ENV 797 - Time Series Analysis for Energy and Environment Applications | Spring 2025

Assignment 4 - Due date 02/11/25

Yuqi Yang

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 file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., “LuanaLima_TSA_A04_Sp25.Rmd”). Then change “Student Name” on line 4 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. 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.\

```r
#Load/install required package here
library(readxl)
library(ggplot2)
library(forecast)
library(tseries)
library(Kendall)
library(trend)
library(cowplot)
```

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”**.

```r
#Importing data set - you may copy your code from A3
raw_energy_data <- read_excel(
path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
skip = 12,
sheet="Monthly Data",
col_names=FALSE)

#Extract the column names from row 11 only
read_col_names <- read_excel(
path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
skip = 10,
n_max = 1,
sheet="Monthly Data",
col_names=FALSE)

#Assign the column names to the data set
colnames(raw_energy_data) <- read_col_names

#Selecting the columns of interest
energy_data <- raw_energy_data[,c(1,5)]
nobs <- nrow(energy_data)

#Remove last year by replacing current data frame
tail(energy_data)
```

```
## # A tibble: 6 x 2
##   Month               ‘Total Renewable Energy Production‘
##   <dttm>                                            <dbl>
## 1 2024-04-01 00:00:00                                751.
## 2 2024-05-01 00:00:00                                762.
## 3 2024-06-01 00:00:00                                758.
## 4 2024-07-01 00:00:00                                746.
## 5 2024-08-01 00:00:00                                751.
## 6 2024-09-01 00:00:00                                695.
```

```
energy_data <- energy_data[1:(nobs-9),]

#update object with number of observations
nobs <- nobs-9

#Tail again to check if the rows were correctly removed
tail(energy_data)
```

```
## # A tibble: 6 x 2
##   Month               ‘Total Renewable Energy Production‘
##   <dttm>                                            <dbl>
## 1 2023-07-01 00:00:00                                716.
## 2 2023-08-01 00:00:00                                713.
## 3 2023-09-01 00:00:00                                673.
## 4 2023-10-01 00:00:00                                694.
## 5 2023-11-01 00:00:00                                682.
## 6 2023-12-01 00:00:00                                721.
```

```
#Create vector t - time index
t <- 1:nobs

#transforming into ts object
```

```
ts_energy_data <- ts(energy_data[t,2], frequency=12,start=c(1973,1))

head(ts_energy_data)
```

```
##          Jan     Feb     Mar     Apr     May     Jun
## 1973 219.839 197.330 218.686 209.330 215.982 208.249
```

Stochastic Trend and Stationarity Tests

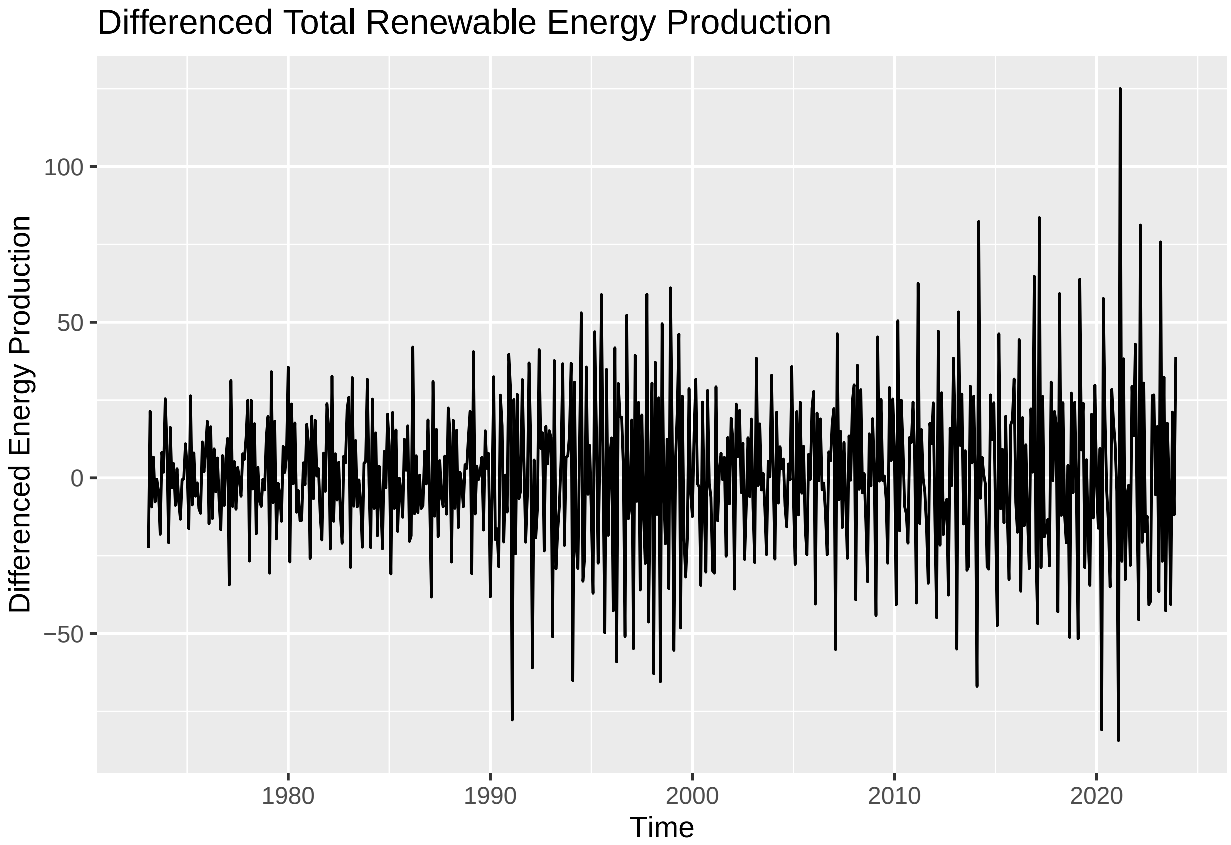
For this part you will work only with the column Total Renewable Energy Production.

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?

```
#Difference the series
diff_energy_data <- diff(ts_energy_data,lag = 1, differences = 1)

#Transform into time series objects
ts_diff_renewable <- ts(diff_energy_data, start = c(1973,1), frequency = 12)

#Plot the differenced series
autoplot(diff_energy_data) +
  ggtitle("Differenced Total Renewable Energy Production") +
  ylab("Differenced Energy Production")
```



The differenced series does not show a trend, but the variance seems to increase over time.

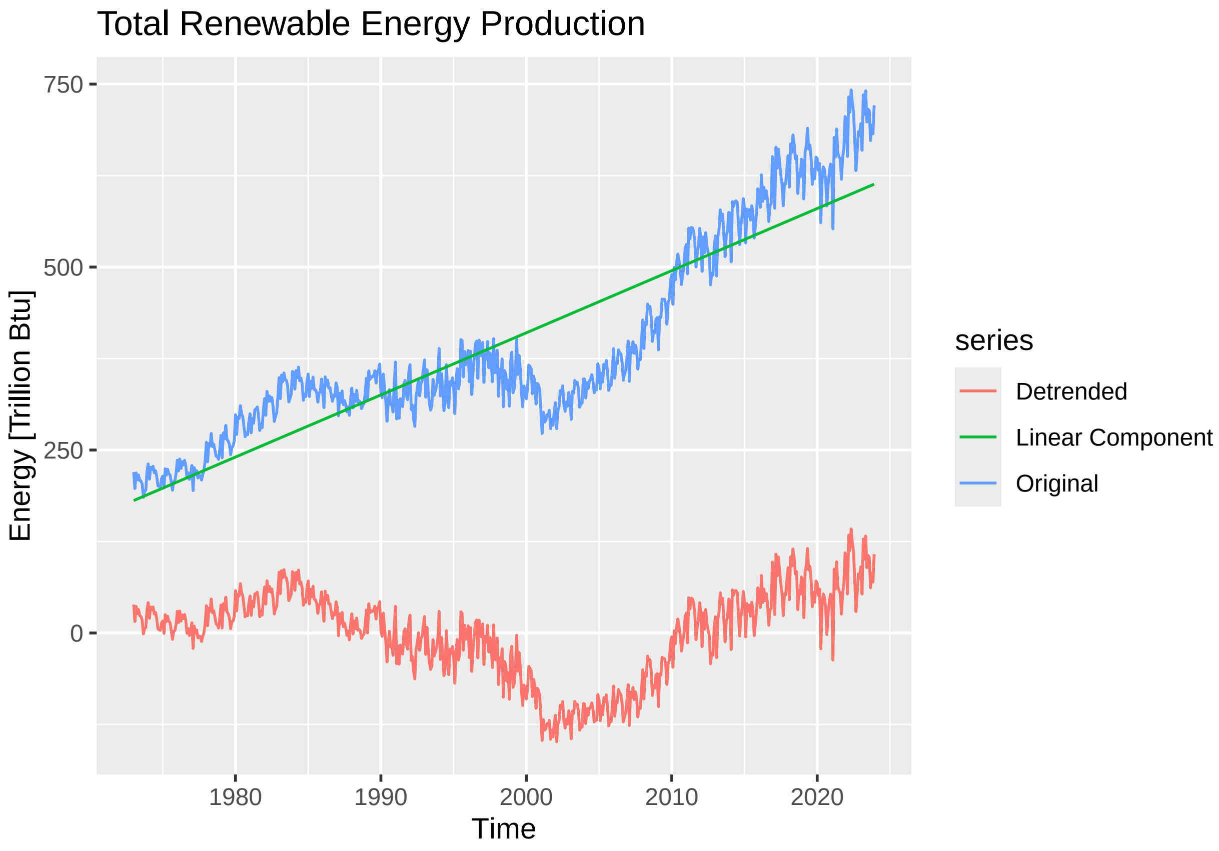
Q2

Copy and paste part of your code for A3 where you run the regression for Total Renewable Energy Production and subtract that from the original series. This should be the code for Q3 and Q4. make sure you use the same name for you time series object that you had in A3, otherwise the code will not work.

```
#Fit a linear trend to the series
linear_trend_model_renewable <- lm(ts_energy_data ~ t)
summary(linear_trend_model_renewable)
```

```
##
## Call:
## lm(formula = ts_energy_data ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.76  -36.24   12.25   41.49  142.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.24534    4.89680   36.81   <2e-16 ***
## t             0.70757    0.01384   51.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.5 on 610 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.8104
## F-statistic:  2613 on 1 and 610 DF,  p-value: < 2.2e-16
```

```
#Save coefficients for further analysis
beta0_renewable <- coef(linear_trend_model_renewable)[1]
beta1_renewable <- coef(linear_trend_model_renewable)[2]

#Detrend the series
linear_trend_renewable <- beta0_renewable + beta1_renewable * t

detrend_renewable <- ts_energy_data - linear_trend_renewable

#Transform into time series objects
ts_linear_renewable <- ts(linear_trend_renewable, start = c(1973,1), frequency = 12)
ts_detrend_renewable <- ts(detrend_renewable, start = c(1973,1), frequency = 12)

#Plot the detrended series
autoplot(ts_energy_data, series="Original") +
  autolayer(ts_detrend_renewable, series="Detrended") +
  autolayer(ts_linear_renewable, series="Linear Component") +
  ggtitle("Total Renewable Energy Production") +
  ylab("Energy [Trillion Btu]")
```



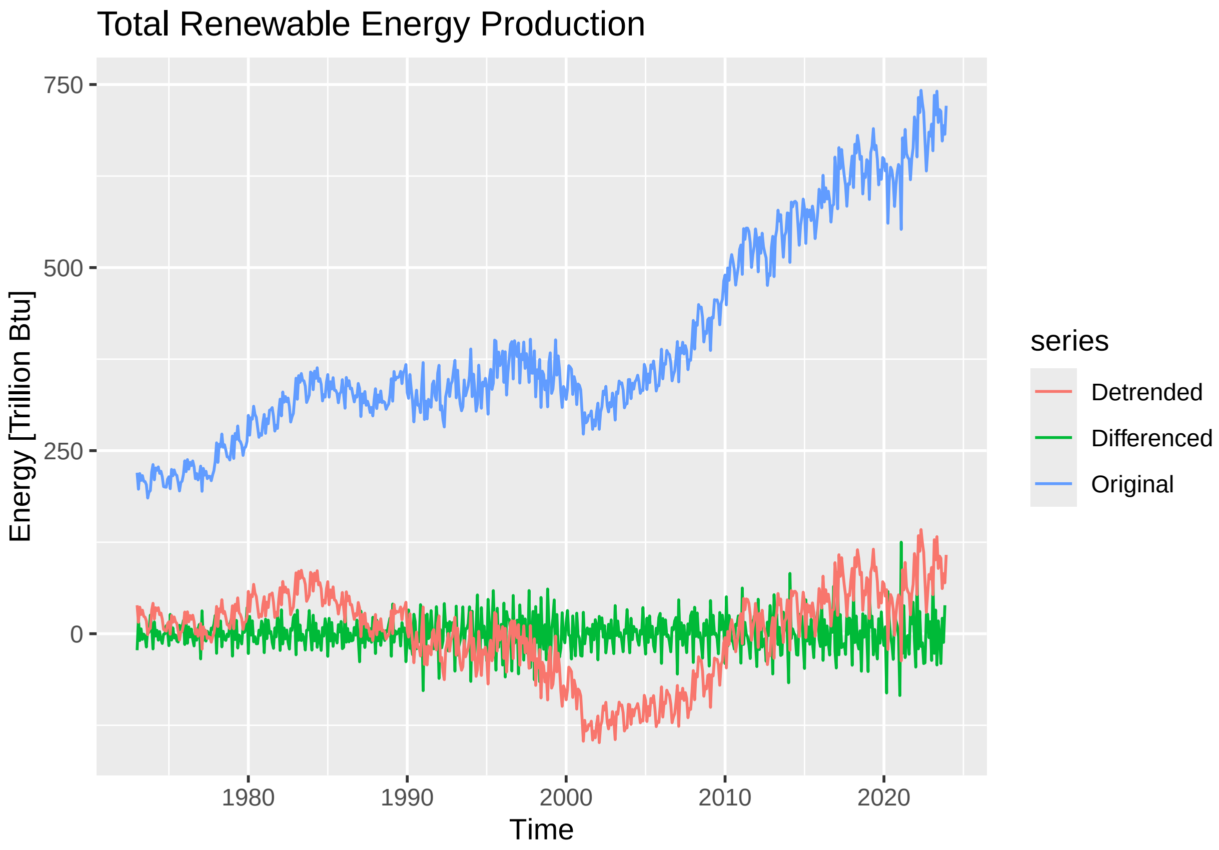
Q3

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 Q2 using linear regression.

Using autoplot() + autolayer() create a plot that shows the three series together. Make sure your plot has a legend. The easiest way to do it is by adding the `series=` argument to each autoplot and autolayer function. Look at the key for A03 for an example on how to use autoplot() and autolayer().

What can you tell from this plot? Which method seems to have been more efficient in removing the trend?

```
# Plot all three series together
autoplot(ts_energy_data, series = "Original") +
  autolayer(ts_diff_renewable, series = "Differenced") +
  autolayer(ts_detrend_renewable, series = "Detrended") +
  ggtitle("Total Renewable Energy Production") +
  ylab("Energy [Trillion Btu]")
```

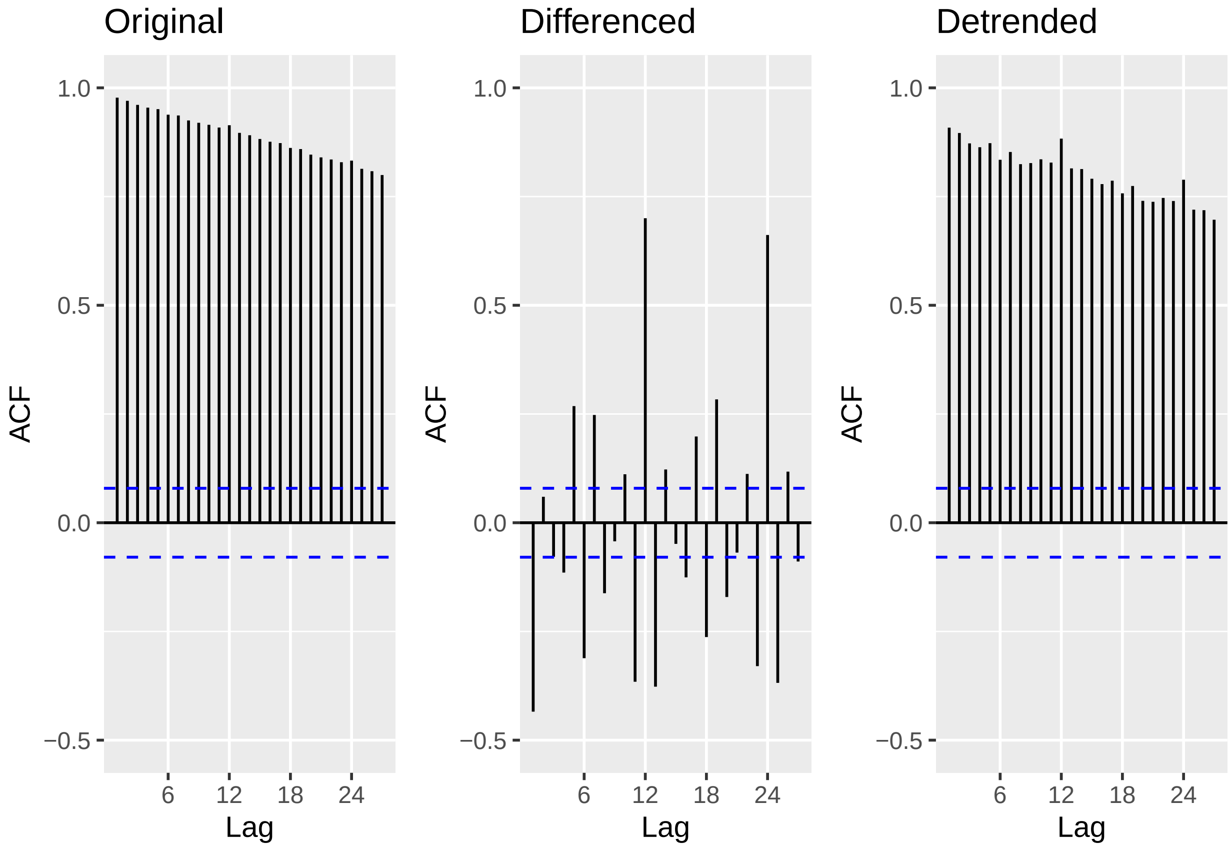


Answer: **The original series** shows a clear upward trend.**The detrended series** removes the linear trend, but certain patterns are still visible, with a general decline from 1985 to 2000, followed by a rise.**The difference series** effectively removes the trend, the values fluctuate around zero, indicating stationarity.

Q4

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

```
plot_grid(
  ggAcf(ts_energy_data) + ggtitle("Original") + ylim(c(-0.5,1)),
  ggAcf(ts_diff_renewable) + ggtitle("Differenced") + ylim(c(-0.5,1)),
  ggAcf(ts_detrend_renewable) + ggtitle("Detrended") + ylim(c(-0.5,1)),
  ncol=3
)
```



Answer: Differencing is more efficient in eliminating the trend, with fewer significant peaks in the ACF of the differenced series, indicating that the trend has been largely eliminated, and the series is more stationary. While the ACF of the detrended series remains very high and retains strong autocorrelation, indicating that the linear regression detrending is not effective in removing the trend.

Q5

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. What is the conclusion from the Seasonal Mann Kendall test? What's the conclusion for the ADF test? Do they match what you observed in Q3 plot? Recall that having a unit root means the series has a stochastic trend. And when a series has stochastic trend we need to use differencing to remove the trend.

```
#Compute the Seasonal Mann-Kendall Test for original series
summary(SeasonalMannKendall(ts_energy_data))
```

```
## Score =  12013 , Var(Score) = 181899
## denominator =  15299.5
## tau = 0.785, 2-sided pvalue =< 2.22e-16
```

```
#Compute the ADF Test for original series
print(adf.test(ts_energy_data,alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_energy_data
## Dickey-Fuller = -1.2251, Lag order = 8, p-value = 0.9024
## alternative hypothesis: stationary
```

> Answer: For the **Seasonal Mann Kendall test**, tau(0.791) is close to 1, indicating a strong positive trend, and p-value (2.22e-16) is much smaller than 0.05, thus, we reject the null hypothesis, meaning this series has a statistically significant upward trend.
> For the **ADF test**, p-value(0.9242) is much greater than 0.05, we fail to reject the null hypothesis, meaning this series is not stationary and has a stochastic trend.
> The results match the Q3 plot. The Seasonal Mann-Kendall test confirms a clear upward trend, which is visible in the original series (blue line). The difference series (green line) in Q3 is much more stationary, confirming that the difference effectively removes the trend, which is consistent with the ADF test showing a stochastic trend.

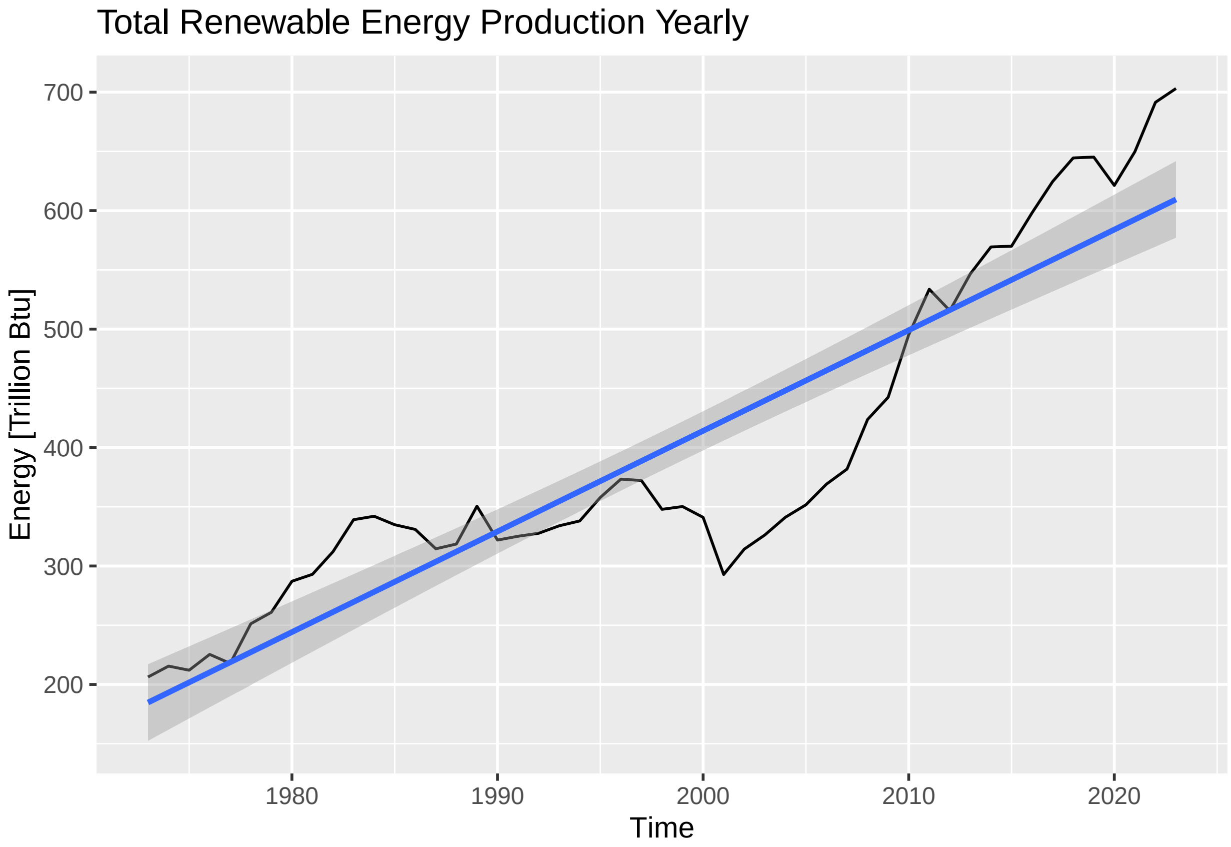
Q6

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. Convert the accumulates yearly series into a time series object and plot the series using autoplot().

```
#Group data in yearly steps instances
energy_data_matrix <- matrix(ts_energy_data,byrow=FALSE,nrow=12)
energy_data_yearly <- colMeans(energy_data_matrix)

my_year <- c(1973:2023)

energy_data_yearly <- data.frame(my_year,"Renewable_Production_Yearly"=energy_data_yearly)

#Convert the accumulates yearly series into a time series object
ts_energy_data_yearly <- ts(energy_data_yearly[,2], frequency=1,start=1973)

#Plot the series
autoplot(ts_energy_data_yearly) +
  ggtitle("Total Renewable Energy Production Yearly") +
```

```r
  ylab("Energy [Trillion Btu]") +
  geom_smooth(method="lm")
```

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



Q7

Apply the Mann Kendall, Spearman correlation rank test and ADF. Are the results from the test in agreement with the test results for the monthly series, i.e., results for Q6?

```r
#Mann-Kendall test
summary(MannKendall(ts_energy_data_yearly))
```

```
## Score =  1033 , Var(Score) = 15158.33
## denominator =  1275
## tau = 0.81, 2-sided pvalue =< 2.22e-16
```

```r
#Spearman Correlation rank test
print(cor.test(ts_energy_data_yearly,my_year,method="spearman"))
```

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

```
##
## data:  ts_energy_data_yearly and my_year
## S = 1852, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9161991
```

```
#ADF test
print(adf.test(ts_energy_data_yearly,alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_energy_data_yearly
## Dickey-Fuller = -1.0646, Lag order = 3, p-value = 0.9196
## alternative hypothesis: stationary
```

> Answer:The test results are consistent with those of the monthly series.
>
> The Mann-Kendall test for the yearly series shows a strong increasing trend as the tau (0.81) is very close to 1 and the p-value (2.22e-16) is much less than 0.05. This is consistent with the results of the Seasonal Mann Kendall test, both of which indicate a statistically significant positive trend over time.
>
> The Spearman correlation rank test on the yearly confirms a strong monotonic increasing trend since the rho (0.916) is very close to 1 and p-value (2.22e-16) is much less than 0.05. This is consistent with the results of the linear regression of the monthly series in Q2 indicating steady growth (slope is positive 0.7) and the results of Seasonal Mann-Kendall confirming a trend of significant growth.
>
> For the ADF test for the annual series, since the p value (0.9196) is much greater than 0.05, we fail to reject the null hypothesis, indicating that the data is non-stationary. This is consistent with the results of the ADF test for the monthly series.