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

Assignment 4 - Due date 02/17/22

Yared S. Asfaw

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(openxlsx)
library(ggplot2)
library(forecast)
library(tseries)
library(Kendall)
library(lubridate)
library(dplyr)
library(tidyverse)
```

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
us_REPC_by_source <- read.xlsx("./Data/Raw/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
startRow=11)

us_TREP <- data.frame(us_REPC_by_source[-1,c(1,5)])

class(us_TREP$Total.Renewable.Energy.Production)
```

```
## [1] "character"
```

```
us_TREP$Total.Renewable.Energy.Production <- as.numeric(us_TREP$Total.Renewable.Energy.Production)
class(us_TREP$Total.Renewable.Energy.Production)
```

```
## [1] "numeric"
```

```
# Renaming the column names for convenience
colnames(us_TREP) <- c("my_date", "TREP")
colnames(us_TREP)
```

```
## [1] "my_date" "TREP"
```

```
# Formatting the my_date column to "Date"
start_date <- as.Date("01/01/1973", format = "%m/%d/%Y") # the date format in the dataset is m/d/y
us_TREP$my_date <- seq(start_date, by= "month", length.out=585)
```

Stochastic Trend and Stationarity Tests

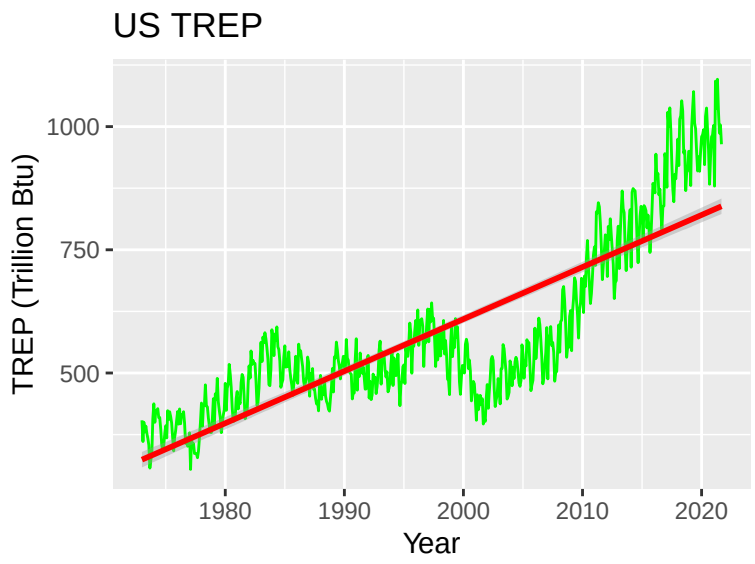
Q1

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?

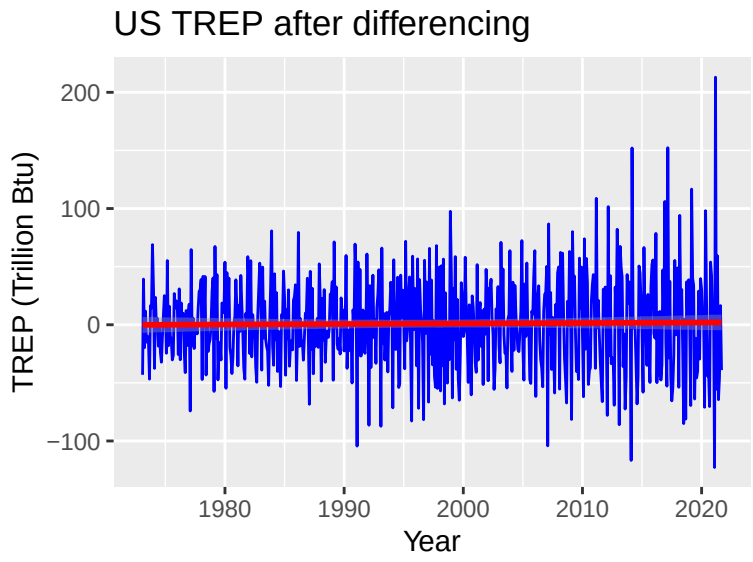
```
# The plot before differencing
ggplot(us_TREP, aes(x=my_date, y=us_TREP[,2]))+
  geom_line(color = "green") +
  geom_smooth(color="red", method = "lm")+
  ylab(label = "TREP (Trillion Btu)") +
  xlab(label = "Year")+
  ggtitle("US TREP")
```

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




```
# Differencing the "Total Renewable Energy Production" series
us_TREP_diff <- diff(us_TREP[,2], lag = 1, differences = 1)

# Converting it to data frame
# Equalizing the number of observations in the original series with the differenced
#series to form data frame
us_TREP1 <- us_TREP[-1,]

df.us_TREP_diff <- data.frame("Date" = as.Date(us_TREP1$my_date), us_TREP_diff)

# Plotting the diffeernced TREP series with the mean horizontal line to check the trend

ggplot(df.us_TREP_diff, aes(x=Date, y=us_TREP_diff))+
  geom_line(color = "blue") +
  geom_smooth(color="red", method = "lm")+
  ylab(label = "TREP (Trillion Btu)") +
  xlab(label = "Year")+
  ggtitle("US TREP after differencing")
```

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



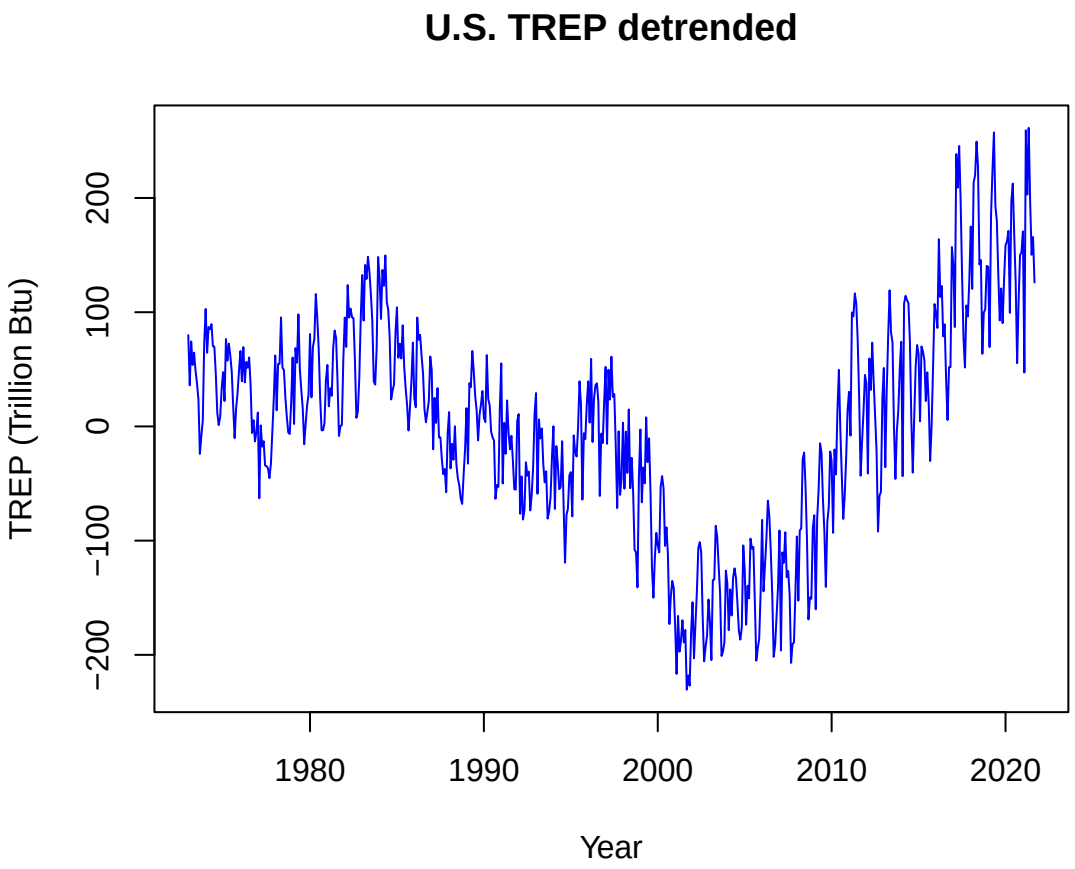
No, the series do not have trend that can be traced after differencing.

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

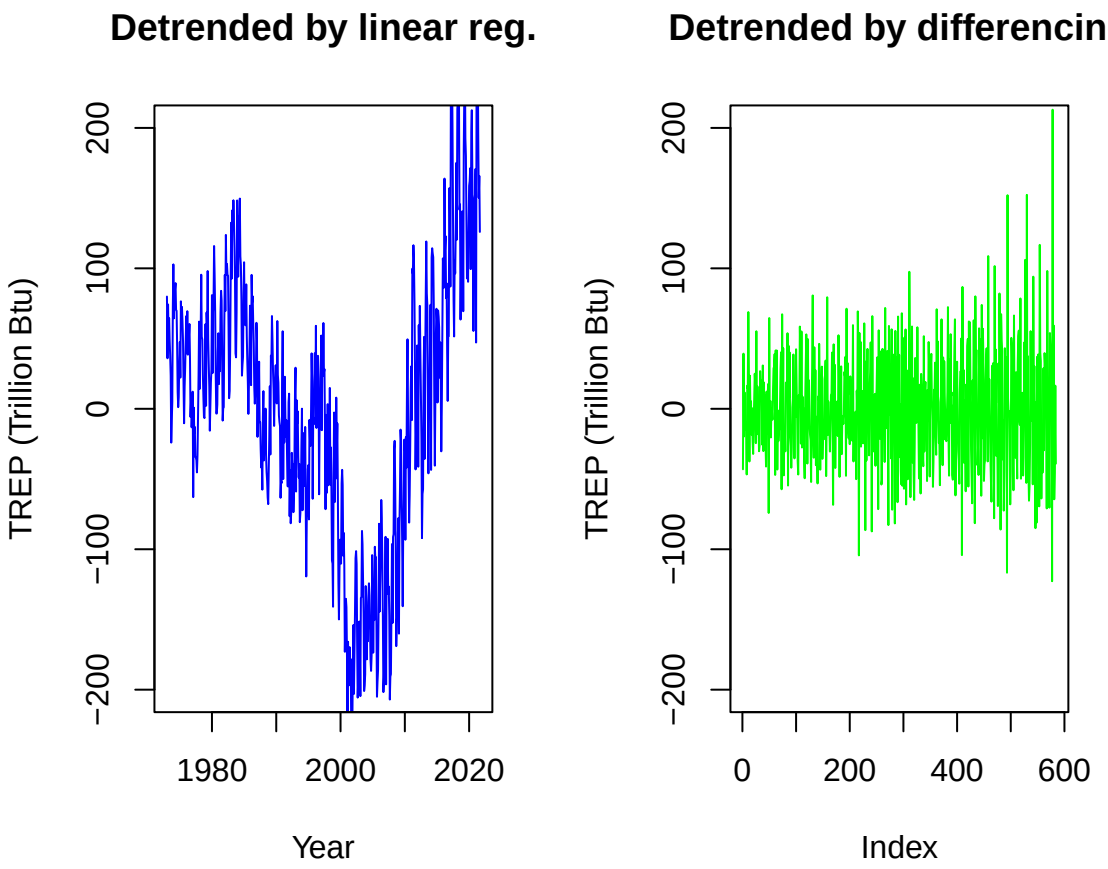
```
# Code from A03 for reference and use
ts_us_TREP <- ts(us_TREP$TREP, start = c(1973, 1), frequency= 12)
t <- c(1:585)
linear_trend_TREP = lm(ts_us_TREP ~ t)
beta0_TREP <- as.numeric(linear_trend_TREP$coefficients[1]) # Intercept
beta1_TREP <- as.numeric(linear_trend_TREP$coefficients[2]) # Slope
detrend_ts_us_TREP <- ts_us_TREP - (beta0_TREP + beta1_TREP * t)

plot(detrend_ts_us_TREP, type="l",col="blue", main="U.S. TREP detrended",
ylab="TREP (Trillion Btu)",
xlab= "Year")
```




```
# Comparing the differenced and the detrended series
par(margin(4,4,4,4));par(mfrow=c(1,2))
```

```
## NULL
```

```
plot(detrend_ts_us_TREP, type="l",col="blue",
     ylim=c(-200,200), xlab= "Year",
ylab="TREP (Trillion Btu)",
main="Detrended by linear reg.")
plot(us_TREP_diff, type="l",col="green",
     ylim=c(-200,200), ylab="TREP (Trillion Btu)",
main="Detrended by differencing")
```



When the TREP series detrended by linear regression, it shows a random movement /pattern of incereasing and decreasing over time whereas when it is differenced, the series doesn't show a trend.

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.

```
# us_TREP is the original series
# detrend_ts_us_TREP is the detrended by regression series
# us_TREP_diff is the differenced series

# Removing the first observations from the original series and the detrended by regrression series
us_TREP1 <- us_TREP[-1,]
detrend_ts_us_TREP1 <- as.data.frame(detrend_ts_us_TREP)
detrend_ts_us_TREP1 <- as.data.frame(detrend_ts_us_TREP1[-1,])
us_TREP_diff1 <- as.data.frame(us_TREP_diff)

combined_series_df <- data.frame(myDate=as.Date((us_TREP1$my_date), format="%Y-%d-%m"),
```

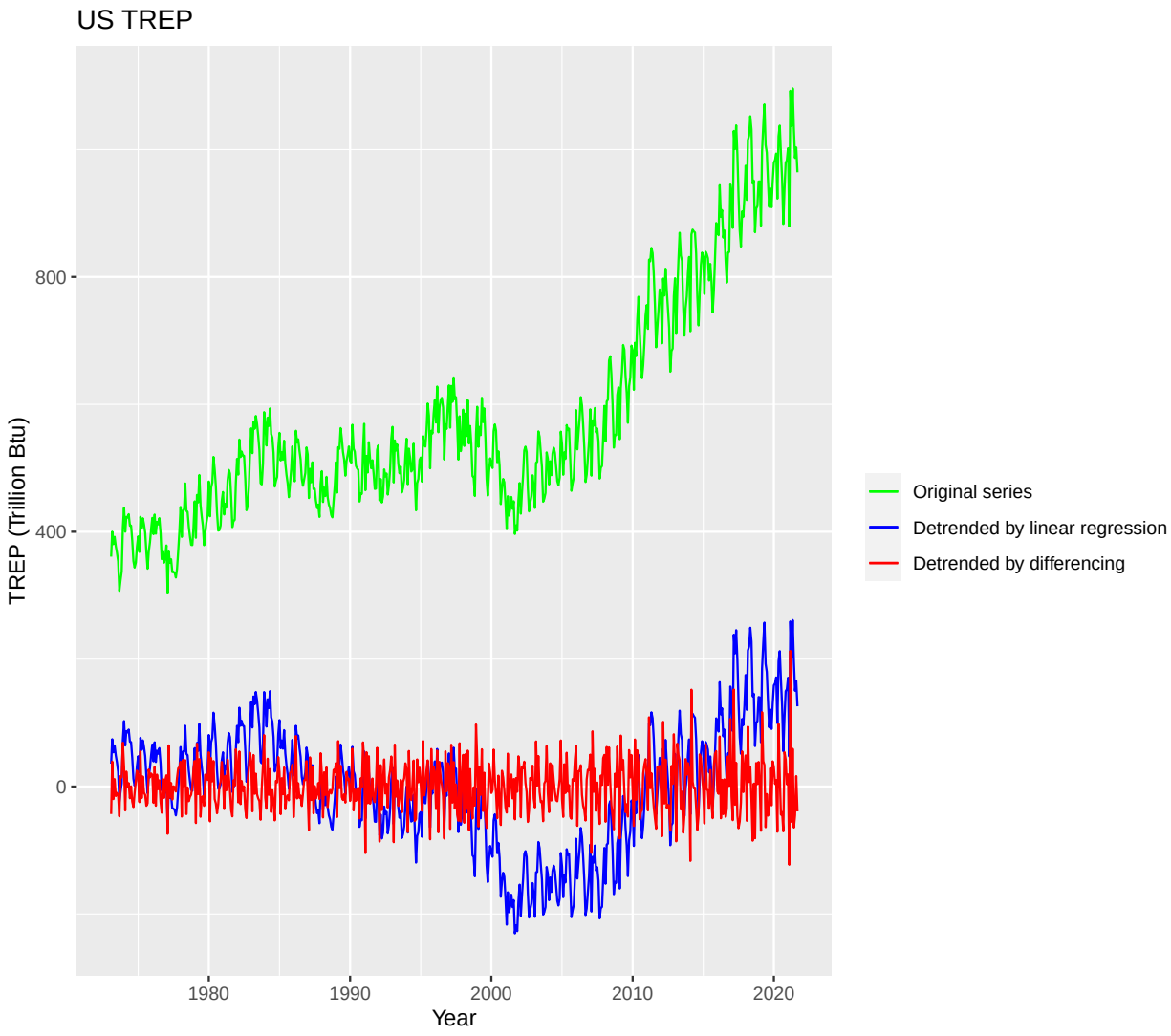
```
Original_TREP=us_TREP1$TREP,
Detrended_by_reg=detrend_ts_us_TREP1$`detrend_ts_us_TREP1[-1, ]`, Detrended_by_differ=us_TREP_diff1$us_
head(combined_series_df)
```

```
##       myDate Original_TREP Detrended_by_reg Detrended_by_differ
## 1 1973-02-01       360.900         35.95655             -43.081
## 2 1973-03-01       400.161         74.33705              39.261
## 3 1973-04-01       380.470         53.76554             -19.691
## 4 1973-05-01       392.141         64.55603              11.671
## 5 1973-06-01       377.232         48.76653             -14.909
## 6 1973-07-01       367.325         37.97902              -9.907
```

Q4

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

```
#Using ggplot to create a line plot for the three series
ggplot(combined_series_df, aes(x=myDate)) +
  geom_line(aes(y= Original_TREP, color="Original_TREP")) +
  geom_line(aes(y=Detrended_by_reg, color="Detrended_by_reg")) +
  geom_line(aes(y=Detrended_by_differ, color="Detrended_by_differ")) +
  labs(color="") +
  scale_color_manual(values = c("Original_TREP" = "green", "Detrended_by_reg" = "blue",
"Detrended_by_differ" = "red"),
labels=c("Original series",
"Detrended by linear regression",
"Detrended by differencing")) +
theme(legend.position = "right") +
ylab(label="TREP (Trillion Btu)") +
xlab(label="Year") +
ggtitle("US TREP")
```



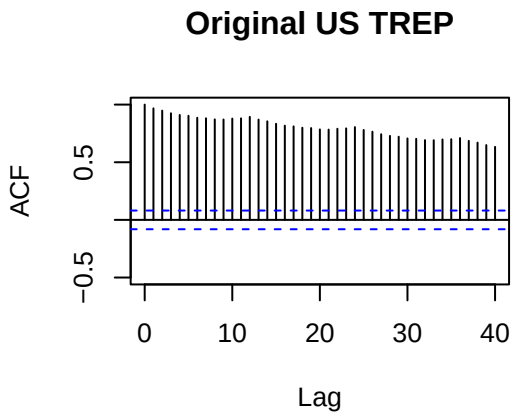
Q5

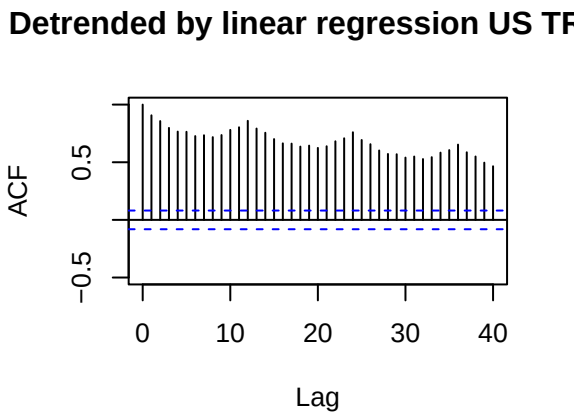
Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the Acf() function to make sure all three y axis have the same limits. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

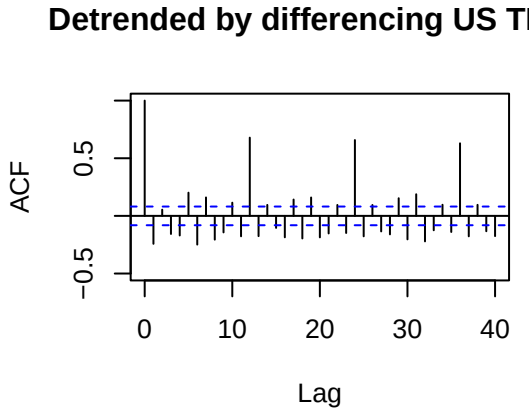
```
# Comparing the ACFs of the original series, detrended by Regression # Series and differenced series
par(mfrow=c(2,2));par(margin(4,4,4,4))
```

```
## NULL
```

```
acf(us_TREP1$TREP, ylim=c(-0.5,1), lag.max = 40, plot = TRUE, main= "Original US TREP")
acf(detrend_ts_us_TREP1, ylim=c(-0.5,1), lag.max = 40,plot = TRUE,
main= "Detrended by linear regression US TREP")
acf(us_TREP_diff1, ylim=c(-0.5,1), lag.max = 40,plot = TRUE,
main= "Detrended by differencing US TREP")
```







When comparing the ACF plots, the original series shows a gradual decline over time, the detrended using linear regression series shows a decline over time with a kind of seasonal pattern, whereas the differenced series shows a total removal of the trend and there is no sign left that show a trend in the series. Thus, the more efficient method in eliminating the trend is the differencing.

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.

```
# Computing the Seasonal Mann-Kendall and ADF Test for the original TREP series
SMKtest <- SeasonalMannKendall(ts_us_TREP) # Seasonal Mann-Kendall test
print("Results for Seasonal Mann Kendall /n")
```

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

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

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

The p-value is less than 0.05 and thus, we reject the null hypothesis that states the series is stationary. This tells that the series exhibits a monotonic trend and this trend is an increasing trend (tau = 0.715). In other words, the TREP has an increasing trend overtime.

```
ADFtest <- adf.test(ts_us_TREP) # Agumented Dickey-Fuller test
print("Results for ADF test /n")
```

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

```
print(ADFtest)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_us_TREP
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161
## alternative hypothesis: stationary
```

The p-value is 0.8161 and the test statistic is -1.4383, and the p-value is geater than 0.05, thus, we fail to reject the null hypotesis (the series is non-stationary). Thus, the TREP series has a stochastic trend or the series exhibits a time-dependent structure and does not have constant variance over time.

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.

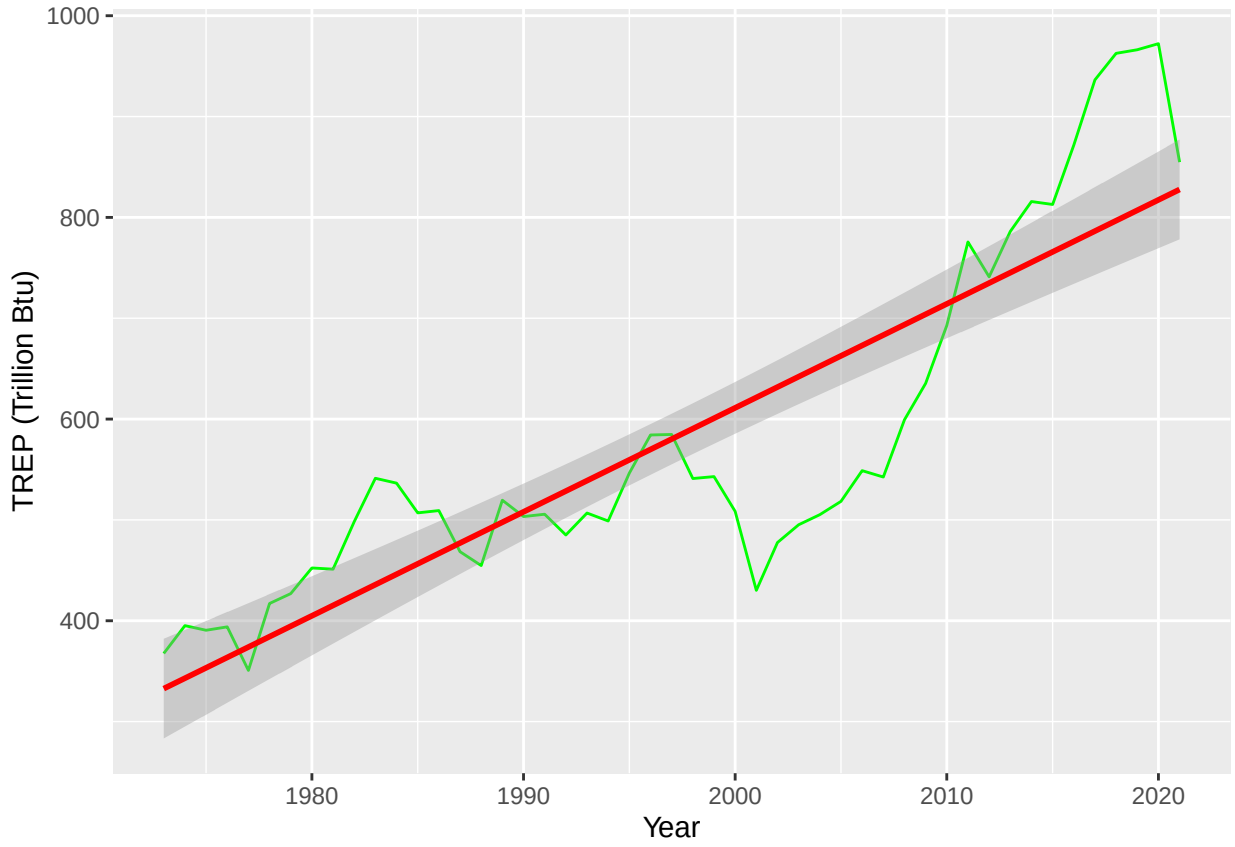
```
# Aggregating the original TREP series by year
# Grouping the data in yearly step instances
us_TREP_matrix <- matrix(us_TREP$TREP, byrow = FALSE, nrow = 12)
```

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

```
us_TREP_yearly <- colMeans(us_TREP_matrix)

my_year <- c(year(first(us_TREP$my_date)):year(last(us_TREP$my_date)))
```

```
# Aggregating the series by year
us_TREP_agg_yearly <- data.frame(my_year, us_TREP_yearly)

ggplot(us_TREP_agg_yearly, aes(x=my_year, y=us_TREP_yearly))+
  geom_line(color="green") +
  geom_smooth(color="red",method="lm") +
  ylab(label="TREP (Trillion Btu)") +
  xlab(label="Year")
```

```
## `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?

```
# Checking deterministic trend with Mann Kendall test
ts_us_TREP_agg_yearly <- as.ts(us_TREP_agg_yearly)
MKtest <- MannKendall(ts_us_TREP_agg_yearly)
print("Results for Mann Kendall /n")
```

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

```
print(summary(MKtest))
```

```
## Score =  -371 , Var(Score) = 106150.3
## denominator =  4753
## tau = -0.0781, 2-sided pvalue =0.25611
## NULL
```

According to the test, the p-value is greater than the significance level 0.05 and thus, lead to the conclusion of failing to reject the null hypothesis (the series is stationary). However, this finding is different from the finding obtained from the Mann Kendall test for the non-aggregated series under question number 6 where we rejected the null hypothesis and conclude that the series has a non-stationary trend. One reason for this difference could be the number of data points/observations in the aggregated series is limited that the test couldn't detect the true trend in the series.The availablity of more data will help the MK test to distingusih the trend the more easily.

```
# Checking stochastic trend with Agumented Dickey-Fuller test
ADFtest_yearly <- adf.test(us_TREP_agg_yearly$us_TREP_yearly)
print("Results for ADF test /n")
```

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

```
print(ADFtest_yearly)
```

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

The p-value (0.49) and the test statistic is -2.2085, and the p-value is geater than 0.05 (the significance level), thus, we fail to reject the null hypotesis (the series is non-stationary or contain unit root). Thus, the TREP series has a time-dependent structure or stochastic trend. This result is similar to the conclusion we reached for the non-aggregated series under question number 6.

```
# Checking deterministic trend with Spearman Correlation Rank Test
sp_cor_test1 <- cor.test(us_TREP_agg_yearly$us_TREP_yearly, my_year, method = "spearman")
print("Results from Spearman Correlation Rank Test")
```

```
## [1] "Results from Spearman Correlation Rank Test"
```

```
print(sp_cor_test1)
```

```
##
##  Spearman's rank correlation rho
```

```
##
## data:  us_TREP_agg_yearly$us_TREP_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
```

The p-value is less than the significance level, thus we reject the null hypothesis which says the series is stationary or doesn't have a trend. Therefore, there is a significant association (rho=0.87) between the TREP series and time and this association is a positive deterministic /an increasing trend over time.