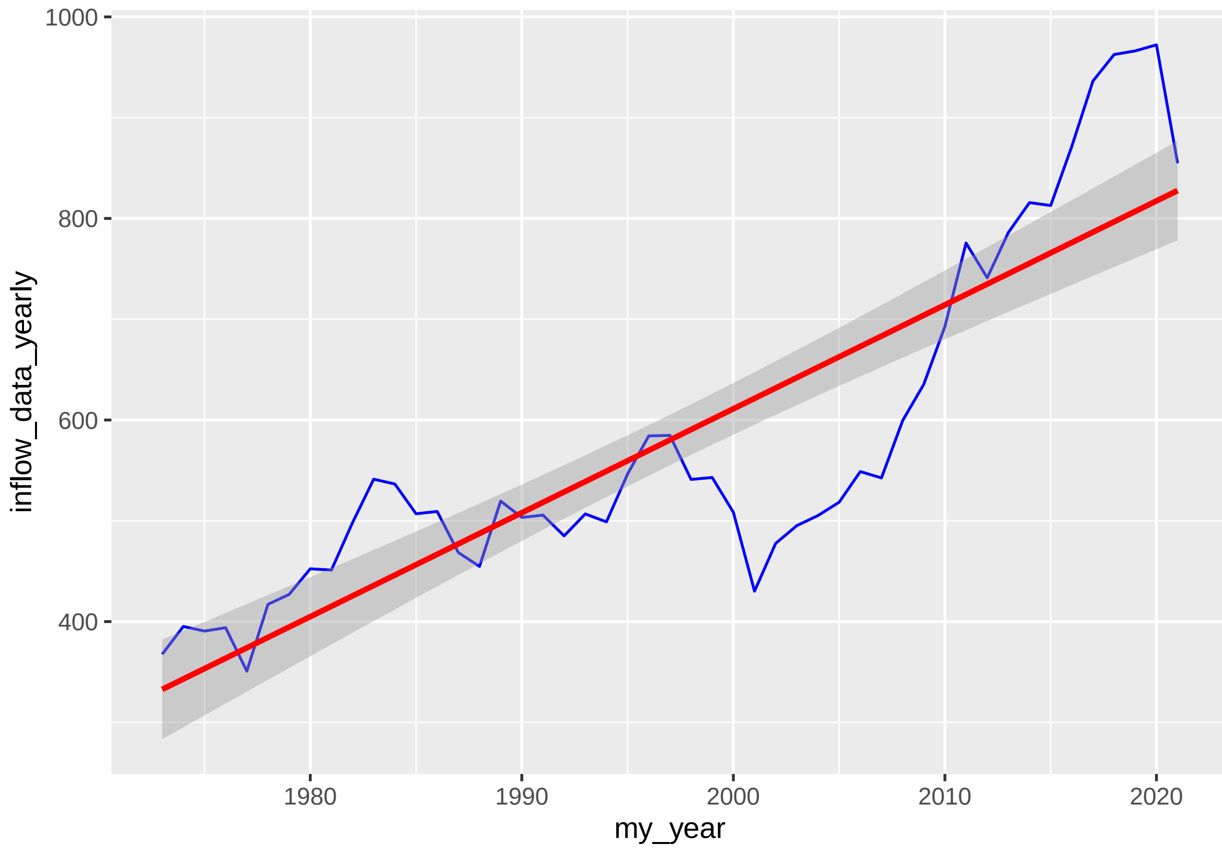
```
inflow_data_yearly <- colMeans(ener_data_matrix)


my_year <- c(year(first(Ener_processed$Date)):year(last(Ener_processed$Date)))
inflow_data_new_yearly <- data.frame(my_year, inflow_data_yearly)
inflow_data_new_yearly
```

```
##    my_year inflow_data_yearly
## 1     1973           367.5781
## 2     1974           395.1543
## 3     1975           390.5934
## 4     1976           393.9292
## 5     1977           350.7473
## 6     1978           417.1201
## 7     1979           426.9045
## 8     1980           452.3618
## 9     1981           451.1407
## 10    1982           498.3031
## 11    1983           541.3011
## 12    1984           536.4885
## 13    1985           507.0013
## 14    1986           509.2615
## 15    1987           468.4839
## 16    1988           454.7295
## 17    1989           519.5548
## 18    1990           503.3353
## 19    1991           505.6488
```

```
## 20    1992           485.0463
## 21    1993           506.8257
## 22    1994           498.9286
## 23    1995           546.4422
## 24    1996           584.2408
## 25    1997           584.7342
## 26    1998           541.0612
## 27    1999           542.9657
## 28    2000           508.4722
## 29    2001           430.1476
## 30    2002           477.5752
## 31    2003           495.2055
## 32    2004           505.2226
## 33    2005           518.4010
## 34    2006           548.8537
## 35    2007           542.5307
## 36    2008           599.2957
## 37    2009           635.4112
## 38    2010           692.8135
## 39    2011           775.6386
## 40    2012           741.0728
## 41    2013           786.0602
## 42    2014           815.7308
## 43    2015           812.8228
## 44    2016           871.6138
## 45    2017           936.3855
## 46    2018           962.6813
## 47    2019           966.2615
## 48    2020           972.2888
## 49    2021           854.7981
```

```
ggplot(inflow_data_new_yearly, aes(x=my_year, y=inflow_data_yearly)) +
            geom_line(color="blue") +
            geom_smooth(color="red",method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Q8

Apply the Mann Kendal, Spearman correlation rank test and ADF. Are the results from the test in agreement with the test results for the non-aggregated series, i.e., results for Q6?

- Results from the Mann-Kendall Score : The lower Score of 854 is due to less observations. However, the value is still a high number according to the z test, reflecting its significance. Like the results for Q6, the positive score indicates a positive trend. The results from the test is in agreement with the results for Q6. The tau value is slightly higher than Q6. This indicates that the increasing trend is clearer with yearly data than looking at the seasonal component. Similar to Q6, the very low p-value gives our confidence about our tau values./
- Results from the Spearman correlation : Spearman's rho is reported as 0.86. The null hypothesis states that rho=o, which indicates no trend. The alternative = rho is not equal to zero. Reject the null hypothesis and accept the alternative due to the low p value.

```
print(summary(MannKendall(inflow_data_yearly)))
```

```
## Score =  854 , Var(Score) = 13458.67
## denominator =  1176
## tau = 0.726, 2-sided pvalue =< 2.22e-16
## NULL
```

```r
sp_rho=cor.test(inflow_data_yearly,my_year,method="spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  inflow_data_yearly and my_year
## S = 2578, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8684694
```