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

Assignment 4 - Due date 02/17/22

Colin Lee

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)
library(readxl)
library(ggplot2)
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.1.2

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

```
library(tseries)

library(Kendall)
library(tidyverse)
```

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

## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4

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

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
mydata <- read.xlsx(file = "/Users/colinlee/Documents/Duke/Spring 2022/ENV790/ENV790_TimeSeriesAnalysis_

mydata <- mydata[5]

colnames(mydata)=c("Total Renewable Energy Production")

head(mydata)
```

```
##   Total Renewable Energy Production
## 1                           403.981
## 2                           360.900
## 3                           400.161
## 4                           380.470
## 5                           392.141
## 6                           377.232
```

Stochastic Trend and Stationarity Tests

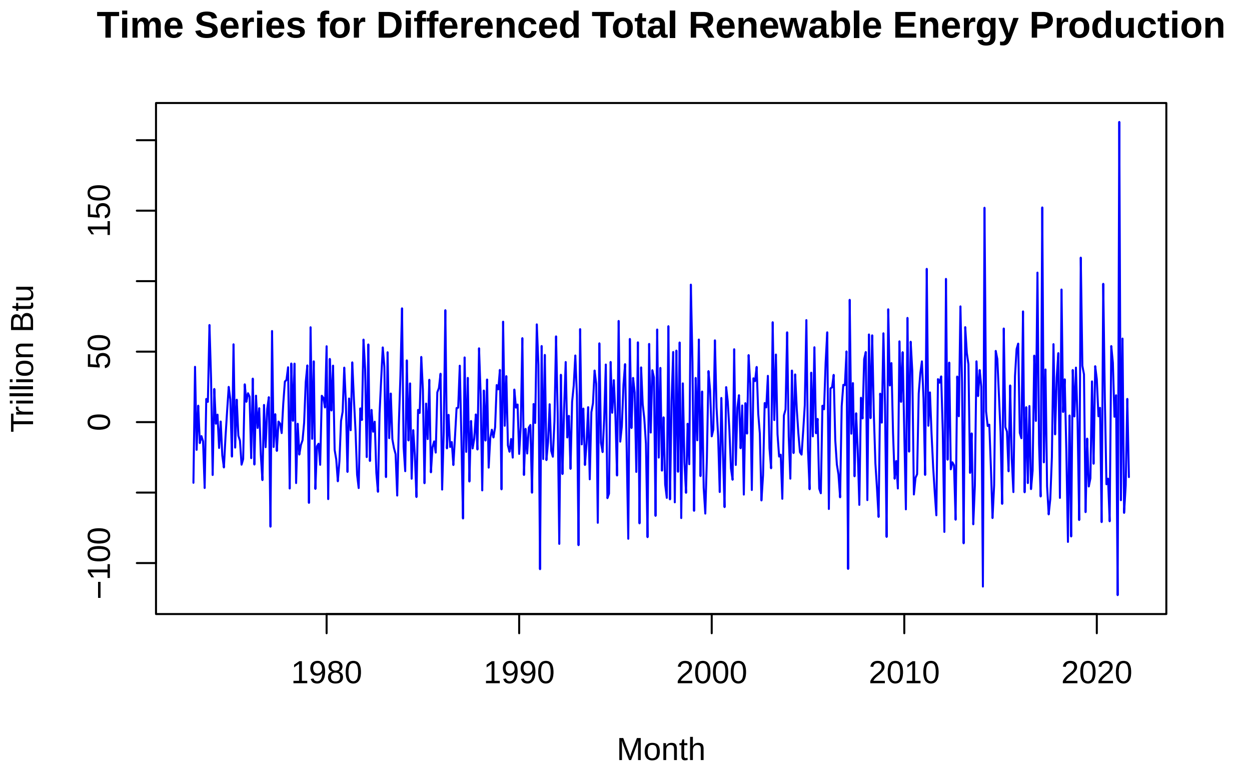
Q1

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.

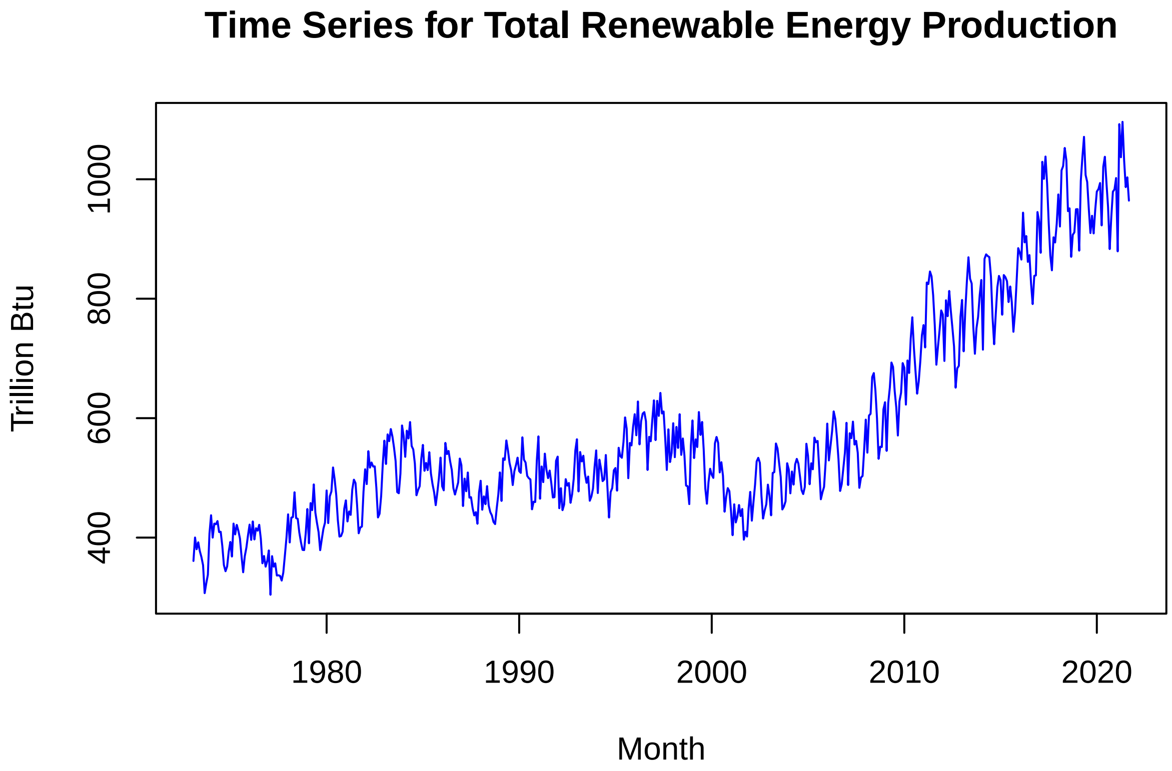
Try differencing at lag 1 only once, i.e., make `lag=1` and `differences=1`. Plot the differenced series Do the series still seem to have trend?

No, with the differenced series, it appears that there is no upwards trend that we see with the undifferenced series.

```
ts_mydata <- ts(mydata, start = c(1973,1), frequency = 12)

mydata_diff <- diff(ts_mydata,lag=1,differences=1)

df_mydata <-
  ts_mydata %>%
  cbind("Differenced Total Renewable Energy Production" = c(NA,as.numeric(mydata_diff))) %>%
  na.omit("Differenced Total Renewable Energy Production")

colnames(df_mydata)=c("Total Renewable Energy Production","Differenced Total Renewable Energy Productio

plot(df_mydata[,"Differenced Total Renewable Energy Production"],type="l",col="blue",ylab="Trillion Btu
title(main="Time Series for Differenced Total Renewable Energy Production")
```



```
plot(df_mydata[,"Total Renewable Energy Production"],type="l",col="blue",ylab="Trillion Btu",xlab = "Mon
title(main="Time Series for Total Renewable Energy Production")
```



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

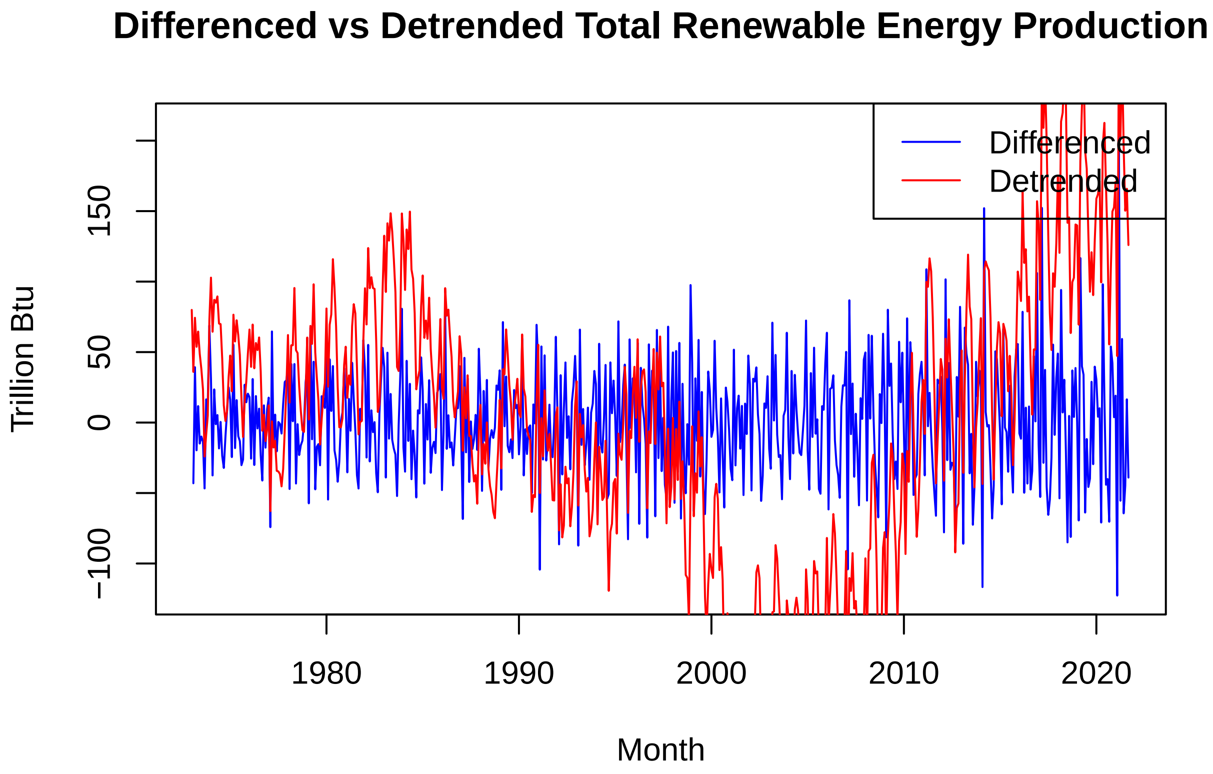
It appears the differencing is much flatter overall trend compared to the detrended data, which has a pretty significant dip, but its regression line has a slope ~0.

```
t <- c(1:nrow(mydata))

linear_trend_model_2=lm(mydata[,"Total Renewable Energy Production"]~t)
summary(linear_trend_model_2)
```

```
##
## Call:
## lm(formula = mydata[, "Total Renewable Energy Production"] ~
##     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
```

```
beta02=as.numeric(linear_trend_model_2$coefficients[1])  #first coefficient is the intercept term or be
beta12=as.numeric(linear_trend_model_2$coefficients[2])  #second coefficient is the slope or beta1

detrend_mydata_renewable <- mydata[,"Total Renewable Energy Production"]-(beta02+beta12*t)

ts_detrended <- ts(detrend_mydata_renewable, start = c(1973,1), frequency = 12)

plot(df_mydata[,"Differenced Total Renewable Energy Production"],type="l",col="blue",ylab="Trillion Btu
lines(ts_detrended,col="red")
title(main="Differenced vs Detrended Total Renewable Energy Production")
legend("topright",legend=c("Differenced", "Detrended"), lty=c("solid","solid"),col=c("blue","red"))
```



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.

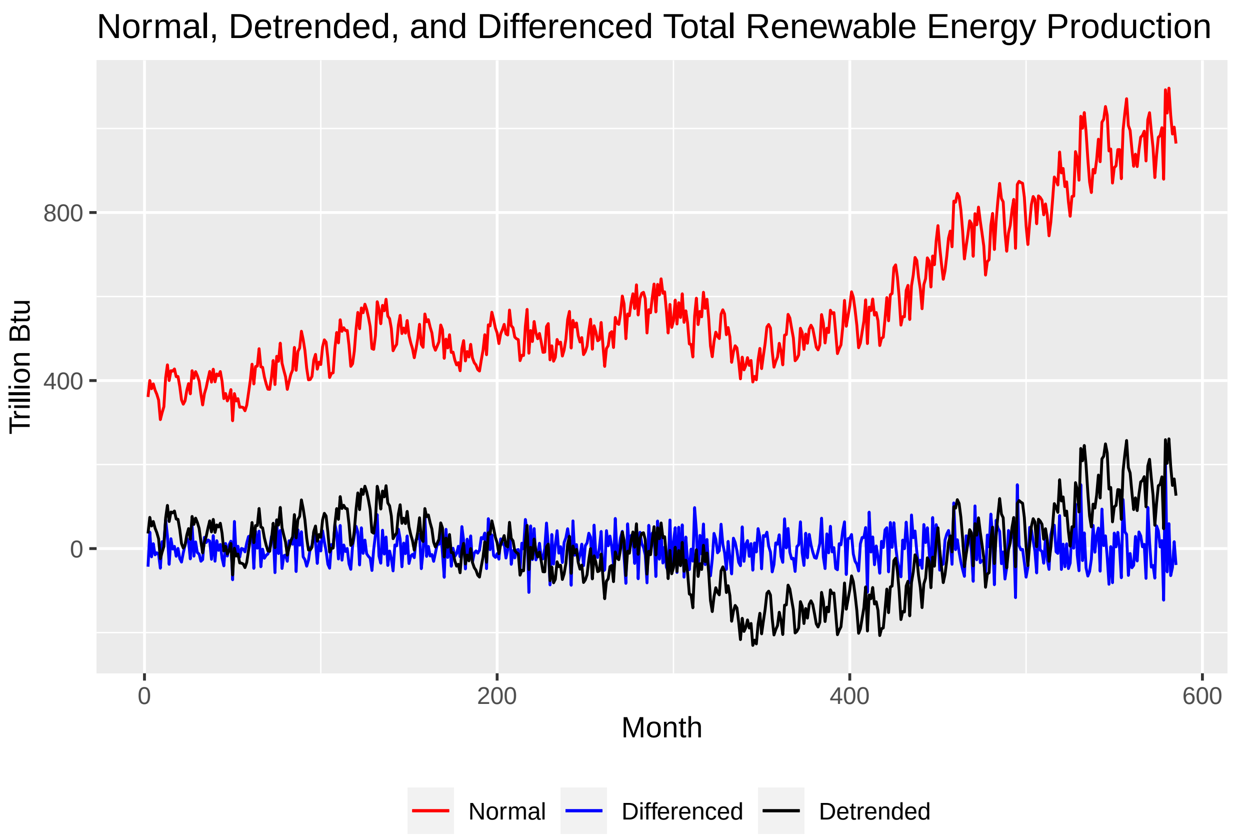
```
#Data frame - remember to not include January 1973
t <- c(1:nrow(mydata))
ts_t <- ts(t, start = c(1973,1), frequency = 12)
t2 <- ts_t[-1]
ts_detrended2 <- ts_detrended[-1]
totaldata <- data.frame((df_mydata[,1]),(df_mydata[,2]), (ts_detrended2),(t2))
colnames(totaldata)=c("Total Renewable Energy Production","Differenced Total Renewable Energy Productio
```

Q4

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

```
ggplot(totaldata) +
  geom_line(aes(x = Month, y = totaldata[,1], color = "1")) +
  geom_line(aes(x = Month, y = totaldata[,2], color = "2")) +
  geom_line(aes(x = Month, y = totaldata[,3], color = "3")) +
  labs(y = "Trillion Btu",
            x = "Month",
            title = "Normal, Detrended, and Differenced Total Renewable Energy Production",color="") +
  scale_color_manual(values = c("1" = "red", "2" = "blue", "3" = "black"),
                                labels=c("Normal", "Differenced", "Detrended")) +
  theme(legend.position = "bottom")
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



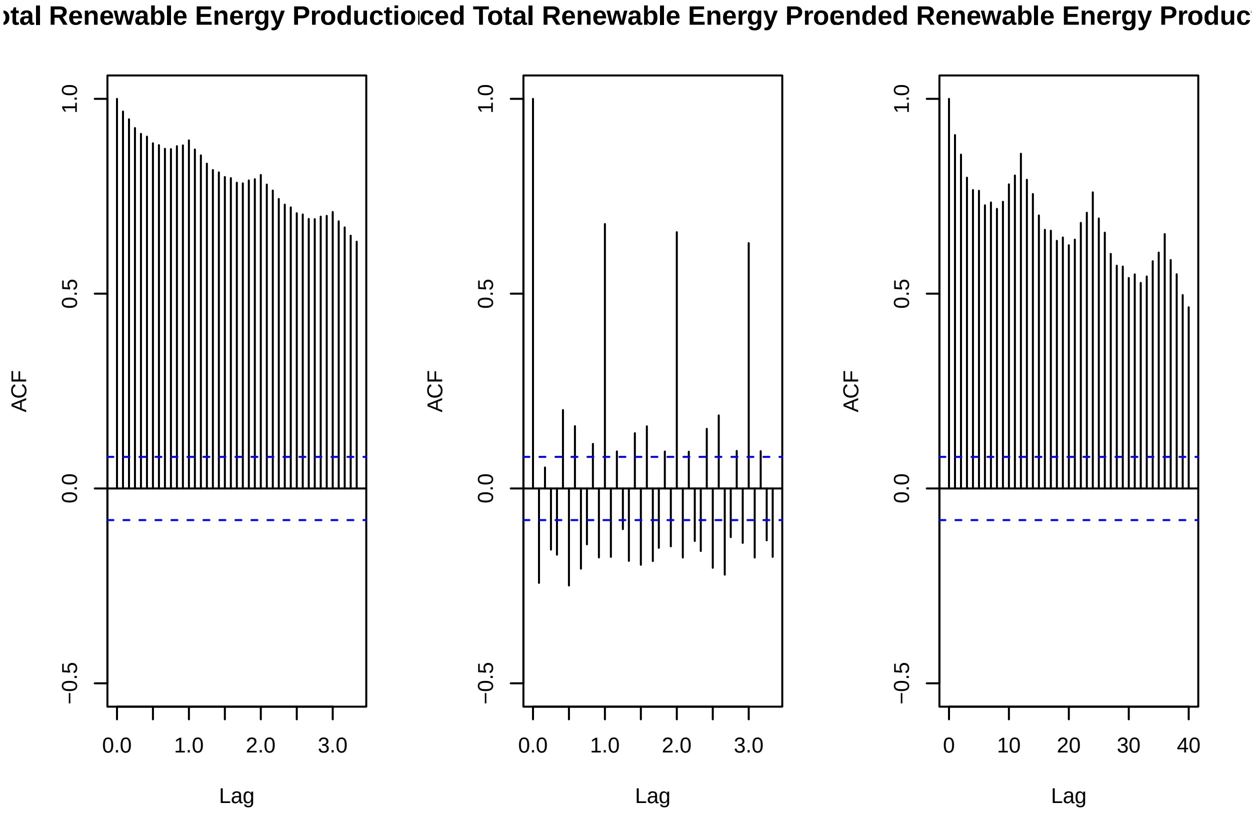
```
#could not get legend to appear for some reason? sorry
```

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?

I think the differencing was most effective in eliminating the trend since the ACF is much more random. The other linear regression ACF still has certain trend characteristics.

```
#Compare ACFs
par(mfrow=c(1,3))
acf(totaldata[,1],lag.max=40,main=paste("Total Renewable Energy Production ACF"),ylim=c(-0.5,1))
acf(totaldata[,2],lag.max=40,main=paste("Differenced Total Renewable Energy Production ACF"),ylim=c(-0.
acf(totaldata[,3],lag.max=40,main=paste("Detrended Renewable Energy Production ACF"),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.

According to the SMK test, we have a high score of 9984 and low p-value, we reject the null hypothesis and Total Renewable Energy Production does follow a trend.

According to the ADF test, we have a p-value of 0.8, indicating that there the series contains a unit root and is not stationary.

This correlates with my understanding that the high autocorrelation of the Total Renewable Energy Production indicates the series is non-stationary, which was proven with ADF. Furthermore, it is clear that the data follows a trend, which is shown by the SMK test.

Thus, it appears that the series is a non-stationary, likely stochastic trend, which makes sense given the graph where the renewable energy production has a fixed increase over time.

```r
SMKtest <- SeasonalMannKendall(ts_mydata)
print("Results for Seasonal Mann Kendall /n")
```

```
## [1] "Results for Seasonal Mann Kendall /n"
```

```r
print(summary(SMKtest))
```

```
## Score =  9984 , Var(Score) = 159104
## denominator =  13968
## tau = 0.715, 2-sided pvalue =< 2.22e-16
## NULL
```

```r
print("Results for ADF test/n")
```

```
## [1] "Results for ADF test/n"
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_mydata
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161
## 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.

```r
mydata_matrix <- matrix(ts_mydata,byrow=FALSE,nrow=12)
```

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

```r
mydata_yearly <- colMeans(mydata_matrix)
```

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?

The results are in agreement with the results from Q6. With regard to the MK test, we also see a high score (not as high since it's years now) and a low p-value of 2.2e-16, indicating the data follows a trend, just like the monthly data in Q6.

Furthermore, Spearman correlation rank test shows that with a p-value also at around 2.2E-16, that the data also follows a trend. Furthermore, the Rho value is 0.868, indicating a strong positive correlation. This does not contradict the outcomes of Q6, as we did not perform a Spearman test in Q6.

The ADF test is also similar to Q6 where we fail to reject the null hypothesis and the time series does contain a unit root, indicating the series is non-stationary.

Overall, the results match those in Q6 and the lack of a seaosonal trend did not make a large difference.

```
print("Results of Mann Kendall on average yearly series")
```

```
## [1] "Results of Mann Kendall on average yearly series"
```

```
print(summary(MannKendall(mydata_yearly)))
```

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

```
#cor.test test statistics
sp_rho=cor.test(mydata_yearly,c(1:ncol(mydata_matrix)),method="spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  mydata_yearly and c(1:ncol(mydata_matrix))
## S = 2578, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8684694
```

```
print("Results for ADF test on yearly data/n")
```

```
## [1] "Results for ADF test on yearly data/n"
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  mydata_yearly
## Dickey-Fuller = -2.2085, Lag order = 3, p-value = 0.4907
## alternative hypothesis: stationary
```